Problèmes faiblement bien posés : discrétisation et applications.

Sabrina Petit-Bergez

UNIVERSITÉ PARIS 13

N° attribué par la bibliothèque
|_|_|_|_|_|_|_|_|_|_|_|_|

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 13

Discipline : **Mathématiques appliquées**

présentée et soutenue publiquement

par

Sabrina PETIT-BERGEZ

le 5 décembre 2006

Titre :

Problèmes faiblement bien posés : discrétisation et applications

JURY

Mme	E. Bécache	Examinatrice
M.	A. Bendali	Rapporteur
M.	M. Crouzeix	Rapporteur
Mme	L. Halpern	Directrice de thèse
M.	O. Lafitte	Examinateur

Remerciements

Je tiens à remercier en premier lieu Laurence Halpern qui m'a encadrée durant toute cette thèse. Je la remercie pour son implication dans mon travail et pour ses nombreuses connaissances scientifiques qu'elle m'a fait partager. Le soutien moral qu'elle m'a accordé a aussi été fondamental pendant ces années de thèse.

Je remercie Abderrahmane Bendali et Michel Crouzeix d'avoir accepté d'être mes rapporteurs. Leurs commentaires avisés me furent très utiles. Je les remercie également pour l'intérêt qu'ils ont accordé à mon travail.

Je remercie également Eliane Bécache et Olivier Lafitte d'avoir accepté de faire partie du jury.

Je tiens à remercier particulièrement Jeffrey Rauch avec qui j'ai eu le plaisir de travailler. Je le remercie de m'avoir fait partager le spectre très large de ses connaissances mathématiques.

Je voudrais également remercier les membres du LAGA pour leur accueil. Le laboratoire a été pour moi un lieu de travail agréable et stimulant. Je remercie plus particulièrement Yolande Jimenez pour son efficacité et sa sympathie.

Je remercie tous les thésards du LAGA pour leur convivialité. Merci à Stéphanie, Laurentiu, Xavier, Christophe, Aurélien, Olivier, Véronique, Assia, Ruben et Sandra pour tous les joyeux déjeuners et pauses que nous avons pris ensemble. Je remercie plus particulièrement Christine Vespa et Ibrahima Cissé pour leur soutien dans la bonne humeur et leurs nombreux conseils.

Enfin, je souhaite remercier toute ma famille pour son soutien sans faille tout au long de mes études. Un grand merci à mes parents et à mon frère Sébastien qui se sont toujours intéressés à ma thèse. Pour finir, un grand merci à Jean-François pour m'avoir soutenue en toutes circonstances.

Table des matières

Introduction

Les problèmes de Cauchy faiblement bien posés sont des problèmes hyperboliques pour lesquels l'existence et l'unicité d'une solution est assurée mais tels que la solution est moins régulière que la donnée initiale. Plus précisément, il s'agit de problèmes tels qu'il y ait une perte de régularité, au sens des espaces de Sobolev, constante entre la donnée initiale et la solution du problème à un instant t quelconque. Ils ont été introduits par Gårding [18] et sont évoqués brièvement dans [15] et [27]. La grande majorité des équations fondamentales issues de la physique conduisent à un problème fortement bien posé, c'est à dire sans perte de régularité par rapport à la donnée initiale. C'est pourquoi les problèmes faiblement bien posés n'ont été qu'extrêmement peu étudiés jusqu'à présent.

Cependant, des problèmes faiblement bien posés sont apparus lors de l'étude des couches parfaitement adaptées de Bérenger [12] ou couches PML (Perfectly Matched Layers). Les couches PML ont été créées afin d'étudier la propagation d'ondes électromagnétiques en domaine non borné. C'est d'abord le mode transverse électrique des équations de Maxwell bidimensionnelles qui a été étudié. Le principe de cette méthode est d'entourer le domaine d'intérêt par une couche. Les équations vérifiées dans la couche sont obtenues en effectuant un découpage en deux composantes du champ magnétique, ces deux composantes n'ayant pas de signification physique, et en introduisant un coefficient d'absorption. L'intérêt de cette manipulation est qu'entre deux couches, les ondes sont parfaitement transmises pour toute fréquence et tout angle d'incidence. Il n'y a aucun problème de réflexion entre deux couches PML. Cependant, les couches PML présentent un inconvénient théorique majeur [1] : elles ne sont plus fortement bien posées, comme l'étaient les équations de Maxwell, mais uniquement faiblement bien posées.

Nous allons maintenant nous intéresser à la discrétisation des équations PML. Comme les équations de Maxwell sont discrétisées habituellement dans l'industrie par un schéma de Yee [41], les équations PML issues des équations Maxwell devront l'être naturellement aussi. L'application du schéma de Yee aux équations PML donne des résultats numériques très satisfaisants. Toutefois, il a été montré [1] que le schéma de Yee pour les équations PML n'est pas stable. La motivation de cette thèse est d'expliquer pourquoi un schéma qui n'est pas stable au sens classique peut tout de

même être convergent quand il est utilisé pour un problème faiblement bien posé.

La première partie de cette thèse est l'étude des schémas numériques pour les problèmes faiblement bien posés. En effet, le problème du schéma de Yee est que la définition d'un schéma stable n'est pas adaptée aux problèmes faiblement bien posés. Nous allons donc reprendre la théorie de la discrétisation par différences finies des problèmes fortement bien posés pour l'adapter aux problèmes faiblement bien posés.

La théorie des schémas numériques pour les problèmes fortement bien posés est assez ancienne et elle a été étudiée par de nombreux auteurs [37], [29] et reprise dans [19], [38]. Le résultat fondamental dans cette théorie est le théorème de Lax-Richtmyer qui donne une condition nécessaire et suffisante de convergence pour un schéma numérique. De plus, la convergence peut être précisée grâce au taux de convergence. Le taux de convergence se mesure sur l'erreur entre la solution exacte du problème de Cauchy et la solution discrète obtenue par un schéma. Cette erreur diminue lorsque les pas d'espace et de temps diminuent et cette diminution est, asymptotiquement, polynomiale. Le taux de convergence est le degré du polynôme en les pas d'espace et de temps qui majore l'erreur. Un autre résultat fondamental de la théorie de la discrétisation des problèmes fortement bien posés est que, pour une donnée initiale suffisamment régulière, le taux de convergence est égal à l'ordre de convergence du schéma, donné par l'erreur de troncature. De nombreux schémas ont pu être étudiés grâce à cette théorie, les plus classiques étant le schéma décentré qui est d'ordre 1, le schéma de Lax-Wendroff qui est d'ordre 2, le schéma de Crank-Nicolson qui est aussi d'ordre 2 mais est implicite et le schéma saute-mouton,d'ordre 2, qui est à deux pas en temps.

Le but de cette thèse est d'étendre les résultats connus dans le cas des problèmes fortement bien posés aux problèmes faiblement bien posés. Nous commencerons donc par donner des nouvelles définitions qui vont être adaptées aux problèmes faiblement bien posés. En effet, la perte de régularité par rapport à la donnée initiale qui apparaît dans le problème continu doit aussi apparaître dans le problème discret. Nous allons donc définir une nouvelle notion de stabilité que nous appellerons la stabilité faible et qui autorisera une perte de régularité de la solution discrète par rapport à la donnée initiale du schéma. Nous ferons de même pour les notions de convergence et de consistance.

Grâce à ces définitions qui sont bien adaptées aux problèmes faiblement bien posés, nous étendrons le théorème de Lax-Richtmyer. Le résultat démontré a une structure identique au cas fortement bien posé, à savoir : une condition suffisante de convergence est que le schéma soit stable et consistant et une condition nécessaire de convergence est que le schéma soit stable. Toutefois, dans le cas des problèmes faiblement bien posés, nous allons devoir relier entre elles toute les pertes de régularité intervenant dans les différentes définitions.

Nous nous intéresserons ensuite au cas particulier des schémas à coefficients constants. Nous donnerons alors une interprétation matricielle des différentes notions introduites. Nous prouverons en particulier que la condition nécessaire mais non suffisante de stabilité portant sur les valeurs propres de la matrice d'amplification du schéma est en fait une condition nécessaire et suffisante de stabilité faible. Puis, nous donnerons une minoration du taux de convergence simple à calculer qui aura pour conséquence fondamentale de montrer que, même pour un problème faiblement bien posé, pour une donnée initiale suffisamment régulière, le taux de convergence est égal à l'ordre de convergence. Nous verrons que, toutefois, la donnée initiale doit être plus régulière que dans le cas d'un problème fortement bien posé.

Nous nous intéresserons alors au calcul d'un taux de convergence optimal, c'est-à-dire, tel que le taux de convergence observé numériquement ne soit pas seulement minoré par le taux de convergence théorique mais en soit proche. Cette étude est beaucoup plus difficile que pour les problèmes fortement bien posés. La difficulté provient du fait que la perte de régularité est constante. En effet, si la perte de régularité à l'instant t est de q_1, la perte de régularité à l'instant $2t$ est aussi de q_1, donc le raisonnement qui conduit à manipuler la solution à l'instant t pour ensuite passer à l'instant $2t$ peut conduire à une perte de régularité $2q_1$. Les démonstrations utilisant ce type de raisonnement vont donc donner des résultats non optimaux. De plus, les équations scalaires ne peuvent être faiblement bien posées, nous n'étudierons donc dans cette thèse que des problèmes matriciels ce qui va poser des problèmes de non commutativité. Nous avons résolu ces problèmes dans le cas de problèmes de Cauchy étudiés sur la droite réelle. Notre étude est basée sur la théorie des perturbations et les développements en série de Puiseux [24] et elle concerne une classe particulière de schémas que nous avons définie. Cette étude conduit au calcul d'un taux de convergence optimal. La majorité des schémas classiques est dans cette classe, à condition de bien définir la discrétisation du terme d'ordre 0. Nous avons étudié plus particulièrement ces schémas d'un point de vue théorique et numérique. Le seul schéma qui nécessite un traitement particulier et plus poussé est le schéma décentré.

Cette première partie conduit donc à une théorie pour la discrétisation des problèmes faiblement bien posés qui est validée numériquement en considérant des schémas classiques.

La seconde partie de cette thèse est l'étude de la stabilité des PML. En effet, les couches PML, qui étaient initialement destinées aux équations de Maxwell ont été étendues à d'autres équations comme les équations d'Euler par exemple [21]. De plus, de nombreuses études numériques ont été effectuées et ces études ont relevé deux types d'instabilité. Le premier type est une instabilité à temps long, qui a été observée dans [4] et le second, observé dans [9], est une instabilité qui s'installe plus rapidement dans la couche PML. L'observation de ces instabilités se traduit par une

croissance anormale en temps de la solution. Le premier type d'instabilité correspond à une croissance polynomiale en temps. En effet, les problèmes PML ne sont que faiblement bien posés donc, comme nous l'avons dit précédemment, il y a une perte de régularité par rapport à la donnée initiale. Toutefois cette perte de régularité est toujours associée à une croissance polynomiale en temps, son degré étant égal à la perte de régularité. Or, cela n'est pas le cas pour les problèmes fortement bien posés, d'où l'observation de cette croissance anormale dans la couche. Le second type d'instabilité, qui apparaît plus rapidement, est dû à une croissance exponentielle en temps. En effet, alors que les équations considérées initialement étaient homogènes, les équations PML ne le sont plus car l'introduction de l'absorption donne un terme d'ordre 0 non nul. Or, pour un problème avec terme d'ordre 0, qu'il soit faiblement ou fortement bien posé, il peut y avoir une croissance exponentielle en temps. Cette croissance qui va apparaître pour certains problèmes PML n'est pas en accord avec les équations étudiées initialement.

De nombreuses solutions ont été proposées pour remédier à ce type d'instabilités. Les premiers résultats [2], [20] sont une modification des équations PML qui rend le problème fortement bien posé. L'inconvénient de ces méthodes est la perte du caractère parfaitement adapté des couches. D'autres méthodes ont alors été proposées [3], [34] mais elles ne prennent pas en compte le second type de stabilité. Enfin, plus récemment, les deux types de stabilité ont été traités dans [22] et [11]. Notre but dans cette thèse est d'étudier les équations proposées initialement par Bérenger et d'expliquer pourquoi, malgré tous les problèmes théoriques, les résultats numériques demeurent satisfaisants. Nous ne nous intéresserons donc pas aux nouvelles formulations PML qui sont fortement bien posées et stables.

Nous commencerons par donner des estimations d'énergie pour les équations PML de Maxwell. Des estimations avaient déjà été obtenues dans [32]. Nous allons reprendre et étendre ces résultats. Nous donnerons, dans le chapitre 7, deux méthodes de démonstration rigoureuses pour ce type d'estimation. L'une est basée sur une approximation par semi-discrétisation des équations, l'autre sur l'étude du symbole. Elle permet de traiter le cas d'une absorption dans les deux directions et a l'avantage de se généraliser simplement au cas des équations de Maxwell tridimensionnelles. De plus, une version discrète de cette méthode va s'appliquer au schéma de Yee. Une conséquence des estimations d'énergie obtenues pour le schéma de Yee est la preuve de sa stabilité faible. Grâce aux résultats de la première partie de cette thèse nous serons donc en mesure de justifier les observations numériques de Bérenger, à savoir que, même si le schéma de Yee est instable au sens classique, il est convergent. De plus, nous minorerons le taux de convergence du schéma.

Dans le dernier chapitre de cette thèse, nous nous intéresserons à la stabilité des PML au sens défini par [9], c'est-à-dire savoir si la croissance en temps de la solution est polynomiale ou exponentielle. Une condition nécessaire de stabilité a été donnée par Hu dans [22] puis démontrée rigoureusement par Bécache *et al* dans [9]. Cette

condition porte sur la direction de la vitesse de phase et de la vitesse de groupe. Nous utilisons l'approximation WKB de l'optique géométrique afin d'obtenir une condition nécessaire de stabilité plus générale. En effet, cette condition s'appliquera aux équations de Maxwell tridimensionnelles qui étaient exclues auparavant et sera valable pour un coefficient d'absorption dans les deux directions et qui pourra être variable. Nous appliquerons ce résultat aux équations de Maxwell et d'Euler.

La seconde partie de cette thèse n'est donc pas uniquement une application de la première partie qui concernait les schémas numériques. Elle contient des estimations d'énergie plus précises que celles nécessaires pour le calcul de la perte de régularité des problèmes continus et discrets. En effet, nous montrons des estimations sans perte de régularité mais uniquement pour une certaine norme. De plus, nous nous intéressons aussi au comportement asymptotique en temps des équations PML.

Première partie

Schémas pour les problèmes faiblement bien posés

Chapitre 1

Le problème de Cauchy

Dans ce chapitre, nous allons donner des résultats généraux sur les problèmes fortement et faiblement bien posés. En utilisant la théorie des perturbations, nous établirons une nouvelle caractérisation des problèmes faiblement bien posés dans le cas de la dimension 1. Nous étudierons alors quelques exemples, en particulier, nous étudierons le cas des équations PML sans terme d'absorption.

Lorsque cela n'est pas précisé, la norme par défaut $\|.\|$ sera la norme euclidienne $\|.\|_2$.

1.1 Caractérisation des problèmes faiblement bien posés

Nous considérons le problème de Cauchy général, sous la forme :

$$\begin{cases} \partial_t U = P(t,x,\partial_x)U, \\ U(0,.) = U^0, \end{cases} \tag{1.1}$$

avec $t > 0$, $x \in \mathbb{R}^d$, $U \in \mathbb{R}^N$.

Définition 1.1 *Le problème de Cauchy est faiblement bien posé s'il existe $q_1 > 0$, $K > 0$, $\alpha \in \mathbb{R}$ tels que pour toute donnée initiale $U^0 \in H^q(\mathbb{R}^d)$ avec $q \geq q_1$, le problème a une unique solution $U \in \mathcal{C}^0(\mathbb{R}^+, H^{q-q_1}(\mathbb{R}^d))$, vérifiant, $\forall t \geq 0$:*

$$\|U(t,.)\|_{H^{q-q_1}(\mathbb{R}^d)} \leq Ke^{\alpha t}\|U^0\|_{H^q(\mathbb{R}^d)}.$$

Si q_1 est le plus petit entier vérifiant cette propriété, on dit alors que le problème de Cauchy est faiblement bien posé de défaut q_1.
Dans le cas où $q_1 = 0$ le problème est dit fortement bien posé.

Un problème faiblement bien posé correspond donc au cas où l'on observe une perte de régularité par rapport à la donnée initiale mais où la norme L^2 de la solution est quand même contrôlée par une certaine norme de la donnée initiale. Cette notion à été introduite par Gårding dans [18].

Les problèmes fortement bien posés à coefficients constants ont été beaucoup étudiés notament dans [27] et [26]. Nous présentons ici les résultats principaux de cette étude.

On considère ici que $P(\partial_x) = \sum_{j=1}^{d} A_j \partial_{x_j} + B$ où $A_j, B \in \mathcal{M}_N(\mathbb{R})$.

1.1.1 Quelques résultats importants sur les problèmes fortement bien posés

Le principe de cette théorie est de se ramener, grâce à un passage à la transformée de Fourier, à une étude matricielle du symbole qui est défini par :

$$P(i\xi) = \sum_{j=1}^{d} i\xi_j A_j + B.$$

En effet, en variables de Fourier, le problème de Cauchy devient :

$$\begin{cases} \partial_t \widehat{U}(t,\xi) = P(i\xi)\widehat{U}(t,\xi), \\ \widehat{U}(0,\xi) = \widehat{U^0}. \end{cases}$$

La transformée de Fourier de la solution est donc :

$$\widehat{U}(t,\xi) = \exp(tP(i\xi))\widehat{U^0}(\xi).$$

Nous avons alors la caractérisation suivante des problèmes fortement bien posés :

Théorème 1.1 *Le problème (1.1) est fortement bien posé si et seulement si :*

$$\exists K > 0,\ \exists \alpha \in \mathbb{R},\ \forall t > 0,\ \forall \xi \in \mathbb{R}^d,\ \| \exp(tP(i\xi)) \| \leq K e^{\alpha t}$$

Dans le cas des problèmes fortement bien posés, l'étude du symbole principal : $P_0(i\xi) = \sum_{j=1}^{s} i\xi_j A_j$ est primordiale. En effet, nous avons le résultat suivant qui est une conséquence du théorème matriciel de Kreiss [27] :

Théorème 1.2 *Le problème de Cauchy (1.1) est fortement bien posé si et seulement si le problème $\partial_t U = P_0(\partial_x)U$ l'est aussi.*

Dans le cas des problèmes fortement bien posés, nous pouvons donc toujours considérer des problèmes homogènes, c'est à dire, où $B = 0$. Nous avons alors la caractérisation suivante :

Théorème 1.3 *Le problème (1.1) est fortement bien posé si et seulement si les deux conditions suivantes sont vérifiées :*

1. *Pour tout $\xi \in \mathbb{R}^d$, $\|\xi\| = 1$, les valeurs propres de $P_0(i\xi)$ sont imaginaires pures.*
2. *Il existe une constante C, telle que pour tout $\xi \in \mathbb{R}^d$, $\|\xi\| = 1$, il existe une matrice inversible $S(\xi)$ vérifiant :*

$$\|S(\xi)\| + \|S^{-1}(\xi)\| \leq C,$$

et telle que la matrice $S(\xi)^{-1}P_0(i\xi)S(\xi)$ est diagonale.

Les cas particuliers suivants sont fortement bien posés :

Proposition 1.1 1. *Si $\forall j \in \{1, \ldots, d\}$, $A_j = A_j^*$, le problème de Cauchy est dit symétrique hyperbolique.*
Un problème symétrique hyperbolique est fortement bien posé.

2. *Si $\forall \xi \in \mathbb{R}^d \setminus \{0\}$, les valeurs propres de $P_0(i\xi)$ sont imaginaires pures et distinctes, le problème de Cauchy est dit strictement hyperbolique.*
Un problème strictement hyperbolique est fortement bien posé.

1.1.2 Etude des problèmes faiblement bien posés

Contrairement au cas fortement bien posé, l'étude du caractère faiblement bien posé ne peut pas se faire sur le symbole principal. En effet, il existe toujours une perturbation d'ordre 0 qui rend le problème mal posé c'est à dire tel que la norme du symbole $\|\exp(tP(i\xi))\|$ croît plus vite que tout polynôme en $\|\xi\|$ (voir [27]).

Toutefois, la caractérisation matricielle sur le symbole reste valable :

Proposition 1.2 *Le problème (1.1) est faiblement bien posé de défaut q_1 si et seulement si :*

$$\exists K > 0, \exists \alpha \in \mathbb{R}, \forall t > 0, \forall \xi \in \mathbb{R}^d, \|\exp(tP(i\xi))\| \leq Ke^{\alpha t}(1 + \|\xi\|)^{q_1}.$$

De plus, on a la caractérisation suivante qui donne une borne supérieure sur le défaut mais qui ne permet pas de le calculer :

Proposition 1.3 *Le problème (1.1) est faiblement bien posé si et seulement si il existe $\alpha > 0$ tel que pour tout ξ, les valeurs propres $\lambda(\xi)$ de $P(i\xi)$ vérifient :*

$$Re\,(\lambda(\xi)) \leq \alpha.$$

De plus, le défaut est majoré par $N - 1$.

PREUVE : Nous reprenons ici la preuve de Kreiss [26].

- Supposons le problème faiblement bien posé. Soit $\lambda(\xi)$ une valeur propre de $P(i\xi)$. Soit $X(\xi)$ un vecteur propre qui lui est associé. Alors $\widehat{U}(t,\xi) = e^{\lambda(\xi)t}X(\xi)$ est la solution du problème dont la transformée de Fourier de la condition initiale est $X(\xi)$. Puisque le problème est faiblement bien posé : $\|\widehat{U}(t,\xi)\| \leq Ke^{\alpha t}(1+\|\xi\|)^{q_1}\|X(\xi)\|$. Ainsi, $e^{Re\,(\lambda(\xi))t} \leq Ke^{\alpha t}(1+\|\xi\|)^{q_1}$, d'où la conclusion en faisant tendre t vers $+\infty$.
- Supposons la partie réelle des valeurs propres du symbole majorée. Nous effectuons la décomposition de Schur de $P(i\xi)$ [28] :

$$S^*(\xi)P(i\xi)S(\xi) = \begin{pmatrix} \lambda_1(\xi) & a_{12}(\xi) & \dots & a_{1,N}(\xi) \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{N-1,N}(\xi) \\ 0 & \dots & 0 & \lambda_N(\xi) \end{pmatrix} = T(\xi) = D(\xi) + B(\xi),$$

avec $S(\xi)$ unitaire, $D(\xi)$ la partie diagonale de $T(\xi)$ et $B(\xi)$ la partie triangulaire supérieure de $T(\xi)$. Nous supposons, quitte à effectuer une permutation, que les valeurs propres sont numérotées de telle sorte que, pour ξ fixé :

$$Re\,(\lambda_1(\xi)) \geq \dots \geq Re\,(\lambda_N(\xi)).$$

Soit $\widehat{V}(t,\xi) = \exp(-tD(\xi))S^*(\xi)\widehat{U}(t,\xi)$. Alors, $\widehat{V}$ vérifie l'équation : $\partial_t\widehat{V}(t,\xi) = \exp(-tD(\xi))B(\xi)\exp(tD(\xi))\widehat{V}(t,\xi)$, avec :

$$(\exp(-tD(\xi))B(\xi)\exp(tD(\xi)))_{m,n} = \begin{cases} a_{m,n}e^{(\lambda_n(\xi)-\lambda_m(\xi))t} & \text{si } m<n \\ 0 & \text{sinon.} \end{cases}$$

De plus, nous avons $\|P(i\xi)\| = \|T(\xi)\| = O(\xi)$ donc, pour $m<n$, $a_{m,n}(\xi) = O(\xi)$. Ainsi, d'après l'ordre choisi sur les valeurs propres, $a_{m,n}(\xi)e^{(\lambda_n(\xi)-\lambda_m(\xi))t} = O(\xi)$. Nous obtenons alors, en résolvant le système triangulaire précédent :

$$\|\widehat{V}(t,\xi)\| \leq K(1+\|\xi\|)^{N-1}(1+t)^{N-1}\|\widehat{V}(0,\xi)\|.$$

Or $\|\widehat{U}(t,\xi)\| = \|\exp(tD(\xi))\widehat{V}(t,\xi)\| \leq e^{\alpha t}\|\widehat{V}(t,\xi)\|$, donc :

$$\|\widehat{U}(t,\xi)\| \leq Ke^{\alpha t}(1+\|\xi\|)^{N-1}(1+t)^{N-1}\|\widehat{U^0}(\xi)\|.$$

Ce qui prouve que le problème est faiblement bien posé de défaut majoré par $N-1$.

□

Corollaire 1.1 *Une condition nécessaire pour qu'un problème soit faiblement bien posé est que les valeurs propres des matrices A_j, $j \in \{1,\dots,d\}$ soient réelles.*

PREUVE : Supposons le problème faiblement bien posé. Soit $j \in \{1, \ldots, d\}$, montrons, que les valeurs propres de A_j sont réelles. Soit μ_0 une valeur propre de A_j. Par continuité des valeurs propres, il existe $\mu(\tau)$ valeur propre de $A - i\tau B$ telle que $\lim_{\tau \to 0} \mu(\tau) = \mu_0$. Considérons le vecteur ξ dont la $j^{\text{ème}}$ composante est $\frac{1}{\tau}$ et les autres composantes sont nulles. Alors $P(i\xi) = \frac{i}{\tau}A_j + B = \frac{i}{\tau}(A - i\tau B)$ et $\lambda(\xi) = \frac{i}{\tau}\mu(\tau)$ est une valeur propre de $P(i\xi)$. Donc, comme le problème est faiblement bien posé, $Re\,(\frac{i}{\tau}\mu(\tau)) \leq \alpha$. Donc $-\frac{1}{\tau} Im\,\mu(\tau) \leq \alpha$. En faisant tendre τ vers zéro par valeurs positives puis par valeurs négatives, nous obtenons $Im\,(\mu_0) = 0$ ce qui est bien le résultat voulu. □

Remarque 1.1 *Cependant, cette condition n'est pas suffisante. En effet, en dimension 1, si* $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ *et* $B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, *les valeurs propres de* $i\xi A + B$ *sont* $i\xi \pm \sqrt{i\xi}$ *qui ne sont pas de partie réelle majorée donc le problème n'est pas faiblement bien posé.*

Les résultats précédents sont des résultats classiques exposé dans [27].

Le corollaire suivant, est dû à Rauch et Taylor [40]. Il permet de se ramener à un problème sans perte de régularité.

Corollaire 1.2 *Supposons le problème faiblement bien posé. Alors, pour tout* $\xi \in \mathbb{R}^d$, *il existe* $M(\xi) \in \mathcal{M}_N(\mathbb{C})$ *mesurable et inversible telle que si* $\hat{W}(t,\xi) = M(\xi)\hat{U}(t,\xi)$, *on a :*

$$\exists K > 0,\ \forall s \in \mathbb{R},\ \forall t > 0\ \|W(t,.)\|_{H^s} \leq K(1+t)^{N-1}\|W^0\|_{H^s}.$$

PREUVE : Nous utilisons les notations de la preuve de la proposition précédente. Nous considérons $R(\xi) = \text{diag}(1, 1 + \|\xi\|, \ldots, (1 + \|\xi\|)^{N-1})$ et nous posons $M(\xi) = R(\xi)\exp(-tD(\xi))S^*(\xi)$. Alors, l'équation vérifiée par $\widehat{W}$ est :

$$\partial_t \widehat{W}(t,\xi) = R(\xi)\exp(-tD(\xi))B(\xi)\exp(tD(\xi))R^{-1}(\xi)\widehat{W}(t,\xi),$$

avec :

$$(R(\xi)\exp(-tD(\xi))B(\xi)\exp(tD(\xi))R^{-1}(\xi))_{m,n}$$
$$= \begin{cases} a_{m,n}e^{(\lambda_n(\xi)-\lambda_m(\xi))t}(1+\|\xi\|)^{m-n} & \text{si } m < n \\ 0 & \text{sinon.} \end{cases}$$

Or, comme la matrice est strictement triangulaire supérieure, $m - n \leq -1$, donc $a_{m,n}(\xi)e^{(\lambda_n(\xi)-\lambda_m(\xi))t}(1+\|\xi\|)^{m-n} = O(1)$. En résolvant l'équation, nous trouvons :

$$\|\widehat{W}(t,\xi)\| \leq K(1+t)^{N-1}\|\widehat{W}(0,\xi)\|,$$

ce qui donne la conclusion. □

Ce corollaire signifie que, pour une certaine norme, les problèmes faiblement bien posés sont fortement bien posés.

1.2 Rappels sur le développement en séries de Puiseux

Nous allons donner, dans ce paragraphe, les résultats sur le développement des fonctions en séries de Puiseux qui seront utiles pour le paragraphe suivant mais aussi pour le calcul de taux de convergence optimal pour une classe particulière de schémas.

Tous les résultats cités sont extraits du livre de Kato [24].

Nous considérons ici une matrice carrée T, que nous allons perturber linéairement. Nous notons T' la perturbation. La matrice étudiée est alors :

$$T(x) = T + xT',$$

avec $x \in D_0 \subset \mathbb{C}$, où D_0 est un ouvert contenant 0.

Remarque 1.2 *Les résultats présentés par Kato sont, en fait, valables pour des perturbations analytiques. Nous ne les appliquerons ici que pour des perturbations linéaires.*

Nous nous intéresserons ici à l'étude des valeurs propres et de la décomposition de Dunford de la matrice $T(x)$ en fonction de la matrice T.

Proposition 1.4

- *Il existe $D_{ex} \subset D_0$, et $s \in \mathbb{N}^*$ tels que pour tout $x \in D_0 \setminus D_{ex}$, le nombre de valeurs propres distinctes de $T(x)$ est égal à s.*
- *D_{ex} est fini.*

Définition 1.2 *Les éléments de D_{ex} sont appelés des points exceptionnels.*

Dans le cas des points qui ne sont pas exceptionnels, l'étude des valeurs propres ne nécessite pas de développement en série de Puiseux.

Théorème 1.4 *Si D est un sous-ensemble simplement connexe de D_0 ne contenant pas de point exceptionnel, alors les valeurs propres distinctes de $T(x)$, $x \in D$ peuvent se mettre sous la forme de fonctions $\lambda_1(x), \ldots \lambda_s(x)$ holomorphes sur D.*

Nous allons maintenant étudier le cas d'un domaine contenant un point exceptionnel. Pour simplifier les notations, nous supposerons que $x = 0$ est un point exceptionnel, $x \in D_{ex}$, et nous considérons un disque D centré en 0 tel que $D \setminus \{0\}$ ne contienne pas de point exceptionnel. Dans ce cas, les séries de Puiseux vont remplacer les séries entières.

Théorème 1.5 *Nous pouvons numéroter les s valeurs propres distinctes de $T(x)$, pour $x \in D \setminus \{0\}$ de la manière suivante : $(\lambda_k^m(x))_{\{0 \leq k \leq p_m - 1, 1 \leq m \leq m_0\}}$, avec $m_0 \in \mathbb{N}^*$ et $p_m \in \mathbb{N}^*$ et tel que pour $j \in \mathbb{N}^*$ et $m \in \{1, \ldots, m_0\}$, il existe des coefficients $\alpha_{j,m} \in \mathbb{C}$, tels que $\lambda_k^m(x)$ admette le développement en série de Puiseux suivant :*

$$\lambda_k^m(x) = \lambda_m + \sum_{j=1}^{+\infty} \alpha_{j,m} e^{2ikj\pi/p_m} x^{j/p_m},$$

où λ_m est une valeur propre de T.

Remarque 1.3 *La valeur de p_m est toujours inférieure à la multiplicité de λ_m comme valeur propre de T. De plus, quitte à remplacer p_m par $ppcm(p_m,\, 1 \leq m \leq m_0)$ et à introduire des termes nuls dans le développement en série de Puiseux, nous pouvons considérer que $p_m = p$ est indépendant de l'indice m.*

Ce résultat est une application de la théorie des fonctions algébriques, exposée dans [25], dans le cas particulier où le polynôme étudié est un polynôme caractéristique.

Nous allons maintenant étudier le comportement de la décomposition de Dunford, c'est à dire la décomposition en somme d'une matrice diagonalisable et d'une matrice nilpotente qui commutent. Nous nous intéressons à cette décomposition car elle permet de calculer aisément les exponentielles de matrices. En effet, si nous considérons la décomposition de Dunford de $T(x)$: $T(x) = N(x) + D(x)$ où $N(x)$ est nilpotente d'indice k, $D(x)$ est diagonalisable et $N(x)$ et $D(x)$ commutent, alors :

$$\exp(T(x)) = \exp(D(x)) \exp(N(x)) = \exp(D(x)) \sum_{j=0}^{k-1} \frac{1}{j!} N(x)^j.$$

Nous allons commencer par exprimer les parties diagonalisables et nilpotentes de la décomposition de Dunford en fonction des valeurs propres et des projections sur les sous espaces-caractéristiques.

Posons $I = \{(m,k),\, 1 \leq m \leq m_0,\, 0 \leq k \leq p_m - 1\}$. Pour $(m,k) \in I$, nous notons $M_k^m(x)$, le sous espace-caractéristique de $T(x)$ pour la valeur propre $\lambda_k^m(x)$. Nous avons alors, pour $x \notin D_{ex}$:

$$\mathbb{C}^N = \bigoplus_{(m,k) \in I} M_k^m(x).$$

Si $P_k^m(x)$ désigne la projection sur $M_k^m(x)$ correspondant à cette décomposition, alors, nous avons :

$$N(x) = \sum_{(m,k) \in I} (T(x) - \lambda_k^m(x)) P_k^m(x) \text{ et } D(x) = \sum_{(m,k) \in I} \lambda_k^m(x) P_k^m(x).$$

Pour développer les parties nilpotentes et diagonalisables, il suffit donc de développer les projections sur les sous-espaces caractéristiques. Pour cela, nous pouvons utiliser le théorème suivant [24] :

Théorème 1.6 *Les notations sont les mêmes que dans le théorème 1.5.*
Il existe un voisinage de 0*, un entier* $r \in \mathbb{N}$ *et une suite de matrices* $A_{j,m} \in \mathcal{M}_N(\mathbb{C})$ *tels que, pour tout* $(m,k) \in I$*, on ait :*

$$P_k^m(x) = \sum_{j=-r}^{\infty} A_{j,m} e^{2i\pi jk/p_m} x^{j/p_m}.$$

Remarque 1.4 *Contrairement aux valeurs propres, les projections sur les sous-espaces caractéristiques ne sont pas, en général, analytiques mais seulement méromorphes.*

En combinant les deux derniers théorèmes, nous voyons que les deux parties de la décomposition de Dunford sont développables en série de Puiseux au voisinage d'un point exceptionnel.

Remarque 1.5 *Ces résultats ne se généralisent pas, a priori, à des perturbations dans* $\mathbb{C}^n$ *avec* $n \geq 2$*. La théorie des perturbations pour des opérateurs dépendant de plusieurs variables complexes a été traitée dans [8]. Une différence notable entre la dimension* $n > 1$ *et la dimension 1 est que l'ensemble des points exceptionnels est une variété analytique de dimension strictement inférieure à* n*. Nous ne pouvons donc plus construire de voisinage de point exceptionnel ne contenant pas d'autres points exceptionnels.*

1.3 Caractérisation dans le cas unidimensionnel

Nous étudions dans cette partie le cas où $d = 1$ et alors $P(i\xi) = i\xi A + B$. Nous donnons, dans ce cas particulier une nouvelle caractérisation des problèmes faiblement bien posés portant sur les valeurs propres, qui sera utile pour étudier la stabilité des schémas numériques. Ici, la théorie des singularités de Kato permet de préciser les valeurs propres. Nous noterons $\sigma(P)$ le spectre de P.

Proposition 1.5 *Une condition nécessaire et suffisante pour que le problème soit faiblement bien posé est que les deux propriétés suivantes soient vérifiées :*

(i) les valeurs propres de A *sont réelles,*

(ii) il existe $C > 0$ *tel que* $\forall z \in \mathbb{C}$, $|z| > C$, $\forall \lambda(z) \in \sigma(P(z))$, $\lambda(z)$ *issue de la branche* $\lambda_0 \in \sigma(A)$*, il existe une fonction* f *bornée sur* $|z| > C$ *telle que :*

$$\lambda(z) = \lambda_0 z + f(z).$$

PREUVE :

- Supposons les propriétés (i) et (ii) vérifiées. Alors, en appliquant (ii) à $z = i\xi$, nous avons :

$$\forall \xi \in \mathbb{R},\ |\xi| > C,\ \forall \lambda(\xi) \in \sigma(P(i\xi))\ Re\,(\lambda(\xi)) \leq Im\,(\lambda_0)\xi + \|f\|_\infty.$$

  Et comme, d'après (i) λ_0 est réel, alors $Re\,(\lambda(\xi))$ est majoré pour $|\xi| > C$. Comme sur le compact $|\xi| \geq C$, $Re\,(\lambda(\xi))$ est majoré par continuité, le problème est faiblement bien posé.
- Supposons le problème faiblement bien posé. Alors :

$$\exists \alpha \in \mathbb{R},\ \forall \xi \in \mathbb{R},\ \forall \lambda(\xi) \in \sigma(P(i\xi)),\ Re\,(\lambda(\xi)) \leq \alpha.$$

  Soit $\mu(\tau)$ une valeur propre de $A + \tau B$. Nous effectuons le développement de $\mu(\tau)$ en série de Puiseux au voisinage de $\tau = 0$. Nous supposons que $\mu(0) = \lambda_0 \in \sigma(A)$. nous savons alors, d'après [24] et d'après le théorème 1.5, qu'il existe $c > 0$ et $p \in \mathbb{N}^*$, tels que si $|\tau| \leq c$:

$$\mu(\tau) = \lambda_0 + \sum_{k=1}^{+\infty} \alpha_k \omega^{kh} \tau^{k/p}$$

  où $\omega = \exp\left(\frac{2\pi i}{p}\right)$, $h \in \{0, \ldots, p-1\}$.
  Si $\tau = 0$ n'est pas un point exceptionnel, la relation précédente est vraie avec $p = 1$ et (ii) est vérifiée. Nous considérerons donc désormais que $\tau = 0$ est un point exceptionnel.
  Alors, en prenant $z = \frac{1}{\tau}$, si $|z| > \frac{1}{c}$, nous avons :

$$\lambda(z) = \lambda_0 z + \sum_{k=1}^{p-1} \alpha_k \omega^{kh} z^{1-k/p} + \sum_{k=p}^{+\infty} \alpha_k \omega^{kh} z^{1-k/p}.$$

  Si $z = i\xi = |\xi| \exp\left(\pm i\frac{\pi}{2}\right)$, alors :

$$\begin{aligned} Re\,(\lambda(i\xi)) \quad = \quad & \sum_{k=1}^{p-1} Re\,(\alpha_k \omega^{kh} \exp(\pm i\pi/2(1-k/p))|\xi|^{1-k/p} \\ & + \sum_{k=p}^{+\infty} Re\,(\alpha_k \omega^{kh} \exp(\pm i\pi/2(1-k/p))|\xi|^{1-k/p}. \end{aligned}$$

  Comme la deuxième somme est bornée pour $|\tau| \leq c$, et que $Re\,(\lambda(i\xi)) \leq \alpha$ et ceci doit être vrai pour toutes les branches, c'est-à-dire pour tous les $h \in \{0, \ldots, p-1\}$, nous devons donc avoir :

$$Re\,(\alpha_k \omega^{kh} \exp(\pm i\pi/2(1-k/p)) \leq 0,\ \forall h \in \{0, \ldots, p-1\},\ \forall k \in \{1, \ldots, p-1\}.$$

Ceci implique alors que $\forall k \in \{1, \ldots, p-1\}$, $\alpha_k = 0$.
Donc :

$$\lambda(z) = \lambda_0 z + \sum_{k=p}^{+\infty} \alpha_k \omega^{kh} z^{1-k/p}, \tag{1.2}$$

d'où la conclusion.

□

Remarque 1.6 *Une formulation équivalente de (ii) est que les valeurs propres de $\mu(\tau)$ de $A + \tau B$ sont dérivables en 0.*

Corollaire 1.3 *Si $\partial_t U = A\partial_x U + BU$ est faiblement bien posé, alors $\forall s \in \mathbb{R}$, $\forall \mu \in \mathbb{R}^*$, $\partial_t U = A\partial_x U + \mu(B + sA)U$ est faiblement bien posé.*

Preuve :

- Si $\mu = 1$:
  Le symbole associé à $\partial_t U = A\partial_x U + \mu(B + sA)U$ est $(i\xi + s)A + B$. Nous prenons $z = i\xi + s$ dans le lemme précédent, alors si $\lambda(\xi) \in \sigma((i\xi + s)A + B)$,

  $$\lambda(\xi) = \lambda_0(i\xi + s) + f(i\xi + s),$$

  où f est bornée. Ainsi $Re\,(\lambda(\xi)) \leq \lambda_0 s - \alpha$ et le problème est faiblement bien posé.
- Si $\mu \neq 1$:
  Il suffit de remplacer ξ par ξ/μ et t par t/μ pour avoir la conclusion.

□

1.4 Exemples

Dans cette partie nous étudierons quelques exemples. Les exemples choisis seront ou bien issus de la physique ou bien purement théoriques.

1.4.1 En dimension 1

Ici nous allons étudier le cas où $P(i\xi) = i\xi A + B \in \mathcal{M}_2(\mathbb{C})$ ou $\mathcal{M}_3(\mathbb{C})$. Ces exemples sont ceux qui vont être étudiés dans les résultats numériques. Nous considérerons toujours, quitte à effectuer un changement de base, que A est sous forme de Jordan.

Problèmes homogènes

Le cas d'un problème homogène, c'est-à-dire, le cas où $B = 0$ est très simple à traiter. En effet, le simple calcul de l'exponentielle d'une matrice de Jordan permet de montrer le résultat suivant :

Proposition 1.6 *Le problème de Cauchy $\partial_t U = A\partial_x U$ est faiblement bien posé de défaut $q_1 = J - 1$ où J désigne la taille maximale des blocs de Jordan de A.*

Matrices de taille 2

Le cas où A est diagonalisable conduisant à un problème fortement bien posé, le seul cas intéressant est le cas où :

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

Proposition 1.7 *Le problème de Cauchy $\partial_t U = A\partial_x U + BU$ est faiblement bien posé si et seulement si $b_{2,1} = 0$. Dans ce cas, le défaut est 1.*

PREUVE : Le symbole est :

$$P(i\xi) = \begin{pmatrix} i\xi\lambda + b_{1,1} & i\xi + b_{1,2} \\ b_{2,1} & i\xi\lambda + b_{2,2} \end{pmatrix}$$

La partie réelle de ses valeurs propres vaut :

$$\frac{b_{1,1} + b_{2,2} \pm X}{2},$$

avec :

$$X^2 = \frac{1}{2}\left((b_{1,1} - b_{2,2})^2 + 4b_{1,2}b_{2,1} + \sqrt{[(b_{1,1} - b_{2,2})^2 + 4b_{1,2}b_{2,1}]^2 + 16b_{2,1}^2\xi^2}\right).$$

Or le problème est faiblement bien posé si et seulement si la partie réelle des valeurs propres du symbole est majorée donc si et seulement si X^2 est borné ce qui n'est le cas que lorsque $b_{2,1} = 0$.
Donc le problème est faiblement bien posé si et seulement si $b_{2,1} = 0$.
De plus, on a, lorsque $b_{2,1} = 0$:

$$\exp(Pt) = \begin{pmatrix} e^{t(b_{1,1}+i\xi\lambda)} & (i\xi + b_{1,2})\frac{e^{t(b_{2,2}+i\xi\lambda)} - e^{t(b_{1,1}+i\xi\lambda)}}{b_{1,1}-b_{2,2}} \\ 0 & e^{t(i\xi\lambda+b_{2,2})} \end{pmatrix},$$

(en prolongeant par continuité quand $b_{1,1} = b_{2,2}$). Donc :

$$\|\exp(Pt)\| \leq Ce^{t\min(b_{1,1},b_{2,2})}(1 + |\xi|).$$

Ainsi, le défaut vaut 1. □

Matrices de taille 3

Pour cette étude, nous utilisons la proposition 1.5 et nous calculons le début des développements en séries de Puiseux en utilisant Maple.

Proposition 1.8

1. *Cas* $A = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$ *:*
   Le problème est faiblement bien posé si et seulement si une des deux conditions suivantes est vérifiée :
   - $B_{3,1} = 0$, $B_{3,2} + B_{2,1} = 0$ *et* $B_{1,1} = B_{3,3}$
   - $B_{3,1} = B_{3,2} = B_{2,1} = 0$

   Dans ce cas, le défaut est $q_1 = 2$.
2. *Cas* $A = \begin{pmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{pmatrix}$ *avec* $\lambda_1 \neq \lambda_2$ *:*
   Le problème est faiblement bien posé si et seulement si $B_{2,1} = 0$.
   Dans ce cas, le défaut est $q_1 = 1$.
3. *Cas* $A = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$ *:*
   Le problème est faiblement bien posé si et seulement si $B_{2,1} = 0$ *et* $B_{3,1}B_{2,3} = 0$.
   Dans ce cas, le défaut est $q_1 = 1$.

Preuve :

- Nous commençons par étudier le caractère faiblement bien posé.
  Nous calculons le développement en série de Puiseux des valeurs propres de $iA + \tau B$. Nous savons qu'il existe $p \in \{1, 2, 3\}$ tel que $\mu(\tau^{1/p})$ soit analytique. Il existe donc j tel que les valeurs propres $\mu(t)$ de $iA + t^j B$ sont analytiques au voisinage de 0 et admettent donc le développement suivant :

  $$\mu(t) = \sum_{n=0}^{+\infty} \alpha_n t^n$$

  Nous avons alors :

  $$\lambda(\xi) = \sum_{n=0}^{+\infty} \alpha_n \xi^{1-n/j}$$

  et, d'après la caractérisation prouvée dans la proposition 1.5 le problème est faiblement bien posé si et seulement si $Re\,(\alpha_1) = Re\,(\alpha_2) = \cdots = Re\,(\alpha_{j-1}) = 0$.

Pour calculer ces premiers coefficients, en utilisant Maple, nous calculons $Q_N(t) = det(iA + t^j B - \sum_{n=0}^{N} \alpha_n t^n)$ où N est bien choisi, et nous annulons les coefficients polynomiaux en t.

1. Pour le premier cas, nous prenons $j = 6$ car dans ce cas, nous n'avons aucune indication sur la valeur de $p \in \{1, 2, 3\}$. En prenant le plus petit multiple commun de ces trois valeurs, on aura bien une fonction analytique mais nous aurons rajouté certains coefficients nuls. Nous avons alors :
   - pour $N = 0$, le coefficient d'ordre 0 de Q_0 est : $(\alpha_0 - i\lambda)^3$.
     Nous prenons donc pour la suite $\alpha_0 = i\lambda$.
   - pour $N = 1$, les coefficients d'ordre 0, 1 et 2 sont nuls et le coefficient d'ordre 3 de Q_1 est : α_1^3.
     Nous prenons donc pour la suite $\alpha_1 = 0$.
   - pour $N = 2$, les coefficients d'ordre 0 à 5 sont nuls et le coefficient d'ordre 6 de Q_2 est : $-\alpha_2^3 - B_{3,1}$.
     Nous prenons donc pour la suite $\alpha_2 = B_{3,1} = 0$.
   - pour $N = 3$, les coefficients d'ordre 0 à 8 sont nuls et le coefficient d'ordre 9 de Q_3 est :$-\alpha_3(\alpha_3^2 - i(B_{3,2} + B_{2,1})$.
     Nous prenons donc pour la suite $\alpha_3 = B_{3,2} + B_{2,1} = 0$.
   - pour $N = 4$, les coefficients d'ordre 0 à 11 sont nuls et le coefficient d'ordre 12 de Q_4 est :$-\alpha_4^3 - iB_{2,1}(B_{3,3} - B_{1,1})$.
     Nous avons donc deux cas à étudier : $\alpha_4 = B_{2,1} = 0$ ou ($\alpha_4 = 0$ et $B_{3,3} = B_{1,1}$).
   - pour $N = 5$, dans les deux cas, les coefficients d'ordre 0 à 14 sont nuls et le coefficient d'ordre 15 de Q_5 est :$-\alpha_5^3$.
     Donc $\alpha_5 = 0$.
2. Pour le deuxième cas, nous prenons $j = 2$ car dans ce cas, comme A n'a pas de valeur propre de multiplicité 3, alors $p = 1$ ou 2. Nous effectuons le même raisonnement que précédemment :
   - pour $N = 0$, le coefficient d'ordre 0 de Q_0 est : $(\alpha_0 - i\lambda_1)^2(\alpha_0 - i\lambda_2)$.
     Nous avons donc deux cas à étudier :$\alpha_0 = i\lambda_1$ ou $\alpha_0 = i\lambda_2$.
   - dans le cas où $\alpha_0 = i\lambda_1$, pour $N = 1$, les coefficients d'ordre 0 et 1 sont nuls et le coefficient d'ordre 2 de Q_1 est : $-i(\lambda_1 - \lambda_2)\alpha_1^2 - (\lambda_1 - \lambda_2)B_{2,1}$.
     Nous prenons donc pour la suite $\alpha_1 = B_{2,1} = 0$.
   - dans le cas où $\alpha_0 = i\lambda_2$, pour $N = 1$, le coefficient d'ordre 0 est nul et le coefficient d'ordre 1 de Q_1 est : $-\alpha_1(\lambda_1 - \lambda_2)^2 = 0$.
     Nous prenons donc pour la suite $\alpha_1 = 0$.
3. Pour le troisième cas, de même que dans le premier cas, nous prenons $j = 6$. Nous avons alors :
   - pour $N = 0$, le coefficient d'ordre 0 de Q_0 est : $(\alpha_0 - i\lambda)^3$.

Nous prenons donc pour la suite $\alpha_0 = i\lambda$.

- pour $N = 1$, les coefficients d'ordre 0, 1 et 2 sont nuls et le coefficient d'ordre 3 de Q_1 est : α_1^3.
  Nous prenons donc pour la suite $\alpha_1 = 0$.
- pour $N = 2$, les coefficients d'ordre 0 à 5 sont nuls et le coefficient d'ordre 6 de Q_2 est : α_2^3.
  Nous prenons donc pour la suite $\alpha_2 = 0$.
- pour $N = 3$, les coefficients d'ordre 0 à 8 sont nuls et le coefficient d'ordre 9 de Q_3 est :$-\alpha_3(\alpha_3^2 - iB_{2,1})$.
  Nous prenons donc pour la suite $\alpha_3 = B_{2,1} = 0$.
- pour $N = 4$, les coefficients d'ordre 0 à 11 sont nuls et le coefficient d'ordre 12 de Q_4 est :$\alpha_4^3 - iB_{3,1}B_{2,3}$.
  Nous prenons donc pour la suite : $\alpha_4 = B_{3,1}B_{2,3} = 0$ ou ($\alpha_4 = 0$ et $B_{3,3} = B_{1,1}$).
- pour $N = 5$, les coefficients d'ordre 0 à 14 sont nuls et le coefficient d'ordre 15 de Q_5 est :$-\alpha_5^3$.
  Donc $\alpha_5 = 0$.

Nous avons bien les résultats voulus en annulant toutes les branches.

- Nous allons maintenant calculer le défaut.
  Nous commençons par calculer l'exponentielle d'une matrice $P \in \mathcal{M}_3(\mathbb{C})$ de polynôme caractéristique $\chi_P(X) = (X-\lambda_1)(X-\lambda_2)(X-\lambda_3)$. Nous supposons d'abord que les trois valeurs propres sont distinctes. Nous prolongerons ensuite les formules par continuité. Nous effectuons la division euclidienne de X^n par $\chi_P(X)$:

$$X^n = \chi_P Q + aX^2 + bX + c.$$

  Nous obtenons alors le système :

$$\left\{ \begin{array}{rcl} \lambda_1^n & = & a\lambda_1^2 + b\lambda_1 + c \\ \lambda_2^n & = & a\lambda_2^2 + b\lambda_2 + c. \\ \lambda_3^n & = & a\lambda_3^2 + b\lambda_3 + c \end{array} \right.$$

  La solution de ce système est alors :

$$a = \frac{(\lambda_2 - \lambda_3)\lambda_1^n + (\lambda_3 - \lambda_1)\lambda_2^n + (\lambda_1 - \lambda_2)\lambda_3^n}{(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_3)},$$

$$b = \frac{(\lambda_3^2 - \lambda_2^2)\lambda_1^n + (\lambda_1^2 - \lambda_3^2)\lambda_2^n + (\lambda_2^2 - \lambda_1^2)\lambda_3^n}{(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_3)},$$

$$c = \frac{\lambda_2\lambda_3(\lambda_2 - \lambda_3)\lambda_1^n + \lambda_1\lambda_3(\lambda_3 - \lambda_1)\lambda_2^n + \lambda_1\lambda_2(\lambda_1 - \lambda_2)\lambda_3^n}{(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_3)}.$$

Donc :

$$\begin{aligned}\exp(tP) &= \frac{1}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)(\lambda_2-\lambda_3)}\\ &\quad\times\big([(\lambda_2-\lambda_3)e^{\lambda_1 t}+(\lambda_3-\lambda_1)e^{\lambda_2 t}+(\lambda_1-\lambda_2)e^{\lambda_3 t}]P^2\\ &\quad+[(\lambda_3^2-\lambda_2^2)e^{\lambda_1 t}+(\lambda_1^2-\lambda_3^2)e^{\lambda_2 t}+(\lambda_2^2-\lambda_1^2)e^{\lambda_3 t}]P\\ &\quad+[\lambda_2\lambda_3(\lambda_2-\lambda_3)e^{\lambda_1 t}+\lambda_1\lambda_3(\lambda_3-\lambda_1)e^{\lambda_2 t}+\lambda_1\lambda_2(\lambda_1-\lambda_2)e^{\lambda_3 t}]I\big)\\ &= f_2(t,\lambda_1,\lambda_2,\lambda_3)P^2+f_1(t,\lambda_1,\lambda_2,\lambda_3)P+f_0(t,\lambda_1,\lambda_2,\lambda_3)I.\end{aligned}$$

Si $\lambda_2=\lambda_1$ et si $\lambda_3\neq\lambda_1$, nous prolongeons la formule par continuité :

$$\begin{aligned}\exp(tP) &= \frac{1}{(\lambda_3-\lambda_1)^2}\big([(1+t(\lambda_3-\lambda_1))e^{\lambda_1 t}-e^{\lambda_3 t}]P^2\\ &\quad+[(-2\lambda_1+t(\lambda_1^2-\lambda_3^2))e^{\lambda_1 t}+2\lambda_1 te^{\lambda_3 t}]P\\ &\quad+[\lambda_3(2\lambda_1-\lambda_3+t\lambda_1(\lambda_3-\lambda_1))e^{\lambda_1 t}-\lambda_1^2e^{\lambda_3 t}]I\big)\\ &= g_2(t,\lambda_1,\lambda_3)P^2+g_1(t,\lambda_1,\lambda_3)P+g_0(t,\lambda_1,\lambda_3)I,\end{aligned}$$

avec, pour $j\in\{1,2,3\}$, $\lim_{\lambda_2\to\lambda_1} f_j(t,\lambda_1,\lambda_3,\lambda_3)=g_j(t,\lambda_1,\lambda_3)$.
Si $\lambda_3=\lambda_2=\lambda_1$, nous obtenons :

$$\begin{aligned}\exp(tP) &= \frac{t^2}{2}e^{\lambda_1 t}P^2+(t-\lambda_1 t^2)e^{\lambda_1 t}P+(1-\lambda_1 t+\frac{\lambda_1^2t^2}{2}e^{\lambda_1 t}I\\ &= h_2(t,\lambda_1)P^2+h_1(t,\lambda_1)P+h_0(t,\lambda_1)I,\end{aligned}$$

avec, pour $j\in\{1,2,3\}$, $\lim_{\lambda_3\to\lambda_1} g_j(t,\lambda_1,\lambda_3)=h_j(t,\lambda_1)$.

Nous appliquons maintenant ces résultats à $P(i\xi)=i\xi A+B$. Si nous notons $\lambda_1(\xi)$, $\lambda_2(\xi)$ et $\lambda_3(\xi)$ ses valeurs propres, nous savons qu'il existe $C>0$ tel que pour $|\xi|>C$, leur nombre est constant et, d'après la proposition 1.5, $\lambda_j(\xi)=i\xi\lambda_j^0+\phi_j(\xi)$, où λ_j^0 valeur propre de A et ϕ_j bornée pour $|\xi|>C$.

(i) $\underline{\lambda_1(\xi)=\lambda_2(\xi)=\lambda_3(\xi)}$.
Alors $\lambda_1^0=\lambda_2^0=\lambda_3^0=\lambda$ donc nous sommes dans le premier ou le troisième cas de la proposition. Nous avons, en effectuant le développement limité pour $|\xi|\to+\infty$ de l'expression trouvée pour le calcul de l'exponentielle :

$$\exp(tP(i\xi)=-\frac{t^2}{2}\xi^2e^{\lambda_1(\xi)t}(A-\lambda I)^2+O(\xi).$$

Comme dans le premier cas, $(A-\lambda I)^2\neq 0$, $q_1=2$. Et pour le troisième cas, $(A-\lambda I)^2=0$ donc $q_1=1$.

(ii) $\underline{\lambda_2(\xi) = \lambda_1(\xi) \text{ et } \lambda_3(\xi) \neq \lambda_1(\xi)}$.

Alors $\lambda_1^0 = \lambda_2^0$.

- Si $\lambda_3(\xi) - \lambda_1(\xi) \to 0$, alors $\lambda_1^0 = \lambda_3^0$ et donc $\lambda_1^0 = \lambda_2^0 = \lambda_3^0 = \lambda$. Et comme $\exp(tP(i\xi) - (h_2(t, \lambda_1(\xi))P^2(i\xi) + h_1(t, \lambda_1(\xi))P(i\xi) + h_0(t, \lambda_1(\xi))I) \to 0$, nous pouvons nous ramener au cas précédent.
- Si $\lambda_3(\xi) - \lambda_1(\xi)$ ne tend pas vers 0.
  - Si $\lambda_1^0 \neq \lambda_3^0$, nous sommes dans le deuxième cas de la proposition et nous avons, en effectuant le développement limité pour $|\xi| \to +\infty$ de l'expression trouvé pour le calcul de l'exponentielle :

$$\exp(tP(i\xi) = -i\xi t e^{\lambda_1(\xi)t}(A - \lambda_1^0 I)(A - \lambda_3^0 I) + O(1).$$

    Or, comme A n'est pas diagonalisable $(A - \lambda_1^0 I)(A - \lambda_3^0 I) \neq 0$ donc $q_1 = 1$.
  - Si $\lambda_1^0 = \lambda_3^0$, l'exponentielle vaut :

$$\begin{aligned} \exp(tP(i\xi) &= -\frac{1}{(\lambda_3(\xi) - \lambda_1(\xi))^2}([-\xi^2(1 - t(\lambda_3(\xi) - \lambda_1(\xi))]e^{\lambda_1(\xi)t} \\ &\quad +\xi^2 e^{\lambda_3(\xi)t}](A - \lambda_1^0)^2 + O(\xi)). \end{aligned}$$

    De plus, par hypothèse $\frac{1}{(\lambda_3(\xi)-\lambda_1(\xi))^2} = O(1)$, donc, si $(A - \lambda_1^0) \neq 0$, ce qui correspond au premier cas, $q_1 = 2$ et si $(A - \lambda_1^0) = 0$, ce qui correspond au troisième cas, $q_1 = 1$.

(iii) $\underline{\lambda_1(\xi) \neq \lambda_2(\xi) \neq \lambda_3(\xi)}$.

- Si $\lambda_3(\xi) - \lambda_1(\xi) \to 0$ et si $\lambda_2(\xi) - \lambda_1(\xi) \to 0$, nous nous ramenons au (i).
- Si $\lambda_2(\xi) - \lambda_1(\xi) \to 0$ et si $\lambda_3(\xi) - \lambda_2(\xi)$ ne tend pas vers 0, nous nous ramenons au (ii).
- Si $\lambda_2(\xi) - \lambda_1(\xi)$, $\lambda_3(\xi) - \lambda_2(\xi)$ et $\lambda_3(\xi) - \lambda_1(\xi)$ ne tendent pas vers 0. Nous avons alors le développement limité suivant :

$$\begin{aligned} \exp(tP(i\xi)) &= \frac{-i\xi^3}{(\lambda_2(\xi) - \lambda_1(\xi))(\lambda_3(\xi) - \lambda_1(\xi))(\lambda_2(\xi) - \lambda_3(\xi))} \\ &\quad \times((e^{\lambda_1(\xi)t}(\lambda_2^0 - \lambda_3^0)(A - \lambda_2^0 I)(A - \lambda_3^0 I) \\ &\quad +(e^{\lambda_2(\xi)t}(\lambda_3^0 - \lambda_1^0)(A - \lambda_3^0 I)(A - \lambda_1^0 I) \\ &\quad +(e^{\lambda_3(\xi)t}(\lambda_1^0 - \lambda_2^0)(A - \lambda_1^0 I)(A - \lambda_2^0 I) + O(1/\xi)). \end{aligned}$$

- Si $\lambda_1^0 = \lambda_2^0 \neq \lambda_3^0$, nous sommes dans le second cas , et nous avons alors $\frac{1}{\lambda_2(\xi)-\lambda_1(\xi)} = O(1)$ et $\frac{1}{(\lambda_3(\xi)-\lambda_1(\xi))(\lambda_3(\xi)-\lambda_1(\xi))} = O(1/\xi^2)$. De plus, comme $(\lambda_2^0 - \lambda_3^0)(A - \lambda_2^0 I)(A - \lambda_3^0 I)_n eq 0$, nous avons $q_2 = 1$.

– Si $\lambda_1^0 = \lambda_2^0 = \lambda_3^0 = \lambda$, alors comme dans le premier cas $(A - \lambda I)^2 \neq 0$, nous avons $q_1 = 2$ et dans le troisième cas $(A - \lambda I)^2 = 0$, nous avons $q_1 = 1$.

Nous avons bien étudié tous les cas, nous avons donc démontré le calcul du défaut pour un problème faiblement bien posé de taille 3.

