Università degli Studi di Udine
Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica
Dottorato di Ricerca in Ingegneria Industriale e dell'Informazione

XVIII Ciclo

Numerical Methods for Integrated Optics

Dottorando
Lorenzo Bolla

Relatore
Professor Michele Midrio

Anno Accademico
2005/2006

Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica
Università degli Studi di Udine
Via delle Scienze, 206
33100 Udine
Italia

Contents

Preface

Before solving a problem, we must first of all set it properly. We have a physical situation on the one hand, with a description (dimensions, values of physical parameters) and a query about this situation, coming from some interested party. The task of our party (the would-be Mathematical Modeler, Computer Scientist and Expert in the Manipulation of Electromagnetic Software Systems) is to formulate a relevant mathematical problem, liable to approximate solution (usually with a computer), and this solution should be in such final form that the query is answered, possibly with some error or uncertainty, but within a controlled and predictable margin. Mathematical modelling is the process by which such a correspondence between a physical situation and a mathematical problem is established.
Alain Bossavit in [Bos98a, page 32]

The present thesis describes the mathematical models and their computer implementation to study and solve the electromagnetic problems faced in the three years of my Ph.D..

In Part I, a novel algorithm to study the propagation of light will be investigated, both in time- and in frequency-domain. Comparisons with commercial software will be worked out and conclusions will follow.

In Part II, two algorithms, to study straight dielectric waveguides and periodic structures, will be discussed. Validations and comparisons with commercial software and literature will be presented.

Finally, in Part III, these algorithms will be applied to the study of a physical device, its design and its experimental validation.

Invaluable source of ideas and suggestions have been all my collegues from the PICCO Project, who supported my interests on novel time-domain algorithms to study photonic crystals, from Photon Design, whose experience and friendship have made learning from them a joy, and from the FUNFOX Project, who allowed me to experiment my theorical studies. In every moment of these experiences, Professor Michele Midrio has been a constant reference and an enthusiastic supporter.

Lorenzo Bolla
January 31, 2006

Prefazione

Prima di risolver un problema, dobbiamo prima di tutto porlo nel modo corretto. Si parte da una situazione fisica, con una descrizione (attraverso valori e dimensioni di certi parametri fisici) e una domanda su tale situazione da parte di qualcuno. L'obiettivo del nostro gruppo (l'aspirante Matematico Applicato, Programmatore ed Esperto in Software per l'Elettromagnetismo) è di formulare un appropriato problema matematico, dotato di una soluzione approssimata (di solito ricavabile con un computer) tale da rispondere alla domanda posta, anche se con un certo margine di errore o incertezza, ma all'interno di margini controllati e predicibili. La creazione di modelli matematici è il processo attraverso il quale si realizza questa corrispondenza tra situazione fisica e problema matematico.
Alain Bossavit in [Bos98a, page 32] – traduzione dell'Autore

La presente tesi descrive i modelli matematici e la relativa implementazione al calcolatore per lo studio di problemi di elettromagnetismo, affrontati nei tre anni del mio dottorato di ricerca.

Nella Parte I, è descritto un originale algoritmo per lo studio della propagazione della luce, sia nel dominio del tempo che della frequenza, e validato attraverso confronto con software commerciale.

Nella Parte II, sono discussi due algoritmi: uno per lo studio di guide d'onda dielettriche a sezione costante, l'altro per lo studio di strutture dielettriche periodiche. Anche in questo caso, sono presentati validazione ed esempi tratti dalla letteratura e da software commerciale.

Infine, nella Parte III, questi algoritmi saranno applicati allo studio, al design e alla validazione sperimentale di un dispositivo reale.

Un'inesauribile fonte di idee e suggerimenti sono stati tutti i miei colleghi del progetto europeo PICCO, per il supporto e l'incitamento nello studio di nuovi algoritmi dedicati alla propagazione di luce in cristalli fotonici, della Photon Design, la cui esperienza ed amicizia ha reso imparare da loro una gioia, e del progetto europeo FUNFOX, per l'opportunità di applicare praticamente i miei studi teorici. In ogni fase di queste esperienze, il Professor Michele Midrio ha rappresentato una presenza costante e una continua fonte di entusiasmo e incitamento.

Lorenzo Bolla
31 Gennaio 2006

I

Propagators

What are the “Propagators”?

Electromagnetic phenomena are fully described by Maxwell equations. As any differential equations, they can be mathematically solved if initial conditions and boundary conditions are defined. We call “propagators” the techniques to solve Maxwell equations, both in the time- and frequency-domain, once a domain, with some boundaries, and some sources have been defined.

1

Discretization Schemes

1.1 Introduction

Many of the most important and widely used algorithms to study electromagnetic problems (the *Finite Different Time Domain*, the *Finite Element* and the *Finite Volume* methods, to cite only a few) are based on some sort of discretization of the four-dimensional space-time domain in which the physical problem is set [BF04].

Some of them are based on the differential formulation of Maxwell equations described in [Max54], others on their integral formulation [Ton00]. The former is independent on the particular reference system in which the equations are defined: nonetheless, to be able to solve them, one needs to be set. The latter is not connected to a particular discretization scheme, but one needs to be defined to be able to solve them.

We can always pass from an integral (or global) representation to a differential one by a *limiting process*, though. This is how from experiments, which are integral processes, differential equations have been induced. While this is very convenient in theory, because many analytical tools are available for differential equations, it is not so useful in practice, when Maxwell equations must be solved on a computer. Its finite nature suggests to avoid this limiting process [Lt03].

In this chapter, we will focus our attention of the integral formulation of Maxwell equations and on the discretization process necessary to solve them. Two steps can be identified:

1. first of all, the three spatial dimensions and the time dimension are divided into some "elementary" geometrical objects (one-, two-, three- and four-dimensional), which, altogether, compose a *space-time mesh*, also called *cell complex*;

2. then, if the physical quantities used in the differential formulation are functions of points in the four-dimensional domain, i.e. functions of points in space and instants in time, in the integral formulations they are strictly connected with geometrical elements: this association must then be defined.

Many algorithms are available, nowadays. We have studied them, looking for their best characteristic and trying to condense them into a unique algorithm.

One source of errors in the Finite Difference Time Domain [Yee66], FDTD for short, for example, is due to the staircase approximation of metallic boundaries, not parallel to one of the coordinate planes in the orthogonal grid used [CW91]. The main effects are:

- numerical dispersion due to the staircase approximation is more severe than that for the FDTD approximation of Maxwell equations on an infinite grid;
- non-physical surface modes are supported in the numerical solution by the sawtooth conducting boundary;
- waves propagating along the conducting surface are slowed down (i.e. the mode of a metallic waveguide with walls parallel to the coordinate axis runs faster than one with walls tilted);
- high frequencies suffer more that low frequencies.

In [CW91] are reported some solutions to treat these boundaries. One of them is to use an unstructured grid, instead of an orthogonal one: this powerful characteristic is included in the algorithm described in the following sections.

1.2 Mesh

1.2.1 Definition and Properties

The definition of "mesh", also known as "simplicial complex", passes through the definition of "simplices".

Definition 1 (Simplex) *Given* $x_0, x_1, \ldots, x_M$ *affine points in an abstract space, an* M*-simplex* σ^M *is the set of points given by:*

$$x = \sum_{i=0}^{M} \lambda_i x_i,$$

where λ_i *are the* barycentric coordinates *(see Appendix C) such that* $\sum_{i=0}^{M} \lambda_i = 1$ *and* $\lambda_i \geq 0$*. We write* $\sigma^M = [x_0, x_1, \ldots x_M]$ *[Tei01].*

In a three-dimensional space, a 0-simplex is a point (or vertex or node or instant in time), a 1-simplex is a line segment (or edge or interval in time), a 2-simplex is a surface (or face, usually a triangle), and a 3-simplex is a volume (or cell, usually a tetrahedron). An *oriented* M-simplex changes sign under a change of orientation, i.e., if $\sigma^M = [x_0, x_1, \ldots, x_M]$ and a permutation of the indices is carrier out, then $[x_{\tau(0)}, x_{\tau(1)}, \ldots, x_{\tau(M)}] = (-1)^\tau \sigma^M$, where τ denotes the total number of permutations needed to restore the original index order. The j-face of a simplex is the set defined by $\lambda_j = 0$. The faces of a 1-simplex $[x_0, x_1]$ are the points $[x_0]$ and $[x_1]$ (0-simplices), the faces of a 2-simplex $[x_0, x_1, x_2]$ are its three edges $[x_0, x_1]$, $[x_1, x_2]$, $[x_2, x_0]$ (1-simplices), and so forth.

Definition 2 (Simplicial Complex) *A* simplicial complex $\mathcal{K}$ *is a collection of simplices such that:*

- *for all* σ^M *belonging to* $\mathcal{K}$*, its faces also belong to* $\mathcal{K}$*;*
- *for any two simplices their intersection is either empty or it is a common face of both.*

We note that the concept of simplicial complex is *independent of a metric*: this will be a key point in understanding why it is the general structure over which the discretized version of Maxwell's equations will be cast.

For a given simplicial complex, let $\mathcal{N}$ be the set of nodes n, $\mathcal{E}$ the set of edges e, $\mathcal{F}$ the set of faces f and $\mathcal{V}$ the set of volumes v [TK04, vR01].

1.2.2 Orientation

In Definition 1, an oriented M-simplex has been defined. Moreover, we can define an *oriented mesh* as a normal mesh, as defined in Definition 2, but with all the M-simplices oriented themselves.

Orientation if very important. While in the differential formulation of Maxwell equations the physical quantities can be vectors (like $\vec{E}$, $\vec{B}$, etc.) defined for each point of the domain, in the integral formulation equations involve always scalar quantities. This does not mean that we are losing information associated with the vectorial nature of quantities in the differential equations: we have simply switched this information from the physical quantities themselves to the reference system we are using to represent them. For example, the vectorial nature of the electric field $\vec{E}$ is translated into the *orientation dependence* of its circuitation along a line: if the line has an orientation, inverting it means changing the sign of the circuitation of $\vec{E}$ along it.

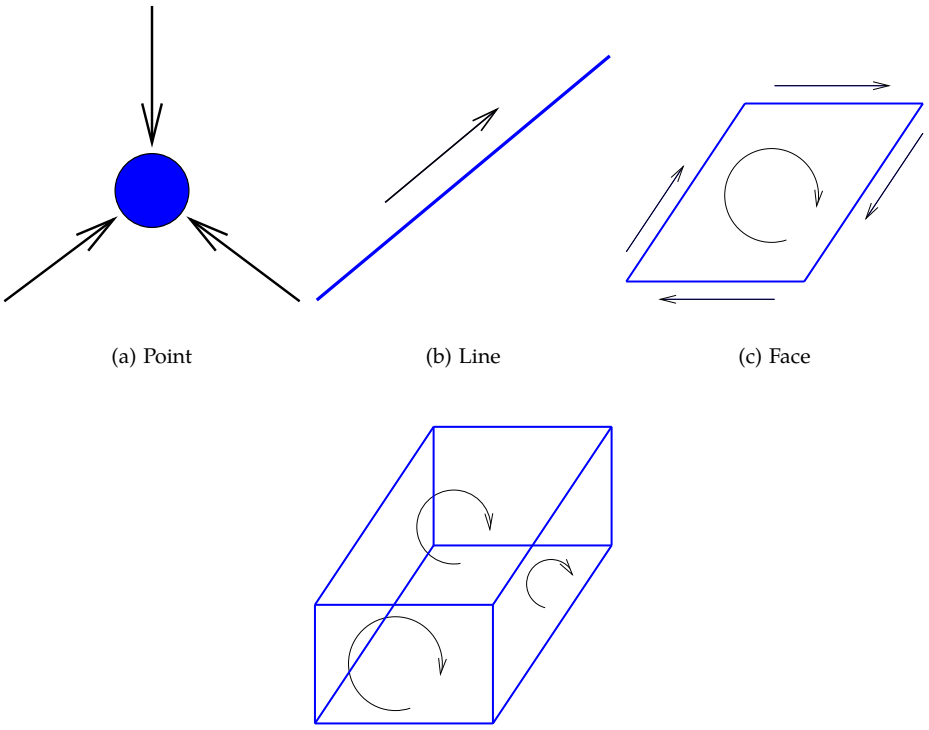
The are two ways to orient geometrical elements [Ton00].

Internal orientation We can think of it as an indication of the oriented direction *along* the geometrical element: it can be defined without the need to "leave" the geometrical element itself (see Figure 1.1). A line is internally oriented by defining a verse along it, i.e. by choosing which are the first and the second nodes of the edge; a surface, by internally orienting its boundary in a *coherent* way, so that each node on its boundary is the first node of an edge and the second of the adjacent edge at the same time; a volume, by internally orienting all its boundary faces in a *coherent* way, matching the orientation of edges between two adjacent faces. Points can be internally oriented too, even if they have null measure: the definition can be made indirectly. From points, we can trace lines going outward or inward, so defining its orientation: conventionally, a point from which all the lines go outward is defined a "source", or *negatively oriented*; a points from which all the lines to inward is defined a "sink", or *positively oriented*. We call *incidence number* between an oriented line and an oriented point the number $+1$ if the orientation of the line agrees with

the orientation of the point, -1 otherwise. A typical example is the definition of one dimensional increment Δ of a function f in the interval $[x, x + h]$:

$$\Delta f(x) = (-1)f(x) + (+1)f(x + h).$$

As long as the interval can be internally oriented from its first point to the last (it is indeed a one-dimensional vector), automatically an internal orientation for its first and second points is defined: note that the $+1$ is associated with the point $x + h$, which is the sink. Analogously, instants and intervals in time can be internally oriented.



(a) Point (b) Line (c) Face

(d) Volume

Figure 1.1: *Internal orientation.*

External orientation We can think of it as the oriented direction *through* the geometrical element: it can be defined only if we suppose to watch the p-dimensional geometrical element from a $(p + 1)$-dimensional point of view

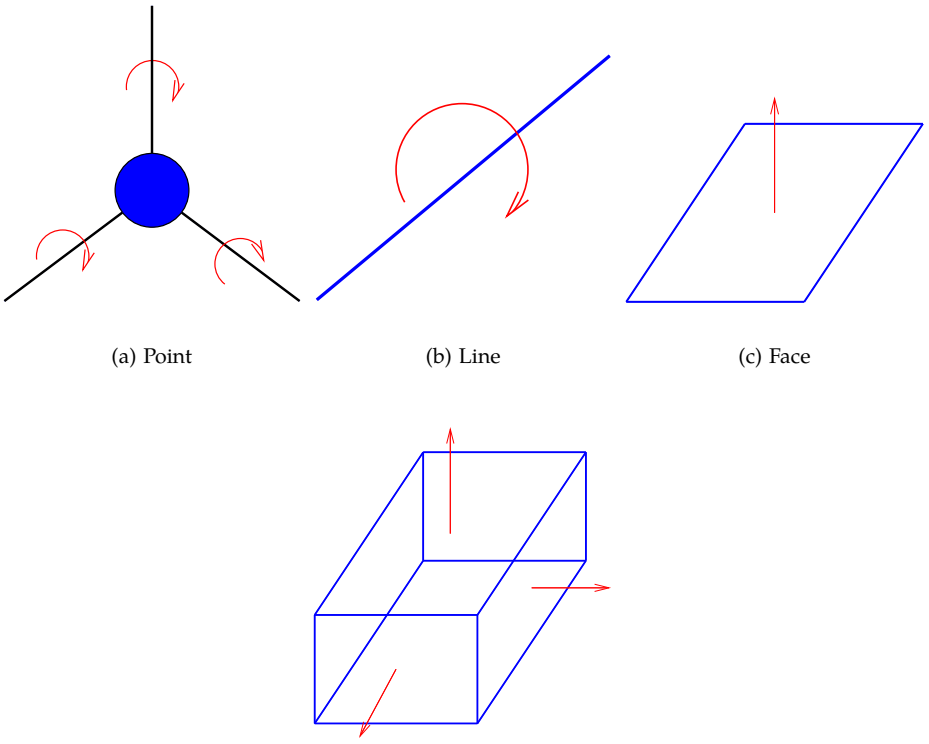
(see Figure 1.2). For example, a surface (2-dimensional element) can be externally oriented by distinguishing its two faces (and this requires to watch it from a 3-dimensional point of view) and defining a verse through it: in other words, an oriented line not lying on the surface can externally orient it by distinguishing which of the two faces of the surface itself is crossed first. The same can be done with a line, which can be externally oriented by defining an internally oriented surface not containing the line itself, or a volume, by defining an internally oriented point inside it. Even a point can be externally oriented, by defining an internally oriented volume containing it. Again, the same is applicable to instants and intervals in time.

Note that internally oriented p-dimensional elements are used to externally orient $(3-p)$-dimensional elements: this will be used in the orientation of the dual mesh (Section 1.2.3).

A word about notation: externally oriented geometrical elements can be distinguished by the internally oriented for the tilde on top. So: $\widetilde{p}$, $\widetilde{e}$, $\widetilde{f}$ and $\widetilde{v}$ are externally oriented point, edge, face and volume, respectively.

An example to distinguish between internal and external orientation can be made thinking about inversion of time. There are physical quantities which are left unchanged by an inversion of time (like the total electric charge inside a volume) and others that change their sign (like the electric charge that pass through a given surface). The first ones are associated with internally oriented instants or externally oriented intervals and the seconds with externally oriented instants or internally oriented intervals [Ton00].

Given the domain to discretize, the choice of the simplicial cell complex is not unique. One very common way of defining a cell complex is by triangles in two dimensions or tetrahedral in three dimensions. These are the easiest geometrical shapes that satisfy the properties in Definition 2 and they are therefore called *simplicial elements*: each more complicated shape (squares, polygons, parallelepipeds, ...) can be simplified into simplicial elements. Moreover, very often, simplicial elements which satisfy another property are used to build simplicial complexes, which are called *Delaunay simplicial complexes*. Their peculiar property is that each circumcenter (or circumsphere in three dimensions) associated to a cell does not contain any other vertices apart from the ones that define the cell itself. This property has a very important consequence on the stability of the algorithms associated with this mesh. Consider, for example, Figure 1.3: a thermal field is defined on the cells t_1 and t_2, and $T_1 > T_2$ are the temperatures measured at points C_1 and C_2, respectively. Heat flows naturally from the warmer zone to the colder, through the common face between cells t_1 and t_2. But if the verse of the segment C_1C_2 is opposed to the verse of the segment going from t_1 to t_2, the heat flux will be negative (i.e., in the opposite direction, from a colder zone to a warmer zone), which is non-physical. This process is going to increase without limit: it's an instability [Ton00]. This is due to the fact that the mesh in Figure 1.3(b) is not a Delaunay mesh.

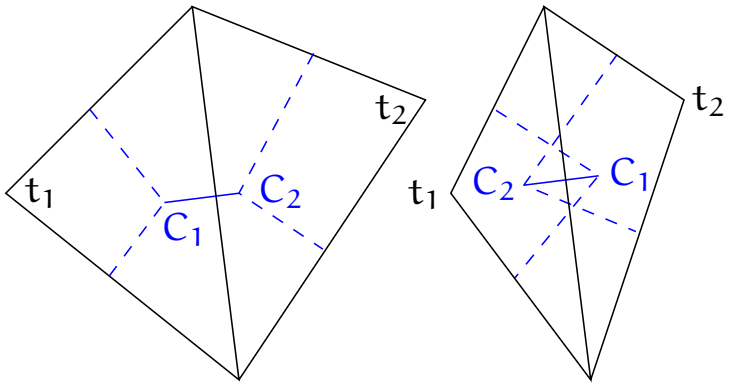


(a) Point (b) Line (c) Face

(d) Volume

Figure 1.2: *External orientation.*

Something very similar can happen in electromagnetic simulations with fluxes and circuitations of vectorial fields.



(a) Delaunay primal mesh.

(b) Non-Delaunay primal mesh.

Figure 1.3: *The simplicial complex choice affects the stability of the algorithm. If a thermal field is defined on the triangles t_1 and t_2 and the temperatures measured at points C_1 and C_2 are $T_1 > T_2$, a thermal flux from the colder cell to the warmer will be simulated in the right-hand side mesh: this is non-physical.*

Another very common three-dimensional mesh is the two-dimensional extruded mesh: see Figure 1.4 for an example. It is made of a two-dimensional mesh, extruded in the third dimension, like a stack of two-dimensional meshes. The advantage is that simple two-dimensional meshing software [web] can be used to model each "floor" of the stack and then it can be very easily extruded in the third dimension like a structured mesh. Two-dimensional extruded grids are best suited to model planar geometries: they will be used to study planar photonic crystal channels (see Section 4.5.2).

1.2.3 Dual Mesh

Once we have defined a cell complex $\mathcal{K}$, orienting its geometrical elements suggests the definition of another cell complex:

$$\widetilde{\mathcal{K}} = \left\{ \widetilde{\mathcal{N}}, \widetilde{\mathcal{E}}, \widetilde{\mathcal{F}}, \widetilde{\mathcal{V}} \right\},$$

made of points $\widetilde{n} \in \widetilde{\mathcal{N}}$, lines $\widetilde{e} \in \widetilde{\mathcal{E}}$, surfaces $\widetilde{f} \in \widetilde{\mathcal{F}}$ and volumes $\widetilde{v} \in \widetilde{\mathcal{V}}$ used to externally orient its volumes, surfaces, lines and points, respectively. This *induced* cell complex is called *dual* cell complex, and the original one is called *primal* cell complex [Ton00]. Some observations can be made:

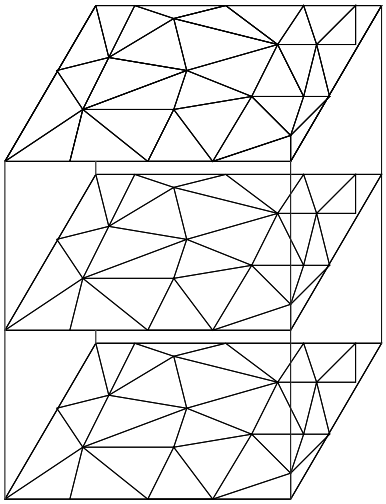


Figure 1.4: *Two-dimensional extruded mesh. Two-dimensional meshes are the "floors" of a three-dimensional stack. Distance between two-dimensional meshes has been exaggerated, for the sake of clarity.*

- for each cell complex, a dual cell complex can be defined;
- there is a surprising "symmetrical" property, for which a vertex in a primal cell complex corresponds to a cell in the dual complex, a line to a surface, a surface to a line and a volume to a point, an instant to an interval and an interval to an instant: an upside-down classification of geometrical elements, as shown by the Diagram 1.1;

$$\begin{array}{ccc} n & \longrightarrow & \widetilde{v} \\ \downarrow & & \downarrow \\ e & \longrightarrow & \widetilde{f} \\ \downarrow & & \downarrow \\ f & \longrightarrow & \widetilde{e} \\ \downarrow & & \downarrow \\ v & \longrightarrow & \widetilde{n} \end{array} \tag{1.1}$$

- the choice of the dual mesh is somehow arbitrary: there is no special need to choice one particular orientation over another possible orientation (even if, conventionally, we use the *pit*'s convention for points and volumes and the *right-hand* convention for lines and surfaces) or to choose a particular

point or line to externally orient a volume or a surface of the primal cell complex. As we'll see later, though, some choices are better than others, therefore some dual meshes are better than others, for stability and ease of equations formulation.

As said before, the choice of the primal cell complex is not unique: neither is the choice of its dual. For primal cell complexes, Delaunay tessellations satisfy some very important properties that have positive drawbacks on the algorithms' stability [CHP94]. Their dual complexes are called *Voronoï* complexes and they are defined by taking the circumcenters (or circumspheres in three-dimensions) of the primal cells as nodes. Each dual edge will be orthogonal to a primal face and each dual face will be orthogonal to a primal edge, for construction. At the limit of very small cells, the couple Delaunay-Voronoï complexes are locally an orthogonal reference system. As we'll in Section 2.2, this greatly simplifies the numerical discretization of Maxwell equations.

Another possible choice is to consider the barycenters of primal cells as the nodes of the dual meshes: it they are connected by straight lines, the dual edges, the resulting dual mesh is called *Poincaré* dual mesh. If, on the other hand, dual edges are piece-lines going from the barycenter of a primal cell to the barycenter of an adjacent primal cell passing by the barycenter of the common primal face, the resulting dual mesh is called *barycentric*.

1.2.4 Matricial Representation

A mesh can be seen as graph, possibly oriented, with edges representing "relations" between the nodes. Therefore, many techniques applicable to graphs can be applied to meshes, as well.

In particular, to represent an oriented mesh *incidence matrices* can be used. Incidence matrices are matrices (usually sparse) made of 0s, 1s and -1s to describe the interconnections between geometrical elements. If two elements i, j are not connected the matrix element (i, j) is 0, otherwise it's ± 1: the sign is positive if the orientation of i agrees with the orientation of j, otherwise is negative.

Mutual interconnections of the simplices in the primal mesh $\mathcal{K}$ can be described by incidence matrices [TK04]:

- $\mathbf{G}$ is the incidence matrix between edges e and nodes n;
- $\mathbf{R}$ is the incidence matrix between faces f and edges e;
- $\mathbf{D}$ is the incidence matrix between volumes v and faces f.

With respect to figure 1.5, the k-row of the matrix $\mathbf{G}$, incidence matrix between

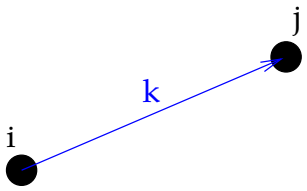


Figure 1.5: Adjacency edges-nodes. Orientation of edges is shown by the arrow sign, while orientation of nodes is conventionally chosen to be positive for sinks and negative for sources.

edges and nodes, is:

$$\mathbf{G} = \begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -1 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \leftarrow \text{k row}$$

$$\uparrow \qquad \uparrow$$
$$\text{i col} \qquad \text{j col}$$

where the $+1$ says that the edge k is entering the node j, i.e. agrees with the "sink" convention for oriented nodes.

Something very similar can be done to build the other matrices.

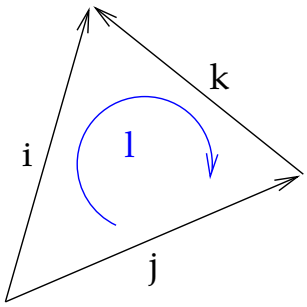


Figure 1.6: Adjacency faces-edges. Orientation of edges is shown by the arrow sign, while faces are internally oriented by the curved arrow. In the picture, the orientation of the edge i agrees with the orientation of the face l, while edges j and k have opposite orientations.

With respect to figure 1.6, the l-row of the matrix $\mathbf{R}$, incidence matrix between faces and edges, is:

$$\mathbf{R} = \begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & \dots & -1 & \dots & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \leftarrow \text{l row}$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$
$$\text{i col} \qquad \text{j col} \qquad \text{k col}$$

where the $+1$ says that the edge $\mathtt{i}$ agrees with the orientation of the face $\mathtt{l}$.

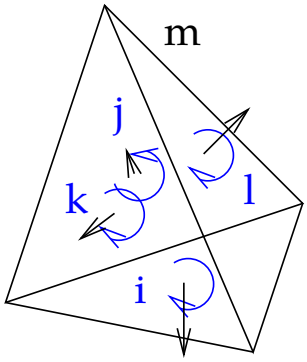


Figure 1.7: *Adjacency volumes-faces. Internal orientation of faces is shown the curved arrow. The volume is conventionally oriented from inside to outside. In the picture, the orientation of the face* $\mathtt{i}$ *agrees with the orientation of the volume* $\mathtt{m}$*, while the other faces disagree.*

Finally, from figure 1.7, the $\mathtt{m}$-row of matrix $\mathbf{D}$, incidence matrix between volumes and faces, is:

$$\mathbf{D} = \begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & \dots & -1 & \dots & -1 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \leftarrow \mathtt{m} \text{ row}$$

$$\begin{matrix} & & \uparrow & & \uparrow & & \uparrow & \uparrow & & \\ & & \mathtt{i} \text{ col} & & \mathtt{j} \text{ col} & & \mathtt{k} \text{ col} & \mathtt{l} \text{ col} & & \end{matrix}$$

where the $+1$ says that the face $\mathtt{i}$ agrees with the orientation of the volume $\mathtt{m}$.

Note also that for a non-pathological meshes:

$$\mathtt{n}_n = \mathtt{n}_{\widetilde{v}} \qquad \mathtt{n}_e = \mathtt{n}_{\widetilde{f}} \qquad \mathtt{n}_f = \mathtt{n}_{\widetilde{e}} \qquad \mathtt{n}_v = \mathtt{n}_{\widetilde{n}}$$
$$\mathtt{n}_{\widetilde{n}} = \mathtt{n}_v \qquad \mathtt{n}_{\widetilde{e}} = \mathtt{n}_f \qquad \mathtt{n}_{\widetilde{f}} = \mathtt{n}_e \qquad \mathtt{n}_{\widetilde{v}} = \mathtt{n}_n$$

where $\mathtt{n}_\mathtt{x}$ denotes the number of geometrical elements $\mathtt{x}$.

Therefore, we have:

$$\mathbf{D} \in \mathtt{n}_v \times \mathtt{n}_f$$
$$\mathbf{R} \in \mathtt{n}_f \times \mathtt{n}_e$$
$$\mathbf{G} \in \mathtt{n}_e \times \mathtt{n}_n$$

and

$$\mathbf{D}^\mathsf{T} \in \mathtt{n}_{\widetilde{e}} \times \mathtt{n}_{\widetilde{n}} \equiv \widetilde{\mathbf{G}}$$
$$\mathbf{R}^\mathsf{T} \in \mathtt{n}_{\widetilde{f}} \times \mathtt{n}_{\widetilde{e}} \equiv \widetilde{\mathbf{R}}$$
$$-\mathbf{G}^\mathsf{T} \in \mathtt{n}_{\widetilde{v}} \times \mathtt{n}_{\widetilde{f}} \equiv \widetilde{\mathbf{D}}.$$

The matrices $\mathbf{D}^T$, $\mathbf{R}^T$ and $-\mathbf{G}^T$[1] describe the interconnections of $\widetilde{\mathcal{K}}$.

1.3 Geometry and Physics

There are different discretization schemes for Maxwell equations. Considering the discretization in time, we can distinguish two big families of discretization schemes [BF04]:

collocated schemes , in which physical quantities are all associated to the same time. In other words, the discretization in time does not depend on the specific physical quantity we are modeling;

uncollocated schemes , in which different physical quantities are associated with different points in time, as if there were different discretization schemes in time for each physical quantity.

The same concept may be applied to the discretization in space, with the added complexity that the domain to discretize is now three-dimensional, instead of one-dimensional. The most clearly visible difference is that the geometrical objects we can use to discretize the space and to which associate the physical quantities can be points, lines, surfaces or volumes: much more freedom than in time, where we just have points (instants) and lines (intervals)!

Again, we can distinguish [BF04]:

unstaggered schemes : they are analogous to collocated schemes in time, in the sense that physical quantities are all associated to the geometrical elements of a unique mesh;

staggered schemes : in which each physical quantity has its own discretization in space. These are the analogous of uncollocated schemes in time.

Until now, we have spoken of "associating physical quantities to geometrical elements": with this expression, we intend to create a map between physical quantities, such as fields, potentials or charges, and geometrical elements, such as points, lines, faces and volumes. The result of the map will be a scalar, used in the integral form of Maxwell equations.

Let's start from the integral form of the Maxwell equations:

$$\begin{cases} \int_{\partial \mathcal{S}} \overrightarrow{E} \cdot d\overrightarrow{l} = -\partial_t \iint_{\mathcal{S}} \overrightarrow{B} \cdot \hat{n} dS \\ \int_{\partial \widetilde{\mathcal{S}}} \overrightarrow{H} \cdot d\overrightarrow{\widetilde{l}} = \partial_t \iint_{\widetilde{\mathcal{S}}} \overrightarrow{D} \cdot \widehat{\widetilde{n}} d\widetilde{S} \end{cases} , \tag{1.2}$$

where, $\partial \mathcal{S}$ denotes the boundary of a surface $\mathcal{S}$, $d\overrightarrow{l}$ a vector tangent to it and $\hat{n}$ a versor orthogonal to $\mathcal{S}$: similar meanings hold for the tilded variables.

[1] The minus sign comes from the assumption that n is oriented as a sink, whereas the boundary of $\widetilde{v}$ is oriented by the outer normal.

We can distinguish two kinds of physical quantities in (1.2). The first equation in (1.2), the *Faraday equation*, reads: "given an oriented surface $\mathcal{S}$, the circuitation of the electric field along its boundary equals the opposite of the derivative in time of the flux of the magnetic vector through it". The second of (1.2), the *Ampère law*, reads: "given an oriented surface $\mathcal{S}$, the circuitation of the magnetic induction equals the derivative in time of the electric induction plus the flux of the current density through it". Many observations can be made:

- Maxwell equations, in the integral form, only relate circuitations to fluxes;
- each equation relate a vector ($\overrightarrow{E}$ or $\overrightarrow{D}$) with a covector ($\overrightarrow{H}$ or $\overrightarrow{B}$): the curl operator transforms a vector to a covector [Bos98a];
- the two equations are only topological relations, in the sense that they depend only on the topological properties of surfaces and boundaries and not on the physical properties of the underlying medium; materials link vectors to covectors by material equations: $\overrightarrow{D} = \epsilon \overrightarrow{E}$ and $\overrightarrow{B} = \mu \overrightarrow{H}$;
- we need an oriented space to make the equation work: the orientation is arbitrary, but it needs to be coherent;
- in the first equation, if the electric field is evaluated at some time t, so must be the partial derivative in time of $\overrightarrow{B}$: the same holds for the second equation with $\overrightarrow{H}$ and $\overrightarrow{D}$. This will drive the choice of a uncollocated discretization scheme in time.

The most spontaneous way for the association of physical quantities to geometrical elements is "circuitations with lines" and "fluxes with surfaces": there is no circuitation without a line and no flux without a surface. We can then define the maps:

$$\begin{aligned}
\text{primary edge } e &\xmapsto{e} \int_e \overrightarrow{E} \cdot d\overrightarrow{l} \\
\text{primary surface } f &\xmapsto{b} \iint_f \overrightarrow{B} \cdot \hat{n} dS \\
\text{primary surface } f &\xmapsto{m} \iint_f \overrightarrow{M} \cdot \hat{n} dS \\
\text{dual line } \widetilde{e} &\xmapsto{h} \int_{\widetilde{e}} \overrightarrow{H} \cdot d\overrightarrow{\widetilde{l}} \\
\text{dual surface } \widetilde{f} &\xmapsto{d} \iint_{\widetilde{f}} \overrightarrow{D} \cdot \hat{n} d\tilde{S} \\
\text{dual surface } \widetilde{f} &\xmapsto{j} \iint_{\widetilde{f}} \overrightarrow{J} \cdot \hat{n} d\tilde{S}
\end{aligned} \tag{1.3}$$

Maxwell equations are topological relations: as long as the mesh satisfies the properties in Definition 2, with the association in (1.3), they can be written to be

strictly exact. So, where are the necessary approximations, always connected to a discretization process? Obviously, they are outside Maxwell equations, in particular in the material equations (see Section 2)[2].

One might argue, why should we care about dividing metric independent, i.e. Maxwell, and metric dependent, i.e. material, equations? There are at least three good answers [Tei01]:

1. many theorems (such as charge conservation, Faraday equation and Ampère law) are automatically fulfilled after discretization, without the need to involve metric concepts, because they only depend on topology;

2. the metric is completely encoded into the material equations, the treatment of curved boundaries and material interfaces can be done in a more systematic manner, without affecting, for example, conservation laws related to the topological equations;

3. the topological (spatial) part of the equations often comprises integer arithmetic only and are more efficiently handled by a computer if, a priori, recognized as such.

Diagram 1.4 shows graphically the relations between the four vectorial fields used in (1.2).

$$\begin{array}{ccc} \overrightarrow{E} & \xrightarrow{\text{Faraday}} & \overrightarrow{B} \\ {\scriptstyle\epsilon}\uparrow & & \downarrow{\scriptstyle\mu} \\ \overrightarrow{D} & \xleftarrow[\text{Ampère}]{} & \overrightarrow{H} \end{array} \tag{1.4}$$

Discretization in time, discretization in space and association of physical quantities with geometrical elements are "decoupled" choices: one particular choice in one, does not limit the choices in the others. This is the reason why there are so many algorithms, more or less similar to each others, based on the same concepts of discretization of Maxwell equations, each sharing the same strengths and weaknesses. An example of an algorithm based on an uncollocated staggered scheme is the *Finite Difference Time Domain* method [Taf00], in the *Yee implementation*[Yee66]. The same algorithm, but implemented using the *forward/backward differences* both in time and in space, becomes a collocated unstaggered scheme [LBF+04]. In the frequency domain, where time doesn't need to be discretized because not present in Maxwell equations, we can list the *Finite Element* method, with the node element base [Jin02], as a staggered scheme, or the *Finite Element* method with the Whitney elements discretization scheme [BK00], as an unstaggered scheme.

[2]It's worth noting that this is not the only possibility: some discretization schemes introduce errors in Maxwell equations, but model exactly material equations. We think that these schemes should not be used because they fail to satisfy identically all the properties encoded in Maxwell equation, like flux conservation, for example. We think that modeling the wrong material is better than modeling the wrong physical equations.

Not all the discretization schemes are equally suited to model electromagnetic problems, though. The geometrical properties of the physical problem suggest somehow the best discretization scheme to use, which is also the most elegant [Max71].

Maxwell equations are two coupled first order partial differential equations: partial derivatives in space of the electric field are related to partial derivative in time of the magnetic field and viceversa. This "chiasmo" in space and time strongly suggests to use an uncollocated staggered scheme, which geometrically suggests the spatial and temporal interconnection of fields. Indeed, this is the best choice, leading to simple equation and, more importantly, to the satisfaction of the equation $\nabla \cdot \vec{B} = 0$ everywhere (and every when) on the grid. The divergenceless of the magnetic field is a fundamental physical property: any mathematical model deviating from this fact is incorrect[3] and can lead to instabilities that are non-physical. *Finite Volume* methods, for example, undergo non-physical attenuation of waves, for the choice of a collocated unstaggered scheme [Taf98]: it is as if the discretized medium itself, where the electromagnetic waves are propagated, were lossy. Losses come from the numerical world, not from the "real" one, though. Collocated scheme are also badly suited to study coupled equations: they naturally lead to uncoupled equations, and the coupling, present in Maxwell's equations, is obtained in these methods only at the expense of elegance in the formalism.

1.3.1 Matricial Representation

In Section 1.2.4, the incidence matrices of the primal and dual simplicial complexes $\mathcal{K}$ and $\widetilde{\mathcal{K}}$ have been defined, from a purely topological point of view. With the map defined in (1.3), applied to Maxwell equations, they can be seen under another light [TK04].

$\mathbf{R}$ and $\widetilde{\mathbf{R}}$ are the *discrete curl operators* on the primal and dual grid, respectively, $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ are the *discrete divergence operators* and $\mathbf{G}$ and $\widetilde{\mathbf{G}}$ are the *discrete gradient operators*[4].

We can grasp a geometrical explanation of these discrete operators thinking that to compute the divergence of a vectorial field we need to define a closed surface, on which the field is defined, and make it smaller and smaller (zero, at limit): the flux through this infinitesimal surface is the divergence at the point that it contains. This is the Gauss theorem. In other words, the divergence is an operator that takes a field defined on a surface and gives a field defined in a volume (infinitesimally small – a point): in the discrete world, it is an operator between a flux through a surface and a volume integral. In matrix form, it must be a $n_f \times n_v$ matrix, like the matrix $\mathbf{D}$. A similar reasoning holds for the gradient, i.e. the matrix $\mathbf{G}$, and

[3] As far as modern physics knows [Mea00].

[4] This symmetry is better explained with a drawing in Appendix B.

the curl, i.e. the matrix $\mathbf{R}$. The Diagram 1.5 shows these relations.

$$\begin{array}{cccc} & \text{primal} & & \text{dual} \\ \text{div} & \mathbf{D} & \longrightarrow & \widetilde{\mathbf{D}} = -\mathbf{G}^{\mathsf{T}} \\ & \downarrow & & \downarrow \\ \text{curl} & \mathbf{R} & \longrightarrow & \widetilde{\mathbf{R}} = \mathbf{R}^{\mathsf{T}} \\ & \downarrow & & \downarrow \\ \text{grad} & \mathbf{G} & \longrightarrow & \widetilde{\mathbf{G}} = \mathbf{D}^{\mathsf{T}} \end{array} \tag{1.5}$$

The same properties that hold for differential operators also hold for their discrete counterparts [SW00]:

$$\nabla \cdot \nabla \times \bullet = 0 \qquad \begin{cases} \mathbf{D}\,\mathbf{R} = 0 & \text{on the primal grid} \\ \widetilde{\mathbf{D}}\,\widetilde{\mathbf{R}} = 0 & \text{on the dual grid} \end{cases}.$$

Due to the additional constraints on the orientation and the numbering on the cell edges and the corresponding dual cell facets, we have the *duality relation* of the two curl operators 1.8:

$$\mathbf{R} = \widetilde{\mathbf{R}}^{\mathsf{T}}$$

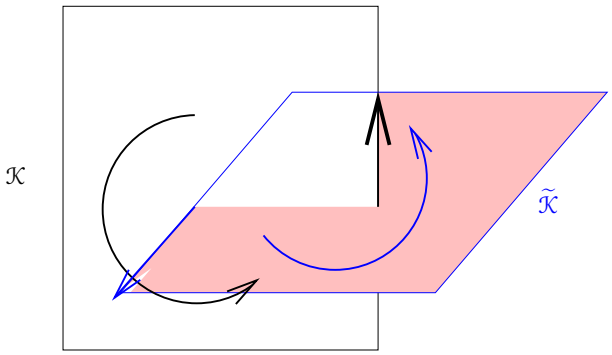


Figure 1.8: *Duality relation of the curl operators: the matrix of the discrete curl operator on the primal mesh is the transpose of the matrix of the discrete curl operator of the dual mesh.*

Combining these two properties, we have the counterpart of the well known identity:

$$\nabla \times \nabla \bullet = 0 \qquad \begin{cases} \mathbf{R}\,\mathbf{G} = 0 & \text{on the primal grid} \\ \widetilde{\mathbf{R}}\,\widetilde{\mathbf{G}} = 0 & \text{on the dual grid} \end{cases}$$

Also the degrees of freedom defined in (1.3) can be recast in matrix form. Let:

$$\begin{aligned} \mathbf{e} &\in \mathfrak{n}_e \times 1 & \mathbf{h} &\in \mathfrak{n}_{\tilde{e}} \times 1 \\ \mathbf{b} &\in \mathfrak{n}_f \times 1 & \mathbf{d} &\in \mathfrak{n}_{\tilde{f}} \times 1 \\ \mathbf{m} &\in \mathfrak{n}_f \times 1 & \mathbf{j} &\in \mathfrak{n}_{\tilde{f}} \times 1 \\ \rho_m &\in \nu \times 1 & \rho_e &\in \mathfrak{n}_{\tilde{v}} \times 1 \end{aligned}$$

where $\mathfrak{n}_n$, $\mathfrak{n}_e$, $\mathfrak{n}_f$ and $\mathfrak{n}_v$ denote the number of nodes, edges, faces and volumes in $\mathcal{K}$ and $\mathfrak{n}_{\tilde{n}}$, $\mathfrak{n}_{\tilde{e}}$, $\mathfrak{n}_{\tilde{f}}$ and $\mathfrak{n}_{\tilde{v}}$ in the $\widetilde{\mathcal{K}}$. ρ_e and ρ_m are the electric and magnetic charge density: they will be used in (1.12) and (1.13). The arrays just defined contain the degrees of freedom associated to the elements of $\mathcal{K}$ and $\widetilde{\mathcal{K}}$, according to the map in (1.3) [TK04]. In particular:

- **e** is associated to primal edges, while **h** to dual edges; they contain the circuitations present in Maxwell equations;
- **b** and **m** are associated to primal faces, while **d** and **j** to dual faces; they contain the fluxes present in Maxwell equations;
- ρ_m is associated to primal cells, while ρ_e to dual cells.

