version du 3 decembre 2004

Thèse

présentée pour obtenir le grade de docteur

de l'Université Louis Pasteur Strasbourg 1

Spécialité : mathématiques appliquées

Michel Mehrenberger

Inégalités d'observabilité et résolution adaptative de l'équation de Vlasov par éléments finis hiérarchiques

Date de soutenance : 15 décembre 2004

Directeur de thèse :	Vilmos KOMORNIK	Professeur à l'Université Louis Pasteur, Strasbourg.
Codirecteur de thèse :	Eric SONNENDRÜCKER	Professeur à l'Université Louis Pasteur, Strasbourg.
Rapporteur Externe :	Albert COHEN	Professeur à l'Université Pierre et Marie Curie, Paris.
Rapporteur Externe :	Alain HARAUX	Professeur à l'Université Pierre et Marie Curie, Paris.
Rapporteur Interne :	Bopeng RAO	Professeur à l'Université Louis Pasteur, Strasbourg.
Examinatrice :	Paola LORETI	Professeur à l'Université La Sapienza, Roma.

Université Louis Pasteur Strasbourg 1

A mon père,

Remerciements

Je remercie vivement Vilmos Komornik pour m'avoir initié à la recherche, pour ses conseils, son aide, sa passion pour les preuves élégantes et courtes, sa facon de voir les maths que j'aime beaucoup, sa culture, toutes les discussions, pour toutes les histoires sur les mathématiciens... Je remercie ensuite Eric Sonnendrücker pour la disponibilité qu'il m'a accordee, pour m'avoir ouvert à un nouveau champ de recherche, pour m'avoir permis de participer au CEMRACS'03, et à bien d'autres manifestations scientifiques, pour toute l'energie qu'il investit dans les projets, la recherche, le rassemblement entre différentes communautés (physiciens, mathématiciens et informaticiens). Je remercie Albert Cohen pour avoir accepté de rapporter en détail sur une grande partie de ma thèse, pour la précision et l'importance de ses remarques, pour l'intérêt qu'il a témoigné pour mon travail. Je remercie également Alain Haraux d'avoir accepté d'être rapporteur et pour ses remarques très précises. Tous mes remerciements à Bopeng Rao pour des discussions intéressantes et pour avoir accepté d'être rapporteur. Je tiens fortement à remercier Paola Loreti, pour avoir accepté d'être membre du jury et pour sa passion pour la recherche.
Je remercie ensuite Nicolas, d'abord pour son travail, qui m'a servi souvent et qui m'a ouvert à des questions qui m'ont passionnées. Je le remercie pour la collaboration et aussi pour les moments de discussion. Je ne saurais assez remercier Martin, pour tout ce que j'ai appris de lui, et à qui je dois beaucoup, pour avoir pu travailler avec lui depuis le CEMRACS'03. Son efficacité, ses compétences, la clarté de ses explications, sa disponibilité m'ont beaucoup aidé. Son expérience, ses idées et son travail ont grandement contribué au développement et à la concrétisation de cette recherche faite en collaboration. Je le remercie aussi pour ses conseils et sa sensibilité. Je remercie vivement Eric Violard et Olivier de s'être intéressés au code adaptatif, de l'avoir parallélisé, et aussi des discussions passionnantes que nous avons pu avoir á ce sujet. Je suis très heureux d'avoir pu être en contact ainsi avec des informaticiens. J'en profite á ce moment pour remercier Mathieu de m'avoir transmis une petite partie de sa passion pour l'informatique. Toute ma reconnaissance à Régine, pour les maths échangées de tous les jours, pour tous les services rendus, son agréable compagnie au bureau 408. Un grand merci à Michaël pour son aide pour les présentations et ses divers conseils, et à Pierre pour tous les dépannages, à Vincent, Stéphanie,...,Etienne,... Je remercie l'IRMA, CALVI, Mme Borell... Je tenais à citer : Michèle Audin, Jean-Yves, Robert. Merci à ma mère pour m'avoir encouragé, à Marie-Elise, pour tout son amour et pour m'avoir supporté durant ce temps, à Blandine, Jo et à toute ma famille.

Résumé

Cette thèse porte d'une part sur l'observabilité de systèmes linéaires faiblement couplés et sur des inégalités de type Parseval et d'autre part sur la convergence de schémas numériques semi-lagrangiens d'ordre élevé, et sur des aspects théoriques et numériques de méthodes adaptatives pour la résolution numérique de l'équation de Vlasov basées sur sur une décomposition de l'espace des phases en cellules.

Observabilité de systèmes couplés. On se donne un problème d'évolution

$$x' = Ax, \quad x(0) = x_0,$$

où A est un opérateur linéaire dans un espace de Hilbert (complexe et séparable). L'observation du système se traduit ici par le fait que la formule

$$P(x_0) := \sqrt{\sum_{j=1}^{m} \int_{I_j} p_j(x(t))^2 \mathrm{d}t}$$

définit une norme équivalente à celle sur H pour la condition initiale. Ici et pour ce qui suit, les I_j sont des intervalles non dégénérés, de longueur finie, et les p_j sont des semi-normes. On parlera aussi d'opérateurs, d'espaces ou de vecteurs observables.
On introduit les *vecteurs propres généralisés* $\{E_{k,\ell}, k \in \mathbb{Z}, \ell = 1, \dots, m_k\}$ solutions de :

$$AE_{k,\ell} = \lambda_k E_{k,\ell} + E_{k,\ell-1},$$

où les complexes (λ_k) sont les valeurs propres de multiplicité m_k.
Des théorèmes d'observabilité sont souvent montrés en supposant que A admet une base de Riesz de vecteurs propres généralisés. Il se trouve que pour certaines applications, cette hypothèse est trop contraignante (par exemple, il se peut que des vecteurs propres se rejoignent à l'infini) ou difficile à montrer, tandis qu'elle est souvent assurée pour un opérateur $\tilde{A}$ qui est compactement proche de A (c'est-à-dire que A s'écrit $\tilde{A} + B$ où B est un opérateur compact). Le théorème suivant permet d'inclure de tels cas.

Théorème 1. Supposons que

- A est compactement proche d'un opérateur qui admet une base de Riesz de vecteurs propres et dont les valeurs propres tendent vers l'infini, sont de partie réelle et de multiplicité bornées,
- A admet une base de Riesz au sens des sous-espaces (e_k),

• A est compactement proche d'un opérateur observable (au sens ci-dessus) sur un espace de codimension finie engendré par certains e_k.
Alors l'espace engendré par les vecteurs e_k observables forme un espace de codimension finie observable pour tous les intervalles J_j de longueur strictement plus grande que I_j.

On retrouve en cas particulier un théorème de Komornik et Loreti en enlevant le terme "compactement proche".
On applique ensuite ce théorème au système suivant :

$$\begin{cases} u_i'' = a_i^2 \Delta u_i - \sum_{j=1}^{m+n} \alpha_{i,j} u_j & \text{dans } \mathbb{R} \times \Omega, i = 1, \dots m, \\ u_i'' = -a_i^2 \Delta^2 u_i - \sum_{j=1}^{m+n} \alpha_{i,j} u_j & \text{dans } \mathbb{R} \times \Omega, i = m+1, \dots, m+n, \\ u_i = 0 & \text{sur } \mathbb{R} \times \Gamma_i, i = 1 \dots m, \\ u_i = \Delta u_i = 0 & \text{sur } \mathbb{R} \times \Gamma_i, i = m+1, \dots m+n, \\ u_i(0) = u_{i0}, u_i'(0) = u_{i1}, & \text{dans } \Omega, i = 1 \dots m+n, \end{cases}$$

avec $\Omega \subset \mathbb{R}^N$ un ouvert non vide de frontière Γ, $a_i \in \mathbb{R}^{+*}$ et $\alpha_{i,j} \in \mathbb{C}$.
Dans le cas d'une boule, et sous des hypothèses classiques sur les domaines d'observation $\Gamma_i \subset \Gamma$, on obtient alors en particulier la

Proposition 2. Les paramètres $(\alpha_{i,j})$ pour lesquels il n'y a pas observabilité frontière (sur tout l'espace) contiennent et sont contenus dans une infinité dénombrable de surfaces de codimension $n+m$.

L'observabilité frontière est donnée ici par l'équivalence entre

$$\sum_{i=1}^{m+n} \int_{I_i} \int_{\Gamma_i} |\partial_\nu u_i|^2 \, d\Gamma dt,$$

et la norme naturelle sur $H_0^1(\Omega)^m \times L^2(\Omega)^m \times H_0^1(\Omega)^n \times H^{-1}(\Omega)^n$.
Ces résultats font partie d'un article paru.

Inégalités de type Parseval. On se donne une suite de réels (λ_n) et on définit, suivant Pólya, sa densité supérieure par la formule

$$D^+ = \lim_{n \to \infty} \frac{n^+(r)}{r},$$

où $n^+(r)$ est le nombre maximum d'occurences de la suite dans un intervalle de longueur r. Dans des problèmes d'observabilité, on demande souvent que la suite d'exponentielles $e^{i\lambda_n t}$ soit de Riesz, c'est-à-dire que l'on ait des estimations du type :

$$c \sum |a_n|^2 \leq \int_I \left| \sum a_n e^{i\lambda_n t} \right|^2 \leq C \sum |a_n|^2,$$

où les constantes c, C sont indépendantes de la suite (a_n). Pour inclure des cas d'exposants λ_n proches, on peut construire une suite de différences divisées. Dans ce cadre, Baiocchi, Komornik et Loreti ont alors obtenu la propriété de suite de Riesz pour les différences divisées sous la condition que l'intervalle I soit de longueur strictement supérieure à $2\pi D^+$.
Le théorème suivant montre l'optimalité de la borne :

Théorème 3. Si la longueur de l'intervalle I est strictement inférieure à $2\pi D^+$, alors la suite de différences divisées associée à $e^{i\lambda_n t}$ ne forme pas une suite de Riesz.

Le théorème fait partie d'un article accepté pour publication.

Convergence de schémas semi-lagrangiens d'ordre élevés. On établit la convergence de certaines classes de schémas semi-lagrangiens d'ordre quelconque pour la résolution numérique d'un système de Vlasov Poisson à une dimension (x, v) avec conditions limites périodiques en x dans le cas (uniforme) de grilles cartésiennes. Au cours de la démonstration, on obtient aussi de nouvelles preuves de deux inégalités d'interpolation dues à Strang (1962) et De Boor (1976) :

Théorème 4. (Reconstruction par polynôme de Lagrange) Soit $n \in \mathbb{N}^*$, $\theta \in \mathbb{R}$ et P l'unique polynôme de degré $2n$ interpolant la fonction $\exp(i\theta\cdot)$ en les points $-n, \ldots, n$. Alors $|P| \leq 1$, sur $[-1, 1]$.

Théorème 5. (Reconstruction par splines) Soit $m \in \mathbb{N}$, $\theta \in \mathbb{R}$ et B_m la B-spline d'ordre m définie par convolution de la fonction indicatrice de l'intervalle $[-\frac{1}{2}, \frac{1}{2}]$. Alors

$$\Phi_m(\alpha) := \left| \sum_{k \in \mathbb{Z}} B_m(k + \alpha) e^{ik\theta} \right|^2$$

admet son maximum en les entiers.

Cette partie a été établie en collaboration avec Nicolas Besse.

Résolution numérique adaptative de l'équation de Vlasov. On présente ici une nouvelle méthode de résolution de l'équation de Vlasov, basée sur une décomposition hiérarchique de l'espace des phases en utilisant des éléments biquadratiques. Des cas tests classiques comme l'amortissement de Landau et des faisceaux semi-gaussiens valident le code implémenté pour l'instant en $1D$, mais la méthode se généralise facilement pour les dimensions supérieures. Généralement les codes numériques adaptatifs génèrent une structure complexe de la localisation des données. La découpe de l'espace des phases en mailles dyadiques permet ici une parallélisation efficace. Des améliorations devront être effectuées pour rendre le code conservatif et pour pouvoir passer à des degrés plus élevés (d'après l'étude précédente, le schéma n'est pas stable pour des degrés plus élevés).

Convergence d'un schéma adaptatif. Suite à ce précédent travail, on cherche à trouver un cadre (plus simplifié) pour prouver mathématiquement la convergence d'un schéma adaptatif basé sur une découpe en cellules. Sous les hypothèses de :

- splitting d'opérateurs (découpe de l'advection en advection horizontale et verticale),
- reconstruction affine par maille (des triangles obtenus en découpant les carrés en deux),

on obtient un critère, qui permet de contrôler les erreurs produites à chaque pas de temps. Les 2 parties précédentes sont issus d'une collaboration avec Martin Campos Pinto.

Publications :

- *Observability of coupled systems*, Acta Math. Hungar. 103 (4) (2004), 321–348.
- *Critical length for a Beurling type theorem*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat., à paraître.
- *Convergence of classes of high order semi-Lagrangian schemes for the Vlasov equation*, avec Nicolas Besse, soumis.

- *Adaptive Numerical Resolution of the Vlasov Equation*, rapport du CEMRACS 2003, avec Martin Campos Pinto, à paraître.
- *Convergence of an adaptive scheme for the one-dimensional Vlasov-Poisson system*, avec Martin Campos Pinto, en préparation.

Table des matières

Chapitre 1

Introduction à des problèmes d'observabilité

Etant donnée une équation aux dérivées partielles (ou un système), les questions que l'on se pose souvent sont : existence de solutions, unicité, régularité et approximation numérique ; on s'intéresse aussi parfois à leur *contrôle*.
Ainsi, la *contrôlabilité exacte* répond à la question suivante : peut-on agir sur un système au moyen d'une force (*le contrôle*) de telle sorte que la solution de ce système soit la solution nulle (ou un état stationnaire) en un temps T donné ? (Si oui, on aura contrôlé le système.)
Une autre question est la *stabilisabilité* : le système peut-il tendre vers zéro (et de quelle manière) sous l'effet d'un certain terme d'amortissement ?
Enfin, on se pose aussi parfois la question de l'*observabilité* : étant donnée une certaine quantité (*l'observation*), peut-on distinguer deux solutions de conditions initiales différentes si on observe juste cette quantité pendant un certain temps ?
Il se trouve que ces trois concepts sont étroitement liés. Le rapport entre l'observabilité et la contrôlabilité est exploité par la célèbre méthode HUM (*Hilbert Uniqueness Method*) de J.-L. Lions, exposée dans un article de revue en 1988 [36] et dans son monographe [35]. De même, on peut lier de manière générale l'observabilité à la stabilisabilité, grâce à *la méthode de stabilisation rapide* de Komornik [28].
Ainsi, toutes ces questions se ramènent souvent à l'établissement de l'observabilité, plus précisément d'*inégalités d'observabilité*.
La *méthode des multiplicateurs*, initialement utilisée dans ce contexte par Ho [22], puis grandement développée par Lions (voir aussi le livre [27]) a permis de répondre à beaucoup de problèmes.
Une autre approche qui a eu un regain d'intérêt ces derniers temps est la *méthode d'analyse non harmonique*. Des résultats plus précis ont pu être établis dans certains cas particuliers. C'est ce point de vue que nous adopterons ici. La méthode est basée sur un développement de la solution en série de Fourier (dans un sens généralisé).
Ainsi l'identité de Parseval (valable pour les vraies séries de Fourier) va pouvoir se trans-

former en une double inégalité dans des cas plus généraux, qui correspond justement aux inégalités d'observabilité désirées. Dans tout ce qui suit, une des questions omniprésentes est de trouver des conditions pour pouvoir établir ces généralisations de l'égalité de Parseval.
Il existe d'autres méthodes pour traiter les problèmes d'observabilité : par exemple, l'analyse micro-locale (qui donne ainsi des résultats plus précis et presque optimaux pour l'équation des ondes [7]) et les estimations de Carleman (très utilisées actuellement, notamment pour traiter des problèmes paraboliques ; pour une référence récente : [40]).
Nous nous limitons à l'étude des équations réversibles en temps (comme l'équation des ondes et non comme l'équation de la chaleur). La plupart des exemples qui vont suivre sont des problèmes mono-dimensionnels, le cas de questions intervenant pour des dimensions plus élevées est mentionné à la fin.
Pour faciliter la lecture, nous allons aussi rappeler plusieurs notions et résultats classiques de la théorie du contrôle, afin de mieux indiquer la place de nos résultats dans ce domaine.

1.1 L'observabilité de l'équation des ondes en dimension 1

On considère le système des ondes suivant :

$$\begin{cases} u'' - \Delta u = 0 & \text{dans } \mathbb{R} \times \Omega = (0,\pi), \\ u|_\Gamma = 0 & \text{sur } \mathbb{R} \times \Gamma = \{0,\pi\}, \\ u(0) = u_0, u'(0) = u_1 & \text{dans } \Omega. \end{cases}$$

E_0 est l'énergie initiale, c'est-à-dire le carré de la norme usuelle de (u_0, u_1) dans $H_0^1(\Omega) \times L^2(\Omega)$. On se demande alors si on peut avoir des inégalités du type

$$cE_0 \leq \int_0^T \int_\Gamma |\partial_\nu u|^2 \mathrm{d}\Gamma \mathrm{d}t \leq c' E_0, \tag{1.1.1}$$

où $c, c' > 0$ sont indépendants des données initiales. Pour simplifier, on écrira

$$\int_0^T \int_\Gamma |\partial_\nu u|^2 \mathrm{d}\Gamma \mathrm{d}t \asymp E_0.$$

Ces inégalités sont appelées *inégalités d'observabilité.*

Remarque 1.1.1. *En fait, dans un langage strict, seule l'inégalité de gauche est appelée inégalité d'observabilité ou* inégalité indirecte *(c'est celle-ci qui permet d'établir un résultat d'unicité), l'inégalité de droite traduit plutôt une sorte de continuité de l'observation par rapport à la norme définie par l'énergie, et est souvent appelée* inégalité de régularité cachée *ou* inégalité directe. *Dans la suite, on parlera (par abus) d'inégalités d'observabilité directe ou indirecte.*

Pour illustration, rappelons le résultat simple suivant :

Théorème 1.1.1. *Si $T > 2\pi$, alors (1.1.1) est valable.*

Preuve. La solution s'écrit

$$u(t,x) = \sum \alpha_k e^{i\gamma_k t} z_k, \tag{1.1.2}$$

où (z_k) est une base orthonormée de l'opérateur de Laplacien-Dirichlet :

$$\begin{cases} -\Delta z_k = \gamma_k^2 z_k & \text{sur } \Omega, \\ z_k|_\Gamma = 0, \\ 0 < \gamma_1 \leq \gamma_2 \cdots \to \infty. \end{cases}$$

On s'intéresse ensuite à des estimations du type :

$$\sum \gamma_k^2 |\alpha_k|^2 \asymp \int_0^T \int_\Gamma \Big| \sum \alpha_k e^{i\gamma_k t} \partial_\nu z_k \Big|^2 \mathrm{d}\Gamma \mathrm{d}t. \tag{1.1.3}$$

Dans cet exemple, nous avons :

$$z_k = \sin kx, \quad \partial_\nu z_k(0) = k, \quad \partial_\nu z_k(\pi) = (-1)^k k.$$

- Si $T = 2\pi$, il s'agit de l'égalité de Parseval.
- Si $T > 2\pi$, les inégalités subsistent.
- L'inégalité *directe* (à droite) est valable pour tout temps $T > 0$.

□

On introduit maintenant quelques notations :

$$\Lambda = (\lambda_n),$$
$$\mathcal{E}(\Lambda) = (e^{i\lambda_n \cdot}).$$

Lorsque l'on ne précise pas, les éléments d'une suite Λ sont tous distincts, et l'ensemble de ses éléments sera aussi noté Λ. En regardant la preuve, et afin de pouvoir avoir d'autres résultats, on définit une *suite de Riesz* :

Définition. *$\mathcal{E}(\Lambda)$ est une suite de Riesz dans $L^2(I)$, si*

$$\int_I \Big| \sum_{k \in K} \alpha_k e^{i\lambda_k t} \Big|^2 \mathrm{d}t \asymp \sum_{k \in K} |\alpha_k|^2.$$

Ainsi on peut se demander pour quelles conditions sur $|I|$ et Λ a-t-on une suite de Riesz ? On obtient facilement que, pour $a \in \mathbb{R}^+\backslash\{0\}$, les conditions :

$$\Lambda \subset a\mathbb{Z} \text{ et } |I| > \frac{2\pi}{a}$$

sont suffisantes.

Désormais, K désignera toujours un ensemble infini dénombrable.

1.2 Un résultat classique d'Ingham

Ingham ([23]), en 1936, a établi une première réponse qui permet déjà de traiter beaucoup de problèmes. Dans le contexte de la thérie du contrôle, ce résultat a été utilisé pour la première fois par Ball et Slemrod [6] (on pourra aussi consulter le livre de Zygmund régulièrement réédité auquels ces auteurs font référence [46]).

Théorème 1.2.1. *Supposons qu'il existe $\gamma_1 > 0$ tel que $\lambda_{n+1} - \lambda_n \geq \gamma_1$ pour tout n (condition de* gap*). Alors $\mathcal{E}(\Lambda)$ est une suite de Riesz dans $L^2(I)$ pour chaque intervalle borné I, avec $|I| > \frac{2\pi}{\gamma_1}$.*

- Ingham a aussi donné un contrexemple si $|I| = \frac{2\pi}{\gamma_1}$:

$$\gamma_1 = 1 \quad \text{et} \quad \Lambda = \{m + \frac{3}{4}, m \in \mathbb{N}\}.$$

- L'écart (*gap*) est nécessaire, d'après ce qui va suivre.

Suites uniformément discrètes : on dit que Λ est *uniformément discret* s'il existe un nombre $\delta > 0$, tel que

$$|\lambda_n - \lambda_m| \geq \delta, \quad \text{pour tous} \quad n \neq m.$$

Remarque 1.2.1. *La définition est valable pour tout $\Lambda \subset \mathbb{C}$. Néanmoins, on considère le cas réel. Dans le cas complexe, on supposera toujours que Λ est contenue dans une bande parallèle à l'axe réel.*

Proposition 1.2.2. *Supposons que $\mathcal{E}(\Lambda)$ est une suite de Riesz dans $L^2(I)$, alors Λ est uniformément discret.*

Preuve. On a :

$$2c \leq \int_I |1 \cdot e^{i\lambda_n t} - 1 \cdot e^{i\lambda_m t}|^2 \mathrm{d}t \leq |\lambda_n - \lambda_m|^2 \int_I t^2 \mathrm{d}t.$$

□

Cependant, la longueur critique ne peut pas s'exprimer seulement en termes de γ_1. On a par exemple le résultat suivant de A. Haraux [19] :

Proposition 1.2.3. *Supposons que $\lambda_{n+1} - \lambda_n \geq \gamma_0 > \gamma_1$, pour tout $n \in \mathbb{Z}$ excepté a un nombre fini pour lesquels*

$$\gamma_1 < \lambda_{n+1} - \lambda_n \leq \gamma_0.$$

Alors la condition $|I| > \frac{2\pi}{\gamma_0}$ convient, qui est meilleure que $|I| > \frac{2\pi}{\gamma_1}$.

1.3 Un théorème de Beurling

On introduit ici une nouvelle quantité afin d'obtenir une condition optimale : la *densité supérieure* d'une suite (λ_n). Elle est définie par la formule

$$D^+ = \lim_{r\to\infty} \frac{n^+(r)}{r},$$

où $n^+(r)$ est égal au plus grand nombre de termes de $\Re(\lambda_n)$ dans un intervalle de longueur r.

Exemple.

$$\Lambda = \alpha\mathbb{Z}, \quad D^+ = \frac{1}{\alpha}.$$

On vient juste de voir que $\mathcal{E}(\Lambda)$ est une suite de Riesz si $\alpha = \frac{2\pi}{|I|}$, i.e.

$$|I| = 2\pi|D^+|.$$

Théorème 1.3.1. *(Beurling [11])*

- *Supposons que $|I| > 2\pi D^+$ et que Λ est uniformément discrète. Alors $\mathcal{E}(\Lambda)$ est une suite de Riesz.*

- *Si $\mathcal{E}(\Lambda)$ est une suite de Riesz alors $|I| \geq 2\pi D^+$ et Λ est uniformément discrète.*

Exemple. *Si $\Lambda = \{n^3 : n \in \mathbb{N}\}$, alors le théorème d'Ingham donne $|I| > 2\pi$ tandis qu'avec le théorème de Beurling, on obtient $|I| > 0$.*
Si $\Lambda = \{2n, 2n + 10^{-3}n \in \mathbb{N}\}$, alors on obtient respectivement $|I| > 2000\pi$ et $|I| > 2\pi$.

1.4 Différences divisées

Il se trouve que dans certaines applications (voir la section suivante), Λ n'est pas uniformément discret. Baiocchi, Komornik et Loreti ont utilisé les *différences divisées* des exponentielles.

Suites relativement uniformément discrètes. On dit que Λ est *relativement uniformément discret*, si Λ est consistué d'un nombre fini de sous-ensembles uniformément discrets.

Exemple. $\Lambda = \{n, n+1/n : n \in \mathbb{N}^*\} = \{n : n \in \mathbb{N}^*\} \dot\cup \{n+1/n : n \in \mathbb{N}^*\}$.

On note d'ailleurs une caractérisation :

Proposition 1.4.1. *Les assertions suivantes sont équivalentes :*

(i) Λ est relativement uniformément discrète ;

(ii) $D^+(\Lambda) < \infty$;

(iii) $n^+(1) < \infty$.

Si l'on suppose que (λ_n) est croissante, on a encore une nouvelle caractérisation :

Proposition 1.4.2. *On a l'équivalence :*

(i) $D^+(\Lambda) < \infty$;

(ii) il existe $\gamma > 0$ et $M > 0$ tels que $\lambda_{n+M} - \lambda_n > M\gamma$ pour tout $n \in \mathbb{Z}$.

Remarque 1.4.3. *La condition (ii) est une condition de "gap" élargie.*

On supposera ici toujours que $D^+ < \infty$. On peut alors réindicer (λ_n) telle que (λ_n) soit croissante.

Différence divisée. On définit alors la différence divisée $[\mu_1, \dots, \mu_k]$ par récurrence :

$$[\mu_j](t) = e^{i\mu_j t} \quad j = 1, \dots, k,$$

$$[\mu_1, \dots, \mu_k] = \begin{cases} \frac{[\mu_1,\dots,\mu_{k-1}]-[\mu_1,\dots,\mu_k]}{\mu_1-\mu_k} & \text{si } \mu_1 \neq \mu_k \\ \frac{\partial}{\partial\mu}[\mu, \mu_1, \dots, \mu_{k-1}]|_{\mu=\mu_1} & \text{si } \mu_1 = \mu_k. \end{cases}$$

Remarquons que l'on peut avoir une formule explicite pour $[\mu_1, \dots, \mu_k]$:

$$[\mu_1, \dots, \mu_j](t) = (it)^{j-1} \int_0^1 \int_0^{s_1} \dots \int_0^{s_{j-2}} \tag{1.4.1}$$

$$\exp(i[s_{j-1}(\mu_j - \mu_{j-1}) + \dots + s_1(\mu_2 - \mu_1) + \mu_1)]t) \mathrm{d}s_{j-1} \dots \mathrm{d}s_1.$$

Il s'agit de la formule de Genocchi–Hermite (voir [21] pour une référence originale ou [10] pour une référence très récente).

Remarque 1.4.4. *Dans le cas où Λ n'est pas relativement uniformément discret, on pourrait définir les différences divisées, qui dépendraient alors de l'énumération choisie. Néanmoins, pour toute énumération, les différences divisées ne pourront pas former de suite de Riesz.*

Chaîne d'exposants proches. Soit alors $\gamma > 0$ tel que l'on ait la condition de "gap" élargie, et soit $0 < \gamma' \leq \gamma$.
On dit que $\lambda_m, \ldots, \lambda_{m+j-1}$ forment une *chaîne d'exposants γ' proches*, si :

$$\begin{cases} \Re(\lambda_m) - \Re(\lambda_{m-1}) \geq \gamma'; \\ \Re(\lambda_n) - \Re(\lambda_{n-1}) < \gamma', \quad n = m+1, \ldots, m+j-1; \\ \Re(\lambda_{m+j}) - \Re(\lambda_{m+j-1}) \geq \gamma'. \end{cases} \tag{1.4.2}$$

Pour $j = 1, \ldots, M$ et $m \in \mathbb{Z}$ tels que $\lambda_m, \ldots, \lambda_{m+j-1}$ forment une suite d'exposants γ' proches, on définit :

$$e_\ell^{\gamma'} = [\lambda_m, \ldots, \lambda_\ell] \quad \ell = m, \ldots, m+j-1. \tag{1.4.3}$$

On vérifie alors que $(e_n^{\gamma'})$ est défini pour tout $n \in \mathbb{Z}$.

Définition. *$(e_n^{\gamma'})$ est alors appelée suite des différences divisées relative à γ' associée à Λ.*

Considérons dans un premier temps que $\Lambda \subset \mathbb{R}$.

Baiocchi, Komornik et Loreti ont alors obtenu dans le cadre des différences divisées le théorème suivant :

Théorème 1.4.1. *(BKL [9]) Soit $\Lambda = \{\lambda_n\}$ une suite croissante (au sens large). Si $|I| > 2\pi D^+$, alors on peut définir M et γ, tels que l'on ait une condition de "gap" élargie, et pour tout $0 < \gamma \leq \gamma'$, les* différences divisées *$(e_n^{\gamma'})$ forment une suite de Riesz dans $L^2(I)$ et les constantes qui interviennent ne dépendent que de I, γ, γ' et M.*

Remarque 1.4.5. *(i) On note que l'on peut choisir arbitrairement γ' entre 0 et γ ; la structure peut être simplifiée si l'on prend des γ' plus petits.*

(ii) Comme M et γ peuvent être directement définies à partir de D^+, on peut alors faire dépendre les constantes uniquement de I et D^+.

Dans le cas complexe, on rappelle que

$$D^+(\Lambda) = D^+(\Re(\Lambda)),$$

et le théorème reste vrai (voir Avdonin et Ivanov, [3]). La preuve dans [3] utilise alors des outils d'analyse complexe.

Remarque 1.4.6. *(i) La définition des différences divisées dans le cas complexe donnée dans [3] est légèrement différente.*

(ii) La dépendance des constantes dans [3] n'est pas explicitée.

1.5 Longueur critique

Nous avons vu que la longueur critique dans le cas des suites de Riesz est donnée par la densité supérieure. Il se trouve que cela reste vrai dans le cas des différences divisées et cela va être montré au chapitre 3.

Théorème 1.5.1. *Si $I < 2\pi D^+$, alors les différences divisées ne forment pas une suite de Riesz.*

Observabilité spectrale. Notons que cette borne peut-être différente de celle donnée par *l'observabilité spectrale*, définie pour Λ par

$$|a_n|^2 \leq C_n \int_J \left|\sum a_k e^{i\lambda_k t}\right|^2.$$

La longueur critique pour l'observabilité spectrale est alors directement liée au rayon de densité grâce aux deux lemmes suivants obtenus par Jaffard et Haraux [20] :

Lemme 1.5.1. *Si on a observabilité spectrale pour Λ' dans $L^2(I)$ alors $\mathcal{E}(\Lambda)$ n'est pas dense dans $L^2(J)$, pour $|J| > |I|$, avec Λ et Λ' différant seulement d'un nombre fini de termes.*

Lemme 1.5.2. *Si $\mathcal{E}(\Lambda)$ n'est pas dense dans $L^2(J)$, alors on a observabilité spectrale dans $L^2(J)$.*

Le rayon de densité est donné par :

$$R(\Lambda) = \sup\{A > 0 \quad : \quad \mathrm{Vect}(\mathcal{E}(\Lambda)) \text{ est dense dans } L^2([-A, A])\}. \tag{1.5.1}$$

La longueur critique pour l'observabilité spectrale est alors donnée par $2R(\Lambda)$.
0n constate souvent dans la pratique que

$$R(\Lambda) = \pi D^+,$$

mais ceci n'est pas toujours le cas ; le rayon peut être calculé au moyen d'une autre densité : la densité de *Beurling–Malliavin*, D_{BM} et on a :

$$R(\Lambda) = \pi D_{BM}.$$

On pourra consulter le livre de Koosis dédié à la démonstration du théorème de Beurling–Malliavin [33].

Exemple. *On peut construire une suite telle que $D_{BM} \leq \epsilon$ et $D^+ = 1$. Pour cela, pour chaque $k \in \mathbb{N}$, on peut prendre de plus en plus d'entiers espacés de 1 proches de 2^k, dans l'intervalle $[2^k - k/\varepsilon, 2^k]$.*

Remarquons enfin que pour $I > 2\pi D^+$, on peut avoir une estimation de la constante pour l'observabilité spectrale :

Proposition 1.5.3. *Pour $n \in K$ et $I > 2\pi D^+$, on peut écrire*

$$c(n,\Lambda)|a_n|^2 \leq \int_I \Big|\sum_{k\in K} a_k e^{i\lambda_k t}\Big|^2 \mathrm{d}t \tag{1.5.2}$$

pour une constante de la forme

$$\begin{aligned} &c(n,\Lambda) = c(M,\gamma,\gamma') \\ &\min(1, |\tilde\lambda_2 - \tilde\lambda_1|, |\tilde\lambda_3 - \tilde\lambda_2| \cdot |\tilde\lambda_2 - \tilde\lambda_1|, \ldots, |\tilde\lambda_j - \tilde\lambda_{j-1}| \ldots |\tilde\lambda_2 - \tilde\lambda_1|), \end{aligned}$$

où $\tilde\lambda_1, \ldots, \tilde\lambda_j$ est la suite d'exposants $\gamma'-$proches. (voir la définition au chapitre 3).

Preuve. Cela vient du fait que si e_m est la suite des différences divisées et $f_m = e^{i\lambda_n \cdot}$, alors on a des relations du type :

$$\begin{aligned} f_m &= e_m \\ f_{m+1} &= (\lambda_{m+1} - \lambda_m)e_{m+1} + e_m \\ f_{m+2} &= (\lambda_{m+2} - \lambda_m e_{m+1})(\lambda_{m+1} - \lambda_m)e_{m+1} + \ldots \end{aligned}$$

□

Remarque 1.5.4. *Dans le cas d'exposants multiples : $\lambda_{k+1} - \lambda_k$, on remplace $|\lambda_{k+1} - \lambda_k|$ par 1.*

1.6 Bases de Riesz au sens des sous-espaces

L'étude de l'observabilité d'un système se traduit souvent par la recherche d'un équivalent de

$$\int_I p(u(t))^2 \mathrm{d}t, \tag{1.6.1}$$

où $u(t)$ est la solution d'un système et p une semi-norme. Nous venons d'exposer le cas particulier de

$$\int_I \Big|\sum a_k e^{i\lambda_k t}\Big|^2 \mathrm{d}t, \tag{1.6.2}$$

qui s'applique à des problèmes en dimension 1. Pour traiter des problèmes de dimension supérieure ou certains systèmes d'équations, on sera amené à considérer des solutions $u(t)$ sous la forme

$$u(t) = \sum a_k e^{i\lambda_k t} e_k, \tag{1.6.3}$$

où (e_k) est une famille de vecteurs dont on va préciser les hypothèses.
On se fixe un espace de Hilbert complexe séparable H et on se donne un système d'évolution :

$$x' = Ax, \quad x(0) = x_0, \tag{1.6.4}$$

où A est un opérateur non borné de domaine $D(A) \subset H$ dense.

Théorie des semi-groupes. Un opérateur A est dit *générateur infinitésimal* d'un groupe (noté e^{tA}), si pour tout $x_0 \in D(A)$, le problème (1.6.4) admet une unique solution continûment différentiable sur $\mathbb{R}^*$ et si l'ensemble résolvant $\rho(A)$ (qui est consistué des complexes λ tels que $(\lambda I - A)^{-1}$ est un opérateur linéaire borné) est non vide (voir Pazy [39], p. 102).

Remarque 1.6.1. *Pour un C_0-*semi*-groupe, la définition est identique en remplacant $\mathbb{R}^*$ par $\mathbb{R}^{+*}$.*

On a une caractérisation des C_0-groupes :

Théorème 1.6.1. *A est générateur infinitésimal d'un C_0-groupe si et seulement si il existe $\omega \in \mathbb{R}$, tel que*

$$\rho(A) \subset \mathbb{R}\backslash[-|\omega|, |\omega|] \quad et \quad \|(\lambda - \omega)(A - \lambda I)^{-n}\| \leq M, \tag{1.6.5}$$

pour tout $n \in \mathbb{N}$ et $\lambda \in \rho(A)$, avec une constante M indépendante de n et λ.

Le cadre qui nous intéresse ici est lorsque la solution se décompose dans une certaine base de vecteurs propres.

Eléments spectraux. Comme en dimension finie, on peut définir les valeurs propres et les vecteurs propres. On rappelle ici quelques définitions.
A est un opérateur *discret* ou *à résolvante compacte* s'il existe $\mu \in \mathbb{C}$ tel que $(\mu - A)^{-1}$ est compact dans H.
Un vecteur non nul $Y \in H$ est appelé *vecteur propre généralisé* de A correspondant à une *valeur propre* λ s'il existe un entier positif n tel que

$$(\lambda - A)^n Y = 0.$$

Le nombre $n_{a\lambda}$ de vecteurs propres généralisés linéairement indépendants est alors appelé la *multiplicité algébrique* de λ.
En particulier, λ est dit *algébriquement simple* si $n_{a\lambda} = 1$.

Remarque 1.6.2. *Pour un opérateur discret, chaque valeur propre est de multiplicité algébrique finie (voir par exemple [16]).*

Un vecteur non nul $Y \in H$ est appelé *vecteur propre* de A si

$$(\lambda - A)Y = 0.$$

Le nombre $n_{g\lambda}$ de vecteurs propres linéairement indépendants est appelé la *multiplicité géométrique* de λ.
(en particulier, λ est dit *géométriquement simple* si $n_{g\lambda} = 1$).

Base orthonormée : un des cas assez fréquents dans les applications est lorsque A est à résolvante compacte et anti-auto-adjoint ($iA = (iA)^*$, où $*$ représente l'adjoint). Dans ce cas (grâce au théorème spectral), A admet une base orthonormée de vecteurs propres (e_k) et l'on peut écrire la solution de (1.6.4) sous la forme

$$x = \sum x_k e^{i\lambda_k t} e_k, \tag{1.6.6}$$

où λ_k est une suite croissante de réels.
On retrouve le cas de l'équation des ondes :

$$u(t,x) = \sum (a_k e^{i\omega_k t} + a_{-k} e^{-i\omega_k t}) e_k(x) \tag{1.6.7}$$

avec (ω_k) une suite croissante positive et (e_k) est une base orthonormée de vecteurs propres de l'opérateur Laplacien.

Bases de Riesz. Pour traiter d'autres applications, il est courant d'affaiblir l'orthonormalité en considérant des familles qui sont images par un isomorphisme de Banach (et non de Hilbert) de bases orthonormées, les bases de Riesz (voir Young pour plus d'informations [44]).
Un cas particulier très courant est le suivant :
il existe une base de Riesz de vecteurs

$$\{E_{k,\ell} : k \in K, \ell = 1, \dots, m_k\} \tag{1.6.8}$$

et des complexes λ_k tels que

$$AE_{K,\ell} = \lambda_k E_{k,\ell} + E_{k,\ell-1},$$

avec $E_{k,0} := 0$ et

- $|\lambda_k| \to \infty$,
- $\sup |\Re\lambda_k| < \infty$,
- $\sup m_k < \infty$.

Sous ces conditions, A est un opérateur à résolvante compacte ; ses valeurs propres sont les λ_k (de multiplicité algébrique finie supérieure ou égale à m_k), associés aux vecteurs

propres généralisés $E_{k,\ell}$ (qui sont vecteurs propres si $\ell = 1$). De plus, A engendre un groupe et une *solution généralisée* (i.e. valable pour tout $x_0 \in H$) est donnée par

$$U(t) = \sum_{k\in K}\sum_{\ell=1}^{m_k} U_{k,\ell}F_{k,\ell}(t) \quad \text{avec} \quad F_{k,\ell}(t) = \sum_{j=0}^{\ell-1}\frac{t^j e^{\lambda_k t}}{j!}E_{k,\ell-j}. \tag{1.6.9}$$

On peut ainsi traiter le cas de couplages cordes-poutres (voir chapitre 2).
Néanmoins, on verra (toujours au chapitre 2) que pour certains autres couplages, les vecteurs propres, se rapprochant à l'infini, ne peuvent pas former de bases de Riesz.

Bases de Riesz au sens des sous-espaces. On définit alors une *base de Riesz au sens des sous-espaces* par : tout $U_0 \in H$ admet une unique décomposition de la forme

$$U_0 = \sum_{k\in K} u_k \quad u_k \in Z_k, \quad \sum \|u_k\|^2 \asymp \|U_0\|^2 \tag{1.6.10}$$

avec (Z_k) une suite d'espace de *dimension finie* uniformément bornée et la convergence de la série a lieu dans H ; une base de Riesz au sens des sous-espaces (relative à Z_k) est alors une famille obtenue en prenant une base sur chaque espace Z_k.

Remarque 1.6.3. *(i) On dira aussi que les sous-espaces Z_k sont en somme de Riesz.*

(ii) Dans le cas où tous les Z_k sont de dimension 1, on retrouve les bases de Riesz.

On dit alors que A est générateur d'un C_0-groupe associé à une base des Riesz au sens des sous-espaces, si :

- A engendre un C_0-groupe ;
- Il existe des sous-espaces Z_k en somme de Riesz stables par A tels que

$$e^{tA}U_0 = \sum_{k\in K} e^{tA}u_k \quad u_k \in Z_k, \quad \text{pour tout } U_0 = \sum_{k\in K} u_k \in H. \tag{1.6.11}$$

Remarque 1.6.4. *Le fait que A engendre un C_0-groupe ne se déduit plus ici de la formule explicite, même avec les conditions que l'on avait fixées précédemment dans le cas de bases de Riesz (l'ensemble résolvant peut être vide), voir chapitre 2.*

Il est intéressant de remarquer que les vecteurs propres peuvent être supposés encore plus faiblement orthogonaux (au sens de Riesz), si l'on rajoute une hypothèse de séparation relativement uniforme des valeurs propres aux endroits où ils manqueraient d'être assez orthogonaux (toujours au sens de Riesz) :

Théorème 1.6.2. *On suppose que le spectre de A est imaginaire pur.*
On a alors :
A est générateur d'un C_0-groupe associé à une base de Riesz au sens des sous-espaces si et seulement si

- *A est générateur d'un C_0-groupe,*

- *il existe des sous-espaces Z_m stables par A engendrés par des vecteurs propres généralisés de A et tels que , la densité supérieure D_m, associée aux valeurs propres de Z_m, est uniformément bornée en m ;*

- *tout $U_0 \in H$ admet une unique décomposition de la forme*

$$U_0 = \sum u_m \quad u_m \in Z_m, \quad \sum \|u_m\|^2 \asymp \|U_0\|^2. \tag{1.6.12}$$

Remarque 1.6.5. *(i) Z_m n'est plus forcément de dimension finie ; si tous les Z_m sont de dimension finie, les densités D_m sont nulles, et l'équivalence est une tautologie (et on peut donc enlever l'hypothèse sur l'irréalité du spectre).*

(ii) A l'inverse, si $Z_m = H$, on peut également considérer que le spectre de A est dans $\mathbb{C}$, grâce à un résultat d'Avdonin [3] (et il s'agit du théorème récent de Guo [18]).

(iii) cette nouvelle caractérisation permet donc d'unifier les deux cas précédents ; pour les cas intermédiaires, le résultat reste valable sur $\mathbb{C}$, si on arrive à expliciter suffisamment les constantes dans [3].

Preuve du théorème 1.6.2. La preuve est une adaptation directe de celle donnée dans [18], en utilisant les estimations précises des différences divisées du théorème BKL.
On a déjà vu que le sens direct est immédiat, si l'on prend pour Z_m des espaces de dimension finie.
Pour l'autre sens, montrons d'abord que l'on a une base de Riesz au sens des sous-espaces formée par les vecteurs propres généralisés. L'hypothèse sur les densités, associée au théorème BKL nous assure qu'il existe un temps T assez grand et des constantes indépendantes de m telles que $(e_{m,k,k'}(t))_{(k,k')\in K_m}$ soit une suite de Riesz dans $L^2(0,T)$, où $(e_{m,k,k'})_{(k,k')\in K_m}$ est la suite des différences divisées associée aux valeurs propres de l'opérateur A restreint à Z_m. On définit $Z_{m,k}$ comme l'espace des vecteurs propres généralisés (de dimension finie) correspondant aux vecteurs $(e_{m,k,k'})$, avec tous les k' tels que $(k,k') \in K_m$.
Soit maintenant

$$u_m = \sum_{k\in K_m} u_{m,k}, \quad u_{m,k} \in Z_{m,k}.$$

On suppose d'abord que l'on a une somme finie ; le cas général se déduira alors par densité. Fixons nous une base orthonormée (ϕ_k) de Z_m. On a alors :

$$e^{At}u_m = \sum_{k\in K_m} e^{At}u_{m,k}.$$

On peut réarranger la somme de telle sorte que

$$e^{At}u_m = \sum_{(k,k')\in K'_m} e_{m,k,k'}(t)v_{m,k,k'},$$

où les $v_{m,k,k'}$ appartiennent aussi à $Z_{m,k}$ et sont formés à partir de $u_{m,k}$.
On en déduit alors, en fixant une base orthonormée (ϕ_ℓ) de Z_m, que :

$$\int_0^T \|e^{At}u_m\|^2 \mathrm{d}t = \sum_\ell \int_0^T \Big| \sum_{(k,k')\in K_m} e_{m,k,k'}(t) < v_{m,k,k'}, \phi_\ell > \Big|^2 \mathrm{d}t. \tag{1.6.13}$$

Il découle :

$$\int_0^T \|e^{At}u_m\|^2 \mathrm{d}t \asymp \sum |v_{m,k,k'}|^2. \tag{1.6.14}$$

En utilisant l'hypothèse de C_0-groupe, on a aussi :

$$\int_0^T \|e^{At}u_m\|^2 \mathrm{d}t \asymp \|u_m\|^2. \tag{1.6.15}$$

On en déduit alors, d'après l'hypothèse sur les Z_m et comme les constantes sont indépendantes de m :

$$\|U_0\|^2 \asymp \sum |v_{m,k,k'}|^2. \tag{1.6.16}$$

En particulier,

$$\|u_{m,k}\|^2 \asymp \sum_{k'} |v_{m,k,k'}|^2, \tag{1.6.17}$$

et donc,

$$\|U_0\|^2 \asymp \sum \|u_{m,k}\|^2.$$

En rajoutant un raisonnement de densité, on en conclut que l'on a bien une base de Riesz au sens des sous-espaces de vecteurs propres généralisés. Enfin, par densité également, on obtient la forme de la solution (1.6.11), ce qui complète la preuve. □

1.7 Exemples de spectres Λ

Equation des ondes. En prenant $\Omega = (0, \pi)$ (cf section 1.1), le spectre est donné par $\Lambda = \mathbb{Z}$, et donc on a $D^+ = 1$.
Avec

$$A_1 := \int_0^T |\partial_\nu u(t,0)|^2 \mathrm{d}t = \int_0^T |\sum k\alpha_k e^{ikt}|^2 \mathrm{d}t,$$

et

$$A_2 := \int_0^T |\partial_\nu u(t,\pi)|^2 \mathrm{d}t = \int_0^T |\sum k\alpha_k (-1)^k e^{ikt}|^2 \mathrm{d}t,$$

l'observabilité se traduit par la recherche d'un équivalent de $A1 + A2$.
Par une application directe du théorème BKL (mais ici l'identité de Parseval suffit), on obtient l'observabilité pour $T > 2\pi$:

$$A_1 + A_2 \asymp \sum k^2 |\alpha_k|^2. \tag{1.7.1}$$

Le terme de droite est d'ailleurs l'énergie totale E_0. Cependant, en remarquant que

$$\int_0^T |\sum k\alpha_k e^{ikt} + k\alpha_k (-1)^k e^{ikt}|^2 \mathrm{d}t \leq 2A_1 + 2A_2,$$

on obtient en fait l'observabilité pour $T > \pi$ ($T > 2\pi$, si l'on observe seulement une extrémité (c'est-à-dire A_1 ou A_2). L'intervalle de temps est alors optimal, puisque pour $T < \pi$, d'après le théorème 1.5.1 (mais ici on peut aussi utiliser l'égalité de Parseval), on ne peut pas avoir

$$c \sum k^2 |\alpha_k|^2 \leq A_1 \leq A_1 + A_2. \tag{1.7.2}$$

Un système de poutres. Considérons le système suivant :

$$\begin{cases} \partial_{tt} u + \Delta^2 u = 0 & \text{dans } \mathbb{R} \times \Omega = (0, \pi), \\ u = \Delta u = 0, & \text{sur } \mathbb{R} \times \Gamma = \{0, \pi\}, \\ u(0) = u_0 \quad u'(0) = u_1, \end{cases}$$

dans le cadre fonctionnel $H_0^1(\Omega) \times H^{-1}(\Omega)$.
On a ici $\Lambda = \{k^2 : k \in \mathbb{N}\}$ et $D^+ = 0$. On obtient alors le même type de résultat, mais cette fois-ci pour $T > 0$.

Un couplage : onde-onde. On considère le système suivant :

$$\begin{cases} \partial_{tt} - \Delta u + Au + Bv = 0 & \text{dans } \mathbb{R} \times \Omega = (0, \pi) \\ \partial_{tt} - \Delta v + Cu + Dv = 0, & \text{dans } \mathbb{R} \times \Omega = (0, \pi) \\ u = 0 \quad \partial_\nu u = 0 & \text{sur } \mathbb{R} \times \Gamma = \{0, \pi\}, \end{cases}$$

où $A, B, C, D \in \mathbb{R}$.
Ici on a $\Lambda = \{\pm\sqrt{-\lambda_k}\}$, avec $-\lambda_k = k^2 + \frac{A+D}{2} \pm \frac{\sqrt{(A-D)^2+4BC}}{2}$. On peut remarquer que :

- Λ n'est pas uniformément discret.

- Il peut y avoir des valeurs propres non réelles (pour des petites valeurs de k) ; dans les cas où $A = D = 0$ et $B = C = \alpha$, les valeurs propres sont néanmoins toutes réelles.

Un couplage : cordes-poutres. En couplant avec l'opérateur de Petrovski, on obtient cette fois-ci :

$$-\lambda_k = \frac{1}{2}(k^4+k^2+A+D+\sqrt{(k^4-k^2+D-A)^2+4BC}),$$
$$-\mu_k = \frac{1}{2}(k^4+k^2+a+D)-\sqrt{(k^4-k^2+D-A)^2+4BC}),$$

et ainsi $\Lambda = \{\pm\sqrt{-\lambda_k}, \pm\sqrt{\mu_k}\}$.
On remarque que

$$\sqrt{-\lambda_k} = k^2 + \frac{D}{2}k^{-2} + O(k^{-6}),$$
$$\sqrt{-\mu_k} = k + \frac{A}{2}k^{-1} + O(k^{-3}).$$

Ainsi $\sqrt{-\lambda_k}$ est proche de $\sqrt{\mu_{k^2}}$.

Remarque 1.7.1. *(i) Les deux systèmes précédents seront étudiés plus en détail au chapitre 2 dans le cas multidimensionnel, par une autre méthode (de perturbation compacte).*

(ii) En dimension $n > 1$, la densité supérieure est infinie ; néanmoins, on peut obtenir des résultats dans le cas d'une boule, en utilisant l'orthogonalité des harmoniques sphériques (on peut aussi faire l'analogie avec le théorème 1.6.2), voir par exemple [29] ; enfin, dans le cas de pavés, il existe des versions multidimensionnelles des théorèmes d'Ingham (les vecteurs propres s'exprimant aussi sous forme d'exponentielles) ; voir [8] et aussi [26].

Observabilité simultanée de cordes. Dans un cas intervenant pour l'observabilité de cordes de longueur $\ell_1, \dots, \ell_N$:

$$\begin{cases} \partial_{tt}u_j - \Delta u_j + Au + Bv = 0 & \text{dans } \mathbb{R}\times(0,\ell_j), \\ u_j(t,0) = u_j(t,\ell_j) = 0 & \text{pour } t\in\mathbb{R}, \\ u_j(0,x) = u_{j0}(x) \text{ et } \partial_t u_j(0,x) = u_{j1}(x) & \text{pour } x\in(0,\ell_j), \\ j = 1,\dots,N, \end{cases}$$

on a

$$\Lambda = \{\frac{k\pi}{\ell_j}, k\in\mathbb{Z}, j = 1\dots, N\}.$$

On obtient alors $D^+ = \frac{\ell_1+\cdots+\ell_N}{\pi}$.

Remarque 1.7.2. *Dans le cas d'observabilité de poutres, les valeurs propres sont données par $\left(\frac{k\pi}{\ell_j}\right)^2$ et l'on a $D^+ = 0$.*

Les solutions s'écrivent sous la forme

$$u_j(t,x) = \sum_{k=1}^{\infty}(b_{j,k}e^{i|\mu_{j,k}|t} + b_{j,-k}e^{-i|\mu_{j,k}|t})\sin\mu_{j,k}x, \tag{1.7.3}$$

avec :

$$\mu_{j,k} := k\pi/\ell_j. \tag{1.7.4}$$

On cherche alors une équivalence pour

$$\int_I \Big|\sum_{j=1}^{N} u_{j,x}(t,0)\Big|^2 \mathrm{d}t, \tag{1.7.5}$$

qui s'écrit

$$\Big|\sum_{j=1}^{N}\sum_{k=1}^{\infty}(b_{j,k}e^{i|\mu_{j,k}|t} + b_{j,-k}e^{-i|\mu_{j,k}|t})\mu_{j,k}\Big|^2, \tag{1.7.6}$$

et qui est donnée par les différences divisées, grâce au théorème BKL pour $|I| > 2\pi D^+$ et qui n'est plus valable pour $|I| < 2\pi D^+$ avec le théorème 1.5.1.

Remarque 1.7.3. *(i) Pour avoir une minoration dans une bonne norme de Sobolev, on utilise ensuite des résultats d'approximation diophantienne. (On pourra consulter [9] et aussi [12], où l'on utilise également des résultats de théorie des nombres après avoir appliqué le théorème d'Ingham.)*

(ii) Dans un cas similaire d'un réseau en arbre de cordes vibrantes, on peut également appliquer cette méthode. Néanmoins, le calcul de la densité est moins évident. On peut s'en sortir en utilisant le principe de min-max (voir [13], [15], [14], [38], [41]).

(iii) On pourrait étudier encore bien d'autres exemples (par exemple des coques [37]) : cette méthode fonctionne bien lorsque l'on a une bonne connaissance du spectre pour des problèmes qui se ramènent à une dimension (pour que la densité supérieure soit finie).

1.8 Observabilité partielle

L'observabilité partielle a aussi été étudiée par Lions [35], [36] (voir aussi [32] pour des résultats plus récents). On donne ici quelques résultats optimaux en dimension 1 qui peuvent se traiter avec la méthode de perturbation compacte du chapitre 2.
Soient $\Omega = (a,b)$ et A, B, C, D des réels.
On considère le système

$$\begin{cases} \partial_{tt}u_1 - \Delta u_1 + Au_1 + Bu_2 = 0 & \text{dans } \mathbb{R}\times\Omega, \\ \partial_{tt}u_2 + \Delta^2 u_2 + Cu_1 + Du_2 = 0 & \text{dans } \mathbb{R}\times\Omega, \\ u_1 = u_2 = \Delta u_2 = 0 & \text{sur } \mathbb{R}\times\Gamma = \{a,b\}, \end{cases}$$

soumis aux conditions initiales :

$$u_i(0) = u_{i0}, \quad u_i'(0) = u_{i1} \quad \text{sur}(a,b), \quad i = 1,2.$$

Si $(u_{10}, u_{11}, u_{20}, u_{21}) \in H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times H^{-1}(\Omega)$, alors le système précédent admet une unique solution faible :

$$u_1 \in C(\mathbb{R}, H_0^1(\Omega)) \cap C^1(\mathbb{R}; L^2(\Omega))$$
$$u_2 \in C(\mathbb{R}, H_0^1(\Omega)) \cap C^1(\mathbb{R}; H^{-1}(\Omega)).$$

On définit l'*énergie initiale* par

$$E_0 := \|u_{10}\|^2_{H_0^1(\Omega)} + \|u_{11}\|^2_{L^2(\Omega)} + \|u_{20}\|^2_{H_0^1(\Omega)} + \|u_{21}\|^2_{H^{-1}(\Omega)}.$$

Maintenant, on étudie l'observabilité partielle, c'est-à-dire que l'on veut estimer :

$$\int_0^T \int_\Gamma |\partial_\nu u_1|^2 \mathrm{d}\Gamma \mathrm{d}t, \tag{1.8.1}$$

avec des conditions initiales satisfaisant :

$$u_{20} = u_{21} = 0.$$

Remarque 1.8.1. *Dans le cas non couplé, si $u_{10} = u_{11} = 0$ et $u_{20} \neq 0$ ou $u_{21} \neq 0$, on ne peut pas espérer avoir observabilité :*

$$E_0 > 0,$$

et (1.8.1) est égal à zéro.

On peut alors établir le théorème suivant :

Théorème 1.8.1. *Pour $|I| > 2(b-a)$ et pour presque tous quadruplets $(A, B, C, D) \in \mathbb{R}^4$, on a :*

$$\int_I |\partial_x u_1(a,t)|^2 \mathrm{d}t \asymp E_0. \tag{1.8.2}$$

Les paramètres exceptionnels (ceux pour lesquels il n'y a pas observabilité) sont donnés par les quadruplets (A, B, C, D) tels qu'il existe $\ell \neq m$ satisfaisant

$$\lambda_k = \mu_\ell \quad \text{ou} \quad \lambda_k = \lambda_\ell \quad \text{ou} \quad \mu_k = \mu_\ell,$$

avec

$$\lambda_k = \frac{1}{2}(k^4 + k^2 + A + D + \sqrt{(k^4 - k^2 + D - A)^2 + 4BC}),$$
$$\mu_k = \frac{1}{2}(k^4 + k^2 + A + D - \sqrt{(k^4 - k^2 + D - A)^2 + 4BC})$$

Si $I < 2(b-a)$, l'inégalité de droite n'est plus valable.

Remarque 1.8.2. *(i) L'optimalité en temps provient du fait que l'on connaît l'optimalité pour l'opérateur non perturbé.*

(ii) Ce théorème peut aussi se montrer en utilisant le théorème BKL ; néanmoins, on obtient alors le temps non optimal $T > 4(b-a)$.

1.9 Observabilité indirecte

L'observabilité indirecte a été développée par Alabau [1], [2]. Dans ce cas, l'observabilité du système perturbé ne peut alors plus être déduite directement de l'observabilité du système non perturbé, puisque celui-ci n'est pas observable. On obtient cette fois-ci les résultats (en dimension 1) grâce au théorème BKL. Soit $\Omega = (a, b)$ et des réels A, B, C et D. On considère le système :

$$\begin{cases} \partial_{tt} u_1 - \Delta u_1 + A u_1 + B u_2 = 0 & \text{dans } \mathbb{R} \times \Omega, \\ \partial_{tt} u_2 - \Delta u_2 + C u_1 + D u_2 = 0 & \text{dans } \mathbb{R} \times \Omega, \\ u_1 = u_2 = 0 & \text{sur } \mathbb{R} \times \{a, b\}, \end{cases}$$

soumis aux conditions initiales :

$$u_i(0) = u_{i0} \quad u_i'(0) = u_{i1} \quad \text{sur } (a, b), \quad i = 1, 2.$$

Si $(u_{10}, u_{11}, u_{20}, u_{21}) \in (H_0^1(\Omega) \times L^2(\Omega))^2$, le système précédent admet une unique solution faible $u = (u_1, u_2)$ telle que $u_1, u_2 \in C(\mathbb{R}, H_0^1(\Omega)) \cap C^1(\mathbb{R}, L^2(\Omega))$.
On note l'énergie initiale :

$$E_0 = \|u_{10}\|^2_{H_0^1(\Omega)} + \|u_{11}\|^2_{L^2(\Omega)} + \|u_{20}\|^2_{H_0^1(\Omega)} + \|u_{21}\|^2_{L^2(\Omega)}.$$

On cherche à estimer

$$\int_0^T \int_\Gamma |\partial_\nu u_1|^2 \mathrm{d}\Gamma \mathrm{d}t.$$

Cette fois-ci on n'impose pas de conditions spéciales pour les conditions initiales. On cherche ici à obtenir les inégalités d'observabilité pour une norme d'énergie affaiblie :

$$\tilde{E_0} = \|u_{10}\|^2_{L^2(\Omega)} + \|u_{11}\|^2_{H^{-1}(\Omega)} + \|u_{20}\|^2_{L^2(\Omega)} + \|u_{21}\|^2_{H^{-1}(\Omega)}.$$

Théorème 1.9.1. *Pour presque tous les paramètres $(A, B, C, D) \in \mathbb{R}^4$, on a*

$$\tilde{E_0} \leq C \int_0^T |\partial_x u_1(0, t)|^2 \mathrm{d}t. \tag{1.9.1}$$

si $T > 4\pi$. Les paramètres exceptionnels sont donnés par :

$$B = 0, \quad (A - D)^2 + 4BC = 0,$$
$$\lambda_k = \mu_\ell,$$

avec

$$\lambda_k = k^2 + \frac{A + D}{2} + \sqrt{((D - A)^2 + 4BC)},$$
$$\mu_k = k^2 + \frac{A + D}{2} - \sqrt{(D - A)^2 + 4BC}.$$

Schéma de preuve. On suppose d'abord que $b \neq 0$ et $AD - BC \neq 0$ La solution s'écrit alors sous la forme

$$u(t,x) = \sum_{k\in K}[(a_k e^{i\sqrt{\lambda_k}t} + b_k e^{-i\sqrt{\lambda_k}t})v_k + c_k e^{i\sqrt{\mu_k}t} + d_k e^{-i\sqrt{\mu_k}t})w_k]\sin kx.$$

Or on a

$$|\sqrt{\lambda_k} - \sqrt{\mu_k}|^2 \asymp \frac{1}{k^2}. \tag{1.9.2}$$

On en déduit que l'estimation à montrer est équivalente à :

$$\sum(|a_k|^2 + |b_k|^2)(v_{k1}^2 + v_{k2}^2) + (|c_k|^2 + |d_k|^2)(w_{k1}^2 + w_{k2}^2)$$
$$\leq c\int_I |\sum_{k\in K}[(a_k e^{i\sqrt{\lambda_k}t} + b_k e^{-i\sqrt{\lambda_k}t})v_{k1} + c_k e^{i\sqrt{\mu_k}t} + d_k e^{-i\sqrt{\mu_k}t})w_{k1}]k|^2 \mathrm{d}t.$$

Or $E_k = (v_{k1}, v_{k2}, w_{k1}, w_{k2})$ est vecteur propre de la matrice :

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k^2 - A & -B & 0 & 0 \\ -C & -k^2 - D & 0 & 0 \end{pmatrix}. \tag{1.9.3}$$

Ainsi, pour toute valeur propre non nulle, (v_{k1}, v_{k2}) et (w_{k1}, w_{k2}) sont vecteurs propres colinéaires de

$$\begin{pmatrix} -k^2 - A & -B \\ -C & -k^2 - D \end{pmatrix}, \tag{1.9.4}$$

et peuvent donc être choisis indépendants de k ; de plus, comme $B \neq 0$, v_{k2} peut s'exprimer en fonction de v_{k1}. L'inégalité à montrer reste donc équivalente à :

$$\sum(|a_k|^2 + |b_k|^2)(v_{k1}^2) + (|c_k|^2 + |d_k|^2)(w_{k1}^2)$$
$$\leq c\int_I |\sum_{k\in K}[(a_k e^{i\sqrt{\lambda_k}t} + b_k e^{-i\sqrt{\lambda_k}t})v_{k1} + c_k e^{i\sqrt{\mu_k}t} + d_k e^{-i\sqrt{\mu_k}t})w_{k1}]k|^2 \mathrm{d}t.$$

Par le théorème BKL et le théorème 1.5.1, en tenant compte de (1.9.2), on en déduit que cette égalité est vraie si $|I| > 4\pi$ et fausse si $|I| < 4\pi$, pour les hautes fréquences (i.e. pour k assez grand). Le théorème BKL donne d'ailleurs aussi le résultat pour toutes les fréquences (il faut enlever le cas de valeurs propres multiples) qui sont données dans le théorème. □

Remarque 1.9.1. *(i) Dans le second cas $\lambda_k = \mu_\ell$, on peut cependant observer dans un espace de codimension finie.*

(ii) On peut aussi découper en termes de haute et de basse fréquence :

$$\int_I |\partial_x u(0,t)|^2 \mathrm{d}t \asymp \int_I |\partial_x u^H(0,t)|^2 \mathrm{d}t + \int_I |\partial_x u^B(0,t)|^2 \mathrm{d}t, \tag{1.9.5}$$

cette inégalité étant valable dès que l'on a observabilité spectrale.