□

Remarque 1.7 *Pour calculer* $\exp(tP(i\xi))$*, nous aurions pu aussi utiliser la formule de Newton [17]. En effet, si nous considérons la fonction* $f(z) = e^z$*, et si* M *est une matrice de taille 3, alors* $f(M)$ *est un polynôme de degré 2 donc est égal à son polynôme d'interpolation de degré 2 passant par 3 points. Donc, si* λ_1*,* λ_2 *et* λ_3 *sont les valeurs propres de* M*, la formule de Newton donne :*

$$\exp(M) = f(\lambda_1)Id + f[\lambda_1, \lambda_2](M - \lambda_1 Id) + f[\lambda_1, \lambda_2, \lambda_3](M - \lambda_1 Id)(M - \lambda_2 Id),$$

où les différences divisées de f *sont définies par la formule de récurrence suivante :*

$$f[\lambda_1] = f(\lambda_1),$$

$$\forall k \geq 1,\ f[\lambda_1, \ldots, \lambda_{k+1}] = \frac{f[\lambda_2, \ldots, \lambda_{k+1}] - f[\lambda_1, \ldots, \lambda_k]}{\lambda_{k+1} - \lambda_1}.$$

Remarque 1.8 *Dans le cas de matrices de taille 2 ou 3, le défaut est égal à la dimension maximale des blocs de Jordan de* A *moins 1 et est indépendant du choix de* B*. Ceci n'est pas vrai dans le cas général. En effet, si nous choisissons :*

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Alors, la taille des blocs de Jordan moins 1 (qui correspond au défaut pour le problème sans terme d'ordre 0) est 1. Alors que si nous choisissons :

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

le défaut est $q_1 = 2$*. En effet, en écrivant* $A = I + E$*, nous avons :*

$$\exp(tP(i\xi)) = \exp(i\xi t)\exp(t(i\xi)E + B).$$

Et comme $E^2 = B^2 = BEB = 0$ *et* $EBE \neq 0$*, nous obtenons :*

$$\exp(tP(i\xi)) = \exp(i\xi t)(Id + t(i\xi E + B) + \frac{t^2}{2} i\xi(EB + BE) - \frac{t^3}{6}\xi^2 EBE).$$

1.4.2 Equations d'Euler pour un modèle diphasique

Les équations d'Euler linéarisées s'écrivent sous la forme :

$$\partial_t \begin{pmatrix} u \\ v \\ w \\ \rho \end{pmatrix} + A_1 \partial_x \begin{pmatrix} u \\ v \\ w \\ \rho \end{pmatrix} + A_2 \partial_y \begin{pmatrix} u \\ v \\ w \\ \rho \end{pmatrix} + A_3 \partial_z \begin{pmatrix} u \\ v \\ w \\ \rho \end{pmatrix} = 0$$

avec :

$$A_1 = \begin{pmatrix} U & 0 & 0 & \kappa/R \\ 0 & U & 0 & 0 \\ 0 & 0 & U & 0 \\ R & 0 & 0 & U \end{pmatrix}, \; A_2 = \begin{pmatrix} V & 0 & 0 & \kappa/R \\ 0 & V & 0 & 0 \\ 0 & 0 & V & 0 \\ 0 & R & 0 & V \end{pmatrix}$$

$$\text{et } A_3 = \begin{pmatrix} W & 0 & 0 & \kappa/R \\ 0 & W & 0 & 0 \\ 0 & 0 & W & 0 \\ 0 & 0 & R & W \end{pmatrix}$$

où (u, v, w) représente le déplacement du fluide au voisinage de l'état constant (U, V, W), ρ la densité du fluide au voisinage de R et $\kappa = \frac{dr}{d\rho}(R)$ où la pression pour le système non linéarisé est $p = r(\rho)$.

Si nous considérons un modèle diphasique liquide-vapeur, entre le passage liquide-vapeur, nous introduisons une zone de mélange dans laquelle l'entropie est l'enveloppe concave de l'entropie des deux phases. Nous obtenons alors une entropie qui est linéaire dans la phase de mélange.

Or nous avons $p = T\frac{\partial s}{\partial \tau}$ où T est la température, s l'entropie et $\tau = 1/\rho$ le volume. Donc, dans le mélange, la pression est constante ainsi $\kappa = 0$. Dans ce cas le symbole est :

$$P(i\xi) = -\begin{pmatrix} i\xi.\mathbf{U} & 0 & 0 & 0 \\ 0 & i\xi.\mathbf{U} & 0 & 0 \\ 0 & 0 & i\xi.\mathbf{U} & 0 \\ i\xi_1 R & i\xi_2 R & i\xi_3 R & i\xi.\mathbf{U} \end{pmatrix}$$

avec $i\xi.\mathbf{U} = i\xi_1 U + i\xi_2 V + i\xi_3 W$, et nous avons :

$$\exp(tP(i\xi)) = \exp(-(i\xi_1 U + i\xi_2 V + i\xi_3 W)t) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -i\xi_1 Rt & -i\xi_2 Rt & -i\xi_3 Rt & 1 \end{pmatrix}$$

En conclusion, nous avons démontré le résultat suivant :

Proposition 1.9 *Dans la zone de mélange, les équations d'Euler linéarisées sont faiblement bien posées de défaut* $q_1 = 1$.

1.4.3 Equation PML pour les équations de Maxwell 2D TE

Les équations de Maxwell TE en deux dimensions s'écrivent sous la forme :

$$\partial_t \begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} + A_1 \partial_x \begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} + A_2 \partial_y \begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} + B \begin{pmatrix} E_x \\ E_y \\ H_z \end{pmatrix} = 0$$

avec :

$$A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/\varepsilon_0 \\ 0 & 1/\mu_0 & 0 \end{pmatrix}, \; A_2 = \begin{pmatrix} 0 & 0 & -1/\varepsilon_0 \\ 0 & 0 & 0 \\ -1/\mu_0 & 0 & 0 \end{pmatrix} \text{ et } B = \begin{pmatrix} \sigma/\varepsilon_0 & 0 & 0 \\ 0 & \sigma/\varepsilon_0 & 0 \\ 0 & 0 & \sigma^*/\mu_0 \end{pmatrix}$$

où (E_x, E_y) représente le champ électrique, H_z le champ magnétique, ε_0 la permitivité du milieu, μ_0 sa perméabilité et σ, σ^*les pertes électriques et magnétiques.

Afin d'étudier la propagation des ondes électromagnétiques en domaine non borné, Bérenger [12] a introduit les couches parfaitement adaptées (PML). Pour ces équations il s'agit de splitter le champ magnétique en deux composantes non physiques et d'introduire une absorption. Il obtient alors le système suivant :

$$\partial_t \begin{pmatrix} E_x \\ E_y \\ H_{zx} \\ H_{zy} \end{pmatrix} + \tilde{A}_1 \partial_x \begin{pmatrix} E_x \\ E_y \\ H_{zx} \\ H_{zy} \end{pmatrix} + \tilde{A}_2 \partial_y \begin{pmatrix} E_x \\ E_y \\ H_{zx} \\ H_{zy} \end{pmatrix} + \tilde{B} \begin{pmatrix} E_x \\ E_y \\ H_{zx} \\ H_{zy} \end{pmatrix} = 0$$

avec :

$$\tilde{A}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/\varepsilon_0 & 1/\varepsilon_0 \\ 0 & 1/\mu_0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; \tilde{A}_2 = \begin{pmatrix} 0 & 0 & -1/\varepsilon_0 & -1/\varepsilon_0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1/\mu_0 & 0 & 0 & 0 \end{pmatrix}$$

$$\tilde{B} = \begin{pmatrix} \sigma_y/\varepsilon_0 & 0 & 0 & 0 \\ 0 & \sigma_x/\varepsilon_0 & 0 & 0 \\ 0 & 0 & \sigma_x^*/\mu_0 & 0 \\ 0 & 0 & 0 & \sigma_y^*/\mu_0 \end{pmatrix}$$

Pour plus de détails sur les équations PML, on pourra consulter la deuxième partie de cette thèse.

Dans [1], Abarbanel et Gottlieb montrent que, même si les équations de Maxwell sont fortement bien posées, le résultat suivant est vrai pour les équations PML :

Proposition 1.10 *Les équations PML avec absorption nulle ne sont que faiblement bien posées de défaut* $q_1 = 1$.

La preuve proposée dans [1] est basée sur le calcul explicite de la transformée de Fourier du champ électromagnétique pour les PML. Toutefois ce résultat est plus général. En effet, nous verrons dans la partie 9.2.1, que le symbole principal du problème PML associé à $\partial_t U - A_1\partial_x U - A_2\partial_y U = 0$ est :

$$P(i\xi) = \begin{pmatrix} i\xi_1 A_1 & i\xi_1 A_1 \\ i\xi_2 A_2 & i\xi_2 A_2 \end{pmatrix},$$

ce qui correspond au symbole des équations PML avec absorption nulle.

Nous avons alors le résultat général suivant :

Proposition 1.11 *Si le problème $\partial_t U - A_1\partial_x U - A_2\partial_y U = 0$ est fortement bien posé, alors le problème PML avec absorption nulle i.e. le problème de symbole $P(i\xi)$ est faiblement bien posé de défaut 1.*

PREUVE : Nous allons calculer $\exp(tP(i\xi))$. Pour cela nous allons résoudre l'équation différentielle $Y'(t) = P(i\xi)Y(t)$. Nous notons $Y(t) = \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix}$ et $Z(t) = Y_1(t) + Y_2(t)$. Nous avons alors :

$$Y'(t) = P(i\xi)Y(t) \Rightarrow Z'(t) = (i\xi_1 A_1 + i\xi_2 A_2)Z(t).$$

Donc $Z(t) = exp((i\xi_1 A_1 + i\xi_2 A_2)t)Z(0)$. Or $Y_1'(t) = i\xi_1 A_1 Z(t) = i\xi_1 A_1 exp((i\xi_1 A_1 + i\xi_2 A_2)t)Z(0)$. Nous avons donc, en faisant de même pour Y_2 :

$$\begin{cases} Y_1(t) &= Y_1(0) + \int_0^t i\xi_1 A_1 exp((i\xi_1 A_1 + i\xi_2 A_2)\tau)Z(0)d\tau \\ Y_2(t) &= Y_2(0) + \int_0^t i\xi_2 A_2 exp((i\xi_1 A_1 + i\xi_2 A_2)\tau)Z(0)d\tau \end{cases}.$$

D'où :

$$Y(t) = Y(0) + \int_0^t \begin{pmatrix} i\xi_1 A_1 \\ i\xi_2 A_2 \end{pmatrix} exp((i\xi_1 A_1 + i\xi_2 A_2)\tau)(Id_N \quad Id_N)d\tau Y(0),$$

soit :

$$\exp(tP(i\xi)) = Id_{2N} + \int_0^t \begin{pmatrix} i\xi_1 A_1 \\ i\xi_2 A_2 \end{pmatrix} exp((i\xi_1 A_1 + i\xi_2 A_2)\tau)(Id_N \quad Id_N)d\tau.$$

Or, le problème $\partial_t U - A_1\partial_x U - A_2\partial_y U = 0$ est fortement bien posé, donc $\|exp((i\xi_1 A_1 + i\xi_2 A_2)\tau)\| \leq K_C e^{\alpha_C t}$. Ainsi,

$$\|\exp(tP(i\xi))\| \leq 1 + \int_0^t \left\| \begin{pmatrix} i\xi_1 A_1 \\ i\xi_2 A_2 \end{pmatrix} \right\| K_C e^{\alpha_C \tau} d\tau.$$

Donc, il existe des constantes K'_C et α'_C telles que $\| \exp(tP(i\xi)) \| \leq K'_C e^{\alpha'_C t}(1+\|\xi\|)$. Ce qui prouve que le problème de symbole $P(i\xi)$ est faiblement bien posé de défaut 1. □

Dans le cas des équations de Maxwell, le splitting ne concerne pas toutes les variables (par exemple, dans le cas de Maxwell TE, le champ électrique n'est pas splitté) donc le symbole du problème PML sans absorption n'est pas de la forme précédente. Nous allons toutefois prouver un résultat similaire pour le modèle PML classique sans absorption associé aux équations de Maxwell tridimensionnelles. Nous allons étudier les équations sous la forme suivante :

$$\begin{cases} \partial_t E_{xy} - \partial_y(H_{zx} + H_{zy}) = 0 \\ \partial_t E_{yz} - \partial_z(H_{xy} + H_{xz}) = 0 \\ \partial_t E_{zx} - \partial_x(H_{yz} + H_{yx}) = 0 \\ \partial_t E_{xz} + \partial_z(H_{yz} + H_{yx}) = 0 \\ \partial_t E_{yx} + \partial_x(H_{zx} + H_{zy}) = 0 \\ \partial_t E_{zy} + \partial_y(H_{xy} + H_{xz}) = 0 \\ \partial_t H_{xy} + \partial_y(E_{zx} + E_{zy}) = 0 \\ \partial_t H_{yz} + \partial_z(E_{xy} + E_{xz}) = 0 \\ \partial_t H_{zx} + \partial_x(E_{yz} + E_{yx}) = 0 \\ \partial_t H_{xz} - \partial_z(E_{yz} + E_{yx}) = 0 \\ \partial_t H_{yx} - \partial_x(E_{zx} + E_{zy}) = 0 \\ \partial_t H_{zy} - \partial_y(E_{xy} + E_{xz}) = 0 \end{cases} \tag{1.3}$$

Nous donnerons plus de détails sur ces équations dans la partie 7.2.3.

Proposition 1.12 *Le modèle PML classique sans absorption associé aux équations de Maxwell tridimensionnelles 1.3 est faiblement bien posé de défaut 1.*

PREUVE : Notons $P(i\xi)$ le symbole des équations 1.3. Il est de la forme : $P(i\xi) = \begin{pmatrix} 0 & M(i\xi) \\ -M(i\xi) & 0 \end{pmatrix}$ où $M(i\xi)$ est de la forme $M(i\xi) = \begin{pmatrix} B_1(i\xi) & B_1(i\xi) \\ B_2(i\xi) & B_2(i\xi) \end{pmatrix}$, avec :

$$B_1(i\xi) + B_2(i\xi) = \begin{pmatrix} 0 & -i\xi_3 & i\xi_2 \\ i\xi_3 & 0 & -i\xi_1 \\ -i\xi_2 & i\xi_1 & 0 \end{pmatrix}.$$

Or nous savons que :

$$\exp(tP(i\xi)) = \begin{pmatrix} \frac{\exp(iM(i\xi)t)+\exp(-iM(i\xi)t)}{2} & \frac{\exp(iM(i\xi)t)-\exp(-iM(i\xi)t)}{2i} \\ -\frac{\exp(iM(i\xi)t)-\exp(-iM(i\xi)t)}{2i} & \frac{\exp(iM(i\xi)t)+\exp(-iM(i\xi)t)}{2} \end{pmatrix}.$$

Il suffit donc de prouver qu'il existe K'_C et α'_C tels que $\| \exp(iM(i\xi)t) \| \leq K'_C e^{\alpha'_C t}(1+\|\xi\|)$ pour avoir le résultat. Or si nous remplaçons, dans la preuve de la proposition

précédente $i\xi_1 A_1$ (resp. $i\xi_2 A_2$) par $iB_1(i\xi)$ (resp. $iB_2(i\xi)$), il suffit de prouver qu'il existe K_C et α_C tels que $\|\exp(i(B_1(i\xi)+B_2(i\xi))t)\| \leq K_C e^{\alpha_C t}$ ce qui correspond au caractère fortement bien posé du problème initial. Or la matrice $i(B_1(i\xi)+B_2(i\xi))$ est anti-symétrique réelle donc elle est diagonalisable dans une base orthonormée et ses valeurs propres sont imaginaires pures. Nous avons donc bien l'estimation $\|\exp(i(B_1(i\xi)+B_2(i\xi))t)\| \leq K_C e^{\alpha_C t}$ ce qui achève la preuve.

□

Chapitre 2

Théorèmes généraux sur les schémas

Nous allons maintenant nous intéresser à la discrétisation par des schémas aux différences finies des problèmes de Cauchy faiblement bien posés. Pour cela, nous devons modifier les définitions usuelles utilisées dans la théorie des schémas aux différences finies pour les problèmes fortement bien posés. En effet, les définitions de stabilité, convergence et consistance ne prennent pas en compte la perte de régularité. Nous allons donc les élargir afin de nous adapter aux problèmes faiblement bien posés.

Une fois que nous aurons donné des définitions satisfaisantes, nous étudierons la convergence des schémas. Pour cela, nous prouverons une extension du théorème de Lax-Richtmyer qui s'applique aux problèmes faiblement bien posés.

2.1 Préliminaires : les outils

Nous utilisons les notations introduites par Strikwerda dans [38].

Nous allons considérer une grille de $\mathbb{R}^d$ de pas d'espace $h_1, \ldots, h_d$, notée $G = h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z}$. Nous notons $h = (h_1, \ldots, h_d)$ et dans le cas de la dimension 1, le pas h_1 sera noté h. Nous allons définir les différents outils utiles pour la suite.

2.1.1 Transformée de Fourier discrète

Nous définissons l'espace L^2 discret sur la grille par :

$$L^2(G) = \{V \in (\mathbb{R}^N)^{\mathbb{Z}^d},\ H \sum_{m=(m_1,\ldots,m_d)\in\mathbb{Z}^d} \|V_m\|^2 < +\infty\}$$

muni de la norme :

$$\|V\|_h^2 = H \sum_{m\in\mathbb{Z}^d} \|V_m\|^2$$

avec $H = \left(\prod_{j=1}^{d} h_j\right)$ et où $\|V_m\|$ désigne la norme euclidienne de V_m dans $\mathbb{R}^N$.

La transformée de Fourier discrète de $V \in L^2(G)$ est, pour $\xi \in \mathcal{D}_d$, où nous posons $\mathcal{D}_d = \prod_{j=1}^{d}[-\frac{\pi}{h_j}, \frac{\pi}{h_j}]$:

$$\mathcal{F}_h(V)(\xi) = \frac{H}{(2\pi)^{d/2}} \sum_{m\in\mathbb{Z}^d} e^{-i\sum_{j=1}^{d} m_j h_j \xi_j} V_m.$$

Nous notons $\mathcal{F}$ la transformée de Fourier continue.
Lorsqu'il n'y a pas d'ambiguïté, nous noterons $\widehat{.}$ les transformées de Fourier discrète et continue.

Nous avons la formule d'inversion suivante :

$$V_m = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}_d} e^{i\sum_{j=1}^{d} m_j h_j \xi_j} \mathcal{F}_h(V)(\xi) d\xi$$

ainsi que l'identité de Parseval, pour $V \in L^2(G)$:

$$\|V\|_h = \|\mathcal{F}_h(V)\|_{L^2(\mathcal{D}_d)}$$

2.1.2 Interpolation

Si $V \in L^2(G)$, nous définissons son interpolée $SV \in L^2(\mathbb{R}^d)$ par :

$$SV(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}_d} e^{ix.\xi} \mathcal{F}_h(v)(\xi) d\xi.$$

Nous avons alors :

$$\widehat{SV}(\xi) = \begin{cases} \mathcal{F}_h(V)(\xi) & si\, \xi \in \mathcal{D}_d, \\ 0 & sinon. \end{cases}$$

2.1.3 Espaces de Sobolev discrets

Si $V \in (\mathbb{R}^N)^{\mathbb{Z}^d}$, nous définissons la dérivée d'ordre q_1 par rapport à la première variable, ..., q_d par rapport à la $d^{\text{ième}}$ variable de V de pas $h_1, \ldots, h_d$ par $D_1^{q_1} \ldots D_d^{q_d} V = ((D_1^{q_1} \ldots D_d^{q_d} V)_m)_{m\in\mathbb{Z}^d}$ où :

$$D_j(V)_m = \frac{V_m - V_{m_1,\ldots,m_{j-1},\ldots,m_d}}{h_j}.$$

Calculons la transformée de Fourier des dérivées :

Lemme 2.1

$$\mathcal{F}_h(D_1^{q_1} \ldots D_d^{q_d} V)(\xi) = e^{i\sum_{j=1}^{d}(h_j \xi_j q_j)/2} \prod_{j=1}^{d} \left(\frac{\sin(h_j\xi_j/2)}{h_j}\right)^{q_j} \mathcal{F}_h(V)(\xi).$$

PREUVE : Il suffit de montrer le résultat pour $\mathcal{F}_h(D_jV)(\xi)$. Nous avons :

$$\mathcal{F}_h(D_jV)(\xi) = \frac{1-e^{-ih_j\xi_j}}{h_j}\mathcal{F}_h(V)(\xi) = e^{ih_j\xi_j/2}\left(\frac{\sin(h_j\xi_j/2)}{h_j}\right)\mathcal{F}_h(V)(\xi).$$

En utilisant $\mathcal{F}_h(D_1^{q_1}\dots D_d^{q_d}) = \mathcal{F}_h(D_1)^{q_1}\dots\mathcal{F}_h(D_d)^{q_d}$, nous avons prouvé le résultat. □

Nous définissons alors l'espace de Sobolev discret :

$$H^q(G) = \{V \in (\mathbb{R}^N)^{\mathbb{Z}^d}, \sum_{q_1+\dots q_d \leq q} \|D_1^{q_1}\dots D_d^{q_d}V\|_h^2\} < \infty,$$

muni de la norme :

$$|||V|||_{h,q}^2 = \sum_{q_1+\dots q_d \leq q} \|D_1^{q_1}\dots D_d^{q_d}V\|_h^2.$$

L'interpolation permet de faire le lien entre les normes discrètes et les normes continues. En effet, nous définissons l'espace suivant :

$$\widetilde{H^q}(G) = \{V \in L^2(G), \int_{\mathcal{D}_d}(1+\|\xi\|^q)^2\|\mathcal{F}_h(V)(\xi)\|^2 d\xi < +\infty\},$$

muni de la norme :

$$\|V\|_{h,q}^2 = \frac{1}{(2\pi)^d}\int_{\mathcal{D}_d}(1+\|\xi\|^q)^2\|\mathcal{F}_h(V)(\xi)\|^2 d\xi.$$

Remarque 2.1 *Nous avons l'habitude de considérer des normes de Sobolev avec* $(1+\|\xi\|^2)^q$*, ici, nous considérerons plutôt* $(1+\|\xi\|^q)^2$ *afin de simplifier les calculs.*

Nous avons alors :

- $\|V\|_{h,q} = \|SV\|_{H^q(\mathbb{R}^d)}$
- $\widetilde{H^q}(G) = \{V \in L^2(G),\, SV \in H^q(\mathbb{R}^d)\}$.

De plus, les normes $\|.\|_{h,q}$ et $|||.|||_{h,q}$ sont équivalentes, en effet, d'après le lemme précédent :

$$\mathcal{F}_h(D_1^{q_1}\dots D_d^{q_d}V)(\xi) \;=\; e^{i\sum_{j=1}^d (h_j\xi_j q_j)/2}\prod_{j=1}^d\left(\frac{\sin(h_j\xi_j/2)}{h_j}\right)^{q_j}\mathcal{F}_h(V)(\xi)$$

Donc :

$$\|D_1^{q_1}\dots D_d^{q_d}V\|_h^2 = \frac{1}{\prod_{j=1}^d h_j^{2q_j}}\int_{\mathcal{D}_d}\prod_{l=1}^d\left(\sin\frac{h_l\xi_l}{2}\right)^{2q_l}\|\mathcal{F}_h(V)(\xi)\|^2 d\xi.$$

Or, pour $x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $\frac{2}{\pi} \leq \frac{|\sin x|}{x} \leq 1$. Ainsi :

$$\int_{\mathcal{D}_d} \prod_{l=1}^{d} \frac{1}{\pi^{2q_l}} \|\xi\|^{2q_l} \|\mathcal{F}_h(V)(\xi)\|^2 d\xi \quad \leq \quad \|D_1^{q_1} \dots D_d^{q_d} V\|_h^2 \leq \int_{\mathcal{D}_d} \prod_{l=1}^{d} \|\xi\|^{2q_l} \|\mathcal{F}_h(V)(\xi)\|^2 d\xi,$$

d'où l'équivalence des deux normes.

Nous avons aussi le lemme suivant :

Lemme 2.2 *Soit $h_m = \min_{1 \leq j \leq d} h_j$ et $h_M = \max_{1 \leq j \leq d} h_j$, alors si $q_1 \leq q_2$, si $0 < h_M \leq h_0$, nous avons :*

$$\|V\|_{h,q_1} \leq \|V\|_{h,q_2} \leq C_{q_1,q_2}(h_0) \left(\sum_{j=1}^{d} \frac{1}{h_j} \right)^{q_2 - q_1} \|V\|_{h,q_1} \leq \frac{C'_{q_1,q_2}(h_0)}{h_m^{q_2-q_1}} \|V\|_{h,q_1}.$$

En particulier, pour $q_1 = 0$ et $q \geq 0$:

$$\|V\|_h \leq \|V\|_{h,q} \leq C_{0,q}(h_0) \left(\sum_{j=1}^{d} \frac{1}{h_j} \right)^{q} \|V\|_h \leq \frac{C'_{0,q}(h_0)}{h_m^{q}} \|V\|_h.$$

Ce lemme montre qu'à h fixé, les espaces $H_{q_1}(G)$ et $H_{q_2}(G)$ sont identiques. Toutefois lorsque les pas tendent vers 0, ces espaces sont différents.

2.1.4 Troncature

Si $U \in L^2(\mathbb{R}^d)$, nous définissons sa troncature $TU \in L^2(G)$ par :

$$(TU)_m = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}_d} e^{i \sum_{j=1}^{d} m_j h_j \xi_j} \mathcal{F}U(\xi) d\xi.$$

On a alors :

$$\mathcal{F}_h(TU)(\xi) = \mathcal{F}U(\xi) \quad \textit{pour} \quad \xi \in \mathcal{D}_d.$$

2.1.5 Evaluation

Si $U \in \mathcal{C}^0(\mathbb{R}^d)$, nous définissons son évaluée par :

$$(EU)_m = U(m_1 h_1, \dots, m_d h_d).$$

On montre dans [38] l'identité :

$$\mathcal{F}_h(EU)(\xi) = \sum_{l \in \mathbb{Z}^d} \mathcal{F}U(\xi_1 + \frac{2\pi l_1}{h_1}, \dots, \xi_d + \frac{2\pi l_d}{h_d}).$$

Lemme 2.3 *Si $U \in H^{q+\alpha}(\mathbb{R}^d)$, $q+\alpha > \frac{d}{2}$ alors :*

$$\|EU - TU\|_{h,q} \leq C_{q,\alpha} \frac{h_M^{q+\alpha}}{h_m^q} \|U\|_{H^{q+\alpha}(\mathbb{R}^d)}.$$

PREUVE : La preuve est une extension à H^q de celle de [38]. Posons $\mathbb{Z}_*^d = \mathbb{Z}^d \setminus (0,\ldots,0)$ et $\xi_{l/h} = \xi + \frac{2\pi l}{h}$.

$$\begin{aligned}
\|EU - TU\|_{h,q}^2 &= \int_{\mathcal{D}_d} (1+\|\xi\|^q)^2 \|\mathcal{F}_h(EU)(\xi) - \mathcal{F}_h(TU)(\xi)\|^2 d\xi \\
&= \int_{\mathcal{D}_d} (1+\|\xi\|^q)^2 \left\| \sum_{l\in\mathbb{Z}_*^d} \hat{U}(\xi_{l/h}) \right\|^2 d\xi
\end{aligned}$$

En majorant $(1+\|\xi\|^q)^2$ par Ch_m^{-2q}, puis en utilisant l'inégalité de Cauchy-Schwartz, nous obtenons :

$$\begin{aligned}
\|EU - TU\|_{h,q}^2 &\leq \frac{C}{h_m^{2q}} \int_{\mathcal{D}_d} \left(\sum_{l\in\mathbb{Z}_*^d} \|\xi_{l/h}\|^{-2(q+\alpha)} \right) \left(\sum_{l\in\mathbb{Z}_*^d} \|\hat{U}(\xi_{l/h})\|^2 \|\xi_{l/h}\|^{2(q+\alpha)} \right) d\xi \\
&\leq \frac{C}{h_m^{2q}} \max_{\xi\in\mathcal{D}_d} \left(\sum_{l\in\mathbb{Z}_*^d} \|\xi_{l/h}\|^{-2(q+\alpha)} \right) \int_{\mathcal{D}_d} \sum_{l\in\mathbb{Z}_*^d} \|\hat{U}(\xi_{l/h})\|^2 \|\xi_{l/h}\|^{2(q+\alpha)} d\xi \\
&\leq \frac{C}{h_m^{2q}} \max_{\xi\in\mathcal{D}_d} \left(\sum_{l\in\mathbb{Z}_*^d} \|\xi_{l/h}\|^{-2(q+\alpha)} \right) \sum_{l\in\mathbb{Z}_*^d} \int_{\mathcal{D}_d} \|\hat{U}(\xi_{l/h})\|^2 \|\xi_{l/h}\|^{2(q+\alpha)} d\xi.
\end{aligned}$$

Nous remarquons maintenant que :

$$\begin{aligned}
\sum_{l\in\mathbb{Z}_*^d} \int_{\mathcal{D}_d} \|\hat{U}(\xi_{l/h})\|^2 \|\xi_{l/h}\|^{2(q+\alpha)} d\xi &= \sum_{l\in\mathbb{Z}_*^d} \int_{\mathcal{D}_d + \frac{2\pi l}{h}} \|\hat{U}(\xi)\|^2 \|\xi\|^{2(q+\alpha)} d\xi \\
&= \int_{\mathbb{R}^d \setminus \mathcal{D}_d} \|\hat{U}(\xi)\|^2 \|\xi\|^{2(q+\alpha)} d\xi \\
&\leq \|U\|_{H^{q+\alpha}(\mathbb{R}^d)}.
\end{aligned}$$

Pour achever la démonstration, il suffit de prouver que $\sum_{l\in\mathbb{Z}_*^d} \|\xi_{l/h}\|^{-2(q+\alpha)} \leq Ch_M^{q+\alpha}$. Pour cela, nous allons utiliser le lemme suivant :

Lemme 2.4 *Pour tous $l = (l_1,\ldots,l_d) \in \mathbb{N}^d$, $\xi \in \mathcal{D}_d$, nous avons :*

$$\|\xi_{l/h}\| \geq \frac{\pi \|l\|}{h_M}.$$

PREUVE : Nous avons, si $\xi \in \mathcal{D}_d$:

$$\frac{(2l_j-1)\pi}{h_j} \leq \xi_j + \frac{2\pi l_j}{h_j} \leq \frac{(2l_j+1)\pi}{h_j}.$$

Donc, si $l_j \neq 0$,

$$\begin{aligned} |\xi_j + \frac{2\pi l_j}{h_j}| &\geq \frac{\pi^2}{h_j^2}\min((2l_j-1),(2l_j+1)) \\ &\geq \frac{\pi^2}{h_j^2}(2|l_j|-1)^2 \\ &\geq \frac{\pi^2}{h_j^2}|l_j|^2. \end{aligned}$$

Cette inégalité étant aussi vraie pour $l_j = 0$, on a donc :

$$\left\|\xi_{l/h}\right\|^2 \geq \pi^2 \sum_{j=1}^{d} \frac{|l_j|^2}{h_j^2} \geq \frac{\pi^2}{h_M^2}\|l\|^2.$$

□

Nous reprenons la démonstration du lemme 2.3. Nous avons alors :

$$\sum_{l\in\mathbb{Z}_*^d} \|\xi_{l/h}\|^{-2(q+\alpha)} \leq \sum_{l\in\mathbb{Z}_*^d} \left(\frac{h_M}{\pi\|l\|}\right)^{2(q+\alpha)} \leq C h_M^{2(q+\alpha)},$$

car $q+\alpha > \frac{d}{2}$. Donc :

$$\|EU - TU\|_{h,q}^2 \leq C\frac{h_M^{2(q+\alpha)}}{h_m^{2q}}\|U\|_{H^{q+\alpha}}.$$

□

Lemme 2.5 *Si $u \in H^{q+\alpha}(\mathbb{R}^d)$, $q+\alpha > \frac{d}{2}$ alors :*

$$\|Eu\|_{h,q} \leq \left(1 + C_{q,\alpha}\frac{h_M^{\alpha+q}}{h_m^q}\right)\|u\|_{H^{q+\alpha}(\mathbb{R}^d)}.$$

PREUVE : Nous écrivons

$$\|Tu\|_{h,q}^2 = \int_{\mathcal{D}_d} (1+\|\xi\|^q)^2 \|\hat{u}\|^2 d\xi \leq \|u\|_{H^{q+\alpha}(\mathbb{R}^d)}^2,$$

et nous concluons en utilisant le lemme précédent. □

2.1.6 Normes d'applications linéaires

Nous définissons, pour T un opérateur linéaire, les normes d'opérateurs linéaires suivantes :

• Si $T : L^2(h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z}) \to L^2(h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z})$, on pose :

$$\|T\|_h = \sup_{v \neq 0} \frac{\|Tv\|_h}{\|v\|_h}.$$

• Si $T : H^q(h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z}) \to L^2(h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z})$, on pose :

$$\|T\|_{h,q} = \sup_{v \neq 0} \frac{\|Tv\|_h}{\|v\|_{h,q}}.$$

• Si $T : H^q(h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z}) \to H^q(h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z})$, on pose :

$$\|T\|_{h,qq} = \sup_{v \neq 0} \frac{\|Tv\|_{h,q}}{\|v\|_{h,q}}.$$

2.2 Définitions

Nous considérons le problème de Cauchy :

$$\begin{cases} \partial_t U = P(t, x, \partial_x)U, \\ U(0, .) = U^0, \end{cases} \tag{2.1}$$

avec $t > 0$, $x \in \mathbb{R}^s$, $U \in \mathbb{R}^N$. Nous considérons le schéma à $q+1$ pas en temps :

$$Q_{-1}V^{n+1} = \sum_{\sigma=0}^{q} Q_\sigma V^{n-\sigma}, \tag{2.2}$$

où Q_{-1} est inversible et pour tout $\sigma \in \{-1, \ldots, q\}$, Q_σ est à coefficients réguliers et bornés par rapport aux variables x et t :

$$(Q_\sigma V)_j = \sum_{-r \leq \nu_j \leq p} A_{\nu,\sigma}(x_{j-\nu}, t_{n-\sigma}, h, k)V_{j-\nu} \text{ pour } \sigma \in \{-1, \ldots, q\}.$$

Dans le cas où le problème de Cauchy est fortement bien posé, le théorème de Lax-Richtmyer affirme que si le schéma est stable et consistant alors il est convergent et que si le schéma est convergent, alors il est stable.

Nous allons montrer l'analogue de ce théorème dans le cas des problèmes faiblement bien posés. Pour cela, nous allons introduire des nouvelles notions de stabilité, consistance et convergence qui feront intervenir un défaut.

Nous noterons k le pas de temps, $h = (h_1, \ldots, h_d)$ les pas d'espace et $h_M = \max_{1\leq j\leq d} h_j$, $h_m = \min_{1\leq j\leq d} h_j$.

Nous rappelons la définition d'un problème faiblement bien posé :

Définition 2.1 *Le problème de Cauchy est faiblement bien posé s'il existe $q_1 > 0$, $K > 0$, $\alpha \in \mathbb{R}$ tels que pour toute donnée initiale $U^0 \in H^q(\mathbb{R}^d)$ avec $q \geq q_1$, le problème a une unique solution $U \in \mathcal{C}^0(\mathbb{R}^+, H^{q-q_1}(\mathbb{R}^d))$, vérifiant, $\forall t \geq 0$:*

$$\|U(t,.)\|_{H^{q-q_1}(\mathbb{R}^d)} \leq K_C e^{\alpha_C t} \|U^0\|_{H^q(\mathbb{R}^d)}.$$

Nous définissons l'opérateur S_h par :

$${}^t(V^{n+q}, \ldots, V^n) = S_h(t_n, t_\nu)\,{}^t(V^{\nu+q}, \ldots, V^\nu).$$

Nous avons alors :

$$S_h(t_n, t_\nu) = \prod_{\mu=1}^{n-\nu} \begin{pmatrix} M_0 & M_1 & \ldots & M_{q-1} & M_q \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix},$$

où $M_\sigma = Q_{-1}^{-1}(t_{n+q+1-\mu}) Q_\sigma(t_{n+q+1-\mu})$.

2.2.1 Stabilité

Nous donnons une définition de la stabilité qui, de manière analogue à la définition des problèmes faiblement bien posés, prend en compte la perte de régularité mais, cette fois, au sens de la norme de Sobolev discrète.

Définition 2.2 *Le schéma est dit* **faiblement stable** *s'il existe $q_2 > 0$, h_0, $k_0 > 0$, α_S, K_S, $C_S \in \mathbb{R}$ tels que, $\forall h$, $0 < h_M \leq h_0$, $\forall k$, $0 < k \leq k_0$, $\forall n$, $0 \leq nk \leq T$, $\forall \nu \leq n$, on a :*

$$\|Q_{-1}^{-1}\|_{h,q_2 q_2} \leq C_S \quad et \quad \|S_h(t_n, t_\nu)\|_{h,q_2} \leq K_S e^{\alpha_S (t_n - t_\nu)}.$$

Si q_2 est le plus petit entier vérifiant cette propriété, nous dirons alors que le schéma est faiblement stable de défaut q_2.

Si le schéma est faiblement stable de défaut q_2, nous avons, par utilisation de la définition :

$$\|V^n\|_h \leq K_S e^{\alpha_S t_n} \|(V^q, \ldots, V^0)\|_{h,q_2}.$$

Dans le cas où $q_2 = 0$, on retrouve la définition de la stabilité.

2.2.2 Consistance

Pour les problèmes faiblement bien posés, la régularité de la donnée initiale est fondamentale. La définition de la consistance que nous proposons en tiendra donc compte.

Définition 2.3 *Si U est la solution continue de (2.1), nous définissons l'***erreur de troncature** $\tau^n(U^0)$, *pour $n \geq q$, par :*

$$k\tau_j^n(U^0) = Q_{-1}U(t_{n+1}, x_j) - \sum_{\sigma=0}^{q} Q_\sigma U(t_{n-\sigma}, x_j).$$

Le schéma est dit **consistant** *s'il existe $q_3 > 0$ et $\theta \in \mathbb{R}$ tels que, pour tout $U^0 \in H^\theta(\mathbb{R}^s)$ et pour tout $n \geq 0$:*

$$\|\tau^n(U^0)\|_{h,q_3} \leq L(t_n)\varepsilon(h_M, k),$$

où $L(t_n)$ est borné sur tout intervalle fini et $\lim_{h_M,k\to 0} \varepsilon(h_M, k) = 0$.
On dit alors que le schéma est q_3-consistant de régularité θ.

Remarque 2.2 *Nous pouvons définir de même τ^n pour $n \leq q-1$ en utilisant le schéma d'initialisation.*

Remarque 2.3 *Dans tout le texte, (α_C, K_C) et (C_S, K_S, α_S) sont des noms de constantes affectés aux définitions de faiblement bien posé et de stabilité. Toutes les autres constantes interviennent au cours de calculs et sont, sans autre mention, données par la phrase "il existe une constante telle que".*

2.2.3 Convergence

La notion de convergence d'un schéma pour un problème faiblement bien posé ne peut pas avoir comme seule hypothèse la convergence en norme L^2 de la donnée initiale du schéma vers la condition initiale du problème continu car, pour prendre en compte la perte de régularité, il fait une convergence plus forte, c'est-à-dire, en une certaine norme de Sobolev.

Définition 2.4 *Le schéma est dit* **convergent** *s'il existe q_4 tel que pour tout $U^0 \in H^{q_4}(\mathbb{R}^d)$, pour tout V^σ, $0 \leq \sigma \leq q$ tel que $\lim_{h_M,k\to 0} \|SV^\sigma - U(\sigma k, .)\|_{H^{q_4}(\mathbb{R}^d)} = 0$, on a :*

$$\lim_{(h_M,k)\to 0} \|U(t_n, .) - SV^n\|_{L^2(\mathbb{R}^d)} = 0.$$

On dit alors que le schéma est q_4-convergent.

2.3 Extension du théorème de Lax Richtmyer

Nous allons, dans cette partie, étendre les démonstrations de Kreiss du théorème de Lax-Richtmyer au cas des problèmes faiblement bien posés.

Dans le cas des problèmes faiblement bien posés, nous voulons connaître la régularité des conditions initiales qui conduisent à un schéma convergent. C'est pourquoi, dans l'extension du théorème de Lax-Richtmyer, nous allons faite intervenir les différents défauts que nous avons définis précédemment.

Etant donné la complexité des notations et des preuves, nous séparons les énoncés en condition nécessaire et condition suffisante.

2.3.1 Condition suffisante de convergence

Théorème 2.1 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, si le schéma est faiblement stable de défaut q_2 et q_3-consistant de régularité θ avec $q_3 \geq q_2 > \frac{d}{2}$, si de plus $(\frac{h_M}{h_m})$ est borné et si, dans le cas $q \neq 0$, $q_4 \geq q_1 + q_2$, alors le schéma est q_4- convergent avec $q_4 > \max(\frac{d}{2} + q_1, q_2)$ et $q_4 \geq \theta$.*

PREUVE : Soit $U^0 \in H^{q_4}(\mathbb{R}^d)$ et V^n solution du schéma de condition initiale $(V^0, \ldots, V^q)$ qui est telle que $\lim_{h_M,k\to 0} \|SV^\sigma - U(\sigma k, .)\|_{H^{q_4}(\mathbb{R}^d)} = 0$. Nous allons étudier $U(t_n, .) - SV^n$.
Nous commençons par évaluer $W^n = EU(t_n, .) - V^n$ qui a un sens car $U(t, .) \in H^{q_4-q_1}(\mathbb{R})$. En effet, le problème est faiblement bien posé de défaut q_1 et comme $q_4 - q_1 > \frac{d}{2}$, $U(t, .) \in \mathcal{C}^0(\mathbb{R}^d)$. Posons :

$$\mathbf{W}^n = {}^t (W^{n+q}, \ldots, W^n)$$

$$\text{et } \mathbf{T}_n = {}^t (Q_{-1}^{-1}(t_{n+q})\tau^{n+q}, \ldots, Q_{-1}^{-1}(t_n)\tau^n),$$

où, pour alléger les notations, $\tau^n = \tau^n(U^0)$. Nous avons alors :

$$\mathbf{W}^{n+1} = S_h(t_{n+1}, t_n)\mathbf{W}^n + k\mathbf{T}_n,$$

d'où :

$$\mathbf{W}^n = S_h(t_n, 0)\mathbf{W}^0 + k\sum_{\nu=0}^{n-1} S_h(t_n, t_{\nu+1})\mathbf{T}_\nu.$$

Nous allons d'abord montrer que W^n tend vers 0.

- Comme le schéma est q_3-consistant, $\tau^n \in H^{q_3}(G)$ donc $\tau^n \in H^{q_2}(G)$. Nous avons alors, par définition de la stabilité :

$$\|S_h(t_n, t_{\nu+1})\mathbf{T}_\nu\|_h \leq C_S K_S e^{\alpha_S(t_n - t_{\nu+1})} \|{}^t(\tau^{\nu+q}, \ldots, \tau^\nu)\|_{h,q_2}.$$

- Pour $0 \leq \sigma \leq q$, nous avons $W^\sigma = EU(t_\sigma, .) - V^\sigma$, or $V^\sigma \in L^2(G)$, $U^0 \in H^{q_4}$, $q_4 - q_1 > \frac{d}{2}$ donc $W^0 \in L^2(G) = H^{q_2}(G)$ (à h fixé). Donc, par définition de la stabilité :
$$\|S_h(t_n, 0)\mathbf{W}^0\|_h \leq K_S e^{\alpha_S t_n} \|\mathbf{W}^0\|_{h,q_2}.$$
Ainsi :
$$\|\mathbf{W}^n\|_h \leq K_S e^{\alpha_S t_n} \|\mathbf{W}^0\|_{h,q_2} + kCK_S \sum_{\nu=0}^{n-1} e^{\alpha_S (t_n - t_{\nu+1})} \|^t(\tau^{\nu+q}, \ldots, \tau^\nu)\|_{h,q_2}. \tag{2.3}$$
Pour montrer que $\|W^n\|_h$ tend vers 0, nous allons analyser successivement les deux termes.
- Montrons que $\lim_{h_M \to 0} \|\mathbf{W}^0\|_{h,q_2} = 0$.
On a, pour $0 \leq \sigma \leq q$:
$$\begin{aligned} \|W^\sigma\|_{h,q_2} &= \|EU(t_\sigma, .) - V^\sigma\|_{h,q_2} \\ &= \|SEU(t_\sigma, .) - SV^\sigma\|_{H^{q_2}} \\ &\leq \|SEU(t_\sigma, .) - U(t_\sigma, .)\|_{H^{q_2}} + \|U(t_\sigma, .) - SV^\sigma\|_{H^{q_2}}. \end{aligned}$$
Comme $q_2 \leq q_4$, nous avons, par hypothèse, $\lim_{h_M \to 0} \|U(t_\sigma, .) - SV^\sigma\|_{H^{q_2}} = 0$.
D'autre part :
$$\|SEU(t_\sigma, .) - U(t_\sigma, .)\|_{H^{q_2}} = \int_{\mathcal{D}_d} \|(\widehat{EU(t_\sigma)} - \widehat{U}(t_\sigma, .))(\xi)\|^2 (1 + \|\xi\|^2)^{q_2} d\xi$$
$$+ \int_{\mathbb{R} \setminus \mathcal{D}_d} \|\widehat{U}(t_\sigma, .)(\xi)\|^2 (1 + \|\xi\|^2)^{q_2} d\xi.$$
Comme $q_4 \geq q_1 + q_2$ par hypothèse, $U(t_\sigma) \in H^{q_4 - q_1} \subset H^{q_2}$ (si $q = 0$ $U(t_\sigma) \in H^{q_4}$), nous avons, par définition de l'intégrale de Lebesgue :
$$\lim_{h_M \to 0} \int_{\mathbb{R} \setminus \mathcal{D}_d} \|\widehat{U}(t_\sigma, .)(\xi)\|^2 (1 + \|\xi\|^2)^{q_2} d\xi = 0,$$
et, nous avons aussi :
$$\begin{aligned} \left(\int_{\mathcal{D}_d} \left\| (\widehat{EU}(t_\sigma) - \hat{U}(t_\sigma))(\xi) \right\|^2 (1 + \|\xi\|^2)^{q_2} d\xi \right)^{\frac{1}{2}} & \\ \leq \|EU(t_\sigma) - TU(t_\sigma)\|_{H^{q_2}} & \\ \leq C_{q_2,\alpha} h^\alpha \|U(t_\sigma)\|_{H^{q_2+\alpha}} \text{ si } q_2 + \alpha > \frac{d}{2}. & \end{aligned}$$

Nous prenons $\alpha > 0$ tel que $\frac{d}{2} < q_2 + \alpha < q_4$, alors, nous avons la convergence de l'estimation précédente vers 0. Donc :
$$\lim_{h_M \to 0} \|\mathbf{W}^0\|_{h,q_2} = 0.$$

- Montrons que $\lim_{h\to 0} \|\mathbf{W}^n\|_{h,q_2} = 0$.
  D'après l'équation (2.3), il suffit de montrer maintenant que :

$$\lim_{h_M\to 0} k\sum_{\nu=0}^{n-1} e^{\alpha_S(t_n-t_{\nu+1})}\|^t(\tau^{\nu+q},\ldots,\tau^{\nu})\|_{h,q_2} = 0,$$

  or, comme $q_2 \leq q_3$, nous avons :

$$k\sum_{\nu=0}^{n-1} e^{\alpha_S(t_n-t_{\nu+1})}\|^t(\tau^{\nu+q},\ldots,\tau^{\nu})\|_{h,q_2} \leq t_n e^{\alpha_S t_n} \sup_{\nu=0,..,n+q-1} \|\tau^{\nu}\|_{h,q_3}.$$

  Et comme $U^0 \in H^{q_4}(\mathbb{R}^d) \subset H^{\theta}(\mathbb{R}^d)$, nous avons, par définition :

$$\sup_{\nu=0,..,n+q-1} \|\tau^{\nu}\|_{h,q_3} \leq L(t_{n+q-1})\varepsilon(h_M,k).$$

  Donc

$$\lim_{h_M\to 0} \sup_{\nu=0,..,n+q-1} \|\tau^{\nu}\|_{h,q_3} = 0.$$

  Ainsi :

$$\lim_{h_M\to 0} \|\mathbf{W}^n\|_{h,q_2} = 0.$$

  Nous allons maintenant montrer que $U(t_n,.) - SV^n$ tend vers 0.
- Nous avons :

$$\|U(t_n,.)-SV^n\|_{L^2(\mathbb{R}^d)} \leq \|U(t_n,.)-SEU(t_n,.)\|_{L^2(\mathbb{R}^d)}+\|SEu(t_n,.)-SV^n\|_{L^2(\mathbb{R}^d)},$$

  or :

$$\|SEU(t_n,.) - SV^n\|_{L^2(\mathbb{R}^d)} = \|EU(t_n,.) - V^n\|_h = \|W^n\|_h \leq \|\mathbf{W}^n\|_h \to 0,$$

  et :

$$\begin{aligned}\|U(t_n,.) - SEU(t_n,.)\|^2_{L^2(\mathbb{R}^d)} &\leq \|TU(t_n,.) - EUu(t_n,.)\|^2_h \\ &\quad +\|\hat{U}(t_n,.)\|^2_{L^2(\mathbb{R}^d\setminus\mathcal{D}_d)} \\ &\leq (C_{0,d/2+\varepsilon}h_M^{d/2+\varepsilon}\|U(t_n,.)\|_{H^{d/2+\varepsilon}})^2 \\ &\quad +\|\hat{U}(t_n,.)\|^2_{L^2(\mathbb{R}^d\setminus\mathcal{D}_d)},\end{aligned}$$

  pour $\varepsilon > 0$, en utilisant le lemme 2.3. Cette dernière quantité tend vers 0 car $U(t_n,.) \in H^{q_4-q_1}$ et $q_4 - q_1 > \frac{d}{2}$. Ainsi :

$$\lim_{h\to 0} \|U(t_n,.) - SV^n\|_{L^2(\mathbb{R}^d)} = 0,$$

  et nous avons prouvé le théorème.

□

2.3.2 Condition nécessaire de convergence

Théorème 2.2 *Si le problème est faiblement bien posé de défaut q_1 et si le schéma est q_4-convergent, et si, de plus, $\|Q_{-1}^{-1}\|_{h,q_2q_2}$ est borné, alors, il est stable de défaut $q_2 \geq \max(q_1, q_4)$ ($q_2 \geq q_1$ si $q = 0$).*

PREUVE : Nous allons raisonner par l'absurde. Supposons le schéma q_4-convergent et non stable de défaut $q_2 \geq \max(q_1, q_4)$, c'est-à-dire :

$$\forall q_2 \geq \max(q_1, q_2),\ \forall h_0,\ k_0 > 0,\ \forall \alpha, K \in \mathbb{R},$$

$$\exists h \leq h_0, \exists k \leq k_0, \exists n, \exists \nu, \|S_h(t_n, t_\nu)\|_{h,q_2} > K e^{\alpha(t_n - t_\nu)}.$$

Donc, nous pouvons construire les suites suivantes : $h_m \to 0$, $k_m \to 0$, $K_m \to +\infty$, n_m, ν_m, telles que $n_m k_m$, $\nu_m k_m \in [0, T]$, et $\mathbf{F}_m$ dans H^{q_2} tel que $\|\mathbf{F}_m\|_{h,q_2} = 1$ vérifiant :

$$\|S_h(t_{n_m}, t_{\nu_m})\mathbf{F}_m\|_h > \frac{1}{2} K_m e^{\alpha_S(t_{n_m} - t_{\nu_m})}.$$

Quitte à démarrer le schéma au pas ν_m, nous pouvons supposer $\nu_m = 0$, on a alors :

$$\|S_h(t_{n_m}, 0)\mathbf{F}_m\|_h > \frac{1}{2} K_m.$$

Soit V_m^n la solution du schéma de condition initiale $\mathbf{V}_m^0 = \frac{\mathbf{F}_m}{K_m}$, alors :

$$\|\mathbf{V}_m^{n_m}\|_h = \left\| S_h(t_{n_m}, 0) \frac{\mathbf{F}_m}{K_m} \right\|_h > \frac{1}{2}.$$

Soit U_m la solution du problème de Cauchy de condition initiale $\frac{(S\mathbf{F}_m)_{q+1}}{K_m}$, alors, par définition du défaut, nous avons pour $0 \leq k \leq q_1$:

$$\begin{aligned}
\|U_m(t, .)\|_{H^k} &\leq K_C e^{\alpha_C t} \left\| \frac{S\mathbf{F_m}}{K_m} \right\|_{H^{q_1 - k}} \\
&\leq \frac{K_C e^{\alpha_C t}}{K_m} \|\mathbf{F}_m\|_{h, q_1 - k} \\
&\leq \frac{K_C e^{\alpha_C t}}{K_m} \|\mathbf{F}_m\|_{h,q_2} \text{ si } q_2 \geq q_1 - k \\
&\leq \frac{K_C e^{\alpha_C t}}{K_m},
\end{aligned}$$

donc, en prenant $k = 0$:

$$\lim_{m \to +\infty} \|U_m(t, .)\|_{L^2} = 0.$$

Or nous avons, pour $0 \leq \sigma \leq q$:

$$\|SV_m^\sigma - U_m(t_\sigma, .)\|_{H^{q_4}} \leq \|SV_m^\sigma\|_{H^{q_4}} + \|U_m(t_\sigma, .)\|_{H^{q_4}} ,$$

(dans le cas où $q = 0$, nous avons $\|SV_m^\sigma - U_m(t_\sigma, .)\|_{H^{q_4}} = 0$) or si $q \neq 0$, comme $q_2 \geq q_1 - q_4$, nous avons, d'après ce qui précède $\|U_m(t_\sigma, .)\|_{H^{q_4}} \to 0$. De plus,

$$\|SV_m^\sigma\|_{H^{q_4}} \leq \left\|S\mathbf{V}_m^0\right\|_{H^{q_2}} \to 0,$$

donc, par q_4-convergence :

$$\lim_{h \to 0} \|U_m(t_{n_m}, .) - SV_m^{n_m}\|_{L^2} = 0,$$

or nous avons aussi :

$$\begin{aligned} \|U_m(t_{n_m}, .) - SV_m^{n_m}\|_{L^2} &\geq \|SV_m^{n_m}\|_{L^2} - \|U_m(t_{n_m}, .)\|_{L^2} \\ &> \frac{1}{2} - \|U_m(t_n, .)\|_{L^2} \\ &> \frac{1}{4} \text{ pour } m \text{ assez grand ,} \end{aligned}$$

d'où une contradiction et la preuve du théorème. □

Chapitre 3

Etude des schémas à un pas pour des équations à coefficients constants

Dans ce chapitre, nous étudions le cas particulier des schémas pour des équations à coefficients constants. Les schémas considérés sont alors, eux aussi, à coefficients constants. Dans ce cas, l'utilisation de la transformée de Fourier discrète permet de se ramener à une étude matricielle. En effet, nous pouvons étudier les notions de stabilité, consistance et convergence grâce à la matrice d'amplification du schéma. De plus, nous préciserons la notion de convergence en évaluant le taux de convergence.

Les techniques utilisées ici sont proches de celles de [38] et [27], mais le fait que le problème ne soit plus fortement bien posé mais uniquement faiblement bien posé pose de nombreuses difficultés supplémentaires. En effet, la perte de régularité entraîne l'introduction d'un facteur polynomial en ξ qui pose problème lorsque l'on étudie des puissances $n^{\text{ièmes}}$. De plus, nous étudions des systèmes matriciels, nous devons donc gérer la non commutativité du symbole principal et du terme d'ordre 0.

Nous considérons le schéma à un pas, à coefficients constants, suivant, où pour $\sigma = 0, -1$, on écrit :

$$Q_{-1}V^{n+1} = Q_0 V^n, \tag{3.1}$$

avec :

$$Q_\sigma = \sum_{-r \leq \nu_j \leq p} A_{\sigma,\nu} F^\nu,$$

où $A_{\sigma,\nu}$ est indépendant de x_j et de t_n et F^ν est défini par :

$$F^\nu V = (V_{m+\nu})_{m \in \mathbb{Z}^d}.$$

Le symbole de Q_σ est donné par :

$$\widehat{Q_\sigma}(\xi) = \sum_{-r \leq \nu_j \leq p} A_{\sigma,\nu} e^{i\nu\xi.h}.$$

Nous pouvons alors effectuer la transformation de Fourier discrète de 3.1 :

$$\widehat{Q_{-1}}\widehat{V^{n+1}}(\xi) = \widehat{Q_0}\widehat{V^n}(\xi).$$

Nous poserons toujours :

$$\widehat{Q} = \widehat{Q_{-1}}^{-1}\widehat{Q_0}.$$

3.1 Interprétation des définitions en termes matriciels

3.1.1 Caractérisations de la stabilité

Nous étudierons deux caractérisations de la stabilité. La première portera sur la matrice d'amplification et la seconde sur ses valeurs propres.

La première caractérisation est l'étude des puissances $n^{\text{ièmes}}$ de la matrice d'amplification. Il s'agit de l'analogue discret de l'étude de $\|e^{P(i\xi)t}\|$ utilisé pour déterminer si un problème continu est faiblement bien posé ou non.

Proposition 3.1 *Le schéma est faiblement stable de défaut q_2 si et seulement si on a les deux propriétés suivantes :*

$$\forall \xi \in \mathcal{D}_d,\ \|\widehat{Q_{-1}}^{-1}(\xi)\|_2 \leq C_S \tag{3.2}$$

et

$$\forall n,\ t_n = nk \in [0,T], \forall \xi \in \mathcal{D}_d, \|\widehat{Q}^n(\xi)\| \leq K_S e^{\alpha_S t_n}(1 + \|\xi\|^{q_2}). \tag{3.3}$$

Preuve :

- Supposons que les propriétés (3.2) et (3.3) sont vérifiées, et estimons la norme de V^n solution du problème discret 3.1 :

$$\begin{aligned} \|V^n\|_h^2 &= \int_{\mathcal{D}_d} \|\widehat{Q}^n(\xi)\widehat{V^0}(\xi)\|^2 d\xi \\ &\leq K_S^2 e^{2\alpha_S t_n} \int_{\mathcal{D}_d} (1 + \|\xi\|^{q_2})^2 \|\widehat{V^0}(\xi)\|^2 d\xi. \end{aligned}$$

Donc :

$$\|V^n\|_h \leq K_S e^{\alpha_S t_n} \|V^0\|_{h,q_2}.$$

De plus, nous avons, pour tout $v \in H^{q_2}(G)$:

$$\begin{aligned} \|Q_{-1}^{-1} v\|_{h,q_2}^2 &= \int_{\mathcal{D}_d} (1 + \|\xi\|^{q_2})^2 \|\widehat{Q_{-1}}^{-1}(\xi)\hat{v}(\xi)\|^2 d\xi \\ &\leq C_S^2 \|v\|_{h,q_2}^2. \end{aligned}$$

Donc :

$$\|Q_{-1}^{-1}\|_{h,q_2 q_2} \leq C_S.$$

Ainsi, le schéma est faiblement stable de défaut q_2.

- Réciproquement, supposons le schéma faiblement stable de défaut q_2.
  En utilisant la définition 2.2 de la stabilité, appliquée pour $\nu = 0$ et en remarquant que dans le cas des schémas à un pas $S_h(t_n, t_0) = Q^n$, en appliquant une transformation de Fourier discrète, nous avons, pour toute donnée initiale V^0 :

$$\int_{\mathcal{D}_d} \|\widehat{Q}^n \widehat{V^0}\|^2 d\xi \leq K_S^2 e^{2\alpha_S t_n} \int_{\mathcal{D}_d} (1 + \|\xi\|^{q_2})^2 \|\widehat{V^0}(\xi)\|^2 d\xi. \tag{3.4}$$

  – Montrons d'abord (3.3) par l'absurde.
  Supposons qu'il existe h_0, $\xi_0 \in \prod_{j=1}^d [-\frac{\pi}{(h_0)_j}, \frac{\pi}{(h_0)_j}]$, n_0, tels que :

$$\|\widehat{Q}^{n_0}(\xi_0)\|_2 > K_S e^{\alpha_S t_{n_0}} (1 + \|\xi_0\|^{q_2}). \tag{3.5}$$

  Pour tout $\varepsilon > 0$, il existe $W_0 \in \mathbb{C}^N$ de norme 1 tel que :

$$\|\widehat{Q}^{n_0}(\xi_0)\| \leq \|\widehat{Q}^{n_0}(\xi_0) W_0\| + \varepsilon. \tag{3.6}$$

  Pour tout $\mu > 0$, nous choisissons, comme condition initiale du schéma, V^0 tel que :

$$\widehat{V^0}(\xi) = W_0 \mathbf{1}_{\|\xi - \xi_0\| \leq \mu}.$$

  Nous avons alors, d'après 3.4 :

$$\frac{1}{2\mu} \int_{\|\xi - \xi_0\| \leq \mu} \|\widehat{Q}^n(\xi) W_0\|^2 d\xi \leq K_S^2 e^{2\alpha_S t_n} \frac{1}{2\mu} \int_{\|\xi - \xi_0\| \leq \mu} (1 + \|\xi\|^{q_2}) \|W_0\|^2 d\xi.$$

  Nous faisons tendre μ vers 0, nous obtenons alors :

$$\|\widehat{Q}^n(\xi_0) W_0\|^2 \leq K_S^2 e^{2\alpha_S t_n} (1 + \|\xi_0\|^{q_2}).$$

  Alors :

$$\|\widehat{Q}^n(\xi_0)\| - \varepsilon \leq K_S e^{\alpha_S t_n} (1 + \|\xi_0\|^{q_2}).$$

Nous choisissons :

$$\varepsilon = \frac{1}{2}(\|\widehat{Q}^{n_0}(\xi_0)\| - K_S e^{\alpha_S} t_{n_0}(1 + \|\xi_0\|^{q_2})).$$

Donc :

$$\|\widehat{Q}^{n}(\xi_0)\| \leq K_S e^{\alpha_S t_n}(1 + \|\xi_0\|^{q_2}).$$

Ce qui contredit 3.5, donc (3.3) est vérifiée.

– Montrons ensuite (3.2) par l'absurde.

Supposons qu'il existe h_0, $\xi_0 \in \prod_{j=1}^{d}[-\frac{\pi}{(h_0)_j}, \frac{\pi}{(h_0)_j}]$ tels que : $\|\widehat{Q_{-1}}^{-1}(\xi_0)\|_2 > C_S$.

Soit $W_0 \in \mathbb{C}^N$ de norme 1 tel que :

$$\|\widehat{Q_{-1}}^{-1}(\xi_0)W_0\| > C_S.$$

Pour tout $\mu > 0$, nous choisissons, comme condition initiale du schéma, V^0 tel que : $\widehat{V_0}(\xi) = W_0 \mathbf{1}_{\|\xi-\xi_0\|\leq\mu}$.

Alors, comme $\|\widehat{Q_{-1}}^{-1}\|_{h,q_2q_2} \leq C_S$, nous avons :

$$\frac{1}{2\mu}\int_{\|\xi-\xi_0\|\leq\mu}(1+\|\xi\|^{q_2})^2\|\widehat{Q_{-1}}^{-1}(\xi)W_0\|^2 d\xi \leq \frac{C_S}{2\mu}\int_{\|\xi-\xi_0\|\leq\mu}(1+\|\xi\|^{q_2})^2\|W_0\|^2 d\xi,$$

et μ tend vers 0 donne :

$$\|\widehat{Q_{-1}}^{-1}(\xi_0)W_0\| \leq C_S.$$

Ce qui est absurde donc (3.2) est vérifiée.

