Mathematical Modelling and Numerical Analysis
Modélisation Mathématique et Analyse Numérique

ESAIM: M2AN
M2AN, Vol. 36, N° 1, 2002, pp. 87–119
DOI: 10.1051/m2an:2002004

ON THE ANALYSIS OF BÉRENGER'S PERFECTLY MATCHED LAYERS FOR MAXWELL'S EQUATIONS

Eliane Bécache[1] and Patrick Joly[1]

Abstract. In this work, we investigate the Perfectly Matched Layers (PML) introduced by Bérenger [3] for designing efficient numerical absorbing layers in electromagnetism. We make a mathematical analysis of this model, first *via* a modal analysis with standard Fourier techniques, then *via* energy techniques. We obtain uniform in time stability results (that make precise some results known in the literature) and state some energy decay results that illustrate the absorbing properties of the model. This last technique allows us to prove the stability of the Yee's scheme for discretizing PML's.

Résumé. Dans ce travail, nous considérons le modèle de couches parfaitement adaptées, dit PML (Perfectly Matched Layers), introduit par Bérenger [3] pour la modélisation de frontières absorbantes en électromagnétisme. Nous menons une analyse mathématique de ce modèle, d'une part par une analyse modale par transformation de Fourier, d'autre part par des techniques énergétiques. Nous obtenons ainsi des résultats de stabilité uniforme en temps (qui précisent des résultats déjà connus dans la littérature) et établissons des résultats de décroissance d'énergie qui illustrent les propriétés d'absorption du modèle. Cette dernière technique permet aussi de démontrer la stabilité du schéma de Yee pour discrétiser les couches absorbantes.

Mathematics Subject Classification. 65M06, 65M12, 35L05, 35L40, 78M20, 35B35.

Received: April 19, 2001.

Introduction

In this paper, we investigate from a mathematical point of view the Perfectly Matched Layers (PML) introduced by Bérenger [3] in order to design efficient numerical absorbing boundary conditions (more exactly absorbing layers) for the computation of time dependent solutions of Maxwell's equations in unbounded domains. To derive the Bérenger's PML model, one first rewrites Maxwell's equations, whose unknowns are the electric and magnetic fields (E, H), in a so called "split form" in the unknowns (E, H^x, H^y), where H^x and H^y are non-physical variables, whose sum gives the magnetic field, $H^x + H^y = H$. One then obtains the "PML model" by adding a very special zero-order absorption term, proportional to some absorption coefficient σ. Finally, an absorbing layer is obtained by replacing the Maxwell's equations by the PML model inside a layer of finite width that surrounds the bounded domain of interest for the computations (the "Maxwell domain").

Keywords and phrases. Absorbing layers, PML, Maxwell's equations, stability, hyperbolic systems, Fourier analysis, energy techniques, Yee's scheme.

[1] INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France. e-mail: eliane.becache@inria.fr; patrick.joly@inria.fr

This absorbing layer has the property to be "perfectly matched". This means that a wave propagating in the "Maxwell domain" does not produce any reflection when it meets the interface with the absorbing layer.

For a mathematician, the first natural reflex (that we are far from being the first to have) is to study the well-posedness of the Cauchy problem associated to the PML model. Therefore, taking into account the abundant literature that has been devoted to the subject, it is natural to wonder when reading the title of this paper if there is something new in the present work. Indeed, we think that this is the case and one of the objective of this introduction is to explain why.

Let us recall that for a short period, starting from the observation [1] that the PML model was a zero order perturbation of a weakly hyperbolic first order system (see [9]), there was a rumor on the fact that the PML model could be ill-posed, despite of the fact that it worked very well numerically. It is now known that it is not the case: the perturbations that could lead to ill-posedness do not correspond to the PML perturbation. More precisely, several authors [2, 10–12, 15, 16] proposed "new PMLs", in fact a reformulation of the original equations of Berenger using new unknowns. A discussion on several of these absorbing layers can be found in [14]. These models appear (in some sense which is not always so clear) as lower-order perturbations of the Maxwell's system, which is strongly hyperbolic. Therefore, the general theory [9] ensures the well posedness of the Cauchy problem associated to the PML equations. Such a result does not depend on the sign of the absorption coefficient σ. However when one makes numerical computations with a negative σ, one obtains exponentially growing solutions and the PML layer is no longer absorbing. That is why we think that a more pertinent – or at least more complete – analysis should take into account the sign of the coefficient σ. This is one of the objectives of this study.

In fact, by speaking of weak (resp. strong) well-posedness [9] of the Cauchy problem, we mean in particular that there exists some constants $K > 0$ and $\alpha \in \mathbb{R}$ such that the solution $U(t)$ satisfies an estimate on the type

$$\|U(.,t)\|_{L^2} \leq K e^{\alpha t} \|U(.,0)\|_{H^s}\,, \tag{1}$$

for $U(.,0)$ given in the Sobolev space H^s, $s > 0$. (resp. $s = 0$). In particular, this makes possible the solution to blow up exponentially when the time goes to infinity. This is something which is of course highly undesirable for an absorbing model and in practice, in numerical calculations, such a phenomenon is often described as an instability and difficult to distinguish from a real ill-posedness. That is why we want to distinguish the notion of well-posedness and the one of stability.

Definition. A system which is a zero order perturbation of a first-order hyperbolic system is weakly (resp. strongly) stable if the solution $U(t)$ satisfies the estimate

$$\|U(.,t)\|_{L^2} \leq K(1+At)^s \|U(.,0)\|_{H^s}\,, \tag{2}$$

for $U(.,0)$ given in the Sobolev space H^s, $s > 0$ (resp. $s = 0$).

In this paper, we pursue the three following objectives:

- By using standard Fourier techniques, as in [1], we show the stability of the PML model for positive σ in a homogeneous domain. We pay particular attention to deriving uniform (in time) *a priori* estimates, which did not appear in the literature up to our knowledge (Sect. 1). We point out the unknowns of the problem which are affected by a loss of regularity due to the weak hyperbolicity of the unperturbed model.
- We show that the stability result can be obtained *via* energy techniques (Sect. 2). More precisely, we show the decay in time of some quadratic energy in the electromagnetic field. For this we exploit the formulation proposed in [15]. This is probably the more original contribution of this paper. Analogous energy techniques also permit us to obtain the well-posedness of the Cauchy problem in the case of a variable σ. However we have not been able to extend in this case the energy decay result.
- We show that the stability results for positive constant σ extends to the numerical solution of the PML equations approximated by the Yee's scheme (which was originally used by Bérenger)(Sect. 3). This is obtained by proving the decay of a discrete energy. This last result is interesting in the sense that it

is known that the numerical approximation of problems which are only weakly well-posed can lead to numerical instabilities. This is not the case with the Yee's scheme.

To conclude this introduction let us emphasize the fact that, if for Maxwell's equations the PML model leads to a stable problem, this is no longer true for other equations. In particular, the instabilities observed for the linearized Euler equations or for the elasticity equations (see [5–7]) are related to an exponential behavior in time, and this is the object of a forthcoming paper.

1. Fourier analysis of the Cauchy problem associated to the constant coefficients Bérenger's system

In this section, our goal is to analyze the well-posedness of the Bérenger's system through properties of its Fourier symbol. We start by recalling, in Section 1.1.1, the usual properties of the 2D Maxwell's equations. In Section 1.1.2, we recall the split form of Maxwell's equations and some known results about this system (namely the Bérenger's system without perturbation). We finally introduce the Bérenger system in Section 1.2. The analysis of this system by Fourier techniques is developed in Section 1.3. After a brief comment about the weak well-posedness (Sect. 1.3.1), we state our two main results (Th. 1.2 and Th. 1.3) in Section 1.3.2. The proof of these results is detailed in Section 1.3.3.

1.1. Preliminary results

Notations. We denote by $(.,.)$ the L^2 scalar product, $\|.\|_{H^s}$ the norm in H^s. The notation $|.|$ is used either for the Euclidean norm of a vector or for the associated matrix norm.

1.1.1. *The 2D Maxwell's equations.*

We consider the two dimensional Maxwell equations for the TE mode in the free space. We denote the magnetic component H_z by H and assume that $\varepsilon_0 = 1$ and $\mu_0 = 1$. Then the TE equations are given by

$$\left\{\begin{array}{l} \partial_t E_x = \partial_y H \\ \\ \partial_t E_y = -\partial_x H \\ \\ \partial_t H = \partial_y E_x - \partial_x E_y \end{array}\right. \iff \left\{\begin{array}{l} \partial_t \vec{E} = \vec{\mathrm{rot}} H \\ \\ \partial_t H = -\mathrm{rot}\vec{E} \end{array}\right. \tag{3}$$

where operators rot and $\vec{\mathrm{rot}}$ are adjoint to each other. We add to (3) initial data (E_x^0, E_y^0, H^0). Let $V = (E_x, E_y, H)^t$, this system is a first-order system that can be written as

$$\partial_t V = P_0(\nabla)V,$$

where the symbol of $P_0(\nabla)$ (see [9] for the definition) is the anti-hermitian matrix given by:

$$P_0(i\vec{k}) = \begin{pmatrix} 0 & 0 & ik_y \\ 0 & 0 & -ik_x \\ ik_y & -ik_x & 0 \end{pmatrix}.$$

The above matrix has three purely imaginary and distinct eigenvalues $(0, \pm i\left|\vec{k}\right|)$, which shows that system (3) is strictly hyperbolic (in particular strongly hyperbolic), [9]. Problem (3) is thus well-posed. Moreover, the estimate $\left|e^{P(i\vec{k})t}\right| \leq C$, where C does not depend on $\vec{k}$ and t implies, *via* Plancherel's theorem, that the solution is stable with respect to the initial data, in L^2-norm, uniformly in time.

Let us recall that in this case, as (3) is a symmetric system, another approach to get well-posedness, still valid for non-constant coefficients, consists in establishing energy estimates. Since operators rot and $\vec{\text{rot}}$ are adjoint to each other, it is easy to check the conservation of energy:

$$(\partial_t H, H) + (\partial_t \vec{E}, \vec{E}) = 0 \Rightarrow \|H\|_{L^2}^2 + \|\vec{E}\|_{L^2}^2 = \|H^0\|_{L^2}^2 + \|\vec{E}^0\|_{L^2}^2. \tag{4}$$

1.1.2. *The split form of Maxwell's equations*

In order to construct his absorbing layers, Bérenger [3] first introduced a new "equivalent form" which consists in rewriting H as

$$H = H^x + H^y, \tag{5}$$

and in replacing (3) by:

$$\begin{cases} \partial_t E_x = \partial_y (H^x + H^y) \\ \partial_t E_y = -\partial_x (H^x + H^y) \\ \partial_t H^x = -\partial_x E_y \\ \partial_t H^y = \partial_y E_x. \end{cases} \tag{6}$$

This system is also the unperturbed form of the Bérenger's system. Then solving (6) with initial data $(E_x^0, E_y^0, (H^x)^0, (H^y)^0)$, one easily shows that $(E_x, E_y, H = H^x + H^y)$ is solution of (3) for the initial data (E_x^0, E_y^0, H^0) with:

$$H^0 = (H^x)^0 + (H^y)^0. \tag{7}$$

In this sense, systems (6) and (3) are equivalent under the compatibility condition (7) for the initial data. To complete our presentation, it is important to study (6) as a more general initial value problem, including the case of initial data which do not satisfy (7). Let $U = (E_x, E_y, H^x, H^y)^t$ the unknown. The symbol of $P(\nabla)$ is

$$P(i\vec{k}) = \begin{pmatrix} 0 & 0 & ik_y & ik_y \\ 0 & 0 & -ik_x & -ik_x \\ 0 & -ik_x & 0 & 0 \\ ik_y & 0 & 0 & 0 \end{pmatrix},$$

and has the characteristic polynomial $\det(P(i\vec{k}) - \lambda I) = \lambda^2(\lambda^2 + |\vec{k}|^2)$. As for P_0, the eigenvalues are still imaginary but not distinct any more. Therefore (6) is hyperbolic but not strictly hyperbolic. Moreover, it is easy to show that $\dim \ker(P(i\vec{k})) = 1$ which is strictly less than the multiplicity of the eigenvalue $\lambda = 0$. Thus $P(i\vec{k})$ cannot be diagonalized, which means that (6) is not strongly hyperbolic either. From [9] we know that the corresponding Cauchy problem is weakly well-posed but not strongly well-posed: there is necessarily a loss of regularity, at least for some initial data. The maximal loss of regularity can be estimated thanks to an estimate on $|e^{P(i\vec{k})t}|$. It can be shown that [1]

$$|\mathrm{e}^{P(i\vec{k})t}| \leq (1 + Ct)(1 + |\vec{k}|), \tag{8}$$

where C is a positive constant. Parseval's theorem implies

$$\|U(.,t)\|_{L^2} \leq (1 + Ct)\|U^0\|_{H^1}. \tag{9}$$

This result can be slightly improved (and also generalized for non-constant coefficients), in a simple way *via* energy estimates, as explained in [4]. Let $H = H^x + H^y$, as we said before, $(\vec{E}, H)$ satisfies (3), thus the classical energy conservation holds

$$\|(H^x + H^y)(.,t)\|_{L^2}^2 + \|\vec{E}(.,t)\|_{L^2}^2 = \|(H^x)^0 + (H^y)^0\|_{L^2}^2 + \|\vec{E}^0\|_{L^2}^2. \tag{10}$$

On the other hand, we have

$$\partial_t(H^y - H^x) = \partial_y E_x + \partial_x E_y$$

$$\Longrightarrow (H^y - H^x)(t) = (H^y - H^x)(0) + \int_0^t (\partial_y E_x + \partial_x E_y)(s)\mathrm{d}s.$$

The space derivatives of $(\vec{E}, H)$ still satisfy Maxwell's equation, we thus get

$$\|\nabla\vec{E}(t)\|_{L^2} + \|\nabla H(t)\|_{L^2} \leq \|\nabla\vec{E}^0\|_{L^2} + \|\nabla H^0\|_{L^2}, \tag{11}$$

and finally

$$\|(H^y - H^x)(t)\|_{L^2} \leq \|(H^y - H^x)^0\|_{L^2} + t\|U^0\|_{H^1}. \tag{12}$$

Remark 1.1. We can notice that, with (10), we obtain strong estimates on $\vec{E}$ and $H = H^x + H^y$. The loss of regularity comes from the estimate on $H^x - H^y$. This will remain true with the perturbation term.