Chapitre 2

Observability of coupled systems

By applying the theory of semigroups, we generalize an earlier result of Komornik and Loreti [30] on the observability of compactly perturbed systems. As an application, we answer a question of the same authors concerning the observability of weakly coupled linear distributed systems.

2.1 Introduction

Consider the evolutionary problem

$$x' = (A + B)x, \qquad x(0) = x_0$$

where A and B are linear operators in a complex separable Hilbert space H. B is supposed to be compact, it is a so-called compact perturbation. We study the observability of the system, that is, given a finite number of seminorms $p_1, \ldots, p_m$ in H or in a subspace of H (the observations) and a finite number of intervals $I_1, \ldots, I_m$ in $\mathbb{R}$, (here every interval is finite and not reduced to a point) we are wondering whether these observations are sufficient to distinguish solutions corresponding to different initial data. More precisely, we ask whether we have

$$c\|x_0\|^2 \leq \sum_{j=1}^{m} \int_{I_j} p_j(x(t))^2 \, dt \tag{2.1.1}$$

with some positive constant c independent of the particular choice of x_0, which may be different at different places. We also study the estimates

$$\sum_{j=1}^{m} \int_{I_j} p_j(x(t))^2 \, dt \leq c\|x_0\|^2.$$

Here we suppose that the unperturbed system (i.e. with $B = 0$) is observable, at least if the initial data belong to a certain finite codimensional subspace, and thus one can ask whether the perturbed system is also observable. In many concrete cases, A is a skew-adjoint operator having a compact resolvent and thus A is diagonalisable with an orthonormal basis which is an excellent framework to study the estimates. However, orthonormal bases don't often resist to compact perturbations. In fact, looking only for norm equivalences, we can extend the framework to bases which are the images of orthonormal ones by a Banach isomorphism (i. e without keeping necessarily the orthogonality) : the Riesz bases. In fact, if there exists a Riesz basis formed by ordinary and generalized eigenvectors of $A + B$, we can, under natural additional assumptions conclude to the observability. Nevertheless, it is not always easy to prove that the perturbed operator admits a Riesz basis of eigenvectors and sometimes it is not even the case. In order to understand this phenomenon, let us consider a class of operators which are stable under a Riesz sum of finite dimensional spaces. To be more precise, fix a doubly indexed Riesz basis $\{e_{k,l} \ : \ k \geq 1, 1 \leq l \leq m_k\}$ with a bounded sequence (m_k) of positive integers, and introduce the finite dimensional spaces

$$Z_k = \text{Vect } \{e_{k,l} \ : \ 1 \leq l \leq m_k\}.$$

Then we build an operator C, stable under the Z_k, by the giving of endomorphisms $A_k : Z_k \to Z_k$:

$$\begin{aligned} D(C) &:= \left\{ x = \sum x_{k,l} e_{k,l} : \sum A_k x_{k,l} e_{k,l} \in H \right\}, \\ Cx &:= \sum A_k x_{k,l} e_{k,l}. \end{aligned}$$

We can show that C is closed and that if an unbounded linear operator is closable and stable under the Z_k then it coincides with C on its domain. Furthermore, the initial value problem

$$\begin{aligned} x'(t) &= Cx(t), \qquad t \in \mathbb{R}, \\ x(0) &= x_0 \in D(C) \end{aligned}$$

has a unique continuously differentiable solution such that

$$\|x(t)\| \leq c\|x_0\|$$

with a constant c, (which may depend on the time t, but remains independent of the initial data x_0), if and only if $\exp(tA_k)$ is bounded (for a certain norm : we can choose an arbitrary norm on each $\mathbb{C}^{m_k}$ since (m_k) is bounded, the same norm in $\mathbb{C}^{m_k}$ and $\mathbb{C}^{m_\ell}$, if $k \neq \ell$, but $m_k = m_\ell$), for each t $\in \mathbb{R}$. We say then that the problem is well posed for C, and that C generates a strongly continuous group (see [39] for a general definition).

For instance, the problem is well posed for a closed operator A if the latter has a Riesz basis of (generalized) eigenvectors with bounded real parts of their eigenvalues. However, this property may be lost in case of compact perturbations :

Example. Setting

$$A_k = \begin{pmatrix} \lambda_k & k(-\lambda_k + \mu_k) \\ 0 & \mu_k \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ & 1/k \end{pmatrix} \begin{pmatrix} \lambda_k & \\ & \mu_k \end{pmatrix} \begin{pmatrix} 1 & 1 \\ & 1/k \end{pmatrix}^{-1}.$$

The problem is well-posed for A if the sequences

$$\Re(\lambda_k), \Re(\mu_k) \text{ and } k(-\lambda_k + \mu_k) \text{ are bounded}$$

(it is a bounded perturbation of a C^0 semigroup), but the eigenvectors

$$e_{k,1}, e_{k,1} + \frac{1}{k} e_{k,2}$$

don't form a Riesz basis. (We see here that bringing together the eigenvalues may lead to the loss of the independence of the eigenvectors at infinity.) In particular, we notice that if

$$k(-\lambda_k + \mu_k) \to 0$$

and

$$\Im(\lambda_k), \Im(\mu_k) \to \infty,$$

then we have a compact perturbation of a skew adjoint operator with a compact resolvant.

In [30], general observability results were established for compactly perturbed operators under the assumption that there exists a Riesz basis of generalized eigenvectors. The purpose of this paper is to extend that result so as to include cases like the above example. We will also give a concrete application where this more general result is essential.

2.2 Observability results

Let $A : D(A) \subset H \to H$ be an unbounded linear operator in a separable Hilbert space H and $B : H \to H$ a continuous linear operator. We suppose that A generates a strongly continuous group S_A. Since B is continuous, $A + B$ also generates a strongly continuous group S_{A+B}. See for example [39].

Let L be a finite-codimensional subspace of H, and $L_H \subset L$ a linear hull generated by a Riesz basis $(e_k)_{k \in K}$ of L

$$Z_L = \text{Vect } \{e_k \ : \ k \in K\}, \quad \overline{Z_L} = L.$$

K denotes here and in the sequel an infinite countable set. Similarly, we will denote by Z_H a linear hull generated by a Riesz basis of H.
Concerning the direct inequality, we assume that :

$$\sum_{j=1}^{m} \int_{I_j} p_j(S_A(t)x_0)^2 dt \leq c \left\|x_0\right\|^2 \text{ for all } x_0 \in Z_L, \tag{2.2.1}$$

and we want to deduce from this the estimate

$$\sum_{j=1}^{m} \int_{J_j} p_j(S_{A+B}(t)x_0)^2 dt \leq c \left\|x_0\right\|^2 \text{ for all } x_0 \in Z_H, \tag{2.2.2}$$

for every choice of intervals J_j.

Concerning the inverse inequality, we assume that

$$c \left\|x_0\right\|^2 \leq \sum_{j=1}^{m} \int_{I_j} p_j(S_A(t)x_0)^2 dt \text{ for all } x_0 \in Z_L. \tag{2.2.3}$$

We then want to deduce

$$c \left\|x_0\right\|^2 \leq \sum_{j=1}^{m} \int_{J_j} p_j(S_{A+B}(t)x_0)^2 dt \text{ for all } x_0 \in Z_{\widetilde{L}}, \tag{2.2.4}$$

where J_j are intervals such that they contain the closure of I_j in their interior, and $Z_{\widetilde{L}}$ is a linear hull generated by a Riesz basis of $\widetilde{L}$, which is a finite codimensional subspace as big as possible, that is,

$$H = \widetilde{L} \oplus \overline{M}$$

where M (respectively $\overline{M}$) is the (respectively closed) linear hull of all vectors $x \in H$ which satisfy for some complex number λ and for some nonnegative integer k the equalities

$$p_j((A + B - \lambda \mathrm{Id}\ )^{\ell} x) = 0, \tag{2.2.5}$$

for all $\ell = 0, \ldots, k$, $j = 1, \ldots, m$, and

$$(A + B - \lambda \mathrm{Id}\ )^k x = 0. \tag{2.2.6}$$

Indeed, we have :

Lemma 2.2.1. *If $x_0 \in M$, then*

$$p_j(S_{A+B}(t)x_0) = 0 \tag{2.2.7}$$

and therefore (2.2.4) *doesn't hold if $x_0 \in M \setminus \{0\}$.*

Concerning the direct equality, we have the following result :

Proposition 2.2.2. *We suppose* (2.2.1)*, then we have* (2.2.2)*.*

Concerning the inverse equality, we have two results. Let us first introduce the following definition.

Definition. $(f_k)_{k\geq 1}$ is a *pseudo-basis* if Vect $\{f_k\}$ is dense in H and if, for every bounded sequence (x_k) such that

$$x_k \in \text{Vect } \{f_j \ : \ j \geq k\},$$

we have

$$x_k \rightharpoonup 0.$$

Lemma 2.2.3. *$\{f_{k,\ell} \ : \ k \geq 1, \ 1 \leq \ell \leq m_k\}$ is a pseudo-basis of H, if there exists a Riesz basis of H $\{e_{k,\ell} \ : \ k \geq 1, 1 \leq \ell \leq m_k\}$ such that*

$$\text{Vect } \{e_{k,\ell} \ : \ 1 \leq \ell \leq m_k\} = \text{Vect } \{f_{k,\ell} \ : \ 1 \leq \ell \leq m_k\} \tag{2.2.8}$$

for each k. In that case, we will say that $\{f_{k,\ell} \ : \ k \geq 1, \ 1 \leq \ell \leq m_k\}$ is a Riesz basis of subspaces in H.

Then we have the following result :

Proposition 2.2.4. *We suppose* (2.2.1), (2.2.3), *that B is compact. Then there exists a finite codimensional subspace $L' \subset L$ such that*

$$c\|x_0\|^2 \leq \sum_{j=1}^{m} \int_{J_j} p_j(S_{A+B}(t)x_0)^2 \ dt \quad \textit{for all} \quad x_0 \in Z_{L'}. \tag{2.2.9}$$

Moreover for every Riesz basis $(f_k)_{k\geq 1}$ such that $Z_L = \text{Vect } \{f_j \ : \ j \geq k'\}$ for some k', we can take $Z_{L'} = \text{Vect } \{f_j \ : \ j \geq k''\}$ with a sufficiently large integer $k'' \geq k'$.

If $A + B$ satisfies some spectral properties, then we will obtain a better result. For this, let us recall, e.g., from [16] that a vector $x \in H$ is called a *generalized eigenvector* with eigenvalue $\lambda \in \mathbb{C}$ of a linear operator C in H if

$$(C - \lambda \text{Id})^m x = 0$$

for some positive integer m. Furthermore, an eigenvalue $\lambda \in \mathbb{C}$ is called *of finite type* if the corresponding generalized eigenvectors form a finite dimensional subspace M, and if

$$H = M \oplus S$$

with M and S stable by C.

Let us now formulate our main result :

Theorem 2.2.5. *Assume that*

- *A is a skew-adjoint operator having a compact resolvant,*
- *B is compact,*
- *$A + B$ has a Riesz basis of subspaces in H of generalized eigenvectors, whose eigenvalues are of finite type,*

– (2.2.1) *and* (2.2.3) *are satisfied with a finite codimensional subspace* Z_L *generated by some generalized eigenvectors of* $A+B$.

Then (2.2.4) *holds true and* M *is finite dimensional.*

Remark 2.2.6. *In particular, this theorem asserts that the cases of non observability coming from such compact perturbations are those for which* $M \neq \{0\}$. *In fact, we can easily see that* $M \neq \{0\}$ *is equivalent to the existence of a non zero vector* $x \in H$ *which satisfies the inequalities*

$$p_j(x) = 0$$

for all $j = 1, \ldots, m$, *and*

$$(A+B)x = \lambda x,$$

for some complex number λ.

Remark 2.2.7. *(i) The proof can be adapted to the case where* A *is only compactly near of an operator which has the* Riesz basis property *(cf chapter* 1*).*

(ii) In order to treat the case of partial observability (see chapter 1*), we mention that we can also apply the theorem to a subspace invariant by the operators.*

We prove the above formulated results in the next section. Then, in the last section of the paper we apply these results in order to answer a question left open in [29], as we will see in remark 2.4.9.

2.3 Proof of the results.

2.3.1 Proof of Lemma 2.2.1

If x_0 satisfies (2.2.6), we obtain :

$$S_{A+B}(t)x_0 = \sum_{j=0}^{k-1} \frac{t^j e^{\lambda t}}{j!}(A+B-\lambda \mathrm{Id}\ )^j x_0,$$

and thus, according to the properties of the seminorms, we get (2.2.7). Now, if $x_0 \in M$, x_0 is a finite linear combination of vectors which satisfy (2.2.6), and the result remains true, thanks to the property of the seminorms for the sums.

2.3.2 Proof of Proposition 2.2.2

We will first prove that

$$\sum_{j=1}^{m} \int_{J_j} p_j(S_A(t)x_0)^2 \, dt \leq c\|x_0\|^2 \text{ for all } x_0 \in Z_H. \tag{2.3.1}$$

We fix a Riesz basis $(e_l)_{l\geq k}$ of L, such that $Z_L = \text{Vect}_{\ l\geq k}(e_j)$ for a certain integer k. We can complete this Riesz basis in order to obtain a Riesz basis $(e_k)_{k\geq 1}$ of H, and we have a decomposition $Z_H = Z_L \oplus M$, with a finite dimensional $M := \text{Vect}_{\ 1\leq j\leq k}(e_j)$, we will also assume that $(e_j)_{1\leq j\leq k}$ is orthonormal (this is always possible by the Gram-Schmidt procedure). We denote by π_1 (resp. π_2) the projection onto the finite dimensional supplement M (resp. onto Z_L), according to the latter decomposition. From (2.2.1), we have, for each $t \in \mathbb{R}$,

$$\int_{I_j} p_j(S_A(s)\pi_2 S_A(t)x_0)^2 \ ds \leq c\|\pi_2 S_A(t)x_0\|^2.$$

Since S_A is a strongly continuous group, there exist numbers ω and M such that

$$\|S_A(t)\| \leq Me^{\omega|t|} \text{ for all } t \in \mathbb{R} \tag{2.3.2}$$

and therefore

$$\int_{I_j} p_j(S_A(s)\pi_2 S_A(t)x_0)^2 \ ds \leq cM^2 e^{2\omega|t|}\|x_0\|^2.$$

Given an interval I, which we will fix later, we integrate this inequality over I :

$$\int_I \int_{I_j} p_j(S_A(s)\pi_2 S_A(t)x_0)^2 \ ds \ dt \leq cM^2 \int_I e^{2\omega|t|} \ dt\|x_0\|^2.$$

Then, applying the Fubini-Tonelli theorem, we have

$$\int_{I_j} \left(\int_I p_j(S_A(s)\pi_2 S_A(t)x_0)^2 dt \right) \ ds \leq cM^2 \int_I e^{2\omega|t|} \ dt\|x_0\|^2.$$

Hence, there exists $s_0 \in I_j$ (which may depend on I) such that

$$\int_I p_j(S_A(s_0)\pi_2 S_A(t)x_0)^2 dt \leq \frac{2cM^2}{|I_j|} \int_I e^{2\omega|t|} \ dt\|x_0\|^2 \tag{2.3.3}$$

On the other hand, we have

$$\pi_1 S_A(t)x_0 = \sum_{l=1}^{k} (S_A(t)x_0|e_l)\, e_l,$$

since the sequence $(e_j)_{1\leq j\leq k}$ is orthonormal. Then, using the inequalities between the arithmetic and quadratic means, we obtain

$$p_j(S_A(s_0)\pi_1 S_A(t)x_0)^2 \leq k \sum_{l=1}^{k} |(S_A(t)x_0|e_l)|^2 \, p_j(S_A(s_0)e_l)^2.$$

Hence, thanks to (2.3.2), we have

$$\int_I p_j(S_A(s_0)\pi_1 S_A(t)x_0)^2 dt \leq kM^2 \int_I e^{2\omega|t|} \ dt \sum_{l=1}^{k} p_j(S_A(s_0)e_l)^2 \|x_0\|^2. \tag{2.3.4}$$

Combining (2.3.3) and (2.3.4) we obtain that

$$\int_I p_j(S_A(s_0)S_A(t)x_0)^2 dt \leq 2\max\left(\frac{2cM^2}{|I_j|}\int_I e^{2\omega t}dt,\ kM^2\int_I e^{2\omega t}dt\sum_{l=1}^{k} p_j(S_A(s_0)e_l)^2\right)\|x_0\|^2.$$

Since $S_A(s_0)S_A(t) = S_A(s_0+t)$, then we have

$$\int_{I+s_0} p_j(S_A(t)x_0)^2 dt \leq 2\max\left(\frac{2cM^2}{|I_j|}\int_I e^{2\omega t}dt,\ kM^2\int_I e^{2\omega t}dt\sum_{l=1}^{k} p_j(S_A(s_0)e_l)^2\right)\|x_0\|^2.$$

Now, let J_j be an interval; we can choose I such that $J_j \subset I + s_0$. For example, if $J_j = (a,b)$ and $I_j = (c,d)$, since $s_0 \in I_j$, we may take $I = (a-d, b-c)$.

So we obtain

$$\int_{J_j} p_j(S_A(t)x_0)^2\ dt \leq c\|x_0\|^2$$

and (2.3.1) follows.

Next we prove (2.2.2). Let $x_0 \in Z_H$. Thanks to (2.3.1), we only have to show that

$$\int_{J_j} p_j(S_{A+B}(t)x_0 - S_A(t)x_0)^2 dt \leq c\|x_0\|^2, \tag{2.3.5}$$

because

$$\int_{J_j} p_j(S_{A+B}(t)x_0)^2\ dt \leq 2\left(\int_{J_j} p_j(S_{A+B}(t)x_0 - S_A(t)x_0)^2\ dt + \int_{J_j} p_j(S_A(t)x_0)^2 dt\right).$$

Suppose at first that $J_j \subset \mathbb{R}^+$. We begin with the following formula from Duhamel (see e. g. Pazy [39] for a proof)

$$S_{A+B}(t)x_0 - S_A(t)x_0 = \int_0^t S_A(t-s)BS_{A+B}(s)x_0\ ds.$$

Hence, putting $J_j = (a,b)$ we have

$$\begin{aligned}\int_{J_j} p_j(S_{A+B}(t)x_0 - S_A(t)x_0)^2\ dt & \\ &= \int_a^b p_j\left(\int_0^t S_A(t-s)BS_{A+B}(s)x_0 ds\right)^2\ dt \\ &\le \int_a^b \left(\int_0^t p_j\left(S_A(t-s)BS_{A+B}(s)x_0\right)\ ds\right)^2\ dt \\ &\le \int_a^b t\int_0^t p_j\left(S_A(t-s)BS_{A+B}(s)x_0\right)^2\ ds\ dt\end{aligned}$$

by using successively the Minkowski inequality for p_j and the Cauchy-Schwarz inequality.

Next, using the Fubini-Tonelli theorem we have

$$\int_{t=a}^{t=b}\int_{s=0}^{s=t} = \int_{t=a}^{t=b}\int_{s=0}^{s=a} + \int_{t=a}^{t=b}\int_{s=a}^{s=t} = \int_{s=0}^{s=a}\int_{t=a}^{t=b} + \int_{s=a}^{s=b}\int_{t=s}^{t=b},$$

and thus

$$\begin{aligned}\int_{J_j} p_j(S_{A+B}(t)x_0 - S_A(t)x_0)^2\ dt & \\ &\le \int_0^a \left(\int_{a-s}^{b-s} t p_j\left(S_A(t)BS_{A+B}(s)x_0\right)^2 dt\right)\ ds \\ &\quad + \int_a^b \left(\int_0^{b-s} t p_j\left(S_A(t)BS_{A+B}(s)x_0\right)^2 dt\right)\ ds \\ &\le c\int_0^a \|BS_{A+B}(s)x_0\|^2\ ds + c\int_a^b \|BS_{A+B}(s)x_0\|^2\ ds,\end{aligned}$$

thanks to (2.3.1). Since B is continuous, we obtain (2.3.5). We recall that we have supposed $J_j = (a,b) \subset \mathbb{R}^+$. Now, if $J_j \subset \mathbb{R}^-$, we proceed alike, by changing t, a, b into $-t, -b, -a$. At last, we conclude to (2.3.5) in the general case, by cutting the interval into two parts, one included in $\mathbb{R}^+$ and the other included in $\mathbb{R}^-$.

2.3.3 Proof of Lemma 2.2.3

Set a bounded sequence $(x_{k,\ell})$ such that

$$x_{k,\ell} \in \text{Vect}\left\{\mathrm{f_{j,i}} : (j,i) \ge (k,\ell)\right\}.$$

(Here we use the lexicographic order). Thanks to (2.2.8), we have

$$x_{k,\ell} \in \text{Vect}\ \left\{\mathrm{e_{j,i}} : (\mathrm{j},\mathrm{i}) \ge (\mathrm{k},1)\right\}$$

Since $\{e_{k,l} : k \geq 1, 1 \leq l \leq m_k\}$ is a Riesz basis, there exists a Banach space automorphism Φ and an orthonormal basis

$$\{u_{k,l} : k \geq 1, 1 \leq l \leq m_k\}$$

such that $\Phi(e_{k,l}) = u_{k,l}$. Thus, we have

$$\Phi^{-1} x_{k,\ell} \in \text{Vect } \{u_{j,i} : (j,i) \geq (k,1)\},$$

that is, we can find numbers $(y_{j,i}^{(k,\ell)})$ such that

$$\Phi^{-1} x_{k,l} = \sum_{(j,i)\geq(k,1)} y_{j,i}^{(k,\ell)} u_{j,i}$$

Now, let $x \in H$ and compute :

$$\begin{aligned}(x_{k,l}|x) &= \left(\Phi^{-1}x_{k,l}|\Phi^* x\right) = \sum_{(j,i)\geq(k,1)} y_{j,i}^{(k,l)} \left(u_{k,l}|\Phi^* x\right) \\ &\leq \left\|\Phi^{-1}x_{k,l}\right\| \left(\sum_{(j,i)\geq(k,1)} \left|(u_{j,i}|\Phi^* x)\right|^2\right)^{1/2}\end{aligned}$$

thanks to the Cauchy-Schwarz inequality. Now, $\Phi^{-1}x_{k,\ell}$ remains bounded and $(u_{j,i}|\Phi^* x)$ is square summable by the Parseval identity.

We obtain therefore that

$$(x_{k,l}|x) \to 0$$

as k tends to infinity. Thus, we have the result, since, thanks to (2.2.8), Vect $\{f_{k,\ell}$ is also dense in H.

2.3.4 Proof of Proposition 2.2.4

We fix a Riesz basis $(f_k)_{k\geq 1}$ such that $Z_L = \text{Vect}_{\ j\geq k'}(f_j)$ for some integer k'. We fix an integer $k \geq k'$, which we will choose later and a vector $x_0 \in \text{Vect}\,\{f_j : j \geq k\}$.

Then we have :

$$\int_{J_j} p_j (S_A(t)x_0)^2 \, dt \leq 2\left(\int_{J_j} p_j (S_A(t)x_0 - S_{A+B}(t)x_0)^2 \, dt + \int_{J_j} p_j (S_{A+B}(t)x_0)^2 \, dt\right).$$

Since

$$S_{A+B}(t)x_0 - S_A(t)x_0 = \int_0^t S_A(t-s) B S_{A+B}(s)x_0 \, ds,$$

we obtain

$$\int_{J_j} p_j(S_{A+B}(t)x_0)^2\ dt \geq \frac{1}{2}\int_{J_j} p_j(S_A(t)x_0)^2\ dt - \int_{J_j} p_j(\int_0^t S_A(t-s)BS_{A+B}(s)x_0\ ds)^2\ dt. \quad (2.3.6)$$

We write $J_j = (a,b)$, and we consider only the case where $J_j \subset \mathbb{R}^+$ (the general case follows with the same argument as in the preceding proof). Thanks to (2.2.1), we have like in the Proposition 2.2.2 :

$$\begin{aligned}\int_{J_j} p_j(\int_0^t S_A(t-s)BS_{A+B}(s)x_0\ ds)^2\ dt &\leq c\int_0^a \|BS_{A+B}(s)x_0\|^2\ ds + c\int_a^b \|BS_{A+B}(s)x_0\|^2\ ds \\ &\leq c\int_0^b \left(\sup_{\substack{x\in \mathrm{Vect}\ \{f_j:j\geq k\}\\ \|x\|\leq 1}} \|BS_{A+B}(s)x\|\right)^2 ds\,\|x_0\|^2. \quad (2.3.7)\end{aligned}$$

Now, for each fixed $s \in \mathbb{R}$, let (x_k) be an approximation of the supremum

$$\sup_{\substack{x\in \mathrm{Vect}\ \{f_j:j\geq k\}\\ \|x\|\leq 1}} \|BS_{A+B}(s)x\|.$$

Since $(f_k)_{k\geq 1}$ is also a pseudo-basis, (x_k) converges weakly to zero. Since B is compact, so is $BS_{A+B}(s)$ and therefore, $BS_{A+B}(s)x_k$ converges strongly to zero. So, we can easily conclude that the approximation and thus the supremum (2.3.4) converges to zero. We also notice that (2.3.4) is dominated by $\|BS_{A+B}(s)\|$, which is integrable as B is continuous. So, by applying Lebesgue's dominated convergence theorem, we obtain that

$$\varepsilon_k := \int_0^b \left(\sup_{\substack{x\in \mathrm{Vect}\ \{f_j:j\geq k\}\\ \|x\|\leq 1}} \|BS_{A+B}(s)x\|\right)^2 ds \to 0 \text{ as } k\to\infty. \quad (2.3.8)$$

Keeping in mind from (2.3.6) and (2.3.7) that :

$$\int_{J_j} p_j(S_{A+B}(t)x_0)^2 dt \geq \frac{1}{2}\int_{J_j} p_j(S_A(t)x_0)^2 dt - c\varepsilon_k\|x_0\|^2.$$

thanks to (2.2.3) and (2.3.8), we can now choose k independent from x_0 such that (2.2.9) holds true with $k'' = k$.

2.3.5 Proof of Theorem 2.2.5

Since the eigenvalues of $A+B$ are of finite type, and since $A+B$ has a Riesz basis of subspaces in H, we know that H decomposes into a direct sum :

$$Z_H = \oplus_{i\geq 1}\text{Ker } (A+B-\lambda_i\text{Id})^{m_i} \tag{2.3.9}$$

with distinct numbers λ_i. The symbol $\oplus$ means here that we have

$$\text{Ker } (A+B-\lambda_i\text{Id})^{m_i} = \text{Vect }_{1\leq l\leq n_i}(e_{i,l}),$$

where the whole sequence $(e_{i,l})$ forms a Riesz basis of H. For further use, we denote by π_λ the projection onto

$$E_\lambda := \oplus_{\substack{i\geq 1\\ \lambda_i\neq\lambda}} \text{Ker } (A+B-\lambda_i\text{Id})^{m_i},$$

according to the decomposition (2.3.9). Now, since Z_L is a finite codimensional space generated by generalized eigenvectors of $A+B$, we may assume, by "diminishing" Z_L if necessary, that Z_L is of the form

$$Z_L = \oplus_{i\geq r}\text{Ker } (A+B-\lambda_i\text{Id})^{m_i}$$

with some integer r (this only weakens our assumption concerning the estimates (2.2.1) and (2.2.3)).

Thanks to Proposition 2.2.4, since $A+B$ has a Riesz basis of subspaces in H of generalized eigenvectors, there exists $r'\geq r$, such that (2.2.9) holds true with

$$Z_{L'} = \oplus_{i\geq r'}\text{Ker } (A+B-\lambda_i\text{Id})^{m_i},$$

which means here that we have

$$\text{Ker } (A+B-\lambda_i\text{Id})^{m_i} = \text{Vect }_{1\leq l\leq n_i}(e_{i,l}),$$

where the whole sequence $(e_{i,l})_{\substack{i\geq r'\\ 1\leq l\leq n_i}}$ forms a Riesz basis of L'.

In order to prove the theorem, we will use a transformation due to Haraux [19] : given $\delta>0$, $\lambda\in\mathbb{C}$ and $x_0\in Z_H$, set

$$I_{\delta,\lambda}(x_0) := x_0 - \frac{1}{\delta}\int_0^\delta e^{-\lambda s}S_{A+B}(s)x_0 ds.$$

We first recall some properties of this transformation.

Lemma 2.3.1.

(a) $I_{\delta,\lambda}S_{A+B}(t)x_0 = S_{A+B}(t)I_{\delta,\lambda}x_0$.

(b) For any seminorm p in Z_H, and for any interval (a,b) we have the estimates

$$\int_a^b p(I_{\delta,\lambda}S_{A+B}(t)x_0)^2\ dt \le c\int_a^{b+\delta} p(S_{A+B}(t)x_0)^2\ dt, \quad \textit{for all } x_0 \in Z_H. \tag{2.3.10}$$

(c) For any $m \in \mathbb{N}^*$, we have the inclusion :

$$I_{\delta,\lambda}\left(\mathrm{Ker}(A+B-\lambda\mathrm{Id})^m\right) \subset \mathrm{Ker}(A+B-\lambda\mathrm{Id})^{m-1}. \tag{2.3.11}$$

Proof.

(a) By uniqueness of the Cauchy problem.

(b) For every fixed $t \in \mathbb{R}$, by setting $x(t) = S_{A+B}(t)x_0$, we have

$$\begin{aligned}
p(I_{\delta,\lambda}x(t))^2 &\le 2p(x(t))^2 + 2p\Big(\frac{1}{\delta}\int_0^\delta e^{-\lambda s}x(t+s)\ ds\Big)^2 \\
&\le 2p(x(t))^2 + \frac{2}{\delta^2}\Big(\int_0^\delta e^{-\lambda s}p(x(t+s))\ ds\Big)^2 \\
&\le 2p(x(t))^2 + \frac{2}{\delta^2}\int_0^\delta |e^{-\lambda s}|^2\ ds\int_0^\delta p(x(t+s))^2\ ds \\
&\le 2p(x(t))^2 + 2\delta^{-1}e^{2|\Re\lambda|\delta}\int_t^{t+\delta} p(x(s))^2\ ds.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_a^b p(I_{\delta,\lambda}x(t))^2\ dt \\
&\quad\le 2\int_a^b p(x(t))^2\ dt + 2\delta^{-1}e^{2|\Re\lambda|\delta}\int_a^b\int_t^{t+\delta} p(x(s))^2\ ds\ dt \\
&\quad= 2\int_a^b p(x(t))^2\ dt + 2\delta^{-1}e^{2|\Re\lambda|\delta}\int_{a-\delta}^{b+\delta}\int_{\max\{a,s-\delta\}}^{\min\{b,s\}} p(x(s))^2\ dt\ ds \\
&\quad\le 2\int_a^b p(x(t))^2\ dt + 2e^{2|\Re\lambda|\delta}\int_{a-\delta}^{b+\delta} p(x(s))^2\ dt,
\end{aligned}$$

and (2.3.20) follows with

$$c = 2 + 2e^{2|\Re\lambda|\delta}.$$

(c) Let $x_0 \in \mathrm{Ker}(A+B-\lambda\mathrm{Id})^m$. Then we have

$$S_{A+B}(t)x_0 = \sum_{j=0}^{m-1}\frac{t^j e^{\lambda t}}{j!}(A+B-\lambda Id)^j x_0,$$

and thus

$$I_{\delta,\lambda}x_0 = \frac{-1}{\delta}\sum_{j=1}^{m-1}\int_0^\delta t^j dt(A+B-\lambda Id)^j x_0,$$

so that

$$(A+B-\lambda Id)^{m-1} I_{\delta,\lambda}x_0 = 0.$$

□

We now prove a deeper property of the Haraux transformation.

Lemma 2.3.2. *For all but countably many $\delta > 0$, we have*

$$\|\pi_\lambda x_0\|^2 \le c\,\|\pi_\lambda I_{\delta,\lambda}(x_0)\|^2, \quad \textit{for all } x_0 \textit{ in } Z_H \tag{2.3.12}$$

Proof. We fix an integer r'' which will be chosen later and we suppose at first that $x_0 \in Z_{L''} := \oplus_{i\ge r''}\mathrm{Ker}\ (A+B-\lambda_i \mathrm{Id})^{m_i}$. We know that A is a skew-adjoint operator having a compact resolvant, thus, we can fix an orthonormal basis $(e_k)_{k\ge1}$ of eigenvectors for A, with purely imaginary eigenvalues μ_k which tend to infinity. We construct a sequence (ε_k) which tends to zero and such that all numbers $\mu_k+\varepsilon_k$ are distinct from λ, and we define a closed operator B_0 by $B_0 e_k = \varepsilon_k e_k$. Now,we have $x_0 = \sum x_k e_k$ and we introduce the Haraux transformation for $A+B_0$:

$$J_{\delta,\lambda}(x_0) := x_0 - \frac{1}{\delta}\int_0^\delta e^{-\lambda s} S_{A+B_0}(s)x_0 ds = \sum x_k a(k,\delta)e_k,$$

with,

$$a(k,\delta) := 1 - \frac{1}{\delta}\int_0^\delta e^{(\mu_k+\varepsilon_k-\lambda)s} ds.$$

The quantity $a(k,\delta)$ tends to 1 as k tends to infinity, and the set of the δ such that there exist $k\in\mathbb{N}$ cancelling $|a(k,\delta)|$ is countable, since $a(k,\delta)$ is analytic in $\delta>0$. Thus for all but countably many $\delta>0$, $\inf_{k\in\mathbb{N}}|a(k,\delta)|$ is strictly positive and thus

$$\|x_0\|^2 \le c\,\|J_{\delta,\lambda}(x_0)\|^2. \tag{2.3.13}$$

Now we have :

$$S_{A+B_0}(t)x_0 - S_{A+B}(t)x_0 = \int_0^t S_{A+B_0}(t-s)(B_0-B)S_{A+B}(s)x_0 ds.$$

Hence, by the Cauchy-Schwarz inequality, we obtain

$$\|J_{\delta,\lambda}(x_0) - I_{\delta,\lambda}(x_0)\|^2 \le \frac{1}{\delta^2}\int_0^\delta e^{-2\Re(\lambda)t}dt\int_0^\delta \left\|\int_0^t S_{A+B_0}(t-s)(B_0-B)S_{A+B}(s)x_0 ds\right\|^2 dt.$$

and thus,

$$\|J_{\delta,\lambda}(x_0) - I_{\delta,\lambda}(x_0)\|^2 \leq c \int_0^\delta \left(\sup_{x \in L'', |x| \leq 1} \|(B - B_0) S_{A+B}(s) x\| \right)^2 ds \, \|x_0\|^2 \quad (2.3.14)$$

Now, collecting (2.3.13) and (2.3.14), we obtain :

$$\|x_0\|^2 \leq c \, \|J_{\delta,\lambda}(x_0)\|^2 \leq 2c \, \|I_{\delta,\lambda}(x_0)\|^2 + 2c \, \|J_{\delta,\lambda}(x_0) - I_{\delta,\lambda}(x_0)\|^2 ,$$

Now, since $A + B$ has a Riesz basis of subspaces in H of generalized eigenvectors, by proceeding like in the preceding proof, we can choose r'', such that

$$\|x_0\|^2 \leq c \, \|I_{\delta,\lambda}(x_0)\|^2 .$$

By increasing r'', if necessary, since λ_i tend to infinity, because B is compact and A has a compact resolvant, we can suppose that $\lambda \neq \lambda_i$ for $i \geq r''$. Thus, for all $x_0 \in Z_{L''}$, $x_0 \in E_\lambda$, $I_{\delta,\lambda} x_0 \in E_\lambda$ and the preceding inequality reduces to (2.3.12). Now, let $z_0 = x_0 + y_0 \in Z_H$ with

$$x_0 \in Z_{L''} \quad \text{and} \quad y_0 \in \oplus_{j<r''} \mathrm{Ker} \, (A + B - \lambda_j \mathrm{Id})^{m_j}$$

Suppose once that

$$\|\pi_\lambda y_0\|^2 \leq c \, \|\pi_\lambda I_{\delta,\lambda}(y_0)\|^2 . \quad (2.3.15)$$

We then obtain the inequality

$$\|\pi_\lambda z_0\|^2 \leq c \, \|\pi_\lambda I_{\delta,\lambda}(x_0)\|^2 + c \, \|\pi_\lambda I_{\delta,\lambda}(y_0)\|^2 .$$

By the tool of a Riesz basis such that some of its members generate

$$\oplus_{j<r''} \mathrm{Ker} \, (A + B - \lambda_j \mathrm{Id})^{m_j}$$

and the others $Z_{L''}$, we obtain :

$$\|\pi_\lambda z_0\|^2 \leq c \, \|\pi_\lambda I_{\delta,\lambda}(z_0)\|^2$$

Now, it remains to prove (2.3.15). Since $\oplus_{j<r''} \mathrm{Ker} \, (A+B-\lambda_j \mathrm{Id})^{m_j}$ is a finite dimensional space, it suffices to verify that

$$\pi_\lambda I_{\delta,\lambda}(y_0) = 0 \Rightarrow \pi_\lambda y_0 = 0. \quad (2.3.16)$$

for all but countably many $\delta > 0$. For proving this, we can restrict us to the case where $y_0 \in \mathrm{Ker}(A + B - \lambda_i \mathrm{Id})^k$, with $\lambda \neq \lambda_i$. Now, if (2.3.16) is not true, we would have

$$\det(\mathrm{Id} - \sum_{j=0}^{k-1} \frac{G_j(\delta)}{j!} (A + B - \lambda \mathrm{Id})^j) = 0, \quad (2.3.17)$$

for uncountably many $\delta > 0$, and thus for $\delta \in \mathbb{C}$, since

$$G_j(\delta) := \frac{1}{\delta} \int_0^\delta e^{(\lambda_i - \lambda)s} s^j \mathrm{d}s, \tag{2.3.18}$$

and thus the left hand side of (2.3.17) are analytic in δ. By taking $\delta = -\tilde{\delta}\overline{\lambda - \lambda_i}$, and letting $\tilde{\delta} \in \mathbb{R}^+$ tend to ∞, the left hand side of (2.3.17) tends to $\det(\mathrm{Id}) = 1 \neq 0$, which yields to a contradiction.

□

We now can prove a weaker form of the estimate (2.1.1).

Lemma 2.3.3. *Set*

$$\pi := \prod_{i=1}^{r} \pi_{\lambda_i}$$

Then

$$c \left\| \pi x_0 \right\|^2 \leq \sum_{j=1}^{m} \int_{J_j} p_j (S_{A+B}(t) x_0)^2 \ dt \ \textit{for all} \ x_0 \in Z_H$$

Proof. Set

$$M = \sum_{k<r'} m_k$$

and fix a sufficiently small $\delta > 0$ so that writing $I_j = (a_j, b_j)$ we have

$$(a_j - M\delta, b_j + M\delta) \subset J_j \quad \text{for} \quad j = 1, \ldots, m.$$

We can choose δ such that the estimate (2.3.12) of the lemma 2.3.2 is satisfied for every λ_k with $k < k'$. Let us introduce the linear operator

$$I = \prod_{k<r'} I_{\delta,\lambda_k}^{m_k}$$

(composition of M linear operators). It follows from the definition of $I_{\delta,\lambda}$ that the factors I_{δ,λ_k} and π_{λ_k} commute. Hence, by a repeated application of the lemma 2.3.2 we obtain that

$$\left\| \pi x_0 \right\|^2 \leq c \left\| \pi I(x_0) \right\|^2 \tag{2.3.19}$$

and on the other hand, by a repeated application of (2.3.10), we obtain :

$$\sum_{j=1}^{m} \int_{I_j} p_j (S_{A+B}(t) I x_0)^2 dt \leq c \sum_{j=1}^{m} \int_{J_j} p_j (S_{A+B}(t) x_0)^2 dt, \quad \forall x_0 \in Z_H \tag{2.3.20}$$

It turns out by a repeated application of (2.3.11) that $I(x_0) \in Z_{L'}$. It follows that $\pi I(x_0) = I(x_0)$ and that (2.2.9) holds true. Thus, we have :

$$c \left\| \pi I(x_0) \right\|^2 = c \left\| I(x_0) \right\|^2 \leq \sum_{j=1}^{m} \int_{I_j} p_j (S_{A+B}(t) I x_0)^2 dt$$

By collecting this, (2.3.19) and (2.3.20), we obtain the result. □

Now we are ready to prove our main theorem.

proof of theorem 2.2.5. We first show that M is finite dimensional. Let $x_0 \in Z_H$ satisfying (2.2.5) and (2.2.6). Thanks to (2.2.6), there exists an integer i such that $x_0 \in \mathrm{Ker}\ (\mathrm{A}+\mathrm{B}-\lambda_{\mathrm{i}}\mathrm{Id}\ )_{\mathrm{i}}^{\mathrm{m}}$. Since (2.2.5) holds, according to lemma 2.2.1, (2.1.1) doesn't hold. Therefore, from (2.2.9), we must have $i < r'$. We then see that M is included in $M' := \oplus_{i<r'}\mathrm{Ker}(A+B-\lambda_i\mathrm{Id})^{m_i}$ and is therefore finite dimensional.

We now fix a supplement S of M in M' and take $Z_{\widetilde{L}} = S \oplus Z_{L'}$ Let $x_0 = y_0 + z_0 \in Z_{\widetilde{L}}$, with $y_0 \in S$ and $z_0 \in Z_{L'}$.

Assume for a moment that

$$\|y_0\|^2 \leq c\sum_{j=1}^{m}\int_{I_j} p_j(S_{A+B}(t)y_0)^2\ dt. \tag{2.3.21}$$

Then

$$\begin{aligned}\|x_0\|^2 &\leq 2\|y_0\|^2 + 2\|z_0\|^2\\ &\leq c\sum_{j=1}^{m}\int_{I_j} p_j(S_{A+B}(t)y_0)^2\ dt + 2\|z_0\|^2\\ &\leq c\sum_{j=1}^{m}\int_{I_j} 2p_j(S_{A+B}(t)x_0)^2 + 2p_j(S_{A+B}(t)z_0)^2\ dt + 2\|z_0\|^2.\end{aligned}$$

(We used in the first step the triangle inequality.) Applying (2.2.9), for z_0, it follows that

$$\|x_0\|^2 \leq c\sum_{j=1}^{m}\int_{I_j} p_j(S_{A+B}(t)x_0)^2\ dt + c\|z_0\|^2.$$

Applying the preceding lemma, since $\pi x_0 = z_0$, we conclude that

$$\|x_0\|^2 \leq c\sum_{j=1}^{m}\int_{J_j} p_j(S_{A+B}(t)x_0)^2\ dt.$$

It remains to prove (2.3.21). We can restrict us to the case where there is a single interval I_j, and $0 \in I_j$. Since $\oplus_{i<r'}\mathrm{Ker}(A+B-\lambda_i\mathrm{Id})^{m_i}$ is finite dimensional, it suffices to prove that

$$p_j(S_{A+B}(t)y_0) = 0 \quad \text{in} \quad I_j \Rightarrow y_0 = 0. \tag{2.3.22}$$

So, we suppose that

$$p_j(S_{A+B}(t)y_0) = 0 \quad \text{in} \quad I_j.$$

We fix $0 < \delta < |I_j|$. Thus, we get

$$p_j(I_{\delta,\lambda}y_0) = 0.$$

The solution has the form

$$S_{A+B}(t)y_0 = \sum_{i<r'} \sum_{\ell=0}^{m_i-1} \frac{t^\ell e^{\lambda_i t}}{\ell!}(A+B-\lambda_i Id)^\ell y_{0,i}$$

with $y_{0,i} \in \mathrm{Ker}(A+B-\lambda_i Id)^{m_i}$.

Let $I_{(i)} := \prod_{\substack{k<r' \\ k\neq i}} I_{\delta,\lambda_k}^{m_k}$. We then have :

$$p_j(I_{(i)}y_0) = 0$$

and

$$I_{(i)}y_0 = \sum_{\ell=0}^{m_i-1} \alpha_{i,\ell}(A+B-\lambda_i \mathrm{Id})^\ell y_{0,i}$$

with some numbers $\alpha_{i,\ell}$.

We have more generally :

$$p_j(S_{A+B}(t)I_{(i)}y_0) = 0 \quad \text{in} \quad I_j \quad \text{for}$$

Now let L_i be defined by $L_i y(t) := y'(t) - \lambda_i y(t)$. Then we have :

$$p_j(L_i S_{A+B}(t)I_{(i)}y_0) = 0.$$

Suppose now that $y_{0,i} \neq 0$ and let ℓ_0 be the first index such that $\alpha_{i,\ell_0} \neq 0$. Thus

$$p_j(L_i^{m_i-1-\ell_0} S_{A+B}(t)I_{(i)}y_0) = 0$$

and

$$L_i^{m_i-1-\ell_0} S_{A+B}(t)I_{(i)}y_0 = \alpha_{i,\ell_0}(A+B-\lambda_i Id)^{m_i-1}y_{0,i}.$$

So

$$p_j((A+B-\lambda_i Id)^{m_i-1}y_{0,i}) = 0$$

We go on :

$$p_j(L_i^{m_i-2-\ell_0} S_{A+B}(t)I_{(i)}y_0) = 0$$

and

$$L_i^{m_i-2-\ell_0} S_{A+B}(t)I_{(i)}y_0 = \alpha_{i,\ell_0} t(A+B-\lambda_i Id)^{m_i-2}y_{0,i} + \alpha_{i,\ell_0+1}(A+B-\lambda_i Id)^{m_i-1}y_{0,i};$$

thus

$$p_j((A+B-\lambda_i Id)^{m_i-2}y_{0,i}) = 0.$$

By recurrence, we then obtain

$$p_j((A+B-\lambda_i Id)^k y_{0,i}) = 0, k = 0, 1, \ldots$$

So we conclude that $y_{0,i} \in M$. Thus, y_0 belongs to M. On the other hand, y_0 belongs to S, so $y_0 = 0$ and we have (2.3.22). □

2.4 Application

As an application of our result, we will improve results of [29] and [30], as we shall see in remarque 2.4.9).

Let $\Omega \subset \mathbb{R}^N$ be a bounded open subset of boundary Γ. We fix two integers m and n, numbers $a_1, \ldots, a_{m+n} > 0$ and complex numbers $\alpha_{i,j}$ $(1 \leq i, j \leq m+n)$. We consider the following system :

$$\begin{cases} u_i'' = a_i^2 \Delta u_i - \sum_{j=1}^{m+n} \alpha_{i,j} u_j \ \text{ in } \mathbb{R} \times \Omega, 1 \leq i \leq m, \\ u_i'' = -a_i^2 \Delta^2 u_i - \sum_{j=1}^{m+n} \alpha_{i,j} u_j \text{ in } \mathbb{R} \times \Omega, m < i \leq m+n, \\ u_i = 0 \text{ on } \mathbb{R} \times \Gamma, 1 \leq i \leq m, \\ u_i = \Delta u_i = 0 \text{ on } \mathbb{R} \times \Gamma, m < i \leq m+n, \\ u_i(0) = u_{i0}, u_i'(0) = u_{i1}, \text{ in } \Omega, 1 \leq i \leq m+n. \end{cases} \tag{2.4.1}$$

We can verify by standard methods that, if $(u_{i0}, u_{i1}) \in H_0^1(\Omega) \times L^2(\Omega)$, for $1 \leq i \leq m$, and $(u_{i0}, u_{i1}) \in H_0^1(\Omega) \times H^{-1}(\Omega)$, for $m < i \leq m+n$, then (2.4.1) has a unique weak solution $u = (u_1, ..., u_m, ..., u_{m+n})$ which satisfies :

$$\begin{aligned} &u_i \in C(\mathbb{R}, H_0^1(\Omega)) \cap C^1(\mathbb{R}, L^2(\Omega)), \quad 1 \leq i \leq m. \\ &u_i \in C(\mathbb{R}, H_0^1(\Omega)) \cap C^1(\mathbb{R}, H^{-1}(\Omega)), \quad m < i \leq m+n. \end{aligned}$$

Let E_0 be the *initial energy* of the solution defined by

$$E_0 := \frac{1}{2}\Big(\sum_{i=1}^{m} \|u_{i0}\|_{H_0^1(\Omega)}^2 + \|u_{i1}\|_{L^2(\Omega)}^2 + \sum_{i=m+1}^{m+n} \|u_{i0}\|_{H_0^1(\Omega)}^2 + \|u_{i1}\|_{H^{-1}(\Omega)}^2\Big).$$

$L^2(\Omega)$ and $H_0^1(\Omega)$ are endowed with the norm :

$$\|v\|_{L^2(\Omega)}^2 = \int_\Omega |v|^2 \, dx, \qquad \|v\|_{H_0^1(\Omega)}^2 = \int_\Omega |\nabla v|^2 \, dx$$

and $H^{-1}(\Omega)$ is endowed with the dual norm of $H_0^1(\Omega)$.

We denote by H the underlying Hilbert space :

$$H := H_0^1(\Omega)^m \times L^2(\Omega)^m \times H_0^1(\Omega)^n \times H^{-1}(\Omega)^n.$$

Let ν be the normal exterior unit vector to Γ, and $\Gamma_1, \ldots, \Gamma_{m+n}$ be open subsets of Γ, $\omega_1, \ldots, \omega_{m+n}$ be open subsets of Ω, $I_1, \ldots, I_{m+n}$ intervals of $\mathbb{R}$.

We look for the internal observability estimates :

$$c_1 E_0 \leq \sum_{i=1}^{m+n} \int_{I_i} \int_{\omega_i} |u_i'|^2 \, dx \, dt \leq c_2 E_0, \tag{2.4.2}$$

and the boundary observability estimates :

$$c_1 E_0 \leq \sum_{i=1}^{m+n} \int_{I_i} \int_{\Gamma_i} |\partial_\nu u_i|^2 \, d\Gamma dt \leq c_2 E_0. \tag{2.4.3}$$

Theorem 2.4.1. *We suppose that* (2.4.2), *respectively* (2.4.3), *holds for every solution* u *satisfying* (2.4.1) *with* $\alpha_{i,j} = 0$. *Then, given any other choice of* $\alpha_{i,j}$, *there exists an orthogonal decomposition of the underlying Hilbert space* H *such that*

$$H = M \oplus L$$

with a finite dimensional space M *satisfying the following conditions :*

(i) for all initial data belonging to L, (2.4.2), *respectively* (2.4.3), *holds for a solution* u *satisfying* (2.4.1) *with this particular choice of* $\alpha_{i,j}$, *this initial data, and intervals* J_j *instead of* I_j, J_j *containing the closure of* I_j *in its interior ;*

(ii) for all initial data belonging to $M \backslash \{0\}$, (2.4.2), *respectively* (2.4.3), *doesn't hold for any solution* u *satisfying* (2.4.1) *with the same choice of* $\alpha_{i,j}$, *and this other initial data.*

Proof of theorem 2.4.1. We rewrite the problem (2.4.1) in the form

$$\begin{aligned} y' &= (A + B)y, \\ y(0) &= y_0 \end{aligned}$$

with

$$y = (u_1, ..., u_m, u_1', ..., u_m', u_{m+1}, ..., u_{m+n}, u_{m+1}', ..., u_{m+n}')$$

and A corresponding to the case $\alpha_{i,j} = 0$.

B then is a compact perturbation of A and A is a skew adjoint operator having a compact resolvent and it generates a group.

Set z_k be an orthonormal basis in $L^2(\Omega)$, satisfying

$$\begin{aligned} -\Delta z_k &= \gamma_k^2 z_k \quad \text{in } \Omega, \\ z_k &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Since $Z_k := \{\beta \cdot z_k, \beta \in \mathbb{C}^{2m+2n}\}$ is stable by $A+B$, we obtain a Riesz basis of subspaces generated by generalized eigenvectors for $A+B$ and we thus can apply the abstract theorem with

$$p_j(x) := \left\|x_j'\right\|_{L^2(\omega_i)},$$

in the case of internal observability, and

$$p_j(x) := \left\|\partial_\nu x_j\right\|_{L^2(\gamma_i)},$$

in the case of boundary observability, for all $j = 1, \ldots, m+n$. □

Example. Let us give a concrete example when the compactly perturbed operator $A+B$ does not have a Riesz basis of eigenvectors. Choosing

$$m = 3, \quad n = 0, \quad a_1 = 2 < a_2 = a_3 = 4.$$

$$(\alpha_{i,j}) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

the eigenvectors of $A+B$ are given up to a multiplicative factor by the following formulae

$$\begin{aligned}
e^+_{k,1} &= (1, 2\gamma_k^2, -1, 2i\gamma_k, 4i\gamma_k^3, -2i\gamma_k)z_k, \\
e^+_{k,2} &= (\delta_k, -1, 0, \lambda_k\delta_k, -\lambda_k, 0)z_k, \\
e^+_{k,3} &= (\delta_k^{-1}, 1, 0, \mu_k\delta_k^{-1}, \mu_k, 0)z_k, \\
e^-_{k,1} &= (1, 2\gamma_k^2, -1, -2i\gamma_k, -4i\gamma_k^3, 2i\gamma_k)z_k, \\
e^-_{k,2} &= (\delta_k, -1, 0, -\lambda_k\delta_k, \lambda_k, 0)z_k, \\
e^-_{k,3} &= (\delta_k^{-1}, 1, 0, -\mu_k\delta_k^{-1}, -\mu_k, 0)z_k,
\end{aligned}$$

where we put :

$$\lambda_k := \sqrt{\left|3\gamma_k^2 - \sqrt{\gamma_k^4+1}\right|}, \quad \text{if} \quad 3\gamma_k^2 - \sqrt{\gamma_k^4+1} < 0,$$

$$\lambda_k := i\sqrt{3\gamma_k^2 - \sqrt{\gamma_k^4+1}}, \quad \text{if} \quad 3\gamma_k^2 - \sqrt{\gamma_k^4+1} \geq 0,$$

$$\mu_k := i\sqrt{3\gamma_k^2 + \sqrt{\gamma_k^4+1}},$$

and

$$\delta_k := \gamma_k^2 + \sqrt{\gamma_k^4+1}$$

for brevity.

Since for example

$$\frac{(e^+_{k,1}|e^+_{k,3})}{\left\|e^+_{k,1}\right\|\left\|e^+_{k,3}\right\|} \to 1,$$

they cannot be normalized so as to form a Riesz basis.

One interesting question, now, is to determine the dimension of the parameters $\alpha_{i,j}$ for which we do not have observability, i.e., for which $M \neq \{0\}$.

Concerning internal observability, we have the following proposition :

Proposition 2.4.2. *The parameters for which $M \neq \{0\}$ form a countable union of hypersurfaces ; hence their set has zero Lebesgue measure.*

Remark 2.4.3. *These special parameters correspond exactly to those which ensure the existence of constant solutions different from zero ; in order not to have such parameters, we must observe*

$$\sum_{i=1}^{m+n} \int_{I_i}\int_{\omega_i} |u_i'|^2\, dx\ dt + \sum_{i=1}^{m+n} \int_{I_i}\int_{\omega_i} |u_i|^2\, dx\ dt$$

instead of

$$\sum_{i=1}^{m+n} \int_{I_i}\int_{\omega_i} |u_i'|^2\, dx\ dt.$$

Proof. We distinguish two cases. If 0 is not an eigenvalue of $A+B$, then it follows from the structure of $A+B$ that every eigenvector of $A+B$ with eigenvalue λ has the form

$$e = \beta^1 z_1 + \cdots + \beta^k z_k, \tag{2.4.4}$$

with a minimal k , where

$$\begin{gathered} z_l \in H_0^1(\Omega),\ z_l \neq 0, \\ -\Delta z_l = \gamma_l^2 z_l \quad \text{in} \quad \Omega, \\ \beta^l \in \mathbb{C}^{2m+2n} \quad \text{with} \\ \beta^l = \left(\beta^l_1, \ldots, \beta^l_m, \beta'^l_1, \ldots, \beta'^l_m, \beta^l_{m+1}, \ldots, \beta^l_{m+n}, \beta'^l_{m+1}, \ldots, \beta'^l_{m+n}\right) \\ \beta'^l_j = \lambda \beta^l_j, \quad j = 1, \ldots, m+n. \end{gathered}$$

We may assume that z_1,..., z_k are linearly independent. We may also assume that the β^ℓ associated with the same γ_j are linearly independent. Otherwise, we can diminish k. Indeed, if, for example $\gamma_1 = \gamma_2 = \gamma_3$, and $\beta^3 = \beta^1 + \beta^2$, we have

$$\beta^1 z_1 + \beta^2 z_2 + \beta^3 z_3 = \beta^1(z_1 + z_3) + \beta^2(z_2 + z_3)$$

and, since $z_1 + z_3$ and $z_2 + z_3$ remain independent and satisfy (2.4)-(2.4), we can use the vectors $z_1 + z_3$ and $z_2 + z_3$ in (2.4.4) instead of z_1, z_2, z_3 : that is, we diminish k. So, since 0 is not an eigenvalue, we have the equivalence :

$$\beta^l_1 = \cdots = \beta^l_{m+n} = 0 \Longleftrightarrow \beta'^l_1 = \cdots = \beta'^l_{m+n} = 0. \tag{2.4.5}$$

If $p_1(e) = \cdots = p_{m+n}(e) = 0$, then

$$\beta_j^1 z_1 + \cdots + \beta_j^k z_k = 0 \quad \text{in} \quad \omega_j, \quad 1 \le j \le m+n.$$

Applying $-\Delta$ repeatedly to these equations, we obtain for each $1 \le j \le m+n$ the linear system

$$(\gamma_1^2)^i \beta_j^1 z_1 + \cdots + (\gamma_k^2)^i \beta_j^k z_k = 0 \quad \text{in} \quad \omega_j, \quad i = 0, \ldots, k-1$$

for the variables $\beta_j^1 z_1$,..., $\beta_j^k z_k$.

If the numbers γ_l are pairwise distinct, the determinant of this system is different from zero, and therefore

$$\beta_j^1 z_1 = \cdots = \beta_j^k z_k = 0 \quad \text{in} \quad \omega_j, \quad 1 \le j \le m.$$

In the general case, we only obtain for every $\gamma > 0$ the equality

$$\sum_{\gamma_\ell = \gamma} \beta_j^\ell z_\ell = 0 \quad \text{in} \quad \omega_j, \quad 1 \le j \le m+n.$$

Now, for each $j = 1, \ldots, m$, putting $u_j(t) = e^{i\gamma_\ell t} \sum_{\gamma_\ell = \gamma} \beta_j^\ell z_\ell$ and $u_p(t) = 0$ for all other $1 \le p \le m+n$, we obtain a solution of (2.4.1) in the uncoupled case $\alpha_{i,j} = 0$. Hence, applying the hypothesis we conclude that

$$\sum_{\gamma_\ell = \gamma} \beta_j^\ell z_\ell = 0 \text{ in } \Omega.$$

We obtain the same conclusion for $j = m+1, \ldots, m+n$ by changing γ_ℓ to γ_ℓ^2 in the definition of $u_j(t)$ above. Since z_1,..., z_k are linearly independent, it follows that

$$\beta_j^1 = \cdots = \beta_j^k = 0.$$

Using (2.4.5) hence we conclude that $e = 0$, which implies, by (2.2.6) that $M = \{0\}$. Now, suppose that 0 is an eigenvalue of $A + B$ and let y_0 be a corresponding nonzero eigenvector. Then the constant function $y(t) := y_0$ solves (2.4.1) and $p_j(y(t)) \equiv 0$ for all $j = 1, \ldots, m+n$. Thus $M \neq \{0\}$.

It remains to prove that the parameters $\alpha_{i,j}$, for which 0 is an eigenvalue of $A+B$ form a countable union E of surfaces of codimension 1. In fact E consists of all matrices $(\alpha_{i,j})$ such that 0 is an eigenvalue of $A + B|_{Z_k}$ for some k, because the subspaces $Z_k, \quad (k = 1, 2, \ldots)$ are stable by $A + B$ and that determines, for some k a hypersurface in $\mathbb{C}^{(m+n)^2}$.

□

Now, consider the case of boundary observability.

Proposition 2.4.4. *The parameters for which $M \neq \{0\}$ are contained in countably many surfaces of codimension $n+m$ of $\mathbb{C}^{(m+n)^2}$.*

Proof. We suppose that $M \neq \{0\}$. We fix an orthonormal basis of the Laplacien-Dirichlet operator. So, keeping in mind the preceding proof, we can find an integer k, and k elements $z_1, \ldots, z_k$ of the fixed orthonormal basis and k nonzero elements $\beta^1, \ldots, \beta^k \in \mathbb{C}^{m+n}$ such that

$$\beta_j^1 \partial_\nu z_1 + \ldots \beta_j^k \partial_\nu z_k = 0, \quad \text{on } \Gamma_j, \text{for } j = 1, \ldots, m+n. \tag{2.4.6}$$

The vectors $\beta^1, \ldots, \beta^k$ also have to satisfy the relations :

$$\Big((\alpha_{i,j}) - \lambda^2 I_{m+n}\Big)\beta^\ell = aG_\ell\beta^\ell \text{for } \ell = 1, \ldots, k, \tag{2.4.7}$$

with

$$a = \begin{pmatrix} a_1^2 & & \\ & \ldots & \\ & & a_{m+n}^2 \end{pmatrix}, \quad G_\ell = \begin{pmatrix} \gamma_\ell^2 I_m & \\ & \gamma_\ell^4 I_n \end{pmatrix}.$$

We keep here the notations of the preceding proof for the definition of γ_ℓ. Suppose once that, for this given sequence $z_1, \ldots, z_k$, the parameters $c_{i,j}$ defined by

$$C := (c_{i,j}) := a^{-1}\left(\alpha_{i,j} - \lambda^2 I_{m+n}\right)$$

are described by at most $(m+n)^2 - (m+n+1)$ parameters (*). Then, we sum over all the countable sequences $z_1, \ldots, z_k$ and we add the the parameter λ to describe all the parameters $\alpha_{i,j}$. So, if we prove (*), we conclude that the exceptional parameters are contained in countably many surfaces of dimension less than or equal $(m+n)^2 - (m+n+1) + 1$, that is, of codimension superior or equal to $m+n$. It remains now to prove (*) ; we distinguish two cases.

Suppose that the vectors

$$\beta^1, \ldots, \beta^k$$

form a free family. For each $j = 1, \ldots, m+n$, there exists a point $x \in \Gamma_j$ where $\partial_\nu z_1(x) \neq 0$ by our hypothesis of observability in the uncoupled case. This allows us to express β_j^1 by the variables $\beta_j^2, \ldots, \beta_j^k$, via the equation (2.4.6). On the other hand, we can suppose that $\beta_1^2 \in \{0, 1\}$ by dividing all the equations (2.4.6) and (2.4.7) by β_1^2, if necessary. This doesn't change the definition of the parameters $c_{i,j}$. Hence, the set of parameters (β_j^ℓ) is described by at most $(k-1)(m+n) - 1$ parameters. For each such choice of the vectors (β_j^ℓ), the parameters $(c_{i,j})$ are the solutions of the linear system

$$C\beta^\ell = G_\ell\beta^\ell, \ell = 1, \ldots, k, \tag{2.4.8}$$

which is the union of $m+n$ uncoupled linear systems

$$c_{i,1}\beta_1^\ell + \cdots + c_{i,m+n}\beta_{m+n}^\ell = \begin{cases} \gamma_\ell^2\beta_i^\ell, & i \leq m \\ \gamma_\ell^4\beta_i^\ell, & i > m \end{cases}, \quad \ell = 1, \ldots, k$$

of rank k for each $i = 1, \dots, m+n$. It follows that the parameters $(c_{i,j})$ form an affine subspace described by $(m+n)(m+n-k)$ parameters. Summarizing, the parameters $(c_{i,j})$ are given by at most

$$(k-1)(m+n) - 1 + (m+n)(m+n-k) = (m+n)^2 - (m+n+1)$$

parameters.