□

Remarque 3.1 *Lorsque q_2 désigne le défaut le meilleur possible, alors la dépendance polynomiale en ξ de $\|\widehat{Q}^n(\xi)\|$ ne peut pas être de degré inférieur à q_2 et réciproquement. La proposition précédente permet donc de calculer le défaut de manière optimale.*

L'avantage de la proposition précédente est qu'elle permet d'évaluer le défaut de stabilité q_2, toutefois elle est difficilement applicable. En effet, elle nécessite le calcul des puissances $n^{\text{ièmes}}$ qui est en général délicat. La seconde caractérisation que l'on donne n'apporte aucune indication sur la valeur de q_2 mais elle a l'avantage de ne porter que sur les valeurs propres.

Proposition 3.2 *On suppose que :*

$$\exists K_0 > 0,\ \exists \theta \in \mathbb{N},\ \forall \xi \in \mathcal{D}_d,\ \|\widehat{Q} - Id\| \leq K_0 k(1 + \|\xi\|)^{\theta}.$$

Alors le schéma est faiblement stable si et seulement si pour tout $\xi \in \mathcal{D}_d$, pour toute valeur propre $\lambda(\xi)$ de $\widehat{Q}$, on a $|\lambda(\xi)| \leq e^{\alpha_S k}$.

Remarque 3.2 *Dans le cas des problèmes fortement bien posés, la condition sur les valeurs propres n'est qu'une condition nécessaire de stabilité. Ici, dans le cas des problèmes faiblement bien posés, nous prouvons qu'elle est nécessaire et suffisante.*

PREUVE :

- Supposons le schéma faiblement stable. Alors :

$$\forall \xi \in \mathcal{D}_d,\ \|\widehat{Q}^n(\xi)\|_2 \leq K_S e^{\alpha_S t_n}(1+\|\xi\|^{q_2}).$$

Soit $\lambda(\xi)$ une valeur propre de $\widehat{Q}(\xi)$, alors :

$$|\lambda(\xi)|^n \leq \|\widehat{Q}^n(\xi)\|_2 \leq K_S e^{\alpha_S t_n}(1+\|\xi\|^{q_2}).$$

Donc

$$|\lambda(\xi)| \leq K_S^{1/n} e^{\alpha_S k}(1+\|\xi\|^{q_2})^{1/n}.$$

Or, pour ξ fixé, $\lim_{n\to+\infty} K_S^{1/n} = 1$ et $\lim_{n\to+\infty}(1+\|\xi\|^{q_2})^{1/n} = 1$.
Donc, en passant à la limite quand n tend vers l'infini, nous obtenons, pour tout $\xi \in \mathcal{D}_d$:

$$|\lambda(\xi)| \leq e^{\alpha_S k}.$$

- Réciproquement, supposons que pour tout $\xi \in \mathcal{D}_d$, $|\lambda| \leq e^{\alpha_S k}$. Nous effectuons la décomposition de Schur de $\widehat{Q}$ (voir [28]) :

$$\widehat{Q} = S^*(\xi)(\Lambda(\xi)+T(\xi))S(\xi),$$

où $\Lambda(\xi) = \begin{pmatrix} \lambda_1(\xi) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N(\xi) \end{pmatrix}$ et $T = \begin{pmatrix} 0 & \dots & t_{i,j}(\xi) \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix}$,
avec S unitaire.
Posons $V^n(\xi) = S(\xi)\widehat{U^n}(\xi)$, alors $V^{n+1}(\xi) = (\Lambda(\xi)+T(\xi))V^n(\xi)$. Nous avons alors :

$$\forall i,\ 1 \leq i \leq N,\ V_i^{n+1}(\xi) = \lambda_i(\xi)V_i^n(\xi) + \sum_{j=i+1}^{N} t_{i,j}(\xi)V_j^n(\xi),$$

donc

$$V_i^n(\xi) = \lambda_i^n(\xi)V_i^0(\xi) + \sum_{l=0}^{n-1} \lambda_i^{n-l-1}(\xi) \sum_{j=i+1}^{N} t_{i,j}(\xi)V_j^l(\xi).$$

Nous allons montrer le lemme calculatoire suivant :
Lemme 3.1

$$\forall i,\ 1 \leq i \leq N,\ \forall \xi \in \mathcal{D}_d,\ |V_i^n(\xi)| \leq C_i(1+\|\xi\|)^{(N-i)\theta} t_n^{N-i} e^{\alpha_S t_n} \max_{r\geq i} |V_r^0(\xi)|.$$

PREUVE : Nous utilisons la structure triangulaire de T et nous procédons par récurrence descendante sur i :

– Pour $i = N$, nous avons $V_N^n(\xi) = \lambda_N^n(\xi)V_N^0(\xi)$, donc d'après l'hypothèse $|V_N^n(\xi)| \leq e^{\alpha_S t_n}|V_N^0(\xi)|$, et $C_N = 1$ convient.

– Soit i tel que $1 \leq i \leq N-1$, supposons le résultat vrai pour tout $j \geq i+1$, nous avons alors :

$$\begin{aligned}
|V_i^n(\xi)| &\leq |\lambda_i(\xi)|^n|V_i^0(\xi)| + \sum_{l=0}^{n-1}|\lambda_i(\xi)|^{n-l-1}\sum_{j=i+1}^{N}|t_{i,j}(\xi)||V_j^l(\xi)| \\
&\leq e^{\alpha_S t_n}|V_i^0(\xi)| \\
&\quad + \sum_{l=0}^{n-1} e^{\alpha_S t_{n-l-1}}\sum_{j=i+1}^{N}|t_{i,j}(\xi)|C_j(1+\|\xi\|)^{(N-j)\theta}t_l^{N-j}e^{\alpha_S t_r}\max_{r\geq j}|V_r^0(\xi)|.
\end{aligned}$$

Nous allons évaluer $|t_{i,j}(\xi)|$ grâce à l'hypothèse sur $\|\widehat{Q}(\xi) - Id\|$. Comme sur $\mathbb{R}^N$, toutes les normes sont équivalentes, il existe $K_1 > 0$ telle que :

$$|t_{i,j}(\xi)| \leq K_1\|\Lambda(\xi) + T(\xi) - Id\| = K_1\|\widehat{Q}(\xi) - Id\| \leq K_1K_0k(1+\|\xi\|)^\theta.$$

Ainsi :

$$\begin{aligned}
|V_i^n(\xi)| &\leq e^{\alpha_S t_n}|V_i^0(\xi)| \\
&\quad + \sum_{l=0}^{n-1} e^{\alpha_S t_{n-1}}\sum_{j=i+1}^{N} K_0K_1kC_j(1+\|\xi\|)^{(N-j+1)\theta}t_l^{N-j}\max_{r\geq j}|V_r^0(\xi)|
\end{aligned}$$

Nous notons $K_i' = K_0K_1\max_{j,\, i+1\leq j\leq N} C_j$. Nous obtenons alors.

$$\begin{aligned}
|V_i^n(\xi)| &\leq e^{\alpha_S t_n}|V_i^0(\xi)| \\
&\quad + e^{\alpha_S t_{n-1}}K_i'k\left(\sum_{l=0}^{n-1}\left(\sum_{j=i+1}^{N} t_l^{N-j}\right)\right)(1+\|\xi\|)^{(N-i)\theta}\max_{r\geq i+1}|V_r^0(\xi)|
\end{aligned}$$

Or, comme, $\sum_{j=i+1}^{N} t_k^{N-j} \leq \sum_{j=i+1}^{N} t_n^{N-(i+1)} \leq (N-i)t_n^{N-(i+1)}$, nous avons :

$$\begin{aligned}
|V_i^n(\xi)| &\leq e^{\alpha_S t_n}|V_i^0(\xi)| \\
&\quad + e^{\alpha_S t_{n-1}}K_i'k\left(\sum_{l=0}^{n-1}(N-i)t_n^{N-i-1}\right)(1+\|\xi\|)^{(N-i)\theta}\max_{r\geq i+1}|V_r^0(\xi)| \\
&\leq e^{\alpha_S t_n}|V_i^0(\xi)| + e^{\alpha_S t_{n-1}}K_i't_n(N-i)t_n^{N-i-1}(1+\|\xi\|)^{(N-i)\theta}\max_{r\geq i+1}|V_r^0(\xi)| \\
&\leq e^{\alpha_S t_n}|V_i^0(\xi)| + C_i e^{\alpha_S t_{n-1}}t_n^{N-i}(1+\|\xi\|)^{(N-i)\theta}\max_{r\geq i+1}|V_r^0(\xi)|.
\end{aligned}$$

Ce qui prouve le lemme, en prenant $C_i = K_i'(N-i)$.

□

Nous avons, en particulier, l'existence d'une constante $K_1 > 0$ telle que

$$\forall i,\ 1 \leq i \leq N,\ \forall \xi \in \mathcal{D}_d,\ |V_i^n(\xi)| \leq K_1(1+\|\xi\|)^{(N-1)\theta} t_n^N e^{\alpha_S t_n} \max_{r \geq i} |V_r^0(\xi)|.$$

Ceci étant vrai pour toute les composantes de V^n, nous obtenons alors :

$$\|V^n(\xi)\| \leq K_2(1+\|\xi\|)^{(N-1)\theta} t_n^N e^{\alpha_S t_n} \|V^0(\xi)\|.$$

Ainsi :

$$\begin{aligned} \|U^n(\xi)\| = \|V^n(\xi)\| &\leq K_2(1+\|\xi\|)^{(N-1)\theta} t_n^N e^{\alpha_S t_n} \|V^0(\xi)\| \\ &\leq K_2(1+\|\xi\|)^{(N-1)\theta} t_n^N e^{\alpha_S t_n} \|U^0(\xi)\|. \end{aligned}$$

Donc

$$\|\widehat{Q}^n(\xi)\| \leq K_2(1+\|\xi\|)^{(N-1)\theta} t_n^N e^{\alpha_S t_n},$$

et cela prouve la stabilité du schéma.

□

Remarque 3.3 *Nous remarquons que la preuve de cette proposition donne une majoration de q_2 : $q_2 \leq (N-1)\theta$.*

Remarque 3.4 *Dans le cas des problèmes fortement bien posés en dimension $d = 1$, une classe particulière de schémas est étudiée : les schémas dissipatifs (voir [19]). Nous rappelons qu'un schéma est dit dissipatif d'ordre $2r$ s'il existe une constante $\delta > 0$ telle que toutes les valeurs propres $\lambda(\xi)$ de $\widehat{Q}(\xi)$ vérifient $|\lambda(\xi)| \leq (1 - \delta(|\xi|h)^{2r})e^{\alpha_S k}$. Dans ce cas, nous savons [19] qu'un schéma d'ordre $2r-1$ qui est dissipatif d'ordre $2r$ est stable.*

Dans le cas d'un problème faiblement bien posé, il est facile de voir que si un schéma est consistant et dissipatif, alors il est faiblement stable. En effet si nous considérons un schéma dissipatif alors la proposition précédente 3.2 implique que le schéma est faiblement stable à condition de prouver la majoration de $\|\widehat{Q} - Id\|$. Or, si nous supposons $\frac{k}{h} = \gamma$ est constant et, comme le schéma est consistant, nous avons $\|\widehat{Q} - \exp(kP(i\xi))\| \leq Kk(1+|\xi|^\rho)$, donc :

$$\begin{aligned} \|\widehat{Q} - Id\| &\leq \|\widehat{Q} - \exp(kP(i\xi))\| + \|\exp(kP(i\xi)) - Id)\| \\ &\leq Kk(1+|\xi|^\rho) + \sum_{n=1}^{+\infty} \frac{\|P(i\xi)\|^n k^n}{n!} \\ &\leq Kk(1+|\xi|^\rho) + \exp(k\|P(i\xi)\|) - 1 \\ &\leq Kk(1+|\xi|^\rho) + \|P(i\xi)\| k \exp(k\|P(i\xi)\|). \end{aligned}$$

Or cette étude se fait pour $|\xi|k \leq \gamma\pi$, *et* $k \leq k_0$, *donc* $\exp(k\|P(i\xi)\|)$ *est borné. Nous avons ainsi :*

$$\|\widehat{Q} - Id\| \leq Kk^r(1 + |\xi|^\rho) + K'(1 + |\xi|)k.$$

Les hypothèses de la proposition 3.2 sont vérifiées donc le schéma est faiblement stable.

3.1.2 Remarques sur l'étude de la consistance

Dans le cas des schémas à coefficients constants, nous n'allons pas étudier l'erreur de troncature τ^n. En effet, nous préférerons étudier les différences de matrices suivantes :

$$\frac{\exp(kP(i\xi)) - \widehat{Q}(\xi)}{k} \text{ et } \exp(t_nP(i\xi)) - \widehat{Q}^n(\xi).$$

Nous allons donner le lien entre ces deux différences et τ^n.

La transformée de Fourier de τ^n est :

$$k\widehat{\tau^n}(U^0)(\xi) = \widehat{Q_{-1}}(\xi)\exp(t_{n+1}P(i\xi))\widehat{U^0}(\xi) - \widehat{Q_0}(\xi)\exp(t_nP(i\xi))\widehat{U^0}(\xi).$$

Ainsi, nous faisons le lien suivant entre τ^n et la première différence :

$$\widehat{\tau^n}(U^0)(\xi) = \widehat{Q_{-1}}(\xi)\frac{\exp(kP(i\xi)) - \widehat{Q}(\xi)}{k}\exp(t_nP(i\xi))\widehat{U^0}(\xi). \tag{3.7}$$

De plus, si nous notons $e_n = (\exp(t_nP(i\xi)) - \widehat{Q}^n(\xi))\widehat{U^0}(\xi)$, nous avons :

$$e_{n+1} = \widehat{Q}(\xi)e_n - k\widehat{Q_{-1}}^{-1}(\xi)\widehat{\tau^n}(U^0)(\xi).$$

Ainsi,

$$e_n = \widehat{Q}(\xi)^n e_0 - k\sum_{j=0}^{n-1}\widehat{Q}(\xi)^{n-j-1}\widehat{Q_{-1}}^{-1}(\xi)\widehat{\tau^k}(U^0)(\xi).$$

D'où le lien suivant entre τ^n et la seconde différence :

$$(\exp(t_nP(i\xi)) - \widehat{Q}^n(\xi))\widehat{U^0}(\xi) = -k\sum_{j=0}^{n-1}\widehat{Q}(\xi)^{n-j-1}\widehat{Q_{-1}}^{-1}(\xi)\widehat{\tau^k}(U^0)(\xi). \tag{3.8}$$

Dans le cas des problèmes faiblement bien posés, les formules (3.7) et (3.8) ne suffisent pas pour le calcul de l'erreur. En effet, la matrice d'amplification qui intervient dans la formule (3.8) donne une perte de régularité q_2 à laquelle va s'ajouter la perte de régularité q_1 provenant de l'exponentielle du symbole dans la formule (3.7). Cela va conduire à des calculs d'erreur non optimaux ce qui n'était pas le cas pour les problèmes fortement bien posés car comme, dans ce cas, $q_1 = q_2 = 0$, il n'y avait pas de problème de cumul des pertes de régularité provenant du problème de Cauchy et du schéma.

3.2 Taux de convergence

3.2.1 Estimation générale du taux de convergence

Nous commençons par étudier de façon générale le taux de convergence i.e. β tel que $\|U(t_n,.) - SV^n\|_{L^2} \leq Ch^\beta$. Les estimations que nous allons obtenir ont l'avantage d'être vraies pour tout type de schéma quelle que soit la dimension mais elles ne sont pas optimales.

Nous supposerons, pour simplifier les notations que le pas de la grille est le même dans toutes les directions : $h_1 = h_2 = \cdots = h_d = h$.

Théorème 3.1 *Nous considérons que le problème de Cauchy est faiblement bien posé de défaut q_1 et que le schéma a pour condition initiale $V^0 = TU^0$ et est stable de défaut q_2.*
Nous supposons de plus qu'il existe une constante $C > 0$ et $\delta \in [0,1]$ tels que pour tout ξ vérifiant $\|\xi\| \leq \frac{\pi}{h^\delta}$:

$$\left\| e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right\|_2 \leq C t_n h^s (1 + \|\xi\|^2)^{\sigma/2}. \tag{3.9}$$

Soit q_4 tel que :

$$\max(q_1, q_2) \leq q_4 \qquad et \qquad s \leq \delta(\max(q_4, \sigma) - \max(q_1, q_2)). \tag{3.10}$$

Alors : $\exists K', \alpha', \forall f \in H^{q_4}(\mathbb{R}^d)$,

$$\|U(t_n,.) - SV^n\|_{L^2} \leq K' e^{\alpha' t_n} t_n h^{\beta_1} \|f\|_{H^{q_4}},$$

où :

$$\beta_1 = \frac{s(q_4 - \max(q_1, q_2))}{\max(q_4, \sigma) - \max(q_1, q_2)}.$$

Preuve :

$$\|U(t_n,.) - SV^n\|_{L^2}^2 = \int_{\mathbb{R}^d} \left\| e^{t_n P(i\xi)} U^0(\xi) - \widehat{SV^n}(\xi) \right\|^2 d\xi.$$

Or, d'après le paragraphe 2.1.2, nous avons :

$$\widehat{SV^n}(\xi) = \begin{cases} \mathcal{F}_h(V^n)(\xi) = \widehat{Q}^n(\xi)\widehat{U^0}(\xi) & si\ \|\xi\| \leq \frac{\pi}{h}, \\ 0 & sinon. \end{cases}$$

Donc :

$$\begin{aligned} \|U(t_n,.) - SV^n\|_{L^2}^2 &= \int_{\|\xi\| \leq \frac{\pi}{h}} \left\| \left(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right) \hat{U^0}(\xi) \right\|^2 d\xi \\ &\quad + \int_{\|\xi\| > \frac{\pi}{h}} \left\| e^{t_n P(i\xi)} \hat{U^0}(\xi) \right\|^2 d\xi. \end{aligned}$$

- Puisque le problème est faiblement bien posé de défaut q_1, nous avons :

$$\int_{\|\xi\|>\frac{\pi}{h}} \left\| e^{t_n P(i\xi)} \hat{f}(\xi) \right\|^2 d\xi \leq K_C^2 e^{2\alpha_C t_n} \int_{\|\xi\|>\frac{\pi}{h}} \left(1+\|\xi^2\|\right)^{q_1} \|\hat{f}(\xi)\|^2 d\xi.$$

- Pour $\eta \in [0, \frac{\pi}{h}]$, nous avons :

$$\begin{aligned}
&\int_{\|\xi\|\leq\frac{\pi}{h}} \left\| \left(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right) \hat{U}^0(\xi) \right\|^2 d\xi \\
&\qquad = \int_{\|\xi\|\leq\eta} \left\| \left(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right) \hat{U}^0(\xi) \right\|^2 d\xi \\
&\qquad + \int_{\eta\leq\|\xi\|\leq\frac{\pi}{h}} \left\| \left(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right) \hat{U}^0(\xi) \right\|^2 d\xi.
\end{aligned}$$

Or, par stabilité, nous avons :

$$\begin{aligned}
&\left\| \left(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right) \hat{U}^0(\xi) \right\| \\
&\qquad \leq \left(\left\| e^{t_n P(i\xi)} \right\| + \left\| \widehat{Q}^n(\xi) \right\| \right) \left\| \hat{U}^0(\xi) \right\| \\
&\qquad \leq 2\max(K_C, K_S) e^{\max(\alpha_C,\alpha_S) t_n} \left(1+\|\xi\|^2\right)^{\max(q_1,q_2)/2} \left\| \hat{U}^0(\xi) \right\|.
\end{aligned}$$

Donc :

$$\begin{aligned}
&\int_{\eta\leq\|\xi\|\leq\frac{\pi}{h}} \left\| \left(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right) \hat{U}^0(\xi) \right\|^2 d\xi \\
&\qquad \leq 4C_1^2 e^{2\alpha' t_n} \int_{\eta\leq\|\xi\|\leq\frac{\pi}{h}} \left(1+\|\xi\|^2\right)^{\max(q_1,q_2)} \left\| \hat{f}(\xi) \right\|^2 d\xi, \quad (3.11)
\end{aligned}$$

avec $C_1 = \max(K_C, K_S)$ et $\alpha' = \max(\alpha_C, \alpha_S)$. Et, si nous choisissons $\eta \leq \frac{\pi}{h^\delta}$, nous pouvons borner la première intégrale (3.11) au moyen de la majoration :

$$\int_{\|\xi\|\leq\eta} \|(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi))\hat{U}^0(\xi)\|^2 d\xi \leq C^2 t_n^2 h^{2s} \int_{\|\xi\|\leq\eta} (1+\|\xi\|^2)^\sigma \|\hat{U}^0(\xi)\|^2 d\xi.$$

Or, si $\|\xi\| \leq \eta$ et si ρ est tel que $0 \leq \rho \leq \sigma$, nous avons :

$$(1+\|\xi\|^2)^\sigma \leq (1+\eta^2)^{\sigma-\rho} (1+\|\xi\|^2)^\rho.$$

Donc :

$$\int_{\|\xi\|\leq\eta} \|(e^{t_n P(i\xi)} - \widehat{Q}^n(\xi))\hat{U}^0(\xi)\|^2 d\xi \leq C^2 t_n^2 h^{2s} (1+\eta^2)^{\sigma-\rho} \int_{\|\xi\|\leq\eta} (1+\|\xi\|^2)^\rho \|\hat{U}^0(\xi)\|^2 d\xi.$$

- Nous avons alors alors, en sommant les inégalités obtenues :

$$\begin{aligned}
&\|U(t_n,.) - SV^n\|_{L^2}^2 \\
&\qquad \leq \; 4C_1^2 e^{2\alpha' t_n} \int_{\eta \leq \|\xi\|} \left(1+\|\xi\|^2\right)^{\max(q_1,q_2)} \left\|\hat{U^0}(\xi)\right\|^2 d\xi \\
&\qquad\quad + C^2 t_n^2 h^{2s} (1+\eta^2)^{\sigma-\rho} \int_{\|\xi\| \leq \eta} \left(1+\|\xi\|^2\right)^{\rho} \left\|\hat{U^0}(\xi)\right\|^2 d\xi \\
&\qquad \leq \; C_2 e^{2\alpha' t_n} t_n^2 \left(\frac{1}{\eta^{2\theta}} \int_{\eta \leq \|\xi\|} \|\xi\|^{2\theta} \left(1+\|\xi\|^2\right)^{\max(q_1,q_2)} \left\|\hat{U^0}(\xi)\right\|^2 d\xi \right. \\
&\qquad\quad \left. + h^{2s} (1+\eta^2)^{\sigma-\rho} \int_{\|\xi\| \leq \eta} \left(1+\|\xi\|^2\right)^{\rho} \left\|\hat{U^0}(\xi)\right\|^2 d\xi \right),
\end{aligned}$$

où $\theta \geq 0$.
Si $\eta = \frac{\pi}{h^{\delta'}}$, $0 \leq \delta' \leq \delta \leq 1$, alors :

$$\begin{aligned}
&\|U(t_n,.) - SV^n\|_{L^2}^2 \\
&\qquad \leq \; C_3 e^{2\alpha' t_n} t_n^2 \left(h^{2\delta'\theta} \int_{\|\xi\| \geq \frac{\pi}{h^{\delta'}}} \left(1+\|\xi\|^2\right)^{\max(q_1,q_2)+\theta} \left\|\hat{U^0}(\xi)\right\|^2 d\xi \right. \\
&\qquad\quad \left. + h^{2(s-\delta'(\sigma-\rho))} \int_{\|\xi\| \leq \frac{\pi}{h^{\delta'}}} \left(1+\|\xi\|^2\right)^{\rho} \left\|\hat{U^0}(\xi)\right\|^2 d\xi \right).
\end{aligned}$$

Nous avons maintenant trois paramètres à ajuster, δ', θ et ρ. Afin que l'intégrale sur $\|\xi\| \geq \frac{\pi}{h^{\delta'}}$ existe, nous prenons $q_4 \geq \max(q_1,q_2) + \theta$. De plus, si on veut ne faire intervenir que la norme de f dans H^{q_4}, il faut prendre $\rho \leq q_4$. Pour avoir une estimation d'erreur optimale, nous devons maximiser $\min(\delta'\theta, s - \delta'(\sigma-\rho))$ sous les contraintes :

$$0 \leq \delta' \leq \delta \leq 1, \quad 0 \leq \rho \leq \min(\sigma, q_4), \quad 0 \leq \theta \leq q_4 - \max(q_1,q_2)$$

L'ensemble des contraintes est non vide car $\max(q_1,q_2) \leq q_4$.
Comme θ n'intervient que dans $\delta'\theta$, on prend $\theta = q_4 - \max(q_1,q_2)$. De plus, à δ' constant, le maximum de $\min(\delta'\theta, s-\delta'(\sigma-\rho))$ est atteint pour $\rho = \min(\sigma, q_4)$. Comme $\sigma - \rho \geq 0$, on prend δ' tel que :

$$\delta'\theta = s - \delta'(\sigma-\rho)$$

c'est-à-dire :

$$\delta' = \frac{s}{\theta + \sigma - \rho} = \frac{s}{q_4 - \max(q_1,q_2) + \sigma - \min(\sigma, q_4)} = \frac{s}{\max(\sigma, q_4) - \max(q_1,q_2)}.$$

Et nous avons alors :

$$\|U(t_n,.) - SV^n\|_{L^2}^2 \leq Ke^{2\alpha' t_n} t_n^2 h^{2\beta_1} \|f\|_{H^{q_4}}^2,$$

où :

$$\beta_1 = \delta'\theta = s - \delta'(\sigma - \rho).$$

Mais il faut que $0 \leq \delta' \leq \delta$ c'est-à-dire $s \leq \delta(\max(\sigma, q_4) - \max(q_1, q_2))$ ce qui donne la condition (3.10).
On a alors le résultat voulu.

□

Si la condition (3.10) n'est pas vérifiée, nous pouvons quand même estimer le taux de convergence grâce à la proposition suivante :

Proposition 3.3 *Si les hypothèses du théorème 3.1 sont vérifiées à l'exception de la condition (3.10) qui est remplacée par l'hypothèse moins restrictive $q_1 \leq q_4$, alors nous obtenons le même résultat avec :*

$$\beta_1' = \min(\delta(q_4 - q_1), s - (\sigma - q_4)^+).$$

PREUVE : Dans la preuve du théorème 3.1, nous prenons $\delta' = \delta$, nous avons alors $\eta = \frac{\pi}{h^\delta}$ et l'estimation :

$$\begin{aligned}\|U(t_n,.) - SV^n\|_{L^2}^2 \quad \leq \quad & C_3 e^{2\alpha' t_n} t_n^2 \left(h^{2\delta\theta} \int_{\|\xi\| \geq \frac{\pi}{h^\delta}} (1 + \|\xi\|^2)^{q_1+\theta} \|\hat{U^0}(\xi)\|^2 d\xi \right. \\ & \left. + h^{2(s-\delta(\sigma-\rho))} \int_{\|\xi\| \leq \frac{\pi}{h^\delta}} (1 + \|\xi\|^2)^{\rho} \|\hat{U^0}(\xi)\|^2 d\xi \right).\end{aligned}$$

Nous prenons alors : $\theta = q_4 - q_1$ et $\rho = \min(\sigma, q_4)$. Nous obtenons alors :

$$\beta_1' = \min(\delta(q_4 - q_1), s - (\sigma - q_4)^+).$$

□

Il n'est pas toujours facile d'évaluer s et σ dans (3.9). La proposition suivante va donner un choix de s et σ qui ne nécessite pas le calcul de puissances $n^{ièmes}$.

Proposition 3.4 *Si il existe $\delta \in [0,1]$ tel que pour tout ξ vérifiant $\|\xi\| \leq \frac{\pi}{h^\delta}$,*

$$\left\| \frac{e^{kP(i\xi)} - \hat{Q}(\xi)}{k} \right\| \leq Ch^r (1 + \|\xi\|^2)^{\rho/2},$$

alors, nous avons (3.9) avec $s = r$ et $\sigma = \rho + q_1 + q_2$. Nous obtenons alors le taux de convergence :

$$\beta_2 = \frac{r(q_4 - \max(q_1, q_2))}{\max(q_4, \rho + q_1 + q_2) - \max(q_1, q_2)}.$$

Remarque 3.5 *L'intérêt majeur de ce résultat est qu'il montre que, comme dans le cas des problèmes fortement bien posés, le taux de convergence est égal à l'ordre du schéma pour une donnée initiale suffisamment régulière. Par contre, nous verrons dans la partie suivante que ces valeurs ne conduisent pas à un taux de convergence optimal.*

PREUVE : Il suffit de développer $e^{t_n P(i\xi)} - \widehat{Q}^n(\xi)$:

$$e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) = \sum_{j=0}^{n-1} e^{k(n-1-j)P(i\xi)} \left(e^{kP(i\xi)} - \hat{Q}(\xi)\right) \widehat{Q}^j(\xi)$$

Nous utilisons alors la stabilité des problèmes continu et discret ainsi que l'hypothèse de la proposition.

$$\begin{aligned} \|e^{t_n P(i\xi)} - \widehat{Q}^n(\xi)\| &\leq \sum_{j=0}^{n-1} K_C e^{\alpha_C t_{n-1-j}} K_S e^{\alpha_S t_j} (1+\|\xi\|^2)^{(q_1+q_2+\rho)/2} k h^r \\ &\leq C e^{\alpha' t_n} t_n h^r (1+\|\xi\|^2)^{(q_1+q_2+\rho)/2}. \end{aligned}$$

□

Remarque 3.6 *Dans le cas d'un problème fortement bien posé la condition (3.10) donne $r \leq \max(q_4, \rho)$ et on a :*

- *$\beta_2 = r$ si $\rho \leq q_4$*
- *$\beta_2 = \frac{r q_4}{\rho}$ si $\rho \geq q_4$*

On retrouve donc le résultat de [38].

En pratique, lorsque l'on fait des estimations numériques d'erreur, on n'utilise pas $\|U(t_n,.) - SV^n\|$, qui ne se lit pas directement sur le tracé, mais on estime $\|EU(t_n,.) - V^n\|$. La proposition suivante donne la vitesse de convergence théorique pour cette erreur :

Proposition 3.5 *On considère que le problème de Cauchy est faiblement bien posé de défaut q_1 et que le schéma a pour condition initiale $V^0 = EU^0 \in L^2((h\mathbb{Z})^d)$ et est stable de défaut q_2.*
On suppose de plus que (3.9) est vérifiée.
Soit q_4 tel que :

$$\max(q_1, q_2) < q_4 - \frac{d}{2} \qquad et \qquad s \leq \delta(\max(q_4, \sigma) - \max(q_1, q_2)). \tag{3.12}$$

Alors : $\exists K_2, \alpha_2, \forall U^0 \in H^{q_4}(\mathbb{R}^d)$,

$$\|EU(t_n,.) - V^n\|_{L^2(G)} \leq K_2 e^{\alpha - 2t_n} t_n h^{\beta_1} \|U^0\|_{H^{q_4}},$$

où :

$$\beta_1 = \frac{s(q_4 - \max(q_1, q_2))}{\max(q_4, \sigma) - \max(q_1, q_2)}.$$

PREUVE : Soit W^n la solution du schéma de condition initiale $W^0 = TU^0$. On a alors :

$$\|EU(t_n,.) - W^n\|_{L^2(G)} \leq \|EU(t_n,.) - TU(t_n,.)\|_{L^2(G)} + \|TU(t_n,.) - W^n\|_{L^2(G)}.$$

Puisque $q_4 - q_1 > \frac{d}{2}$, nous majorons la première intégrale grâce au lemme 2.3.

$$\|EU(t_n,.) - W^n\|_{L^2(G)} \leq C_{0,q_4-q_1} h^{q_4-q_1} \|U(t_n,.)\|_{H^{q_4-q_1}(\mathbb{R}^d)} + \left\| \widehat{TU}(t_n,.) - \widehat{W^n} \right\|_{L^2(\mathcal{D}_d)}.$$

Puisque le problème de Cauchy est faiblement bien posé de défaut q_1, nous avons :

$$\|EU(t_n,.) - W^n\|_{L^2(G)} \leq C_{0,q_4-q_1} h^{q_4-q_1} K_C e^{\alpha_C t_n} \|U^0\|_{H^{q_4-q_1}(\mathbb{R}^d)} + \left\| \widehat{TU}(t_n,.) - \widehat{W^n} \right\|_{L^2(\mathcal{D}_d)}.$$

On majore la seconde intégrale en utilisant les définitions de T et S.

$$\|EU(t_n,.) - W^n\|_{L^2(G)} \leq C_{0,q_4-q_1} h^{q_4-q_1} K_C e^{\alpha_C t_n} \|U^0\|_{H^{q_4-q_1}(\mathbb{R}^d)} + \left\| \widehat{U}(t_n,.) - \widehat{SW^n} \right\|_{L^2(\mathcal{D}_d)}.$$

Nous appliquons alors le théorème 3.1 à W^n.

$$\|EU(t_n,.) - W^n\|_{L^2(G)} \leq C_1 t_n e^{\alpha_1 t_n} h^{\min(q_4-q_1,\beta_1)} \|U^0\|_{H^{q_4-q_1}(\mathbb{R}^d)}.$$

En utilisant la condition (3.10), on voit que $\beta_1 \leq q_4 - q_1$, il ne reste donc plus qu'à évaluer $\|W^n - V^n\|_{L^2(G)}$. Nous avons par définition :

$$\|W^n - V^n\|^2_{L^2(G)} = \int_{\mathcal{D}_d} \left\| \widehat{Q}^n(\xi) \left(\widehat{EU^0}(\xi) - \widehat{TU^0}(\xi) \right) \right\|^2 d\xi.$$

Par la condition de stabilité,

$$\|W^n - V^n\|^2_{L^2(G)} \leq K_S^2 e^{2\alpha_S t_n} \int_{\mathcal{D}_d} (1 + \|\xi\|^2)^{q_2} \left\| \widehat{EU^0}(\xi) - \widehat{TU^0}(\xi) \right\|^2 d\xi.$$

Soit, en revenant en variables physiques :

$$\|W^n - V^n\|^2_{L^2(G)} \leq K_S^2 e^{2\alpha_S t_n} \|EU^0 - TU^0\|^2_{h,q_2}.$$

Puisque $q_4 - q_1 > \frac{s}{2}$, nous utilisons le lemme 2.3.

$$\|W^n - V^n\|^2_{L^2(G)} \leq K_S^2 e^{2\alpha_S t_n} C^2_{q_2,q_4-q_2} h^{2(q_4-q_2)} \|U^0\|^2_{H^{q_4}(\mathbb{R}^d)}.$$

Et comme, d'après (3.12), nous avons $\beta_1 \leq q_4 - q_2$, nous avons prouvé la proposition. □

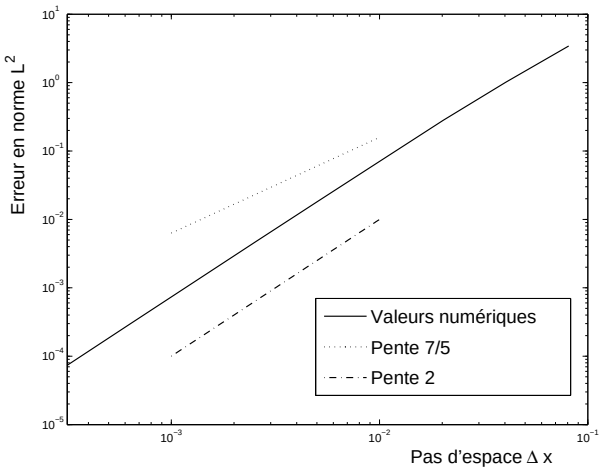
3.2.2 Un exemple

Nous allons étudier le problème monodimensionnel $\partial_t U = A\partial_x U + BU$ avec :

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}.$$

Dans ce cas, nous avons, d'après la proposition 1.8, $q_1 = 2$. Nous utilisons un schéma de Lax-Wendroff, que l'on précisera ultérieurement, pour discrétiser cette équation. Nous pouvons alors montrer que $q_2 = 2$. De plus, les paramètres r et σ intervenant dans la proposition 3.4 valent $r = 2$ et $\sigma = 3$. Donc, si nous appliquons le schéma à une donnée initiale $H^{11/2}$, le taux de convergence calculé dans la proposition 3.4 est de 7/5.

Nous allons maintenant comparer cette valeur avec les résultats numériques.



Nous remarquons alors que le taux de convergence n'est pas 7/5 mais plutôt 2. Dans la partie suivante, nous allons améliorer le résultat de la proposition 3.4 pour arriver à un taux de convergence qui sera en accord avec les résultats numériques.

3.2.3 Calcul optimal du taux de convergence dans le cas monodimensionnel

Nous allons étudier dans cette partie le cas particulier de la dimension 1 qui permet d'obtenir une estimation optimale du taux de convergence. Nous supposerons donc ici que $d = 1$. Nous noterons $\Delta x = h$ et $\Delta t = k$.

Le calcul optimal du taux de convergence ne s'applique qu'à une classe particulière de schémas. Toutefois, lorsque nous choisissons bien la discrétisation du terme d'ordre 0, la plupart des schémas classiques font partie de cette classe. Nous appellerons "schémas développables en P" les schémas de cette classe.

Définition 3.1 *Un schéma est dit développable en P à l'ordre N_0 s'il existe deux fonctions $z = z(\Delta t, \Delta x, \xi) \in \mathbb{C}$ et $R_{N_0}(\Delta t, \Delta x, \xi)$ tels que la matrice d'amplification du schéma soit de la forme :*

$$\widehat{Q}(\Delta t, \Delta x, \xi) = \sum_{k=0}^{N_0} \frac{P(z)^k \Delta t^k}{k!} + \Delta t^{N_0+1} R_{N_0}(\Delta t, \Delta x, \xi),$$

où : $R_{N_0}(\Delta t, \Delta x, \xi)$ ainsi que toutes ses dérivées par rapport à Δt commutent avec $P(z)$ et $\exists h_0 > 0$, $\exists k_0 > 0$, $\exists \delta \in [0, 1[$, $\exists \varepsilon > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$, $\forall z$, $|z| \leq \frac{\pi}{\Delta x^\delta}$, $\|\widehat{Q}(\xi) - Id\| \leq 1 - \varepsilon$.

Nous allons calculer une nouvelle valeur de s et σ qui interviennent dans le théorème 3.1. Nous séparerons le calcul en deux estimations : la première porte sur $\|\widehat{Q}^n - \exp(t_n P(z))\|$ et la seconde sur la différence d'exponentielles $\|\exp(t_n P(z)) - \exp(t_n P(i\xi))\|$.

Première estimation

Proposition 3.6 *Nous supposons qu'il existe $h_0 > 0$ et $k_0 > 0$ tels que :*

- *le schéma est développable en P à l'ordre N_0,*
- $\forall k \geq 0$, $\exists B_k > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$,

$$\left\| \frac{\partial^k R_{N_0}(\Delta t, \Delta x, \xi)}{\partial \Delta t^k} \right\| \leq B_k (1 + |\xi|^{N_0+k+1}), \tag{3.13}$$

- $\exists C > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$,

$$|z| \leq C|\xi|, \tag{3.14}$$

• *le quotient* $\frac{\Delta t}{\Delta x} = \gamma$ *est constant et le schéma est faiblement stable de défaut* q_2.
Alors, nous avons :

$$\exists K > 0,\, \forall \Delta x,\, \Delta x \leq h_0,\, \forall \Delta t,\, \Delta t \leq k_0,\, \forall \xi,\, |\xi| \leq \frac{\pi}{\Delta x^\delta},\, \forall z,\, |z| \leq \frac{\pi}{\Delta x^\delta},$$

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq K t_n e^{\alpha_S t_n} \Delta t^{N_0} (1 + |\xi|)^{N_0+1+q_2}$$

PREUVE : Nous allons exprimer $\widehat{Q}^n - \exp(t_n P(z) t_n)$ au moyen de l'inégalité des accroissements finis.
Nous notons $M(s) = \sum_{k=1}^{N_0} \frac{P(z)^k s^k}{k!} + s^{N_0+1} R_{N_0}(s, \Delta x, \xi)$. Nous avons alors $\widehat{Q} = Id + M(\Delta t)$.

• D'après les hypothèses, $\|M(\Delta t)\| < 1$ donc nous pouvons écrire :

$$\widehat{Q}^n = \exp\left(n \log(Id - M(\Delta t))\right) = \exp\left(t_n \times \frac{1}{\Delta t} \log(Id - M(\Delta t))\right),$$

et comme $M(\Delta t)$ et $P(z)$ commutent, par hypothèse, nous avons :

$$\widehat{Q}^n = \exp(t_n P(z)) \exp\left(t_n (\frac{1}{\Delta t} \log(Id + M(\Delta t)) - P(z)\right).$$

Posons :

$$f(t) = \frac{1}{t} \log(Id + M(t)) - P(z).$$

Puisque :

$$f(0) = P(z) - P(z) = 0 \text{ par hypothèse,}$$

nous avons :

$$\widehat{Q}^n - \exp(t_n P(z)) = \exp(t_n P(z)) \left(\exp(t_n f(\Delta t)) - \exp(t_n f(0))\right).$$

Par l'inégalité des accroissements finis,

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq \Delta t . t_n \sup_{t \in [0, \Delta t]} \|\exp(t_n P(z)) f'(t) \exp(t_n f(t))\|.$$

Or f' commute avec $P(z)$, ainsi :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq \Delta t . t_n \sup_{t \in [0, \Delta t]} \|f'(t) \exp(t_n P(z)) \exp(t_n f(t))\|.$$

De plus, par définition de f,

$$\exp(t_n P(z)) \exp(t_n f(t)) = \widehat{Q}^n.$$

Donc, comme le schéma est stable de défaut q_2,

$$\sup_{t\in[0,\Delta t]} \|\exp(t_n P(z))\exp(t_n f(t))\| \leq \sup_{t\in[0,\Delta t]} \|\widehat{Q}^n\| \leq K_S e^{\alpha_S t_n}(1+|\xi|)^{q_2}.$$

Nous avons alors :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq \Delta t.t_n K_S e^{\alpha_S t_n}(1+|\xi|)^{q_2} \sup_{t\in[0,\Delta t]} \|f'(t)\|. \tag{3.15}$$

Nous allons maintenant exprimer f'.

- Nous avons

$$f'(t) = -\frac{1}{t^2}\log(Id+M(t)) + \frac{1}{t}M'(t)(Id+M(t))^{-1}.$$

Posons $g(t) = M'(t)(Id+M(t))^{-1}$. Alors, nous pouvons exprimer f' en fonction de g :

$$f'(t) = -\frac{1}{t^2}\int_0^t g(s)ds + \frac{1}{t}g(t).$$

Lemme 3.2

$$g(0) = P(z) \text{ et } \forall n,\ 1 \leq n \leq N_0 - 1,\ g^{(n)}(0) = 0.$$

PREUVE :

– Commençons par calculer $g(0)$. D'après la définition d'un schéma développable en P, nous avons $M(0) = 0$ et $M'(0) = P(z)$ donc $g(0) = P(z)$.

– Calculons $g^{(n)}(0)$ pour $n \geq 1$:
Nous avons : $g(s)(Id + M(s)) = M'(s)$ donc :

$$\sum_{k=0}^{n} C_n^k g^{(k)}(s)\frac{d^{(n-k)}}{ds^{(n-k)}}(Id+M(s)) = M^{(n+1)}(s).$$

Donc :

$$g^{(n)}(s)(Id+M(s)) = M^{(n+1)}(s) - \sum_{k=0}^{n-1} C_n^k g^{(k)}(s)M^{(n-k)}(s),$$

et comme $M(0) = 0$:

$$g^{(n)}(0) = M^{(n+1)}(0) - \sum_{k=0}^{n-1} C_n^k g^{(k)}(0)M^{(n-k)}(0).$$

De plus, d'après la définition d'un schéma développable en P, si $1 \leq k \leq N_0$,

$$M^{(k)}(0) = k!\frac{P(z)^k}{k!} = P(z)^k.$$

Montrons alors le lemme par récurrence sur n.

– Pour $n = 1$:

$$g'(0) = M''(0) - g(0)M'(0) = P(z)^2 - P(z) \times P(z) = 0.$$

– Soit $1 \leq n \leq N_0 - 1$, supposons que $\forall 1 \leq k \leq n-1$, $g^{(k)}(0) = 0$, alors :

$$g^{(n)}(0) = M^{(n+1)}(0) - g(0)M^{(n)}(0) = P(z)^{n+1} - P(z) \times P(z)^n = 0.$$

Nous avons donc bien $\forall 1 \leq n \leq N_0 - 1$, $g^{(n)}(0) = 0$.

□

- Nous pouvons alors écrire :

$$f'(t) = -\frac{1}{t^2}\int_0^t \left(g(s) - \sum_{k=0}^{N_0-1} \frac{g^{(k)}(0)}{k!}s^k\right) ds + \frac{1}{t}\left(g(t) - \sum_{k=0}^{N_0-1} \frac{g^{(k)}(0)}{k!}t^k\right). \tag{3.16}$$

Or, par la formule de Taylor avec reste intégral, nous avons :

$$g(t) - \sum_{k=0}^{N_0-1} \frac{g^{(k)}(0)}{k!}t^k = \frac{t^N}{(N-1)!}\int_0^1 (1-s)^{(N_0-1)} g^{(N_0)}(st)ds. \tag{3.17}$$

Exprimons maintenant $M^{(n)}$.

Lemme 3.3

$$\forall n \geq 0,\ \exists K_n > 0,\ \forall s \in [0, \Delta t],\ \|M^{(n)}(s)\| \leq K_n(1 + |\xi|^n).$$

PREUVE : Nous avons :

$$M^{(n)}(s) = \sum_{k=n}^{N_0} \frac{P(z)^k}{(k-n)!}s^{k-n} + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!}{(N_0+1-j)!}s^{N_0+1-j}\frac{\partial^{n-j}R_{N_0}(s,\xi)}{\partial s^{n-j}}.$$

Donc, d'après les hypothèses et (3.13) et (3.14) :

$$\begin{aligned}\|M^{(n)}(s)\| &\leq \sum_{k=n}^{N_0} \frac{(1+|\xi|^k)}{(k-n)!}s^{k-n} \\ &\quad + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!}{(N_0+1-j)!}s^{N_0+1-j}B_{n-j}(1+|\xi|^{N_0+n-j+1}).\end{aligned}$$

Comme $s \in [0, \Delta t]$,

$$\begin{aligned}\|M^{(n)}(s)\| &\leq \sum_{k=n}^{N_0} \frac{(1+|\xi|^k)}{(k-n)!}\Delta t^{k-n} \\ &\quad + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!}{(N_0+1-j)!}\Delta t^{N_0+1-j}B_{n-j}(1+|\xi|^{N_0+n-j+1}).\end{aligned}$$

De plus, nous avons supposé que $\frac{\Delta t}{\Delta x} = \gamma$, donc :

$$\begin{aligned}
\|M^{(n)}(s)\| &\leq \sum_{k=n}^{N_0} \frac{(1+|\xi|^k)\gamma^{k-n}}{(k-n)!}\Delta x^{k-n} \\
&\quad + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!\gamma^{N_0+1-j}}{(N_0+1-j)!}\Delta x^{N_0+1-j} B_{n-j}(1+|\xi|^{N_0+n-j+1}) \\
&\leq \sum_{k=n}^{N_0} \frac{\gamma^{k-n}}{(k-n)!}\Delta x^{k-n} + |\xi^n| \sum_{k=n}^{N_0} \frac{|\xi|^{k-n}\gamma^{k-n}}{(k-n)!}\Delta x^{k-n} \\
&\quad + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!\gamma^{N_0+1-j}}{(N_0+1-j)!}\Delta x^{N_0+1-j} B_{n-j} \\
&\quad |\xi^n| + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!\gamma^{N_0+1-j}}{(N_0+1-j)!}\Delta x^{N_0+1-j} B_{n-j}|\xi|^{N_0-j+1}
\end{aligned}$$

Or, nous avons $|\xi| \leq \frac{\pi}{\Delta x^\delta}$, donc :

$$\begin{aligned}
\|M^{(n)}(s)\| &\leq \sum_{k=n}^{N_0} \frac{\gamma^{k-n}}{(k-n)!}\Delta x^{k-n} + |\xi|^n \sum_{k=n}^{N_0} \frac{\pi^{k-n}\gamma^{k-n}}{(k-n)!}\Delta x^{(k-n)(1-\delta)} \\
&\quad + \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!\gamma^{N_0+1-j}}{(N_0+1-j)!}\Delta x^{N_0+1-j} B_{n-j} \\
&\quad + |\xi|^n \sum_{j=0}^{\min(n,N_0+1)} \frac{(N_0+1)!\gamma^{N_0+1-j}}{(N_0+1-j)!}\Delta x^{(N_0+1-j)(1-\delta)} B_{n-j}\pi^{N_0-j+1}
\end{aligned}$$

Donc, il existe h_0 assez petit et il existe une constante K_n tels que si $\Delta x < h_0$, nous avons :

$$\|M^{(n)}(s)\| \leq K_n(1+|\xi|^n).$$

□

Montrons que $\forall n \geq 0$, $\exists C_n > 0$, $\forall s \in [0, \Delta t]$, $\|g^{(n)}(s)\| \leq C_n(1+|\xi|)^{n+1}$:
– Pour $n = 0$:

$$\|g(s)\| = \|M'(s)(Id + M(s))^{-1}\|.$$

Or, comme $\|\hat{Q}(\xi) - Id\| \leq 1 - \varepsilon$ d'après la définition des schémas développables en P.

$$\|(Id + M(s))^{-1}\| \leq \frac{1}{1 - \|M(s)\|} \leq \frac{1}{\varepsilon}.$$

Ainsi, grâce au lemme précédent :

$$\|g(s)\| \leq \frac{K_1}{\varepsilon}(1+|\xi|).$$

– Soit $n \geq 0$, supposons que $\forall k \leq n$, $\|g^{(n)}(s)\| \leq C_n(1+|\xi|^{n+1})$. Nous avons :

$$\begin{aligned}\|g^{(n)}(s)\| = \; & \Big\|M^{(n+1)}(s)(Id+M(s))^{-1} \\ & -\sum_{k=0}^{n-1} C_n^k g^{(k)}(s) M^{(n-k)}(s)(Id+M(s))^{-1}\Big\|\end{aligned}$$

Et, d'après le lemme précédent et l'hypothèse de récurrence,

$$\begin{aligned}\|g^{(n)}(s)\| \leq & \left(K_{n+1}(1+|\xi|^{n+1}) + \sum_{k=0}^{n-1} C_n^k C_k (1+|\xi|^{k+1}) K_{n-k}(1+|\xi|^{n-k})\right)\frac{1}{\varepsilon} \\ \leq & \; C_n(1+|\xi|)^{n+1}.\end{aligned}$$

Nous avons donc :

$$\|g^{(N_0)}(s)\| \leq C_{N_0}(1+|\xi|^{N_0+1}).$$

D'où, d'après (3.17) :

$$\|g(t) - \sum_{k=0}^{N_0-1} \frac{g^{(k)}(0)}{k!} t^k\| \leq K' t^{N_0}(1+|\xi|^{N_0+1}).$$

Ainsi, en remplaçant dans (3.16) :

$$\|f'(t)\| \leq K'' t^{N_0-1}(1+|\xi|^{N_0+1}).$$

- Nous pouvons maintenant remplacer l'estimation de f' dans (3.15), nous obtenons alors :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq \Delta t.t_n K_S e^{\alpha_S t_n}(1+|\xi|)^{q_2} K'' t^{N_0-1}(1+|\xi|^{N_0+1}).$$

Donc :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq K t_n e^{\alpha_S t_n} \Delta t^{N_0}(1+|\xi|)^{N_0+1+q_2}.$$

□

Seconde estimation

Le but de cette partie est d'estimer la différence d'exponentielles $\|\exp(t_n P(z)) - \exp(t_n P(i\xi))\|$. Nous allons étudier, de manière plus générale, la différence $\|\exp(t_n P(z_1)) - \exp(t_n P(z_2))\|$ en utilisant la théorie des perturbations exposée dans [24].

Nous commençons par développer en série de Puiseux $\exp(tP(z))$.

Proposition 3.7 *Il existe $C > 0$, $r \in \mathbb{N}$ et $A_{k,j,n} \in \mathcal{M}_N(\mathbb{C})$ tels que pour tout $|z| \geq C$, pour tout $t \geq 0$, $P(z) = zA + B$ a s valeurs propres distinctes notées $\lambda_k(z)$ et on a :*

$$\exp(tP(z)) = \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r}^{+\infty} A_{k,j,n} z^{-n/p} \right) \right) \exp(\lambda_k(z)t) \right), \tag{3.18}$$

où le développement est normalement convergent.

Preuve :

- Soit $M(\tau) = A + \tau B$. Nous étudions $M(\tau)$ au voisinage de $\tau = 0$. Nous nous plaçons sur un voisinage de 0 ne contenant pas de point exceptionnel à part, éventuellement, 0.
  Nous écrivons la décomposition de Dunford de $M(\tau)$: $M(\tau) = N(\tau) + D(\tau)$ où N est nilpotente, D est diagonalisable et N et D commutent.
  Soient $\mu_k(\tau)$ les s valeurs propres distinctes de $M(\tau)$ et $P_k(\tau)$ la projection sur le sous-espace caractéristique associé à $\mu_k(\tau)$.
  Alors $D(\tau) = \sum_{k=1}^{s} \mu_k(\tau) P_k(\tau)$ et $N(\tau) = \sum_{k=1}^{s} N_k(\tau)$,où la somme correspond à la décomposition de l'espace en somme directe de sous-espaces caractéristiques.
  On a alors, d'après [24] (voir partie 1.2), un développement de la forme :

$$P_k(\tau) = \sum_{n=-r}^{+\infty} \alpha_{n,k} \tau^{n/p} \text{ et } N_k(\tau) = \sum_{n=-r}^{+\infty} \beta_{n,k} \tau^{n/p}.$$

- Posons $\tau = \frac{1}{z}$, alors $P(z) = zA + B = zM(\tau) = \frac{1}{\tau} N(\tau) + \frac{1}{\tau} D(\tau)$. Nous avons donc :

$$\begin{aligned}
\exp(tP(z)) &= \exp\left(\frac{t}{\tau} N(\tau)\right) \exp\left(\frac{t}{\tau} D(\tau)\right) \\
&= \left(\sum_{k=1}^{s} \exp\left(\frac{t}{\tau} N_k(\tau)\right) \right) \left(\sum_{k=1}^{s} \exp\left(\frac{t}{\tau} \mu_k(\tau)\right) P_k(\tau) \right) \\
&= \left(\sum_{k=1}^{s} \sum_{j=0}^{N} \frac{t^j}{\tau^j j!} N_k(\tau)^j \right) \left(\sum_{k=1}^{s} \exp\left(\frac{t}{\tau} \mu_k(\tau)\right) P_k(\tau) \right).
\end{aligned}$$

En introduisant les développements de N_k et P_k, et en réarrangeant les sommes, nous obtenons une expression de la forme :

$$\exp(tP(z)) = \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r}^{+\infty} A_{k,j,n} \tau^{n/p} \right) \right) \exp\left(\frac{t}{\tau} \mu_k(\tau) \right) \right).$$

Si $\lambda_k(z)$ est valeur propre de $P(z)$, alors $\lambda_k(z) = \frac{1}{z}\mu_k(\frac{1}{z})$, d'où la conclusion.

□

A partir de ce développement en série de Puiseux, on va déterminer le défaut q_1 du problème de Cauchy.

Proposition 3.8 *Si le problème de Cauchy est faiblement bien posé et si on a (3.18) avec $z = i\xi$, alors le défaut est exactement $q_1 = \frac{r_0}{p}$, où $r_0 = \inf\{r \geq 0, \exists k \in \{1, \ldots, s\}, \exists j \in \{1, \ldots, N\}, A_{k,j,-r} \neq 0\}$.*

PREUVE :

- Montrons d'abord que $q_1 \leq \frac{r_0}{p}$.
  Pour $|\xi| \geq C$, nous avons :

$$\exp(tP(i\xi)) = \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} A_{k,j,n} (i\xi)^{-n/p} \right) \right) \exp(\lambda_k(i\xi)t) \right).$$

  Donc :

$$\| \exp(tP(i\xi)) \| \leq \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} \|A_{k,j,n}\| |\xi|^{-n/p} \right) \right) \exp(Re\,(\lambda_k(i\xi))t) \right).$$

  Or, comme le problème est faiblement bien posé, il existe un $\alpha > 0$ tel que $Re\,(\lambda_k(i\xi)) \leq \alpha$ et ainsi, comme la série est normalement convergente :

$$\begin{aligned} \| \exp(tP(i\xi)) \| &\leq |\xi|^{r/p} e^{\alpha t} \sum_{k=1}^{s} \sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} \|A_{k,j,n}\| |\xi|^{-\frac{n+r}{p}} \right) \\ &\leq K |\xi|^{r/p} e^{\alpha t} (1 + t^N) \text{ pour } |\xi| \geq C. \end{aligned}$$

  On a donc :

$$q_1 \leq \frac{r_0}{p}.$$

- Procédons maintenant par l'absurde et supposons que $q_1 < \frac{r_0}{p}$.
  Comme $\exp(tP(i\xi))$ s'écrit :

$$\exp(tP(i\xi)) = (i\xi)^{r/p} \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} A_{k,j,n} (i\xi)^{-\frac{n+r_0}{p}} \right) \right) \exp(\lambda_k(i\xi)t) \right),$$

et que le problème est faiblement bien posé de défaut q_1, nous avons :

$$\left\| \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} A_{k,j,n} (i\xi)^{-\frac{n+r_0}{p}} \right) \right) \exp(\lambda_k(i\xi)t) \right) \right\| \leq K e^{\alpha t} |\xi|^{q_1 - r_0/p},$$

ce qui entraîne que :

$$\lim_{|\xi| \to +\infty} \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} A_{k,j,n} (i\xi)^{-\frac{n+r_0}{p}} \right) \right) \exp(\lambda_k(i\xi)t) \right) = 0.$$

Or, d'après la convergence normale de la série,

$$\sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j \left(\sum_{n=-r_0}^{+\infty} A_{k,j,n} (i\xi)^{-\frac{n+r_0}{p}} \right) \right) \exp(\lambda_k(i\xi)t) \right)$$

$$= \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j A_{k,j,-r_0} \right) \exp(\lambda_k(i\xi)t) \right) + o(1).$$

Donc, pour tout $t \geq 0$,

$$\lim_{|\xi| \to +\infty} \sum_{k=1}^{s} \left(\left(\sum_{j=0}^{N} t^j A_{k,j,-r_0} \right) \exp(\lambda_k(i\xi)t) \right) = 0.$$

Ainsi, à t fixé, on a une combinaison linéaire d'exponentielles bornées qui tend vers 0, donc $\forall t \geq 0$, $\sum_{j=0}^{N} t^j A_{k,j,-r_0} = 0$. Et ainsi $\forall k, j$, $A_{k,j,-r_0} = 0$ ce qui contredit la définition de r_0. Nous avons donc $q_1 \geq \frac{r_0}{p}$, ce qui conclut la démonstration de la proposition.

□

Pour évaluer la différence d'exponentielles, nous allons utiliser une formule de Taylor. Nous commençons donc par évaluer la dérivée de l'exponentielle.

Proposition 3.9 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, alors il existe $C > 0$ tel que si $|z| \geq C$, alors, il existe $K > 0$ tel que :*

$$\left\| \frac{d}{dz} \exp(tP(z)) \right\| \leq K |z|^{q_1} (1 + t^{N+1}) e^{\alpha t}.$$

Preuve : Nous utilisons les notations des propositions précédentes.

- Le développement en série de Puiseux des valeurs propres et la proposition 1.5 donnent, pour $|z| \geq C$:

$$\lambda_k(z) = z\lambda_k^0 + \sum_{j=0}^{+\infty} \alpha_j z^{-j/p}$$

Donc λ_k est dérivable pour $|z| \geq C$, et :

$$\frac{d}{dz}\lambda_k(z) = \lambda_k^0 + \sum_{j=0}^{+\infty} \left(-\frac{j}{p}\right) \alpha_j z^{-j/p-1}.$$

Donc, pour $|z| \geq C$, $\frac{d}{dz}\lambda_k(z) \leq K$. Ainsi :

$$\begin{aligned} \left|\frac{d}{dz}\exp(\lambda_k(z)t)\right| &= \left|\frac{d}{dz}\lambda_k(z).t.\exp(\lambda_k(z)t)\right| \\ &\leq Kte^{\alpha t} \end{aligned}$$

- Nous avons, en utilisant la formule (3.18) :

$$\begin{aligned} \frac{d}{dz}\exp(tP(z)) &= -\sum_{k=1}^{s}\sum_{j=0}^{N} t^j \left[\left(\sum_{n=-r_0}^{+\infty} A_{k,j,n}\frac{n}{p}z^{-n/p-1}\right)\exp(\lambda_k(z)t)\right. \\ &\quad \left. + \left(\sum_{n=-r_0}^{+\infty} A_{k,j,n}z^{-n/p}\right)\frac{d}{dz}\exp(\lambda_k(z)t)\right] \end{aligned}$$

D'où :

$$\begin{aligned} \left\|\frac{d}{dz}\exp(tP(z))\right\| &\leq \sum_{k=1}^{s}\sum_{j=0}^{N} t^j \left[\left(\sum_{n=-r_0}^{+\infty} \|A_{k,j,n}\|\frac{n}{p}|z|^{-n/p-1}\right)e^{\alpha t}\right. \\ &\quad \left. + \left(\sum_{n=-r_0}^{+\infty} \|A_{k,j,n}\||z|^{-n/p}\right)Kte^{\alpha t}\right] \\ &\leq |z|^{r_0/p-1}\sum_{k=1}^{s}\sum_{j=0}^{N} t^j \left[\left(\sum_{n=-r_0}^{+\infty} \|A_{k,j,n}\|\frac{n}{p}|z|^{(r_0-n)/p-1}\right)e^{\alpha t}\right. \\ &\quad \left. + |z|^{r_0/p-1}\left(\sum_{n=-r_0}^{+\infty} \|A_{k,j,n}\||z|^{(r_0-n)/p}\right)Kte^{\alpha t}\right] \end{aligned}$$

Or d'après la convergence normale de la série (3.18), les deux séries sont bornées pour $|z| \geq C$. Nous avons alors :

$$\left\|\frac{d}{dz}\exp(tP(z))\right\| \leq K_1(1+t^N)|z|^{r_0/p-1}e^{\alpha t} + K_2(1+t^N)|z|^{r/p}te^{\alpha t},$$

or $q_1 = \frac{r_0}{p}$, donc, si $|z| \geq C$:

$$\left\| \frac{d}{dz} \exp(tP(z)) \right\| \leq K(1+t^{N+1})|z|^{q_1} e^{\alpha t}.$$

□

Nous arrivons maintenant à l'estimation de la différence d'exponentielles :

Proposition 3.10 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, alors il existe $C > 0$ tel que, pour tous $|z_1|$, $|z_2| \geq C$, tels que pour tout $z \in [z_1, z_2]$, $|z| \geq C$, alors :*

$$\| \exp(tP(z_1)) - \exp(tP(z_2)) \| \leq K(1+t^{N+1}) e^{\alpha t} (|z_1| + |z_2|)^{q_1} |z_1 - z_2|.$$

PREUVE : Nous pouvons utiliser l'inégalité des accroissements finis :

$$\| \exp(tP(z_1)) - \exp(tP(z_2)) \| \quad \leq \quad \|z_1 - z_2\| \sup_{z \in [z_1, z_2]} \| \frac{d}{dz} \exp(tP(z)) \|$$

En utilisant la proposition précédente, nous obtenons :

$$\begin{aligned} \| \exp(tP(z_1)) - \exp(tP(z_2)) \| &\leq K(1+t^{N+1}) \|z_1 - z_2\| e^{\alpha t} \sup_{z \in [z_1, z_2]} |z|^{q_1} \\ &\leq K(1+t^{N+1}) \|z_1 - z_2\| e^{\alpha t} (|z|_1 + |z_2|)^{q_1} \end{aligned}$$

□

Remarque 3.7 *Soit $\varepsilon > 0$ alors si $|z_1 - z_2| \leq \varepsilon$ et si $|z_1| \geq C$ et $|z_2| \geq C + \varepsilon$, alors $|sz_1 + (1-s)z_2| \geq |z_2| - s|z_1 - z_2| \geq C + \varepsilon - s\varepsilon \geq C$ et l'hypothèse est vérifiée.*

Conclusion

En utilisant les propositions 3.6 et 3.10, on montre que les valeurs suivantes de s et de σ conviennent pour l'utilisation du théorème 3.1.