1.2. **The Bérenger's PML system**

The PML equations in a layer parallel to the y axis. We now consider the problem posed in a Bérenger's PML layer, with damping in one direction, namely the x-direction:

$$\begin{cases} \partial_t E_x = \partial_y(H^x + H^y) & (a) \\ \partial_t E_y + \sigma E_y = -\partial_x(H^x + H^y) & (b) \\ \partial_t H^x + \sigma H^x = -\partial_x E_y & (c) \\ \partial_t H^y = \partial_y E_x & (d) \end{cases} \tag{13}$$

with $H = H^x + H^y$, and where we make the assumption that σ is a non-zero coefficient. This system can be rewritten as

$$\partial_t U + \Sigma U = P(\nabla)U, \quad \text{with } \Sigma = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 \\ 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{14}$$

Other PML systems. System (13) is a particular case of the following more general system:

$$\begin{cases} \partial_t E_x + \sigma_y E_x = \partial_y (H^x + H^y) & (a) \\ \partial_t E_y + \sigma_x E_y = -\partial_x (H^x + H^y) & (b) \\ \partial_t H^x + \sigma_x H^x = -\partial_x E_y & (c) \\ \partial_t H^y + \sigma_y H^y = \partial_y E_x & (d) \end{cases} \tag{15}$$

which can be written in the form (14) with

$$\Sigma = \begin{pmatrix} \sigma_y & 0 & 0 & 0 \\ 0 & \sigma_x & 0 & 0 \\ 0 & 0 & \sigma_x & 0 \\ 0 & 0 & 0 & \sigma_y \end{pmatrix}.$$

- One recovers the equations in a layer parallel to the y axis by choosing $\sigma_x = \sigma$, $\sigma_y = 0$.
- One obtains the equations in a layer parallel to the x axis by choosing $\sigma_x = 0$, $\sigma_y = \sigma$.
- One obtains the equations in a corner domain with $\sigma_x \neq 0$ and $\sigma_y \neq 0$.

Most of our analysis will concern the system (13) in a layer parallel to the y axis. However, we shall give some results on the general system (15) in Sections 2.3.2 and 2.4.2.

1.3. **Fourier analysis of system (13)**

System (13) is a zero-order perturbation of the weakly hyperbolic system (6). Our aim in this section is not only to prove that the initial value problem associated to (13) is nevertheless (weakly) well-posed but also to give precise uniform estimates on the solution, in fact stability results in the sense of definition , for positive σ. The analysis is related to the properties of the matrix symbol of the operator $P(\nabla) - \Sigma$:

$$P_\sigma(i\vec{k}) = P(i\vec{k}) - \Sigma = \begin{pmatrix} 0 & 0 & ik_y & ik_y \\ 0 & -\sigma & -ik_x & -ik_x \\ 0 & -ik_x & -\sigma & 0 \\ ik_y & 0 & 0 & 0 \end{pmatrix}, \tag{16}$$

whose characteristic polynomial is:

$$q(\lambda; \sigma, \vec{k}) = \lambda^4 + 2\sigma\lambda^3 + (\sigma^2 + |k|^2)\lambda^2 + 2\sigma k_y^2 \lambda + k_y^2 \sigma^2. \tag{17}$$

1.3.1. *Weak well-posedness*

Checking that (13) is weakly well-posed amounts to proving that the characteristic equation

$$q(\lambda; \sigma, \vec{k}) = 0, \tag{18}$$

has no solution $\lambda_j(\sigma, \vec{k})$ such that

$$\Re e \lambda_j(\sigma, \vec{k}) \to +\infty \quad \text{when} \quad |\vec{k}| \to +\infty.$$

To prove this it is sufficient to study equation (18) for large $|\vec{k}|$. It is useful to introduce the following dimensionless quantities:

$$\varepsilon = \frac{\sigma}{|\vec{k}|}, \quad \vec{K} = \frac{\vec{k}}{|\vec{k}|}, \quad \mu = \frac{\lambda}{|\vec{k}|}\cdot \tag{19}$$

One easily sees that, solving $q(\lambda;\sigma,\vec{k}) = 0$ is equivalent to:

$$Q(\mu;\varepsilon,\vec{K}) = \mu^2(\mu^2+1) + 2\varepsilon\mu(\mu^2+K_y^2) + \varepsilon^2(\mu^2+K_y^2) = 0, \tag{20}$$

which shows that the eigenvalues $\lambda_j(\sigma,\vec{k}), 1 \le j \le 4$, of $P_\sigma(i\vec{k})$ are given by:

$$\lambda_j(\sigma,\vec{k}) = |\vec{k}|\ \mu_j(\frac{\sigma}{|\vec{k}|}, \frac{\vec{k}}{|\vec{k}|}), \tag{21}$$

where the $\mu_j(\varepsilon,\vec{K}), 1 \le j \le 4$, are the roots of equation (20). As $\mu_j(\varepsilon,\vec{K})$ is clearly odd with respect to ε, one deduces that

$$\lambda_j(-\sigma,\vec{k}) = -\lambda_j(\sigma,\vec{k}). \tag{22}$$

Let $(\mu_j^0)_{j=1..4}$ be the four roots for the unperturbed characteristic equation,

$$Q(\mu_j^0, 0, \vec{K}) = 0 \Longleftrightarrow (\mu_j^0)^2((\mu_j^0)^2+1) = 0.$$

For the two single roots $\mu_1^0 = i$ and $\mu_2^0 = -i$, the implicit function theorem applies, and one can show that the two branches of solutions of (20), issued of points $(\mu_j^0, 0)$, $j = 1, 2$, are given by:

$$\mu_j(\varepsilon,\vec{K}) = \pm i - \varepsilon K_x^2 + O(\varepsilon^2).$$

This shows that for large $|\vec{k}|$, one has for these branches:

$$\lambda_j(\vec{k},\sigma) = \pm|\vec{k}|i - \sigma\frac{k_x^2}{|\vec{k}|^2} + O(\frac{1}{|\vec{k}|}), \quad j = 1, 2.$$

Therefore,

$$\Re e\lambda_j(\vec{k},\sigma) = -\sigma\frac{k_x^2}{|k|^2} + O(\frac{1}{|\vec{k}|})$$

does not tend to infinity. It is even negative for a positive σ, which reveals an absorption phenomenon for high frequencies.

In the neighborhood of the double root $\mu_j^0 = 0$, $j = 3, 4$, introducing

$$\nu = \varepsilon\mu = \frac{\lambda}{\sigma}, \tag{23}$$

the characteristic equation becomes:

$$\varepsilon^2\nu^2(\nu+1)^2 + (\nu+K_y^2)^2 + K_x^2K_y^2 = 0. \tag{24}$$

We notice that ν depends on $\vec{K}/\varepsilon$, which shows that:

$$\lambda_j(\sigma, \vec{k}) = \sigma \nu_j(\frac{\vec{k}}{\sigma}) \cdot \tag{25}$$

For $\varepsilon = 0$, (24) becomes

$$(\nu + K_y^2)^2 + K_x^2 K_y^2 = 0,$$

and by implicit function theorem, one shows that

$$\nu_j(\varepsilon, \vec{K}) = -K_y^2 \pm i K_x K_y + O(\varepsilon^2), \quad j = 3, 4$$

which implies

$$\lambda_j(\vec{k}, \sigma) = \sigma(-\frac{k_y^2}{|\vec{k}|^2} \pm i \frac{k_y k_x}{|\vec{k}|^2}) + O(\frac{1}{|\vec{k}|^2}), \quad j = 3, 4.$$

Therefore

$$\Re e \lambda_j(\vec{k}, \sigma) = -\sigma \frac{k_y^2}{|k|^2} + O(\frac{1}{|\vec{k}|^2}), \quad j = 3, 4,$$

and again this is not only bounded, but also negative for a positive σ.

1.3.2. *Stability and a priori estimates*

In the sequel, we consider the Cauchy problem made of equations (13) written for $(x, y) \in \mathbb{R}^2$ and $t > 0$ coupled with the initial condition:

$$\begin{cases} E_x(0) = E_x^0 \\ E_y(0) = E_y^0 \\ H^x(0) = (H^x)^0 \\ H^y(0) = (H^y)^0 \end{cases} \tag{26}$$

and we denote by U the vector solution

$$U = (E_x, E_y, H^x, H^y)^t,$$

and by U^0 the corresponding vector of initial data. In this section, we state two theorems which will be proved in Section 1.3.3. Our first main result is a stability result for system (13).

Theorem 1.2. *For any $U^0 \in (H^1(\mathbb{R}^2))^4$, problem (13) admits a unique solution which satisfies the a priori estimate:*

$$\|U(., t)\|_0 \leq C_\sigma \, \|U^0\|_1, \tag{27}$$

where $C_\sigma = C \min(1, 1/\sigma)$, ($\sigma > 0$).

The *weak* well-posedness appears in the loss of regularity between the initial data and the solution at time t. We can note that this estimate is not optimal when σ tends to zero (*cf.* (9) when $\sigma = 0$). This can be made more precise in our next result.

Theorem 1.3. *The electromagnetic field $(\vec{E}, H)$, deriving from the solution $U = (E_x, E_y, H^x, H^y)$ (with $H = H^x + H^y$) of problem (13) satisfies the following estimate*

$$\|E(t)\|_{L^2} + \|H(t)\|_{L^2} \leq C \left\|U^0\right\|_{L^2}, \tag{28}$$

and the splitted unknowns H^x and H^y are estimated by:

$$\|H^x(t)\|_{H^{-1}} + \|H^y(t)\|_{H^{-1}} \leq C\ t \left\|U^0\right\|_{L^2}, \tag{29}$$

where C is a positive constant, independent of σ, ($\sigma > 0$).

This result shows that the physical unknowns satisfy the usual estimate, as for Maxwell's equations and it points out that the loss of regularity only affects the split fields.

Remark 1.4. For a negative σ, one would obtain similar estimates with a bound which grows exponentially in time:

$$\|E(t)\|_{L^2} + \|H(t)\|_{L^2} \leq Ce^{|\sigma|t} \left\|U^0\right\|_{L^2}$$

$$\|H^x(t)\|_{H^{-1}} + \|H^y(t)\|_{H^{-1}} \leq C\ e^{|\sigma|t} \left\|U^0\right\|_{L^2}.$$

1.3.3. *Proof of the stability and a priori estimates*

In this section, we give the proofs of Theorems 1.2 and 1.3 stated in Section 1.3.2. The basic ingredient for the stability result is the fact that all the solutions of the characteristic equation (18) have a negative real part (Lem. 1.5 (i)). Then the difficulty for obtaining *a priori* estimates amounts to derive uniform estimates (in $\vec{k}$ and t) of the matrices $\mathrm{e}^{P_\sigma(i\vec{k})t}$. This is made difficult by the fact that $P_\sigma(i\vec{k})$ is not always diagonalizable (Lem. 1.5 (iii)) and demands additional technical work.

The main technical lemma concerns the study of the eigenvalues of the matrices $P_\sigma(i\vec{k})$.

Lemma 1.5. (i) *$P_\sigma(i\vec{k})$ admits two couples of complex conjugate eigenvalues $(\lambda_1(\sigma,\vec{k})$, $\lambda_2(\sigma,\vec{k}) = \overline{\lambda}_1(\sigma,\vec{k}))$, and $(\lambda_3(\sigma,\vec{k})$, $\lambda_4(\sigma,\vec{k}) = \overline{\lambda}_3(\sigma,\vec{k}))$, that are continuous functions in $\vec{k}$ and satisfying:*

$$\forall \vec{k} \in R^2, \quad -\sigma \leq \Re e\ \lambda_j(\sigma,\vec{k}) \leq 0. \tag{30}$$

Moreover, one can choose $\lambda_1(\sigma,\vec{k})$ such that

$$\exists C \geq 0,\ \forall \vec{k} \in R^2, \quad \left|\Im m\ \lambda_1(\sigma,\vec{k})\right| \leq C|\sigma| \tag{31}$$

while

$$\left|\Im m\ \lambda_3(\sigma,\vec{k})\right| \to +\infty,\ \text{when } |\vec{k}| \to +\infty.$$

(ii) *For all $\vec{k} \neq (\pm\sigma, \pm\sigma)^t$, $P_\sigma(i\vec{k})$ can be diagonalized. Moreover:*
   (a) *If $k_x k_y \neq 0$, the four eigenvalues are not real and distinct.*
   (b) *In the case $k_x = 0$, $P_\sigma(i\vec{k})$ has two purely imaginary eigenvalues $\lambda = \pm i\left|\vec{k}\right|$ and one double real eigenvalue $\lambda = -\sigma$.*
   (c) *In the case $k_y = 0$, $P_\sigma(i\vec{k})$ has two complex conjugate eigenvalues $\lambda = -\sigma \pm i\left|\vec{k}\right|$ and one double real eigenvalue $\lambda = 0$.*

(iii) *For* $\vec{k} \in \mathcal{N}_\sigma \equiv \{(\pm\sigma, \pm\sigma)^t\}$, $P_\sigma(i\vec{k})$ *has two Jordan blocks of dimension* 2 *associated to the double eigenvalues,* $\lambda_1 = \sigma(-1/2 + i\sqrt{3}/2)$ *and* $\lambda_2 = \overline{\lambda}_1$.

Proof.

(i) The proof of point (i) is quite long and postponed in Appendix A.

(ii) From the form of equation (24) which involves the sum of squares, we see that for $k_x k_y \neq 0$, the eigenvalues are not real. Therefore there are two couples of complex conjugate eigenvalues, $(\nu_1, \nu_2 = \overline{\nu}_1)$ and $(\nu_3, \nu_4 = \overline{\nu}_3)$ with $\Im m \nu_j \neq 0$. Let us assume that we have multiple eigenvalues, $\nu_1 = \nu_3 = a + ib$, then (18) could be rewritten

$$(\nu - \nu_1)^2 (\nu - \overline{\nu}_1)^2 = 0.$$

By identification with (24), we get

$$\begin{cases} a = -\dfrac{1}{2}; & 4a^2 + 2\left|\nu_1\right|^2 = 1 + \dfrac{1}{\varepsilon^2}, \\ -4a\left|\nu_1\right|^2 = 2\dfrac{K_y^2}{\varepsilon^2}; & \left|\nu_1\right|^4 = \dfrac{K_y^2}{\varepsilon^2}, \end{cases}$$

from which we deduce

$$\nu_1 = -\frac{1}{2} + i\frac{\sqrt{3}}{2}, \quad K_y^2 = K_x^2 = \varepsilon^2 = \frac{1}{2},$$

which corresponds to

$$\lambda = \sigma(-\frac{1}{2} + i\frac{\sqrt{3}}{2}), \quad \left|k\right|^2 = 2\sigma^2 \; ; k_x^2 = k_y^2 = \sigma^2.$$

We conclude that if $k_x^2 \neq \sigma^2$ or $k_y^2 \neq \sigma^2$, $P_\sigma(i\vec{k})$ has four distinct eigenvalues, and thus can be diagonalized. This proves (ii)-(a). The particular cases (b) and (c) are trivial.