Note that scalar electric and magnetic potential respectively can be associated at primal and dual nodes: they are not present in Maxwell equations, though not included in (1.3).

Finally, using the discrete versions of differential operators and the arrays of degrees of freedom just defined, we can write Maxwell equations in discrete form:

$$\mathbf{R}^\mathsf{T}\,\mathbf{h} = \partial_t \mathbf{d} + \mathbf{j} \qquad \text{Ampere's Law} \tag{1.6}$$

$$\mathbf{R}\,\mathbf{e} = -\partial_t \mathbf{b} + \mathbf{m} \qquad \text{Faraday's Law} \tag{1.7}$$

with the discrete Gauss laws:

$$\mathbf{D}\,\mathbf{b} = -\partial_t \rho_m \qquad \text{Magnetic Gauss' Law} \tag{1.8}$$

$$\mathbf{G}^\mathsf{T}\,\mathbf{d} = -\partial_t \rho_e \qquad \text{Electric Gauss' Law} \tag{1.9}$$

the discrete constitutive equations (see Chapter 2):

$$\mathbf{h} = \mathbf{M}_\mu\,\mathbf{b} \qquad \text{Magnetic Constitutive Equation} \tag{1.10}$$

$$\mathbf{e} = \mathbf{M}_\epsilon\,\mathbf{d} \qquad \text{Electric Constitutive Equation} \tag{1.11}$$

and the discrete conservation equations:

$$\mathbf{G}^\mathsf{T}\,\mathbf{j} = -\partial_t \rho_e \qquad \text{Electric Charge Conservation} \tag{1.12}$$

$$\mathbf{D}\,\mathbf{m} = -\partial_t \rho_m \qquad \text{Magnetic Charge Conservation} \tag{1.13}$$

1.4 Stability of Space Discretization

The Fourier method is used to analyze the dispersive, dissipative and isotropy errors of various spatial and time discretizations applied to Maxwell equations on multi-dimensional grids [Liu96]. Dissipation causes the attenuation of wave amplitude and dispersion causes incorrect wave propagating speed: these errors may depend on the direction of wave propagation with respect to the grid. The most troublesome aspect is that these errors are cumulative in nature: after propagating for a long distance or time, the solution can be greatly affected and sometimes becomes non-physical.

The Fourier method has been widely used in the past centuries to solve partial differential equations: here, it is applied to analyze the solutions found by other means, i.e. by the discretization method.

As long as we want to study the influence of the grid on the solutions, we are not interested on the particular dielectric objects inside the domain. To keep things simple, suppose to study the propagation of light in homogeneous space. It is well known that the exact solution of Maxwell equations in this hypothesis consists of the superposition of linear, non-dispersive and non-dissipative harmonics waves:

$$\overrightarrow{F} = \begin{Bmatrix} \overrightarrow{D} \\ \overrightarrow{B} \end{Bmatrix} e^{\imath(\overrightarrow{k}\cdot\overrightarrow{r}-\omega t)}. \tag{1.14}$$

In a matrix representation, like the one in (1.6) and (1.7), the numerical solution depends on the properties of both the eigenvalues and the eigenvectors of the operator (circulant) matrix. Each eigenvector corresponds to a *discrete* harmonic. However, for the same reason why the maximum and minimum values of a discretized sinusoidal function do not coincide with the same continuous function, discrete eigenvectors usually differ from real ones.

Let's write the generic eigenvector as:

$$\overrightarrow{F} = \overrightarrow{\mathcal{F}}(t)e^{\imath\overrightarrow{k}\cdot\overrightarrow{r}},$$

where the space dependence is explicitly stated. Maxwell equations can now be recast in a form like:

$$d_t\overrightarrow{\mathcal{F}} = \mathbf{G}_s\overrightarrow{\mathcal{F}}. \tag{1.15}$$

$\mathbf{G}_s$ is called the *spatial amplification matrix* and is a function of phase speed, wavenumber and grid-spacings. The solution of (1.15) has a solution that varies as $e^{\lambda t}$, where λ are the eigenvalues of $\mathbf{G}_s$. Note that the exact solution varies as $e^{-\imath\omega t}$, with $\omega = \kappa c$ and $\kappa = \left\|\overrightarrow{k}\right\|$. In general, $\lambda = \lambda_R + \imath\lambda_I$ is a complex number: its real part determines the dissipative error of the spatial discretization and its imaginary part determines the dispersive error.

Let:

$$\widetilde{c} = -\frac{\lambda_I}{\kappa}$$

be the numerical phase speed. In a time interval Δt, the harmonic wave undergoes a phase shift $\phi = -\kappa c \Delta t$, while the numerical wave a phase shift:

$$\widetilde{\phi} = -\kappa \widetilde{c} \Delta t = \frac{\widetilde{c}}{c} \phi.$$

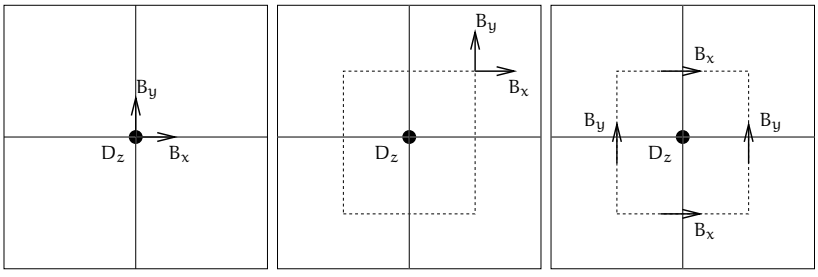
The ratio $\frac{\widetilde{c}}{c}$ is called the *normalized numerical phase speed* and $\frac{\widetilde{\phi}}{\phi}$ the *normalized numerical phase shift*.

The *numerical dissipative error* in the time interval Δt is $e^{\lambda_R \Delta t}$.

As said in (1.3), the unknowns must be associated to the geometrical elements of the grid: this association determines the characteristics of the algorithms. Closely following the analysis in [Liu96], we will first investigate Cartesian grids and then hexagonal grids, whose geometry is closely related to unstructured grids. For both of them, the spatial discretization scheme can be unstaggered, collocated staggered and uncollocated staggered.

1.4.1 Cartesian Grids

In the unstaggered grid, the unknowns $\overrightarrow{D}$ and $\overrightarrow{B}$ are placed on the points of the primal grid. There is no dual grid. In the collocated staggered grid $\overrightarrow{D}$ is placed on the points of the primal grid and $\overrightarrow{B}$ on the points of the dual grid. Finally, in the uncollocated staggered grid $\overrightarrow{D}$ is placed on the points of the primal grid and $\overrightarrow{B}$ on the midpoints of the dual edges. See Figure 1.9 for a two-dimensional example of the three schemes. Note that for all the schemes, the number of unknowns is the same, except on the boundaries of the domain.



(a) Unstaggered.

(b) Collocated staggered.

(c) Uncollocated staggered.

Figure 1.9: *Placement of unknowns on two-dimensional Cartesian grids.*

The two-dimensional Maxwell equation, for the TM case, for example, is:

$$\begin{cases} \partial_t D_z &= \partial_x H_y - \partial_y H_x \\ \partial_t B_x &= -\partial_y E_z \\ \partial_t B_y &= \partial_x E_z \end{cases} . \tag{1.16}$$

For each scheme, we have:

- unstaggered grid:

$$\begin{aligned} d_t\ D_z{}^{j,k} &= \frac{H_y{}^{j+1,k} - H_y{}^{j-1,k}}{2\Delta x} - \frac{H_x{}^{j,k+1} - H_x{}^{j,k-1}}{2\Delta y} \\ d_t\ B_x{}^{j,k} &= -\frac{E_z{}^{j,k+1} - E_z{}^{j,k-1}}{2\Delta y} \\ d_t\ B_y{}^{j,k} &= \frac{E_z{}^{j+1,k} - E_z{}^{j-1,k}}{2\Delta x}; \end{aligned} \tag{1.17}$$

- collocated staggered grid:

$$\begin{aligned} d_t\ D_z{}^{j,k} &= \frac{\begin{array}{c} H_y{}^{j+1/2,k+1/2} + H_y{}^{j+1/2,k-1/2} \\ - H_y{}^{j-1/2,k+1/2} - H_y{}^{j-1/2,k-1/2} \end{array}}{2\Delta x} \\ &\quad - \frac{\begin{array}{c} H_x{}^{j+1/2,k+1/2} + H_x{}^{j-1/2,k+1/2} \\ - H_x{}^{j+1/2,k-1/2} - H_x{}^{j+1/2,k-1/2} \end{array}}{2\Delta y} \\ d_t\ B_x{}^{j+1/2,k+1/2} &= -\frac{E_z{}^{j+1,k+1} + E_z{}^{j,k+1} - E_z{}^{j+1,k} - E_z{}^{j,k}}{2\Delta y} \\ d_t\ B_y{}^{j+1/2,k+1/2} &= \frac{E_z{}^{j+1,k+1} + E_z{}^{j+1,k} - E_z{}^{j,k+1} - E_z{}^{j,k}}{2\Delta x}; \end{aligned} \tag{1.18}$$

- uncollocated staggered grid:

$$\begin{aligned} d_t\ D_z{}^{j,k} &= \frac{H_y{}^{j+1/2,k} - H_y{}^{j-1/2,k}}{\Delta x} - \frac{H_x{}^{j,k+1/2} - H_x{}^{j,k-1/2}}{\Delta y} \\ d_t\ B_x{}^{j,k+1/2} &= -\frac{E_z{}^{j,k+1} - E_z{}^{j,k}}{\Delta y} \\ d_t\ B_y{}^{j+1/2,k} &= \frac{E_z{}^{j+1,k} - E_z{}^{j,k}}{2\Delta x}; \end{aligned} \tag{1.19}$$

where $F^{j,k}$ stands for the value of D_z, B_x or B_y evaluated at the points $j\Delta y$ and $k\Delta z$, Δy and Δz being the gridspacings in Figure 1.9.

The schemes (1.17) and (1.19) involve the same number of operations, while (1.18) involves about twice the number of operations. The scheme (1.17) divides

the system into two independent set of unknowns, which sometimes can lead to undesirable numerical oscillations: this is commonly referred to as the *odd-even decoupling* or the *chessboard decoupling*. There is no decoupling in the other two schemes.

For all the three schemes, the eigenvalues of each corresponding matrix **G** are pure imaginary or zero, implying that they are all non-dissipative, but dispersive.

The normalized numerical phase speed can be easily obtained substituting (1.14) into the previous equations. We obtain:

$$\frac{\widetilde{c}}{c} = \begin{cases} \left[\frac{\sin^2\frac{\xi}{2}}{\kappa^2\Delta x^2} + \frac{\cos^2\frac{\eta}{2}}{\kappa^2\Delta y^2}\right]^{1/2} & \text{unstaggered} \\ 2\left[\frac{\cos^2\frac{\eta}{2}\sin^2\frac{\xi}{2}}{\kappa^2\Delta x^2} + \frac{\cos^2\frac{\xi}{2}\sin^2\frac{\eta}{2}}{\kappa^2\Delta y^2}\right]^{1/2} & \text{collocated staggered} \\ 2\left[\frac{\sin^2\frac{\xi}{2}}{\kappa^2\Delta x^2} + \frac{\cos^2\frac{\eta}{2}}{\kappa^2\Delta y^2}\right]^{1/2} & \text{uncollocated staggered,} \end{cases} \tag{1.20}$$

where $\xi = k_x x$ and $\eta = k_y y$. Let $\theta = \tan^{-1}\frac{k_y}{k_x}$ be the direction of wave propagation. We can note that the normalized numerical phase speed depends on the direction of propagation, i.e. it is not isotropic. By defining the number of gridpoints per wavelength as $N = 2\pi/\kappa/\Delta s$, with $\Delta s = \Delta x = \Delta y$ in the case of a uniform grid spacing, we can rewrite the (1.20) as:

$$\frac{\widetilde{c}}{c} = \begin{cases} \frac{1}{2}\frac{N}{\pi}\left[\sin^2\frac{2\pi\cos\theta}{N} + \cos^2\frac{2\pi\sin\theta}{N}\right]^{1/2} & \text{u.} \\ \frac{N}{\pi}\left[\cos^2\frac{\pi\sin\theta}{N}\sin^2\frac{\pi\cos\theta}{N} + \cos^2\frac{\pi\cos\theta}{N}\sin^2\frac{\pi\sin\theta}{N}\right]^{1/2} & \text{c. s.} \\ \frac{N}{\pi}\left[\sin^2\frac{\pi\cos\theta}{N} + \cos^2\frac{\pi\sin\theta}{N}\right]^{1/2} & \text{u. s.,} \end{cases} \tag{1.21}$$

From the (1.21) we can note that the unstaggered grid requires twice the number of points per wavelength in each direction in order to have the same numerical phase speed as the uncollocated staggered grid. In other words, for a given grid spacing Δs, the error for the high frequency modes is greater than for the low frequency modes, because we have less points per wavelength.

The error, defined as $1 - \frac{\widetilde{c}}{c}$, is anisotropic (1.11): greatest along the axes ($\theta = 0, \pi/2, \pi, 3\pi/2$) and least along the diagonals ($\theta = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$) for both the unstaggered and uncollocated staggered grids. The opposite is true for the collocated staggered grid. Note that, even if a proper choice of the discretization scheme in time can improve the overall dispersive error, it cannot completely cancel it. This error depends only on the spatial discretization: the medium "grid" is anisotropic itself.

We can define the *isotropy error* as the difference between the maximum and the minimum values of the normalized numerical phase speed and we find that, for $N = 20$, it is 0.8% on the unstaggered grid, 0.4% on the collocated staggered grid and 0.2% on the uncollocated staggered grid. In other words, to have the same isotropy error of 0.1% for the three schemes we need $N = 58, 41, 29$ grid points per wavelength, respectively.

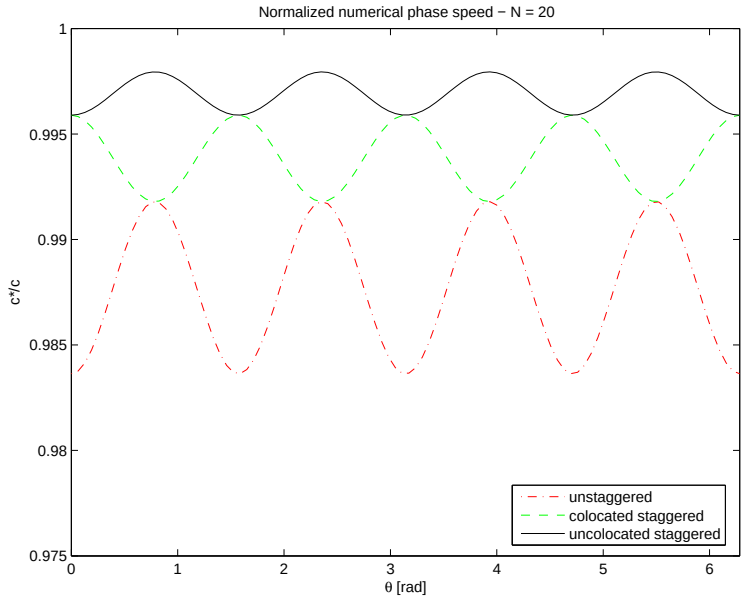
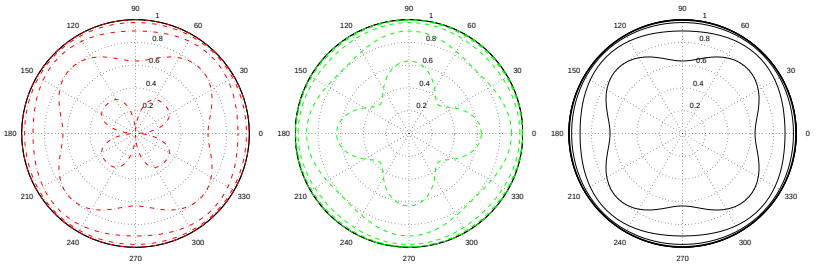


Figure 1.10: *Comparison of the normalized numerical phase speeds for Cartesian grids.*



(a) Unstaggered (b) Collocated staggered (c) Uncollocated staggered

Figure 1.11: *Polar diagrams of the normalized numerical phase speed for Cartesian grids and* $\mathsf{N} = 2, 4, 8, 16$.

The dependence of the isotropy error on N can be described expanding in Taylor series the equations (1.21):

$$1 - \frac{\widetilde{c}}{c} = \begin{cases} \left(\frac{1}{2} + \frac{1}{6}\cos 4\theta\right)\frac{\pi^2}{N^2} + \cdots & \text{unstaggered} \\ \left(\frac{1}{4} - \frac{1}{12}\cos 4\theta\right)\frac{\pi^2}{N^2} + \cdots & \text{collocated staggered} \\ \left(\frac{1}{8} + \frac{1}{24}\cos 4\theta\right)\frac{\pi^2}{N^2} + \cdots & \text{uncollocated staggered.} \end{cases}$$

The leading dispersive errors are proportional to N^{-2}, i.e. doubling the coarseness of the grid reduces the dispersive error by a factor 4.

From these considerations it should be clear that the uncollocated staggered grid is the best choice:

- it is more computationally efficient, as shown in (1.19);
- it is 4 and 2 times more precise than the unstaggered and collocated staggered schemes, respectively, from a dispersive error point of view, as shown in (1.21).

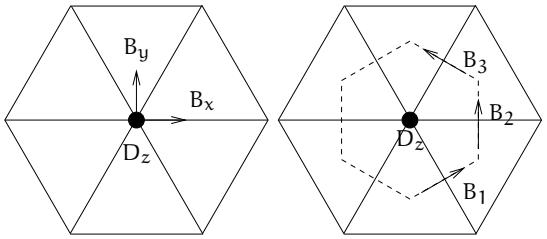
1.4.2 Hexagonal Grids

The major deficiencies of conventional schemes come from their one dimensional approach in which each spatial operator is approximated by employing data only along one coordinate line. Since only a five point Cartesian stencil is involved in each discretization, it is not surprising that all three schemes described in the previous subsection exhibit some anisotropy.

In this subsection we'll analyze a more efficient and accurate scheme, not based on an orthogonal grid, but on a 7-point stencil on regular hexagonal or triangular grid. This can also give a hint on the accuracy of unstructured grids, which are topologically equivalent to this one.

We also use the information of the previous subsection, in which we concluded that the uncollocated staggered scheme is the most efficient, and we'll limit our study to the same scheme, but applied to the new triangular grid.

Figure 1.12 shows two possible hexagonal discretization schemes: unstaggered or staggered. We will focus our attention only on the staggered one, which is the topological equivalent of the discretization scheme described in Chapter 1.



(a) Unstaggered. (b) Staggered.

Figure 1.12: *Placement of unknowns on two-dimensional hexagonal grids.*

With respect to Figure 1.12(b), (1.16) can be discretized as follows:

$$\begin{aligned}
d_t\ D_z{}^{j,k} &= \frac{2\left(\begin{array}{c} {H_1}^{j+1/4,k-\sqrt{3}/4} - {H_1}^{j-1/4,k+\sqrt{3}/4} + \\ {H_2}^{j+1/2,k} - {H_2}^{j-1/2,k} + \\ {H_3}^{j+1/4,k+\sqrt{3}/4} - {H_3}^{j-1/4,k-\sqrt{3}/4} \end{array}\right)}{3\Delta s} \\
d_t\ {B_1}^{j+1/4,k-\sqrt{3}/4} &= \frac{{E_z}^{j+1/2,k-\sqrt{3}/2} - {E_z}^{j,k}}{\Delta s} \\
d_t\ {B_2}^{j+1/2,k} &= \frac{{E_z}^{j+1,k} - {E_z}^{j,k}}{\Delta s} \\
d_t\ {B_3}^{j+1/4,k+\sqrt{3}/4} &= \frac{{E_z}^{j+1/2,k+\sqrt{3}/2} - {E_z}^{j,k}}{\Delta s}.
\end{aligned} \tag{1.22}$$

Again, the eigenvalues of the corresponding matrix $\mathbf{G}_s$ are pure imaginary or zero, implying that this scheme is non-dissipative, but dispersive.

Applying the Fourier analysis to the above equations, we obtain the normalized numerical phase speed:

$$\frac{\widetilde{c}}{c} = \frac{\sqrt{8/3}}{\kappa\Delta s}\left[\sin^2\frac{\xi}{2} + \sin^2\left(\frac{\xi}{4} - \frac{\sqrt{3}\eta}{4}\right) + \sin^2\left(\frac{\xi}{4} + \frac{\sqrt{3}\eta}{4}\right)\right]^{1/2} \tag{1.23}$$

and the isotropy error:

$$1 - \frac{\widetilde{c}}{c} = \frac{1}{8}\frac{\pi^2}{N^2} - \left(\frac{7}{1152} + \frac{1}{720}\cos 6\theta\right)\frac{\pi^4}{N^4} + \cdots. \tag{1.24}$$

Even if the isotropy error is still leaded by a term proportional to N^{-2}, Figure 1.13 shows that it doesn't depend anymore on the direction of the wave propagation.

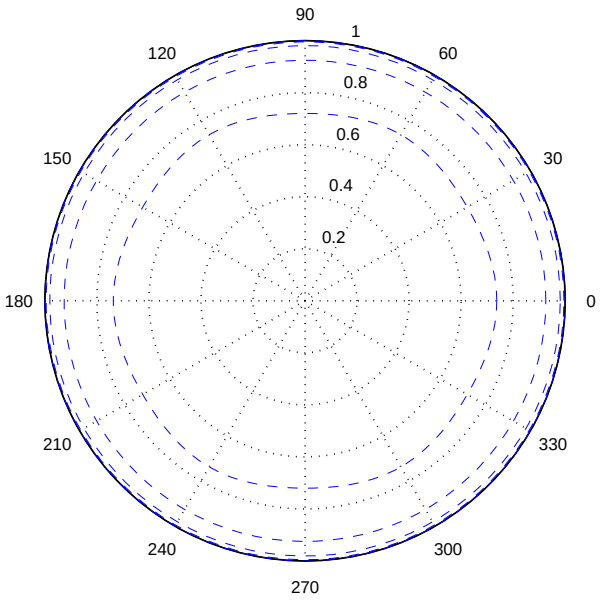


Figure 1.13: *Polar diagram of the normalized numerical phase speed for hexagonal grids.*

Quite surprisingly, its value is exactly the average of its Cartesian counterpart, as shown in Figure 1.14.

The anisotropy appears in the fourth order term, which is two orders of magnitude smaller than in the best Cartesian grid.

From these results, we conclude that uncollocated staggered hexagonal grids are clearly superior to Cartesian grids from a numerical error point of view: Cartesian grids require less operations and, if implemented on a computer, lead to a faster algorithm. A complete analysis of unstructured grids is prohibitive, because an analytical expression for the isotropy error is not possible: for each possible unstructured grids, an eigenvalue problem should be solved, to extract its characteristics. Being topologically equivalent to hexagonal grids, though, we can expect them to present the same good numerical errors characteristics. Moreover, they permit to discretize some parts of the computational domain with a coarser grid: this is a clear advantage over both Cartesian and hexagonal grids, whose gridspacing is constant everywhere. This becomes a speed advantage in a computer implementation.

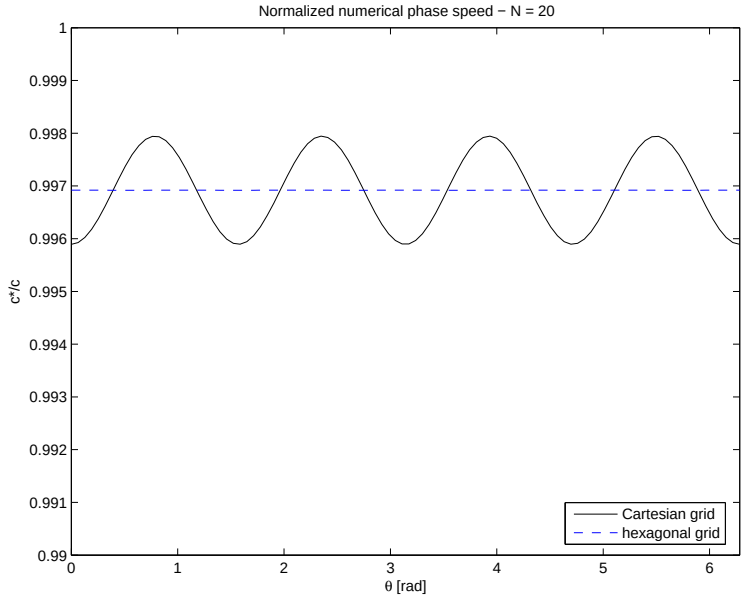


Figure 1.14: *Comparison of the normalized numerical phase speed for uncollocated staggered grids: Cartesian and hexagonal.*

2

Material equations

2.1 Introduction

In Section 1.3, we have described how to associate physical quantities to the geometrical elements of the computational mesh. We have also pointed out that Maxwell equations are purely topological relations, exactly satisfied by the degrees of freedom chosen, while errors, unavoidably introduced by the discretization process, are caused only by the non-perfect modeling of material equations. The resulting material matrices, also called *mass matrices* in Finite Elements context and *discrete Hodge operators* in differential geometry, greatly affects the stability and the accuracy of time- and frequency-domain algorithms [SSW02, SW98].

In the next sections we'll describe how to build the material matrices, given a primal mesh, for different dual meshes: in particular, for the Voronoï, Poincaré and barycentric dual grids.

2.2 Voronoï Dual Mesh

As described in Section 1.2.3, given a primal simplicial Delaunay mesh, its Voronoï dual mesh is built connecting the spherocenters of the primal cells. Some properties hold:

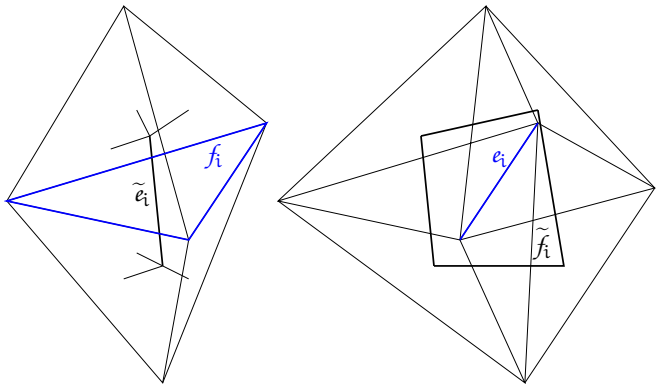
- dual edges are orthogonal to primal faces;
- primal edges are orthogonal to dual faces;
- dual volumes are the loci of nearest points to the corresponding primal point.

These properties represent, as we'll see later, an advantage in modeling Maxwell equations, but they represent tough geometrical constrains that meshing software must satisfy in order to generate proper meshes [She].

Consider a Delaunay primal mesh and its Voronoï dual mesh and let:

$$\begin{aligned} e_i &= \overrightarrow{E} \cdot \widehat{p}_i |e_i| \\ h_i &= \overrightarrow{H} \cdot \widehat{s}_i |\widetilde{e}_i| \\ b_i &= \overrightarrow{B} \cdot \widehat{s}_i \left| f_i \right| \\ d_i &= \overrightarrow{D} \cdot \widehat{p}_i \left| \widetilde{f}_i \right|, \end{aligned} \tag{2.1}$$

where $\widehat{p}_i$ ($\widehat{s}_i$) is the primal (dual) edge versor, be the realization of the mapping discussed in (1.3). The integral is not necessary because edges are considered straight and $\overrightarrow{E}$ is considered piecewise constant along each edge. From the definition of Voronoï dual mesh, $\widehat{p}_i$ is orthogonal to $\widehat{s}_i$. See Figure 2.1.



(a) The dual edge $\widetilde{e}_i$ is orthogonal to the primal face f_i.

(b) The edge e_i is orthogonal to the dual face $\widetilde{f}_i$

Figure 2.1: *Voronoï dual mesh.*

Maxwell equations read:

$$\begin{cases} {}^{n+1/2}b_i &= {}^{n-1/2}b_i - \Delta t \sum_{j \in \mathcal{B}_{f_i}} {}^{n}e_j \\ {}^{n+1}d_i &= {}^{n-1}d_i + \Delta t \sum_{j \in \mathcal{B}_{\widetilde{f}_i}} {}^{n+1/2}h_j . \end{cases} \tag{2.2}$$

$\mathcal{B}_{f_i}$ is the set of the indices that identifies the border edges of the primal face f_i:

$$\mathcal{B}_{f_i} = \left\{ j \ : \ e_j \in \partial f_i \right\}$$

where ∂f denotes the boundary of the face f, and $\mathcal{B}_{\widetilde{f}_i}$ is the set of the indices that identifies the border edges of the dual face $\widetilde{f}_i$:

$$\mathcal{B}_{\widetilde{f}_i} = \left\{ j \ : \ \widetilde{e}_j \in \partial \widetilde{f}_i \right\}$$

Thank to the properties of Voronoï meshes, the constitutive equations are straightforward. From (2.1), consider the i^{th} primal face, and the corresponding i^{th} (orthogonal) dual edge; we can write:

$$h_i = \frac{|\widetilde{e}_i|}{|f_i| \mu_i} b_i \tag{2.3}$$

where μ_i is the magnetic permeability associated to the primal face f_i: it is supposed piecewise constant on it. $|\widetilde{e}|$ and $\left|\widetilde{f}\right|$ are the measures of the dual edge $\widetilde{e}$ and the dual face $\widetilde{f}$, respectively; as we can note, material matrices require a metric to be defined [Bos98a]: they are not purely topological relations! We have chosen to use the classical Euclidean norm, as measure.

The other constitutive equation, for the i^{th} primal edge and the i^{th} dual face, reads:

$$e_i = \frac{|e_i|}{\left|\widetilde{f}_i\right| \epsilon_i} d_i \tag{2.4}$$

where ϵ_i is the electric permeability associated to the dual face $\widetilde{f}_i$: it is supposed piecewise constant on it.

Note that the value of h_i (e_i) only depends on the value of b_i (d_i) on the corresponding (dual) face. Rewriting (2.3) and (2.4) in matrix form, we have:

$$\mathbf{h} = \mathbf{M}_\mu \, \mathbf{b}$$
$$\mathbf{e} = \mathbf{M}_\epsilon \, \mathbf{d},$$

where $\mathbf{h}$, $\mathbf{b}$, $\mathbf{e}$ and $\mathbf{d}$ are the arrays of all the h_i, b_i, e_i and d_i in (2.2) and $\mathbf{M}_\epsilon$ and $\mathbf{M}_\mu$ are square diagonal matrices whose dimensions are the number of dual and primal faces, respectively:

$$\mathbf{M}_\mu \in n_f \times n_f \qquad\qquad \mathbf{M}_\epsilon \in n_{\widetilde{f}} \times n_{\widetilde{f}}.$$

2.3 Poincaré Dual Mesh

Given a primal triangular mesh, the Poincaré dual mesh is built connecting the barycenters of the primal volumes. Comparing to the Voronoï dual mesh, we can note:

1. the dual points are always inside the primal volumes: this makes the scheme more stable in the time-domain;

2. the dual edges are not orthogonal to the primal faces and the primal edges are not orthogonal to the dual faces: this makes the constitutive relations more complicated to implement.

Consider a 2D triangular primal mesh, as in Figure 2.2 and let's limit to the study of the TE polarization[1]: let E_z and D_z be associated to the nodes of the primal grid and $\overrightarrow{H}_t$ and $\overrightarrow{B}_t$ to the edges of the dual grid.

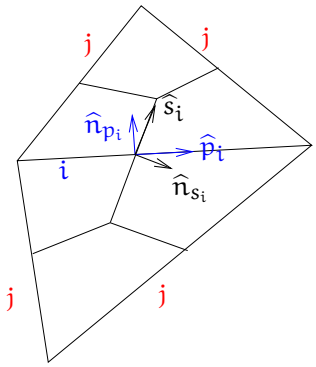


Figure 2.2: *Poincaré dual mesh: primal (dual) edges are not orthogonal to dual (primal) faces. In red, identified with* j*, are the primal faces (actually edges in the two-dimensional mesh) sharing a common vertical primal edge (actually a point) with the primal face (actually an edge)* i*.*

Let:

$$\begin{aligned} e_i &= \overrightarrow{E} \cdot \widehat{p}_i |e| \\ h_i &= \overrightarrow{H} \cdot \widehat{s}_i |\widetilde{e}_i| \\ b_i &= \overrightarrow{B} \cdot \widehat{n}_{p_i} \left|f_i\right| \\ d_i &= \overrightarrow{D} \cdot \widehat{n}_{s_i} \left|\widetilde{f}_i\right| \end{aligned} \tag{2.5}$$

where $\widehat{p}_i$ ($\widehat{s}_i$) is the primal (dual) edge versor and $\widehat{n}_{p_i}$ ($\widehat{n}_{s_i}$) is the versor orthogonal to the primal (dual) face. As noted before, $\widehat{n}_{p_i}$ is not parallel to $\widehat{s}_i$ and $\widehat{n}_{s_i}$ is not parallel to $\widehat{p}_i$ as in Voronoï dual grids.

Maxwell equations, being topological relations, are the same as for the Voronoï dual mesh, in (2.2).

Constitutive equations are more complicated because $\overrightarrow{H}_t \nparallel \overrightarrow{B}_t$. The electric constitutive equation (2.4), valid in the Voronoï case, holds also in this case: this is because we are dealing now only with the TE case, for which the electric field component E_z is orthogonal to dual faces (which lie on the xy plane). For the

[1] for the TM polarization, the Duality Theorem can be applied [Som98]; the extension to the 3D case is straightforward, too.

magnetic constitutive equation, recall that we have to write:

$$\mathbf{h} = \mathbf{M}_\mu \, \mathbf{b},$$

where $\mathbf{h}$ and $\mathbf{b}$ are the arrays of all the h_i and b_i in (2.2) and $\mathbf{M}_\mu$ is a square matrix whose dimension is the number of dual edges (or primal faces).

This can be done in three steps [Taf98].

1. Consider the primal face f_i and all the faces f_j sharing with it one edge, like in Figure 2.2. Write $\overrightarrow{B}_j$, the magnetic field on the face f_j, as a linear combination of b_i and b_j, magnetic fluxes through f_i and f_j:

$$\begin{cases} \overrightarrow{B}_j \cdot \overrightarrow{N}_{p_i} = \overrightarrow{B} \cdot \overrightarrow{N}_{p_i} = b_i \\ \overrightarrow{B}_j \cdot \overrightarrow{N}_{p_j} = \overrightarrow{B} \cdot \overrightarrow{N}_{p_j} = b_j. \end{cases}$$

   where $\overrightarrow{N}_{p_i} = \hat{n}_{p_i} |f_i|$ and the same for $\overrightarrow{N}_{p_j}$. These relations come from the imposition of flux conservation.

   The right-hand sides in (1) are known from Faraday equation (the first in (2.2)), so we can calculate $\overrightarrow{B}_j$:

$$\overrightarrow{B}_j = \mathbf{N}^{-1} \begin{bmatrix} b_i \\ b_j \end{bmatrix} \qquad \text{where} \qquad \mathbf{N} = \begin{bmatrix} N^x_{p_i} & N^y_{p_i} \\ N^x_{p_j} & N^y_{p_j} \end{bmatrix}.$$

2. Write $\overrightarrow{B}$ as a linear combination of $\overrightarrow{B}_j$:

$$\overrightarrow{B}_i = \frac{1}{W} \sum_j w_j \overrightarrow{B}_j,$$

   where $W = \sum_j w_j$ and $w_j = \left\| \hat{z} \cdot \left(\overrightarrow{N}_{p_i} \times \overrightarrow{N}_{p_j} \right) \right\|$ are weighting coefficients whose value is twice the area of the triangle identified by the edges i and j.

3. Finally, compute $\overrightarrow{H}$ from the continuum constitutive relation $\overrightarrow{H} = \mu^{-1} \overrightarrow{B}$ and then h_i, from (2.5):

$$\begin{aligned} h_i &= \overrightarrow{H} \cdot \hat{s}_i |\widetilde{e}_i| \\ &= H_x |\widetilde{e}_i| \cos\beta + H_y |\widetilde{e}_i| \sin\beta \\ &= \mu_x^{-1} B_x |\widetilde{e}_i| \cos\beta + \mu_y^{-1} B_y |\widetilde{e}_i| \sin\beta \\ &= \frac{1}{W} \sum_j w_j \left[\mu_x^{-1} B_x |\widetilde{e}_i| \cos\beta + \mu_y^{-1} B_y |\widetilde{e}_i| \sin\beta \right] \\ &= \frac{1}{W} \sum_j w_j \begin{bmatrix} \mu_x^{-1} |\widetilde{e}_i| \cos\beta \, (k_{11} b_i + k_{12} b_j) \\ + \mu_y^{-1} |\widetilde{e}_i| \sin\beta \, (k_{21} b_i + k_{22} b_j) \end{bmatrix} \\ &= \frac{|\widetilde{e}_i|}{W} \sum_j w_j \begin{bmatrix} \left(\mu_x^{-1} k_{11} \cos\beta + \mu_y^{-1} k_{21} \sin\beta \right) b_i \\ + \left(\mu_y^{-1} k_{12} \cos\beta + \mu_y^{-1} k_{22} \sin\beta \right) b_j \end{bmatrix} \end{aligned} \tag{2.6}$$

where β is the angle between the dual edge $\widetilde{e}_i$ and the x-axis and $\mathbf{K} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \mathbf{N}^{-1}$.

In matricial form, (2.6) reads:

$$\mathbf{h} = \mathbf{M}_\mu \, \mathbf{b},$$

where

$$\mathbf{M}_\mu = \left[\mathbf{M}_{\mu_{ij}}\right] = \begin{cases} \frac{|\widetilde{e}|}{W} \sum_j w_j \left(\mu_x^{-1} k_{11} \cos\beta + \mu_y^{-1} k_{21} \sin\beta\right) & \text{for } i = j \\ \frac{|\widetilde{e}|}{W} w_j \left(\mu_x^{-1} k_{12} \cos\beta + \mu_y^{-1} k_{22} \sin\beta\right) & \text{for } i \neq j \end{cases} \tag{2.7}$$

It is worth noting that:

- these three steps are fully explicit, i.e. we don't ever need to solve a linear system to compute the unknowns we need;
- the final matrix $\mathbf{M}_\mu$ is not diagonal, as explicitly stated in (2.7): this is not a problem, neither in the time- nor in the frequency-domain;
- (2.6) is valid also for anisotropic materials: just use a tensor $\mu = \begin{bmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{yx} & \mu_{yy} \end{bmatrix}$. Note also that dealing with an anisotropic material is equivalent in introducing a local change of metric. In fact, for the simple case of a coordinate system oriented with the crystallographic axis of the anisotropic material[2]:

$$\begin{aligned} h_i &= \left(\mu^{-1} \overrightarrow{B}\right) \widehat{s}_i |\widetilde{e}_i| \\ &= \left(\mu_x^{-1} s_{ix} B_x + \mu_y^{-1} s_{iy} B_y\right) |\widetilde{e}_i| \\ &= \overrightarrow{B} \, \overrightarrow{s}'_i |\widetilde{e}_i| \end{aligned}$$

where $\overrightarrow{s}'_i = \left(\mu_x^{-1} s_{ix}, \mu_y^{-1} s_{iy}\right)$. The metric, as said before, is needed when dealing with material equations.

2.4 Barycentric Dual Mesh

Given a triangular primal mesh, its barycentric dual is made of:

- dual points that are the barycenters of the primal volumes;
- dual edges that are piece-lines made by two straight lines whose contact point is the barycenter of the corresponding primal face, and connect the dual points;
- dual surfaces having as a border the dual lines surrounding the corresponding primal edge.

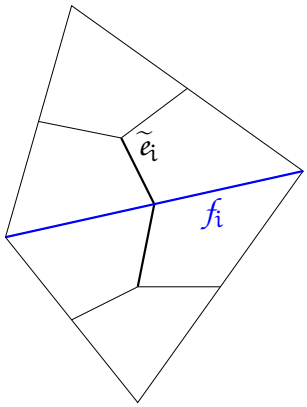


Figure 2.3: *Barycentric dual mesh: dual edges are piece-lines to connect barycenters of adjacent primal cells.*

Since primal and dual meshes are not orthogonal, as in the Poincaré mesh, the constitutive equation requires a similar mathematical procedure to be worked out. The advantage of this mesh is its superior stability and the possibility to use any kind of triangular grid, not only a Delaunay one [Mar00].

Closely following [Mar00], we have implemented the material matrices for the two-dimensional mesh and both the polarizations. It has been used to study dispersive materials, as described in Section 3.5

2.5 Particular Material Equations

In the previous sections the conventional material equations have been discretized for different dual meshes. We'll show now how more complex materials can be modeled with the present method. Another example can be found in Section 3.5.

2.5.1 Ohm Losses

The description of the Ohm losses can be easily included in (1.6) and (1.7), as follows:

$$\begin{cases} \mathbf{R}^{\top}\,\mathbf{h} = \partial_t \mathbf{d} + \mathbf{j} + \Omega_{\mathbf{EN}}\mathbf{e} \\ \mathbf{R}\,\mathbf{e} = -\partial_t \mathbf{b} + \mathbf{m} - \Omega_{\mathbf{MN}}\mathbf{h} \end{cases} .$$

According to the association of the degrees of freedom to the geometrical elements of the mesh, shown in Section 1.3, supposing that given a primal edge, the electric field is uniform on it and the conductibility is uniform over the corresponding dual face, we can easily build the matrices $\Omega_{\mathbf{EN}}$ and $\Omega_{\mathbf{MN}}$ with the

[2] If not, a simple rotation of the reference axis can bring us to this condition

procedures described in the previous sections for the material matrices, depending on the particular dual grid chosen. The only problem arises in the time domain: in fact, the matrix/operator $\Omega_{\mathbf{EN}}$ ($\Omega_{\mathbf{MN}}$) must link a physical quantity defined on the primal instants (the electric field) with one defined on the dual instants (the magnetic field). Hence, it must be estimated by some means of time approximation. The easiest way to do it, which keeps the algorithm stable, is to approximate the electric field at the time-step $n + 1/2$ as the mean value of the instants before and after it:

$$^{n+1/2}\mathbf{e} \simeq \frac{^{n}\mathbf{e} + {}^{n+1}\mathbf{e}}{2}.$$

The same reasoning holds for the magnetic conductivity matrix $\Omega_{\mathbf{MN}}$[3].