Théorème 3.2 *On suppose que :*

- *le problème de Cauchy est faiblement bien posé de défaut q_1,*
- *le quotient $\frac{\Delta t}{\Delta x} = \gamma$ est constant et le schéma est faiblement stable de défaut q_2,*
- *le schéma est développable en P à l'ordre N_0,*
- *$\forall k \geq 0$, $\exists B_k > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$, $\left\| \frac{\partial^k R_{N_0}(\Delta t, \Delta x, \xi)}{\partial \Delta t^k} \right\| \leq B_k(1 + |\xi|^{N_0+k+1})$,*
- *$\exists C > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$, $|z| \leq C|\xi|$,*

- $\forall |\xi| \leq \frac{\pi}{\Delta x^\delta}$, $|z - i\xi| \leq K\Delta x^r |\xi|^\rho$.
  Alors, il existe $K > 0$ *tel que :*

$$\|\widehat{Q}^n - \exp(t_n P(i\xi))\| \leq K e^{\alpha t_n}(\Delta t^{N_0} + \Delta x^r)(1 + |\xi|)^{\max(N_0+1+q_2, \rho+q_1)}.$$

PREUVE : Nous écrivons :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq \|\widehat{Q}^n - \exp(t_n P(z))\| + \|\exp(t_n P(z)) - \exp(t_n P(i\xi))\|.$$

Or, d'après la proposition 3.6, nous avons :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq K t_n e^{\alpha_S t_n} \Delta t^{N_0} (1 + |\xi|)^{N_0+1+q_2}.$$

Nous appliquons alors la proposition 3.10 en prenant $z_1 = i\xi$ et $z_2 = z$. Nous obtenons :

$$\|\exp(tP(z_1)) - \exp(tP(z_2))\| \leq K'(1 + t^{N+1}) e^{\alpha t} (|\xi|)^{q_1} \Delta x^r |\xi|^\rho.$$

D'où :

$$\|\widehat{Q}^n - \exp(t_n P(z))\| \leq K'' e^{\alpha t}(\Delta t^{N_0} + \Delta x^r)(1 + |\xi|)^{\max(N_0+1+q_2, \rho+q_1)}.$$

□

L'estimation obtenue dans le théorème précédent est plus fine que celle obtenue à la proposition 3.4. En utilisant cette estimation dans le théorème 3.1, nous obtiendrons un taux de convergence plus grand. Nous verrons dans le chapitre suivant que, sur les exemples classiques de schémas, les conditions d'application du théorème 3.2 ne sont pas restrictives, à condition de bien choisir la discrétisation du terme d'ordre 0 et que de plus, le taux de convergence obtenu est optimal.

Chapitre 4

Application à divers schémas et résultats numériques

Dans ce chapitre, nous allons appliquer les résultats précédents aux schémas à un pas classiques. Pour chaque schéma, nous allons étudier la stabilité et le taux de convergence. Nous comparerons alors les résultats obtenus avec des résultats numériques.

Afin de pouvoir appliquer le théorème 3.2, nous choisirons une discrétisation du terme d'ordre 0 qui rende le schéma étudié développable en P, ce qui n'est pas toujours le cas pour la discrétisation standard.

Nous posons :

$$\gamma = \frac{\Delta t}{\Delta x}$$

Dans tout ce chapitre, on supposera que γ est une constante.

4.1 Schéma de Lax-Wendroff

Le schéma de Lax Wendroff homogène (sans terme d'ordre 0) est obtenu par "équation équivalente"'. En revanche, la discrétisation du terme d'ordre 0 n'est pas intrinsèque, et nous verrons qu'un choix astucieux permet d'obtenir un schéma développable en P.

4.1.1 Ecriture d'un schéma de Lax-Wendroff

Nous commençons par appliquer la méthode qui conduit à l'obtention d'un schéma de Lax-Wendroff usuel.

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \partial_t U(t_n, x_j) + \frac{\Delta t}{2} \partial_{tt} U(t_n, x_j) + O(\Delta t^2)$$

$$\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = \partial_x U(t_n, x_j) + O(\Delta x^2).$$

Donc :

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} - A\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = BU(t_n, x_j) + \frac{\Delta t}{2}\partial_{tt} U(t_n, x_j) + O(\Delta t^2) + O(\Delta x^2).$$

Or en dérivant l'équation en temps et en espace, nous obtenons :

$$\partial_{tt} U = A\partial_{tx} U + B\partial_t U \text{ et } \partial_{tx} U = A\partial_{xx} U + B\partial_x U.$$

Donc :

$$\begin{aligned} \partial_{tt} U &= A^2\partial_{xx} U + AB\partial_x U + B\partial_t U \\ &= A^2\partial_{xx} U + (AB + BA)\partial_x U + B^2 U. \end{aligned}$$

D'où :

$$\begin{aligned} \frac{U_j^{n+1} - U_j^n}{\Delta t} - A\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} &= BU(t_n, x_j) \\ &\quad + \frac{\Delta t}{2}\left(A^2\partial_{xx} U + (AB + BA)\partial_x U + B^2 U\right) \\ &\quad + O(\Delta t^2) + O(\Delta x^2) \\ &= BU(t_n, x_j) + \frac{\Delta t}{2}A^2\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2} \\ &\quad + \frac{\Delta t}{2}(AB + BA)\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} + \frac{\Delta t}{2}B^2 U(t_n, x_j) \\ &\quad + O(\Delta t^2) + O(\Delta x^2). \end{aligned}$$

Nous n'explicitons pas, pour le moment le terme d'ordre 0, nous obtenons alors le schéma suivant :

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \left(A + \frac{\Delta t}{2}(AB + BA)\right)\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} + \frac{\Delta t}{2}A^2\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2} + f(U^n),$$

où $f(U^n)$ est une dicrétisation du terme d'ordre 0 telle que $f(U(t_n, x_j)) = BU(t_n, x_j) + \frac{\Delta t}{2}B^2 U(t_n, x_j) + O(\Delta t^2) + O(\Delta x^2)$.

Le symbole du schéma est alors, en posant $\gamma = \frac{\Delta t}{\Delta x}$:

$$\begin{aligned}
\widehat{Q}(\xi) &= Id + i\gamma\left(A + \frac{\Delta t}{2}(AB+BA)\right)\sin(\xi\Delta x) \\
&\quad -2\gamma^2 A^2(\sin(\xi\Delta x/2))^2 + \Delta t\widehat{f} \\
&= Id + \Delta t\left(i\frac{\sin(\xi\Delta x)}{\Delta x}A\right) + 2\Delta t^2\left(i\frac{\sin(\xi\Delta x/2)}{\Delta x}A + \frac{\sin(\xi\Delta x)}{4\sin(\xi\Delta x/2)}B\right)^2 \\
&\quad -\frac{\Delta t^2\sin(\xi\Delta x)^2}{8(\sin(\xi\Delta x/2))^2}B^2 + \Delta t\widehat{f} \\
&= Id + \Delta t\left(i\frac{\sin(\xi\Delta x)}{\Delta x}A\right) + 2\Delta t^2\left(i\frac{\sin(\xi\Delta x/2)}{\Delta x}A + \frac{\cos(\xi\Delta x/2)}{2}B\right)^2 \\
&\quad -\frac{\Delta t^2\cos(\xi\Delta x/2)^2}{2}B^2 + \Delta t\widehat{f} \\
&= Id + \Delta t\left(i\frac{\sin(\xi\Delta x)}{\Delta x}A\right) + \frac{\Delta t^2\cos(\xi\Delta x/2)^2}{2}\left(2i\frac{\tan(\xi\Delta x/2)}{\Delta x}A + B\right)^2 \\
&\quad -\frac{\Delta t^2\cos(\xi\Delta x/2)^2}{2}B^2 + \Delta t\widehat{f}.
\end{aligned}$$

Pour que le terme d'ordre 2 soit développable en P, nous posons : $z = 2i\frac{\tan(\xi\Delta x/2)}{\Delta x}$, on a alors :

$$\begin{aligned}
\widehat{Q}(\xi) &= Id + \Delta t(\cos(\xi\Delta x/2))^2(zA+B) + \frac{\Delta t^2\cos(\xi\Delta x/2)^2}{2}(zA+B)^2 \\
&\quad -\frac{\Delta t^2\cos(\xi\Delta x/2)^2}{2}B^2 - \Delta t(\cos(\xi\Delta x/2))^2 B + \Delta t\widehat{f}.
\end{aligned}$$

Nous choisissons alors :

$$\widehat{f} = (\cos(\xi\Delta x/2))^2 B + \frac{\Delta t\cos(\xi\Delta x/2)^2}{2}B^2,$$

ce qui donne :

$$f(U^n) = \left(Id + \frac{\Delta t}{2}B\right)B\frac{U^n_{j+1} + U^n_j + U^n_{j-1}}{4}.$$

Nous montrons alors le résultat suivant :

Proposition 4.1 *Le schéma de Lax-Wendroff suivant :*

$$\begin{aligned}
\tfrac{U^{n+1}_j - U^n_j}{\Delta t} &= \left(A + \tfrac{\Delta t}{2}(AB+BA)\right)\tfrac{U^n_{j+1} - U^n_{j-1}}{\Delta x} + \tfrac{\Delta t}{2}A^2\tfrac{U^n_{j+1} - 2U^n_j + U^n_{j-1}}{\Delta x^2} \\
&\quad + \left(Id + \tfrac{\Delta t}{2}B\right)B\tfrac{U^n_{j+1} + U^n_j + U^n_{j-1}}{4}
\end{aligned} \tag{4.1}$$

est développable en P.

PREUVE : Nous avons déjà montré que :

$$\widehat{Q}(\Delta t, \Delta x, \xi) = Id + \Delta t(\cos(\xi\Delta x/2))^2 P(z) + \frac{\Delta t^2 \cos(\xi\Delta x/2)^2}{2} P(z)^2,$$

où :

$$z = 2i\frac{\tan(\xi\Delta x/2)}{\Delta x}.$$

Donc :

$$\widehat{Q}(\Delta t, \Delta x, \xi) = Id + \Delta t P(z) + \frac{\Delta t^2}{2} P(z)^2 + \Delta t^3 R_2((\Delta t, \Delta x, \xi),$$

avec :

$$\begin{aligned} R_2((\Delta t, \Delta x, \xi) &= \frac{1}{\Delta t}(1 - \cos(\xi\Delta x/2)^2)\left(P(z) + \frac{1}{2}P(z)^2\right) \\ &= \frac{1}{\Delta t}\left(-\sum_{k=1}^{+\infty} \frac{(-1)^k \xi^{2k} \Delta t^{2k}}{\gamma^{2k} 2^{2k} k!}\right)\left(P(z) + \frac{1}{2}P(z)^2\right) \\ &= -\left(\sum_{k=1}^{+\infty} \frac{(-1)^k \xi^{2k} \Delta t^{2k-1}}{\gamma^{2k} 2^{2k} k!}\right)\left(P(z) + \frac{1}{2}P(z)^2\right). \end{aligned}$$

Donc R_2 est bien défini et il est clair que R_2 et ses dérivées commutent avec $P(z)$.

De plus,

$$\|\widehat{Q}(\xi) - Id\| \leq \Delta t\|P(z)\| + \frac{\Delta t^2}{2}\|P(z)\|^2.$$

Or $|\xi\Delta x| \leq \pi\Delta x^{1-\delta} \leq \pi h_0^{1-\delta}$ si $\delta < 1$. Donc, pour h_0 assez petit :

$$|z| = 2\frac{|\tan(\xi\Delta x/2)|}{\Delta x} \leq C\Delta x^{\delta}.$$

Donc :

$$\|\widehat{Q}(\xi) - Id\| \leq C(\Delta t(1 + \Delta x^{\delta}) + \Delta t^2(1 + \Delta x^{\delta})^2) \leq 1 - \varepsilon,$$

ce qui achève la démonstration. □

4.1.2 Etude de la stabilité

Proposition 4.2 *Si le problème de Cauchy est faiblement bien posé, le schéma de Lax-Wendroff (4.1) est faiblement stable sous la condition CFL :*

$$\forall\lambda \in \sigma(A),\ |\lambda\gamma| \leq 1.$$

L'intérêt de ce résultat est que la condition de stabilité ne fait intervenir que les valeurs propres de A qui sont aisément calculables.

PREUVE : Nous allons montrer la stabilité en utilisant la caractérisation de la proposition 3.2.

En effet, nous avons bien $\|\widehat{Q} - Id\| \leq K\Delta t(1 + \|\xi\|)^\theta$ avec $\theta = 1$. Il suffit donc de montrer que pour toute valeur propre $\mu(\xi)$ de $\widehat{Q}(\xi)$, nous avons $|\mu(\xi)| \leq e^{\alpha_S \Delta t}$.

Nous allons utiliser la proposition 1.5 qui permet d'avoir des renseignements sur les valeurs propres de $P(z)$.

Soit $\mu(\xi)$ une valeur propre de $\widehat{Q}$, elle s'écrit :

$$\mu(\xi) = 1 + \Delta t(\cos(\xi\Delta x/2))^2\lambda(z) + \frac{\Delta t^2 \cos(\xi\Delta x/2)^2}{2}\lambda(z)^2,$$

où $\lambda(z)$ est une valeur propre de $P(z)$. D'après la proposition 1.5 et comme le problème est faiblement bien posé, nous savons qu'il existe $C > 0$ tel que, si $|z| \geq C$:

$$\lambda(z) = \lambda_0 z + f(z),$$

avec f bornée.

$$\begin{aligned}
\mu(\xi) &= 1 + \Delta t(\cos(\xi\Delta x/2))^2\lambda(z) + \frac{\Delta t^2 \cos(\xi\Delta x/2)^2}{2}\lambda(z)^2 \\
&= 1 + \Delta t(\cos(\xi\Delta x/2))^2\lambda_0 z + \frac{\Delta t^2 \cos(\xi\Delta x/2)^2}{2}\lambda_0^2 z^2 \\
&\quad + \Delta t(\cos(\xi\Delta x/2))^2 f(z) + \frac{\Delta t^2 \cos(\xi\Delta x/2)^2}{2} f(z)^2 \\
&\quad + \Delta t^2 \cos(\xi\Delta x/2)^2 \lambda_0 z f(z) \\
&= 1 + i\gamma \sin(\xi\Delta x)\lambda_0 - 2\gamma^2 \sin(\xi\Delta x/2)^2\lambda_0^2 \\
&\quad + \Delta t \left((\cos(\xi\Delta x/2))^2 + \frac{\Delta t \cos(\xi\Delta x/2)^2}{2} f(z) + i\gamma \sin(\xi\Delta x)\lambda_0 \right) f(z).
\end{aligned}$$

Comme, pour $|z| \geq C$, $f(z) \leq M$, nous avons :

$$\begin{aligned}
|\mu(\xi)| &\leq |1 + i\gamma \sin(\xi\Delta x)\lambda_0 - 2\gamma^2 \sin(\xi\Delta x/2)^2\lambda_0^2| + \Delta t \left(1 + \frac{\Delta t}{2} M + \gamma|\lambda_0| \right) M \\
&\leq |1 + i\gamma \sin(\xi\Delta x)\lambda_0 - 2\gamma^2 \sin(\xi\Delta x/2)^2\lambda_0^2| + K\Delta t.
\end{aligned}$$

Or sous la condition CFL, nous avons $|1 + i\gamma \sin(\xi\Delta x)\lambda_0 - 2\gamma^2 \sin(\xi\Delta x/2)^2\lambda_0^2| \leq 1$ car ce terme correspond à un schéma de Lax-Wendroff classique pour l'équation $\partial_t u = \lambda_0 \partial_x u$. Nous avons donc montré que pour $|z| \geq C$:

$$|\mu(\xi)| \leq 1 + K\Delta t \leq e^{Kt}.$$

Lorsque $|z| \leq C$, le résultat est clair car alors $\lambda(z)$ est bornée.
Nous avons donc bien la stabilité du schéma de Lax-Wendroff. □

4.1.3 Taux de convergence

Théorème 4.1 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, et si le schéma de Lax-Wendroff (4.1) est faiblement stable de défaut q_2 sous la condition CFL $\forall \lambda \in \sigma(A)$, $|\lambda\gamma| \leq 1$, alors, pour une donnée initiale dans H^{q_4}, le taux de convergence pour le schéma de Lax-Wendroff (4.1) est :*

$$\beta = \frac{2(q_4 - \max(q_1, q_2))}{\max(q_4, 3 + \max(q_1, q_2)) - \max(q_1, q_2)}.$$

PREUVE : Nous allons appliquer le théorème 3.1. Il suffit donc de montrer la relation (3.9) :

$$\left\| e^{P(i\xi)t_n} - \widehat{Q}^n(\xi) \right\| \leq C t_n h^s (1 + \|\xi\|^2)^{\sigma/2},$$

avec $s = 2$ et $\sigma = 3 + \max(q_1, q_2)$.
Pour cela,nous allons appliquer le théorème 3.2 et montrer que $N_0 = r = 2$ et $\rho = 3$ ce qui permettra de conclure.

- Montrons que : $\forall j \geq 0$, $\exists B_k > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$, $\left\| \frac{\partial^j R_{N_0}(\Delta t, \Delta x, \xi)}{\partial \Delta t^j} \right\| \leq B_k(1 + |\xi|^{3+j})$.
  Nous avons :

$$\begin{aligned}
\left\| \frac{\partial^j R_2(\Delta t, \Delta x, \xi)}{\partial \Delta t^j} \right\| &= \left\| - \left(\sum_{k \geq (j+1)/2} \frac{(-1)^k \xi^{2k} (2k-1)! \Delta t^{2k-1-j}}{\gamma^{2k} 2^{2k} k! (2k-1-j)!} \right) \right. \\
&\quad \left. \times \left(P(z) + \frac{1}{2} P(z)^2 \right) \right\| \\
&\leq |\xi|^{j+1} \left(\sum_{k \geq (j+1)/2} \frac{(2k-1)! (|\xi| \Delta t)^{2k-1-j}}{\gamma^{2k} 2^{2k} k! (2k-1-j)!} \right) \\
&\quad \times C(1 + |\xi|^2) \\
&\leq B_j (1 + |\xi|^{3+j}).
\end{aligned}$$

- Montrons que $\exists C > 0$, $\forall \Delta x$, $\Delta x \leq h_0$, $\forall \Delta t$, $\Delta t \leq k_0$, $\forall \xi$, $|\xi| \leq \frac{\pi}{\Delta x^\delta}$, $|z| \leq C|\xi|$. Nous avons $|z| = 2\frac{|\tan(\xi \Delta x/2)|}{\Delta x}$. Or pour x assez petit, $|\tan(x)| \leq 2|x|$. En prenant alors $x = \xi \Delta x/2$ et en remarquant que $|x| \leq \pi \Delta x^{1-\delta}/2 \leq \pi h_0^{1-\delta}/2$, nous avons, pour h_0 assez petit $|z| \leq 2|\xi|$.

- Montrons que $\forall|\xi| \leq \frac{\pi}{\Delta x^\delta}$, $|z - i\xi| \leq K\Delta x^2|\xi|^3$ ce qui montrera que $r = 2$ et $\rho = 3$.

$$z = 2i\frac{\tan(\xi\Delta x/2)}{\Delta x} = i\xi + O(\xi^3\Delta x^2),$$

ce qui donne la conclusion.

□

Remarque 4.1 *Le taux que l'on a obtenu ici est meilleur que celui du théorème 3.4 qui prenait $s = 2$ et $\sigma = 3 + q_1 + q_2$. Les résultats numériques vont montrer que ce taux est optimal.*

4.1.4 Résultats numériques

Les explications sur les calculs numériques concernent tous les schémas étudiés. Nous allons effectuer des calculs numériques sur trois exemples :

- **Exemple 1 :** $q_1 = 2$

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}$$

- **Exemple 2 :** $q_1 = 1$

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

- **Exemple 3 :** $q_1 = 2$

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 1 & 0 \\ -2 & 1 & -1 \\ 0 & 2 & 1 \end{pmatrix}$$

Nous allons effectuer toutes les simulations numériques sur des matrices de taille 3. Nous allons donc avoir besoin de la valeur de q_2 pour pouvoir comparer les résultats numériques et théoriques.

Proposition 4.3 *Si A, $B \in \mathcal{M}_3(\mathbb{R})$ sont de la forme donnée dans la proposition 1.8, et si nous considérons un schéma faiblement stable de matrice d'amplification telle que $\widehat{Q}(\xi) - \mu(\xi)Id = i\xi\Delta t(A - \lambda Id) + O(\Delta t)$ pour toute valeur propre $\mu(\xi)$ de $\widehat{Q}(\xi)$ et où λ est une valeur propre de A, alors le schéma est faiblement stable de défaut $q_2 = q_1$.*

PREUVE : De même que dans la proposition 1.8, si μ_1, μ_2, μ_3, désignent les valeurs propres de $\widehat{Q}$, nous avons :

- si $\mu_1 \neq \mu_2 \neq \mu_3$,

$$\begin{aligned}\widehat{Q}^n &= \frac{\mu_1^n}{(\mu_2-\mu_1)(\mu_3-\mu_1)}(\widehat{Q}-\mu_2 Id)(\widehat{Q}-\mu_3 Id)\\ &+\frac{\mu_2^n}{(\mu_2-\mu_1)(\mu_2-\mu_3)}(\widehat{Q}-\mu_3 Id)(\widehat{Q}-\mu_1 Id)\\ &+\frac{\mu_3^n}{(\mu_3-\mu_1)(\mu_3-\mu_2)}(\widehat{Q}-\mu_1 Id)(\widehat{Q}-\mu_2 Id),\end{aligned}$$

- si $\mu_1 = \mu_2 \neq \mu_3$,

$$\widehat{Q}^n = -n\frac{\mu_1^{n-1}}{(\mu_3-\mu_1)}(\widehat{Q}-\mu_1 Id)(\widehat{Q}-\mu_3 Id)+\frac{\mu_3^n}{(\mu_3-\mu_1)^2}(\widehat{Q}-\mu_1 Id)^2,$$

- si $\mu_1 = \mu_2 = \mu_3$,

$$\widehat{Q}^n = \frac{n(n-1)}{2}\mu_1^{n-2}(\widehat{Q}-\mu_1 Id)^2.$$

Nous calculons alors, comme dans la proposition 1.8 le développement limité de $\widehat{Q}^n$, en remplaçant $P(i\xi)-\lambda(\xi)$ par $\frac{\widehat{Q}(\xi)-\mu(\xi)}{\Delta t}$ qui vérifient bien les mêmes propriétés par hypothèse. En utilisant $n\Delta t = t_n$, nous obtenons alors un résultat analogue à celui de la proposition 1.8.

□

Proposition 4.4 *Sous condition CFL, le schéma de Lax-Wendroff (4.1) vérifie les hypothèses de la proposition précédente.*

PREUVE : Nous avons :

$$\widehat{Q}(\xi)-\mu(\xi)Id = \Delta t(\cos(\xi\Delta x/2))^2(P(z)-\lambda(z)Id)+\frac{\Delta t^2(\cos(\xi\Delta x/2))^2}{2}(P(z)^2-\lambda(z)^2 Id)$$

Or, d'après la condition CFL, $(\cos(\xi\Delta x/2))^2 = 1+O(\xi^2\Delta t^2)$, de plus $P(z) = i\xi A + O(1)$ par définition, et $\lambda(z) = i\xi\lambda + O(1)$, donc $\widehat{Q}(\xi)-\mu(\xi)Id = i\xi\Delta t(A-\lambda Id)+O(\Delta t)$.

□

Nous allons maintenant effectuer les calculs numériques.

Sur chacun de ces exemples, nous allons calculer le taux de convergence numérique. Pour cela, nous allons utiliser différentes données initiales. Ces données

initiales sont toutes à support compact dans $[0,1]$ et elles ont pour régularités respectives 3/2, 5/2, 7/2, 9/2 et 11/2.

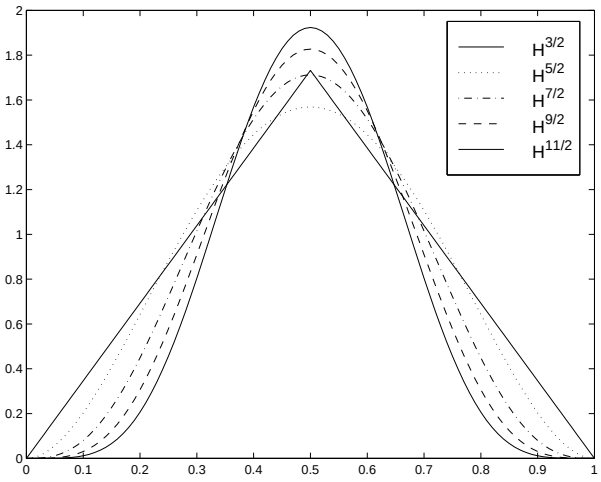


FIG. 4.1 – Conditions initiales

Pour chaque donnée initiale, nous calculons l'erreur à l'instant $t = 0,9$ sous la condition CFL $\gamma = 0,9$ pour des pas d'espace variant entre $10^{-3,5}$ et 10^{-2}. Ces valeurs sont suffisamment petites pour représenter le comportement asymptotique de l'erreur. Le taux de convergence numérique est alors la pente de cette droite.

Nous comparerons les résultats numériques avec le taux général β_2 du théorème 3.4 et avec le taux optimisé β calculé dans les exemples de schémas traités.

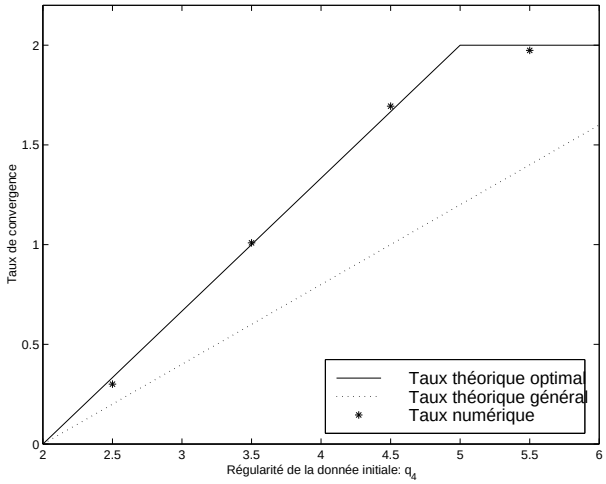


FIG. 4.2 – Schéma de Lax-Wendroff, exemple 1 : $q_2 = 2$

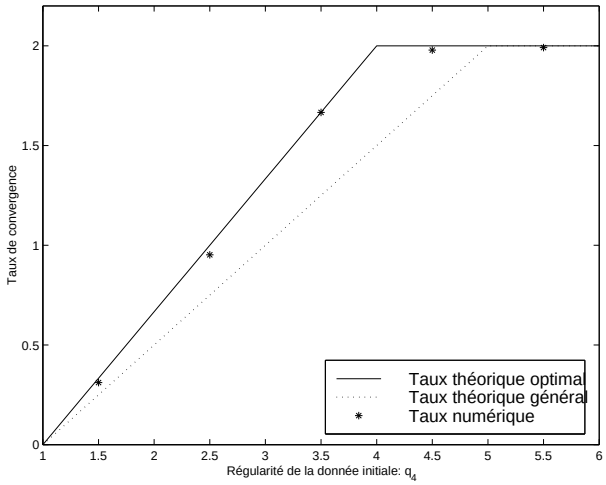


FIG. 4.3 – Schéma de Lax-Wendroff, exemple 2 : $q_2 = 1$

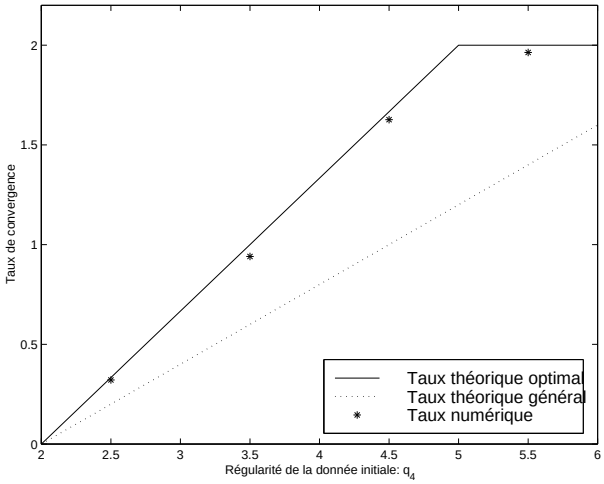


FIG. 4.4 – Schéma de Lax-Wendroff, exemple 3 : $q_2 = 2$

Nous remarquons que le taux calculé au théorème 4.1 est bien optimal alors que celui calculé au théorème 3.4 n'est que minimal.

4.1.5 Remarque sur le schéma de Lax-Wendroff usuel

Si nous discrétisons les termes $U(t_n, x_j)$ par la discrétisation standard U_j^n, nous obtenons le schéma de Lax-Wendroff usuel :

$$\begin{aligned}\frac{U_j^{n+1} - U_j^n}{\Delta t} &= \left(A + \frac{\Delta t}{2}(AB + BA)\right)\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} \\ &\quad + \frac{\Delta t}{2}A^2\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2} + BU_j^n + \frac{\Delta t}{2}B^2U_j^n.\end{aligned}$$

Ce schéma n'est pas développable en P toutefois, le taux de convergence observé numériquement est égal à celui calculé dans le théorème 3.2. Il semble donc que la condition d'être développable en P ne soit pas nécessaire pour que le taux de convergence soit celui calculé dans le théorème 3.2 mais uniquement suffisante.

4.2 Schéma de Crank-Nicolson

4.2.1 Ecriture d'un schéma de Crank-Nicolson

Nous utilisons ici la discrétisation standard du terme d'ordre 0. Le schéma de Crank-Nicolson est donc le suivant :

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = A\frac{U_{j+1}^n - U_{j-1}^n + U_{j+1}^{n+1} - U_{j-1}^{n+1}}{4\Delta x} + B\frac{U_j^n + U_j^{n+1}}{2} \tag{4.2}$$

La matrice d'amplification du schéma est alors :

$$\widehat{Q}(\xi) = \left(Id - \frac{\Delta t}{2}(i\frac{\sin(\xi\Delta x)}{\Delta x}A + B) \right)^{-1} \left(Id + \frac{\Delta t}{2}(i\frac{\sin(\xi\Delta x)}{\Delta x}A + B) \right).$$

Posons $z = i\frac{\sin(\xi\Delta x)}{\Delta x}$, alors $\widehat{Q}(\xi)$ s'écrit :

$$\widehat{Q}(\xi) = \left(Id - \frac{\Delta t}{2}P(z) \right)^{-1} \left(Id + \frac{\Delta t}{2}P(z) \right).$$

Proposition 4.5 *Le schéma de Crank-Nicolson (4.2) est développable en P.*

PREUVE :

Nous développons $\left(Id - \frac{\Delta t}{2}P(z)\right)^{-1}$ en série de Neumann ce qui est licite car $\frac{\Delta t}{2}\|P(z)\| \leq \frac{\Delta t}{2}(|\xi|\|A\| + \|B\|) \leq \frac{1}{2}(\gamma\pi h_0^{1-\delta}\|A\| + \gamma h_0\|B\|) < 1$ pour h_0 assez petit.

$$\begin{aligned}
\widehat{Q}(\xi) &= \left(Id - \frac{\Delta t}{2}P(z) \right)^{-1} \left(Id + \frac{\Delta t}{2}P(z) \right) \\
&= \left(\sum_{k=0}^{+\infty} \frac{\Delta t^k}{2^k} P(z)^k \right) \left(Id + \frac{\Delta t}{2}P(z) \right) \\
&= Id + \sum_{k=1}^{+\infty} \frac{\Delta t^k}{2^{k-1}} P(z)^k \\
&= Id + P(z) + \frac{\Delta t}{2}P(z)^2 + \Delta t^3 R_2,
\end{aligned}$$

avec :

$$R_2 = \sum_{k=3}^{+\infty} \frac{\Delta t^{k-3}}{2^{k-1}} P(z)^k = \sum_{k=0}^{+\infty} \frac{\Delta t^k}{2^{k+2}} P(z)^{k-3}.$$

Et R_2 et ses dérivées commutent bien avec $P(z)$. De plus, nous avons :

$$\begin{aligned}
\|\widehat{Q}(\xi) - Id\| &\leq \sum_{k=1}^{+\infty} \frac{\Delta t^k}{2^{k-1}} \|P(z)\|^k \\
&= \Delta t\|P(z)\| \left(1 - \frac{\Delta t}{2}\|P(z)\| \right)^{-1}.
\end{aligned}$$

Or, pour tout $\eta > 0$, il existe $h_0 > 0$ tel que $\frac{\Delta t}{2}\|P(z)\| \leq \eta$, ainsi

$$\|\widehat{Q}(\xi) - Id\| \leq \frac{2\eta}{1-\eta}.$$

Donc, il existe $\varepsilon > 0$ tel que pour h_0 assez petit nous avons :

$$\|\widehat{Q}(\xi) - Id\| < 1 - \varepsilon,$$

ce qui achève la démonstration.

□

4.2.2 Etude de la stabilité

Proposition 4.6 *Si le problème de Cauchy est faiblement bien posé, le schéma de Crank-Nicolson (4.2) est stable.*

PREUVE : Nous utilisons la même méthode de démonstration et les mêmes notations que pour le schéma de Lax-Wendroff. Soit $\mu(\xi)$ une valeur propre de $\widehat{Q}$, elle s'écrit :

$$\mu(\xi) = \frac{1 + \frac{\Delta t}{2}\lambda(z)}{1 - \frac{\Delta t}{2}\lambda(z)}$$

où $\lambda(z)$ est une valeur propre de $P(z)$. En utilisant la proposition 1.5, nous avons :

$$\begin{aligned}\mu(\xi) &= \frac{1 + \frac{\Delta t}{2}\lambda_0 z + \frac{\Delta t}{2}f(z)}{1 - \frac{\Delta t}{2}\lambda_0 z - \frac{\Delta t}{2}f(z)} \\ &= \frac{1 + i\frac{\gamma}{2}\lambda_0 \sin(\xi\Delta x) + \frac{\Delta t}{2}f(z)}{1 - i\frac{\gamma}{2}\lambda_0 \sin(\xi\Delta x) - \frac{\Delta t}{2}f(z)}.\end{aligned}$$

Donc :

$$|\mu(\xi)| = \frac{|1 + \frac{\Delta t}{2}\frac{f(z)}{1+i\frac{\gamma}{2}\lambda_0 \sin(\xi\Delta x)}|}{|1 - \frac{\Delta t}{2}\frac{f(z)}{1-i\frac{\gamma}{2}\lambda_0 \sin(\xi\Delta x)}|}.$$

Or, nous avons, pour $|z| \geq C$:

$$\left|\frac{\Delta t}{2}\frac{f(z)}{1 + i\frac{\gamma}{2}\lambda_0 \sin(\xi\Delta x)}\right| \leq \frac{\Delta t}{2}M \text{ et } \left|\frac{\Delta t}{2}\frac{f(z)}{1 - i\frac{\gamma}{2}\lambda_0 \sin(\xi\Delta x)}\right| \leq \frac{\Delta t}{2}M,$$

où $M = \sup_{|z|\geq C} |f(z)|$. Ainsi, si $\Delta t < \frac{2}{M}$:

$$\begin{aligned}|\mu(\xi)| &\leq \frac{1 + \frac{\Delta t}{2}M}{1 - \frac{\Delta t}{2}M} \\ &\leq \exp(\frac{3M\Delta t}{2}).\end{aligned}$$

Et, de même que pour Lax-Wendroff, lorsque $|z| \leq C$, comme $\lambda(z)$ est bornée, le résultat est vrai.
Nous avons donc bien montré la stabilité du schéma de Crank-Nicolson. □

4.2.3 Taux de convergence

Théorème 4.2 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, et si le schéma de Crank-Nicolson (4.2) est faiblement stable de défaut q_2, alors, pour une donnée initiale dans H^{q_4}, le taux de convergence pour le schéma de Crank-Nicolson (4.2) est :*

$$\beta = \frac{2(q_4 - \max(q_1, q_2))}{\max(q_4, 3 + \max(q_1, q_2)) - \max(q_1, q_2)}.$$

PREUVE : De même que pour le schéma de Lax-Wendroff, il suffit de montrer que $N_0 = r = 2$ et $\rho = 3$.

- Montrons que $\exists C > 0, \forall \Delta x,\ \Delta x \leq h_0, \forall \Delta t,\ \Delta t \leq k_0, \forall \xi,\ |\xi| \leq \frac{\pi}{\Delta x^\delta},\ |z| \leq C|\xi|$. Par définition, $|z| = \frac{|\sin(\xi \Delta x|}{\Delta x} \leq |\xi|$. Donc $C = 1$ convient.
- Montrons que : $\forall j \geq 0, \exists B_k > 0, \forall \Delta x,\ \Delta x \leq h_0, \forall \Delta t,\ \Delta t \leq k_0, \forall \xi,\ |\xi| \leq \frac{\pi}{\Delta x^\delta},\ \left\| \frac{\partial^j R_{N_0}(\Delta t, \Delta x, \xi)}{\partial \Delta t^j} \right\| \leq B_k(1 + |\xi|^{3+j})$.
  Nous avons :

$$\begin{aligned}
\left\| \frac{\partial^j R_{N_0}(\Delta t, \Delta x, \xi)}{\partial \Delta t^j} \right\| &= \left\| \sum_{k=j}^{+\infty} \frac{k! \Delta t^{k-j}}{2^{k+2}(k-j)!} P(z)^{k-3} \right\| \\
&\leq \sum_{k=j}^{+\infty} \frac{k! \Delta t^{k-j}}{2^{k+2}(k-j)!} C(1 + |\xi|^{k-3}) \\
&\leq C'(1 + |\xi|^{j+3}) \sum_{k=j}^{+\infty} \frac{k!(|\xi| \Delta t)^{k-j}}{2^{k+2}(k-j)!} \\
&\leq B_j(1 + |\xi|^{j+3}).
\end{aligned}$$

- Montrons que $\forall |\xi| \leq \frac{\pi}{\Delta x^\delta},\ |z - i\xi| \leq K \Delta x^2 |\xi|^3$ ce qui montrera que $r = 2$ et $\rho = 3$.

$$z = i \frac{\sin(\xi \Delta x)}{\Delta x} = i\xi + O(\xi^3 \Delta x^2),$$

  ce qui donne la conclusion.

□

4.2.4 Résultats numériques

Nous effectuons les mêmes calculs que pour le schéma de Lax-Wendroff.

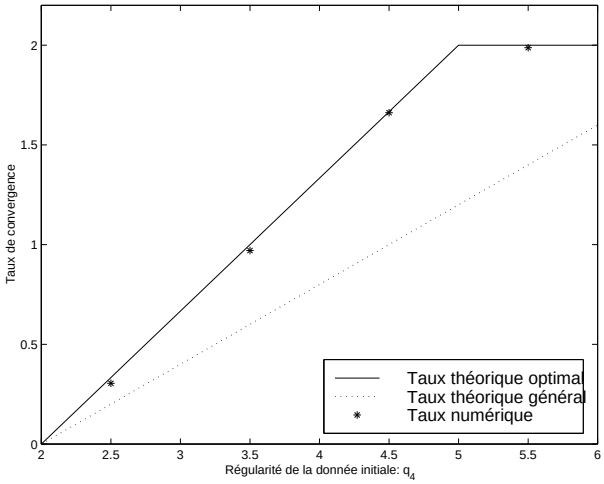


FIG. 4.5 – Schéma de Crank-Nicolson, exemple 1 : $q_2 = 2$

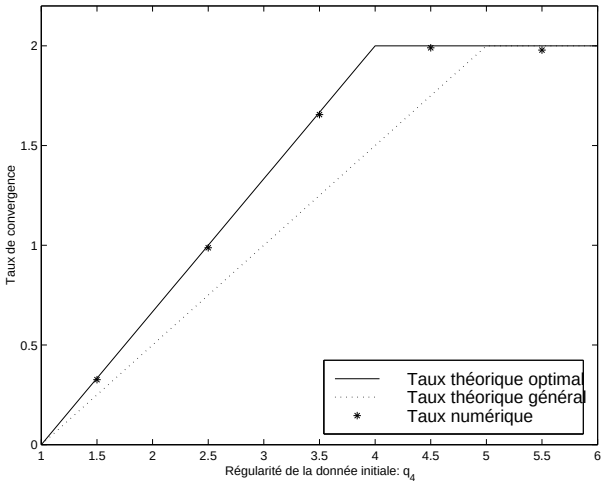


FIG. 4.6 – Schéma de Crank-Nicolson, exemple 2 : $q_2 = 1$

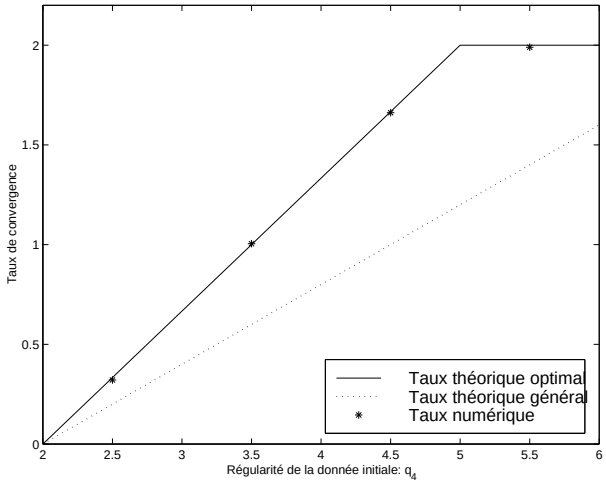


FIG. 4.7 – Schéma de Crank-Nicolson, exemple 3 : $q_2 = 2$

4.2.5 Présentation rapide d'un autre exemple : le schéma Box-Scheme

Nous allons étudier le schéma Box-Scheme suivant :

$$\begin{aligned} \frac{(V_j^{n+1}+V_{j+1}^{n+1})-(V_j^n+V_{j+1}^n)}{2\Delta t} &= A\frac{(V_{j+1}^{n+1}-V_j^{n+1})+(V_{j+1}^n-V_j^n)}{2\Delta x} \\ &+ B\frac{(V_j^{n+1}+V_{j+1}^{n+1})+(V_j^n+V_{j+1}^n)}{4}. \end{aligned} \tag{4.3}$$

Nous avons alors :

$$\left(Id - \Delta t\frac{e^{i\xi\Delta x}-1}{\Delta x(1+e^{i\xi\Delta x})}A - \frac{\Delta t}{2}B\right)\widehat{V^{n+1}} = \left(Id + \Delta t\frac{e^{i\xi\Delta x}-1}{\Delta x(1+e^{i\xi\Delta x})}A + \frac{\Delta t}{2}B\right)\widehat{V^n}.$$

D'où :

$$\widehat{Q}(\xi) = \left(Id + \frac{\Delta t}{2}P(z)\right)^{-1}\left(Id - \frac{\Delta t}{2}P(z)\right),$$

avec ici :

$$z = \frac{2(e^{i\xi\Delta x}-1)}{\Delta x(1+e^{i\xi\Delta x})}.$$

Ce schéma a une forme analogue au schéma de Crank-Nicolson, et nous pouvons donc montrer les résultats suivants :

Proposition 4.7 *Si le problème de Cauchy est faiblement bien posé, le schéma Box-Scheme (4.3) est stable.*

Théorème 4.3 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, et si le schéma Box-Scheme (4.3) est faiblement stable de défaut q_2, alors, pour une donnée initiale dans H^{q_4}, le taux de convergence pour le schéma Box-Scheme (4.3) est :*

$$\beta = \frac{2(q_4 - \max(q_1, q_2))}{\max(q_4, 3 + \max(q_1, q_2)) - \max(q_1, q_2)}.$$

4.3 Schéma de Lax-Friedrichs

Nous allons étudier dans cette partie un schéma de Lax-Friedrichs sous la condition $\frac{\Delta t}{\Delta x} = \gamma$ constant, alors le schéma de Lax-Friedrichs est d'ordre 1. La discrétisation classique du terme d'ordre 0 va donner un schéma développable en P. Nous avons le schéma de Lax-Friedrichs suivant :

$$\frac{V_j^{n+1} - \frac{1}{2}(V_{j+1}^n + V_{j-1}^n)}{\Delta t} = A\frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x} + BV_j^n. \tag{4.4}$$

Nous avons alors :

$$\widehat{Q}(\xi) = \cos(\xi\Delta x)Id + \Delta t P(z),$$

avec $z = i\frac{\sin(\xi\Delta x)}{\Delta x}$.

Proposition 4.8 *Le schéma de Lax-Friedrichs (4.4) est développable en P.*

PREUVE : Nous écrivons le schéma sous la forme :

$$\begin{aligned} \widehat{Q}(\Delta t, \Delta x, \xi) &= \cos(\xi\Delta x)Id + \Delta t P(z) \\ &= Id + \Delta t P(z) + \Delta t^2 R_1. \end{aligned}$$

avec :

$$\begin{aligned} R_1 &= \frac{1}{\Delta t^2}(1 - \cos(\xi\Delta x))Id \\ &= \sum_{k=0}^{+\infty} \frac{\xi^{2(k+1)}(2k)!}{\gamma^{2(k+1)}(2(k+1))!}\Delta t^{2k}, \end{aligned}$$

qui a bien les propriétés voulues. De plus, nous avons :

$$\|\widehat{Q}(\xi) - Id\| \leq |\cos(\xi\Delta x) - 1| + \Delta t\|P(z)\|.$$

Or $|\xi\Delta x| \leq \pi\Delta x^{1-\delta} \leq \pi h_0^{1-\delta}$ si $\delta < 1$.
Donc, pour h_0 assez petit, nous avons la conclusion. □

4.3.1 Etude de la stabilité

Proposition 4.9 *Si le problème de Cauchy est faiblement bien posé, le schéma de Lax-Friedrichs (4.4) est stable sous la condition CFL :*

$$\forall \lambda \in \sigma(A),\ |\lambda\gamma| \leq 1.$$

PREUVE : De même que pour les schémas précédents :

$$\mu(\xi) = \cos(\xi\Delta x) + \Delta t(\lambda_0 z + f(z)).$$

Donc, pour $|z| \geq C$:

$$\|\mu(\xi)\| \leq |\cos(\xi\Delta x) + i\gamma\lambda_0 \sin(\xi\Delta x)| + M\Delta t.$$

Donc, sous condition CFL, nous avons bien :

$$\|\mu(\xi)\| \leq 1 + M\Delta t \leq e^{M\Delta t},$$

ce qui prouve la stabilité. □

4.3.2 Taux de convergence

Théorème 4.4 *Si le problème de Cauchy est faiblement bien posé de défaut q_1, et si le schéma de Lax-Friedrichs (4.4) est faiblement stable de défaut q_2 sous la condition CFL $\forall \lambda \in \sigma(A)$, $|\lambda\gamma| \leq 1$, alors, pour une donnée initiale dans H^{q_4}, le taux de convergence pour le schéma de Lax-Friedrichs (4.4) est :*

$$\beta = \frac{(q_4 - \max(q_1, q_2))}{\max(q_4, 2 + \max(q_1, q_2)) - \max(q_1, q_2)}.$$

PREUVE : De même que pour le schéma de Lax-Wendroff, il suffit de montrer que $N_0 = r = 1$ et $\rho = 2$.

- Evaluons les dérivées de R_1 :

$$\begin{aligned} \left\| \frac{\partial^j R_1(\Delta t, \Delta x, \xi)}{\partial \Delta t^j} \right\| &= \left\| \sum_{k=j}^{+\infty} \frac{\xi^{2(k+1)}(2k)!}{\gamma^{2(k+1)}(2(k+1))!(2k-j)!} \Delta t^{2k-j} \right\| \\ &\leq B_j(1 + |\xi|^{j+2}). \end{aligned}$$

- Montrons que $\forall |\xi| \leq \frac{\pi}{\Delta x^\delta}$, $|z - i\xi| \leq K\Delta x^2 |\xi|^3$ ce qui montrera que $r = 1$ et $\rho = 2$.

$$z = i\frac{\sin(\xi\Delta x)}{\Delta x} = i\xi + O(\xi^3\Delta x^2) = i\xi + O(\xi^2\Delta x) \text{ si } \delta < 1,$$

ce qui donne la conclusion.

□

4.4 Schéma décentré

Supposons que nous avons écrit la matrice A sous forme de blocs :

$$A = \begin{pmatrix} A_+ & 0 \\ 0 & A_- \end{pmatrix},$$

où A_+ de dimension $N_+ \times N_+$ (resp. A_- de dimension $N_- \times N_-$) est constitué des blocs de Jordan de valeurs propres positives (resp. strictement négatives).

Nous décentrons les dérivées en x selon les caractéristiques, et discrétisons de façon standard le terme d'ordre 0. Le schéma décentré s'écrit alors :

$$\frac{V_j^{n+1} - V_j^n}{\Delta t} = \begin{pmatrix} A_+ & 0 \\ 0 & 0 \end{pmatrix} \frac{V_{j+1}^n - V_j^n}{\Delta x} + \begin{pmatrix} 0 & 0 \\ 0 & A_- \end{pmatrix} \frac{V_j^n - V_{j-1}^n}{\Delta x} + BV_j^n. \qquad (4.5)$$

Sa matrice d'amplification est donc :

$$\widehat{Q}(\xi) = Id + \Delta t \begin{pmatrix} \frac{e^{i\xi\Delta x}-1}{\Delta x} A_+ & 0 \\ 0 & \frac{1-e^{-i\xi\Delta x}}{\Delta x} A_- \end{pmatrix} + \Delta t B.$$

Nous ne pouvons pas ici utiliser les résultats du chapitre précédent, car $\widehat{Q}$ n'est pas développable en P. Il dépend en fait de deux paramètres, comme nous allons le voir maintenant. Pour étudier les valeurs propres de $\widehat{Q}$, nous posons

$$\mathcal{P}(\xi, u) = \xi \begin{pmatrix} \frac{e^{iu}-1}{u} A_+ & 0 \\ 0 & \frac{1-e^{-iu}}{u} A_- \end{pmatrix} + B,$$

avec $u = \xi\Delta x$. Nous avons alors $\widehat{Q}(\xi) = Id + \Delta t \mathcal{P}(\xi, \xi\Delta x)$, et nous étudions les valeurs propres de $\mathcal{P}(\xi, u)$, où nous supposerons $\xi \geq 0$ et $u \geq 0$ (quitte à échanger les signes). Nous faisons le changement de variables

$$x = \frac{u}{\sin u} \frac{1}{\xi}, \; y = \frac{1 - \cos u}{u},$$

et définissons la matrice

$$M(x, y) = \begin{pmatrix} (i-y)A_+ & 0 \\ 0 & (i+y)A_- \end{pmatrix} + xB.$$

Alors $\mathcal{P}(\xi, \xi\Delta x) = \frac{1}{x} M(x, y)$. Lorsque $\xi \to +\infty$ et $\xi\Delta x \to 0$, ce qui est le cas qui nous intéresse, $x \sim \frac{1}{\xi} \to 0$ et $y \sim \frac{\xi\Delta x}{2} \to 0$. La matrice M est donc une perturbation de iA mais, comme cette perturbation dépend de deux variables, la théorie des séries de Puiseux appliquée précédemment n'est plus valable. Nous allons obtenir un résultat asymptotique à l'aide du théorème de Seidenberg-Tarski. La matrice B est décomposée par blocs suivant les blocs A^+ et A^-, $B = \left(\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right)$.

Lemme 4.1 *Il existe une transformation analytique $\tilde{S}$ dans les variables (x,y) telle que :*

$$\tilde{S}^{-1}(x,y)M(x,y)\tilde{S}(x,y) = \begin{pmatrix} \widetilde{M}_{11}(x,y) & 0 \\ 0 & \widetilde{M}_{22}(x,y) \end{pmatrix} = \widetilde{M}(x,y),$$

avec $\tilde{S}(x,y) = \begin{pmatrix} Id & xS(x,y) \\ 0 & Id \end{pmatrix} \begin{pmatrix} Id & 0 \\ xT(x,y) & Id \end{pmatrix}$.

PREUVE : L'analyticité de la transformation résulte de théorèmes généraux, puisque la matrice M est elle-même analytique [14]. Nous allons écrire explicitement la transformation, car nous aurons besoin de connaître les premiers termes du développement. Nous éliminons d'abord le bloc supérieur droit. Remarquons d'abord que $\left(\begin{smallmatrix} I & xS \\ 0 & I \end{smallmatrix}\right)^{-1} = \left(\begin{smallmatrix} I & -xS \\ 0 & I \end{smallmatrix}\right)$. Nous calculons pour toute matrice S,

$$\begin{pmatrix} Id & -xS(x,y) \\ 0 & Id \end{pmatrix} M(x,y) \begin{pmatrix} Id & xS(x,y) \\ 0 & Id \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

Avec :

$$\begin{aligned}
R_{11} &= (i-y)A_+ + xB_{11} - x^2S(x,y)B_{21}, \\
R_{12} &= [(i-y)A_+ + xB_{11} - x^2S(x,y)B_{21}]xS(x,y) + xB_{12} - xS(x,y)[(i+y)A_- + xB_{22}], \\
R_{21} &= xB_{21}, \\
R_{22} &= x^2S(x,y)B_{21} + (i+y)A_- + xB_{22}.
\end{aligned}$$

Nous cherchons S sous la forme $S = \sum_{k=0}^{+\infty} S_k(y)x^k$, de façon que $R_{12} = 0$. Nous définissons sur l'espace vectoriel complexe des matrices $N_+ \times N_-$, l'opérateur Δ par $\Delta X = (i-y)A_+X - (i+y)XA_-$. Puisque $(i-y)A_+$ et $-(i+y)A_-$ n'ont pas les mêmes valeurs propres, Δ est un isomorphisme. Nous obtenons les relations de récurrence suivantes :

$$\begin{aligned}
&\Delta S_0(y) + B_{12} = 0, \\
&\Delta S_1(y) + B_{21}S_0(y) - S_0(y)B_{22} = 0, \\
&\Delta S_k(y) + B_{11}S_{k-1}(y) - \sum_{i+j=k-2} S_j(y)B_{21}S_j(y) - S_{k-1}(y)B_{22},\ k \geq 2,
\end{aligned}$$

qui définissent la suite S_k de façon unique. Nous définissons maintenant les opérateurs $\Delta^+X = A_+X - XA_-$ et $\Delta^-X = A_+X + XA_-$, et nous décomposons S_0 sous la forme $\sum_{j=0}^{+\infty} S_0^j y^j$. Nous obtenons la relation de récurrence suivante :

$$\begin{aligned}
&\Delta^+ S_0^0 = iB_{12}, \\
&\Delta^+ S_0^j = i\Delta^- S_0^{j-1},\ j \geq 1.
\end{aligned} \tag{4.6}$$

ceci nous détermine le développement en série de S dans les variables x et y. En effectuant le même calcul pour T, nous obtenons

$$\begin{pmatrix} Id & 0 \\ -xT(x,y) & Id \end{pmatrix} M(x,y) \begin{pmatrix} Id & 0 \\ xT(x,y) & Id \end{pmatrix} = \begin{pmatrix} R'_{11} & 0 \\ R'_{21} & R'_{22} \end{pmatrix}.$$

Avec :

$$\begin{aligned} R'_{11} &= R_{11}, \\ R'_{21} &= x(-TR_{11} + B_{21} + R_{22}TR_{22}), \\ R'_{22} &= x(B_{21} + R_{22}TR_{22}). \end{aligned}$$

La première ligne de blocs n'est donc pas affectée par cette opération, et nous ne travaillons que sur la deuxième ligne, ou nous devons définir T de façon à annuler R'_{21}, ce qui se fait comme précédemment. □

$\widetilde{M}_{11}(x,y)$ est donnée par R_{11}, et si nous regardons les premiers termes du développement en séries, nous obtenons

$$\widetilde{M}_{11}(x,y) = (i-y)A_+ + xB_{11} - x^2 S_0^0 B_{21} - x^2 y S_0^1 B_{21} + \cdots \tag{4.7}$$

Nous notons $\tilde{\mu_1}(x,y)$ les valeurs propres de $\widetilde{M}_{11}(x,y)$ et $\lambda_0(A_+)$ est une valeur propre de A_+. Nous avons alors :

- $\tilde{\mu_1}(0,y) = (i-y)\lambda_0(A_+)$ donc $Re\,(\tilde{\mu_1}(0,y)) = -y\lambda_0(A_+) \geq 0$,
- $\tilde{\mu_1}(x,0)$ est une valeur propre de $iA_+ + xB$ ainsi, comme le problème est faiblement bien posé, $Re\,(\tilde{\mu_1}(x,0)) \leq \alpha x$.

Nous avons le même résultat pour les valeurs propres $\tilde{\mu_2}$ de $\widetilde{M}_{22}(x,y)$. Donc, toute valeur propre $\tilde{\mu}(x,y)$ de $M(x,y)$, vérifie :

- $Re\,(\tilde{\mu}(0,y)) \geq 0$,
- $Re\,(\tilde{\mu}(x,0)) \leq \alpha x$.

Théorème 4.5 *Il existe $\varepsilon > 0$, $\beta > 0$ et $\delta > 0$ tels que pour tout (x,y) tel que $\|(x,y)\| < \varepsilon$, pour toute valeur propre $\tilde{\mu}(x,y)$ de $M(x,y)$,*

$$Re\,(\tilde{\mu}(x,y)) \leq \alpha x + \beta(xy)^\delta.$$

PREUVE : Elle repose sur le théorème de Seidenberg-Tarski sur les ensembles algébriques [14] que nous rappelons ici, car son énoncé est très simple. Un ensemble semi-algébrique de $\mathbb{R}$ est une union finie d'ensembles définis par des égalités ou inégalités polynômiales. S'il n'y a pas d'inégalités, on parle d'ensembles algébriques.

Théorème de Seidenberg-Tarski L'image d'un ensemble semi-algébrique par une application polynômiale est semi-algébrique.

Nous utiliserons ici une forme équivalente : la projection d'un ensemble semi-algébrique sur un sous-espace est semi-algébrique. Nous nous plaçons dans $\mathbb{R}^4 =$

$\{(x, y, Re\,(\tilde{\mu}), Im\,(\tilde{\mu})\}$. Le polynôme caractéristique de M y définit une surface algébrique de dimension 2. Soit

$$\mathcal{G} = \{(x, y, \tilde{\mu}_1), \exists \tilde{\mu}_2 \text{ avec } \tilde{\mu}_1 + i\tilde{\mu}_2 \text{ valeur propre de } M\}$$

$\mathcal{G}$ est semi-algébrique dans $\mathbb{R}^4$ par le théorème de Seidenberg-Tarski. Définissons maintenant

$$\mathcal{H} = \{(\varepsilon, \eta, x, y, \tilde{\mu}_1) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{G},\ \tilde{\mu}_1 \geq \alpha x + \eta, x \geq 0, y \geq 0, xy \leq \varepsilon\}.$$

$\mathcal{H}$ est un sous-ensemble semi-algébrique de $\mathbb{R}^5$. Supposons $\mathcal{H}$ non vide, et regardons sa projection $\mathcal{H}_{\varepsilon,\eta}$ dans le plan ε, η. Elle est aussi semi-algébrique. Par définition, $\mathcal{H}$ ne rencontre la droite $\varepsilon = 0$ qu'en 0. Puisqu'il est donné par un nombre fini d'équations algébriques et d'inéquations de la forme $P_i(\varepsilon, \eta) \leq 0$, on peut faire un développement de Puiseux dans chacune des inégalités, et obtenir une collection d'inégalités du type $\eta \leq \beta_i \varepsilon^{\delta_i}$. On a donc

$$\exists \beta > 0, \delta > 0, \forall (\varepsilon, \eta) \in \mathcal{H}_{\varepsilon,\eta},\ \eta \leq \beta \varepsilon^{\delta}. \tag{4.8}$$

Montrons maintenant que pour tout $(x, y, \tilde{\mu}_1)$ dans $\mathcal{G}$, $\tilde{\mu}_1 \leq 2\beta(xy)^{\delta}$. Procédons par l'absurde et supposons qu'il existe $(x, y, \tilde{\mu}_1)$ dans $\mathcal{G}$, avec $\tilde{\mu}_1 > 2\beta(xy)^{\delta}$. Choisissons dans (4.8) $\eta = \tilde{\mu}_1$ et $\varepsilon = xy$. On a alors $\tilde{\mu}_1 \leq \beta(xy)^{\delta}$, ce qui contredit l'hypothèse. □

Ici, nous avons $xy = \frac{1-\cos(\xi\Delta x)}{\sin(\xi\Delta x)\xi} \sim \frac{\Delta x}{2}$ et $x \sim \frac{1}{\xi}$, donc :

$$Re\,(\widetilde{\mu}(x, y)) \leq \alpha' \frac{1}{\xi} + \beta' \Delta x^{\delta}.$$

Donc si μ est une valeur propre de $\mathcal{P}$, alors :

$$Re\,(\mu(x, y)) \leq \alpha'' + \beta'' \xi \Delta x^{\delta} \leq C.$$

Les parties réelles des valeurs propres de $\mathcal{P}$ sont bornées.