(iii) This result is obtained by simple computation and shows that the Jordan form of the matrix $P_\sigma(i\vec{k}_0)$ with $\vec{k}_0 = (\pm\sigma, \pm\sigma)^t$ is:

$$J = \begin{pmatrix} \nu_1 & 1 & 0 & 0 \\ 0 & \nu_1 & 0 & 0 \\ 0 & 0 & \overline{\nu}_1 & 1 \\ 0 & 0 & 0 & \overline{\nu}_1 \end{pmatrix}$$

with $\nu_1 = -\dfrac{1}{2} + i\dfrac{\sqrt{3}}{2}$. □

To illustrate Lemma 1.5 we have represented in Figure 1 the curves described by the $\nu_j(\varepsilon, \vec{K}) = \lambda_j(\varepsilon, \vec{K})/\sigma$ when ε is fixed ($\varepsilon = 1$) and $\vec{K} = (\cos\zeta, \sin\zeta), 0 \leq \zeta \leq 2\pi$ describes the unit circle.

The proof of Theorem 1.2 will be based on estimates of $\mathrm{e}^{P_\sigma(i\vec{k})\, t}$. We first treat the neighborhood of the four points $\vec{k}_0 \in \mathcal{N}_\sigma = \{(\pm\sigma, \pm\sigma)^t\}$, for which $P_\sigma(i\vec{k})$ is not diagonalizable (Lem. 1.7). For this we shall use a technical lemma about the exponential of matrices (which is not trivial only when one considers non commuting matrices, A and K).

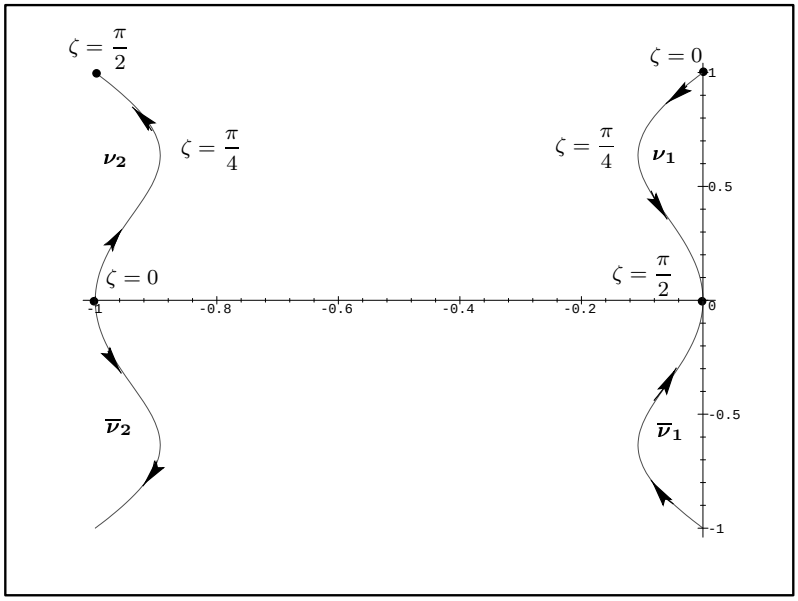

FIGURE 1. Location of the eigenvalues (for ζ varying).

Lemma 1.6. *Let A a matrix satisfying $\left|\mathrm{e}^{At}\right| \leq C\mathrm{e}^{-\lambda t}$, for some positive constant C, and some real number λ, then for any matrix K, one has*

$$\left|\mathrm{e}^{(A+K)t}\right| \leq C\mathrm{e}^{-(\lambda - C|K|)t}.$$

Proof. Let us set $w(t) = \mathrm{e}^{(A+K)t}u_0$. We have

$$\left\{ \begin{array}{l} w'(t) = (A+K)w \\ w(0) = u_0 \end{array} \right. \iff \left\{ \begin{array}{l} w'(t) - Aw = Kw \\ w(0) = u_0 \end{array} \right.$$

therefore:

$$w(t) = \mathrm{e}^{At}u_0 + \int_0^t \mathrm{e}^{A(t-s)}Kw(s)\mathrm{d}s$$

from which we deduce

$$\begin{aligned} &|w(t)| \leq C\mathrm{e}^{-\lambda t}\,|u_0| + C\int_0^t \mathrm{e}^{-\lambda(t-s)}\,|K|\,|w(s)|\,\mathrm{d}s \\ &\Rightarrow \mathrm{e}^{\lambda t}\,|w(t)| \leq C\,|u_0| + C\,|K|\int_0^t \mathrm{e}^{\lambda s}\,|w(s)|\,\mathrm{d}s. \end{aligned}$$

Gronwall's lemma then yields

$$|w(t)| \leq C\mathrm{e}^{-(\lambda - C|K|)t}\,|u_0|\,,$$

which concludes the proof. □

We then observe that we can write

$$P_\sigma(i\vec{k}) = \sigma \mathcal{M}(\vec{K}/\varepsilon),$$

with

$$\mathcal{M}(\vec{v}) = \begin{pmatrix} 0 & 0 & iv_y & iv_y \\ 0 & -1 & -iv_x & -iv_x \\ 0 & -iv_x & -1 & 0 \\ iv_y & 0 & 0 & 0 \end{pmatrix},$$

where $\vec{v} = \vec{K}/\varepsilon = \vec{k}/\sigma$. The eigenvalues of $\mathcal{M}(\vec{v})$ are the $\nu_j(\vec{v})$ defined as the solutions of (24) and related to the eigenvalues of P_σ through (25).

Lemma 1.7. *There exists $\rho > 0$ such that for any $\vec{v}_0 \in \mathcal{N}_1 = \{(\pm 1, \pm 1)^t\}$*

$$|\vec{v} - \vec{v}_0| < \rho \Longrightarrow \left|\mathrm{e}^{P_\sigma(i\vec{k})t}\right| = \left|\mathrm{e}^{\sigma\mathcal{M}(\vec{v})t}\right| \leq C\mathrm{e}^{-\sigma/8t}.$$

Proof. Thanks to the Jordan form of $P_\sigma(i\vec{k}_0)$ (*cf.* Lem. 1.5 (iii)) we have

$$\left|\mathrm{e}^{P_\sigma(i\vec{k}_0)t}\right| \leq C\left|\mathrm{e}^{\sigma Jt}\right| \leq C_0(1+\sigma t)\mathrm{e}^{-\frac{1}{2}\sigma t} \leq C_0(1+4\mathrm{e})\mathrm{e}^{-\frac{1}{4}\sigma t},$$

where we have used $x\mathrm{e}^{-x} \leq \mathrm{e}$. On the other hand, by continuity of $\vec{v} \to \mathcal{M}(\vec{v})$ we know there exists $\rho > 0$ such that

$$\begin{aligned} \forall\ \vec{v}_0 \in \mathcal{N}_1 \quad |\vec{v} - \vec{v}_0| < \rho \quad &\Longrightarrow \quad |\mathcal{M}(\vec{v}) - \mathcal{M}(\vec{v}_0)| \leq \frac{1}{8C_0(1+4e)} \\ &\Longrightarrow \quad \left|P_\sigma(i\vec{k}) - P_\sigma(i\vec{k}_0)\right| \leq \frac{\sigma}{8C_0(1+4e)}. \end{aligned}$$

It is then very easy to conclude thanks to Lemma 1.6 (take $A = P_\sigma(i\vec{k}_0)$, $K = P_\sigma(i\vec{k}) - P_\sigma(i\vec{k}_0)$, $C = C_0(1+4e)$ and $\lambda = \sigma/4$). □

We set

$$\mathcal{D}_1 = \bigcup_{\vec{v}_0 \in \mathcal{N}_1} B(\vec{v}_0, \rho), \quad \mathcal{D}_2 = \mathbb{R}^2 \setminus \mathcal{D}_1,$$

where $B(\vec{v}_0, \rho)$ is the open ball of center $\vec{v}_0$ and of radius ρ, ρ being given by Lemma 1.7, see Figure 2. Lemma 1.7 shows that $e^{\sigma\mathcal{M}(\vec{v})t}$ is uniformly bounded on $\mathcal{D}_1$.

It remains to analyze its behavior for $\vec{v} \in \mathcal{D}_2$. From Lemma 1.5, we know that for any $\vec{v} \in \mathcal{D}_2$, $\mathcal{M}(\vec{v})$ is diagonalizable, which means that there exists a family of invertible matrices $S_\sigma(\vec{v})$ so that

$$\mathcal{M}(\vec{v}) = S(\vec{v})\Delta(\vec{v})S(\vec{v})^{-1},$$

where

$$\Delta(\vec{v}) = \mathrm{diag}\{\nu_j(\vec{v})\}.$$

Therefore

$$\mathrm{e}^{\sigma\mathcal{M}(\vec{v})\ t} = S(\vec{v})\ \mathrm{e}^{\sigma\Delta(\vec{v})\ t}\ S(\vec{v})^{-1}. \tag{32}$$

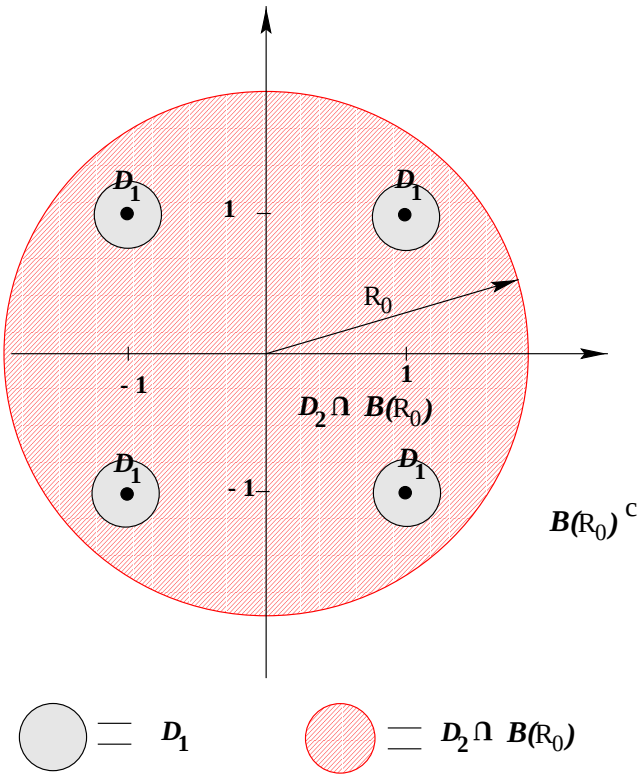

FIGURE 2. $\mathcal{D}_1$: the domain near the points where $P_\sigma(i\vec{k})$ is not diagonalizable, $\mathcal{D}_2$: the complementary domain ($\mathcal{D}_2 = (\mathcal{D}_2 \cap \overline{B}_{R_0}) \cup \overline{B}^c_{R_0}$).

In Lemma 1.5 (i) we proved that the eigenvalues have a negative real part; this implies that

$$\left|e^{\sigma \mathcal{M}(\vec{v})\, t}\right| \leq |S(\vec{v})|\ \left|S(\vec{v})^{-1}\right|. \tag{33}$$

In the sequel, we shall choose the matrix $S(\vec{v})$ as:

$$S(\vec{v}) = [S_1\ S_2\ S_3\ S_4] \tag{34}$$

where the S_j are the normalized eigenvectors of $\mathcal{M}(\vec{v})$, so that $|S(\vec{v})| = 1$. One can check that:

$$S_j = \frac{\widehat{S}_j}{\left|\widehat{S}_j\right|},$$

where

$$\widehat{S}_j = \begin{pmatrix} -v_x v_y \nu_j(\vec{v}) \\ -(\nu_j(\vec{v})+1)((\nu_j(\vec{v}))^2 + (v_y)^2) \\ i v_x ((\nu_j(\vec{v}))^2 + (v_y)^2) \\ -i (v_y)^2 v_x \end{pmatrix}. \tag{35}$$

Before concluding, we need a result on their behavior for large $|\vec{v}|$. Using the expansion of $\nu_j(\vec{v})$ for large $\vec{v}$ in (35), one can check the following result:

Lemma 1.8. *There exist two constants $R_0 > 0$ and $C_1 > 0$ such that for any $\vec{v}$ such that $|\vec{v}| \geq R_0$ then the matrix $S^{-1}(\vec{v})$ satisfies:*

$$S^{-1}(\vec{v}) = |\vec{v}| S_{-1}^{-1}(\vec{K}) + R(\vec{v}), \qquad \vec{K} = \vec{v}/|\vec{v}| \tag{36}$$

where

$$S_{-1}^{-1}(K) = \begin{pmatrix} iK_x & iK_y & 0 & 0 \\ -iK_x & -iK_y & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{37}$$

and

$$|R(\vec{v})| \leq C_1. \tag{38}$$

Let $\overline{B}_{R_0}$ the ball of radius R_0, with R_0 defined in Lemma 1.8. In the sequel, we assume that R_0 is chosen large enough, so that $\mathcal{D}_1 \subset \overline{B}_{R_0}$ and $\mathcal{D}_2 \cap \overline{B}_{R_0}^c = \overline{B}_{R_0}^c$, see Figure 2. We are now able to estimate the exponential of the matrix $\sigma \mathcal{M}(\vec{v})t$ for $\vec{v}$ in $\mathcal{D}_2$:

Lemma 1.9. *(i) There exists a constant $C_2 = C_2(R_0) > 0$ such that for any $\vec{v} \in \mathcal{D}_2 \cap \overline{B}_{R_0}$, we have*

$$\left| e^{\sigma \mathcal{M}(\vec{v})t} \right| \leq C_2. \tag{39}$$

(ii) There exists a constant $C_3 > 0$ such that for any $\vec{v} \in \overline{B}_{R_0}^c$, we have

$$\left| e^{\sigma \mathcal{M}(\vec{v})t} \right| \leq C_3 \sqrt{1 + |\vec{v}|^2}. \tag{40}$$

Proof. We use that the matrices $S(\vec{v})$ and $S^{-1}(\vec{v})$ are continuous functions of $\vec{v}$ in $\mathcal{D}_2$ [8].