Suppose now, that the vectors $\beta^1, \dots, \beta^k$ are linked and consider a relation with a minimum of indices, say $1, \dots, r+1$, by rearrangering the indices if necessary (r is less than or equal to the rank of the system of vectors). We recall that the β^ℓ associated with the same γ_j are linearly independent. Thus, by rearrangering again the indices, we may assume that $\gamma_{r+1} \neq \gamma_1$. In order to determine the parameters $c_{i,j}$, we just consider the relations (2.4.7) for $\ell = 1, \dots, r+1$. (In reality, the $c_{i,j}$ should also satisfy the other relations from (2.4.6) and (2.4.7), but that will diminish the numbers of parameters which give the $c_{i,j}$ still further). Now we can suppose that

$$\beta^{r+1} = \beta^1 + \cdots + \beta^r, \tag{2.4.9}$$

by multiplying each relation (2.4.7) for $\ell = 1, \dots, r+1$ by a suitable multiplicative factor.

From this, we also can suppose that $\beta_1^2 \in \{0, 1\}$. Indeed, we only have to divide all the relations we need (i.e : (2.4.7) for $\ell = 1, \dots, r+1$ and (2.4.9)) by β_1^2, if necessary. Again, this doesn't change the definition of the $c_{i,j}$. So, we first choose the $(r-1)(n+m) - 1$ parameters for $\beta^2, \dots, \beta^r$. Then, since $G_{r+1} \neq G_1$ holds, β^1 is determined by the compatibility condition :

$$(G_{r+1} - G_1)\beta^1 + \cdots + (G_{r+1} - G_r)\beta^r = 0,$$

from (2.4.7). Here, we have implicitely supposed that $r \geq 2$. In fact r cannot be equal to 1, according to the preceding equality. Hence, the set of parameters (β_j^ℓ) is described by at most $(r-1)(m+n) - 1$ parameters. In each such (β_j^ℓ), the parameters $(c_{i,j})$ are the solutions of the linear system (2.4.8) with $k = r$. Repeating the above arguments, we obtain $(r-1)(m+n) - 1 - (m+n)(m+n-r) = (m+n)^2 - (m+n+1)$ again.

□

Remark 2.4.5. *We have considered the case where the parameters $\alpha_{i,j}$ belong to $\mathbb{C}$, in order to simplify the discussion. In the real case, the proof is identical for the real eigenvalues λ ; the management of the complex eigenvalues seems to complicate the situation.*

Now, if we do not couple the Petrovsky and wave systems, and if we observe in a common region for all the equations, there are not exceptional parameters :

Proposition 2.4.6. *If $n = 0$ or $m = 0$ and if $\bigcap_1^{m+n} \Gamma_i$ has nonempty interior, then there are no parameters for which $M \neq \{0\}$.*

Proof. The condition of the intersection ensures that β^i are linked. On the other hand, we may suppose, following the proof of the last proposition, that the β^ℓ corresponding to the same γ_i are independent. In fact, even the vectors β^ℓ corresponding to different γ_i are independent. Indeed, G_ℓ is a multiple of the identity matrix and therefore the β^ℓ are now eigenvectors corresponding to different eigenvalues and have no other choice than being independent. So the β^i cannot be linked, that is : there is no exceptional parameters. □

Remark 2.4.7. *If $\bigcap_1^{m+n}\Gamma_i$ has empty interior, then there may exist special parameters. For example, consider the case : $n=2,\quad m=0,\quad N=1,\quad \Omega=]0,\pi[,\quad \Gamma_1=\{0\},\quad \Gamma_2=\{\pi\},\quad a_1=a_2=1$. We then have $u_1=2sinx+sin2x,\quad u_2=2sinx-sin2x$ satisfy the system (2.4.1) with $\alpha_{1,1}=\alpha_{2,2}=\frac{5}{2}$ and $\alpha_{2,1}=\alpha_{1,2}=-\frac{3}{2}$, and $\partial_\nu u_1(0,t)=\partial_\nu u_2(\pi,t)=0$*

Now we look at the special case where Ω is a ball.

Proposition 2.4.8. *We suppose that Ω is a ball. Then, if $n\geq 1$ and $m\geq 1$, the parameters for which $M\neq\{0\}$ contain countably many surfaces of codimension $m+n$.*

Proof. If Ω is a ball, we recall that each eigenfunction of the Laplacian-Dirichlet operator is given by the product of a radius function with an hyperspherical harmonic, and for each such hyperspherical harmonic, there exist countably many independent eigenfunctions of the Laplacian-Dirichlet operator. Thus, we can choose $n+m+1$ eigenfunctions z_k corresponding to different γ_k such that the $\partial_\nu z_k$ are colinear on $\partial\Omega$. So, the set of the exceptional values contains the set $\mathcal{E}$ of the parameters $\alpha_{i,j}$ such that there exists $\mu\in\mathbb{C}$

Indeed, if these equations are satisfied, we can choose $m+n+1$ nonzero vectors $\beta^1,\ldots,\beta^{m+n+1}$ which agree with (2.4.7). Now, these $m+n+1$ vectors of $\mathbb{C}^{m+n}$ are automatically linked, and thanks to the colinearity of the $\partial_\nu z_k$ on $\partial\Omega$„ the other condition (2.4.6) is also satisfied.

Now, it remains to prove that the set $\mathcal{E}$ contain a variety of codimension $m+n$ (*). Suppose at first that the set E of the parameters $\alpha_{i,j}$ such that

$$\det\left((\alpha_{\mathrm{i,j}})-\mathrm{aG}_\ell\right)=0\quad\text{for }\ell=1,\ldots,m+n+1$$

contains non isolated points(**). Then, if these $n+m+1$ equations are independent, that is, if the differentials of the functions defining these equations evaluated at some point of E are independent linear forms, then E is a variety of dimension $(n+m)^2-(n+m+1)$. In the general case, we can consider a non isolated point x_0 of E where the rank of these linear forms is maximal (we take the maximum along all the non isolated points of E). Then the rank r remains constant in a neighborhood of x_0, because x_0 is not isolated, and E will contain a variety of codimension r, thanks to the constant rank theorem; thus, in any case, E contains a variety of codimension $m+n+1$. Now, each element of $\mathcal{E}$ is the sum of an element of E and an arbitrary multiple of the identity, say μI_{m+n}. So, in order to prove (*), we must prove in a way that the parameter μ is independent of

$(n+m)^2-(n+m+1)$ parameters which defines the variety of codimension $n+m+1$ included in E. So, if we can choose a non isolated point x in E such that I_{m+n} (which represents a tangent vector corresponding to the parameter μ) does not belong to the tangent space of E at the point x, then the tangent space of $\mathcal{E}$ at the point x will be of enough dimension to have (*) and (**) at the same time. So the proposition will be proved if we find an example of such x. Following the case $n=1$ and $m=1$ in [29], we can find $\alpha_{1,1},\alpha_{1,n+1},\alpha_{n+1,1},\alpha_{n+1,n+1}$ such that

$$\det\begin{pmatrix}\alpha_{1,1}-a_1^2\gamma_\ell^2 & \alpha_{1,2}\\ \alpha_{2,1} & \alpha_{2,2}-a_{n+1}^2\gamma_\ell^4\end{pmatrix}=0 \quad \text{for } \ell=1,2,3.$$

and Now, we take for the other parameters : $\alpha_{i,j}=0$ if $i\neq j$,$\alpha_{2,2}=\gamma_4^2,\cdots=\alpha_{n,n}=\gamma_{n+2}^2$ and $\alpha_{n+2,n+2}=\gamma_{n+3}^4,\ldots,\alpha_{n+m}=\gamma_{n+m+1}^4$. We can easily verify that with this choice $x=\alpha_{i,j}$, (2.4) is satisfied and x is also not isolated, since the parameters $\alpha_{1,1},\alpha_{1,n+1},\alpha_{n+1,1},\alpha_{n+1,n+1}$ form a surface of dimension 2. On the other hand, I_{m+n} doesn't belong to tangent space of E. In fact, if it would be the case, we would have :

$$\mathrm{tr}\big(\mathrm{Com}(\mathrm{A}-\mathrm{G}_\ell)\big)=0 \quad \text{for } \ell=1,\ldots,n+m+1.$$

In particular, for $\ell=1,2,3$, we would obtain

$$\alpha_{n+1,n+1}-a_{n+1}^2\gamma_\ell^4+\alpha_{1,1}-a_1^2\gamma_\ell^2,$$

as the γ_ℓ are all distinct and that is impossible.

□

Remark 2.4.9. *Concerning the internal observability, proposition 2.4.2 was established when $m=0$ and $a_1<a_2<\cdots<a_m$ in [30].*
In the boundary case, in comparison to proposition 2.4.4, Komornik et Loreti proved in [32] (and also in [29], when $n=m=1$) that the parameters $\alpha_{i,j}$ for which $M\neq\{0\}$ are contained in countably many hypersurfaces (surfaces of codimension 1), under the condition that $a_1<a_2<\cdots<a_m$ and $a_{m+1}<\ldots a_{m+n}$. Concerning the example of the ball, they also showed in [29] that the parameters for which $M\neq\{0\}$ contain countably many surfaces of codimension 2, when $n=m=1$ and the question of the exact dimension of the set of the "exceptional parameters" was left open.

Chapitre 3

Critical length for a Beurling type theorem

In a recent paper [9] C. Baiocchi, V. Komornik and P. Loreti obtained a generalisation of Parseval's identity by means of divided differences. We give here a proof of the optimality of that theorem.

3.1 Introduction

Let us give a sequence $(\lambda_n)_{n\in\mathbb{Z}}$ of real numbers and a non-degenerated interval I of finite length ($0 < |I| < \infty$). We can define the upper density D^+ by the formula

$$D^+ := \lim_{r\longrightarrow\infty} \frac{n^+(r)}{r},$$

where $n^+(r)$ denotes the biggest number of occurences of the sequence (λ_n) contained in an interval of length r. This limit is well defined (see [9]).

We say that $(\lambda_n)_{n\in\mathbb{Z}}$ is uniformly discrete if it satisfies for a certain $\delta > 0$ the "gap condition"

$$|\lambda_n - \lambda_m| > \delta \quad \text{for all} \quad n \neq m. \tag{3.1.1}$$

Then we have a celebrated theorem which gives the critical length for a generalisation of the Parseval identity (see e.g. [42]) :

Theorem 3.1.1. *Let $(\lambda_n)_{n\in\mathbb{Z}}$ be a uniformly discrete sequence. For I of length $|I| > 2\pi D^+$, $(e^{i\lambda_n \cdot})$ forms a Riesz sequence, that is there exist two constants $c_1, c_2 > 0$, such that*

$$c_1 \sum_{n=-\infty}^{\infty} |b_n|^2 \leq \int_I |f(t)|^2 \mathrm{d}t \leq c_2 \sum_{n=-\infty}^{\infty} |b_n|^2 \tag{3.1.2}$$

for every sum

$$f(t) = \sum_{n=-\infty}^{\infty} b_n e^{i\lambda_n t} \tag{3.1.3}$$

with square summable coefficients b_n. On the other hand, $(e^{i\lambda_n \cdot})$ doesn't form a Riesz sequence anymore if $|I| < 2\pi D^+$.

For certain applications in theoretical control, we have to consider sequences which do not satisfy the gap condition (3.1.1). The question is now the following : what happens if we do not have this condition anymore ? In fact, if $(e^{i\lambda_n \cdot})$ forms a Riesz sequence, then (λ_n) is uniformly discrete (see [42]).

One idea was then to use the divided differences (which will be further explained). This approach was introduced by Ullrich in [43] in some particular cases, then a general answer was given by Baiocchi, Komornik and Loreti in [9]. The theorem takes the form :

Theorem 3.1.2. *If $|I| > 2\pi D^+$, then the divided differences (e_n) form a Riesz sequence.*

The question we will discuss is the following : what happens if $|I| < 2\pi D^+$?

Always in [9], the authors indicated that $2\pi D^+$ is really the optimal length, that is, if $|I| < 2\pi D^+$, the divided differences don't form a Riesz sequence and that one could find a proof by adapting a method of [34]. We propose here to prove this result by adapting a simpler method developed in [17].

3.2 Main result

In order to formulate the result announced in the introduction in a more precise way, we have to define the divided differences. Assume first that $D^+ < \infty$. We can then suppose by a rearranging argument that $(\lambda_n)_{n\in\mathbb{Z}}$ is an increasing sequence. At this stage, we have the following characterization (see [9]) :

Proposition 3.2.1. *Given an increasing sequence (λ_n), we have $D^+ < \infty$ if and only if there exist $\gamma > 0$ and $M > 0$ such that*

$$\lambda_{n+M} - \lambda_n > M\gamma \text{ for every } n \in \mathbb{Z}. \tag{3.2.1}$$

We now define the divided differences like in [9].

For each $j \geq 1$ and all reals $\lambda_1, \ldots, \lambda_j$, we can define :

$$[\lambda_1, \ldots, \lambda_j](t) := (it)^{j-1} \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{j-2}} \tag{3.2.2}$$

$$\exp(i[s_{j-1}(\lambda_j - \lambda_{j-1}) + \cdots + s_1(\lambda_2 - \lambda_1) + \lambda_1)]t)\mathrm{d}s_{j-1}\ldots\mathrm{d}s_1.$$

If $\lambda_1, \ldots, \lambda_j$ are pairwise distinct, then (3.2.2) is equivalent to the familiar definition

$$[\lambda_1, \ldots, \lambda_j](t) := \sum_{p=1}^{j} [\prod_{q=1}^{j}{}' (\lambda_p - \lambda_q)]^{-1} \exp(i\lambda_p t)$$

where the sign $'$ in the products indicates the omission of the zero factor corresponding to $p = q$.

Now let $\gamma > 0$ such that (3.2.1) holds. We fix a number $0 < \gamma' \leq \gamma$. For $j = 1, \ldots, M$ and $m \in \mathbb{Z}$ we say that $\lambda_m, \ldots, \lambda_{m+j-1}$ forms a $\gamma'-$closed exponents chain if

$$\begin{cases} \lambda_m - \lambda_{m-1} \geq \gamma', \\ \lambda_n - \lambda_{n-1} < \gamma' \quad \text{for } n = m+1, \ldots, m+j-1, \\ \lambda_{m+j} - \lambda_{m+j-1} \geq \gamma' \end{cases}$$

For $j = 1, \ldots, M$ and $m \in \mathbb{Z}$ such that $\lambda_m, \ldots, \lambda_{m+j-1}$ forms a $\gamma'-$closed exponents chain, then we define the divided differences by :

$$e_\ell = [\lambda_m, \ldots, \lambda_\ell],$$

for $\ell = m, \ldots, m+j-1$.

Then we verify that e_n is well defined for each $n \in \mathbb{Z}$. The sequence (e_n) is called the sequence of divided differences (relative to γ'), associated to the sequence (λ_n).

We can now formulate the theorem that we want to prove :

Theorem 3.2.2. *Let (λ_n) be a sequence such that we have (3.2.1) for a certain γ. Then for any $0 < \gamma' \leq \gamma$, the sequence of divided differences relative to γ' doesn't form a Riesz sequence in $L^2(I)$ for $|I| < 2\pi D^+$, that is there don't exist constants $c_1, c_2 > 0$, such that*

$$c_1 \sum_{n=-\infty}^{\infty} |a_n|^2 \leq \int_I |f(t)|^2 \mathrm{d}t \leq c_2 \sum_{n=-\infty}^{\infty} |a_n|^2$$

for every finite sum

$$f(t) = \sum_{n=-\infty}^{\infty} a_n e_n(t).$$

By *finite sum* we mean a sum having only a finite number of nonzero elements.

Now what happens if $D^+ = \infty$? The definition of the divided differences then depends on the enumeration of the sequence (λ_n) (see [3]). In that case, for any enumeration, the divided differences do not form a Riesz sequence. Indeed, if $D^+ = \infty$, choose enough and

not too much elements of (λ_n) according to a fixed enumeration of (λ_n) such that the upper density D_0 of these elements satisfies : $|I| < 2\pi D_0 < \infty$. Now, the sequence of divided differences coming from the selected elements of (λ_n) is included in the sequence of divided differences of the whole sequence (λ_n) and doesn't form a Riesz sequence thanks to the theorem 3.2.2, so the whole sequence doesn't form a Riesz sequence, either.

3.3 Proof

In order to prove the theorem, we use a method developed in [17].

Let be given a sequence (λ_n), a real $\gamma > 0$ and an integer M such that we have (3.2.1). We recall that $D^+ < \infty$. Let $0 < \gamma' \leq \gamma$. We suppose that the sequence of divided differences (e_n) relative to γ' forms a Riesz sequence in $L^2(I)$. We want to prove that $|I| \geq 2\pi D^+$.

We introduce some notations. We write $\Lambda := (\lambda_n)$. We introduce the trigonometric system over $L^2(I)$ given by $f_n := \exp(\frac{i2\pi n\cdot}{|I|})$ and we write $\Gamma := (\gamma_n) = (\frac{2\pi n}{|I|})$. We also write $n^+_\Lambda(r)$, $D^+(\Lambda)$ to precise that we talk about the upper density associated to Λ. We thus want to show that $|I| \geq 2\pi D^+(\Lambda)$. We can already remark that

$$D^+(\Gamma) = \frac{|I|}{2\pi}.$$

For $y \in \mathbb{R}$ and $r > 0$, we note $\Lambda_r := \Lambda \cap (y - r, y + r)$ and $\Gamma_r := \Gamma \cap (y - r, y + r)$.

The result will be deduced from a comparison theorem :

Theorem 3.3.1. *For every $\varepsilon > 0$, there exists $R > 0$ such that for all $r \geq 0$ and for all $y \in \mathbb{R}$, we have*

$$(1 - \varepsilon)\text{Card } (\Lambda_r) \leq \text{Card } (\Gamma_{r+R}). \tag{3.3.1}$$

Before proving this theorem, let show how it applies to prove Theorem 3.2.2.

From the inequality (3.3.1), we obtain

$$(1 - \varepsilon)n^+_\Lambda(r) \leq n^+_\Gamma(r + R),$$

then

$$(1 - \varepsilon)D^+(\Lambda) \leq \lim_{r \longrightarrow \infty} \frac{n^+_\Gamma(r + R)}{r + R} \cdot \frac{r + R}{r}.$$

Thus

$$(1 - \varepsilon)D^+(\Lambda) \leq D^+(\Gamma) = \frac{|I|}{2\pi}$$

Since ε can be arbitrarily small, we obtain effectively that

$$|I| \geq 2\pi D^+(\Lambda).$$

It now remains to prove the comparison theorem. The strategy here is to introduce associated finite dimensional spaces, to define an operator between these spaces and the inequality (3.3.1) will be derived from the estimation of the trace of this operator obtained in two different manners.

So we consider, for $y \in \mathbb{R}$ and $r > 0$ the linear hull V_r of the vectors e_n with $\lambda_n \in \Lambda_r$ (we recall that $e_n = [\lambda_m, \ldots, \lambda_n]$). Similarly, we define the linear hull W_r of the vectors f_n with $\gamma_n \in \Gamma_r$ (we recall that $f_n = e^{i\gamma_n \cdot}$).

Since Λ_r and Γ_r are finite sets, these spaces are effectively finite dimensional. Then we define the orthogonal projections

$$P_r : L^2(I) \longrightarrow V_r$$

and

$$Q_{r+R} : L^2(I) \longrightarrow W_{r+R}.$$

Denoting by i the injection

$$i : V_r \hookrightarrow L^2(I),$$

then we can define the endomorphism S_r of V_r by

$$S_r = P_r \circ Q_{r+R} \circ i.$$

The aim is to estimate the trace of S_r (which will be denoted by tr (S_r)) in two different manners in order to obtain (3.3.1).

Lemma 3.3.2. *For every $R > 0$, $r \geq 0$ and $y \in \mathbb{R}$, we have*

$$|\text{tr } (S_r)| \leq \text{Card } (\Gamma_{r+R}).$$

Proof. We have

$$\|S_r\| \leq \|P_r\| \|Q_{r+R}\| \leq 1.$$

Thus the eigenvalues of S_r have their moduli less than 1. So, we have :

$$|\text{tr } (S_r)| \leq \text{rang}(S_r) \leq \dim W_{r+R}.$$

Since $\dim W_{r+R} = \text{Card } (\Gamma_{r+R})$, the lemma follows. □

In order to obtain the inverse inequality, we use a homogeneous approximation lemma.

We recall e.g. from [44] a lemma about Riesz sequences.

Lemma 3.3.3. *If (g_n) is a Riesz sequence in a Hilbert space H, then it admits a biorthogonal bounded sequence.*

Then we apply this lemma to the divided differences sequence (e_j) and we call by (φ_j) the associated biorthogonal sequence. It is then possible to express the trace of S_r in terms of (φ_j).

Lemma 3.3.4. *We have*

$$\text{tr } (S_r) = \text{Card } (\Lambda_r) + \sum_{\lambda_j \in \Lambda_r} ((Q_{r+R} - \text{Id })e_j | P_r \varphi_j).$$

Proof. Using the biorthogonal sequence we have :

$$\begin{aligned} \text{tr } (S_r) &= \sum_{\lambda_j \in \Lambda_r} (S_r e_j | \varphi_j) \\ &= \sum_{\lambda_j \in \Lambda_r} (Q_{r+R} e_j | P_r \varphi_j) \\ &= \sum_{\lambda_j \in \Lambda_r} (e_j | P_r \varphi_j) + \sum_{\lambda_j \in \Lambda_r} ((Q_{r+R} - \text{Id })e_j | P_r \varphi_j). \end{aligned}$$

Since $P_r e_j = e_j$, we obtain that $(e_j | P_r \varphi_j) = (P_r e_j | \varphi_j) = 1$ and the result follows. □

Then we use the following homogeneous approximation lemma :

Lemma 3.3.5. *For every $\varepsilon > 0$, there exists $R > 0$ such that for all $r > 0$, $y \in \mathbb{R}$ and ℓ such that $\lambda_\ell \in \Lambda_r$, we have*

$$\| (Q_{r+R} - \text{Id }) e_\ell \| \leq \varepsilon.$$

Proof. Since the trigonometric system (f_p) of $L^2(I)$ is orthonormal and since Q_{r+R} is an orthogonal projection over W_{r+R}, we obtain :

$$\| (Q_{r+R} - \text{Id }) e_\ell \|^2 = \sum_{|\gamma_p - y| > r+R} |(e_\ell | f_p)|^2.$$

Now we have :

$$(e_\ell | f_p) = \int_I g(t) e^{i\lambda_\ell t} e^{-i\gamma_p t} \mathrm{d}t$$

with

$$g(t) = [\lambda_m - \lambda_\ell, \ldots, \lambda_\ell - \lambda_\ell](t).$$

Integrating by parts over $I = (a, b)$ we obtain

$$(e_\ell | f_p) = [\frac{1}{i\lambda_\ell - i\gamma_p} g(t) e^{i\lambda_\ell t} e^{-i\gamma_p t}]_a^b - \int_I \frac{1}{i\lambda_\ell - i\gamma_p} g'(t) e^{i\lambda_\ell t} e^{-i\gamma_p t} \mathrm{d}t.$$

Now, by a direct computation from the formula (3.2.2), given an integer $r \geq 1$ and reals $\mu_1, \ldots, \mu_r$, we have :

$$[\mu_1, \ldots, \mu_r]'(t) \leq \frac{(r-1)t^{r-2}}{(r-1)!} + (|\mu_r - \mu_{r-1}| + \cdots + |\mu_2 - \mu_1| + |\mu_1|)\frac{t^{r-1}}{(r-1)!}$$

Thus, in our case, thanks to the $\gamma'-$closed exponents property, we have :

$$|g'(t)| \leq (\ell - m)\frac{t^{\ell-m-1}}{(\ell-m)!} + (\ell - m)\gamma' \frac{t^{\ell-m}}{(\ell-m)!}$$

At this stage, we can find a constant C depending only on γ', M, a and b such that

$$|(e_\ell | f_p)|^2 \leq \frac{C}{|\lambda_\ell - \gamma_p|^2}.$$

Recalling that $\lambda_\ell \in \Lambda_r$, we obtain :

$$\begin{aligned} \|(Q_{r+R} - \mathrm{Id}\ )e_\ell\|^2 &\leq \sum_{|\gamma_p - y| > r+R} \frac{C}{|\lambda_\ell - y + y - \gamma_p|^2} \\ &\leq \sum_{|\gamma_p - y| > r+R} \frac{C}{||y - \gamma_p| - r|^2} \\ &\leq \sum_{p \in \mathbb{Z}} \frac{C}{|\frac{2\pi|p|}{|I|} + R|^2}. \end{aligned}$$

Since this last expression doesn't depend on r, y and tends to 0 as $R \to 0$, the lemma follows. □

Now we can finish the proof of Theorem 3.3.1.

Proof of Theorem 3.3.1. Let $\varepsilon > 0$. By combining the two preceding lemmas, since (φ_j) is bounded, we obtain

$$\mathrm{tr}\ (S_r) \geq (1 - \varepsilon)\mathrm{Card}\ (\Lambda_r).$$

Then the theorem follows from Lemma 3.3.2. □

Bibliographie

[1] F. Alabau, *Observabilité frontière de systèmes faiblement couplés*, C. R. Acad. Sci. Paris Sér. I **333** (2001), 645–650.

[2] F. Alabau-Boussouira, *Two-level energy method for boundary observability*, SIAM J. Control Optim., 42 (2002), 871–905.

[3] S. A. Avdonin et S. A. Ivanov, *Exponential Riesz bases of subspaces and divided differences*, St. Petersburg Math. J. 13 (3) (2002), 339–351.

[4] S. A. Avdonin et W. Moran, *Simultaneous control problems for systems of elastic strings and beams*, Systems & Control Letters 44 (2001), 147–155.

[5] S. A. Avdonin et S.A. Ivanov, *Ingham type inequalities and Riesz bases of divided differences*, J. Appl. Math. Comput. Sci.

[6] J. Ball et M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Comm. Pure Appl. Math. 37 (1979), 555–587.

[7] C. Bardos, G. Lebeau et J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim. 30 (1992), 1024–1065.

[8] C. Baiocchi, V. Komornik et P. Loreti, *Ingham type theorems and applications to control theory*, Bol. Un. Mat. Ital. B (8) 2 (1999), no. 1, 33–63.

[9] C. Baiocchi, V. Komornik et P. Loreti, *Ingham-Beurling type theorems with weakened gap conditions*, Acta Math. Hungar. 97 (1-2) (2000), 55–95.

[10] B. J. C. Baxter et R. Brummelhuis, *Exponential brownian motion and divided differences*, disponible sur http ://cato.tzo.com/brad/baxter.html.

[11] J. N. J. W. L. Carleson et P. Malliavin, éditeurs, *The Collected Works of Arne Beurling*, Volume 2, Birkhäuser, 1989.

[12] E. Crépeau at C. Prieur, *Control of a clamped-free beam by a piezoelectric actuator*, Rapport LAAS N°04101, Février 2004, 12p.

[13] René Dáger, *Observation and control of vibrations in tree-shaped network of strings*, SIAM J. Control Optim., Vol.43, No. 2 (2004), pp. 590-623.

[14] R. Dáger, E. Zuazua, *Controllability of tree-shaped networks of strings*, C. R. Acad. Sci. Paris Sér/ I332 (2001), 1087–1092.

[15] R. Dáger, E. Zuazua, *Spectral boundary controllability of network of strings*, C. R. Acad. Sci. Paris Sér/ I334 (2002), 545–550.

[16] I. Gohberg, S. Goldberg et M. A. Kaashoek, *Classes of linear operators*, Birkhäuser-Verlag, 1990.

[17] K. Grőchenig et H. Razafinjatovo, *On Landau's necessary density conditions for sampling and interpolation of band-limited functions*, J. London Math. Soc (2) 54 (1996), 557–565.

[18] Bao-Zhu Guo et Yu Xie, *A sufficient condition on Riesz basis with parentheses of non-selfadjoint operator and application to a serially connected string system under joint feedbacks*, SIAM, Journal on Control and Optimization, to appear.

[19] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl. 68 (1989), 457–465.

[20] A. Haraux et S. Jaffard, *Pointwise and spectral control of plate vibrations*, Rev. Mat. Iberoamericana 7 (1991), no. 1, 1–24.

[21] C. Hermite, *Sur la formule d'interpolation de Lagrange*, Journal für die Reine und Angewendte Mathematik 84 (1878), 70–79, disponible sur le site web www.math.technion.ac.il/hat.

[22] Lop Fat Ho, *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Paris Sér. I Math. 302 (1986), no. 12, 443–446.

[23] A. E. Ingham, *Some trigonometrical inequalities with applications in the theory of series*, Math. Z. 41 (1936), 367-379.

[24] S.Jaffard, *A Density Criterion for Frames of Complex Exponentials*, Michigan Math. J.38 (1991), 339–348.

[25] J.P. Kahane, *Pseudo-Périodicité et séries de fourier lacunaires*, Ann. scienc. Ec. Norm. Sup (1962), 93–150.

[26] V. Komornik, *On the exact internal controllability of a Petrowsky system*, J. Math. Pures Appl. (9) 71 (1992), 331–342.

[27] V. Komornik, *Exact Controllability and stabilization. The multiplier method*, Masson, Paris and John Wiley & Sons, Chicester, 1994.

[28] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim. 35 (1997), 1591–1613.

[29] V. Komornik et P. Loreti, *Ingham type theorems for vector-valued functions and observability of coupled linear systems*, SIAM J. Control Optim. 37 (1998), 461–485.

[30] V. Komornik et P. Loreti, *Observability of compactly perturbed systems*, J. Math. Anal. Appl. 243 (2000), 409–428.

[31] V. Komornik et P. Loreti, *Boundary observability of compactly perturbed systems*, Conference on Control of Distributed Parameter Systems (Graz, 2001).

[32] V. Komornik et P. Loreti, *Partial observability of coupled linear systems*, Acta Math. Hungar. 86 (1–2) (2000), 49–74.

[33] P. Koosis, *Lecons sur le théorème de Beurling et Malliavin*, Les publications CRM, Montreal (1996).

[34] H. J. Landau, *Necessary density conditions for sampling and interpolation of certain entire functions*, Acta Math. 117 (1967), 37–52.

[35] J. L. Lions, *Controlabilité exacte, perturbations et stabilisation de systèmes distribués I-II*, Rech. Math. Appl. 8, Masson, Paris 1988.

[36] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., **30** (1988),1–68.

[37] P. Loreti, *Exact controllability of shells in minimal time*, Rend. Mat. Acc. Lincei (9) 12 (2001), 43–48.

[38] S. Nicaise, *Spectre des réseaux topologiques finis*, Bull. Sc. math., 2° série, 111, (1987), P. 401–413.

[39] A.Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, 1983.

[40] Kim-Dang Phung, *Remarques sur l'observabilité pour l'équation de Laplace*, ESAIM : COCV, August 2003, Vol. 9, pp. 621–635.

[41] J. P. Roth, *Une généralisation de la formule de Poisson*, Publication mathématique n°23 de l'Université de Haute-Alsace, Mulhouse, 1984.

[42] K. Seip, *On the Connection between Exponential Bases and Certains Related Sequences in* $L^2(-\pi,\pi)$, Journal of Functional Analysis 130 (1995), 131–160.

[43] D. Ullrich, *Divided differences and system of nonharmonic Fourier series*, Proc. Amer. Math. Soc. 80 (1980), 47–57.

[44] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[45] R. M. Young, *On a class of Riese-Fischer sequences*, Proc. Amer. Math. Soc. 126 (1998), 1139–1142.

[46] A. Zygmund, *Trigonometric series*, Vol. I, II, troisième edition, Cambridge University Press (2002).

Chapitre 4

Introduction à l'équation de Vlasov

On rappelle ici des généralités à propos de l'équation de Vlasov. Le plan des chapitres suivants est indiqué à la fin de ce chapitre.

4.1 Contexte physique

Notre environnement est essentiellement composé de matière qui est sous forme de solide, de liquide ou de gaz. Cependant, il existe un *quatrième état* de la matière appelé *plasma*, qui est un gaz de particules chargées (ions ou électrons) avec une densité suffisamment grande pour que les particules suivent une loi statistique et assez faible pour que les interactions binaires soient négligeables par rapport aux forces coulombiennes de longue portée. Pour produire du plasma, on peut augmenter la température (mais pas trop, si la densité est faible). Le plasma dépend en fait d'un *ratio* densité/température.

Quelques quantités physiques. Afin de mieux décrire les dépendances entre les quantités, on introduit les *échelles caractéristiques*.

• *Fréquence plasma* (échelle de temps) : si un groupe d'électrons est brutalement déplacé dans une direction, une force électrostatique de direction opposée est créée. Afin de restaurer l'état d'équilibre, les particules oscillent autour d'une position d'équilibre avec une fréquence qui est la fréquence plasma ω_p, définie par :

$$\omega_p = \Big(\frac{n_0 q^2}{\varepsilon_0 m_* \varepsilon}\Big)^{1/2}, \tag{4.1.1}$$

où n_0 est la densité de charge, q la charge, m_ε la masse de l'électron et ε_0 la permittivité de l'air.

• *Longueur de Debye* (échelle de longueur) : la sphère d'influence d'une particule caractérisée par le fait qu'en dehors de cette sphère, l'individualité des particules disparaît, est appelée *sphère de Debye*, et son rayon λ_D est appelé longueur de Debye et est donné par la formule : b

$$\frac{1}{\lambda_D^2} = \sum_\alpha \frac{n_{0,\alpha} q_\alpha^2}{\varepsilon_0 K_B T_\alpha}, \tag{4.1.2}$$

où l'indice α correspond au type de particules, $n_{0,\alpha}$ est la densité, T_α la température, et K_B une constante (la constante de Boltzmann).
On définit ensuite le *paramètre plasma* g par

$$g = \frac{1}{n_0 \lambda_D^3}.$$

Il correspond à l'inverse du nombre de particules contenues dans la sphère de Debye. Dans le cas d'un *plasma Vlasov*, on fait l'hypothèse :

$$g \to 0,$$

ce qui signifie que l'on a une infinité de particules dans la sphère de Debye.
Comme

$$n_0 \lambda_D^3 = O\left(T\sqrt{\frac{T}{n_0}}\right),$$

afin d'avoir un plasma Vlasov, on peut soit augmenter la température, soit diminuer la densité.
On fait aussi l'hypothèse :

$$\frac{\nu_c}{\omega_P} << 1,$$

où ν_c est la fréquence des collisions, ce qui veut dire que l'on néglige les collisions.

Remarque 4.1.1. *(i) On note ici l'importance de l'ordre de grandeur des quantités physiques. Il est impossible de tenir compte de toutes les quantités exactement ; l'approximation est nécessaire pour pouvoir mener des calculs.*

(ii) Les hypothèses que l'on peut faire ne sont parfois valables que pour un certain temps. Ainsi, négliger les collisions, n'est plus envisageable au bout d'un temps long ; il faut alors changer de modèle.

(iii) On n'a introduit que les quantités physiques les plus importantes. Il existe également une vaste gamme de plasmas (voir [34],[15], par exemple).

Le plasma dans la vie quotidienne. Le plasma est assez rare dans la vie (on peut en trouver par exemple dans les tubes à néon), mais il représente 99 pour cent de la matière de l'univers.
Il est utilisé pour la *fusion thermonucléaire* ; grâce à un champ magnétique élevé, le plasma

peut être confiné à de hautes températures. Les phénomènes physiques intervenant dans de telles expérimentations sont modélisés par les équations de Vlasov. Une application prometteuse de la fusion thermonucléaire est d'obtenir une nouvelle forme d'énergie qui est propre et renouvelable.

4.2 Modélisation

L'évolution en temps d'ensembles de particules est modélisée par la théorie cinétique. Les particules peuvent être des objets de nature très différentes dépendant de la situation physique. Par exemple, les particules peuvent être des atomes ou des molécules dans un gaz neutre, des ions dans un plasma. En dynamique des étoiles, les particules sont des étoiles et dans le cas cosmologique, les particules sont des galaxies, voire même des amas de galaxies. Les modèles mathématiques des systèmes de particules sont souvent décrits par les équations des fluides et les équations cinétiques.

Modèle cinétique : il s'agit d'un modèle statistique où les systèmes de particules sont décrits par la fonction de distribution :

$$f = f(t, x, p),$$

qui représente la densité de particules à une position dans l'espace-temps donnée $(t, x) \in \mathbb{R} \times \mathbb{R}^3$, avec une quantité de mouvement $p \in \mathbb{R}^3$. Une fonction de distribution contient beaucoup d'informations ; les quantités macroscopiques sont déduites à partir de celle-ci.

Modèle fluide : les quantités qui décrivent le système ne dépendent pas de la quantité de mouvement p, mais seulement de la position dans l'espace-temps (t, x). Notons que ce modèle n'est pas toujours suffisant pour décrire la physique et peut engendrer aussi des phénomènes non physiques.

Validité du modèle. L'évolution en temps du système est déterminé par les interactions entre les particules qui dépend de la situation physique. Par exemple, le principal mécanisme pour l'évolution en temps d'un gaz neutre est la collision entre les particules (équation de Boltzmann). Pour un plasma, l'interaction est donnée par les charges électriques (système de Vlasov-Maxwell) et pour les étoiles, l'interaction est gravitationnelle (système d'Einstein-Vlasov). Bien sûr, des combinaisons de ces processus d'interaction sont aussi considérées, mais dans beaucoup de situations, un de ces processus domine fortement et les autres sont alors négligés.

4.3 Dérivation de l'équation pour la fonction de distribution.

Dans le cas non relativiste, on remplace le moment p (appelé aussi impulsion) par la vitesse v. La fonction de distribution $f(t,x,v)$ est une fonction positive. On suppose que la position x et la vitesse v appartiennent à $\mathbb{R}^d$ (auparavant d valait 3 ; en général, les dimensions $d = 1, 2$ ou 3 peuvent avoir une signification physique).

Régularité minimale : comme d'un point de vue physique, $f(t,x,v)\mathrm{d}x\mathrm{d}v$ représente la probabilité de présence de particules dans un élement de volume $\mathrm{d}x\mathrm{d}v$, au temps t et à la position (x,v) de l'espace des phases, on fera l'hypothèse assez réaliste sur la régularité de f :

$$f(t,\cdot) \in L^1_{loc}(\mathbb{R}^d \times \mathbb{R}^d).$$

Caractéristiques. On introduit les caractéristiques (ou encore trajectoires) $(X(t),V(t))$ qui sont données par la loi de Newton. Dans le cas de la dynamique classique (non relativiste), on a :

$$\dot{X} = V \quad , \quad \dot{V} = F(t,X(t)) = -\nabla_x \varphi(t,X),$$

si la force F dérive d'un potentiel ϕ.
La notation $\dot{u}$ désigne ici la dérivée de u par rapport au temps.
Comme f décrit l'évolution statistique du système de particules, on demande à ce que f soit *constante le long des caractéristiques* :

$$\frac{\mathrm{d}}{\mathrm{d}t} f\big(t,X(t),V(t)\big) = \partial_t f + V(t)\cdot\nabla_x f + F(t,X(t))\cdot\nabla_v f$$

La fonction de distribution f satisfait donc l'équation de transport, aussi appelée équation de Vlasov :

$$\partial_t f + v\cdot\nabla_x f + F(t,x)\cdot\nabla_v f = 0. \tag{4.3.1}$$

On a utilisé la notation :

$$\nabla_x f = \big(\partial_{x_1} f, \ldots, \partial_{x_d} f\big)$$

et on utilisera aussi :

$$\begin{aligned}
&\nabla_x \cdot f = \partial_{x_1} f + \cdots + \partial_{x_d} f,\\
&\Delta_x f = \partial^2_{x_1} f + \cdots + \partial^2_{x_d} f,\\
&|x|^2 = x_1^2 + \cdots + x_d^2,\\
&a\cdot b = a_1 b_1 + \cdots + a_d b_d,
\end{aligned}$$

$$\nabla_x \times (F_1,F_2,F_3) = \textbf{rot}\ {}_x F = (\partial_{x_2}F_3 - \partial_{x_3}F_2, \partial_{x_3}F_1 - \partial_{x_1}F_3, \partial_{x_1}F_2 - \partial_{x_2}F_1),$$

et si on remplace x par v ou x,v, cela revient à remplacer $x_1,\ldots,x_d$ par $v_1,\ldots,v_d$ ou $x_1,\ldots,x_d,v_1,\ldots,v_d$, lorsque cela a un sens.

Approximation par champ moyen. L'équation de Vlasov est en général couplée avec une équation pour le calculs des champs qui dépendent de certaines moyennes de la fonction de distribution (quantités macroscopiques). Ainsi, dans le cas où le couplage est donné par l'équation de Poisson pour le champ électrique, on a

$$\nabla_x \cdot F = \rho \quad \text{où} \quad \rho = \int_{\mathbb{R}^d} f(t,x,v)\mathrm{d}v. \tag{4.3.2}$$

Les équations (4.3.1) et (4.3.2) correspondent au système de Vlasov-Poisson.

4.4 Quelques propriétés des solutions.

Considérons une solution de l'équation de transport (4.3.1).

Conservation de la masse. En intégrant formellement (4.3.1) par parties par rapport à v, et si la *densité de courant* est définie par :

$$j(t,x) = \int_{\mathbb{R}^d} f(t,x,v)v\mathrm{d}v,$$

on obtient

$$\partial_t \rho(t,x) + \nabla_x \cdot j(t,x) = 0 \tag{4.4.1}$$

Ce calcul peut être justifié lorsque l'on suppose que f décroit suffisamment. L'équation (4.4.1) exprime alors la *conservation locale* de la masse (ou du nombre de particules). En intégrant encore une fois par rapport à x, on obtient :

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d\times\mathbb{R}^d} f(t,x,v)\mathrm{d}x\mathrm{d}v = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\rho(t,x)\mathrm{d}x = 0$$

Cette relation traduit la *conservation globale* de la masse.

Conservation de l'énergie. Considérons une solution de :

$$\partial_t f + v\cdot\nabla_x f - \nabla_x\varphi\cdot\nabla_v f = 0 \tag{4.4.2}$$

En multipliant l'équation (4.4.2) par $\frac{|v|^2}{2}$ et en intégrant par rapport à x et v, on obtient :

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{|v|^2}{2}f(t,x,v)\mathrm{d}x\mathrm{d}v - \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\nabla_x\varphi\cdot\nabla_v f\frac{|v|^2}{2}\mathrm{d}x\mathrm{d}v = 0,$$

qui, par intégrations par parties, combiné avec (4.4.1) et $-\Delta\phi = \rho$, donne la conservation de l'énergie

$$\frac{d}{dt}\Big[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(t,x,v)\frac{|v|^2}{2}\mathrm{d}x\mathrm{d}v + \int_{\mathbb{R}^d}|\nabla_x\varphi(t,x)|^2\mathrm{d}x\Big] = 0. \tag{4.4.3}$$

Préservation de la mesure. Considérons une solution de (4.4.2). Alors l'application $(x,v) \to (X(s,t,x,v), V(s,t,x,v))$ donnée par les *caractéristiques*

$$\partial_t X = V, \quad \partial_t V = -\nabla_x \varphi(t,X), \quad (X,V)(s,s,x,v) = (x,v)$$

préserve les mesures, ce qui veut dire ici que le jacobien de la transformation est unitaire (son déterminant vaut 1 en valeur absolue). Cela peut-être déduit du flot d'un système hamiltonien.

Forme conservative. On considère le *flux*

$$a(t,x,v) = (v, F(t,x))$$

de l'équation de transport (4.3.1) ; on a alors :

$$\nabla_{x,v} \cdot a = 0.$$

L'équation de transport (4.3.1) se réécrit alors sous forme conservative :

$$\partial_t f + \nabla_{x,v}(af) = 0.$$

4.5 Les modèles de Vlasov-Maxwell/Poisson

On considère un plasma non collisionnel (le plasma Vlasov), qui est une collection de particules pour lesquelles les collisions sont assez rares ; l'interaction principale est donnée par leurs charges. On suppose ici que le plasma n'est composé que d'une seule espèce de particules. La masse de la particule est normalisée à 1. En théorie cinétique, le cadre le plus général des équations modélisant un plasma non collisionnel est alors le *système relativiste de Vlasov-Maxwell* :

$$\partial_t f + \hat{v} \cdot \nabla_x f + (E(t,x) + \hat{v} \times B(t,x)) \cdot \nabla_v f = 0, \tag{4.5.1}$$

$$\partial_t E + j = c\nabla_x \times B, \quad \nabla_x \cdot E = \rho, \tag{4.5.2}$$

$$\partial_t B = -c\nabla_x \times E, \nabla_x \cdot B = 0. \tag{4.5.3}$$

E et B sont les champs électriques et magnétiques et $\hat{v}$ est la vitesse relativiste :

$$\hat{v} = \frac{v}{\sqrt{1 + \frac{|v|^2}{c^2}}}, \tag{4.5.4}$$

où c est la vitesse de la lumière. La densité de charge ρ et la densité de courant j sont donnés par :

$$\rho = \int_{\mathbb{R}^3} f \mathrm{d}v, \quad j = \int_{\mathbb{R}^3} \hat{v} f \mathrm{d}v. \tag{4.5.5}$$

L'équation (4.5.1) est l'équation de Vlasov relativiste, et (4.5.2)-(4.5.3) sont les équations de Maxwell.
Un cas spécial en dimension trois est obtenu en considérant des conditions initiales symétriques. Pour de telles données, on peut montrer que le champ magnétique doit être constant. On déduit de l'équation de Faraday $\nabla_x \times E = -\partial_t B$ que le champ électrique est le gradient d'un potentiel φ. Dans ce cas, le système de Vlasov-Maxwell relativiste s'écrit :

$$\partial_t f + \hat{v} \cdot \nabla_x f + \beta E(t,x) \cdot \nabla_v f = 0, \tag{4.5.6}$$
$$E = -\nabla_x \varphi, \quad -\Delta_x \varphi = \rho. \tag{4.5.7}$$

Ici $\beta = 1$ et le champ magnétique constant est mis à zéro (un champ magnétique constant n'a pas de signification dans la discussion). Le système a alors un sens pour toutes données initiales et est appelé système de Vlasov-Poisson. Un autre cas intéressant est la limite classique obtenue en faisant tendre la vitesse de la lumière c vers l'infini dans (4.5.1)–(4.5.3), donnant ainsi :

$$\partial_t f + v \cdot \nabla_x f + \beta E(t,x) \cdot \nabla_v f = 0, \tag{4.5.8}$$
$$E = -\nabla_x \varphi, \quad -\Delta_x \varphi = \rho, \tag{4.5.9}$$

où $\beta = 1$. On pourra consulter Schaeffer [65] pour une dérivation rigoureuse de ce résultat. Il s'agit là de l'équation de Vlasov-Poisson non relativiste. Le cas $\beta = 1$ correspond à des forces répulsives (comme le cas du plasma) tandis qu'en prenant $\beta = -1$, il s'agit d'un modèle avec forces attractives pour le système d'auto-gravitation de Newton (intervenant pour les étoiles).

4.6 Solutions classiques pour le système de Vlasov-Poisson

On considère le système de Vlasov-Poisson suivant :

$$\partial_t f + v \cdot \nabla_x f - \nabla_x \phi \cdot \nabla_v f = 0 \quad t > 0, \quad x, v \in \mathbb{R}^d, \tag{4.6.1}$$
$$-\Delta_x \phi = \gamma \rho = \gamma \int_{\mathbb{R}^d} f(t,x,v) \mathrm{d}v, \tag{4.6.2}$$
$$\tag{4.6.3}$$

en dimension $d = 3$ et $\gamma = 1$ (cas de plasmas ou électrostatique) ou $\gamma = -1$ (cas gravitationnel).
Un ingrédient principal est alors de commencer avec une solution qui est initialement à support compact en vitesse et de contrôler la taille de son support. On définit ainsi la quantité :

$$Q(t) := 1 + \sup\{|v| : \exists (t,x) \in (0,t) \times \mathbb{R}^d, \text{ tel que } f(t,x,v) \neq 0\}.$$

On obtient alors le théorème suivant :

Théorème 4.6.1. *Soit f_0 une fonction C^1 positive à support compact. Alors le problème de Cauchy pour (4.6.1) a une unique solution C^1 et*

$$Q(t) \leq C_p(1+t)^p \quad avec \quad p > \frac{33}{17},$$

où C_p est une constante indépendante de t.

La preuve peut être trouvée dans [41] et se déduit de l'étude du schéma itératif suivant :

$$\begin{cases} \partial_t f_{n+1} + v \cdot \nabla_x f_{n+1} - \nabla_x \varphi_n \partial_v f_{n+1} = 0, \\ -\Delta_x \varphi_n = \gamma \int_{\mathbb{R}^3} f_n \mathrm{d}v, \\ f_{n+1}(t=0,\cdot,\cdot) = f_0, \end{cases}$$

qui est résolu à chaque pas de temps par la méthode des caractéristiques (ou méthode semi-lagrangienne, qui sera présentée par la suite). Afin de pouvoir passer à la limite, on doit prouver des bornes uniformes sur le champ E et ses dérivées, et celles-ci peuvent être facilement obtenues si l'on a une estimation uniforme sur la taille du support de f.

4.7 Schémas numériques

L'équation de Vlasov a la particularité de faire intervenir beaucoup de variables (6 pour l'espace des phases en dimension 3 plus le temps). La discrétisation de l'équation de Vlasov suscite beaucoup d'intérêt de la part des physiciens. Ainsi, la simulation numérique permet de compléter, de guider et d'améliorer les tests expérimentaux souvent très coûteux dans le domaine de la physique des plasmas (comme la construction de lasers, de tokamaks ou de réacteurs nucléaires).

Méthodes PIC. La résolution numérique de l'équation de Vlasov est souvent réalisée par des méthodes *particule-maillage*, plus communément appelées méthodes PIC (Particle-in-Cell) (voir par exemple le livre de Birdsall et Langdon [19]). L'idée de telles méthodes est de considérer l'évolution de points assez nombreux et bien choisis dans l'espace des phases. Les trajectoires de N "particules" sont données grâce à un champ de force F, qui est calculé avec un maillage (de l'espace physique). Autrement dit, la fonction de distribution $f(t,x,v)$ est discrétisée par :

$$f_N(t,x,v) = \sum_{i,j} \omega_i(t) \delta_{x_i(t)}(x) \otimes \delta_{v_i(t)}(v), \tag{4.7.1}$$

avec

$$\delta_{x_i}(x) = \begin{cases} 1 & \text{si } x = x_i, \\ 0 & \text{ailleurs}, \end{cases} \tag{4.7.2}$$

où $\omega_i > 0$ est le poids de la particule i dont la position $(x_i(t), v_i(t))$ est solution du système de caractéristiques :

$$\begin{cases} \dot{x_i}(t) = v_i(t), \quad i = 1, \dots, N, \\ \dot{v_i}(t) = F(t, x_i(t), v_i(t)), \quad i = 1, \dots, N, \\ \omega_i(t) = \omega_i(0) \quad i = 1, \dots, N, \end{cases} \tag{4.7.3}$$

Le champ de force de Lorentz est alors discrétisé dans l'espace physique (ce qui explique le C dans PIC). Un intérêt majeur ici est que l'on n'a pas de maillage de l'espace des phases. Cependant, le taux de convergence n'est que de l'ordre de $1/\sqrt{N}$, où N désigne le nombre de particules.

Méthodes eulériennes. Récemment, avec la montée en puissance des calculateurs et de la gestion de la mémoire, les méthodes basées sur une grille de l'espace des phases ont regagné de l'intérêt. On distingue en particulier les méthodes de volumes finis et les méthodes semi-lagrangiennes : soit des volumes, soit des points sont advectés sur une grille le long des courbes caractéristiques (on utilise le fait que la fonction de distribution est constante sur les courbes caractéristiques pour mettre à jour les nouvelles valeurs de la fonction de distribution à chaque pas de temps).
Beaucoup de travail a été accompli (voir par exemple [34] ou [15], et les références incluses), pour avoir des schémas précis et pour conserver les propriétés de l'équation de Vlasov.
Remarquons que grâce à l'advection par les caractéristiques, on n'a pas la contrainte habituelle de restriction forte sur le pas de temps, comme c'est généralement le cas pour des schémas habituels de différences finies ou de volumes finis.

Remarque 4.7.1. *Les méthodes semi-lagrangiennes sont très utilisées dans la communauté de prédiction météo (voir les articles de la revue* Mounthly Review*) et restent toujours sujet de recherche (pour une référence récente : voir [58]).*

Méthodes adaptatives. Les caractéristiques de l'équation de Vlasov peuvent par endroits se rapprocher de plus en plus avec le temps (on parle d'*enroulement des caractéristiques*). Ce phénomène introduit donc des régions de fort gradient pour la fonction de distribution, qui ne peuvent être résolues qu'avec une taille de maille très fine (pour le cas de méthodes eulériennes). D'un autre côté, cela peut être une vraie perte de temps et de mémoire pour d'autres régions où la fonction de distribution est nulle. Ainsi, des méthodes adaptatives qui ne gardent que le nombre nécessaire de points de grille, tout en restant suffisamment précises sont bienvenues pour être développées et utilisées.

Remarque 4.7.2. *Bien sûr, il existe encore beaucoup d'autres méthodes pour la résolution de l'équation de Vlasov (ainsi, on peut résoudre dans l'espace de Fourier, voir par exemple [50]). Le lecteur pourra consulter par exemple [36], [37], [34], [15] et les références incluses.*

4.8 Plan

Dans les chapitres suivants, on s'intéresse à quelques questions concernant la résolution numérique de l'équation de Vlasov.

Chapitre 5 **:** on étudie le système de Vlasov-Poisson dans le cas monodimensionnel et périodique. A partir d'un résultat d'existence de Cooper et Klimas, on redémontre des résultats de régularité. On détaille ensuite l'erreur de discrétisation en temps pour un schéma semi-Lagrangien avec splitting en temps.
Les résultats de cette partie sont alors utiles pour le chapitre 7.

Chapitre 6 **:** on s'intéresse ici à une discrétisation uniforme de l'espace des phases, mais avec une reconstruction d'ordre élevée. Grâce à des inégalités sur des polynômes (qui sont redémontrées), on obtient alors la convergence de schémas d'ordre arbitraire.
Ce chapitre est une copie de l'article soumis en collaboration avec Nicolas Besse ; il peut être lu de manière indépendante.

Chapitre 7 **:** on donne un résultat de convergence pour un schéma adaptatif avec une reconstruction affine sur chaque triangle (obtenu en coupant un carré en deux).
L'étude entreprise dans ce chapitre, ainsi que dans le chapitre 5 contribuent à un article en préparation en collaboration avec Martin Campos Pinto.

Chapitre 8 **:** une méthode adaptative basée sur une reconstruction biquadratique est développée et des résultats numériques obtenus dans des simulations classiques de Vlasov sont exhibés.
Ce travail est issu d'une collaboration initiée au CEMRACS'03 avec Martin Campos Pinto. La parallélisation a été développée par Eric Violard et Olivier Hoenen en 2004.

Chapitre 5

Le système de Vlasov-Poisson périodique en dimension 1

Dans le précédent chapitre, on a mentionné un résultat d'existence et de régularité pour le système de Vlasov-Poisson en dimension 3. On se propose ici de redémontrer des estimations de régularité des solutions dans le cas plus simple du système de Vlasov-Poisson périodique en dimension 1 en se basant sur un résultat d'existence de Cooper-Klimas. Ces estimations seront utilisées par la suite pour l'étude de schémas semi-Lagrangiens, avec splitting en temps d'ordre 2. Grâce à l'effet régularisant du champ de Poisson, on obtient alors une erreur en temps d'ordre deux, pour une donnée initiale lipschitzienne, ce qui améliore des résultats de [16], où la fonction initiale était supposée de classe C^2 (voir aussi [15]). La discrétisation spatiale de l'espace des phases sera plus spécifiquement étudiée aux chapitres 6 et 7, avec respectivement des schémas uniformes d'ordre élevés et un schéma adaptatif, avec reconstruction affine par maille.
Cette étude préliminaire sera utile pour le chapitre 7.

Notations. On utilisera les notations $f(t,x,v)$, $f(t)(x,v)$, $f(t)$ et f pour la fonction de distribution, suivant le contexte, et de même, on notera $E(t,x)$, $E(t)$ ou E le champ électrique.

5.1 Enoncés des théorèmes de Cooper-Klimas

On rappelle ici quelques résultats préliminaires, issus de l'article de Cooper et Klimas ([27]).

Soit $W(T)$ l'espace des fonctions $A(t,x)$ sur $[0,T]\times\mathbb{R}$ telles que A et $\partial_x A$ soient continues

sur $[0,T]\times\mathbb{R}$ et telles que

$$\sup|A(t,x)|<\infty \quad \text{et} \quad \sup|\partial_x A(t,x)|<\infty,$$

où le supremum est pris sur $[0,T]\times\mathbb{R}$.
Le lemme suivant donne alors la propriété de conservation de mesure :

Lemme 5.1.1. *Soit $A\in W(T)$, alors le système*

$$\begin{cases} \dot{X}=V, \quad \dot{V}=-A, \\ X(0)=x_0, \quad V(0)=v_0, \end{cases} \tag{5.1.1}$$

admet une unique solution $(X(t),V(t))$ et l'application $(x_0,v_0)\to(X(t),V(t)):\mathbb{R}^2\to\mathbb{R}^2$ est bijective, continûment différentiable et de jacobien $J(t)\equiv 1$.

Soit L un nombre strictement positif qui désignera la *période*.
On considère maintenant le système :

$$\partial_x\mathcal{E}(t,x)=\int_{\mathbb{R}}[n(v)-f(t,x,v)]\mathrm{d}v, \tag{5.1.2a}$$

$$\partial_t\mathcal{E}(t,x)=\int_{\mathbb{R}}[f(t,x,v)-n(v)]v\mathrm{d}v, \tag{5.1.2b}$$

$$\int_0^L \mathcal{E}(0,x)\mathrm{d}x=0, \tag{5.1.2c}$$

avec $f(x,v,t)$ constante le long des courbes caractéristiques

$$\dot{X}=V, \quad \dot{V}=-\mathcal{E}(t,X(t)). \tag{5.1.3}$$

L'équation (5.1.2a) est l'équation de Poisson, tandis que (5.1.2b) est l'équation d'Ampère et n représente la distribution du fond d'ions.
Soit σ une fonction continue décroissante sur $\mathbb{R}^+$ telle que

$$\int_{\mathbb{R}^+}\sigma(v)v\mathrm{d}v<\infty.$$

Le théorème d'existence pour les solutions périodiques est alors donné par :

Théorème 5.1.1. *Soient $f_0(x,v)$ $L-$périodique en x et $n(v)\geq 0$ des fonctions continues vérifiant :*

(a) $|f_0(x,v)|\leq\sigma(|v|)$,

(b) $\int_{\mathbb{R}} n(v)|v|\mathrm{d}v<\infty$,

(c) $\int_0^x\int_{\mathbb{R}}|f_0(y,v)-n(v)|\mathrm{d}v\mathrm{d}y$ *uniformément borné en* x,

(d) $\int_0^L\int_{\mathbb{R}}[n(v)-f_0(x,v)]\mathrm{d}v\mathrm{d}x=0$.

Alors il existe un unique couple $(\mathcal{E},f)$ $L-$périodique en x, solution de (5.1.2) telle que $\mathcal{E}(t,x),\partial_x\mathcal{E}(t,x),\partial_t\mathcal{E}(t,x)$ sont continues sur $(0,\infty)\times\mathbb{R}$ et bornées sur $[0,T]\times\mathbb{R}$ pour chaque $T>0$ et $f(t,x,v)$ est continue sur $[0,\infty]\times\mathbb{R}^2$.

On a alors aussi une information supplémentaire sur la moyenne du champ

$$\overline{\mathcal{E}}(t) = \frac{1}{L}\int_0^L \mathcal{E}(t,x)\mathrm{d}x,$$

et sur le premier moment en vitesse de la fonction de distribution

$$\overline{\mathcal{V}}(t) = \int_0^L \int_{\mathbb{R}} [f(t,x,v) - n(v)]v\mathrm{d}v\mathrm{d}x.$$

Théorème 5.1.2. *Sous les hypothèses du théorème 5.1.1, en notant* $\alpha^2 = \int_{\mathbb{R}} n(v)\mathrm{d}v$, *on a :*

$$\begin{aligned} \overline{\mathcal{E}}(t) &= \overline{\mathcal{V}}_0 \alpha^{-1} \sin \alpha t, \\ \overline{\mathcal{V}}(t) &= \overline{\mathcal{V}}_0 \cos \alpha t, \end{aligned} \tag{5.1.4}$$

où

$$\overline{\mathcal{V}}_0(t) = \frac{1}{L}\int_0^L \int_{\mathbb{R}} [f_0(t,x,v) - n(v)]v\mathrm{d}v\mathrm{d}x.$$

Remarque 5.1.2. *En intégrant l'équation d'Ampère (5.1.2b) en* x, *on obtient facilement :*

$$\partial_t \overline{\mathcal{E}}(t) = \mathcal{V}(t).$$

La difficulté est alors d'obtenir

$$\partial_t \overline{\mathcal{V}}(t) = -\alpha^2 \overline{\mathcal{E}}(t),$$

comme f *n'est que supposée continue. L'argument utilisé dans [27] est de faire porter le temps sur le domaine d'intégration, par un changement de variables sur les caractéristiques. Notons aussi que pour établir cette dernière égalité la relation d'Ampère n'est pas utilisée.*

Le théorème d'existence se montre par un argument de point fixe sur le champ $\mathcal{E}$.
Plus précisément, on note $e(x)$ une primitive (définie à une constante additive près) de

$$\int_{\mathbb{R}} [n(v) - f_0(x,v)]\mathrm{d}v \tag{5.1.5}$$

En partant de $A_0(t,x) := e(x)$, on construit alors une suite de solutions approximantes

$$A_j(t,x) := \mathcal{L}(A_{j-1})(t,x), \tag{5.1.6}$$

et on montre que ces itérées convergent alors uniformément sur chaque intervalle de longueur finie vers une solution du théorème 5.1.1.
L'application $\mathcal{L}$ pour laquelle on applique le point fixe est définie de la manière suivante. D'après le lemme 5.1.1, pour $A \in X(T)$, l'application $(x_0, v_0) \to (x(t), v(t))$ admet un inverse que l'on note $\phi_A(t,x,v)$.
L'application $\mathcal{L}$ est alors donnée par :

$$\mathcal{L}(A)(t,x) = e(x) + \int_0^t \int_{\mathbb{R}} [f_0(\phi_A(s,x,v)) - n(v)]v\mathrm{d}v\mathrm{d}s. \tag{5.1.7}$$

5.2 Formulation du problème

On donne ici plusieurs formulations pour le système de Vlasov-Poisson périodique. Nous verrons plus précisément dans la section suivante que toutes ces formulations sont équivalentes.
On considère le système de Vlasov-Poisson :

$$\partial_t f + v\partial_x f - \frac{eE_s}{m_e}\partial_v f = 0$$
$$\partial_x E_s = \frac{e}{\varepsilon_0}\big(n_0 - \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v\big)$$

décrivant l'évolution de la fonction de distribution pour des électrons ayant une masse m_e et une charge électrique $-e$ soumis au champ électrique *auto-consistant* E_s.
Le paramètre n_0 correspond ici à une densité de charge de fond (on parle aussi de *fond neutralisant*).
Parfois, on rajoute un champ *appliqué* E_a :

$$\partial_t f + v\partial_x f - \frac{e(E_s+E_a)}{m_e}\partial_v f = 0,$$
$$\partial_x E_s = \frac{e}{\varepsilon_0}\big(n_0 - \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v\big).$$

Remarque 5.2.1. *En prenant un champ appliqué linéaire (c'est-à dire de la forme $E_a(x) = ax$, avec a une constante), on peut en fait se ramener à un système sans champ appliqué :*

$$\partial_t f + v\partial_x f - \frac{e(E_s+E_a)}{m_e}\partial_v f = 0,$$
$$\partial_x (E_s+E_a) = a + \frac{e}{\varepsilon_0}n_0 - \frac{e}{\varepsilon_0}\int_{\mathbb{R}} f(t,x,v)\mathrm{d}v.$$

Le fond neutralisant fait alors office de champ appliqué (et vice-versa).