Théorème 4.6 *Si le problème de Cauchy est faiblement bien posé, si les pas de temps et d'espace sont tels que pour toute valeur propre λ de A, $|\gamma\lambda| \leq 1$, le schéma décentré est faiblement stable.*

PREUVE : La même démonstration que dans le théorème 4.5 montre que pour toute valeur propre de M correspondant à une valeur propre positive de A, il existe $\alpha > 0$, $\beta > 0$, et $\delta > 0$ tels que, pour x et y assez petits, on ait

$$|\tilde{\mu}(x, y) - (i - y)\lambda_0(A^+)| \leq \alpha x + \beta(xy)^{\delta} \tag{4.9}$$

En effet si nous posons $\Phi(x,y) = \tilde{\mu}(x,y) - (i-y)\lambda_0(A^+)$, nous avons $\Phi(0,y) = 0$, $\Phi(x,0) = \tilde{\mu}(x,0) - i\lambda_0(A^+)$. Puisque Φ est continue et que $\Phi(0,0) = 0$, il existe un α positif tel que pour x et y suffisamment petits, on ait

$$\begin{array}{l} Re\,(\Phi(0,y)) = 0,\ -\alpha x \leq Re\,(\Phi(x,0)) \leq \alpha x, \\ Im\,(\Phi(0,y)) = 0,\ -\alpha x \leq Im\,(\Phi(x,0)) \leq \alpha x. \end{array}$$

On en déduit (4.9) comme dans la preuve du théorème (4.5). Pour démontrer la stabilité, nous utilisons maintenant la proposition 3.2. D'abord nous avons $\|\widehat{Q} - Id\| \leq K\Delta t(1+\|\xi\|)^\theta$ avec $\theta = 1$. Il suffit donc de montrer que toute valeur propre $\mu(\xi)$ de $\hat{Q}$ vérifie bien

$$|\mu(\xi)| \leq e^{\alpha\Delta t}.$$

pour un certain α. Or $\mu(\xi) = 1 + \frac{\Delta t}{x}\tilde{\mu}(x,y)$, et si $\tilde{\mu}$ correspond à une valeur propre positive de A, nous pouvons écrire par (4.9)

$$\tilde{\mu}(x,y) = (i-y)\lambda_0(A^+) + R(x,y),$$

où le reste $R(x,y)$ est majoré en module par $\alpha x + \beta(xy)^\delta$. Nous pouvons écrire alors

$$\begin{array}{rcl} \mu(\xi) & = & 1 + \dfrac{\Delta t}{x}((i-y)\lambda_0(A^+) + R(x,y)), \\ & = & 1 - \gamma(1-\cos\xi\Delta x)\lambda_0(A^+) + i\gamma\sin\xi\Delta x + \Delta t\dfrac{R(x,y)}{x} \end{array}$$

Nous reconnaissons à droite de l'égalité le coefficient d'amplification du schéma décentré pour l'équation scalaire, et donc sous la condition $\gamma\lambda_0(A^+) \leq 1$, on a

$$|\mu(\xi)| \leq 1 + \Delta t(\alpha + \beta\frac{(xy)^\delta}{x})$$

et $\frac{(xy)^\delta}{x} \sim \xi(\frac{\Delta x}{2})^\delta$ qui est borné si on a choisi $\xi\Delta x^\delta \leq \pi$. Nous avons donc prouvé que toute valeur propre $\mu(\xi)$ de $\hat{Q}$ correspondant à une valeur propre positive de A vérifie bien

$$|\mu(\xi)| \leq e^{\alpha\Delta t}.$$

Le résultat est le même pour les autres valeurs propres.

□

Corollaire 4.1 *Si le problème de Cauchy est faiblement bien posé, si les pas de temps et d'espace sont tels que pour toute valeur propre λ de A, $|\gamma\lambda| \leq 1$, le schéma décentré est convergent pour une donnée initiale de régularité $q_4 \geq \max(q_1,q_2)$ et le taux de convergence du schéma est minoré par :*

$$\beta_2 = \frac{q_4 - \max(q_1,q_2)}{\max(q_4, 2+q_1+q_2) - \max(q_1,q_2)}.$$

PREUVE : Nous allons appliquer la proposition 3.4. Comme nous avons prouvé que le schéma est stable, il suffit de montrer que $r = 1$ et $\rho = 2$. Or, nous avons :

$$\begin{aligned}\widehat{Q}(\xi) &= Id + \Delta t(\xi A + 0(\xi^2\Delta x)) + \Delta t B \\ &= Id + \Delta t P(i\xi) + O(\xi^2\Delta x^2) \\ &= \exp(\Delta t P(i\xi)) + O(\xi^2\Delta x^2).\end{aligned}$$

Nous avons donc bien $r = 1$ et $\rho = 2$ ce qui achève la preuve. □

Si dans la définition 3.1, nous remplaçons $P(z)$ par $\mathcal{P}(\xi, \xi\Delta x)$, alors le schéma décentré vérifie les hypothèses de la proposition 3.6 avec $N_0 = 1$. Nous obtenons donc l'estimation analogue à celle obtenue dans le cas des schémas développables en P :

$$\|\widehat{Q}^n(\xi) - \exp(t_n\mathcal{P}(\xi, \xi\Delta x))\| \leq Ke^{\alpha_S t_n \Delta t}(1 + |\xi|)^{2+q_2}.$$

Par contre, la seconde estimation 3.10 ne s'obtient plus comme dans le cas des schémas développables en P . Nous avons même une forte présomption qu'il existe des cas où elle n'est pas valable. Considérons en effet les matrices

$$A^+ = \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix},\ A^- = -A^+,\ B_{11} = B_{22} = 0,\ B_{21} = \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},\ B_{12} = \begin{pmatrix}1 & 0\\ 2 & 0\end{pmatrix}.$$

Alors on peut calculer les valeurs propres $\tilde{\mu}(x, y)$ comme plus haut, et une au moins d'entre elles est de la forme

$$\tilde{\mu}(x, y) = i + y + \frac{1-i}{2}x\sqrt{y} + \cdots,$$

et n'est pas dérivable en 0 dans la variable y. C'est pourquoi la différence

$$\|\exp(\frac{t_n}{x}M(x, y)) - \exp(\frac{t_n}{x}M(x, 0))\|$$

ne pourra pas être bornée de manière analogue au cas des schémas développables en P. Le schéma décentré a donc un comportement inattendu. Le traitement de ce comportement est en cours.

Chapitre 5

Etude de schémas multipas pour des équations à coefficients constants

Dans cette partie, nous allons considérer des schémas à $q+1$ pas en temps avec $q \geq 1$ à coefficients constants :

$$Q_{-1}V^{n+1} = \sum_{\sigma=0}^{q} Q_\sigma V^{n-\sigma}.$$

Nous nous ramenons à un schéma à un pas en notant : :

$$\mathbf{V}^n = {}^t(V^{n+q}, \ldots, V^n) \text{ et } \widehat{\mathbf{V}}^n = {}^t(\widehat{V}^{n+q}, \ldots, \widehat{V}^n).$$

Nous avons alors :

$$\widehat{\mathbf{V}}^{n+1} = \widehat{Q}\widehat{\mathbf{V}}^n,$$

avec :

$$\widehat{Q} = \begin{pmatrix} (\widehat{Q_{-1}})^{-1}\widehat{Q_0} & (\widehat{Q_{-1}})^{-1}\widehat{Q_1} & \ldots & \ldots & (\widehat{Q_{-1}})^{-1}\widehat{Q_q} \\ Id & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & Id & 0 \end{pmatrix}.$$

Nous allons adapter les résultats du chapitre 3 aux schémas multipas et les appliquer au schéma saute-mouton.

5.1 Un résultat de stabilité

Nous montrons ici un résultat de stabilité qui n'utilise plus, contrairement à la proposition 3.2, l'hypothèse $\|\widehat{Q} - Id\| \leq K\Delta t(1+\|\xi\|)^\theta$ qui n'est pas vérifiée pour le schéma de saute-mouton par exemple. Le résultat suivant s'appliquera aux schémas multipas de type "point milieu" pour la discrétisation en temps.

Proposition 5.1 *Nous considérons un schéma de la forme :*

$$\frac{V^{n+1} - V^{n-1}}{\Delta t} = RV^n,$$

où R est un opérateur en espace $R = R(D^+, D^-)$. On suppose qu'il existe $h_0, k_0 > 0$ tels que :

- $\exists K > 0,\ \forall h_j \leq h_0,\ \forall \xi \in \mathcal{D}_d,\ \|\widehat{R}\| \leq K(1 + \|\xi\|)$,
- *il existe α_S tel que pour tout $\Delta t \leq k_0$ et pour toute valeur propre μ de $\widehat{Q}$, $|\mu| \leq e^{\alpha_S \Delta t}$,*
- *le schéma de démarrage vérifie : $\forall h_j \leq h_0,\ \forall \xi \in \mathcal{D}_d,\ \ \|\widehat{V^1}\| \leq K(1+\|\xi\|)\|\widehat{V^0}\|$.*

Alors le schéma est faiblement stable.

Preuve : Ici $\hat{Q}(\xi) = \begin{pmatrix} \Delta t\hat{R}(\xi) & Id \\ Id & 0 \end{pmatrix}$. Nous effectuons la décomposition de Schur de $\widehat{R}$:

$$\widehat{R} = S^*(\Lambda + T)S,$$

où $\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{pmatrix}$ et $T = \begin{pmatrix} 0 & t_{1,2} & \dots & t_{1,N} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & t_{N-1,N} \\ 0 & \dots & \dots & 0 \end{pmatrix}$, avec S unitaire. Posons $\widehat{W^n} = S\widehat{V^n}$, alors $\widehat{W^{n+1}} = \widehat{W^{n-1}} + \Delta t(\Lambda + T)\widehat{W^n}$. Nous avons alors :

$$\forall 1 \leq i \leq N-1,\ \widehat{W_i^{n+1}} = \widehat{W_i^{n-1}} + \Delta t\lambda_i\widehat{W_i^n} + \Delta t \sum_{j=i+1}^{N} t_{i,j}\widehat{W_j^n},$$

$$\widehat{W_N^{n+1}} = \widehat{W_N^{n-1}} + \Delta t\lambda_N\widehat{W_N^n}.$$

Nous montrons le lemme suivant :

Lemme 5.1 *Si pour tout $n \geq 1$, $U^{n+1} = MU^n + NU^{n-1} + F_n$, alors, pour tout $n \geq 1$, nous avons :*

$$\begin{pmatrix} U^n \\ U^{n-1} \end{pmatrix} = \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix}^{n-1} \begin{pmatrix} U^1 \\ U^0 \end{pmatrix} + \sum_{k=0}^{n-2} \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix}^{n-k-2} \begin{pmatrix} F_{k+1} \\ 0 \end{pmatrix}.$$

PREUVE : Nous procédons par récurrence sur n :

- Pour $n = 1$, le résultat est clairement vrai.
- Supposons le résultat vrai au rang $n \in \mathbb{N}^*$. Nous avons :

$$\begin{aligned}
\begin{pmatrix} U^{n+1} \\ U^n \end{pmatrix} &= \begin{pmatrix} MU^n + NU^{n-1} + F_n \\ U^n \end{pmatrix} \\
&= \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix} \begin{pmatrix} U^n \\ U^{n-1} \end{pmatrix} + \begin{pmatrix} F_n \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix}^n \begin{pmatrix} U^1 \\ U^0 \end{pmatrix} + \sum_{k=0}^{n-2} \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix}^{n-k-1} \begin{pmatrix} F_{k+1} \\ 0 \end{pmatrix} \\
&\quad + \begin{pmatrix} F_n \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix}^n \begin{pmatrix} U^1 \\ U^0 \end{pmatrix} + \sum_{k=0}^{n-1} \begin{pmatrix} M & N \\ Id & 0 \end{pmatrix}^{n-k-1} \begin{pmatrix} F_{k+1} \\ 0 \end{pmatrix}.
\end{aligned}$$

Nous avons ainsi prouvé le lemme. □

En utilisant le lemme précédent, nous avons :

$$\begin{aligned}
\begin{pmatrix} \widehat{W_i^n} \\ \widehat{W_i^{n-1}} \end{pmatrix} &= \begin{pmatrix} \Delta t\lambda_i & 1 \\ 1 & 0 \end{pmatrix}^{n-1} \begin{pmatrix} \widehat{W_i^1} \\ \widehat{W_i^0} \end{pmatrix} \\
&\quad + \sum_{k=0}^{n-2} \begin{pmatrix} \Delta\lambda_i & 1 \\ 1 & 0 \end{pmatrix}^{n-k-2} \begin{pmatrix} \Delta t \sum_{j=i+1}^{N} t_{i,j} \widehat{W_j^{k+1}} \\ 0 \end{pmatrix}.
\end{aligned}$$

Or, en diagonalisant, nous obtenons :

$$\begin{pmatrix} \alpha & 1 \\ 1 & 0 \end{pmatrix}^n = \frac{1}{\mu_1 - \mu_2} \begin{pmatrix} \mu_1^{n+1} - \mu_2^{n+1} & \mu_1\mu_2(\mu_1^n - \mu_2^n) \\ \mu_1^n - \mu_2^n & \mu_1\mu_2(\mu_1^{n-1} - \mu_2^{n-1}) \end{pmatrix}.$$

Avec : $\mu_{1,2} = \frac{\alpha \pm \sqrt{\alpha^2+4}}{2}$.

Nous avons donc :

$$\begin{aligned}
\widehat{W_i^n} &= \frac{\mu_1^n - \mu_2^n}{\mu_1 - \mu_2} W_i^1 + \mu_1\mu_2(\mu_1^{n-1} - \mu_2^{n-1})\widehat{W_i^0} \\
&\quad + \sum_{k=0}^{n-2} \frac{\mu_1^{n-k-1} - \mu_2^{n-k-1}}{\mu_1 - \mu_2} \left(\Delta t \sum_{j=i+1}^{N} t_{i,j} \widehat{W_j^{k+1}} \right) \\
&= \frac{\mu_1^n - \mu_2^n}{\mu_1 - \mu_2} (\widehat{W_i^1} + \mu_2 \widehat{W_i^0}) - \mu_2^n \widehat{W_i^0} + \sum_{k=0}^{n-2} \frac{\mu_1^{n-k-1} - \mu_2^{n-k-1}}{\mu_1 - \mu_2} \left(\Delta t \sum_{j=i+1}^{N} t_{i,j} \widehat{W_j^{k+1}} \right),
\end{aligned}$$

où $\mu_{1,2} = \frac{\alpha \pm \sqrt{\alpha^2+4}}{2}$ avec $\alpha = \Delta t \lambda_i$.

Or $\widehat{Q} = \begin{pmatrix} \Delta t \widehat{R} & Id \\ Id & 0 \end{pmatrix}$, donc les $\mu_{1,2}$ sont les valeurs propres de $\widehat{Q}$, donc par hypothèse :

$$|\mu_{1,2}| \leq e^{\alpha_S \Delta t}.$$

Ainsi, si nous considérons $\|\xi\|\Delta t < \varepsilon$, alors :

$$\Delta t|\lambda_i| \leq \Delta t \|\widehat{R}\| \leq K\Delta t(1 + \|\xi\|) \leq K(h_0 + \varepsilon).$$

Donc, pour h_0 et ε assez petit, il existe $C > 0$ tel que :

$$|\mu_1 - \mu_2| = |\sqrt{\Delta t^2 \lambda_i^2 + 4}| \geq \sqrt{C}.$$

Donc :

$$\left| \frac{\mu_1^n - \mu_2^n}{\mu_1 - \mu_2} \right| \leq \frac{2e^{\alpha_S t_n}}{\sqrt{C}}.$$

Si nous considérons maintenant $\varepsilon > 0$, pour $\|\xi\|\Delta t \geq \varepsilon$:

$$\left| \frac{\mu_1^n - \mu_2^n}{\mu_1 - \mu_2} \right| = \left| \sum_{k=0}^{n-1} \mu_1^k \mu_2^{n-1-k} \right| \leq n e^{\alpha_S t_n} \leq \frac{t_n}{\Delta t} e^{\alpha_S t_n} \leq t_n \frac{\|\xi\|}{\varepsilon} e^{\alpha_S t_n}.$$

Nous avons donc prouvé l'existence d'une constante C' telle que :

$$\left| \frac{\mu_1^n - \mu_2^n}{\mu_1 - \mu_2} \right| \leq C'(1 + \|\xi\|) e^{\alpha_S t_n}.$$

Nous pouvons alors montrer le lemme suivant :

Lemme 5.2 $\forall i \in \{1, \ldots, N\}$,

$$|\widehat{W_i^n}| \leq C_i (1 + \|\xi\|)^{N-i+1} t_n^{N-i} e^{\alpha_S t_n} \max_{r \geq i} (|\widehat{W_r^0}| + |\widehat{W_r^1}|).$$

PREUVE : Nous procédons par récurrence descendante sur i :

- Pour $i = N$,

$$\widehat{W_i^n} = \frac{\mu_1^n - \mu_2^n}{\mu_1 - \mu_2} \widehat{W_i^1} + \mu_1 \mu_2 (\mu_1^{n-1} - \mu_2^{n-1}) \widehat{W_i^0}.$$

Donc :

$$\begin{aligned} |\widehat{W_i^n}| &= C'(1 + |\xi|) e^{\alpha_S t_n} |\widehat{W_i^1}| + 2e^{\alpha_S t_{n+1}} |\widehat{W_i^0}| \\ &\leq C_N (1 + |\xi|) e^{\alpha_S t_n} \max_{r \geq i} (|\widehat{W_r^0}| + |\widehat{W_r^1}|). \end{aligned}$$

- Soit $1 \leq i \leq N-1$, supposons le résultat vrai pour tout $j \geq i+1$, nous avons alors :

$$\begin{aligned}
|\widehat{W_i^n}| &= C'(1+|\xi|)e^{\alpha_S t_n}(|\widehat{W_i^1}| + e^{\alpha_S \Delta t}|\widehat{W_i^0}|) + e^{\alpha_S \Delta t}|\widehat{W_i^0}| + \left(\sum_{k=0}^{n-2} C'(1+\|\xi\|)e^{\alpha_S t_{n-k-1}}\right. \\
&\quad \left.\times \left(\Delta t \sum_{j=i+1}^{N} t_{i,j} C_j (1+\|\xi\|)^{N-j+1} t_n^{N-j} e^{\alpha_S t_n} \max_{r \geq j}(|\widehat{W_r^0}| + |\widehat{W_r^1}|)\right)\right) \\
&\leq C_{i+1}(1+\|\xi\|)^{N-i+1} t_n^{N-i} e^{\alpha_S t_n} \max_{r \geq i}(|\widehat{W_r^0}| + |\widehat{W_r^1}|).
\end{aligned}$$

□

Donc :

$$\|\widehat{W^n}\| \leq C(1+\|\xi\|)^{N+1} t_n^N e^{\alpha_S t_n}(\|\widehat{W^1}\| + \|\widehat{W^0}\|).$$

Or $\forall k$, $\|\widehat{W^k}\| = \|\widehat{V^k}\|$ et comme $\|\widehat{V^1}\| \leq K(1+\|\xi\|)\|\widehat{V^0}\|$, nous avons :

$$\|\widehat{V^n}\| \leq C(1+\|\xi\|)^{N+2} t_n^N e^{\alpha_S t_n} \|\widehat{V^0}\|.$$

Ce qui achève la preuve.

□

5.2 Taux de convergence

Le théorème 3.1 ainsi que toutes les propositions qui en découlent restent valables mais, contrairement aux schémas à un pas, le problème n'est plus d'évaluer :

$$\left\| e^{t_n P(i\xi)} - \widehat{Q}^n(\xi) \right\|,$$

mais :

$$\left\| \begin{pmatrix} e^{t_{n+q} P(i\xi)} \widehat{U^0} \\ \vdots \\ e^{t_n P(i\xi)} \widehat{U^0} \end{pmatrix} - \widehat{Q}^n(\xi) \begin{pmatrix} \widehat{V^q} \\ \vdots \\ \widehat{V^0} \end{pmatrix} \right\|.$$

L'analogue du théorème 3.1 est alors :

Théorème 5.1 *On considère que le problème de Cauchy est faiblement bien posé de défaut q_1 et que le schéma est faiblement stable de défaut q_2.*
On suppose de plus que :

$$\exists 0 \leq \delta \leq 1,\ \forall \|\xi\| \leq \frac{\pi}{h^\delta},$$

$$\left\| \begin{pmatrix} e^{t_{n+q}P(i\xi)}\widehat{U^0} \\ \vdots \\ e^{t_nP(i\xi)}\widehat{U^0} \end{pmatrix} - \widehat{Q}^n(\xi) \begin{pmatrix} \widehat{V^q} \\ \vdots \\ \widehat{V^0} \end{pmatrix} \right\| \leq Ct_nh^s(1+\|\xi\|^2)^{\sigma/2}\|\widehat{U^0}\|.$$

Soit q_4 tel que :

$$\max(q_1,q_2) \leq q_4 \qquad et \qquad s \leq \delta(\max(q_4,\sigma) - \max(q_1,q_2)). \tag{5.1}$$

Alors : $\exists K, \alpha', \forall f \in H^{q_4}(\mathbb{R})$,

$$\|U(t_n,.) - SV^n\|_{L^2} \leq Ke^{\alpha' t_n}t_nh^{\beta_1}\|U^0\|_{H^{q_4}},$$

où :

$$\beta_1 = \frac{s(q_4 - \max(q_1,q_2))}{\max(q_4,\sigma) - \max(q_1,q_2)}.$$

Nous cherchons maintenant, comme dans le cas des schémas à un pas, à calculer une valeur de σ qui donne un taux de convergence optimal.

Nous considérons à partir de maintenant que $d = 1$. Comme dans le chapitre 3, nous notons $\Delta x = h$.

Définition 5.1 *Un schéma multipas est dit développable en P s'il existe $\widetilde{Q}$ tel que :*

$$\sum_{\sigma=0}^{q} (\widehat{Q_{-1}})^{-1}\widehat{Q_\sigma}\widetilde{Q}^{q-\sigma} = \widetilde{Q}^{q+1}$$

et $\widetilde{Q}$ est développable en P au sens de la définition 3.1 .

Nous allons alors prouver le théorème, analogue au théorème 3.2, suivant :

Théorème 5.2 *On suppose que :*

- *le problème de Cauchy est faiblement bien posé de défaut q_1,*
- *$\gamma = \frac{\Delta t}{\Delta x}$ est constant et le schéma est faiblement stable de défaut q_2,*
- *le schéma multipas est développable en P et on note R_{N_0} le reste du développement de $\widetilde{Q}$,*
- *$\forall k \geq 0, \exists B_k > 0, \forall \Delta x, \Delta x \leq h_0, \forall \Delta t, \Delta t \leq k_0, \forall \xi, |\xi| \leq \frac{\pi}{\Delta x^\delta}, \left\| \frac{\partial^k R_{N_0}(\Delta t, \Delta x, \xi)}{\partial \Delta t^k} \right\| \leq B_k(1+|\xi|^{N_0+k+1})$,*
- *$\exists C > 0, \forall \Delta x, \Delta x \leq h_0, \forall \Delta t, \Delta t \leq k_0, \forall \xi, |\xi| \leq \frac{\pi}{\Delta x^\delta}, |z| \leq C|\xi|$,*
- *$\forall |\xi| \leq \frac{\pi}{\Delta x^\delta}, |z - i\xi| \leq K\Delta x^r|\xi|^\rho$, où $z = z(\Delta t, \Delta x, \xi)$ est introduit dans la définition 3.1.*
- *les schémas d'initialisation sont tels que : $\forall \sigma \in \{0,\ldots,q\}$,*

$$\|\widetilde{Q}^\sigma\widehat{U^0} - \widehat{V^\sigma}\| \leq C\Delta t^{N_0}(1+|\xi|)^{N_0+1}.$$

Alors :

$$\left\| \begin{pmatrix} e^{t_{n+q}P(i\xi)}\widehat{U^0} \\ \vdots \\ e^{t_nP(i\xi)}\widehat{U^0} \end{pmatrix} - \widehat{Q}^n(\xi) \begin{pmatrix} \widehat{V^q} \\ \vdots \\ \widehat{V^0} \end{pmatrix} \right\| \leq Ke^{t_n\alpha}(\Delta t^{N_0}+\Delta x^r)(1+|\xi|)^{\max(N_0+1+q_2,\rho+q_1)}\|\widehat{U^0}\|.$$

PREUVE :

- Nous avons :

$$\widehat{Q} \begin{pmatrix} \widetilde{Q}^{n+q} \\ \vdots \\ \widetilde{Q}^n \end{pmatrix} = \begin{pmatrix} \widetilde{Q}^{n+q+1} \\ \vdots \\ \widetilde{Q}^{n+1} \end{pmatrix}.$$

  Donc :

$$\widehat{Q}^n \begin{pmatrix} \widetilde{Q}^q \\ \vdots \\ Id \end{pmatrix} = \begin{pmatrix} \widetilde{Q}^{n+q} \\ \vdots \\ \widetilde{Q}^n \end{pmatrix}.$$

  Nous avons alors, en utilisant les hypothèses portant sur $\widetilde{Q}$:

$$\begin{aligned} \left\| \begin{pmatrix} e^{t_{n+q}P(i\xi)}\widehat{U^0} \\ \vdots \\ e^{t_nP(i\xi)}\widehat{U^0} \end{pmatrix} - \widehat{Q}^n(\xi) \begin{pmatrix} \widetilde{Q}^q \\ \vdots \\ Id \end{pmatrix} \widehat{U^0} \right\| &= \left\| \begin{pmatrix} (e^{t_{n+q}P(i\xi)} - \widetilde{Q}^{n+q})\widehat{U^0} \\ \vdots \\ (e^{t_nP(i\xi)} - \widetilde{Q}^n)\widehat{U^0} \end{pmatrix} \right\| \\ &\leq Ke^{\alpha t_{n+q}}(\Delta t^{N_0} + \Delta x^r)(1+|\xi|)^{\max(N_0+1+q_2,\rho+q_1)}\|\widehat{U^0}\|. \end{aligned}$$

- Il reste donc à évaluer :

$$\left\| \widehat{Q}^n(\xi) \begin{pmatrix} \widehat{V^q} \\ \vdots \\ \widehat{V^0} \end{pmatrix} - \widehat{Q}^n(\xi) \begin{pmatrix} \widetilde{Q}^q \\ \vdots \\ Id \end{pmatrix} \widehat{U^0} \right\|.$$

  Or, comme le schéma est faiblement stable de défaut q_2, nous avons :

$$\begin{aligned} &\left\| \widehat{Q}^n(\xi) \begin{pmatrix} \widehat{V^q} \\ \vdots \\ \widehat{V^0} \end{pmatrix} - \widehat{Q}^n(\xi) \begin{pmatrix} \widetilde{Q}^q \\ \vdots \\ Id \end{pmatrix} \widehat{U^0} \right\| \\ &\qquad \leq K_S e^{\alpha_S t_n}(1+|\xi|)^{q_2} \left\| \begin{pmatrix} \widehat{V^q} \\ \vdots \\ \widehat{V^0} \end{pmatrix} - \begin{pmatrix} \widetilde{Q}^q \\ \vdots \\ Id \end{pmatrix} \widehat{U^0} \right\| \\ &\qquad \leq K_S e^{\alpha_S t_n}(1+|\xi|)^{q_2} \max_{0\leq\sigma\leq q} \|\widehat{V^\sigma} - \widetilde{Q}^\sigma\widehat{U^0}\| \\ &\qquad \leq K_S e^{\alpha_S t_n}(1+|\xi|)^{q_2}\Delta t^{N_0}(1+|\xi|)^{N_0+1}, \end{aligned}$$

en utilisant l'hypothèse sur le schéma d'initialisation.
Nous avons donc la conclusion.

□

5.3 Etude d'un exemple : le schéma de saute-mouton

5.3.1 Ecriture d'un schéma de saute-mouton

Nous utilisons la discrétisation classique du terme d'ordre 0. Nous obtenons le schéma suivant :

$$\frac{V_j^{n+1} - V_j^{n-1}}{2\Delta t} = A\frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x} + BV_j^n. \tag{5.2}$$

Nous avons alors :

$$\begin{pmatrix} \widehat{V^{n+1}} \\ \widehat{V^n} \end{pmatrix} = \begin{pmatrix} 2\Delta t\left(i\frac{\sin(\xi\Delta x)}{\Delta x}A + B\right) & Id \\ Id & 0 \end{pmatrix} \begin{pmatrix} \widehat{V^n} \\ \widehat{V^{n-1}} \end{pmatrix}.$$

Donc, la matrice d'amplification du schéma est :

$$\widehat{Q}(\xi) = \begin{pmatrix} 2\Delta t P(z) & Id \\ Id & 0 \end{pmatrix},$$

avec :

$$z = i\frac{\sin(\xi\Delta x)}{\Delta x}.$$

5.3.2 Etude de la stabilité

Pour étudier la stabilité du schéma de saute-mouton, nous allons appliquer la proposition 5.1.

Proposition 5.2 *Sous la condition CFL :*

$$\forall\lambda \in \sigma(A),\ \gamma|\lambda| \leq 1$$

et pour tout schéma d'initialisation tel que :

$$\|\widehat{V^1}\| \leq K(1 + |\xi|)\|\widehat{V^0}\|,$$

le schéma de saute-mouton est faiblement stable.

PREUVE : Avec les notations de la proposition 5.1, on a $\widehat{R} = 2P(z)$. Montrons que les quatre hypothèses sont vérifiées.

- Nous avons :

$$\|\widehat{R}\| \leq 2(|z|\|A\| + \|B\|) \leq 2(|\xi|\|A\| + \|B\|) \leq K(1 + |\xi|).$$

- Soit μ une valeur propre de $\widehat{Q}$. Alors, il existe $\lambda(z)$ valeur propre de $P(z)$ telle que :

$$\mu^2 - 2\lambda(z)\Delta t\mu - 1 = 0.$$

Donc :

$$\mu = \lambda(z)\Delta t \pm \sqrt{\lambda(z)^2\Delta t^2 + 1}.$$

Or, nous savons qu'il existe C tel que si $|z| \geq C$, $\lambda(z) = \lambda_0 z + f(z)$ avec f bornée. Nous avons donc :

$$\begin{aligned}\mu &= (\lambda_0 z + f(z))\Delta t \pm \sqrt{(\lambda_0 z + f(z))^2\Delta t^2 + 1} \\ &= i\lambda_0\gamma\sin(\xi\Delta x) + \Delta t f(z) \\ &\quad \pm\sqrt{-\lambda_0^2\gamma^2\sin(\xi\Delta x)^2 + 2i\lambda_0\gamma\sin(\xi\Delta x)\Delta t f(z) + f(z)^2\Delta t^2 + 1}.\end{aligned}$$

Or, nous avons :

$$\begin{aligned}&\lambda_0^2\gamma^2\sin(\xi\Delta x)^2 + 2\lambda_0\gamma\sin(\xi\Delta x)\Delta t f(z) + f(z)^2\Delta t^2 + 1 \\ &\qquad = 1 - \lambda_0^2\gamma^2\sin(\xi\Delta x)^2 + O(\Delta t) \text{ uniformément en } \xi.\end{aligned}$$

Donc :

$$\mu = i\lambda_0\gamma\sin(\xi\Delta x) \pm \sqrt{1 - \lambda_0^2\gamma^2\sin(\xi\Delta x)^2} + O(\Delta t).$$

Or, $|i\lambda_0\gamma\sin(\xi\Delta x) \pm \sqrt{1 - \lambda_0^2\gamma^2\sin(\xi\Delta x)^2}| = 1$, donc :

$$|\mu| \leq 1 + M\Delta t \leq e^{M\Delta t},$$

ce qui prouve bien le deuxième point.

- Enfin le dernier point est vérifié par hypothèse.

La proposition 5.1 montre donc que le schéma de saute-mouton est faiblement stable.

□

Remarque 5.1 *Dans le cas de problèmes fortement bien posés, la condition CFL doit être stricte pour obtenir la stabilité du schéma. Toutefois le comportement n'est pas le même pour $\gamma|\lambda| > 1$ et pour $\gamma|\lambda| = 1$. En effet, l'instabilité provenant de $\gamma|\lambda| > 1$ est une croissance exponentielle de $\|\widehat{Q}^n\|$ en fonction de n, qui est toujours une instabilité au sens de la stabilité faible. Cependant, dans le cas $\gamma|\lambda| = 1$, la croissance observée est linéaire en n, ce qui peut se ramener à une croissance linéaire en ξ, qui n'est pas exclue dans la définition de la stabilité faible. Le graphique suivant*

donne l'erreur en norme L^2 en fonction du pas d'espace. Le schéma d'initialisation est un schéma de Lax-Wendroff, l'équation traitée est celle de l'exemple 1 et la donnée initiale est $H^{11/2}$. Nous avons pris $\gamma = 1$ et nous observons que le schéma est bien convergent.

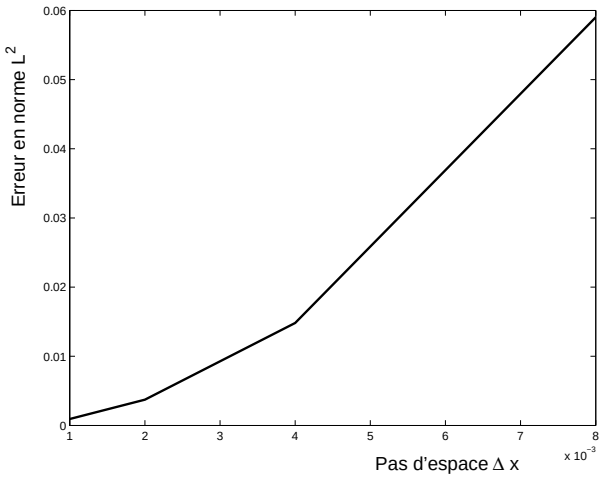


FIG. 5.1 – Schéma de saute-mouton pour $\gamma|\lambda| = 1$

5.3.3 Taux de convergence

Nous allons calculer un taux de convergence optimal en utilisant le théorème 5.2.

Théorème 5.3 *Si les hypothèses suivantes sont vérifiées :*

- *le problème de Cauchy est faiblement bien posé de défaut q_1,*
- *le schéma de saute-mouton (5.2) est faiblement stable de défaut q_2 sous la condition CFL stricte $\forall \lambda \in \sigma(A)$, $\gamma|\lambda| < 1$,*
- *le schéma d'initialisation est tel que :*
  $\| \exp(\Delta t P(i\xi))\widehat{U^0} - \widehat{V^1}\| \leq C\Delta t^2(1+|\xi|)^3$.

Alors, pour une donnée initiale dans H^{q_4}, le taux de convergence pour le schéma de saute-mouton (5.2) est :

$$\beta = \frac{2(q_4 - \max(q_1, q_2))}{\max(q_4, 3 + \max(q_1, q_2)) - \max(q_1, q_2)}.$$

Remarque 5.2 *L'hypothèse sur le schéma d'initialisation nécessite uniquement que le schéma soit d'ordre 1. En effet, si c'est le cas et si nous notons $\widehat{Q}_{init}$ la transformée de Fourier du schéma d'initialisation, nous avons : $\| \frac{\exp(\Delta t P(i\xi)) - \widehat{Q}_{init}}{\Delta t} \| \leq K\Delta t(1+|\xi|)^2$.*

PREUVE :

- Nous cherchons $\widetilde{Q}$ vérifiant l'équation :

$$\widetilde{Q}^2 - 2\Delta t P(z)\widetilde{Q} - Id = 0.$$

Nous cherchons un développement en P de $\widetilde{Q}$ avec $a_{k,j} = \delta_{k,j}a_k$, c'est à dire :

$$\widetilde{Q} = Id + \sum_{k=1}^{+\infty} a_k P(z)^k \Delta t^k.$$

Nous posons $a_0 = 1$, nous avons alors :

$$\widetilde{Q}^2 = \sum_{k=0}^{+\infty} \left(\sum_{i+j=k} a_i a_j \right) P(z)^k \Delta t^k.$$

En ré-injectant dans l'équation vérifiée par $\widetilde{Q}$, nous obtenons :

$$\sum_{k=0}^{+\infty} (\sum_{i+j=k} a_i a_j) P(z)^k \Delta t^k - 2\sum_{k=1}^{+\infty} a_{k-1} P(z)^k \Delta t^k - Id = 0.$$

Ce qui donne la relation de récurrence suivante :

$$\begin{cases} a_0^2 - 1 = 0 \\ \sum_{i+j=k} a_i a_j - 2a_{k-1} = 0 \text{ pour } k \geq 1. \end{cases}$$

Donc pour $k \geq 1$, $2a_k = 2a_{k-1} - \sum_{i+j=k,\, i,j\neq k} a_i a_j$ ainsi les coefficients sont bien définis.
De plus, la relation de récurrence est indépendante de la dimension, donc, si nous posons $x = \Delta t P(z)$, les a_k sont les coefficients de la décomposition en série entière de $x + \sqrt{x^2+1}$ qui a pour rayon de convergence 1. Donc le développement de $\widetilde{Q}$ converge normalement pour $\|\Delta t P(z)\| < 1$ ce qui est le cas lorsque $\delta < 1$.

- Comme les a_k sont les coefficients de la décomposition en série entière de $x + \sqrt{x^2+1}$, nous avons $a_1 = 1$ et $a_2 = \frac{1}{2}$, donc $\widetilde{Q} = Id + \Delta t P(z) + \frac{\Delta t^2}{2} P(z)^2 + \Delta t^3 R_2$, avec $R_2 = \sum_{k=0}^{+\infty} a_{k+3} P(z)^{k+3} \Delta t^k$. Cela prouve que $N_0 = 2$.
De plus, $\frac{\partial^j R_2}{\partial \Delta t^j} = \sum_{k=j}^{+\infty} a_{k+3} P(z)^{k+3} \frac{k!}{(k-j)!} \Delta t^{k-j}$. En faisant le même raisonnement que pour les schémas étudiés précédemment, nous trouvons : $\left\| \frac{\partial^j R_2}{\partial \Delta t^j} \right\| B_j (1 + |\xi|^{j+3})$.
- Nous avons : $z = i\frac{\sin(\xi \Delta x)}{\Delta x} = i\xi + O(\xi^3 \Delta x^2)$, donc $r = 2$ et $\rho = 3$.

- Pour étudier le schéma d'initialisation, nous commençons par remarquer que :

$$\|\widetilde{Q} - \exp(\Delta t P(z))\|\| \leq C\Delta t^3(1+|\xi|)^3.$$

Donc :

$$\begin{aligned}\|\widetilde{Q}\widehat{U^0} - \widehat{V^1}\| &\leq \|\widetilde{Q}\widehat{U^0} - \exp(\Delta t P(z))\widehat{U^0}\| \\ &\quad + \|\exp(\Delta t P(z))\widehat{U^0} - \exp(\Delta t P(i\xi))\widehat{U^0}\| \\ &\quad + \|\exp(\Delta t P(i\xi))\widehat{U^0} - \widehat{V^1}\| \\ &\leq C\Delta t^2(1+|\xi|)^3\|\widehat{U^0}\|.\end{aligned}$$

Ce qui achève la preuve.

□

5.3.4 Résultats numériques

Nous effectuons les mêmes calculs que précédemment. Le schéma d'initialisation est un schéma de Lax-Wendroff. Nous obtenons alors les résultats suivants :

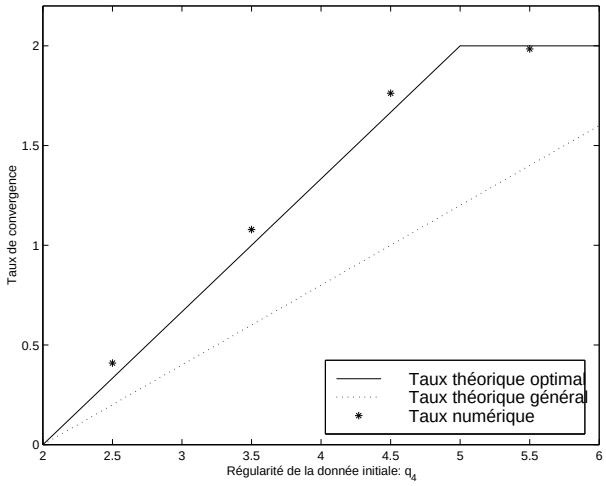


FIG. 5.2 – Schéma de saute-mouton, exemple 1 : $q_2 = 2$

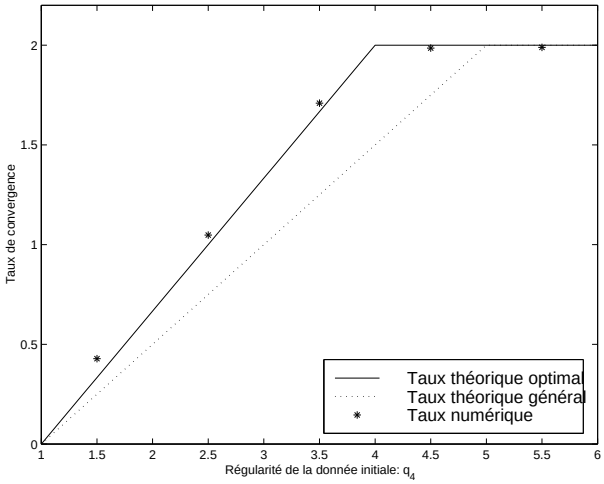


FIG. 5.3 – Schéma de saute-mouton, exemple 2 : $q_2 = 1$

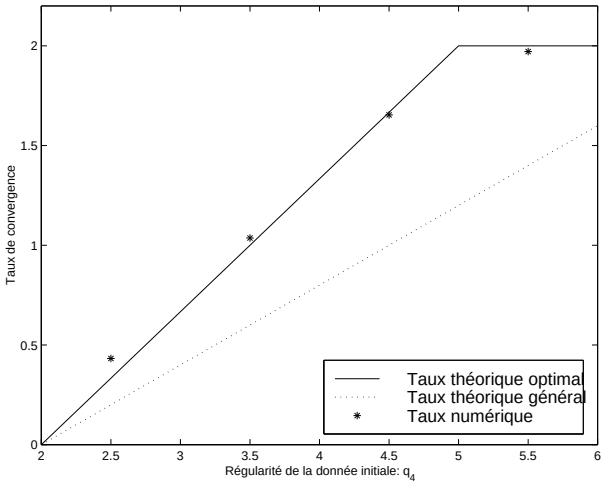


FIG. 5.4 – Schéma de saute-mouton, exemple 3 : $q_2 = 2$

Remarque 5.3 *Le taux de convergence est analogue lorsque le schéma d'initialisation est un schéma de Crank-Nicolson ou un schéma décentré.*

Conclusion

Dans cette partie, nous avons donné des définitions de stabilité, consistance et convergence qui sont adaptées à la perte de régularité observée lorsque nous étudions des problèmes faiblement bien posés. Avec ces définitions nous obtenons une condition nécessaire et suffisante de convergence qui est une extension du théorème de Lax-Richtmyer. Ce théorème est valable sous des hypothèses très larges : problème à coefficients variables, espace de dimension quelconque et schéma à un pas ou multipas.

Nous nous sommes ensuite intéressés au cas particulier des problèmes à coefficients constants. Nous avons obtenu une minoration du taux de convergence pour des schémas quelconques. Grâce à cette minoration, nous avons retrouvé un résultat essentiel de la théorie des schémas pour les problèmes fortement bien posés : pour une donnée initiale suffisamment régulière, le taux de convergence du schéma est égal à son ordre de convergence. Toutefois, ce résultat peut être amélioré. En effet, le taux de convergence obtenu dans le cas d'une donnée initiale peu régulière n'est pas proche de celui observé numériquement.

Afin de calculer de manière optimale le taux de convergence, nous avons défini une nouvelle classe de schémas pour laquelle nous avons démontré une nouvelle expression du taux de convergence valable dans le cas de la dimension 1. Cette nouvelle valeur est très satisfaisante. En effet, on retrouve le taux de convergence observé numériquement. De plus, la classe de schémas étudiée est suffisamment large pour que, lorsque nous choisissons correctement la discrétisation du terme d'ordre 0, les schémas classiques en fassent partie. Seul le schéma décentré nécessite un traitement particulier.

Nous avons donc atteint notre objectif, à savoir, donner une nouvelle théorie des schémas numériques pour les problèmes faiblement bien posés qui explique pourquoi des schémas instables au sens de la définition usuelle peuvent converger quand même. Il nous reste à améliorer le calcul du taux de convergence. En effet, il serait intéressant de pouvoir calculer le taux de convergence de manière optimale dans le cas multidimensionnel et d'étendre la classe de schémas pour lequel le résultat est valable. Une piste intéressante serait d'utiliser la théorie des décréments [7].

De plus, dans cette thèse, nous n'avons étudié que le problème de Cauchy. Une

autre perspective est donc d'étendre la théorie développée ici dans le cas de problèmes aux limites et de voir si une extension de la théorie GKS au cas des problèmes faiblement bien posés est possible.

Dans la seconde partie de cette thèse nous allons étudier plus particulièrement les problèmes PML qui étaient la motivation de la première partie. Nous allons étudier différentes causes d'instabilité de ces problèmes et l'une d'entre elles sera le caractère faiblement bien posé. Sur l'exemple des équations de Maxwell, nous calculerons le défaut du problème continu et le défaut de stabilité faible du schéma de Yee. Nous avons donc un cas concret d'application des résultats obtenus dans cette partie.

Deuxième partie

Stabilité des PML

Chapitre 6

Généralités

6.1 Les premières PML de Bérenger

Les couches parfaitement adaptées (Perfectly Matched Layers) ont été introduites par Bérenger dans [12]. Le but de cette méthode est d'étudier la propagation d'ondes électromagnétiques en domaine non borné. Le problème étant que, lorsque l'on se restreint au domaine d'intérêt, des conditions aux limites sont nécessaires.

Le principe de la technique de Bérenger est d'introduire une couche autour du domaine d'intérêt dans laquelle les équations sont modifiées. Cette couche ne va pas influer sur les équations à l'intérieur du domaine d'intérêt.

Cette méthode a acquis une énorme popularité car elle ne pose pas de problème de condition aux limites, elle traite le problème des coins et elle est simple à mettre en oeuvre.

6.1.1 Ecriture des équations

Nous considèrons les équations de Maxwell transverses électriques (TE) dans un milieu de permittivité ε_0, de perméabilité μ_0 et de pertes électriques et magnétiques σ et σ^*. Si (E_x, E_y) représente le champ électrique et H_z le champ magnétique, les équations sont alors les suivantes :

$$\begin{cases} \varepsilon_0 \partial_t E_x - \partial_y H_z + \sigma E_x = 0 \\ \varepsilon_0 \partial_t E_y + \partial_x H_z + \sigma E_y = 0 \\ \mu_0 \partial_t H_z - \partial_y E_x + \partial_x E_y + \sigma^* H_z = 0. \end{cases}$$

Pour obtenir les équations PML, nous effectuons deux opérations. La première est le splitting. Il s'agit de séparer le champ magnétique en deux composantes non physiques afin de n'avoir qu'une dérivée partielle en espace dans chaque équation. Nous avons alors les équations :

$$\begin{cases} \varepsilon_0 \partial_t E_x - \partial_y (H_{zx} + H_{zy}) + \sigma E_x = 0 \\ \varepsilon_0 \partial_t E_y + \partial_x (H_{zx} + H_{zy}) + \sigma E_y = 0 \\ \mu_0 \partial_t H_{zx} + \partial_x E_y + \sigma^* H_{zx} = 0 \\ \mu_0 \partial_t H_{zy} - \partial_y E_x + \sigma^* H_{zy} = 0. \end{cases}$$

Nous remarquons que si $H_z = H_{zx} + H_{zy}$, en sommant les deux dernières équations, nous retrouvons la solution des équations de Maxwell.

La deuxième manipulation est une modification de l'absorption (ou une introduction de l'absorption dans le cas du vide). Nous obtenons alors les équations PML :

$$\begin{cases} \varepsilon_0 \partial_t E_x - \partial_y (H_{zx} + H_{zy}) + \sigma_y E_x = 0 \\ \varepsilon_0 \partial_t E_y + \partial_x (H_{zx} + H_{zy}) + \sigma_x E_y = 0 \\ \mu_0 \partial_t H_{zx} + \partial_x E_y + \sigma_x^* H_{zx} = 0 \\ \mu_0 \partial_t H_{zy} - \partial_y E_x + \sigma_y^* H_{zy} = 0. \end{cases} \tag{6.1}$$

6.1.2 Etude de la réflexion

L'étude de la réflexion se fait en étudiant la propagation des ondes planes entre deux milieux PML de coefficients d'absorption différents séparés par l'interface $x = 0$:

$$PML(\sigma_{x1}, \sigma_{x1}^*, \sigma_{y1}, \sigma_{y1}^*) \;\Big|\; PML(\sigma_{x2}, \sigma_{x2}^*, \sigma_{y2}, \sigma_{y2}^*)$$

Nous supposons que l'on a :

$$\frac{\sigma_{x1}}{\varepsilon_0} = \frac{\sigma_{x1}^*}{\mu_0}, \; \frac{\sigma_{x2}}{\varepsilon_0} = \frac{\sigma_{x2}^*}{\mu_0}, \; \frac{\sigma_{y1}}{\varepsilon_0} = \frac{\sigma_{y1}^*}{\mu_0}, \; \frac{\sigma_{y2}}{\varepsilon_0} = \frac{\sigma_{y2}^*}{\mu_0}.$$

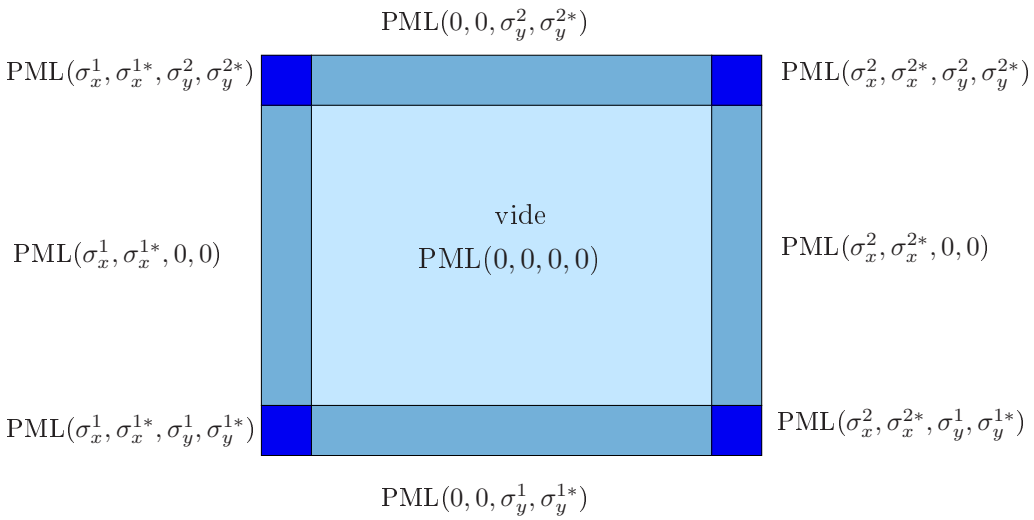
Ces relations sont analogues à $\frac{\sigma}{\varepsilon_0} = \frac{\sigma^*}{\mu_0}$, ce qui pour les équations de Maxwell, signifie que l'impédance du milieu est égale à celle du vide et qu'il n'y a pas de réflexion entre le vide et le milieu pour une onde se propageant perpendiculairement à l'interface.

Nous supposons de plus :

$$\sigma_{y1} = \sigma_{y2}, \; \sigma_{y1}^* = \sigma_{y2}^*.$$

Alors, il n'y pas de réflexion à l'interface $x = 0$ et cela est vrai pour toute onde plane, quelles que soient sa fréquence et son incidence.

Si nous voulons étudier les équations PML dans le vide, le schéma ci-dessous donne donc des couches sans aucune réflexion aux interfaces.



6.1.3 Etude de l'absorption

L'étude par ondes planes effectuée dans le cas précédent où l'interface est $x = 0$, montre que les ondes sont absorbées exponentiellement lorsque $\sigma_x \neq 0$ et $\sigma_x^* \neq 0$ ce qui est bien le cas dans la configuration précédente. Le problème de la condition aux limites au bord extérieur de la couche va donc être de moindre importance.

6.1.4 Un exemple pratique

Pour mettre en oeuvre la méthode des PML de manière concrète, il reste à déterminer la taille de la couche absorbante ainsi que la forme du coefficient d'absorption.

Dans [4], la propagation des ondes électromagnétiques est étudiée dans une bande de direction, l'axe des x. Le domaine d'intérêt est pour $x \in [-50, 50]$. Dans ce domaine, les équations considérées sont les équations PML avec absorption nulle. Pour $50 < |x| < 60$, nous considérons les équations PML avec une absorption σ_x non nulle. Il s'agit de la partie "couche absorbante" du domaine.

FIG. 6.1 – Domaine d'étude

Le coefficient d'absorption doit être continu. La forme standard [4] du coefficient d'absorption dans la couche absorbante est un polynôme de degré compris entre 2 et 4. Ici, nous représentons le cas d'un polynôme de degré 3.

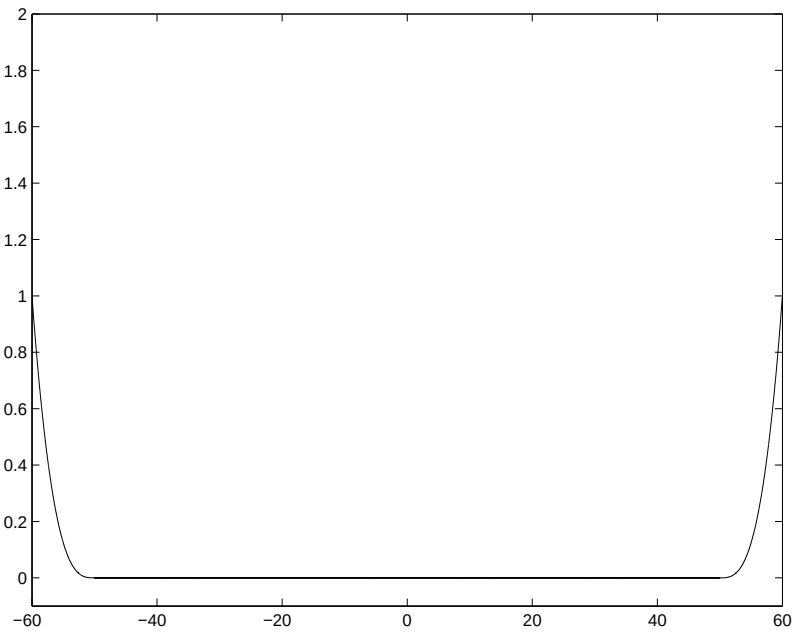


FIG. 6.2 – Absorption σ_x

6.2 Application à d'autres équations

6.2.1 Equations de Maxwell tridimensionnelles

Bérenger a très rapidement étendu les résultats obtenus pour les cas des équations de Maxwell TE en dimension 2 aux équations de Maxwell tridimensionnelles [13]. La technique utilisée reste analogue au cas de la dimension deux, à savoir une étude par ondes planes.

6.2.2 Equations d'Euler linéarisées

Les techniques utilisées pour l'électromagnétisme ont alors été étendues par Hu [21] aux équations d'Euler linéarisées avec flot uniforme. Dans cet article, Hu montre, toujours en manipulant les ondes planes, qu'une transformation identique à celle de Bérenger, à savoir splitting puis absorption, conduit bien, dans le cas des équations d'Euler à des couches parfaitement adaptées.

6.2.3 Les PML comme changement complexe de variable

Si les PML se généralisent facilement à d'autres équations que celles de l'électromagnétisme pour lesquelles elles ont été crées, c'est parce que la transformation effectuée ne dépend pas des équations. En effet, nous pouvons obtenir les équations PML à partir d'un changement complexe de variables [16].

En effet si nous considérons le système général hyperbolique dans $\mathbb{R}^2$ suivant :

$$\partial_t U - A_1 \partial_x U - A_2 \partial_y U = 0,$$

où A_1, $A_2 \in \mathcal{M}_N(\mathbb{R})$.

Les équations PML associées à ce problème avec absorption dans la direction des x sont alors :

$$\begin{cases} \partial_t U^1 - A_1 \partial_x (U^1 + U^2) + \sigma_x U^1 = 0 \\ \partial_t U^2 - A_2 \partial_y (U^1 + U^2) = 0, \end{cases}$$

où $\sigma_x = \sigma_x(x) > 0$ désigne l'absorption dans la direction des x.

Nous appliquons alors au système précédent une transformation de Laplace en temps. Les équations fréquentielles en la variable duale du temps notée ω sont alors :

$$-i\omega \mathcal{L}(U^1) - A_1 \partial_x (\mathcal{L}(U^1) + \mathcal{L}(U^2)) + \sigma_x \mathcal{L}(U^1) = 0 \tag{6.2}$$

$$-i\omega \mathcal{L}(U^2) - A_2 \partial_y (\mathcal{L}(U^1) + \mathcal{L}(U^2)) = 0. \tag{6.3}$$

En multipliant (6.2) par $-i\omega$ et (6.3) par $-i\omega + \sigma_x$, et en ajoutant les deux équations, nous obtenons l'équation suivante :

$$-i\omega(-i\omega + \sigma_x)\mathcal{L}(U^1 + U^2) + i\omega A_1 \partial_x \mathcal{L}(U^1 + U^2) - (-i\omega + \sigma_x) A_2 \partial_y \mathcal{L}(U^1 + U^2) = 0.$$

En divisant cette équation par $-i\omega + \sigma_x$ et en posant $U = U^1 + U^2$, nous avons :

$$-i\omega\mathcal{L}(U) - \frac{i\omega}{i\omega - \sigma_x} A_1 \partial_x \mathcal{L}(U) - A_2 \partial_y \mathcal{L}(U) = 0.$$

Nous retrouvons donc la transformée de Laplace de l'équation de départ où seule la dérivée partielle par rapport à la variable x a été remplacée par $\frac{i\omega}{i\omega - \sigma_x}\partial_x$. Cette transformation correspond au changement de variable :

$$x \to x - \frac{1}{i\omega} \int_0^x \sigma_x(u) du. \tag{6.4}$$

L'utilisation de ce changement de variable permet de construire des PML pour des équations diverses et de les étudier de façon assez générale.

6.3 Les problèmes rencontrés

6.3.1 La régularité

La transformée de Fourier des solutions des équations PML pour les équations de Maxwell TE sans absorption a été calculée de manière exacte par Abarbanel et Gottlieb [1]. En effectuant ce calcul, ils ont observé que le problème n'était que faiblement bien posé de défaut 1. Ainsi, il existe une perturbation d'ordre 0 tel que le problème devienne mal posé ce qui pose, a priori, beaucoup de difficultés. Toutefois, la perturbation d'ordre 0 correspondant à l'absorption dans les équations PML a une forme précise et dans ce cas le problème PML reste faiblement bien posé. Dans [9], un résultat général prouve que, sous certaines hypothèses, le problème PML est faiblement bien posé. Nous énoncerons ce résultat ultérieurement dans le chapitre 9. Toutefois, ce résultat ne permet pas le calcul du défaut. Nous rappelons que nous avons montré dans la proposition 1.11, que, de manière générale, le défaut associé à un problème PML sans absorption valait 1. Nous calculerons dans la partie 7 le défaut associé aux équations de Maxwell PML avec absorption.

La perte de régularité entraîne obligatoirement une croissance polynomiale en temps. En effet, lorsque nous calculons la transformée de Fourier de la solution, des termes en ξt apparaissent. La partie polynomiale en ξ va donner la perte de régularité et la partie en t la croissance polynomiale en temps. Ce type d'instabilité s'observe numériquement à temps long. En effet, Abarbanel, Gottlieb et Hesthaven ont étudié numériquement les équations PML associées aux équations de Maxwell dans [4]. Ils observent que, longtemps après que l'onde soit totalement sortie du domaine, une onde se créée à l'intérieur des couches absorbantes et cette onde va être propagée dans le domaine d'intérêt. Ce type de comportement est dû au caractère faiblement bien posé des équations. Nous expliquerons au début de la partie 7, les différentes

méthodes pour y remédier, puis nous donnerons des estimations d'énergie sans perte de régularité pour les équations de Maxwell PML proposées par Bérenger.

6.3.2 L'instabilité asymptotique

Les équations fortement hyperboliques traitées sont homogènes : il n'y a donc pas de croissance exponentielle en temps dans les estimations d'énergie qui sont donc de la forme : $\|U(t,.)\|_{L^2} \leq K\|U^0\|_{L^2}$. Dans les équations PML, le terme d'ordre 0 crée a priori une croissance exponentielle en temps, correspondant à une instabilité asymptotique. Nous parlerons de stabilité, au sens de [9], lorsque la croissance ne sera que polynomiale en temps. Nous ne pouvons pas exclure la croissance polynomiale en temps, car, comme nous l'avons expliqué précédemment, elle est liée à la perte de régularité. Ces questions de stabilité ont été étudiées dans [9], [11] et [6].

Ce type d'instabilité a été observé numériquement dans [22] et [9]. Il ne s'agit plus d'instabilités à temps long mais toujours d'une croissance en temps anormale provenant de la couche absorbante. La cause de cette instabilité est liée à la direction de la vitesse de groupe et de la vitesse de phase. Nous expliquerons de manière plus détaillée ce type de problèmes dans le chapitre 9. De plus, nous généraliserons l'étude faite dans [9] au cas de problèmes ayant une absorption à coefficients variables.

Chapitre 7

Estimations d'énergie pour les équations de Maxwell PML

Le but de ce chapitre est de donner des estimations d'énergie sur les solutions des équations de Maxwell PML. Ces estimations d'énergie sont indispensables pour pouvoir appliquer la théorie des schémas exposée dans la première partie de cette thèse. En effet, elles vont permettre de calculer le défaut q_1 du problème PML.

Nous allons commencer par décrire les différentes méthodes proposées par les auteurs pour éviter ce problème. En effet, il existe des techniques permettant de modifier les équations PML afin de les rendre fortement bien posées. Toutefois, ces méthodes peuvent présenter des inconvénients.

Nous présenterons ensuite les estimations d'énergie déjà connues et expliquerons quels sont les avantages des méthodes présentées dans cette thèse.

7.1 Motivations

7.1.1 Des équations PML fortement bien posées

Nous allons nous intéresser ici aux modifications des équations PML conduisant à un problème fortement bien posé mais pour lesquelles le comportement en temps de la solution n'a pas été particulièrement étudié. Nous présentons les résultats obtenus pour les équations de Maxwell et d'Euler.

La première méthode a été proposée dans [2]. Elle concerne les équations de Maxwell dans un milieu avec pertes. Les nouvelles équations PML proposées dans cet article sont constituées d'une perturbation d'ordre 0 des équations de Maxwell ainsi que de deux variables auxiliaires vérifiant des équations différentielles (au lieu d'équations aux dérivées partielles). Comme les équations différentielles ne font qu'ajouter des termes nuls dans le symbole et qu'une perturbation d'ordre 0 d'un

problème fortement bien posé est toujours fortement bien posée, les nouvelles équations obtenues sont fortement bien posées.

La même méthode a été utilisée dans [20] pour les équations d'Euler. Le modèle PML de départ est celui qui a été proposé par Hu [21]. Hesthaven montre que ce problème n'est que faiblement bien posé et le modifie ensuite.

L'inconvénient majeur de cette méthode est que le caractère parfaitement adapté est perdu. Toutefois, les résultats numériques restent satisfaisants.

Pour résoudre ce problème, une méthode proche de celle proposée dans [20] a été étudiée dans [3] pour les équations d'Euler linéarisées à flot constant. Les nouvelles équations sont constituées des équations initiales auxquelles un terme d'ordre 0 a été ajouté, ainsi que de deux nouvelles variables intervenant dans des équations différentielles ordinaires et d'une autre variable intervenant dans une équation aux dérivées partielles.