(i) The continuity of $S^{-1}(\vec{v})$ in $\mathcal{D}_2 \cap \overline{B}_{R_0}$ implies that there exists a constant $C_4(R_0) > 0$ such that

$$\forall\ \vec{v} \in \mathcal{D}_2 \cap \overline{B}_{R_0}, \quad \left| S^{-1}(\vec{v}) \right| \leq C_4(R_0).$$

Using (33) and $|S(\vec{v})| = 1$, we get (39).

(ii) For large $|\vec{v}|$, Lemma 1.8 implies that

$$\forall\ \vec{v} \in \mathcal{D}_2,\ |\vec{v}| \geq R_0 \Longrightarrow \left| S^{-1}(\vec{v}) \right| \leq |\vec{v}| \left| S_{-1}^{-1}(\vec{K}) \right| + C_1 \leq C_5 \sqrt{1 + |\vec{v}|^2}$$

since $S_{-1}^{-1}(\vec{K})$ is obviously bounded. This gives (40). □

Proof of Theorem 1.2. We conclude the proof of Theorem 1.2 *via* Plancherel's theorem since, $\widehat{U}(\vec{k},t)$ denoting the space Fourier transform of $U(.,t)$,

$$\widehat{U}(\vec{k},t) = \mathrm{e}^{P_\sigma(i\vec{k})t}\ \widehat{U}^0(\vec{k}). \tag{41}$$

Collecting the results of Lemmas 1.7 and 1.9, one easily sees that there exists a constant $C>0$ such that

$$\left|\mathrm{e}^{\sigma\mathcal{M}(\vec{v})t}\right| \leq C\sqrt{1+|\vec{v}|^2}, \quad \forall \vec{v}\in\mathbb{R}^2,$$

which yields

$$\left|\widehat{U}(\vec{k},t)\right| \leq C\sqrt{1+|\vec{v}|^2}\left|\widehat{U}^0(\vec{k})\right| = C\frac{\sqrt{\sigma^2+|\vec{k}|^2}}{\sigma}\left|\widehat{U}^0(\vec{k})\right| \leq C_\sigma\sqrt{1+|\vec{k}|^2}\left|\widehat{U}^0(\vec{k})\right|,$$

with $C_\sigma = \min(1,1/\sigma)$.

□

Proof of Theorem 1.3. The loss of regularity came from the behavior of the matrix exponential for large $\vec{v}$. The idea to get a finer result is to split the domain of integration into the three domains, $\mathcal{D}_1$, $\mathcal{D}_2\cap\overline{B}_{R_0}$ and $\mathcal{D}_2\cap\overline{B}^c_{R_0}\equiv\overline{B}^c_{R_0}$, and to use Lemma 1.8 for estimating $S^{-1}(\vec{v})$ in the domain $\overline{B}^c_{R_0}$.

Parseval's theorem implies

$$\|U(t)\|_0^2 = \int\left|\widehat{U}(\vec{k},t)\right|^2\mathrm{d}\vec{k} = I_1+I_2+I_3, \tag{42}$$

where

$$\begin{cases} I_1 = \displaystyle\int_{\vec{k}/\sigma\in\mathcal{D}_1}\left|\widehat{U}(\vec{k},t)\right|^2\mathrm{d}\vec{k} = \int_{\vec{k}/\sigma\in\mathcal{D}_1}\left|\mathrm{e}^{P_\sigma(i\vec{k})t}\ \widehat{U}^0(\vec{k})\right|^2\mathrm{d}\vec{k} \\ I_2 = \displaystyle\int_{\vec{k}/\sigma\in\mathcal{D}_2\cap\overline{B}_{R_0}}\left|\widehat{U}(\vec{k},t)\right|^2\mathrm{d}\vec{k} = \int_{\vec{k}/\sigma\in\mathcal{D}_2\cap\overline{B}_{R_0}}\left|\mathrm{e}^{P_\sigma(i\vec{k})t}\ \widehat{U}^0(\vec{k})\right|^2\mathrm{d}\vec{k} \\ I_3 = \displaystyle\int_{\vec{k}/\sigma\in\overline{B}^c_{R_0}}\left|\widehat{U}(\vec{k},t)\right|^2\mathrm{d}\vec{k} = \int_{\vec{k}/\sigma\in\overline{B}^c_{R_0}}\left|\mathrm{e}^{P_\sigma(i\vec{k})t}\ \widehat{U}^0(\vec{k})\right|^2\mathrm{d}\vec{k}. \end{cases}$$

- Lemma 1.7 implies that there is a constant $C>0$ such that

$$I_1 \leq C\int_{\vec{k}/\sigma\in\mathcal{D}_1}\left|\widehat{U}^0(\vec{k})\right|^2\mathrm{d}\vec{k}. \tag{43}$$

- Lemma 1.9-(i) implies that there is a constant $C_2>0$ such that

$$I_2 \leq C_2\int_{\vec{k}/\sigma\in\mathcal{D}_2\cap\overline{B}_{R_0}}\left|\widehat{U}^0(\vec{k})\right|^2\mathrm{d}\vec{k}. \tag{44}$$

- It remains to estimate I_3, which requires to estimate

$$\widehat{U}(\vec{k},t) = \mathrm{e}^{P_\sigma(i\vec{k})t}\ \widehat{U}^0(\vec{k}),$$

for large $\vec{k}$ ($|\vec{k}| \geq \sigma R_0$). In this domain, $P_\sigma(i\vec{k})$ can be diagonalized so that, using (32), we can write:

$$\widehat{U}(\vec{k},t) = S(\vec{v})\ \mathrm{e}^{\sigma\Delta(\vec{v})\ t}\ S(\vec{v})^{-1}\widehat{U}^0(\vec{k}), \quad (\vec{v} = \frac{\vec{k}}{\sigma})$$

that can be decomposed, thanks to (36) as

$$\widehat{U}(\vec{k},t) = V(\vec{k},t) + W(\vec{k},t),$$

with

$$V(\vec{k},t) = |\vec{v}|S(\vec{v})\ \mathrm{e}^{\sigma\Delta(\vec{v})\ t}\ S_{-1}^{-1}(\vec{K})\ \widehat{U}^0(\vec{k}),$$

and

$$W(\vec{k},t) = S(\vec{v})\ \mathrm{e}^{\sigma\Delta(\vec{v})\ t}\ R(\vec{v})\ \widehat{U}^0(\vec{k}).$$

From (38) and Lemma 1.5-(i), we get

$$\left|W(\vec{k},t)\right| \leq C\left|\widehat{U}^0(\vec{k})\right|.$$

Let us look more carefully at $V(\vec{k},t)$. Denoting by $S_{ij}(\vec{v})$ the entries of $S(\vec{v})$ and taking into account the particular form of S_{-1}^{-1} (37) and the fact that $\lambda_2(\sigma,\vec{k}) = \overline{\lambda_1}(\sigma,\vec{k})$, we get:

$$V(\vec{k},t) = (i\vec{K}.\vec{E}^0)\ |\vec{v}| \begin{pmatrix} S_{11}(\vec{v})e^{\lambda_1(\sigma,\vec{k})t} - S_{12}(\vec{v})e^{\overline{\lambda_1}(\sigma,\vec{k})t} \\ S_{21}(\vec{v})e^{\lambda_1(\sigma,\vec{k})t} - S_{22}(\vec{v})e^{\overline{\lambda_1}(\sigma,\vec{k})t} \\ S_{31}(\vec{v})e^{\lambda_1(\sigma,\vec{k})t} - S_{32}(\vec{v})e^{\overline{\lambda_1}(\sigma,\vec{k})t} \\ S_{41}(\vec{v})e^{\lambda_1(\sigma,\vec{k})t} - S_{42}(\vec{v})e^{\overline{\lambda_1}(\sigma,\vec{k})t} \end{pmatrix}. \tag{45}$$

Looking at the coefficients of matrix S, we get for large $|\vec{v}|$

$$\begin{cases} |S_{11}(\vec{v})| \leq \dfrac{C}{|\vec{v}|},\ |S_{12}(\vec{v})| \leq \dfrac{C}{|\vec{v}|},\ |S_{21}(\vec{v})| \leq \dfrac{C}{|\vec{v}|},\ |S_{22}(\vec{v})| \leq \dfrac{C}{|\vec{v}|} \\ \left|S_{31}(\vec{v}) - \dfrac{i}{2}\right| \leq \dfrac{C}{|\vec{v}|}, \quad \left|S_{32}(\vec{v}) - \dfrac{i}{2}\right| \leq \dfrac{C}{|\vec{v}|}, \\ \left|S_{41}(\vec{v}) + \dfrac{i}{2}\right| \leq \dfrac{C}{|\vec{v}|}, \quad \left|S_{42}(\vec{v}) + \dfrac{i}{2}\right| \leq \dfrac{C}{|\vec{v}|}. \end{cases} \tag{46}$$

We thus notice that the behavior of the two first components of $V(\vec{k},t)$ differ from the one of the two last components. This is due to the fact that for $1 \leq i,j \leq 2$, $|\vec{v}|S_{ij}(\vec{v})$ remains bounded. More precisely, this leads to the following expression

$$V(\vec{k},t) = i\vec{K}.\vec{E}^0\ |\vec{v}|\sin(\Im m\lambda_1 t)\mathrm{e}^{\Re e\lambda_1 t}(0\ 0\ 1\ -1)^t + V_R(\vec{k},t), \tag{47}$$

where

$$\left|V_R(\vec{k},t)\right| \leq C\left|\vec{E}^0\right|,$$

from which we deduce immediately

$$\left|V_1(\vec{k},t)\right| + \left|V_2(\vec{k},t)\right| \leq C\left|\vec{E}^0\right|.$$

One can also notice on (47) that the sum of the two last components of the principal part vanishes so that we get

$$\left|V_3(\vec{k},t) + V_4(\vec{k},t)\right| \leq C\left|\vec{E}^0\right|.$$

These two results, together with (43) and (44) yields easily (28). We also deduce from (46) and (45) another estimate for the splitted unknowns H^x and H^y:

$$\begin{aligned}\left|V_3(\vec{k},t)\right| + \left|V_4(\vec{k},t)\right| &\leq C|\vec{v}|\left|\vec{E}^0(\vec{k})\right||\sin(\Im m\lambda_1 t)| = C\frac{\left|\vec{k}\right|}{\sigma}\left|\vec{E}^0(\vec{k})\right||\sin(\Im m\lambda_1 t)| \\ &\leq C\left|\vec{k}\right| t\left|\vec{E}^0(\vec{k})\right|,\end{aligned}$$

the last inequality deriving from $|\sin x| \leq |x|$ and (31). This shows that for large $\vec{k}$, namely for $\vec{k}/\sigma \in \overline{B}^c_{R_0}$, one has

$$\frac{1}{\sqrt{1+|\vec{k}|^2}}\left(\left|\widehat{H}^x(\vec{k},t)\right| + \left|\widehat{H}^y(\vec{k},t)\right|\right) \leq Ct\left|\widehat{U}^0(\vec{k})\right|$$

which is still true in the other domains, in which $\widehat{H}^x$ and $\widehat{H}^y$ remain bounded. This finally implies (29), which ends the proof of Theorem 1.3. □

2. Analysis of the Bérenger system by energy techniques

2.1. **Introduction**

As we recalled in Section 1.1.1, the analysis of Maxwell's system, which is symmetric, can be done with the help of energy estimates. It is thus tempting to try to obtain analogous estimates for the PML system. This is in fact not so trivial. Recently, Vacus and Metral [10] proposed an analyze of the initial value problem associated to (13) *via* the semi-group theory (Hille-Yosida's theorem) which implies to establish appropriate energy estimates. Their estimates, which differ from the ones we shall present in this paper, point out a difference in the behaviors of $\vec{E}, H$ and H^x. More precisely they replace the Bérenger's formulation posed in the unknown $U = (\vec{E}, H^x, H^y)$ by an equivalent formulation in the unknowns $(\vec{E}, H, G)$ where $G = \sigma H^x$ and show that for initial data $(\vec{E}^0, H^0, G^0) \in H^1 \times H^1 \times L^2$, there exist positive constants $C_1, C_2 > 0$ such that

$$\|\vec{E}(t)\|_{H^1} + \|H(t)\|_{H^1} + \|\sigma H^x(t)\|_{L^2} \leq C_1 e^{C_2 t}(\|\vec{E}^0\|_{H^1} + \|H^0\|_{H^1} + \|(\sigma H^x)^0\|_{L^2}). \tag{48}$$

However this result does not distinguish between $\sigma > 0$ and $\sigma < 0$, and the estimate on H^x is lost when σ tends to zero.

We have chosen here to analyze (13) by considering the formulation of the PML model proposed by Zhao-Cangellaris in [15]. This formulation is equivalent to the Bérenger's formulation in a sense that we will make precise in Section 2.2. In Section 2.3, we shall come back to the case of a constant $\sigma > 0$ already considered in Section 1 and will establish an energy decay result. In Section 2.4, we shall show how to obtain energy estimates when σ can be a discontinuous function of x. We first consider the model with a layer parallel to the y axis (Sect. 2.4.1) and then the model in a corner (Sect. 2.4.2).

2.2. **Zhao-Cangellaris's formulation and its equivalence with Bérenger's formulation**

2.2.1. *Zhao-Cangellaris's formulation in a layer parallel to the y axis*

Bérenger's formulation involves a splitting of the unknown H, which leads to a splitting of the usual space differential operator rot. The idea of Zhao and Cangellaris [15] is to restore the usual operator by introducing a new unknown. Let us present how to go from the formulation of Bérenger to the one of Zhao and Cangellaris.