It should be clear that this problem only arises in the time-domain. In the frequency domain, the easiest way to deal with Ohm losses is to accept a complex value for the electric and magnetic permeabilities. In formulas:

$$\begin{cases} \nabla \times \overrightarrow{H} = -\imath\omega\epsilon\overrightarrow{E} + \sigma_e\overrightarrow{E} \\ \nabla \times \overrightarrow{E} = \imath\omega\mu\overrightarrow{H} + \sigma_m\overrightarrow{H} \end{cases} \implies \begin{cases} \nabla \times \overrightarrow{H} = -\imath\omega\bar{\epsilon}\overrightarrow{E} \\ \nabla \times \overrightarrow{E} = \imath\omega\bar{\mu}\overrightarrow{H} \end{cases}$$

where $\bar{\epsilon} = \epsilon + \imath\frac{\sigma_e}{\omega}$ and $\bar{\mu} = \mu + \imath\frac{\sigma_m}{\omega}$. Obviously, also the material matrices $\mathbf{M}_\epsilon$ and $\mathbf{M}_\mu$ will be complex.

For later convenience, let $\mathbf{P}_m = \mathbf{M}_\mu\,\Omega_{MN}$ and $\mathbf{P}_e = \mathbf{M}_\epsilon\,\Omega_{EN}$.

2.5.2 PML

PMLs are used to model materials with infinite extent, i.e. to model the radiation of light as if the boundaries of the computational domain were at infinite distance from the source. There are mainly two formulations to model the PMLs:

coordinate stretching : in which a local change of metric is introduced where attenuation is needed, so that any plane wave impinging at the interface undergoes no reflection, but exponentially decays at zero as propagating inside the PMLs;

Uniaxial PML : in which an anisotropic lossy material, but fully Maxwellian, is defined, with the property of absorbing, without reflections, any plane wave impinging to it.

Even if both the formulations have been successfully implemented in many algorithms, we have decided to implement the second. The main reason is that the U-PML are a fully Maxwellian medium, while the medium coming out from the coordinate stretching is not: Maxwell equations hold in the U-PMLs, but not in

[3]Even if the magnetic conductivity can appear a useless physical quantity, it is not, because it is used in the modelling of PML losses

the coordinate stretched system and new equation must be implemented[4]. Having a scheme of degrees of freedom and geometrical elements fully functional to solve Maxwell equations it seems a waste of time to introduce a new one only where PMLs are needed: sometimes, laziness is a good quality for scientists and programmers.

Uniaxial PMLs are an anisotropic lossy medium. Maxwell equations in this medium read:

$$\begin{cases} \nabla \times \overrightarrow{E} = -\imath\omega\underline{\mu}\overrightarrow{H} \\ \nabla \times \overrightarrow{H} = \imath\omega\underline{\epsilon}\overrightarrow{E} + \overrightarrow{J} \end{cases},$$

where $\underline{\mu}$ and $\underline{\epsilon}$ are tensors defined as:

$$\underline{\mu} = \mu\,\underline{s}$$
$$\underline{\epsilon} = \epsilon\,\underline{s}$$

and:

$$\underline{s} = \underline{s}_x\,\underline{s}_y\,\underline{s}_z$$
$$\underline{s}_x = \begin{bmatrix} s_x^{-1} & 0 & 0 \\ 0 & s_x & 0 \\ 0 & 0 & s_x \end{bmatrix}$$
$$\underline{s}_y = \begin{bmatrix} s_y & 0 & 0 \\ 0 & s_y^{-1} & 0 \\ 0 & 0 & s_y \end{bmatrix}$$
$$\underline{s}_z = \begin{bmatrix} s_z & 0 & 0 \\ 0 & s_z & 0 \\ 0 & 0 & s_z^{-1} \end{bmatrix}$$

and, finally:

$$s_x = k_x + \frac{\sigma_x}{\imath\omega\epsilon_0}$$
$$s_y = k_y + \frac{\sigma_y}{\imath\omega\epsilon_0}$$
$$s_z = k_z + \frac{\sigma_z}{\imath\omega\epsilon_0}.$$

In the previous sections, we have shown how to describe both anisotropic and lossy materials, so we have all we need to describe U-PMLs, at least in the frequency-domain.

[4]In particular, Gauss laws fail to hold: at the interface between two different media inside the PMLs, the normal components of both $\overrightarrow{E}$ and $\overrightarrow{D}$ are continuous, instead of just $\overrightarrow{D}$. The same is valid for $\overrightarrow{H}$ and $\overrightarrow{B}$.

In the time-domain, the presence of the frequency ω in the definition of the absorbing coefficients introduces a *non-locality* in time. Material equations in the frequency-domain, are valid in the form:

$$\overrightarrow{D}(\omega) = \epsilon(\omega)\overrightarrow{E}(\omega)$$
$$\overrightarrow{B}(\omega) = \mu(\omega)\overrightarrow{H}(\omega)$$

that, in the time-domain, become a convolution:

$$\overrightarrow{D}(t) = \int \epsilon(t-\tau)\overrightarrow{E}(\tau)d\tau$$
$$\overrightarrow{B}(t) = \int \mu(t-\tau)\overrightarrow{H}(\tau)d\tau$$

The consequence is that the value of $\overrightarrow{D}$ at the time t depends on the values of $\overrightarrow{E}$ for all the previous[5] times τ.

Using some notation that will be explained in Chapter 3, we can write Maxwell equations in the time-domain, which hold in the PMLs:

$$\begin{cases} {}^{n+1}\mathbf{d} = {}^{n}\mathbf{d} + \Delta t\mathbf{R}^{T}\ {}^{n+1/2}\mathbf{h} - \Delta t\Omega_{EN}\ {}^{n+1/2}\mathbf{e} - {}^{n+1/2}\mathbf{j} \\ {}^{n+1}\mathbf{e} = {}^{n}\mathbf{e} + \mathbf{M}_{\epsilon}\left({}^{n+1}\mathbf{d} - {}^{n}\mathbf{d}\right) - \Delta t\Omega_{EC}\ {}^{n+1/2}\mathbf{e} \\ {}^{n+3/2}\mathbf{b} = {}^{n+1/2}\mathbf{b} - \Delta t\mathbf{R}\ {}^{n+1}\mathbf{e} - \Delta t\Omega_{MN}\ {}^{n+1}\mathbf{h} + {}^{n+1}\mathbf{m} \\ {}^{n+3/2}\mathbf{h} = {}^{n+1/2}\mathbf{h} + \mathbf{M}_{\mu}\left({}^{n+3/2}\mathbf{b} - {}^{n+1/2}\mathbf{b}\right) - \Delta t\Omega_{MC}\ {}^{n+1}\mathbf{h} \end{cases} \tag{2.8}$$

where:

$$\Omega_{MN} = \sigma\frac{|f|}{|\widetilde{e}|} \qquad \text{Magnetic Ohm losses}$$
$$\Omega_{MC} = \frac{\sigma}{\mu_0} \qquad \text{Magnetic PML losses}$$
$$\Omega_{EN} = \sigma\frac{|\widetilde{f}|}{|e|} \qquad \text{Electric Ohm losses}$$
$$\Omega_{EC} = \frac{\sigma}{\epsilon_0} \qquad \text{Electric PML losses}$$

Note that where $\sigma = 0$, (2.8) becomes (3.1): we don't need two different equations to deal with PMLs.

[5]Only the previous times, not the future, because causality always holds.

3

Time-Domain

In Chapter 1, we have discussed about the discretization in space of Maxwell equations, for an unstructured grid. The time-domain version of Maxwell equations, though, present a derivatives in time that have to be discretized as well. Luckily enough, being time one-dimensional, the discretization scheme can only be collocated or uncollocated, i.e. with all the physical quantities associated to the same time or with a time grid for each of them, as described in Section 1.3.

For the reasons that will be explained in Section 3.4, we have chosen, for our algorithm, an uncollocated scheme, already widely used in electromagnetic simulation software, like the `FDTD` : the *leapfrog timestepping*.

3.1 Leapfrog Timestepping

In [Taf00] a precise description of the leapfrog timestepping, applied to a Cartesian grid, is given. Applying it to our unstructured grid is not different [BF04].

A more detailed description of the method is given in Section 3.4, but it's worth seeing now how it applies to our algorithm.

Let's divide the time in discrete points, or instants, $t_0, t_1, \ldots, t_n$: this automatically identifies time intervals $[t_0, t_1], [t_1, t_2], \ldots, [t_{n-1}, t_n]$. Dual instants $\widetilde{t_0}, \widetilde{t_1}, \ldots, \widetilde{t_n}$ are chosen at the midpoints of primal intervals (Figure 3.1).

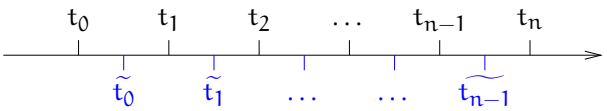


Figure 3.1: *Discretization of time: dual instants (in blue) are the midpoints of primal intervals.*

We arbitrarily associate the electric field $\overrightarrow{E}$ to primal instants; the Faraday equation in (1.2) tells us to associate the partial derivative $\partial_t \overrightarrow{B}$ to the same primal instants. Using a central difference scheme to discretize the derivative in time at the

first order, we have:

$$d_t \overrightarrow{B}|_{t_n} = \frac{\overrightarrow{B}|_{t_{n+1/2}} - \overrightarrow{B}|_{t_{n-1/2}}}{t_{n+1/2} - t_{n-1/2}}.$$

$\overrightarrow{B}$ is evaluated at timesteps $t_{-1/2}, t_{1/2}, \ldots, t_{n-1/2}, t_{n+1/2}$, i.e. on dual instants.

A word about notation: from now on, the value of a field $\overrightarrow{F}$ evaluated at the timestep t_n is written as:

$$\overrightarrow{F}|_{t_n} = {}^n\overrightarrow{F}.$$

3.2 Sources

Before being able to write the complete set of the discretized Maxwell equations in the time-domain, we need to implement sources.

There are many possible sources of an electromagnetic field: in our algorithm, we have implemented two [BF04].

Current sources : they are used to model antennas and dipoles. They are very easy to model, because they are already in Maxwell equations, identified by the electric current $\overrightarrow{J}$ and magnetic current $\overrightarrow{M}$.

Field sources : they are used if the electromagnetic field distribution is known on a surface inside or on the boundary of the domain[1]. This is a very common situation: one of the most interesting problems is to study the behavior of a device whose input is provided by some modes of an input waveguide: in this case, the field distribution is known at the input waveguide cross section and we want to compute the field inside the device under test. Modeling these sources is not a trivial task, though. The *Equivalence Theorem* [Som98] helps us. It reads that the electromagnetic field inside a region R, generated by a field distribution $\overrightarrow{E}_i, \overrightarrow{H}_i$ at its boundary dR, is equivalent to the one generated by two surface currents $\overrightarrow{J}_s$ and $\overrightarrow{M}_s$ on dR, whose linear densities are given by:

$$\overrightarrow{J}_s = \hat{n} \times \overrightarrow{H}_i \qquad\qquad \overrightarrow{M}_s = \overrightarrow{E}_i \times \hat{n},$$

where $\hat{n}$ is the normal versor to dR.

With the choice of degrees of freedom explained in (1.3), $\overrightarrow{J}_s$ and $\overrightarrow{M}_s$ are naturally associated with primal and dual surfaces respectively, as fluxes through them (see Figure 3.2). Note that we can pass from a linear current density to a flux through a surface using the Dirac δ function:

$$J(x, y, z) = J_s(x, y)\delta(z).$$

[1] Strictly speaking, just the tangential part of the electromagnetic field is sufficient

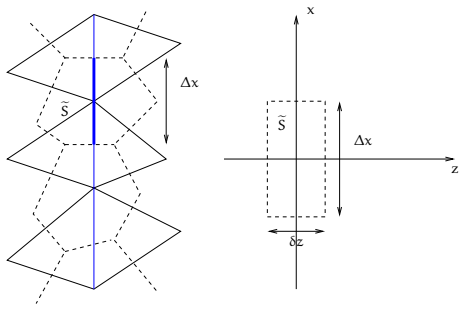


Figure 3.2: *Electric sources are associated to the dual surfaces, magnetic sources to the primal surfaces. If the electromagnetic field distribution is known on the blue line, we can use the* Equivalence Theorem *to build equivalent sources: at limit, we can think of the line (in two dimensions) over which the equivalent sources are defined as an infinitesimally thin surface.*

It is as if we thought to reduce the surfaces, through which we are computing the fluxes, to segments: the limiting process gives linear densities. Finally, in the coordinate system of Figure 3.2, we can write:

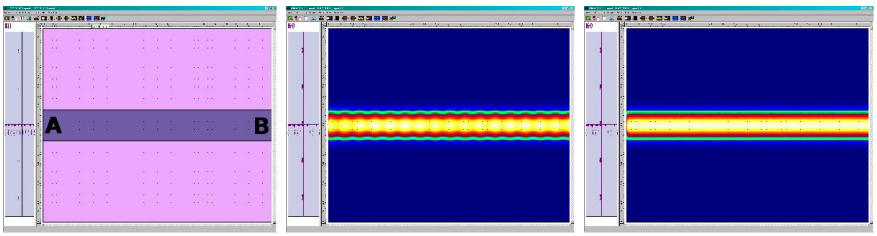
$$\int_{\tilde{S}} \overrightarrow{J} \cdot \hat{y} d\tilde{S} = 2\hat{n} \times \overrightarrow{H} \cdot \hat{y}.$$

Note that if we only decide to excite one of the $\overrightarrow{J}_s$ and $\overrightarrow{M}_s$ fields, we'll excite two electromagnetic waves, propagating in opposite directions. This is accounted for by the factor 2 in the equation above. To have a wave propagating in one direction, we need to excite both $\overrightarrow{J}_s$ and $\overrightarrow{M}_s$ at the same time.

Figure 3.3 shows an example of field sources usage: they have been applied to tune the PMLs and achieve optimal performance (see Section 2.5.2 for the description of PMLs). As long as there is no optimal design parameters for the PMLs, but they depend on the particular domain [Taf00], we have studied a problem whose solution is known analytically: a simple dielectric slab waveguide.

The input source is provided by the fundamental mode of the waveguide, computed by FIMMWAVE [FIMb]: the optimal PMLs parameters are computed imposing the minimum reflection at the right hand side facet. To estimate reflections we define the *transmission coefficient* and the *reflection coefficient* of the device as:

$$T_j = \int_B \overrightarrow{E}_B \times \overrightarrow{H}_{m_i}^* \cdot \hat{x} dB$$
$$R_j = \int_A (\overrightarrow{E}_A - \overrightarrow{E}_{m_i}) \times \overrightarrow{H}_{m_i}^* \cdot \hat{x} dA = \int_A \overrightarrow{E}_A \times \overrightarrow{H}_{m_i}^* \cdot \hat{x} dA - 1,$$



(a) Waveguide. (b) Badly tuned PMLs. (c) Finely tuned PMLs.

Figure 3.3: *PMLs tuning by field sources: if finely tuned, PMLs give no reflections to the input and output facets of the waveguide.*

where x is the axis of the waveguide, A is the left-hand side facet of the waveguide, B the right-hand side facet and $\left\{\overrightarrow{E}_{m_i}, \overrightarrow{H}_{m_i}\right\}$ is the m_i mode of the waveguide.

We have estimated that we obtain negligible reflections, below 40dB, for a PML thickness of about 1λ and parabolic losses profile.

3.3 Matricial Representation

We are now ready to write the complete discretized Maxwell equation in the time domain [TK04].

Starting from (1.6) and (1.7), with the material equations ((1.11) and (1.10)) and using the leapfrog timestepping to discretize the partial derivative in time, we obtain:

$$\begin{cases} {}^{n+1}\mathbf{d} = {}^{n}\mathbf{d} + \Delta t \mathbf{R}^{T}\ {}^{n+1/2}\mathbf{h} - {}^{n+1/2}\mathbf{j} \\ {}^{n+1}\mathbf{e} = \mathbf{M}_{\epsilon}\ {}^{n+1}\mathbf{d} \\ {}^{n+3/2}\mathbf{b} = {}^{n+1/2}\mathbf{b} - \Delta t \mathbf{R}\ {}^{n+1}\mathbf{e} + {}^{n+1}\mathbf{m} \\ {}^{n+3/2}\mathbf{h} = \mathbf{M}_{\mu}\ {}^{n+3/2}\mathbf{b} \end{cases}, \tag{3.1}$$

where the vectors are defined in Section 1.3 and the matrices $\mathbf{M}_{\epsilon}$ and $\mathbf{M}_{\mu}$ are defined in Chapter 2. Ohm losses can also be included: see Section 2.5.1.

The initial conditions are ${}^{0}\mathbf{d}$, ${}^{1/2}\mathbf{j}$, ${}^{1/2}\mathbf{h}$ and ${}^{1}\mathbf{m}$: all the other fields, at every timestep, are *explicitly* computed from these.

We can note that in the simple formulation of (3.1), the $\mathbf{e}$ and the $\mathbf{h}$ vectors are *auxiliary* fields: the constitutive equations are so easily expressed that they can be included in the Ampere and Faraday equations:

$$\begin{cases} {}^{n+1}\mathbf{d} &= {}^{n}\mathbf{d} + \Delta t \mathbf{R}^{T} \mathbf{M}_{\mu}\ {}^{n+1/2}\mathbf{b} - {}^{n}\mathbf{j} \\ {}^{n+3/2}\mathbf{b} &= {}^{n+1/2}\mathbf{b} - \Delta t \mathbf{R}\, \mathbf{M}_{\epsilon}\ {}^{n+1}\mathbf{d} + {}^{n+1/2}\mathbf{m} \end{cases}$$

or, viceversa, we can consider $\mathbf{d}$ and $\mathbf{b}$ auxiliary and write:

$$\begin{cases} {}^{n+1}\mathbf{e} &= {}^{n}\mathbf{e} + \Delta t \mathbf{M}_\epsilon \, \mathbf{R}^T \, {}^{n+1/2}\mathbf{h} - \mathbf{M}_\epsilon \, {}^{n}\mathbf{j} \\ {}^{n+3/2}\mathbf{h} &= {}^{n+1/2}\mathbf{h} - \Delta t \mathbf{M}_\mu \, \mathbf{R} \, {}^{n+1}\mathbf{e} + \mathbf{M}_\mu \, {}^{n+1/2}\mathbf{m} \end{cases} . \tag{3.2}$$

We can further simplify Maxwell equations obtaining the discrete Helmholtz equation. Derive in time (1.6) and substitute it in (1.7), using (1.10) and (1.11). We obtain:

$$\mathbf{M}_\epsilon \, \mathbf{R}^T \, \mathbf{M}_\mu \, \mathbf{R} \, \mathbf{e} = -\partial_t^2 \mathbf{e}. \tag{3.3}$$

With the leapfrog timestepping, this scheme is stable only if the eigenvalues of the matrix $\mathbf{A} = \mathbf{M}_\epsilon \, \mathbf{R}^T \, \mathbf{M}_\mu \, \mathbf{R}$ are real and positive [Liu96]. As we'll show later, a sufficient condition is that the material matrices $\mathbf{M}_\epsilon$ and $\mathbf{M}_\mu$ are symmetric and positive definite [SSW02, SW98]. This condition is easily satisfied for Voronoï grids, where the material matrices are diagonal, but special care must be taken for other kinds of dual grids.

3.4 Stability of Time Discretization

Following [Liu96], we can study the stability and accuracy of the time discretization. Like spatial discretization, it can also introduce errors: however, these errors are only associated with dispersion and dissipation and they are isotropic.

The equation 1.15 can be diagonalized, projecting $\overrightarrow{\mathcal{F}}$ on the eigenspace:

$$d_t f = \lambda f. \tag{3.4}$$

For a given time discretization, the amplification factor σ, defined as the ratio of f at two adjacent time levels

$$\sigma = {}^{n+1}f / {}^{n}f$$

can be expressed as a function of λ[2]

Substituting the eigenvalues of $\mathbf{G}_s$ into σ, one obtains the combined errors of spatial and time discretization and 1.15 becomes:

$${}^{n+1}\overrightarrow{\mathcal{F}} = \mathbf{G} \, {}^{n}\overrightarrow{\mathcal{F}}.$$

$\mathbf{G}$ is called total amplification matrix. The eigenvalues of $\mathbf{G}$ give the amplification factor σ. The modulus of σ determines the dissipative error and the stability of the algorithm and the argument determines the combined phase shift $\widetilde{\phi} = \angle\sigma$.

[2]For example, with a leapfrog timestepping, 3.4 becomes:

$${}^{n+1}f - {}^{n}f = \Delta t \lambda \, {}^{n+1/2}f = \frac{\Delta t \lambda}{2}\left({}^{n+1}f + {}^{n}f\right) \qquad \Longrightarrow \qquad {}^{n+1}f = \underbrace{\left(\frac{1+\frac{\lambda\Delta t}{2}}{1-\frac{\lambda\Delta t}{2}}\right)}_{\sigma} {}^{n}f.$$

Time integration of Maxwell equations can be solved in many ways, both explicitly and implicitly. While implicit schemes are usually unconditionally stable (for example, the ADI timestepping [Taf00]), they require the solution of a linear system at each timestep, which can be computational too intensive. On the other hand, explicit schemes are conditionally stable but they don't require any solution of a linear problem. The choice between the two schemes must be done with the problem one wants to solve in mind. For example, if one is interested in studying the resonances of a particular device, it would be useful to look at a big interval in time, so that all the transient waves can decay and only the resonant ones survive: this would be best simulated with an implicit scheme. An explicit scheme is best suited to study, for example, the propagation of waves into a waveguide or the coupling coefficient of a coupler: the time duration of the simulation is only set by the physical length of the device, not by the accuracy required.

We will only focus on explicit schemes. The most used ones are the *leapfrog timestepping* (already mentioned in Section 3.1) and the *Runge-Kutta* methods.

Leapfrog timestepping it is a second-order method and can be divided in:

1. staggered: if electric and magnetic fields are associated to two different time grids, staggered by half a timestep. So:

$$^{n+1}f = {}^{n}f + \Delta t d_t \, {}^{n+1/2}f; \tag{3.5}$$

2. unstaggered: if electric and magnetic fields are both associated to the same time grid:

$$^{n+1}f = {}^{n-1}f + 2\Delta t d_t \, {}^{n}f. \tag{3.6}$$

Runge-Kutta methods : they use the fields computed at intermediate instants between two consecutive timesteps, in order to increase accuracy, but are computationally more intensive than the leapfrog methods:

1. third order:

$$\begin{aligned} {}^{n+1/3}f &= {}^{n}f + \tfrac{1}{3}\Delta t \, {}^{n}d_t f \\ {}^{n+2/3}f &= {}^{n}f + \tfrac{2}{3}\Delta t \, {}^{n+1/3}d_t f \\ {}^{n+1}f &= {}^{n}f + \tfrac{1}{4}\Delta t \left(3 \, {}^{n+2/3}d_t f + {}^{n}d_t f\right) \end{aligned} \tag{3.7}$$

2. forth order:

$$\begin{aligned} {}^{n+1/2}\tilde{f} &= {}^{n}f + \tfrac{1}{2}\Delta t \, {}^{n}d_t f \\ {}^{n+1/2}\hat{f} &= {}^{n}f + \tfrac{1}{2}\Delta t \, {}^{n+1/2}d_t \tilde{f} \\ {}^{n+1}\bar{f} &= {}^{n}f + \Delta t \, {}^{n+1/2}d_t \hat{f} \\ {}^{n+1}f &= {}^{n}f + \tfrac{1}{6}\Delta t \left({}^{n}d_t f + 2 \, {}^{n+1/2}d_t \tilde{f} + 2 \, {}^{n+1/2}d_t \hat{f} + {}^{n+1}d_t \bar{f}\right) \end{aligned} \tag{3.8}$$

Using the equation (3.4) and the above equations, the amplification factors can be calculated. For example, for the unstaggered leapfrog timestepping (3.6) we can write:

$$^{n+1}f = {}^{n}f + \lambda\Delta t\ {}^{n+1/2}f$$

and

$$^{n-1}f = {}^{n}f - \lambda\Delta t\ {}^{n-1/2}f.$$

Summing each member:

$$\begin{aligned} {}^{n+1}f + {}^{n-1}f &= 2\ {}^{n}f + \lambda\Delta t({}^{n+1/2}f + {}^{n-1/2}f) \\ &= \left[2 + (\lambda\Delta t)^2\right]\ {}^{n}f \end{aligned}$$

therefore:

$$\begin{aligned} \sigma + \frac{1}{\sigma} &= 2 + (\lambda\Delta t)^2 \\ \sigma_{1,2} &= (\lambda\Delta t) \pm \sqrt{(\lambda\Delta t) + 1} \end{aligned}$$

The same procedure can be applied to the other time integration schemes, obtaining:

$$\sigma = \begin{cases} 1 + \frac{1}{2}(\lambda\Delta t)^2 \pm \sqrt{\left[1 + \frac{1}{2}(\lambda\Delta t)^2\right]^2 - 1} & \text{for (3.5)} \\ (\lambda\Delta t) \pm \sqrt{(\lambda\Delta t)^2 + 1} & \text{for (3.6)} \\ 1 + \lambda\Delta t + \frac{1}{2}(\lambda\Delta t)^2 + \frac{1}{6}(\lambda\Delta t)^3 & \text{for (3.7)} \\ 1 + \lambda\Delta t + \frac{1}{2}(\lambda\Delta t)^2 + \frac{1}{6}(\lambda\Delta t)^3 + \frac{1}{24}(\lambda\Delta t)^4 & \text{for (3.8)} \end{cases} \tag{3.9}$$

Stability is obtained if $|\sigma| < 1$. In the hypothesis of $\lambda = \imath\lambda_I$, purely imaginary, so that possible losses, physical or numerical, that could stabilize the algorithm are not taken into account, for the staggered leapfrog algorithm:

$$|\sigma| < 1 \iff \left|1 - \frac{1}{2}\lambda^2\Delta t^2\right| < 1 \iff \lambda\Delta t \in [-2, 2] - \{0\}. \tag{3.10}$$

The method is conditionally stable (see Figure 3.4). For Cartesian grids, the ratio between the maximum timestep and maximum distance between two adjacent gridpoints multiplied by the speed of light is called *Courant factor* S: conditional stability can be written in terms of S, as $S \leq 1$. In Appendix D a geometrical interpretation of the Courant factor is given. The stability criterion in (3.10) is equivalent to the Courant stability, but it is applicable to any kind of grids.

On the other hand, the unstaggered leapfrog algorithm is unconditionally unstable, because $|\sigma| = 1$[3]:

$$|\sigma|^2 = (\lambda_I \Delta t)^2 + \left(1 - (\lambda_I \Delta t)^2\right) = 1 \quad \forall \Delta t.$$

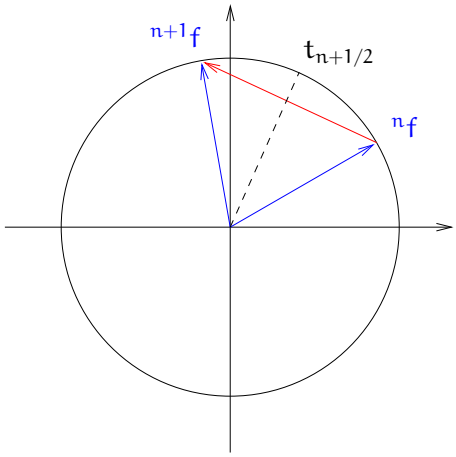


Figure 3.4: *The staggered leapfrog timestepping is conditionally stable. The stability condition $|\sigma| < 1$ can be geometrically interpreted noting that ${}^{n+1}\mathbf{f} = \sigma\, {}^{n}\mathbf{f}$: if $|\sigma| > 1$ than the vectors $\mathbf{f}$ will diverge from the origin indefinitely. The condition $|\sigma| = 1$ is the limit: a simple numerical error (due to the finite precision arithmetics of a computer) can induce instability. In red is the vector $\Delta t\, {}^{n+1/2}\mathbf{f}'$, as computed from the leapfrog timestepping.*

Finally, both the Runge-Kutta methods are conditionally stable. For an imaginary λ, the two leapfrog timestepping schemes are non-dissipative but the Runge-Kutta methods are. All the methods are dispersive.

From the above considerations, one could think that the forth order Runge-Kutta method is the most accurate time integration method for electromagnetic problems. This is correct only if we don't consider the spatial discretization: errors due to time and space discretizations can cancel out. Indeed, for central difference spatial discretization, used in our algorithm, leapfrog timestepping is more accurate than Runge-Kutta.

The staggered leapfrog timestepping in time combined with a staggered central difference in space is non-dissipative and the least dispersive: applied to Cartesian grid, it is called *Yee scheme* [Yee66]. It is the choice for our algorithm.

[3]Theoretically, the condition $|\sigma| = 1$ is the limit between stability and instability: in the world of finite precision arithmetic of computers, though, even very small numerical errors can introduce instability if the equality condition holds. It is usually safer to impose $|\sigma| < 1$, strictly.

3.5 Dispersive and Negative Index Material Example

As an example of a very complex problem that can be modelled with the time-domain version of our algorithm, consider the propagation of light into a Negative Index Material [BMS04].

Negative Index Materials (NIM) are materials with simultaneously negative permettivity and permeability. Veselago [Ves68] predicted that lossless materials with negative ϵ and μ would exhibit unusual properties, such as negative index refraction $n = -\sqrt{\epsilon\mu}$, antiparallel wavevector $\overrightarrow{k}$ and Pointing vector $\overrightarrow{S}$, antiparallel phase $\overrightarrow{v}_p$ and group $\overrightarrow{v}_g$ velocities. Furthermore, if these materials are uniform, $\overrightarrow{k}$, $\overrightarrow{E}$ and $\overrightarrow{H}$ form a left-handed set of vectors: therefore, these materials are also called *left-handed materials* (LHM).

The phenomenon of negative refraction follows these unusual properties. The refraction of a monochromatic wave impinging from a conventional (positive index) material (PIM) with a certain angle θ_i to a negative index material (NIM), will be "the wrong side" with respect to the normal of the interface. In formulas, Snell's law will read, as usual:

$$n_{PIM} \sin\theta_i = n_{NIM} \sin\theta_t$$

but, being $n_{NIM} < 0$, θ_t will be negative, i.e. on "the wrong side" of the normal to the PIM/NIM interface.

Materials with simultaneously negative ϵ and μ over a fixed range of frequencies have been suggested [Pen96] [Pen00] and manufactured in the microwave regime [SPV+00], or just simulated at optical frequencies for a two-dimensional photonic crystal [FES03]. In all the cases, the negative refraction effect is obtained by the so called "metamaterials", i.e. materials with sub-wavelength variations whose, globally, give the desired physical characteristics.

In Figure 3.5, a typical example of application of a negative index material is shown. From a PIM, a point source emits a monochromatic wave. The upper half plane is filled with a NIM. The negative refraction of the emitted wave creates an *image source* in the NIM half plane, symmetric to the source. As long as negative refraction also affects evanescent waves, the focusing is perfect and, in theory, a perfect point-like image can be created. This can not be achieved with conventional lenses, because the finite extension of the lenses themselves and the loss of information brought by evanescent waves, for which conventional lenses fail to work, limit the focusing power of the lens [BW02].

The simulation has been carried out with a dispersive version of our algorithm. Kramers-Konig relations impose that a negative index material must be dispersive in order to not violate causality [Jac99]. Besides, even if only monofrequency sources are present in the simulation, at regime, in the transient time, when sources are "switched on", many frequencies are present: naively setting negative ϵ and μ into the algorithm makes it unstable. A fully dispersive algorithm must be employed.

We have implemented dispersive materials for the two-dimensional time-domain algorithm, with barycentric dual grids (see Section 2.4). The model chosen is the

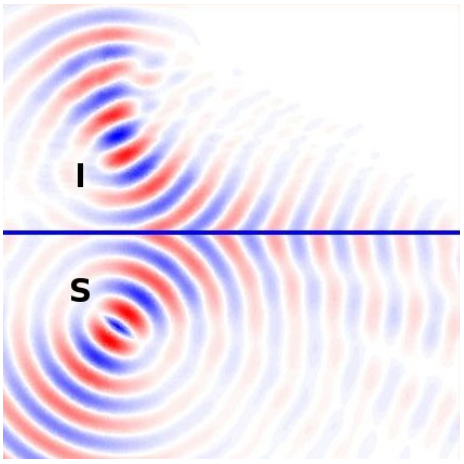


Figure 3.5: *Negative refraction and perfect lens effect. The lower half plane is a PIM and the upper half plane is a NIM, with the same absolute value of the refractive index. Perfect self focusing is achieved. "S" stands for "Source", "I" for "Image". We can see some power reflected by the interface, at a frequency for which PIM and NIM are not matched.*

one described in [Taf00] of Drude media.

The algorithm has been employed to study more in detail the negative refraction phenomenon, with particular attention to a time-limited, i.e. broadband, input field. Each frequency in the input spectrum will see a different refractive index for the NIM, due to its dispersive nature, some will be negatively refracted, some positively and some totally reflected [BMS04]. The spatial spreading of frequencies due to this phenomenon has been studied to be used as an optical multiplexer/demultiplexer [WMKK02].

4

Frequency-Domain

4.1 From Time- to Frequency-Domain

Starting from (3.2), it is interesting to study how the system behaves for a single input frequency, in the time-domain. At first sight, this can appear to be not very useful because the strength of time-domain algorithms is to be able to model the spectral response of a system for a broadband input source. Nonetheless, it is useful, because it represents the first conceptual step toward the formulation of the discretized Maxwell equations in the frequency-domain. It will also lead us to some important observations.

Given a source at the single frequency ω, all the fields will show a dependence in time of the form $e^{\imath\omega t}$:

$$\mathbf{e} = \mathbb{R}\left[\mathbf{e}_{\mathbb{C}} e^{\imath\omega t}\right] \qquad\qquad \mathbf{h} = \mathbb{R}\left[\mathbf{h}_{\mathbb{C}} e^{\imath\omega t}\right].$$

This hypothesis allows us to rewrite (3.2), so to have:

$$\begin{cases} \mathbf{h}_{\mathbb{C}} e^{\imath\omega\Delta t/2} = \mathbf{h}_{\mathbb{C}} e^{-\imath\omega\Delta t/2} - \Delta t \mathbf{R}_e \mathbf{e}_{\mathbb{C}} - \Delta t \mathbf{P}_m \mathbf{h}_{\mathbb{C}} e^{-\imath\omega\Delta t/2} \\ \mathbf{e}_{\mathbb{C}} e^{\imath\omega\Delta t} = \mathbf{e}_{\mathbb{C}} + \Delta t \mathbf{R}_m \mathbf{h}_{\mathbb{C}} e^{\imath\omega\Delta t/2} - \Delta t \mathbf{P}_e \mathbf{e}_{\mathbb{C}} - \Delta t \mathbf{M}_\epsilon \mathbf{j}_{\mathbb{C}}, \end{cases} \tag{4.1}$$

where $\mathbf{R}_e \triangleq \mathbf{M}_\mu \mathbf{R}$, $\mathbf{R}_m \triangleq \mathbf{M}_\epsilon \mathbf{R}^T$ and $\mathbf{P}_e$ and $\mathbf{P}_m$ account for the Ohm losses, as explained in 2.5.

Now, we can explicitly write $\mathbf{h}_{\mathbb{C}}$ from one of the two equations in (4.1) and substitute it in the other. Therefore, we obtain:

$$\begin{cases} \mathbf{h}_{\mathbb{C}} = \underbrace{\left(\mathbf{I} - \mathbf{I} e^{-\imath\omega\Delta t} + \Delta t \mathbf{P}_m e^{-\imath\omega\Delta t}\right)^{-1}}_{\widetilde{\mathbf{D}}} \left(-\Delta t \mathbf{R}_e e^{-\imath\omega\Delta t}\right) \mathbf{e}_{\mathbb{C}} \\ \underbrace{\left(\mathbf{I}\frac{1 - e^{-\imath\omega\Delta t}}{\Delta t} + \mathbf{P}_e e^{-\imath\omega\Delta t} + \Delta t \mathbf{R}_m \widetilde{\mathbf{D}}^{-1} \mathbf{R}_e e^{-\imath\omega\Delta t}\right)}_{\mathbf{D}} \mathbf{e}_{\mathbb{C}} = \mathbf{M}_\epsilon \mathbf{j}_{\mathbb{C}}, \end{cases} \tag{4.2}$$

where $\mathbf{I}$ is the identity matrix.

The matrix $\mathbf{D}$ is worth a further study. For $\Delta t \to 0$:

$$\lim_{\Delta t \to 0} \frac{1 - e^{\imath \omega \Delta t}}{\Delta t} = \lim_{\Delta t \to 0} \underbrace{\frac{1 - \cos \omega \Delta t}{\Delta t}}_{\sim \frac{(\omega \Delta t)^2}{2 \Delta t} \to 0} + \imath \underbrace{\frac{\sin \omega \Delta t}{\Delta t}}_{\to \omega} = \imath \omega$$

and

$$\lim_{\Delta t \to 0} \Delta t \widetilde{\mathbf{D}}^{-1} = \left(\mathbf{I} \frac{1 - e^{-\imath \omega \Delta t}}{\Delta t} + \mathbf{P}_m e^{-\imath \omega \Delta t} \right)^{-1} = (\imath \omega \mathbf{I} + \mathbf{P}_m)^{-1}$$

therefore:

$$\mathbf{D} = \imath \omega \mathbf{I} + \mathbf{P}_e + \mathbf{R}_m \left(\imath \omega \mathbf{I} + \mathbf{P}_m \right)^{-1} \mathbf{R}_e. \tag{4.3}$$

This expression, which is the direct analogous of (3.3) but with sources and in the frequency-domain, has been obtained first discretizing Maxwell equations and then making the timestep Δt go to zero. The same expression can be obtained starting from the frequency-domain form of Maxwell equations and then discretize them only in space: time derivatives, in fact, are not present anymore. In formulas:

$$\begin{cases} \nabla \times \overrightarrow{E} = \imath \omega \mu \overrightarrow{H} + \sigma^* \overrightarrow{H} \\ \nabla \times \overrightarrow{H} = -\imath \omega \epsilon \overrightarrow{E} - \sigma \overrightarrow{E} + \overrightarrow{J} \end{cases}$$

$$\begin{cases} \mathbf{R}\, \mathbf{e} = \imath \omega \mathbf{M}_\mu^{-1} \mathbf{h} + \mathbf{M}_\mu^{-1} \mathbf{P}_m \mathbf{h} \\ \mathbf{R}^T \mathbf{h} = -\imath \omega \mathbf{M}_\epsilon^{-1} \mathbf{e} - \mathbf{M}_\epsilon^{-1} \mathbf{P}_e \mathbf{e} + \mathbf{j} \end{cases}, \tag{4.4}$$

therefore:

$$\begin{cases} \overrightarrow{H} = (\imath \omega \mu + \sigma^*)^{-1} \nabla \times \overrightarrow{E} \\ \nabla \times (\imath \omega \mu + \sigma^*)^{-1} \nabla \times \overrightarrow{E} + (\imath \omega \epsilon + \sigma) \overrightarrow{E} = \overrightarrow{J} \end{cases}$$

$$\begin{cases} \mathbf{h} = (\imath \omega \mathbf{I} + \mathbf{P}_m)^{-1} \mathbf{M}_\mu \mathbf{R}\, \mathbf{e} \\ \underbrace{\left(\mathbf{M}_\epsilon \mathbf{R}^T (\imath \omega \mathbf{I} + \mathbf{P}_m)^{-1} \mathbf{M}_\mu \mathbf{R} + \imath \omega \mathbf{I} + \mathbf{P}_e \right)}_{\mathbf{D}} \mathbf{e} = \mathbf{M}_\epsilon \mathbf{j} \end{cases}, \tag{4.5}$$

where both the continuum version of Maxwell equations and its discrete counterpart are shown for comparison. $\mathbf{D}$ is the discrete form of the rot-rot operator in the Helmholtz equation. Note that its definition in (4.5) is the same as in (4.3): one is obtained by a limiting process from the time-domain, the other by direct discretization of Helmholtz equation in the frequency-domain.

Even if the result is the same, the first procedure allows us to study what happens to the system if the timestep Δt changes, till the limit value of $\Delta t = 0$. We already know from Section 3.4 that the system is conditionally stable: there

is a Δt_{max} so that $\forall \Delta t > \Delta t_{max}$ the system is unstable. From a mathematical point of view, the instability is due to the fact that at least one of the eigenvalues of the matrix $\mathbf{D}$ has modulus greater than one. This can be seen in the following procedure.

Rewrite (3.2) as:

$$^{n+1}\mathbf{v} = {}^{n}\mathbf{v} + \Delta t \mathbf{M}\ {}^{n}\mathbf{v} + {}^{n}\mathbf{u}, \tag{4.6}$$

where:

$$^{n}\mathbf{v} \triangleq \begin{bmatrix} {}^{n}\mathbf{e} \\ {}^{n+1/2}\mathbf{h} \end{bmatrix}$$
$$^{n}\mathbf{u} \triangleq \begin{bmatrix} -\mathbf{M}_\epsilon\ {}^{n}\mathbf{j} \\ \mathbf{M}_\mu\ {}^{n+1/2}\mathbf{m} + \Delta t \mathbf{M}_\mu \mathbf{R} \mathbf{M}_\epsilon\ {}^{n}\mathbf{j} \end{bmatrix}$$

and

$$\mathbf{M} \triangleq \begin{bmatrix} 0 & \mathbf{M}_\epsilon \mathbf{R}^T \\ -\mathbf{M}_\mu \mathbf{R}^T & -\Delta t \mathbf{M}_\mu \mathbf{R} \mathbf{M}_\epsilon \mathbf{R}^T \end{bmatrix}.$$

We can also write (4.6) as follows, which closely resembles the formalism used in [Liu96]:

$$^{n+1}\mathbf{v} = \widetilde{\mathbf{M}}\ {}^{n}\mathbf{v} + {}^{n}\mathbf{u}, \tag{4.7}$$

where $\widetilde{\mathbf{M}} = \mathbf{I} + \Delta t \mathbf{M}$.

In the hypothesis of sources $\mathbf{u}$ with finite amplitude, the equation (4.7) must give solutions with finite amplitude: otherwise, it wouldn't be physically acceptable. Therefore, all the eigenvalues of $\widetilde{\mathbf{M}}$ must lie inside the unitary circle:

$$\forall \lambda_{\widetilde{\mathbf{M}}} : |\lambda| \leq 1. \tag{4.8}$$

In formulas, the eigenvalues λ satisfy:

$$\begin{aligned} \det\left(\widetilde{\mathbf{M}} - \lambda \mathbf{I}\right) &= 0 \\ \begin{vmatrix} \mathbf{I}(1-\lambda) & \Delta t \mathbf{M}_\epsilon \mathbf{R}^T \\ -\mathbf{M}_\mu \mathbf{R} & \mathbf{I}(1-\lambda) - \Delta t^2 \mathbf{M}_\mu \mathbf{R} \mathbf{M}_\epsilon \mathbf{R}^T \end{vmatrix} &= 0 \\ (1-\lambda)^{2N} + (1-\lambda)^{N} \Delta t^2 \det\left(\mathbf{M}_\mu \mathbf{R} \mathbf{M}_\epsilon \mathbf{R}^T\right) &= 0 \end{aligned}$$

where $2N$ is the dimension of the matrix $\mathbf{M}$.

Note that the particular Δt that satisfies the condition in (4.8) is said to satisfy the *Courant condition*. Besides, for $\Delta t = 0$, the eigenvalues are the $2N$ roots of unity, therefore they lie on the unitary circle. Finally, the eigenvalue $\lambda = 1$ is always present (see Figure 4.1).