Une méthode plus algébrique a été proposée par Rahmouni dans [34] et [35] pour les équations d'Euler linéarisées. Le principe de cette méthode est d'utiliser des transformations algébriques pour obtenir un modèle fortement bien posé qui est non local et ensuite de le localiser en introduisant une nouvelle inconnue. Le système obtenu alors est symétrique hyperbolique, ce qui était une propriété importante du système initialement étudié. De plus, cette méthode présente l'avantage d'être généralisable à d'autres systèmes hyperboliques. Enfin, ce modèle est parfaitement adapté lorsque le coefficient d'absorption n'est pas constant.

7.1.2 Revue des estimations d'énergie connues

Des estimations d'énergie ont été prouvées dans [32], [31] et [10] pour des équations PML associées aux équations de Maxwell.

Dans [32], les équations PML considérées sont celles de Bérenger pour une absorption uniquement dans la direction de l'axe des x, mais les variables sont : $(E_x, E_y, H_z, \sigma_x H_{zx})$. La norme H^1 de la solution à l'instant t est alors contrôlée par la norme H^1 de la donnée initiale, une croissance en temps étant autorisée.

Les PML considérés dans [31] sont les PML proposées par Bérenger dans le cas des équations de Maxwell tridimensionnelles mais le terme d'ordre 0 a été régularisé par convolution. Cette manipulation permet d'avoir des estimations d'énergie sans pertes pour la norme $(L^2(\mathbb{R}^3))^6 \times (H^{-1}(\mathbb{R}^3))^6$, les six premières variables représentant les trois composantes du champ électrique qui sont chacune splittées en deux composantes non physiques, et de même pour les six dernières avec le champ magnétique. Des estimations similaires ont été obtenues pour les équations d'Euler linéarisées.

Enfin, des estimations d'énergie ont été proposées dans [10] pour la formulation de Zhao-Cangellaris [33] des équations PML pour les équations de Maxwell bidimensionnelles. Cette formulation a été créée dans le but d'obtenir une formulation

varitionnelle des équations PML qui pourra permettre d'appliquer une méthode d'éléments finis pour l'étude numérique. Cette formulation est une perturbation d'ordre 0 des équations de Maxwell, elle est donc fortement bien posée. Des estimations d'énergie dans lesquelles toutes les constantes sont explicitées sont données. Une autre étude est faite pour le modèle initial proposé par Bérenger dans le cas d'une absorption constante dans la direction de l'axe x. Deux estimations sont alors obtenues, elles sont de la forme : $\|(E_x, E_y, H_z)(t,.)\|_{L^2} \leq C_1 \|(E_x^0, E_y^0, H_{zx}^0, H_{zy}^0)\|_{L^2}$ et $\|(H_{zx}, H_{zy})(t,.)\|_{H^{-1}} \leq C_2 t \|(E_x^0, E_y^0, H_{zx}^0, H_{zy}^0)\|_{L^2}$.

7.1.3 Méthodes proposées

Dans cette thèse, nous allons présenter deux méthodes différentes qui conduisent à des estimations d'énergie. Notre but est d'étudier les équations PML proposées initialement par Bérenger et non pas les problèmes fortement bien posés proposés par la suite. En effet, la question à laquelle nous souhaitons répondre est : pourquoi, malgré les problèmes théoriquement rencontrés, les équations de Bérenger conduisent-elles à des résultats numériques satisfaisants ? Toutefois, même si nous ne modifions pas les équations, nous n'allons pas étudier les variables standards $(E_x, E_y, H_{zx}, H_{zy})$ mais des variables qui en découlent.

La première méthode est basée sur le symbole principal des équations PML. Son avantage sur la méthode proposée dans [10] est qu'elle permet d'étudier le cas d'une absorption variable et qui est dans les deux directions d'espace. De plus, par rapport à la méthode de [32], qui n'avait pas non plus traité le problème d'une absorption dans les deux directions, cette méthode se généralise aux équations de Maxwell tridimentionnelles et sera aussi appliquable au schéma de Yee (voir chapitre 8).

La seconde méthode permet de redémontrer le résultat obtenu dans [32] de manière plus rigoureuse et plus détaillée. Cette méthode est basée sur une version semi-discrétisée des équations qui vont approcher le problème continu.

7.2 Etude du symbole

Dans cette partie, nous allons donner des estimations d'énergie formelles sans perte pour les équations de Maxwell PML à coefficients variables. Nous étudierons le symbole d'un problème prenant en compte les variables issues du champ électromagnétique ainsi que certaines de leurs dérivées. Nous choisirons ces variables de manière à obtenir un problème fortement bien posé.

Pour justifier l'étude sur le symbole alors que l'absorption est à coefficients variables, nous aurons besoin d'un théorème de Kreiss.

7.2.1 Théorème de Kreiss

Nous allons utiliser un cas particulier du théorème suivant issu de [27]. Ce théorème est l'analogue du théorème 1.2 dans le cas de coefficients variables.

Théorème 7.1 *Nous considérons un système hyperbolique du premier ordre :*

$$\partial_t U = P(x,t,\partial_x)U + C(x,t)U$$

où $x \in \mathbb{R}^s$, $t \geq 0$ et C régulière et bornée. Nous supposons que le problème est fortement hyperbolique, c'est-à-dire qu'il existe une matrice hermitienne régulière $H(x,t,k)$ définie positive telle que :

$$H(x,t,k)P(x,t,ik) + P^*(x,t,ik)H(x,t,k) = 0$$

Alors le problème de Cauchy est fortement bien posé.

Remarque 7.1 *Cette définition d'un problème fortement hyperbolique est cohérente avec celle donnée dans la première partie pour les problèmes à coefficients constants. En effet, dans ce cas, l'existence d'une telle matrice hermitienne caractérise le fait d'être fortement bien posé.*

Ici, nous n'allons appliquer ce théorème que dans le cas où P est à coefficients constants et seul C sera à coefficients variables. Dans ce cas particulier, un problème fortement hyperbolique au sens précédent est un problème dont la partie principale est fortement hyperbolique.

De plus, nous pouvons préciser l'estimation d'énergie en regardant la dépendance par rapport à C. En appliquant la démonstration de [27], nous avons :

$$\|U(t)\|_{L^2} \leq K e^{(K' + \|C\|_\infty)t\|} \|U^0\|_{L^2}.$$

7.2.2 Equations de Maxwell PML en dimension 2

Nous considérons ici, pour simplifier les calculs, des équations PML adimensionnées : $\varepsilon_0 = \mu_0 = 1$. Les absorptions vérifient alors : $\sigma_x = \sigma_x^*$ et $\sigma_y = \sigma_y^*$.

Dans [32], Métral et Vacus montrent, dans le cas d'une absorption suivant l'axe O_x i.e. $\sigma_y = 0$, des estimations d'énergie sans perte pour les variables $(E_x, E_y, H_z, G = \sigma_x H_{zx})$, provenant des équations PML de Bérenger, en norme $\left(H^1(\mathbb{R}^2)\right)^3 \times L^2(\mathbb{R}^3)$. Nous allons montrer ici ces estimations de manière formelle en utilisant le symbole. L'intérêt de cette méthode est qu'elle se généralise au cas de la dimension 3.

Proposition 7.1 *Supposons que σ_x, $\sigma_y \in W^{1,\infty}(\mathbb{R}^2)$, nous avons l'estimation d'énergie suivante pour les solutions de (6.1) homogénéisées :*

$$\exists C_1,\, C_2 > 0,\, \forall t \geq 0,\, \forall (E_x^0, E_y^0, H_z^0, H_{zx}^0) \in \left(H^1(\mathbb{R}^2)\right)^3 \times L^2(\mathbb{R}^3),$$
$$\forall \sigma_x(x,y),\, \sigma_y(x,y) \in \mathcal{C}_C^\infty(\mathbb{R}^2),$$
$$\|(E_x, E_y, H_z)(t,.)\|_{H^1}^2 + \|(\sigma_y E_x, \sigma_x E_y, \sigma_x H_{zx} + \sigma_y H_{zy})(t,.)\|_{L^2}^2$$
$$\leq C_1 e^{(C_2 + 2(\|\sigma_x\|_{W^{1,\infty}} + \|\sigma_y\|_{W^{1,\infty}} + \|\sigma_x\|_{W^{1,\infty}} \|\sigma_y\|_{W^{1,\infty}}))t}$$
$$\times \left(\|(E_x^0, E_y^0, H_z^0)\|_{H^1}^2 + \|(\sigma_y E_x^0, \sigma_x E_y^0, \sigma_x H_{zx}^0 + \sigma_y H_{zy}^0)\|_{L^2}^2\right).$$

PREUVE : La méthode consiste à écrire les systèmes vérifiés par les champs et leurs dérivées et à se ramener à un problème fortement bien posé. Plus précisément, nous posons :

$U^1 = {}^t(E_x, E_y, H_z)$, $U^2 = \partial_x U^1$, $U^3 = \partial_y U^1$, et $V = {}^t(\sigma_y E_x, \sigma_x E_y, \sigma_x H_{zx} + \sigma_y H_{zy})$,

avec $H_z = H_{zx} + H_{zy}$. Soit $\widetilde{P}(\partial)$, l'opérateur correspondant aux équations de Maxwell :

$$\widetilde{P}(\partial) = \begin{pmatrix} 0 & 0 & -\partial_y \\ 0 & 0 & \partial_x \\ -\partial_y & \partial_x & 0 \end{pmatrix}.$$

Nous considérons les matrices B et $C(x,y)$:

$$B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } C = \begin{pmatrix} \sigma_x(x,y) & 0 & 0 \\ 0 & \sigma_y(x,y) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Alors, nous avons les équations suivantes :

$$\begin{cases} \partial_t U^1 + \widetilde{P}(\partial) U^1 + C(x,y) U^1 + BV = 0 \\ \partial_t U^2 + \widetilde{P}(\partial) U^2 + \partial_x C(x,y) U^1 + C(x,y) U^2 + B\partial_x V = 0 \\ \partial_t U^3 + \widetilde{P}(\partial) U^3 + \partial_y C(x,y) U^1 + C(x,y) U^3 + B\partial_y V = 0 \end{cases}$$

et nous ajoutons une équation sur V :

$$\begin{aligned} \partial_t V \;+\; & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\sigma_x(x,y)\sigma_y(x,y) \end{pmatrix} U^1 + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \sigma_x(x,y) \\ 0 & \sigma_x(x,y) & 0 \end{pmatrix} U^2 \\ + & \begin{pmatrix} 0 & 0 & -\sigma_y(x,y) \\ 0 & 0 & 0 \\ -\sigma_y(x,y) & 0 & 0 \end{pmatrix} U^3 \\ + & \begin{pmatrix} \sigma_y(x,y) & 0 & 0 \\ 0 & \sigma_x(x,y) & 0 \\ 0 & 0 & \sigma_x(x,y) + \sigma_y(x,y) \end{pmatrix} V = 0. \end{aligned}$$

Montrons que le système constitué de ces quatre équations est fortement bien posé.

La partie principale de ce système d'équations en la variable (U^1, U^2, U^3, V) est bien à coefficients constants et son symbole est $P(i\xi)$:

$$P(i\xi) = \begin{pmatrix} \widetilde{P}(i\xi) & 0 & 0 & 0 \\ 0 & \widetilde{P}(i\xi) & 0 & i\xi_1 B \\ 0 & 0 & \widetilde{P}(i\xi) & i\xi_2 B \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Comme $\widetilde{P}(i\xi)$ est le symbole des équations de Maxwell, nous savons qu'il existe $\widetilde{S}(\xi)$ telle que $\widetilde{D}(\xi) = \widetilde{S}(\xi)\widetilde{P}(i\xi)\widetilde{S}(\xi)^{-1}$ est diagonale de valeurs propres imaginaires pures et $\|\widetilde{S}(\xi)\| + \|\widetilde{S}(\xi)^{-1}\|$ est borné.

Remarque 7.2 *Nous pouvons calculer explicitement $\widetilde{D}$ et $\widetilde{S}$. Nous obtenons :*

$$\widetilde{D} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & i\|\xi\| & 0 \\ 0 & 0 & i\|\xi\| \end{pmatrix} \text{ et } \widetilde{S} = \frac{1}{\|\xi\|} \begin{pmatrix} \xi_1 & -\xi_2 & \xi_2 \\ \xi_2 & \xi_1 & -\xi_1 \\ 0 & \|\xi\| & \|\xi\| \end{pmatrix}.$$

Posons :

$$M_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \text{ et } M_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Alors, nous avons :

$$\widetilde{P}(i\xi)M_1 = i\xi_1 B \text{ et } \widetilde{P}(i\xi)M_2 = i\xi_2 B.$$

Nous considérons la matrice de passage :

$$S(\xi) = \begin{pmatrix} \widetilde{S}(\xi) & 0 & 0 & 0 \\ 0 & \widetilde{S}(\xi) & 0 & \widetilde{S}(\xi)M_1 \\ 0 & 0 & \widetilde{S}(\xi) & \widetilde{S}(\xi)M_2 \\ 0 & 0 & 0 & Id \end{pmatrix}.$$

Son inverse est :

$$S(\xi)^{-1} = \begin{pmatrix} \widetilde{S}(\xi)^{-1} & 0 & 0 & 0 \\ 0 & \widetilde{S}(\xi)^{-1} & 0 & -M_1 \\ 0 & 0 & \widetilde{S}(\xi)^{-1} & -M_2 \\ 0 & 0 & 0 & Id \end{pmatrix}.$$

et nous avons :

$$
\begin{aligned}
&S(\xi)P(i\xi)S(\xi)^{-1}\\
&\quad= \begin{pmatrix}
\widetilde{D}(\xi) & 0 & 0 & 0\\
0 & \widetilde{D}(\xi) & 0 & -\widetilde{S}(\xi)\widetilde{P}(i\xi)M_1 + i\xi_1\widetilde{S}(\xi)B\\
0 & 0 & \widetilde{D}(\xi) & -\widetilde{S}(\xi)\widetilde{P}(i\xi)M_2 + i\xi_2\widetilde{S}(\xi)B\\
0 & 0 & 0 & 0
\end{pmatrix}\\
&\quad= \begin{pmatrix}
\widetilde{D}(\xi) & 0 & 0 & 0\\
0 & \widetilde{D}(\xi) & 0 & 0\\
0 & 0 & \widetilde{D}(\xi) & 0\\
0 & 0 & 0 & 0
\end{pmatrix}.
\end{aligned}
$$

Donc $S(\xi)P(i\xi)S(\xi)^{-1}$ est diagonale à coefficients imaginaires pures et $\|S(\xi)\| + \|S(\xi)^{-1}\|$ est borné. Ainsi le problème en les variables (U^1, U^2, U^3, V) est fortement bien posé et comme la norme infinie du coefficient du terme d'ordre 0 est majorée par $\|\sigma_x\|_{W^{1,\infty}} + \|\sigma_y\|_{W^{1,\infty}} + \|\sigma_x\|_{W^{1,\infty}}\|\sigma_y\|_{W^{1,\infty}}$, en utilisant le théorème de Kreiss, nous obtenons le résultat voulu. □

7.2.3 Equations de Maxwell PML en dimension 3

Nous considérons ici les équations PML pour les équations de Maxwell en dimension 3 avec une absorption dans la direction de l'axe Ox :

$$
\left\{
\begin{array}{l}
\partial_t E_{xy} - \partial_y(H_{zx} + H_{zy}) = 0\\
\partial_t E_{xz} + \partial_z(H_{yz} + H_{yx}) = 0\\
\partial_t E_{yz} - \partial_z(H_{xy} + H_{xz}) = 0\\
\partial_t E_{yx} + \partial_x(H_{zx} + H_{zy}) + \sigma E_{yx} = 0\\
\partial_t E_{zx} - \partial_x(H_{yz} + H_{yx}) + \sigma E_{zx} = 0\\
\partial_t E_{zy} + \partial_y(H_{xy} + H_{xz}) = 0\\
\partial_t H_{xy} + \partial_y(E_{zx} + E_{zy}) = 0\\
\partial_t H_{xz} - \partial_z(E_{yz} + E_{yx}) = 0\\
\partial_t H_{yz} + \partial_z(E_{xy} + E_{xz}) = 0\\
\partial_t H_{yx} - \partial_x(E_{zx} + E_{zy}) + \sigma H_{yx} = 0\\
\partial_t H_{zx} + \partial_x(E_{yz} + E_{yx}) + \sigma H_{zx} = 0\\
\partial_t H_{zy} - \partial_y(E_{xy} + E_{xz}) = 0.
\end{array}
\right. \tag{7.1}
$$

Nous avons alors l'estimation suivante :

Théorème 7.2

$$\exists C_1,\, C_2 > 0,\, \forall t \geq 0,\, \forall \sigma(x,y,z) \in \mathcal{C}_C^\infty(\mathbb{R}^3),$$

$$\forall (E_x^0, E_y^0, E_z^0, H_x^0, H_y^0, H_z^0, E_{yx}^0, E_{zx}^0, H_{zx}^0, H_{yx}^0) \in \left(H^2(\mathbb{R}^3)\right)^{10},$$

$$\|(E_x, E_y, E_z, H_x, H_y, H_z)(t,.)\|_{(H^2)^6} + \|\sigma(E_x, E_{yx}, E_{zx}, H_x, H_{zx}, H_{yx})(t,.)\|_{(H^2)^6}$$
$$\leq C_1 e^{(C_2+2\|\sigma\|_{W^{1,\infty}})t}$$
$$\times \left(\|(E_x^0, E_y^0, E_z^0, H_x^0, H_y^0, H_z^0)\|_{(H^2)^6} + \|\sigma(E_x^0, E_{yx}^0, E_{zx}^0, H_x^0, H_{zx}^0, H_{yx}^0)\|_{(H^2)^6}\right).$$

Remarque 7.3 *Nous avons choisi de prendre l'absorption dans une seule direction afin de simplifier l'écriture et de réduire le nombre d'équations à étudier. Toutefois, cette méthode se généralise au cas d'une absorption dans les trois directions.*

PREUVE : Nous posons :

$$U^1 = (E_x, E_y, E_z, H_x, H_y, H_z),$$

$$U^2 = \partial_x U^1,\ U^3 = \partial_y U^1,\ U^4 = \partial_z U^1,\ U^5 = \partial_{xx} U^1,\ U^6 = \partial_{xy} U^1,\ U^7 = \partial_{xz} U^1,$$

$$V^1 = {}^t(0, \sigma E_{yx}, \sigma E_{zx}, 0, \sigma H_{yx}, \sigma H_{zx}),\ V^2 = \partial_x V^1,\ V^3 = \partial_y V^1,\ V^4 = \partial_z V^1,$$

$$V^5 = {}^t(\partial_{xx}(\sigma E_x) + \partial_{xy}(\sigma E_{yx}) + \partial_{xz}(\sigma E_{zx}), 0, 0, \partial_{xx}(\sigma H_x) + \partial_{xy}(\sigma H_{yx}) + \partial_{xz}(\sigma H_{zx})),$$

$$V^6 = {}^t(\partial_{xy}(\sigma E_x) + \partial_{yy}(\sigma E_{yx}) + \partial_{yz}(\sigma E_{zx}), 0, 0, \partial_{xy}(\sigma H_x) + \partial_{yy}(\sigma H_{yx}) + \partial_{yz}(\sigma H_{zx})),$$

$$V^7 = {}^t(\partial_{xz}(\sigma E_x) + \partial_{yz}(\sigma E_{yx}) + \partial_{zz}(\sigma E_{zx}), 0, 0, \partial_{xz}(\sigma H_x) + \partial_{yz}(\sigma H_{yx}) + \partial_{zz}(\sigma H_{zx})).$$

Soit $\widetilde{P}(\partial)$, l'opérateur correspondant aux équations de Maxwell :

$$\widetilde{P}(\partial) = \begin{pmatrix} 0 & 0 & 0 & 0 & \partial_z & -\partial_y \\ 0 & 0 & 0 & -\partial_z & 0 & \partial_x \\ 0 & 0 & 0 & \partial_y & -\partial_x & 0 \\ 0 & -\partial_z & \partial_y & 0 & 0 & 0 \\ \partial_z & 0 & -\partial_x & 0 & 0 & 0 \\ -\partial_y & \partial_x & 0 & 0 & 0 & 0 \end{pmatrix}.$$

- Nous avons les équations suivantes :

$$\begin{cases} \partial_t U^1 + \widetilde{P}(\partial) U^1 + V^1 = 0 \\ \partial_t U^2 + \widetilde{P}(\partial) U^2 + V^2 = 0 \\ \partial_t U^3 + \widetilde{P}(\partial) U^3 + V^3 = 0 \\ \partial_t U^4 + \widetilde{P}(\partial) U^4 + V^4 = 0. \end{cases}$$

- Pour les composantes (U^5, U^6, U^7), nous avons :

$$\begin{cases} \partial_t U^5 + \widetilde{P}(\partial) U^5 + \partial_x V^2 = 0 \\ \partial_t U^6 + \widetilde{P}(\partial) U^6 + \partial_y V^2 = 0 \\ \partial_t U^7 + \widetilde{P}(\partial) U^7 + \partial_z V^2 = 0. \end{cases}$$

Il faut donc exprimer les dérivées premières de V^2 en fonction des $(U^j, V^j)_{j\in\{1..7\}}$. Or, nous avons :

$$\begin{aligned}
\partial_x V^2 &= \partial_{xx} \begin{pmatrix} 0 \\ \sigma E_{yx} \\ \sigma E_{zx} \\ 0 \\ \sigma H_{yx} \\ \sigma H_{zx} \end{pmatrix} \\
&= \widetilde{Q}(\partial)\partial_x \begin{pmatrix} 0 \\ \sigma E_{yx} \\ \sigma E_{zx} \\ 0 \\ \sigma H_{yx} \\ \sigma H_{zx} \end{pmatrix} + \begin{pmatrix} \partial_{xy}(\sigma E_{yx}) + \partial_{xz}(\sigma E_{zx}) \\ 0 \\ 0 \\ \partial_{xy}(\sigma H_{yx}) + \partial_{xz}(\sigma H_{zx}) \\ 0 \\ 0 \end{pmatrix} \\
&= \widetilde{Q}(\partial)V^2 + V^5 - \begin{pmatrix} \partial_{xx}(\sigma E_x) \\ 0 \\ 0 \\ \partial_{xx}(\sigma H_x) \\ 0 \\ 0 \end{pmatrix},
\end{aligned}$$

où l'on a posé :

$$\widetilde{Q}(\partial) = \begin{pmatrix} 0 & -\partial_y & -\partial_z & 0 & 0 & 0 \\ 0 & \partial_x & 0 & 0 & 0 & 0 \\ 0 & 0 & \partial_x & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\partial_y & -\partial_z \\ 0 & 0 & 0 & 0 & \partial_x & 0 \\ 0 & 0 & 0 & 0 & 0 & \partial_x \end{pmatrix}.$$

Or :

$$\begin{aligned}
\begin{pmatrix} \partial_{xx}(\sigma E_x) \\ 0 \\ 0 \\ \partial_{xx}(\sigma H_x) \\ 0 \\ 0 \end{pmatrix} &= \partial_{xx}\sigma \begin{pmatrix} E_x \\ 0 \\ 0 \\ H_x \\ 0 \\ 0 \end{pmatrix} + 2\partial_x\sigma \begin{pmatrix} \partial_x E_x \\ 0 \\ 0 \\ \partial_x H_x \\ 0 \\ 0 \end{pmatrix} + \sigma \begin{pmatrix} \partial_{xx} E_x \\ 0 \\ 0 \\ \partial_{xx} H_x \\ 0 \\ 0 \end{pmatrix} \\
&= \partial_{xx}\sigma T U^1 + 2\partial_x\sigma T U^2 + \sigma T U^5,
\end{aligned}$$

où :

$$T = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Donc :

$$\partial_x V^2 = \widetilde{M}(\partial)V^2 + V^5 - \partial_{xx}\sigma T U^1 - 2\partial_x \sigma T U^2 - \sigma T U^5.$$

En faisant de même pour les autres dérivées, nous obtenons les équations :

$$\left\{\begin{array}{l} \partial_t U^5 + \widetilde{P}(\partial)U^5 + \widetilde{Q}(\partial)V^2 + V^5 - \partial_{xx}\sigma T U^1 - 2\partial_x\sigma T U^2 - \sigma T U^5 = 0 \\ \partial_t U^6 + \widetilde{P}(\partial)U^6 + \widetilde{Q}(\partial)V^3 + V^6 - \partial_{xy}\sigma T U^1 - \partial_x\sigma T U^3 - \partial_y\sigma T U^2 - \sigma T U^6 = 0 \\ \partial_t U^7 + \widetilde{P}(\partial)U^7 + \widetilde{Q}(\partial)V^4 + V^7 - \partial_{xz}\sigma T U^1 - \partial_x\sigma T U^4 - \partial_z\sigma T U^2 - \sigma T U^7 = 0. \end{array}\right.$$

- Nous posons :

$$A = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right).$$

Nous avons alors :

$$\left\{\begin{array}{l} \partial_t V^1 + \sigma A U^2 + \sigma V^1 = 0 \\ \partial_t V^2 + \sigma A U^5 + \partial_x \sigma A U^2 + \sigma V^2 + \partial_x \sigma V^1 = 0 \\ \partial_t V^3 + \sigma A U^6 + \partial_y \sigma A U^2 + \sigma V^3 + \partial_y \sigma V^1 = 0 \\ \partial_t V^4 + \sigma A U^7 + \partial_z \sigma A U^2 + \sigma V^4 + \partial_z \sigma V^1 = 0 \end{array}\right. .$$

• Nous calculons maintenant $\partial_t V^5$:

$$\begin{aligned}
\partial_t V^5 &= \begin{pmatrix} \partial_{xx}(\sigma\partial_t E_x) + \partial_{xy}(\sigma\partial_t E_{yx}) + \partial_{xz}(\sigma\partial_t E_{zx}) \\ 0 \\ 0 \\ \partial_{xx}(\sigma\partial_t H_x) + \partial_{xy}(\sigma\partial_t H_{yx}) + \partial_{xz}(\sigma\partial_t H_{zx}) \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \partial_{xx}(\sigma(\partial_y H_z - \partial_z H_y)) + \partial_{xy}(\sigma(-\partial_x H_z - \sigma E_{yx})) + \partial_{xz}(\sigma(\partial_x H_y - \sigma E_{zx})) \\ 0 \\ 0 \\ \partial_{xx}(\sigma(\partial_z E_y - \partial_y E_z)) + \partial_{xy}(\sigma(\partial_x E_z - \sigma H_{yx})) + \partial_{xz}(\sigma(-\partial_x E_y - \sigma H_{zx})) \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} a_5 \\ 0 \\ 0 \\ b_5 \\ 0 \\ 0 \end{pmatrix}.
\end{aligned}$$

Nous calculons a_5 :

$$\begin{aligned}
a_5 &= \partial_{xx}\sigma.\partial_y H_z + \partial_x\sigma.\partial_{xy} H_z - \partial_{xx}\sigma.\partial_z H_y - \partial_x\sigma.\partial_{xz} H_y - \partial_{xy}\sigma.\partial_x H_z - \partial_y\sigma.\partial_{xx} H_z \\
&\quad -\partial_{xy}\sigma.\sigma E_{yx} - \partial_x\sigma.\partial_y(\sigma E_{yx}) - \partial_y\sigma.\partial_x(\sigma E_{yx}) - \sigma.\partial_{xy}(\sigma E_{yx}) + \partial_{xz}\sigma.\partial_x H_y \\
&\quad +\partial_z\sigma.\partial_{xx} H_y - \partial_{xz}\sigma.\sigma E_{zx} - \partial_x\sigma.\partial_z(\sigma E_{zx}) - \partial_z\sigma.\partial_x(\sigma E_{zx}) - \sigma.\partial_{xz}(\sigma E_{zx}),
\end{aligned}$$

et nous faisons de même pour b_5. Nous obtenons alors :

$$\begin{aligned}
\partial_t V^5 &= C_1U^2 + C_2U^3 + C_3U^4 + C_4U^5 + C_5U^6 + C_6U^7 + C_7V^1 + C_8V^2 \\
&\quad +C_9V^3 + C_{10}V^4 - \sigma \begin{pmatrix} \partial_{xy}(\sigma E_{yx}) + \partial_{xz}(\sigma E_{zx}) \\ 0 \\ 0 \\ \partial_{xy}(\sigma H_{yx}) + \partial_{xz}(\sigma H_{zx}) \\ 0 \\ 0 \end{pmatrix}.
\end{aligned}$$

Avec :

$$C_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & \partial_{xz}\sigma & -\partial_{xy}\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\partial_{xz}\sigma & \partial_{xy}\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, C_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \partial_{xx}\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\partial_{xx}\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$C_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & -\partial_{xx}\sigma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \partial_{xx}\sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, C_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & \partial_z\sigma & -\partial_y\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\partial_z\sigma & \partial_y\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$C_5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \partial_x\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\partial_x\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, C_6 = \begin{pmatrix} 0 & 0 & 0 & 0 & -\partial_x\sigma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \partial_x\sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$C_7 = \begin{pmatrix} 0 & -\partial_{xy}\sigma & -\partial_{xz}\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\partial_{xy}\sigma & -\partial_{xz}\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$C_8 = \begin{pmatrix} 0 & -\partial_y\sigma & -\partial_z\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\partial_y\sigma & -\partial_z\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$C_9 = \begin{pmatrix} 0 & -\partial_x\sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\partial_x\sigma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, C_{10} = \begin{pmatrix} 0 & 0 & -\partial_x\sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\partial_x\sigma \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Nous obtenons alors un système de la forme :

$$\begin{cases} \partial_t V^5 &- C_1U^2 - C_2U^3 - C_3U^4 - C_4U^5 - C_5U^6 - C_6U^7 - C_7V^1 - C_8V^2 \\ &- C_9V^3 - C_{10}V^4 + \sigma V^5 - \sigma\partial_{xx}\sigma TU^1 - 2\sigma\partial_x\sigma TU^2 - \sigma^2 TU^5 = 0 \\ \partial_t V^6 &- D_1U^2 - D_2U^3 - D_3U^4 - D_4U^5 - D_5U^6 - D_6U^7 - D_7V^1 - D_8V^2 \\ &- D_9V^3 - D_{10}V^4 + \sigma V^6 - \sigma\partial_{xy}\sigma TU^1 - \sigma\partial_x\sigma TU^3 - \sigma\partial_y\sigma TU^2 - \sigma^2 TU^6 = 0 \\ \partial_t V^7 &- E_1U^2 - E_2U^3 - E_3U^4 - E_4U^5 - E_5U^6 - E_6U^7 - E_7V^1 - E_8V^2 \\ &- E_9V^3 - E_{10}V^4 + \sigma V^7 - \sigma\partial_{xz}\sigma TU^1 - \sigma\partial_x\sigma TU^4 - \sigma\partial_z\sigma TU^2 - \sigma^2 TU^7 = 0. \end{cases}$$

• La partie principale de ce système est à coefficients constants et son symbole $P(i\xi)$ est :

$$\left(\begin{array}{cccc|cccc|ccc|ccc}
\widetilde{P}(i\xi) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \widetilde{P}(i\xi) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \widetilde{P}(i\xi) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \widetilde{P}(i\xi) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & \widetilde{P}(i\xi) & 0 & 0 & 0 & \widetilde{Q}(i\xi) & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \widetilde{P}(i\xi) & 0 & 0 & 0 & \widetilde{Q}(i\xi) & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \widetilde{P}(i\xi) & 0 & 0 & 0 & \widetilde{Q}(i\xi) & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right).$$

De même que pour la dimension 2, comme $\widetilde{P}(i\xi)$ est le symbole des équations de Maxwell tridimensionnelles, nous savons qu'il existe $\widetilde{S}(\xi)$ telle que $\widetilde{D}(\xi) = \widetilde{S}(\xi)\widetilde{P}(i\xi)\widetilde{S}(\xi)^{-1}$ soit diagonale de valeurs propres imaginaires pures et $\|\widetilde{S}(\xi)\| + \|\widetilde{S}(\xi)^{-1}\|$ soit borné.

Nous considèrons alors la matrice de passage :

$$S(\xi) = \left(\begin{array}{ccc|ccc|c|ccc|ccc}
\ddots & & & & & & & & & & & \\
& \widetilde{S}(\xi) & & & & & & & & & & \\
& & \ddots & & & & & & & & & \\
& & & \ddots & & & & & & & & \\
\hline
& & & \ddots & & & & \ddots & & & & \\
& & & & \widetilde{S}(\xi) & & & & \widetilde{S}(\xi)A & & & \\
& & & & & \ddots & & & & \ddots & & \\
\hline
& & & & & & Id & & & & & \\
\hline
& & & & & & & \ddots & & & & \\
& & & & & & & & Id & & & \\
& & & & & & & & & \ddots & & \\
\hline
& & & & & & & & & & \ddots & \\
& & & & & & & & & & & Id \\
& & & & & & & & & & & \ddots
\end{array}\right).$$

Son inverse est alors :

$$S^{-1}(\xi) = \left(\begin{array}{c|c|c|c|c} \widetilde{S}^{-1}(\xi) & & & & \\ \hline & \widetilde{S}^{-1}(\xi) & -A & & \\ \hline & & Id & & \\ \hline & & & Id & \\ \hline & & & & Id \end{array}\right),$$

et nous avons alors, en remarquant que $-\widetilde{P}(i\xi)A + \widetilde{Q}(i\xi) = 0$:

$$\begin{aligned} & S(\xi)P(i\xi)S(\xi)^{-1} \\ &= \left(\begin{array}{c|c|c|c|c} \widetilde{S}(\xi)\widetilde{P}(i\xi)\widetilde{S}^{-1}(\xi) & & & & \\ \hline & \widetilde{S}(\xi)\widetilde{P}(i\xi)\widetilde{S}^{-1}(\xi) & -\widetilde{S}(\xi)\widetilde{P}(i\xi)M + \widetilde{S}(\xi)\widetilde{Q}(i\xi) & & \\ \hline & & 0 & & \\ \hline & & & 0 & \\ \hline & & & & 0 \end{array}\right) \\ &= \left(\begin{array}{c|c|c|c|c} \widetilde{S}(\xi)\widetilde{P}(i\xi)\widetilde{S}^{-1}(\xi) & & & & \\ \hline & \widetilde{S}(\xi)\widetilde{P}(i\xi)\widetilde{S}^{-1}(\xi) & 0 & & \\ \hline & & 0 & & \\ \hline & & & 0 & \\ \hline & & & & 0 \end{array}\right). \end{aligned}$$

Et nous concluons comme dans le cas de la dimension 2 que le problème en les variables $(U^j, V^j)_{j\in\{1\dots7\}}$ est fortement bien posé ce qui prouve le théorème.

□

Remarque 7.4 *Dans le cas d'une absorption dans les trois directions, il n'y a plus de termes nuls dans V^1 mais des termes analogues à ceux considérés dans V pour le cas de Maxwell en dimension 2. Il en est de même pour V^5, V^6 et V^7. De plus, les six dérivées secondes doivent être étudiées et plus seulement les trois qui font intervenir la variable x.*

7.3 Etude par semi-discrétisation

Le but de cette partie est d'établir une autre preuve des estimations d'énergie en construisant une approximation de la solution continue par une semi-discrétisation.

7.3.1 Discrétisation des équations PML

Notations

Nous considèrons une grille carrée sur $\mathbb{R}^2$ de pas $h > 0$.
Nous définissons les dérivées discrètes et les opérateurs de translation :

Définition 7.1 *Si $\nu = (\nu_1, \nu_2) \in \mathbb{Z}^2$ et $U \in \mathbb{R}^{\mathbb{Z}^2}$ alors :*

$$(D_x^+ U)_\nu = \frac{U_{\nu_1+1,\nu_2} - U_\nu}{h}$$

$$(D_x^- U)_\nu = \frac{U_\nu - U_{\nu_1-1,\nu_2}}{h}$$

$$(D_y^+ U)_\nu = \frac{U_{\nu_1,\nu_2+1} - U_\nu}{h}$$

$$(D_y^- U)_\nu = \frac{U_\nu - U_{\nu_1,\nu_2-1}}{h}$$

$$(T_x U)_\nu = U_{\nu_1+1,\nu_2}$$

$$(T_y U)_\nu = U_{\nu_1,\nu_2+1}.$$

Nous définissons ensuite les espaces suivants :

Définition 7.2 *Nous appelons espace mixte discret sur la grille $G = (h\mathbb{Z})^2$, l'espace $\mathcal{V}^q(G)$, $q \in \mathbb{N}^*$, défini par :*

$$\mathcal{V}^q(G) = (H^q(G))^3 \times H^{q-1}(G),$$

où les espaces de Sobolev discrets sont définis dans la partie 2.1. Cet espace est muni de la norme :

$$\|(U_1, U_2, U_3, U_4)\|^2_{h,\mathcal{V}^q} = \|U_1\|^2_{h,H^q} + \|U_2\|^2_{h,H^q} + \|U_3\|^2_{h,H^q} + \|U_4\|^2_{h,H^{q-1}}.$$

Nous notons $\mathcal{V}(G) = \mathcal{V}^1(G)$.

Remarque 7.5 *Les espaces mixtes correspondent à la discrétisation de l'espace introduit dans [32].*

Equations discrétisées

Nous considèrons l'opérateur $A(\sigma) : L^2(G) \to L^2(G)$ défini pour σ suffisamment régulier, tel que si $u \in L^2(G)$,

$$\mathcal{F}_h(A(\sigma).U) = \hat{\sigma} * \mathcal{F}_h(U).$$

Nous considérons alors le problème discret suivant :

$$\frac{dE_x}{dt} - D_y^+ H_z = F_1, \tag{7.2}$$

$$\frac{dE_y}{dt} + D_x^+ H_z + A(\sigma).E_y = F_2, \tag{7.3}$$

$$\frac{dH_z}{dt} + D_x^- E_y - D_y^- E_x + G = F_3, \tag{7.4}$$

$$\frac{dG}{dt} + A(\sigma).D_x^+ E_y + A(\sigma).G + A(\varphi).E_y = F_4, \tag{7.5}$$

où F_1,F_2, F_3, F_4, σ et φ sont données et suffisament régulières.

7.3.2 Lemmes calculatoires

Lemme 7.1 *Pour tout $\sigma \in L^\infty(\mathbb{R}^2)$, pour tout $U \in L^2(G)$, nous avons :*

$$\|A(\sigma).U\|_h \leq \|\sigma\|_\infty \|U\|_h.$$

PREUVE : D'après le théorème de Parseval cité dans la partie 2.1, nous avons

$$\begin{aligned} \|A(\sigma).U\|_h^2 &= \|\mathcal{F}_h(A(\sigma).U)\|_{L^2} \\ &= \int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2} |\hat{\sigma} * \mathcal{F}_h(U)(\xi)|^2 d\xi \\ &= \int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2} |\widehat{\sigma.SU}(\xi)|^2 d\xi, \end{aligned}$$

où S désigne l'interpolée définie dans la partie 2.1.2. Nous avons alors :

$$\begin{aligned} \|A(\sigma).U\|_h^2 &\leq \|\sigma.SU\|_{L^2}^2 \\ &\leq \|\sigma\|_\infty^2 \|SU\|_{L^2}^2 \\ &\leq \|\sigma\|_\infty^2 \|U\|_h^2. \end{aligned}$$

□

Lemme 7.2 *Pour tout $\sigma \in L^\infty(\mathbb{R}^2)$, pour tout $U \in L^2(G)$,*

$$\|D_x^+ A(\sigma).U\|_h^2 \leq C(\|\partial_x \sigma\|_\infty^2 \|U\|_h^2 + \|\sigma\|_\infty^2 \|D_x^+ U\|_h^2)$$

PREUVE :

$$\begin{aligned} \|D_x^+ A(\sigma).U\|_h^2 &= \|\mathcal{F}_h(D_x^+ A(\sigma).U)\|_{L^2} \\ &= \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} \left|\frac{e^{i\xi_1 h}-1}{h}\right|^2 |\widehat{\sigma.SU}(\xi)|^2 d\xi. \end{aligned}$$

Or $|e^{i\xi_1 h} - 1| = 2|\sin \xi_1 h/2| \leq |\xi_1| h$, ainsi :

$$\begin{aligned} \|D_x^+ A(\sigma).U\|_h^2 &\leq \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} |\xi_1|^2 |\widehat{\sigma.SU}(\xi)|^2 d\xi \\ &\leq \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} |\widehat{\partial_x(\sigma.SU)}(\xi)|^2 d\xi \\ &\leq \|\partial_x(\sigma.SU)\|_{L^2}^2 \\ &\leq 2(\|(\partial_x\sigma).SU)\|_{L^2}^2 + \|\sigma\partial_x(SU)\|_{L^2}^2) \\ &\leq 2(\|\partial_x\sigma\|_\infty^2 \|U\|_h^2 + \|\sigma\|_\infty^2 \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} |\xi_1|^2 |\mathcal{F}_h(U)(\xi)|^2 d\xi). \end{aligned}$$

Or, comme nous l'avons déjà utilisé dans le chapitre 2.1, si $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\frac{|\sin x|}{|x|} \geq \frac{2}{\pi}$,

$$\begin{aligned} \|D_x^+ U\|_h^2 &= \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} \left(\frac{\sin(h\xi_1/2)}{h/2}\right)^2 |\mathcal{F}_h(U)(\xi)|^2 d\xi \\ &\geq \frac{4}{\pi^2} \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} |\xi_1|^2 |\mathcal{F}_h(U)(\xi)|^2 d\xi. \end{aligned}$$

Donc

$$\|D_x^+ A(\sigma).U\|_h^2 \leq C(\|\partial_x\sigma\|_\infty^2 \|U\|_h^2 + \|\sigma\|_\infty^2 \|D_x^+ U\|_h^2). \tag{7.6}$$

□

Montrons également l'analogue de la formule de Leibnitz suivant :

Lemme 7.3

$$(D_x^+)^n A(\sigma).U = \sum_{k=0}^n C_n^k T_x^{n-k} A(d_{x-}^{n-k}\sigma).(D_x^+)^k U,$$

où :

$$d_{x-}\sigma(x,y) = \frac{\sigma(x,y) - \sigma(x-h,y)}{h}.$$

PREUVE : Montrons cette formule par récurrence sur n.
• Pour $n = 0$ la formule est vraie.
• Pour $n = 1$, nous avons :

$$\begin{aligned}
\mathcal{F}_h(D_x^+ A(\sigma).U)(\xi) &= \left(\frac{e^{i\xi_1 h}-1}{h}\right)\mathcal{F}_h(A(\sigma).U)(\xi)\\
&= \left(\frac{e^{i\xi_1 h}-1}{h}\right)\hat{\sigma} * \mathcal{F}_h(U)(\xi)\\
&= \int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2}\left(\frac{e^{i\xi_1 h}-e^{i\tau_1 h}+e^{i\tau_1 h}-1}{h}\right)\hat{\sigma}(\xi-\tau)\mathcal{F}_h(U)(\tau)d\tau\\
&= e^{i\xi_1 h}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2}\left(1-\frac{e^{-i(\xi_1-\tau_1)h}}{h}\right)\hat{\sigma}(\xi-\tau)\mathcal{F}_h(U)(\tau)d\tau\\
&\quad+\int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2}\left(\frac{e^{i\tau_1 h}-1}{h}\right)\hat{\sigma}(\xi-\tau)\mathcal{F}_h(U)(\tau)d\tau\\
&= e^{i\xi_1 h}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2}\widehat{d_{x-}\sigma}(\xi-\tau)\mathcal{F}_h(U)(\tau)d\tau\\
&\quad+\int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2}\hat{\sigma}(\xi-\tau)\mathcal{F}_h(D_x^+U)(\tau)d\tau\\
&= e^{i\xi_1 h}\widehat{d_{x-}\sigma} * \mathcal{F}_h(U)(\xi)+\hat{\sigma} * \mathcal{F}_h(D_x^+U)(\xi)\\
&= e^{i\xi_1 h}\mathcal{F}_h(A(d_{x-}\sigma).U)(\xi)+\mathcal{F}_h(A(\sigma).D_x^+U)(\xi)\\
&= \mathcal{F}_h\big(T_x A(d_{x-}\sigma).U+A(\sigma).D_x^+U\big)(\xi).
\end{aligned}$$

D'où :

$$D_x^+ A(\sigma).U = T_x A(d_{x-}\sigma).U + A(\sigma).D_x^+ U,$$

ce qui est la formule voulue.
• Soit $n \in \mathbb{N}^*$, supposons la formule vraie au rang n. En appliquant l'opérateur D_x^+ ainsi que la formule au rang 1, nous obtenons :

$$\begin{aligned}
(D_x^+)^{n+1}A(\sigma).U &= \sum_{k=0}^{n} C_n^k T_x^{n-k} D_x^+(A(d_{x-}^{n-k}\sigma).(D_x^+)^k U)\\
&= \sum_{k=0}^{n} C_n^k T_x^{n-k+1} A(d_{x-}^{n-k+1}\sigma).(D_x^+)^k U)+\sum_{k=0}^{n} C_n^k T_x^{n-k} A(d_{x-}^{n-k}\sigma).(D_x^+)^{k+1}U)\\
&= \sum_{k=0}^{n+1} C_{n+1}^k T_x^{n+1-k} D_x^+(A(d_{x-}^{n+1-k}\sigma).(D_x^+)^k U).
\end{aligned}$$

D'où le lemme. □

7.3.3 Existence d'une solution

D'après le théorème de Cauchy-Lipschitz linéaire appliqué sur $L^2(G)^4$, il suffit de montrer que l'application linéaire :

$$L : \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} \mapsto \begin{pmatrix} -D_y^+ U_3 \\ D_x^+ U_3 + A(\sigma).U_2 \\ D_x^- U_2 - D_y^- U_1 + U_4 \\ A(\sigma).D_x^+ U_2 + A(\sigma).U_4 + A(\varphi).U_2 \end{pmatrix},$$

est continue sur $L^2(G)^4$.

Or nous avons :

- $\|D_y^+ U_3\|_h \leq \frac{2}{h}\|U_3\|_h$,
- d'après le lemme 7.1, $\|D_x^+ U_3 + A(\sigma).U_2\|_h \leq \frac{2}{h}\|U_3\|_h + \|\sigma\|_{L^\infty}\|U_2\|_h$,
- $\|D_x^- U_2 - D_y^- U_1 + U_4\|_h \leq \frac{2}{h}(\|U_2\|_h + \|U_1\|_h) + \|U_4\|_h$,
- $\|A(\sigma).D_x^+ U_2 + A(\sigma).U_4 + A(\varphi).U_2\|_h \leq \frac{2}{h}\|\sigma\|_{L^\infty}\|U_2\|_h + \|\sigma\|_{L^\infty}\|U_4\|_h + \|\varphi\|_{L^2}\|U_2\|_h$, en utilisant le lemme 7.1.

Nous avons donc :

$$\|LU\|_h \leq C(h)\|U\|_h.$$

Donc, d'après le théorème de Cauchy-Lipschitz :

Proposition 7.2 *Le système formé des équations (7.2), (7.3), (7.4), (7.5), admet une unique solution :*

$$(E_x, E_y, H_z, G) \in \mathcal{C}^1([0, +\infty[, L^2(G)^4).$$

7.3.4 Estimations d'énergie

Première estimation d'énergie

Nous utilisons la même méthode que celle de [32] pour le problème continu. Montrons le résultat suivant :

Théorème 7.3 *Soient σ, $\varphi \in \mathcal{C}_c^\infty$, alors il existe $C_1 = C_1(\|\sigma\|_{1,\infty}, \|\varphi\|_\infty)$, C_2 et C_3 constantes tels que :*
quelles que soient F_1, F_2, F_3, $F_4 \in \mathcal{C}([0, +\infty[, \mathcal{V}(G))$, $(E_x^0, E_y^0, H_z^0, G^0) \in \mathcal{V}(G)$,
si (E_x, E_y, H_z, G) est la solution du système (7.2) à (7.5) de condition initiale $(E_x^0, E_y^0, H_z^0, G^0)$, alors pour tout $t \in \mathbb{R}$:

$$\|(E_x, E_y, H_z, G)\|_{h,\mathcal{V}}^2 \leq C_2\|(E_x^0, E_y^0, H_z^0, G^0)\|_{h,\mathcal{V}}^2 e^{C_1 t} + C_3 \int_0^t \|(F_1, F_2, F_3, F_4)\|_{h,\mathcal{V}}^2(s) e^{C_1(t-s)} ds.$$

PREUVE : Nous commençons par faire apparaître une énergie :

– $E_x\times$ (7.2) donne :

$$\frac{1}{2}\frac{d}{dt}E_x^2 - E_x.D_y^+H_z = E_x.F_1, \tag{7.7}$$

– $E_y\times$ (7.3) donne :

$$\frac{1}{2}\frac{d}{dt}E_y^2 + E_y.D_x^+H_z + E_y.A(\sigma).E_y = E_y.F_2, \tag{7.8}$$

– $H_z\times$ (7.4) donne :

$$\frac{1}{2}\frac{d}{dt}H_z^2 + H_z.D_x^-E_y - H_z.D_y^-E_x + H_z.G = H_z.F_3, \tag{7.9}$$

– $\lambda G\times$ (7.5) donne :

$$\frac{\lambda}{2}\frac{d}{dt}G^2 + \lambda G.A(\sigma).D_x^+E_y + \lambda G.A(\sigma).G + \lambda G.A(\varphi).E_y = \lambda G.F_4, \tag{7.10}$$

– $D_x^+E_x \times D_x^+$ (7.2) donne :

$$\frac{1}{2}\frac{d}{dt}(D_x^+E_x)^2 - D_x^+E_x.D_x^+D_y^+H_z = D_x^+E_x.D_x^+F_1 \tag{7.11}$$

– $D_x^+E_y \times D_x^+$ (7.3) donne :

$$\frac{1}{2}\frac{d}{dt}(D_x^+E_y)^2 + D_x^+E_y.D_x^+D_x^+H_z + D_x^+E_y.D_x^+A(\sigma).E_y = D_x^+E_y.D_x^+F_2, \tag{7.12}$$

– $D_x^-H_z \times D_x^-$ (7.4) donne :

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}(D_x^-H_z)^2 + D_x^-H_z.D_x^-D_x^-E_y - D_x^-H_z.D_x^-D_y^-E_x + D_x^-H_z.D_x^-G \\ = D_x^-H_z.D_x^-F_3,\end{aligned} \tag{7.13}$$

– $D_y^+E_x \times D_y^+$ (7.2) donne :

$$\frac{1}{2}\frac{d}{dt}(D_y^+E_x)^2 - D_y^+E_x.D_y^+D_y^+H_z = D_y^+E_x.D_y^+F_1, \tag{7.14}$$

– $D_y^+E_y \times D_y^+$ (7.3) donne :

$$\frac{1}{2}\frac{d}{dt}(D_y^+E_y)^2 + D_y^+E_y.D_x^+D_y^+H_z + D_y^+E_y.D_y^+A(\sigma).E_y = D_y^+E_y.D_y^+F_2, \tag{7.15}$$

– $D_y^-H_z \times D_y^-$ (7.4) donne :

$$\begin{aligned}\tfrac{1}{2}\tfrac{d}{dt}(D_y^-H_z)^2 + D_y^-H_z.D_x^-D_y^-E_y - D_y^-H_z.D_y^-D_y^-E_x + D_y^-H_z.D_y^-G \\ = D_y^-H_z.D_y^-F_3,\end{aligned} \tag{7.16}$$

Soit $\mathcal{E}(t) = \frac{1}{2}(\|E_x\|^2_{h,H^1} + \|E_y\|^2_{h,H^1} + \|H_z\|^2_{h,H^1} + \lambda\|G\|^2_h)$.
En sommant les équations (7.7) à (7.16) sur tout l'espace et en utilisant

$$< A, D_+B >_h = - < D_-A, B >,$$

nous obtenons :

$$\begin{aligned}
&\tfrac{d}{dt}\mathcal{E}(t) + < E_y, A(\sigma).E_y >_h + < H_z, G >_h + \lambda < G, A(\sigma).D_x^+ E_y >_h \\
&+\lambda < G, A(\sigma).G >_h + \lambda < G, A(\varphi).E_y >_h + < D_x^+ E_y, D_x^+ A(\sigma).E_y >_h \\
&+ < D_x^- H_z, D_x^- G >_h + < D_y^+ E_y, D_y^+ A(\sigma).E_y >_h + < D_y^- H_z, D_y^- G >_h \\
&= < E_x, F_1 >_h + < E_y, F_2 >_h + < H_z, F_3 >_h + \lambda < G, F_4 >_h \\
&+ < D_x^+ E_x, D_x^+ F_1 >_h + < D_x^+ E_y, D_x^+ F_2 >_h + < D_x^- H_z, D_x^- F_3 >_h \\
&+ < D_y^+ E_x, D_y^+ F_1 >_h + < D_y^+ E_y, D_y^+ F_2 >_h + < D_y^- H_z, D_y^- F_3 >_h .
\end{aligned}$$

Or, nous avons :

$$< D_x^- H_z, D_x^- G >_h + < D_y^- H_z, D_y^- G >_h = - < (D_x^+ D_x^- + D_y^+ D_y^-) H_z, G >_h,$$

mais, d'après (7.2) :

$$D_y^+ H_z = \frac{d}{dt} E_x - F_1 \quad \text{donc} \quad D_y^+ D_y^- H_z = \frac{d}{dt} D_y^- E_x - D_y^- F_1.$$

Et, d'après (7.3) :

$$D_x^+ H_z = F_2 - \frac{d}{dt} E_y - A(\sigma).E_y,$$

donc

$$D_x^+ D_x^- H_z = D_x^- F_2 - \frac{d}{dt} D_x^- E_y - D_x^- A(\sigma).E_y.$$

Ainsi :

$$\begin{aligned}
&< D_x^- H_z, D_x^- G >_h + < D_y^- H_z, D_y^- G >_h \\
&\quad = < \tfrac{d}{dt}(D_x^- E_y - D_y^- E_x) + D_x^- A(\sigma).E_y, G >_h \\
&\quad\quad + < D_y^- F_1 - D_x^- F_2, G >_h .
\end{aligned}$$

Et comme, d'après (7.5) nous avons :

$$\begin{aligned}
&< \tfrac{d}{dt} G, D_x^- E_y - D_y^- E_x >_h \\
&\quad + < A(\sigma).D_x^+ E_y + A(\sigma).G + A(\varphi).E_y, D_x^- E_y - D_y^- E_x >_h \\
&\quad = < F_4, D_x^- E_y - D_y^- E_x >_h,
\end{aligned}$$

nous avons :

$$\begin{aligned}
&\tfrac{d}{dt}\mathcal{E}(t) + < E_y, A(\sigma).E_y >_h + < H_z, G >_h + \lambda < G, A(\sigma).D_x^+ E_y >_h \\
&+\lambda < G, A(\sigma).G >_h + \lambda < G, A(\varphi).E_y >_h + < D_x^+ E_y, D_x^+ A(\sigma).E_y >_h \\
&+ < D_y^+ E_y, D_y^+ A(\sigma).E_y >_h + \tfrac{d}{dt} < G, D_x^- E_y - D_y^- E_x >_h + < D_x^- A(\sigma).E_y, G >_h \\
&+ < A(\sigma).D_x^+ E_y + A(\sigma).G + A(\varphi).E_y, D_x^- E_y - D_y^- E_x >_h \\
&= < E_x, F_1 >_h + < E_y, F_2 >_h + < H_z, F_3 >_h + \lambda < G, F_4 >_h + < D_x^+ E_x, D_x^+ F_1 >_h \\
&+ < D_x^+ E_y, D_x^+ F_2 >_h + < D_x^- H_z, D_x^- F_3 >_h + < D_y^+ E_x, D_y^+ F_1 >_h \\
&+ < D_y^+ E_y, D_y^+ F_2 >_h + < D_y^- H_z, D_y^- F_3 >_h + < F_4, D_x^- E_y - D_y^- E_x >_h \\
&+ < D_x^- F_2 - D_y^- F_1, G >_h .
\end{aligned}$$

Posons :

$$U(t) = \|E_x, E_y, H_z, G\|^2_{h,\mathcal{V}} \quad \text{et} \quad F(t) = \|F_1, F_2, F_3, F_4\|^2_{h,\mathcal{V}}.$$

Alors en utilisant les lemmes (7.1) et (7.2) ainsi que :

$$| < A, B >_h | \leq \frac{1}{2}(\|A\|^2_h + \|B\|^2_h),$$

nous obtenons une inégalité de la forme :

$$\mathcal{E}(t) \leq K_1(1+\lambda)U(t) + K_2(1+\lambda)F(t) - \frac{d}{dt} < G, D_x^- E_y - D_y^- E_x >_h,$$

où $K_1 = K_1(\|\sigma\|_{1,\infty}, \|\phi\|_\infty)$ et K_2 constante.
Nous intégrons alors en temps :

$$\begin{aligned} \mathcal{E}(t) + < G, D_x^- E_y - D_y^- E_x >_h (t) \quad &\leq \quad \mathcal{E}(0) + < G, D_x^- E_y - D_y^- E_x >_h (0) \\ &\quad + K_1(1+\lambda) \int_0^t U(s) ds \\ &\quad + K_2(1+\lambda) \int_0^t F(s) ds. \end{aligned} \tag{7.17}$$

De plus, nous avons :

$$\begin{aligned} \mathcal{E}(t) + < G, D_x^- E_y - D_y^- E_x >_h (t) \quad &\leq \quad \frac{1}{2}\|E_x, E_y, H_z\|^2_{h,H^1} + \frac{\lambda}{2}\|G\|^2_h \\ &\quad + \frac{1}{2}(\|G\|^2_h + \|D_x^- E_y\|^2_h) + \frac{1}{2}(\|G\|^2_h + \|D_y^- E_x\|^2_h) \\ &\leq \quad \|E_x, E_y, H_z\|^2_{h,1} + \frac{\lambda+2}{2}\|G\|^2_h \\ &\leq \quad \frac{\lambda+2}{2} U(t). \end{aligned} \tag{7.18}$$

Et nous avons aussi, en prenant $\mu > 0$:

$$\begin{aligned} \mathcal{E}(t) + < G, D_x^- E_y - D_y^- E_x >_h (t) \quad &\geq \quad \frac{1}{2}\|E_x, E_y, H_z\|^2_{h,1} + \frac{\lambda}{2}\|G\|^2_h \\ &\quad - \frac{1}{2}\left(\frac{1}{\mu}\|G\|^2_h + \mu\|D_x^- E_y\|^2_h\right) \\ &\quad - \frac{1}{2}\left(\frac{1}{\mu}\|G\|^2_h + \mu\|D_y^- E_x\|^2_h\right) \\ &\geq \quad \frac{1-\mu}{2}\|E_x, E_y, H_z\|^2_{h,1} + (\frac{\lambda}{2} - \frac{1}{\mu})\|G\|^2_h. \end{aligned} \tag{7.19}$$

Nous choisissons :

$$\mu = \frac{1 - \lambda + \sqrt{(\lambda - 1)^2 + 8}}{2}.$$

Nous avons alors :

$$\frac{1-\mu}{2} = \frac{\lambda}{2} - \frac{1}{\mu},$$

et si nous choisissons $\lambda > 2$, on a alors $0 < \mu < 1$. Posons :

$$K_3 = \frac{\lambda + 2}{2} \quad \text{et} \quad K_4 = \frac{1-\mu}{2} > 0.$$

Nous avons alors d'après (7.18) et (7.19) :

$$K_4 U(t) \leq \mathcal{E}(t) + < G, D_x^- E_y - D_y^- E_x >_h (t) \leq K_3 U(t).$$

Et, en réinjectant dans (7.17) nous obtenons :

$$K_4 U(t) \leq K_3 U(0) + K_1' \int_0^t U(s) ds + K_2' \int_0^t F(s) ds.$$

Et par le lemme de Gronwall :

$$\begin{aligned} U(t) &\leq \frac{1}{K_4}(K_3 U(0) + K_2' \int_0^t F(s) ds) \\ &\quad + \int_0^t \frac{K_1'}{K_4^2} \left(K_3 U(0) + K_2' \int_0^s F(\tau) d\tau \right) \exp\left(\frac{K_1'}{K_4}(t-s) \right) ds \\ &\leq \frac{K_3}{K_4} U(0) \exp\left(\frac{K_1'}{K_4} t \right) + \frac{K_2'}{K_4} \int_0^t F(s) \exp\left(\frac{K_1'}{K_4}(t-s) \right) ds. \end{aligned}$$

D'où l'estimation voulue.