Pour la suite, on va considérer le système de Vlasov-Poisson mono-dimensionnel avec une masse, une charge et une permittivité normalisée à 1.
On considère toujours L un nombre strictement positif, qui représente la période.

Solutions généralisées. En suivant [27], une *solution généralisée* du système de Vlasov-Poisson est une paire de fonctions $(E(t,x), f(t,x,v))$ L–périodiques en x telles que :

(a) $E(t,x), \partial_x E(t,x)$ et $\partial_t E(t,x)$ sont continues sur $[0,L]\times(0,\infty)$ et bornées sur $[0,L]\times[0,T]$ pour chaque $T>0$; $f(t,x,v)$ est continue sur $[0,\infty[\times\mathbb{R}^2$.

(b) $f(t,x,v)$ est constant le long des *courbes caractéristiques* :

$$\dot{X} = v, \quad \dot{V} = E(t,X).$$

(c) E et f satisfont :

$$\partial_x E(t,x) = \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v - 1, \tag{5.2.1}$$

$$\partial_t E(t,x) = \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v - \frac{1}{L}\int_0^L \int_{\mathbb{R}} v f_0(x,v)\mathrm{d}v\mathrm{d}x, \tag{5.2.2}$$

et

$$\int_0^L E(0,x)\mathrm{d}x = 0. \tag{5.2.3}$$

(d) $f(0,x,v) = f_0(x,v)$.

Autres formulations. On introduit d'abord la *fonction de Green* $G(x,y)$ qui est explicitement donnée par :

$$G(x,y) = \begin{cases} x(1-\frac{y}{L}), & 0 \le x \le y \\ y(1-\frac{x}{L}), & y \le x \le L \end{cases} \tag{5.2.4}$$

et le *noyau de Green* K :

$$K(x,y) = -\partial_x G(x,y) = \begin{cases} \frac{y}{L} - 1, & 0 \le x < y, \\ \frac{y}{L}, & y < x \le L. \end{cases} \tag{5.2.5}$$

On définit ensuite les autres équations pour le champ électrique :

$$\partial_t E(t,x) = \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v - \frac{1}{L}\int_0^L \int_{\mathbb{R}} v f(t,x,v)\mathrm{d}v\mathrm{d}x, \tag{5.2.6}$$

$$E(t,x) = \int_0^L K(x,y)\Big(\int_{\mathbb{R}} f(t,y,v)dv - 1\Big)dy, \tag{5.2.7}$$

$$\int_0^L E(t,x)\mathrm{d}x = 0, \tag{5.2.8}$$

ce qui nous amène à 4 *formulations*

- $[I]$ $\quad (a)+(b)+(5.2.1)+(5.2.2)+(5.2.3)+(d)$,
- $[II]$ $\quad (a)+(b)+(5.2.1)+(5.2.6)+(5.2.3)+(d)$,
- $[III]$ $\quad (a)+(b)+(5.2.1)+(5.2.8)+(d)$,
- $[IV]$ $\quad (a)+(b)+(5.2.7)+(d)$,

la première correspondant à la *solution généralisée* définie au paragraphe précédent.

Formulations au sens des distributions l'hypothèse $(b)+(d)$ peut aussi être remplacée par l'équation de Vlasov

$$\partial_t f + v\partial_x f + E\partial_v f = 0, \quad f(0,x,v) = f_0(x,v), \tag{5.2.9}$$

dans $\mathcal{D}'([0,T]\times[0,L]\times\mathbb{R})$ qui signifie ici

$$\int_0^T\int_0^L\int_{\mathbb{R}} f(\partial_t\phi + v\partial_x\phi + E(t,x)\partial_v\phi)\mathrm{d}v\mathrm{d}x\mathrm{d}t + \int_0^L\int_{\mathbb{R}} f_0\phi(0,x,v)\mathrm{d}v\mathrm{d}x = 0, \tag{5.2.10}$$

pour toute fonction test $\phi(t,x,v)$ infiniment dérivable à support compact en (t,v) et $L-$périodique en x.
Ceci conduit à de nouvelles *formulations* notées I'–IV'.

Quantités physiques : on définit
la *masse* du système :

$$\int_0^L\int_{\mathbb{R}} f(x,v)\mathrm{dvdx}, \tag{5.2.11}$$

le *premier moment de vitesse* :

$$\frac{1}{L}\int_0^L\int_{\mathbb{R}} vf(x,v)\mathrm{d}v\mathrm{d}x, \tag{5.2.12}$$

la *densité de charge* :

$$\rho(t,x) := \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v - 1, \tag{5.2.13}$$

la *densité de courant* :

$$j(t,x) := \int_{\mathbb{R}} vf(t,x,v)\mathrm{d}v. \tag{5.2.14}$$

Remarque 5.2.2. *(i) On peut mentionner aussi que l'équation (5.2.6) ou (5.2.2) est appelée la loi d'Ampère, tandis que l'équation (5.2.1) désigne la loi de Poisson.*

(ii) L'équation (5.2.8) est une condition de champ moyen nul.

5.3 Résultats d'existence

Pour l'étude des schémas numériques, on utilisera plus spécifiquement la formulation $[IV]$; cette formulation étant différente de celle utilisée dans l'article de Cooper-Klimas, précédemment mentionnée, on détaille ici l'équivalence entre toutes les formulations pour pouvoir se rattacher au théorème de Cooper-Klimas.
On a alors le théorème suivant :

Théorème 5.3.1. *Supposons que f_0 est continue dans $\mathbb{R}\times\mathbb{R}$, L-périodique en x et que*

$$|f(x,v)| \le \sigma(|v|), \quad \textit{pour tout } x, v \in \mathbb{R}, \tag{5.3.1}$$

où la fonction majorante σ est supposée être continue, monotone décroissante sur $\mathbb{R}^+$, et telle que

$$\int_{\mathbb{R}^+} \sigma(v) v \mathrm{d}v < \infty.$$

Alors il existe un unique couple (E,f) qui est une solution généralisée *; de plus, les 4 formulations sont équivalentes.*

Preuve. Ce résultat va découler de l'article de Cooper et Klimas [27].
On supposera toujours implicitement (sauf mention du contraire) que (E,f) est L périodique en x et vérifie (a).

• [**I**] Le théorème 5.1.1 s'applique en prenant

$$n(v) = \frac{1}{L}\int_0^L f_0(x,v)\mathrm{d}x.$$

En utilisant (5.2.1), E est $L-$périodique si et seulement si :

$$\frac{1}{L}\int_0^L \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v\mathrm{d}x = 1. \tag{5.3.2}$$

Donc toutes les 4 formulations vérifient la *conservation de la masse* (5.3.2).

• [**I**] $\Rightarrow$ [**II**, **III**] La solution généralisée vérifie $\overline{\mathcal{V}}_0 = 0$ et donc d'après le théorème 5.1.2, elle satisfait aussi (5.2.8) et (5.2.6).

• [**III**] $\Leftrightarrow$ [**IV**] Ensuite, une solution vérifie (5.2.1) et (5.2.8) si et seulement si elle vérifie (5.2.7). En effet, si on a (5.2.1) et (5.2.8), en écrivant $E = -\partial_x\phi$, le potentiel ϕ est solution de

$$\begin{array}{ll} -\partial_{xx}\phi = \int_{\mathbb{R}} f(t,x,v)\mathrm{d}v - 1, & x\in[0,L], \quad t\ge 0, \\ \phi(t,0) = \phi(t,L), & t\ge 0. \end{array} \tag{5.3.3}$$

Ce problème de Poisson avec un terme source continu admet une unique solution qui s'exprime à l'aide du noyau de Green et l'on en déduit l'expression du champ sous la forme (5.2.7). Réciproquement, si E est donné par (5.2.7), on vérifie qu'en prenant

$$\phi(t,x) = \int_0^L G(x,y)\left(\int_{\mathbb{R}} f(t,y,v)\mathrm{d}v - 1\right)\mathrm{d}y, \tag{5.3.4}$$

on a bien $E = -\partial_x\phi$; comme ϕ est alors la solution de (5.3.3), E vérifie alors bien (5.2.1) et (5.2.8).

• [**II**] $\Rightarrow$ [**III**] Puis, si une solution vérifie (5.3.2)+(5.2.3) (ce qui est vrai pour chaque

formulation) et (5.2.6), elle vérifie aussi (5.2.8).

• $[*] \Leftrightarrow [*]'$ On obtient ensuite l'équivalence entre $(b)+(d)$ et la formulation au sens des distributions (5.2.10).
Pour le sens direct, on utilise la relation :

$$\int_0^T \int_0^L \int_{\mathbb{R}} \frac{\mathrm{d}}{\mathrm{d}t}\Big(f(t,X(t),V(t))\Big)\phi(t,X(t),V(t))\mathrm{d}v\mathrm{d}x\mathrm{d}t = 0, \tag{5.3.5}$$

où $(X(t),V(t))$ sont les caractéristiques et un changement de variables sur les caractéristiques (en appliquant le lemme 5.1.1). Pour la réciproque, on considère par exemple une suite de fonctions tests de la forme $\psi_n(t)\theta(x,v)$, où θ est une fonction test quelconque et $\psi_n(t)$ est une approximation de la fonction continue affine par morceaux qui vaut 1 en 0 et 0 pour $t \geq \frac{1}{n}$. On en déduit alors que $f(0,x,v) = f_0(x,v)$ dans $\mathcal{D}'([0,L]\times\mathbb{R})$ et donc on a (d). On obtient alors, à nouveau avec le lemme 5.1.1, que :

$$\int_0^T \int_0^L \int_{\mathbb{R}} (f(t,X(t),V(t))\frac{\mathrm{d}}{\mathrm{d}t}\Big(\phi(t,X(t),V(t))\Big)\mathrm{d}v\mathrm{d}x\mathrm{d}t + \int_0^L \int_{\mathbb{R}} f(0,x,v)\phi(0,x,v)\mathrm{d}v\mathrm{d}x = 0. \tag{5.3.6}$$

Avec les notation habituelles $X(t) = X(t;0,x,v)$ et $V(t) = V(t;0,x,v)$ (les caractéristiques sont uniques), et pour une fonction test quelconque $\psi(t,x,v)$, on peut prendre ϕ sous la forme :

$$\phi(t,x,v) = \psi(t;X(0;t,x,v),V(0;t,x,v)), \tag{5.3.7}$$

et il s'ensuit que :

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t,X(t),V(t)) = 0, \tag{5.3.8}$$

dans $\mathcal{D}'([0,T]\times[0,L]\times\mathbb{R})$, et ceci implique bien (b).

• [**III**] $\Rightarrow$ [**II**] Grâce à cette nouvelle formulation, on obtient dans $\mathcal{D}'([0,T]\times[0,L])$ la *conservation locale de la masse*

$$\partial_t \rho + \partial_x j = 0, \tag{5.3.9}$$

en intégrant (5.2.9) par rapport à v.
L'intégration par rapport à v au sens des distributions consiste ici à utiliser une suite de fonctions tests de la forme $\psi(t,x)\rho_n(v)$, où ρ_n est une régularisée de la fonction indicatrice de l'intervalle $[-n,n]$; l'intégrabilité en v de la fonction de distribution $f(t,x,v)$ permet alors de justifier le passage à la limite.
Etant donnée ϕ infiniment dérivable et à support compact dans $[0,T]$, et $K_n(x,y)$ une régularisée en y de $K(x,y)$, on obtient :

$$\int_0^L \int_0^T \rho(t,y)\phi'(t)K_n(x,y)\mathrm{d}t\mathrm{d}y = -\int_0^L \int_0^T j(t,y)\partial_y K_n(x,y)\mathrm{d}t\mathrm{d}y. \tag{5.3.10}$$

Au sens des distributions, on a :

$$\partial_y K(x,y) = \frac{1}{L} - \delta(x-y). \tag{5.3.11}$$

En faisant tendre n vers l'infini, on obtient alors (5.2.6), si la solution vérifie (5.2.7) et $(b)+(d)$.

• [**IV**] $\Rightarrow$ [**I**] Enfin, dans le cas où l'on a (5.2.8) et (b)+(5.2.1)+(d), on obtient d'après le théorème 5.1.2 et la remarque 5.1.2 avec les notations correspondantes :

$$\partial_t \overline{\mathcal{V}} = -\alpha^2 \int_0^L \mathcal{E}(t,x)\mathrm{d}x. \tag{5.3.12}$$

Comme le membre de droite est nul d'après (5.2.8), et que $\overline{\mathcal{V}}(0,x) = 0$ d'après (d), on en conclut (5.2.2). □

Remarque 5.3.1. *On peut obtenir l'existence dans des cas plus faibles (voir P. L. Lions, B. Perthame [53] pour le cas $3D$), mais alors la forme caractéristique n'est plus toujours assurée ; l'unicité peut aussi être perdue.*

Pour la suite, on supposera toujours qu'au moins les hypothèses du théorème 5.3.1 sont vérifiées.

5.4 Quelques propriétés de l'équation de Vlasov

L'équation de Vlasov satisfait de nombreuses propriétés. On détaille ici pour le système que l'on considère les propriétés les plus essentielles, qui ont déjà été données à titre d'introduction dans le cas du système $3D$.

Courbes caractéristiques Les solutions du système d'équations différentielles ordinaires :

$$\dot{X}(t) = V(t), \quad \dot{V}(t) = E(t, X(t)), \tag{5.4.1}$$

soumises aux conditions initiales :

$$X(t_0) = x, \quad V(t_0) = v, \tag{5.4.2}$$

sont communément nommées *courbes caractéristiques*, et on écrit :

$$X(t; t_0, x, v), \quad V(t; t_0, x, v).$$

D'après la définition des solutions généralisées, on voit que la fonction de distribution f est *constante le long des caractéristiques.*
La solution du système de Vlasov-Poisson s'écrit alors :

$$f(t,x,v) = f(t, X(t;t,x,v)) = f_0(X(0;t,x,v)). \tag{5.4.3}$$

Remarque 5.4.1. *Cette propriété est essentielle pour la résolution numérique par une méthode de type semi-lagrangienne.*
Ainsi connaissant une approximation à un temps t, on peut en déduire une approximation au temps $t + \Delta t$, grâce à la formule :

$$f(t + \Delta t, x, v) = f(t + \Delta t, X(t + \Delta t; t + \Delta t, x, v)) = f(t, X(t; t + \Delta t, x, v)). \quad (5.4.4)$$

Positivité : de cette expression (5.4.3), on en déduit que si f_0 est positive, alors elle reste positive, ce qui est le cas pour les applications physiques (f représente une distribution de particules).

Principe du maximum : grâce à (5.4.3), on obtient aussi :

$$\inf_{x,v\in\mathbb{R}} f_0(x, v) \leq f(x, v) \leq \sup_{x,v\in\mathbb{R}} f_0(x, v). \quad (5.4.5)$$

Conservation de la masse. En intégrant l'équation de Vlasov (5.2.9) au sens des distributions, on obtient dans $\mathcal{D}'(0, T)$,

$$\frac{d}{dt}\Big(\int f(t, x, v)\mathrm{d}x\mathrm{d}v\Big) = 0,$$

ce qui exprime la conservation de la masse.
En intégrant l'équation de Poisson (5.2.1) par rapport à x, on obtient aussi

$$E(t, L) - E(t, 0) = \int f\mathrm{d}x\mathrm{d}v - L = 0, \quad (5.4.6)$$

puisque E est supposé L-périodique.

Conservation du premier moment. A partir de (5.2.2) et (5.2.6), on a :

$$\int (f_0 - f)v\mathrm{d}x\mathrm{d}v = \frac{\mathrm{d}}{\mathrm{d}t}\int_0^L E(t, x)\mathrm{d}x, \quad (5.4.7)$$

qui est nul grâce à (5.2.8).

Remarque 5.4.2. *(i) On peut aussi dériver cette conservation de l'équation de Vlasov, en multipliant l'équation par v, en intégrant et en utilisant (5.2.8).*

(ii) Les moments d'ordre supérieur peuvent ne pas être définis, si f_0 n'est pas assez régulière.

(iii) L'équation de Vlasov conserve aussi l'énergie (cf 4.4.3) :

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\int_0^L \int_{\mathbb{R}} f(t, x, v)\mathrm{d}v\mathrm{d}x + \int_0^L \big|E(t, x)\big|^2\mathrm{d}x\Big). \quad (5.4.8)$$

Préservation de mesures : d'après [27], par exemple (cf lemme 5.1.1), l'application

$$(x, v) \to (X(t; s, x, v), V(t; s, x, v)) \tag{5.4.9}$$

est bijective et son jacobien $J(t, s)$ vérifie : $|J(t, s)| \equiv 1$.
On en déduit, en utilisant ce changement de variable, que :

$$\int \beta(f(t, x, v))\mathrm{d}x\mathrm{d}v = \int \beta(f_0(x, v))\mathrm{d}x\mathrm{d}v, \tag{5.4.10}$$

pour chaque fonction continue β telle que l'intégrale soit bien définie.
En particulier, (5.4.10) implique la conservation de toutes les normes L^p.

5.5 Espaces de fonctions

Comment se comporte la régularité de la fonction de distribution

$$f : [0, T] \times [0, L] \times \mathbb{R} \to \mathbb{R}$$

suivant la condition initiale

$$f_0 : [0, L] \times \mathbb{R} \to \mathbb{R}?$$

L'équation de Vlasov, contrairement à certaines équations hyperboliques a la propriété de propager la régularité.
Ainsi, il est bien connu que si f_0 est m–fois continûment différentiable, alors il en est de même pour f (voir par exemple au chapitre suivant où l'on rappelle un tel théorème [théorème 6.2.1]).
On rappelle ici les principaux espaces qui servent à mesurer la régularité pour donner une vue générale. Néanmoins, on se limitera par la suite à des espaces plus spécifiques, pour lesquels on étudiera plus en détail l'évolution de la régularité dans la section §*Estimations a priori*.

Régularité. Notons que la notion de régularité n'est pas sujet à une définition bien établie. Classiquement, elle est liée au nombre de fois que la la fonction est dérivable. Avec l'apparition des distributions, on peut donner un sens à la dérivée de beaucoup de fonctions (toutes les fonctions localement intégrables). On cherche alors plutôt à borner les quantités :

$$\|f(\cdot + h) - f(\cdot)\|, \tag{5.5.1}$$

ou plus généralement (pour des ordres de régularité plus élevée) les *différences finies* :

$$\|\Delta_h^n f\|, \tag{5.5.2}$$

par une certaine puissance de h multiplié par une constante C. L'opérateur de différences finies d'ordre n est défini récursivement par

$$\Delta_h^1 f(x) = f(x+h) - f(x), \quad \Delta_h^n f(x) = \Delta_h^1(\Delta_h^{n-1}) f(x). \tag{5.5.3}$$

▷ Le choix de la norme $\|\cdot\|$ permet donc de considérer plusieurs types de régularité (pour un même ordre de régularité). Ainsi, suivant la norme que l'on considère, une fonction pourra avoir cet ordre de régularité ou non.

▷ Pour une régularité donnée, on peut dire aussi qu'une fonction est plus régulière qu'une autre si la constante C qui intervient pour l'une est plus petite que la constante pour l'autre.

▷ On peut en général lier (et parfois caractériser) l'ordre de régularité d'une fonction avec le degré d'approximation (linéaire ou non) de celle-ci par des classes d'éléments plus simples (en général : approximation polynômiale par morceaux), voir le paragraphe *Erreurs de projections* de ce chapitre. ▷ Ainsi, étant donné la fonction initiale f_0, on a intérêt à choisir la classe d'éléments pour que son ordre de régularité (donc d'approximation, si on a une telle caractérisation) soit le plus élevé possible.

▷ Néanmoins, pour un problème d'évolution (dépendant du temps), il ne suffit pas de regarder la donnée initiale. Il faut aussi savoir comment évolue la régularité avec le temps. Pour choisir une méthode numérique adaptée, il semble donc essentiel de comprendre comment se propage la régularité.

Remarque 5.5.1. *Une seconde manière de considérer la régularité est de donner des conditions de petitesse des coefficients d'une fonction exprimée dans une base. Ainsi par exemple, pour une base orthonormée (e_n) dans un espace de Hilbert, on peut définir des les classes V_α où les éléments s'écrivent sous la forme*

$$U = \sum_{k\in K} u_k e_k, \quad \sum |u_k|^2 k^{2\alpha} < \infty, \tag{5.5.4}$$

et le membre de droite permet de mesurer la régularité.

On rappelle maintenant les espaces de fonctions classiques utilisés en théorie des équations aux dérivées partielles et de l'approximation pour mesurer la régularité.

Notations. On considère une fonction $f : \mathbb{R}^d \to \mathbb{R}$, où la dimension d est un nombre entier strictement positif.
Ω désigne un domaine de $\mathbb{R}^d$.
Soit m un entier positif (en général : le degré de différentiation ou de régularité).
Pour tout multi-entier $\alpha = (\alpha_1, \ldots, \alpha_d)$, on définit

$$|\alpha| = \alpha_1 + \cdots + \alpha_d, \qquad \partial^\alpha = \partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d}.$$

Les fonctions de classe $\mathbf{C^m}$. On définit $C^m(\Omega)$ comme étant l'espace des fonctions continues dont les dérivées partielles $\partial^\alpha f$ sont bornées et continues pour tout $|\alpha| \leq m$. Cet espace est équipé de la norme :

$$\|f\|_{C^m(\Omega)} := \sup_{x\in\Omega} |f(x)| + \sum_{|\alpha|=m} \sup_{x\in\Omega} |\partial^\alpha f(x)|, \tag{5.5.5}$$

et forme ainsi un espace de Banach.
On rajoutera un indice c pour désigner les sous-espaces contenant les fonctions à support compact dans Ω, et L_x pour désigner les fonctions L-périodiques par rapport à la variable x. On a ainsi ici par exemple $C^m_c(\Omega), C^m_{2\pi_x}(\Omega)$. Notons que $C^m_{c,2\pi_x}(\Omega)$ est l'ensemble des fonctions sont périodiques par rapport à x et à support compact par rapport aux autres variables.

Espaces de Hölder. Pour traiter des dérivées d'ordre non entier, on définit $C^s(\Omega)$, pour $m < s < m+1$, l'espace des fonctions continues vérifiant

$$\sup_{x\in\Omega} |\partial^\alpha f(x+h) - \partial^\alpha f(x)| \leq C|h|^s, \quad |\alpha| = m. \tag{5.5.6}$$

Il s'agit d'un espace de Banach pour la norme :

$$\|f\|_{C^s(\Omega)} := \sup_{x\in\Omega} |f(x)| + \sum_{|\alpha|=m} \sup_{x\neq y\in\Omega} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|} \tag{5.5.7}$$

Remarque 5.5.2. *$C^s(\Omega)$ se note parfois aussi $W^{s,\infty}$ lorsque s est non entier.*

Espaces de Sobolev. Pour mesurer la régularité en moyenne, il est naturel d'introduire les espaces de Sobolev $W^{m,p}(\Omega)$ qui consistent en les distributions $f \in L^p$ dont les dérivées partielles jusqu'à l'ordre m sont dans L^p, avec un entier $1 \leq p \leq \infty$. Cet espace muni de la norme

$$\|f\|_{W^{m,p}} := \|f\|_{L^p} + |f|_{W^{m,p}}, \quad |f|_{W^{m,p}} := \sum_{|\alpha|=m} \|\partial^\alpha f\|_{L^p}. \tag{5.5.8}$$

est aussi un espace de Banach.

Remarque 5.5.3. *On peut noter que les espaces C^m et $W^{m,\infty}$ ont la même norme; pourtant ces espaces sont différents.*

Espaces de Sobolev fractionnaires. Similairement aux espaces de Hölder, on peut définir pour $m < s < m+1$, les espaces $W^{s,p}$ qui sont les distributions f telles que $\|f\|_{W^{s,p}} = \|f\|_{L^p} + |f|_{W^{s,p}}$ soit fini, avec

$$|f|_{W^{s,p}} := \sum_{|\alpha|=m} \int_{\Omega\times\Omega} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|^{(s-m)p+d}} \mathrm{d}x\mathrm{d}y, \quad p < \infty \tag{5.5.9}$$

Espaces de Lipschitz. On définit l'espace $\mathrm{Lip}(s, L_p(\Omega))$

$\triangleright$ pour $m > 0$ entier et $p = \infty$, il s'agit du cas limite $s = m + 1$ pour les espaces de Hölder ; c'est aussi $W^{m,\infty}$ pour une norme équivalente.

$\triangleright$ dans les autre cas, on prend pour définition $W^{s,p}$.

Espaces de Besov. Une description encore plus subtile de la régularité peut se faire à l'aide des espaces de Besov $B^\alpha_q(L_p(\Omega))$, où $p, q \geq 1$ sont les paramètres *primaire* et *secondaire*. Ils sont définis par les distributions f vérifiant

$$|f|_{B^\alpha_\infty(L_p(\Omega))} := \sup_{t>0} t^{-\alpha} \omega_{\lfloor\alpha\rfloor+1}(f,t)_p < \infty, \tag{5.5.10}$$

et

$$|f|_{B^\alpha_q(L_p(\Omega))} := \left(\int_0^\infty [t^{-\alpha} \omega_{\lfloor\alpha\rfloor+1}(f,t)_p]^q \frac{\mathrm{dt}}{t} \right)^{1/q} < \infty, \quad q < \infty \tag{5.5.11}$$

Il s'agit d'un espace de Banach en prenant la norme :

$$\|f\|_{B^s_{p,q}} := \|f\|_{L^p} + |f|_{B^s_{p,q}}. \tag{5.5.12}$$

On a utilisé ici le r-ième *module de continuité dans* $L^p(\Omega)$ défini par :

$$\omega_r(f,t)_p := \sup_{|h|\leq t} \|\Delta^r_h(f,\cdot)\|_{L_p(\Omega)}. \tag{5.5.13}$$

Remarque 5.5.4. *Le* sup *est pris sur tous les vecteurs dans* $\mathbb{R}^d$ *de norme euclidienne* $|h| \leq t$ *et on a pris comme convention que* $\Delta^r_h(f,x) = 0$ *si un des points* $x, \ldots, x + rh$ *n'est pas dans* Ω.

Remarque 5.5.5. *(i) On a* $W^{s,p} = B^s_{p,p}$ *si et seulement si* s *n'est pas un entier ou* $p = 2$.

(ii) Ainsi, tous les espaces que l'on vient de décrire se résument par :

$$B^s_{p,q}, \quad W^{m,p} \quad \text{et} \quad C^m, \tag{5.5.14}$$

avec $s \in \mathbb{R}^+$, $m \in \mathbb{N}$, $1 \leq p, q \leq \infty$ *et les seuls espaces communs sont* $H^m := B^m_{2,2} = W^{m,2}$.

Fonctions en temps. Etant donné un espace de Banach X, on peut considérer des fonctions $f(t,\cdot) \in X$; on peut alors redéfinir tous les espaces précédents en remplacant la valeur absolue des fonctions à valeurs réelles par la norme de l'espace X.
On note alors par exemple $C(0,T,X), L^\infty(0,T,X), |f|_{L^\infty(0,t;X)}$.

Inclusions. Les espaces $B^s_{p,q}$, $W^{m,p}$, et C^m vérifient des relations d'inclusions.

$\triangleright$ Pour p fixé et q arbitraire, les espaces $W^{s,p}$ et $B^s_{p,q}$ deviennent de plus en plus petits, lorsque s augmente.

$\triangleright$ Pour s et q fixé et dans le cas de domaines bornés, ces espaces s'agrandissent avec $1/p$.

$\triangleright$ Pour s et p fixés, les espaces $B^s_{p,q}$ grandissent avec q.

Pour une description plus précise, en prenant $s_1 \geq s_2 > 0$, $1 \leq p_1 \leq p_2$, on a les injections de Sobolev et de Besov suivantes :

$\triangleright$ Pour $s_1 - s_2 \geq d(1/p_1 - 1/p_2)$ et $p_2 \neq \infty$, on a $W^{s_1,p_1} \subset W^{s_2,p_2}$

$\triangleright$ Pour $s_1 - s_2 \geq d(1/p_1 - 1/p_2)$, on a $B^{s_1}_{p_1,p_2} \subset B^{s_2}_{p_2,p_2}$.

Si les espaces de Besov traitent bien le cas $p_2 = \infty$, la discussion est plus délicate dans le cas des espaces de Sobolev :

$\triangleright$ Si $s_1 - s_2 > d/p$, on a $W^{s_1,p} \subset C^{s_2} \subset W^{s_2,\infty}$.

$\triangleright$ Si $s_2 = s_1 - d/p$ n'est pas un entier, on a $W^{s_1,p} \subset C^{s_2} = W^{s_2,\infty}$.

$\triangleright$ Si $s_2 = s_1 - d/p$ est un entier et $p > 1$, on n'a pas de telle inclusion.

$\triangleright$ Si $p = 1$ et m est un entier, on a $W^{m+d,1} \subset C^m \subset W^{m,\infty}$.

Remarque 5.5.6. *Le domaine Ω doit satisfaire certaines hypothèses pour satisfaire ces inclusions ; elles sont toujours vérifiées si Ω est borné de frontière C^∞ ; mais pour la plupart, il suffit que Ω ait la* propriété de cône *(on pourra consulter [1] ou [72] pour plus de détails).*

On regarde maintenant plus précisément les espaces de fonctions auxquels on s'intéressera pour l'étude de la convergence du schéma numérique adaptatif du chapitre 7.
Le lemme suivant est contenu dans les inclusions précédentes, on donne ici une preuve élémentaire.

Lemme 5.5.7. *Supposons que $g \in W^{2,1}(\Omega)$, où Ω est un domaine rectangulaire, alors g est continue.*

Preuve. Pour ne pas alourdir les calculs, on prend $\Omega = [0,1]^2$. Supposons d'abord, pour simplifier que g s'annule en $(0,0)$; on obtient alors :

$$\begin{aligned} g(x,v) &= g(x,0) + \int_0^v \partial_v g(x,w)\, dw \\ &= g(x,0) + \int_0^v \partial_v g(0,w)\, dw + \int_0^x \int_0^v \partial^2_{xv} g(u,w)\, dw\, du \\ &= \int_0^x \partial_x g(u,0)\, du + \int_0^v \partial_v g(0,w)\, dw + \int_0^x \int_0^v \partial^2_{xv} g(u,w)\, dw\, du, \end{aligned}$$

et donc

$$\|g\|_{L^\infty([0,1]^2)} \leq \|g(\cdot,0)\|_{W^{1,1}([0,1])} + \|g(0,\cdot)\|_{W^{1,1}([0,1])} + \|g\|_{W^{2,1}([0,1]^2)}.$$

De la même manière, à partir de

$$\partial_x g(x,0) = \partial_x g(x,v) - \int_0^v \partial^2_{xv} g(x,w)\, dw,$$

on a l'inégalité de trace

$$\|g(\cdot,0)\|_{W^{1,1}([0,1])} \leq \|g\|_{W^{2,1}([0,1]^2)},$$

qui nous mène à

$$\|g\|_{L^\infty([0,1]^2)} \leq 4\|g\|_{W^{2,1}([0,1]^2)},$$

si g n'est pas forcément nulle en $(0,0)$. Maintenant, pour voir que g est continue, considérons une suite de fonctions continues (ϕ_n) qui tend vers g dans $W^{2,1}$ (les fonctions continues sont denses dans L^1 et donc dans $W^{2,1}$) ; d'après l'inégalité précédente appliquée à $\phi_n - g$, on obtient la convergence dans L^∞, et donc g hérite de la continuité de la suite (ϕ_n). □

Les fonctions de $W^{1,\infty}$ sont aussi continues ; on rappelle en fait la caractérisation suivante (voir par exemple [21]) :

Proposition 5.5.8. *Une fonction g appartient à $W^{1,\infty}(\Omega)$ si et seulement si elle est lipschitzienne.*

5.6 Erreurs de projection.

Une des étapes importantes dans un schéma semi-lagrangien, est la projection sur un maillage. L'erreur commise pour une projection qui préserve des polynômes (cas particulier : une interpolation), entre une fonction et sa fonction projetée, peut être analysée par la régularité dans les espaces de Sobolev de la fonction en question.
On rappelle ici les notations et un théorème énoncé dans [23].
Ce théorème sera utilisé dans le cadre de l'espace $W^{2,1}$ au chapitre 7 ; il peut servir aussi à démontrer les estimations (6.3.6) et (6.3.12) du chapitre 6.

Notations. Deux sous-ensembles Ω et $\hat{\Omega}$ de $\mathbb{R}^d$ sont dits *affinement équivalents*, s'il existe une application affine inversible

$$F : \hat{x} \in \mathbb{R}^d \to F(\hat{x}) = B\hat{x} + b \in \mathbb{R}^d, \tag{5.6.1}$$

telle que

$$\Omega = F(\hat{\Omega}). \tag{5.6.2}$$

On a alors une correspondance entre les points $x \in \Omega$ et $\hat{x} \in \hat{\Omega}$:

$$\hat{x} \in \Omega \to x = F(\hat{x}) \in \Omega, \tag{5.6.3}$$

et aussi entre les fonctions définies sur Ω et $\hat{\Omega}$:

$$(\hat{v} : \Omega \to \mathbb{R}) \to (v = \hat{v} \circ F^{-1} : \Omega \to \mathbb{R}). \tag{5.6.4}$$

Ainsi, on a :

$$\hat{v}(\hat{x}) = v(x), \tag{5.6.5}$$

pour tous les points x et $\hat{x}$ en correspondance 5.6.3, et pour toutes les fonctions v et $\hat{v}$ en correspondance 5.6.4.

On peut alors énoncer un théorème classique dans le cadre des espaces de Sobolev (cf [23, théorème 15.3]).

Théorème 5.6.1. *Soit $\hat{\Omega}$ un domaine ouvert borné connexe de $\mathbb{R}^d$.*
Supposons que pour des entiers $k \geq 0$ et $m \geq 0$ et des nombres $p, q \in [0, \infty]$, on ait :

$$W^{k+1,p}(\hat{\Omega}) \subset W^{m,p}(\hat{\Omega}), \tag{5.6.6}$$

et soit $\hat{\Pi}$ une application linéaire de $W^{k+1,p}(\hat{\Omega})$ dans $W^{m,p}(\hat{\Omega})$ vérifiant :

$$\hat{\Pi}\hat{p} = \hat{p}, \tag{5.6.7}$$

pour tout polynôme $\hat{p}$ de $\hat{\Omega}$ dans $\mathbb{R}$, de degré $\leq k$.
Pour tout ouvert Ω affinement équivalent à $\hat{\Omega}$, l'application Π_Ω est définie par :

$$\widehat{\Pi_\Omega v} = \hat{\Pi}\hat{v}, \tag{5.6.8}$$

pour toutes les fonctions $\hat{v} \in W^{k+1,p}(\hat{\Omega})$ et $v \in W^{k+1,p}(\Omega)$ en correspondance (5.6.4).
Alors, il existe une constante $C(\hat{\Pi}, \hat{\Omega})$, telle que, pour tous les ensembles Ω affinement équivalents, on ait :

$$|v - \Pi_\Omega v|_{W^{m,q}(\Omega)} \leq C(\hat{\Pi}, \hat{\Omega})\{mes(\Omega)\}^{1/q-1/p}\frac{h^{k+1}}{\rho^m}|v|_{W^{k+1,p}(\Omega)}, \tag{5.6.9}$$

pour tout $v \in W^{k+1,p}(\Omega)$, où $mes(\Omega)$ désigne le volume de Ω, h est le diamètre (extérieur) de Ω défini par

$$h = diam(\Omega) := \sup_{x,y \in \mathbb{R}^d \cap \Omega} \|x - y\| \tag{5.6.10}$$

(la norme $\|\cdot\|$ est ici la norme euclidienne sur $\mathbb{R}^d$), et ρ est le diamètre intérieur :

$$\rho := \sup\{diam(S); S \subset \Omega \text{ est une boule}\}. \tag{5.6.11}$$

5.7 Estimations à priori

On étudie maintenant plus en détail l'évolution de certaines normes au cours du temps. Plus précisément, on dérive déjà dans un premier temps des estimations qui sont disponibles, sans hypothèses supplémentaires sur la fonction de distribution (hormis celles mentionnées ci-après). Puis, en supposant que la fonction initiale est dans $W^{1,\infty}$, on obtient alors les estimations nécessaires pour l'étude de l'erreur de discrétisation en temps du schéma semi-Lagrangien avec splitting en temps, utilisé dans les chapitres 6 et 7.
Notons que pour le schéma adaptatif, on supposera aussi que la fonction de distribution est dans $W^{2,1}$, ceci pour contrôler les erreurs de projections (comme mentionné dans la section précédente). Il semble donc alors aussi intéressant (même si on n'utilisera pas ce résultat) de voir comment évolue la semi-norme $|f(t)|_{W^{2,1}}$ au cours du temps. Si on est assuré que f est dans $W^{2,1}$ pour tout temps fini (on le montrera pour simplifier sous la condition plus forte $f_0 \in W^{2,\infty}$), on a alors une estimation du type :

$$|f(t)|_{W^{2,1}} \leq e^{\alpha T^2+\beta T}|f_0|_{W^{2,1}}, \quad 0 \leq t \leq T. \tag{5.7.1}$$

où les constantes α, β sont explicitement données et ne dépendent que de la norme infinie de f_0, de son support en vitesse et de la période L.

On considère pour la suite que la donnée initiale f_0 vérifie en outre :

- $f_0 \geq 0$.
- f_0 à support compact.

On cherche à avoir des informations supplémentaires sur la solution généralisée f suivant la régularité de la condition initiale f_0.
On introduit alors la constante $Q(t)$ qui est définie comme la taille du support de $f(t)$ dans la direction v :

$$Q(t) := 1 + \sup\{|v| : \exists x,\, \exists \tau \in [0,t],\, f(t,x,v) > 0\}. \tag{5.7.2}$$

Remarque 5.7.1. *(i) Cette quantité a été introduite dès 1977. Batt ([5]) avait remarqué que le contrôle de cette quantité ajouté à l'existence d'une solution locale permettait d'obtenir un théorème d'existence globale.*
(ii) Comme f_0 est à support compact, on est assuré que $Q(0)$ est fini.

Notations. On note ici $\Omega = [0, L] \times \mathbb{R}$ le domaine de l'espace des phases.
Pour la suite, la lettre C désignera une constante (qui peut varier à chaque occurence), et on écrira $C(t)$, si elle dépend du temps t, ou C_T, si elle dépend du temps final T. On utilisera parfois aussi la notation $\lesssim$:

$$a \lesssim b \tag{5.7.3}$$

signifie qu'il existe une constante C indépendante de a et b telle que $a \leq Cb$.

Premières estimations. Sans imposer d'hypothèse supplémentaire sur les données, on a déjà (automatiquement) les estimations suivantes :

Lemme 5.7.2. *La fonction de distribution $f(t,x,v)$ reste positive pour tout $t>0$, son support en vitesse vérifie*

$$Q(T) \leq Q(0) + 2LT, \tag{5.7.4}$$

et on a :

$$\|f(t)\|_{L^1(\Omega)} = \|f_0\|_{L^1(\Omega)} = L, \quad \|f(t)\|_{L^\infty(\Omega)} = \|f_0\|_{L^\infty(\Omega)}. \tag{5.7.5}$$

Le champ électrique satisfait

$$\|E(t)\|_{L^\infty([0,L])} \leq 2L, \tag{5.7.6}$$

$$|E(t)|_{W^{1,\infty}([0,L])} \leq 2(Q(0)+2LT)\|f_0\|_{L^\infty(\Omega)}, \tag{5.7.7}$$

$$\|\partial_t E(t)\|_{L^1([0,L])} \leq 2L + 2Q(0) + 4LT, \tag{5.7.8}$$

et

$$\|\partial_t E(t)\|_{L^\infty([0,L])} \leq \|f_0\|_{L^\infty(\Omega)}(Q(0)+2LT)(Q(0)+2LT+1). \tag{5.7.9}$$

Preuve. Comme f est constante le long des courbes caractéristiques, on en déduit que f reste positive pour tout temps, que f vérifie le principe du maximum ; on obtient alors (5.7.5) et

$$\|\rho\|_{L^1} \leq 2L. \tag{5.7.10}$$

En utilisant (5.2.7), on en déduit une borne sur le champ électrique.

$$\|E(t)\|_{L^\infty([0,L])} \leq 2L\|K\|_{L^\infty([0,L])} \leq 2L. \tag{5.7.11}$$

On dérive aussi une borne pour $Q(T)$ à partir de la forme de f définie via les caractéristiques :

$$Q(T) = 1 + \sup_{0\leq t\leq T,(x,v)\in\Omega_0} |V(\tau;0,x,v)|, \tag{5.7.12}$$

où $\Omega_0 := \{(x,v) : f_0(x,v) > 0\}$, et donc

$$Q(T) \leq Q(0) + T\|E\|_{L^\infty(0,T;L^\infty)} \leq Q(0) + 2LT. \tag{5.7.13}$$

Ainsi, avec la relation de Poisson (5.2.1), on a :

$$\|\partial_x E(t)\|_{L^\infty([0,L])} \leq 2Q(T)\|f_0\|_{L^\infty(\Omega)}, \tag{5.7.14}$$

ce qui donne le résultat.
Enfin, (5.7.8) et (5.7.9) s'obtiennent en utilisant la loi d'Ampère (5.2.6).
Pour établir (5.7.9), on calcule ainsi :

$$\|\partial_t E(t)\|_{L^\infty(\Omega)} \leq \|f_0\|_{L^\infty(\Omega)} \left(Q(T) + 2\int_0^{Q(T)} |v|\mathrm{d}v \right) \leq \|f_0\|_{L^\infty(\Omega)}(Q(T)+Q(T)^2), \tag{5.7.15}$$

tandis que pour établir (5.7.8), on calcule :

$$\|\partial_t E(t)\|_{L^1(\Omega)} \leq \|f(t)\|_{L^1(\Omega)} + \frac{Q(T)}{L}\|f_0\|_{L^1(\Omega)} = 2L\left(1+\frac{Q(T)}{L}\right). \tag{5.7.16}$$

□

Régularité $W^{1,\infty}$. Les précédentes estimations étaient valables en supposant seulement que f_0 est continue.
On suppose maintenant que f_0 est dans $W^{1,\infty}(\Omega)$, ce qui est équivalent à supposer que f_0 est lipschitzienne. Alors la régularité $W^{1,\infty}$ est conservée pour $f(t)$; ce résultat est bien connu (cf [16], par exemple). On donne ici une preuve de ce résultat.
On rappelle d'abord l'inégalité de Gronwall :

Lemme 5.7.3. *Soit α, β et y trois fonctions continues sur un segment $[a,b]$, à valeurs positives et vérifiant :*

$$y(t) \leq \alpha(t) + \int_a^t \beta(s)y(s)\mathrm{d}s. \tag{5.7.17}$$

Alors, pour tout $t \in [a,b]$, on a :

$$y(t) \leq \alpha(t) + \int_a^t \alpha(s)\beta(s)e^{\int_s^t \beta(u)\mathrm{d}u}\mathrm{d}s. \tag{5.7.18}$$

En particulier, lorsque α et β ne dépendent pas de t, on a :

Proposition 5.7.4. *Soient $T > 0$, $\alpha, \beta \in \mathbb{R}^+$ et y une fonction positive sur $\mathbb{R}^+$ vérifiant :*

$$y(t) \leq \alpha + \beta \int_0^t y(s)\mathrm{d}s, \quad t \in [0,T]. \tag{5.7.19}$$

Alors, pour tout $t \in [0,T]$, on a :

$$y(t) \leq \alpha e^{\beta t}. \tag{5.7.20}$$

On a alors le lemme suivant.

Lemme 5.7.5. *Supposons que $f_0 \in W^{1,\infty}(\Omega)$, alors $f(t) \in W^{1,\infty}(\Omega)$.*

Preuve. La donnée initiale est lipschitzienne, donc il existe un nombre $K > 0$ tel que :

$$|f_0(x,v) - f_0(y,w)| \leq K(|x-y| + |v-w|). \tag{5.7.21}$$

En utilisant la forme caractéristique de f, on obtient alors :

$$\begin{aligned} |f(t,x,v) - f(t,y,w)| \leq K(&|X(-t;0,x,v) - X(-t;0,y,w)| \\ &+ |V(-t;0,x,v) - V(-t;0,y,w)|) \end{aligned} \tag{5.7.22}$$

On calcule ensuite

$$\begin{aligned} X(-t;0,x,v) - X(-t;0,y,w) - (x-y) &= \int_0^{-t} \dot{X}(s;0,x,v) - \dot{X}(s;0,y,w)\mathrm{d}s \\ &= \int_0^{-t} V(s;0,x,v) - V(s;0,y,w)\mathrm{d}s, \end{aligned} \tag{5.7.23}$$

et de même

$$V(-t;0,x,v) - V(-t;0,y,w) - (v-w) = \int_0^{-t} \dot{V}(s;0,x,v) - \dot{V}(s;0,y,w)\mathrm{d}s$$
$$= \int_0^{-t} E(s, X(s;0,x,v)) - E(s, X(s;0,y,w))\mathrm{d}s. \quad (5.7.24)$$

En utilisant la borne $W^{1,\infty}$ du champ E, on obtient :

$$|E(s,X(s;0,x,v)) - E(s,X(s;0,y,w))|$$
$$\leq |E(s)|_{W^{1,\infty}([0,L])}\ |X(s;0,x,v) - X(s;0,y,w)|$$
$$\leq A_T |X(s;0,x,v) - X(s;0,y,w)|, \quad (5.7.25)$$

avec

$$A_T := 2(Q(0) + 2LT)\|f_0\|_{L^\infty(\Omega)}. \quad (5.7.26)$$

En notant maintenant

$$e_x(s) := |X(s;0,x,v) - X(s;0,y,w)|, \quad e_v(s) := |V(s;0,x,v) - V(s;0,y,w)|, \quad (5.7.27)$$

on obtient

$$e_x(-t) \leq |x-v| + \int_{-t}^0 e_v(s)\mathrm{d}s, \quad e_v(-t) \leq |y-w| + A_T \int_{-t}^0 e_x(s)\mathrm{d}s, \quad (5.7.28)$$

et donc en notant $e := e_x + e_v$, on a

$$e(-t) \leq e(0) + (1+A_T)\int_0^t e(-s)\mathrm{d}s. \quad (5.7.29)$$

En utilisant l'inégalité de Gronwall de la proposition 5.7.4, on en déduit alors que

$$e(-t) \leq e^{(1+A_T)t} e(0), \quad (5.7.30)$$

et donc, avec (5.7.22), on obtient

$$|f(t,x,v) - f(t,y,w)| \leq K e^{(1+A_T)t}(|x-y| + |v-w|), \quad (5.7.31)$$

ce qui nous assure que $f(t) \in W^{1,\infty}(\Omega)$. □

Estimations $W^{1,p}$. On suppose maintenant que la donnée initiale f_0 est dans $W^{1,\infty}$, ainsi d'après le lemme précédent, il en est de même pour $f(t)$. On donne alors différentes estimations qui ne pouvaient pas être obtenues lorsque f_0 était juste supposée continue. Le domaine Ω étant borné en x, et comme on a vu que le support de f reste également borné par une constante C_T, on sait alors que $f(t)$ est dans $W^{1,p}$, pour tout $1 \leq p \leq \infty$. On utilise les arguments de Glassey ([41]), pour le lemme suivant.

Lemme 5.7.6. *Si $f_0 \in W^{1,\infty}(\Omega)$ alors pour $1 \leq p \leq \infty$, $T > 0$ et $t \in [0,T]$, on a :*

$$|f(t)|_{W^{1,p}(\Omega)} \leq |f_0|_{W^{1,p}(\Omega)} e^{\beta_T T}, \tag{5.7.32}$$

avec

$$\beta_T := (2Q(0) + 2LT)\|f_0\|_{L^\infty(\Omega)} + 1. \tag{5.7.33}$$

On a aussi

$$\|\partial_t f(t)\|_{L^p(\Omega)} \leq (Q(0) + 2LT + 2L)|f_0|_{W^{1,p}(\Omega)} e^{\beta_T T}. \tag{5.7.34}$$

Pour le champ électrique, les quantités

$$|\partial_t E(t)|_{W^{1,p}([0,L])} \quad et \quad \|\partial_t \partial_t E(t)\|_{L^p([0,L])}, \tag{5.7.35}$$

sont majorées par

$$(2Q(0) + 4LT)^{1-1/p}(Q(0) + 2LT + 2L)|f_0|_{W^{1,p}(\Omega)} e^{\beta_T T}, \tag{5.7.36}$$

et on a également

$$|E(t)|_{W^{2,p}([0,L])} \leq (2Q(0) + 4LT)^{1-1/p}|f_0|_{W^{1,p}(\Omega)} e^{\beta_T T}. \tag{5.7.37}$$

Preuve. On dérive (5.2.9) par rapport à v

$$\partial_t(\partial_v f) + v\partial_x(\partial_v f) + E\partial_v(\partial_v f) = -\partial_x f, \tag{5.7.38}$$

au sens des distributions.
On définit ensuite φ par $\varphi_{x,v}(s) := \partial_v f(s, X(s), V(s))$ avec

$$X(s) := X(s;0,x,v), \quad V(s) := V(s;0,x,v). \tag{5.7.39}$$

On obtient alors

$$\dot{\varphi}_{x,v}(s) = -\partial_x f(s, X(s), V(s)),$$

au sens des distributions, de telle sorte que

$$|\varphi_{x,v}(t)| \leq |\varphi_{x,v}(0)| + \int_0^t |\partial_x f(s, X(s), V(s))|\, ds,$$

cette inégalité ayant bien un sens pour presque tous t, x et v, puisque l'on sait que $f(s) \in W^{1,\infty}(\Omega)$, pour $s \in [0,T]$.
On a ainsi

$$|\partial_v f(t, X(t), V(t))| \leq |\partial_v f_0(x,v)| + \int_0^t |\partial_x f(s, X(s), V(s))|\mathrm{d}s. \tag{5.7.40}$$

De la même manière, à partir de ∂_x(5.2.9) (on utilise cette notation pour dire que l'on dérive (5.2.9) par rapport à v)

$$\partial_t(\partial_x f) + v\partial_x(\partial_x f) + E\partial_v(\partial_x f) = -\partial_x E \partial_v f, \tag{5.7.41}$$

on obtient

$$\begin{aligned}|\partial_x f(t,X(t),V(t))| &\\ &\leq |\partial_x f_0(x,v)| + \int_0^t |E(s)|_{W^{1,\infty}(\Omega)} |\partial_v f(s,X(s),V(s))| \mathrm{d}s \\ &\leq |\partial_x f_0(x,v)| + (2Q(0)+2LT)\|f_0\|_{L^\infty(\Omega)} \int_0^t |\partial_v f(s,X(s),V(s))| \mathrm{d}s,\end{aligned}$$

c'est-à-dire

$$|\partial_x f(t,X(t),V(t))| \leq |\partial_x f_0(x,v)| + (\beta_T - 1)\int_0^t |\partial_v f(s,X(s),V(s))| \mathrm{d}s. \tag{5.7.42}$$

En notant $y(t) := |\partial_x f(t,X(t),V(t))| + |\partial_v f(t,X(t),V(t))|$, on a donc

$$y(t) \leq y(0) + \beta_T \int_0^t y(s) \mathrm{d}s, \tag{5.7.43}$$

en sommant les inégalités (5.7.40) et (5.7.42).
Il découle alors de l'inégalité de Gronwall de la proposition 5.7.4 :

$$y(t) \leq y(0) e^{\beta_T T}. \tag{5.7.44}$$

• Si $p = \infty$, on a alors

$$y(t) \leq |f_0|_{W^{1,\infty}(\Omega)} e^{\beta_T T}. \tag{5.7.45}$$

pour presque tous $x, v \in \Omega$ et tout $t \in [0,T]$.
Comme l'application $(x,v) \to (X(t;0,x,v), V(t;0,x,v))$ est bijective, on en déduit (5.7.32).

• Si $p < \infty$, on élève (5.7.44) à la puissance p et on intègre ; grâce à la préservation de la mesure, on en déduit alors également (5.7.32).
Avec l'équation de Vlasov (5.2.9), on obtient

$$\|\partial_t f\|_{L^p(\Omega)} \leq Q(T)\|\partial_x f\|_{L^p(\Omega)} + \|E\|_{L^\infty([0,L])}\|\partial_v f\|_{L^p(\Omega)}, \tag{5.7.46}$$

et (5.7.34) en découle.
En ce qui concerne les estimations sur le champ électrique, elles sont obtenues grâce aux relations d'Ampère (5.2.6) et de Poisson (5.2.1).
Ainsi, en dérivant l'équation de Poisson (5.2.1) par rapport à t, on obtient :

$$\partial_t \partial_x E(t,x) = \int_{\mathbb{R}} \partial_t f(t,x,v) \mathrm{d}v, \tag{5.7.47}$$

et aussi

$$\partial_t^2 E(t,x) = \int_{\mathbb{R}} \partial_t f(t,x,v) \mathrm{d}v, \tag{5.7.48}$$

en dérivant la relation d'Ampère. Si l'on dérive (5.2.1) par rapport à x, on trouve de plus

$$\partial_x^2 E(t,x) = \int_{\mathbb{R}} \partial_x f(t,x,v) \mathrm{d}v. \tag{5.7.49}$$

Pour $p < \infty$, on a d'après l'inégalité de Jensen :

$$\int_0^L \left| \int_{-Q(T)}^{Q(T)} 2Q(T)\partial_t f(t,x,v)\frac{\mathrm{d}v}{2Q(T)} \right|^p \mathrm{d}x \leq \int_\Omega (2Q(T)\partial_t f(t,x,v))^p \frac{\mathrm{d}v}{2Q(T)}\mathrm{d}x \quad (5.7.50)$$

et donc :

$$\|\partial_t \partial_x E(t)\|_{L^p\Omega} = \|\partial_t^2 E(t)\|_{L^p(\Omega)} \leq (2Q(T))^{1-1/p}\|\partial_t f(t)\|_{L^p(\Omega)}, \quad (5.7.51)$$

ce qui donne la majoration (5.7.36).
La dernière estimation s'obtient de manière similaire, et pour $p = \infty$, on obtient le résultat par majoration directe.

□

Régularité $W^{2,\infty}$. On suppose maintenant que la donnée initiale est dans $W^{2,\infty}(\Omega)$; comme dans le cas $W^{1,\infty}$, cette régularité est préservée par f.
Rappelons (cf [24] par exemple) que l'espace $W^{2,\infty}$ peut-être caractérisé par :

Lemme 5.7.7. *Une fonction g appartient à $W^{2,\infty}(\Omega)$, si et seulement si $g \in C^0(\Omega)$ vérifie :*

$$|g(x+2h_x, v+2h_v) - 2g(x+h_x, v+h_v) + g(x,v)| \leq K(|h_x|+|h_v|)^2, \quad (5.7.52)$$

pour tous les réels x, v, h_x, h_v tels que l'expression soit bien définie, la constante K étant indépendante de ces réels.

On peut alors adapter la démonstration précédente du cas $W^{1,\infty}$, pour obtenir :

Lemme 5.7.8. *Si $f_0 \in W^{2,\infty}(\Omega)$, alors $f(t) \in W^{2,\infty}(\Omega)$, pour tout $t \in [0,T]$, et*

$$|f(t)|_{W^{2,\infty}(\Omega)} \leq C_T. \quad (5.7.53)$$

Preuve. On utilise à nouveau la forme caractéristique de f, et on va montrer (5.7.52) pour $f(t)$.
Au vu de (5.7.52), on ne restreint pas la généralité en supposant que $x \in [0, L]$.
Pour simplifier, on notera $z = (x, v)$, $h = (h_x, h_v)$ et

$$Z(t;s,z) := \big(X(t;s,x,v), V(t;s,x,v)\big), \quad (5.7.54)$$

et pour un élément de $z = (x,v) \in \mathbb{R}^2$, on note $|z| = |x| + |v|$.
On peut aussi supposer que $|h| \leq 1$.
On a

$$\begin{aligned}\Delta_h^2 f(t)(x,v) &:= f(t, z+2h) - 2f(t, z+h) + f(t,z) \\ &= f_0(Z(-t;0,z+2h) - 2f_0(Z(-t;0,x+h)) + f_0(Z(-t;0,x)). \quad (5.7.55)\end{aligned}$$

On pose ensuite

$$Z := Z(-t;0,z), \quad H := Z(-t;0,z+h) - Z(0;Z(-t;0,z)), \tag{5.7.56}$$

de telle sorte que

$$\Delta_h^2 f(t)(z) \;=\; f_0(Z(-t;0,z+2h)) - 2f_0(Z+H) + f_0(Z) \;=\; F_1 + F_2, \tag{5.7.57}$$

avec

$$F_1 := f_0(Z(-t;0,z+2h)) - f_0(Z+2H), \tag{5.7.58}$$

et

$$F_2 := f_0(Z+2H) - 2f_0(Z+H) + f_0(Z). \tag{5.7.59}$$

Comme $f_0 \in W^{2,\infty}(\Omega)$, d'après la précédente caractérisation, on obtient déjà que

$$\big|F_2\big| \leq |H|^2. \tag{5.7.60}$$

Or, d'après l'inégalité (5.7.30) du lemme 5.7.5, qui s'applique ici puisque $f_0 \in W^{2,\infty}(\Omega) \subset W^{1,\infty}(\Omega)$, on obtient

$$|H| \leq C|h|, \tag{5.7.61}$$

et ceci donne donc

$$|F_2| \leq C|h|^2. \tag{5.7.62}$$

Il reste donc à étudier l'autre terme.
A nouveau, puisque $f_0 \in W^{1,\infty}(\Omega)$ est lipschitzienne, on a

$$|F_1| \leq C|Z(-t;0,z+2h) - Z - 2H|. \tag{5.7.63}$$

On calcule alors

$$\begin{aligned}\Delta_Z(-t) &:= Z(-t;0,z+2h) - Z - 2H \\ &= Z(-t;0,z+2h) - 2Z(-t;0,z+h) + Z(-t;0,z).\end{aligned} \tag{5.7.64}$$

Puis, on continue avec

$$\begin{aligned}\Delta_X(-t) &:= X(-t;0,z+2h) - 2X(-t;0,z+h) + X(-t;0,z) \\ &= \int_0^{-t} \dot{X}(s;0,z+2h) - 2\dot{X}(s;0,z+h) + \dot{X}(s;0,z)\mathrm{d}s \\ &= \int_0^{-t} \Delta_V(s)\mathrm{d}s,\end{aligned} \tag{5.7.65}$$

avec

$$\Delta_V(s) := V(s;0,z+2h) - 2V(s;0,z+h) + V(s;0,z). \tag{5.7.66}$$

On calcule donc ensuite

$$\begin{aligned}\Delta_V(-t) &:= \int_0^{-t} \dot{V}(s;0,z+2h) - 2\dot{V}(s;0,z+h) + \dot{V}(s;0,z)\mathrm{d}s \\ &= \int_0^{-} tE(s,x+2h_x)X(s;0,z+2h) - 2E(s,x+h_x)X(s;0,z+h) \\ &\qquad + E(s,x)X(s;0,z)\mathrm{d}s.\end{aligned} \tag{5.7.67}$$

En utilisant la notation $E_i := E(s, x + ih_x)$ et $X_i := X(s; 0, z + ih)$, pour $i = 0, 1, 2$. On voit que le terme sous l'intégrale se décompose de la manière suivante :

$$\begin{aligned} E_2X_2 - 2E_1X_1 + E_0X_0 &= (E_2 - E_1)X_2 + (E_0 - E_1)X_0 + E_1(X_2 - 2X_1 + X_0) \\ &= (E_2 - E_1)X_2 + (E_0 - E_1)X_0 + E_1\Delta_X(s), \end{aligned} \quad (5.7.68)$$

et

$$\begin{aligned} (E_2 - E_1)X_2 + (E_0 - E_1)X_0 &= (E_2 - E_1)(X_2 - X_1) \\ &\quad + (E_0 - E_1)(X_0 - X_1) + X_1(E_2 - 2E_1 + E_0). \end{aligned} \quad (5.7.69)$$

Or on a

$$|X_1(E_2 - 2E_1 + E_0)| \leq C_T|X_1|h_x^2, \quad (5.7.70)$$

puisque l'on sait que

$$\big|E(s)\big|_{W^{2,\infty}([0,L])} \leq C_T, \quad (5.7.71)$$

pour tout $s \in [0, T]$, d'après le lemme 5.7.6. On a aussi

$$|X_1| \leq |X_1 - X_0| + L, \quad (5.7.72)$$

comme $|X_0| = |x| \leq L$.
D'autre part, on a

$$|X_0 - X_1| \leq C_T|h| \quad |X_2 - X_1| \leq C_T|h|, \quad (5.7.73)$$

à nouveau d'après (5.7.30), et il en découle que

$$|X_1| \leq L + C_T|h| \leq C_T, \quad (5.7.74)$$

puisque l'on avait supposé $|h| \leq 1$.
En reprenant (5.7.70), on en déduit bien que

$$|X_1(E_2 - 2E_1 + E_0)| \leq C_T|h|^2. \quad (5.7.75)$$

On sait aussi que

$$\big|E(s)\big|_{W^{1,\infty}([0,L])} \leq C_T, \quad (5.7.76)$$

pour tout $s \in [0, T]$, d'après le lemme 5.7.2. Ainsi, on a également

$$|E_2 - E_1| \leq C_T|h| \quad |E_1 - E_0| \leq C_T|h|. \quad (5.7.77)$$

Finalement, en reprenant (5.7.69), à l'aide de (5.7.75), (5.7.73) et de la précédente inégalité, on est assuré de l'estimation

$$\big|(E_2 - E_1)X_2 + (E_0 - E_1)X_0\big| \leq C_T|h|^2, \quad (5.7.78)$$

et puis, comme on a toujours $|E(t)|_{L^\infty([0,L])} \leq C_T$ d'après (5.7.2), on en déduit de (5.7.68) que

$$\big|E_2X_2 - 2E_1X_1 + E_0X_0\big| \leq C_T. \quad (5.7.79)$$

On peut alors appliquer l'inégalité de Gronwall de la proposition 5.7.4,

$$|\Delta_X(-t)| + |\Delta_V(-t)| \leq C_T h^2 \left(\int_0^t |\Delta_X(-s)| + |\Delta_V(-s)| \mathrm{d}s \right), \tag{5.7.80}$$

pour obtenir

$$|\Delta_Z(-t)| \leq C_T |h|^2, \tag{5.7.81}$$

et donc $|F_1| \leq C_T |h|^2$. Finalement, on a (5.7.53). □

Remarque 5.7.9. *(i) On pourrait continuer l'étude pour des ordres supérieurs.*

(ii) Avec les injections de Sobolev, on en déduit que si f_0 est de classe C^2, alors $f(t)$ est de classe C^1 ; en fait $f(t)$ reste de classe C^2 (voir le théorème 6.2.1 et ses références).