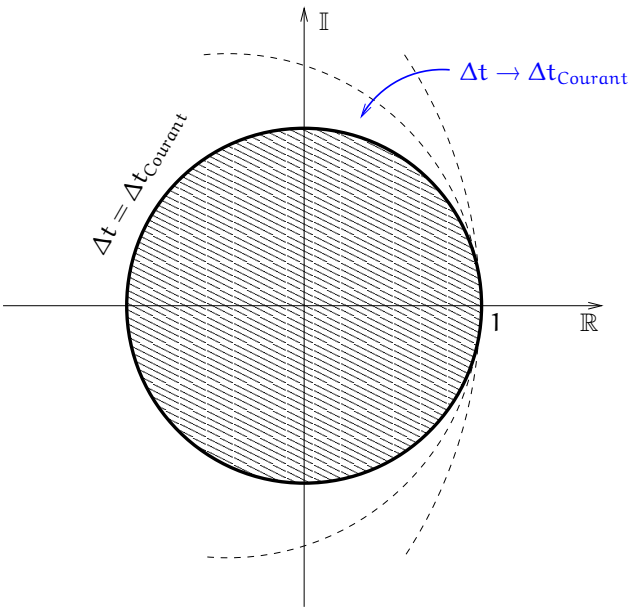


Figure 4.1: *Region of stability of the time-domain algorithm as* Δt *changes. Note that each circle passes through the point* $(1, 0)$.

4.2 Check of the Stationary State

To check (4.1), (3.2) can be used, using a monofrequency source and the following procedure.

Let $\mathbf{v} \triangleq \left[\begin{smallmatrix}\mathbf{e}\\ \mathbf{h}\end{smallmatrix}\right]$ and:

$$\mathbf{v}_1 = \mathbf{v}(t_1)$$
$$\mathbf{v}_2 = \mathbf{v}(t_2)$$

be two values of $\mathbf{v}$ at different instants t_1 and t_2. If the source is monofrequency, even the solution, if the system is linear, has the same frequency dependence, hence it can be written as: $\mathbf{v} = \mathbb{R}[\mathbf{v}_{\mathbb{C}} e^{\imath\omega t}]$. Therefore:

$$\mathbf{v}_1 = \mathbb{R}[\mathbf{v}_{\mathbb{C}}] = \mathbf{v}_{\mathbb{R}} \cos\omega t_1 - \mathbf{v}_{\mathbb{I}} \sin\omega t_1$$
$$\mathbf{v}_2 = \mathbb{I}[\mathbf{v}_{\mathbb{C}}] = \mathbf{v}_{\mathbb{R}} \cos\omega t_2 - \mathbf{v}_{\mathbb{I}} \sin\omega t_2$$

or, in matrix form:

$$\begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{bmatrix} = \underbrace{\begin{bmatrix}\cos\omega t_1 & -\sin\omega t_1\\ \cos\omega t_2 & -\sin\omega t_2\end{bmatrix}}_{\mathbf{L}} \begin{bmatrix}\mathbf{v}_{\mathbb{R}}\\ \mathbf{v}_{\mathbb{I}}\end{bmatrix}.$$

Finally, $\mathbf{v}_{\mathbb{C}}$ can be computed from the two sampled values of $\mathbf{v}$ in the time-domain, by inverting the matrix $\mathbf{L}$:

$$\mathbf{v}_{\mathbb{C}} = \begin{bmatrix}\mathbf{v}_{\mathbb{R}}\\ \mathbf{v}_{\mathbb{I}}\end{bmatrix} = \mathbf{L}^{-1} \begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{bmatrix}.$$

Obviously, t_1 and t_2 must be chosen so that $\mathbf{L}^{-1}$ exists.

Then, $\mathbf{v}_{\mathbb{C}}$ can be substituted in (4.1): the result, which should be zero in the stationary state, is called *residual*. From the simulations, we note that the residual correctly tends to zero as the number of timesteps computed in the time-domain grows (and the source's spectrum becomes more and more similar to a monofrequency source, after the initial transient state[1]). The residual tends to zero more slowly in the PMLs regions: this is expected, as the eigenvectors which represent the modes in the PMLs tends to grow exponentially and the numerical cancellation is slower.

4.3 Matricial Representation

From (4.4), we can pass to the matricial form:

$$\mathbf{M}\,\mathbf{v} = \mathbf{u}, \tag{4.9}$$

[1] Its bandwidth can be significantly reduced by filtering the source with a *raised cosine* function.

where:

$$\mathbf{v} \triangleq \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix}, \quad \mathbf{M} \triangleq \begin{bmatrix} \imath\omega\mathbf{M}_\epsilon^{-1} + \mathbf{M}_\epsilon^{-1}\,\mathbf{P}_e & \mathbf{R}^T \\ \mathbf{R} & -\left(\imath\omega\mathbf{M}_\mu^{-1} + \mathbf{M}_\mu^{-1}\,\mathbf{P}_m\right) \end{bmatrix}, \quad \mathbf{u} \triangleq \begin{bmatrix} \mathbf{j} \\ \mathbf{0} \end{bmatrix}.$$

Let's write (4.9), explicitly showing its frequency dependence:

$$\left(\imath\omega\mathbf{I} + \widetilde{\mathbf{M}}\right)\mathbf{v} = \mathbf{u} \tag{4.10}$$

and step back for a moment to the time-domain:

$$d_t\mathbf{v} = -\widetilde{\mathbf{M}}\,\mathbf{v} + \mathbf{u}. \tag{4.11}$$

(4.11) can be solved with the *state variables* method, i.e. by writing the solution as a linear combination of the eigenvectors $\mathbf{e}_k$ of the matrix $\widetilde{\mathbf{M}}$:

$$\mathbf{v} = \sum_k \alpha_k \mathbf{e}_k. \tag{4.12}$$

Therefore, it reads:

$$\sum_k d_t\alpha_k\mathbf{e}_k = -\sum_k \lambda_k\alpha_k\mathbf{e}_k + \mathbf{u}.$$

Dot-multiplying both members by $\mathbf{e}_k{}^*$ and using the orthonormality of the eigenvectors, we have:

$$d_t\alpha_k = -\lambda_k\alpha_k + u_k,$$

where $u_k = \mathbf{u}\cdot\mathbf{e}_k{}^*$. We can rearrange the terms, multiplied by $e^{\lambda_k t}$, to obtain:

$$d_t\left(\alpha_k e^{\lambda_k t}\right) = u_k e^{\lambda_k t}$$

and finally:

$$\alpha_k = \frac{u_k}{\imath\omega + \lambda_k}\left(e^{\imath\omega t} - e^{-\lambda_k t}\right).$$

For the solution (4.12) to be limited in time, α_k in (4.3) can not diverge with time: therefore, $\mathbb{R}[\lambda_k] \geq 0$. See figure 4.2.

Compare the two figures in 4.3. One represents the region of stability of the time-domain algorithm; the other one, the region of stability of the frequency-domain. One can switch between the two by the transformation $z \mapsto e^{-z}$ and its inverse. Obviously, this is the transformation applied to (3.2) to obtain (4.1).

One last thing is worth noting. The stability problem found in the time-domain is not an issue in the frequency domain: nonetheless, it becomes an *existence* and *well-posedness* problem as the mesh becomes denser and denser. It can be shown [BF04] that for a very dense grid, both in space and in time, the eigenvalues of the frequency-domain system tend to collapse into the origin, causing numerical problems. These numerical problems have a physical explanation. We can find it

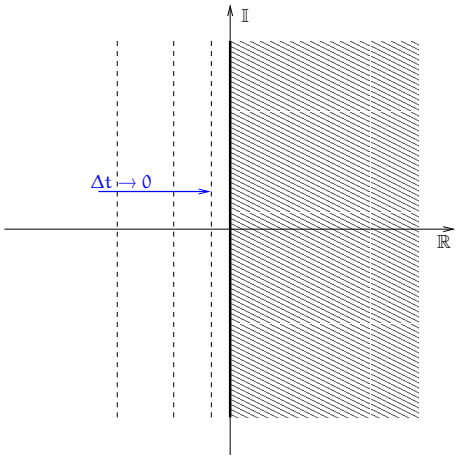


Figure 4.2: *Region of stability of the frequency-domain algorithm.*

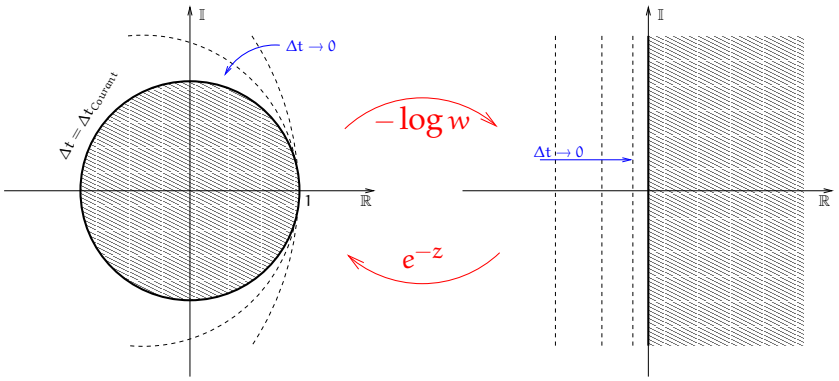


Figure 4.3: *Map from time- to frequency-domain.*

in the *Uniqueness Theorem* [Som98], which states the need of losses for the existence (and uniqueness) of a solution of Maxwell equations in the frequency-domain. The situation can be understood if we think that in a lossless domain (both material and radiation losses being zero), a (monofrequency) sinusoidal source pumping power into it is going to increase the electromagnetic field without limit. We need losses to have physically meaningful results.

To avoid this problem, we have implemented two kind of losses (see Chapter 2):

1. Ohm losses;
2. PMLs, in the U-PML implementation [Taf00], adapted to unstructured grids.

4.4 Solvers

Solving the frequency domain problem in (4.2), has now become finding the solution of a set of linear equations, i.e. a linear algebra problem, of the form:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b}. \tag{4.13}$$

It's well worth studying more in detail the different strategies to solve such problems [Num].

Thank to the formulation used to define the problem, the linear system obtained is *sparse*, in the sense that only a small number of its matrix elements a_{ij} are nonzero. This is a direct consequence of the locality of Maxwell equations (and its discretized form). The value of the fields in a particular gridpoint depends only on the value on adjacent gridpoints, so that a_{ij} will be nonzero only if the nodes (or other geometrical elements) i and j are topologically close to each other. It is wasteful to use general methods of linear algebra on such problems, because most of the $\mathcal{O}\left[N^3\right]$ (with N dimension of the matrix $\mathbf{A}$) arithmetic operations, devoted to solving the set of equations or inverting the matrix, involve zero operands. Furthermore, the dimension of the system matrix can be so large as to tax the available memory space, unfruitfully storing zero elements. Usually, the choice is between speed or memory saving, when looking for a sparse solver routine.

There are two different families of methods to solve sparse linear systems as (4.13): *direct* and *iterative*.

4.4.1 Direct Methods

Direct methods try to factor the system matrix $\mathbf{A}$ into the product of matrices with a particular shape, for which the solution of a linear system is more easily implemented.

The most common direct method, which works both for dense and sparse linear systems, is the *LU decomposition* [Mat, Wik, Num]. It consists in the factorization of

the matrix **A** into the product of two matrices **L** and **U**:

$$\mathbf{A} = \mathbf{L}\,\mathbf{U},$$

where **L** is *lower triangular* and **U** is *upper triangular*. Then, the linear system (4.13) can be written:

$$\mathbf{A}\,\mathbf{x} = (\mathbf{L}\,\mathbf{U})\;\mathbf{x} = \mathbf{L}\;(\mathbf{U}\,\mathbf{x}) = \mathbf{b} \tag{4.14}$$

and solved, by first solving it for the vector **y**:

$$\mathbf{L}\,\mathbf{y} = \mathbf{b} \tag{4.15}$$

and then solving:

$$\mathbf{U}\,\mathbf{x} = \mathbf{y}. \tag{4.16}$$

The advantage of breaking down (4.13) into (4.14) is that (4.15) and (4.16) can be easily solved, thank to the particular element pattern of **L** and **U**. The equation (4.15) can be solved by *forward substitution* while (4.16) by *backward substitution*.

Another great advantage of direct methods is that, once the factorization of **A** has been worked out, the system (4.13) can be solved very efficiently for as many right-hand sides as we want. This is particularly useful in our algorithm, where the matrix **A** is determined by the particular physics and geometry of the problem to study: once factored, we can study the response of the device to as many input waves as we want (represented by the array **b**) with very little computational effort.

There are fundamentally two different algorithms to work out the factorization: the Doolittle [Wik] and the Crout [Num] algorithms. Leaving the details to the cited references, it's worth noting that LU factorization is neither always possible nor unique:

- from a mathematical point of view, all the principal minors of **A** must be nonzero, in which case **A** is invertible[2]. The factorization is not unique, unless we require that the diagonal elements of **L** are ones;
- from a numerical point of view, we have two problems:

  stability : backward substitution can be numerically unstable if the choice of a pivot element is not done properly;

  memory : if **A** is sparse, **L** and **U** are not. This is why, for sparse matrices, there are routines that compute an *approximate* LU factorization, with the additional constraint of keeping the number of nonzero elements of the factors small. These routines are called *Incomplete LU decomposition* routines, ILU for short [NAG]: they must be used in conjunction with an iterative method to find the required solution, though.

[2]Indeed, finding the inverse $\mathbf{A}^{-1}$ once its factors **L** and **U** are known is trivial: just solve the linear system (4.13) N times, with different **b** of the form $\mathbf{b} = [0, \ldots, 1, \ldots, 0]$.

One of the most common and efficient implementation of the LU decomposition is the UMFPACK set of routines [UMF]. It is based on the Unsymmetric-pattern Multi-Frontal method [ADD96] and it works for sparse unsymmetric matrices, as the ones met in our problem, both real and complex. Before factoring the matrix, a permutation of rows is done so as to reduce the fill-in of the factor matrices: qualitatively, this procedure tries its best to condense all the elements of the matrix near its main diagonal, by swapping rows and columns. From a geometrical point of view, as long as each row represents an edge in the primal mesh, the procedure consists in renumbering the edges of the mesh so that adjacent edges have adjacent numbers. This procedure, which improves locality of variables in the computer's memory, increases the computational speed also in the time-domain, by a better use of the processor's cache.

The main disadvantage of direct methods is their memory requirements. As said before, if $\mathbf{A}$ is sparse, $\mathbf{L}$ and $\mathbf{U}$ are not and, for large linear systems, the memory required to store them can be prohibitive. Experience teaches us that two-dimensional problems are well suited to be solved directly, but three-dimensional ones are not. Even if in industrial environment direct methods are the primary choice because of their greater stability, more memory efficient methods must be looked for.

4.4.2 Iterative Methods

The term "iterative methods" refers to a wide range of techniques that use successive approximations to obtain more accurate solutions to a linear system at each step. There are two fundamental families of iterative methods [BBC+94].

Stationary methods Older, simpler but not very effective, they can be expressed in the simple form:

$$\mathbf{E}\ {}^{n+1}\mathbf{x} = \mathbf{F}\ {}^{n}\mathbf{x} + \mathbf{b}, \tag{4.17}$$

where $\mathbf{A} = \mathbf{E} - \mathbf{F}$. $\mathbf{E}$ is easily invertible, while $\mathbf{F}$ is called *remainder*. $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{b}$ do not depend on n. Methods belonging to this family are: the *Jacobi method*, the *Gauss-Seidel method*, the *Successive Over-relaxation method* (SOR) and the *Symmetric Successive Over-relaxation method* (SSOR) [BBC+94].

It can be noted that (4.17) closely resembles (4.7): time-domain solution of Maxwell equations with monofrequency sources is a sort of relaxation method of the corresponding frequency-domain problem.

Non-stationary methods Newer, more complex to understand and interpret, but very effective, they differ from the stationary methods because the computation involves information that changes at each iteration. *Conjugate Gradient methods* (CG), with all the existing variants Conjugate Gradient Squared (CGS), Biconjugate Gradient (BiCG), Biconjugate Gradient Stabilized (BiCGStab), Conjugate Gradient for Normal Equations (CGNE) and *Minimum Residual methods*

(MINRES), with all the existing variants, Symmetric LQ (SYMMLQ), Generalized Minimum Residual (GMRES) and Quasi-Minimal Residual (QMR), to name just a few, all belong to this family [BBC+94].

Each methods is only applicable or performs effectively only on a particular class of matrices.

Jacobi method works reasonably well only for strongly dominant matrices. It's more often used as a preconditioner for a more effective iterative method.

Gauss-Seidel, SOR and SSOR represent an improvement over the Jacobi method, but they are overcome by non-stationary methods.

CG works for symmetric positive definite systems. As very clearly explained in [She94], the method is based on the idea of minimizing the function:

$$f(x) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\,\mathbf{x} - \mathbf{x}^T \mathbf{b}$$

f is minimized when its gradient:

$$\nabla f = \mathbf{A}\,\mathbf{x} - \mathbf{b}$$

is zero, which is equivalent to (4.13). The minimization is carried out by generating a succession of search directions $\mathbf{p}_k$ and improved minimizers $\mathbf{x}_k$. At each iteration, a quantity α_k is found that minimizes $f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$ and $\mathbf{x}_{k+1}$ is set to the new point $\mathbf{x}_k + \alpha_k \mathbf{p}_k$. $\mathbf{p}_k$ and $\mathbf{x}_k$ are chosen in such a way that $\mathbf{x}_{k+1}$ minimizes f over the whole vector space of directions already taken, $\{p_1, \ldots, p_k\}$. After N iterations the procedure gives the minimizer of the whole vector space, i.e. the solution to (4.13).

While this is theoretically true, in practical implementation on a computer, round-off errors tend to disturb the convergence which is usually achieved with more than N iterations. Practically, the convergence is slower, the bigger is the *condition number* of the matrix [Num].

CGS, BiCG and BiCGStab are modified versions of the Conjugate Gradient to work with non-symmetric matrices.

GMRES is applicable to non-symmetric matrices. Its complexity grows linearly with the iteration number, unless a "restarted" version is used.

QMR is applicable to non-symmetric matrices and converges more smoothly than the Biconjugate Gradient. It has the big advantage that it improves the residual at each iteration, even of a small amount, when other methods stagnate or diverge.

4.4.3 Preconditioning

The rate of convergence of iterative methods depends greatly on the spectrum of the coefficient matrix. Hence, to improve convergence, a second matrix, called *preconditioner*, is used to transform the coefficient matrix into one with a favorable spectrum: the construction and application cost of applying the preconditioner is usually overcome by the improved convergence. In formulas, instead of (4.13), one tries to solve:

$$\mathbf{M}^{-1}\,\mathbf{A}\,\mathbf{x} = \mathbf{M}^{-1}\,\mathbf{b}, \tag{4.18}$$

which has the same solution of the original system, but the new coefficient matrix $\mathbf{M}^{-1}\,\mathbf{A}$ may have a favorable spectrum. $\mathbf{M}$ is the preconditioner.

Preconditioning can be split in left and right, to preserve the symmetry of the coefficient matrix [Saa00].

$$\mathbf{M}_1^{-1}\,\mathbf{A}\,\mathbf{M}_2^{-1}\;(\mathbf{P}_2\mathbf{x}) = \mathbf{M}_1^{-1}\,\mathbf{j}.$$

There are different possible preconditioners [BBC+94]:

1. Jacobi: $\mathbf{M} = \mathrm{diag}\,(\mathbf{A})$.
2. SSOR: write $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$ with $\mathbf{L}$ strictly lower triangular, $\mathbf{U}$ strictly upper triangular and $\mathbf{D}$ diagonal; then $\mathbf{M}_1 = \mathbf{D}\,\left(\mathbf{I} + w\mathbf{D}^{-1}\,\mathbf{L}\right)$ and $\mathbf{M}_2 = \mathbf{I} + w\mathbf{D}^{-1}\,\mathbf{U}$ with $0 < w < 2$. The optimal value of w can reduce the number of iterations to a lower order.
3. Incomplete LU
4. Incomplete Cholesky: valid only for symmetric matrices.
5. Divergenceless preconditioner [HSZH97]: this preconditioner is the matricial version of the equation $\nabla \cdot \overrightarrow{\mathrm{D}} = 0$. Discretized, the equation reads:

   $$\mathbf{D}\,\mathbf{d} = \mathbf{D}\,\mathbf{M}_\epsilon^{-1}\,\mathbf{e} = \mathbf{P}\,\mathbf{e} = 0,$$

   with $\mathbf{D}$ discrete divergence. The linear system is equivalent to:

   $$\underbrace{(\mathbf{A} + \mathbf{P})}_{\widetilde{\mathrm{A}}}\,\mathbf{x} = \mathbf{b} \tag{4.19}$$

   because $\mathbf{P}\,\mathbf{x} = 0$.

   Highlighting the preconditioner:

   $$\left(\mathbf{I} + \mathbf{P}\,\mathbf{A}^{-1}\right)\,\mathbf{A}\,\mathbf{x} = \left(\mathbf{I} + \mathbf{P}\,\mathbf{A}^{-1}\right)\,\mathbf{b} = \mathbf{b} + \mathbf{P}\,\mathbf{x} = \mathbf{b}, \tag{4.20}$$

   where the preconditioner is $\mathbf{I} + \mathbf{P}\,\mathbf{M}^{-1}$.

   The matrix $\widetilde{\mathbf{A}}$ has better eigenvalues than $\mathbf{A}$.

   Note that this preconditioner can only be useful in three-dimensional simulations: in two-dimensions $\nabla \cdot \overrightarrow{\mathrm{D}} \equiv 0$, that is $\mathbf{D} \equiv 0$.

4.4.4 Looking for the Best Methods

All the techniques shown above are not equally effective in solving our specific electromagnetic problem, as formulated in (4.9) and (4.5).

To choose the best one, we need to examine more in detail the mathematical properties of the system matrix we need to solve.

Symmetry

A symmetric system matrix has several advantages over a non symmetric one [SSW02, SW98]:

- it uses less memory, because you only need to store half of it;
- iterative solvers are more efficient on symmetric matrices;
- without sources the problem (4.9) becomes a generalized eigenvalue problem, with all the eigenvalues real and nonnegative, and a lot of efficient routines to find them exist.

For our specific problem, making the system matrix symmetric depends on the choice of the dual grid.

For the sake of clarity, recall (4.5):

$$\mathbf{D}\,\mathbf{e} = \mathbf{M}_\epsilon\,\mathbf{j} \tag{4.21}$$

with $\mathbf{D} = \mathbf{A} - \omega^2\mathbf{I}$ and $\mathbf{A} = \mathbf{M}_\epsilon\,\mathbf{R}^T\,\mathbf{M}_\mu\,\mathbf{R}$ and $\mathbf{e}$ and $\mathbf{j}$ accordingly: for the following reasoning, only the shape of $\mathbf{D}$ is important.

Note that $\mathbf{D}$ is symmetric if and only if $\mathbf{A}$ is symmetric. Let $\mathbf{A} = \mathbf{M}_\epsilon\,\widetilde{\mathbf{A}}$, with $\widetilde{\mathbf{A}} = \mathbf{R}^T\,\mathbf{M}_\mu\,\mathbf{R}$.

Voronoï dual grid With this orthogonal dual grid, described in 2.2:

- $\widetilde{\mathbf{A}}$ is symmetric because $\mathbf{M}_\mu$ is symmetric;
- $\mathbf{A}$ is *not* symmetric because $\mathbf{M}_\epsilon\,\widetilde{\mathbf{A}} \neq \widetilde{\mathbf{A}}\,\mathbf{M}_\epsilon$: $\mathbf{M}_\epsilon$ and $\widetilde{\mathbf{A}}$ does not commute.

We can make the system 4.21 symmetric, by letting $\mathbf{M} = \mathbf{M}_\epsilon^{-1}\,\mathbf{D} = \mathbf{A} - \omega^2\mathbf{M}_\epsilon^{-1}$ and $\mathbf{A} = \mathbf{R}^T\,\mathbf{M}_\mu\,\mathbf{R}$.

Note that we can always have symmetric material matrices for orthogonal grids [SW98].

Poincaré and barycentric dual grids For non-orthogonal grids, "$\overrightarrow{D}^i$ and $\overrightarrow{E}_i$ are not collinear and all the three contravariant components are needed to calculate one covariant component $\overrightarrow{E}_i$" [SW98]. In other words, there is coupling between neighboring components and material matrices are not diagonal.

Different interpolation schemes gives different system matrices:

- Gedney's interpolation scheme [Taf00, page 511]: it gives a non-symmetric matrix;
- Schuhmann's interpolation scheme:
  1. for coordinate grids [SW98]:
     (a) using primary grid edges;
     (b) using dual grid edges;
  2. for unstructured triangular grids [SSW02], using Whitney forms;

  they all give symmetric matrices;
- the "Finite Element way" using Whitney forms (weak form representation) [BK00, SSW02]: it gives a symmetric matrix[3].

The interpolation schemes described in 2.3 and 2.4 both give non-symmetric matrices.

Indefiniteness

Indefinite matrices are matrices that have both negative and positive eigenvalues. Positive definite matrices have all positive eigenvalues. Negative definite, all negative.

Conjugate gradient methods work well with positive definite matrices [She94], but they usually fail with indefinite matrices.

The system matrix $\mathbf{D}$, as defined in the previous subsections, is positive definite if and only if:

$$\forall \mathbf{x}: \quad \mathbf{x}^T \mathbf{D} \mathbf{x} = \mathbf{x}^T \left(\mathbf{A} - \omega^2 \mathbf{I} \right) \mathbf{x} = \mathbf{x}^T \mathbf{A} \mathbf{x} - \omega^2 \left\| \mathbf{x} \right\|^2 > 0$$

Let's try to answer to the following questions.

1. is $\mathbf{A}$ positive definite?
   (a) $\mathbf{M}_\epsilon$ and $\mathbf{M}_\mu$ are positive definite for stability [CMP04, SSW02];
   (b) for any matrix $\mathbf{B}$, $\mathbf{B}^T \mathbf{B}$ is positive definite: in fact,

   $$\forall \mathbf{x}, \mathbf{x}^T \left(\mathbf{B}^T \mathbf{B} \right) \mathbf{x} = \left(\mathbf{B} \mathbf{x} \right)^T \left(\mathbf{B} \mathbf{x} \right) = \left\| \mathbf{B} \mathbf{x} \right\|^2 > 0;$$

   (c) if $\mathbf{A}$ and $\mathbf{B}$ are positive definite, then $\mathbf{C} = \mathbf{A} \mathbf{B}$ is positive definite.

   Therefore, if we write:

$$\begin{aligned} \mathbf{A} &= \mathbf{M}_\epsilon \mathbf{R}^T \left(\mathbf{M}_\mu^{1/2} \mathbf{M}_\mu^{1/2} \right) \mathbf{R} \\ &= \mathbf{M}_\epsilon \left(\left(\mathbf{M}_\mu^{1/2} \right)^T \mathbf{R} \right)^T \left(\mathbf{M}_\mu^{1/2} \mathbf{R} \right) \end{aligned}$$

[3] Note that Whitney forms can be used for both Poincaré and barycentric dual grids. In fact, they are used to interpolate $\mathbf{e}$ and $\mathbf{b}$, which are defined on orthogonal grids anyhow. What is different between Poincaré and barycentric dual grids are only the material matrices.

and if $\mathbf{M}_\mu^{1/2}$ is symmetric (for example, diagonal, as in the Voronoï dual grids), then $\mathbf{A}$ is positive definite.

2. is $\mathbf{D}$ positive definite? Only for certain values of ω, smaller than $\bar{\omega}$:

$$\omega \leq \bar{\omega} = \left(\frac{\mathbf{x}^T \mathbf{A}\, \mathbf{x}}{\|\mathbf{x}\|^2} \right)^{1/2}.$$

At high frequencies the system is indefinite and most of the conjugate gradient methods won't work. This is the reason why solving low frequency or, at limit, stationary problems (like magnetostatic), is computationally much easier than problems at optical frequencies.

As an example, the quadratic form in the very simple case of a 2×2 linear system is shown in Figure 4.4 [She94]. We can appreciate the geometric explanation of indefiniteness of a matrix, as the presence of at least one non-positive eigenvalue. In the figure, this is represented by a direction along which the concavity of the quadratic form is negative: the three-dimensional plot, in this case, becomes a saddle, instead of a paraboloid.

The more different the eigenvalues of $\mathbf{A}$ are, the more elliptic the section of the paraboloid is, the easiest is that the quadratic form associated with $\mathbf{D}$ becomes indefinite, subtracting $\omega^2\mathbf{I}$. The best condition would be to have all the eigenvalues of $\mathbf{A}$ with the same modulus: this is what a preconditioner usually tries to do.

Another problem of indefinite matrices is that ILU preconditioning is unstable [Ben02]: other preconditioners, usually slower, but stable, must be used.

Hermitianity

The matrix from the frequency-domain problem can not be hermitian, because all hermitian matrices have real values on the diagonal, while we have $\mathbf{M}_\epsilon^{-1}$ on the diagonal, which is complex if PMLs are present (4.5).

Nevertheless, if we build the matrix so that it is symmetric, it becomes:

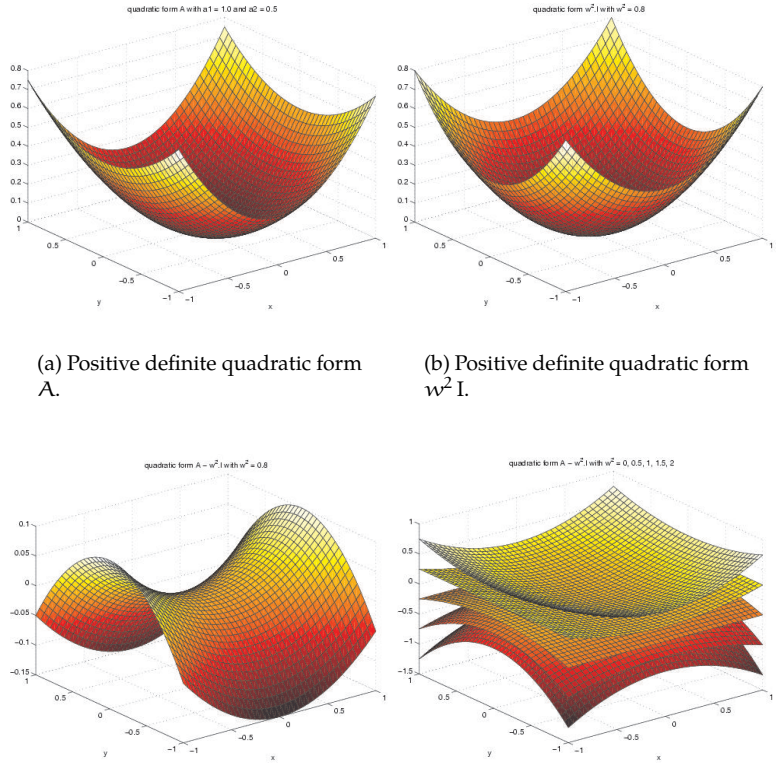
- real, if no PMLs are present, so that all its eigenvalues are real;
- complex, if PMLs are present, so that some of the eigenvalues are complex.

Some Results

Here are collected some results for different kind of matrices arising from the frequency-domain problem. The problems differ for the choice of the parameters listed in Table 4.1:

In Figure 4.5, the eigenvalues of the system matrix are plotted [4].

[4] Only the eigenvalues with maximum and minimum real and imaginary part are actually plotted: finding them all would have been computationally too expensive.



(a) Positive definite quadratic form A.

(b) Positive definite quadratic form w^2 I.

(c) Indefinite quadratic form $A - w^2$ I.

(d) From positive definite to indefinite changing w.

Figure 4.4: *Quadratic forms for a simple 2×2 linear system.*

property	possible values
primal grid	unstructured, structured
λ	$\lambda_0 = 1.55\mu m$, $2\lambda_0$
PML	present, not present

Table 4.1: *Parameters used to test different direct and iterative solvers.*

From Figure 4.5(a) we can note that the eigenvalues of the system matrix tend to spread as the frequency of the problem grows. In fact, for $\lambda = 2\lambda_0$ the eigenvalues are much closer to each others. As a consequence, the problem is "less indefinite", if not positive definite, and iterative methods perform better.

From Figure 4.5(b), we can note that changing the topology of the mesh does not change drastically the position of the eigenvalues. We deduce that, from a stability and ease of solving point of view, structured grids and unstructured grids are similar. The advantage of structured grids is, obviously, the bigger flexibility of the mesh itself. Eigenvalues of the structured grid are more clustered for this reason.

Finally, Figure 4.5(c) shows that the eigenvalues for a problem without PMLs are all real and positive, as expected.

The algorithm has been applied to both two-dimensional and three-dimensional problems. As said before, most of two-dimensional problems can be extremely efficiently solved with direct methods. If iterative methods are applied, most of them tend to give decent performances (not comparable to direct methods, though): only few do not converge to the exact result.

For three-dimensional problems, the situation is drastically different. For the much increased complexity of the problem (more non-zero elements fraction, bigger matrices, divergenceless property of the magnetic field not automatically satisfied, more challenging preconditioners), only very few iterative methods seem to converge. In particular, many experiments have shown that the most effective iterative solver is the QMR with Jacobi preconditioning.

QMR is very effective for non-Hermitian sparse linear systems [FN91, FN94, FG92, KMR01], so for the kind of problem we face in our formulation;

Jacobi preconditioning is effective for strongly diagonal matrices because it uses as a preconditioner $\mathbf{P}$ just the diagonal of the system matrix.

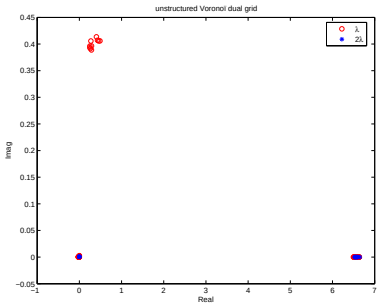
It doesn't achieve good convergence speed, but it's definitely stable and it is well defined also from strongly indefinite matrices, while others preconditioners (like the ILU) are not.

4.5 Examples

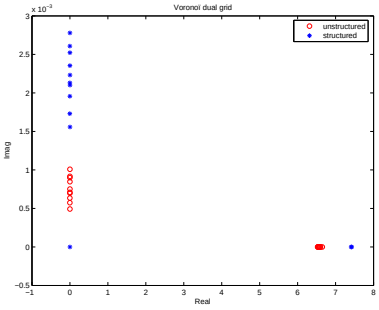
4.5.1 2-D

Photonic crystals are a typical example of a two-dimensional problem in which the present method is particularly effective. The particular circular shape of the scatterers of the photonic crystal used enhances the advantage of having an unstructured grid over a structured one [CW91].

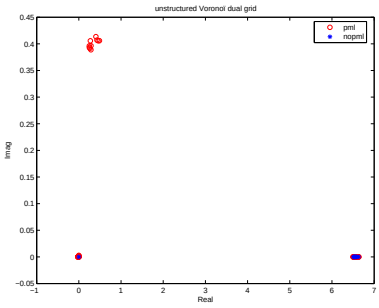
Figure 4.6(a) shows the mesh for a two-dimensional photonic crystal channel. The dielectric scatterers are air holes in a dielectric substrate (refractive index $n =$



(a) Larger λ means lower frequency and a simpler problem to solve: the eigenvalues are closer to each other.



(b) Structured and unstructured grids give approximately the same eigenvalues, but for unstructured grid they are more clustered.



(c) Without PMLs all the eigenvalues are real.

Figure 4.5: *Eigenvalues of the system matrix, for different possible parameters, listed in Table 4.1.*

3.0), with lattice period $\Lambda = 0.5\mu m$ and radius $R = 0.35\Lambda$. Only the TE polarization is studied, for which the photonic crystal presents a bandgap[5].

Figure 4.6(b) shows the system matrix obtained for the frequency domain problem, for a Voronoï dual grid. We can note that it is sparse, with a symmetric non-zero pattern and strongly diagonal. Its dimension is 49795×49795, with 346532 non-zero elements: only 0.1398% of the elements are non-zero. Its sparsity pattern is a direct indication of the node numbering used by the meshing software used to generate the mesh [She].

Figure 4.6(c) and Figure 4.6(d) show the convergence of two iterative methods used to solve the problem: the BiCGStab and the GMRES, as implemented in the PETSc library [SBK+]. The stopping criterion is a value of the residual below 10^{-7}. We can note that GMRES achieves convergence within 4000 steps (a number of iterations about 1.15% of the dimension of the system matrix!) with the Eisenstat [SBK+] and the ILU preconditioning: the convergence is fairly exponential in these cases. For other preconditioners, the convergence is not so rapid, and sub-exponential. Without preconditioning the algorithm tends to stagnate. The BiCGStab, on the other hand, achieves convergence faster (after about 3000 steps, with the same preconditioners), but it's much more irregular.

Note that the memory requirements for same problem allow to solve it using direct methods [UMF]. In this case, the performances of the algorithm are definitely superior: the direct solver gives an exact[6] solution in few seconds, compared to few minutes of the iterative solver.

Another interesting example is the study of a photonic crystal Y-junction: see Figure 4.7. The lattice is triangular with a period $\Lambda = 430nm$, made of holes of radius $R = 0.30\Lambda$ in a substrate of GaAs/AlGaAs. The big problem of these devices is that for the classical triangular lattice of air holes on a silicon substrate, bends are usually lossy. Light must then be "helped" to bend by arranging the holes in a proper way. Thank to its high speed when a direct solver is chosen, the two-dimensional algorithm has been employed with the help of the commercial optimizer Kallistos [Kal], to verify the best position for the scatterers and achieve the highest power transmission[7]. The simulations, even if two-dimensional, have been used to model a real three-dimensional structure, using the *effective index method*. The complete three-dimensional structure has also been built, by the University of St. Andrews, and a very promising 75% transmission has been measured over $110nm$ at $\lambda = 1.55\mu m$ [AWDK05].

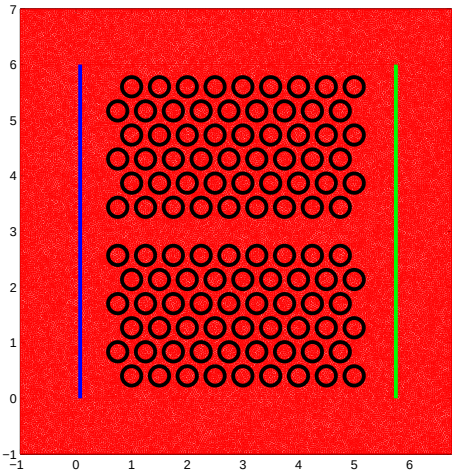
4.5.2 3-D

Two examples of three-dimensional devices are reported here.

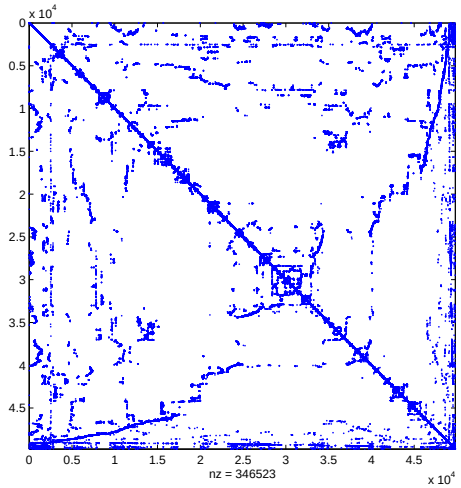
[5]The value of the refractive index have no specific meaning: it has been chosen so that the photonic crystal has a bandgap for the given λ, Λ and R. Studying a highly resonant frequency – like one in the bandgap – is a tougher problem than a non-resonant one.

[6]Exact, within the numerical precision of the computer used.

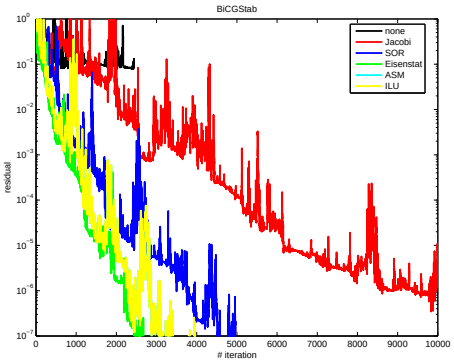
[7]The first optimization has been carried out with FIMMWAVE [FIMb] and Kallistos [Kal].



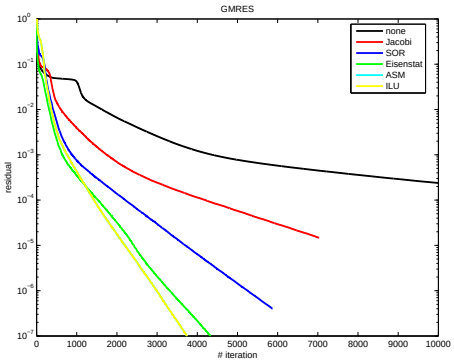
(a) Primary mesh. In black are the circular scatterers, in blue, on the left, a plane wave source at a fixed λ and in green, on the right, a sensor to collect the output fields.



(b) System matrix. Only 0.1398% of the elements are non-zero.

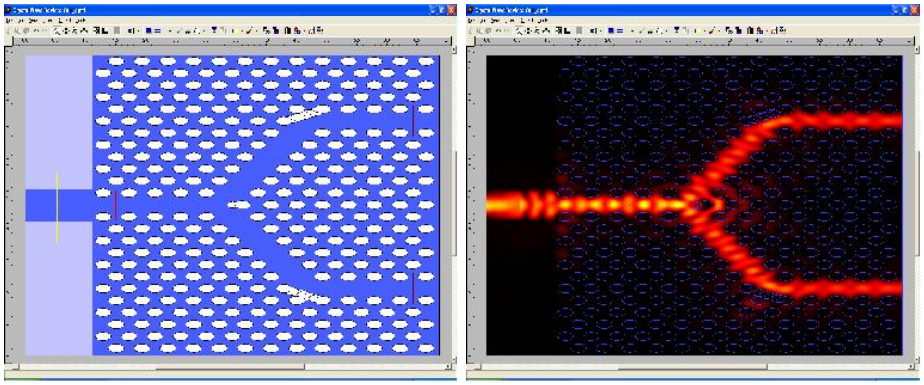


(c) Convergence of the BiCGStab algorithm to the exact solution for different preconditioners. The best result is achieved with the Eisenstat preconditioner, a particular type of SOR preconditioner [SBK+].



(d) Convergence of the GMRES algorithm to the exact solution for different preconditioners. Note that the convergence is much smoother than for the BiCGStab.

Figure 4.6: *Two-dimensional example: photonic crystal channel.*



(a) Optimized position of the scatterers. The lattice period is $\Lambda = 430\text{nm}$ and the radius of the scatterers is $R = 0.30\Lambda$.

(b) Power in the device. Note the "steering" effect of the holes in the Y-junction bends.

Figure 4.7: *Two-dimensional example: a photonic crystal Y-junction.*

Metallic Waveguide

First, the propagation in a metallic waveguide is studied. This device, extremely simple from a theoretical point of view, permits to compare the analytical solution with the computed one and have an exact reference.

The waveguide, shown in Figure 4.8, is a rectangular waveguide, with cross section of $\sqrt{3} \times 1\,\mu\text{m}^2$ and a length of $6\,\mu\text{m}$, full of air. The working wavelength is $\lambda = \sqrt{3}\,\mu\text{m}$, for which the waveguide is monomodal. The metallic walls are perfect and PMLs are present only at the beginning and at the end, to simulate an infinitely long waveguide. The computational domain correspond to the waveguide itself. The fundamental mode $\text{TE}_{1,0}$, known from theory, is excited from one end.