□

Estimation d'énergie des dérivées

Dans cette partie nous considérerons que $F_1 = F_2 = F_3 = F_4 = 0$ et que $\varphi = 0$. Nous avons alors le résultat suivant :

Théorème 7.4 *Soient $\sigma \in \mathcal{C}_c^\infty$, $p \in \mathbb{N}$, $q \in \mathbb{N}^*$, alors il existe $A_{p,q} = A_{p,q}(\|\sigma\|_{q,\infty})$ et $B_{p,q} = B_{p,q}(\|\sigma\|_{1,\infty})$ tels que :*
quels que soient $(E_x^0, E_y^0, H_z^0, G^0) \in \mathcal{V}^{q+p}(G)$,
si (E_x, E_y, H_z, G) est la solution du système (7.2) à (7.5) avec $F_1 = F_2 = F_3 = F_4 = 0$ et $\varphi = 0$ de condition initiale $(E_x^0, E_y^0, H_z^0, G^0)$, alors pour tout $t \in \mathbb{R}$:

$$\| \frac{d^p}{dt^p}(E_x, E_y, H_z, G) \|_{h,\mathcal{V}^q}^2 \leq A_{p,q} \| (E_x^0, E_y^0, H_z^0, G^0) \|_{h,\mathcal{V}^{p+q}}^2 \exp(B_{p,q} t).$$

PREUVE :

- Nous commençons par montrer le résultat pour $p = 0$ en effectuant une récurrence sur q.
  - Pour $q = 1$, il suffit d'appliquer le résultat précédent et de prendre : $A_{0,1} = C_2$ et $B_{0,1} = C_1$.
  - Soit $q \in \mathbb{N}^*$, supposons l'estimation vraie au rang q.
    Soient q_1, q_2 tels que $q_1 + q_2 = q$. Nous appliquons $(D_x^+)^{q_1}(D_y^+)^{q_2}$ au système formé des équations (7.2) à (7.5) ainsi que le lemme (7.3).Nous avons alors les résultats suivants :

    L'équation (7.2) devient :

$$\frac{d}{dt}(D_x^+)^{q_1}(D_y^+)^{q_2}E_x - D_y^+(D_x^+)^{q_1}(D_y^+)^{q_2}H_z = 0.$$

    L'équation (7.3) devient :

$$\begin{aligned}&\tfrac{d}{dt}(D_x^+)^{q_1}(D_y^+)^{q_2}E_y + D_x^+(D_x^+)^{q_1}(D_y^+)^{q_2}H_z + A(\sigma).(D_x^+)^{q_1}(D_y^+)^{q_2}E_y\\ &\quad = -\textstyle\sum_{0\le k\le q_1, 0\le l\le q_2,(k,l)\neq(q_1,q_2)} C_{q_1}^k C_{q_2}^l T_x^{q_1-k}T_y^{q_2-l}A(d_{x-}^{q_1-k}d_{y-}^{q_2-l}\sigma)(D_x^+)^k(D_y^+)^l E_y.\end{aligned}$$

    L'équation (7.4) devient :

$$\frac{d}{dt}(D_x^+)^{q_1}(D_y^+)^{q_2}H_z + D_x^-(D_x^+)^{q_1}(D_y^+)^{q_2}E_y - D_y^-(D_x^+)^{q_1}(D_y^+)^{q_2}E_x + (D_x^+)^{q_1}(D_y^+)^{q_2}G = 0.$$

    L'équation (7.5) devient :

$$\begin{aligned}&\tfrac{d}{dt}(D_x^+)^{q_1}(D_y^+)^{q_2}G + A(\sigma).D_x^+(D_x^+)^{q_1}(D_y^+)^{q_2}E_y\\ &\quad + q_1 T_x A(d_{x-}\sigma)(D_x^+)^{q_1}(D_y^+)^{q_2}E_y + A(\sigma).(D_x^+)^{q_1}(D_y^+)^{q_2}G\\ &\qquad = -\textstyle\sum_{(k,l)\in I}\left(C_{q_1}^k C_{q_2}^l T_x^{q_1-k}T_y^{q_2-l}A(d_{x-}^{q_1-k}d_{y-}^{q_2-l}\sigma)(D_x^+)^{k+1}(D_y^+)^l E_y\right)\\ &\qquad\quad -\textstyle\sum_{(k,l)\in I}\left(C_{q_1}^k C_{q_2}^l T_x^{q_1-k}T_y^{q_2-l}A(d_{x-}^{q_1-k}d_{y-}^{q_2-l}\sigma)(D_x^+)^k(D_y^+)^l G\right),\end{aligned}$$

    où $I = \{(k,l),\ 0 \le k \le q_1, 0 \le l \le q_2, (k,l) \neq (q_1,q_2), (q_1-1,q_2)\}$.
    Nous utilisons alors l'équation initiale de variables $(D_x^+)^{q_1}(D_y^+)^{q_2}(E_x, E_y, H_z, G)$ avec $\varphi = q_1 d_{x-}\sigma$ (l'opérateur de translation T_x ne modifie rien) et avec le second membre apparaissant dans les équations précédentes.
    Le théorème 7.3 donne alors :

$$\begin{aligned}\|(D_x^+)^{q_1}(D_y^+)^{q_2}(E_x,E_y,H_z,G)\|_{h,\mathcal{V}}^2 \;&\le\; C_2\|(D_x^+)^{q_1}(D_y^+)^{q_2}(E_x^0,E_y^0,H_z^0,G^0)\|_{h,\mathcal{V}}^2 e^{C_1 t}\\ &\quad + C_3\int_0^t \|(F_1,F_2,F_3,F_4)\|_{h,\mathcal{V}}^2(s)e^{C_1(t-s)}ds,\end{aligned}$$

où $C_1 = C_1(\|\sigma\|_{1,\infty}, \|\varphi\|_\infty)$. Or :

$$d_{x-}\sigma(x,y) = \frac{1}{h}\int_{x-h}^{x} \partial_x\sigma(\xi,y)d\xi.$$

Donc :

$$\|\varphi\|_\infty \leq C\|\sigma\|_{1,\infty}.$$

D'après la forme de C_1, nous avons : $C_1 = C_1(\|\sigma\|_{1,\infty})$. De plus, d'après le lemme (7.1), nous avons :

$$\begin{aligned}\|A(d_{x-}^{q_1-k}d_{y-}^{q_2-l}\sigma)(D_x^+)^k(D_y^+)^l U\|_h &\leq d_{x-}^{q_1-k}d_{y-}^{q_2-l}\sigma\|_\infty\|(D_x^+)^k(D_y^+)^l U\|_h \\ &\leq \|\sigma\|_{q_1-k+q_2-l,\infty}\|U\|_{h,k+l}.\end{aligned}$$

Nous avons alors :

$$\begin{aligned}\|F_2\|_h &\leq K_1\|\sigma\|_{q,\infty}\|E_y\|_{h,q-1}, \\ \|D_x^+F_2\|_h &\leq K_2\|\sigma\|_{q+1,\infty}\|E_y\|_{h,q}, \\ \|D_y^+F_2\|_h &\leq K_3\|\sigma\|_{q+1,\infty}\|E_y\|_{h,q},\end{aligned}$$

donc

$$\|F_2\|_{h,H^1} \leq K_4\|\sigma\|_{q+1,\infty}\|E_y\|_{h,q},$$

et

$$\|F_4\|_h \leq K_5\|\sigma\|_{q,\infty}(\|E_y\|_{h,H^q} + \|G\|_{h,q-1}),$$

d'où, comme $F_1 = F_3 = 0$:

$$\|(F_1,F_2,F_3,F_4)\|_{h,\mathcal{V}} \leq K\|\sigma\|_{q+1,\infty}\|(E_x,E_y,H_z,G)\|_{h,\mathcal{V}^q},$$

et par hypothèse de récurrence :

$$\|(F_1,F_2,F_3,F_4)\|_{h,\mathcal{V}}^2 \leq K^2\|\sigma\|_{q+1,\infty}^2 A_{0,q}\|(E_x^0,E_y^0,H_z^0,G^0)\|_{h,\mathcal{V}^q}^2\exp(B_{0,q}t).$$

Ainsi, nous obtenons :

$$\begin{aligned}\|(D_x^+)^{q_1}(D_y^+)^{q_2}(E_x,E_y,H_z,G)\|_{h,\mathcal{V}}^2 &\leq C_2\|(D_x^+)^{q_1}(D_y^+)^{q_2}(E_x^0,E_y^0,H_z^0,G^0)\|_{h,\mathcal{V}}^2 e^{C_1 t} \\ &\quad + C_3\int_0^t \big(K^2 A_{0,q}\|\sigma\|_{q+1,\infty}^2 \\ &\quad \times \|(E_x^0,E_y^0,H_z^0,G^0)\|_{h,\mathcal{V}^q}^2 e^{B_{0,q}s + C_1(t-s)}\big)\,ds \\ &\leq A'_{0,q+1}\|(E_x^0,E_y^0,H_z^0,G^0)\|_{h,\mathcal{V}^{q+1}}^2\exp(B'_{0,q}t).\end{aligned}$$

Ceci étant vrai pour tout q_1, q_2 tels que $q_1 + q_2 = q$, nous avons bien le résultat au rang $q+1$.
Nous avons alors le théorème pour $p = 0$.

– Soit $p \in \mathbb{N}$, nous supposons le résultat vrai au rang p pour tous les q. Comme σ ne dépend pas du temps, appliquer l'opérateur $\frac{d^{p+1}}{dt^{p+1}}$ aux équations ne pose pas de problème. Il reste à étudier les conditions initiales.
D'après l'équation (7.2), nous avons :

$$\frac{d^{p+1}}{dt^{p+1}} E_x = D_y^+ \frac{d^p}{dt^p} H_z,$$

donc :

$$\| \frac{d^{p+1}}{dt^{p+1}} E_x \|_{h,q}(t) \leq \| \frac{d^p}{dt^p} H_z \|_{h,q+1}(t).$$

Ainsi, en faisant de même avec les autres équations :

$$\| \frac{d^{p+1}}{dt^{p+1}} (E_x, E_y, H_z, G) \|_{h,\mathcal{V}^q}(0) \leq C(\|\sigma\|_{1,\infty}) \| \frac{d^p}{dt^p} (E_x, E_y, H_z, G) \|_{h,\mathcal{V}^{q+1}}(0),$$

et par hypothèse de récurrence :

$$\| \frac{d^{p+1}}{dt^{p+1}} (E_x, E_y, H_z, G) \|_{h,\mathcal{V}^q}(0) \leq C(\|\sigma\|_{1,\infty}) A_{p,q} \| (E_x^0, E_y^0, H_z^0, G^0) \|^2_{h,\mathcal{V}^{p+q}} e^{B_{p,q} t}.$$

d'où la conclusion.

□

7.3.5 Solution régulière du problème continu

Interpolation de la solution discrète

Dans cette partie, nous considérons des conditions initiales du problème continu $E_x^0, E_y^0, H_z^0, G^0 \in \mathcal{C}_c^\infty$. Les données initiales du problème discret notées aussi E_x^0, E_y^0, H_z^0, G^0 sont les évaluées des données continues aux points de la grille.

Nous notons :
$W_1^h = SE_x$, $W_2^h = SE_y$, $W_3^h = SH_z$ et $W_4^h = SG$.
Nous avons alors :

Proposition 7.3

1. $\lim_{h\to 0}(\partial_t W_1^h - \partial_y W_3^h) = 0$,
2. $\lim_{h\to 0} \partial_t(W_2^h + \partial_x W_3^h + \sigma W_4^h) = 0$,
3. $\lim_{h\to 0} \partial_t(W_3^h + \partial_x W_2^h - \partial_y W_1^h + W_4^h) = 0$,
4. $\lim_{h\to 0} \partial_t(W_4^h + \sigma \partial_x W_2^h + \sigma W_4^h) = 0$.

Toutes ces limites étant uniformes en espace.

Preuve :

1.

$$(\partial_t W_1^h - \partial_y W_3^h))(x,y) = \frac{1}{2\pi}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}(\partial_t \mathcal{F}_h(E_x)(\xi) - i\xi_2 \mathcal{F}_h(H_z)(\xi)d\xi,$$

or :

$$0 = \mathcal{F}_h(\frac{dE_x}{dt} - D_y^+ H_z)(\xi) = \partial_t \mathcal{F}_h(E_x)(\xi) - \frac{e^{ih\xi_2}-1}{h}\mathcal{F}_h(H_z)(\xi),$$

donc :

$$(\partial_t W_1^h - \partial_y W_3^h))(x,y) = \frac{1}{2\pi}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\left(\frac{e^{ih\xi_2}-1}{h} - i\xi_2\right)\mathcal{F}_h(H_z)(\xi)d\xi,$$

or :

$$\left|\frac{e^{ih\xi_2}-1}{h} - i\xi_2\right| \leq \xi_2^2 h,$$

donc :

$$\begin{aligned}
|(\partial_t W_1^h - \partial_y W_3^h))(x,y)| &\leq \frac{h}{2\pi}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} \xi_2^2 |\mathcal{F}_h(H_z)(\xi)|d\xi \\
&\leq \frac{h}{2\pi}\left(\frac{2\pi}{h}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \frac{\xi_2^4}{(1+\xi_2)^3}d\xi_2\right)^{1/2} \\
&\quad \times \left(\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} (1+\xi_2^3)^2 |\mathcal{F}_h(H_z)(\xi)|^2 d\xi\right)^{1/2} \\
&\leq C\sqrt{h}\|H_z\|_{h,3} \\
&\leq C'(t)\sqrt{h}\|(E_x^0, E_y^0, H_z^0, G^0)\|_{h,\mathcal{V}^3}.
\end{aligned}$$

Nous utilisons alors un cas particulier du lemme 2.5 :

Lemme 7.4 *Si $U \in \mathcal{C}_c^\infty$ et si EU est l'évaluée de U aux points de la grille alors :*

$$\|EU\|_{h,q} \leq C\|U\|_{H^{q+2}}$$

Donc :

$$|(\partial_t W_1^h - \partial_y W_3^h))(x,y)| \leq K(t)\sqrt{h},$$

où la constante K dépend des données initiales du problème continu mais pas de h.

Nous avons alors la conclusion.

2. De même que pour le point précédent, nous avons :

$$\begin{array}{l}(\partial_t W_2^h + \partial_x W_3^h + \sigma W_2^h))(x,y) = \\ \quad \frac{1}{2\pi}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\left(-\xi_1 - \frac{e^{ih\xi_1}-1}{h}\right)\mathcal{F}_h(H_z)(\xi)d\xi \\ \quad +\frac{1}{2\pi}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\left[\sigma(x,y)\mathcal{F}_h(E_y)(\xi) - \hat{\sigma} * \mathcal{F}_h(E_y)(\xi)\right]d\xi.\end{array}$$

Le premier terme se traite comme précédemment, il reste à étudier le second terme :

$$I = \left|\frac{1}{2\pi}\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\left[\sigma(x,y)\mathcal{F}_h(E_y)(\xi) - \hat{\sigma} * \mathcal{F}_h(E_y)(\xi)\right]d\xi\right|.$$

Nous avons alors :

$$\begin{aligned} I &= \frac{1}{2\pi}\left|\sigma(x,y)\int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\widehat{SE_y}(\xi)d\xi - \int_{[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\widehat{\sigma.SE_y}d\xi\right| \\ &= \frac{1}{2\pi}\left|\sigma(x,y) * 2\pi SE_y(x,y) - \int_{\mathbb{R}^2} e^{i(x\xi_1+y\xi_2)}\widehat{\sigma.SE_y}(\xi)d\xi\right. \\ &\quad \left. + \int_{\mathbb{R}^2\setminus[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\widehat{\sigma.SE_y}(\xi)d\xi\right| \\ &= \frac{1}{2\pi}\left|\int_{\mathbb{R}^2\setminus[-\frac{\pi}{h},\frac{\pi}{h}]} e^{i(x\xi_1+y\xi_2)}\widehat{\sigma.SE_y}(\xi)d\xi\right|. \end{aligned}$$

Or :

$$\begin{aligned} |\widehat{\sigma.SE_y}(\xi)| &= |\hat{\sigma} * \widehat{SE_y}(\xi)| \\ &= \left|\int_{\mathbb{R}^2} \hat{\sigma}(\tau)\widehat{SE_y}(\xi-\tau)d\tau\right| \\ &\leq \left(\int_{\mathbb{R}^2}(1+|\tau|^2)^3|\hat{\sigma}(\tau)|^2\frac{1}{(1+|\tau|^2)^3(1+|\xi-\tau|^2)^3}d\tau\right)^{1/2} \\ &\quad \times\left(\int_{\mathbb{R}^2}(1+|\xi-\tau|^2)^3|\widehat{SE_y}(\xi-\tau)|^2d\tau\right)^{1/2}. \end{aligned}$$

Mais, nous avons :
– Si $|\tau| > \frac{1}{2}|\xi|$:

$$\frac{1}{(1+|\tau|^2)(1+|\xi-\tau|^2)} \leq \frac{1}{(1+|\tau|^2)} \leq \frac{1}{(1+\frac{|\xi|^2}{4})}.$$

– Si $|\tau| \leq \frac{1}{2}|\xi|$:

$$\frac{1}{(1+|\tau|^2)(1+|\xi-\tau|^2)} \leq \frac{1}{(1+|\xi-\tau|^2)}.$$

Or $|\xi-\tau| \geq |\xi| - |\tau| \geq \frac{1}{2}|\xi|$ donc :

$$\frac{1}{(1+|\tau|^2)(1+|\xi-\tau|^2)} \leq \frac{1}{(1+\frac{|\xi|^2}{4})}.$$

Ainsi :

$$\begin{aligned} |\widehat{\sigma.SE_y}(\xi)| &\leq \frac{1}{(1+\frac{|\xi|^2}{4})}\|\sigma\|_{H^3}\|\|SE_y\|_{H^3} \\ &\leq \frac{1}{(1+\frac{|\xi|^2}{4})}\|\sigma\|_{H^3}\|\|E_y\|_{h,H^3} \\ &\leq C(t)\frac{1}{(1+\frac{|\xi|^2}{4})}, \end{aligned}$$

où C dépend des données initiales et de σ mais pas de h. Nous avons donc une majoration indépendante de h par une fonction intégrable sur $\mathbb{R}^2$.
Comme $\lim_{h\to 0} \chi_{\mathbb{R}^2\setminus[-\frac{\pi}{h},\frac{\pi}{h}]} = 0$ presque partout, nous avons par le théorème de convergence dominée de Lebesgue :

$$\lim_{h\to 0}\int_{\mathbb{R}^2\setminus[-\frac{\pi}{h},\frac{\pi}{h}]} |\widehat{\sigma.SE_y}(\xi)|d\xi = 0.$$

Et comme

$$I \leq \frac{1}{2\pi}\int_{\mathbb{R}^2\setminus[-\frac{\pi}{h},\frac{\pi}{h}]} |\widehat{\sigma.SE_y}(\xi)|d\xi,$$

nous avons la concusion.

3. et 4. Les deux limites suivantes se montrent avec des inégalités de la même forme que celles des cas déjà traités.

□

Existence d'une solution régulière

Nous allons maintenant faire converger les W^h de manière à obtenir une solution au problème continu.

Nous commençons par étudier la convergence sur les compacts :

Proposition 7.4 *Pour tout $p, q_1, q_2 \in \mathbb{N}$,pour tout $T > 0$, pour tout $j \in \{1,2,3,4\}$ et pour tout compact K de $\mathbb{R}^2$, il existe $W_j \in \mathcal{C}^\infty([0,T] \times K)$ et une suite (h_k) tendant vers 0 tels que :*

$$\lim_{k \to +\infty} \sup_{[0,T]\times K} \left| \frac{\partial^{p+q_1+q_2}}{\partial t^p \partial x^{q_1} \partial y^{q_2}} W^j_{h_k} - \frac{\partial^{p+q_1+q_2}}{\partial t^p \partial x^{q_1} \partial y^{q_2}} W^j \right| = 0.$$

De plus, nous avons, sur $[0,T] \times K$:

$$\left\{ \begin{array}{rcl} \partial_t W_1 - \partial_y W_3 & = & 0 \\ \partial_t W_2 + \partial_x W_3 + \sigma W_2 & = & 0 \\ \partial_t W_3 + \partial_x W_2 - \partial_y W_1 + W_4 & = & 0 \\ \partial_t W_4 + \sigma \partial_x W_2 + \sigma W_4 & = & 0. \end{array} \right.$$

Preuve : D'après le théorème d'Ascoli, il suffit de majorer :

$$\left| \frac{\partial^{p+q_1+q_2}}{\partial t^p \partial x^{q_1} \partial y^{q_2}} W^j_h \right|,$$

pour avoir l'existence de la limite. Nous le faisons pour $j = 1$ les autres cas étant semblables :

$$\begin{aligned} \left| \frac{\partial^{p+q_1+q_2}}{\partial t^p \partial x^{q_1} \partial y^{q_2}} W^1_h \right| &= \left| \frac{1}{2\pi} \int_{\mathbb{R}^2} e^{i(x\xi_1 + y\xi_2)} \widehat{\frac{\partial^{p+q_1+q_2}}{\partial t^p \partial x^{q_1} \partial y^{q_2}} W^1_h}(t,\xi) d\xi \right| \\ &= \frac{1}{2\pi} \left| \int_{\mathbb{R}^2} e^{i(x\xi_1 + y\xi_2)} (i\xi_1)^{q_1} (i\xi_2)^{q_2} \frac{\partial^p}{\partial t^p} \widehat{W^h_1}(t,\xi) d\xi \right| \\ &= \frac{1}{2\pi} \left| \int_{\mathbb{R}^2} e^{i(x\xi_1 + y\xi_2)} (i\xi_1)^{q_1} (i\xi_2)^{q_2} \frac{\partial^p}{\partial t^p} \mathcal{F}_h(E_x)(t,\xi) d\xi \right| \\ &\leq \frac{1}{2\pi} \left(\int_{\mathbb{R}^2} \frac{1}{(1+|\xi|^2)^3} d\xi \right)^{1/2} \\ &\quad \times \left(\int_{\mathbb{R}^2} (1+|\xi|^2)^3 |\xi_1|^{2q_1} |\xi_2|^{2q_2} \left| \frac{\partial^p}{\partial t^p} \mathcal{F}_h(E_x)(t,\xi) \right|^2 d\xi \right)^{1/2} \\ &\leq C \| \frac{\partial^p}{\partial t^p} E_x \| h, H^{q_1+q_2+3} \\ &\leq K(T). \end{aligned}$$

D'où l'existence d'une limite extraite. La proposition précédente montre que les équations sont vérifiées. □

Maintenant, nous allons étendre cette solution à l'espace entier.

Théorème 7.5 *Si $E_x^0, E_y^0, H_z^0, G^0, \sigma \in \mathcal{C}_c^\infty$ alors le système*

$$\left\{\begin{array}{rcl} \partial_t E_x - \partial_y H_z &=& 0 \\ \partial_t E_y + \partial_x H_z + \sigma E_y &=& 0 \\ \partial_t H_z + \partial_x E_y - \partial_y E_x + G &=& 0 \\ \partial_t G + \sigma \partial_x E_y + \sigma G &=& 0, \end{array}\right.$$

avec $E_x(0,.) = E_x^0$, $E_y(0,.) = E_y^0$, $H_z(0,.) = H_z^0$ et $G(0,.) = G^0$,
admet une solution appartenant à $\mathcal{C}^\infty([0,T]\times\mathbb{R}^2)$. De plus cette solution vérifie l'estimation d'énergie de Métral et Vacus [32] :

$$\|(E_x, E_y, H_z, G)\|_{\mathcal{V}}^2 \leq C_2 \|(E_x^0, E_y^0, H_z^0, G^0)\|_{h,\mathcal{V}}^2 e^{C_1 t},$$

où $C_1 = C_1(|\sigma\|_{1,\infty})$ et $\|(E_x, E_y, H_z, G)\|_{\mathcal{V}}^2 = \|(E_x, E_y, H_z)\|_{H^1}^2 + \|G\|_{L^2}^2$.

Preuve :

- Soit $K_n = [-n,n]^2$ compact de $\mathbb{R}^2$. Nous construisons par récurrence une solution sur $\mathbb{R}^2$ telle que la solution sur K_{n+1} prolonge la solution sur K_n.
  - Pour $n = 1$, nous construisons la solution comme extraction $(h_{\varphi_1(j)})$ des (W_h).
  - Soit $n \in \mathbb{N}^*$. Supposons que nous ayons construit une solution sur K_n par extraction $(h_{\varphi_n(j)})$.
    Nous considèrons les suites $(W_{h_{\varphi_n(j)}})$ qui vérifient les mêmes hypothèses que (W_h). Nous pouvons donc, comme précédemment, leur appliquer le théorème d'Ascoli : nous trouvons une suite $(W_{h_{\varphi_n(\varphi(j))}})$ qui converge vers une solution sur K_{n+1}. Comme cette suite est extraite de la précédente, cette solution et la précédente coïncident sur K_n.

  Nous avons ainsi construit une solution sur $\mathbb{R}^2$ que nous appelons $W = (W_1, W_2, W_3, W_4)$.
- La solution W est $\mathcal{C}^\infty$ sur K_n° pour tout $n \in \mathbb{N}^*$ donc $W \in \mathcal{C}^\infty(\mathbb{R}^2)$ et est solution forte du problème.
- Montrons que $W \in \mathcal{V}$. Nous allons montrer que $W_1 \in H^1(\mathbb{R}^2)$, les autres composantes se traitant de la même manière.
  Sur $[0,T]\times K_n$, nous avons :
  - $\lim_{k\to+\infty} \left| W^1_{h_{\varphi_n(k)}} - W^1 \right| = 0$,
  - $\lim_{k\to+\infty} \left| \frac{\partial}{\partial x} W^1_{h_{\varphi_n(k)}} - \frac{\partial}{\partial x} W^1 \right| = 0$,
  - $\lim_{k\to+\infty} \left| \frac{\partial}{\partial y} W^1_{h_{\varphi_n(k)}} - \frac{\partial}{\partial y} W^1 \right| = 0$,
  - $\left| W^1_{h_{\varphi_n(k)}} - W^1 \right| \leq C_1$,
  - $\left| \frac{\partial}{\partial x} W^1_{h_{\varphi_n(k)}} - \frac{\partial}{\partial x} W^1 \right| \leq C_2$,

– $\left|\frac{\partial}{\partial y}W^1_{h_{\varphi_n(k)}} - \frac{\partial}{\partial y}W^1\right| \leq C_3$.

Or, comme les constantes sont intégrables sur K_n, nous avons, par le théorème de convergence dominée de Lebesgue :

$$\lim_{k\to+\infty} \|W^1_{h_{\varphi_n(k)}} - W^1\|_{H^1(K_n)} = 0.$$

Ainsi pour k assez grand :

$$\|W^1\|_{H^1(K_n)} \leq 2\|W^1_{h_{\varphi_n(k)}}\|_{H^1(K_n)} \leq 2\|W^1_{h_{\varphi_n(k)}}\|_{H^1}.$$

Or, d'après le lemme (7.4) et le théorème (7.3) :

$$\begin{aligned}\|W^1_h\|^2_{H^1} &= \int_{[-\frac{\pi}{h},\frac{\pi}{h}]^2} (1+|\xi|^2)|\mathcal{F}_h(E_x)(\xi)|^2 d\xi \\ &\leq C\|E_x\|^2_{h,H^1} \\ &\leq K(T)\|(E^0_x, E^0_y, H^0_z, G^0)\|^2_{V^3}.\end{aligned}$$

Nous avons donc montré que $\|W^1\|_{H^1(K_n)}$ est majoré par une constante indépendante de n donc :

$$W^1 \in H^1(\mathbb{R}^2).$$

– Nous avons maintenant toutes les hypothèses de régularité justifiant les calculs de [32], l'estimation d'énergie est donc vérifiée.

□

7.3.6 Régularisation du probleme continu

Nous allons maintenant prendre des hypothèses de régularité plus faibles sur les conditions initiales et l' absorption. Nous notons $\mathcal{V}$ l'espace $\mathcal{V} = (H^1(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)$ et $\|.\|_{\mathcal{V}}$ la norme sur cet espace. Nous avons alors le résultat suivant :

Théorème 7.6 *Si $(E^0_x, E^0_y, H^0_z, G^0) \in (H^1(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)$ et $\sigma \in W^{1,\infty}(\mathbb{R}^2)$ alors le système*

$$\left\{\begin{array}{rcl} \partial_t E_x - \partial_y H_z &=& 0 \\ \partial_t E_y + \partial_x H_z + \sigma E_y &=& 0 \\ \partial_t H_z + \partial_x E_y - \partial_y E_x + G &=& 0 \\ \partial_t G + \sigma\partial_x E_y + \sigma G &=& 0, \end{array}\right.$$

avec $E_x(0,.) = E^0_x$, $E_y(0,.) = E^0_y$, $H_z(0,.) = H^0_z$ et $G(0,.) = G^0$
admet une solution appartenant à $L^2(0,T,\mathcal{V}) \cap \mathcal{C}^0(0,T,(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2))$. De plus cette solution vérifie l'estimation d'énergie de Métral et Vacus :

$$\|(E_x, E_y, H_z, G)\|^2_{\mathcal{V}} \leq C_2\|(E^0_x, E^0_y, H^0_z, G^0)\|^2_{\mathcal{V}} e^{C_1 t},$$

où $C_1 = C_1(|\sigma\|_{1,\infty})$ et $\|(E_x, E_y, H_z, G)\|^2_{\mathcal{V}} = \|(E_x, E_y, H_z)\|^2_{H^1} + \|G\|^2_{L^2}$

PREUVE :

– Nous régularisons les données initiales et l'absorption :
Soient $E^0_{x,n}, E^0_{y,n}, H^0_{z,n}, G^0_n, \sigma_n \in \mathcal{C}^\infty_c$ telles que :

$$\lim_{n\to+\infty} \|E^0_{x,n} - E^0_x\|_{H^1} = 0,$$

$$\lim_{n\to+\infty} \|E^0_{y,n} - E^0_y\|_{H^1} = 0,$$

$$\lim_{n\to+\infty} \|H^0_{z,n} - H^0_z\|_{H^1} = 0,$$

$$\lim_{n\to+\infty} \|G^0_{,n} - G^0\|_{L^2} = 0.$$

$$\lim_{n\to+\infty} \|\sigma_n - \sigma\|_{W^{1,\infty}} = 0$$

Soit $W_n = (E_{x,n}, E_{y,n}, H_{z,n}, G_n)$ la solution de :

$$\left\{\begin{array}{rcl} \partial_t E_{x,n} - \partial_y H_{z,n} &=& 0 \\ \partial_t E_{y,n} + \partial_x H_{z,n} + \sigma_n E_{y,n} &=& 0 \\ \partial_t H_{z,n} + \partial_x E_{y,n} - \partial_y E_{x,n} + G_n &=& 0 \\ \partial_t G_n + \sigma_n \partial_x E_{y,n} + \sigma_n G_n &=& 0, \end{array}\right.$$

avec les conditions initiales : $E_{x,n}(0,.) = E^0_{x,n}$, $E_{y,n}(0,.) = E^0_{y,n}$, $H_{z,n}(0,.) = H^0_{z,n}$ et $G_n(0,.) = G^0_n$.

– Montrons que la suite (W_n) est de Cauchy dans $\mathcal{C}^0(0,T,\mathcal{V})$.
Soient $p, q \in \mathbb{N}$, alors $W_p - W_q$ vérifie :

$$\left\{\begin{array}{l} \partial_t(E_{x,p} - E_{x,q}) - \partial_y(H_{z,p} - H_{z,q}) = 0 \\ \partial_t(E_{y,p} - E_{y,q}) + \partial_x(H_{z,p} - H_{z,q}) + \sigma_p(E_{y,p} - E_{y,q}) = (\sigma_q - \sigma_p)E_{y,q} \\ \partial_t(H_{z,p} - H_{z,q}) + \partial_x(E_{y,p} - E_{y,q}) - \partial_y(E_{x,p} - E_{x,q}) + (G_p - G_q) = 0 \\ \partial_t(G_p - G_q) + \sigma_p\partial_x(E_{y,p} - E_{y,q}) + \sigma_p(G_p - G_q) \\ \qquad = (\sigma_q - \sigma_p)\partial_x E_{y,q} + (\sigma_q - \sigma_p)G_q. \end{array}\right.$$

Nous pouvons alors appliquer l'estimation d'énergie obtenue dans le cas régulier (le second membre se traite comme dans le cas discret). Nous obtenons :

$$\begin{aligned}
\|W_p - W_q\|_{\mathcal{V}}^2 &\leq C_2\|W_p^0 - W_q^0\|_{\mathcal{V}}^2 e^{C_1 t} \\
&\quad + C_3 \int_0^t \Big(\|(\sigma_q - \sigma_p)E_{y,q}\|_{H^1}^2 \\
&\quad + \|(\sigma_q - \sigma_p)\partial_x E_{y,q}\|_{L^2}^2 + \|(\sigma_q - \sigma_p)G_q\|_{L^2}^2 \Big)(s) e^{C_1(t-s)} ds \\
&\leq C_2\|W_p^0 - W_q^0\|_{\mathcal{V}}^2 e^{C_1 t} \\
&\quad + C_3\|(\sigma_q - \sigma_p)\|_{1,\infty} \int_0^t \left(\|W_q\|_{\mathcal{V}}^2 \right)(s) e^{C_1(t-s)ds} \\
&\leq C_2\|W_p^0 - W_q^0\|_{\mathcal{V}}^2 e^{C_1 t} \\
&\quad + C_3\|(\sigma_q - \sigma_p)\|_{1,\infty} \int_0^t \left(C_2' \|W_q^0\|_{\mathcal{V}}^2 e^{C_1' s} \right) e^{C_1(t-s)ds},
\end{aligned}$$

où $C_1 = C_1(\|(\sigma_q - \sigma_p)\|_{1,\infty})$ et $C_1' = C_1'(\|\sigma_q\|_{1,\infty})$. Or nous avons :
- $\lim_{p,q\to+\infty} \|W_p^0 - W_q^0\|_{\mathcal{V}} = 0$,
- $\lim_{p,q\to+\infty} \|(\sigma_q - \sigma_p)\|_{1,\infty} = 0$,
- $\|W_q^0\|_{\mathcal{V}} \leq 2\|W^0\|_{\mathcal{V}}$ pour q assez grand,
- C_1 et C_1' sont à dépendance polynomiale donc ils sont bornés en p et q.

Donc

$$\lim_{p,q\to+\infty} \sup_{t\in[0,T]} \|W_p - W_q\|_{\mathcal{V}} = 0.$$

Ainsi (W_n) est de Cauchy dans $\mathcal{C}^0(0,T,\mathcal{V})$.

- Nous utilisons maintenant un cas particulier des espaces définis dans [30] :

$$\mathcal{W}(0,T) = \{u, u \in L^2(0,T,\mathcal{V}), \partial_t u \in L^2(0,T,L^2(\mathbb{R}^2)^4)\},$$

muni de la norme

$$\|u\|_{\mathcal{W}(0,T)}^2 = \|u\|_{L^2(0,T,\mathcal{V})}^2 + \|\partial_t u\|_{L^2(0,T,L^2(\mathbb{R}^2)^4)}^2.$$

Montrons que (W_n) est de Cauchy dans $\mathcal{W}(0,T)$.
- Comme (W_n) est de Cauchy dans $\mathcal{C}^0(0,T,\mathcal{V})$ alors (W_n) est de Cauchy dans $L^2(0,T,\mathcal{V})$.
- Nous utilisons les équations pour montrer que $(\partial_t W_n)$ est de Cauchy dans $L^2(0,T,L^2(\mathbb{R}^2)^4)$. Par exemple, pour la composante $E_{y,n}$, nous avons :

$$\begin{aligned}
\|\partial_t(E_{y,p} - E_{y,q})\|_{L^2} &= \|\partial_x(H_{z,p} - H_{z,q}) + \sigma_p(E_{y,p} - E_{y,q}) - (\sigma_q - \sigma_p)E_{y,q}\|_{L^2} \\
&\leq (\|\sigma_p\|_{W^{1,\infty}} + \|\sigma_p - \sigma_q\|_{W^{1,\infty}})\|W_p - W_q\|_{\mathcal{V}} \\
&\leq 6\|\sigma_p\|_{W^{1,\infty}}\|W_p - W_q\|_{\mathcal{V}},
\end{aligned}$$

pour p, q assez grands. Comme (W_n) est de Cauchy dans $L^2(0,T,\mathcal{V})$ alors $(\partial_t W_n)$ est de Cauchy dans $L^2(0,T,L^2(\mathbb{R}^2)^4)$.

Donc (W_n) converge vers W dans $\mathcal{W}(0,T)$.

– Montrons que W vérifie les équations voulues. Nous allons montrer uniquement la deuxième équation les autres étant semblables :
Soit $n \in \mathbb{N}$,

$$\begin{aligned}
&\|\partial_t E_y + \partial_x H_z + \sigma E_y\|_{L^2(0,T,L^2(\mathbb{R}^2))} \\
&\quad = \|\partial_t(E_y - E_{y,n}) + \partial_x(H_z - H_{z,n}) + \sigma(E_y - E_{y,n}) \\
&\qquad +(\sigma - \sigma_n)E_{y,n}\|_{L^2(0,T,L^2(\mathbb{R}^2))} \\
&\quad \leq (1 + \|\sigma\|_{W^{1,\infty}} + \|\sigma - \sigma_n\|_{W^{1,\infty}}\|W - W_n\|_{\mathcal{W}(0,T)}.
\end{aligned}$$

Donc en faisant tendre n vers $+\infty$, nous avons :

$$\|\partial_t E_y + \partial_x H_z + \sigma E_y\|_{L^2(0,T,L^2(\mathbb{R}^2))} = 0.$$

D'où l'équation voulue.

– Il reste à étudier les conditions initiales. D'après [30], nous savons que :

$$W \in \mathcal{C}^0(0,T,(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)),$$

et que l'application :

$$\begin{array}{rcl}
\mathcal{W}(0,T) & \mapsto & \mathcal{C}^0(0,T,(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)), \\
u & \rightarrow & u
\end{array}$$

est continue. Donc :

$$\|W - W^n\|_{\mathcal{C}^0(0,T,(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2))} \leq C\|W - W_n\|_{\mathcal{W}(0,T)}.$$

Ainsi

$$\begin{array}{rcl}
\|W(0,.) - W^0\|_{(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)} & \leq & \|W(0,.) - W_n^0\|_{(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)} \\
& & +\|W_n - W^0\|_{(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2)} \\
& \leq & \|W - W^n\|_{\mathcal{C}^0(0,T,(H^{1/2}(\mathbb{R}^2))^3 \times L^2(\mathbb{R}^2))} \\
& & +\|W_n - W^0\|_{\mathcal{V}}.
\end{array}$$

Comme les deux termes du second membre tendent vers 0, nous avons :

$$W(0,.) = W^0$$

Nous avons donc prouvé le théorème. □

Chapitre 8

Schémas de Yee

Les résultats montrés dans la partie précédente impliquent que les équations PML proposées par Bérenger pour les équations de Maxwell TE sont faiblement bien posées de défaut 1. Nous allons maintenant nous intéresser à la discrétisation de ces équations et au calcul du défaut *via* des estimations d'énergie.

Les équations de Maxwell sont discrétisées classiquement par un schéma de Yee. Nous rappelons dans la première partie de ce chapitre que ce schéma est stable. Le schéma de Yee peut aussi être utilisé pour discrétiser les équations PML de Maxwell TE. Nous cherchons alors des estimations d'énergie pour ce schéma qui nous permettront de voir que le schéma est faiblement stable de défaut 1.

Des estimations d'énergie ont été prouvées dans [10] mais elles sont valables dans le cas de la formulation de Zhao-Cangellaris qui est fortement bien posée.

Dans cette partie, G désignera la grille $G = \Delta x\mathbb{Z} \times \Delta y\mathbb{Z}$.

8.1 Schéma de Yee pour les équations de Maxwell

Nous allons étudier dans cette partie la stabilité du schéma de Yee pour les équations de Maxwell :

$$\left\{ \begin{array}{l} \partial_t E_x - \partial_y H_z = 0 \\ \partial_t E_y + \partial_x H_z = 0 \\ \partial_t H_z + \partial_x E_y - \partial_y E_x = 0 \end{array} \right.$$

Le schéma de Yee [41], étudié dans [39], est le schéma d'ordre 2 suivant :

$$\left\{ \begin{array}{l} \frac{(E_x)_{i,j}^{n+1} - (E_x)_{i,j}^{n}}{\Delta t} - \frac{(H_z)_{i,j+1/2}^{n+1/2} - (H_z)_{i,j-1/2}^{n+1/2}}{\Delta y} = 0 \\ \frac{(E_y)_{i,j}^{n+1} - (E_y)_{i,j}^{n}}{\Delta t} + \frac{(H_z)_{i+1/2,j}^{n+1/2} - (H_z)_{i-1/2,j}^{n+1/2}}{\Delta x} = 0 \\ \frac{(H_z)_{i,j}^{n+1/2} - (H_z)_{i,j}^{n-1/2}}{\Delta t} + \frac{(E_y)_{i+1/2,j}^{n} - (E_y)_{i-1/2,j}^{n}}{\Delta x} - \frac{(E_x)_{i,j+1/2}^{n} - (E_x)_{i,j-1/2}^{n}}{\Delta y} = 0. \end{array} \right. \tag{8.1}$$

Les conditions initiales sont E_x^0, E_y et $H_z^{-1/2}$.
Nous posons :

$$\gamma_x = \frac{\Delta t}{\Delta x} \text{ et } \gamma_y = \frac{\Delta t}{\Delta y}.$$

Nous avons alors la stabilité du schéma de Yee sous condition CFL :

Théorème 8.1 *Sous la condition CFL $\gamma_x^2 + \gamma_y^2 \leq 1$, le schéma de Yee (8.1) est stable.*

PREUVE : Posons :

$$k_1 = \frac{2}{\Delta x} \sin\left(\frac{\xi_1 \Delta x}{2}\right) \text{ et } k_2 = \frac{2}{\Delta y} \sin\left(\frac{\xi_2 \Delta y}{2}\right).$$

La matrice d'amplification du schéma de Yee par rapport à la variable $(\widehat{E_x^n}, \widehat{E_y^n}, \widehat{H_z^{n-1/2}})$ est :

$$\widehat{Q} = R_0^{-1} S_0,$$

où nous avons posé :

$$R_0 = \begin{pmatrix} 1 & 0 & -ik_2\Delta t \\ 0 & 1 & ik_1\Delta t \\ 0 & 0 & 1 \end{pmatrix} \text{ et } S_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ ik_2\Delta t & -ik_1\Delta t & 1 \end{pmatrix}.$$

Nous avons alors :

$$\widehat{Q} = \begin{pmatrix} 1 - k_2^2\Delta t^2 & k_1k_2\Delta t^2 & ik_2\Delta t \\ k_1k_2\Delta t^2 & 1 - k_1^2\Delta t^2 & -ik_1\Delta t \\ ik_2\Delta t & -ik_1\Delta t & 1 \end{pmatrix}.$$

Soit P la matrice :

$$P = \begin{pmatrix} ik_1\Delta t & ik_2\Delta t(1+\lambda_+) & ik_2\Delta t(1+\lambda_-) \\ ik_2\Delta t & -ik_1\Delta t(1+\lambda_+) & -ik_1\Delta t(1+\lambda_-) \\ 0 & \lambda_+ & \lambda_- \end{pmatrix},$$

avec :

$$\lambda_\pm = \frac{-|k|^2\Delta t^2 \pm \sqrt{|k|^4\Delta t^4 - 4|k|^2\Delta t^2}}{2}.$$

Nous avons alors :

$$\widehat{Q} = P \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1+\lambda_+ & 0 \\ 0 & 0 & 1+\lambda_- \end{pmatrix} P^{-1}.$$

Nous remarquons que, si $\|k\|^2\Delta t^2 \leq 4$ alors $|1+\lambda_\pm| = 1$. Or $\|k\|^2\Delta t^2 \leq 4 \Leftrightarrow \gamma_x^2 + \gamma_y^2 \leq 1$ donc, sous condition CFL, les valeurs propres de la matrice d'amplification $\widehat{Q}$ sont de module 1. Pour prouver la stabilité du schéma, il suffit donc de montrer que $\|P\|$ et $\|P^{-1}\|$ sont bornées.

Pour cela, nous commençons par remarquer que :

$$ {}^tPP = \begin{pmatrix} -\Delta t^2\|k\|^2 & 0 & 0 \\ 0 & \lambda_+ - (1+\lambda_+)^2\Delta t^2\|k\|^2 & 0 \\ 0 & 0 & \lambda_- - (1+\lambda_-)^2\Delta t^2\|k\|^2 \end{pmatrix}. $$

Posons $A^2 = -\Delta t^2\|k\|^2$, $B^2 = \lambda_+^2 - (1+\lambda_+)^2\Delta t^2\|k\|^2$ et $C^2 = \lambda_-^2 - (1+\lambda_-)^2\Delta t^2\|k\|^2$. Soit :

$$ R = \begin{pmatrix} \frac{ik_1\Delta t}{A} & \frac{ik_2\Delta t(1+\lambda_+)}{B} & \frac{ik_2\Delta t(1+\lambda_-)}{C} \\ \frac{ik_2\Delta t}{A} & \frac{-ik_1\Delta t(1+\lambda_+)}{B} & \frac{-ik_1\Delta t(1+\lambda_-)}{C} \\ 0 & \frac{\lambda_+}{B} & \frac{\lambda_-}{C} \end{pmatrix}. $$

Alors $R^{-1} = {}^t R$. Comme changer P en R, ne change pas la diagonalisation de $\widehat{Q}$, il suffit donc d'évaluer $\|R\|$. Nous avons :

$$ \begin{aligned} |B|^2 &= |\lambda_+^2 - (1+\lambda_+)^2\Delta t^2\|k\|^2| \\ &= |\lambda_+^2 + (1+\lambda_+)\lambda_+^2| \text{ car } \lambda_+^2 + \|k\|^2\Delta t^2\lambda_+ + \|k\|^2\Delta t^2 = 0 \\ &= |\lambda_+|^2|\lambda_+ + 2| \\ &= \|k\|^2\Delta t^2|\lambda_+ + 2| \\ &= \|k\|^2\Delta t^2\sqrt{4 - \|k\|^2\Delta t^2} \\ &\geq 2\sqrt{1-\gamma_x^2-\gamma_y^2}\|k\|^2\Delta t^2 \\ &\geq K\|k\|^2\Delta t^2. \end{aligned} $$

D'où $|B| \geq \|k\|\Delta t$ et de même, $|C| \geq \|k\|\Delta t$. Et on a alors $\|R\| \leq K$ ce qui prouve la stabilité du schéma de Yee.

□

8.2 Schéma de Yee pour les équations de Maxwell PML

Le schéma de Yee pour les équations PML de variables $(E_x, E_y, H_z = H_{zx} + H_{zy}), G = \sigma H_{zx}$ avec une absorption constante dans la direction des x ($\sigma_x = \sigma$) et nulle dans la direction des y ($\sigma_y = 0$) est le suivant :

$$\left\{\begin{array}{l}
\frac{(E_x)_{i,j}^{n+1}-(E_x)_{i,j}^{n}}{\Delta t}-\frac{(H_z)_{i,j+1/2}^{n+1/2}-(H_z)_{i,j-1/2}^{n+1/2}}{\Delta y}=0\\
\frac{(E_y)_{i,j}^{n+1}-(E_y)_{i,j}^{n}}{\Delta t}+\frac{(H_z)_{i+1/2,j}^{n+1/2}-(H_z)_{i-1/2,j}^{n+1/2}}{\Delta x}+\sigma\frac{(E_y)_{i,j}^{n+1}+(E_y)_{i,j}^{n}}{2}=0\\
\frac{(H_z)_{i,j}^{n+1/2}-(H_z)_{i,j}^{n-1/2}}{\Delta t}+\frac{(E_y)_{i+1/2,j}^{n}-(E_y)_{i-1/2,j}^{n}}{\Delta x}-\frac{(E_x)_{i,j+1/2}^{n}-(E_x)_{i,j-1/2}^{n}}{\Delta y}+\frac{(G)_{i,j}^{n+1/2}+(G)_{i,j}^{n-1/2}}{2}=0\\
\frac{(G)_{i,j}^{n+1/2}-(G)_{i,j}^{n-1/2}}{\Delta t}+\sigma\frac{(E_y)_{i+1/2,j}^{n}-(E_y)_{i-1/2,j}^{n}}{\Delta x}+\sigma\frac{(G)_{i,j}^{n+1/2}+(G)_{i,j}^{n-1/2}}{2}=0.
\end{array}\right. \tag{8.2}$$

Nous allons montrer une estimation d'énergie discrète analogue à celle de la proposition 7.1. Pour cela, nous avons besoin de savoir comment se comporte un schéma stable sous l'effet d'une perturbation d'ordre 0. Nous utiliserons donc le résultat suivant qui est la version discrète du fait qu'un problème fortement bien posé l'est toujours lorsqu'on lui rajoute une perturbation d'ordre 0.

Proposition 8.1 *Soit $\widehat{Q_0}$ tel que $\|\widehat{Q_0}^n\| \leq K_S e^{\alpha_S t_n}$ et $\|\widetilde{Q}\| \leq M$. Alors, si $\widehat{Q} = \widehat{Q_0} + \Delta t\widetilde{Q}$, on a :*

$$\widehat{Q}^n \leq K_S e^{(\alpha_S + \Delta t K_S M)t_n}.$$

PREUVE : Soit $\widehat{V^n}$ la suite définie par : $\widehat{V^{n+1}} = \widehat{Q}\widehat{V^n}$, alors :

$$\widehat{V^{n+1}} = \widehat{Q_0}\widehat{V^n} + \Delta t\widetilde{Q}\widehat{V^n}.$$

Donc :

$$\widehat{V^n} = \widehat{Q_0}^n\widehat{V^0} + \Delta t\sum_{j=0}^{n-1}\widehat{Q_0}^{n-1-j}\widetilde{Q}\widehat{V^j}.$$

Définissons la suite (W_k^n) par :

$$\left\{\begin{array}{rcl} W_0^n & = & \widehat{Q_0}^n\widehat{V^0} \\ W_{k+1}^n & = & \Delta t\sum_{j=0}^{n-1}\widehat{Q_0}^{n-1-j}\widetilde{Q}W_k^j. \end{array}\right.$$

Alors, si la série existe, nous avons : $\widehat{V^n} = \sum_{k=0}^{+\infty} W_k^n$.

Montrons, par récurrence sur n, que :

$$\|W_k^n\| \leq K_s^{k+1}\frac{(\Delta t M)^k t_n^k}{k!}e^{\alpha_S t_n}\|\widehat{V^0}\|.$$

- Pour $n = 0$, nous avons :

$$W_0^n = \widehat{V^0} \text{ et } \forall k \geq 1,\ W_k^n = 0.$$

Le résultat est donc vrai.

- Soit $n \in \mathbb{N}^*$, supposons le résultat vrai pour tout $j \leq n-1$. Montrons le résultat au rang n par récurrence sur k.
  - Pour $k = 0$, nous avons, par hypothèse :

$$\|W_0^n\| \leq \|\widehat{Q_0}^n\| \|\widehat{V^0}\| \leq K_S e^{\alpha_S t_n} \|\widehat{V^0}\|$$

  - Soit $k \in \mathbb{N}$, supposons le résultat vrai au rang k.

$$\begin{aligned}
\|W_{k+1}^n\| &\leq \Delta t \sum_{j=0}^{n-1} K_S e^{\alpha_S t_{n-1-j}} M \|W_k^j\| \\
&\leq \Delta t K_S M \sum_{j=0}^{n-1} e^{\alpha_S t_{n-1-j}} \left(K_s^{k+1} \frac{(\Delta t M)^k t_j^k}{k!} e^{\alpha_S t_j} \right) \|\widehat{V^0}\| \\
&\leq K_S^{k+2} \frac{(\Delta t M)^{k+1}}{k!} e^{\alpha_S t_{n-1}} \left(\sum_{j=0}^{n-1} t_j^k \right) \|\widehat{V^0}\|.
\end{aligned}$$

Or :

$$\sum_{j=0}^{n-1} t_j^k \leq \int_0^{t_n} s^k ds \leq \frac{t_n^{k+1}}{k+1}.$$

Donc :

$$\|W_{k+1}^n\| \leq K_S^{k+2} \frac{(\Delta t M)^{k+1}}{(k+1)!} e^{\alpha_S t_{n-1}} t_n^{k+1} \|\widehat{V^0}\|.$$

D'où le résultat au rang n.

Nous avons alors :

$$\begin{aligned}
\sum_{k=0}^{+\infty} \|W_k^n\| &\leq K_S e^{\alpha_S t_n} \sum_{k=0}^{+\infty} \frac{(K_S \Delta t M t_n)^k}{k!} \|\widehat{V^0}\| \\
&\leq K_S e^{(\alpha_S + \Delta t K_S M) t_n} \|\widehat{V^0}\|.
\end{aligned}$$

D'où l'existence de la série et l'estimation :

$$\|\widehat{Q}^n \widehat{V^0}\| = \|\widehat{V^n}\| \leq K_S e^{(\alpha_S + \Delta t K_S M) t_n} \|\widehat{V^0}\|.$$

Et ainsi :

$$\|\widehat{Q}^n\| \leq K_S e^{(\alpha_S + \Delta t K_S M) t_n}.$$

□

Nous montrons maintenant l'estimation discrète suivante :

Théorème 8.2 *Sous la condition CFL :*

$$\left(\frac{\Delta t}{\Delta x}\right)^2 + \left(\frac{\Delta t}{\Delta y}\right)^2 \leq 1,$$

il existe $K_S > 0, \alpha_S > 0$,*tels que pour tout* $(E_x^0, E_y^0, H_z^{-1/2}, G^{-1/2}) \in (H^1(G))^3 \times L^2(G)$,

$$\begin{aligned}\|(E_x^n, E_y^n, H_z^{n-1/2}, G^{n-1/2})\|_{(H^1(G))^3\times L^2(G)} \\ \leq K_S e^{\alpha_S t_n} \|(E_x^0, E_y^0, H_z^{-1/2}, G^{-1/2})\|_{(H^1(G))^3\times L^2(G)}.\end{aligned}$$

Preuve :

- Après transformation de Fourier, le schéma de Yee devient :

$$\left\{\begin{array}{l}\frac{\widehat{E_x^{n+1}}-\widehat{E_x^n}}{\Delta t} - ik_2\widehat{H_z^{n+1/2}} = 0 \\ \frac{\widehat{E_y^{n+1}}-\widehat{E_y^n}}{\Delta t} + ik_1\widehat{H_z^{n+1/2}} + \sigma\frac{\widehat{E_y^{n+1}}+\widehat{E_y^n}}{2} = 0 \\ \frac{\widehat{H_z^{n+1/2}}-\widehat{H_z^{n-1/2}}}{\Delta t} + ik_1\widehat{E_y^n} - ik_2\widehat{E_x^n} + \frac{\widehat{G^{n+1/2}}+\widehat{G^{n-1/2}}}{2} = 0 \\ \frac{\widehat{G^{n+1/2}}-\widehat{G^{n-1/2}}}{\Delta t} + ik_1\sigma\widehat{E_y^n} + \sigma\frac{\widehat{G^{n+1/2}}+\widehat{G^{n-1/2}}}{2} = 0,\end{array}\right.$$

où on a posé :

$$k_1 = \frac{2}{\Delta x}\sin\left(\frac{\xi_1\Delta x}{2}\right) \text{ et } k_2 = \frac{2}{\Delta y}\sin\left(\frac{\xi_2\Delta y}{2}\right).$$

Soit :

$$\widehat{U^n} = (\widehat{E_x^n}, \widehat{E_y^n}, \widehat{H_z^{n-1/2}}, ik_1\widehat{E_x^n}, ik_1\widehat{E_y^n}, ik_1\widehat{H_z^{n-1/2}}, ik_2\widehat{E_x^n}, ik_2\widehat{E_y^n}, ik_2\widehat{H_z^{n-1/2}}, \widehat{G^{n-1/2}}).$$

Posons :

$$R = \begin{pmatrix} 1 & 0 & -ik_2\Delta t \\ 0 & 1+\frac{\sigma\Delta t}{2} & ik_1\Delta t \\ 0 & 0 & 1 \end{pmatrix} = R_0 + \Delta t\widetilde{R},$$

avec :

$$R_0 = \begin{pmatrix} 1 & 0 & -ik_2\Delta t \\ 0 & 1 & ik_1\Delta t \\ 0 & 0 & 1 \end{pmatrix} \text{ et } \widetilde{R} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{\sigma}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-\frac{\sigma\Delta t}{2} & 0 \\ ik_2\Delta t & -ik_1\Delta t & 1 \end{pmatrix} = S_0 + \Delta t\widetilde{S},$$

avec :

$$R_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ ik_2\Delta t & -ik_1\Delta t & 1 \end{pmatrix} \text{ et } \widetilde{S} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{\sigma}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Nous avons alors :

$$\mathcal{R}\widehat{U^{n+1}} = \mathcal{S}\widehat{U^n},$$

avec :

$$\mathcal{R} = \left(\begin{array}{ccc|ccc|ccc|c} & R & & & 0 & & & 0 & & \begin{matrix}0\\0\\\Delta t/2\end{matrix} \\ \hline & 0 & & & R & & & 0 & & \begin{matrix}0\\0\\ik_1\Delta t/2\end{matrix} \\ \hline & 0 & & & 0 & & & R & & \begin{matrix}0\\0\\ik_2\Delta t/2\end{matrix} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1+\sigma\Delta t/2 \end{array}\right) = \mathcal{R}_0 + \Delta t\widetilde{\mathcal{R}},$$

et

$$\mathcal{S} = \left(\begin{array}{ccc|ccc|ccc|c} & S & & & 0 & & & 0 & & \begin{matrix}0\\0\\-\Delta t/2\end{matrix} \\ \hline & 0 & & & S & & & 0 & & \begin{matrix}0\\0\\-ik_1\Delta t/2\end{matrix} \\ \hline & 0 & & & 0 & & & S & & \begin{matrix}0\\0\\-ik_2\Delta t/2\end{matrix} \\ \hline 0 & 0 & 0 & 0 & -\sigma\Delta t & 0 & 0 & 0 & 0 & 1-\sigma\Delta t/2 \end{array}\right) = \mathcal{S}_0 + \Delta t\widetilde{\mathcal{S}},$$

où l'on a posé :

$$\mathcal{R}_0 = \left(\begin{array}{ccc|ccc|ccc|c} & R_0 & & & 0 & & & 0 & & \begin{matrix}0\\0\end{matrix} \\ \hline & 0 & & & R_0 & & & 0 & & \begin{matrix}0\\0\\ik_1\Delta t/2\end{matrix} \\ \hline & 0 & & & 0 & & & R_0 & & \begin{matrix}0\\0\\ik_2\Delta t/2\end{matrix} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

et

$$\widetilde{\mathcal{R}} = \left(\begin{array}{ccc|ccc|ccc|c} & \widetilde{R} & & & 0 & & & 0 & & \begin{matrix}0\\0\\1/2\end{matrix} \\ \hline & 0 & & & \widetilde{R} & & & 0 & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline & 0 & & & 0 & & & \widetilde{R} & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma/2 \end{array}\right),$$

et de même pour $\mathcal{S}_0$ et $\widetilde{\mathcal{S}}$.
Le symbole du schéma de Yee est alors : $\widehat{Q} = \mathcal{R}^{-1}\mathcal{S}$.

- Montrons que $\widehat{Q} = \widehat{Q_0} + \Delta t\widetilde{Q}$ avec $\widehat{Q_0} = \mathcal{R}_0^{-1}\mathcal{S}_0$ et $\|\widetilde{Q}\|$ borné. Nous avons :

$$\begin{aligned} \widehat{Q} &= \mathcal{R}^{-1}\mathcal{S}_0 + \Delta t\mathcal{R}^{-1}\widetilde{S} \\ &= (Id + \Delta t\mathcal{R}_0^{-1}\mathcal{R})^{-1}\mathcal{R}_0^{-1}\mathcal{S}_0 + \Delta t\mathcal{R}^{-1}\widetilde{S}. \end{aligned}$$

  - Montrons qu'il existe K_1 tel que $\|\mathcal{R}^{-1}\| \leq K_1$.
    Nous avons :

$$\mathcal{R}^{-1} = \left(\begin{array}{ccc|ccc|ccc|c} & R^{-1} & & & 0 & & & 0 & & -\frac{\Delta t}{2(1+\sigma\Delta t/2)}R^{-1}\begin{pmatrix}0\\0\\1\end{pmatrix} \\ \hline & 0 & & & R^{-1} & & & 0 & & -\frac{ik_1\Delta t}{2(1+\sigma\Delta t/2)}R^{-1}\begin{pmatrix}0\\0\\1\end{pmatrix} \\ \hline & 0 & & & 0 & & & R^{-1} & & -\frac{ik_2\Delta t}{2(1+\sigma\Delta t/2)}R^{-1}\begin{pmatrix}0\\0\\1\end{pmatrix} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/(1+\sigma\Delta t/2) \end{array}\right),$$

    et

$$R^{-1} = \begin{pmatrix} 1 & 0 & ik_2\Delta t \\ 0 & \left(1+\frac{\sigma\Delta t}{2}\right)^{-1} & -ik_1\Delta t\left(1+\frac{\sigma\Delta t}{2}\right)^{-1} \\ 0 & 0 & 1 \end{pmatrix}.$$

    Or sous condition CFL, $|k_1\Delta t| \leq 1$ et $|k_2\Delta t| \leq 1$. Donc il existe K_1 tel que $\|\mathcal{R}^{-1}\| \leq K_1$.
    Nous avons montré ainsi que $\|\mathcal{R}^{-1}\widetilde{S}\|$ est borné.

– Montrons qu'il existe K_2 tel que $\|(Id+\Delta t\mathcal{R}_0^{-1}\mathcal{R})^{-1}-Id\| \leq K_2\Delta t$. Comme, de même que pour $\mathcal{R}^{-1}$, il existe K_3 tel que $\|\mathcal{R}_0^{-1}\| \leq K_3$, cela démontrera la décomposition de $\widehat{Q}$.
Il existe M tel que $\widetilde{R} \leq M$, donc si $\Delta t \leq \frac{1}{K_3 M}$, alors $\|\Delta t\mathcal{R}_0^{-1}\mathcal{R}\| < 1$. Or :

$$\begin{aligned}
\|(Id+\Delta t\mathcal{R}_0^{-1}\mathcal{R})^{-1}-Id\| &= \left\|\sum_{k=1}^{+\infty}(-\Delta t)^k(\mathcal{R}_0^{-1}\mathcal{R})^k\right\| \\
&\leq \Delta t\sum_{k=0}^{+\infty}\Delta t^k\|\mathcal{R}_0^{-1}\mathcal{R}\|^k \\
&\leq \frac{\Delta t K_3 M}{1+\Delta t K_3 M} \\
&\leq K_2\Delta t.
\end{aligned}$$

- Il reste à montrer que le schéma de symbole $\widehat{Q_0}$ est stable. En effet, le proposition 8.1 montrera que le schéma de symbole $\widehat{Q}$ est stable et, étant donnée la définition de $\widehat{U^n}$, nous aurons l'estimation discrète voulue. Nous avons :

$$\mathcal{R}_0^{-1}\mathcal{S}_0 = \left(\begin{array}{ccc|ccc|ccc|c}
\multicolumn{3}{c|}{R_0^{-1}S_0} & \multicolumn{3}{c|}{0} & \multicolumn{3}{c|}{0} & \begin{matrix}0\\0\\0\end{matrix} \\ \hline
\multicolumn{3}{c|}{0} & \multicolumn{3}{c|}{R_0^{-1}S_0} & \multicolumn{3}{c|}{0} & -ik_1\Delta tz \\ \hline
\multicolumn{3}{c|}{0} & \multicolumn{3}{c|}{0} & \multicolumn{3}{c|}{R_0^{-1}S_0} & -ik_2\Delta tz \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right),$$

où $z = R_0^{-1}\begin{pmatrix}0\\0\\1\end{pmatrix}$.
Nous remarquons que $R_0^{-1}S_0$ est le symbole du schéma de Yee pour les équations de Maxwell. Nous savons donc, d'après la preuve du théorème 8.1 qu'il existe une matrice de passage P telle que $R_0^{-1}S_0 = PDP^{-1}$ avec $\|P\| \leq K$, $\|P^{-1}\| \leq K$ et D matrice diagonale à éléments diagonaux de module inférieur ou égal à 1.

Posons :

$$\mathcal{P} = \left(\begin{array}{ccc|ccc|ccc|c} & P & & & 0 & & & 0 & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline & 0 & & & P & & & 0 & & Z_1 \\ \hline & 0 & & & 0 & & & P & & Z_2 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

alors :

$$\mathcal{P}^{-1} = \left(\begin{array}{ccc|ccc|ccc|c} & P^{-1} & & & 0 & & & 0 & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline & 0 & & & P^{-1} & & & 0 & & -P^{-1}Z_1 \\ \hline & 0 & & & 0 & & & P^{-1} & & -P^{-1}Z_2 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

et

$$\mathcal{P}^{-1}\mathcal{R}_0^{-1}\mathcal{S}_0\mathcal{P} = \left(\begin{array}{ccc|ccc|ccc|c} & D & & & 0 & & & 0 & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline & 0 & & & D & & & 0 & & P^{-1}\left((R_0^{-1}S_0 - Id)Z_1 - ik_1\Delta tz\right) \\ \hline & 0 & & & 0 & & & D & & P^{-1}\left((R_0^{-1}S_0 - Id)Z_2 - ik_2\Delta tz\right) \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right).$$

Nous choisissons $Z_1 = -\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ et $Z_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, alors :

$$\mathcal{P}^{-1}\mathcal{R}_0^{-1}\mathcal{S}_0\mathcal{P} = \left(\begin{array}{ccc|ccc|ccc|c} & D & & & 0 & & & 0 & & \begin{matrix}0\\0\\0\end{matrix} \\ \hline & 0 & & & D & & & 0 & & 0 \\ \hline & 0 & & & 0 & & & D & & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

qui est une matrice diagonale à coefficients de module inférieur ou égal à 1. Comme $\|\mathcal{P}\| \leq K$, $\|\mathcal{P}^{-1}\| \leq K$, on a montré que le schéma de symbole $\widehat{Q_0}$ est stable, ce qui achève la preuve.