Estimations $W^{2,1}$. La technique utilisée pour les estimations $W^{1,p}$ s'applique également pour des estimations d'ordre plus élevés.
Ainsi, en dérivant l'équation de Vlasov k fois par rapport à x et ℓ fois par rapport à v :

$$\begin{aligned} &\partial_t(\partial_x^{(k)}\partial_v^\ell f) + v\partial_x(\partial_x^{(k)}\partial_v^\ell f)) + E\partial_v(\partial_x^{(k)}\partial_v^\ell f)) \\ &\qquad = -\sum_{j=1}^{k} \binom{k}{j} \partial_x^{(j)} E \partial_x^{(k-j)} \partial_v^{(\ell+1)} f - \partial_x^{(k+1)} \partial_v^{\ell-1} f, \end{aligned} \tag{5.7.82}$$

cela nous permet d'obtenir

$$\begin{aligned} \partial_x^{(k)}\partial_v^\ell f(t, X(t), V(t)) &= \partial_x^{(k)}\partial_v^{(\ell)} f(0,x,v) - \int_0^t \partial_x^{(k+1)}\partial_v^{(\ell-1)} f(s, X(s), V(s)) \mathrm{d}s \\ &\quad - \sum_{j=1}^{k} \binom{k}{j} \int_0^t \partial_x^{(j)} E(s, X(s)) \partial_x^{(k-j)} \partial_v^{(\ell+1)} f(s, X(s), V(s)) \mathrm{d}s, \end{aligned} \tag{5.7.83}$$

Dans le cas particulier où l'on cherche une estimation $W^{2,1}$, l'étude est simplifiée, et l'on obtient les résultats suivants :

Lemme 5.7.10. *On suppose que $f_0 \in W^{2,1}(\Omega)$, alors,*

$$|f(t)|_{W^{2,1}(\Omega)} \leq |f_0|_{W^{2,1}(\Omega)} e^{\alpha_T T}, \tag{5.7.84}$$

avec

$$\alpha_T := 8(Q(0) + 2LT)\|f_0\|_{L^\infty(\Omega)} + 1. \tag{5.7.85}$$

Preuve. On sait que $f(t) \in W^{2,1}(\Omega)$, d'après par exemple [64]. On a précisément :

$$\begin{aligned} \partial_x^2 f(t, X(t), V(t)) = \partial_x^2 f_0(x,v) - \int_0^t &2\partial_x E(s, X(s)) \partial_x \partial_v f(s, X(s), V(s)) \\ &- \partial_x^2 E(s, X(s)) \partial_v f(s, X(s), V(s)) \mathrm{d}s, \end{aligned} \tag{5.7.86}$$

$$\partial_x\partial_v f(t,X(t),V(t)) = \partial_x\partial_v f_0(x,v) - \int_0^t \partial_x^2 f(s,X(s),V(s)) \\ - \partial_x E(s,X(s))\partial_v^2 f(s,X(s),V(s))\mathrm{d}s, \quad (5.7.87)$$

et

$$\partial_v^2 f(t,X(t),V(t)) = \partial_v^2 f_0(x,v) - \int_0^t \partial_x\partial_v f(s,X(s),V(s))\mathrm{d}s. \quad (5.7.88)$$

Or, en utilisant que le changement de variable $(x,v) \to (X(s),V(s))$ est de déterminant 1 (il s'agit de la préservation de la mesure), on obtient

$$\int_\Omega \int_0^t \partial_x^2 E(s,X(s))\partial_v f(s,X(s),V(s))\mathrm{d}s\mathrm{d}x\mathrm{d}v \\ = \int_0^t \int_\Omega \partial_x^2 E(s,x)\partial_v f(s,x,v)\mathrm{d}x\mathrm{d}v\mathrm{d}s \\ = -\int_0^t \int_\Omega \partial_x E(s,x)\partial_x\partial_v f(s,x,v)\mathrm{d}x\mathrm{d}v\mathrm{d}s, \quad (5.7.89)$$

la dernière relation provenant d'une intégration par parties, en utilisant le fait que

$$\partial_x E(s,L)\partial_v f(s,L,v) = \partial_x E(s,0)\partial_v f(s,0,v),$$

puisque le champ E et la fonction de distribution f sont L–périodiques en x. Ainsi, grâce à cette relation, et en intégrant (5.7.86),(5.7.87) et (5.7.88) par rapport à x et v, on obtient, en utilisant à nouveau la préservation de la mesure,

$$\|\partial_x^2 f(t)\|_{L^1(\Omega)} \le \|\partial_x^2 f_0\|_{L^1(\Omega)} + 4\int_0^t |E(s)|_{W^{1,\infty}} \|\partial_x\partial_v f(s)\|_{L^1(\Omega)}\mathrm{d}s, \quad (5.7.90)$$

et également

$$\|\partial_x\partial_v f(t)\|_{L^1(\Omega)} \le \|\partial_x\partial_v f_0\|_{L^1(\Omega)} + \int_0^t \|\partial_x^2 f(s)\|_{L^1(\Omega)} \\ + |E(s)|_{W^{1,\infty}([0,L])} \|\partial_v^2 f(s)\|_{L^1(\Omega)}\mathrm{d}s, \quad (5.7.91)$$

et

$$\|\partial_v^2 f(t)\|_{L^1(\Omega)} \le \|\partial_v^2 f_0\|_{L^1(\Omega)} + \int_0^t \|\partial_x\partial_v f(s)\|_{L^1(\Omega)}\mathrm{d}s. \quad (5.7.92)$$

En sommant les relations (5.7.90),(5.7.91) et (5.7.92), on obtient

$$|f(t)|_{W^{2,1}(\Omega)} \le |f_0|_{W^{2,1}(\Omega)} + \int_0^t (4|E(s)|_{W^{1,\infty}([0,L])} + 1)|f(s)|_{W^{2,1}(\Omega)}. \quad (5.7.93)$$

En utilisant la majoration $W^{1,\infty}$ du champ électrique (5.7.7) et l'inégalité de Gronwall de la proposition (5.7.4), on en déduit le résultat. □

Remarque 5.7.11. *En utilisant l'équation de Vlasov, on en déduit également des majorations du même type sur* $|\partial_t f(t)|_{W^{1,1}(\Omega)}$ *et* $\|\partial_t^2 f(t)\|_{L^1(\Omega)}$ *et on peut aussi borner* $|E|_{W^{3,1}([0,L])}$, $|\partial_t E|_{W^{2,1}([0,L])}$, $|\partial_t^2 E|_{W^{1,1}([0,L])}$ *et* $\|\partial_t^3 E\|_{L^1([0,L])}$.

Synthèse. Le lemme suivant résume les estimations utilisées ou mentionnées pour le chapitre 7

Lemme 5.7.12. *On a les estimations :*

$$\begin{array}{lcl} Q(T) & \leq & C_T \\ \|E\|_{L^\infty(0,T;W^{1,\infty}(\Omega))} & \leq & C_T \end{array}$$

Si $f_0 \in W^{1,\infty}(\Omega)$, on a de plus :

$$\begin{array}{lcl} \|f\|_{L^\infty(0,T;W^{1,\infty}(\Omega))} & \leq & C_T \\ \|E\|_{L^\infty(0,T;W^{2,\infty}([0,L]))} & \leq & C_T \\ \|\partial_t E\|_{L^\infty(0,T;W^{1,\infty}([0,L]))} & \leq & C_T \\ \|\partial^2_{tt} E\|_{L^\infty(0,T;L^\infty([0,L]))} & \leq & C_T \end{array}$$

Si $f_0 \in W^{2,\infty}(\Omega)$, on obtient :

$$\|f\|_{L^\infty(0,T;W^{2,1}(\Omega))} \leq C_T,$$

cette dernière constante ne dépendant que de la norme $W^{2,1}(\Omega)$ et de la norme $L^\infty(\Omega)$ (et non de la norme $W^{2,\infty}(\Omega)$).

5.8 Un schéma général de splitting en temps

On donne ici une forme générale d'un schéma de splitting en temps, on étudie l'erreur induite de discrétisation en temps, et on indique comment on décomposera l'erreur de discrétisation en espace pour les schémas plus spécifiques qui seront étudiés aux chapitres 6 et 7.
Considérons un système qui s'écrit sous la forme :

$$u_t = Au. \tag{5.8.1}$$

L'idée du splitting est de décomposer A en $A = A_1 + A_2$, et de résoudre successivement

$$u_t = A_1 u, \tag{5.8.2}$$

puis

$$u_t = A_2 u. \tag{5.8.3}$$

Avec un splitting d'ordre 1, on obtient

$$u(t + \Delta t) = e^{A_2 \Delta t} e^{A_1 \Delta t} u(t). \tag{5.8.4}$$

Pour un splitting de Strang d'ordre 2, on résoud (5.8.2) sur un demi-pas de temps, puis (5.8.3) sur un pas de temps, et enfin (5.8.2) sur un demi-pas de temps :

$$u(t + \Delta t) = e^{A_1 \Delta t/2} e^{A_2 \Delta t} e^{A_1 \Delta t/2} u(t). \tag{5.8.5}$$

Notons que la solution exacte est donnée par

$$u(t+\Delta t) = e^{(A_1+A_2)\Delta t}u(t). \tag{5.8.6}$$

En approchant (5.8.6) par (5.8.4), on commet une erreur d'ordre 1 :

$$\begin{aligned}(\mathrm{Id} + \mathrm{A}_1\Delta\mathrm{t} + \mathrm{A}_1^2\Delta\mathrm{t}^2/2 + \mathrm{O}(\Delta\mathrm{t}^3))(\mathrm{Id} + \mathrm{A}_2\Delta\mathrm{t} + \mathrm{A}_2^2\Delta\mathrm{t}^2/2 + \mathrm{O}(\Delta\mathrm{t}^3)) \\ = \mathrm{Id} + (\mathrm{A}_1 + \mathrm{A}_2)\Delta\mathrm{t} + (\mathrm{A}_1^2 + 2\mathrm{A}_1\mathrm{A}_2 + \mathrm{A}_2^2)\Delta\mathrm{t}^2/2 + \mathrm{O}(\Delta\mathrm{t}^3).\end{aligned} \tag{5.8.7}$$

On remarque d'ailleurs que l'approximation est d'ordre supérieur à deux si et seulement si A_1 et A_2 commutent ; dans ce cas il n'y a en fait plus d'approximation : (5.8.4) coincide avec (5.8.6).
Pour le splitting d'ordre 2, en approchant cette fois-ci (5.8.6) par (5.8.5), on commet une erreur d'ordre 2 :

$$\begin{aligned}&(\mathrm{Id} + A_1\Delta t/2 + A_1^2\Delta t^2/8 + O(\Delta t^3))(\mathrm{Id} + A_2\Delta t + A_2^2\Delta t^2/2 + O(\Delta t^3)) \\ &(\mathrm{Id}+A_1\Delta t/2+A_1^2\Delta t^2/8+O(\Delta t^3)) = \mathrm{Id}+(A_1+A_2)\Delta t+(A_1^2+A_1A_2+A_2A_1+A_2^2)\Delta t^2/2+O(\Delta t^3)\end{aligned} \tag{5.8.8}$$

Un avantage du *splitting* en temps est que les opérateurs de transport vont consister ici en une succession d'advections à coefficients constants. Afin de simplifier les notations, on va traiter ici seulement le cas d'un *splitting* d'ordre deux ; pour une généralisation au cas de degrés plus élevés (ce qui peut être important, si la discrétisation spatiale est aussi d'ordre élevée, cf [33]), voir la remarque 6.4.9.

Opérateurs de transport. On introduit les deux opérateurs de transports. Soit $\mathcal{T}_x$ défini par

$$\mathcal{T}_x g(x,v) = g(x - v\Delta t/2, v)$$

pour chaque fonction de distribution g. Il s'agit ici d'une advection constante le long de la direction x.
L'advection correspondante le long de la direction v est plus complexe, car elle utilise le champ électrique qui dépend non linéairement de la fonction de distribution. Ainsi, on définit $\mathcal{T}_v^{\tilde{g}}$ par :

$$\mathcal{T}_v^{\tilde{g}} g(x,v) = g(x, v - \Delta t E_{\tilde{g}}(x)),$$

où le champ électrique $E_{\tilde{g}}$ est donné par :

$$E_{\tilde{g}}(x) = \int_0^L K(x,y)\Big(\int_{\mathbb{R}} \tilde{g}(y,v)\mathrm{d}v - 1\Big)\mathrm{d}y.$$

On a ici remplacé la fonction de distribution $f(t)$ dans (5.2.1) par une fonction arbitraire $\tilde{g}$; cela nous sera utile plus tard pour l'analyse. Cependant, pour simplifier, on écrira :

$$\mathcal{T}_v g := \mathcal{T}_v^g g. \tag{5.8.9}$$

Remarque 5.8.1. *(i) Les lettres x et v dans $\mathcal{T}_x$ et $\mathcal{T}_v$ ne seront jamais considérées comme des variables ; ils font juste référence à l'advection en x ou en v.*

(ii) Les opérateurs $\mathcal{T}_x$ et $\mathcal{T}_v^g$ sont linéaires, par contre l'opérateur $\mathcal{T}_v$ défini par (5.8.9) ne l'est pas.

Erreur de discrétisation en temps. En partant de $f(t^n)$ solution exacte de l'équation de Vlasov (5.4.3) au temps t^n, l'approximation de $f(t^{n+1})$ inhérente au time-splitting est donnée par :

$$\mathcal{S}f(t^n) := \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n).$$

Lemme 5.8.2. *Supposons que $f_0 \in W^{1,\infty}(\Omega)$, alors*

$$\|f(t^{n+1}) - \mathcal{S}f(t^n))\|_{L^\infty(\Omega)} \leq C_T\Delta t^3.$$

Preuve. Soit (x,v) fixé dans $\mathbb{R}^2$. Comme f est constante le long des caractéristiques, on a

$$f(t^{n+1},x,v) = f(t^n, X(t^n), V(t^n)) \quad \text{avec} \quad (X,V)(s) = (X,V)(s;t^{n+1},x,v),$$

D'autre part, on calcule :

$$\begin{aligned}\mathbb{S}f(t^n)(x,v) = \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)(x,v) &= \mathcal{T}_x\mathcal{T}_v f(t^n, x - v\Delta t/2, v)\\ &= \mathcal{T}_x f(t^n, x - v\Delta t/2, v - \Delta t E_{\mathcal{T}_x f(t^n)}(x - v\Delta t/2)) = f(t^n, X^n, V^n),\end{aligned}$$

avec

$$\begin{cases} X^n := x - v\Delta t + \frac{\Delta t^2}{2} E_{\mathcal{T}_x f(t^n)}(x - v\Delta t/2) \\ V^n := v - \Delta t E_{\mathcal{T}_x f(t^n)}(x - v\Delta t/2). \end{cases}$$

D'après le lemme 5.7.12 , on sait que f est dans $L^\infty(0,T;W^{1,\infty})$, donc, on a :

$$\|f(t^{n+1}) - \mathcal{S}f(t^n))\|_{L^\infty} \leq \max_n |f(t^n,\cdot,\cdot)|_{W^{1,\infty}}(|X(t^n) - X^n| + |V(t^n) - V^n|), \quad (5.8.10)$$

avec

$$\max_n |f(t^n,\cdot,\cdot)|_{W^{1,\infty}} \leq \|f\|_{L^\infty([0,T];W^{1,\infty}(0,L))} \leq C_T. \quad (5.8.11)$$

Il suffit donc de montrer que

$$\max(|X^n - X(t^n)|, |V^n - V(t^n)|) \leq C_T\Delta t^3. \quad (5.8.12)$$

On va d'abord montrer quelques inégalités concernant le champ

$$E_X(t) := E(t, X(t)).$$

On a

$$\dot{E}_X(t) = \partial_t E(t,X(t)) + \partial_x E(t,X(t))V(t), \quad (5.8.13)$$

$$\ddot{E}_X(t) = \partial^2_{tt}E(t,X(t)) + \partial^2_{xx}E(t,X(t))V(t)^2 + \partial^2_{tx}E(t,X(t))V(t) + \partial_x E(t,X(t))E_X(t) \quad (5.8.14)$$

et donc

$$|\ddot{E}_X(t)| \leq |\partial^2_{tt}E(t,X(t))| + |\partial^2_{xx}E(t,X)|V^2 + |\partial^2_{tx}E(t,X)V| + |\partial_x E(t,X)E_X(t)| \quad (5.8.15)$$

et on sait d'après le lemme 5.7.12 que toutes ces quantités sont bornées sur $(0,T)$; on est donc assuré que :

$$|E_X(t)| + |\dot{E}_X(t)| + |\ddot{E}_X(t)| \leq C_T, \tag{5.8.16}$$

pour tout $t \in [0,T]$, ce qui exprime que E_X est dans $W^{2,\infty}([0,T])$.
On décompose maintenant :

$$\begin{aligned} X^n - X(t^n) &= X(t^{n+1}) - X(t^n) - v\Delta t + \tfrac{\Delta t^2}{2} E_{\mathcal{T}_x f(t^n)}(x - v\Delta t/2) \\ &= E_1 + \tfrac{\Delta t^2}{2}(E_2 + E_3), \end{aligned}$$

avec

$$\begin{aligned} E_1 &:= X(t^{n+1}) - X(t^n) - v\Delta t + \tfrac{\Delta t^2}{2} E_X(t^{n+1/2}) \\ E_2 &:= E(t^{n+1/2}, x - v\Delta t/2) - E_X(t^{n+1/2}) \\ E_3 &:= E_{\mathcal{T}_x f(t^n)}(x - v\Delta t/2) - E(t^{n+1/2}, x - v\Delta t/2) \end{aligned}$$

On décompose aussi :

$$V^n - V(t^n) = \quad V(t^{n+1}) - V(t^n) - \Delta t E_{\mathcal{T}_x f(t^n)}(x - v\Delta t/2) = \quad F_1 + \Delta t(E_2 + E_3).$$

avec

$$F_1 := V(t^{n+1}) - V(t^n) - \Delta t E_X(t^{n+1/2}). \tag{5.8.17}$$

Il suffit donc de montrer que :

$$|E_1| \leq C_T \Delta t^3, \quad |E_2| \leq C_T \Delta t^2, \quad |E_3| \leq C_T \Delta t^2 \quad \text{et} \quad |F_1| \leq C_T \Delta t^3. \tag{5.8.18}$$

Pour le premier terme, on a :

$$\begin{aligned} E_1 &= \int_{t^n}^{t^{n+1}} (V(t) - v)\mathrm{d}t + \tfrac{\Delta t^2}{2} E_X(t^{n+1/2}) \\ &= \int_{t^n}^{t^{n+1}} (V(t) - V(t^{n+1}))\mathrm{d}t - \int_{t^n}^{t^{n+1}} \int_t^{t^{n+1}} E_X(t^{n+1/2})\, ds\, dt \\ &= \int_{t^n}^{t^{n+1}} \int_t^{t^{n+1}} \big(-E_X(s) + E_X(t^{n+1/2})\big)\, ds\, dt. \end{aligned}$$

Or, grâce à (5.8.16), on a

$$|E_X(t^{n+1/2}) - E_X(s)| \leq |\dot{E}_X|_{L^\infty([0,T])} |t^{n+1/2} - s| \leq C_T \Delta t,$$

ce qui assure

$$|E_1| \leq C_T \Delta t^3. \tag{5.8.19}$$

Pour le second terme, on écrit :

$$\begin{aligned} |E_2| &= \big|E(t^{n+1/2}, x - v\Delta t/2) - E(t^{n+1/2}, X(t^{n+1/2}))\big| \\ &\leq |\partial_x E(t^{n+1/2})|_{L^\infty([0,L])} |X(t^{n+1/2}) - x + v\Delta t/2| \\ &\leq C_T |X(t^{n+1/2}) - x + v\Delta t/2|. \end{aligned}$$

Or, on a

$$\begin{aligned} X(t^{n+1/2}) - x + v\Delta t/2 &= X(t^{n+1/2}) - X(t^{n+1}) + v\Delta t/2 \\ &= \textstyle\int_{t^{n+1/2}}^{t^{n+1}} (v - V(s))\mathrm{d}t \\ &= \textstyle\int_{t^{n+1/2}}^{t^{n+1}} (V(t^{n+1}) - V(s))\mathrm{d}t \\ &= \textstyle\int_{t^{n+1/2}}^{t^{n+1}} \int_t^{t^{n+1}} E_X(s)\mathrm{d}s\mathrm{d}t, \end{aligned}$$

et donc

$$|X(t^{n+1/2}) - x + v\Delta t/2| \leq \|E_X\|_{L^\infty([0,T])} \Delta t^2 \leq C_T \Delta t^2, \tag{5.8.20}$$

avec (5.8.16), ce qui permet d'établir :

$$|E_2| \leq C_T \Delta t^2. \tag{5.8.21}$$

On s'intéresse maintenant au troisième terme

$$E_3 = \Big| \int_0^L K(x - v\Delta t/2, y) \int_{\mathbb{R}} [f(t^n, y - v\Delta t/2, v) - f(t^{n+1/2}, y, v)]\, dv\, dy \Big|, \tag{5.8.22}$$

Comme $\|K\|_{L^\infty([0,L])} \leq 1$, on en déduit que

$$|E_3| \leq \int_0^L \Big| \int_{\mathbb{R}} f(t^n, y - v\Delta t/2, v) - f(t^{n+1/2}, y, v)\, dv \Big|\, dy. \tag{5.8.23}$$

Or

$$\begin{aligned} f(t^n, y - v\Delta t/2, v) - f(t^{n+1/2}, y, v) &= \int_0^{\Delta t/2} \frac{\mathrm{d}}{\mathrm{d}s} f(t^n + \Delta t/2 - s, y - vs, v)\, \mathrm{d}s \\ &= \int_0^{\Delta t/2} (-\partial_t f - v\partial_x f)(t^n + \Delta t/2 - s, y - vs, v)\mathrm{d}s \\ &= \int_0^{\Delta t/2} (-E(t^n + \Delta t/2 - s, y - vs)\partial_v f(t^n + \Delta t/2 - s, y - vs, v)\mathrm{d}s. \end{aligned} \tag{5.8.24}$$

Une majoration directe donnerait $|E_3| \leq C_T \Delta t$.
Afin d'obtenir $|E_3| \leq C_T \Delta t^2$, on utilise une intégration par parties :

$$\begin{aligned} &\int_{\mathbb{R}} - s\partial_x E(t^n + \Delta t/2 - s, y - vs) f(t^n + \Delta t/2 - s, y - vs, v)]\mathrm{d}v \\ &\quad = \int_{\mathbb{R}} E(t^n + \Delta t/2 - s, y - vs) \frac{\mathrm{d}}{\mathrm{d}v} [f(t^n + \Delta t/2 - s, y - vs, v)]\mathrm{d}v \\ &\quad = \int_{\mathbb{R}} E(t^n + \Delta t/2 - s, y - vs)\ (-s\partial_x f + \partial_v f)(t^n + \Delta t/2 - s, y - vs, v)]\mathrm{d}v, \end{aligned} \tag{5.8.25}$$

et donc

$$\begin{aligned} &\int_{\mathbb{R}} -E(t^n + \Delta t/2 - s, y - vs)\partial_v f(t^n + \Delta t/2 - s, y - vs, v)\mathrm{d}v \\ &\qquad = \int_{\mathbb{R}} [-E(t^n + \Delta t/2 - s, y - vs)\ s\partial_x f(t^n + \Delta t/2 - s, y - vs, v) \\ &\qquad\qquad + s\partial_x E(t^n + \Delta t/2 - s, y - vs) f(t^n + \Delta t/2 - s, y - vs, v)]\mathrm{d}v. \end{aligned} \tag{5.8.26}$$

En intégrant (5.8.24) par rapport à v et en utilisant le théorème de Fubini-Tonelli, on obtient alors :

$$\begin{aligned}\int_{\mathbb{R}}[f(t^n, y - v\Delta t/2, v) - f(t^{n+1/2}, y, v)]\mathrm{d}v \\ = \int_0^{\Delta t/2}\int_{\mathbb{R}}[-E(t^n + \Delta t/2 - s, y - vs)\; s\partial_x f(t^n + \Delta t/2 - s, y - vs, v) \\ + s\partial_x E(t^n + \Delta t/2 - s, y - vs) f(t^n + \Delta t/2 - s, y - vs, v)]\mathrm{d}v\mathrm{d}s. \quad (5.8.27)\end{aligned}$$

En reprenant (5.8.23), on en déduit :

$$\begin{aligned}|E_3| \le \|E\|_{L^\infty([0,T];L^\infty([0,L]))}\int_0^L\int_0^{\Delta t/2}\int_{\mathbb{R}} s\partial_x f(t^{n+1/2} - s, y - vs, v)\mathrm{d}v\mathrm{d}s\mathrm{d}y \\ + |E|_{L^\infty([0,T];W^{1,\infty}([0,L]))}\int_0^L\int_0^{\Delta t/2}\int_{\mathbb{R}} sf(t^{n+1/2} - s, y - vs, v)\mathrm{d}v\mathrm{d}s\mathrm{d}y, \quad (5.8.28)\end{aligned}$$

et donc

$$\begin{aligned}|E_3| \le \big(\|E\|_{L^\infty([0,T];L^\infty([0,L]))}\|f\|_{L^\infty([0,T],W^{1,\infty}(\Omega))} \\ + |E|_{L^\infty([0,T];W^{1,\infty}([0,L]))}\|f\|_{L^\infty([0,T],L^\infty(\Omega))}\big)L2Q(T)\Delta t^2 \le C_T\Delta t^2. \quad (5.8.29)\end{aligned}$$

Pour le dernier terme F_1, on écrit :

$$\begin{aligned}F_1 =& V(t^{n+1}) - V(t^{n+1/2}) + V(t^{n+1/2}) - V(t^n) - \Delta t E_X(t^{n+1/2}) \\ & \int_0^{\Delta t/2} E_X(t^{n+1} - s) + E_X(t^n + s)\, ds - \Delta t E_X(t^{n+1/2}) \\ & = \int_0^{\Delta t/2} E_X(t^{n+1} - s) - E_X(t^{n+1/2}) + E_X(t^n + s) - E_X(t^{n+1/2})\, ds \\ & = \int_0^{\Delta t/2}\int_s^{\Delta t/2} (\dot{E}_X(t^{n+1} - u) - \dot{E}_X(t^n + u))\, du\, ds,\end{aligned}$$

et donc

$$|F_1| \le \int_0^{\Delta t/2}\int_s^{\Delta t/2} 2\|\ddot{E}_X\|_{L^\infty([0,T])} \le C_T\Delta t^3. \quad (5.8.30)$$

□

Discrétisation spatiale et schéma numérique. Après chaque advection, on va faire une projection sur la fonction advectée ; cette projection peut dépendre du temps (projection adaptative). Ainsi, on écrira Π_0 pour la première projection de la donnée initiale, Π_1^{n+1}, Π_2^{n+1} et Π_3^{n+1} pour les projections suivantes ; le schéma numérique général de splitting en temps s'écrit alors :

$$f^{n+1} = \Pi_3^{n+1}\mathcal{T}_x\Pi_2^{n+1}\mathcal{T}_v\Pi_1^{n+1}\mathcal{T}_x f^n,$$

et

$$f^0 = \Pi_0 f_0.$$

Pour simplifier, on écrira :

$$\tilde{\mathcal{T}}_x^{n+1} := \Pi_3^{n+1}\mathcal{T}_x,$$
$$\tilde{\mathcal{T}}_x^{n+1/2} := \Pi_1^{n+1}\mathcal{T}_x.$$

On définit l'advection en v avec un champ exact :

$$\mathcal{T}_{v,n+1}^{E} g(x,v) = g(x, v - \Delta t E_{\mathcal{T}_x f(t^n)}), \tag{5.8.31}$$

l'advection en v avec un champ approché :

$$\mathcal{T}_{v,n+1}^{E_h} g(x,v) = g(x, v - \Delta t E_h^n), \tag{5.8.32}$$

où le champ numérique E_h^n au temps t^n est une approximation du champ E^n défini par :

$$E^n(x) = \int K(x,y)(\int \tilde{\mathcal{T}}_x^{n+1/2} f^n(y,v)\mathrm{d}v - 1)\mathrm{d}y. \tag{5.8.33}$$

Le paramètre h est un paramètre de discrétisation de l'intervalle $[0, L]$, pour le calcul du champ approché E_h^n. On utilisera alors aussi les notations :

$$\tilde{\mathcal{T}}_{v,n+1}^{E} = \Pi_2^{n+1}\mathcal{T}_{v,n+1}^{E},$$
$$\tilde{\mathcal{T}}_{v,n+1}^{E_h} = \Pi_2^{n+1}\mathcal{T}_{v,n+1}^{E_h},$$

Le schéma s'écrit alors :

$$f^{n+1} = \tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n,$$

La question que l'on se pose alors est la suivante : quel type de projections doit-on prendre afin d'avoir une approximation précise de la fonction de distribution f ?
Dans les deux chapitres suivants, on va prouver la convergence dans des cas particuliers. Dans le cas d'un algorithme uniforme, on va bénéficier de la régularité de la solution exacte f.
Pour décomposer l'erreur, remarquons d'abord que l'on a les relations :

$$\tilde{\mathcal{T}}_x^{n+1}\mathcal{T}_v\mathcal{T}_x f(t^n) = \Pi_3^{n+1}\mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)$$
$$\tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E}\mathcal{T}_x f(t^n) = \tilde{\mathcal{T}}_x^{n+1}\Pi_2^{n+1}\mathcal{T}_{v,n+1}^{E}\mathcal{T}_x f(t^n) = \tilde{\mathcal{T}}_x^{n+1}\Pi_2^{n+1}\mathcal{T}_v\mathcal{T}_x f(t^n)$$
$$\tilde{\mathcal{T}}_x^{n+1}\Pi_2^{n+1}\mathcal{T}_{v,n+1}^{E}\tilde{\mathcal{T}}_x^{n+1/2} f(t^n) = \tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E}\tilde{\mathcal{T}}_x^{n+1/2} f(t^n) = \tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E}\Pi_1^{n+1}\tilde{\mathcal{T}}_x f(t^n)$$
$$\tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f(t^n) = \tilde{\mathcal{T}}_x^{n+1}\Pi_2^{n+1}\mathcal{T}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f(t^n).$$

L'erreur

$$e^n := f(t^n) - f^n$$

se décompose alors sous la forme :

$$e^{n+1} = e_1^{n+1} + e_2^{n+1} + e_3^{n+1} + e_4^{n+1} + e_5^{n+1} + e_6^{n+1} \tag{5.8.34}$$

avec

$$\begin{aligned}
e_1^{n+1} &= f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\\
e_2^{n+1} &= (I - \Pi_3^{n+1})\mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\\
e_3^{n+1} &= \tilde{\mathcal{T}}_x^{n+1}(I - \Pi_2^{n+1})\mathcal{T}_v\mathcal{T}_x f(t^n)\\
e_4^{n+1} &= \tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E}(I - \Pi_1^{n+1})\mathcal{T}_x f(t^n)\\
e_5^{n+1} &= \tilde{\mathcal{T}}_x^{n+1}\Pi_2^{n+1}(\mathcal{T}_{v,n+1}^{E} - \mathcal{T}_{v,n+1}^{E_h})\tilde{\mathcal{T}}_x^{n+1/2} f(t^n)\\
e_6^{n+1} &= \tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2}(f(t^n) - f^n).
\end{aligned}$$

et pour étudier le cinquième terme, on va écrire :

$$[\mathcal{T}_{v,n+1}^{E} - \mathcal{T}_{v,n+1}^{E_h}]g = g(t, x, v - \Delta t E_{\mathcal{T}_x f(t^n)}) - g(t, x, v - \Delta t E_h^n), \tag{5.8.35}$$

où l'on décompose

$$E_{\mathcal{T}_x f(t^n)} - E_h^n = E_{\mathcal{T}_x f(t^n)} - E^n + E^n - E_h^n, \tag{5.8.36}$$

et

$$\begin{aligned}
(E_{\mathcal{T}_x f(t^n)} - E^n)(x) = \int K(x,y)\Big(\int [\tilde{T}_x^{n+1/2}(f(t^n) - f^n)\\
+ (I - \Pi_1^{n+1})T_x f(t^n)](y,v)\Big)\mathrm{d}v - 1\Big)\mathrm{d}y,
\end{aligned} \tag{5.8.37}$$

en remarquant que $\tilde{\mathcal{T}}_x^{n+1/2} f(t^n) = \Pi_1^{n+1}\mathcal{T}_x f(t^n)$.
D'autre part, dans le cas adaptatif, on va prendre avantage de la régularité de la solution numérique.
On utilise cette fois-ci les relations :

$$\begin{aligned}
\mathcal{T}_x\mathcal{T}_{v,n+1}^{E}\mathcal{T}_x f(t^n) &= \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\\
\mathcal{T}_x\mathcal{T}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n &= \mathcal{T}_x\mathcal{T}_{v,n+1}^{E_h}\Pi_1^{n+1}\mathcal{T}_x f^n\\
\mathcal{T}_x\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n &= \mathcal{T}_x\Pi_2^{n+1}\mathcal{T}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n\\
\tilde{\mathcal{T}}_x^{n+1}\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n &= \Pi_3^{n+1}\mathcal{T}_x\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n,
\end{aligned}$$

et on décompose alors l'erreur en

$$e^{n+1} = e_1^{n+1} + \varepsilon_2^{n+1} + \varepsilon_3^{n+1} + \varepsilon_4^{n+1} + \varepsilon_5^{n+1} + \varepsilon_6^{n+1} \tag{5.8.38}$$

avec

$$
\begin{aligned}
e_1^{n+1} &= f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n) \\
\varepsilon_2^{n+1} &:= \mathcal{T}_x\mathcal{T}_{v,n+1}^{E}\mathcal{T}_x(f(t^n) - f^n) \\
\varepsilon_3^{n+1} &:= \mathcal{T}_x(\mathcal{T}_{v,n+1}^{E} - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f^n \\
\varepsilon_4^{n+1} &:= \mathcal{T}_x\mathcal{T}_{v,n+1}^{E_h}(I - \Pi_1^{n+1})\mathcal{T}_x f^n \\
\varepsilon_5^{n+1} &:= \mathcal{T}_x(I - \Pi_2^{n+1})\mathcal{T}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n \\
\varepsilon_6^{n+1} &:= (I - \Pi_3^{n+1})\mathcal{T}_x\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x^{n+1/2} f^n,
\end{aligned}
$$

et cette fois-ci, pour étudier le troisième terme, on décompose l'erreur sur le champ électrique de la manière suivante :

$$
(E_{\mathcal{T}_x f(t^n)} - E^n)(x) = \int K(x,y)([\int T_x(f(t^n) - f^n) + (I - \Pi_1^{n+1})T_x f^n](y,v))\mathrm{d}v - 1)\mathrm{d}y. \tag{5.8.39}
$$

Chapitre 6

Convergence of classes of high-order semi-Lagrangian schemes for the Vlasov-Poisson system

Dans ce chapitre, on présente quelques classes de schémas semi-Lagrangiens pour résoudre le système de Vlasov-Poisson dans l'espace des phases sur des maillages uniformes. On démontre que la fonction de distribution $f(t,x,v)$ et le champ électrique $E(t,x)$ convergent en norme L^2, avec un taux en

$$\mathcal{O}\left(\Delta t^2 + \frac{h^{m+1}}{\Delta t}\right),$$

où m est le degré de la reconstruction polynômiale, Δt et h sont respectivement les paramètres de discrétisation en temps et dans l'espace des phases. Cette étude complète les résultats de [15].

Collaboration Ce travail fait l'objet d'un article soumis avec Nicolas Besse. Il s'agit ici d'une copie de l'article.

6.1 Introduction

In this paper the numerical analysis of semi-Lagrangian methods started in [16, 17] is continued. The high-order error estimates announced in the paper [16], are rigorously proved in the present one. If a lot of papers present satisfactory numerical results using semi-Lagrangian methods [7, 18, 26, 69], few rigorous mathematical results on convergence analysis of semi-Lagrangian methods have been stated. Although interesting a *priori* estimates have been pointed out (cf [10, 11, 33]), a lot of work still remains to give complete

and rigorous results in more general situations. The more difficult step in the convergence analysis of semi-Lagrangian method is to obtain a stability result for the interpolation operators. If stability results in L^∞ norm seem inaccessible for high-order interpolation operators because of the Runge phenomena (artificial oscillations whose amplitude increases with the degree of the polynomial in the case of Lagrange interpolation, appear at the edges of finite elements), a more appropriate mathematical framework is L^2 stability. This paper is organized as follows. In the first part we present the continuous problem. In the second part we expose the discrete problem and the numerical scheme to solve it. Then we study the convergence of our numerical scheme. The already known inequalities, crucial for the L^2 stability, are given, with proofs which are somewhat new, self-contained and can be read independantly in the appendix A and B.

6.2 The continuous problem

Denoting by $f(t,x,v) \geq 0$ the distribution function of electrons in phase-space (with mass normalized to one and charge to plus one), and by $E(t,x)$ the self consistent electric field, the adimensional Vlasov-Poisson system reads

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + E(t,x)\frac{\partial f}{\partial v} = 0, \tag{6.2.1}$$

$$\frac{dE}{dx}(t,x) = \rho(t,x) = \int_{-\infty}^{+\infty} f(t,x,v)dv - 1, \tag{6.2.2}$$

where x and v are independent variables. We consider a periodic plasma of period L. Hence in (6.2.1) et (6.2.2) we have $x \in [0,L]$, $v \in \mathbb{R}$, $t \geq 0$, and the functions f and E satisfy the periodic boundary conditions

$$f(t,0,v) = f(t,L,v) \qquad v \in \mathbb{R} \quad t \geq 0, \tag{6.2.3}$$

and

$$E(t,0) = E(t,L) \iff \frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f(t,x,v)dvdx = 1, \ t \geq 0, \tag{6.2.4}$$

which means that the plasma is globally neutral. In order to have a well posed problem we add to equations (6.2.1)-(6.2.4) a zero-mean electrostatic condition :

$$\int_0^L E(t,x)dx = 0, \quad t \geq 0, \tag{6.2.5}$$

and a initial condition

$$f(0,x,v) = f_0(x,v), \quad x \in [0,L], \quad v \in \mathbb{R}. \tag{6.2.6}$$

Besides, assuming that the electric field E is smooth enough we can solve equations (6.2.1), (6.2.3) et (6.2.6) in the classical sense as follows. For the existence, the uniqueness

and the regularity of the solutions of the following differential system we refer the reader to [20].

We consider the first order differential system

$$\begin{aligned} \frac{dX}{dt}(t;s,x,v) &= V(t;s,x,v), \\ \frac{dV}{dt}(t;s,x,v) &= E(t,X(t;s,x,v)) \end{aligned} \tag{6.2.7}$$

and denote by $t \to (X(t;s,x,v), V(t;s,x,v))$ the characteristic curves, which are the solution of (6.2.7) with the initial condition

$$X(s;s,x,v) = x \qquad V(s;s,x,v) = v. \tag{6.2.8}$$

Then the solution of problem (6.2.1), (6.2.6) is given by

$$f(t,x,v) = f_0(X(0;t,x,v), V(0;t,x,v)), \quad x,\ v \in \mathbb{R}\ \ t \geq 0. \tag{6.2.9}$$

We note that the periodicity in x of $f_0(x,v)$, and $E(t,x)$ implies the periodicity in x of $f(t,x,v)$. Moreover as

$$\left| \frac{\partial(X,V)}{\partial(x,v)} \right| = 1,$$

we get

$$\frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f(t,x,v)dvdx = \frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f_0(x,v)dvdx = 1.$$

Therefore, according to the previous considerations, an equivalent form of the Vlasov-Poisson periodic problem is to find a couple (f,E) , smooth enough, periodic with respect to x, with period L, and solving the equations (6.2.2), (6.2.7), (6.2.8), and (6.2.9).

If we introduce the electrostatic potential $\phi = \phi(t,x)$ such that $E(t,x) = -\partial_x \phi(t,x)$, and if we denote by $G = G(x,y)$ the fundamental solution of the Laplacian operator in one dimension with periodic boundary condition, we obtain

$$E(t,x) = \int_0^L K(x,y) \left(\int_{-\infty}^{+\infty} f(t,y,v)dv - 1 \right) dy,$$

where

$$K(x,y) = -\partial_x G(x,y) = \begin{cases} (\dfrac{y}{L} - 1) & 0 \leq x < y \\ \dfrac{y}{L} & y < x \leq L. \end{cases}$$

6.2.1 Existence, uniqueness and regularity of the solution of the continuous problem

In this section we recall a theorem of existence of classical solutions for the Vlasov-Poisson system. The following theorem gives the existence, the uniqueness and the regularity of the classical solutions, global in time, of the Vlasov-Poisson periodic system in one dimension.

Theorem 6.2.1. *Assuming that $f_0 \in \mathscr{C}^1_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$ (continuously differentiable functions which are periodic with respect to x and compactly supported with respect to v), positive, periodic with respect to the variable x with period L, and $Q(0) \leq R$ with $R > 0$ and $Q(t)$ defined as follows*

$$Q(t) = 1 + \sup\left\{|v| : \exists x \in [0, L], \quad \tau \in [0, t] \mid f(\tau, x, v) \neq 0\right\},$$

and

$$\frac{1}{L}\int_0^L \int_{-\infty}^{+\infty} f_0(x, v) dv dx = 1,$$

then the periodic Vlasov-Poisson system has a unique classical solution (f, E) periodic in x, with period L, for all time t in $[0, T]$, such that

$$f \in \mathscr{C}^1_b\left(0, T; \mathscr{C}^1_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)\right)$$

$$E \in \mathscr{C}^1_b\left(0, T; \mathscr{C}^1_{b,per_x}(\mathbb{R})\right),$$

and there exists a constant $C = C(R, f_0)$ dependent of R and f_0 such that

$$Q(T) \leq CT.$$

Moreover if we assume that $f_0 \in \mathscr{C}^m_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$, then

$$(f, E) \in \mathscr{C}^m_b\left(0, T; \mathscr{C}^m_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)\right) \times \mathscr{C}^m_b\left(0, T; \mathscr{C}^m_{b,per_x}(\mathbb{R})\right),$$

for all finite time T.

Proof. For the proof we refer the reader to the articles [27], [41], [47], [48], [49], [53], [63], [65], [73] □

6.3 The discrete problem.

6.3.1 Approximation spaces and interpolation operators

Let $\Omega = [0, L] \times [-R, R]$, with $R > Q(T)$, and $\mathcal{M}_h$ a cartesian mesh of the phase-space Ω. Then $\mathcal{M}_h$ is given by a first increasing sequence $(x_i)_{i \in \{0, \ldots, N_x\}}$ of the interval $[0, L]$ and

a second increasing sequence $(v_i)_{i\in\{0,\dots,N_v\}}$ of the interval $[-R,R]$.
Let $\Delta x_i = x_{i+1} - x_i$ the physical space set and $\Delta v_i = v_{i+1} - v_i$ the velocity space set. In order to simplify the convergence analysis we suppose that $\Delta x_i = \Delta x = L/(N_x+1)$, and $\Delta v_i = \Delta v = 2R/(N_v+1)$ where $N_x, N_v \in \mathbb{N}$. Then we define $h = \max\{\Delta x, \Delta v\}$.
We fix an ending time T, and we consider a uniform time discretization (t^n) of the interval $[0,T]$, with a time step $\Delta t = t^{n+1} - t^n$. On each point $(x_i, v_j) \in \mathcal{M}_h$, we will compute an approximation $f^n(x_i, v_j)$ of the exact distribution function $f(t^n, x_i, v_j)$ at time $t^n = n\Delta t$, and the approximation function f^n is then given on each point of $\mathbb{R}_x \times \mathbb{R}_v$ thanks to an interpolation operator $\mathscr{R}_h$, as we shall see below.
For each function g defined on all the points $(x_i, v_j) \in \mathcal{M}_h$, we will write $g_{i,j} := g(x_i, v_j)$, and we complete the sequence on $\mathbb{Z}\times\mathbb{Z}$ by periodicity, setting $g_{i,j} := g_{i \bmod N_x+1, j \bmod N_v+1}$.
The sequence (x_i, v_j) will also be defined on the whole set $\mathbb{Z}\times\mathbb{Z}$, by $x_i := i\Delta x$ and $v_j := -R + j\Delta v$.
We denote by $\mathcal{P}_{L,R}$ the set of all the functions L–periodic in x and $2R$–periodic in v. We notice that for $g \in \mathcal{P}_{L,R}$, we have $g_{i,j} = g(x_i, v_j)$, for all indices $(i,j) \in \mathbb{Z}\times\mathbb{Z}$. In fact, we will see that all the functions which we consider on $\mathbb{R}_x \times \mathbb{R}_v$ will belong to $\mathcal{P}_{L,R}$.

Lagrange interpolation

In order to avoid the Runge phenomena near the edge of the finite elements, we will only interpolate on a middle cell, by using symmetric Lagrange bases.
Let $x_{i+1/2} = (x_i + x_{i+1})/2$, then we define the cells $C_{i+1/2,j+1/2}$ and $C_{i,j}$ by

$$C_{i+1/2,j+1/2} = (x_i, x_{i+1}) \times (v_j, v_{j+1}), \quad C_{i,j} = (x_{i-1/2}, x_{i+1/2}) \times (v_{j-1/2}, v_{j+1/2}).$$

Therefore, we introduce the characteristic functions $\chi_{i+1/2,j+1/2}$ and $\chi_{i,j}$ defined by

$$\chi_{i+1/2,j+1/2}(x,v) = \begin{cases} 1 \text{ if } (x,v) \in C_{i+1/2,j+1/2}, \\ 0 \text{ otherwise} \end{cases}$$

$$\chi_{i,j}(x,v) = \begin{cases} 1 \text{ if } (x,v) \in C_{i,j}, \\ 0 \text{ otherwise.} \end{cases}$$

Let P_m be the Lagrange polynomial space of order less or equal to m in one dimension.
Let X_h, be the approximation space defined by

$$X_h = \left\{ f \in C(\Omega) \cap H^1(\Omega) \cap \mathcal{P}_{L,R} \mid f_{|C_{i+1/2,j+1/2}} \in P_m \otimes P_m,\ i \in \mathbb{Z},\ j \in \mathbb{Z} \right\},$$

if m is odd, and

$$X_h = \left\{ g \in C(\Omega) \cap H^1(\Omega) \cap \mathcal{P}_{L,R} \mid f_{|C_{i,j}} \in P_m \otimes P_m,\ i \in \mathbb{Z}/N_x,\ j \in \mathbb{Z}/N_v \right\},$$

if m is even.
In one dimension, if m is odd, we define the symmetric $m+1$ interpolation points on (x_i, x_{i+1}), by

$$x_{i-(m-1)/2}, \dots, x_{i+(m+1)/2},$$

and if m is even, the $m+1$ interpolation points on $(x_{i-1/2}, x_{i+1/2})$ are this time given by

$$x_{i-m/2}, \ldots, x_{i+m/2},$$

and are also symmetric with respect to the interval $(x_{i-1/2}, x_{i+1/2})$.
We denote by $\{\ell^i_{k,\Delta x}\}_{k\in\{0,\ldots,m\}}$ the Lagrange basis associated to the interval (x_i, x_{i+1}) (resp. $(x_{i-1/2}, x_{i+1/2})$) that is

$$\ell^i_{k,\Delta_x}(z) = \prod_{\substack{s=0\\ s\neq k}}^{s=m} \frac{(z-z_s)}{(z_k-z_s)},$$

with the interpolation points $z_s = x_{i-(m-1)/2} + s\Delta x$ (respectively $z_s = x_{i-m/2} + s\Delta x$). We define similarly $\ell^j_{k,\Delta v}$. Now, we define the interpolation operator $\mathcal{L}_h$ as follows. We consider a function g defined on the grid $\mathcal{M}_h$.
If m is odd we define $\mathcal{L}_h$ as

$$\mathcal{L}_h g(x,v) = \sum_{i\in\mathbb{Z}}\sum_{j\in\mathbb{Z}} \mathcal{L}_h^{i+1/2,j+1/2} g(x,v)\chi_{i+1/2,j+1/2}(x,v) \tag{6.3.1}$$

where we have put

$$\mathcal{L}_h^{i+1/2,j+1/2} g(x,v) = \mathcal{L}_h g(x,v)_{|C_{i+1/2,j+1/2}} = \sum_{k=i-(m-1)/2}^{i+(m+1)/2} \sum_{l=j-(m-1)/2}^{j+(m+1)/2} g_{k,l}\ell^i_{k,\Delta x}(x)\ell^j_{l,\Delta v}(v) \tag{6.3.2}$$

and if m is even we define $\mathcal{L}_h$ as

$$\mathcal{L}_h g(x,v) = \sum_{i\in\mathbb{Z}}\sum_{j\in\mathbb{Z}} \mathcal{L}_h^{i,j} g(x,v)\chi_{i,j}(x,v) \tag{6.3.3}$$

where we have this time

$$\mathcal{L}_h^{i,j} g(x,v) = \mathcal{L}_h g(x,v)_{|C_{i,j}} = \sum_{k=i-m/2}^{i+m/2} \sum_{l=j-m/2}^{j+m/2} g_{k,l}\ell^i_{k,\Delta x}(x)\ell^j_{l,\Delta v}(v), \tag{6.3.4}$$

with $g_{k,l} = g(x_{k \mod N_x+1}, v_{l \mod N_v+1})$.
$\mathcal{L}_h$ is a continuous linear operator from $C^{m+1}(\Omega)\cap W^{m+1,p}(\Omega)$, $1\le p\le\infty$, into X_h and the following interpolation error estimates hold (see [23, 62]).

$$||f-\mathcal{L}_h f||_{C(\Omega)} \le Ch^{m+1} \max_{|\alpha|=m+1} \sup_{(x,v)\in\Omega} |D^\alpha f(x,v)|, f\in C^{m+1}(\Omega)\cap\mathcal{P}_{L,R}. \tag{6.3.5}$$

$$||f-\mathcal{L}_h f||_{W^{k,p}(\Omega)} \le Ch^{m+1-k}|f|_{W^{m+1,p}(\Omega)},\ k=0,1,\ f\in W^{m+1,p}(\Omega)\cap\mathcal{P}_{L,R}. \tag{6.3.6}$$

B-splines interpolation. In this section we define the space of B-splines of order $m+1$ and the approximation space Y_h. Let m and r be two positive integers, then we define $\mathcal{B}_{m+1,\Delta_x}$, the linear space of the B-spline functions of order $m+1$ relative to Δx as

$$\mathcal{B}_{m+1,\Delta_x} = \left\{ s(x)\in C^{m-1}(\mathbb{R}),\quad D^{m+1}s(x) = 0,\ \forall x\in(x_i,x_{i+1}), \forall i\in\mathbb{Z} \right\},$$

if $m+1$ is even, and

$$\mathcal{B}_{m+1,\Delta_x} = \left\{ s(x) \in C^{m-1}(\mathbb{R}), \quad D^{m+1}s(x) = 0, \ \forall x \in (x_{i-1/2}, x_{i+1/2}), \forall i \in \mathbb{Z} \right\},$$

if $m+1$ is odd.
In the same way we define $\mathcal{B}_{r+1,\Delta_v}$, the space of B-spline functions of order $r+1$ relative to Δv. Therefore the space of B-spline functions in two dimensions is defined as the tensor product of the spaces $\mathcal{B}_{m+1,\Delta_x}$ and $\mathcal{B}_{r+1,\Delta_v}$.

$$\begin{aligned} \mathcal{B}_{m+1,r+1,\Delta_x,\Delta_v} &= \mathcal{B}_{m+1,\Delta_x} \otimes \mathcal{B}_{r+1,\Delta_x} \\ &= \left\{ s(x,v) = s_1(x)s_2(v) : \quad s_1 \in \mathcal{B}_{m+1,\Delta_x}, \ s_2 \in \mathcal{B}_{r+1,\Delta_v} \right\}. \end{aligned}$$

Let us note that $\mathcal{S}_{m+1,r+1,\Delta_x,\Delta_v} \subset W^{k,p}(\mathbb{R}^2)$ with $k = \min(r,m)$ and $1 \leq p \leq \infty$. We suppose that $r = m$, then $\mathcal{B}_{m+1,h}$ denotes the two-dimensional B-spline functions space of order $m+1$. Hence Y_h is defined by

$$Y_h = \left\{ f \in C^{m-1}(\Omega) \cap \mathcal{P}_{L,R}, \ f \in \mathcal{B}_{m+1,h} \right\},$$

and the interpolation operator $\mathcal{S}_h$ is defined by

$$\mathcal{S}_h f = \sum_{i \in \mathbb{Z}} \sum_{j \in \mathbb{Z}} \gamma_{i,j}(f) B_{m+1}(x/\Delta x - i) B_{m+1}(v/\Delta v - j) \tag{6.3.7}$$

where the unidimensional B-spline B_m, of order m is recursively defined by

$$B_m(\cdot) = (\underbrace{B * \ldots * B}_{m \text{ times}})(\cdot) = \int B_{m-1}(\cdot - u) B(u) du,$$

with

$$B_1(u) = B(u) = \begin{cases} 1 & -1/2 \leq u \leq 1/2 \\ 0 & \text{elsewhere.} \end{cases}$$

The coefficients $\gamma_{i,j}(f)$ are solutions of the linear system

$$f_{i,j} = \sum_{k,l} \gamma_{k,l}(f) B_{m+1}(i-k) B_{m+1}(j-l), \tag{6.3.8}$$

for $(i,j) \in [0, N_x] \times [0, N_v]$, and $\gamma_{i,j}(f) := \gamma_{i \mod N_x+1, j \mod N_v+1}(f)$, for all indices in $\mathbb{Z} \times \mathbb{Z}$.
Now we recall some useful properties on B-splines interpolation.

i)

$$\mathcal{B}_{m+1,h} = \text{Span}\left\{ B_{m+1}(./\Delta x - i) B_{m+1}(./\Delta v - j); \ i \in \mathbb{Z}, j \in \mathbb{Z} \right\} \tag{6.3.9}$$

ii)

$$\mathcal{B}_{m+1,h} \subset W^{m,p} \quad 1 \leq p \leq \infty \quad 0 \leq k \leq m \tag{6.3.10}$$

iii) stability :

$$\|\mathcal{S}_h f\|_{L^p(\Omega)} \leq C \|f\|_{L^p(\Omega)}, \quad \forall f \in L^p(\Omega) \cap \mathcal{P}_{L,R}, \quad 1 \leq p \leq \infty \tag{6.3.11}$$

iv) Consistency and optimal accuracy : For $1 \leq p \leq \infty$ and $0 \leq k \leq m$

$$\|\mathcal{S}_h f - f\|_{W^{k,p}(\Omega)} \leq C h^{m+1-k} |f|_{W^{m+1,p}(\Omega)}, \quad \forall f \in W^{m+1,p}(\Omega) \cap \mathcal{P}_{L,R} \tag{6.3.12}$$

v) The matrix $\left[\sum_{k,l} B_{m+1}(i-k) B_{m+1}(j-l)\right]_{i,j=0}^{N_x,N_v}$ is positive and definite, so that the solution of the linear system (6.3.8) exists and is unique.

vi)

$$\sum_i B_m(\cdot/h - i) = 1, \quad \int B_m(u) du = 1$$

vii) The B-spline $B_{m,i}(\cdot) = B_m(\cdot/h - i)$ is constructed on the points

$$\{x_{i-m/2}, \ldots, x_{i+m/2}\}.$$

6.3.2 The numerical scheme

We recall some notations of the preceding chapter in order to describe the numerical scheme and the further analysis. Thus, we define the transport operator in the x-direction $\mathcal{T}_x$

$$\mathcal{T}_x g(x,v) := g(x - v\Delta t/2, v),$$

for each distribution function g. The transport operator in the v-direction $\mathcal{T}_v$ is given by

$$\mathcal{T}_v g := \mathcal{T}_v^g g,$$

where $\mathcal{T}_v^{\tilde{g}}$ reads

$$\mathcal{T}_v^{\tilde{g}} g(x,v) := g(x, v - \Delta t E_{\tilde{g}}(x)),$$

with the electric field $E_{\tilde{g}}$ defined by

$$E_{\tilde{g}}(x) := \int_0^L K(x,y) \left(\int_{-R}^R \tilde{g}(y,v) dv - 1 \right) \mathrm{d}y, \quad x \in [0,L), \quad E_{\tilde{g}}(x+L) = E_{\tilde{g}}(x), \quad x \in \mathbb{R}.$$

Then, we define

$$\tilde{\mathcal{T}}_x := \mathscr{R}_h \mathcal{T}_x, \quad \tilde{\mathcal{T}}_v := \mathscr{R}_h \mathcal{T}_v,$$

where the operator $\mathscr{R}_h$ is either $\mathcal{L}_h$, if we use Lagrange elements, or $\mathcal{S}_h$ in the case we use B-splines. If g is defined on $\mathcal{M}_h$, then $\mathscr{R}_h g$ belongs to $Z_h \subset \mathcal{P}_{L,R} \cap C(\mathbb{R}_x \times \mathbb{R}_v)$, where Z_h denotes X_h or Y_h. Therefore the numerical scheme reads as follows.
First we suppose that we know $f^n \in Z_h$ at time t^n

S1 : We compute a first half backward advection in the x-direction of an increment $v\Delta t/2$, for each coordinate point $(x,v) \in \mathcal{M}_h$

$$f^n(x - v\Delta t/2, v).$$

Then, the new approximation after this first fractional time step is given by the approximation in the space Z_h defined on $\mathcal{M}_h$:

$$f_1^{n+1} := \tilde{\mathcal{T}}_x f^n$$

S2 : We compute an approximation of the electric field from the previous solution, for each coordinate point $(x, v) \in \mathcal{M}_h$:

$$E^{n+1/2}(x) := E_{f_1^{n+1}}(x)$$

S3 : We compute a backward advection in the v-direction of an increment $\Delta t E^{n+1/2}(x)$: for each coordinate point $(x, v) \in \mathcal{M}_h$

$$f_1^{n+1}(x, v - E^{n+1/2}(x)\Delta t).$$

Then, the new approximation after this second fractional step is given by

$$f_2^{n+1} := \tilde{\mathcal{T}}_v f_1^{n+1}.$$

S4 : We repeat the step (S1) and the new approximation at time t^{n+1} is given by

$$f^{n+1} = \tilde{\mathcal{T}}_x f_2^{n+1}.$$

In summary the numerical scheme is given by

$$f^{n+1} = \tilde{\mathcal{T}}_x \tilde{\mathcal{T}}_v \tilde{\mathcal{T}}_x f^n,$$

with $f^0 = \mathscr{R}_h f_0$, the discretization of the initial data, f_0. We can also compute the electric field at time t^n, which is given for $(x, v) \in \mathcal{M}_h$ by

$$E^n(x) := E_{f^n}(x).$$

6.3.3 A convergence theorem

Here we state the main theorem which gives convergence and error estimates for our scheme.

Theorem 6.3.1. *Assuming that $f_0 \in C^{m+1}_{c,L_x}(\mathbb{R}_x \times \mathbb{R}_v)$, positive, periodic with respect to the variable x with period L. Then the numerical solution of the Vlasov-Poisson system (f^n, E^n), computed by the numerical scheme exposed in section 6.3.2 converges towards the solution (f, E) of the periodic Vlasov-Poisson system on Ω, and there exists a constant $C = C\left(||f||_{C^2(0,T;C^{m+1}(\Omega))}\right)$ independent of Δt and h such that*

$$\|f(t^n) - f^n\|_{L^2(\Omega)} + \|E(t^n) - E^n\|_{L^\infty([0,L])} \leq C\left(\Delta t^2 + \frac{h^{m+1}}{\Delta t}\right)$$

.

We notice that it suffices to have the convergence in Ω, since we know that f is L–periodic in x and that $f(t, \cdot, v) = 0$ for $0 \leq t \leq T$, $|v| > R$, and in the sequel, we modify for convenience the definition of $f(t)$ outside Ω, such that $f(t) \in \mathcal{P}_{L,R}$, as it is the case for f^n. Since $R > Q(T)$, the new function f has the same regularity as the old one.

6.4 Proof of the convergence theorem

6.4.1 Idea of the proof

We remark that $f(t^{n+1}) = \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)$, and we then decompose the error

$$e^n := f(t^n) - f^n$$

as follows (as it was already announced in the preceding chapter) :

$$e^{n+1} = e_1^{n+1} + e_2^{n+1} + e_3^{n+1} + e_4^{n+1} + e_5^{n+1} + e_6^{n+1}$$

with

$$\begin{aligned}
e_1^{n+1} &= f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n) \\
e_2^{n+1} &= (I - \mathscr{R}_h)\mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n) \\
e_3^{n+1} &= \tilde{\mathcal{T}}_x(I - \mathscr{R}_h)\mathcal{T}_{v,n+1}^E T_x f(t^n) \\
e_4^{n+1} &= \tilde{\mathcal{T}}_x\mathscr{R}_h(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n) \\
e_5^{n+1} &= \tilde{\mathcal{T}}_x\tilde{\mathcal{T}}_{v,n+1}^{E_h}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n) \\
e_6^{n+1} &= \tilde{\mathcal{T}}_x\tilde{\mathcal{T}}_{v,n+1}^{E_h}\tilde{\mathcal{T}}_x(f(t^n) - f^n),
\end{aligned}$$

where we have put

$$\mathcal{T}_{v,n+1}^E g(x,v) = g(x, v - \Delta t E_{\mathcal{T}_x f(t^n)}),$$

for the advection with the exact electric field, and

$$\mathcal{T}_{v,n+1}^{E_h} g(x,v) = g(x, v - \Delta t E_{\tilde{\mathcal{T}}_x f^n}),$$

for the advection with the approximate one.
In order to prove the theorem, we shall use Fourier analysis tools. More precisely, we will classically define a discrete norm $\|\cdot\|_{L_h^2}$ satisfying $\|e^n\|_{L^2} \leq C\|e^n\|_{L_h^2}$, together with the stability condition

$$\|e^{n+1}\|_{L_h^2} \leq (1 + C\Delta t)\|e^n\|_{L_h^2} + C(\Delta t^3 + h^{m+1})$$

which ensures the convergence by using a Gronwall inequality. The plan of the proof is the following : we first introduce some notations relative to the Fourier analysis, we then state and prove the useful estimates that the interpolation operator $\mathscr{R}_h$ satisfies and finally we conclude to the proof of the theorem, from these estimates according to the above error decomposition.

6.4.2 Notations and definitions

We will use the indices $(i,j) \in \mathbb{Z}^2$ corresponding to the points $\mathbf{z}_{i,j} := (x_i, v_j)$, of the phase space, and the indices $\boldsymbol{\omega} = (\omega_x, \omega_v) \in \mathbb{Z}^2$, corresponding to the points of the

Fourier transform space $\mathbf{k}(\boldsymbol{\omega}) = (k_x(\omega), k_v(\omega))$, with $k_x = 2\pi\omega_x/L$ and $k_v = 2\pi\omega_v/(2R)$. We set also $\mathbf{N} = (N_x, N_v)$ and $\mathbf{0} = (0,0)$. Now, we define a grid-function f by a sequence $f = \left(f_{i,j}\right)_{(i,j)\in\mathbb{Z}\times\mathbb{Z}}$, periodic with respect to i and j with respectively the period $N_x + 1$ and $N_v + 1$. If f is a function defined on the points of $\mathcal{M}_h$, we can associate to it a grid-function $\tilde{f}$ defined by $\tilde{f}_{i,j} := f(x_i, v_j)$ for all $(i,j) = \mathbf{0}, \ldots, \mathbf{N}$ and we will write also f instead of $\tilde{f}$, in order not to make the notations two heavy. Let $L^2_h(\Omega)$ be the set of all the grid-functions equipped with the discrete norm $\|.\|_{L^2_h(\Omega)}$

$$\|f\|_{L^2_h(\Omega)} = \left(\Delta x \Delta v \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} |f_{i,j}|^2\right)^{1/2}.$$

As in the continuous case we can define a discrete scalar product $\langle .,. \rangle_{L^2_h(\Omega)}$ as follows. Let f and g be two grid-functions of $L^2_h(\Omega)$ the scalar product $\langle f, g\rangle_{L^2_h(\Omega)}$ is defined as

$$\langle f, g\rangle_{L^2_h(\Omega)} = \Delta x \Delta v \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} f_{i,j}\overline{g_{i,j}}.$$

The discrete Fourier transform is classically defined by

$$\widehat{f}(\boldsymbol{\omega}) = \frac{1}{|\Omega|^{1/2}} \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} f_{i,j} e^{-i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle}, \quad \boldsymbol{\omega} = \mathbf{0}, \ldots, \mathbf{N}. \tag{6.4.1}$$

We then have

$$f_{i,j} = \frac{1}{|\Omega|^{1/2}} \sum_{\boldsymbol{\omega}=0}^{\mathbf{N}} \widehat{f}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle}, \quad (i,j) = \mathbf{0}, \ldots, \mathbf{N}, \tag{6.4.2}$$

and also the Parseval formula

$$\|f\|^2_{L^2_h(\Omega)} = \Delta x \Delta v \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} |f_{i,j}|^2 = \sum_{\boldsymbol{\omega}=0}^{\mathbf{N}} \left|\widehat{f}(\boldsymbol{\omega})\right|^2,$$

which means that the transformation is unitary. Instead of making run ω_x from 0 to N_x (resp. N_v), by an obvious change of variable, ω_x goes from $-N_x/2$ to $N_x/2$, if N_x is even, or from $-(N_x - 1)/2$ to $(N_x + 1)/2$, if N_x is odd, and the same holds for N_v. In order to simplify notations, without loss of generality we suppose that N_x and N_v are even and we will then write for example

$$\sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} = \sum_{|\omega_x|\leq N_x/2} \sum_{|\omega_v|\leq N_v/2} = \sum_{\omega_x=-N_x/2}^{N_x/2} \sum_{\omega_v=-N_v/2}^{N_v/2}$$

We now define a shifted quantity, which will be useful in the stability analysis. Given a function $g \in \mathcal{P}_{L,R}$, and two vectors

$$\boldsymbol{\alpha} := (\alpha_0, \cdots, \alpha_j, \cdots, \alpha_{N_v}) \text{ and } \boldsymbol{\beta} := (\beta_0, \cdots, \beta_i, \cdots, \beta_{N_x})$$

satisfying $0 \leq \alpha_j, \beta_i < 1$, $\forall (i,j) = \mathbf{0}, \ldots \mathbf{N}$, we define

$$g_{i+\alpha_j, j+\beta_i} := g(x_i + \alpha_j \Delta x, v_j + \beta_i \Delta v), \quad (i,j) = \mathbf{0}, \ldots \mathbf{N}$$

and the shifted quantity $\| \cdot \|_{L^2_h, \Delta_h^{\boldsymbol{\alpha},\boldsymbol{\beta}}}$ by

$$\|g\|_{L^2_h, \Delta_h^{\boldsymbol{\alpha},\boldsymbol{\beta}}} := \left(\Delta x \Delta v \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} |g_{i+\alpha_j, j+\beta_i}|^2 \right)^{1/2}.$$

The definition is in fact valid for $\alpha_j, \beta_i \in \mathbb{R}$, since $g \in \mathcal{P}_{L,R}$. This quantity will practically be applied to the computation of the discrete norm of $\mathcal{T}_x g$ and $\mathcal{T}^{E_h}_{v,n+1}$, since we have

$$\|\mathcal{T}_x g\|_{L^2_h} = \|g\|_{L^2_h, \Delta_h^{\boldsymbol{\alpha},\mathbf{o}}},$$

and

$$\|\mathcal{T}^{E_h}_{v,n+1} g\|_{L^2_h} = \|g\|_{L^2_h, \Delta_h^{\mathbf{o},\boldsymbol{\beta}}},$$

with

$$\alpha_j = -v_j \Delta t/2, \quad \beta_i = -E^{n+1/2}(x_i) \Delta t, \quad (i,j) = \mathbf{0}, \ldots \mathbf{N},$$

for $g \in \mathcal{P}_{L,R}$.