After the discretization of Figure 4.9(a), the system matrix dimension is 174839×174839, with 2005449 non-zero elements (see Figure 4.9(b)). The "block-diagonal" non-zero pattern is easily explained by the three-dimensional mesh used and the numeration of nodes chosen: a two-dimensional extruded mesh, already mentioned in Section 1.2.2. The numbering of nodes is fixed by the meshing software [She] in each "floor", and this explains the internal nonzero pattern of each block in the matrix, but it's ordered between one "floor" and the next: this explains the block diagonal pattern. In a sense, each single layer is a block of the matrix, plus some entries that are the edges linking the "floors" each others.

Figure 4.9(c) shows the convergence of the GMRES algorithm with two possible preconditioners: Jacobi and ILU. Neither of them give good results: after ten thousand steps, the residual is still above 10^{-5}. A better convergence, even if more irregular, is given by the QMR algorithm with Jacobi preconditioner, shown in Fig-

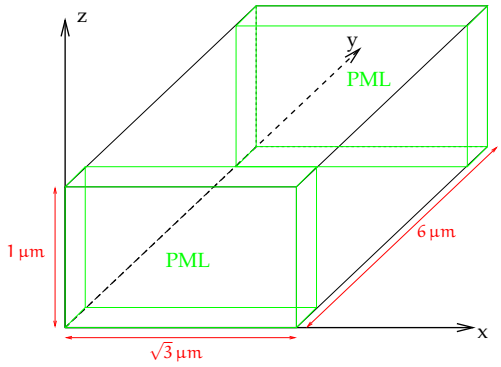


Figure 4.8: *Three-dimensional example: geometry of a metallic waveguide.*

ure 4.9(d). It achieves the same residual of the GMRES, but in one tenth of the steps.

The solution obtained is shown and compared to the theoretical one in Figure 4.10(a) and Figure 4.10(b). The standard deviation σ, defined as the square root of the mean value of the squared distance between theoretical and computed values, sampled in N points, is never above 1%. In formulas:

$$\sigma = \sqrt{\frac{1}{N}\sum_{N}\left(\left|E_z^{\text{computed}}\right| - \left|E_z^{\text{theoretical}}\right|\right)^2} \leq 1\% \tag{4.22}$$

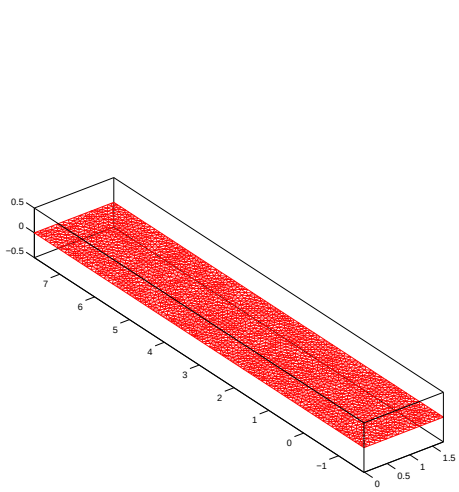
Planar Photonic Crystal Channel

A more challenging example is shown in Figure 4.11. It is the simulation of a Gaussian pulse propagating inside a three-dimensional photonic crystal channel.

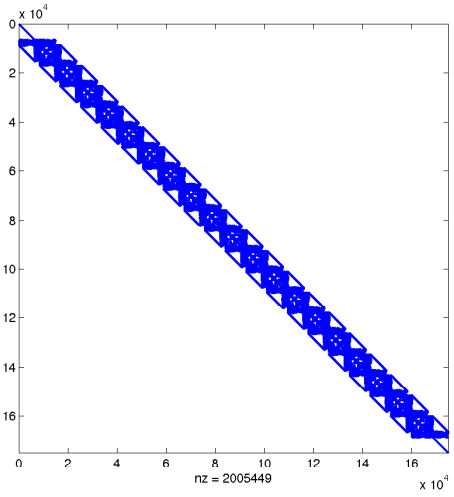
The photonic crystal is made of a 1 µm thick dielectric slab ($n = 3.0$) with 1 µm deep air holes. A line defect uses the bandgap of the structure to guide the light through the device.

The absolute value of the z-component of electric field computed is shown in Figure 4.11(c).

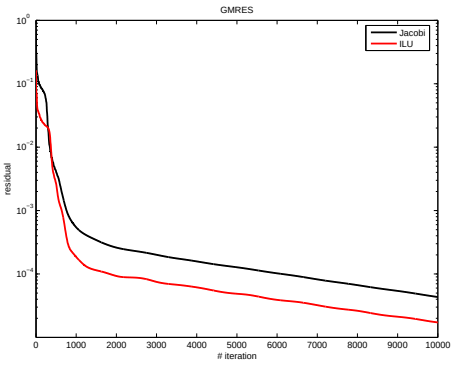
Figure 4.11(a) shows the system matrix: again, the numbering of each layer is visible. Note that the non-zero elements are 8 millions: this is a device by far bigger than any commercial software is able to simulate. Thanks to the present formulation, the unstructured grid and an efficient iterative method, the solution is found after 5000 steps, with a precision below 10^{-6}. The whole simulation took few hours over a Pentium III computer.



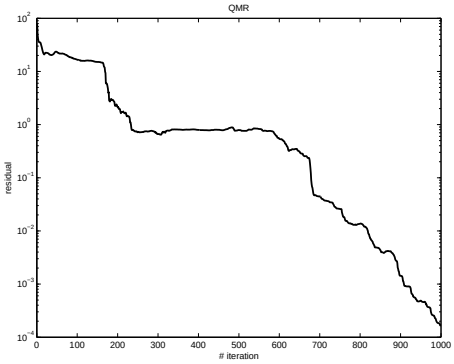
(a) Example of a mesh layer in the extruded three-dimensional grid.



(b) System matrix.

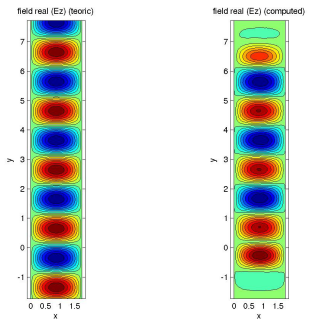


(c) GMRES convergence.

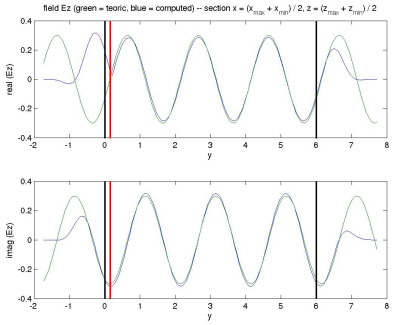


(d) QMR with Jacobi preconditioner convergence.

Figure 4.9: *Three-dimensional example: the metallic waveguide.*



(a) $\mathbb{R}[E_z]$ in a z-constant section: theoretical and computed.



(b) $\mathbb{R}[E_z]$ in a xz-constant section: theoretical and computed. Black lines shows the PMLs; the red line, the $TE_{1,0}$ mode source. Comparison is meaninful only outside the PMLs.

Figure 4.10: *Three-dimensional example: resulting fields for the metallic waveguide.*

4.6 Validation of the method

Validation of the method has been worked out by comparison with the commercial software FIMMPROP [FIMa] by Photon Design [Pho]: the analysis is particular useful because FIMMPROP uses a completely different algorithm, based on the Eigenmode Expansion technique.

The choice of the problems for the validation has been guided by the will to test situations where our algorithm has advantages over the Eigenmode Expansion technique. For this reason, photonic crystals have been chosen.

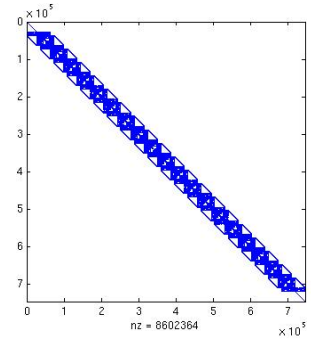
The validation is done by the study of three problems:

free-space propagation : this problem has been used as a "calibration" test, to have a feel of convergence and accuracy as well as of the normalization constants between the two algorithms;

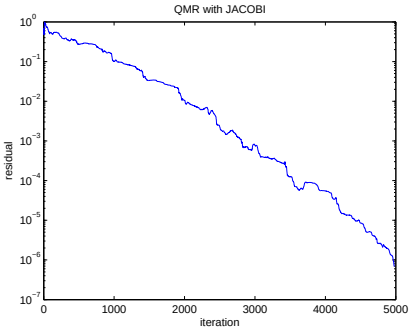
single scatterer : this problem has been used to highlight possible advantages of the present method in case of circular scatterers, where staircase approximation (done by the Eigenmode Expansion technique) can lead to bad performances;

photonic crystal channel : this is a complete test, for which the Eigenmode Expansion technique is not the best solution.

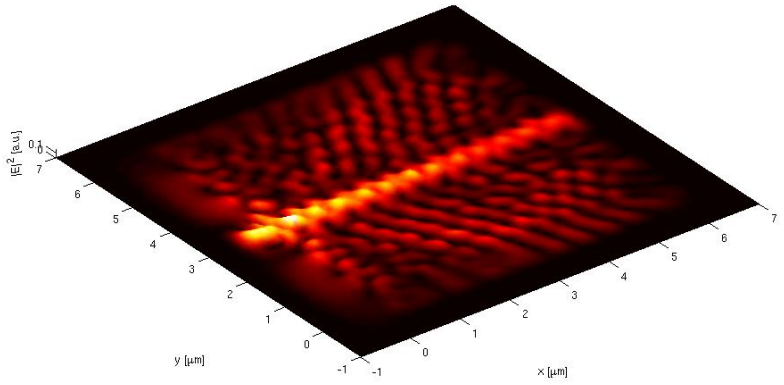
Figure 4.12, Figure 4.13 and Figure 4.14 show the domain of the three validation tests. The yellow line on the left-hand side is always the input, which is a Gaussian



(a) System matrix: it presents more than 8 million non-zeros.



(b) QMR with Jacobi convergence.



(c) $|E_z|$ on a z-constant section.

Figure 4.11: *Three-dimensional example: planar photonic crystal channel.*

pulse. The red lines are *output sensors*, in which fields are collected and compared.

Comparison is done by the calculation of the standard deviation, defined in (4.22).

4.6.1 Free-space propagation

The first validation test is studied mainly to calibrate the two algorithms, i.e. to find possible numerical constants implicitly included in the computed fields, "tune" the PMLs and test the results on a very simple problem. Figure 4.12(a) shows the computational domain: a $6 \times 1.5\,\mu m^2$ empty rectangle. The source is a Gaussian beam, $w = 1\,\mu m$ width, at $\lambda = 1.55\,\mu m$: with this choice of λ and w, diffraction is clearly visible. Fields are collected on the red lines, with the most left-hand sided used as a reference for the other one.

Figure 4.12(b) shows the quadratic error between FIMMPROP and our algorithm, as a function of the grid density. The different curves on the graph refer to different PMLs parameters in FIMMPROP.

We can note that the results tend to converge to the same value, within 0.2% of tolerance for a 30 points-per-wavelength mesh density. For a mesh 15 points-per-wavelength density, the standard deviation is already below 1%: the algorithm tends to converge very rapidly to a good solution and then any refinement is very slow. The reason can also be a non-perfect convergence of FIMMPROP's result, which showed difficulties with PMLs.

The field collected at the output sensor is shown in Figure 4.12(c): we can note its Gaussian shape, larger than at the input for the effect of diffraction. The little wobbling at the boundaries of the domain confirms the non-perfect tuning of the PMLs.

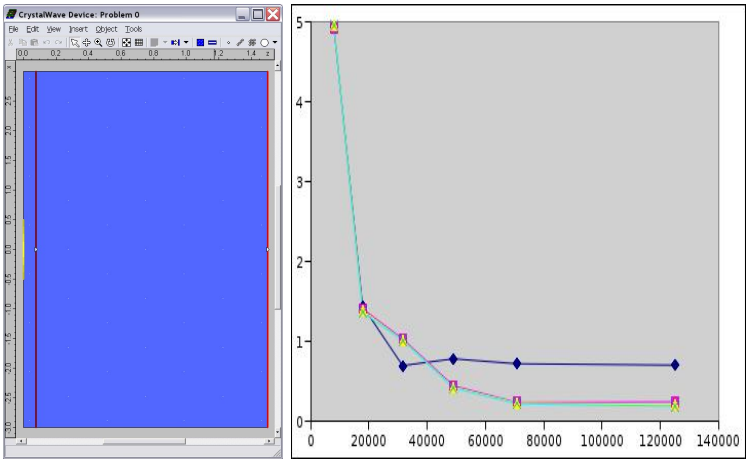
4.6.2 Single scatterer

The second validation test, whose computational domain is shown in Figure 4.13, is the diffraction of a Gaussian pulse by a dielectric scatterer. The domain's geometry is the same as in 4.6.1. The dielectric scatterer is a circle filled with air, in a dielectric domain with refractive index $n = 3.0$.

This time, a preliminary PMLs optimization in FIMMPROP has been done, and the results are shown in Figure 4.13(b). Different lines refer to different PML thicknesses[8].

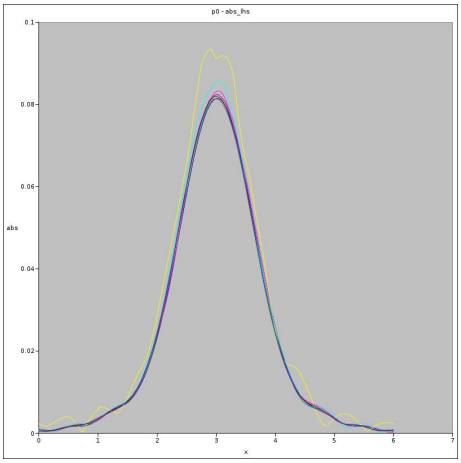
Figure 4.13(c) and Figure 4.13(d) show the convergence to the reference solution as a function of the mesh density and the output field. Again, the result converge quickly within 1% and then any mesh refinement brings little improvement. Note the shadow in the output field, due to the presence of the dielectric scatterer.

[8]It is not always true that the thicker the PML, the best the result: in algorithms based on modal expansion, the exponentially increasing modes inside the PMLs can lead to numerical instabilities. The tradeoff is, of course, between the numerical instability of a too thick PML layer and the poor absorption of a too thin one.



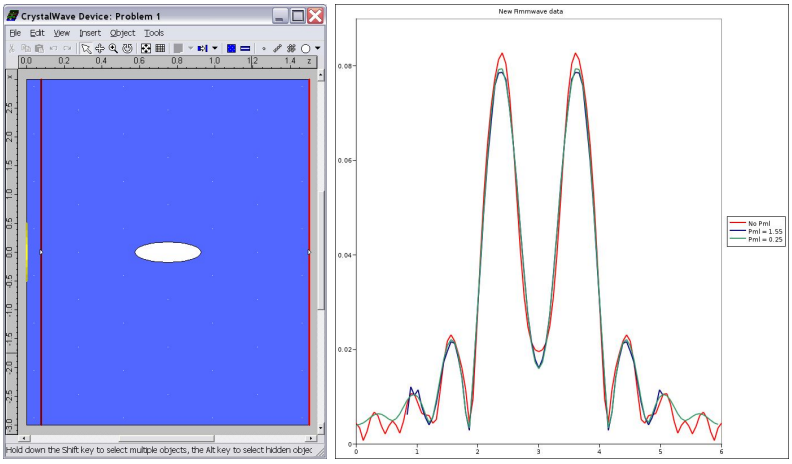
(a) Computational domain.

(b) Convergence to the reference solution.



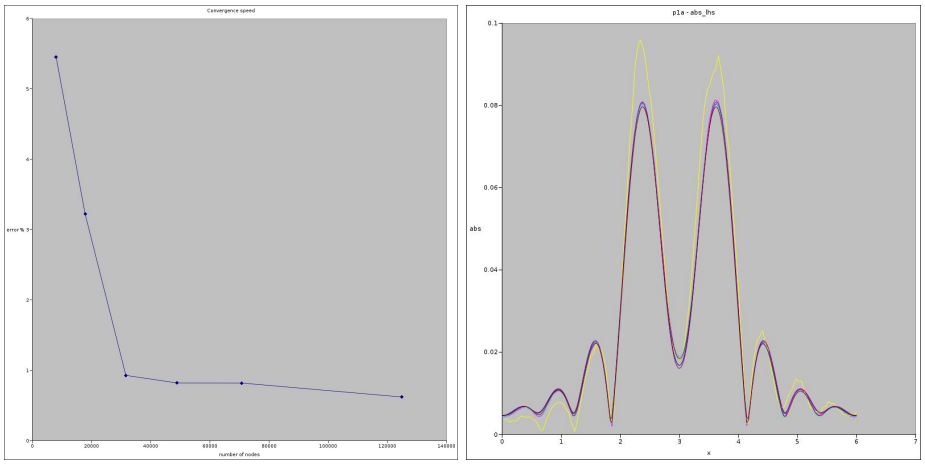
(c) Absolute value of the magnetic field on the most right-hand sided output sensor, for different mesh grids.

Figure 4.12: *Free-space propagation.*



(a) Computational domain.

(b) FIMMPROP PMLs tuning.



(c) Convergence to the reference solution.

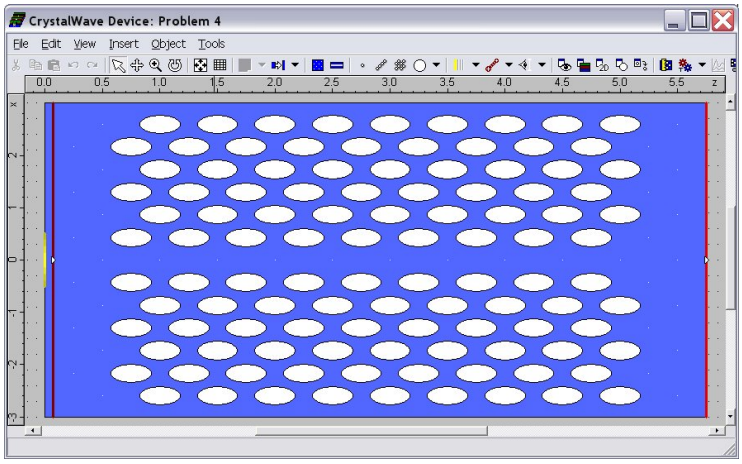
(d) Absolute value of the magnetic field on the most right-hand sided output sensor, for different mesh grids.

Figure 4.13: *Single scatterer.*

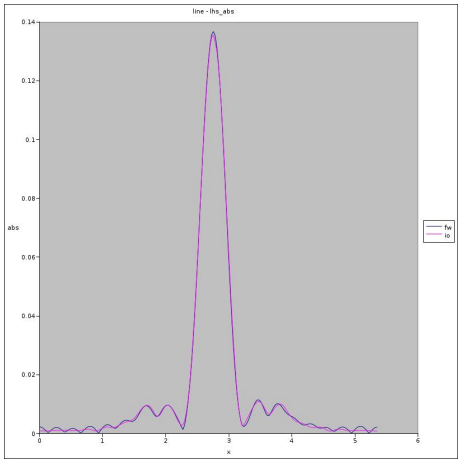
4.6.3 Photonic Crystal Channel

Finally, Figure 4.14 shows a complete photonic crystal channel. The domain is $6 \times 5.75\mu m^2$, filled with 108 scatterers as the one studied in 4.6.2, arranged in a triangular lattice of periodicity $\Lambda = 0.5\mu m$. A line defect is left in the lattice to guide the light through the device.

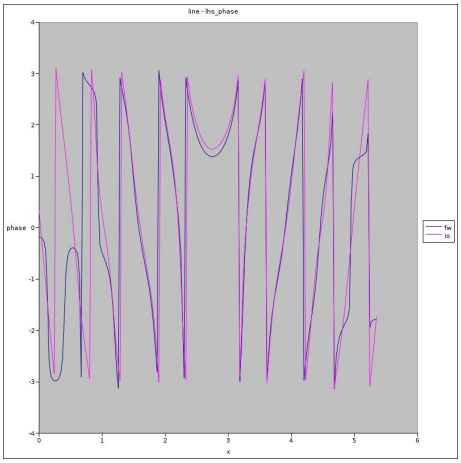
Figure 4.14(b) and Figure 4.14(c) show the absolute value of the z-component of the magnetic field and its phase at the output sensor. Comparison with FIMMPROP is very good, except near the boundaries where, to increase the speed and the accuracy of FIMMPROP, PMLs have been removed. Therefore, comparison makes sense only away from the boundaries, where the results of the two algorithms are in good agreement.



(a) Computational domain.



(b) Absolute value of the magnetic field at the output sensor.



(c) Phase of the magnetic field at the output sensor.

Figure 4.14: *Photonic crystal channel.*

Conclusions

The present algorithm represents a novel method to discretize and solve Maxwell equations. It shares some similarities with the *Finite Integration Technique* and the *Finite Element* method [BK00], but it's original in its integral approach to describe and discretize the problem.

Simulations have shown that it is an effective way to solve electromagnetic problems, both in the time- and frequency-domain. Benchmarks presented in section 4.6.1 and followings show a good accuracy of the results, if compared with commercial software.

On the other hand, the Author is convinced that the present method is very effective and superior to other simpler algorithms like `FDTD` or `FEM` only in two-dimensional simulation of frequency-domain problems. In the time-domain, the advantage of having an unstructured grid, which seems to be the only clear advantage over a conventional `FDTD` algorithm, does not pay the increased complexity of the algorithm, which directly effects the possibility of an efficient implementation on a computer. For three-dimensional frequency-domain problems, memory requirements are still too stringent to be applied for very complex problems. Other techniques must be used, that makes more assumptions on the problems in order to simplify it. This is the case, for example, of the *Multiple Scattering* method [TM97], which, making the hypothesis of perfectly circular scatterers, reduces the problem to an expansion over a efficient basis functions set and translate it into an eigenvalue problem, or the *Eigenmode Expansion* technique, used in FIMMPROP [FIMa].

A two-dimensional frequency-domain implementation of the algorithm, implemented by the Author, is commercially available from Photon Design [Pho] .

II

Mode Solvers

What are the "Mode Solvers"?

One of the most powerful approaches to study optical systems is the eigenmode decomposition: the electromagnetic propagation is expressed as a set of definite-frequency time-harmonic modes. In the absence of nonlinear effects, all the optical phenomena can be described as a super-imposition of these modes with an accuracy directly related to the number of modes used in the expansion and the completeness of the set of modes, seen as a basis functions set. All the possible algorithm formulations to find modes share a common characteristic: they reduce the problem to find the eigenvalues and eigenvectors of a given matrix, which represents the geometry and physics of the device under study.

These algorithms are what we call "Mode Solvers".

In the next chapters, we will describe two particular algorithms: the first, suitable to study waveguides with an invariant cross section along the direction of propagation, based on the Finite Difference Method, the second, suitable to study periodic structures, also known as photonic crystal, based on the Plane Wave Expansion technique. They somehow complete themselves, covering the greatest majority of the common problems found in optical systems.

5

Finite Difference Method

5.1 Introduction

The Finite Difference Method is well suited to study z-invariant devices, where z is the direction of propagation of the electromagnetic field, like the one shown in Figure 5.1.

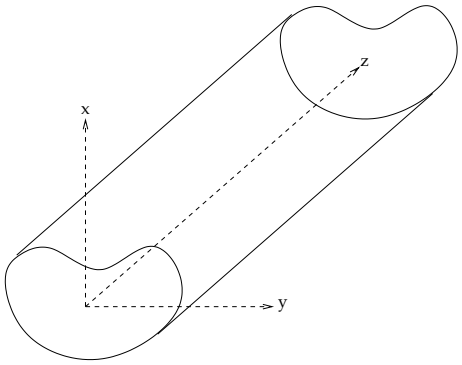


Figure 5.1: *z-invariant waveguide with a generic cross section.*

Maxwell equations, without sources, in the harmonic regime, for a linear non-magnetic medium, read [Som98]:

$$\begin{cases} \nabla \times \overrightarrow{E} = -\imath\omega\mu_0 \overrightarrow{H} \\ \nabla \times \overrightarrow{H} = \imath\omega\epsilon \overrightarrow{E} \end{cases} . \tag{5.1}$$

From the second equation we can write $\overrightarrow{E} = \frac{1}{\imath\omega\epsilon}\nabla \times \overrightarrow{H}$ and obtain the Helmholtz equation for the $\overrightarrow{H}$ field:

$$\nabla \times \frac{1}{\epsilon}\nabla \times \overrightarrow{H} - \omega^2\mu_0 \overrightarrow{H} = 0. \tag{5.2}$$

The same can be done for the $\overrightarrow{E}$ field, obtaining:

$$\nabla \times \nabla \times \overrightarrow{E} - \omega^2 \mu_0 \epsilon \overrightarrow{E} = 0. \tag{5.3}$$

Where ϵ is homogeneous, from Helmholtz equation we obtain the Laplace equation, valid for both $\overrightarrow{H}$ and $\overrightarrow{E}$ fields:

$$\nabla^2 \overrightarrow{F} + \omega^2 \mu_0 \epsilon \overrightarrow{F} = 0, \tag{5.4}$$

where $\overrightarrow{F} = \left\{ \begin{smallmatrix} \overrightarrow{H} \\ \overrightarrow{E} \end{smallmatrix} \right\}$ and we have used $\nabla \times \nabla \times \bullet = \nabla\nabla \cdot \bullet - \nabla^2 \bullet$ and $\nabla \cdot \overrightarrow{F} = 0$ (i.e. $\nabla \cdot \overrightarrow{H} = 0$ or $\nabla \cdot \overrightarrow{E} = 0$).

Now, a z-dependence of the type $e^{-\imath\beta z}$ is supposed for both $\overrightarrow{E}$ and $\overrightarrow{H}$: this is the z-dependence of any solution of the wave equation in a translational symmetric system in the z direction. Moreover, any mode can be described simply by its transversal coordinates x and y [Som98]. Therefore, in Cartesian coordinates, where $\overrightarrow{H} = (H_x, H_y, H_z)$ and $\overrightarrow{E} = (E_x, E_y, E_z)$, Laplace equation (5.4) becomes:

$$\begin{cases} \partial_x^2 F_x + \partial_y^2 F_x + (\omega^2 \mu_0 \epsilon - \beta^2) F_x = 0 \\ \partial_x^2 F_y + \partial_y^2 F_y + (\omega^2 \mu_0 \epsilon - \beta^2) F_y = 0 \end{cases}, \tag{5.5}$$

where, again, F stands for H or E.

Note that F_x and F_y are decoupled. Coupling between the two transverse components only arises where ϵ is non-homogeneous and it's stronger where bigger changes in ϵ are present. This can be easily seen also theoretically: in low index contrast devices modes are *almost purely* TE or TM, while in high index contrast devices they are *quasi-* TE or TM. See Part III.

Helmholtz equations (5.2) and (5.3), alone, are not equivalent to Maxwell equations: they have solutions that are not solutions of Maxwell equations, which are called *spurious solutions*. To avoid them, a supplementary condition must be imposed: $\nabla \cdot \overrightarrow{H} = 0$ (or $\nabla \cdot \overrightarrow{D} = 0$). This can be noted taking the divergence of (5.2) (or (5.3)): it is clear that, without this condition, an arbitrary field $\overrightarrow{H}$ (or $\overrightarrow{E}$), even if non-divergenceless, would be a zero frequency solution. This is, obviously, unphysical, because the divergenceless condition must hold for all frequencies.

Other approximations can be imposed to the previous equations, to reduce the complexity of the problem and simplify the algorithm to implement on a computer. The most often used simplification, is to suppose that TE and TM polarization are decoupled: the resulting algorithm is the so called *semivectorial*, described in Section 5.2. It is very easy to implement and very effective to study low index contrast devices. Unluckily, to study high index contrast devices, such as polarization rotators (see Chapter 7), a fully vectorial mode solver is needed, which takes into account the TE/TM coupling. It is described in Section 5.3.

5.2 Semivectorial Mode Solver

The semivectorial mode solver that we have implemented is a very simple, yet accurate, method, which can be used to have a first insight into the problem, when an absolute accuracy is not needed. It is based on the algorithm described in [Ste88].

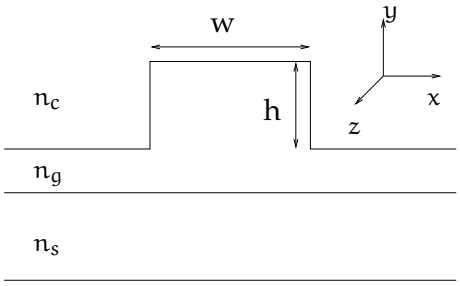


Figure 5.2: *Typical ridge waveguide, with a cladding, a guiding layer and a substrate. The reference system is also shown.*

Ridge waveguides like the one in Figure 5.2 can be studied: the refractive index profile precludes an analytical solution which is, strictly speaking, fully vectorial. In fact, TE and TM polarizations can not be distinguished: as pointed out in the Section 5.1, this is particularly true the higher the refractive index discontinuities are.

The simplicity of the present method, which is both its strength and its weakness, derives from some hypothesis on the geometry and the physics of the problem:

1. the refractive index is piecewise constant inside the domain, with the discontinuities occurring only at horizontal or vertical planes[1];
2. TE and TM polarizations are considered as decoupled, instead of being coupled into a fully vectorial solution: this allows to formulate two separate eigenproblems, one for each polarization, of smaller dimension than the full vectorial problem, but introduces the already mentioned approximations.

Starting from Maxwell equations in the frequency domain for a linear sourceless medium, we have:

$$\begin{cases} \nabla \times \overrightarrow{E} = -\imath\omega\mu\overrightarrow{H} \\ \nabla \times \overrightarrow{H} = \imath\omega\epsilon\overrightarrow{E} \end{cases} \tag{5.6}$$

and electric and magnetic Gauss laws:

$$\begin{cases} \nabla \cdot \overrightarrow{D} = 0 \\ \nabla \cdot \overrightarrow{B} = 0 \end{cases} . \tag{5.7}$$

[1]The algorithm is implemented in Cartesian coordinates.

Combining them together, we obtain Helmholtz equation for the electric field:

$$\nabla \times \nabla \times \overrightarrow{E} = \nabla\nabla \cdot \overrightarrow{E} - \nabla^2 \overrightarrow{E} = \omega^2 \epsilon \mu \overrightarrow{E} = k^2 \overrightarrow{E}. \tag{5.8}$$

The first hypothesis leads to the first simplification. Inside the regions where the refractive index is constant, from the first equation in (5.7), we have:

$$\nabla \cdot \overrightarrow{D} = \nabla \cdot \left(\epsilon \overrightarrow{E}\right) = \nabla\epsilon \cdot \overrightarrow{E} + \epsilon \nabla \cdot \overrightarrow{E} = 0,$$

from which we can write:

$$\nabla \cdot \overrightarrow{E} = \frac{\nabla\epsilon}{\epsilon} \cdot \overrightarrow{E}.$$

Where ϵ is piecewise constant, $\nabla\epsilon = 0$ and $\nabla \cdot \overrightarrow{E} = 0$ is satisfied: elsewhere, this is not strictly true, but the smaller the gradient, the smaller the refractive index discontinuities and the less important this approximation.

With this hypothesis, (5.8) becomes:

$$\nabla_T^2 \overrightarrow{E} + k^2 \overrightarrow{E} = \beta^2 \overrightarrow{E}, \tag{5.9}$$

where we have supposed a z dependence of the $\overrightarrow{E}$ fields of the type $e^{-\imath\beta z}$ and $\nabla_T^2 \triangleq \partial_x^2 + \partial_y^2$.

(5.9) does not contain any coupling between the x and y components of the $\overrightarrow{E}$ field, therefore it holds independently for each polarization:

$$\nabla_T^2 \overrightarrow{E} + k^2 \overrightarrow{E} = \beta^2 \overrightarrow{E} \qquad \Rightarrow \qquad \begin{cases} \nabla_T^2 E_x + k^2 E_x = \beta^2 E_x \\ \nabla_T^2 E_y + k^2 E_y = \beta^2 E_y \end{cases}. \tag{5.10}$$

We use now the second hypothesis: for the TE polarization, the electric field is $\overrightarrow{E} = (E_x, 0, E_z)$ and the first equation in (5.10) holds; for the TM polarization, $\overrightarrow{E} = (0, E_y, E_z)$ and the second holds.

Helmholtz equations (5.10) must now be discretized somehow, to be able to solve them on a computer. Discretization is done substituting the second order derivatives by finite differences on an orthogonal grid, like the one in Figure 5.3. Each internal interface is placed half-way between adjacent grid lines and each internal grid point is placed at the center of a rectangular cell of constant refractive index. Refractive index changes are only allowed at cell boundaries.

On the domain boundaries, we can impose different conditions:

- *Perfect Electric Conductor* (PEC): null tangential $\overrightarrow{E}$ and $\partial_n \overrightarrow{E}$, with $\hat{n}$ versor normal to the boundary;
- *Perfect Magnetic Conductor* (PMC): null tangential $\overrightarrow{H}$ and $\partial_n \overrightarrow{H}$, with $\hat{n}$ versor normal to the boundary;

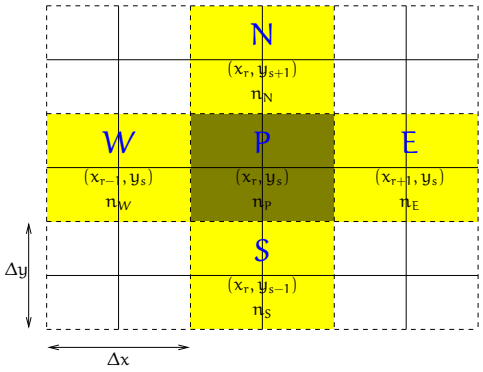


Figure 5.3: *Cell structure (in dashed lines) of the finite difference grid (in solid lines): refractive indices are constant within each cell, with discontinuities only allowed at cell boudaries. Cells have dimensions $\Delta x \times \Delta y$.*

- *Perfectly Matched Layers* (PMLs) or *Transparent Boundary Conditions* (TBC) to simulate free space.

For the TE polarization, the second order discretization of the space derivatives reads:

$$\begin{aligned}\Delta x^2 \left(\partial_x^2 E_x\right)_{r,s}^{(P)} &\simeq 2K_{r-1,s}^{(W)} E_{r-1,s} \\ &\quad - \left[2 + \left(K_{r,s}^{WP} - K_{r-1,s}^{W}\right) + \left(K_{r,s}^{EP} - K_{r+1,s}^{E}\right)\right] E_{r,s} \\ &\quad + 2K_{r+1,s}^{(E)} E_{r+1,s} \\ \Delta y^2 \left(\partial_y^2 E_x\right)_{r,s}^{(P)} &\simeq E_{r,s-1} - 2E_{r,s} + E_{r,s+1}.\end{aligned} \tag{5.11}$$

Note that the E_x component is continuous across horizontal interfaces and discontinuous across vertical interfaces. Factors K are defined:

$$\begin{aligned}K_{r+1,s}^{(E)} &\triangleq \frac{k_{r+1,s}^2}{k_{r,s}^2 + k_{r+1,s}^2} \\ K_{r+1,s}^{(EP)} &\triangleq \frac{k_{r,s}^2}{k_{r,s}^2 + k_{r+1,s}^2} \\ K_{r-1,s}^{(W)} &\triangleq \frac{k_{r-1,s}^2}{k_{r-1,s}^2 + k_{r,s}^2} \\ K_{r+1,s}^{(WP)} &\triangleq \frac{k_{r,s}^2}{k_{r-1,s}^2 + k_{r,s}^2}.\end{aligned}$$

Substituting of (5.11) into (5.10) and letting $\mathbf{E}_{TE}$ be the column vector with all the E_x values, one for each node of the discretization grid, we can write the matricial

eigenvalue equation:

$$\mathbf{A}_{TE}\,\mathbf{E}_{TE} = \beta_{TE}^2\,\mathbf{E}_{TE}. \tag{5.12}$$

The order in which the E_x values are stored in the $\mathbf{E}_{TE}$ vector is arbitrary: the problem is not dependent on the particular numbering of the grid nodes. Efficiency of numerical methods used to solve the eigenvalue problem (5.12) depends on the numbering, though: for example, the ILU decomposition of the sparse matrix $\mathbf{A}_{TE}$ tend to be more efficient if its non-zero values are closer to the main diagonal. Thus, any node numbering satisfying this condition is preferred in this case: luckily, most numerical routines to solve eigenvalue problems internally reorder $\mathbf{A}_{TE}$ to achieve better results.

A very similar reasoning can be applied to the TM polarization, for which the E_y component is continuous across vertical interfaces and discontinuous across horizontal interfaces. We end up with a matricial equation very similar to (5.12):

$$\mathbf{A}_{TM}\,\mathbf{E}_{TM} = \beta_{TM}^2\,\mathbf{E}_{TM}$$

that can be solved in the same way.

5.3 Vectorial Mode Solver

As said in Section 5.2, a semivectorial mode solver is not accurate enough to study HIC devices: hence, we have implemented a fully vectorial mode solver.

The algorithm chosen is the one presented in [LSSU94], that we are going to analyze more in detail here. Based on the Finite Difference approach, it is well known to ensure the absence of spurious modes, by a proper formulation of the physical equations and of the boundary conditions. It has also been proven to achieve a considerable precision on the computed effective index. This is a key point in today's designs: hundreds of microns long devices are quite common and errors on the effective indices as small as 10^{-5} at optical frequencies ($\lambda = 1.5\mu m$) results in a very high overall phase error.

Let's consider a z-invariant waveguide and discretize its xy cross section with a non-uniform Cartesian mesh, like in Figure 5.4. Association of physical quantities to geometrical elements is done supposing the magnetic field $\vec{H}$ attached to the nodes of the mesh and the refractive index constant inside each rectangle of the mesh: therefore, discontinuities can only occur at the edges. As said to deduce (5.4), inside each rectangle of the mesh, the decoupled version of the Helmholtz equation holds. Let's call $H|_W = H_W$ the magnetic field evaluated at point W (a similar nomenclature holds for P, N, S and E) and expand it in Taylor series around the point P to obtain a discretized form of $\partial_x^2 H_W$:

$$H_W = H_P + \partial_x H|_w(-w) + \frac{1}{2}\partial_x^2 H|_w(-w)^2 + \mathcal{O}\left[w^3\right],$$

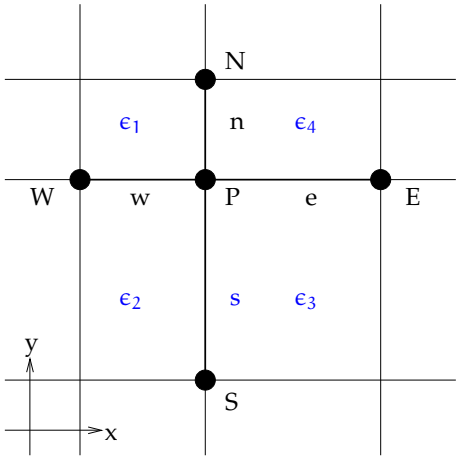


Figure 5.4: *Non-uniform Cartesian mesh for the vectorial mode solver.*

therefore:

$$\partial_x^2 H|_w = \frac{2H_W}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_w + \mathcal{O}\left[w^3\right]. \tag{5.13}$$

Analogous expressions hold for H_S, H_E, H_N and for ∂_y^2. Note that the first order derivatives are not discretized: the reason will be clear after the discussion about the boundary conditions.

Substituting (5.13) into (5.5), we obtain the discretized form, for both H_x and H_y:

$$\begin{aligned}
&\frac{2H_W}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_w + \frac{2H_N}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_n + \epsilon_1 k^2 H_P = \beta^2 H_P \\
&\frac{2H_W}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_w + \frac{2H_S}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_s + \epsilon_2 k^2 H_P = \beta^2 H_P \\
&\frac{2H_E}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_e + \frac{2H_S}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_s + \epsilon_3 k^2 H_P = \beta^2 H_P \\
&\frac{2H_E}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_e + \frac{2H_N}{w^2} - \frac{2H_P}{w^2} + \frac{2}{w}\partial_x H|_n + \epsilon_4 k^2 H_P = \beta^2 H_P.
\end{aligned} \tag{5.14}$$

The condition of divergenceless $\overrightarrow{H}$ is imposed through the boundary conditions. H_z and E_z are both continuous everywhere on the xy plane; therefore:

$$\nabla \cdot \overrightarrow{H} = 0 \qquad \Rightarrow \qquad H_z = \frac{1}{\imath\beta}\left(\partial_x H_x + \partial_y H_y\right) \tag{5.15}$$

and, from the second of Maxwell equations (5.1):

$$\nabla \times \overrightarrow{H} = \imath\omega\epsilon\overrightarrow{E} \qquad \Rightarrow \qquad E_z = \frac{1}{\imath\omega\epsilon}\left(\partial_x H_y - \partial_y H_x\right) \tag{5.16}$$

Discontinuities of ϵ can only occur on the edges of the mesh, so the boundaries between two different values of ϵ can be:

1. horizontal boundaries: on which H_x and H_z are continuous, and 5.15 becomes:
$$d_y H_y|_n = d_y H_y|_s,$$
while, 5.16 becomes:
$$\epsilon_n d_y H_x|_s - \epsilon_s d_y H_x|_n = (\epsilon_n - \epsilon_s) d_y H_x;$$

2. vertical boundaries: on which H_y and H_z are continuous, and 5.15 and 5.16 become, respectively:
$$\partial_x H_x|_w = \partial_x H_x|_e$$
and
$$\epsilon_e \partial_x H_y|_w - \epsilon_w \partial_x H_y|_e = (\epsilon_e - \epsilon_w) \partial_x H_y.$$

With these four boundary conditions and (5.14), the requested coupled differential equations in H_x and H_y can be obtained. Just note that, for the H_x component, multiplying the first of (5.14) by $\epsilon_2 n/2$ and summing it to the second multiplied by $\epsilon_1 s/2$, we obtain the left hand side of the second boundary condition. Similar arithmetical manipulations can be done, to obtain all the terms in the boundary conditions and to eliminate the first order derivative in 5.14. Finally, two coupled equations in H_x and H_y can be written, in the form:

$$\begin{cases} a_{xxW} H_{xW} + a_{xxE} H_{xE} + a_{xxN} H_{xN} + a_{xxS} H_{xW} + a_{xxP} H_{xP} + \\ a_{xyW} H_{yW} + a_{xyP} H_{yP} + a_{xyE} H_{yE} = \beta^2 H_{xP} \\ a_{yyW} H_{yW} + a_{yyE} H_{yE} + a_{yyN} H_{yN} + a_{yyS} H_{yW} + a_{yyP} H_{yP} + \\ a_{yxW} H_{xW} + a_{yxP} H_{xP} + a_{yxE} H_{xE} = \beta^2 H_{yP} \end{cases}.$$

In a more compact matricial form:

$$\mathbf{A}\,\mathbf{H} = \begin{bmatrix} \mathbf{A}_{xx} & \mathbf{A}_{xy} \\ \mathbf{A}_{yx} & \mathbf{A}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{H}_x \\ \mathbf{H}_y \end{bmatrix} = \beta^2 \begin{bmatrix} \mathbf{H}_x \\ \mathbf{H}_y \end{bmatrix} = \beta^2 \mathbf{H},$$

where $\mathbf{A_{ij}}$ contains all the coefficients ${a_{ij,rs}}^2$, $i, j = x, y$, and $\mathbf{H}_x$ ($\mathbf{H}_y$) all the x (y) components of the magnetic field, one for each grid point. Note that the coupling between H_x and H_y is caused by the elements of the off-diagonal matrices $\mathbf{A}_{xy}$ and $\mathbf{A}_{yx}$: in semivectorial algorithms or in homogeneous regions, they are zero.