□

8.3 Application des résultats de la première partie

Nous allons, en appliquant les résultats de la première partie de cette thèse, montrer que le schéma de Yee est convergent et minorer son taux de convergence.

Nous considérons les équations de Maxwell PML en dimension 2 avec absorption constante dans la direction des x et nulle dans la direction des y et nous considérons les variables E_x, E_y, H_z et $G = \sigma H_{zx}$:

$$\begin{cases} \partial_t E_x - \partial_y H_z = 0 \\ \partial_t E_y + \partial_x H_z + \sigma E_y = 0 \\ \partial_t H_z + \partial_x E_y - \partial_y E_x + G = 0 \\ \partial_t G + \sigma \partial_x E_y + \sigma G = 0. \end{cases} \tag{8.3}$$

Nous discrétisons ces équations par le schéma de Yee (8.2).

Proposition 8.2 *Le problème (8.3) est faiblement bien posé de défaut $q_1 = 1$ et le schéma de Yee (8.2) est faiblement stable de défaut $q_2 = 1$.*
De plus, si la donnée initiale du schéma est de régularité H^{q_4}, le taux de convergence du schéma est minoré par : $\frac{q_4 - 1}{\max(q_4, 4) - 1}$.

Preuve :

- Nous appliquons le résultat de la proposition 7.1 à $\sigma_x = \sigma$ constant et $\sigma_y = 0$. Nous obtenons :

$$\begin{aligned}&\|(E_x, E_y, H_z)(t,.)\|^2_{H^1} + \|(\sigma E_y, G)(t,.)\|^2_{L^2}\\ &\quad\leq C_1 e^{(C_2+2\sigma)t}\left(\|(E_x^0, E_y^0, H_z^0)\|^2_{H^1} + \|(\sigma E_y^0, G^0)\|^2_{L^2}\right),\end{aligned}$$

  d'où l'existence de constantes K_C et α_C telles que : $\|(E_x, E_y, H_z, G)(t,.)\|_{L^2} \leq K_C e^{\alpha_C t}\|(E_x^0, E_y^0, H_z^0, G^0)\|_{H^1}$, ce qui prouve que le problème est faiblement bien posé de défaut $q_1 = 1$.
- Une conséquence immédiate du théorème 8.2 est que :

$$\|(E_x^n, E_y^n, H_z^{n-1/2}, G^{n-1/2})\|_{(L^2(G)^4)} \leq K_S e^{\alpha_S t_n}\|(E_x^0, E_y^0, H_z^{-1/2}, G^{-1/2})\|_{(H^1(G))^4},$$

  ce qui prouve que le schéma est faiblement stable de défaut 1.
- Pour évaluer le taux de convergence, nous allons utiliser la proposition 3.4. Il suffit alors de montrer que $r = 1$ et $\rho = 2$ pour avoir le résultat.
  Nous posons, comme dans la preuve précédente $k_1 = \frac{2}{\Delta x}\sin\left(\frac{\xi_1\Delta x}{2}\right)$ et $k_2 = \frac{2}{\Delta y}\sin\left(\frac{\xi_2\Delta y}{2}\right)$. La matrice d'amplification du schéma de Yee est alors :

$$\widehat{Q} = \begin{pmatrix} 1-k_2^2\Delta t^2 & \frac{2k_1\Delta t^2 k_2}{2+\sigma\Delta t} & ik_2\Delta t & \frac{-2ik_2\Delta t^2}{2+\sigma\Delta t} \\ \frac{2k_1\Delta t^2 k_2}{2+\sigma\Delta t} & 1-\frac{2\Delta t(2\sigma+\sigma^2\Delta t+2k_1^2\Delta t)}{(2+\sigma\Delta t)^2} & \frac{-2ik_1\Delta t}{2+\sigma\Delta t} & \frac{4ik_1\Delta t^2}{(2+\sigma\Delta t)^2} \\ ik_2\Delta t & \frac{-2ik_1\Delta t}{2+\sigma\Delta t} & 1 & \frac{-2\Delta t}{2+\sigma\Delta t} \\ 0 & \frac{-2ik_1\sigma\Delta t}{(2+\sigma\Delta t} & 0 & 1-\frac{2\sigma\Delta t}{2+\sigma\Delta t} \end{pmatrix}.$$

  Or le symbole du problème continu est :

$$P(i\xi) = \begin{pmatrix} 0 & 0 & i\xi_2 & 0 \\ 0 & -\sigma & -i\xi_1 & 0 \\ i\xi_2 & -i\xi_2 & 0 & -1 \\ 0 & -i\xi_1\sigma & 0 & -\sigma \end{pmatrix}.$$

  Donc :

$$\begin{aligned}\widehat{Q} &= Id + \Delta t P(i\xi) + O(\Delta t^2\|\xi\|^2)\\ &= \exp(\Delta t P(i\xi)) + O(\Delta t^2\|\xi\|^2),\end{aligned}$$

ce qui prouve bien que $r = 1$ et $\rho = 2$. □

Ce résultat implique que le schéma de Yee est d'ordre 1 ce qui ne devrait pas être le cas, d'après sa construction. Cependant, le schéma de Yee est une approximation de $(E_x(t_n,.), E_y(t_n,.), H_z(t_{n-1/2},.), G(t_{n-1/2},.))$ et pas de la solution à l'instant t_n. Comme l'approximation de $(E_x(t_n,.), E_y(t_n,.), H_z(t_n,.), G(t_n,.))$ par

$(E_x^n, E_y^n, H_z^{n-1/2}, G^{n-1/2})$ n'est que d'ordre 1, il est normal que l'on obtienne un ordre 1.

Pour remédier à ce problème, nous allons approcher $(E_x(t_n,.), E_y(t_n,.), H_z(t_n,.), G(t_n,.))$ par $(E_x^n, E_y^n, \frac{H_z^{n-1/2}+H_z^{n+1/2}}{2}, \frac{G^{n-1/2}+G^{n+1/2}}{2})$ qui est bien d'ordre 2. De plus, il est clair que le schéma reste faiblement stable de défaut $q_2 = 1$.

Proposition 8.3 *Si nous posons* $(E_x^n, E_y^n, H_z^n, G^n) = (E_x^n, E_y^n, \frac{H_z^{n-1/2}+H_z^{n+1/2}}{2}, \frac{G^{n-1/2}+G^{n+1/2}}{2})$, *alors nous définissons bien un schéma à un pas en temps. Le taux de convergence du schéma est minoré par :* $\frac{2(q_4-1)}{\max(q_4,5)-1}$ *pour une donnée initiale de régularité* H^{q_4}.

PREUVE : Nous notons $V^n = {}^t(E_x^n, E_y^n, H_z^n, G^n)$ et $W^n = {}^t(E_x^n, E_y^n, H_z^{n-1/2}, G^{n-1/2})$. Alors, nous savons que $\widehat{W^{n+1}} = \widehat{Q}\widehat{W^n}$, où $\widehat{Q}$ est la matrice d'amplification du schéma de Yee calculé précédemment. Nous allons calculer $\widetilde{Q}$ tel que $\widehat{V^{n+1}} = \widetilde{Q}\widehat{V^n}$.
Nous posons :

$$M_+ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ et } M_- = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Nous avons alors :

$$\begin{aligned} \widehat{V^{n+1}} &= (M_+ + \frac{1}{2}M_-)\widehat{W^{n+1}} + \frac{1}{2}M_-\widehat{W^{n+2}} \\ &= \left((M_+ + \frac{1}{2}M_-)\widehat{Q} + \frac{1}{2}M_-\widehat{Q}^2\right)\widehat{W^n} \\ &= \left((M_+ + \frac{1}{2}M_-)\widehat{Q} + \frac{1}{2}M_-\widehat{Q}^2\right)\left(M_+\widehat{V^n} + \begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n-1/2}} \\ \widehat{G^{n-1/2}} \end{pmatrix}\right) \end{aligned}$$

Or, nous avons :

$$\begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n-1/2}} \\ \widehat{G^{n-1/2}} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n+1/2}} \\ \widehat{G^{n+1/2}} \end{pmatrix} = 2\begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^n} \\ \widehat{G^n} \end{pmatrix}. \tag{8.4}$$

Nous allons calculer ${}^t(0, 0, \widehat{H_z^{n+1/2}}, \widehat{G^{n+1/2}})$.

Nous avons alors :

$$\begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n+1/2}} \\ \widehat{G^{n+1/2}} \end{pmatrix} = M_-\widehat{W^{n+1}} = M_-\widehat{Q}\widehat{W^n}$$

$$= M_-\widehat{Q}\left(M_+\widehat{V^n} + M_-\begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n-1/2}} \\ \widehat{G^{n-1/2}} \end{pmatrix}\right).$$

Donc, en utilisant (8.4), nous obtenons :

$$\left(Id + M_-\widehat{Q}M_-\right)\begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n-1/2}} \\ \widehat{G^{n-1/2}} \end{pmatrix} + M_-\widehat{Q}M_+\widehat{V^n} = 2M_-\widehat{V^n}.$$

Ainsi :

$$\begin{pmatrix} 0 \\ 0 \\ \widehat{H_z^{n-1/2}} \\ \widehat{G^{n-1/2}} \end{pmatrix} = \left(Id + M_-\widehat{Q}M_-\right)^{-1}(2M_- - M_-\widehat{Q}M_+)\widehat{V^n}.$$

Nous obtenons alors :

$$\widehat{V^{n+1}} = \left((M_+ + \frac{1}{2}M_-)\widehat{Q} + \frac{1}{2}M_-\widehat{Q}^2\right)\left(M_+ + \left(Id + M_-\widehat{Q}M_-\right)^{-1}(2M_- - M_-\widehat{Q}M_+)\right)\widehat{V^n}.$$

Nous avons donc :

$$\widetilde{Q} = \left((M_+ + \frac{1}{2}M_-)\widehat{Q} + \frac{1}{2}M_-\widehat{Q}^2\right)\left(M_+ + \left(Id + M_-\widehat{Q}M_-\right)^{-1}(2M_- - M_-\widehat{Q}M_+)\right),$$

et le calcul explicite de $\widetilde{Q}$ ainsi qu'un développement limité donnent :

$$\widetilde{Q} = Id + \Delta t P(i\xi) + \frac{\Delta t^2}{2}P(i\xi)^2 + O(\Delta t^3\|\xi\|^3 = \exp(tP(i\xi)) + O(\Delta t^3\|\xi\|^3.$$

Nous avons donc prouvé que $r = 2$ et $\rho = 3$. La proposition 3.4 permet de conclure.

□

Chapitre 9

Stabilité WKB

Le but de cette partie est de donner une condition nécessaire de stabilité analogue au théorème 9.2 donnée dans [9] mais qui s'applique dans des cas plus généraux.

Ce travail a été effectué avec Jeffrey Rauch.

9.1 Le problème de la stabilité

9.1.1 Origine de l'instabilité

Le phénomène d'instabilité observé numériquement a été expliqué par Hu dans [22] puis, plus récemment, dans [23]. Il est dû à une direction opposée de la vitesse de phase et de la vitesse de groupe.

En effet, si nous reprenons le changement de variables (6.4), et si nous considérons une onde plane de la forme $e^{i(\xi_1 x+\xi_2 y-\omega t)}$, alors elle devient : $e^{i(\xi_1 x+\xi_2 y-\omega t)}e^{-\frac{\xi_1}{\omega}\int_0^x \sigma_x(u)du}$. L'amplitude de cette onde ne va donc décroître exponentiellement que si : $-\frac{\xi_1}{\omega}\int_0^x \sigma_x(u) < 0$. Ainsi, si le coefficient d'absorption est strictement positif et si $x > 0$, l'onde sera absorbée exponentiellement lorsque $\frac{\xi_1}{\omega} > 0$, ce qui peut se traduire en terme de vitesse de phase par : la composante dans la direction de l'axe des x de la vitesse de phase est positive. Or nous avons choisi de prendre $x > 0$, c'est-à-dire de placer la couche absorbante à droite du domaine d'intérêt. Les ondes étudiées vont donc se diriger de la gauche vers la droite, ce qui signifie que la composante dans la direction de l'axe des x de leur vitesse de groupe est positive. En conséquence, nous pouvons remarquer que si la vitesse de groupe et la vitesse de phase ne sont pas dans la même direction selon l'axe des x, l'onde ne sera pas absorbée exponentiellement dans la couche, ce qui posera des problèmes.

Dans [9], il a été montré de manière rigoureuse que la cohérence de la direction des vitesses de phase et de groupe était une condition nécessaire de stabilité. Nous rappellerons l'énoncé de ce résultat dans la partie suivante.

9.1.2 Des équations PML fortement bien posées et stables

De nouvelles équations PML ont été proposées dans [22] pour résoudre le problème de la stabilité pour les équations d'Euler linéarisées. La méthode consiste à appliquer une transformation de Fourier en espace et de Laplace en temps et d'utiliser ensuite le changement complexe de variables qui correspond aux équations PML. L'introduction d'une nouvelle variable permettant de se ramener aux composantes non splittées va ensuite conduire à un problème stable. Toutefois ce problème n'est pas fortement bien posé. Les équations vont donc être modifiées grâce à l'introduction d'un petit paramètre qui va rendre le problème symétrique hyperbolique donc fortement bien posé. Toutefois, l'inconvénient du paramètre introduit est qu'il fait perdre le caractère parfaitement adapté. Cependant les expériences numériques montrent que, lorsque le paramètre est choisi suffisamment petit, les résultats obtenus avec et sans l'introduction de ce paramètres sont très proches. Les équations uniquement stabilisées sont utilisées en pratique à moins que l'étude ne nécessite la conservation du caractère symétrique des équations d'Euler.

Des équations parfaitement adaptées, fortement bien posées et stables ont aussi été données pour les équations de Maxwell dans [11]. Le changement de variables complexes a été effectué sur les équations de Zhao-Cangellaris et des estimations d'énergie ont prouvé que les champs étaient bornés indépendamment du temps ce qui démontre la stabilité. Une autre technique a été proposée dans [5] pour les équations de Maxwell tridimensionnelles. Les équations obtenues ne concernent que les variables physiques non splittées.

Une méthode plus générale est présentée dans [6]. Elle concerne les systèmes symétriques hyperboliques quelconques et est basée sur les PML proposés par Hagstrom. Le principe est d'introduire un terme parabolique dans les équations puis de jouer sur les coefficients pour stabiliser le problème. Cette méthode est appliquée aux cas des équations de Maxwell TM et d'Euler linéarisées.

9.2 Etude de la stabilité pour des équations PML générales

9.2.1 Rappel des définitions et des principaux résultats antérieurs

Nous considérons le système dans $\mathbb{R}^2$ suivant :

$$\partial_t U - A_1 \partial_x U - A_2 \partial_y U = 0, \tag{9.1}$$

où A_1, $A_2 \in \mathcal{M}_N(\mathbb{R})$.

Les équations PML associées à ce problème avec absorption dans la direction des x sont alors :

$$\begin{cases} \partial_t U^1 - A_1 \partial_x (U^1 + U^2) + \sigma_x U^1 = 0 \\ \partial_t U^2 - A_2 \partial_y (U^1 + U^2) = 0, \end{cases} \tag{9.2}$$

où $\sigma_x > 0$ désigne l'absorption dans la direction des x.

Nous avons [9] alors le résultat suivant :

Théorème 9.1 *Sous les hypothèses suivantes :*

- *le problème (9.1) est fortement hyperbolique*
- *le symbole associé à (9.1) admet un nombre constant N_e de valeurs propres simples non nulles et la valeur propre 0 avec la multiplicité $N - N_e$.*

Alors le problème (9.2) est faiblement bien posé.

Dans la définition d'un problème fortement ou faiblement bien posé, une croissance exponentielle en temps de la solution est autorisée et elle peut être considérée comme une instabilité. C'est pourquoi, une nouvelle définition de stabilité est introduite par Bécache, Fauqueux et Joly dans laquelle seule une croissance polynomiale en temps est autorisée.

Définition 9.1 *Le système (9.2) est dit stable s'il est faiblement bien posé de défaut $s \geq 0$ et si il existe $K > 0$ tel que pour toute donnée initiale U^0 :*

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

Il est facile d'obtenir alors la caractérisation analogue à celle des problèmes faiblement bien posés suivante :

Proposition 9.1 *Un problème de Cauchy de symbole $P(i\xi)$ est stable si et seulement si pour tout ξ, les valeurs propres $\lambda(\xi)$ de $P(i\xi)$ vérifient :*

$$Re\,(\lambda(\xi)) \leq 0$$

Une condition nécessaire de stabilité est alors démontrée dans [9] :

Théorème 9.2 *Sous les mêmes hypothèses que pour le théorème 9.1 une condition nécessaire de stabilité pour le problème (9.2) est que les valeurs propres $\lambda(\xi)$ non nulles de $P(i\xi) = i\xi_1 A_1 + i\xi_2 A_2$ vérifient :*

$$\forall \xi,\ \|\xi\| = 1,\ \frac{\xi_1}{\lambda(\xi)} . \partial_{\xi_1} \lambda(\xi) \geq 0.$$

La démonstration de ce théorème est basée sur un développement limité en $\frac{\sigma_x}{\|\xi\|}$ proche de 0. Le cas d'une absorption variable dans les deux directions n'est donc pas traité.

Ce résultat s'applique aux équations de l'élastodynamique et de Maxwell bidimensionnelles pour lesquelles les valeurs propres sont bien simples.

9.2.2 Résultat général

Nous considérons maintenant le système dans $\mathbb{R}^d$ suivant :

$$\partial_t U - \sum_{l=1}^{d} A_l \partial_{x_l} U = 0,$$

avec $A_l \in \mathcal{M}_N(\mathbb{R})$. Dans les exemples pratiques, nous prendrons $d = 2$ ou 3.

Nous étudions les équations PML avec absorptions variables dans toutes les directions :

$$\partial_t U^l - A_l \partial_{x_l}(\textstyle\sum_{j=1}^{d} U^j) + \sigma_l(x_l) U^l = 0, \quad 1 \leq l \leq d. \tag{9.3}$$

Pour $l \in \{1, \ldots, d\}$ nous notons $\widetilde{A}_l \in \mathcal{M}_{Nd}(\mathbb{R})$ les matrices correspondant aux équations PML :

$$\widetilde{A}_l = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 \\ \vdots & & & & \vdots \\ A_l & \ldots & \ldots & \ldots & A_l \\ \vdots & & & & \vdots \\ 0 & \ldots & \ldots & \ldots & 0 \end{pmatrix} \leftarrow l^{\text{iéme}} \text{ ligne },$$

et, pour $x = (x_1, \ldots, x_d)$, $B(x) \in \mathcal{M}_{Nd}(\mathbb{R})$ le terme d'ordre 0 est la matrice diagonale suivante :

$$B(x) = \begin{pmatrix} \sigma_1(x_1) Id_N & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_d(x_d) Id_N \end{pmatrix}.$$

Nous notons $\widetilde{P}(i\xi)$ le symbole principal correspondant à l'opérateur associé aux équations (9.3) et $P(i\xi)$ le symbole du problème initial (9.1) :

$$\widetilde{P}(i\xi) = \sum_{l=1}^{d} i\xi_l \widetilde{A}_l \text{ et } P(i\xi) = \sum_{l=1}^{d} i\xi_l A_l.$$

Nous fixons pour toute la suite $\xi \neq 0$ et nous considérons $\lambda(\xi)$ une valeur propre non nulle de $\widetilde{P}(i\xi)$. Le lemme suivant prouve que $\lambda(\xi)$ est aussi une valeur propre de $P(i\xi)$.

Lemme 9.1

$$\forall \xi \in \mathbb{R}^d,\ det(\widetilde{P}(i\xi) - \lambda Id_{Nd}) = (-\lambda)^{N(d-1)} det(P(i\xi) - \lambda Id_N).$$

PREUVE :

$$
\begin{aligned}
det(\widetilde{P}(i\xi)-\lambda Id_{Nd}) &= \begin{vmatrix} i\xi_1A_1-\lambda Id_N & \dots & i\xi_1A_1 \\ \vdots & \ddots & \vdots \\ i\xi_dA_d & \dots & i\xi_dA_d-\lambda Id_N \end{vmatrix} \\
&= \begin{vmatrix} P(i\xi)-\lambda Id_N & \dots & \dots & P(i\xi)-\lambda Id_N \\ i\xi_2A_2 & i\xi_2A_2-\lambda Id_N & \dots & i\xi_2A_2 \\ \vdots & & \ddots & \vdots \\ i\xi_dA_d & \dots & \dots & i\xi_dA_d-\lambda Id_N \end{vmatrix} \quad (L_1 \leftarrow \sum_{l=1}^{d} L_l) \\
&= det(P(i\xi)-\lambda Id_N) \begin{vmatrix} Id_N & \dots & \dots & Id_N \\ i\xi_2A_2 & i\xi_2A_2-\lambda Id_N & \dots & i\xi_2A_2 \\ \vdots & & \ddots & \vdots \\ i\xi_dA_d & \dots & \dots & i\xi_dA_d-\lambda Id_N \end{vmatrix} \\
&= det(P(i\xi)-\lambda Id_N) \begin{vmatrix} Id_N & 0 & \dots & 0 \\ i\xi_2A_2 & -\lambda Id_N & \dots & 0 \\ \vdots & & \ddots & \vdots \\ i\xi_dA_d & 0 & \dots & -\lambda Id_N \end{vmatrix} \quad (C_j \leftarrow C_j - C_1) \\
&= (-\lambda)^{N(d-1)} det(P(i\xi)-\lambda Id_N).
\end{aligned}
$$

□

Nous cherchons alors une solution $U=(U^1,\dots,U^d)$ de (9.3) obtenue par l'approximation de l'optique géométrique (voir [36]) :

$$U^\varepsilon \sim e^{\Phi/\varepsilon}\sum_{j=0}^{+\infty}\varepsilon^j a_j(t,x).$$

Le principe de l'optique géométrique est de chercher une approximation de la solution à haute fréquence ou approximation WKB, d'où l'introduction de la phase $\frac{\Phi}{\varepsilon}$ qui va être grande pour ε petit. De plus, cette approximation est d'ordre infini en ε, dans le sens où, les coefficients a_j sont tels que si $U_N^\varepsilon = e^{\Phi/\varepsilon}\sum_{j=0}^{N}\varepsilon^j a_j(t,x)$, $\partial_t U_N^\varepsilon - \sum_{l=1}^{d}\widetilde{A}_l\partial_{x_l}U_N^\varepsilon + BU_N^\varepsilon = O(\varepsilon^N)$. Donc nous allons chercher à construire U^ε qui soit une solution approchée de (9.3).

Théorème 9.3 *Nous supposons que le problème $\partial_t U - \sum_{l=1}^{d} A_l\partial_{x_l}U = 0$ vérifie les hypothèses suivantes :*

- *le problème est fortement bien posé,*
- *$det(P(i\xi)-XId) = (-X)^{\alpha_0}\prod_{k=1}^{s}(\lambda_k(\xi)-X)^{\alpha_k}$ avec, si $k\neq 0$, $\lambda_k(\xi)\neq 0$ et $dim\, ker(P(i\xi)-\lambda_k(\xi)Id)=\alpha_k$.*

Soit $k \geq 0$ et $\xi = (\xi_1, \ldots, \xi_d) \in \mathbb{R}^d \setminus \{(0, \ldots, 0)\}$ fixés, nous considérons la fonction suivante :

$$U^\varepsilon = e^{\Phi/\varepsilon} \sum_{j=0}^{+\infty} \varepsilon^j a_j(t, x, \xi),$$

avec $\Phi(t, x, \xi) = t.\lambda_k(\xi) + ix.\xi$ et les coefficients a_j vérifient la relation de récurrence suivante :

$$a_0 \in ker(\widetilde{P}(i\xi) - \lambda_k(\xi) Id).$$

$$\forall j \geq 0,\ (\lambda_k(\xi) Id - \widetilde{P}(i\xi)) a_{j+1}(t, x, \xi) + (\partial_t - \sum_{l=1}^{d} \widetilde{A}_l \partial_{x_l} + B) a_j(t, x, \xi) = 0.$$

Alors U^ε est une solution approchée de l'équation (9.3).

De plus, si nous notons $ker(\widetilde{P}(i\xi) - \lambda_k(\xi) Id) = Vect(V_k^1(\xi), \ldots, V_k^{\alpha_k}(\xi))$, alors $a_0(t, x, \xi) = \sum_{j=1}^{\alpha_k} \mu_k^j(t, x, \xi)) V_k^j$, et il existe des fonctions scalaires $w_1^{kj}(\xi)$ et $w_2^{kj}(\xi)$ telles que les μ_k^j vérifient les équations de transport suivantes :

$$\partial_t \mu_k^j - \sum_{l=1}^{d} w_l^{kj} \partial_{x_l} \mu_k^j + \beta_{kj}(x) \mu_k^j = 0,$$

avec $\widetilde{\Pi}_{kj}(\xi) B(x) \widetilde{\Pi}_{kj}(\xi) = \beta_{kj}(x, \xi) \widetilde{\Pi}_{kj}(\xi)$ où $\widetilde{\Pi}_{kj}(\xi)$ désigne la projection sur le sous-espace engendré par $V_k^j(\xi)$.

Preuve :

Nous commençons par déterminer les équations vérifiées par les coefficients a_j. Pour cela, nous allons identifier les termes en ε^j obtenus par injection de U^ε dans (9.3).
Nous avons :

$$\begin{aligned} &\partial_t U^\varepsilon - \textstyle\sum_{l=1}^{d} \widetilde{A}_l \partial_{x_l} U^\varepsilon + B(x) U^\varepsilon \\ &\quad = e^{\Phi/\varepsilon} \Big(\tfrac{1}{\varepsilon} (\lambda_k(\xi) a_0 - \widetilde{P}(i\xi)) a_0 \\ &\qquad + \textstyle\sum_{j=0}^{+\infty} (\partial_t a_j + \lambda_k(\xi) a_{j+1} - \sum_{l=1}^{d} \widetilde{A}_l (\partial_{x_l} a_j + i\xi_l a_{j+1}) + B a_j) \varepsilon^j \Big). \end{aligned}$$

D'où les relations suivantes :

- $(\lambda_k(\xi) Id - \widetilde{P}(i\xi)) a_0 = 0$,
- $\forall j \geq 0,\ (\lambda_k(\xi) Id - \widetilde{P}(i\xi)) a_{j+1} + (\partial_t - \sum_{l=1}^{d} \widetilde{A}_l \partial_{x_l} + B(x, y)) a_j = 0$.

Nous allons maintenant nous intéresser au coefficient dominant a_0. Le premier point donne :

$$a_0 \in ker(\widetilde{P}(i\xi) - \lambda_k(\xi) Id). \tag{9.4}$$

Le second point donne, pour j=0, l'équation suivante :

$$(\lambda_k(\xi)Id - \widetilde{P}(i\xi))a_1 + (\partial_t - \sum_{l=1}^{d} \widetilde{A}_l \partial_{x_l} + B(x,y))a_0 = 0. \tag{9.5}$$

Nous allons maintenant retranscrire l'hypothèse portant sur les espaces propres de $P(i\xi)$ à ceux de $\widetilde{P}(i\xi)$. Le lemme précédent montre que $det(\widetilde{P}(i\xi) - XId) = (-X)^{N(d-1)+\alpha_0} \prod_{k=1}^{s} (\lambda_k(\xi) - X)^{\alpha_k}$. De plus, si $V(\xi)$ est un vecteur propre pour $P(i\xi)$ associé à la valeur propre $\lambda(\xi) \neq 0$, alors, comme énoncé dans [9], le vecteur $\left(\frac{\xi_1}{\lambda(\xi)} A_1 V(\xi), \ldots, \frac{\xi_d}{\lambda(\xi)} A_d V(\xi)\right)$ est un vecteur propre pour $\widetilde{P}(i\xi)$ associé à la valeur propre $\lambda(\xi)$. Ainsi, nous avons bien $dim\, ker(\widetilde{P}(i\xi) - \lambda_k(\xi)Id) = \alpha_k$.

Nous considérons alors la décomposition en sous-espaces caractéristiques de $\widetilde{P}(i\xi)$:

$$\bigoplus_{k=1}^{s} ker\left(\widetilde{P}(i\xi) - \lambda_k(\xi)Id\right) \oplus ker\left(\widetilde{P}(i\xi)^{\alpha_0+N(d-1)}\right) = \mathbb{C}^{dN}.$$

Pour $1 \leq k \leq s$, nous considérons $(V_k^1(\xi), \ldots, V_k^{\alpha_k}(\xi))$ une base de $ker(\widetilde{P}(i\xi) - \lambda_k(\xi)Id)$. Nous considérons alors la décomposition suivante :

$$\bigoplus_{k=1}^{s} \bigoplus_{j=1}^{\alpha_k} Vect(V_k^j) \oplus ker(\widetilde{P}(i\xi)^{\alpha_0+N(d-1)}) = \mathbb{C}^{dN}. \tag{9.6}$$

Nous considérons la projection $\widetilde{\Pi}_{kj}$ sur $Vect(V_k^j(\xi))$ qui correspond à la décomposition précédente.

Lemme 9.2

$$\widetilde{\Pi}_{kj}(\lambda_k(\xi)Id - \widetilde{P}(i\xi)) = 0.$$

Preuve :

Soit $X \in \mathbb{C}^{dN}$, nous écrivons sa décomposition selon la somme (9.6) :

$$X = \sum_{l=1}^{s} \sum_{j=1}^{\alpha_l} X_{j,l} + X_0.$$

Alors :

$$\begin{aligned}(\lambda_k(\xi)Id - \widetilde{P}(i\xi))X &= \sum_{l=1}^{s} \sum_{j=1}^{\alpha_l} (\lambda_k(\xi) - \lambda_l(\xi))X_{j,l} + \lambda_k(\xi)X_0 - \widetilde{P}(i\xi)X_0 \\ &= \sum_{l\in\{1\ldots s\}, l\neq k} \sum_{j=1}^{\alpha_l} (\lambda_k(\xi) - \lambda_l(\xi))X_{j,l} + \lambda_k(\xi)X_0 - \widetilde{P}(i\xi)X_0,\end{aligned}$$

et cette écriture correspond toujours à la décomposition (9.6).

Donc :

$$\widetilde{\Pi}_{kj}(\lambda_k(\xi)Id - \widetilde{P}(i\xi))X = 0,$$

ce qui prouve le lemme.

□

Donc, en appliquant $\widetilde{\Pi}_{kj}$ à l'équation (9.5), nous obtenons :

$$\widetilde{\Pi}_{kj}(\partial_t - \sum_{l=1}^{d} \widetilde{A}_l \partial_{x_l} + B(x))a_0 = 0.$$

Or d'après, (9.4), $a_0 \in ker(\widetilde{P}(i\xi) - \lambda_k(\xi)Id)$, donc :

$$a_0 = \sum_{j=1}^{\alpha_k} \mu_k^j(t,x)V_k^j,$$

et nous avons :

$$\widetilde{\Pi}_{kj}a_0 = \mu_k^j(t,x)V_k^j$$

donc :

$$\widetilde{\Pi}_{kj}\partial_t a_0 = (\partial_t \mu_k^j)V_k^j.$$

Posons, pour $1 \leq l \leq d$, $\widetilde{\Pi}_{kj}\widetilde{A}_l V_k^j = w_l^{kj} V_k^j$ et $\widetilde{\Pi}_{kj}B(x)V_k^j = \beta_{kj}(x)V_k^j$. Alors, par définition de $\widetilde{\Pi}_{kj}$, nous avons $\widetilde{\Pi}_{kj}B(x)\widetilde{\Pi}_{kj} = \beta_{kj}(x)\widetilde{\Pi}_{kj}$.

Nous avons alors :

$$\widetilde{\Pi}_{kj}\widetilde{A}_l\partial_{x_l}a_0 = w_l^{kj}\partial_{x_l}\mu_k^j V_k^j \text{ et } \widetilde{\Pi}_{kj}B(x)\partial_x a_0 = \beta_{kj}(x)\mu_k^j V_k^j.$$

Nous obtenons alors l'équation de transport suivante vérifiée par μ_k^j :

$$\partial_t \mu_k^j - \sum_{l=1}^{d} w_l^{kj}\partial_{x_l}\mu_k^j + \beta_{kj}(x)\mu_k^j = 0. \tag{9.7}$$

□

Corollaire 9.1 *Avec les notations et les hypothèses du théorème précédent, une condition nécessaire de stabilité pour le problème (9.3) est :*

$$\forall k \in \{1,\ldots,s\},\ \forall j \in \{1,\ldots,\alpha_k\}, \forall \xi \neq 0,\ \exists x \in \mathbb{R}^d,\ \exists a \in \mathbb{R}^d,\ \exists t > 0,$$

$$Re\,(\int_0^t \beta_{kj}(x + a(\tau - t), \xi))d\tau > 0.$$

PREUVE : La solution de l'équation de transport (9.7) est :

$$\begin{aligned}\mu_k^j(t,x) &= \exp\left(-\int_0^t \beta_{kj}(x_1+w_1^{kj}(\tau-t),\ldots,x_d+w_d^{kj}(\tau-t))d\tau\right)\\ &\quad\times\mu_k^j(0,x_1-w_1t,\ldots,x_d-w_dt).\end{aligned}$$

Donc, si la condition donnée dans l'énoncé n'est pas vérifiée, il existe $k\in\{1,\ldots,s\}$, $j\in\{1,\ldots,\alpha_k\}$ et $\xi\neq 0$ tels que pour tout $x\in\mathbb{R}^d$, pour tout $a\in\mathbb{R}^d$, $Re\,(\int_0^t \beta_{kj}(x+a(\tau-t),\xi))<0$. Donc, en prenant $a=w^{kj}$, $|\mu_k^j(t,x)|$ croit exponentiellement en temps. Nous avons donc construit une solution approchée dont le terme d'ordre 0 croit exponentiellement en temps, ce qui contredit la stabilité.

□

9.3 Etude de cas particuliers

Les avantages du corollaire 9.1 sur le théorème 9.2 sont les suivants :

- les coefficients d'absorption ne sont pas constants mais variables,
- l'absorption peut être prise dans les deux directions,
- les valeurs propres non nulles du problème initial ne sont pas nécessairement de multiplicité 1.

9.3.1 Critère géométrique de Bécache, Fauqueux et Joly

Nous allons maintenant montrer que nous retrouvons bien la même condition nécessaire que [9], sous l'hypothèse que les valeurs propres non nulles du problème initial sont de multiplicité 1.

Dans cette partie, nous considérerons que $d=2$ et nous notons $x=x_1$ et $y=x_2$.

Nous supposons maintenant que $\forall j\neq 0$, $\alpha_j=1$ et que $\sigma_y=0$ et nous allons calculer les β définis précédemment.

Nous fixons $\lambda_0(\xi)$ valeur propre non nulle de $P(i\xi)$.

Soit $\lambda(\xi,z)$ une valeur propre de $\widetilde{P}(i\xi)-zB(x,y)$ telle que $\lambda(\xi,0)=\lambda_0(\xi)$.

Comme les valeurs propres non nulles de $\widetilde{P}(i\xi)$ sont de multiplicité 1, $\lambda(\xi,z)$ est analytique en z au voisinage de 0.

Or si $\widetilde{\Pi}(\xi,z)$ désigne la projection sur le sous-espace propre associé à $\lambda(\xi,z)$, $\widetilde{\Pi}(\xi,z)-zB(x,y)$ est aussi analytique au voisinage de 0 et nous avons :

$$\left(\widetilde{P}(i\xi)-zB(x,y)-\lambda(\xi,z)\right)\widetilde{\Pi}(\xi,z)=0.$$

En dérivant cette équation par rapport à z, nous obtenons :

$$\left(-B(x,y)-\frac{d\lambda}{dz}(\xi,z)\right)\widetilde{\Pi}(\xi,z)+\left(\widetilde{P}(i\xi)-zB(x,y)-\lambda(\xi,z)\right)\frac{d\widetilde{\Pi}}{dz}(\xi,z)=0.$$

En appliquant $\widetilde{\Pi}(\xi,z)$ à gauche de l'équation et en utilisant le lemme 9.2, nous obtenons :

$$\widetilde{\Pi}(\xi,z)\left(-B(x,y)-\frac{d\lambda}{dz}(\xi,z)\right)\widetilde{\Pi}(\xi,z)=0.$$

Donc :

$$\begin{aligned}\widetilde{\Pi}(\xi,z)B(x,y)\widetilde{\Pi}(\xi,z) &= -\widetilde{\Pi}(\xi,z)\frac{d\lambda}{dz}(\xi,z)\widetilde{\Pi}(\xi,z)\\ &= -\frac{d\lambda}{dz}(\xi,z)\widetilde{\Pi}(\xi,z).\end{aligned}$$

Nous prenons alors $z=0$ et nous obtenons :

$$\beta(x,y)=-\frac{d\lambda}{dz}(\xi,0).$$

De plus, nous avons l'équation :

$$det(\lambda(\xi,z)Id-\widetilde{P}(i\xi)+zB(x,y))=0.$$

Or,

$$\begin{aligned}det(\lambda(\xi,z)Id-\widetilde{P(i\xi)}+zB(x,y)) &= \begin{vmatrix}\lambda(\xi,z)Id-i\xi_1A_1+z\sigma_xId & -i\xi_1A_1\\ -i\xi_2A_2 & \lambda(\xi,z)Id-i\xi_2A_2\end{vmatrix}\\
&= \tfrac{(\lambda(\xi,z)+z\sigma_x)^N}{\lambda(\xi,z)^N}\begin{vmatrix}\lambda(\xi,z)Id-i\xi_1\frac{\lambda(\xi,z)}{\lambda(\xi,z)+z\sigma_x}A_1 & -i\xi_1\frac{\lambda(\xi,z)}{\lambda(\xi,z)+z\sigma_x}A_1\\ -i\xi_2A_2 & \lambda(\xi,z)Id-i\xi_2A_2\end{vmatrix}\\
&= \tfrac{(\lambda(\xi,z)+z\sigma_x)^N}{\lambda(\xi,z)^N}\\
&\quad\times\begin{vmatrix}\lambda(\xi,z)Id-i\xi_1\frac{\lambda(\xi,z)}{\lambda(\xi,z)+z\sigma_x}A_1-i\xi_2A_2 & \lambda(\xi,z)Id-i\xi_1\frac{\lambda(\xi,z)}{\lambda(\xi,z)+z\sigma_x}A_1-i\xi_2A_2\\ -i\xi_2A_2 & \lambda(\xi,z)Id-i\xi_2A_2\end{vmatrix}\\
&= \tfrac{(\lambda(\xi,z)+z\sigma_x)^N}{\lambda(\xi,z)^N}det(\lambda(\xi,z)Id-i\xi_1\tfrac{\lambda(\xi,z)}{\lambda(\xi,z)+z\sigma_x}A_1-i\xi_2A_2)\begin{vmatrix}Id & Id\\ -i\xi_2A_2 & \lambda(\xi,z)Id-i\xi_2A_2\end{vmatrix}\\
&= \tfrac{1}{\lambda(\xi,z)^N}det((\lambda(\xi,z)+z\sigma_x)\lambda(\xi,z)Id-i\xi_1\lambda(\xi,z)A_1-i\xi_2(\lambda(\xi,z)+z\sigma_x)A_2)\\
&\quad\times\begin{vmatrix}Id & O\\ -i\xi_2A_2 & \lambda(\xi,z)Id\end{vmatrix}\\
&= det((\lambda(\xi,z)+z\sigma_x)\lambda(\xi,z)Id-i\xi_1\lambda(\xi,z)A_1-i\xi_2(\lambda(\xi,z)+z\sigma_x)A_2)\\
&= det((\lambda(\xi,z)+z\sigma_x)\lambda(\xi,z)Id-P(i\xi_1\lambda(\xi,z),i\xi_2(\lambda(\xi,z)+z\sigma_x)).\end{aligned}$$

Ainsi, nous avons l'équation implicite :

$$\lambda(\xi,z)=\lambda_0\left(\xi_1\frac{\lambda(\xi,z)}{(\lambda(\xi,z)+z\sigma_x)},\xi_2\right).$$

Nous dérivons alors cette relation par rapport à z :

$$\frac{d\lambda}{dz}(\xi,z)=\xi_1\frac{\frac{d\lambda(\xi,z)}{dz}(\lambda(\xi,z)+z\sigma_x)-\lambda(\xi,z)(\frac{d\lambda(\xi,z)}{dz}+\sigma_x)}{(\lambda(\xi,z)+z\sigma_x)^2}\partial_{\xi_1}\lambda_0\left(\xi_1\frac{\lambda(\xi,z)}{(\lambda(\xi,z)+z\sigma_x)},\xi_2\right).$$

En prenant $z=0$, nous obtenons :

$$\frac{d\lambda}{dz}(\xi,0)=\xi_1\frac{\frac{d\lambda(\xi,0)}{dz}\lambda_0(\xi)-\lambda_0(\xi)(\frac{d\lambda(\xi,0)}{dz}+\sigma_x)}{\lambda_0(\xi)^2}\partial_{\xi_1}\lambda_0\left(\xi_1,\xi_2\right),$$

d'où :

$$\frac{d\lambda}{dz}(\xi,0)=-\xi_1\frac{\sigma_x}{\lambda_0(\xi)}\partial_{\xi_1}\lambda_0\left(\xi_1,\xi_2\right),$$

et

$$\beta(x,y)=\xi_1\frac{\sigma_x}{\lambda_0(\xi)}\partial_{\xi_1}\lambda_0\left(\xi_1,\xi_2\right).$$

Comme $\sigma_x>0$ est constant, la condition nécessaire de stabilité du corollaire 9.1 est équivalente à :

$$\frac{\xi_1}{\lambda_0(\xi)}\partial_{\xi_1}\lambda_0\left(\xi_1,\xi_2\right)\geq 0$$

qui est bien celle obtenue dans le théorème 9.2 de Bécache, Fauqueux et Joly.

9.3.2 Vitesse de groupe

Dans ce paragraphe, nous allons préciser, sous les mêmes hypothèses que dans le paragraphe précédent, les coefficients w_1 et w_2 intervenant dans les équations de transport (9.7) et faire apparaître la vitesse de groupe.

Avec les notations précédentes, nous avons :

$$\widetilde{\Pi}\widetilde{A_1}\widetilde{\Pi}=w_1\widetilde{\Pi}.$$

0r, d'après l'équation (9.4), nous avons :

$$\left(i\xi_1\widetilde{A_1}+i\xi_2\widetilde{A_2}-\lambda(\xi)Id\right)\widetilde{\Pi}=0.$$

Donc, en dérivant par rapport à ξ_1 :

$$\left(i\widetilde{A_1}-\frac{d\lambda(\xi)}{d\xi_1}Id\right)\widetilde{\Pi}+\left(i\xi_1\widetilde{A_1}+i\xi_2\widetilde{A_2}-\lambda(\xi)Id\right)\frac{d\widetilde{\Pi}}{d\xi_1}=0.$$

Et, en appliquant $\widetilde{\Pi}$, à gauche, nous obtenons :

$$iw_1=\frac{d\lambda(\xi)}{d\xi_1}.$$

Ainsi, nous avons $(w_1,w_2)=-i\nabla_\xi\lambda(\xi)=-V_g(\xi)$ où V_g désigne la vitesse de groupe.

9.4 Application aux équations de Maxwell PML de Bérenger

9.4.1 En dimension 2

Nous considérons les équations (6.1) homogénéisées, c'est-à-dire avec $\varepsilon_0 = \mu_0 = 1$ et $\sigma_x^* = \sigma_x$, $\sigma_y^* = \sigma_y$. Les valeurs propres de $P(i\xi)$, symbole des équations de Maxwell non PML, sont 0 et $\pm i\|\xi\|$ qui sont toutes de multiplicité 1. Nous calculons donc les projections $\widetilde{\Pi}_\pm$ sur les sous-espaces propres de $\widetilde{P}(i\xi)$ associés aux valeurs propres $\pm i\|\xi\|$. Nous obtenons alors :

$$\beta_+ = \beta_- = \frac{\sigma_x(x)\xi_1^2 + \sigma_y(y)\xi_1^2}{\|\xi\|^2}.$$

Donc, lorsque nous prenons des coefficients d'absorption positifs (ce qui est requis pour les PML), la condition nécessaire de stabilité est vérifiée.

9.4.2 En dimension 3

Nous considérons ici la version plus générale des équations (7.1). En effet, nous considérons que le coefficient d'absorption est dans les trois directions :

$$\begin{cases} \partial_t E_{xy} - \partial_y(H_{zx} + H_{zy}) + \sigma_y E_{xy} = 0 \\ \partial_t E_{xz} + \partial_z(H_{yz} + H_{yx}) + \sigma_z E_{xz} = 0 \\ \partial_t E_{yz} - \partial_z(H_{xy} + H_{xz}) + \sigma_z E_{yz} = 0 \\ \partial_t E_{yx} + \partial_x(H_{zx} + H_{zy}) + \sigma_x E_{yx} = 0 \\ \partial_t E_{zx} - \partial_x(H_{yz} + H_{yx}) + \sigma_x E_{zx} = 0 \\ \partial_t E_{zy} + \partial_y(H_{xy} + H_{xz})) + \sigma_y E_{zy} = 0 \\ \partial_t H_{xy} + \partial_y(E_{zx} + E_{zy}) + \sigma_y H_{xy} = 0 \\ \partial_t H_{xz} - \partial_z(E_{yz} + E_{yx}) + \sigma_z H_{xz} = 0 \\ \partial_t H_{yz} + \partial_z(E_{xy} + E_{xz}) + \sigma_z H_{yz} = 0 \\ \partial_t H_{yx} - \partial_x(E_{zx} + E_{zy}) + \sigma_x H_{yx} = 0 \\ \partial_t H_{zx} + \partial_x(E_{yz} + E_{yx}) + \sigma_x H_{zx} = 0 \\ \partial_t H_{zy} - \partial_y(E_{xy} + E_{xz}) + \sigma_y H_{zy} = 0. \end{cases}$$

Les valeurs propres de $P(i\xi)$ sont 0 et $\pm i\|\xi\|$ qui sont toutes de multiplicité 2.

Dans cet exemple, les hypothèses du théorème 9.2 ne sont plus vérifiées. Cela montre l'intérêt pratique du théorème 9.3.

Nous calculons donc les projections $\widetilde{\Pi}_{\pm 1,2}$ sur les sous-espaces engendrés par les vecteurs propres de $\widetilde{P}(i\xi)$ associés aux valeurs propres $\pm i\|\xi\|$. Nous obtenons alors :

$$\beta_{+1} = \beta_{+2} = \beta_{-1} = \beta_{-2} = \frac{\sigma_x(x)\xi_1^2 + \sigma_y(y)\xi_1^2 + \sigma_z(z)\xi_3^2}{\|\xi\|^2}.$$

Nous avons alors la même conclusion que dans le cas de la dimension 2.

9.5 Application aux équations d'Euler PML

Nous allons appliquer ce résultat aux équations d'Euler simplifiées dont la version PML a été proposée dans [20]. Les équations d'Euler sont de la forme :

$$\left\{\begin{array}{l} \partial_t u + M\partial_x u + \partial_x p = 0, \\ \partial_t v + M\partial_x v + \partial_y p = 0, \\ \partial_t p + \partial_x u + M\partial_x p + \partial_y v = 0, \end{array}\right.$$

où (u, v) désigne la vitesse, p la pression et M est le nombre de Mach.

Les équations PML sont alors :

$$\left\{\begin{array}{l} \partial_t u_1 + \partial_x(p_1 + p_2) + \sigma_x u_1 = 0, \\ \partial_t u_2 + M\partial_x(u_1 + u_2) + \sigma_x u_2 = 0, \\ \partial_t v_1 + \partial_y(p_1 + p_2) + \sigma_y v_1 = 0, \\ \partial_t v_2 + M\partial_x(v_1 + v_2) + \sigma_x v_2 = 0, \\ \partial_t p_1 + \partial_x(u_1 + u_2) + M\partial_x(p_1 + p_2) + \sigma_x p_1 = 0, \\ \partial_t p_2 + \partial_y(v_1 + v_2) + \sigma_y p_2 = 0. \end{array}\right. \tag{9.8}$$

Le symbole du problème initial est :

$$P(i\xi) = \begin{pmatrix} i\xi_1 M & 0 & i\xi_1 \\ 0 & i\xi_1 M & i\xi_2 \\ i\xi_1 & i\xi_2 & i\xi_1 M \end{pmatrix}.$$

Ces valeurs propres sont $i\xi_1 M$, $i(\xi_1 M + \|\xi\|)$ et $i(\xi_1 M - \|\xi\|)$.

Les hypothèses du théorème 9.3 ne sont vérifiées que pour $|M| < 1$. En effet, si $|M| \geq 1$, il existe $(\xi_1, \xi_2) \neq (0, 0)$ tel que $i(\xi_1 M + \|\xi\|) = 0$. Donc la valeur propre 0 n'est pas de multiplicité constante. Nous considérerons donc à partir de maintenant que $|M| < 1$.

Nous notons $\widetilde{\Pi}_0$ (resp. $\widetilde{\Pi}_\pm$) la projection sur le sous-espace engendré par le vecteur propre de $\widetilde{P}(i\xi)$ associé à $i\xi_1 M$ (resp. $i(\xi_1 M \pm \|\xi\|)$). Le calcul des β donne :

$$\beta_0 = \sigma_x \text{ et } \beta_\pm = \frac{(\xi_1^2\sigma_x(M^2-1) - \sigma_y\xi_2^2)\|\xi\| \pm \xi_1\xi_2^2 M(\sigma_y - \sigma_x)}{(\xi_1^2(M^2-1) - \xi_2^2)\|\xi\|}.$$

Nous allons étudier le signe de $\beta_\pm$ et en déduire la stabilité ou non du problème (9.8).

Nous posons :

$$\boxed{\begin{array}{c} a_\pm = \mp M(\sigma_y - \sigma_x), \qquad b = (\sigma_x(M^2-1) + \sigma_y), \qquad c = -\sigma_y \\ f_\pm(z) = a_\pm z^3 + bz^2 - a_\pm z + c \\ z_\pm = \frac{-b \pm \sqrt{3a^2 + b^2}}{3a} \end{array}}$$

Proposition 9.2 *Pour $|M| < 1$, s'il existe un point (x, y) tel que $\sigma_x(x) \neq \sigma_y(y)$, si toutes les conditions suivantes sont violées,*

(i) $\frac{b}{3|a_\pm|} \geq 1$ *et* $b > 0$,

(ii) $\left|\frac{b}{3a_\pm}\right| \geq 1$ *et* $b < 0$ *et* $a_\pm < 0$ *et* $f_\pm(z_-) < 0$,

(iii) $\left|\frac{b}{3a_\pm}\right| \geq 1$ *et* $b < 0$ *et* $a_\pm > 0$ *et* $f_\pm(z_+) < 0$,

(iv) $\left|\frac{b}{3a_\pm}\right| < 1$ *et* $a_\pm > 0$ *et* $f_\pm(z_-) < 0$,

(v) $\left|\frac{b}{3a_\pm}\right| < 1$ *et* $a_\pm < 0$ *et* $f_\pm(z_+) < 0$.

alors, le problème (9.8) est instable.

Remarque 9.1 *Si $\sigma_x = \sigma_y$, alors $\beta_\pm = \sigma_x$ donc la condition nécessaire de stabilité est vérifiée. Nous ne pouvons donc pas conclure.*

PREUVE : Nous remarquons que, pour $|M| < 1$, $\beta_\pm$ est du signe opposé à celui de $(\xi_1^2\sigma_x(M^2-1) - \sigma_y\xi_2^2)\|\xi\| \pm \xi_1\xi_2^2 M(\sigma_y - \sigma_x)$. En passant en coordonnées polaires, nous devons donc étudier le signe de la fonction :

$$\phi_\pm(r,\theta) = r^3[(\cos\theta)^2\sigma_x(M^2-1) - \sigma_y(\sin\theta)^2 \pm \cos\theta(\sin\theta)^2 M(\sigma_y - \sigma_x)] = r^3 f_\pm(z),$$

où nous avons posé $z = \cos(\theta)$. Une étude de fonction montre que $f_\pm$ est négative sur $[-1, 1]$ si et seulement si l'une des conditions suivantes est vérifiée :

- $\left|\frac{b}{3a}\right| \geq 1$ et $b > 0$,
- $\frac{b}{3a} \geq 1$ et $b < 0$ et $a < 0$ et $f(z_-) < 0$,
- $\frac{b}{3a} \geq 1$ et $b < 0$ et $a > 0$ et $f(z_+) < 0$,
- $\frac{b}{3a} < 1$ et $a > 0$ et $f(z_-) < 0$,
- $\frac{b}{3a} < 1$ et $a < 0$ et $f(z_+) < 0$.

Donc si les cinq conditions précédentes sont violées le problème est instable.

□

Le cas $|M| \geq 1$ n'est pas traité par notre construction, nous travaillons à une généralisation.

Conclusion

Dans cette partie, nous avons étudié les équations PML proposées par Bérenger. Nous avons donné des estimations d'énergie sans perte de régularité valables pour des coefficients variables dans les deux directions pour les équations de Maxwell en dimension 2. Nous avons généralisé ces estimations aux équations de Maxwell en dimension 3 ainsi qu'au schéma de Yee. Pour le calcul de ces estimations, nous ne nous sommes pas intéressés au comportement en temps mais uniquement à la perte de régularité.

Nous nous sommes ensuite concentrés sur le comportement en temps des solutions de équations PML. Nous avons effectué un développement asymptotique des solutions grâce à l'approximation de l'optique géométrique, ce qui nous a conduit à une condition nécessaire de stabilité ayant des hypothèses très larges. En effet, cette condition est valable pour des problèmes PML pour lesquels l'absorption est à coefficients variables, cette absorption peut être dans les deux directions. De plus nos hypothèses sur les valeurs propres du symbole s'appliquent encore aux équations de Maxwell en dimension 3.

Une perspective de ce travail est la généralisation des estimations d'énergie à d'autres types de problèmes notamment à l'élastodynamique. De plus, nous avons toujours considéré un domaine infini et nous n'avons pas étudié l'influence de la réflexion créée au bord du domaine. En effet, malgré l'introduction des couches PML, il faut tout de même se donner une condition aux limites au bord des couches introduites artificiellement. Nous savons que les couches PML qui entourent le domaine d'intérêt absorbent exponentiellement les ondes mais il faudrait étudier plus précisément l'influence de la réflexion au bord des couches.

Bibliographie

[1] Abarbanel Saul et Gottlieb David
A mathematical analysis of the PML method,
J. Comput. Phys., 134, 357-363, 1997.

[2] Abarbanel Saul et Gottlieb David,
On the construction and analysis of absorbing layers in CEM,
Appl. Numer. Math., 27, 331-340, 1998.

[3] Abarbanel S., Gottlieb D. et Hesthaven J. S.,
Well-posed perfectly matched layers for advective acoustics,
J. Comput. Phys., 154, 266-283, 1999.

[4] Abarbanel S., Gottlieb D. et Hesthaven J. S.,
Long time behavior of the perfectly matched layer equations in computational electromagnetics,
J. Sci. Comput., 17, 405-422, 2002.

[5] Abarbanel S., Gottlieb D. et Hesthaven J. S.,
Non-Linear PML Equations for Time Dependent Electromagnetics in Three Dimensions
J. Sci. Comput., 28, 125-137, 2006.

[6] Appelö D., Hagstrom T. et Kreiss G.,
Perfectly matched layers for hyperbolic systems : General formulation, well-posedness and stability ,
preprint.

[7] Arnold V. I.,
Lectures on bifurcations and versal families ,
Uspehi Mat. Nauk, 27, 1972, 119-184.

[8] Baumgärtel H.
Analytic perturbation theory for matrices and operators,
Birkhäuser Verlag, 1985.

[9] Bécache E, Fauqueux S,et Joly P,
Stability of perfectly matched layers, group velocities and anisotropic waves,
J. Comput. Phys., 188, 399-433, 2003.

[10] Bécache Eliane et Joly Patrick, *On the analysis of Bérenger's perfectly matched layers for Maxwell's equations*,
M2AN Math. Model. Numer. Anal., 36, 87-119, 2002.

[11] Bécache Eliane, Petropoulos Peter G. et Gedney Stephen D.,
On the long-time behavior of unsplit perfectly matched layers,
IEEE Trans. Antennas and Propagation, 52, 1335-1342, 2004.

[12] Berenger Jean-Pierre
A perfectly matched layer for the absorption of electromagnetic waves ,
J. Comput. Phys., 114, 185-200 , 1994.

[13] Berenger Jean-Pierre,
Three-dimensional perfectly matched layer for the absorption of electromagnetic waves,
J. Comput. Phys., 127, 363-379, 1996.

[14] Bierstone Edward et Milman Pierre,
Semianalytic and subanalytic sets,
Publications mathématiques de l'IHES, tome 67, 5-42, 1988.

[15] Chazarain Jacques et Piriou Alain,
Introduction à la théorie des équations aux dérivées partielles linéaires,
Gauthier-Villars, 1981.

[16] Collino Francis et Monk Peter,
The perfectly matched layer in curvilinear coordinates,
SIAM J. Sci. Comput.,2061-2090, 1998.

[17] Crouzeix Michel et Mignot Alain L,
Analyse numérique des équations différentielles,
Masson, 1984.

[18] Gårding Lars,
Linear hyperbolic partial differential equations with constant coefficients,
Acta Math., 85, 1-62, 1951.

[19] Gustafsson Bertil, Kreiss Heinz-Otto et Oliger Joseph,
Time dependent problems and difference methods,
Pure and Applied Mathematics (New York), 1995.

[20] Hesthaven J. S.,
On the analysis and construction of perfectly matched layers for the linearized Euler equations,
J. Comput. Phys., 142, 129-147, 1998.

[21] Hu Fang Q., *On absorbing boundary conditions for linearized Euler equations by a perfectly matched layer*,
J. Comput. Phys., 129, 201-219, 1996.

[22] Hu Fang Q.
A stable, perfectly matched layer for linearized Euler equations in unsplit physical variables,
J. Comput. Phys., 173, 455-480, 2001.

[23] Hu Fang Q.
Development of PML Absorbing Boundary Condition for Computational Aeroacoustics : A Progress Review ,
Proceedings of Euromech Colloquium 467, to appear in a special issue of Computers and Fluids.

[24] Kato Tosio,
Perturbation theory for linear operators,
Springer-Verlag New York, Inc., New York, 1966.

[25] Knopp Konrad
Theory of Functions. II. Applications and Continuation of the General Theory,
Dover Publications, New York, 1947.

[26] Kreiss Heinz-Otto et Busenhart Hedwig-Ulmer,
Time-dependent partial differential equations and their numerical solution,
Birkhäuser Verlag, 2001.

[27] Kreiss Heinz-Otto et Lorenz Jens,
Initial-boundary value problems and the Navier-Stokes equations,
Academic Press Inc., 1989.

[28] Lang Serge
Algebra,
Graduate Texts in Mathematics,2002.

[29] Lax P. D. et Richtmyer R. D.,
Survey of the stability of linear finite difference equations,
Comm. Pure Appl. Math., 267-293, 1956.

[30] Lions J.L. et Magenes E.
Problèmes aux limites non homogènes et applications. Vol 1
Dunod, 1968.

[31] Lions Jacques-Louis, Métral Jérôme et Vacus Olivier,
Well-posed absorbing layer for hyperbolic problems,
Numer. Math., 92, 535-562, 2002.

[32] Métral Jérôme et Vacus Olivier,
Caractère bien posé du problème de Cauchy pour le système de Bérenger,
C. R. Acad. Sci. Paris Sér. I Math., 10, 847–852, 1999.

[33] Petropoulos Peter G., Zhao Li et Cangellaris Andreas C., *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, 184-208, 1998.

[34] Rahmouni Adib,
Un modèle PML bien posé pour les équations d'Euler linéarisées,
C. R. Acad. Sci. Paris Sér. I Math., 331, 159-164, 2000.

[35] Rahmouni Adib N., *An algebraic method to develop well-posed PML models. Absorbing layers, perfectly matched layers, linearized Euler equations*,
J. Comput. Phys., 197, 99-115, 2004.

[36] Rauch Jeffrey,
Lectures on Geometric Optics,
http ://www.math.lsa.umich.edu/ rauch/myresearch.html.

[37] Richtmyer Robert D. et Morton, K. W.,
Difference methods for initial-value problems,
Interscience Publishers John Wiley & Sons, Inc., New York-London-Sydney, 1967.

[38] Strikwerda John C, *Finite difference schemes and partial differential equations*,
Wadsworth & Brooks/Cole Advanced Books & Software, 1989.

[39] Taflov A.
Computational electrodynamics, the finite difference time domain approach,
Artech House, 1995.

[40] Taylor Michael E.,
Pseudodifferential operators ,
Princeton University Press, 1981.

[41] Yee K.S.,
Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media,
IEEE Trans. Antennas and Propagation, 302-307, 1966.

Problèmes faiblement bien posés : discrétisation et applications.

Résumé :
Dans cette thèse, nous nous intéressons à la discrétisation par des schémas aux différences finies de problèmes faiblement bien posés. Nous donnons de nouvelles définitions qui prennent en compte la perte de régularité apparaissant dans les problèmes faiblement bien posés et nous étendons la condition nécessaire et suffisante de convergence de Lax-Richtmyer. Nous définissons une classe de schémas pour laquelle nous calculons le taux de convergence optimal. Ces calculs reposent sur la théorie des perturbations et le développement en série de Puiseux. Nous illustrons numériquement nos résultats.
Dans un deuxième temps, nous nous intéressons à un cas particulier de problèmes faiblement bien posés : les couches parfaitement adaptées de Bérenger ou PML. Nous donnons des estimations d'énergie pour les équations de Maxwell que nous étendons au schéma de Yee. Enfin, nous étudions le comportement asymptotique en temps de la solution d'une équation PML en utilisant l'approximation de l'optique géométrique.

Weakly well posed problems : discretization and applications.

Abstract :
In this work, we study the discretization by finite difference schemes of weakly well posed problems. We draw new definitions treating the loss of regularity appearing in weakly well posed problems et we extand the necessary and sufficiency convergence condition of Lax-Richtmyer. Using perturbation theory and Puiseux series development, we evaluate the convergence rate of schemes belonging to a particular class. We give numerical results.
Secondly, we study a particular case of weakly well posed problems : the perfectly matched layers of Bérenger. We give energy estimates for Maxwell's equations and their extension to Yee scheme. Finally, we designe the asymptotic behaviour of the solution to a PML equation using the geometric optic approximation.

Discipline : Mathématiques appliquées

Mots clés : Problèmes faiblement bien posés, schémas aux différences finies, perfectly matched layers.

LAGA - Institut Galilée - Université Paris 13
99, avenue J.B. Clément - 93430 Villetaneuse