Let (E_x, E_y, H^x, H^y) solution of the Bérenger's system (13), where $\sigma = \sigma(x)$ with initial conditions (E_x^0, E_y^0, $(H^x)^0$, $(H^y)^0$). Applying ∂_t to (c) , $\partial_t + \sigma$ to (d) and adding the two terms gives

$$\partial_t(\partial_t + \sigma)H^x + \partial_t(\partial_t + \sigma)H^y + \partial_t\partial_x E_y - (\partial_t + \sigma)\partial_y E_x = 0.$$

Since σ does not depend on y, the operators $\partial_t + \sigma$ and ∂_y commute and we get, setting $H = H^x + H^y$,

$$\partial_t\left((\partial_t + \sigma)H + \partial_x E_y\right) - \partial_y(\partial_t + \sigma)E_x = 0.$$

In order to transform the last term into a time derivative, we introduce a new variable $\widetilde{E}_x$ satisfying

$$(\partial_t + \sigma)E_x = \partial_t \widetilde{E}_x, \tag{49}$$

and we then obtain

$$\partial_t\left((\partial_t + \sigma)H + \partial_x E_y - \partial_y \widetilde{E}_x\right) = 0.$$

If we make the assumption that at $t = 0$:

$$(\partial_t H)^0 + \sigma H^0 + \partial_x E_y^0 - \partial_y \widetilde{E}_x^0 = 0, \tag{50}$$

we get the Zhao Cangellaris formulation

$$\begin{cases} \partial_t E_x - \partial_y H = 0 & (a) \\ \partial_t E_y + \sigma E_y + \partial_x H = 0 & (b) \\ \partial_t H + \sigma H + \partial_x E_y - \partial_y \widetilde{E}_x = 0 & (c) \\ (\partial_t + \sigma)E_x = \partial_t \widetilde{E}_x. & (d) \end{cases} \tag{51}$$

Note that (49) does not completely define $\widetilde{E}_x$. We have to prescribe the initial value for $\widetilde{E}_x$. We shall choose:

$$\widetilde{E}_x^0 = E_x^0, \tag{52}$$

which in particular implies that $\widetilde{E}_x = E_x$ everywhere $\sigma = 0$. Moreover, from ((13),(c) + (d)), written at time $t = 0$, we get, at least for sufficiently smooth solutions:

$$(\partial_t H)^0 = \partial_x E_y^0 - \partial_y E_x^0 - \sigma(H^x)^0,$$

from which we deduce that condition (50) is equivalent to:

$$\sigma(H^y)^0 + \partial_y(E_x^0 - \widetilde{E}_x^0) = 0.$$

With the choice (52), this reduces to:

$$\sigma(H^y)^0 = 0.$$

This condition appears as a compatibility condition for the initial data in order to pass from Bérenger's formulation to Zhao-Cangellaris'one.

Reciprocally, assume that $(\widetilde{E}_x, E_x, E_y, H)$ satisfies (51) with initial conditions $(\widetilde{E}_x^0, E_x^0, E_y^0, H^0)$, we first define H^y such that

$$\partial_t H^y = \partial_y E_x \tag{53}$$

choosing $(H^y)^0$ such that

$$\sigma(H^y)^0 = \partial_y E_x^0 - \partial_y \widetilde{E}_x^0 \tag{54}$$

(*i.e.* $\sigma(H^y)^0 = 0$ if $E_x^0 = \widetilde{E}_x^0$) and in a second step, we define H^x by $H^x = H - H^y$. It is then trivial to check that (E_x, E_y, H^x, H^y) is a solution of (13). The two formulations are thus "equivalent" in the sense that one can deduce a solution of one of them from the solution of the other one.

Remark 2.1.

- The initial motivation for using formulation (51) was to be able to write a variational formulation and to use finite elements for the numerical approximation.
- In [11], the authors showed that this system (not exactly this one but the one corresponding to the (TM) mode) could be reinterpreted as a zero order perturbation of the usual Maxwell's system, which is strongly well-posed. The general theory [9] permits to conclude that (51) remains strongly well-posed. This is the point we shall point out in a more explicit way in the next sections.
- One could be surprised by the fact that a weakly well-posed system (namely (13)) is "equivalent" to a strongly well-posed one (namely (51)). In fact, as we pointed out above, "equivalence"only means that one can easily deduce the solution of one system from the solution of the other one. However, their respective solutions are different and do not have the same regularity. This is for instance particularly clear from equation (53): we need E_x in H^1 (at least $\partial_y E^x$ in L^2) in order to get H^x in L^2.
- In some sense, the idea of restoring the usual Maxwell's operator (keeping as unknowns E and H) is similar to the one used in [10] (see Sect. 3.1) The difference comes from the choice of the fourth unknown. In [10], it was $G = \sigma H^x$ which was "completely part of the system", whereas in (51) the new unknown $\widetilde{E}_x$ is only related to E_x thanks to (d), *i.e.*, it is simply defined as a primitive in time of E_x. This difference is essential, since the loss of regularity in the first approach came from G, whereas in this second approach the regularity on $\widetilde{E}_x$ will be the same than on E_x, and it will lead to a strongly well-posed formulation, as we will see later.

In order to restore the classical Maxwell's operator, it is useful to adopt new notations for E_x and $\widetilde{E}_x$. This will make the calculations of the next sections more readable.

- E_x is denoted $E_x^\star$,
- $\widetilde{E}_x$ is denoted E_x.

Then, Zhao-Cangellaris's formulation can be rewritten as follows

$$\begin{cases} \partial_t E_x^\star - \partial_y H = 0 & (a) \\ \partial_t E_y + \sigma E_y + \partial_x H = 0 & (b) \\ \partial_t H + \sigma H + \partial_x E_y - \partial_y E_x = 0 & (c) \\ (\partial_t + \sigma) E_x^\star = \partial_t E_x. & (d) \end{cases} \tag{55}$$

One can notice that (55) can also be written as follows

$$\begin{cases} \partial_t \vec{E}^\star - \vec{\mathrm{rot}} H = 0 & (a) \\ \partial_t H^\star + \mathrm{rot} \vec{E} = 0 & (b) \end{cases} \tag{56}$$

if we introduce the new variables $E_y^\star$ and $H^\star$ such that

$$\begin{cases} (\partial_t + \sigma) H = \partial_t H^\star & (i) \\ (\partial_t + \sigma) E_y = \partial_t E_y^\star & (ii) \\ H^\star(0) = H(0), \quad E_y^\star(0) = E_y(0) \end{cases} \tag{57}$$

or equivalently, using (55)-(d)

$$\begin{cases} \partial_t \vec{E} + \Sigma \vec{E} = \partial_t \vec{E}^\star + \Sigma^\star \vec{E}^\star & (i) \\ (\partial_t + \sigma) H = \partial_t H^\star & (ii) \end{cases} \tag{58}$$

with

$$\Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix} \; ; \quad \Sigma^\star = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}. \tag{59}$$

2.2.2. *The PML model in a corner domain*

Until now, we have only considered the case of a damping in the x direction, corresponding to a layer parallel to y. It is straightforward to get the results for a layer parallel to x. On the other hand, the situation is changed when considering a corner. In this case, it is necessary to introduce a damping in each direction, $\sigma_x = \sigma_x(x)$ and $\sigma_y = \sigma_y(y)$. The Zhao-Cangellaris formulation can then be written in the following matricial form:

$$\begin{cases} \partial_t \vec{E}^\star - \vec{\mathrm{rot}} H = 0 & (a) \\ \partial_t H^\star + \mathrm{rot} \vec{E} = 0 & (b) \end{cases} \tag{60}$$

where the link between variables with and without star is:

$$\begin{cases} \partial_t \vec{E} + \Sigma \vec{E} = \partial_t \vec{E}^\star + \Sigma^\star \vec{E}^\star & (i) \\ \partial_{tt}^2 H + (\sigma_x + \sigma_y) \partial_t H + \sigma_x \sigma_y H = \partial_{tt}^2 H^\star & (ii) \end{cases} \tag{61}$$

with

$$\Sigma = \begin{pmatrix} \sigma_y & 0 \\ 0 & \sigma_x \end{pmatrix} \; ; \quad \Sigma^\star = \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix} \tag{62}$$

and we assume all the initial data to be zero. Note that in this case H and $H^\star$ are related through a second-order equation and not a first-order one as previously.

2.3. **Energy decay results for $\sigma > 0$ constant**

2.3.1. *Formulation in a layer parallel to the y axis*

We come back to the 2D Cauchy problem associated to (13), rewritten in the form (56). The first energy that we consider is not the classical one, but involves first order time derivatives of the electromagnetic field:

$$\mathcal{E}_1(t) = \frac{1}{2}\left(\|\partial_t E_x\|_{L^2}^2 + \|\partial_t E_y\|_{L^2}^2 + \|\partial_t H + \sigma H\|_{L^2}^2 + \left\|\partial_t (E_y - E_y^\star)\right\|_{L^2}^2 \right). \tag{63}$$

Lemma 2.2. *For a positive constant value of σ, the energy $\mathcal{E}_1$ of the solution of (13) satisfies the identity:*

$$\partial_t \mathcal{E}_1 = -2\sigma \|\partial_t E_y\|_{L^2}^2 \leq 0 \tag{64}$$

which means in particular that it is a decreasing function of time.

Proof. In order to eliminate $E_x^\star$, we apply to the first equation (55)-(a) the operator $\partial_t + \sigma$ and get, since for constant σ (σ independent of y is sufficient) $(\partial_t + \sigma)\partial_y = \partial_y(\partial_t + \sigma)$:

$$(a)' \quad \partial_{tt}^2 E_x - \partial_y(\partial_t + \sigma)H = 0.$$

Multiplying this equation by $\partial_t E_x$ gives

$$(i) \quad \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\partial_t E_x\|_{L^2}^2 - (\partial_y(\partial_t + \sigma)H, \partial_t E_x)_{L^2} = 0.$$

We then apply $\partial_t + \sigma$ to (55)-(b) and multiply it by $\partial_t E_y$:

$$\begin{gathered} ((\partial_t + \sigma)^2 E_y, \partial_t E_y)_{L^2} + ((\partial_t + \sigma)\partial_x H, \partial_t E_y)_{L^2} = 0 \\ \Downarrow \\ (ii) \quad \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\partial_t E_y\|_{L^2}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\sigma E_y\|_{L^2}^2 + (2\sigma\partial_t E_y, \partial_t E_y)_{L^2} + ((\partial_t + \sigma)\partial_x H, \partial_t E_y)_{L^2} = 0. \end{gathered}$$

We now differentiate (55)-(c) with respect to t and multiply it by $(\partial_t + \sigma)H$ which gives (using an integration by parts)

$$\begin{gathered} (\partial_t(\partial_t + \sigma)H, (\partial_t + \sigma)H)_{L^2} + (\partial_x \partial_t E_y - \partial_y \partial_t E_x, (\partial_t + \sigma)H)_{L^2} = 0 \\ \Downarrow \\ \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|(\partial_t + \sigma)H\|_{L^2}^2 + (\partial_x \partial_t E_y, (\partial_t + \sigma)H)_{L^2} - (\partial_y \partial_t E_x, (\partial_t + \sigma)H)_{L^2} = 0 \\ \Downarrow \\ (iii) \quad \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|(\partial_t + \sigma)H\|_{L^2}^2 - (\partial_t E_y, \partial_x(\partial_t + \sigma)H)_{L^2} + (\partial_t E_x, \partial_y(\partial_t + \sigma)H)_{L^2} = 0. \end{gathered}$$

When σ is constant, we have $(\partial_t+\sigma)\partial_x = \partial_x(\partial_t+\sigma)$ and thus, adding (i), (ii) and (iii), we get:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t E_x\right\|_{L^2}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\partial_t E_y\right\|_{L^2}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\sigma E_y\right\|_{L^2}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|(\partial_t+\sigma)H\right\|_{L^2}^2 + 2\sigma\left\|\partial_t E_y\right\|_{L^2}^2 = 0$$

which implies (64) (one uses (57)-(ii)). □

Remark 2.3. For a varying positive σ, for instance σ a sufficiently smooth function of x, we do not have anymore $(\partial_t+\sigma)\partial_x = \partial_x(\partial_t+\sigma)$ but $\partial_x(\partial_t+\sigma) = (\partial_t+\sigma)\partial_x + \partial_x\sigma$. Adding (i), (ii) and (iii) yields

$$\partial_t\mathcal{E}_1 + (2\sigma\partial_t E_y, \partial_t E_y)_{L^2} = \int (\partial_x\sigma)\ \partial_t E_y\ H\ \mathrm{d}x.$$

It is easy, *via* Gronwall's lemma, to obtain an exponential (in time) bound of $\mathcal{E}_1$ that involves the L^∞ norm of $\partial_x\sigma$. However, it is not at all obvious to get a decay in time: this question remains open.

2.3.2. *The PML model in a corner domain*

We consider the PML model in a corner domain, with positive constant damping coefficients $\sigma_x > 0$ and $\sigma_y > 0$, given by (60)-(61). We only give the result, which corresponds to Lemma 2.2. It is obtained in a similar way.

Lemma 2.4. *For positive constant damping coefficients, the solution of (60)-(61) satisfies the following energy decay:*

$$\mathcal{E}_2(t) \le \mathcal{E}_2(s), \quad \forall t \ge s$$

where $\mathcal{E}_2$ is the second-order energy defined as:

$$\mathcal{E}_2(t) = \frac{1}{2}\left\|\partial_{tt}^2\vec{E}\right\|_{L^2}^2 + \frac{1}{2}\left\|\sigma_y\partial_t E_x\right\|_{L^2}^2 + \frac{1}{2}\left\|\sigma_x\partial_t E_y\right\|_{L^2}^2 + \frac{1}{2}\left\|(\partial_t+\sigma_x)(\partial_t+\sigma_y)H\right\|_{L^2}^2. \tag{65}$$

2.4. Energy estimates based on Zhao-Cangellaris's formulation in the case of general variable coefficients

The superiority of energy techniques, when available, over Fourier techniques, lies in the fact that they can be applied to the case of non constant, even non smooth, coefficients. Our conjecture is that the energy decay result of Section 2.3 should be generalizable to non constant σ (including discontinuous σ). However, we have already seen in Remark 2.3 that such a generalization is not straightforward. As a matter of fact, we have not been able to produce an energy decay in this case. However, energy techniques allow us to prove the well-posedness in L^2 of the Bérenger system under the only assumption that $\sigma \in L^\infty$. This is the object of the next section.

2.4.1. *The PML model in a layer parallel to the y axis*

In this section, σ is supposed to be a positive bounded function of x. In this case, one can show an energy estimate, with a zero-order energy defined as:

$$\widetilde{\mathcal{E}}_0(t) = \frac{1}{2}\left(\left\|E_x\right\|_{L^2}^2 + \left\|E_y\right\|_{L^2}^2 + \left\|H\right\|_{L^2}^2 + \left\|E_x - E_x^\star\right\|_{L^2}^2\right). \tag{66}$$

Note that $\widetilde{\mathcal{E}}_0$ differs from the energy $\mathcal{E}_0$ introduced in the case of a constant σ: H is replaced by $H^\star$ and $E_x - E_x^\star$ is replaced by $E_y - E_y^\star$.