5.4 Examples and validation

As an example of a typical device that can be studied with the described mode solvers, consider the rib waveguide presented in [Loh97].

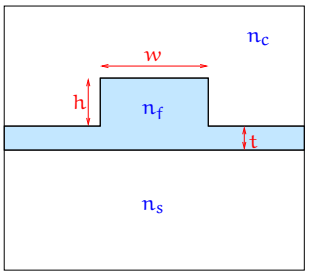


Figure 5.5: *Rib waveguide profile, taken from [Loh97].*

The domain is shown in Figure 5.5, where: $n_s = 3.34$, $n_f = 3.44$, $n_c = 1.0$, $w = 2\mu m$, $h = 1.1\mu m$ and $t = 0.2\mu m$.

In Figure 5.6 and Figure 5.7, results are shown for both the algorithms, the semivectorial of Section 5.2 and the fully vectorial of Section 5.3, respectively. The first four guided modes are plotted, two TE and two TM. The accordance between the two algorithms, in this case, is very good (see Table 5.1), within 0.02%. They also agree with the results reported in [Loh97].

	TE_1	TE_2	TM_1	TM_2
Semivectorial	3.387372	3.331755	3.388132	3.327159
Vectorial	3.388086	3.325017	3.387138	3.331010

Table 5.1: *Effective indices for the first four modes of the rib waveguide in Figure 5.5, computed by both the semivectorial and vectorial mode solvers.*

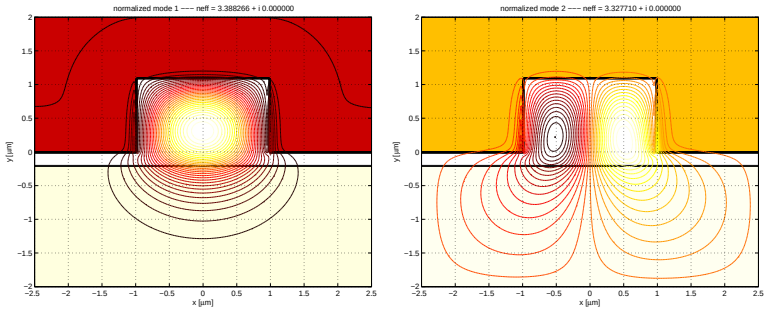
Special care has been taken to make the domain large enough so that the guided modes are far from the boundaries: in fact, no PMLs are present.

In this example, agreement between the semivectorial and the vectorial mode solver is very good: it could make the Reader think that the two algorithms are equivalent and the results are *always* similar. This is not true. It only happens because, for the particular choice of the waveguide's structure, the guided modes, even if not purely TE nor TM, are *strongly* quasi-TE and quasi-TM, with very small longitudinal E_z and H_z components: ignoring them is a very small approximation.

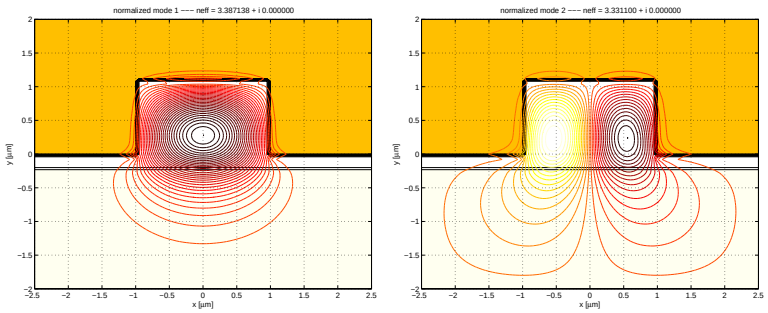
In the next example, a waveguide is studied, where a fully vectorial mode solver is needed. The device is shown in Figure 5.8. It is a buried rectangular waveguide of Si ($n_{core} = 3.45$) in SiO_2 ($n_{cladding} = 1.445$), with $w = 500nm$ and $h = 220nm$.

Table 5.2 shows the results obtained with the two algorithms. They differ greatly. In particular, the results given by the semivectorial mode solver are definitely inaccurate: it predicts that all the four modes are guided, instead of only the first two

[2]The explicit value of the coefficients $a_{ij,rs}$ can be found in [LSSU94].

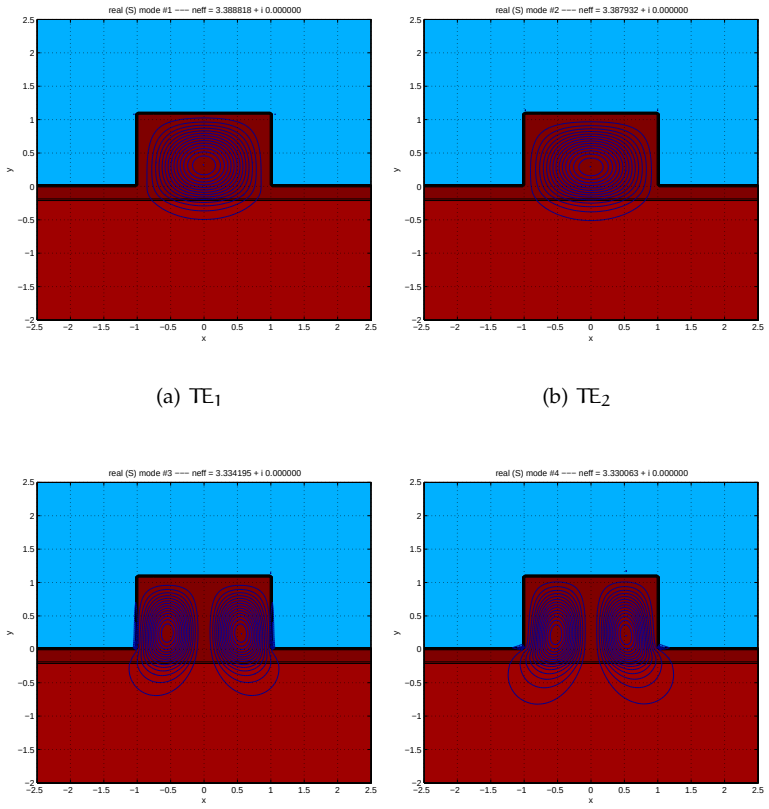


(a) TE_1 (b) TE_2



(c) TM_1 (d) TM_2

Figure 5.6: *Magnetic fields of the first four guided modes, computed by the semivectorial mode solver.*



(a) TE_1 (b) TE_2

(c) TM_1 (d) TM_2

Figure 5.7: *Longitudinal component of the Poynting vector of the first four guided modes, computed by the vectorial mode solver.*

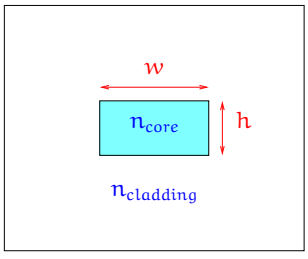


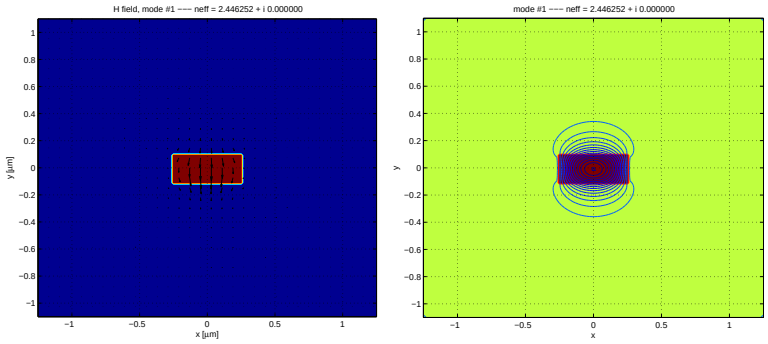
Figure 5.8: *Buried rectangular waveguide profile.*

TE and the first TM modes [Pir]: this is correctly predicted by the vectorial mode solver.

	TE_1	TE_2	TM_1	TM_2
Semivectorial	2.633596	2.263625	2.532738	1.976528
Vectorial	2.446252	1.525369	1.800634	1.333947

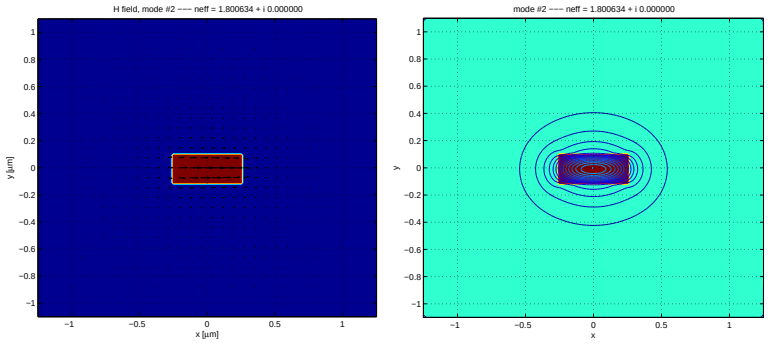
Table 5.2: *Effective indices for the first four modes of the buried rectangular waveguide in Figure 5.8, computed by both the semivectorial and vectorial mode solvers. Note that the* TM_2 *mode is guided for the semivectorial mode solver and not guided for the vectorial: the semivectorial is wrong.*

Figure 5.9 shows the magnetic field for the first two guided modes of the device. More examples for the use of the vectorial mode solver can be found in Chapter 7.



(a) First TE mode: magnetic field.

(b) First TE mode: contour plot of the H_y component.



(c) First TM mode: magnetic field.

(d) First TM mode: contour plot of the H_x component.

Figure 5.9: *Magnetic field for the first two guided modes for the device in Figure 5.8.*

6

Plane Wave Expansion Technique

6.1 Introduction

The Plane Wave Expansion (PWE) technique is a powerful algorithm to study the propagation of light in periodic media. Its main advantage is to use a formulation so that periodic boundary conditions are automatically satisfied: the problem is then reduced to the solution of an eigenvalue problem, whose dimension depends on the desired accuracy of the solution.

First, a brief sketch of periodic structures, also known as photonic crystal, is given, mainly to define the terminology used in the rest of the chapter. Then, the PWE algorithm will be investigated and explained. Finally, a comparison between the implemented algorithm and both freely and commercially available software is presented.

6.2 Photonic Crystals

Photonic crystals are non-homogeneous dielectric structures, whose refractive index discontinuities are periodically spaced of distances comparable with the propagating radiation's wavelength. Periodicity can be one-dimensional, two-dimensional or three-dimensional [Yab93].

One-dimensional photonic crystals , also called *multilayers*, are widely used in integrated optics as phase compensators, mirrors, tunable filters and fiber-waveguide couplers.

Two-dimensional photonic crystals are used as planar waveguides, if propagation of light is *through* the crystal, or as photonic crystal fibers, if the propagation of light is *along* the crystal.

Three-dimensional photonic crystals are present in nature, as diamonds, for examples, but are not practically used because of the difficulty to realize them.

Propagation of light in periodic structures is mathematically described by the Bloch, or Floquet, theorem [Flo83]: basically, the guided modes can be written as the product of a plane wave and a function, periodic over the fundamental cell of the lattice. Periodicity denies some frequencies to be guided through certain directions of propagation, therefore creating an anisotropic medium. Moreover, some frequencies can be denied to propagate for every possible direction: in this case, the crystal is said to present a *complete bandgap*.

Periodic structures like photonic crystals can be geometrically described by the concept of *Bravais lattice*. A Bravais lattice is an infinite array of discrete points with an *arrangement* and an *orientation* that appear *exactly* the same from whichever of the points the array is viewed [Kit95]. In more rigorous words, for a three-dimensional lattice, it consists of all the points with position vectors:

$$\overrightarrow{R} = n_1\overrightarrow{R_1} + n_2\overrightarrow{R_2} + n_3\overrightarrow{R_3},$$

with n_i, $i = 1, 2, 3$, integer number. The vectors $\overrightarrow{R_i}$, $i = 1, 2, 3$, are called *primitive vectors* and must be linearly independent, i.e. non-coplanar (see Figure 6.1).

For a given Bravais lattice, the choice of the primitive vectors is not unique: indeed, there are infinite valid choices (see Figure 6.1(b)).

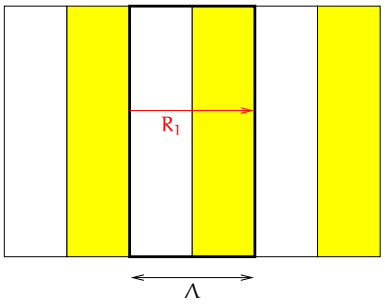
Since all points are equivalent, the Bravais lattice is infinite in extent. This is clearly an approximation in the description of a real crystal, which is never infinite: if it's large enough, though, the vast majority of the points will be too far from the boundaries to be affected by them and the description will be accurate.

The volume of space which, when translated through all the vectors in the Bravais lattice, just fills all of the space without overlapping or leaving voids is called *primitive cell*. Again, there is no unique choice of the primitive cell for a given Bravais lattice, but, as long as every primitive cell must contain exactly one lattice point (unless it is so positioned that there are points on its surface) and the density of points in the lattice is constant, the volume of the primitive cell is constant for each possible choice.

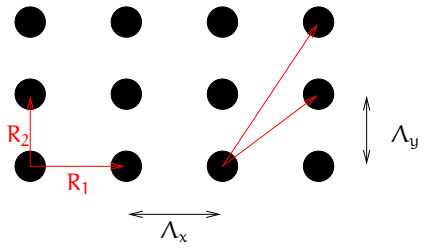
The most frequently used primitive cell, given a Bravais lattice, is the *Wigner-Seitz* primitive cell. It is defined as the region of space which is closer to a point of the lattice than to every other point: see Figure 6.2. Its construction resembles closely the construction of the Voronoï diagram of a Delaunay mesh (see Section 1.2.2).

To describe a real crystal both the description of the underlying Bravais lattice and of the arrangements of atoms inside each unit cell are needed. The latter is called *basis* and a crystal is sometimes referred to as a *lattice with a basis*. For example, Figure 6.3 shows a crystal with a honeycomb arrangement of atoms: it is not a Bravais lattice if the unit cell contains just one atom (the orientation uniformity is missing), but it is if the unit cell is made of two atoms.

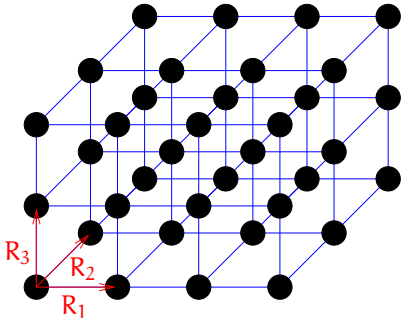
Given a Bravais lattice, another important concept is its *reciprocal lattice*, defined as the set of all the wave vectors $\overrightarrow{G}$ that yield plane waves with the same periodicity of the lattice. Analytically, $\overrightarrow{G}$ belongs to the reciprocal lattice of a Bravais lattice of



(a) One-dimensional. The primitive cell is in bold lines.



(b) Two-dimensional square lattice. Two possible choices of the primitive vectors are shown.



(c) Three-dimensional cubic lattice.

Figure 6.1: *Examples of photonic crystals geometries.*

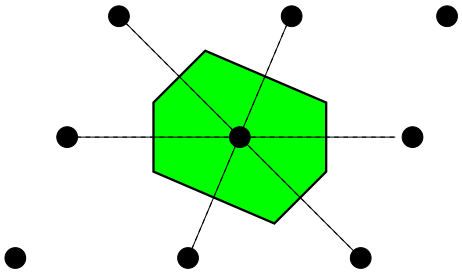


Figure 6.2: *The Wigner-Seitz cell for a two-dimensional Bravais lattice.*

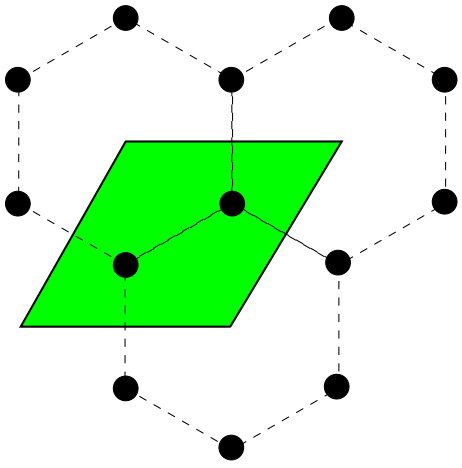


Figure 6.3: *Honeycomb lattice: a Bravais lattice with a two-points basis. In green, the primitive cell: it contains two atoms.*

points $\overrightarrow{R}$ if:

$$e^{\imath\overrightarrow{G}\cdot\left(\overrightarrow{\tau}+\overrightarrow{R}\right)} = e^{\imath\overrightarrow{G}\cdot\overrightarrow{\tau}} \tag{6.1}$$

for any $\overrightarrow{\tau}$ and for all $\overrightarrow{R}$ in the Bravais lattice.

6.1 can be rewritten factoring out $e^{\imath\overrightarrow{G}\cdot\overrightarrow{\tau}}$, to obtain:

$$e^{\imath\overrightarrow{G}\cdot\overrightarrow{R}} = 1 \tag{6.2}$$

for all $\overrightarrow{R}$ in the Bravais lattice.

It's worth noting that the reciprocal lattice of a Bravais lattice is itself a Bravais lattice. To prove it, just note that the reciprocal lattice, in three dimensions, can be generated by the primitive vectors $\overrightarrow{G_i}$, $i = 1, 2, 3$:

$$\overrightarrow{G}_1 = 2\pi\frac{\overrightarrow{R}_2 \times \overrightarrow{R}_3}{\overrightarrow{R}_1 \cdot \left(\overrightarrow{R}_2 \times \overrightarrow{R}_3\right)}$$

$$\overrightarrow{G}_1 = 2\pi\frac{\overrightarrow{R}_3 \times \overrightarrow{R}_1}{\overrightarrow{R}_1 \cdot \left(\overrightarrow{R}_2 \times \overrightarrow{R}_3\right)}$$

$$\overrightarrow{G}_1 = 2\pi\frac{\overrightarrow{R}_1 \times \overrightarrow{R}_2}{\overrightarrow{R}_1 \cdot \left(\overrightarrow{R}_2 \times \overrightarrow{R}_3\right)}$$

and that they satisfy the orthogonality condition:

$$\overrightarrow{G}_i \cdot \overrightarrow{R}_j = 2\pi\delta_{ij},$$

with δ_{ij} the Kronecker delta function. Therefore, any vector $\overrightarrow{k}$ in the reciprocal lattice space can be written as a linear combination of $\overrightarrow{G}_i$, $i = 1, 2, 3$, with coefficients k_i, $i = 1, 2, 3$:

$$\overrightarrow{k} = k_1\overrightarrow{G}_1 + k_2\overrightarrow{G}_2 + k_3\overrightarrow{G}_3$$

and any vector $\overrightarrow{R}$ in the Bravais lattice space can be written as:

$$\overrightarrow{R} = n_1\overrightarrow{R}_1 + n_2\overrightarrow{R}_2 + n_3\overrightarrow{R}_3$$

with $n_i \in \mathbb{Z}$. Therefore:

$$\overrightarrow{k} \cdot \overrightarrow{R} = 2\pi\left(k_1n_1 + k_2n_2 + k_3n_3\right).$$

For $e^{\imath\overrightarrow{k}\cdot\overrightarrow{R}}$ to be unity for any choice of $\overrightarrow{R}$, $\overrightarrow{k} \cdot \overrightarrow{R}$ must be 2π times an integer number for each choice of the integers n_i. This requires that k_i are integers as well. This concludes the proof. □

From this consideration, we can define all the entities, defined for Bravais lattices, on the reciprocal lattice:

- the reciprocal lattice of a reciprocal lattice turns out to be the original Bravais lattice;
- the Wigner-Seitz primitive cell of a reciprocal lattice is called *first Brillouin zone*: it contains all the "meaningful" wavevectors to study the propagation of light into the crystal, in the sense that all the other wavevectors can be deduced from these ones by periodicity or linear combination.

6.3 Plane Wave Expansion Algorithm

Eigen-decomposition of electromagnetic systems can be approached in many different ways. The most common are:

`FDTD` **simulations** : even if this is a time-domain technique, informations in the frequency-domain can be extracted by Fourier transforming the response of the system to a time-varying input signal. The peaks in the spectrum determines the eigenfrequencies of the system [CYH95].

Eigenmode expansion : the fields are expressed as the sum of a finite number of functions of a complete basis set (for example, plane waves) and Maxwell equations are recast into an eigenvalue problem [JJ01, TM97];

Transfer matrix method : for each frequency, the transfer matrix, relating field amplitudes at the beginning and at the end of the system, is computed; this yields the transmission spectrum directly and the modes of the system as the eigenvector of the transfer matrix itself.

Each of them has its own peculiar advantages and disadvantages.

Time-domain simulations oppose the implementation triviality of the algorithm to the difficulty of obtaining accurate results. The biggest problem is that the time-varying input signal must not be orthogonal to any mode of the system, otherwise orthogonal modes will not be excited and their eigenfrequecies will not be visible in the spectral response of the system. Another problem is connected with the computational resources needed by these algorithms. Frequency resolution Δf is directly proportional to the total duration of the simulation T, as stated by the "uncertainty principle" of the Fourier transform $T\Delta f \sim 1$: therefore, very long simulations are needed to have the sufficient spectral resolution and distinguish quasi-degenerate modes.

Eigenmode expansion algorithms, on the other hand, don't suffer these limitations, but present other problems, almost always connected to the convergence of the solution of the eigenvalue problem. Ordinary matricial eigenproblems scale as $\mathcal{O}\left[N^3\right]$, with N the matrix dimension, and need $\mathcal{O}\left[N^2\right]$ storage: in this case, N is also the number of plane waves retained in the expansion. This becomes the bottleneck even for relatively small problems and different solutions have been investigated. We'll talk about them later. Another problem, especially connected to the choice of plane waves as the basis set, is the description of discontinuities in

the dielectric function. Plane waves poorly describe step functions, unless a very large number of them is employed. Clever smoothing of the dielectric function can alleviate the problem.

The transfer matrix approach is somehow hybrid between the time-domain and the eigenmode expansion methods. Even if it tries to get "the best from two worlds" it is only practically applicable when the system is made of many simple blocks, whose transfer matrices are easily obtainable. Otherwise, more complex time- and frequency-domain methods are needed to study each block of the system and then all the blocks need to be combined together to give the total transfer matrix of the system.

We'll describe here the Plane Wave Expansion (`PWE`) technique, which is a particular type of eigenmode expansion algorithm.

`PWE` is a fully vectorial, three-dimensional algorithm used to determine the definite-frequency eigenstates of Maxwell equations in arbitrary periodic dielectric structures. Birefringence and intrinsic losses can be treated as well [JJ01].

Starting from Maxwell equations (5.1) in the time-domain, for a linear medium, and the magnetic Gauss law (5.7) for a non-magnetic medium, we can rewrite them as:

$$\nabla \times \frac{1}{\epsilon} \nabla \times \overrightarrow{H} = -\frac{1}{c^2} d_t^2 \overrightarrow{H} \tag{6.3}$$

$$\nabla \cdot \overrightarrow{H} = 0. \tag{6.4}$$

Suppose an harmonic time-dependence of the type $e^{-\imath\omega t}$ for the $\overrightarrow{H}$ field: we are looking for definite frequency eigenmodes. Apply the Bloch theorem: we are studying periodic systems. Then we can write the field $\overrightarrow{H}$ as:

$$\overrightarrow{H} \triangleq e^{\imath\left(\overrightarrow{k}\cdot\overrightarrow{x}-\omega t\right)} \overrightarrow{H}_{\overrightarrow{k}}, \tag{6.5}$$

where $\overrightarrow{k}$ is the Bloch wavevector and $\overrightarrow{H}_{\overrightarrow{k}}$ is a periodic function field, completely defined in the unit cell, i.e. the Bloch function.

Therefore, substituting 6.5 in 6.3, we obtain:

$$\mathbf{A}_{\overrightarrow{k}} \left[\overrightarrow{H}_{\overrightarrow{k}}\right] = \frac{\omega^2}{c^2} \overrightarrow{H}_{\overrightarrow{k}}, \tag{6.6}$$

where $\mathbf{A}_{\overrightarrow{k}}[\bullet]$ is the positive semi-definite Hermitian operator:

$$\mathbf{A}_{\overrightarrow{k}}[\bullet] = \left(\nabla + \imath\overrightarrow{k}\right) \times \frac{1}{\epsilon} \left(\nabla + \imath\overrightarrow{k}\right) \times \bullet.$$

Being defined on the unit cell, $\overrightarrow{H}_{\overrightarrow{k}}$ has compact support, therefore the eigenvalues of (6.6) are a discrete set of eigenfrequencies $\omega_n(\overrightarrow{k})$, forming a continuous band structure, function of $\overrightarrow{k}$.

Solving (6.6) on a computer requires some sort of discretization, to reduce the infinite-dimensional eigenvalue problem to a finite one: this reduction is the source of all the approximations and errors of the algorithm. In frequency-domain algorithms, the solution is usually written as a linear combination of some basis vectors $\overrightarrow{b}_m$, truncated at some integer N:

$$\overrightarrow{H}_{\vec{k}} = \sum_{m=1}^{\infty} h_m \overrightarrow{b}_m \simeq \sum_{m=1}^{N} h_m \overrightarrow{b}_m. \tag{6.7}$$

Obviously, the equal holds for a complete basis with N equal to infinity and accuracy grows with N.

Substituting 6.7 in 6.6 we obtain an ordinary generalized eigenproblem of the form:

$$\mathbf{A}\,\mathbf{h} = \left(\frac{\omega}{c}\right)^2 \mathbf{B}\,\mathbf{h}, \tag{6.8}$$

where $\mathbf{h}$ is the column vector of the basis coefficients h_m, $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ matrices whose entries are:

$$A_{lm} = \int \overrightarrow{b}_l \cdot \left(\mathbf{A}_{\vec{k}} \left[\overrightarrow{b}_m\right]\right)^* \qquad\qquad B_{lm} = \int \overrightarrow{b}_l \cdot \overrightarrow{b}_m{}^*,$$

where integrals are intended as done over the unit cell.

Also the divergenceless condition (6.4) must be satisfied, in order to get rid of the zero-frequency spurious modes otherwise introduced (see Section 5.1). To satisfy this condition, we have two possibilities:

1. add other constrains to the eigenproblem (6.8): this increases the dimension of the problem and the computational requirements;
2. choose the basis functions so that each of them satisfies (6.4): each linear combination of these functions will then satisfy the condition as well, automagically.

We prefer to adopt the second possibility and, looking for the best basis functions, we can find other properties that they should have:

- they should form a compact representation, so that a reasonable number of basis functions yields a good accuracy and convergence to the exact solution is fast;
- it should be easy to compute $\mathbf{A}\,\mathbf{h}$ and $\mathbf{B}\,\mathbf{h}$.

In the `PWE` method, the basis functions are plane waves:

$$b_m = e^{\imath \vec{G}_m \cdot \vec{x}},$$

where $\overrightarrow{G}_m$ are some reciprocal lattice vectors (see Section 6.2).

For example, for a three-dimensional lattice described by the lattice vectors $\vec{R}_1$, $\vec{R}_2$ and $\vec{R}_3$ and reciprocal lattice vectors $\vec{G}_1$, $\vec{G}_2$ and $\vec{G}_3$, $b_{m_1,m_2,m_3} = e^{\imath \sum_j m_j \vec{G}_j \cdot \vec{x}}$, with $m_j = -\lceil N_j/2 \rceil + 1, \ldots, \lfloor N_j/2 \rfloor$ and $N = N_1 N_2 N_3$.

Note that with this choice, the basis functions are not vectors: as a consequence, the basis coefficients must be vectors, to be able to describe a vectorial field as $\vec{H}$. Therefore:

$$\begin{aligned} \vec{H}(\vec{x}) &= \vec{H}\left(\sum_k n_k \frac{\vec{R}_k}{N_k}\right) \\ &= \sum_{\{m_j\}} \vec{h}_{\{m_j\}} e^{\imath \sum_{j,k} m_j \vec{G}_j \cdot n_k \vec{R}_k / N_k} \\ &= \sum_{\{m_j\}} \vec{h}_{\{m_j\}} e^{2\pi\imath \sum_j m_j n_j / N_j} \end{aligned} \tag{6.9}$$

Note that (6.9) is precisely the three-dimensional DFT (Discrete Fourier Transform) of the coefficients $\vec{h}_{\{m_j\}}$. Many very efficient FFT (Fast Fourier Transform) algorithms are available, which can be used to compute it very efficiently, in $\mathcal{O}[N \log N]$ time, instead of $\mathcal{O}\left[N^2\right]$ as a normal matrix-vector product.

Is the choice of plane waves as basis a good choice? Let's look back at the requirements listed before and discuss them one by one.

- The divergence condition (6.4), in the plane waves world, becomes a simple product:

$$\nabla \cdot \vec{H} = 0 \qquad \longrightarrow \qquad \forall m: \quad \vec{h}_m \cdot \left(\vec{k} + \vec{G}_m\right) = 0.$$

  This is easily satisfied if, for each reciprocal lattice vector $\vec{G}_m$, we decide to write $\vec{h}_m$ as a linear combination of two vectors $\{\vec{u}_m, \vec{v}_m\}$ orthogonal to $\vec{G}_m$: $\vec{h}_m = h_m^{(1)} \vec{u}_m + h_m^{(2)} \vec{v}_m$. Then the basis is intrinsically transverse and the divergenceless condition is satisfied. Moreover, the eigenproblem's dimensions reduce to the $2N$ unknowns $h_m^{(1)}$, $h_m^{(2)}$, instead of $3N$.

- The compact representation is probably the biggest limitation of plane waves as basis functions. A large number of plane waves is needed to describe discontinuous functions, as the dielectric function often is, in conventional optical problems: this leads to both long computational times and slow convergence of the solution, if large refractive index discontinuities are present. Smoothing the dielectric function can alleviate the problem.

- **A h** and **B h** can be easily computed. **B** is simply the identity matrix, thanks to the orthonormality of plane waves, while **A** can be computed in the reciprocal lattice space via FFT , avoiding the expensive curl operators:

$$A_{lm} = -\left(\vec{k} + \vec{G}_l\right) \times \texttt{IFFT}\left[\widetilde{\epsilon^{-1}} \texttt{FFT}\left[\left(\vec{k} + \vec{G}_m\right) \times \bullet\right]\right], \tag{6.10}$$

where $\widetilde{\epsilon^{-1}}$ is an effective inverse dielectric tensor.

6.3.1 Improvements on the Algorithm

Non-uniform grid

One clear disadvantage of plane waves, that is not possible to solve with clever tricks in the algorithm, is that the discretization of the dielectric function is done by a uniform grid. Sometimes it would be useful to be able to increase the accuracy of the solution only in some parts of the domain, i.e. increase the grid density or the number of plane waves, for example where the dielectric function discontinuities are higher. This can be achieved only changing the basis functions. A typical choice, in this case, is the traditional finite-element basis, formed of localized functions on an unstructured mesh.

Other limitations of the plane wave approach can be partially solved by looking with more attention at 6.8 and 6.9.

Inversion symmetry

The basis coefficients vector **h** is, in general, complex. This is due to the fact that the matrix **A** is not real symmetric and its eigenvalues are complex. **A** can be real symmetric only if we suppose that the *inversion symmetry* property holds for the dielectric function, i.e.:

$$\epsilon\left(-\overrightarrow{x}\right) = \epsilon\left(\overrightarrow{x}\right).$$

In this case, the `FFT` applied to ϵ in (6.10) operates on a real and even function resulting in a real and even matrix **A**. Eigenvalue-finders algorithms for real and symmetric matrices require half the storage, less than half the time (because better algorithms for real numbers than for complex are available) and usually less iterations to achieve convergence (because of the reduced dimension of the problem). The spatial fields in a domain that satisfies the inversion symmetry problem, on the other hand, satisfy:

$$\overrightarrow{H}_{\overrightarrow{k}}\left(\overrightarrow{x}\right) = \overrightarrow{H}_{\overrightarrow{k}}\left(\overrightarrow{-x}\right)^{*}.$$

Refractive index discontinuities

The second limitation is that plane waves describe badly discontinuities of the dielectric function. Simply applying the `IFFT` to the inverse dielectric function in 6.10 leads to suboptimal convergence [VP94].

This can be better understood with an example. Using plane waves to series expand a step function, i.e. Fourier transforming it, leads to an infinite number of expansion terms needed to reconstruct the function. As long as a uniformly convergent series of continuous functions, like plane waves, always yields a continuous function, then the Fourier series of the discontinuous step function cannot be uniformly convergent. The convergence is just of the mean, overshooting and

undershooting here and there the actual values, at the simple discontinuity. This is known as *Gibbs phenomenon* (see Figure 6.4). As the number of expansion terms is increased, the overshoots and undershoots increase, while moving closer to the discontinuity: for a dielectric function, it could also happen that it becomes negative, which is physically meaningful only in the creation of plasma phenomenon Section 3.5[1].

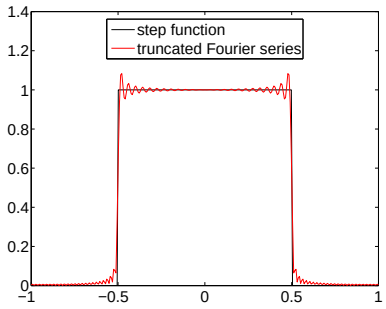


Figure 6.4: *Gibbs phenomenon: the overshooting of the truncated Fourier series is due to the missing harmonics.*

A perfect solution to this problem is not available – truncating the series expansion we are wasting information that we can not reconstruct, – but we can formulate the eigenvalue problem so that it is less sensible to it.

The are two possible formulations [VP94].

$\overrightarrow{H}$ **formulation** , which is the one described in the previous Section 6.3. Helmholtz equation is written using the $\overrightarrow{H}$ field and discretized using plane waves, yielding the eigenproblem in (6.8). This formulation is usually preferred, because it leads to a simple eigenvalue problem, not a generalized one. The only difficulty is represented by the factor $\frac{1}{\varepsilon}$ between the curl operators[2], that leads to the definition of the matrix $\mathbf{A}$. There are two methods to define it:

1. inverse expansion method: $\mathbf{A}$ is the Toeplitz matrix [Mat] obtained from the Fourier transform of $\frac{1}{\varepsilon}$:

$$\mathbf{A} = \text{Toeplitz}\left(\mathcal{F}\left[\frac{1}{\varepsilon}\right]\right);$$

[1]Something similar also happens in the `FDTD` algorithm, when we try to propagate a step function. Taking a one-dimensional propagation as an example, if the Courant factor is not exactly equal to one the numerical grid is a dispersive material [Taf00] and, as the step function propagates, it is modified by the numerical medium. After few timesteps, the plane waves that made up the step function are out of phase and something similar to the Gibbs phenomenon is clearly visible. The coarser the grid, the more visible the phenomenon.

[2]This problem has been already met in (5.2).

2. Ho's method [HCS90]: **A** is the inverse matrix of the Toeplitz matrix obtained from the Fourier transform of ϵ:

$$\mathbf{A} = \left(\text{Toeplitz}\left(\mathcal{F}\left[\varepsilon\right]\right)\right)^{-1}.$$

Both formulations are identical in infinite-dimensional systems, but in truncated systems the first gives a slower convergence than the second. The disadvantage of the Ho's method is that taking the inverse of the initial (symmetric) Toeplitz matrix breaks the symmetry of the **A** matrix and the eigenproblem becomes a generalized eigenvalue problem. Interestingly enough, the Ho's formulation is exactly equivalent to the next $\overrightarrow{E}$ formulation.

$\overrightarrow{E}$ **formulation** . Helmholtz equation is written using the $\overrightarrow{E}$ field and discretized using plane waves. Now, the matrix **A** is the matrix corresponding to the operator $\mathbf{A}_{\overrightarrow{k}}\left[\bullet\right] = \left(\overrightarrow{k} + \overrightarrow{G}\right) \times \left(\overrightarrow{k} + \overrightarrow{G}\right) \times \bullet$ and the matrix **B** is the Toeplitz matrix obtained from the Fourier transform of ϵ. The eigenproblem is now a generalized Hermitian eigenvalue problem. Compared to the $\overrightarrow{H}$ formulation, this leads to a faster convergence, but, even if the number of iterations are less, each iteration involves more floating points operations. Which method to use depends on the dimension of the whole problem and on the incident polarization.

This distinction closely resembles the teaching from *Effective Medium* theory, about the best way of smoothing the refractive index to improve convergence in similar eigenvalue problems: indeed, smoothing of the refractive index is another way to alleviate the plane wave expansion problem, as well. There are two effective ways of smoothing the dielectric function, depending of polarization of the incident light, relative to the versor $\hat{n}$ normal to the discontinuity surface. For $\overrightarrow{E} \parallel \hat{n}$, better results are achieved by averaging the inverse of ϵ; for $\overrightarrow{E} \perp \hat{n}$, it's better to take the inverse of the average of ϵ. If the incident wave is neither TE nor TM, one can average these two "averages" weighting the sum by a projection operator P, that takes into account the hybrid input polarization [JJ01].

This averaging method would improve convergence also in the FDTD algorithm, but it's never been applied there, to the Author's knowledge: this can be due to the difficulty to define the notion of "direction of propagation" in an algorithm which is intrinsically "isotropic", in the sense that do not suppose any particular preferred direction.

If the dielectric discontinuities have particular properties, more effective methods than a simple smoothing can be employed to improve convergence.

For example, if the dielectric discontinuities are piecewise constant and parallel to the principal dielectric axes, the crystal is said to fulfill the *perpendicularity condition* [Lal98]. See Figure 6.5, for an example of such a structure.

In this case $\overrightarrow{E}_x$ is a continuous function in y and z and discontinuous in x. Moreover, only ϵ_x acts on $\overrightarrow{E}_x$, so that the product $\epsilon_x \overrightarrow{E}_x$ is continuous in x. There-

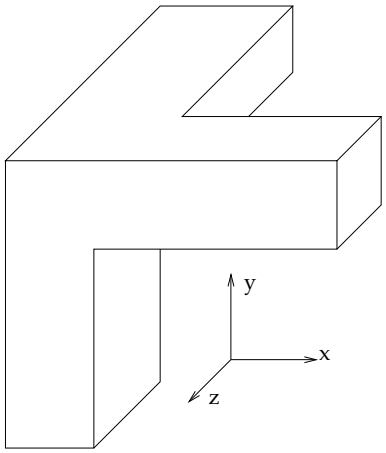


Figure 6.5: *Elementary cell of a three-dimensional square-rod structure, satisfying the perpendicularity condition (taken from [Lal98]).*

fore, the $\overrightarrow{E}$ formulation is better employed with the y and z components and the $\overrightarrow{H}$ formulation is better employed with the x component.

If perpendicular condition does not hold, other methods, like *super-Gaussian* functions [VP94], may become competitive to smooth the dielectric discontinuities.

Finally, speed improvements are achievable thinking that, when we are interested in studying the band diagram of a photonic crystal, we are usually focused on the first bands, at lower frequencies: therefore, we are only interested in the first few eigenvalues of the operator defined in 6.6. While finding all the eigenvalues of given matrix is a very complex problem, which scales with the cube of the dimension of the matrix ($\mathcal{O}\left[N^3\right]$), there are very efficient algorithms to find only the first few eigenvalues, where "first" is defined according to a given order relation: for example, the eigenvalues with the lowest real part or imaginary part or modulus. In this case, the complexity is super-linear ($\mathcal{O}\left[N^p\right]$ with $1 < p < 2$).

These algorithms are also very handy in finding the confined modes of a defect, embedded in an almost perfect lattice. Studying a defect with the plane expansion method is a more complicated task than studying band structures. To define it, we need a super-cell made of many fundamental cells of the crystal and one defect: the bigger the super-cell, the more accurate the results. In fact, as long as the boundary conditions in the PWE are intrinsically periodic, we are not actually studying one defect, but an infinite number of periodic defects in the super-cell lattice. To be confident that the results are correct, we must take care that all the defects are decoupled, i.e. that the super-cell is big enough. Obviously, bigger super-cells require more plane waves in the expansion and produce many more bands in the band diagram. Finding the resonant frequencies of the defect can involve finding tens of modes before them, which is a resource consuming task. It is possible,

though, to reformulate the problem 6.6 so that the first eigenvalues are around a given "target frequency" ω_0. Just let $\omega = \omega_0 + \omega'$ and rewrite 6.6:

$$\begin{aligned}\mathbf{A}_{\vec{k}}\left[\overrightarrow{H}_{\vec{k}}\right] &= \frac{(\omega_0+\omega')^2}{c^2}\overrightarrow{H}_{\vec{k}} \\ &= \frac{\omega_0^2 + 2\omega_0\omega' + \omega'^2}{c^2}\overrightarrow{H}_{\vec{k}} \\ \mathbf{A}'_{\vec{k}}\left[\overrightarrow{H}_{\vec{k}}\right] &= \frac{\omega'}{c^2}(\omega' + 2\omega_0)\overrightarrow{H}_{\vec{k}},\end{aligned}$$

with $\mathbf{A}'_{\vec{k}}[\bullet] = \mathbf{A}_{\vec{k}}[\bullet] - \frac{\omega_0^2}{c^2}$. The first few eigenvalues will be around the target frequency ω_0, which can be tuned to be around the interesting resonant frequency.

Unluckly, the operator in 6.11 has a much smaller condition number that the one in 6.6, therefore iterative methods converge more slowly: this is usually a convenient price to pay if a good estimate of ω_0 is available.

6.4 Examples and Validation

The present algorithm, as implemented by the Author, called BandSolver from now on, has been validated by comparison with two other programs, taken as references:

1. MPB [MPB], developed by Steven G. Johnson, based on the same Plane Wave Expansion algorithm;
2. CrystalWave [Cry], developed by Photon Design [Pho], based on the `FDTD` algorithm.

The validation tests are shown in Table 6.1: the first three tests are two-dimensional photonic crystals, the last one is a three-dimensional photonic crystal planar waveguide.

For all the band diagrams, where not explicitly specified, we have the wavevectors in abscissa and the normalized frequency, defined as $F \triangleq f\,c/\Lambda$, in ordinate. Λ is the lattice constant. Error is defined as the deviation from the reference:

$$\text{Error}[\%] = \frac{F_{\text{BandSolver}} - F_{\text{reference}}}{F_{\text{BandSolver}}} \times 100$$

Test 1

The structure considered in this test is a two-dimensional infinite triangular lattice of dielectric rods in air (Figure 6.6(a)). The lattice vectors are:

$$\overrightarrow{R}_1 = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \qquad \overrightarrow{R}_2 = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$$

The radius of the rods is $R = 0.2\Lambda$, their refractive index is $n_{cyl} = \sqrt{12}$.