6.4.3 Interpolation properties

In this section we establish the estimates which are valid for the interpolation operator $\mathscr{R}_h$, and which will be useful for the proof of the convergence.
The interpolation operator $\mathscr{R}_h$ is a linear projection (we have $\mathscr{R}_h \circ \mathscr{R}_h = \mathscr{R}_h$) which already satisfies an optimal accuracy property.

Lemma 6.4.1. *Assume that $g \in W^{m+1,2}(\Omega) \cap \mathcal{P}_{L,R}$, then*

$$\|(I - \mathscr{R}_h) g\|_{L^2(\Omega)} \leq C h^{m+1} |g|_{W^{m+1,2}(\Omega)}, \tag{6.4.3}$$

and

$$\|(I - \mathscr{R}_h) g\|_{W^{1,2}(\Omega)} \leq C h^{1} |g|_{W^{2,2}(\Omega)}. \tag{6.4.4}$$

Proof. These inequalities are subproducts of (6.3.6) if $\mathscr{R}_h = \mathcal{L}_h$, and of (6.3.12) if $\mathscr{R}_h = \mathcal{S}_h$, and rely on the fact that $\mathscr{R}_h$ preserves the polynomial of degree $\leq m$ (see theorem 5.6.1). □

We will measure the error in the discrete L^2–norm, instead of the L^2–norm; thus it would be useful that these two norms are equivalent, and this is effectively the case for the operators $\mathscr{R}_h$, that we consider.

Lemma 6.4.2. *For each $g \in \mathcal{P}_{L,R}$, we have $\|\mathscr{R}_h g\|_{L^2_h} \leq \|g\|_{L^2_h}$ and*

$$c\|g\|_{L^2_h} \leq \|\mathscr{R}_h g\|_{L^2(\Omega)} \leq C\|g\|_{L^2_h}.$$

Proof. Since $\|g\|_{L^2_h} = \|\mathscr{R}_h g\|_{L^2_h}$, because $\mathscr{R}_h$ interpolates on the grid $\mathcal{M}_h$, we already have the first inequality and also

$$c_h\|g\|_{L^2_h} \leq \|\mathscr{R}_h g\|_{L^2(\Omega)} \leq C_h\|g\|_{L^2_h},$$

with constants c_h and C_h depending on h. The main point to prove here, is that the constants involved can be made independent of h.
We consider at first the Lagrange interpolation, and that m is even (the proof is similar if m is odd). We use a change of variable in the following computation

$$\begin{aligned}
\int_{C_{i,j}} |\mathscr{R}_h g|^2 \mathrm{d}x\mathrm{d}v &= \int_{C_{i,j}} |\mathcal{L}^{i,j}_h g(x,v)|^2 \mathrm{d}x\mathrm{d}v \\
&= \int_{C_{i,j}} \Big| \sum_{|(k,l)|\leq \mathbf{m/2}} g_{i+k,j+l} \ell^i_{i+k,\Delta x}(x) \ell^j_{j+l,\Delta v}(v)\Big|^2 \mathrm{d}x\mathrm{d}v \\
&= \Delta x \Delta v \int_{[-1/2,1/2]^2} \Big| \sum_{|(k,l)|\leq \mathbf{m/2}} g_{i+k,j+l} \ell^0_{k,1}(x) \ell^0_{l,1}(v)\Big|^2 \mathrm{d}x\mathrm{d}v \\
&=: \Delta x \Delta v G(g_{i-m/2,j-m/2}, \ldots, g_{i+m/2,j+m/2}).
\end{aligned}$$

The quantity G is the square of a norm over $\mathbb{R}^{m+1}$, which is thus equivalent to the euclidian one, with constants c, C depending on m, but not on h. We thus obtain

$$\begin{aligned}
\int_\Omega |\mathscr{R}_h g|^2 &\leq \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} \int_{C_{i,j}} |\mathscr{R}_h g|^2 \mathrm{d}x\mathrm{d}v \\
&\leq C^2 \Delta x \Delta v \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} \sum_{(k,l)\leq \mathbf{m/2}} |g_{i+k,j+l}|^2 \\
&\leq C^2 \Delta x \Delta v (2m+1)^2 \sum_{(i,j)=\mathbf{0}}^{\mathbf{N}} |g_{i,j}|^2,
\end{aligned}$$

and the converse is also true, since we have

$$|g_{i,j}|^2 \leq \sum_{(k,l)\leq \mathbf{m/2}} |g_{i+k,j+l}|^2 \leq \frac{1}{c^2} \int_{C_{i,j}} |\mathscr{R}_h g|^2 \mathrm{d}x\mathrm{d}v.$$

Now, let us consider the case where we interpolate with splines. The same argument doesn't hold here directly, since the interpolation is not local. We will use here Fourier analysis in order to get the proof. However, the above treatment can already be applied to the grid-function $\gamma := (\gamma_{i,j}(g))$: we have

$$c\|\gamma\|_{L^2_h} \leq |\mathscr{R}_h g|_{L^2(\Omega)} \leq C\|\gamma\|_{L^2_h}. \tag{6.4.5}$$

Indeed, by supposing also for convenience that $m+1$ is even, we get

$$\int_{C_{i,j}} |\mathscr{R}_h g|^2 \mathrm{d}x\mathrm{d}v = \int_{C_{i,j}} \Big| \sum_{|(k,\ell)|\leq(\mathbf{m+1})/2} \gamma_{i+k,j+l} B_{m+1}(x/\Delta x - k) B_{m+1}(v/\Delta v - l)\Big|^2 \mathrm{d}x\mathrm{d}v,$$

since B_{m+1} has its support on the interval $[-(m+1)/2,(m+1)/2]$, and by considering again a change of variables, we get :

$$\int_{C_{i,j}} |\mathscr{R}_h g|^2 \mathrm{d}x\mathrm{d}v = \Delta x \Delta v \int_{[-1/2,1/2]^2} \Big| \sum_{|(k,\ell)|\leq(\mathbf{m+1})/2} \gamma_{i+k,j+l} B_{m+1}(x-k) B_{m+1}(v-l)\Big|^2 \mathrm{d}x\mathrm{d}v,$$

and the right hand side defines again the square of a norm, since we know that the matrix $\Big[\sum_{k,l} B_{m+1}(i-k)B_{m+1}(j-l)\Big]_{i,j=0}^{m+1/2,m+1/2}$ is invertible. Since all the quantities are now independent of h, we obtain the equivalence (6.4.5). It remains to prove that

$$c_1 |\gamma|_{L_h^2} \leq |g|_{L_h^2} \leq c_2 |\gamma|_{L_h^2}.$$

We will consider this relation in the discrete Fourier space, that means that we will establish that

$$c_1 |\widehat{\gamma}|_{L_h^2} \leq |\widehat{g}|_{L_h^2} \leq c_2 |\widehat{\gamma}|_{L_h^2}.$$

By substituting

$$f_{i,j} = \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} \quad \text{and} \quad \gamma_{k,l} = \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{\gamma}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{k,l}\rangle}$$

in (6.3.8) and using the notations

$$\boldsymbol{\alpha}_{i,j} = (i,j),\ \boldsymbol{\xi}(\boldsymbol{\omega}) = (\xi_x(\omega_x), \xi_v(\omega_v)) = (k_x(\omega_x)\Delta x, k_v(\omega_v)\Delta v)$$

we get

$$\begin{aligned}
&\frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} \\
&\qquad = \frac{1}{|\Omega|^{1/2}} \sum_{k,l} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{\gamma}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{k,l}\rangle} B_{m+1}(i-k) B_{m+1}(j-l) \\
&\qquad = \frac{1}{|\Omega|^{1/2}} \sum_{k,l} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{\gamma}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} e^{-i\langle \boldsymbol{\xi}(\boldsymbol{\omega}), \boldsymbol{\alpha}_{i,j} - \boldsymbol{\alpha}_{k,l}\rangle} B_{m+1}(i-k) B_{m+1}(j-l) \\
&\qquad = \frac{1}{|\Omega|^{1/2}} \sum_{p,q} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{\gamma}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} e^{-i\langle \boldsymbol{\xi}(\boldsymbol{\omega}), \boldsymbol{\alpha}_{p,q}\rangle} B_{m+1}(p) B^{m+1}(q)
\end{aligned}$$

and it follows

$$\frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} = \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{\gamma}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} \mathcal{D}(\boldsymbol{\omega}) \tag{6.4.6}$$

where $\mathcal{D}(\boldsymbol{\omega})$ is defined by

$$\mathcal{D}(\boldsymbol{\omega}) = \sum_{p,q} e^{-i\langle \boldsymbol{\xi}(\boldsymbol{\omega}), \boldsymbol{\alpha}_{p,q}\rangle} B_{m+1}(p) B_{m+1}(q).$$

By multiplying (6.4.6) with its conjugate and $\Delta x \Delta v$ we get

$$\begin{aligned} &\frac{\Delta x \Delta v}{|\Omega|} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \sum_{|\boldsymbol{\omega}'| \le \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) \overline{\widehat{f}(\boldsymbol{\omega}')} e^{i\langle \mathbf{k}(\boldsymbol{\omega}) - \mathbf{k}(\boldsymbol{\omega}'), \mathbf{z}_{i,j}\rangle} \\ &= \frac{\Delta x \Delta v}{|\Omega|} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \sum_{|\boldsymbol{\omega}'| \le \mathbf{N}/2} e^{i\langle \mathbf{k}(\boldsymbol{\omega}) - \mathbf{k}(\boldsymbol{\omega}'), \mathbf{z}_{i,j}\rangle} \widehat{\gamma}(\boldsymbol{\omega}) \overline{\widehat{\gamma}(\boldsymbol{\omega}')} \mathcal{D}(\boldsymbol{\omega}) \overline{\mathcal{D}(\boldsymbol{\omega}')}. \end{aligned}$$

As

$$\frac{\Delta x \Delta v}{|\Omega|} \sum_{i=0}^{N_x} \sum_{j=0}^{N_v} e^{i\langle \mathbf{k}(\boldsymbol{\omega}) - \mathbf{k}(\boldsymbol{\nu}), \mathbf{z}_{i,j}\rangle} = \delta_{\boldsymbol{\omega},\boldsymbol{\nu}} \tag{6.4.7}$$

we find the following relation which will also be useful later,

$$\sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} |f(\boldsymbol{\omega})|^2 = \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} |\widehat{\gamma}(\boldsymbol{\omega})|^2 |\mathcal{D}(\boldsymbol{\omega})|^2, \tag{6.4.8}$$

with

$$|\mathcal{D}(\boldsymbol{\omega})| = |\phi_{m+1}(0, \xi_x(\omega_x))| \cdot |\phi_{m+1}(0, \xi_v(\omega_v))|, \quad \phi_{m+1}(\alpha, \theta) := \sum_{p \in \mathbb{Z}} B_{m+1}(p + \alpha) e^{i\theta p}$$

The proof is then done, if we can establish that

$$\inf_{\theta \in \mathbb{R}} |\phi_{m+1}(0, \theta)| > 0 \quad \text{and} \quad \sup_{\theta \in \mathbb{R}} |\phi_{m+1}(0, \theta)| < \infty,$$

and this property of the B–splines is detailed at the end of the appendix (lemma 6.6.8).

□

The following lemma is crucial; it states the main property that the operator $\mathscr{R}_h$ satisfies.

Lemma 6.4.3. *For each $g \in \mathcal{P}_{L,R}$, and $\boldsymbol{\alpha} \in \mathbb{R}^{N_v}$ and $\boldsymbol{\beta} \in \mathbb{R}^{N_x}$, we have*

$$\|\mathscr{R}_h g\|_{L^2_h, \Delta_h^{\boldsymbol{\alpha}, \mathbf{0}}} \le \|g\|_{L^2_h},$$

and

$$\|\mathscr{R}_h g\|_{L^2_h, \Delta_h^{\mathbf{0}, \boldsymbol{\beta}}} \le \|g\|_{L^2_h}.$$

Proof. **The Lagrange case**
The first step consists in computing $\|\mathcal{L}_h f\|_{L^2_h,\Delta^{\boldsymbol{\alpha},0}_h}$ and $\|\mathcal{L}_h f\|_{L^2_h,\Delta^{0,\boldsymbol{\beta}}_h}$. We recall that $\mathbf{N} = (N_x, N_v)$ and $\mathbf{z}_{k,l} = (x_k, v_l)$. By writing $f_{k,l}$ using its Fourier series decomposition

$$f_{k,l} = \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{k,l}\rangle},$$

and introducing this latter formula in the definition of the interpolation operator $\mathcal{L}_h$ represented by the equations (6.3.1)-(6.3.4), with the following notations

$$p(i) = \begin{cases} i & \text{if } m \text{ is even} \\ i+1/2 & \text{if } m \text{ is odd} \end{cases}, \quad q(i) = \begin{cases} i & \text{if } m \text{ is even} \\ i+1/2 & \text{if } m \text{ is odd} \end{cases},$$

$$d_b(m) = \begin{cases} m/2 & \text{if } m \text{ is even} \\ (m-1)/2 & \text{if } m \text{ is odd} \end{cases}, \quad d_e(m) = \begin{cases} m/2 & \text{if } m \text{ is even} \\ (m+1)/2 & \text{if } m \text{ is odd,} \end{cases}$$

we obtain

$$\begin{aligned}
(\mathcal{L}_h f)_{i+\alpha_j,j} &= \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \sum_{k=i-d_b(m)}^{i+d_e(m)} \sum_{l=j-d_b(m)}^{j+d_e(m)} \widehat{f}(\boldsymbol{\omega}) \ell_k(x_i+\alpha_j\Delta x)\ell_l(v_j) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{k,l}\rangle} \\
&= \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \sum_{k=i-d_b(m)}^{i+d_e(m)} \widehat{f}(\boldsymbol{\omega}) \ell_k(x_i+\alpha_j\Delta x) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{k,j}\rangle} \\
&= \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) \varrho(\alpha_j, \omega_x) e^{-i\langle \mathbf{k}(\boldsymbol{\omega}), (d_b,0)\rangle} e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle} \qquad (6.4.9)
\end{aligned}$$

where

$$\varrho(\alpha_j, \omega_x) = \sum_{k=0}^{m} \ell_k((d_b(m)+\alpha_j)\Delta x) e^{ik_x(\omega_x)x_k}.$$

Hence we get

$$\begin{aligned}
\|\mathcal{L}_h f\|^2_{L^2_h,\Delta^{\boldsymbol{\alpha},0}_h} &= \Delta x \Delta v \sum_{i=0}^{N_x}\sum_{j=0}^{N_v} (\mathcal{L}_h f)_{i+\alpha_j,j} \overline{(\mathcal{L}_h f)}_{i+\alpha_j,j} \\
&= \frac{\Delta x \Delta v}{|\Omega|} \sum_{i=0}^{N_x}\sum_{j=0}^{N_v} \sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} \sum_{|\boldsymbol{\omega}'|\leq \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega})\overline{\widehat{f}(\boldsymbol{\omega}')} \varrho(\alpha_j,\omega_x)\overline{\varrho(\alpha_j,\omega'_x)} \\
&\qquad e^{-i\langle \mathbf{k}(\boldsymbol{\omega})-\mathbf{k}(\boldsymbol{\omega}'), (d_b,0)\rangle} e^{i\langle \mathbf{k}(\boldsymbol{\omega})-\mathbf{k}(\boldsymbol{\omega}'), \mathbf{z}_{i,j}\rangle}.
\end{aligned}$$

Since

$$\frac{\Delta x}{|L|} \sum_{i=0}^{N_x} e^{i(k_x(\omega_x)-k_x(\omega'_x))x_i} = \delta_{\omega_x,\omega'_x} \qquad (6.4.10)$$

it follows

$$\|\mathcal{L}_h f\|^2_{L^2_h,\Delta_h^{\alpha,0}} = \frac{\Delta v}{|2R|}\sum_{j=0}^{N_v}\sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2}\sum_{|\omega'_v|\le N_v/2}\widehat{f}(\boldsymbol{\omega})\overline{\widehat{f}(\omega_x,\omega'_v)}|\varrho(\alpha_j,\omega_x)|^2 e^{i(k_v(\omega_v)-k_v(\omega'_v))v_j}.$$

Since

$$\frac{\Delta v}{|2R|}\sum_{j=0}^{N_v} e^{i(k_v(\omega_v)-k_v(\omega'_v))v_j} = \delta_{\omega_v,\omega'_v} \tag{6.4.11}$$

we get

$$\begin{aligned}
\|\mathcal{L}_h f\|^2_{L^2_h,\Delta_h^{\alpha,0}} &\le \sup\{|\varrho(\alpha,\omega_x)|^2,\ |\omega_x|\le N_x/2,\ 0\le\alpha\le 1\}\\
&\times \frac{\Delta v}{|2R|}\sum_{j=0}^{N_v}\sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2}\sum_{|\omega'_v|\le N_v/2}\widehat{f}(\boldsymbol{\omega})\overline{\widehat{f}(\omega_x,\omega'_v)}e^{i(k_v(\omega_v)-k_v(\omega'_v))v_j}\\
&\le \sup\{|\varrho(\alpha,\omega_x)|^2,\ |\omega_x|\le N_x/2,\ 0\le\alpha\le 1\}\sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2}|\widehat{f}(\boldsymbol{\omega})|^2\\
&\le \sup\{|\varrho(\alpha,\omega_x)|^2,\ |\omega_x|\le N_x/2,\ 0\le\alpha\le 1\}\|f\|^2_{L^2_h(\Omega)}
\end{aligned}$$

In the same way we obtain

$$\|\mathcal{L}_h f\|^2_{L^2_h,\Delta_h^{0,\beta}} \le \sup\{|\varrho(\beta,\omega_v)|^2,\ |\omega_v|\le N_v/2,\ 0\le\beta\le 1\}\|f\|^2_{L^2_h(\Omega)}.$$

The second step is to show that

$$\sup\{|\varrho(\alpha,\omega_x)|^2,\ 0\le\alpha\le 1,\ |\omega_x|\le N_x/2\}\le 1,$$

and

$$\sup\{|\varrho(\beta,\omega_v)|^2,\ 0\le\beta\le 1,\ |\omega_v|\le N_v/2\}\le 1,$$

where

$$\varrho(\alpha,\omega_x)=\sum_{k=0}^{m}\ell_k((d_b+\alpha)\Delta x)e^{ik_x x_k}\quad\text{and}\quad\varrho(\beta,\omega_v)=\sum_{l=0}^{m}\ell_l((d_b+\beta)\Delta v)e^{ik_v v_l}.$$

Without loss of generality we can suppose that $L=2R=2\pi$, and $\Delta x=\Delta v=\Delta y$, so that we have to prove that

$$\sup\left\{|\varrho(y,\omega)|,\ y\in[-\Delta y(m-1)/2,\Delta y(m+1)/2],\ \omega\in\mathbb{Z}\right\}\le 1$$

where $\varrho(y,\omega)=\sum_{k=0}^{m}\ell_k(y)e^{i\omega y_k}$. If we set $\theta=\omega\Delta y$ and $y=\eta\Delta y$, with $0\le\eta\le m$, since $\varrho(\eta,\theta)$ is periodic in θ, now we have to show that

$$\sup\left\{|\varrho(\eta,\theta)|,\ \eta\in[(m-1)/2,(m+1)/2],\ \theta\in[0,2\pi]\right\}\le 1. \tag{6.4.12}$$

Finally if we make the change of variables $\xi=1-\cos\theta$ and $\zeta=\eta-m/2$ we obtain from theorem 6.5.5 and 6.5.12 in appendix 6.5 that

$$
\begin{aligned}
|\varrho^{\theta}_{2n}(\zeta)|^2 = |\varrho_{2n}(\zeta,\xi)|^2 = 1 - \zeta^2\left(1-\zeta^2\right)\cdots\left(n^2-\zeta^2\right)\xi^{n+1} \\
\left[c_0 + c_1\xi\left(1-\zeta^2\right) + \cdots\right. \\
\left. + c_{n-1}\xi^{n-1}\left(1-\zeta^2\right)\cdots\left((n-1)^2-\zeta^2\right)\right], \\
m = 2n, \\
|\varrho^{\theta}_{2n-1}(\zeta)|^2 = |\varrho_{2n-1}(\zeta,\xi)|^2 = 1 - \left((1/2)^2-\zeta^2\right)\cdots\left(((2n-1)/2)^2-\zeta^2\right)\xi^{n} \\
\left[d_0 + d_1\xi\left((1/2)^2-\zeta^2\right) + \ldots\right. \\
\left. + d_{n-1}\xi^{n-1}\left((1/2)^2-\zeta^2\right)\cdots\left(((2n-3)/2)^2-\zeta^2\right)\right], \\
m = 2n+1.
\end{aligned}
$$

with

$$
c_i = \frac{2^{n+i+1}}{(2n)!(2i+1)!(n+1+i)} > 0, \quad \text{and} \quad d_i = \frac{2^{n+i}}{(2n-1)!(2i)!(n+i)} > 0.
$$

These inequalities then imply (6.4.12).

The B-splines case

Now we want to compute $\|\mathcal{S}_h f\|_{L^2_h,\Delta^{\boldsymbol{\alpha},0}_h}$ and $\|\mathcal{S}_h f\|_{L^2_h,\Delta^{0,\boldsymbol{\beta}}_h}$. Let us start by $\|\mathcal{S}_h f\|_{L^2_h,\Delta^{\boldsymbol{\alpha},0}_h}$. We have

$$
\begin{aligned}
(\mathcal{S}_h f)_{i+\alpha_j,j} &= \sum_{k,l}\gamma_{k,l}(f)B_{m+1}(i+\alpha_j-k)B_{m+1}(j-l) \\
&= \frac{1}{|\Omega|^{1/2}}\sum_{k,l}\sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2}\widehat{\gamma}(\boldsymbol{\omega})e^{i\langle \mathbf{k}(\boldsymbol{\omega}),\mathbf{z}_{k,l}\rangle}B_{m+1}(i+\alpha_j-k)B_{m+1}(j-l) \\
&= \frac{1}{|\Omega|^{1/2}}\sum_{k,l}\sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2}\widehat{\gamma}(\boldsymbol{\omega})e^{i\langle \mathbf{k}(\boldsymbol{\omega}),\mathbf{z}_{i+\alpha_j,j}\rangle} \\
&\quad e^{-i\langle \mathbf{k}(\boldsymbol{\omega}),\mathbf{z}_{i+\alpha_j-k,j-l}\rangle}B_{m+1}(i+\alpha_j-k)B_{m+1}(j-l)
\end{aligned}
$$

so that

$$
(\mathcal{S}_h f)_{i+\alpha_j,j} = \frac{1}{|\Omega|^{1/2}}\sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2}\widehat{\gamma}(\boldsymbol{\omega})e^{i\langle \mathbf{k}(\boldsymbol{\omega}),\mathbf{z}_{i+\alpha_j,j}\rangle}\varrho(\omega_x,\alpha_j)D(\omega_x)D(\omega_v) \tag{6.4.13}
$$

where

$$
\varrho(\omega_x,\alpha_j) = \frac{\displaystyle\sum_k e^{i\xi_x(\omega_x)(i+\alpha_j-k)}B_{m+1}(i+\alpha_j-k)}{\displaystyle\sum_p e^{i\xi_x(\omega_x)p}B_{m+1}(p)} \tag{6.4.14}
$$

and

$$D(\omega_x) = \sum_p e^{i\xi_x(\omega_x)p} B_{m+1}(p).$$

By multiplying (6.4.13) with its conjugate, and summing on i and j, thanks to (6.4.10) we get

$$\|\mathcal{S}_h f\|^2_{L^2_h,\Delta_h^{\boldsymbol{\alpha},0}} = \frac{\Delta v}{|2R|} \sum_{\substack{|\boldsymbol{\omega}|\le \mathbf{N}/2 \\ |\omega_v'|\le N_v/2 \\ j=0,\dots,N_v}} \widehat{\gamma}(\boldsymbol{\omega})\overline{\widehat{\gamma}(\omega_x,\omega_v')} |D(\omega_x)|D(\omega_v)\overline{D(\omega_v')} \\ |\varrho(\alpha_j,\omega_x)|^2 e^{i(k_v(\omega_v)-k_v(\omega_v'))v_j}$$

which becomes using (6.4.11) and (6.4.8)

$$\|\mathcal{S}_h f\|^2_{L^2_h,\Delta_h^{\boldsymbol{\alpha},0}} \le \sum_{|\boldsymbol{\omega}|\le \mathbf{N}/2} \varrho_{\sup}(\omega_x)|\widehat{\gamma}(\boldsymbol{\omega})|^2 |\mathcal{D}(\boldsymbol{\omega})|^2 \le \varrho_{\sup} \|f\|^2_{L^2_h}$$

with

$$\varrho_{\sup} = \sup\{|\varrho(\alpha_j,\omega_x)|^2,\ 0\le\alpha_j\le 1,\ \omega_x\in\mathbb{Z}\}.$$

From theorem 6.6.1 in appendix 6.6, we get $|\varrho_{\sup}| \le 1$ and consequently we obtain

$$\|\mathcal{S}_h f\|_{L^2_h,\Delta_h^{\boldsymbol{\alpha},0}} \le \|f\|_{L^2_h(\Omega)} \qquad \text{and} \qquad \|\mathcal{S}_h f\|_{L^2_h,\Delta_h^{0,\boldsymbol{\beta}}} \le \|f\|_{L^2_h(\Omega)}$$

□

6.4.4 A priori estimates

The three following lemmae give information about the regularity of the transport operators, in the norms which will be useful for the further proof.

Lemma 6.4.4. *Let $p \in \mathbb{N}$ (in the application $p = 0, 1$ or $p = m+1$) and $g \in W^{p,2}(\Omega) \cap \mathcal{P}_{L,R}$, then we have*

$$\|\mathcal{T}_x g\|_{W^{p,2}(\Omega)} \le (1 + (\frac{\Delta t}{2})^p)|g|_{W^{p,2}(\Omega)}.$$

If we suppose that $g \in W^{1,2}(\Omega) \cap \mathcal{P}_{L,R}$, we get

$$\|\mathcal{T}^E_{v,n+1} g\|_{W^{1,2}(\Omega)} \le (1 + C_T\Delta t)|g|_{W^{1,2}(\Omega)},$$

where the constant C_T only depends on $\|f(t^n)\|_{L^\infty(\Omega)}$ and L.

Proof. The first equality directly follows from the fact that g is L–periodic in x, by using the Fubini-Tonelli theorem, computing the successive derivatives :

$$\|\mathcal{T}_x g\|^2_{W^{p,2}(\Omega)} = \int_0^L \int_{-R}^R |\partial_x^p g(x - v\Delta t/2, v)|^2 \\ + |(-\Delta t/2)^p \partial_x^p g(x - v\Delta t/2, v) + \partial_v^p g(x - v\Delta t/2, v)|^2 \mathrm{d}v\mathrm{d}x,$$

and applying at last the inequality $\sqrt{a^2+b^2} \leq |a|+|b|$, and the triangular inequality for the L^2–norm, we get finally

$$\begin{aligned}\|\mathcal{T}_x g\|_{W^{p,2}(\Omega)} &\leq \|\partial_x^p g(x-v\Delta t/2,v)\|_{L^2(\Omega)} + \|[(-\Delta t/2)^p\partial_x^p + \partial_v^p]g(x-v\Delta t/2,v)\|_{L^2(\Omega)} \\ &\leq (1+(\Delta t/2)^p)\|\partial_x^p g(x-v\Delta t/2,v)\|_{L^2(\Omega)} + \|\partial_v^p g(x-v\Delta t/2,v)\|_{L^2(\Omega)} \\ &\leq (1+(\Delta t/2)^p)|g|_{W^{p,2}(\Omega)}.\end{aligned}$$

Let us look at the second one. We first compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dx}}\mathcal{T}^E_{v,n+1}g(x,v) &= \frac{\mathrm{d}}{\mathrm{dx}}g(x,v-\Delta t E(x)) \\ &= \partial_x g(x,v-\Delta t E(x)) - \Delta t E'(x)\partial_v g(x,v-\Delta t E(x)), \\ \frac{\mathrm{d}}{\mathrm{dv}}\mathcal{T}^E_{v,n+1}g(x,v) &= \partial_v g(x,v-\Delta t E(x)),\end{aligned}$$

and we have

$$E'(x) = \int_{-R}^{R} \mathcal{T}_x f(t^n,x,v)\mathrm{d}v - 1,$$

and thus

$$|E'(x)| \leq 2R\|f(t^n)\|_{L^\infty(\Omega)} + 1,$$

which leads similarly to

$$\|\mathcal{T}^E_{v,n+1}g\|_{W^{1,2}(\Omega)} \leq (1+\Delta t(2R\|f(t^n)\|_{L^\infty(\Omega)}+1))|g|_{W^{1,2}(\Omega)}$$

□

The management of the advection with the numerical electric field is more delicate, we get here only the weaker estimates :

Lemma 6.4.5. *Suppose that $g \in W^{1,\infty}(\Omega)\cap\mathcal{P}_{L,R}$, we then have*

$$\|\mathcal{T}^{E_h}_{v,n+1}g\|_{W^{1,2}(\Omega)} \leq C(1+\Delta t C_T(1+|e^n|_{L^2(\Omega)}))|g|_{W^{1,\infty}(\Omega)},$$

and if $g \in W^{2,\infty}(\Omega)\cap\mathcal{P}_{L,R}$, we also have

$$\|\mathcal{T}^E_{v,n+1}g - \mathcal{T}^{E_h}_{v,n+1}g\|_{W^{1,2}(\Omega)} \leq C_T\Delta t|e^n|_{L^2(\Omega)}(|g|_{W^{1,\infty}(\Omega)}+|g|_{W^{2,\infty}(\Omega)}).$$

Proof. We can already remark that we have from the preceding lemma that

$$\|\mathcal{T}^E_{v,n+1}g\|_{W^{1,2}(\Omega)} \leq C_T|g|_{W^{1,2}(\Omega)} \leq C_T|g|_{W^{1,\infty}(\Omega)},$$

with a constant depending on $\|f(t^n)\|_{L^\infty(\Omega)}$; we can also obtain it directly, with a constant depending on $\|f(t^n)\|_{L^2(\Omega)}$: by computing, it yields

$$\|\frac{\mathrm{d}}{\mathrm{dx}}\mathcal{T}^E_{v,n+1}g(x,v)\|_{L^2(\Omega)} \leq \|\partial_x g\|_{L^2(\Omega)} + (\int_0^L \Delta t^2 |E'(x)|^2 \mathrm{d}x 2R)^{1/2} |g|_{W^{1,\infty}(\Omega)} \text{ and}$$

$$\begin{aligned}\int_0^L |E'(x)|^2 \mathrm{d}x &\leq 8R^2 \int_0^L (\int_{-R}^R \mathcal{T}_x f(t^n,x,v)\frac{\mathrm{d}v}{2R})^2 \mathrm{d}x + 2L \\ &\leq 8R^2 \int_0^L \int_{-R}^R |\mathcal{T}_x f(t^n,x,v)|^2 \frac{\mathrm{d}v}{2R}\mathrm{d}x + 2L \\ &= 4R\|f(t^n)\|^2_{L^2(\Omega)} + 2L.\end{aligned}$$

thanks to a Jensen inequality, and we get

$$\|\mathcal{T}^E_{v,n+1}g\|^2_{W^{1,2}(\Omega)} \leq C[1+\Delta t(\|f(t^n)\|_{L^2(\Omega)}+1)]|g|^2_{W^{1,\infty}(\Omega)},$$

Now, we begin to prove the second inequality. We have this time

$$\frac{\mathrm{d}}{\mathrm{d}x}[\mathcal{T}^E_{v,n+1}g(x,v) - \mathcal{T}^{E_h}_{v,n+1}g(x,v)] = E_1 - \Delta t(E_2+E_3),$$

and

$$\frac{\mathrm{d}}{\mathrm{d}v}[\mathcal{T}^E_{v,n+1}g(x,v) - \mathcal{T}^{E_h}_{v,n+1}g(x,v)] = E_4,$$

with

$$\begin{aligned}E_1 :=& \partial_x g(x, v-\Delta t E(x)) - \partial_x g(x, v-\Delta t E_h(x)),\\ E_2 :=& [E'(x) - E'_h(x)]\partial_v g(x, v-\Delta t E_h(x)),\\ E_3 :=& E'(x)[\partial_v g(x, v-\Delta t E(x)) - \partial_v g(x, v-\Delta t E_h(x))],\\ E_4 :=& \partial_v g(x, v-\Delta t E(x)) - \partial_v g(x, v-\Delta t E_h(x)),\end{aligned}$$

and

$$E'_h(x) = \int_{-R}^R \mathcal{T}_x f^n(x,v)\mathrm{d}v - 1.$$

Proceeding as before, we already get

$$\int_0^L |E'(x) - E'_h(x)|^2 \mathrm{d}x \leq 4R\|f(t^n) - f^n\|_{L^2(\Omega)},$$

which implies that

$$|E_2|_{L^2(\Omega)} \leq C\|e^n\|_{L^2(\Omega)}|g|_{W^{1,\infty}(\Omega)}.$$

Now, let $h \in \{\partial_x g, \partial_v g\}$. We then have

$$\begin{aligned}\int_\Omega |h(x,v-\Delta t E(x)) - h(x,v-\Delta t E_h(x))|^2 \mathrm{d}x\mathrm{d}v &= \Delta t^2 \int_\Omega |\int_{E(x)}^{E_h(x)} \partial_v h(x,v-\Delta t s)\mathrm{d}s|^2 \mathrm{d}x\mathrm{d}v \\ &\leq C\Delta t^2 \int_0^L |E(x) - E_h(x)|^2 \mathrm{d}x |\partial_v h|^2_{L^\infty(\Omega)} \\ &\leq C\Delta t^2 \|e^n\|^2_{L^2(\Omega)} |\partial_v h|^2_{L^\infty(\Omega)},\end{aligned}$$

thanks to a Jensen inequality. We have thus the result. It remains now to prove the first inequality, which reads, by introducing $E'(x)$ and by using the estimate for E_2 :

$$|\mathcal{T}_{v,n+1}^{E_h} g|_{W^{1,2}(\Omega)} \leq |\partial_x g|_{L^2(\Omega)} + \Delta t |E' \partial_v g|_{L^2(\Omega)} + |(E' - E_h')\partial_v g|_{L^2(\Omega)} + |\partial_v g|_{L^2(\Omega)}$$
$$\leq C|g|_{W^{1,\infty}(\Omega)}(1 + C_T \Delta t(1 + \|e^n\|_{L^2(\Omega)})).$$

□

We now also give a useful lemma for higher order estimates :

Lemma 6.4.6. *Suppose that $f \in \mathcal{P}_{L,R} \cap W^{m+1,\infty}(\Omega)$. We then get :*

$$\|\mathcal{T}_x g\|_{W^{m+1,\infty}(\Omega)} \leq (1 + (\Delta t/2)^p)\|g\|_{W^{m+1,\infty}(\Omega)},$$

and

$$|\mathcal{T}_{v,n+1}^{E} g|_{W^{m+1,2}(\Omega)} \leq C_T \|g\|_{W^{m+1,\infty}(\Omega)},$$

Proof. For the first inequality, we proceed like in the $W^{m+1,2}$ case of the first lemma. By computing the different derivatives, since E has its $k \leq m+1$ bounded derivatives, because it is the case for $f(t^n)$, we get the other inequality. □

Now we recall a lemma which manages the time discretization error :

Lemma 6.4.7. *We have :*

$$\|f(t^{n+1}) - \mathcal{T}_x \mathcal{T}_v \mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \leq C_T \Delta t^3.$$

Proof. See lemma 5.8.2, since we have

$$\|f(t^{n+1}) - \mathcal{T}_x \mathcal{T}_v \mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \leq C\|f(t^{n+1}) - \mathcal{T}_x \mathcal{T}_v \mathcal{T}_x f(t^n)\|_{L^\infty(\Omega)}.$$

□

We can also get an estimate for the coupling error between the Vlasov and Poisson equation in the L^2–norm :

Lemma 6.4.8. *We have*

$$\|[\mathcal{T}_{v,n+1}^{E} - \mathcal{T}_{v,n+1}^{E_h}]\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \leq C_T \Delta t \|e^n\|_{L^2(\Omega)}.$$

Proof. We have :

$$\|[\mathcal{T}_{v,n+1}^{E} - \mathcal{T}_{v,n+1}^{E_h}]\mathcal{T}_x f(t^n)\|_{L^2(\Omega)}^2 = \int_{-R}^{R}\int_{0}^{L} |\mathcal{T}_x f(t^n, x, v - E(x)\Delta t) - \mathcal{T}_x f(t^n, x, v - E_h(x)\Delta t)|^2 \mathrm{d}x\mathrm{d}v,$$

and since $|\mathcal{T}_x f(t^n)|_{W^{1,\infty}(\Omega)} \leq C_T$, we get :

$$\|\mathcal{T}^E_{v,n+1} - \mathcal{T}^{E_h}_{v,n+1}\mathcal{T}_x f(t^n)\|^2_{L^2(\Omega)} \leq C_T \Delta t^2 \int_0^L |E_h(x) - E(x)|^2 \mathrm{d}x \leq C_T \Delta t^2 \|e^n\|^2_{L^2(\Omega)},$$

the last inequality has already been used in lemma 6.4.5. □

We now can treat all the terms of the decomposition and conclude to the proof.

Proof of theorem. We now use the discrete norm. We begin with

$$\begin{aligned}\|e_1^{n+1}\|_{L^2_h} = \|f(t^{n+1}) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\|_{L^2_h} &\leq C\|\mathscr{R}_h(f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n))\|_{L^2(\Omega)} \\ &\leq \|(\mathscr{R}_h - I)(f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n))\|_{L^2(\Omega)} + \|f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &\leq C_T h^{m+1} + C_T \Delta t^3 \|e^n\|^2_{L^2(\Omega)},\end{aligned}$$

thanks to the lemma 6.4.7. The second term is in fact null, but if we want to use only the relations which are stated in the different lemmae from the section *interpolation properties*, we compute

$$\begin{aligned}\|e_2^{n+1}\|_{L^2_h} &= \|(I - \mathscr{R}_h)\mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\|_{L^2_h} \\ &\leq C\|\mathscr{R}_h(I - \mathscr{R}_h)\mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &= C\|(\mathscr{R}_h - \mathscr{R}_h \circ \mathscr{R}_h)\mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} = 0,\end{aligned}$$

since $\mathscr{R}_h$ is a projection. For the third term, we write

$$\begin{aligned}\|e_3^{n+1}\|_{L^2_h} &= \|\mathscr{R}_h\mathcal{T}_x(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{L^2_h} \\ &\leq \|\mathcal{T}_x(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{L^2_h} \\ &\leq C\|\mathscr{R}_h\mathcal{T}_x(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &\leq C\|(\mathscr{R}_h - I)\mathcal{T}_x(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} + C\|\mathcal{T}_x(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &\leq Ch\|\mathcal{T}_x(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{W^{1,2}(\Omega)} + C\|(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &\leq Ch\|(I - \mathscr{R}_h)\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{W^{1,2}(\Omega)} + Ch^{m+1}\|\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{W^{m+1,2}(\Omega)} \\ &\leq Ch^{m+1}\|\mathcal{T}^E_{v,n+1}\mathcal{T}_x f(t^n)\|_{W^{m+1,2}(\Omega)} \\ &\leq C_T h^{m+1}|\mathcal{T}_x f(t^n)|_{W^{m+1,\infty}(\Omega)} \leq C_T h^{m+1}|f(t^n)|_{W^{m+1,\infty}(\Omega)} \leq C_T h^{m+1}.\end{aligned}$$

Then we have

$$\begin{aligned}\|e_5^{n+1}\|_{L^2_h} &= \|\mathscr{R}_h\mathcal{T}_x\mathscr{R}_h\mathcal{T}^{E_h}_{v,n+1}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{L^2_h} \\ &\leq \|\mathcal{T}_x\mathscr{R}_h\mathcal{T}^{E_h}_{v,n+1}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{L^2_h} \\ &\leq \|\mathscr{R}_h\mathcal{T}^{E_h}_{v,n+1}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{L^2_h} \\ &\leq C\|(\mathscr{R}_h - I)\mathcal{T}^{E_h}_{v,n+1}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} + C\|\mathcal{T}^{E_h}_{v,n+1}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &\leq C_T h\|\mathcal{T}^{E_h}_{v,n+1}(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{W^{1,2}(\Omega)} + C\|(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} \\ &\leq C_T h(1 + \Delta t\|e^n\|_{L^2(\Omega)})\|(I - \mathscr{R}_h)\mathcal{T}_x f(t^n)\|_{W^{1,\infty}(\Omega)} + Ch^{m+1}\|\mathcal{T}_x f(t^n)\|_{W^{m+1,2}(\Omega)} \\ &\leq C_T h^{m+1}(1 + \Delta t\|e^n\|_{L^2(\Omega)})\|\mathcal{T}_x f(t^n)\|_{W^{m+1,\infty}(\Omega)} \\ &\leq C_T h^{m+1}(1 + \Delta t\|e^n\|_{L^2(\Omega)}),\end{aligned}$$

and since we also have

$$\|e^n\|_{L^2(\Omega)} \leq \|\mathscr{R}_h e^n\|_{L^2(\Omega)} + \|(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)} \leq C\|e^n\|_{L_h^2} + C_T h^{m+1},$$

we obtain

$$\|e_5^{n+1}\|_{L_h^2} \leq C_T h^{m+1}(1+\Delta t\|e^n\|_{L_h^2}).$$

For the coupling between the Vlasov and the Poisson equation in the discrete norm, we get :

$$\begin{aligned}
&\|e_4^{n+1}\|_{L_h^2} = \|\mathscr{R}_h\mathcal{T}_x\mathscr{R}_h(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)\|_{L_h^2}\\
&\leq \|\mathcal{T}_x\mathscr{R}_h(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)\|_{L_h^2}\\
&\leq \|(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)\|_{L_h^2}\\
&\leq \|\mathscr{R}_h(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)\|_{L^2(\Omega)}\\
&\leq \|(I-\mathscr{R}_h)(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)\|_{L^2(\Omega)} + \|(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)\|_{L^2(\Omega)}\\
&\leq |(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f(t^n)|_{W^{1,2}(\Omega)} + C_T\Delta t\|e^n\|_{L^2(\Omega)}\\
&\leq C_T\Delta t|e^n|_{L^2(\Omega)}\|\mathcal{T}_x f(t^n)\|_{W^{m+1,\infty}(\Omega)} + C_T\Delta t\|e^n\|_{L_h^2(\Omega)} + C_T h^{m+1} \leq C_T\Delta t|e^n|_{L_h^2(\Omega)} + C_T h^{m+1}.
\end{aligned}$$

Finally, the last term reads :

$$\begin{aligned}
&\|e_6^{n+1}\|_{L_h^2} = \|\mathscr{R}_h\mathcal{T}_x\mathscr{R}_h\mathcal{T}_{v,n+1}^{E_h}\mathscr{R}_h\mathcal{T}_x e^n\|_{L_h^2}\\
&\leq \|\mathcal{T}_x\mathscr{R}_h\mathcal{T}_{v,n+1}^{E_h}\mathscr{R}_h\mathcal{T}_x e^n\|_{L_h^2}\\
&\leq \|\mathcal{T}_{v,n+1}^{E_h}\mathscr{R}_h\mathcal{T}_x e^n\|_{L_h^2}\\
&\leq \|\mathcal{T}_x e^n\|_{L_h^2} \leq \|\mathcal{T}_x\mathscr{R}_h e^n\|_{L_h^2} + \|\mathcal{T}_x(I-\mathscr{R}_h)f(t^n)\|_{L_h^2}\\
&\leq \|e^n\|_{L_h^2} + C\|\mathscr{R}_h\mathcal{T}_x(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)}\\
&\leq \|e^n\|_{L_h^2} + C\|(\mathscr{R}_h - I)\mathcal{T}_x(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)} + C\|\mathcal{T}_x(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)}\\
&\leq \|e^n\|_{L_h^2} + Ch\|\mathcal{T}_x(I-\mathscr{R}_h)f(t^n)\|_{W^{1,2}(\Omega)} + C\|(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)}\\
&\leq \|e^n\|_{L_h^2} + C_T h|(I-\mathscr{R}_h)f(t^n)|_{W^{1,2}(\Omega)} + C\|(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)}\\
&\leq \|e^n\|_{L_h^2} + C_T h^{m+1}|f(t^n)|_{W^{m+1,2}(\Omega)} \leq \|e^n\|_{L_h^2} + C_T h^{m+1}
\end{aligned}$$

Now, by putting all together, we obtain :

$$\|e^{n+1}\|_{L_h^2} \leq (1+C_T\Delta t)\|e^n\|_{L_h^2} + C_T(\Delta t^3 + h^{m+1}),$$

and by applying a Gronwall inequality, we have then

$$\|e^n\|_{L_h^2} \leq (1+C_T\Delta t)^{T/\Delta t}\|e^0\|_{L_h^2} + C_T(\Delta t^2 + h^{m+1}/\Delta t).$$

We remark that $e^0 = 0$ (or we can use an argument as above, if we do not want to use the fact that $\|\mathscr{R}_h g\|_{L_h^2} = \|g\|_{L_h^2}$), and we then finally get

$$\begin{aligned}
\|e^n\|_{L^2(\Omega)} \leq \|\mathscr{R}_h e^n\|_{L^2(\Omega)} + \|(I-\mathscr{R}_h)f(t^n)\|_{L^2(\Omega)}&\\
\leq C\|e^n\|_{L_h^2} + h^{m+1}|f(t^n)|_{W^{m+1,2}(\Omega)}&\\
\leq C_T(\Delta t^2 + h^{m+1}/\Delta t).&
\end{aligned}$$

It remains to give a bound for the electric field. We have at first

$$|(E^n)'(x) - E'(x)| \leq \int_{-R}^{R} |f^n(x,v) - f(x,v)| \mathrm{d}v,$$

and thus thanks to a Jensen inequality, we get

$$|(E^n)'(x) - E'(x)|^2 \leq C \int_{-R}^{R} |(f^n - f)(x,v)|^2 \mathrm{d}v,$$

we integrate with respect to x,

$$\int_0^L |(E^n)'(x) - E'(x)|^2 \mathrm{d}x \leq C |e^n|^2_{L^2(\Omega)},$$

which gives the convergence of the derivative of the electric field in the L^2–norm. Now, in order to get the convergence in the L^∞–norm, we have, since $|K(x,y)| \leq 1$:

$$|E^n(x) - E(x)| \leq \int_\Omega |e^n(x,v)| \mathrm{d}x \mathrm{d}v \leq C |e^n|_{L^2(\Omega)}.$$

□

Remark 6.4.9. *Let us notice that our convergence analysis can be extended to higher splitting formulae in order to reach order $N \geq 3$ in time. To build approximations of order N in time (cf [39, 40, 78]) we consider splitting schemes of the form*

$$f(t + \Delta t) = \mathcal{T}_x^{\alpha_1}(\Delta t) \circ \mathcal{T}_v^{\beta_1}(\Delta t) \circ \ldots \circ \mathcal{T}_x^{\alpha_i}(\Delta t) \circ \mathcal{T}_v^{\beta_i}(\Delta t) \circ \ldots \circ \mathcal{T}_x^{\alpha_k}(\Delta t) \circ \mathcal{T}_v^{\beta_k}(\Delta t) f(t)$$

where

$$\mathcal{T}_x^{\alpha_i}(\Delta t) = \exp\left(\alpha_i \Delta t \mathcal{L}_x\right), \qquad \mathcal{T}_v^{\beta_i}(\Delta t) = \exp\left(\beta_i \Delta t \mathcal{L}_v\right),$$

with

$$\mathcal{L}_x = -v \cdot \partial_x, \quad \mathcal{L}_v = -E \cdot \partial_v.$$

The actions of $\mathcal{T}_x^{\alpha_i}(\Delta t)$ and $\mathcal{T}_v^{\beta_i}(\Delta t)$ on the function $f(x,v)$ are respectively

$$\mathcal{T}_x^{\alpha_i}(\Delta t) f(x,v) = f(x - \alpha_i v \Delta t, v), \quad \mathcal{T}_v^{\alpha_i}(\Delta t) f(x,v) = f(x, v - \beta_i E(t,x) \Delta t).$$

6.5 Lagrange interpolation

6.5.1 Introduction

Let us give an integer n, an interval $[-n, n]$ and the unique polynomial P_n of degree less or equal to $2n$ such that $P_n(j) = (-1)^j$ for $j = -n, \ldots, n$. What can we say about its L_∞–norm in the interval considered ? For $n = 1$, we see that $\sup_{[-1,1]} |P_1| = 1$. For

$n = 2$, the situation already changes (see figure 6.1) : we notice that the derivative at -1 is strictly positive and thus the polynomial admits a minimum less than -1 in the interval $[-2, -1]$. But, if we restrict our study to the middle interval $[-1, 1]$, we remark that $\sup_{[-1,1]} P_2 = 1$. Indeed, thanks to parity, the derivative vanishes once in $]-2, -1[$, once in $]1, 2[$ and once at 0 and can no more vanish, since it is of degree 3.

In fact, this situation is true for each integer n :

Proposition 6.5.1. *Let $n \in \mathbb{N}^*$ and $P_n \in \mathbb{R}_{2n}[X]$ such that $P_n(j) = (-1)^j$ for $j = -n, \ldots, n$. Then*

$$\sup_{[-1,1]} |P_n| = 1.$$

On the other hand, we cannot hope to have a longer interval :

Proposition 6.5.2. *Let $n \in \mathbb{N}^*$ and $P_n \in \mathbb{R}_{2n}[X]$ such that $P_n(j) = (-1)^j$ for $j = -n, \ldots, n$. Then there exist a number $0 \leq \varepsilon \leq 1$ such that for $k = 1, \ldots, n-1$,*

$$|P_n(k + \varepsilon)| > 1$$

The first result admits a generalization in the complex case, in the following sense :

Theorem 6.5.3. *Let $n \in \mathbb{N}^*$, $\theta \in \mathbb{R}$ and $\varrho^\theta_{2n} \in \mathbb{C}_{2n}[X]$ such that $\varrho^\theta_{2n}(j) = \exp(ij\theta)$ for $j = -n, \ldots, n$. Then*

$$\sup_{[-1,1]} |\varrho^\theta_{2n}| = 1$$

We also have the result in case of odd degrees :

Theorem 6.5.4. *Let $n \in \mathbb{N}$, $\theta \in \mathbb{R}$ and $\varrho^\theta_{2n-1} \in \mathbb{C}_{2n-1}[X]$ such that $\varrho^\theta_{2n-1}(j) = \exp(ij\theta)$ for $j = -n + 1/2, \ldots, n - 1/2$. Then*

$$\sup_{[-1/2,1/2]} |\varrho^\theta_{2n-1}| = 1$$

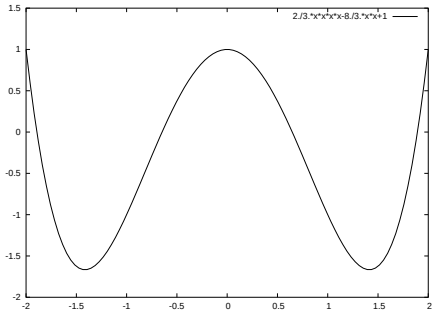


Fig. 6.1 – Graph of polynom P_2 of degree 4 such that $P_2(j) = (-1)^j$, $j = -2 \ldots 2$.

Theorem 6.5.3 and 6.5.4 were respectively proven for $n \leq 2$ and $n \leq 3$ in [15]. For the general case, the author conjectured in [15] the following formula without giving an explicit formula for the coefficients c_i :

Theorem 6.5.5. *Let $n \in \mathbb{N}^*$, $\theta \in \mathbb{R}$ and $\varrho_{2n}^{\theta} \in \mathbb{C}_{2n}[X]$ such that $\varrho_{2n}^{\theta}(j) = \exp(ij\theta)$ for $j = -n, \ldots, n$. By introducing $\xi := 1 - \cos(\theta)$, we then obtain :*

$$|\varrho_{2n}^{\theta}(\zeta)|^2 = 1 - \zeta^2(1-\zeta^2)\ldots(n^2-\zeta^2)\xi^{n+1} \\ [c_0 + c_1\xi(1-\zeta^2)c_{n-1}\xi^{n-1}(1-\zeta^2)\ldots((n-1)^2-\zeta^2)]$$

with positive numbers $c_i, i = 0, \ldots, n-1$ which are given by

$$c_i = \frac{2^{n+i+1}}{(2n)!(2i+1)!(n+i+1)}$$

This result which readily induces theorem 6.5.3 was in fact already known. Let us explain the historic. In fact theorem 6.5.3 was proven since 1962 by Strang in [70], but without establishing the formula (6.5.5) [see subsection 6.5.3 for an application of having such a formula]. Recently, (in 2001), in a preprint of [46], Steinberg and Hong obtained the same result, but this time with formula (6.5.5) (and a similar formula in the odd case). We present here another proof of (6.5.5), based on divided differences and algebraic polynomials. The proof is not shorter but avoids as many computations as in [46]. Concerning the odd case, we give a very short proof of 6.5.4 derived from the even case ; a formula is also derived by following [70].

6.5.2 Proof

We prove at first proposition 6.5.1 and 6.5.2. For this, we use the following lemma :

Lemma 6.5.6. *let $n \geq 1$ and $P_n \in \mathbb{R}_{2n}[X]$ such that $P_n(j) = (-1)^j$ for $j = -n, \ldots, n$. Then, the following formula stands*

$$P_n(x) + P_n(x+1) = A \prod_{\ell=-n}^{n-1} (x - \ell) \tag{6.5.1}$$

with a real A such that $(-1)^n A > 0$.

Proof. We know that $P_n(x) + P_n(x+1)$ is a polynomial of degree $2n$ which vanishes at the integers $-n, \ldots, n-1$ and thus we have this factorization. On the other hand, for each integer $k = -n, \ldots, n-1$, since $P_n(k)P_n(k+1) = -1$, P_n changes sign at least once in the interval $]k, k+1[$; but P_n is of degree $2n$, so P_n changes sign exactly one time in the interval $]k, k+1[$ and nowhere else (otherwise P_n would have more than $2n$ roots). So near infinity, P_n is of same sign than $P_n(n) = (-1)^n$ and behaves like $\frac{A}{2}x^{2n}$. So we have $(-1)^n A > 0$. □

Now, we turn to the proof of propositions 6.5.1 and 6.5.2. Let k be an integer with $-n \leq k \leq n-1$. From formula (6.5.1), we obtain :

$$P_n(k+h) + P_n(k+1+h) = Ah \prod_{\ell=-n}^{k-1} (k-\ell) \prod_{\ell=1}^{n-k-1} (-\ell) + o(h). \tag{6.5.2}$$

Thus there exists $\varepsilon > 0$ small enough such that $P_n(k+\varepsilon) + P_n(k+1+\varepsilon)$ has the same sign than $A(-1)^{n-k-1}$, that is $(-1)^{k+1}$, since $(-1)^n A$ is strictly positive. From (6.5.2), with $k=0$, we also obtain that $P_n'(0) + P_n'(1) < 0$. We know (by unicity of $P_n \in \mathbb{R}_{2n}[X]$ such that $P_n(j) = (-1)^j$ for $j = -n, \ldots, n$) that P_n is even, and so $P_n'(0) = 0$, $P_n'(1) < 0$ and $P_n(1+\varepsilon) < P_n(1) = -1$, by diminishing ε if necessary.
Now, we can conclude that

$$-1 > P_n(1+\varepsilon) > -P_n(2+\varepsilon) > \cdots > (-1)^{n+1} P_n(n+\varepsilon)$$

and the proposition 6.5.2 follows. We also deduce that we have an extremum on each interval $]k, k+1[$, for $k = 1, \ldots, n-1$. So P_n' must vanish $n-1$ times in the interval $]1, \infty[$. Since P_n' is odd, P_n must vanish $n-1$ times in $]-\infty, -1[$ and once at 0. On the other hand, P_n is of degree $2n$, so P_n' can vanish nowhere else. Thus, the extrema of P_n in the interval $[-1, 1]$ can only be at the points $-1, 0$ or 1 and proposition 6.5.1 is proven. So, we have treated the real case. Let us turn to the complex case.

We first give an explicit formula for ϱ_{2n}^θ which will be useful in the sequel. We recall that $\xi = 1 - \cos(\theta)$.

Lemma 6.5.7. *Let $\varrho_{2n}^\theta \in \mathbb{C}_{2n}[X]$ such that $\varrho_{2n}^\theta(j) = \exp(ij\theta)$ for $j = -n \ldots, n$. We have then :*

$$\varrho_{2n}^\theta(z) = \varrho_{2n-2}^\theta(z) + z(z^2-1)\ldots(z^2-(n-1)^2)(\varphi(z-n) + \overline{\varphi}(z+n)), \tag{6.5.3}$$

with $\varphi = \frac{(-2)^{n-1}}{(2n-1)!}\xi^{n-1}(e^{i\theta}-1)$.

Proof. We know that such a decomposition holds by expressing ϱ_{2n}^θ in the basis : $1, z, z(z-1), z(z-1)(z+1)\ldots$, and since

$$\overline{\varrho_{2n}^\theta(-z)} = \varrho_{2n}^{-\theta}(-z) = \varrho_{2n}^\theta(z). \tag{6.5.4}$$

The coefficient φ is given by the divided difference : $\varphi = \varrho_{2n}^\theta[0, 1, \ldots, n-1, -1, \ldots n+1, n]$ which can also be expressed in terms of $\varrho_{2n}^\theta(k) = \exp(ik\theta)$:

$$\varphi = \sum_{k=-n+1}^{n} \prod_{j=-n+1, j\neq k}^{n} \frac{1}{k-j} \exp(ik\theta)$$

Since $\prod_{j=-n+1}^{k-1}(k-j) = (n-1+k)!$ and $\prod_{j=k+1}^{n}(k-j) = (-1)^{n-k}(n-k)!$, we obtain that

$$\varphi = \sum_{k=0}^{2n-1} \frac{(-1)^{2n-1-k}}{k!(2n-1-k)!} e^{i(-n+1+k)\theta} = \frac{(e^{i\omega}-1)^{2n-1}}{(2n-1)!} e^{i(-n+1)\theta},$$

and

$$(e^{i\theta}-1)^{2n-2}e^{i(-n+1)\theta} = (e^{i\theta/2}-e^{-i\theta/2})^{2n-2} = (-2)^{n-1}\xi^{n-1},$$

since $2\sin^2\theta/2 = \xi$, and this gives the expression of φ. □

In order to prove the algebraic formula (6.5.5), we first give a less precise form of it.