Theorem 2.5. *We have the following energy estimate*

$$\widetilde{\mathcal{E}}_0(t) \le \widetilde{\mathcal{E}}_0(0)e^{2\|\sigma\|_\infty t}.$$

Proof. We multiply (56,(a)) by $\vec{E}$, (56,(b)) by H, add the two resulting equalities and integrate in space. Thanks to Green's formula (rot is the adjoint of $\vec{rot}$) we get:

$$(\partial_t E_x^\star, E_x)_{L^2} + (\partial_t E_y^\star, E_y)_{L^2} + (\partial_t H^\star, H)_{L^2} = 0. \tag{67}$$

We then multiply (57)-(ii) with E_y and (57)-(i) with H to rewrite the two last terms as

$$\begin{aligned} (\partial_t E_y^\star, E_y)_{L^2} &= (\partial_t E_y + \sigma E_y, E_y)_{L^2} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|E_y\|_{L^2}^2 + (\sigma E_y, E_y)_{L^2} \\ (\partial_t H^\star, H)_{L^2} &= (\partial_t H + \sigma H, H)_{L^2} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|H\|_{L^2}^2 + (\sigma H, H)_{L^2}. \end{aligned} \tag{68}$$

It remains to treat the term $\partial_t E_x^\star$ with E_x. This product can be rewritten, thanks to (55)(d), as:

$$(\partial_t E_x^\star, E_x)_{L^2} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|E_x\|_{L^2}^2 - (\sigma E_x^\star, E_x)_{L^2}, \tag{69}$$

and (67) becomes

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\|E_x\|_{L^2}^2 + \|E_y\|_{L^2}^2 + \|H\|_{L^2}^2) - (\sigma E_x^\star, E_x)_{L^2} + (\sigma E_y, E_y)_{L^2} + (\sigma H, H)_{L^2} = 0. \tag{70}$$

Equation (55)(d) is also equivalent to $\partial_t(E_x - E_x^\star) = \sigma E_x^\star$ that we multiply by $E_x - E_x^\star$ to get

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|E_x - E_x^\star\|_{L^2}^2 &= (\sigma E_x^\star, E_x - E_x^\star)_{L^2} \\ &= (\sigma(E_x^\star - E_x), E_x - E_x^\star)_{L^2} + (\sigma E_x, E_x)_{L^2} - (\sigma E_x, E_x^\star)_{L^2}. \end{aligned} \tag{71}$$

This gives

$$-(\sigma E_x^\star, E_x)_{L^2} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|E_x - E_x^\star\|_{L^2}^2 + (\sigma(E_x - E_x^\star), E_x - E_x^\star)_{L^2} - (\sigma E_x, E_x)_{L^2}. \tag{72}$$

Combining (70) and (72), we get:

$$\left\{ \begin{array}{l} \dfrac{1}{2}\dfrac{\mathrm{d}}{\mathrm{d}t}\left(\|E_x\|_{L^2}^2 + \|E_y\|_{L^2}^2 + \|H\|_{L^2}^2 + \|E_x - E_x^\star\|_{L^2}^2\right) = \\ -(\sigma E_y, E_y)_{L^2} - (\sigma H, H)_{L^2} - (\sigma(E_x - E_x^\star), E_x - E_x^\star)_{L^2} + (\sigma E_x, E_x)_{L^2}. \end{array} \right. \tag{73}$$

The right-hand side can be bounded by $2\|\sigma\|_\infty \widetilde{\mathcal{E}}_0$. Therefore, after integration in time

$$\widetilde{\mathcal{E}}_0(t) \leq \widetilde{\mathcal{E}}_0(0) + 2\|\sigma\|_\infty \int_0^t \widetilde{\mathcal{E}}_0(s)\mathrm{d}s.$$

Applying Gronwall's lemma yields

$$\widetilde{\mathcal{E}}_0(t) \leq \widetilde{\mathcal{E}}_0(0) e^{2\|\sigma\|_\infty t}.$$

□

Remark 2.6. In fact, the reader will remark that our result is independent of the sign of σ. In fact, restricting ourselves to $\sigma > 0$ does not really help in the proof. For instance, in the right-hand side of (73), the first three terms would be negative (and would then contribute to an energy decay) but the last one would be positive.

2.4.2. *The PML model in a corner domain*

In what follows, we shall assume that the product $\sigma_x\sigma_y$ remains positive everywhere and introduce the new first order energy:

$$\mathcal{E}_0^{xy} = \frac{1}{2}\left(\|\vec{E}^\star - \vec{E}\|_0^2 + \|\vec{E}\|^2 + \|H\|_{L^2}^2 + (\sigma_x\sigma_y\widetilde{H},\widetilde{H})_{L^2}\right),$$

where

$$\widetilde{H}(.,t) = \int_0^t H(.,s)\mathrm{d}s. \tag{74}$$

It is still possible to get an energy estimate and we show that:

Theorem 2.7. *The solution of (60) satisfies the following estimate:*

$$\mathcal{E}_0^{xy}(t) \leq \mathcal{E}_0^{xy}(0)\ \mathrm{e}^{3(\|\sigma_x\|_\infty + \|\sigma_y\|_\infty)t}. \tag{75}$$

Proof. Since (60) has the usual structure, one proceeds as usual and get

$$(\partial_t\vec{E}^\star,\vec{E}) + (\partial_t H^\star, H) = 0.$$

The first term can be rearranged as:

$$(\partial_t\vec{E}^\star,\vec{E})_{L^2} = (\partial_t\vec{E},\vec{E})_{L^2} + (\partial_t(\vec{E}^\star - \vec{E}),\vec{E})_{L^2}.$$

Using $\vec{E} = \vec{E}^\star - \vec{E} + 2\vec{E}$, we get

$$(\partial_t\vec{E}^\star,\vec{E})_{L^2} = (\partial_t\vec{E},\vec{E})_{L^2} + (\partial_t(\vec{E}^\star - \vec{E}),\vec{E}^\star - \vec{E})_{L^2} + 2(\partial_t(\vec{E}^\star - \vec{E}),\vec{E})_{L^2}.$$

That is to say, thanks to (61)-(i)

$$(\partial_t\vec{E}^\star,\vec{E})_{L^2} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\|\vec{E}^\star - \vec{E}\|_{L^2}^2 + \|\vec{E}\|_{L^2}^2) + 2(\Sigma\vec{E} - \Sigma^\star\vec{E}^\star,\vec{E})_{L^2}. \tag{76}$$

In order to rewrite the second term, using the definition (74) and integrating (61)-(ii) in time, we get (since everything vanish at time $t = 0$)

$$\partial_t H + (\sigma_x + \sigma_y)H + \sigma_x\sigma_y\widetilde{H} = \partial_t H^\star.$$

We thus have, since $H = \partial_t\widetilde{H}$

$$\begin{aligned}(\partial_t H^\star, H)_{L^2} &= (\partial_t H, H)_{L^2} + ((\sigma_x + \sigma_y)H, H)_{L^2} + (\sigma_x\sigma_y\widetilde{H},\partial_t\widetilde{H})_{L^2} \\ &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\|H\|_{L^2}^2 + (\sigma_x\sigma_y\widetilde{H},\widetilde{H})_{L^2}\right) + ((\sigma_x + \sigma_y)H, H)_{L^2}.\end{aligned} \tag{77}$$

Adding (76) and (77), we finally get:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_0^{xy} = 2(\Sigma\vec{E} - \Sigma^\star\vec{E}^\star,\vec{E})_{L^2} - ((\sigma_x + \sigma_y)H, H)_{L^2}. \tag{78}$$

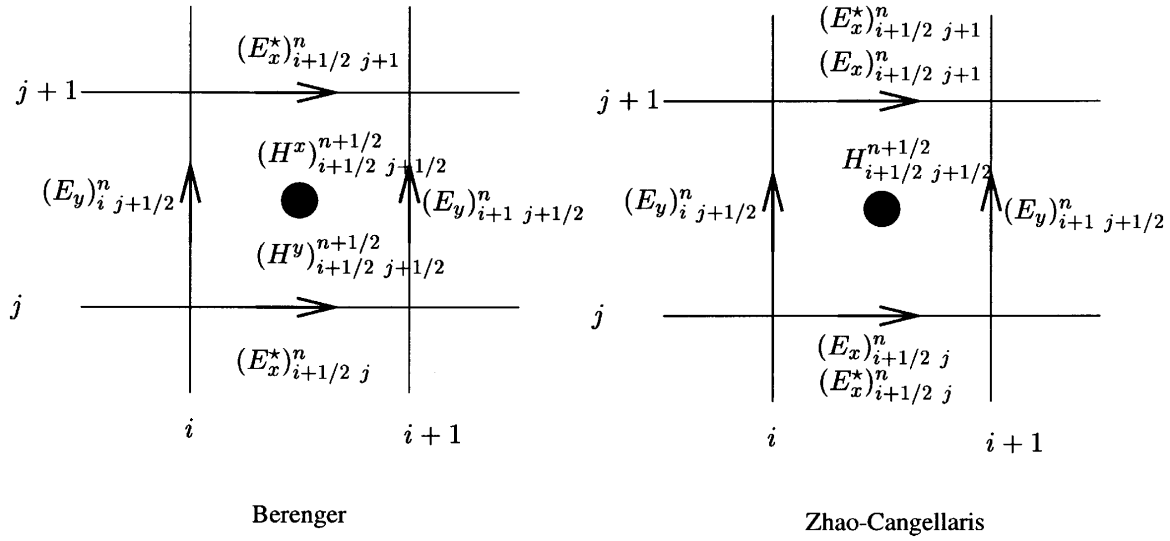

FIGURE 3. The discrete cell.

Writing $\Sigma\vec{E} - \Sigma^\star\vec{E}^\star = (\Sigma - \Sigma^\star)\vec{E} - \Sigma^\star(\vec{E}^\star - \vec{E})$, we obtain:

$$2(\Sigma\vec{E} - \Sigma^\star\vec{E}^\star, \vec{E})_{L^2} = 2((\Sigma - \Sigma^\star)\vec{E}, \vec{E})_{L^2} - 2(\Sigma^\star(\vec{E}^\star - \vec{E}), \vec{E})_{L^2}.$$

Since $|\Sigma - \Sigma^\star| \leq \|\sigma_x\|_\infty + \|\sigma_y\|_\infty$ and $|\Sigma^\star| \leq \max(\|\sigma_x\|_\infty, \|\sigma_y\|_\infty)$, we deduce that:

$$\begin{aligned} 2(\Sigma\vec{E} - \Sigma^\star\vec{E}^\star, \vec{E})_{L^2} &\leq 2(\|\sigma_x\|_\infty + \|\sigma_y\|_\infty)\|\vec{E}\|^2_{L^2} + \max(\|\sigma_x\|_\infty, \|\sigma_y\|_\infty)(\|\vec{E}^\star - \vec{E}\|^2_{L^2} + \|\vec{E}\|^2_{L^2}) \\ &\leq 3(\|\sigma_x\|_\infty + \|\sigma_y\|_\infty)(\|\vec{E}\|^2_{L^2} + \|\vec{E}^\star - \vec{E}\|^2_{L^2}). \end{aligned}$$

Therefore, (78) implies:

$$\mathcal{E}_0^{xy}(t) \leq \mathcal{E}_0^{xy}(0) + 3(\|\sigma_x\|_\infty + \|\sigma_y\|_\infty)\int_0^t \mathcal{E}_0^{xy}(s)\mathrm{d}s,$$

and, as usual, Gronwall's lemma permits to conclude. □

3. Energy estimates for the Yee's scheme

The energy estimates can be extended to the standard discrete centered Yee's scheme. Applied to Zhao-Cangellaris's formulation (55) it consists in approximating E_x and $E_x^\star$ on the horizontal edges, E_y on the vertical edges and H at the cells center, see Figure 3. We define the usual centered difference operators (k being either an integer, $k = n$, or a half integer, $k = n + 1/2$)

$$(D_{\Delta t}U)^k = \frac{U^{k+1/2} - U^{k-1/2}}{\Delta t}, \tag{79}$$

and in the same way, with obvious notations, $D_{\Delta x}$ and $D_{\Delta y}$. We set

$$\overline{U}^k = \frac{U^{k+1/2} + U^{k-1/2}}{2}, \tag{80}$$

and we also introduce the discrete centered operator approximating $\partial_t + \sigma$:

$$(D^{\sigma}_{\Delta t}U)^n_{\alpha\beta} = (D_{\Delta t}U)^n_{\alpha\beta} + \sigma_\alpha \frac{U^{n+1/2}_{\alpha\beta} + U^{n-1/2}_{\alpha\beta}}{2} = (D_{\Delta t}U)^n_{\alpha\beta} + \sigma_\alpha \overline{U}^n_{\alpha\beta} \tag{81}$$

where if α is an integer $\alpha = i$, σ_i is an approximation of $\sigma(x_i)$ ($x_i = i\Delta x$), and if α is a half integer, $\alpha = i+1/2$, $\sigma_{i+1/2} = \dfrac{\sigma_i + \sigma_{i+1}}{2}$. With these notations, the approximation of (55) with Yee's scheme is the following:

$$\begin{cases} (D_{\Delta t}E^{\star}_x)^{n+1/2}_{i+1/2\ j} - (D_{\Delta y}H)^{n+1/2}_{i+1/2\ j} = 0 & (a) \\ (D^{\sigma}_{\Delta t}E_y)^{n+1/2}_{i\ j+1/2} + (D_{\Delta x}H)^{n+1/2}_{i\ j+1/2} = 0 & (b) \\ (D^{\sigma}_{\Delta t}H)^{n}_{i+1/2\ j+1/2} + (D_{\Delta x}E_y)^{n}_{i+1/2\ j+1/2} - (D_{\Delta y}E_x)^{n}_{i+1/2\ j+1/2} = 0 & (c) \\ (D^{\sigma}_{\Delta t}E^{\star}_x)^{n+1/2}_{i+1/2\ j} = (D_{\Delta t}E_x)^{n+1/2}_{i+1/2\ j}. & (d) \end{cases} \tag{82}$$

We define the 1D-discrete scalar product:

$$(U,V) = \sum_{\alpha} U_\alpha V_\alpha, \quad \forall (U,V) \in (l^2(\alpha))^2,$$

where α is either integer or half-integer. When we want to distinguish on which grid we make the summation, we add an index $(U,V)_{l^2(i)}$ on the grid and $(U,V)_{l^2(i+1/2)}$ on the shifted grid. The 2D discrete scalar product will be denoted $((U,V))_h$ (or when needed with an index $l^2(\alpha)\times l^2(\beta)$).