Name	Dimension	Description		From
Test 1	2-D	Triangular lattice of rods in air	$R/\Lambda = 0.2$, $\epsilon_{sub} = 1.0$, $\epsilon_{cyl} = 12.0$	[MPB]
Test 2	2-D	Square lattice of rods in air	$R/\Lambda = 0.2$, $\epsilon_{sub} = 1.0$, $\epsilon_{cyl} = 12.0$	[JJ00]
Test 3	2-D	Triangular lattice of holes in dielectric (line defect)	$R/\Lambda = 0.35$, $\epsilon_{sub} = 10.24$, $\epsilon_{cyl} = 1.0$	[Cry]
Test 4	3-D	Triangular lattice (planar crystal waveguide)		[JJ00]

Table 6.1: *Validation tests.*

Figure 6.6(b) shows the reference results: the figure is taken from [MPB]. The k-path studied is $\Gamma - M - K - \Gamma$, in order to scan all the possible directions of the irreducible Brillouin zone [Kit95]. One bandgap for the TE polarization is visible for a normalized frequency F between 0.82 and 0.86, while two bandgaps for the TM polarization are visible for $0.28 < F < 0.44$ and $0.56 < F < 0.59$.

The results obtained with BandSolver are reported in Table 6.2, Figure 6.6(c) and Figure 6.6(d), for TE and TM polarizations. We can note the very good agreement with reference results, within 1.6%.

Figure 6.6(e) and Figure 6.6(f) show the z-component of the magnetic field of the Bloch mode for the TM polarization at the k-point M, first band.

Test 2

The structure considered in this test is a two-dimensional infinite square lattice of dielectric rods in air (Figure 6.7(a)). The lattice vectors are:

$$\overrightarrow{R}_1 = (1, 0), \quad \overrightarrow{R}_2 = (0, 1)$$

The radius of the rods is $R = 0.2\Lambda$, their refractive index is $n_{cyl} = \sqrt{12}$.

Figure 6.7(b) shows the reference results: the figure is taken from [JJ00]. The k-path studied is $\Gamma - X - K - \Gamma$, in order to scan all the possible directions of the irreducible Brillouin zone. No TE bandgaps are visible, while one large TM bandgap is present for $0.28 < F < 0.42$.

The results obtained with our implementation are reported in Table 6.3, Figure 6.7(c) and Figure 6.7(d), for TE and TM polarizations. We can note the very good agreement with reference results, within 1.2%. The small bandgaps in the TE polarization are not "real", but only the consequence of the discretization in the k-path.

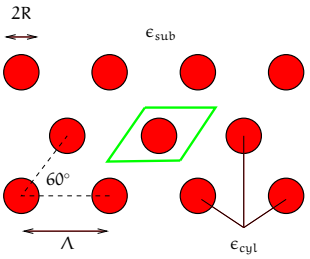
Band	k-point	$F_{BandSolver}$	F_{MPB}	Error [%]
1	Γ	0	0	0
	M	0.457	0.46	-0.6
	K	0.482	0.49	-1.6
2	Γ	0.57	0.565	0.8
	M	0.47	0.47	0
	K	0.55	0.55	0
3	Γ	0.785	0.79	-0.6
	M	0.68	0.55	-1.4
	K	0.55	0.55	0

(a) TE polarization.

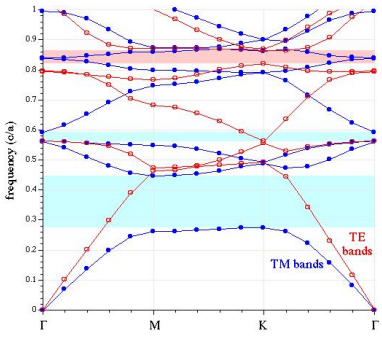
Band	k-point	$F_{BandSolver}$	F_{MPB}	Error [%]
1	Γ	0	0	0
	M	0.262	0.26	0.76
	K	0.275	0.275	0
2	Γ	0.563	0.565	-0.36
	M	0.446	0.446	0
	K	0.49	0.49	0
3	Γ	0.563	0.565	-0.36
	M	0.549	0.55	-0.18
	K	0.49	0.49	0

(b) TM polarization.

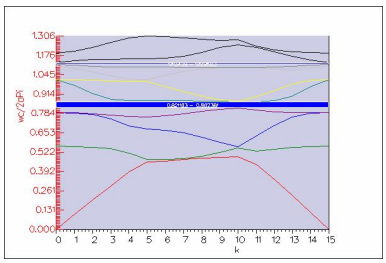
Table 6.2: *Test 1: TE and TM results. Note that the MPB results values have been extracted graphically from the available graph, so accuracy is not better that* 0.01*. Overall accordance is within* 1.6%*.*



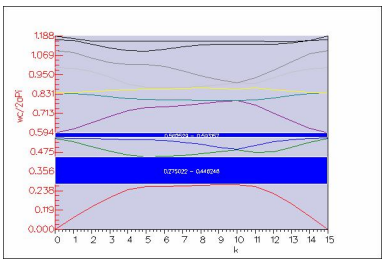
(a) Triangular lattice. In green, the fundamental cell.



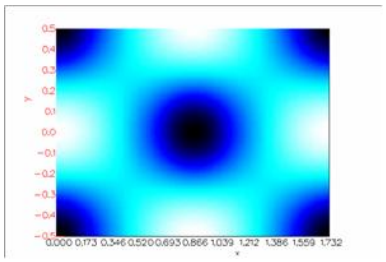
(b) Reference results from [MPB].



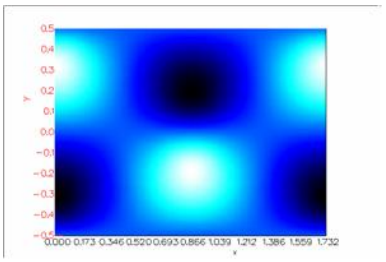
(c) BandSolver's TE band diagram.



(d) BandSolver's TM band diagram.



(e) BandSolver's $\mathbb{R}[H_z]$ of the Bloch mode at k-point M, first band



(f) BandSolver's $\mathbb{I}[H_z]$ of the Bloch mode at k-point M, first band

Figure 6.6: *Test 1.*

Band	k-point	$F_{BandSolver}$	F_{MPB}	Error [%]
1	Γ	0	0	0
	M	0.413	0.42	-0.7
	K	0.497	0.5	-0.6
2	Γ	0.55	0.565	1.2
	M	0.438	0.44	0.4
	K	0.586	0.59	-0.7
3	Γ	0.765	0.77	-0.7
	M	0.636	0.64	-0.6
	K	0.586	0.59	-0.7

(a) TE polarization.

Band	k-point	$F_{BandSolver}$	F_{MPB}	Error [%]
1	Γ	0	0	0
	M	0.243	0.24	1.2
	K	0.281	0.28	0.4
2	Γ	0.551	0.55	0.2
	M	0.417	0.42	-0.7
	K	0.495	0.50	-1.0
3	Γ	0.552	0.55	0.4
	M	0.554	0.555	-0.2
	K	0.495	0.50	-1.0

(b) TM polarization.

Table 6.3: *Test 2: TE and TM results. Note that the MPB results values have been extracted graphically from the available graph, so accuracy is not better that* 0.01*. Overall accordance is within* 1.2%*.*

Figure 6.7(e) and Figure 6.7(f) show the z-component of the magnetic field of the Bloch mode for the TM polarization at the k-point X, first band.

Test 3

The structure considered in this test is a line defect in a two-dimensional triangular lattice of air holes in a dielectric substrate (Figure 6.9(a)). The lattice vectors are:

$$\overrightarrow{R}_1 = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right), \quad \overrightarrow{R}_2 = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$$

The radius of the rods is $R = 0.309\Lambda$, their refractive index is $n_{sub} = 3.2$. Only the TE polarization is considered.

First, the bandgap of an infinite lattice, without defect, is computed. Results are shown in Figure 6.8(a), as computed by CrystalWave, and Figure 6.8(b), with BandSolver. Numerical comparisons are given in Table 6.4.

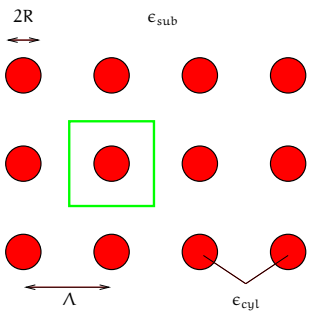
	First bandgap – low F	First bandgap – high F
BandSolver	0.226	0.289
CrystalWave	0.221	0.281

Table 6.4: *Comparison between BandSolver and CrystalWave, on the boundaries of the first bandgap for the TE polarization, for the complete lattice. Accordance is within 2%.*

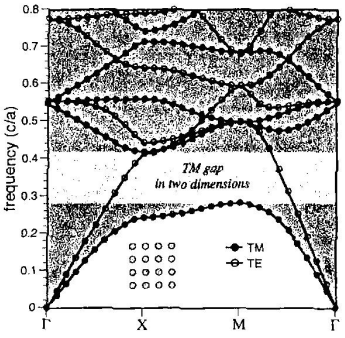
CrystalWave algorithm consists in propagating a broadband plane wave through the crystal, collecting the result and Fourier transforming it to get the spectral response of the device. It implicitly studies just one direction of propagation through the crystal, the one corresponding to the input planewave wavevector. In the example, it corresponds to k-point X. This procedure is somehow limiting, because in doesn't allow the user to study the full bandgap of the structure: a lattice could, in fact, present a not-complete bandgap, i.e. a bandgap only for certain directions, but non for all, and being misinterpreted for a complete one. Our algorithm, on the other hand, scanning all the directions of the irreducible Brillouin zone, avoids this problem and gives a complete grasp of the device's band structure.

Figure 6.8(a) shows the power transmitted through the device as a function of frequency. We can note a broad range of frequencies for which the transmission is zero: this is a bandgap in the direction of propagation of the input planewave. It correspond to the first bandgap (shown in blue) in Figure 6.8(b).

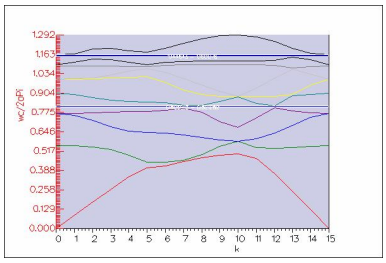
With these informations, the line defect can now be studied. To model it, a super-cell, as shown in Figure 6.9(a) in green, is taken and only the directions parallel to the channel's axis are studied, i.e. the wavevectors between Γ and X. The results obtained with BandSolver are reported in Figure 6.9(c), and they have to be compared with Figure 6.9(b). Note that CrystalWave shows unitary power for the frequencies inside the bandgap: they are guided lossless by the line defect



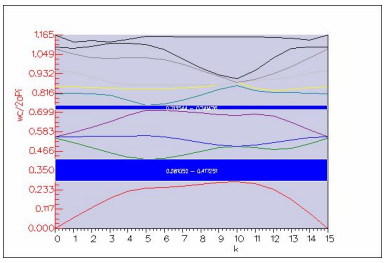
(a) Square lattice. In green, the fundamental cell.



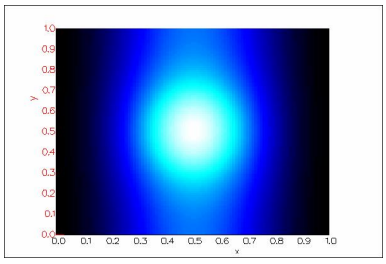
(b) Reference results from [JJ00].



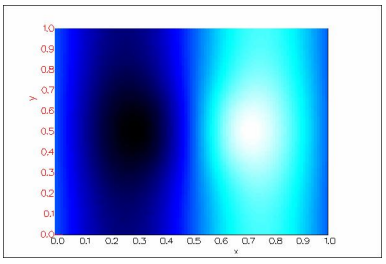
(c) BandSolver's TE band diagram.



(d) BandSolver's TM band diagram.

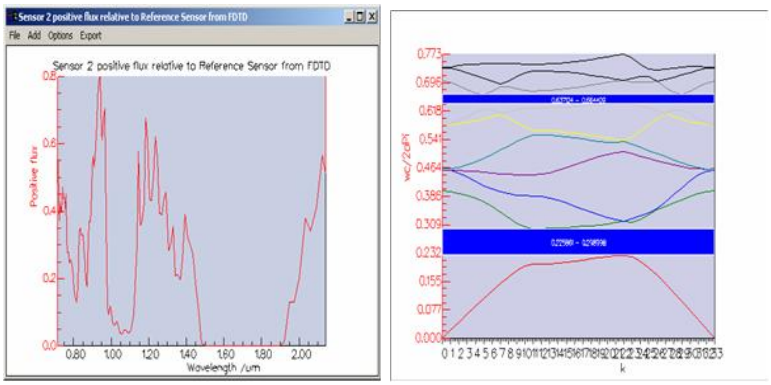


(e) BandSolver's $\mathbb{R}[H_z]$ of the Bloch mode at k-point X, first band



(f) BandSolver's $\mathbb{I}[H_z]$ of the Bloch mode at k-point X, first band

Figure 6.7: *Test 2.*



(a) Reference results computed by CrystalWave.

(b) BandSolver's TE band diagram.

Figure 6.8: *Test 3: complete lattice.*

through the crystal. The small deep at the middle of the bandgap corresponds to the crossing point between the even guided mode (whose characteristic has negative slope) and the odd mode (whose characteristic has positive slope) in Figure 6.9(c). It is due to the fact that, for that particular wavevector, the phase velocity of the excited mode (which is the even mode, in CrystalWave) is almost zero (i.e., its characteristic is almost flat): in a time-domain algorithm as CrystalWave's, this means that power takes very long time to get to the end of the device and to be computed in the FFT . Lasting the simulation longer would reduce the deep.

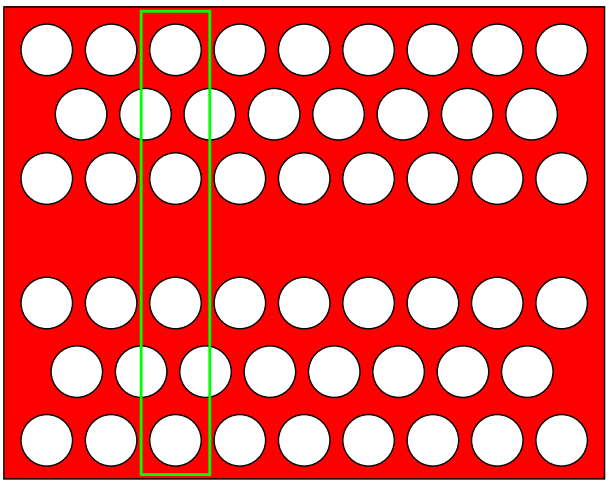
The z-component of the magnetic field of the fundamental guided mode in the super-cell is shown in Figure 6.9(d) and Figure 6.9(e).

Test 4

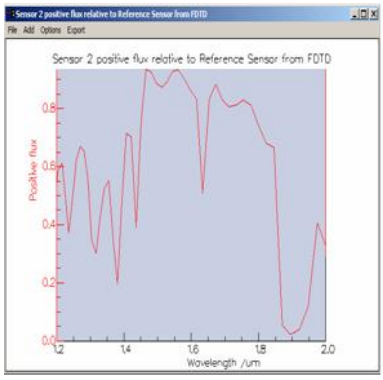
The structure considered in this test is a planar photonic crystal waveguide, made of a triangular lattice of dielectric rods in air, refractive index $n_{cil}^{high} = 3$, height $H = \Lambda$, and refractive index $n_{cil}^{low} = 2$ infinitely above and below, radius $R = 0.3\Lambda$. See Figure 6.10.

Figure 6.10(b) shows the reference results and Figure 6.10(c) the computed ones: Table 6.5 shows that the accordance is again very good.

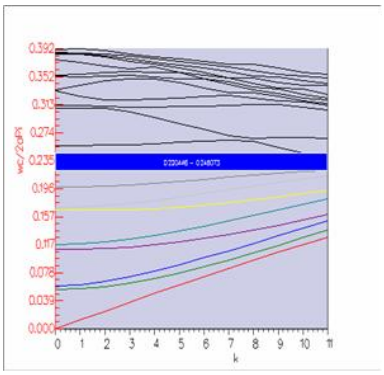
As an example, Figure 6.11 shows the profile of the real part of the Bloch mode at the reciprocal lattice point M, first band. The fields are plotted on an horizontal section passing through the high index slab and on a vertical section passing through the center of a unit cell.



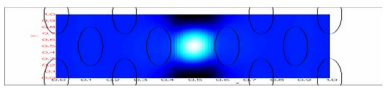
(a) Line defect. In green, the super-cell.



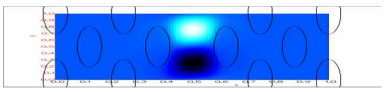
(b) Reference results computed by CrystalWave.



(c) BandSolver's TE band diagram.

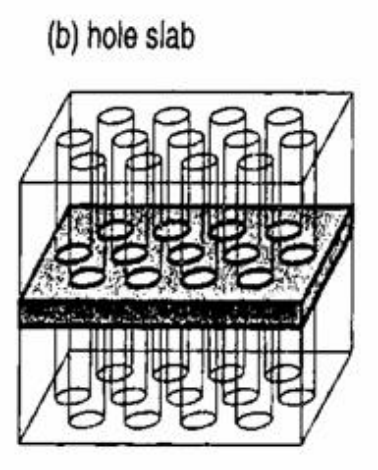


(d) BandSolver's Real H_z.

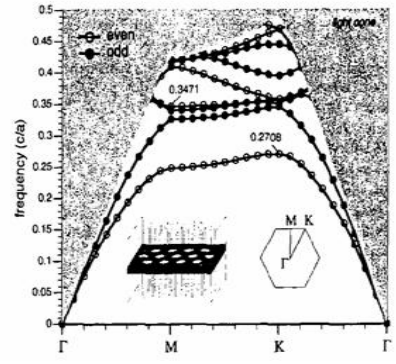


(e) BandSolver's Imag H_z.

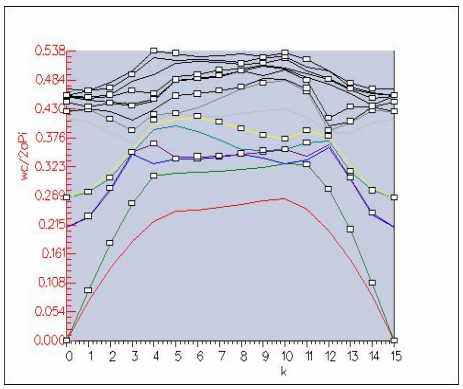
Figure 6.9: *Test 3: line defect.*



(a) Triangular lattice from [JJ00].



(b) Reference results from [JJ00].

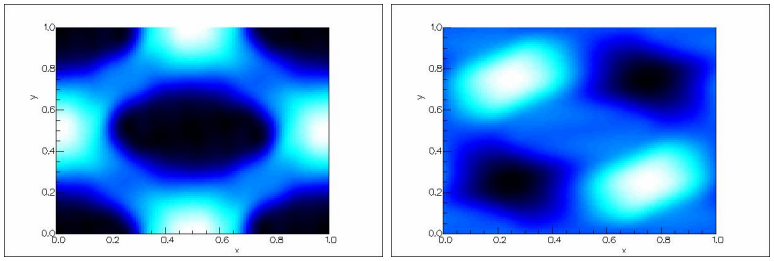


(c) BandSolver's band diagram.

Figure 6.10: *Test 4.*

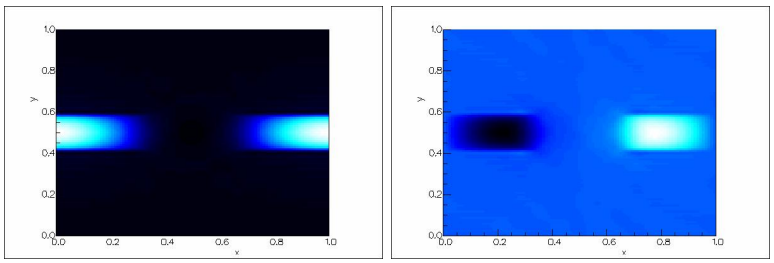
Band	k-point	$F_{our\ algorithm}$	F_{MPB}	Error [%]
1	Γ	0	0	0
	M	0.237	0.24	-1.3
	K	0.268	0.27	-0.7
2	Γ	0	0	0
	M	0.312	0.32	2.6
	K	0.328	0.34	-3.7
3	Γ	0.215	n/a	n/a
	M	0.328	0.34	-3.7
	K	0.328	0.34	-3.7

Table 6.5: *Test 4: comparison between the reference results and our algorithm's result. Overall accordance is within* 3.7%.



(a) Horizontal section, x component

(b) Horizontal section, y component



(c) Vertical section, x component

(d) Vertical section, y component

Figure 6.11: *Test 4: Real part of the Bloch mode, at k-point* M*, first band, computed by BandSolver. Values on the axis are normalized to the lattice constant* Λ.

A commercial implementation of BandSolver is available from Photon Design [Pho].

III

HIC Devices

What are the "HIC Devices"?

High Index Contrast (HIC) dielectric waveguides are optical devices in which the refractive indices of core and cladding differ greatly, much more than in conventional optical fibers. HIC devices exhibit, for this reason, highly confined optical modes and allow for waveguides to be spaced closely together without inducing crosstalk and the propagating field to be guided around sharp bends with minimal radiative loss [WH05]. This goes in the same direction as the need of integration of more and more functionalities inside optical chips, their grown in complexity and the will to reduce prices [HKR03].

However, as the index contrast is increased, the differences between the lateral boundary conditions for TE and TM modes become more pronounced, causing critical design parameters, such as propagation constants, coupling coefficients and scattering losses to be polarization dependent.

To allow polarization independent performance, a necessary feature for a standard single-mode-fiber-based communication link, the polarization sensitivity can be circumvented by the implementation of a polarization diversity scheme [Mad00]. Such an approach requires the input polarization, from the optical fiber, to be split into two orthogonal components. Then, rotation of one of them allows one single polarization to be realized on chip, on two parallel paths of identical devices. A possible scheme is reported in Figure 6.12.

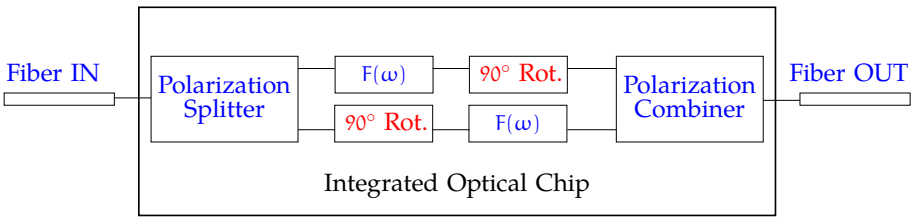


Figure 6.12: *Polarization diversity scheme. The function* $F(\omega)$ *performs some kind of operations on a polarized input signal. The splitter and the rotator before it ensure that the input polarization is the one the function was designed to work with. A second rotator is needed in the recombination stage. Note that the two rotation stages are split on the two arms of the scheme, to make their side effects symmetric (losses, crosstalk, etc.).*

In the next chapter, the study of a polarization rotator (the block in red in Figure 6.12) is reported.

7

Polarization Rotator

7.1 Introduction

Building a polarization rotator is probably the most challenging step in implementing a polarization diversity scheme. While several concepts for polarization splitters are available ([KAD05], [TCB+03], [WMG+04]), available polarization rotators can be subdivided in two big families, based on two opposite physical principles.

Adiabaticity , or mode-evolution principle: an input birefringent waveguide, in which the effective index of one polarization is greater than the other one, is adiabatically transformed into another birefringent waveguide, for which the effective indices associated to orthogonal polarizations are swapped. If the transition is adiabatic, power into one mode do not couple to other modes: if the higher index mode is excited at the input, power stays on the higher index mode, that, at the output, has the orthogonal polarization. There are many ways to achieve this transition, all of them sharing the characteristic that the transition from the input waveguide to the output one must be done asymmetrically: this is needed to prevent symmetry from forcing one polarization to stay in the same polarized mode. A very simple example of polarization rotator in the microwave regime is a rectangular waveguide twisted of 90° around its axis: if the twist is slow enough, the input TE mode (for example, with higher effective index) is "bended" into the output TM mode (again with the higher effective index) without reflections, and viceversa. In planar integrated optics, the twisting of a dielectric waveguide is not feasible, but "effective twisting" structures can be realized, based on the same idea: see [WH05] for a successful adiabatic design. The advantages of adiabaticity are a fabrication tolerant and inherently broadband design. The main disadvantage is a very long device, up to hundreds of microns for optical frequencies, necessary to achieve adiabaticity.

Mode coupling : an input mode, with a given polarization, is used to excite two or more modes of a device which support modes with hybrid polarization and different propagation constants. Power flows from one hybridly polarized

mode to the others and, after a fixed propagation length (the beat length), polarization is rotated as desired. To have complete rotation, i.e. no crosstalk, the hybrid modes must be phase matched, the hybrid modes linearly polarized at $\pm 45°$ and the total device length finely tuned. This results in a fabrication intolerant and wavelength sensitive device. Many authors have presented different designs: see [HSN+00], [TF96], [LHYH98], [KBM+05], [CdSHF03], [ERYJ04], for example. The main advantage over adiabatic devices is that such devices can be very short: in [KBM+05] a $1.6\mu m$ long polarization rotator is presented, which is also reasonably broadband. In fact, making a very short device has the double advantage of saving space on the optical chip and making the device broadband: if mode coupling only happens on a short distance, then the "filtering" property of the device is relaxed and so its frequency selectivity.

The device studied in this chapter is based on mode coupling: the advances of technological processes have convinced the Author of the possibility to achieve the same performances of adiabatic devices – in theory, 100% conversion – in a much shorter device. Moreover, a short polarization rotator could be used for other integrated devices, such as Solc filters [Sol65], and the same principle could be used to rotate the polarization of an arbitrary angle, not only 90°: definitely, a more flexible device than the adiabatic one.

The key difficulty in realizing it is to create a waveguide that supports two orthogonally polarized modes, A and B in Figure 7.1, at $\pm 45°$, with different effective indices $n_{eff}^{\pm 45}$. The resulting waveguide behaves as a half-wave plate: the input mode, polarized at 0° ($A + B$), or 90° ($A - B$), excites both of them with the same amount of power and, after a length $L_\pi = \frac{\lambda}{2|n_{eff}^{+45} - n_{eff}^{-45}|}$, corresponding to a π phase shift of the two modes A and B, the polarization has been rotated by 90°, as desired.

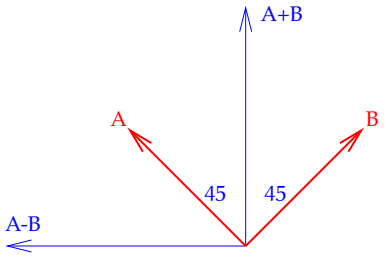


Figure 7.1: *$\pm 45°$ linearly polarized modes. An input vertically polarized mode $A + B$ excites both A and B: after L_π, A and B are phase shifted by π and the resulting field $A - B$ is horizontally polarized.*

This condition has been previously realized by etching a rib waveguide with one straight and one slanted sidewall [ERYJ04] [ERY03]. The resulting device is hundreds of microns long, though: not better, in length, than an adiabatic one.

The effect of the slanted wall is not strong enough to induce a large birefringence $\Delta n = n_{eff}^{+45} - n_{eff}^{-45}$, necessary for a small L_π.

Inspired from [CJN04], the effect of a slanting edge in the waveguide can be enhanced using a one-dimensional photonic crystal. The idea consists in using fabrication techniques typically used in photonic crystal devices to etch deep air slots, slated at a certain angle, into a conventional ridge waveguide.

The complete device, fabricated by the University of St. Andrews, is presented in [KBM+05] and [Bol05]. After a short description of the wafer, the design and the optimization process will be described and, finally, simulated and measured results will be compared.

7.2 Description of the Wafer

The InGaAsP heterostructure (grown by Alcatel) used in the experiments consists of a $0.3\mu m$-thick InP top cladding layer and a $0.522\mu m$ InGaAsP (Q 1.22) core layer, followed by InP lower cladding Figure 7.2.

A set of $230nm$ slots with a $650nm$ period was written using electron-beam lithography (Raith Elphy Plus/Leo 1530) in a $200nm$ thick layer of polymethymethacrylate. The pattern was then transferred into a hard silica mask using reactive ion etching with fluorine chemistry (CHF3). The silica mask was created using commercially available hydrogen silsesquioxane (HSQ) resist that was simply applied through spin coating. Baking this resist at high temperatures ($\sim 400°C$) partially transforms the film into silica.

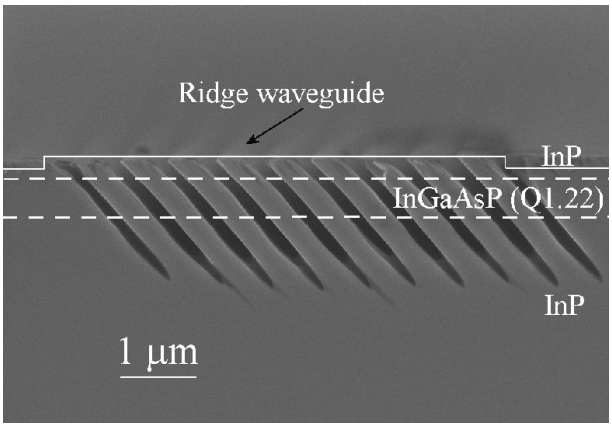
The deep etching of the slots was performed using a high voltage low current chemically assisted ion beam etching (CAIBE) regime at a temperature of $\sim 200°C$. The sample was mounted on a slanted holder in order to etch the slots at the desired angle of 45 degrees.

The access input/output ridge waveguides ($5\mu m$ wide) were defined using photolithography. The shallow etching ($100nm$) necessary for operation in a single mode (as shown in 7.3) was realized using a second stage of CAIBE etching with the photoresist as the protective mask. Finally, the sample was cleaved into die of approximately $1mm$ length.

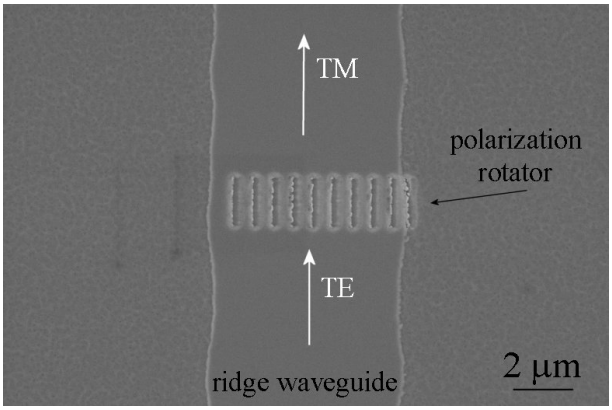
7.3 Design

Given the wafer described in the previous section, a suitable design for the polarization rotator must be studied. The optimized design will be, clearly, strongly wafer-dependent.

The waveguide chosen is a ridge waveguide, as the one shown in Figure 7.3. The ridge width is $5\mu m$, etched by $100nm$ from the air top: these dimensions ensure the monomodality of the waveguide. In Figure 7.4 the first TE and first TM modes of the waveguide without the air slots are shown, computed using the vectorial mode solver described in 5.3. All the other higher modes are cutoff.



(a) Cross section.



(b) Top view.

Figure 7.2: *Experimental device. Compare it with Figure 7.3.*

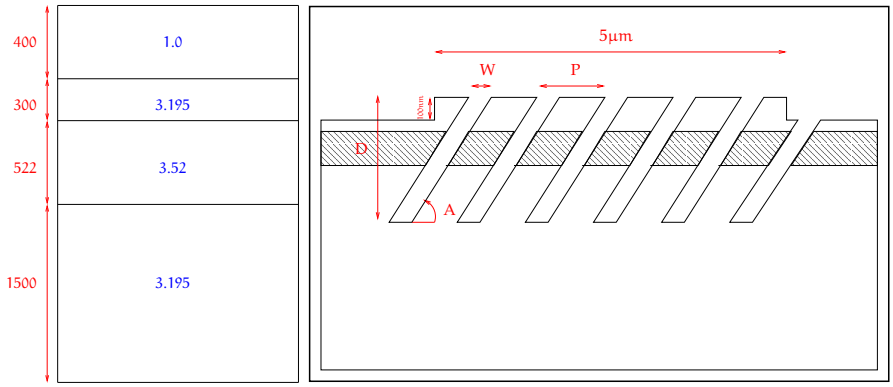


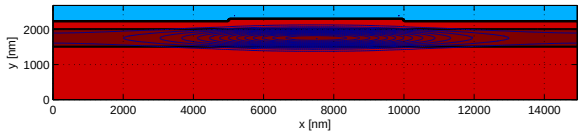
Figure 7.3: *Wafer's cross section. Compare it with Figure 7.2. All the dimensions are in* nm *if not explicitly stated. The optimization parameters are shown:* W*, the air slots width,* A*, the air slots etch angle and* P*, the air slots periodicity. The etch depth* D *is fixed at* 1.6μm.

In Figure 7.3, the optimization parameters are shown. There are four degrees of freedom in the design of the air slots:

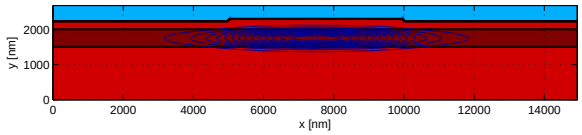
1. W, the width;
2. P, the periodicity;
3. A, the etch angle;
4. D, the etch depth.

In choosing them, to simplify the design, some design hints are taken into account:

- the fabrication process can only etch at some given angles: only angles between 40° and 50° are practically feasible;
- the periodicity P is always greater than the air slots width W, by definition, but a $P \gg W$ resembles the case of a common slab waveguide, which is well know to support only pure TE and pure TM modes: no ±45° linearly polarized modes are expected in this case;
- the larger the air slots, the more mismatched the device will be with the input ridge waveguide: all the light that is reflected at the interface between the input waveguide and the polarization rotator is lost, hence not converted;
- the deeper the air slots, the better, because coupling with the substrate is prevented: a feasible working depth of 1.6μm has been chosen and fixed from now on for the optimization: in Figure 7.9 this choice is justified;



(a) First TE mode.



(b) First TM mode.

Figure 7.4: *Contour plot of the longitudinal component of the Poynting vector, for the two ridge waveguide modes. The ridge width (5µm) and depth (100nm) are chosen to have monomodal operation.*

- the aspect ratio between the etching depth and the air slots W/D (or the etching depth and the periodicity P/D) is critical for the quality of the air slots. Very small aspect ratios are difficult to achieve: non-straight walls, wrong widths and depths are the consequences;
- the rotation effect increases with the number of the air slots along the cross section of the ridge waveguide.

Each design parameter has a range of possible values, given by the fabrication process and the hints above: they are reported in Table 7.1.

Parameter	From	To
W	$260\mathrm{nm}$	$280\mathrm{nm}$
A	40°	50°
P	$600\mathrm{nm}$	$700\mathrm{nm}$

Table 7.1: *Optimization parameters.*

The optimization has been done iterating on all the possible combinations of these parameters, looking for a global maximum of a certain cost function. In our case, the cost function has been thought to weight the most important characteristic that the device must have to work properly. In particular, let A and B be the two modes of the structure, more similar to the ideal $\pm 45^\circ$ linearly polarized modes, polarized at angles α_A and α_B respectively. Then:

1. the polarization angle difference between A and B must be equal to: $\alpha_A - \alpha_B = \pm 90^\circ$;
2. the polarization angle sum between A and B must be equal to: $\alpha_A + \alpha_B = 0^\circ$ or 180°;
3. the mode power P_A and P_B (normalized to unity) must be concentrated mainly in the core.

The first two points refer to the fact that to have 100% conversion the modes must be linearly polarized at $\pm 45^\circ$. Their importance weights 0.4 each in the cost function. The last point endsures that the modes are guided by the core and not by the cladding. This weights 0.2. The final weight function C is:

$$\begin{aligned} C \triangleq\ & 0.4\left(\frac{|\alpha_A - \alpha_B|}{90}\right) \\ & + 0.4\left(1 - \frac{\alpha_A + \alpha_B}{180}\right) \\ & + 0.2\left(\frac{P_A + P_B}{2}\right) \end{aligned} \tag{7.1}$$

and it can have values between 0 and 1.

Figure 7.5 shows the cost function C for different choices of the optimization parameters. As long as C is a function of three parameters W, A and P, for ease of visualization two-dimensional sections are shown: one for constant A and one for P.

We can note that there is a maximum for $W = 266\text{nm}$[1], $A = 45°$ and $P = 650\text{nm}$: the cost function is $C = 0.97$. The two modes obtained in this case are shown in Figure 7.6. They have the effective indices:

$$n_{eff}^{+45} = 3.046 \qquad n_{eff}^{-45} = 2.604$$

and the polarization angles:

$$\alpha^{+45} = 45.1° \qquad \alpha^{-45} = 42.5°.$$

The expected beat length is $L_\pi = 1.5\mu\text{m}$ at $\lambda = 1300\text{nm}$.

7.4 Results

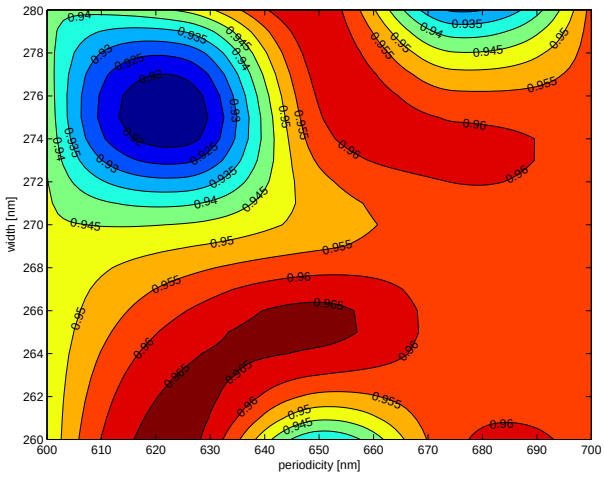
A tunable laser operating between 1250nm and 1365nm was used to launch light into the device via a microscope objective. The polarization of the incoming light was set to either TE or TM, with an analyzer on the output side. Devices with polarization rotator sections varying from 0μm (blank waveguide) to 4μm in length were tested.

Figure 7.7(a) presents the experimental dependence of the output power in TE polarization on the length of the polarization rotator for both TE and TM incoming polarizations. The optimal length at which the output polarization is rotated by ~ 95% with respect to the incoming polarization is approximately 1.6μm for this type of slanted grating. At a length of about 3.2μm, the incoming polarization is rotated through 180 degrees returning the output light to the initial polarization.

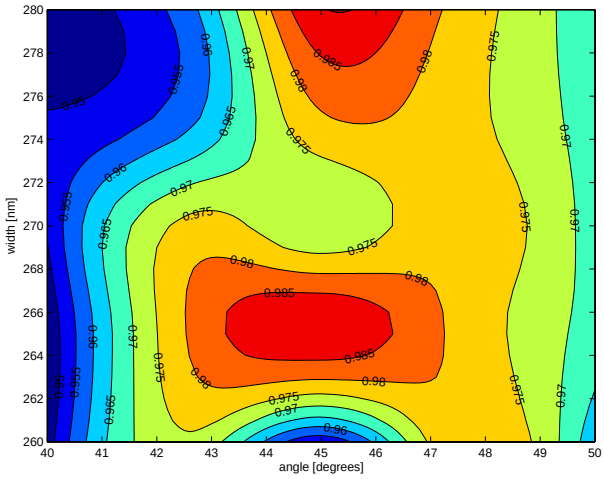
A good agreement between the simulation performed with a full 3D-FDTD and the experimental results can be seen in Figure 7.7(a), assuming a slot width of 270nm. This width is slightly larger than the experimentally determined width of 230nm, although there is some experimental error due to the variation of slot width with etch depth. In the following, we therefore refer to an "effective" width of 270nm for the air slots.

Figure 7.7(b) depicts the absolute normalized output power as a function of the analyzer angle for a blank waveguide and a waveguide containing a 1.5μm long polarization rotator (the incoming light was TE polarized for both cases). As can be seen from the figure, the maximum output power (100%) is in TE polarization (0 degrees) for a blank waveguide. At the same time, 96% of the output light is in TM polarization (~ 90°) with only 4% remaining in TE polarization for the waveguide

[1]The actual parameter passed to the fabrication process is $W = 270\text{nm}$, within the fabrication tolerances.



(a) P-W section at $A = 45^\circ$ constant.



(b) A-W section at $P = 650\text{nm}$ constant.

Figure 7.5: *Cost function* C *as defined in* (7.1)*, for different choices of the optimization parameters. The chosen maximum is for* $W = 266\text{nm}$*,* $A = 45^\circ$ *and* $P = 650\text{nm}$*, where* $C = 0.97$*.*

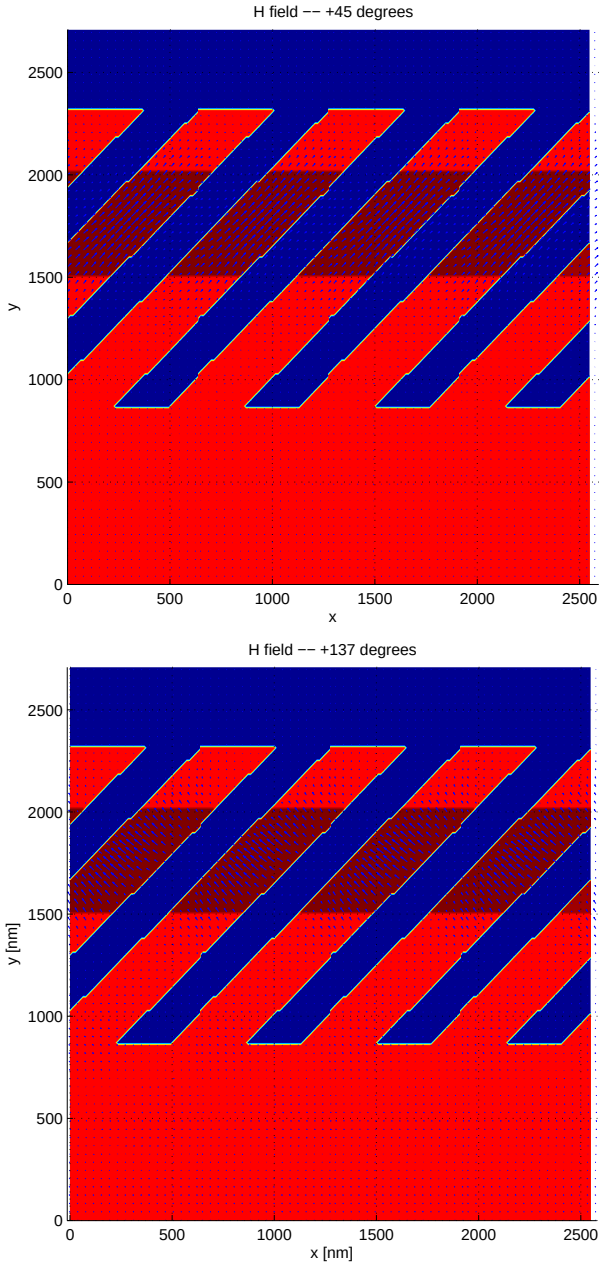
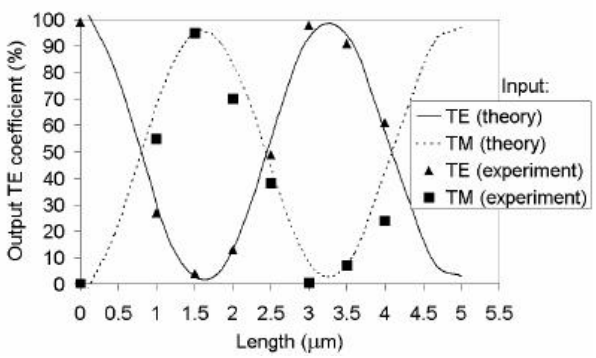
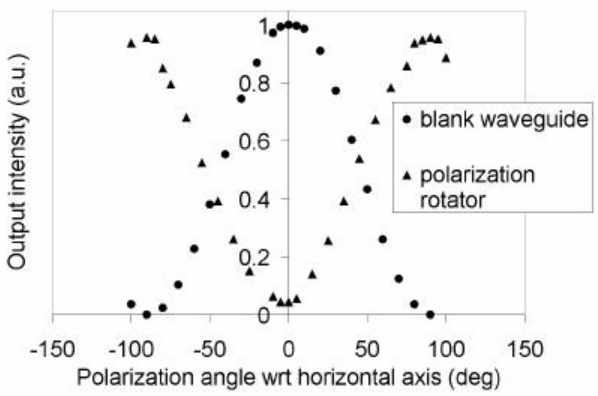


Figure 7.6: *Magnetic field of the optimum modes.*



(a) TE fraction of the output light *versus* the length of the polarization converter for both TE and TM input polarizations. The dots are the experimental data, the lines are the results of the `3D-FDTD`



(b) Output power *versus* polarization angle for TE incoming light.