Proposition 6.5.8. *We have the formula :*

$$|\varrho_{2n}^{\theta}(x)|^2 = 1 - x^2(1-x^2)\dots(n^2-x^2)[C_0(\xi) + C_1(\xi)(1-x^2) + \dots + C_{n-1}(\xi)(1-x^2)\dots((n-1)^2-x^2)], \quad (6.5.5)$$

where C_j are polynomial in ξ (for $j = 0,\dots,n-1$) and are given by

$$C_j(\xi) = (-1)^{n+1+j}\sum_{k=1}^{j+1}\frac{R_n'(k)}{\alpha_{k,j}}, \quad (6.5.6)$$

where the polynomial R_n is defined by

$$R_n(x^2) := |\varrho_{2n}^{\theta}(x)|^2 - 1$$

and $\alpha_{k,j} := k^2\prod_{\ell=0\ell\neq k}^{j+1}(k^2-\ell^2)^2\prod_{\ell=j+2}^{n}(k^2-\ell^2)$

Proof. Since $\varrho_{2n}^{\theta}\in\mathbb{C}_{2n}[X]$ and since $\varrho_{2n}^{\theta}\overline{\varrho_{2n}^{\theta}}$ is even (from 6.5.4), we deduce that $\varrho_{2n}^{\theta}\overline{\varrho_{2n}^{\theta}}\in\mathbb{C}_{2n}[X^2]$. Now, we decompose ϱ_{2n}^{θ} in the basis $x^2, x^2(x^2-1),\dots,x^2\dots(x^2-n^2), x^2\dots(x^2-n^2)(x^2-1),\dots,x^2\dots(x^2-n^2)(x^2-1)\dots(x^2-(n-1)^2)$. We remark that the coefficients of $x^2, x^2(x^2-1),\dots,x^2\dots(x^2-(n-1)^2)$ are null, since

$$\varrho_{2n}^{\theta}\overline{\varrho_{2n}^{\theta}}(j) - 1 = 0, \quad j = 0,\dots,n.$$

In order to calculate the other coefficients, we fix $1 > \varepsilon > 0$ and we decompose ϱ_{2n}^{θ} in the basis $x^2, x^2(x^2-1),\dots,x^2\dots(x^2-n^2), x^2\dots(x^2-n^2)(x^2-1-\varepsilon),\dots,x^2\dots(x^2-n^2)(x^2-1-\varepsilon)\dots(x^2-(n-1)^2-\varepsilon)$. Now, all the factors are distinct, and we can use the divided differences for evaluating the coefficients, and as ε tends to 0, we obtain :

$$(-1)^{n+1+j}C_j = \lim_{\varepsilon\to 0}\sum_{k=1}^{j+1}\frac{R_n(k+\varepsilon)}{\alpha_{k,j,\varepsilon}},$$

with $\alpha_{k,j,\varepsilon} := \prod_{\ell=0}^{n}(k^2+\varepsilon-\ell^2)\prod_{\ell=1\ell\neq k}^{j+1}(k^2+\varepsilon-\ell^2-\varepsilon)$, since $R_n(j) = 0$, for $j = 0,\dots,n$, and the formula (6.5.6) follows. It remains to prove that these coefficients can be expressed as polynomials in ξ. In fact, they are even in θ, since $\varrho_{2n}^{\theta}\overline{\varrho_{2n}^{\theta}}$ is real and thus they are polynomial in $\cos\theta$ (the remaining term in $\sin\theta$ disappears with parity) and therefore in ξ. □

In order to prove (6.5.5), we first prove :

Lemma 6.5.9. *The C_j are polynomial in ξ of degree less or equal than $n+1+j$.*

Proof. From (6.5.6), we know that C_j is a linear combination of

$$R'_n(1), \dots, R'_n((j+1)^2), \tag{6.5.7}$$

where the coefficients are independent of θ. Now, we have :

$$\varrho^\theta_{2n}(x) - \exp(-i\theta)\varrho^\theta_{2n}(x+1) = (\varphi + \overline{\varphi})(1-\exp(-i\theta))w(x), \tag{6.5.8}$$

where w stands for :

$$w(x) = x(x+n)\prod_{k=1}^{n-1}(x^2-k^2),$$

and φ is defined in lemma 6.5.7. Indeed the left-hand side vanishes at the points $-n, -n+1, \dots, n-1$ and is of degree less or equal than $2n$ and the leading coefficient of ϱ^θ_{2n} is given by $\varphi+\overline{\varphi}$, from lemma 6.5.7. Now, by differentiating the relation $R_n(x^2) = \varrho^\theta_{2n}(x)\overline{\varrho^\theta_{2n}(x)} - 1$, and evaluating at an integer $k = 0, \dots, n-1$, we obtain :

$$2(k+1)R'_n((k+1)^2) = e^{i(k+1)\theta}\overline{\varrho^{\theta\prime}_{2n}(k+1)} + e^{-i(k+1)\theta}\varrho^{\theta\prime}_{2n}(k+1).$$

and by doing the same with (6.5.8), we have :

$$\varrho^{\theta\prime}_{2n}(k+1) = \exp(i\theta)\varrho^{\theta\prime}_{2n}(k) - (\exp(i\theta)-1)(\varphi+\overline{\varphi})w'(k).$$

We combine the two preceding equalities :

$$2(k+1)R'_n((k+1)^2) = e^{ik\theta}\overline{\varrho^{\theta\prime}_{2n}(k)} + e^{-ik\theta}\varrho^{\theta\prime}_{2n}(k) \tag{6.5.9}$$

$$+ 2(\cos(k\theta) - \cos((k+1)\theta))(\varphi+\overline{\varphi})w'(k) \tag{6.5.10}$$

$$= 2kR'_n(k^2) + c\xi^n(-(e^{-i\theta}+e^{i\theta})^{k+1} + \sum_{\ell=-k}^{k} a_\ell e^{i\ell\theta}), \tag{6.5.11}$$

with numbers $c, a_{-k}, \dots, a_k$ independent of ω, since $\varphi + \overline{\varphi} = \frac{(-2\xi)^n}{(2n-1)!}$ From the last expression, we see by induction that $R'_n(k+1)$ is a polynom of ξ and cannot be of degree greater than $n+k+1$. That means, with (6.5.7), that the C_j are polynomials of ξ of degree less than $n+j+1$. □

By using another decomposition, we obtain the converse :

Lemma 6.5.10. *The C_j are polynomial in ξ of degree greater or equal than $n+1+j$ (or are null).*

Proof. We first prove by induction that $|\varrho_{2n}^{\theta}|^2$ can be decomposed in sums and products of monoms of the form $\xi(z^2-k^2)$, where $k=0,\dots,n$. So, with (6.5.3), we compute :

$$|\varrho_{2n}^{\theta}(z)|^2 = |\varrho_{2n-2}^{\theta}(z)|^2 + S_n(z),$$

where $S_n(z)$ is a linear combination of terms which are products of such monoms together with products of

$$z(e^{i\theta}-1)(z-k)(e^{-i\theta}-1)(z-\ell). \tag{6.5.12}$$

with integers k,ℓ. Since $|e^{i\theta}-1|^2 = \xi^2+\xi(2-\xi) = 2\xi$, the preceding term reads

$$-2\xi z^2(k+\ell)$$

together with odd terms which are simplified in S_n, since ϱ_{2n}^{θ} is even, thus we have such a decomposition. Now, a product of monoms of the form $\xi(z^2-k^2)$ can be expressed in the basis $1, z^2, z^2(z^2-1),\dots,z^2\dots(z^2-n^2), z^2\dots(z^2-n^2)(z^2-1),\dots,z^2\dots(z^2-n^2)(z^2-1)\dots(z^2-(n-1)^2)$ and its coordinate along a vector $z^2\dots(z^2-n^2)(z^2-1)\dots(z^2-k^2)$ is either null or of degree greater or equal than $n+1+k$, if the product contains at least $n+1+k$ such monoms. We have just seen that $|\varrho_{2n}^{\theta}|^2$ is a sum of products of such monoms ; so, the lemma follows. □

Now, from the two preceding lemma, we have the existence of numbers c_j such that $C_j = c_j\xi^{n+1+j}$; it remains to prove that these c_j are positive. We could deduce it from the real case ($\theta=\pi$), but in fact we can also have an explicit formula for the c_j.

Lemma 6.5.11. *The c_j are positive numbers ; they are given by*

$$c_j = \frac{2^{n+1+j}}{(2n)!(2j+1)!(n+j+1)}. \tag{6.5.13}$$

Proof. We obtain from (6.5.6) and (6.5.9) :

$$c_j = \frac{(-1)^{n+1+j}\Pi_{n+1+j}R_n'((j+1)^2)}{a_{j,n}},$$

where $\Pi_{n+1+j}R_n'((j+1)^2)$ stands for the coordinate of $R_n'((j+1)^2)$ over ξ^{n+1+j}, and

$$\begin{aligned} a_{j,n} &:= \prod_{k=0}^{j}((j+1)^2-k^2)^2 \prod_{\ell=j+2}^{n}((j+1)^2-\ell^2) \\ &= \frac{(-1)^{n-j-1}(2j+1)!(n-j-1)!(n+j+1)!}{2j+2}. \end{aligned}$$

On the other hand, from (6.5.9), we deduce that

$$\Pi_{n+1+j}R_n'((j+1)^2) = \frac{w'(j)2^{n+1+j}(-1)^{n+1+j}}{(2n)!2(j+1)}.$$

Now, we compute :

$$w'(j) = (-1)^{n-1-j}(j+n)(n-1-j)!(n-1+j)!$$

By collecting all together, we obtain the expression of c_j. □

So, the theorem 6.5.5 is proven and thus theorem 6.5.3 also. We deduce now theorem 6.5.4 from theorem 6.5.3.

Proof. With the notations of theorems 6.5.3 and 6.5.4, we have, from Neville's algorithm :

$$\varrho^{\theta}_{2n+1}(\zeta) = \frac{(\zeta+n+1/2)\varrho^{\theta}_{2n}(\zeta+1/2) - (\zeta-n-1/2)\varrho^{\theta}_{2n}(\zeta-1/2)}{2n+1}$$

So, if $\zeta \in [-1/2, 1/2]$, we obtain from theorem 6.5.3 : $|\varrho^{\theta}_{2n+1}(\zeta)| \leq 1$. □

On the other hand, we can also obtain a formula, for the odd case :

Theorem 6.5.12. *Let $n \in \mathbb{N}^*$, $\theta \in \mathbb{R}$ and $\varrho^{\theta}_{2n-1} \in \mathbb{C}_{2n-1}[X]$ such that $\varrho^{\theta}_{2n-1}(j) = \exp(ij\theta)$ for $j = -n+1/2, \ldots, n-1/2$. By introducing $\xi := 1-\cos(\theta)$, we then obtain :*

$$|\varrho^{\theta}_{2n-1}(\zeta)|^2 = 1 - \left((1/2)^2-\zeta^2\right)\ldots\left(((2n-1)/2)^2-\zeta^2\right)\xi^n \\ \left[d_0 + d_1\xi\left((1/2)^2-\zeta^2\right) + \ldots + d_{n-1}\xi^{n-1}\left((1/2)^2-\zeta^2\right)\ldots\left(((2n-3)/2)^2-\zeta^2\right)\right]$$

with positive numbers d_i, $i = 0, \ldots n-1$, which are given by

$$d_i = \frac{2^{n+i}}{(2n-1)!(2i)!(n+i)}$$

Proof. Following [70], in a first time we will compute the derivative of the modulus of ϱ^{θ}_{2n-1} by expressing it in function of the real part $\mathcal{R}$ and the imaginary part $\mathcal{I}$ of ϱ^{θ}_{2n-1}. To be more precise we compute $\partial_\theta|\varrho^{\theta}_{2n-1}|^2(\zeta) = 2\left(\mathcal{I}\partial_\theta\mathcal{I} + \mathcal{R}\partial_\theta\mathcal{R}\right)(\zeta,\theta)$ where

$$\mathcal{R}(\zeta,\theta) = \sum_{k=-n+1/2}^{n-1/2} \ell_k(\zeta)\cos(k\theta), \quad \mathcal{I}(\zeta,\theta) = \sum_{k=-n+1/2}^{n-1/2} \ell_k(\zeta)\sin(k\theta),$$

and

$$\ell_k(\zeta) = \prod_{\substack{i=-n+1/2 \\ i\neq k}}^{n-1/2} \frac{(\zeta-i)}{(k-i)}$$

Let us compute $\mathcal{R}$, $\mathcal{I}$, $\partial_\theta\mathcal{R}$ and $\partial_\theta\mathcal{I}$. If we express the Lagrange interpolation polynomial in the Newton basis and using the notation

$$\varpi_{n-1/2}(\zeta) = (\zeta^2-(1/2)^2)\ldots(\zeta^2-(n-1/2)^2)$$

we get for $\zeta \in [0, 1/2]$ and $\theta \in [0, 2\pi]$

$$\mathcal{R}(\zeta,\theta) = \mathcal{R}_{-1/2} + (\zeta+1/2)\Delta\mathcal{R}_{-1/2} + \frac{\varpi_{1/2}(\zeta)}{2!}\Delta^2\mathcal{R}_{-3/2} + \cdots \\ + \frac{\varpi_{n-3/2}(\zeta)(\zeta+n-1/2)}{(2n-1)!}\Delta^{2n-1}\mathcal{R}_{-n+1/2}$$

and

$$\mathcal{I}(\zeta,\theta) = \mathcal{I}_{-1/2} + (\zeta+1/2)\Delta\mathcal{R}_{-1/2} + \frac{\varpi_{1/2}(\zeta)}{2!}\Delta^2\mathcal{I}_{-3/2} + \cdots + \frac{\varpi_{n-3/2}(\zeta)(\zeta+n-1/2)}{(2n-1)!}\Delta^{2n-1}\mathcal{I}_{-n+1/2}.$$

Let $f(\zeta)$ be an even function, let us show that $\Delta^{2n-1}f_{-n+1/2} = 0$:

$$\begin{aligned}
\Delta^{2n-1}f_{-n+1/2} &= \sum_{l=0}^{2n-1}\binom{2n-1}{l}(-1)^{2n-1-l}f_{l-n+1/2} \\
&= \sum_{l=0}^{n-1}\binom{2n-1}{l}(-1)^{2n-1-l}f_{l-n+1/2} \\
&+ \sum_{l=n}^{2n-1}\binom{2n-1}{l}(-1)^{2n-1-l}f_{l-n+1/2} \\
&= \sum_{l=0}^{n-1}\binom{2n-1}{l}(-1)^{2n-1-l}f_{l-n+1/2} \\
&+ \sum_{l=0}^{n-1}\binom{2n-1}{2n-l-1}(-1)^{l}f_{n-l-1/2} \\
&= \sum_{l=0}^{n-1}\left(\begin{array}{c}2n-1\\ l\end{array}\right)(-1)^l\left(f_{k-1/2-l} - f_{l+1/2-k}\right) = 0.
\end{aligned}$$

Let us compute $\Delta^{2k-2}\mathcal{R}_{-k+1/2}$.

$$\begin{aligned}
\Delta^{2k-2}\mathcal{R}_{-k+1/2} &= \sum_{l=0}^{2k-2}\left(\begin{array}{c}2k-2\\ l\end{array}\right)(-1)^{2k-2-l}\mathcal{R}_{l-k+1/2} \\
&= \sum_{l=0}^{2k-2}\left(\begin{array}{c}2k-2\\ l\end{array}\right)(-1)^{2k-2-l}\sum_{j=-n+1/2}^{n-1/2}\ell_j(l-k+1/2)\cos(j\theta) \\
&= \Re e\left\{e^{i(-k+1/2)\theta}\sum_{l=0}^{2k-2}\left(\begin{array}{c}2k-2\\ l\end{array}\right)(-1)^{2k-2-l}e^{il\theta}\right\} \\
&= (-1)^{k-1}\cos(\theta/2)2^{2k-2}(\sin(\theta/2))^{2k-2}.
\end{aligned}$$

We deduce that $\mathcal{R}(\zeta,\theta)$ can be rewritten as

$$\mathcal{R}(\zeta,\theta) = \cos(\theta/2)\left[1 + \sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)(-1)^k 2^{2k}\sin^{2k}(\theta/2)}{(2k)!}\right]$$

Doing the same kind of computation for $\Delta^{2k-2}\mathcal{I}_{-k+1/2}$ and $\Delta^{2k-1}\mathcal{I}_{-k+1/2}$ we obtain the

following expression for $\mathcal{I}$.

$$\begin{aligned}\mathcal{I}(\zeta,\theta) &= -\sin(\theta/2)+2(\zeta+1/2)\sin(\theta/2)+\sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k}(-1)^{k+1}\sin^{2k+1}(\theta/2)}{(2k)!}\\ &+ \sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)(\zeta+k+1/2)2^{2k+1}(-1)^{k}\sin^{2k+1}(\theta/2)}{(2k+1)!}\\ &= \zeta\sum_{k=0}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k+1}(-1)^{k}\sin^{2k+1}(\theta/2)}{(2k+1)!}\end{aligned}$$

Now let us differentiate $\mathcal{I}$ with respect to θ :

$$\begin{aligned}\partial_\theta\mathcal{I}(\zeta,\theta) &= -\frac{1}{2}\cos(\theta/2)+(\zeta+1/2)\cos(\theta/2)\\ &+\sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k-1}(-1)^{k+1}(2k+1)\sin^{2k}(\theta/2)\cos(\theta/2)}{(2k)!}\\ &+\sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)(\zeta+k+1/2)2^{2k}(-1)^{k}(2k+1)\sin^{2k}(\theta/2)\cos(\theta/2)}{(2k+1)!}\\ &= \zeta\cos\left(\frac{\theta}{2}\right)+\sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k}(-1)^{k}\sin^{2k}(\frac{\theta}{2})}{(2k+1)!}\\ &(-(1/2)(2k+1)\cos(\theta/2)+\cos(\theta/2)(\zeta+k+1/2))\\ &= \zeta\cos(\theta/2)\left[1+\sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k}(-1)^{k}\sin^{2k}(\theta/2)}{(2k)!}\right]=\zeta\mathcal{R}\end{aligned}$$

Now let us differentiating $\mathcal{R}$ with respect to θ.

$$\begin{aligned}\partial_\theta\mathcal{R}(\zeta,\theta) &= -\frac{\sin(\frac{\theta}{2})}{2}\sum_{k=0}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k}(-1)^{k}\sin^{2k}(\frac{\theta}{2})}{(2k)!}\\ &+\cos\left(\frac{\theta}{2}\right)\sum_{k=0}^{n-1}\frac{\varpi_{k-1/2}(\zeta)2^{2k}(-1)^{k}k\sin^{2k-1}(\frac{\theta}{2})\cos(\frac{\theta}{2})}{(2k)!}\end{aligned}$$

In the second sum of the previous relation we replace $\cos^2(\theta/2)$
by $(1-\sin^2(\theta/2))$. Then we rewrite the sum in k from 0 to $n-2$.

Then we get

$$\begin{aligned}\partial_\theta \mathcal{R}(\zeta,\theta) = \frac{\varpi_{n-3/2}(\zeta)(n-1/2)^2 2^{2n-1}(-1)^n \sin^{2n-1}(\theta/2)}{(2n-1)!} \\ + \sum_{k=0}^{n-2} \frac{\varpi_{k-1/2}(\zeta) 2^{2k-1}(-1)^{k+1} \sin^{2k+1}(\theta/2)}{(2k)!} \\ + \sum_{k=0}^{n-2} \frac{\varpi_{k-1/2}(\zeta) 2^{2k}(-1)^{k+1} k \sin^{2k+1}(\theta/2)}{(2k)!} \\ + \sum_{k=0}^{n-2} \frac{\varpi_{k-1/2}(\zeta) 2^{2k+1}(-1)^{k+1} \sin^{2k+1}(\theta/2)}{(2k+1)!}\end{aligned}$$

Reducing all the sums we obtain

$$\begin{aligned}\partial_\theta \mathcal{R}(\zeta,\theta) &= -\frac{\varpi_{n-3/2}(\zeta)(n-1/2)^2 2^{2n-1}(-1)^{n+1} \sin^{2n-1}(\theta/2)}{(2n-1)!} \\ &+ \sum_{k=0}^{n-2} \frac{\varpi_{k-1/2}(\zeta) 2^{2k}(-1)^k k \sin^{2k+1}(\theta/2)}{(2k)!} \overbrace{\left(-\left(k+\frac{1}{2}\right) - \frac{2(\zeta^2-(k+\frac{1}{2})^2)}{2k+1}\right)}^{-\frac{2\zeta^2}{2k+1}} \\ &= \frac{\varpi_{n-1/2}(\zeta) 2^{2n-1}(-1)^{n+1} \sin^{2n-1}(\theta/2)}{(2n-1)!} - \zeta\mathcal{I}\end{aligned}$$

Collecting the expression of $\mathcal{I}$, $\mathcal{R}$, $\partial_\theta\mathcal{I}$ and $\partial_\theta\mathcal{R}$ we obtain

$$\begin{aligned}\partial_\theta |\varrho^\theta_{2n_1}(\zeta)|^2 = -2^{2n}\frac{(-1)^n \varpi_{n-1/2}(\zeta)}{(2n-1)!} \sin^{2n-1}\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right) \\ \left[1 + \sum_{k=1}^{n-1} \frac{\varpi_{k-1/2}(\zeta)(-1)^k 2^{2k}\sin^{2k}(\theta/2)}{(2k)!}\right]\end{aligned}$$

If we integrate the previous expression we get

$$\begin{aligned}|\varrho^\theta_{2n-1}(\zeta)|^2 &= g(\zeta) + -2^{2n+1}\frac{(-1)^n}{2n!}\varpi_{n-1/2}(\zeta)\sin^{2n}\left(\frac{\theta}{2}\right) \\ &- \frac{(-1)^n \varpi_{n-1/2}(\zeta)}{(2n-1)!}\sum_{k=1}^{n-1}\frac{\varpi_{k-1/2}(\zeta)(-1)^k 2^{2(k+n)+1}\sin^{2(k+l)}(\theta/2)}{(2k)!2(k+l)}\end{aligned}$$

As $|\varrho_{2n-1}(\zeta,\theta=0)|^2 = 1$ we get $g(\zeta)=1$ and finally by setting $\xi = 1-\cos\theta = 2\sin^2(\theta/2)$ we find the desired result

$$|\varrho^\theta_{2n-1}(\zeta)|^2 = 1 - (-1)^n \varpi_{n-1/2}(\zeta)\xi^n\left(\sum_{k=0}^{n-1}\frac{2^{k+n+1}(-1)^k\xi^k\varpi_{k-1/2}(\zeta)}{(2n-1)!(2k)!(2(k+n))}\right)$$

□

6.5.3 Another stability region ?

We can ask us whether there exist other regions where the modulus $|\varrho_{2n}^{\theta}|$ is less than 1. In fact, by taking $\theta = \pi$, we see on figure 6.2 that at many places, we have $|\varrho_{2n}^{\pi}| > 1$ and this modulus can take very high values. On the other hand, for some values of z,

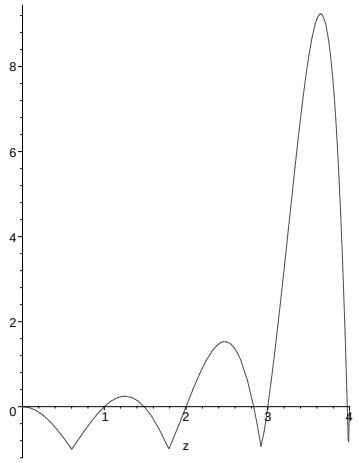


Fig. 6.2 – Graph of $|\varrho_8(z)| - 1$ (where ϱ_8 is the polynom of degree 8 such that $\varrho_8(j) = (-1)^j$, $j = -4 \dots 4$), for $z = 0, \dots, 4$.

which are a little smaller than the integers, the modulus $|\varrho_{2n}^{\theta}(z)|$ can be less than 1 for many values of θ and is not far from 1 for the other values : see on figure 6.3.

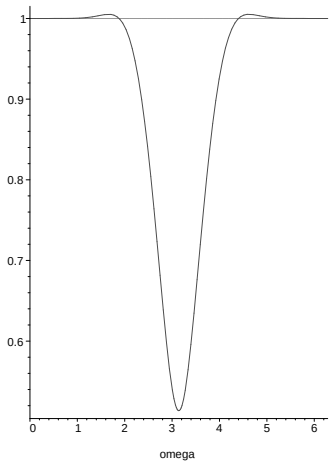


Fig. 6.3 – Graph of $|\varrho_8^{\theta}(1.9)|$ (where ϱ_8^{θ} is the polynomial of degree 8 such that $\varrho_8(j) = \exp(ij\omega)$, $j = -4 \dots 4$), for $\theta = 0, \dots, 2\pi$.

In [33], we can see some $3D$ plots in θ and ζ which give the *stability zone* numerically for small values of n.
Thanks to formula (6.5.5) and (6.5.12), we can however prove the following conjectures,

which permit to determine the optimality of the *stability region* for numerical purposes (*computer proof* done for $n \leq 82$ by Steinberg and Hong),

Conjecture 6.5.13 (Steinberg-Hong, even case). *Define $n \in \mathbb{N}$, $\theta \in \mathbb{R}$ and $\varrho_{2n}^{\theta} \in \mathbb{C}_{2n}[X]$ such that $\varrho_{2n}^{\theta}(j) = \exp(ij\theta)$ for $j = -n, \ldots, n$. Then, if $\zeta \notin [-1, 1]$ (and ζ is not an integer between $-n$ and n), there exists $\theta \in \mathbb{R}$ such that*

$$|\varrho_{2n}^{\theta}(\zeta)| > 1$$

Conjecture 6.5.14 (Steinberg-Hong, odd case). *Define $n \in \mathbb{N}$, $\theta \in \mathbb{R}$ and $\varrho_{2n-1}^{\theta} \in \mathbb{C}_{2n-1}[X]$ such that $\varrho_{2n-1}^{\theta}(j) = \exp(ij\theta)$ for $j = -n + 1/2, \ldots, n - 1/2$. Then, if $\zeta \notin [-1/2, 1/2]$ (and $\zeta \notin \{i + 1/2\}$ where i is an integer between $-n$ and $n - 1$), there exists $\theta \in \mathbb{R}$ such that*

$$|\varrho_{2n-1}^{\theta}(\zeta)| > 1$$

In the even case, they choose θ such that

$$z^2(1 - \cos\omega) = 8$$

This is possible for $|z| \geq 2$ (but for $|z|$ between 1 and 2, we easily see that from 6.5.5 that for $\theta \neq 0$ near 0, we have $|\varrho_{2n}^{\theta}(z)| > 1$). So it suffices to prove the following conjecture :

Conjecture 6.5.15. *Let $n \geq 3$ and $|z| \geq 2$; we then have :*

$$r := \sum_{k=0}^{n-1} \frac{2^{4k} z^{2(n-k-1)}}{(n+1+k)(2k+1)! \prod_{m=1}^{k}(m^2 - z^2)} < 0$$

For $n = 3$, they obtained :

$$r = -\frac{13}{180}(16 - z^2)(z^2 - \frac{16}{13})$$

and for $3 < n \leq 82$, they make the change $b = z^2 - 2^2$ and observed that r, as polynomial in b has only negative coefficients.

Remark 6.5.16. *The fine analysis done for the Lagrange reconstruction on uniform grid doesn't apply if we apply the same scheme on unstructured meshes (triangulation) by using high-order Lagrange interpolation on triangle, and the determination of the stability region in that case remains an interesting open question.*

6.6 B-splines interpolation

6.6.1 Introduction

Let us introduce the well known B-spline of order m recursively defined by

$$B_m(\cdot) = \int B_{m-1}(u - \cdot)B_0(u)\mathrm{d}u, \quad m \geq 1 \tag{6.6.1}$$

with $B_0(u) = 1$, if $-\frac{1}{2} \leq u \leq \frac{1}{2}$ and $B_0(u) = 0$ elsewhere. Now, fix a real θ and consider the following sum :

$$\sum_{k\in\mathbb{Z}} B_m(k+\alpha)e^{ik\theta}$$

What can we say about its L_∞-norm in α ?
In fact, the $B-$splines enjoy the following property :

Theorem 6.6.1. *Let $\theta \in \mathbb{R}$ and define $\Phi_m(\alpha) = |\sum_{k\in\mathbb{Z}} B_m(k+\alpha)e^{ik\theta}|^2$. Then, Φ_m admits its maximum on the integers.*

A similar problem was treated in 1976 by De Boor [22], and in 1991, this problem was solved in a unified context [51], by deriving geometric information from the *Euler spline curves* defined by $\alpha \to \Phi_m(\alpha, \theta)$; the authors of [51] gave also other monoticity properties of the symbol Φ_m, generalizing some results of Schoenberg obtained in 1969. We also mention the use of such a symbol in [60] and in [3]. We give here a new proof of theorem 6.6.1 by exploiting informations about the successive derivatives of Φ_m.

6.6.2 Proof

We know that B_m is infinitely differentiable, except at the points $\frac{1}{2} + k$, with $k = -m, \ldots, m-1$ if m is even and at the points $k = -m, \ldots, m$ if m is odd, where the function is only $m-1$ times differentiable. We will widely use the relation

$$B_{m+1}{}'(x) = B_m(x + \frac{1}{2}) - B_m(x - \frac{1}{2}). \tag{6.6.2}$$

In particular, we have $B_1{}'(0^+) = -1$ and $B_1'(-1^+) = 1$.
The useful properties of Φ_m are summarized in the following lemma.

Lemma 6.6.2. *Φ_m is :*

- *$m-1$ times differentiable*
- *infinitely differentiable in $]0,1[$ if m is even, and in $]-\frac{1}{2},\frac{1}{2}[$ if m is odd.*

- $1-$*periodic,*
- *even and* $\Phi(\frac{1}{2}+\cdot)$ *is even.*
- $\Phi'_m(x) = (1-e^{i\theta})\Phi_{m-1}(x-\frac{1}{2})$

Proof. direct from the properties of the $B-$splines. □

Now, we will use the notation $^{[p]}$ for the derivative or order p.

Lemma 6.6.3. *Let* p, q, m *such that* : $p > m$ *or* $q > m$ *or*

$$0 \leq p < m \quad 0 \leq q < m \quad \textit{and } p+q \textit{ odd,}$$

then,

$$\sum_{k,\ell} B_m^{[p]}(k) B_m^{[q]}(\ell) \cos(k-\ell)\theta = 0.$$

Proof. If $p > m$, then $B_m^{[p]} = 0$ and the result follows (similarly if $q > m$). Now, we can assume that

$$m > p > q \geq 0.$$

We prove then the result by induction on p. If $p = q+1$, we use that the derivative of the function

$$|\sum_{k\in\mathbb{Z}} B_m^{[p]}(k+\alpha) e^{ik\theta}|^2$$

is null for $\alpha = 0$, since the latter function is even. For the general case, we use the decomposition :

$$(B_m^{[q]} B_m^{[q]})^{[p-q]} = B_m^{[p]} B_m^{[q]} + \sum_{r=1}^{p-q} \begin{pmatrix} p-q \\ r \end{pmatrix} B_m^{[p-r]} B_m^{[q+r]}.$$

By taking the corresponding sum, the term on the left vanishes because $p-q$ is odd, and on the other hand, the terms on the left, except

$$\sum_{k,\ell} B_m^{[p]}(k) B_m^{[q]}(\ell) \cos(k-\ell)\theta.$$

are also null by induction, and so the latter term is also null and the result is established. □

Lemma 6.6.4. *If* $0 \leq \alpha \leq \frac{1}{2}$, $m \geq 0$ *and* $-\pi \leq \theta \leq \pi$, *then*

$$\phi_{m,\frac{1}{2}}(\alpha) := \sum_{k\in\mathbb{Z}} B_m(k+\alpha+\frac{1}{2}) \cos(k+\frac{1}{2})\theta > 0,$$

and

$$\phi_{m,0}(\alpha) := \sum_{k\in\mathbb{Z}} B_m(k+\alpha) \cos(k\theta) > 0$$

Proof. We prove this fact by induction. For $m = 0$, the result is true, since $\phi_{0,0}(\alpha) = 1 > 0$ (resp. $1 + \cos(\theta) > 0$), if $0 \leq \alpha < 1$ (resp. $\alpha = 1$) and $\phi_{0,\frac{1}{2}}(\alpha) = \cos(\theta/2) > 0$ (resp. $2\cos(\theta/2) > 0$), if $0 < \alpha \leq \frac{1}{2}$ (resp. $\alpha = 0$).
Now, for $m \geq 1$, we have :

$$\begin{aligned}\phi_{m,0}(\alpha) &= \sum_{k\in\mathbb{Z}} \int_{\alpha-1/2}^{\alpha+1/2} B_{m-1}(k+\alpha+u)\cos(k\theta)\mathrm{d}u \\ &= \int_{-\alpha+1/2}^{\alpha+1/2} \cdots + \int_{\alpha-1/2}^{-\alpha+1/2} \ldots\end{aligned}$$

Now, we have :

$$\int_{\alpha-1/2}^{-\alpha+1/2} \cdots = 2\int_0^{1/2-\alpha} \phi_{m-1,0}(k+u)\mathrm{d}u > 0,$$

by induction, and

$$\int_{-\alpha+1/2}^{0} \cdots = \int_0^{\alpha} \sum_{k\in\mathbb{Z}} B_{m-1}(k+\frac{1}{2}+u)\mathrm{d}u\cos(k+1)\theta.$$

So, by putting together, we obtain that

$$\int_{-\alpha+1/2}^{\alpha+1/2} \cdots = 2\int_0^{\alpha} \phi_{m-1,1/2}(k+u)\mathrm{d}u\cos(\theta/2) > 0,$$

which gives the result for $\phi_{m,0}$.
We now look similarly at $\phi_{m,\frac{1}{2}}(\alpha)$,

$$\phi_{m,\frac{1}{2}}(\alpha) = \int_{-\alpha+1/2}^{\alpha+1/2} \cdots + \int_{\alpha-1/2}^{-\alpha+1/2} \ldots,$$

which yields :

$$\int_{\alpha-1/2}^{-\alpha+1/2} \cdots = 2\int_0^{1/2-\alpha} \phi_{m-1,\frac{1}{2}}(k+u)\mathrm{d}u > 0,$$

and :

$$\int_{-\alpha+1/2}^{\alpha+1/2} \cdots = 2\int_0^{\alpha} \phi_{m-1,0}(k+u)\mathrm{d}u\cos(\theta/2) > 0,$$

and so :

$$\phi_{m,\frac{1}{2}}(\alpha) > 0.$$

□

Remark 6.6.5. *The* $B-$spline symbol $\phi_m := \phi_{m,0}(0)$ *has been widely used in [60]; the above inequality for* $\alpha = 0$ *has been proven in [66], and the proof of this lemma is an adaptation of proposition* 3.1 *in [51].*

Now, thanks to the lemma, we can express the other derivatives in term of the $B-$spline symbol.

Lemma 6.6.6. *For m, q such that $m+q$ is odd, we have the formula :*

$$\sum_{k,\ell\in\mathbb{Z}} B_m^{[m]}(k+\alpha)B_m^{[q]}(\ell+\alpha)\cos(k-\ell)\theta$$
$$= (\cos(\theta)-1)(-|1-e^{i\theta}|^2)^{(q+m-1)/2}\phi_{m-q,0}(0).$$

Proof. We consider

$$A := \sum_{k,\ell\in\mathbb{Z}} B_m^{[m]}(k+\alpha)B_m^{[q]}(\ell+\alpha)\cos(k-\ell)\theta.$$

By algebraic manipulation, we obtain :

$$A = \sum_{k,\ell\in\mathbb{Z}} B_0(k+\alpha)B_{2m}^{[q+m]}(\ell+\alpha)\cos(k-\ell)\theta.$$

Yet,

$$A = (\cos(\theta)-1)(-|1-e^{i\theta}|^2)^{(q+m-1)/2}\phi_{m-q,0}(0).$$

□

From lemma 6.6.3, 6.6.4 and 6.6.6, we obtain that :

Lemma 6.6.7. *We have :*

$$\Phi_{2m+1}^{[2m+1+2\ell]}(0^+)(-1)^{\ell+m-1} > 0.$$

for $\ell = 0, \ldots, m-1$, and similarly :

$$\Phi_{2m}^{[2m+1+2\ell]}((-\frac{1}{2})^+)(-1)^{\ell+m} > 0,$$

for $\ell = 0, \ldots, m-1$.

We can now summarize the results and conclude to the proof of theorem 6.6.1.

Proof of theorem 6.6.1. We can yet derive the variations of the function Φ_m. It turns out that the even derivatives of Φ_m are alternatively decreasing on $[0, \frac{1}{2}]$ and increasing on $[\frac{1}{2}, 1]$ or increasing on $[0, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, 1]$ (see figure 6.4), in the case where m is odd (we can make a similar treatment, if m is even, by considering the interval $]-\frac{1}{2}, \frac{1}{2}[$). Indeed, we can obtain it by induction. More precisely, we see that $\Phi_m^{[2m]}$ is strictly positive and so, $\Phi_m^{[2m-2]}$ is decreasing on $[0, \frac{1}{2}]$ and increasing on $[\frac{1}{2}, 1]$, since $\Phi_m^{[2m-1]}$ is increasing on $]0, 1[$ and $\Phi_m^{[2m-1]}(\frac{1}{2}) = 0$ (we recall that $\Phi_m(\frac{1}{2}+\cdot)$ is even and infinitely differentiable

in $]0,1[$). Now suppose that, for a given ℓ, $\Phi_m^{[2m-2\ell]}$ is increasing on $]0,\frac{1}{2}[$ and decreasing on $]\frac{1}{2},1[$ (the alternating case is ommited and can be treated similarly). So, $\Phi_m^{[2m-2\ell-1]}$ is convex on $]0,\frac{1}{2}[$ and concave on $]\frac{1}{2},1[$. Since we then have $\Phi_m^{[2m-2\ell-1]}(0^+) \leq 0$, we conclude that $\Phi_m^{[2m-2\ell-1]}$ is strictly negative on $]0,\frac{1}{2}[$ and similarly strictly positive on $]\frac{1}{2},1[$, that is the monotocity alterns, as we had to show. As a consequence, by using the former induction, we get that Φ_m is decreasing on $]0,\frac{1}{2}[$ and increasing on $]\frac{1}{2},1[$, that means that Φ_m admits its maximum on the integers. □

We mention here that we have needed another property of the B–splines, discovered in [51], namely :

Lemma 6.6.8. *We define* $\phi(\alpha,\theta) := |\sum_{k\in\mathbb{Z}} B_{m+1}(k+\alpha)e^{i\theta}|$, *then we have*

$$0 < \phi(0,\pi) \leq \phi(0,\theta) = \phi(0,0) = 1, \quad \theta \in \mathbb{R}.$$

Proof. The proof can be found in [51], the lemma 6.6.4 can also be useful to prove it. □

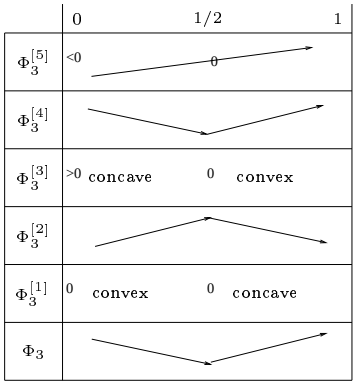


Fig. 6.4 – Variation of Φ_m for $m = 3$

Chapitre 7

Convergence d'un schéma adaptatif pour le système de Vlasov-Poisson en dimension 1

En se basant sur l'étude précédemment développée par Nicolas Besse [16], où la convergence d'un schéma numérique semi-Lagrangien a été prouvée, pour un maillage non structuré, on montre ici la convergence d'un schéma adaptatif, où le maillage évolue dans le temps.

Collaboration. L'étude entreprise dans ce chapitre, ainsi que dans le chapitre 5 contribuent à un article en préparation en collaboration avec Martin Campos Pinto.

7.1 Introduction

On s'intéresse ici à l'analyse numérique d'une méthode adaptative pour la résolution numérique de l'équation de Vlasov, basée sur une méthode semi-Lagrangienne. Plusieurs auteurs se sont penchés sur la convergence de méthodes numériques pour le système de Vlasov-Poisson ou de Vlasov-Poisson-Fokker-Planck. Ainsi P.-A. Raviart et G.-H. Cottet [28] ont présenté une analyse mathématique d'une méthode particulaire pour résoudre le système de Vlasov-Poisson en une dimension. S. Wollman a poursuivi cette étude en dimension une [76], puis en dimension trois [77]. J. Schaeffer a prouvé la convergence d'un schéma aux différences finies pour le système de Vlasov-Poisson-Fokker-Planck, en dimension une. Francis Filbet [35] a montré la convergence d'un schéma de volumes finis pour le système de Vlasov-Poisson, et Nicolas Besse [16] a montré la convergence de plusieurs classes de schémas semi-Lagrangiens (on renvoie à l'introduction du chapitre 6 pour plus de détails).

Concernant les méthodes adaptatives, des schémas adaptatifs ont été développés dans un contexte d'éléments finis, dès la fin des années 1970 (voir par exemple Babushka et Rheinboldt [4]). Ils ont été intensément développés dans un contexte de différences finies, d'éléments finis et de volumes finis (voir par exemple Verfürth [75], Erikkson, Estep, Hansbo et Johnson [31] et aussi Becker et Rannacher [6]). Ces algorithmes sont basés sur des *estimateurs d'erreurs a posteriori*, qui consistent à dériver des indicateurs locaux d'erreurs à partir de la solution courante calculée. Ces indicateurs sont alors utilisés pour raffiner localement ou déraffiner la discrétisation au pas de calcul suivant. Néanmoins, des preuves de convergences sont apparues plus tardivement (voir Doerfler [30] pour l'équation de Poisson, et aussi plus récemment Morin, Nocetto et Siebert [57]). Plus spécifiquement, pour les problèmes d'évolution, une difficulté importante est que les singularités (ou les zones de fort gradient) peuvent non seulement apparaître, mais aussi bouger en fur et à mesure que le temps avance, et donc le maillage adaptatif doit être rafraîchi à chaque étape de temps. Une possibilité est alors d'utiliser des maillages mobiles, ou d'adapter le maillage en changeant de métrique (voir [2]). Une autre possibilité, est d'utiliser des méthodes *multirésolution*, qui sont apparues dans ce contexte dès les années 1980. L'idée principale est d'utiliser une hiérachie de grilles fixes emboîtées, de différents niveaux de résolution, et qui offrent donc la possibilité de sélectionner localement le niveau de discrétisation approprié. Dans le contexte de schémas de volumes finis appliqués à la dynamique des fluides, cette approche a été ainsi développée avec les travaux de Berger, Oliger, M. et P. Collela [8], [9]. Des schémas numériques basés sur les ondelettes ont été développés au début des années 1990, en suivant l'idée empirique que les indicateurs locaux d'erreurs sont directement donnés par la taille des coefficients d'ondelettes : un grand coefficient indique des fluctuations importantes de la solution sur le support de l'ondelette et suggère donc de raffiner l'approximation en rajoutant des ondelettes à des niveaux plus fins dans cette région. Ainsi, dans ce contexte des problèmes d'évolution, la première méthode adaptative utilisant des ondelettes et basée sur l'idée qu'un indicateur pour le pas de temps suivant est donné par la taille des coefficients d'ondelettes de la solution numérique courante a été proposée par Maday, Perrier et Ravel [54] (voir aussi les travaux de Bertoluzza [12], [13]). Actuellement, les méthodes adaptatives se sont développées dans de nombreux contextes. Les discrétisations plus classiques en analyse numérique sont données par des valeurs aux points, des moyennes sur des cellules ou des espaces d'éléments finis. Il est alors possible d'adapter une décomposition multi-échelle à ce genre de discrétisation et de dériver des algorithmes adaptatifs pour des problèmes d'évolution (voir Harten [43]). Ainsi, dans le contexte des volumes finis (discrétisation par des moyennes sur les cellules), Cohen, Kaber, Mueller et Postel [25] ont obtenu la convergence d'un schéma adaptatif.
En ce qui concerne le système de Vlasov-Poisson, d'après nos connaissances, il n'existe pas de résultat de convergence pour un schéma adaptatif. Après avoir étudié numériquement une méthode adaptative basée sur une interpolation par éléments finis biquadratiques hiérarchiques (voir chapitre 8), on démontre ici la convergence d'un schéma adaptatif dans un cadre plus simplifié : la reconstruction est supposée affine par maille (obtenue en coupant chaque carré dyadique en deux), et on utilise un splitting en temps de l'advection. Le plan de ce chapitre est alors le suivant. Les hypothèses sur la fonction de distribution sont donnés-ci après. On décrit ensuite le maillage, puis l'espace d'approximation. On donne ensuite des résultats préliminaires sur les erreurs de projection, puis sur l'évolution

de certaines quantités par le transport. Une stratégie de prédiction de maillage est par la suite établie. Puis, on donne des résultats de stabilité. L'algorithme proprement dit est donné en dernière partie ; on décompose ensuite l'erreur, en utilisant aussi des résultats de la partie 5 pour la discrétisation en temps. le théorème de convergence est alors énoncé en conclusion.

On suppose ici que la donnée initiale est positive et dans $W^{2,1} \cap W^{1,\infty}(\Omega)$, avec

$$\Omega = [0, L] \times \mathbb{R},$$

à support compact en v et périodique en x, de période L.
La fonction de distribution solution de l'équation de Vlasov-Poisson décrite au chapitre 5 est L–périodique et reste à support compact en vitesse sur un intervalle de temps $[0, T]$.
On rappelle que l'on a une majoration de son support qui est donnée par

$$Q(f(t)) \leq Q(t) := Q(0) + 2Lt. \tag{7.1.1}$$

On fait une discrétisation en temps

$$0 = t^0 < t^1 < \cdots < t^N = T \quad N \in \mathbb{N}$$

de l'intervalle $[0, T]$, avec $\Delta t = t^{n+1} - t^n$.

La solution $f(t^n)$ est approchée par une solution numérique f^n, qui est représentée à l'aide d'un maillage adaptatif $\mathcal{M}_n$ que l'on va préciser.

Domaine d'espace. Le *domaine physique* est alors donné par

$$\Omega_R = [0, L[\times[-R, R[,$$

avec R suffisament grand.
D'après la majoration précédente, on va prendre en fait :

$$R \geq Q(T), \tag{7.1.2}$$

on est alors assuré que $f(t, x, v)$ est nulle en dehors de $[0, T] \times \bar{\Omega}_R$, et donc il est suffisant de connaitre f sur Ω_R, en la prolongeant par zéro en v et par périodicité en x.
On fixe aussi une constante assez grande γ_T ; on considèrera alors des domaines du type

$$\Omega_r = [0, L[\times[-r, r[, \quad \frac{R}{\gamma_T} \leq r \leq \gamma_T R. \tag{7.1.3}$$

On utilisera des domaines numériques

$$\Omega^n := [0, L] \times [-r_n, r_n], \tag{7.1.4}$$

avec

$$r_n \geq Q(t^n) \quad r_n \leq \gamma_T R, \tag{7.1.5}$$

de sorte que $r = r_n$ vérifie (7.1.3), en supposant que $\gamma_T \geq 1/Q(0)$.
Pour simplifier, on prendra en fait $R = Q(T)$ et $r_n = Q(t^n)$.
On introduit aussi le *domaine de calcul* $[0,1[^2$, et on note θ l'application de changement de domaine :

$$\theta : \Omega_R \to [0,1[^2,$$

que l'on prolonge sur $\mathbb{R}^2$ en la définissant similairement sur les rectangles translatés qui forment une partition de $\mathbb{R}^2$.
Pour toute fonction g (du domaine physique), on considèrera $\hat{g}$ (du domaine de calcul) par la relation

$$\hat{g}(x,v) = g(\theta^{-1}(x,v)), \tag{7.1.6}$$

et réciproquement, pour toute fonction $\check{g}$, on définit $\check{g}$ par :

$$\check{g}(x,v) = g(\theta(x,v)), \tag{7.1.7}$$

g correspondant au domaine de calcul et $\check{g}$ au domaine physique.
Enfin, pour un ensemble $\omega \subset \mathbb{R}^2$, on définit $\hat{\omega}$ et $\check{\omega}$ par

$$\hat{\omega} = \theta(\omega) \quad \check{\omega} = \theta^{-1}(\omega), \tag{7.1.8}$$

$\hat{\omega}$ est à considérer dans le domaine de calcul, tandis que $\check{\omega}$ est à considérer dans le domaine physique.
Lorsque l'on considère r au lieu de R, on utilisera les notations θ_r et θ_r^{-1} ; mais pour ne pas alourdir les notations, on utilisera aussi les symboles $\hat{a}$ et $\check{a}$, a étant un élément de $\mathbb{R}^2$, un ensemble ou une fonction.

7.2 Description du maillage

Mailles. On définit d'abord les mailles dyadiques de niveau de résolution $j \in \mathbb{N}$

$$\mathcal{M}_{j,k,l} := [k\,2^{-j}, (k+1)\,2^{-j}[\times [l\,2^{-j}, (l+1)\,2^{-j}[,$$

pour tous $k, l \in \mathbb{Z}$.
On note également l'ensemble des mailles dyadiques de niveau j par :

$$\mathcal{M}_j := \{\, [k\,2^{-j}, (k+1)\,2^{-j}[\times [l\,2^{-j}, (l+1)\,2^{-j}[:\ k,\ l \in \mathbb{Z} \,\},$$

Etant donnée une maille (dyadique) α, on note $\ell(\alpha)$ son niveau, ou sa *résolution*. On fixe maintenant un entier $j_0 \geq 0$ le niveau de résolution grossier. On supposera toujours que $j \geq j_0$.

Arbre de mailles. Afin de décrire le maillage dyadique, il est classique d'introduire un *arbre de mailles*. Ainsi, on définit d'abord
les *enfants* $\mathcal{C}(\alpha)$ d'une maille α :

$$\mathcal{C}(\alpha) := \{\ \beta \in \mathcal{M}_{\ell(\alpha)+1}\ :\ \beta \subset \alpha\ \}, \tag{7.2.1}$$

et aussi les *ancêtres* $\mathcal{P}(\alpha)$ de cette même maille :

$$\mathcal{P}(\alpha) := \{\ \beta \in \cup_{j_0 \le i \le \ell(\alpha)-1} \mathcal{M}_i\ :\ \beta \supset \alpha\ \},$$

Un arbre Λ est alors un ensemble de mailles vérifiant

$$\mathcal{M}_{j_0} \subset \Lambda \quad \text{et} \quad \bigcup_{\beta \in \mathcal{P}(\alpha)} \mathcal{C}(\beta) \subset \Lambda \quad \text{pour tout } \alpha \in \Lambda,$$

Remarque 7.2.1. *Si on définit la* mère *d'une maille α de niveau $j > j_0$, comme étant l'unique élément de $\mathcal{M}_{j-1}$ contenant α, et les* soeurs *de α comme étant les enfants de la mère de α, excepté la maille α elle-même, cela signifie que toutes les mailles de niveau j_0 soient contenues dans Λ, et que toute maille de Λ de niveau $j > j_0$ doit contenir sa mère et ses soeurs (qui sont au nombre de trois).*

Maillage adaptatif. On définit alors ensuite une *feuille* d'un arbre Λ, comme étant une maille dont les enfants ne sont pas dans Λ.
Un *maillage adaptatif* $\mathcal{M}$ est alors l'ensemble des feuilles d'un arbre Λ.

Partition. Comme chaque maille d'un arbre n'est jamais partiellement raffinée, c'est-à-dire que tous ses enfants ou aucun de ses enfants sont dans l'arbre, et comme toutes les mailles de résolution grossière sont dans l'arbre, le maillage adaptatif forme une partition de $\mathbb{R}^2$.
On peut alors définir pour tout point de $\mathbb{R}^2$, son *niveau de maillage*, comme étant le niveau de l'unique maille qui la contient.

Remarque 7.2.2. *Pour tous les points dyadiques $0 < a \le 1$ (c'est-à-dire que chaque coordonnée est de la forme $k2^{-}j$ avec $k \in \mathbb{Z}$ et $j \in \mathbb{N}$), on peut aussi définir le* niveau logique *; il s'agit du plus petit entier j tel que $2^j a$ ait ses coordonnées entières et dont l'une est impaire ; on complète par périodicité pour les points dyadiques de $a \in \mathbb{R}$.*

Maillage gradué. Un maillage adaptatif est *gradué* si pour tout couple (α, β) formé par deux mailles *voisines* (i.e. qui partagent un même côté) ont une arête commune), on a :

$$|\ell(\alpha) - \ell(\beta)| \le 1, \tag{7.2.2}$$

c'est-à-dire que le long d'une ligne horizontale, ou verticale, le niveau de maillage ne peut pas augmenter ou diminuer de plus d'unité entre chaque changement de niveau.

Demi-mailles. Toute maille α d'un maillage adaptatif est composé de deux triangles obtenus en découpant la maille α suivant une diagonale du carré. Le choix de la diagonale se fait de la manière suivante.
On considère l'unique maille β de niveau $\ell(\alpha)-1$ qui contient la maille α, i. e. :

$$\beta \in \mathcal{M}_{\ell(\alpha)-1}, \quad \alpha \subset \beta \tag{7.2.3}$$

(il s' agit en fait de la mère de α si α n'est pas de résolution grossière). La diagonale considérée pour α est alors choisie comme celle qui forme une partie de l'ensemble formé par les deux diagonales de β.
Le triangle supérieur de la maille α se note alors α^+ et α^- désigne le triangle inférieur.
Lorsque l'on ne cherche pas à distinguer la demi-maille supérieure ni inférieure, on parlera de demi maille α^*.
On note également le niveau des demi-mailles :

$$\ell(\alpha^+) := \ell(\alpha^-) := \ell(\alpha^*) = \ell(\alpha). \tag{7.2.4}$$

Enfin, lorsqu'une demi-maille α^* possède un noeud contraint (voir définition ci-après, elle ne peut qu'en posséder un seul), elle le partage avec une unique autre demi-maille que l'on note α_*, et l'union $\alpha_* \cup \alpha^*$ sera noté α_*^* ; dans le cas où il n' y a pas de noeud contraint, on notera également $\alpha_*^* := \alpha^*$.

Domaine borné. Tout maillage adaptatif gradué que l'on considère est défini sur $[0,1[^2$, périodisé en x et prolongé en v de telle sorte que le maillage défini sur $\mathbb{R}^2$ soit adaptatif gradué (ceci peut donc demander de rajouter des mailles).

Notations.
On note :

$\mathcal{E}(\mathcal{M})$ l'ensemble des arêtes d'un maillage $\mathcal{M}$.

$\Lambda(\mathcal{M})$ l'arbre associé à $\mathcal{M}$.

$J(\mathcal{M})$ le niveau maximum des mailles de $\mathcal{M}$.

$|\alpha|$, $|\gamma|$ la longueur du côté d'une maille α, ou d'une arête γ.

$a \lesssim b$ signifie qu'il existe une constante C telle que $a \leq Cb$.

$\mathring{A}$ (resp. $\overline{A}$) désigne l'intérieur (resp. l'adhérence) d'un ensemble A.

Désormais, lorsque l'on parlera de maillage, ou de maillage adaptatif, on supposera toujours qu'il s'agit d'un maillage adaptatif gradué.

Noeuds.
On note encore :

$\Gamma(\alpha)$ l'ensemble des sommets d'une maille α.

$\Gamma(\mathcal{M}) := \cup_{\alpha\in\mathcal{M}}\Gamma(\alpha)$ l'ensemble des *noeuds* de $\mathcal{M}$.

7.3 Espace d'approximation

On note $C^0_{L,r}(\mathbb{R}^2)$, l'ensemble des fonctions continues sur $\mathbb{R}^2$, L–périodiques en x et nulles pour $|v| \geq r$. Etant donné un maillage $\mathcal{M}$, on considère l'ensemble de ces fonctions qui sont affines sur chaque maille triangulaire de $\mathcal{M}$:

$$V^r_{\mathcal{M}} := \{\ f \in C^0_{L,r}(\mathbb{R}^2)\ :\ f_{|\check{\alpha}^+} \text{ et } f_{|\check{\alpha}^-} \in \Pi_1\ \forall\alpha \in \mathcal{M}\ \}.$$

(Π_1 désigne l'ensemble des polynômes de degré inférieur ou égal à 1).
L'indice r en exposant signifie que $\check{\alpha}^+ := \theta_r^{-1}(\alpha^+)$ (et de même pour α^-).

Noeuds contraints et noeuds actifs. Pour un maillage $\mathcal{M}$, il peut exister un noeud qui est un point de l'adhérence d'une maille, sans être un noeud de celle-ci (cf figure 7.1). Un tel noeud est appelé *noeud contraint*.
Les autres noeuds sont appelés *noeuds actifs*. L'ensemble $\Gamma_a(\mathcal{M})$ des noeuds actifs peut aussi être défini par :

$$\Gamma_a(\mathcal{M}) := \{a \in \Gamma(\mathcal{M})\ :\ \text{ si } \beta \in \mathcal{M} \text{ satisfait } a \in \beta,\ \text{alors } a \in \Gamma(\beta)\}.$$

Pour un maillage $\mathcal{M}$ et une application f telle que $\hat{f}$ est définie sur tous les noeuds de $\mathcal{M}$, il existe une fonction g telle que $\hat{g}$ coincide avec $\hat{f}$ aux noeuds de $\mathcal{M}$ et vérifiant

$$g_{|\check{\alpha}^+} \text{ et } g_{|\check{\alpha}^-} \in \Pi_1\ \forall\alpha \in \mathcal{M}.$$

Néanmoins, g n'est pas forcément continue. Pour que g soit continue, il faut et suffit que $\hat{g}$ coincide avec $\hat{f}$ aux noeuds actifs de $\mathcal{M}$ et que la valeur de $\hat{g}$ à chaque noeud contraint $\check{a}$ soit égale à la valeur

$$\lim_{x\in\check{\alpha}\to\check{a}} g_{|\check{\alpha}}(x),$$

où α est l'unique maille sur laquelle a n'est pas un noeud ; cette valeur est en fait ici égale à la demi-somme des valeurs aux deux noeuds (actifs) de α, extrémités de l'arête contenant a.

Interpolation. On peut donc maintenant définir l'opérateur d'interpolation

$$P^r_{\mathcal{M}} : C^0_{L,r}(\mathbb{R}^2) \to V^r_{\mathcal{M}},$$

qui à une fonction continue f L–périodique en x et nulle pour $|v| \geq r$, associe l'unique élément de $V^r_{\mathcal{M}}$ vérifiant :

$$P^r_{\mathcal{M}} f(\check{a}) = f(\check{a}) \quad a \in \Gamma_a(\mathcal{M}). \tag{7.3.1}$$

La définition suivante prolonge cet opérateur à $C^0(\mathbb{R}^2)$:

$$P^r_{\mathcal{M}} : C^0(\mathbb{R}^2) \to C^0(\mathbb{R}^2), \quad f \to g$$

où g est l'unique élément de $C^0(\mathbb{R}^2)$ vérifiant :

$$g_{|\check{\alpha}^+} \text{ et } g_{|\check{\alpha}^-} \in \Pi_1 \ \forall \alpha \in \mathcal{M} \tag{7.3.2}$$

et (7.3.1).

Remarque 7.3.1. *Remarquons que l'on a ici (malgré la définition des noeuds contraints), une triangulation conforme : on a une partition du maillage en triangles dont les sommets ne sont jamais des noeuds contraints. Par contre, l'interpolation dans les mailles (carrées) ou les demi-mailles, elle, n'est pas conforme : on est amené à changer la valeur aux sommets du carré qui sont justement les noeuds contraints, pour avoir la continuité.*

Remarque 7.3.2. *La continuité ne semble pas nécessaire pour obtenir des résultats intéressants en pratique ; néanmoins, elle est ici utile dans notre analyse pour établir des bornes d'erreurs* a priori.

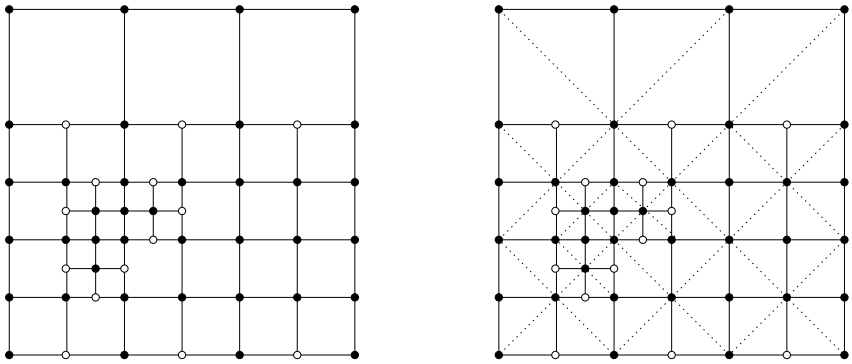


Fig. 7.1 – Un maillage (gradué, adaptatif) $\mathcal{M}$.
$\mathcal{M}_t$ représente $\mathcal{M}$ avec les demi-mailles triangulaires en pointillé.
Les noeuds contraints sont en blanc et les noeuds actifs en noir.

Fonctions affines par morceaux A chaque pas de temps, la fonction de distribution numérique est affine par morceaux sur une partition de l'espace des phases et elle est continue.
La solution transportée, par des opérateurs d'advection spécifiques (splitting directionnel et champ approché affine par morceaux), va rester continue et affine sur une triangulation,

formée cette fois-ci de trapèzes ou de triangles.
Notons V l'espace de toutes les fonctions f affines par morceaux sur une triangulation $\mathcal{T}(f)$.

$$V = \{f \in C^0(\mathbb{R}^2) f_{|T} \in \Pi_1 \quad \forall T \in \mathcal{T}(f)\}, \tag{7.3.3}$$

où

- $\mathbb{R}^2 = \cup_{T\in\mathcal{T}(f)} T$
- Pour tout $T \in \mathcal{T}(f)$, T est un polygone.
- Pour $T_1, T_2 \in \mathcal{T}(f)$ distincts, on a $\mathring{T}_1 \cap \mathring{T}_2 = \emptyset$.

7.4 Mesure de l'erreur de projection

La solution transportée par des opérateurs d'advection est un élément de V, elle est ensuite projetée sur un maillage (adaptatif gradué).
On cherche alors à estimer l'erreur de projection :

$$\|P^r_{\mathcal{M}} g - g\|_{L^\infty}, \tag{7.4.1}$$

pour un élément $g \in V$.

Interpolation affine dans les espaces de Sobolev. En appliquant le théorème 5.6.1, on obtient une information sur l'erreur d'interpolation locale en norme L^∞, lorsque la fonction est deux fois dérivable pour la norme L^1 :

Proposition 7.4.1. *Soit T un triangle et $g \in W^{2,1}(T)$.*
On considère une hauteur de T et sa base correspondante et on note l_1 (resp. l_2) la plus petite (resp. la plus grande) longueur entre la base et la hauteur.
Alors

$$\|g - P_T g\|_{L^\infty(T)} \leq C\Big(\frac{l_2}{l_1}\Big)^2 |g|_{W^{2,1}(T)}, \tag{7.4.2}$$

la constante C étant indépendante du triangle T considéré, et P_T est l'interpolation affine sur T.

Preuve. On reprend les notations du théorème 5.6.1.
On a l'inclusion $W^{2,1}(T) \subset L^\infty(T)$, d'après le lemme 5.7.12.
D'autre part, comme le triangle est contenu dans le rectangle qui a pour côtés une hauteur donnée de T et sa base correspondante, on obtient :

$$(\mathrm{mes}(T))^{-1} \cdot h^2 \leq (l_1 \cdot l_2/2)^{-1} \cdot (l_1^2 + l_2^2) = 2\Big(\Big(\frac{l_1}{l_2}\Big)^2 + \Big(\frac{l_2}{l_1}\Big)^2\Big) \leq 4\Big(\frac{l_2}{l_1}\Big)^2, \tag{7.4.3}$$

et on en déduit donc le résultat. □

Remarque 7.4.2. *De manière générale, dès que l'on prend famille de triangles ayant leurs angles minorés par une constante strictement positive, on obtient une inégalité du type*

$$\|g - P_T g\|_{L^\infty(T)} \leq C|g|_{W^{2,1}(T)}, \tag{7.4.4}$$

où la constante C ne dépend pas du triangle T choisi dans la famille considérée.

Masses de Dirac. Si l'on considère $g \in V$, alors $g \in W^{1,\infty}(\mathbb{R}^2)$. Cependant, les dérivées secondes de g sont des combinaisons linéaires de distribution de Dirac sur les arêtes. Ainsi, en particulier, si g n'est pas une fonction affine sur un domaine ω, g n'appartient pas à $W^{2,1}(\omega)$.
Cependant, on peut néanmoins obtenir un analogue de la proposition 7.4.1 dans ce cadre plus faible.

Notations. Pour une arête $\gamma \in \mathcal{E}(\mathcal{T}(f))$, on note

- $[g]_\gamma$, le saut de valeurs d'une fonction g constante localement de chaque côté de γ.
- $n = (n_x, n_v)$ une normale unitaire de l'arête γ.

La norme $W^{2,1,*}$. On introduit d'abord une suite ρ_n régularisante, définie par :

$$\rho_n(x,v) = \zeta(\frac{x}{n})\zeta(\frac{v}{n}), \tag{7.4.5}$$

où ζ est une fonction de classe C^∞, à support compact vérifiant :

$$\zeta(x) = 1 \text{ si } |x| \leq 1, \quad \zeta(x) = 0 \text{ si } |x| \geq 2. \tag{7.4.6}$$

On définit alors $g_n = \rho_n * g$, où $*$ désigne le produit de convolution.
On a alors la proposition suivante :

Proposition 7.4.3. *Soit $\omega \subset \mathbb{R}^2$ un ouvert borné, alors*

$$|\partial_{xx} g_n|_{L^1(\omega)}, |\partial_{xv} g_n|_{L^1(\omega)}, |\partial_{vx} g_n|_{L^1(\omega)}, |\partial_{vv} g_n|_{L^1(\omega)} \text{ convergent,} \tag{7.4.7}$$

et on note leurs limites

$\int_\omega |\partial_{xx} g|$, $\int_\omega |\partial_{xv} g|$, $\int_\omega |\partial_{vx} g|$, $\int_\omega |\partial_{vv} g|$ *qui vérifient :*

$$\begin{cases} \int_\omega |\partial^2_{xx} g| = \sum |\gamma \cap \omega| \ \left| n_x [\partial_x g]_\gamma \right| \\ \int_\omega |\partial^2_{xv} g| = \sum |\gamma \cap \omega| \ \left| n_v [\partial_x g]_\gamma \right| \\ \int_\omega |\partial^2_{vx} g| = \sum |\gamma \cap \omega| \ \left| n_x [\partial_v g]_\gamma \right| \\ \int_\omega |\partial^2_{vv} g| = \sum |\gamma \cap \omega| \ \left| n_v [\partial_v g]_\gamma \right| \end{cases}, \tag{7.4.8}$$

où la somme porte sur toutes les arêtes $\gamma \in \mathcal{E}(\mathcal{T}(f))$, telles que $\gamma \cap \omega \neq \emptyset$.
De plus, on a

$$\int_\omega |\partial^2_{xv} g| = \int_\omega |\partial^2_{vx} g|. \tag{7.4.9}$$

Enfin, en notant

$$|g|_{W^{2,1,*}(\omega)} = \int_\omega |\partial^2_{xx} g| + 2|\partial^2_{xv} g| + |\partial^2_{vv} g|. \tag{7.4.10}$$

$|g_n|_{W^{2,1}(\omega)}$ *converge vers* $|g|_{W^{2,1,*}(\omega)} < \infty$.