Properties of the discrete differential operators

The discrete differential operators satisfy similar properties than their continuous version. In particular, discrete integrations by parts yields

$$\begin{aligned} (D_{\Delta x}U, V)_{l^2(\alpha)} &= -(U, D_{\Delta x}V)_{l^2(\alpha+1/2)}, \quad \forall U \in l^2(\alpha+1/2), \ \ \forall V \in l^2(\alpha) \\ (D_{\Delta y}U, V)_{l^2(\beta)} &= -(U, D_{\Delta y}V)_{l^2(\beta+1/2)}, \quad \forall U \in l^2(\beta+1/2), \ \ \forall V \in l^2(\beta) \end{aligned} \tag{83}$$

$$\left(\left((D_{\Delta t}U)^{n+1/2}, \frac{U^n + U^{n+1}}{2} \right) \right)_h = \frac{1}{2\Delta t}(\left\| U^{n+1} \right\|^2_h - \|U^n\|^2_h), \quad \forall U \in l^2(\alpha)\times l^2(\beta) \tag{84}$$

$$D^{\sigma}_{\Delta t} D_{\Delta t} = D_{\Delta t} D^{\sigma}_{\Delta t}. \tag{85}$$

Since σ is always independent on y, we have:

$$D^{\sigma}_{\Delta t} D_{\Delta y} = D_{\Delta y} D^{\sigma}_{\Delta t}. \tag{86}$$

If σ is constant:

$$D^{\sigma}_{\Delta t} D_{\Delta x} = D_{\Delta x} D^{\sigma}_{\Delta t}. \tag{87}$$

Remark 3.1. The scalar rotational is approximated with

$$\mathrm{rot}_h\vec{E} = D_{\Delta x}E_y - D_{\Delta y}E_x \tag{88}$$

and the vectorial rotational with:

$$\vec{\mathrm{rot}}_h H = (D_{\Delta y}H, -D_{\Delta x}H)^t. \tag{89}$$

From (83), we thus have

$$\left(\left(\mathrm{rot}_h\vec{E}, V\right)\right)_h = -\left((E_y, D_{\Delta x}V)\right)_h + \left((E_x, D_{\Delta y}V)\right)_h \equiv \left(\left(\vec{E}, \vec{\mathrm{rot}}_h V\right)\right)_h. \tag{90}$$

The scheme thus preserves the structure of the continuous system, with discrete versions of the rotational operators remaining adjoint of each other, which is of course essential to get discrete energy estimates. Note that the equation (82)-(c) can be written with these notations in the convenient form:

$$(D^\sigma_{\Delta t}H)^n + \mathrm{rot}_h\vec{E}^n = 0. \tag{91}$$

Equivalence between Yee's scheme applied to Zhao-Cangellaris's formulation and Yee's scheme applied to Bérenger's formulation

Using the same notations, we approximate equations (13) (where we replace E_x with $E^\star_x$) and with $H = H^x + H^y$, we get

$$\begin{cases} (D_{\Delta t}E^\star_x)^{n+1/2}_{i+1/2\ j} - (D_{\Delta y}H)^{n+1/2}_{i+1/2\ j} = 0 & (a)\\ (D^\sigma_{\Delta t}E_y)^{n+1/2}_{i\ j+1/2} + (D_{\Delta x}H)^{n+1/2}_{i\ j+1/2} = 0 & (b)\\ (D^\sigma_{\Delta t}H^x)^{n}_{i+1/2\ j+1/2} + (D_{\Delta x}E_y)^{n}_{i+1/2\ j+1/2} = 0 & (c)\\ (D_{\Delta t}H^y)^{n}_{i+1/2\ j+1/2} - (D_{\Delta y}E^\star_x)^{n}_{i+1/2\ j+1/2} = 0 & (d)\\ (H^x)^{n+1/2}_{i+1/2\ j+1/2} + (H^y)^{n+1/2}_{i+1/2\ j+1/2} = H^{n+1/2}_{i+1/2\ j+1/2}. & (e) \end{cases} \tag{92}$$

This numerical scheme is nothing but the finite difference scheme considered by Bérenger [3]. Equations (92)-(a) and (92)-(b) are the same than (82)-(a) and (82)-(b). If we apply $D_{\Delta t}$ to (92)-(c) and $D^\sigma_{\Delta t}$ to (92)-(d), and we add these two relations, we get (using (85) and (86))

$$(D_{\Delta t}D^\sigma_{\Delta t}H)^{n+1/2}_{i+1/2\ j+1/2} + (D_{\Delta t}D_{\Delta x}E_y)^{n+1/2}_{i+1/2\ j+1/2} - (D_{\Delta y}D^\sigma_{\Delta t}E^\star_x)^{n+1/2}_{i+1/2\ j+1/2} = 0.$$

This leads us to set

$$(D^\sigma_{\Delta t}E^\star_x)^{n+1/2}_{i+1/2\ j} = (D_{\Delta t}E_x)^{n+1/2}_{i+1/2\ j}$$

which is exactly (82)(d). Assuming that $(D^\sigma_{\Delta t}H + D_{\Delta x}E_y - D_{\Delta y}E_x)^0 = 0$, we thus get (82)(c). It is straightforward to show the reciprocal. Therefore Bérenger's and Zhao-Cangellaris's formulations are also equivalent in the discrete level, when using the Yee's scheme.

A discrete energy decay result

We set the discrete energy:

$$\left\{\begin{array}{rl} \mathcal{E}_1^{n+1/2} = & \dfrac{1}{2\Delta t}\Bigg(\Big\|(D_{\Delta t}E_x)^{n+1/2}\Big\|_h^2 + \Big\|(D_{\Delta t}E_y)^{n+1/2}\Big\|_h^2 + \left\|\sigma\dfrac{E_y^n+E_y^{n+1}}{2}\right\|_h^2 \\ & \\ & + \Big(\big((D^\sigma_{\Delta t}H)^{n+1},(D^\sigma_{\Delta t}H)^n\big)\Big)_h\Bigg). \end{array}\right. \tag{93}$$

Theorem 3.2. *We have the following identity:*

$$\mathcal{E}_1^{n+1/2} - \mathcal{E}_1^{n-1/2} + 2\sigma\left\|\frac{(D_{\Delta t}E_y)^{n+1/2}+(D_{\Delta t}E_y)^{n-1/2}}{2}\right\|_h^2 = 0, \tag{94}$$

which shows in particular that

$$\mathcal{E}_1^{n+1/2} \leq \mathcal{E}_1^{n-1/2}. \tag{95}$$

Proof. Recall that for the continuous problem, the energy decreasing was obtained applying the following $((\partial_t+\sigma)(a),\partial_t E_x)_0+((\partial_t+\sigma)(b),\partial_t E_y)_0+(\partial_t(c),\partial_t H+\sigma H)_0$. The corresponding discrete operation consists in:

(i) Applying $D^\sigma_{\Delta t}$ to (82)-(a). Equation (a) is written at time $n+1/2$, and the operator $D^\sigma_{\Delta t}$ will shift it at time n:

$$D^\sigma_{\Delta t}D_{\Delta t}E_x^\star - D^\sigma_{\Delta t}D_{\Delta y}H = 0.$$

Using (85), (86) and (82)-(d), this is equivalent to

$$(D^2_{\Delta t}E_x)^n - (D_{\Delta y}D^\sigma_{\Delta t}H)^n = 0 \tag{96}$$

that we multiply with the centered discrete version of $\partial_t E_x$, *i.e.* with $(\overline{D_{\Delta t}E_x})^n$, to get

$$\Big(\big((D^2_{\Delta t}E_x)^n,(\overline{D_{\Delta t}E_x})^n\big)\Big)_h - \Big(\big((D_{\Delta y}D^\sigma_{\Delta t}H)^n,(\overline{D_{\Delta t}E_x})^n\big)\Big)_h = 0. \tag{97}$$

(ii) Applying $D^\sigma_{\Delta t}$ to (82)-(b). We proceed as in (i), and this time we use (87) since σ is constant, thus we get

$$((D^\sigma_{\Delta t})^2E_y)^n + (D_{\Delta x}D^\sigma_{\Delta t}H)^n = 0. \tag{98}$$

We multiply (98) with $(\overline{D_{\Delta t}E_y})^n$ to get

$$\Big(\big(((D^\sigma_{\Delta t})^2E_y)^n,(\partial_{\Delta t}E_y)^n\big)\Big)_h - \Big(\big((D_{\Delta x}D^\sigma_{\Delta t}H)^n,(\overline{D_{\Delta t}E_y})^n\big)\Big)_h = 0. \tag{99}$$

(iii) The last equation (82)-(c) (or equivalently (91)) is not written as (a) and (b) at time $n+1/2$ but at time n, it is thus necessary to consider the mean-value of (c) at n and $n+1$:

$$(\overline{D^\sigma_{\Delta t}H})^{n+1/2} + \mathrm{rot}_h(\overline{\vec{E}})^{n+1/2} = 0.$$

We apply $D_{\Delta t}$ to this equation and multiply it by $(D^\sigma_{\Delta t}H)^n$:

$$\left((\overline{D_{\Delta t}D^\sigma_{\Delta t}H})^n,(D^\sigma_{\Delta t}H)^n\right)_h+\left(\left(\mathrm{rot}_h(\overline{(D_{\Delta t}\vec{E})})^n,(D^\sigma_{\Delta t}H)^n\right)\right)_h=0. \tag{100}$$

(iv) Adding (97), (99) and (100). Thanks to (90) we see that

$$\left(\left(\mathrm{rot}_h(\overline{D_{\Delta t}\vec{E}})^n,(D^\sigma_{\Delta t}H)^n\right)\right)_h=\left(\left((\overline{D_{\Delta t}\vec{E}})^n,\vec{\mathrm{rot}}_h(D^\sigma_{\Delta t}H)^n\right)\right)_h$$

thus, as in the continuous case, terms containing space derivatives vanish and it remains

$$\left\{\begin{array}{l}\mathcal{S}=0,\quad\text{with}\quad\mathcal{S}=\mathcal{S}^1+\mathcal{S}^2+\mathcal{S}^3\\ \mathcal{S}^1=\left((D^2_{\Delta t}E_x)^n,(\overline{D_{\Delta t}E_x})^n\right)_h\\ \mathcal{S}^2=\left(((D^\sigma_{\Delta t})^2E_y)^n,(\overline{D_{\Delta t}E_y})^n\right)_h\\ \mathcal{S}^3=\left((\overline{D_{\Delta t}D^\sigma_{\Delta t}H})^n,(D^\sigma_{\Delta t}H)^n\right)_h.\end{array}\right. \tag{101}$$

We now explicit the $\mathcal{S}^j$. From (84), it is straightforward to get

$$\mathcal{S}^1=\frac{1}{2\Delta t}\left(\left\|(D_{\Delta t}E_x)^{n+1/2}\right\|^2_h-\left\|(D_{\Delta t}E_x)^{n-1/2}\right\|^2_h\right). \tag{102}$$

There is also no difficulty to rewrite the third term as

$$\mathcal{S}^3=\frac{1}{2\Delta t}\left(\left((D^\sigma_{\Delta t}H)^{n+1},(D^\sigma_{\Delta t}H)^n\right)_h-\left((D^\sigma_{\Delta t}H)^n,(D^\sigma_{\Delta t}H)^{n-1}\right)_h\right). \tag{103}$$

Concerning the second term, we first develop:

$$\begin{aligned}((D^\sigma_{\Delta t})^2E_y)^n&=(D^\sigma_{\Delta t}D_{\Delta t}E_y)^n+\sigma(\overline{(D^\sigma_{\Delta t}E_y)})^n\\&=(D^2_{\Delta t}E_y)^n+2\sigma(\overline{D_{\Delta t}E_y})^n+\sigma^2\frac{\overline{E_y}^{n+1/2}+\overline{E_y}^{n-1/2}}{2}.\end{aligned}$$

We multiply this expression with $(\overline{D_{\Delta t}E_y})^n$, and rearrange the last term:

$$\sigma^2\left(\left(\frac{\overline{E_y}^{n+1/2}+\overline{E_y}^{n-1/2}}{2},(\overline{D_{\Delta t}E_y})^n\right)\right)_h=\frac{\sigma^2}{2\Delta t}\left(\left\|\overline{E_y}^{n+1/2}\right\|^2_h-\left\|\overline{E_y}^{n-1/2}\right\|^2_h\right)$$

to finally get:

$$\left\{\begin{array}{rl}\mathcal{S}^2=&\frac{1}{2\Delta t}\left(\left\|(D_{\Delta t}E_y)^{n+1/2}\right\|^2_h-\left\|(D_{\Delta t}E_y)^{n-1/2}\right\|^2_h\right.\\&\left.+\sigma^2\left\|\overline{E_y}^{n+1/2}\right\|^2_h-\sigma^2\left\|\overline{E_y}^{n-1/2}\right\|^2_h\right)+2\sigma\left\|(\overline{D_{\Delta t}E_y})^n\right\|^2_h.\end{array}\right. \tag{104}$$

Expressing that $\mathcal{S}=0$ leads to (94), and (95) follows immediately since $\sigma>0$. □

It remains to prove that $\mathcal{E}_1^{n+1/2}$ defines a positive energy.

Theorem 3.3. *The quantity $\mathcal{E}_1^{n+1/2}$ can be written as*

$$\mathcal{E}_1^{n+1/2} = \frac{1}{2\Delta t}\left(\left(\left((D_{\Delta t}\vec{E})^{n+1/2}, K_h D_{\Delta t}\vec{E})^{n+1/2}\right)\right)_h + \left\|\sigma\overline{E_y}^{n+1/2}\right\|_h^2 + \left\|\overline{(D^\sigma_{\Delta t}H)}^{n+1/2}\right\|_h^2\right) \tag{105}$$

with $K_h = I - \dfrac{\Delta t^2}{4} rot_h^\star rot_h$. Therefore under the CFL condition

$$\frac{\Delta t^2}{4}\left\|rot_h^\star rot_h\right\| < 1 \tag{106}$$

the quantity $\mathcal{E}_1^{n+1/2}$ defines an energy, the discrete problem (82) is thus well posed and the scheme is stable.