Figure 7.7: *Comparison between `3D-FDTD` and experimental data.*

with the 1.5µm long polarization rotator. This represents an extinction ratio of 14dB.

For a given width of air slots, the optimal performance of the device strongly depends on the air/semiconductor width ratio (this ratio is ~ 1 : 1.5 in the case depicted in Figure 7.7(a)). Indeed, if the semiconductor width increases, the grating tends to resemble a uniform slab waveguide with tilted walls and the form birefringence is much reduced. The structure then approximates the type of asymmetric waveguide studied previously [TCB+03]. Increasing the semiconductor width for fixed air slots reduces the form birefringence due to a decreased difference between the effective indices of the two orthogonal eigenmodes. This leads to a longer beat length as illustrated in Figure 7.8. For a fixed width of air slots, there is an almost cubic dependence of the beat length on the air/semiconductor ratio. In the other limit, for very small semiconductor width, the form birefringence increases, but the interface reflectivity also goes up due to increased mismatch with the input waveguide. Device fabrication also becomes technologically more demanding for large air/semiconductor ratios. This simulated trend is supported by experimental evidence. A device with a 1 : 4 air/semiconductor ratio and the other parameters unchanged (270nm of air slots effective width, 1080nm semiconductor width) showed an optimal length of 30µm Figure 7.8.

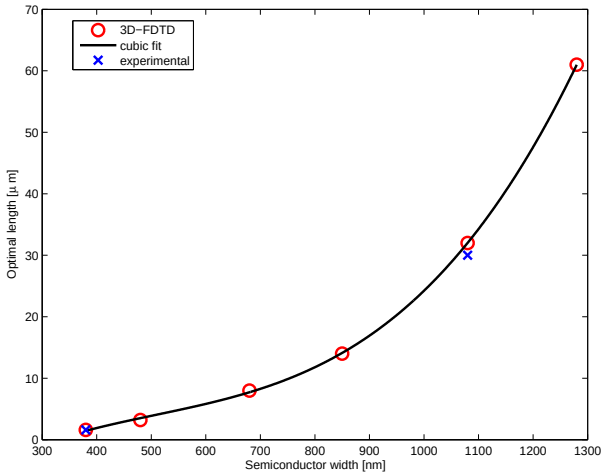


Figure 7.8: *Simulated dependence of the beat length on the width of the semiconductor section for a 270nm width of the air slots. Filled dots: simulated values. Solid line: cubic fit. Crosses: experimental values.*

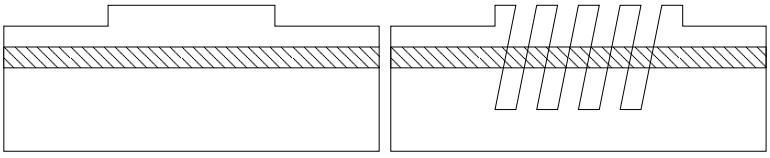
In terms of transmitted power, we currently observe a loss of 3dB (compared to a ridge waveguide without polarization rotator). The corresponding theoretically predicted losses are around 1.5dB for deep (~ 2µm) slanted slots with perfectly parallel sidewalls and 3dB for similar slots with conical walls, as shown in Figure

7.9.

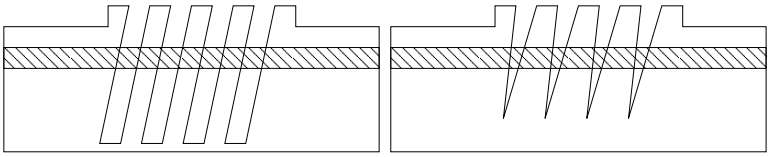
Thus the main source of the loss is considered to lie in the imperfect, irregular or conical shape of the holes as well as insufficiently deep etching. Both of these problems result in undesirable out-of-plane scattering. These issues, however, can be minimized by optimizing the etching regime and achieving deeper, more parallel slots. As technology continuously improves, better slots should be achievable in the near future. The remaining loss of 1.5dB is due to the abrupt interface between the slotted and unslotted sections. Further improvements in the interface design, such as more adiabatic transitions, will reduce these losses to acceptable levels.

With respect to bandwidth, it is worth noting that the experiments were performed at several different wavelengths in the 1290 to 1330nm range, with no compromise in performance; this clearly indicates broadband operation, which is also supported by the predicted wavelength window of 200nm in Figure 7.10(b), in which the extinction ratio is better than 15dB.

Finally, in Figure 7.10(a) the polarization coefficient, defined as the ratio between the power in one polarization over the total power, as a function of the polarization rotator length is shown. We can note that after 1.6μm almost all the power has passed from TE to TM. 1.6μm, as predicted by the 3D-FDTD , slightly differs from the 1.5μm predicted by the mode solver, but greatly agrees with experimental results.



(a) No holes. (b) Normal holes.



(c) Deeper holes. (d) Conical holes.

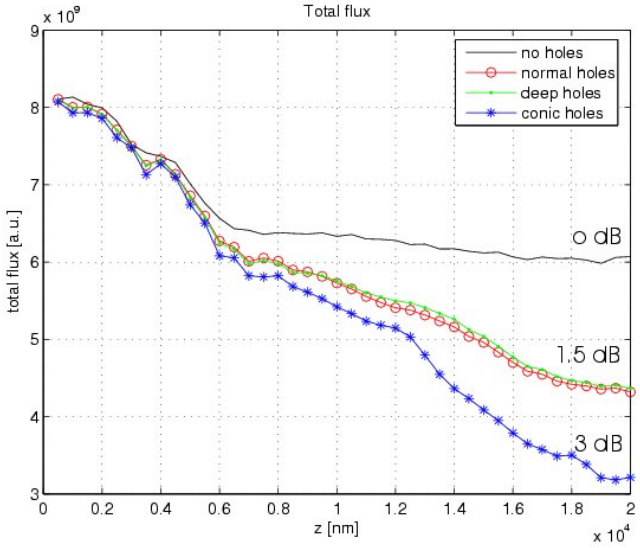
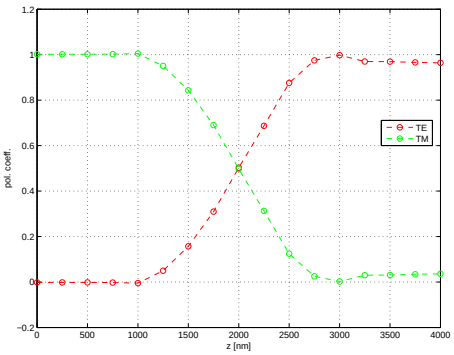
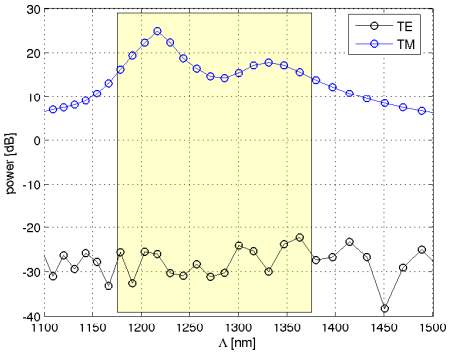


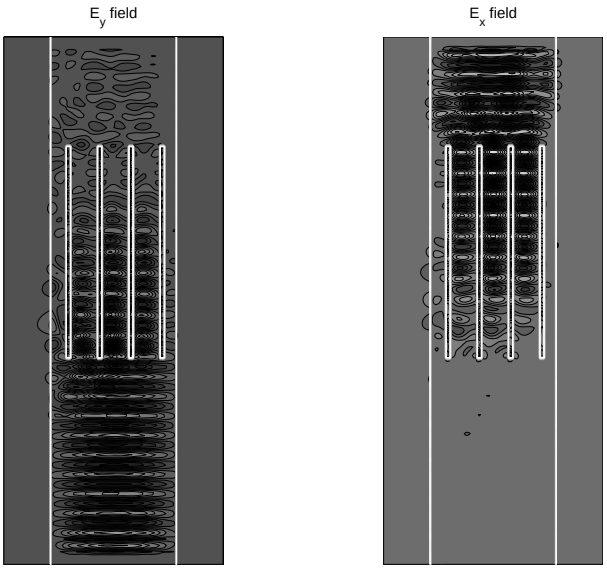
Figure 7.9: *Observed experimental losses are around* 3dB*, while the* 3D-FDTD *predicted losses are around* 1.3dB *for deep slanted slots with perfectly parallel sidewalls. In the graph, losses for other kind of air slots are shown. It can be noted that* 3dB *losses are found for air slots with conical walls, like the one in Figure 7.2. Moreover, making the holes deeper has no effect:* 1.6µm *etching is enough for losses. Note also that the initial part of the graph as little meaning: not a fundamental mode of the input ridge waveguide is inputted, therefore some power is radiated in the first* 10µm*.*



(a) Polarization coefficient (defined as the ratio of the power in one polarization over the total power), for a TE input field.



(b) Spectrum of the device: we achieve 15dB extinction ratio over 200nm of bandwidth (yellow region).



(c) Field plot (not in scale). The input TE polarization is transformed into the output TM polarization.

Figure 7.10: *3D-FDTD results.*

IV

Appendices

A

Notation and Nomeclature

Throughout the text, a coherent notation has been employed. Table A.1, Table A.2, Table A.3, Table A.4 show the notation used.

Element	Example
Vector	$\overrightarrow{v}$
Vector in time-domain	$\widetilde{\overrightarrow{v}}$
Versor	$\widehat{v}$
Norm	$\left\|\overrightarrow{v}\right\|$
Dot Product	$\overrightarrow{x} \cdot \overrightarrow{y}$
Cross Product	$\overrightarrow{x} \times \overrightarrow{y}$
Gradient	$\nabla \bullet$
Curl	$\nabla \times \bullet$
Divergence	$\nabla \cdot \bullet$
Laplacian	$\nabla^2 \bullet$
Transverse Laplacian	$\nabla_T^2 \bullet$

Table A.1: *Notation for vector calculus.*

Element	Example
Array	$\mathbf{x}$
Matrix	$\mathbf{M}$
Transpose Matrix	$\mathbf{A}^{\mathsf{T}}$
Hermitian Matrix	$\mathbf{A}^{*}$
Operator	$\mathbf{A}\left[\bullet\right]$
Fourier Transform	$\mathcal{F}\left[f\right]$
Tensor	$\underline{\epsilon}$
Generic Set	$\mathcal{A}, \mathcal{B}, \ldots$
Empty Set	$\{\}$
Positive Integers	$\mathbb{N}$
Integers	$\mathbb{Z}$
Rationals	$\mathbb{Q}$
Reals	$\mathbb{R}$
Imaginary Numbers	$\mathbb{I}$
Complex Numbers	$\mathbb{C}$

Table A.2: *Notation for matricial calculus and sets.*

Element	Example
Conjugate	z^{*}
Absolute Value	$\left\|z\right\|$
Argument	$\angle z$
Real Part	$\mathbb{R}\left[z\right]$
Imaginary Part	$\mathbb{I}\left[z\right]$
Discretization in space-time of a function f	${}^{n}f^{i,j,k}$
Evaluation of a function f	$f\vert_{x=x_0}$

Table A.3: *Notation for mixed mathematical functions.*

Element	Example
Instant in time	t
Interval in time	τ
Point and Node	n
Line and Edge	e
Face and Surface	f
Volume and Cell	v
Dual element	$\widetilde{x}$
External orientation	$\widetilde{x}$
Measure	$\left\|x\right\|$

Table A.4: *Notation for geometrical elements.*

B

Maxwell House

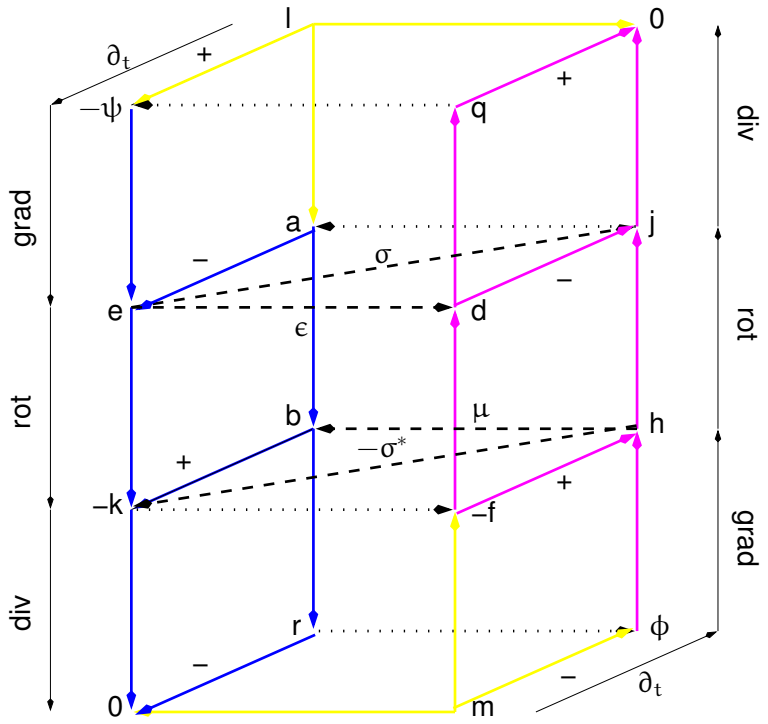


Figure B.1: *Maxwell's House*

Figure B.1, inspired from [Bos98a], tries to represent graphically the symmetry of Maxwell equations. The nodes of the grid represent physical quantities, while edges are the relations between them. The notation used is the same as in [Bos98a].

In Table B.1, the equations that can be read from the diagram are listed.

Vector Potentials	$b = \nabla \times a$	$d = -\nabla \times f$
	$e = -\partial_t a - \nabla \psi$	$h = -\partial_t f - \nabla \phi$
Gauss Law	$\nabla \cdot d = q$	$\nabla \cdot b = r$
Faraday Equation	$\partial_t b + \nabla \times e = -k$	$-\partial_t d + \nabla \times h = j$
Ampère Law	$-\partial_t d + \nabla \times h = j$	$\partial_t b + \nabla \times e = -k$
Charge Conservation	$\nabla \cdot j + \partial_t q = 0$	$\nabla \cdot k + \partial_t r = 0$
Lorenz Gauge	$\nabla \cdot a + 1/c^2 \partial_t \psi = 0$	$\nabla \cdot f + 1/c^2 \partial_t \phi = 0$

Table B.1: *Equations from Figure B.1.*

C

Barycentric Coordinates

Barycentric coordinates are triples of numbers (t_1, t_2, t_3) corresponding to masses placed at the vertices of a reference triangle $A_1A_2A_3$. These masses then determine a point P, which is the geometric centroid of the three masses and is identified with coordinates (t_1, t_2, t_3). The vertices of the triangle are given by $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ [Mat, Cox69].

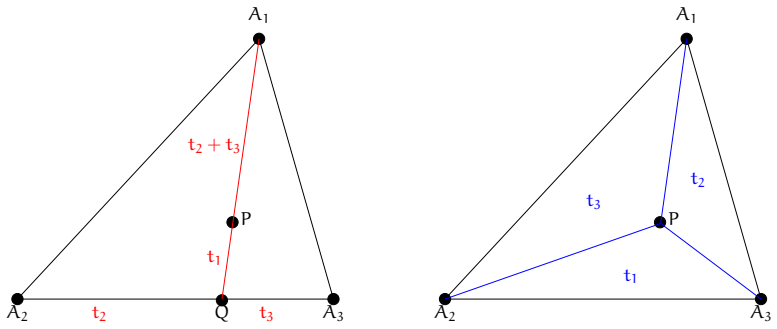


Figure C.1: *Barycentric coordinates.*

To find the barycentric coordinates for an arbitrary point P, find t_2 and t_3 from the point P at the intersection of the line A_1P with the side A_2A_3, and then determine t_1 as the mass at A_1 that will balance a mass $t_2 + t_3$ at Q, thus making P the centroid (see Figure C.1, left). Furthermore, the areas of the *oriented* triangles A_1A_2P, A_1A_3P and A_2A_3P are proportional to the barycentric coordinates t_3, t_2 and t_1 of P (see Figure C.1, right).

Barycentric coordinates are homogeneous, so:

$$(t_1, t_2, t_3) = (\mu t_1, \mu t_2, \mu t_3),$$

for $\mu \neq 0$ and normalized:

$$t_1 + t_2 + t_3 = 1,$$

so that the coordinates give the areas of the oriented subtriangles normalized by the area of the original triangle. Therefore, they are also called *areal coordinates*.

Barycentric coordinates for a number of common centers are summarized in Table C.1

circumcenter	$(a^2(b^2+c^2-a^2), b^2(c^2+a^2-b^2), c^2(a^2+b^2-c^2))$
incenter	(a, b, c)
center of mass	$(1, 1, 1)$

Table C.1: *Barycentric coordinates for some common centers.* a, b *and* c *are the side lenghts of the triangle.*

A point P is internal to a triangle T if all its three barycentric coordinates, with respect to T, are positive. If one of them if 0, the point lies on the perimeter of T.

C.1 Interpolation of a Nodal Function

Given a function f defined on the points A_1, A_2 and A_3 of the triangle T, called a *nodal function*, we can linearly interpolate it on all the points internal to T, using the barycentric coordinates of $P = (t_1, t_2, t_3)$. In formulas:

$$f(P) = t_1 f(A_1) + t_2 f(A_2) + t_3 f(A_3).$$

This ensure that the function f is continuous across two adjacent triangles and that its greatest value lies on one of the vertices of T. Geometrically, the value of f in P is the sampled value of the linearized version of f, a plane, which passes through the points $(x_{A_1}, y_{A_1}, f(A_1))$, $(x_{A_2}, y_{A_2}, f(A_2))$ and $(x_{A_3}, y_{A_3}, f(A_3))$ (see Figure C.2).

This also leads us to define a *barycentric function* associated to the point P, i-th node of a given mesh $\mathcal{K}$, as the function λ^i, piecewise linear on the whole domain where $\mathcal{K}$ is defined, equal to 1 in P and 0 on all the other nodes of the mesh. So, the position x of the point P is:

$$x = \sum_{i \in \{A_1, A_2, A_3\}} \lambda^i(x) x_i$$

and

$$1 = \sum_{i \in \{A_1, A_2, A_3\}} \lambda^i(x)$$

The λ^is are nothing else than the *hat functions* or *Lagrange elements of polynomial degree 1*, of Finite Element theory [BM89].

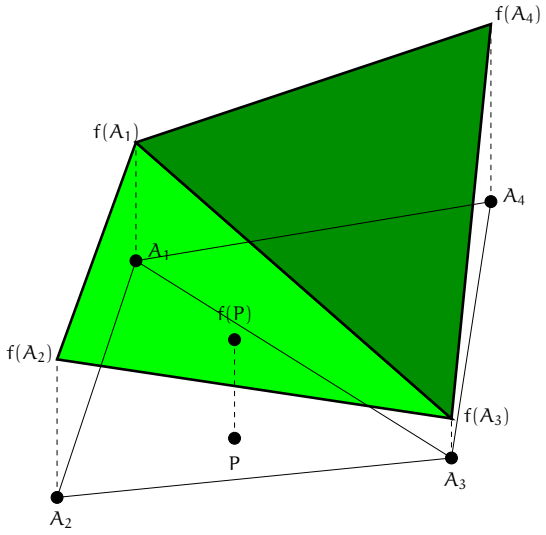


Figure C.2: *Interpolation of a nodal function using barycentric coordinates. Continuity of* f *is ensured by the common edge* A_1A_3.

C.2 Interpolation of an Edge Function

Let f be a vectorial field, whose values are only known as line integrals over the edges of a given mesh $\mathcal{K}$. The values of f in each point inside the mesh can be computed following this procedure.

Let $x_i x_j x_k$ be a triangle, and x, y two points inside it. Using barycentric function defined above, we can write:

$$x = \sum_n \lambda^n(x)\, x_n$$
$$y = \sum_m \lambda^m(y)\, x_m$$

Let xy be the oriented segment that goes from y to x: $xy = y - x$. We can write:

$$\begin{aligned}
y - x &= y - \sum_n \lambda^n(x)\, x_n \\
&= \sum_n \lambda^n(x)\,(y - x_n) \\
&= \sum_n \lambda^n(x) \left(\sum_m \lambda^m(y)\, x_m - x_n \right) \\
&= \sum_n \lambda^n(x) \sum_m \lambda^m(y)\,(x_m - x_n) \\
&= \left[\lambda^i(x)\,\lambda^j(x) - \lambda^j(x)\,\lambda^i(x)\right](x_j - x_i) + \\
&\quad \left[\lambda^j(x)\,\lambda^k(x) - \lambda^k(x)\,\lambda^j(x)\right](x_k - x_j) + \\
&\quad \left[\lambda^k(x)\,\lambda^i(x) - \lambda^i(x)\,\lambda^k(x)\right](x_i - x_k) \\
&= c_{ij}(x_j - x_i) + c_{jk}(x_k - x_j) + c_{ki}(x_i - x_k)
\end{aligned}$$

The coefficients c_{ij} have also a geometrical interpretation (see Figure C.3):

$$\begin{aligned}
c_{ij} &= \frac{\text{Area}(x, y, x_k)}{\text{Area}(x_i, x_j, x_k)} \\
c_{jk} &= \frac{\text{Area}(x_i, x, y)}{\text{Area}(x_i, x_j, x_k)} \\
c_{ki} &= \frac{\text{Area}(y, x_j, x)}{\text{Area}(x_i, x_j, x_k)}
\end{aligned}$$

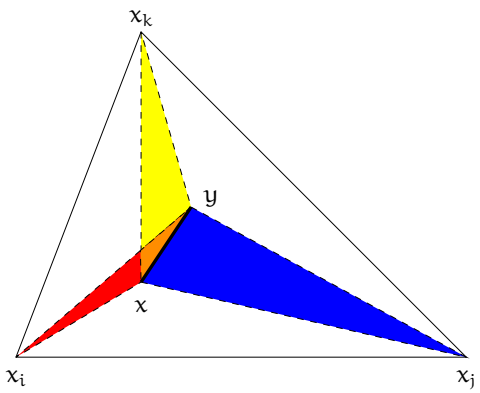


Figure C.3: *Interpolation of an edge function.*

Note that the operator Area returns the *oriented area* of a given triangle, according to the order of its vertices.

Now, as for the barycentric coordinates, we can interpolate the value of an edge function f on the edge xy, if we know its values over the edges x_ix_j, x_jx_k and x_kx_i, by the following formula:

$$f(xy) = c_{ij}f(x_ix_j) + c_{jk}f(x_jx_k) + c_{ki}f(x_kx_i)$$

If we finally want to know the value of the vector defined at a given point x, we can use this procedure to compute the value of its components parallel to the coordinate axes: just choose $y = x + (dx, 0)$ or $y = x + (0, dy)$ (in two dimensions) to find the components along x and y and divide by dx or dy, respectively. The choice of the values dx and dy depends on the particular precision wanted and available.

This procedure closely resembles the definition of *edge-elements* in Finite Element theory [BM89], [LS95], [Web93], [Bos98b], [LLC97], [SLC01], [Web99], [BK00]

D

Geometrical Interpretation of the Courant Factor

Consider the one-dimensional wave equation [Num]:

$$\partial_t^2 u = v^2 \partial_x^2 u.$$

Discretized with a central difference scheme both in time and in space, the resulting equation is conditionally stable, according to the *Courant* stability factor $S \leq 1$:

$$\frac{|v|\Delta t}{\Delta x} = S.$$

From the figure D.1, all the events in the space-time cone (shown in blue) influence the value of u at the point $\left(x^i,\ {}^n t \right)$. The aperture of the cone is velocity of propagation of the events. The cone is called *cone of events*.

The discretization in the space-time leads to a stable algorithm only if the cone of events for the discretized space includes the cone of events for the real space. If not, we can qualitatively say that the discretized differential equation is not "fast enough" to bring the informations to the real physical equation: information is lost and instabilities arise.

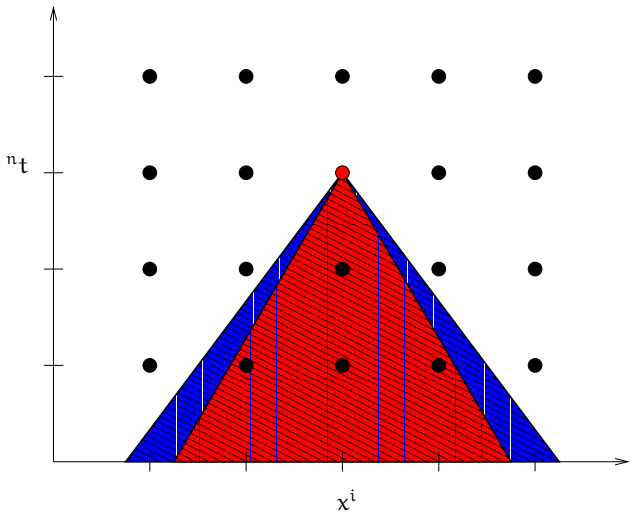


Figure D.1: *Space-time cone of events in* $(x^{i}, {}^{n}t)$ *for the discretized (in blue) and the real (red) space. All the points in the blue cone affect the value of the field in the red point, vertex of the cone. In this example, the stability is assured because the discretized cone includes the real one.*

List of Figures

List of Tables

Bibliography

[ADD96] P. R. Amestoy, T. A. Davis, and I. S. Duff. An approximate minimum degree ordering algorithm. *SIAM Journal of Matrix Analysis and Applications*, 17(4):886–905, 1996.

[AWDK05] M. Ayre, T. J. Lijun Wu, T. Davies, and T. F. Krauss. Experimental verification of numerically optimized photonics crystal injector, Y-splitter, and bend. *IEEE Journal on Selected Areas in Communications*, 23(7):1390–1395, July 2005.

[BBC+94] R. Barrett, M. Berry, T. F. Chan, J. Demmet, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, PA, 1994. `http://www.netlib.org/templates/Templates.html`.

[Ben02] M. Benzi. Preconditioning Techniques for Large Linear Systems: A Survery. *Journal of Computational Physics*, 182:418–477, 2002.

[BF04] L. Bolla and T. Felici. A New Discretisazion Scheme for Frequency Domain Electromagnetics. In *Progress in electromagnetics research symposium (PIERS 2004 Proceedings)*, Pisa, March 2004.

[BK00] A. Bossavit and L. Kettunen. Yee-like Schemes on Staggered Cellular Grids: A Synthesis Between FIT and FEM Approaches. *IEEE Transactions on Magnetics*, 36(4):861–867, July 2000.

[BM89] A. Bossavit and I. Mayergoyz. Edge-elements for scattering problems. *IEEE Transactions on Magnetics*, 25:2816–2821, July 1989.

[BMS04] L. Bolla, M. Midrio, and C. G. Someda. Energy flow in negative index materials. *Chinese Optics Letters*, 2(7):428–430, July 2004.

[Bol05] L. Bolla. Polarization Rotators. Technical report, FUNFOX Project, January 2005.

[Bos98a] A. Bossavit. *Computational Electromagnetism – Variational Formulations, Complementarity, Edge Elements*. Academic Press, Boston, 1998.

[Bos98b] A. Bossavit. How weak is the "Weak Solution" in Finite Element Methods? *IEEE Transactions on Magnetics*, 34(5):2429–2432, September 1998.

[BW02] M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Cambridge University Press, Cambridge, UK, 7th edition, 2002.

[CdSHF03] D. Correia, J. P. da Silva, and H. E. Hernández-Figueroa. Genetic Algorithm and Finite-Element Design of Short Single-Section Passive Polarization Converter. *IEEE Photonics Technology Letters*, 15(7):915–917, July 2003.

[CHP94] J. C. Cavendish, C. A. Hall, and T. A. Porsching. A complementary volume approach for modelling three-dimensional Navier-Stokes equations using dual Delaunay/Voronoï tessellations. *Journal of Numerical Methods Heat Fluid Flow*, 4:329–345, 1994.

[CJN04] J. Cai, J. Jiang, and G. P. Nordin. Ultra-short waveguide polarization converter using a sub-wavelenght grating. In *Integrated Photonics Research Topical Meetings*, 2004. presentation IFG2.

[CMP04] L. Codecasa, V. Minerva, and M. Politi. Use of Baricentric Dual Grids for the Solution of Frequency Domain Problems by FIT. *IEEE Transactions on Magnetics*, 40(2):1414–1419, March 2004.

[Cox69] H. S. M. Coxeter. *Introduction to Geometry*. John Wiley & Sons, New York, 1969.

[Cry] Photon Design's CrystalWave website. `www.photond.com/products/crystalwave.htm`.

[CW91] A. C. Cangellaris and D. B. Wright. Analysis of the Numerical Error Caused by the Stair-Stepped Approximation of a Conducting Boundary in FDTD Simulations of Electromagnetic Phenomena. *IEEE Transaction on Antennas and Propagation*, 39(10):1518–1525, 1991.

[CYH95] C. T. Chan, Q. L. Yu, and K. M. Ho. Order-N spectral method for electromagnetic waves. *Physical Review B*, 51(23), 1995.

[ERY03] H. El-Refaei and D. Yevick. An Optimized InGaAsP/InP Polarization Converter Employing Asymmetric Rib Waveguides. *Journal of Lightwave Technology*, 21(6):1544–1548, July 2003.

[ERYJ04] H. El-Refaei, D. Yevick, and T. Jones. Slanted Rib Waveguide InGaAsP-InP Polarization Converters. *Journal of Lightwave Technology*, 22(5):1352–1357, May 2004.

[FES03] S. Foteinopoulou, E. N. Economou, and C. M. Soukoulis. Refraction in Media with a Negative Refractive Index. *Physical Review Letters*, 90(10), March 2003.

[FG92] R. W. Freund and G. H. Golub. Iterative Solution of Linear Systems. *Acta Numerica*, 1:57–100, 1992.

[FIMa] Photon Design's FIMMPROP website. `www.photond.com/products/fimmprop.htm`.

[FIMb] Photon Design's FIMMWAVE website. `www.photond.com/products/fimmwave.htm`.

[Flo83] G. Floquet. Sur les équations différentielles linéaries à coefficients périodiques. *Annales Scientifiques de l'École Normale Supérieure*, 12:47–88, 1883.

[FN91] R. W. Freund and N. M. Nachtigal. QMR: a Quasi-Minimal Residual Method for Non-Hermitian Linear Systems. *Numerische Mathematik*, 60:315–339, 1991.

[FN94] R. W. Freund and N. M. Nachtigal. An implementation of the QMR method based on coupled-two term recurrences. *SIAM Journal of Scientific Computing*, 15:313–337, 1994.

[HCS90] K. M. Ho, C. T. Chan, and C. M. Soukoulis. Existence of a Photonic Gap in Periodic Dielectric Structures. *Physical Review Letters*, 65(25):3152–3155, 1990.

[HKR03] H. A. Haus, L. C. Kimerling, and M. Romagnoli. Application of high index contrast technology to integrated optical devices. *EXP*, 3(4), December 2003. `http://exp.telecomitalialab.com`.

[HSN+00] J. Z. Huang, R. Scarmozzino, G. Nagy, M. J. Steel, and R. M. Osgood. Realization of a compact and single-mode optical passive polarization converter. *IEEE Photonics Technology Letters*, 12(3):317, 2000.

[HSZH97] G. Hebermehl, R. Schlundt, H. Zscheile, and W. Heinrich. Improved Numerical Solutions for the Simulation of Microwave Circuits. *WIAS Preprint*, 309, January 1997. `http://www.wias-berlin.de/publications/preprints/309/`.

[Jac99] J. D. Jackson. *Classical Electrodynamics*. John Wiley & Sons, 3rd edition, 1999.

[Jin02] Jianming Jin. *The Finite Element Method in Electromagnetics*. John Wiley & Sons, New York, 2002.

[JJ00] S. G. Johnson and J. D. Joannopoulos. *Photonic Crystals: the road from theory to practice*. Kluwer Academic Press, 2000.

[JJ01] S. G. Johnson and J. D. Joannopoulos. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Optics Express*, 8(3):173–190, 2001.

[KAD05] I. Kiyat, A. Aydinli, and N. Dagli. A compact silicon-on-insulator polarization splitter. *IEEE Photonic Technology Letters*, 17:100–102, 2005.

[Kal] Lambda-tek's Kallistos website. `www.lambda-tek.com/software_products.htm`.

[KBM+05] M. V. Kotlyar, L. Bolla, M. Midrio, L. O'Faolain, and T. F. Krauss. Compact polarization converter in InP-based material. *Optics Express*, 13(13):5040–5045, June 2005.

[Kit95] C. Kittel. *Introduction to Solid State Physics*. John Wiley & Sons, New York, 7th edition, 1995.

[KMR01] M. Kilmer, E. Miller, and C. Rappaport. QMR-based Projection Techniques for the Solution of Non-Hermitian Systems with Multiple Right Hand Sides. *SIAM Journal of Scientific Computing*, 2001.

[Lal98] P. Lalanne. Effective properties and band structures of lamellar subwavelength crystals: Plane-wave method revisited. *Physical Review B*, 58(15):9801–9807, 1998.

[LBF+04] A. Lavrinenko, P. I. Borel, L. H. Frandsen, M. Thorhauge, A. Harpoth, M. Kristensen, and T. Niemi. Comprehensive FDTD modelling of photonic crystal waveguide components. *Optics Express*, 234(2), 2004.

[LHYH98] W. W. Lui, T. Hirono, K. Yokayama, and W. P. Huang. Polarization rotation in semiconductor bending waveguides: a coupled-mode theory formulation. *Journal of Lightwave Technology*, 16:929, 1998.

[Liu96] Yen Liu. Fourier Analysis of Numerical Algorithms for the Maxwell Equations. *Journal of Computational Physics*, 124:396–416, 1996.

[LLC97] Jin-Fa Lee, R. Lee, and A. Cangellaris. Time-Domain Finite-Element Methods. *IEEE Transactions on Antennas and Propagation*, 45(3):430–442, March 1997.

[Loh97] M. Lohmeyer. Wave-matching-method for mode analysis of dielectric waveguides. *Optical and Quantum Electronics*, 29(9):907–922, 1997.

[LS95] Jin-Fa Lee and Z. Sacks. Whitney Elements Time Domain (WETD) Methods. *IEEE Transactions on Magnetics*, 31(3):1325–1329, May 1995.

[LSSU94] P. Lüsse, P. Stuwe, J. Schule, and H.-G. Unger. Analysis of Vectorial Mode Fields in Optical Waveguides by a New Finite Difference Method. *Journal of Lightwave Technology*, 12(3):487–494, March 1994.

[Lt03] Lao-tzu. *Tao tê ching*. Adelphi, Milano, 6th edition, 2003.

[Mad00] C. K. Madsen. Optical all-pass filters for polarization mode dispersion compensation. *Optics Letters*, 25(12):878–880, June 2000.

[Mar00] M. Marrone. Computational Aspects of Cell Method in Electrodynamics. In *Progress in Electromagnetic Research – PIERS*, 2000.

[Mat] MathWorld website. `http://mathworld.wolfram.com`.

[Max71] J. C. Maxwell. On the Mathematical Classification of Physical Quantities. In *Proceedings of the London Mathematical Society*, volume 3, pages 224–232, 1871. `http://www.dic.units.it/perspage/discretephysics/papers/RELATED/MaxwellRemarks.pdf`.

[Max54] J. C. Maxwell. *Treatise on Electricity and Magnetism*. Dover, New York, 3rd edition, 1954.

[Mea00] C. A. Mead. *Collective Electrodynamics - Quantum Foundation of Electromagnetism*. The MIT Press, Cambridge, Massachusetts, 2000.

[MPB] MIT's MPB website. `http://ab-initio.mit.edu/mpb`.

[NAG] NAG website. `http://www.nag.com`.

[Num] Numerical Recipies website. `http://www.nr.com`.

[Pen96] J. B. Pendry. Extremely Low Frequency Plasmons in Metallic Mesostructures. *Physical Review Letters*, 76:4773–4776, June 1996.

[Pen00] J. B. Pendry. Negative Refraction Makes a Perfect Lens. *Physical Review Letters*, 85:3966–3969, October 2000.

[Pho] Photon Design website. `www.photond.com`.

[Pir] Pirelli Labs Optical Innovation design team. private communication.

[Saa00] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2000. `http://www-users.cs.umn.edu/~saad/books.html`.

[SBK+] B. Smith, S. Balay, M. Knepley, H. Zhang, and V. Eijkhout. PETSc: Portable Extensible Toolkit for Scientific Computation. `http://www-unix.mcs.anl.gov/petsc/petsc-2/index.html`.

[She] J. R. Shewchuk. Triangle website. `http://www.cs.cmu.edu/~quake/triangle.html`.

[She94] J. R. Shewchuk. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, 1994. `http://www-2.cs.cmu.edu/~jrs/jrspapers.html#cg`.

[SLC01] Din-Kow Sun, Jin-Fa Lee, and Z. Cendes. Construction of nearly orthogonal Nedelec bases for rapid convergence with multilevel preconditioned solvers. *SIAM Journal of Scientific Computing*, 23(4):1053–1076, October 2001.

[Sol65] I. Solc. The birefringent filter. *Journal of Optical Society of America*, 55:621–625, 1965.

[Som98] C. G. Someda. *Electromagnetic Waves*. Chapman & Hall, London, 1998.

[SPV+00] D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz. Composite Medium with Simultaneously Negative Permettivity and Permeability. *Physical Review Letters*, 84:4184, May 2000.

[SSW02] R. Schuhmann, P. Schmidt, and T. Weiland. A New Whitney-Based Material Operator for the Finite-Integration Technique on Triangular Grids. *IEEE Transactions on Magnetics*, 38(2):409–412, March 2002.

[Ste88] M. S. Stern. Semivectorial Polarised Finite Difference Method for Optical Waveguides with Arbitrary Index Profiles. *IEE Proceedings in Optoelectronics*, 135(1):56–63, February 1988.

[SW98] R. Schuhmann and T. Weiland. Stability of the FDTD Algorithm on Nonhorthogonal Grids Related to the Spatial Interpolation Scheme. *IEEE Transactions on Magnetics*, 34:2751–2754, September 1998.

[SW00] R. Schuhmann and T. Weiland. The Nonorthogonal Finite Integration Technique Applied to 2D- and 3D-Eigenvalue Problems. *IEEE Transactions on Magnetics*, 36(4), July 2000.

[Taf98] A. Taflove. *Advances in Computational Electrodynamics: the Finite-Difference Time-Domain Method*. Arthec House, Norwood, 1998.

[Taf00] A. Taflove. *Computational Electrodynamics: the Finite-Difference Time--Domain Method*. Arthec House, Norwood, 2000.

[TCB+03] D. Taillaert, H. Chong, P. I. Borel, L. H. Frandsen, R. M. De La Rue, and R. Baetz. A compact two dimensional grating coupler used as a polarization splitter. *IEEE Photonic Technology Letters*, 15:1249–1251, 2003.

[Tei01] F. L. Teixeira. Geometric Aspects of the Simplicial Discretization of Maxwell's Equations. In *Progress in Electromagnetic Research – PIER 32*, volume 8, pages 171–188, 2001.

[TF96] V. P. Tzolov and M. Fontaine. A passive polarization converter free of longitudinally-periodic structure. *Optics Communications*, 127(7), 1996.

[TK04] F. Trevisan and L. Kettunen. Geometric Interpretation of Discrete Approaches to Solving Magnetostatics. *IEEE Transactions on Magnetics*, 40, March 2004.

[TM97] G. Tayeb and D. Maystre. Rigorous theoretical study of finite-size two-dimensional photonic crystals doped by microcavities. *Journal of Optical Society America A*, 14:3323–3332, 1997.

[Ton00] E. Tonti. Formulazione finita dell'elettromagnetismo partendo dai fatti sperimentali. Technical report, Scuola Nazionale Dottorandi di Elettronica, Palazzo Antonini, Università degli Studi di Udine, Giugno 2000.

[UMF] UMFPACK website. `http://www.cise.ufl.edu/research/sparse/umfpack`.

[Ves68] V. G. Veselago. The electrodynamics of substances with simultaneously negative values of ϵ and μ. *Soviet Physics - Uspekhi*, 10:509–514, 1968.

[VP94] P. R. Villeneuve and M. Pichè. Photonic bandgaps: what is the best numerical representation of periodic structures? *Journal of Modern Optics*, 41(2):241–256, 1994.

[vR01] U. van Rienen. Frequency domain analysis of waveguides and resonantors with FIT on non-orthogonal triangular grids. In *Progress in Electromagnetic Research – PIER 32*, pages 357–381, 2001.

[web] Mesh Generation Software website. `http://www-users.informatik.rwth-aachen.de/~roberts/software.html`.

[Web93] J. P. Webb. Edge Elements and What They can do for You. *IEEE Transactions on Magnetics*, 29(2):1460–1465, March 1993.

[Web99] J. P. Webb. Hierarchal Vector Basis Functions of Arbitrary Order for Triangular and Tetrahedral Finite Elements. *IEEE Transactions on Antennas and Propagation*, 47(8):1244–1253, August 1999.

[WH05] M. R. Watts and H. A. Haus. Integrated mode-evolution-based polarization rotators. *Optics Letters*, 30(2):138–140, January 2005.

[Wik] Wikipedia website. `http://www.wikipedia.org`.

[WMG+04] L. Wu, M. Mazilu, J.-F. Gallet, T. F. Krauss, A. Jugessur, and R. M. De La Rue. Planar photonic crystal polarization splitter. *Optics Letters*, 29:1260–1262, 2004.

[WMKK02] L. Wu, M. Mazilu, T. Karle, and T. F. Krauss. Superprism Phenomena in Planar Photonic Crystals. *IEEE Journal of Quantum Electronics*, 38:915–918, 2002.

[Yab93] E. Yablonovitch. Photonic band-gap structures. *Journal of Optical Society America B*, 10(3), 1993.

[Yee66] K. S. Yee. Numerical Solution of Initial Boundary Value Problems Involving Maxwell's equations in Isotropic Media. *IEEE Transaction on Antennas and Propagation*, 14:302–307, March 1966.

Index