Proof. This is quite classical: see for instance [13]. It essentially relies on the relation $(a,b) = \dfrac{1}{4}\|a+b\|^2 - \dfrac{1}{4}\|a-b\|^2$, applied to the scalar product:

$$\left(\left((D^\sigma_{\Delta t}H)^{n+1}, (D^\sigma_{\Delta t}H)^n\right)\right)_h = \frac{1}{4}\left\|(D^\sigma_{\Delta t}H)^{n+1} + (D^\sigma_{\Delta t}H)^n\right\|_h^2 - \frac{1}{4}\left\|(D^\sigma_{\Delta t}H)^{n+1} - (D^\sigma_{\Delta t}H)^n\right\|_h^2.$$

□

Remark 3.4. It is well known (see for instance [13]) that the CFL condition (106) is equivalent to

$$\frac{\sqrt{2}\Delta t}{h} < 1. \tag{107}$$

Appendix A. Proof of lemma 1.5-(i)

• In the two particular cases where there exists real eigenvalues, the computation is straightforward. If $K_y = 0$ (and $K_x = \pm 1$), the roots are $\nu = 0$ with multiplicity 2 and $\nu = -1 \pm i/\varepsilon$, which correspond to $\lambda = 0$ (double) and $\lambda = -\sigma \pm i\,|k|$. If $K_x = 0$ (and $K_y = \pm 1$), the roots are $\nu = -1$ (double) and $\nu = \pm i/\varepsilon$, which correspond to $\lambda = -\sigma$ (double) and $\lambda = \pm i\,|k|$. In these two cases the real parts are 0 and $-\sigma$.

• If $K_x K_y \neq 0$, the roots are distincts and complex conjugates. We present the proof in the case $1 - 2K_y^2 > 0$, otherwise, it is sufficient to exchange the roles of ν_1 and ν_3. We set: $\nu_1(\vec{K}/\varepsilon) = a(\vec{K}/\varepsilon) + ib(\vec{K}/\varepsilon)$, $\nu_3 = \alpha(\vec{K}/\varepsilon) + i\beta(\vec{K}/\varepsilon)$, $\nu_2 = \overline{\nu}_1$, $\nu_4 = \overline{\nu}_3$.

Reduction to the solution of an equation of degree 3

The roots satisfy the relations:

$$\begin{cases} (i)\ \ a + \alpha = -1, & (ii)\ \ |\nu_1|^2 + |\nu_3|^2 + 4a\alpha = 1 + \dfrac{1}{\varepsilon^2} \\ (iii)\ \ a\,|\nu_3|^2 + \alpha\,|\nu_1|^2 = -\dfrac{K_y^2}{\varepsilon^2}, & (iv)\ \ |\nu_1|^2\,|\nu_3|^2 = \dfrac{K_y^2}{\varepsilon^2}. \end{cases} \tag{108}$$

If we substitute $\alpha = -1-a$ in (ii), we get $|\nu_1|^2+|\nu_3|^2-\frac{1}{\varepsilon^2} = 1+4a+4a^2 = (1+2a)^2$. This shows in particular that $|\nu_1|^2+|\nu_3|^2-\frac{1}{\varepsilon^2} \geq 0$ and thus from (ii) that $4a\alpha-1 \leq 0$, and from (i) we have $4a\alpha-1 = (1+2\alpha)(1+2a)$. In the following we choose a such that $1+2a \geq 0$ (and $1+2\alpha \leq 0$). We introduce the notations:

$$\begin{cases} u = \sqrt{|\nu_1|^2+|\nu_3|^2-1/\varepsilon^2} \equiv \sqrt{v} \equiv 2a+1 & (i) \\ S = |\nu_1|^2+|\nu_3|^2 = v+\dfrac{1}{\varepsilon^2} & (ii) \\ D = |\nu_1|^2-|\nu_3|^2 & (iii) \end{cases} \tag{109}$$

we have

$$|\nu_1|^2 = \frac{S+D}{2} \quad ; \quad |\nu_3|^2 = \frac{S-D}{2},$$

and equations (108)-(iii) and (iv) become

$$a\frac{S-D}{2}+\alpha\frac{S+D}{2} = -\frac{K_y^2}{\varepsilon^2}, \qquad S^2-D^2 = 4\frac{K_y^2}{\varepsilon^2}\cdot$$

The first relation permits to express D with respect to u:

$$D = \frac{2}{2a+1}\left(\frac{K_y^2}{\varepsilon^2}-\frac{S}{2}\right) = \frac{2K_y^2/\varepsilon^2-u^2-1/\varepsilon^2}{u}, \tag{110}$$

and substituting this expression in the second relation leads to solve an equation of degree 3 in v:

$$Q(v) = 0 \quad \text{where } Q(x) = -\varepsilon^4x^3-\varepsilon^2(2-\varepsilon^2)x^2-(1-2\varepsilon^2)x+(1-2K_y^2)^2. \tag{111}$$

A study of function Q shows that it has a unique positive real root which satisfies

$$0 \leq v \leq 1.$$

Therefore, one gets

$$a(\vec{K}/\varepsilon) \in [-1/2, 0] \quad \text{and} \qquad \alpha(\vec{K}/\varepsilon) \in [-1, -1/2],$$

which proves (30).

It remains to prove that the imaginary part of ν_1 is bounded, *i.e.* that there exists a constant $C > 0$ such that

$$|b(\vec{K}/\varepsilon)| \leq C,$$

which is equivalent to prove that $|\nu_1|$ is bounded (since the real part is bounded). From relation (109)-(ii) together with $v \in [0,1]$, we have

$$|\nu_1|^2+|\nu_3|^2 = v+\frac{1}{\varepsilon^2} \leq 1+\frac{1}{\varepsilon^2},$$

so that it is clear that $|\nu_1|$ and $|\nu_3|$ are bounded for ε large enough, for instance $\varepsilon^2 > 1/2$. In the case $\varepsilon^2 \leq 1/2$, we use an explicit expression for the solution v.

Determination of the roots of equation (111), when $\varepsilon^2 \leq 1/2$

In this paragraph, we want to give explicitly the expression of the unique positive root of Q, v. This leads to compute the roots of

$$\frac{Q(x)}{-\varepsilon^4} = x^3 + \frac{(2-\varepsilon^2)}{\varepsilon^2}x^2 + \frac{(1-2\varepsilon^2)}{\varepsilon^4}x - \frac{(1-2K_y^2)^2}{\varepsilon^4}.$$

We set $a_2 = \dfrac{(2-\varepsilon^2)}{\varepsilon^2}$, $a_1 = \dfrac{(1-2\varepsilon^2)}{\varepsilon^4}$, $a_0 = -\dfrac{(1-2K_y^2)^2}{\varepsilon^4}$ and we make the change of unknown $x = y - a_2/3$, which gives an equation in y in the form:

$$y^3 + 3py + 2q = 0, \quad \text{where} \quad p = \frac{1}{3}(a_1 - \frac{1}{3}a_2^2) \quad \text{and} \quad 2q = -\frac{1}{3}a_2a_1 + a_0 + \frac{2}{27}a_2^3.$$

The discriminant of this equation is $D = p^3 + q^2$ (Cardan's formulas). We know that if $D > 0$, there is one real and two complex conjugates roots, and if $D < 0$ there are three real roots. Here, we have:

$$\begin{aligned} D &= -\frac{4K_y^2(1-K_y^2)}{\varepsilon^8}\left(\frac{(1+\varepsilon^2)^3}{27\varepsilon^2} - K_y^2 + K_y^4\right) \\ &= -\frac{4K_y^2K_x^2}{\varepsilon^8}\left(\frac{(1+\varepsilon^2)^3}{27\varepsilon^2} - \frac{1}{4} + (\frac{1}{2} - K_y^2)^2\right) \\ &= -\frac{4K_y^2K_x^2}{\varepsilon^8}\left(\frac{(4+\varepsilon^2)(1-2\varepsilon^2)^2}{108\varepsilon^2} + (\frac{1}{2} - K_y^2)^2\right) \quad \leq 0. \end{aligned}$$

In conclusion, Q admits three real roots. We set $\Delta = -q + i\sqrt{|D|}$. The three real roots are

$$y_1 = 2\Re e(\Delta^{1/3}), \quad y_2 = -\Re e(\Delta^{1/3}) - \sqrt{3}\Im m(\Delta^{1/3}), \quad y_3 = -\Re e(\Delta^{1/3}) + \sqrt{3}\Im m(\Delta^{1/3}),$$

and $x_i = y_i - a_2/3$. After some computations we get

$$\Delta = \frac{r}{\varepsilon^4}\left(1 - 2s^2 + 2is\sqrt{1-s^2}\right) \equiv \frac{r}{\varepsilon^4}(\cos\theta + i\sin\theta),$$

with

$$r = \frac{(1+\varepsilon^2)^3}{27\varepsilon^2} \quad , \quad s = \frac{|K_y|\sqrt{1-K_y^2}}{\sqrt{r}} \in [0,1].$$

We know from the study of function that two roots among the x_i are negative and the one we are interested in is positive. We will show that *the unique positive root is* x_1. We notice that $\sin\theta \geq 0$ thus $\theta \in [0,\pi]$ and

$$\Delta^{1/3} = \frac{r^{1/3}}{\varepsilon^{4/3}}(\cos\frac{\theta}{3} + i\sin\frac{\theta}{3}) = \frac{(1+\varepsilon^2)}{3\varepsilon^2}(\cos\frac{\theta}{3} + i\sin\frac{\theta}{3}),$$

where $\theta/3 \in [0,\pi/3]$ therefore $\Re e(\Delta^{1/3}) \geq 0$ and $\Im m(\Delta^{1/3}) \geq 0$. We deduce that $y_2 < 0$ which implies that $x_2 < 0$. We now look at x_3:

$$x_3 \leq 0 \Longleftrightarrow \frac{(1+\varepsilon^2)}{3\varepsilon^2}(\sqrt{3}\sin\frac{\theta}{3} - \cos\frac{\theta}{3}) - \frac{(2-\varepsilon^2)}{3\varepsilon^2} \leq 0 \Longleftrightarrow \sqrt{3}\sin\frac{\theta}{3} - \cos\frac{\theta}{3} \leq \frac{2-\varepsilon^2}{1+\varepsilon^2}.$$

Using $\varepsilon^2 < 1/2$ we have $\dfrac{2-\varepsilon^2}{1+\varepsilon^2} > 1$. It is thus sufficient to prove that $\sqrt{3}\sin\dfrac{\theta}{3} - \cos\dfrac{\theta}{3} \leq 1$, which is true since $\theta/3 \in [0, \pi/3]$. In conclusion we have shown that $x_2 < 0$ and $x_3 < 0$ thus $x_1 \geq 0$.

Finally the solution $v \in]0, 1]$ has the following expression

$$v = \frac{1}{3\varepsilon^2}\left(2(1+\varepsilon^2)\cos\frac{\theta}{3} - (2-\varepsilon^2)\right) \text{ where } \quad \theta = \arccos\left(1 - 2K_y^2(1-K_y^2)\frac{27\varepsilon^2}{(1+\varepsilon^2)^3}\right). \tag{112}$$

It is then easy to prove that $v(\vec{K}/\varepsilon)$ is continuous in $\varepsilon \in [0, \dfrac{\sqrt{2}}{2}]$ and that

$$v(\vec{K}/\varepsilon) = (1-2K_y^2)^2 + v_R(\vec{K}/\varepsilon)\varepsilon^2,$$

where $v_R(\vec{K}/\varepsilon)$ is uniformly bounded. From this expression, we deduce that $\left|\nu_1(\vec{K}/\varepsilon)\right|$ is also bounded, while $\left|\nu_3(\vec{K}/\varepsilon)\right|$ tends to infinity when ε tends to 0.

References

[1] S. Abarbanel and D. Gottlieb, A mathematical analysis of the PML method. *J. Comput. Phys.* **134** (1997) 357–363.

[2] S. Abarbanel and D. Gottlieb, On the construction and analysis of absorbing layers in CEM. *Appl. Numer. Math.* **27** (1998) 331–340.

[3] J.P. Bérenger, A Perfectly Matched Layer for the Absorption of Electromagnetic Waves. *J. Comput. Phys.* **114** (1994) 185–200.

[4] F. Collino and P. Monk, Conditions et couches absorbantes pour les équations de Maxwell, in G. Cohen and P. Joly, *Aspects récents en méthodes numériques pour les équations de Maxwell*, Eds. École des Ondes, Chapter 4, INRIA, Rocquencourt (1998).

[5] J.W. Goodrich and T. Hagstrom, A comparison of two accurate boundary treatments for computational aeroacoustics. AIAA Paper-1585 (1997).

[6] J.S. Hesthaven, On the Analysis and Construction of Perfectly Matched Layers for the Linearized Euler Equations. *J. Comput. Phys.* **142** (1998) 129–147.

[7] F.Q. Hu, On absorbing boundary conditions for linearized euler equations by a perfectly matched layer. *J. Comput. Phys.* **129** (1996) 201–219.

[8] T. Kato, *Perturbation Theory for Linear Operators*. Springer (1995).

[9] H.-O. Kreiss and J. Lorenz, Initial-Boundary Value Problems and the Navier-Stokes Equations, in *Pure Appl. Math.* **136**, Academic Press, Boston, USA (1989).

[10] J. Métral and O. Vacus, Caractère bien posé du problème de Cauchy pour le système de Bérenger. *C.R. Acad. Sci.* I Math. **10** (1999) 847–852.

[11] P.G. Petropoulos, L. Zhao and A.C. Cangellaris, A reflectionless sponge layer absorbing boundary condition for the solution of Maxwell's equations with high-order staggered finite difference schemes. *J. Comput. Phys.* **139** (1998) 184–208.

[12] A.N. Rahmouni, *Des modèles PML bien posés pour divers problèmes hyperboliques*. Ph.D. thesis, Université Paris Nord-Paris XIII (2000).

[13] Allen Taflove, *Computational electrodynamics: the finite-difference time-domain method*. Artech House (1995).

[14] E. Turkel and A. Yefet, Absorbing PML boundary layers for wave-like equations. *Appl. Numer. Math.* **27** (1998) 533–557.

[15] L. Zhao and A.C. Cangellaris, A General Approach for the Development of Unsplit-Field Time-Domain Implementations of Perfectly Matched Layers for FDTD Grid Truncation. *IEEE Microwave and Guided Letters* **6** May 1996.

[16] R.W. Ziolkowski, Time-derivative lorentz material model-based absorbing boundary condition. *IEEE Trans. Antennas Propagation* **45** (1997) 1530–1535.