Preuve. L'approximation g_n de g est infiniment dérivable ; en particulier, elle appartient à $W^{2,1}(\omega)$.
On suppose d'abord que ω ne contient qu'une seule arête γ (ou une partie d'arête : $\omega \cap \gamma$), et on note ω^+ et ω^- les deux régions de ω séparées par γ. On calcule alors

$$\begin{aligned} A &:= \int_\omega \big|\partial_{xx} g_n(x,v)\big| \mathrm{d}x\mathrm{d}v \\ &= \int_\omega \Big| \int_{\mathbb{R}^2} \partial_x \rho_n(t_x,t_v) \partial_x g(x-t_x, v-t_v) \mathrm{d}t_x \mathrm{d}t_v \Big| \mathrm{d}x\mathrm{d}v \\ &= \int_\omega \Big| \int_{\omega^+_{x,v}} \partial_x \rho_n \partial_x g^+ \mathrm{d}t_x \mathrm{d}t_v + \int_{\omega^-_{x,v}} \partial_x \rho_n \partial_x g^- \mathrm{d}t_x \mathrm{d}t_v \Big| \mathrm{d}x\mathrm{d}v, \end{aligned}$$

où $\omega^+_{x,v}$ (resp. $\omega^-_{x,v}$) désigne l'ensemble des (t_x,t_v) vérifiant $(x-t_x, v-t_v) \in \omega^+$ (resp. $(x-t_x, v-t_v) \in \omega^-$), et $\partial_x g^+$ et $\partial_x g^-$ sont les valeurs de $\partial_x g$ de part et d'autre de l'arête γ ; elles sont indépendantes de x, v, t_x et t_v.
On applique alors la formule de Stokes, et on obtient :

$$A = \int_\omega \Big| \int_{\gamma_{x,v}} \rho_n [\partial_x g]_\gamma \mathrm{d}\Gamma \Big| \mathrm{d}x\mathrm{d}v,$$

où $\gamma_{x,v}$ défini par

$$\gamma_{x,v} = \{(t_x,t_v) : (x-t_x, v-t_v) \in \gamma \cap \omega\},$$

et l'intégrale est entendue pour la mesure $\mathrm{d}\Gamma$ par rapport à $\gamma_{x,v}$.
Comme ρ_n est positive, en faisant un changement de variable, en appliquant le théorème de Fubini-Tonelli, et en refaisant un changement de variable, on obtient :

$$\begin{aligned} A &= \int_\omega \int_{\gamma\cap\omega} \rho_n(x-\cdot, v-\cdot) \mathrm{d}\Gamma [\partial_x g]_\gamma \mathrm{d}x\mathrm{d}v, \\ &= \int_{\gamma\cap\omega} \int_\omega \rho_n(x-\cdot, v-\cdot) [\partial_x g]_\gamma \mathrm{d}x\mathrm{d}v\mathrm{d}\Gamma, \\ &= \int_{\gamma\cap\omega} \int_\omega \rho_n(x,v) [\partial_x g]_\gamma \mathrm{d}x\mathrm{d}v\mathrm{d}\Gamma = |\gamma \cap \omega| \ \big|[\partial_x g]_\gamma\big|. \end{aligned}$$

Maintenant, si on a plusieurs arêtes disjointes, on obtient le résultat en sommant. Dans le cas où ω intersecte plusieurs supports à la fois, on peut rendre l'intégrale sur un domaine

assez petit contenant les intersections de telle sorte qu'on obtient le résultat lorsque $n \to \infty$.
Les autres cas se traitent alors de la même manière.
Pour obtenir (7.4.9), on peut utiliser le fait que le saut de la dérivée le long de la tangente τ à une arête γ est nul.
En effet, si on note a et b deux points distincts de γ, alors la valeur absolue du gradient le long de la tangente τ de chaque côté de γ vaut

$$\frac{|f(b) - f(a)|}{|b - a|}, \tag{7.4.11}$$

où $|b - a|$ est la distance euclidienne entre a et b.
Comme

$$\partial_\tau f = \tau_x \partial_x f + \tau_v \partial_v f, \quad |\tau_x| = |n_v| \quad \text{et } |\tau_v| = |n_x|, \tag{7.4.12}$$

on obtient bien

$$|n_v [\partial_x f]_\gamma| = |n_x [\partial_v f]_\gamma|, \tag{7.4.13}$$

c'est-à-dire (7.4.9) en multipliant par $|\gamma \cap \omega|$ et en sommant. □

Remarque 7.4.4. *Dans le cas où la fonction est continue, on a pu définir la masse des dérivées secondes par l'intégrale d'une masse de Dirac ; si la fonction n'est plus continue, les dérivées secondes engendrent des dérivées de masse de Dirac que l'on ne peut plus intégrer.*

Interpolation affine dans V. On obtient alors une estimation d'erreur de projection pour les éléments de V.

Lemma 7.4.5. *Pour chaque triangle T, et $g \in V$, l'interpolation locale affine P_T satisfait*

$$\|g - P_T g\|_{L^\infty(T)} \leq C \Big(\frac{l_2}{l_1}\Big)^2 |g|_{W^{2,1,*}(T)} \tag{7.4.14}$$

où C est une constante qui ne dépend pas du triangle T, et l_1 (resp. l_2) désigne la plus petite (resp. la plus grande longueur) entre une hauteur de T et sa base correspondante.

Remarque 7.4.6. *Comme les dérivées secondes ne sont pas des fonctions de L^1, il est possible d'avoir :*

$$|g|_{W^{2,1,*}(\bar{T})} > |g|_{W^{2,1,*}(T)} \tag{7.4.15}$$

Preuve du lemme 7.4.5. Soit $g \in V$.
On introduit la suite régularisée

$$g_n := g * \rho_n \in W^{2,1}(T).$$

D'après la proposition 7.4.3, $|g_n|_{W^{2,1}(T)}$ converge vers $|g|_{W^{2,1,*}(T)}$.
D'autre part, on a :

$$\|P_T g_n - P_T g\|_{L^\infty(T)} = \|P_T(g_n - g)\|_{L^\infty(T)} \leq \|g_n - g\|_{L^\infty(T)}. \tag{7.4.16}$$

Comme g est continue, g_n converge uniformément vers g sur tout compact de $\mathbb{R}^2$; en particulier, la quantité de droite de (7.4.16) tend vers zéro.
Le résultat est alors établi en passant à la limite dans l'inégalité (7.4.1) valable pour g_n, □

En ce qui concerne l'interpolation proprement dite sur un maillage adaptatif, qui diffère à cause des noeuds contraints, on a alors le résultat suivant :

Lemme 7.4.7. *Pour tout maillage (adaptatif gradué) $\mathcal{M}$ et pour chaque maille $\alpha \in \mathcal{M}$, et chaque $g \in V$, $P^r_{\mathcal{M}}$ satisfait*

$$\|g - P^r_{\mathcal{M}} g\|_{L^\infty(\overline{\alpha^*_*})} \leq C\sigma_{L,r} |g|_{W^{2,1,*}(\mathring{\alpha}^*_*)} \tag{7.4.17}$$

en notant

$$\sigma_{L,r} := \max\Big(\Big(\frac{L}{r}\Big)^2, \Big(\frac{r}{L}\Big)^2\Big). \tag{7.4.18}$$

La constante C ne dépend pas de α, ni de L et r.

Preuve. S'il n'y a pas de de noeud contraint, il s'agit du lemme 7.4.5, que l'on applique directement au triangle rectangle ouvert $\mathring{\alpha}^*$, en prenant pour base et hauteur un côté perpendiculaire ; comme le quotient entre le côté parallèle à l'axe des x et celui parallèle à l'axe des v est donné par $\frac{L}{2r}$, on a bien le résultat, puisque

$$\sigma_{L,2r} \leq 2\sigma_{L,r}. \tag{7.4.19}$$

Maintenant, dans le cas où il y a un noeud contraint, il s'agit en fait également du lemme 7.4.5, que l'on applique cette fois-ci au triangle ouvert $\mathring{\alpha}^*_*$.
En effet, on a interpolation affine sur ce triangle. Cette fois-ci la base et la hauteur que l'on considère sont à nouveau parallèles aux axes et le rapport entre les deux est soit L/r, soit son inverse et donc le résultat subsiste. □

Courbure totale. La quantité $W^{2,1,*}$ pour un élément de V précédement définie sur des ouverts, peut se prolonger en une mesure. On introduit maintenant une autre quantité qui sera plus pratique dans la suite de l'analyse.
Ainsi, pour tout élément de V et pour tout ensemble $\omega \subset \mathbb{R}^2$, on définit la *courbure totale*

$$|f|_{\star(\omega)} := \sum_{\gamma \in \mathcal{E}(f)} |\gamma \cap \omega| \; \|[Df]_\gamma\|, \tag{7.4.20}$$

avec :

$$\|[Df]_\gamma\| := |[\partial_x f]_\gamma| + |[\partial_v f]_\gamma|. \tag{7.4.21}$$

La notation $|\gamma \cap \omega|$ doit être comprise comme la mesure de Haussdorff monodimensionnelle de $\gamma \cap \omega$ (et donc est bien défini pour tout ensemble de $\mathbb{R}^2$).
Remarquons que cette courbure définit une mesure. Plus particulièrement la propriété capitale dont nous aurons besoin est la suivante :

Lemme 7.4.8. *Pour tout $g \in V$ et pour un nombre fini d'ensembles disjoints*

$$A_i \quad i = 1, \dots, n,$$

avec un entier $n \in \mathbb{N}$, on a :

$$|f|_{\star \cup_{i=1}^n A_i} = |f|_{\star A_1} + \cdots + |f|_{\star A_n} \tag{7.4.22}$$

Preuve. Par récurrence, on se ramène à traiter le cas $n = 2$, et il suffit alors de voir que

$$|\gamma \cap A_1 \cap A_2| = |\gamma \cap A_1| + |\gamma \cap A_2|, \tag{7.4.23}$$

ce qui est vrai, puisque la mesure de Haussdorff monodimensionnelle est une mesure. □

Le lemme suivant donne alors l'équivalence entre cette nouvelle mesure et la quantité $W^{2,1,*}$ définie sur les ouverts.

Lemme 7.4.9. *Pour tout $g \in V$, et tout ouvert borné ω, on a :*

$$|g|_{*(\omega)} \leq |g|_{W^{2,1,*}(\omega)} \leq 4|g|_{*(\omega)}. \tag{7.4.24}$$

Preuve. L'inégalité de droite s'obtient directement.
Pour l'autre inégalité, on a

$$\left|[\partial_x g]_\gamma\right| \leq \left|n_x[\partial_x g]_\gamma\right| + \left|n_v[\partial_x g]_\gamma\right| \tag{7.4.25}$$

et

$$\left|[\partial_v f]_\gamma\right| \leq \left|n_x[\partial_v g]_\gamma\right| + \left|n_v[\partial_v g]_\gamma\right|, \tag{7.4.26}$$

puisque $|n_x| + |n_v| = \sqrt{1 + 2|n_x|\ |n_v|} \geq 1$.
Ainsi, en multipliant les inégalités précédentes par $|\gamma \cap \omega|$ et en sommant sur les arêtes γ, on obtient :

$$|g|_{*(\omega)} \leq \int_\omega |\partial^2_{xx} g| + \int_\omega |\partial^2_{xv} g| + \int_\omega |\partial^2_{vx} g| + \int_\omega |\partial^2_{vv} g|, \tag{7.4.27}$$

ce qui donne bien le résultat en utilisant (7.4.9). □

Courbure maximale. On a vu que l'erreur d'interpolation sur une maille donnée est contrôlée par la courbure totale sur cette maille. L'erreur sur tout le maillage est alors contrôlée par le maximum de la courbure sur toutes ses mailles.
Pour $g \in V$, et pour un maillage adaptatif $\mathcal{M}$, on définit ainsi *la courbure maximale*

$$\mu_r(\mathcal{M}, g) := \sup_{\alpha \in \mathcal{M}} |g|_{\star(\check{\alpha})}. \tag{7.4.28}$$

L'indice r indique ici, à nouveau, que $\check{\alpha} := \theta_r^{-1}(\alpha)$.

Remarque 7.4.10. *Cette quantité peut-être infinie si on suppose seulement que $g \in V$. (mais bien finie si g est aussi supposée dans $C^0_{L,s}(\mathbb{R}^2)$, pour s arbitrairement grand).*

D'après le lemme 7.4.7, on a alors :

Corollaire 7.4.11. *Pour $g \in V$, $P^r_{\mathcal{M}}$ vérifie :*

$$\|g - P^r_{\mathcal{M}} g\|_{L^\infty(\mathbb{R}^2)} \leq C\sigma_{L,r}\mu(\mathcal{M}, g). \tag{7.4.29}$$

Preuve. Soit un point $a \in \mathbb{R}^2$, $\hat{a}$ appartient donc à une (unique) demi-maille α^*, d'une maille $\alpha \in \mathcal{M}$.
Donc

$$|(g - P^r_{\mathcal{M}} g)(a)| \leq \|g - P^r_{\mathcal{M}} g\|_{L^\infty(\check{\alpha}^*_*)} \leq C\sigma_{L,r}|g|_{W^{2,1,*}(\mathring{\alpha}^*_*)}, \tag{7.4.30}$$

d'après le lemme 7.4.7. Ainsi, on est assuré que :

$$|(g - P^r_{\mathcal{M}} g)(a)| \leq C\sigma_{L,r}|g|_{*,\mathring{\alpha}^*_*} \tag{7.4.31}$$

en utilisant le lemme 7.4.9.
Or on a

$$\mathring{\check{\alpha}}^*_* \subset \check{\alpha} \cup \check{\beta}, \tag{7.4.32}$$

où $\beta \in \mathcal{M}$ est définie par l'unique maille contenant α_*.
En utilisant la propriété de mesure de la courbure, on obtient :

$$|g|_{*,\mathring{\alpha}^*_*} \leq |g|_{*,\check{\alpha}} + |g|_{*,\check{\beta}} \leq 2\mu_r(\mathcal{M}, g), \tag{7.4.33}$$

et le résultat en découle. □

7.5 Transport de la régularité

A chaque pas de temps, on transporte une solution définie sur un maillage (adaptatif gradué) et on la reprojette sur un nouveau maillage.
On considère maintenant une maille quelconque, et on cherche à voir quelle est l'erreur que l'on commet en interpolant sur cette maille.
On a vu que la courbure totale sur cette maille de la solution transportée contrôle justement cette erreur d'interpolation, elle est alors elle même contrôlée par la courbure totale sur la maille advectée en arrière de la solution avant le transport.
On regarde ici un cas générique d'advection qui sera par la suite appliqué à nos opérateurs d'advection habituels.

Opérateur de transport. On considère ici un opérateur d'advection de la forme générale :

$$\mathcal{F}_v : (x, v) \to (x, v + G(x))$$

et un opérateur linéaire de transport

$$\mathcal{T}^v : g \in V \to g \circ \mathcal{F}_v^{-1}.$$

Le *champ* G est supposé affine par morceaux. On suppose que G est l'interpolation affine d'une fonction $\tilde{G} \in W^{2,\infty}([0, L])$, et que G est périodique de période L.
De manière similaire, on considère un opérateur d'advection suivant l'axe des x :

$$\mathcal{F}_x : (x, v) \to (x + G(v), v)$$

et un opérateur linéaire de transport

$$\mathcal{T}^x : g \in V \to g \circ \mathcal{F}_x^{-1}.$$

Dans ce cas, G est supposé affine sur $\mathbb{R}$. Lorsque l'on ne distingue pas $\mathcal{T}^x$ et $\mathcal{T}^v$, on écrira $\mathcal{T}$, et on emploiera de même la notation $\mathcal{F}$, au lieu de $\mathcal{F}_x$ et $\mathcal{F}_v$.

Remarque 7.5.1. *On prendra $G(x) = \Delta t E_g(x)$ pour $\mathcal{F}_v$, $G(v) = \Delta t/2$ pour $\mathcal{F}_x$, et les éléments g seront pris dans un espace d'approximation $V_{\mathcal{M}}^r \subset V$.*

On rappelle la quantité

$$Q(g) := \sup\{1 + |v| : \exists x,\ g(x, v) \neq 0\}$$

qui borne le support de g suivant l'axe des vitesses v.

Transport d'un triangle. Le lemme préliminaire suivant est utile pour calculer l'évolution de la courbure transportée par l'opérateur d'advection.

Lemme 7.5.2. *On considère un triangle T, de sommets A, B et C tel que T soit rectangle en A, avec (AC) (resp. (AB)) perpendiculaire à l'axe des x (resp. l'axe des v).*
On note A', B' et C' les images de A, B et C par $\mathcal{F}_v^{-1}$ (sommets de $T' = \mathcal{F}_v^{-1}(T)$) et on suppose que G est affine sur T_x, où T_x est la projection du triangle T sur l'axe des x.
On a alors les estimations suivantes, concernant le transport des longueurs :

$$A'B' = AB \tag{7.5.1}$$

$$A'C' \leq (1 + |G|_{W^{1,\infty}(T_x)}|T_x|)AC \tag{7.5.2}$$

$$B'C' \leq (1 + |G|_{W^{1,\infty}(T_x)}|T_x|)BC \tag{7.5.3}$$

et concernant le transport des sauts de gradients :

$$\|[D(\mathcal{T}g)]_{A'C'}\| \leq (1 + |G|_{W^{1,\infty}(T_x)})\|[D(g)]_{AC}\|, \tag{7.5.4}$$

$$\|[D(\mathcal{T}g)]_{B'C'}\| \leq (1 + |G|_{W^{1,\infty}(T_x)})\|[D(g)]_{BC}\|, \tag{7.5.5}$$

et

$$\|[D(\mathcal{T}g)]_{A'B'}\| \leq \|[D(g)]_{AB}\| + |g|_{W^{1,\infty}(T)}|[\partial_x G]_{x_A}|, \tag{7.5.6}$$

où $|[\partial_x G]_{x_A}|$ *est le saut de discontinuité de* G *(affine par morceaux) en l'abscisse* x_A *de* A.

Remarque 7.5.3. *On a un énoncé équivalent pour* $\mathcal{F}_x$.

Preuve. Remarquons d'abord que $A'B' = AB$.
Si l'on note $v_{A'}$ l'ordonnée de A' image de A par $\mathcal{T}$, et de même $v_{C'}$, v_A et v_C, on obtient :

$$v_{C'} = v_{A'} - G(x_C) + G(x_A), \tag{7.5.7}$$

Avec l'inégalité triangulaire $A'C' \leq AC + |v_{C'} - v_{A'}|$, on en déduit que :

$$A'C' \leq (1 + |G|_{W^{1,\infty}(T_x)}|T_x|)AC. \tag{7.5.8}$$

De même, on a

$$B'C' \leq |x_C - x_B| + |v_{C'} - v_{B'}| \leq BC + |v_{C'} - v_{B'}|, \tag{7.5.9}$$

et donc

$$B'C' \leq (1 + |G|_{W^{1,\infty}(T_x)}|T_x|)BC. \tag{7.5.10}$$

- On utilise maintenant les formules (7.5.28) et (7.5.29).

Pour le côté $A'C'$, G' est constant et on en déduit donc :

$$|[\partial_x \mathcal{T}g]_{A'C'}| \leq |[\partial_x g]_{AC}| + |G|_{W^{1,\infty}(T_x)}|[\partial_v g]_{AC}|. \tag{7.5.11}$$

On a aussi :

$$|[\partial_v \mathcal{T}g]_{A'C'}| \leq |[\partial_v g]_{AC}| \tag{7.5.12}$$

et donc

$$\|[D(\mathcal{T}g)]_{A'C'}\| \leq (1 + |G|_{W^{1,\infty}(T_x)})\|[D(g)]_{AC}\|. \tag{7.5.13}$$

Le côté $B'C'$ se traite de la même manière.
Pour le côté $A'B'$, le calcul est différent.
Comme g est affine sur $A'B'$ et continue, on a :

$$|[\partial_v \mathcal{T}g]_{A'B'}| = |[\partial_v g]_{AB}| = 0. \tag{7.5.14}$$

D'autre part, comme $\partial_x g$ ne dépend pas de v, on a aussi :

$$|[\partial_x g(x, v - G(x))]_{A'B'}| = |[\partial_x g]_{A'B'}|. \tag{7.5.15}$$

Enfin, en notant x^+ et x^-, des coordonnées suivant x de points de part et d'autre de $[A'B']$, qui sont dans deux mailles triangulaires voisines, on a :

$$\begin{aligned}\partial_x G(x^+)\partial_v g(x^+, v - G(x^+)) - \partial_x G(x^-)\partial_v g(x^-, v - G(x^-)) \\ = (\partial_x G(x^+) - \partial_x G(x^-))\partial_v g(x, v - G(x)) \quad x \in \{x^+, x^-\}\end{aligned} \tag{7.5.16}$$

Donc, en posant $|[\partial_x G]_{x_A}| = |\partial_x G(x^+) - \partial_x G(x^-)|$, on obtient :

$$|[\partial_x G(x)\partial_v g(x, v - G(x))]_{A'B'}| \leq |g|_{W^{1,\infty}(\mathcal{F}^{-1}(T))}|[\partial_x G]_{x_A}|. \tag{7.5.17}$$

On a ainsi :

$$\|[D(\mathcal{T}g)]_{A'B'}\| \leq \|[D(g)]_{AB}\| + |g|_{W^{1,\infty}(\mathcal{F}^{-1}(T))}|[\partial_x G]_{x_A}|. \tag{7.5.18}$$

□

Transport d'une maille. Le lemme suivant permet de contrôler la courbure de la solution transportée $\mathcal{T}f$ sur une région ω formée de triangles dont les extrémites sont des noeuds d'interpolation pour $\tilde{G}$. par rapport à la courbure de f sur le domaine d'influence $\mathcal{F}^{-1}(\omega)$.

Lemme 7.5.4. *L'opérateur de transport $\mathcal{T}$ préserve la structure affine par morceaux :*

$$g \in V \Rightarrow \mathcal{T}g \in V. \tag{7.5.19}$$

et le support en vitesse vérifie :

$$Q(\mathcal{T}^x g) = Q(g). \tag{7.5.20}$$

$$Q(\mathcal{T}^v g) \leq Q(g) + \|\tilde{G}\|_{L^\infty}. \tag{7.5.21}$$

Pour $r > 0$, et en définissant un nombre $r_{\mathcal{T}}$ vérifiant :

$$r_{\mathcal{T}^x} \geq r \quad r_{\mathcal{T}^v} \geq r + \|\tilde{G}\|_{L^\infty(\mathbb{R})}, \tag{7.5.22}$$

on a alors

$$g \in C^0_{L,r}(\mathbb{R}^2) \Rightarrow \mathcal{T}g \in C^0_{L,r_{\mathcal{T}}}(\mathbb{R}^2). \tag{7.5.23}$$

De plus, les estimations locales suivantes sont satisfaites :

$$|\mathcal{T}g|_{W^{1,\infty}(\omega)} \leq (1 + |G|_{W^{1,\infty}(\omega_z)})|g|_{W^{1,\infty}(\mathcal{F}^{-1}(\omega))}, \tag{7.5.24}$$

et

$$\begin{aligned} |\mathcal{T}g|_{\star(\omega)} \leq (1 + |G|_{W^{1,\infty}(\omega_z)})^2 |g|_{\star(\mathcal{F}^{-1}(\omega))} \\ + |\omega_z|\ |\tilde{G}|_{W^{2,\infty}(\omega_z)}|g|_{W^{1,\infty}(\mathcal{F}^{-1}(\omega))} \end{aligned}, \tag{7.5.25}$$

où $\omega_z := \omega_x$, si $\mathcal{T} = \mathcal{T}^v$ et $\omega_z := \omega_v$, si $\mathcal{T} = \mathcal{T}^x$, ($\omega_x$ étant la projetée de ω sur l'axe des x, et de même pour ω_v).
Notons que l'on a également : $|G|_{W^{1,\infty}(\omega_x)} \leq |\tilde{G}|_{W^{1,\infty}(\omega_x)}$

Preuve. Comme l'opérateur $\mathcal{T}^x$ ne fait pas augmenter le support en v, on a (7.5.20). D'autre part, à partir de $\mathcal{T}^v g(x, v) = g(x, v - G(x))$,
on obtient

$$Q(\mathcal{T}^v g) \leq \|G\|_{L^\infty} + Q(g). \tag{7.5.26}$$

Or une projection affine fait décroître la norme L^∞, et on obtient donc :

$$Q(\mathcal{T}^v g) \leq \|\tilde{G}\|_{L^\infty} + Q(g). \tag{7.5.27}$$

On suppose maintenant que $\mathcal{T} = \mathcal{T}^v$, mais le raisonnement est le même si $\mathcal{T} = \mathcal{T}^x$.
Pour presque tout (x, v), on a :

$$\partial_x \mathcal{T} g(x,v) = \partial_v g(x, v - G(x)) G'(x) + \partial_x g(x, v - G(x)), \tag{7.5.28}$$

et

$$\partial_v \mathcal{T} g(x) = \partial_v g(x, v - G(x)); \tag{7.5.29}$$

ceci nous assure que

$$|\mathcal{T} g|_{W^{1,\infty}(\omega)} \leq |g|_{W^{1,\infty}(\mathcal{F}^{-1}(\omega))}(1 + \|G'\|_{W^{1,\infty}(\omega_x)}). \tag{7.5.30}$$

D'autre part, G est dérivable par morceaux ; en appliquant le théorème des accroissements finis pour $\tilde{G}$, sur chaque intervalle où G est dérivable, on en déduit que

$$|G|_{W^{1,\infty}(\omega_x)} \leq |\tilde{G}|_{W^{1,\infty}(\omega_x)}. \tag{7.5.31}$$

Quitte à rajouter des mailles (ce qui ne change pas la courbure), on peut supposer que ω est formé de triangles $T = (ABC)$ dont l'un des côtés AB est parallèle à l'axe des vitesses v, un autre, AC est parallèle à l'axe des x, et que G est affine sur T_x, projection de T sur l'axe des positions x.
L'image d'un triangle T par le transport $\mathcal{T}$ est alors un nouveau triangle ; l'image du côté parallèle à l'axe des vitesses reste parallèle et de même longueur.
En particulier, si $g \in V$, alors $\mathcal{T} g \in V$.
Le domaine ω est alors composée de triangles T précédemment décrits.
En sommant toutes les contributions, en utilisant le lemme précédent, on obtient :

$$|\mathcal{T} g|_{\star(\omega)} \leq (1 + |G|_{W^{1,\infty}(\mathcal{F}^{-1}(\omega))})^2 |g|_{\star(\mathcal{F}^{-1}(\omega))} + |g|_{W^{1,\infty}(\mathcal{F}^{-1}(\omega))} (\sum_{x_i \in \omega_x} |[\partial_x G]_{x_i}|), \tag{7.5.32}$$

où (x_i) est une suite croissante et G est affine par morceaux sur les intervalles (x_i, x_{i+1}).
En appliquant le théorème des accroissements finis, il existe une suite (y_i) croissante telle que $\tilde{G}'(y_i) = [\partial_x G]_{x_i}$ ($\tilde{G}$ est bien dérivable, comme elle est dans $W^{2,\infty}$, sa dérivée est lipschitzienne) ; on a $x_{i-1} < y_i < x_i$.
On obtient alors :

$$\begin{aligned} \sum_{x_i \in \omega_x} |[\partial_x G]_{x_i}| &= \sum_{y_i \in \omega_x} |G'(y_{i+1}) - G'(y_i)| \\ &= \sum_{y_i \in \omega_x} |\int_{y_i}^{y_{i+1}} G''(x) \mathrm{d}x| \leq \sum_{y_i \in \omega_x} \int_{y_i}^{y_{i+1}} |G''(x)| \mathrm{d}x \\ &= |G|_{W^{2,1}(\omega_x)} \leq |\omega_x| \; |G|_{W^{2,\infty}(\omega_x)}, \end{aligned} \tag{7.5.33}$$

et (7.5.25) s'ensuit. □

7.6 Prédiction du maillage

On cherche à construire un nouveau maillage (adaptatif gradué) *adapté* à la solution transportée. Ainsi, on souhaite que pour chaque maille l'erreur de projection soit petite. On a vu que celle-ci est contrôlée par la courbure maximale de la solution transportée de ce nouveau maillage prédit.
Dans notre prédiction, cette dernière quantité est contrôlée par la courbure maximale de la solution avant le transport qui peut être rendue petite par adaptation de mailles.

Advection de maillage. Etant donné un maillage $\mathcal{M}$, un champ d'advection $\mathcal{F}$ et un nombre $r > 0$, on construit un *maillage advecté* $\mathbb{T}(\mathcal{M}, \mathcal{F}, r)$.
Le nombre r sert ici pour le support en v des fonctions qui seront définies sur un tel maillage $\mathcal{M}$.
On définit alors un nombre $r_{\mathcal{T}}$ vérifiant (7.5.22) de telle sorte que si une fonction g appartient à $C^0_{L,r}(\mathbb{R})$, alors la fonction advectée $\mathcal{T}g$ appartient à $C^0_{L,r_{\mathcal{T}}}(\mathbb{R})$, d'après le lemme 7.5.4.
Le maillage $\mathbb{T}(\mathcal{M}, \mathcal{F}, r)$ est alors obtenu par l'algorithme suivant :

- Pour chaque maille α de niveau j_0, calculer la position arrière $\mathcal{F}^{-1}(\check{c}_\alpha)$ du centre c_α de la maille, avec ici $\check{c}_\alpha := \theta^{-1}_{r_{\mathcal{T}}}(c_\alpha)$.

- Si le niveau de cette position est strictement supérieur au niveau de la maille α, recommencer avec les quatre filles.

- Si le niveau est inférieur ou égal au niveau de la maille, ajouter cette maille au maillage.

- Périodiser en x, compléter en v et vérifier la structure d'arbre et la propriété d'arbre gradué.

Domaine d'influence. Le lemme suivant donne alors une information sur l'image réciproque d'une maille de ce maillage prédit.

Lemme 7.6.1. *Pour chaque maille $\alpha \in \mathbb{T}(\mathcal{M}, \mathcal{F}, r)$, on définit le domaine d'influence*

$$\mathcal{B}(\alpha) := \{\beta \in \mathcal{M},\ \mathcal{F}(\check{\beta}) \cap \mathring{\check{\alpha}} \neq \emptyset\}. \tag{7.6.1}$$

Ici, on sous-entend que $\check{\beta} := \theta_r^{-1}(\beta)$ tandis que $\check{\alpha} := \theta^{-1}_{r_{\mathcal{T}}}(\alpha)$.
Si

$$\frac{r_{\mathcal{T}}}{r} + \frac{L}{2r}|G|_{W^{1,\infty}([0,L])} \leq 3/2, \quad \mathcal{F} = \mathcal{F}_v, \tag{7.6.2}$$

ou si

$$\max(1 + \frac{2r_{\mathcal{T}}|G|_{W^{1,\infty}(\mathbb{R})}}{L}, \frac{r_{\mathcal{T}}}{r}) \leq 3/2, \quad \mathcal{F} = \mathcal{F}_x, \tag{7.6.3}$$

et si $\mathcal{M}$ est gradué (ce qui est toujours supposé), alors $\mathcal{B}(\alpha)$ contient au plus sept mailles β dont le niveau satisfait :

$$\ell(\beta) \leq \ell(\alpha) + 2. \tag{7.6.4}$$

Preuve. Soit c_α le centre de la maille α.
Supposons d'abord que cette maille est de niveau j_0.
Le point $(x^*, v^*) := (\hat{\mathcal{F}^{-1}}(\check{c_\alpha}))$ appartient à une unique maille $\alpha^* \in \mathcal{M}$ (ici, $\hat{\mathcal{F}^{-1}}$ signifie $\theta_r^{-1}(\mathcal{F}^{-1})$).
D'après la construction du maillage prédit, on a $\ell(\alpha) \geq \ell(\alpha^*)$.
Ainsi, comme $\ell(\alpha^*) \geq j_0$, on obtient $\ell(\alpha^*) = j_0 = \ell(\alpha)$.
Si maintenant α est de niveau strictement plus grand que j_0, la maille α a été raffinée. S'il s'agit d'un raffinement dû au respect de la structure d'arbre gradué, cette maille provient donc d'une maille mère qui est de niveau j_0 ou qui a été raffinée.
On voit alors qu'il existe une maille ancêtre $\tilde{\alpha}$ à α telle que tous les ancêtres jusqu'à $\tilde{\alpha}$ soient raffinées par la structure d'arbre gradué et telle que $\tilde{\alpha}$ soit de niveau j_0 ou $\tilde{\alpha}$ soit raffinée car le centre de la maille de sa mère advecté en arrière provient d'une maille de niveau strictement supérieur à celui de $\tilde{\alpha}$.
Il suffit alors de montrer le lemme pour cette maille $\tilde{\alpha}$, puisque

$$\ell(\tilde{\alpha}) \leq \ell(\alpha) \quad \mathcal{B}(\alpha) \subset \mathcal{B}(\tilde{\alpha}). \tag{7.6.5}$$

Par commodité, on renote α cette maille $\tilde{\alpha}$.
Ainsi, on considère deux cas : (I) soit α est de niveau j_0, et dans ce cas, on a vu que $\ell(\alpha^*) = \ell(\alpha)$; (II) soit le niveau de α est strictement plus grand que le niveau grossier j_0 et la maille α provient alors du raffinement d'une maille λ.
Pour le cas (II), le point $(\tilde{x}, \tilde{v}) := \hat{\mathcal{F}^{-1}}(\check{c_\lambda})$ appartient à une unique maille λ^* qui vérifie cette fois-ci bien

$$\ell(\lambda^*) > \ell(\lambda) = \ell(\alpha) - 1. \tag{7.6.6}$$

On a aussi

$$|c_\lambda - c_\alpha| = 2^{-\ell(\alpha)-1}, \tag{7.6.7}$$

en notant, pour un élément $z = (z_1, z_2) \in \mathbb{R}^2$ $|z| := \max(|z_1|, |z_2|)$.
On calcule maintenant. On se place d'abord dans le cas où $\mathcal{T} = \mathcal{T}^v$.
En partant de $z = (x, v)$, on a

$$\check{z} = (Lx, 2r_{\mathcal{T}}(v - \frac{1}{2})), \tag{7.6.8}$$

puis

$$\mathcal{F}^{-1}(\check{z}) = (Lx, 2r_{\mathcal{T}}(v - \frac{1}{2}) - G(Lx)), \tag{7.6.9}$$

et enfin

$$\begin{aligned}\hat{\mathcal{F}}^{-1}(\check{z}) = (x, \frac{r_{\mathcal{T}}}{r}(v - \frac{1}{2}) + \frac{1}{2} - \frac{1}{2r}G(Lx)) \\ = (x, \frac{r_{\mathcal{T}}}{r}v + \frac{1}{2}(\frac{r_{\mathcal{T}}}{r} - 1) - \frac{1}{2r}G(Lx)).\end{aligned} \tag{7.6.10}$$

En utilisant (7.6.7) et (7.6.10), on obtient alors :

$$|\hat{\mathcal{F}^{-1}}(\check{c_\lambda}) - \hat{\mathcal{F}^{-1}}(\check{c_\alpha})| \leq 2^{-\ell(\alpha)-1} \max(1, \frac{r_\mathcal{T}}{r} + \frac{L}{2r}|G|_{W^{1,\infty}([0,L])}), \tag{7.6.11}$$

et donc

$$|\hat{\mathcal{F}^{-1}}(\check{c_\lambda}) - \hat{\mathcal{F}^{-1}}(\check{c_\alpha})| \leq 2^{-\ell(\alpha)}\ 3/4, \tag{7.6.12}$$

d'après (7.6.2).
Pour le cas où $\mathcal{T} = \mathcal{T}_x$, on reprend les mêmes calculs, et on obtient :

$$\hat{\mathcal{F}}^{-1}(\check{z}) = (x - \frac{1}{L}G(2r_\mathcal{I}(v - \frac{1}{2})), \frac{r_\mathcal{T}}{r}v + \frac{1}{2}(\frac{r_\mathcal{T}}{r} - 1)), \tag{7.6.13}$$

et donc

$$|\hat{\mathcal{F}^{-1}}(\check{c_\lambda}) - \hat{\mathcal{F}^{-1}}(\check{c_\alpha})| \leq 2^{-\ell(\alpha)-1} \max(1 + \frac{2r_\mathcal{T}|G|_{W^{1,\infty}(\mathbb{R})}}{L}, \frac{r_\mathcal{T}}{r}), \tag{7.6.14}$$

ce qui donne également (7.6.12), d'après (7.6.3).
Si, maintenant, le niveau de $\hat{\mathcal{F}}^{-1}(\check{c}_\alpha)$ est inférieur ou égal à $\ell(\alpha) - 2$, en particulier, son niveau est strictement supérieur à j_0, et donc comme $\hat{\mathcal{F}}^{-1}(\check{c}_\lambda)$ est de niveau supérieur à $\ell(\alpha)$, et comme $\mathcal{M}$ est gradué, on en déduit que l'écart entre $\hat{\mathcal{F}}^{-1}(\check{c}_\alpha)$ et $\hat{\mathcal{F}}^{-1}(\check{c}_\lambda)$ doit être supérieur à $2^{-\ell(\alpha)-1}$, ce qui est impossible, d'après l'inégalité précédente.
On a donc dans tous les cas ((I) et (II)) :

$$\ell(\alpha) - 1 \leq \ell(\alpha^*) \leq \ell(\alpha). \tag{7.6.15}$$

Considérons maintenant un point quelconque $z = (x, v)$ tel que $w := \hat{\mathcal{F}}(\check{z}) \in \alpha$, où cette fois-ci $\hat{\mathcal{F}} := \theta_{r_\mathcal{T}}(\mathcal{F})$ et $\check{z} := \theta_r^{-1}(z)$.
On a donc :

$$|w - c_\alpha| \leq 2^{-\ell(\alpha)-1}. \tag{7.6.16}$$

En remarquant que $z = \hat{\mathcal{F}}^{-1}(\check{w})$ et $(x^*, v^*) = \hat{\mathcal{F}}^{-1}(\check{c_\alpha})$, on peut réappliquer les inégalités précédentes que l'on avait utilisé pour $w = c_\lambda$, pour obtenir :

$$|(x, v) - (x^*, v^*)| \leq 2^{-\ell(\alpha)}3/4. \tag{7.6.17}$$

Supposons maintenant que le niveau de α^* est $\ell(\alpha) - 1$.
Dans ce cas, le niveau des cellules voisines est au plus $\ell(\alpha)$, et donc, comme x^* est le milieu de α_x, grâce à (7.6.17), et comme le maillage $\mathcal{M}$ est gradué, $\mathcal{B}(\alpha)$ ne peut contenir que 3 mailles et le niveau des autres mailles peut être $\ell(\alpha) - 1$, $\ell(\alpha)$ ou $\ell(\alpha) - 2$, donc inférieur à $\ell(\alpha)$.
Si maintenant, on a $\ell(\alpha^*) = \ell(\alpha)$, alors le cardinal de $\mathcal{B}(\alpha)$ est inférieur ou égal à 7 (voir figure 7.2).
Le niveau des autres mailles peut être $\ell(\alpha) - 1$, $\ell(\alpha)$ (pour ces cas il y a 2 mailles au total), $\ell(\alpha) + 1$ ou $\ell(\alpha) + 2$ (il peut y avoir alors 3, 5 ou 7 mailles). □

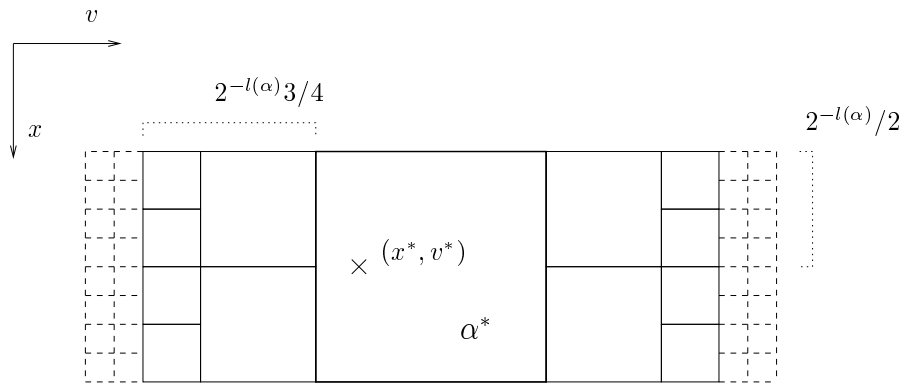


FIG. 7.2 – configuration de cardinalité maximale pour $\mathcal{B}(\alpha)$.

Transport de la courbure maximale. Grâce aux lemmes 7.5.4 et 7.6.1, on peut alors contrôler la courbure de la solution advectée pour le maillage ainsi prédit, en fonction de la courbure de la solution avant le transport.
Cette propriété est cruciale. En effet, c'est la courbure (maximale) de la solution advectée qui contrôle l'erreur pour la projection sur le maillage transporté. Cependant, cette courbure n'est pas disponible numériquement, puisqu'on ne connait pas numériquement la solution advectée exactement. Par contre, on peut calculer numériquement la courbure de la solution avant le transport. Ainsi, si celle-ci est suffisamment petite, grâce à cette propriété, la courbure de la solution advectée sera également petite, et donc l'erreur de projection sur le maillage transporté aussi.
On rappelle que la courbure maximale est définie par

$$\mu_r(\mathcal{M}, g) := \sup_{\alpha \in \mathcal{M}} |g|_{\star(\check{\alpha})}.$$

On aura aussi besoin de

$$\pi_r(M, g) := \sup_{\alpha \in \mathcal{M}} |\alpha| |g|_{W^{1,\infty}(\check{\alpha})}. \tag{7.6.18}$$

On a alors le

Lemme 7.6.2. *Sous la condition (7.6.2) ou (7.6.3) et en rappelant que $\mathcal{M}$ est gradué, on a :*

$$\mu_{r_{\mathcal{T}^v}}(\mathbb{T}(\mathcal{M}, \mathcal{F}_v, r), \mathcal{T}^v g) \leq 7(1 + |G|_{W^{1,\infty}([0,L])})^2 \mu_r(\mathcal{M}, g) + 4|\tilde{G}|_{W^{2,\infty}([0,L])}\ \pi_r(\mathcal{M}, g), \tag{7.6.19}$$

$$\mu_{r_{\mathcal{T}^x}}(\mathbb{T}(\mathcal{M}, \mathcal{F}_x, r), \mathcal{T}^x g) \leq 7(1 + |G|_{W^{1,\infty}(\mathbb{R})})^2 \mu_r(\mathcal{M}, g), \tag{7.6.20}$$

et

$$\pi_{r_{\mathcal{T}^v}}(\mathbb{T}(\mathcal{M}, \mathcal{F}_v, r), \mathcal{T}^v g) \leq 4(1 + |G|_{W^{1,\infty}([0,L])}) \pi_r(\mathcal{M}, g), \tag{7.6.21}$$

$$\pi_{r_{\mathcal{T}^x}}(\mathbb{T}(\mathcal{M}, \mathcal{F}_x, r), \mathcal{T}^x g) \leq 4(1 + |G|_{W^{1,\infty}(\mathbb{R})}) \pi_r(\mathcal{M}, g), \tag{7.6.22}$$

Preuve. Soit $\alpha \in \mathbb{T}(\mathcal{M}, \mathcal{F}, r)$.
Supposons d'abord que $\mathcal{T} = \mathcal{T}^v$.
On suppose pour ne pas alourdir les calculs que les extrémités de la maille sont des points d'interpolation pour $\tilde{G}$, mais le résultat reste vrai sans cette hypothèse et se montre de la même manière.
D'après le lemme 7.5.4, on a alors :

$$\begin{aligned} |\mathcal{T}g|_{\star(\check{\alpha})} \leq (1 + |\tilde{G}|_{W^{1,\infty}(\check{\alpha_x})})^2 |g|_{\star(\mathcal{F}^{-1}(\check{\alpha}))} \\ + |\check{\alpha_x}| \ |\tilde{G}|_{W^{2,\infty}(\check{\alpha}_x)} |g|_{W^{1,\infty}(\mathcal{F}^{-1}(\check{\alpha}))} \end{aligned}, \tag{7.6.23}$$

avec ici $\check{\alpha} := \theta^{-1}_{r_{T^v}}(\alpha)$.
Or, du lemme 7.6.1, grâce à (7.6.2) et en utilisant aussi la propriété de mesure (7.4.8), on en déduit que

$$|g|_{\star(\mathcal{F}^{-1}(\check{\alpha}))} \leq \sum |g|_{\star(\check{\beta})}, \tag{7.6.24}$$

où la somme porte sur au plus 7 mailles $\beta \in \mathcal{M}$.
Ainsi, on a :

$$|g|_{\star(\mathcal{F}^{-1}(\check{\alpha}))} \leq 7\mu(\mathcal{M}, g). \tag{7.6.25}$$

D'autre part, on a de même :

$$|\check{\alpha}_x| \ |g|_{W^{1,\infty}(\mathcal{F}^{-1}(\check{\alpha}))} \leq \max_{\beta} |\check{\alpha}_x| \ |g|_{W^{1,\infty}(\check{\beta})},$$

où la somme porte toujours sur au plus 7 mailles de $\mathcal{M}$.
Or, $|\check{\alpha_x}| = L|\alpha_x| = L|\alpha|$, et d'après (7.6.4), on a

$$|\alpha| = 2^2 2^{-(\ell(\alpha)+2)} \leq 2^2 2^{-\ell(\beta)} = 4|\beta|, \tag{7.6.26}$$

et donc

$$|\check{\alpha}_x| \ |g|_{W^{1,\infty}(\mathcal{F}^{-1}(\check{\alpha}))} \leq 4\pi(\mathcal{M}, g). \tag{7.6.27}$$

Avec (7.6.2) et (7.6.23), on en déduit alors (7.6.19).
Soit maintenant toujours $\alpha \in \mathbb{T}(\mathcal{M}, \mathcal{F})$; on a alors

$$|\alpha| \ |\mathcal{T}g|_{W^{1,\infty}(\check{\alpha})} \leq (1 + |G|_{W^{1,\infty}([0,L])}) |\alpha| \ |g|_{W^{1,\infty}(\mathcal{F}^{-1}(\check{\alpha}))},$$

d'après (7.5.30).
Et donc, comme $|\alpha_x| = |\alpha|$, on obtient d'après (7.6.27),

$$|\alpha| \ |\mathcal{T}g|_{W^{1,\infty}(\check{\alpha})} \leq 4(1 + |G|_{W^{1,\infty}([0,L])}) \pi_r(\mathcal{M}, g). \tag{7.6.28}$$

et (7.6.21) en découle.
On suppose maintenant que $\mathcal{T} = \mathcal{T}^x$.
D'après le lemme 7.5.4, on a cette fois-ci

$$|\mathcal{T}g|_{\star(\check{\alpha})} \leq (1 + |G|_{W^{1,\infty}(\check{\alpha_v})})^2 |g|_{\star(\mathcal{F}^{-1}(\check{\alpha}))}, \tag{7.6.29}$$

puisque G est supposée affine, on a $\tilde{G} = G$ donc $|\tilde{G}|_{W^{2,\infty}} = |G|_{W^{2,\infty}} = 0$.
Comme on a de même (7.6.25), on en déduit (7.6.20).
Enfin, (7.6.22) s'obtient de la même manière que (7.6.21). □

Adaptation de maillage. On a vu que la courbure maximale sur la solution transportée est contrôlée par la courbure sur la solution avant le transport qui est définie sur un maillage dyadique.
Il faut donc s'assurer, pour éviter que les erreurs ne s'accumulent de trop, d'avoir une courbure assez petite sur le maillage dyadique.
Cela est toujours possible, en raffinant localement, puisque la courbure s'exprime en fonction de la longueur des arêtes du maillage (dyadique) ; ainsi en raffinant une maille, on divise la courbure maximale.
Etant donnée une paire $(\mathcal{M},\ g \in V^r_{\mathcal{M}})$, on peut donc la modifier en $(\tilde{\mathcal{M}},\ \tilde{g} = P^r_{\tilde{\mathcal{M}}} g)$, afin qu'elle satisfasse

$$\pi_r(\tilde{\mathcal{M}}, \tilde{g}) + \mu_r(\tilde{\mathcal{M}}, \tilde{g}) \leq \varepsilon \tag{7.6.30}$$

et

$$\|g - \tilde{g}\|_{L^\infty} \lesssim \varepsilon, \tag{7.6.31}$$

où ε est une constante strictement positive prescrite.
Ce nouveau maillage $\tilde{\mathcal{M}}$ sera noté $\mathbb{A}(M, g, \varepsilon, r)$. On vient de voir qu'il peut toujours être obtenu par des raffinements locaux ; la fonction $\tilde{g}$ peut être prise égale à g dans ce cas. Notons également que l'on peut aussi (et il est souhaitable) déraffiner le maillage ; le seul souci est de garder les majorations (7.6.30) et (7.6.31).

7.7 Stabilité

Stabilité des opérateurs de transports. Remarquons d'abord que les opérateurs de transports sont bien stables pour la norme infinie et aussi pour la norme $W^{1,\infty}$.

Lemme 7.7.1. *On se fixe un champ $E \in W^{1,\infty}([0, L])$, et une advection $\mathcal{T}_v$ de la forme :*

$$\mathcal{T}_v g(x, v) = g(x, v + \Delta t E(x)). \tag{7.7.1}$$

On a alors :

$$\|\mathcal{T}_x g\|_{L^\infty(\Omega)} \leq \|g\|_{L^\infty(\mathbb{R}^2)}, \quad \|\mathcal{T}_v g\|_{L^\infty(\mathbb{R}^2)} \leq \|g\|_{L^\infty(\mathbb{R}^2)}, \tag{7.7.2}$$

pour tous $g, \in V$.
On a également

$$|\mathcal{T}_x g|_{W^{1,\infty}(\mathbb{R}^2)} \leq (1 + \Delta t/2)|g|_{W^{1,\infty}(\mathbb{R}^2)}, \tag{7.7.3}$$

pour tout $g \in V$.
Enfin,

$$|\mathcal{T}_v g|_{W^{1,\infty}(\Omega)} \leq (1 + \Delta t |E|_{W^{1,\infty}[0,L]})|g|_{W^{1,\infty}(\mathbb{R}^2)}, \tag{7.7.4}$$

pour tout $g \in V$.

Preuve. On a :

$$|\mathcal{T}_x g(x, v)| = |g(x - v\Delta t/2, v)| \leq \|g\|_{L^\infty(\Omega)}$$
$$\text{et} \quad |\mathcal{T}_v^{\tilde{g}} g(x, v)| = |g(x, v + E(x)\Delta t)| \leq \|g\|_{L^\infty(\Omega)}, \tag{7.7.5}$$

ce qui donne les estimations pour la norme infinie.
D'autre part, grâce au lemme 7.5.4, en prenant $G(v) = \frac{\Delta t}{2} v$, dans le cas $\mathcal{T} = \mathcal{T}^x$ on obtient (7.7.3) et si on prend $G(x) = \Delta t E(x)$, dans le cas $\mathcal{T} = \mathcal{T}^v$, l'inégalité (7.7.4) est vérifiée. □

Stabilité L^∞ de la projection L'opérateur de projection vérifie la propriété de stabilité suivante pour la norme L^∞ :

Lemme 7.7.2. *l'opérateur d'interpolation $P^r_{\mathcal{M}}$ vérifie*

$$\|P^r_{\mathcal{M}} g\|_{L^\infty(\breve{\alpha}^*_*)} \leq \|g\|_{L^\infty(\breve{\alpha}^*_*)}, \tag{7.7.6}$$

pour toute fonction $g \in C^0(\Omega)$, et toute demi-maille α^* d'une maille α quelconque de $\mathcal{M}$.

Preuve. $P^r_{\mathcal{M}} g$ est affine sur $\breve{\alpha}^*_*$, donc son maximum sur $\overline{\breve{\alpha}^*_*}$ est atteint en les sommets qui sont des points d'interpolation, c'est-à-dire des valeurs de g.
On en déduit alors le résultat. □

Stabilité $W^{1,\infty}$ de la projection. Au niveau des dérivées, on a également une propriété de stabilité, on regarde d'abord le cas d'une interpolation affine sur un triangle dont deux côtés sont parallèles aux axes.

Lemme 7.7.3. *Soit T un triangle dont deux de ses côtés sont parallèles aux axes. On a :*

$$|\partial_x P_T g|_{L^\infty(T)} \leq |\partial_x g|_{L^\infty(T)}, \tag{7.7.7}$$

$$|\partial_v P_T g|_{L^\infty(T)} \leq |\partial_v g|_{L^\infty(T)}, \tag{7.7.8}$$

et donc

$$|P_T g|_{W^{1,\infty}(T)} \leq |g|_{W^{1,\infty}(T)}, \tag{7.7.9}$$

pour tout $g \in V$.

Preuve. $\partial_x P_T g$ est constant sur la maille T. notons γ l'arête de T parallèle à l'axe des x, et soient $A = (A_x, A_v)$ et $B = (B_x, B_v)$ ses deux extrémités.
On a alors :

$$|\partial_x P_T g|_{L^\infty(\alpha)} = |g(B) - g(A)|/(|\gamma|). \tag{7.7.10}$$

Maintenant, $g(\cdot, A_v)$ est continue, affine sur des intervalles $I_j = [x_j, x_{j+1}]$, qui forment une partition de l'intervalle γ_x (la projection de γ sur l'axe des x) ; on peut donc appliquer l'inégalité des accroissements finis sur chaque intervalle I_j, pour obtenir :

$$|g(x_{i+1}, A_v) - g(x_i, A_v)| \leq |\partial_x g|_{L^\infty(T)} |I_j|.$$

En utilisant l'inégalité triangulaire, et en sommant les précédentes inégalités, on obtient :

$$|g(B) - g(A)| \leq |\partial_x g|_{L^\infty(T)} |\gamma|, \tag{7.7.11}$$

ce qui permet d'établir (7.7.7), d'après (7.7.10).
La démonstration est similaire pour (7.7.8). □

Remarque 7.7.4. *On ne peut pas établir ce genre d'inégalités pour une triangulation quelconque. On peut considérer comme contrexemple le cas d'une maille triangulaire α où les sommets sont $A := (-1, 0)$, $B := (0, 1/2)$ et $C := (1, 0)$, avec l'unique fonction g continue affine sur chaque demi-plan $x \geq 0$ et $x \leq 0$, qui vaut 0 en A, 1 en $(0, 0)$, 2 en B et 0 en C. On a alors :*

$$\partial_x g = 1/2, \quad \partial_v g = 1 \quad \textit{pour } x \leq 0, \quad v \in \mathbb{R}$$

$$\partial_x g = -1/2, \quad \partial_v g = 1 \quad \textit{pour } x \geq 0, \quad v \in \mathbb{R}$$

et d'autre part

$$\partial_x P_\alpha g = 0, \quad \partial_v P_\alpha g = 2 \quad \textit{pour } (x, v) \in \alpha,$$

où P_α est la projection affine sur la maille α, et donc

$$|\partial_v P_T g|_{L^\infty(\alpha)} = 2 > 1 = |\partial_v g|_{L^\infty(\alpha)}.$$

On donne maintenant un résultat sur la stabilité de la projection sur un maillage adaptatif.

Lemme 7.7.5. *On suppose que $2r = L$.*
Pour tout $g \in V$, pour tout maillage (adaptatif gradué) $\mathcal{M}$ et pour toute demi-maille α^* d'une maille quelconque $\alpha \in \mathcal{M}$, on a alors :

$$|P^r_{\mathcal{M}} g|_{W^{1,\infty}(\bar{\alpha}^*_*)} \leq |g|_{W^{1,\infty}(\bar{\alpha}^*_*)}. \tag{7.7.12}$$

Preuve. Si il n' y a pas de noeud contraint, c'est-à-dire si $\alpha^* = \alpha^*_*$, il s'agit du lemme 7.7.3 appliqué au triangle $T = \alpha^*$, et l'hypothèse $2r = L$ est superflue.
Dans le cas où il y a un noeud contraint, c'est-à dire lorsque $\alpha^*_* = \alpha^* \cup \alpha_*$, avec cette fois-ci α^* et α_* des mailles distinctes, on note b le noeud contraint, c l'autre point commun à $\bar{\alpha}^*$ et $\bar{\alpha}_*$. Les points a et d sont alors respectivement les autres sommets de α^* et α_*.
On suppose ici que (bc) est parallèle à l'axe des x (on raisonnera de même lorsque (bc) est parallèle à l'axe des v), et on note $\Delta x = |b - c|$ et $\Delta v = |b - a|$, de telle sorte que l'on a :

$$\Delta v / \Delta x = 2r/L = 1. \tag{7.7.13}$$

On a alors :

$$2\Delta x \Delta v (|\partial_x P^r_{\mathcal{M}} g| + |\partial_v P^r_{\mathcal{M}} g|) = \Delta v |2g(c) - g(a) - g(d)| + \Delta x |g(d) - g(a)|. \tag{7.7.14}$$

D'autre part, en appliquant le lemme 7.7.3 respectivement aux triangles α^* et α_*, on obtient :

$$|P_{\alpha^*}g|_{W^{1,\infty}(\alpha^*)} = \frac{1}{2\Delta x\Delta v}(2\Delta v|g(c)-g(b)| + 2\Delta x|g(a)-g(b)| \\ \leq |g|_{W^{1,\infty}(\alpha^*)} \leq |g|_{W^{1,\infty}(\alpha_*^*)}, \quad (7.7.15)$$

et

$$|P_{\alpha_*}g|_{W^{1,\infty}(\alpha_*)} = \frac{1}{2\Delta x\Delta v}(2\Delta v|g(c)-g(b)| + 2\Delta x|g(d)-g(b)| \\ \leq |g|_{W^{1,\infty}(\alpha_*)} \leq |g|_{W^{1,\infty}(\alpha_*^*)}, \quad (7.7.16)$$

Ainsi, en prenant le maximum du membre de gauche dans les deux inégalités précédentes, on en déduit que :

$$2\Delta v|g(c)-g(b)| + 2\Delta x\max(|g(a)-g(b)|,|g(c)-g(b)|) \leq 2\Delta x\Delta v|g|_{W^{1,\infty}(\alpha_*^*)}. \quad (7.7.17)$$

Pour établir (7.7.12), il suffit donc de vérifier que :

$$\Delta v|2g(c)-g(a)-g(d)| + \Delta x|g(d)-g(a)| \\ \leq 2\Delta v|g(c)-g(b)| + 2\Delta x\max(|g(a)-g(b)|,|g(c)-g(b)|). \quad (7.7.18)$$

Quitte à considérer $g-g(b)$, on peut se ramener au cas où $g(b)=0$, et grâce à (7.7.13), il reste à vérifier que :

$$|2g(c)-g(a)-g(d)| + |g(d)-g(a)| \leq 2|g(c)| + 2\max(|g(a)|,|g(c)|), \quad (7.7.19)$$

ce qui découle du calcul :

$$\begin{aligned} &|2g(c)-g(a)-g(d)| + |g(d)-g(a)| \\ &= |(g(c)-g(a))+(g(c)-g(d))| + |(g(c)-g(a))-(g(c)-g(d))| \\ &= 2\max(|g(c)-g(a)|,|g(c)-g(d)|) \leq 2|g(c)| + 2\max(|g(a)|,|g(d)|). \end{aligned}$$

Lorsque (bd) est parallèle à l'axe des v, il suffit d'échanger les rôles joués par x et v. □

Remarque 7.7.6. *(i) La stabilité $W^{1,\infty}$ de la projection facilite ici notre preuve de convergence ; elle va permettre de contrôler la norme $W^{2,\infty}$ du champ électrique de manière globale indépendamment de l'adaptation choisie. Le contrôle de cette dernière quantité semble essentiel, car la majoration sur la courbure maximale que l'on a, fait intervenir justement celle-ci.*

(ii) Il se trouve que cette stabilité n'est pas vraie pour n'importe qu'elle triangulation. Ainsi, si l'on considère le cas où les demi-mailles sont toujours orientés dans le même sens, on peut trouver un maillage adaptatif et une fonction de V telle que l'interpolation ne soit pas stable. On peut considérer l'exemple de la figure 7.3 : on a un noeud contraint en a. Si on prend pour g la fonction affine définie par les triangles (abc), (abd) et (bec) la fonction projetée $P_{\mathcal{M}}g$ sera affine sur (abc), mais vaudra 1 en b (à cause du noeud

contraint). Sur le triangle (abc) *(en supposant que ses côtés sont de longueur* 1*), on a alors :*

$$|\partial_x P_{\mathcal{M}} g| + |\partial_v P_{\mathcal{M}} g| = |\frac{g(d)+g(e)}{2} - g(b)| + |g(c) - g(b)| = 11 + 10 = 21, \qquad (7.7.20)$$

tandis que sur chacun des triangles (abd)*,* (abc) *et* (bce)*, le maximum de* $|g|_W^{1,\infty}$ *est donné sur le triangle* (abc) *et vaut* $20 < 21$*. Il est possible de prolonger* g *sur* $\mathbb{R}^2$ *en une fonction affine dont la norme* $W^{1,\infty}$ *ne dépasse pas* 20 *en dehors de ces trois triangles, de telle sorte que l'on a :*

$$|P_{\mathcal{M}} g|_{W^{1,\infty}((abc))} > |g|_{W^{1,\infty}(\mathbb{R}^2)}. \qquad (7.7.21)$$

(iii) Lorsque l'on a un noeud contraint dans une maille, on n'a pas stabilité de chacune des dérivées partielles séparément.

Lorsque la condition $2r/L = 1$ n'est pas vérifiée, on ne peut pas espérer avoir stabilité

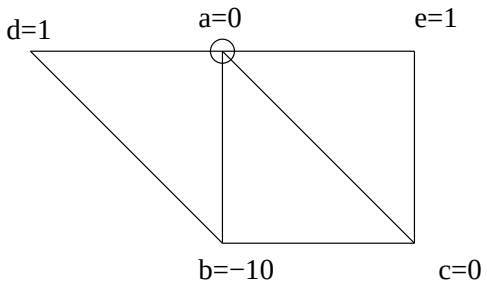


FIG. 7.3 – Une triangulation non $W^{1,\infty}$–stable, avec un noeud contraint en a.
En tournant l'arête (ac) en l'arête (be), on obtient la stabilité.

de la projection pour la norme $W^{1,\infty}$.
Néanmoins, on peut trouver une norme équivalente qui elle reste stable. Pour cela, il suffit de considérer pour un domaine $\omega \subset \mathbb{R}^2$, et pour $g \in W^{1,\infty}(\omega)$, la norme

$$|g|_{W^{1,\infty}_{L,r}(\omega)} := L|\partial_x g|_{L^\infty(\omega)} + 2r|\partial_v g|_{L^\infty(\omega)} \qquad (7.7.22)$$

Cette norme est bien équivalente à l'ancienne :

$$|g|_{W^{1,\infty}_{L,r}(\omega)} \leq \max(L, 2r)|g|_{W^{1,\infty}(\omega)}, \qquad (7.7.23)$$

et

$$\min(L, 2r)|g|_{W^{1,\infty}(\omega)} \leq |g|_{W^{1,\infty}_{L,r}(\omega)}. \qquad (7.7.24)$$

On vérifie, alors que l'on a bien stabilité pour cette nouvelle norme (qui est bien équivalente à l'ancienne) :

Lemme 7.7.7. *Pour tout* $g \in V$*, pour tout maillage (adaptatif gradué)* $\mathcal{M}$ *et pour toute demi-maille* α^* *d'une maille quelconque* $\alpha \in \mathcal{M}$*, on a alors :*

$$|P^r_{\mathcal{M}} g|_{W^{1,\infty}_{L,r}(\check{\alpha}^*_*)} \leq |g|_{W^{1,\infty}_{L,r}(\check{\alpha}^*_*)}. \qquad (7.7.25)$$

Preuve. On reprend les calculs du lemme précédent. Si l'on suppose que (bc) est parallèle à l'axe des x, l'analogue de (7.7.18) est :

$$L\Delta v|2g(c)-g(a)-g(d)|+2r\Delta x|g(d)-g(a)| \\ \leq 2L\Delta v|g(c)-g(b)|+4r\Delta x\max(|g(a)-g(b)|,|g(c)-g(b)|), \quad (7.7.26)$$

tandis que si (bc) est parallèle à l'axe des v, il s'agit de vérifier :

$$2r\Delta x|2g(c)-g(a)-g(d)|+L\Delta v|g(d)-g(a)| \\ \leq 4r\Delta x|g(c)-g(b)|+2L\Delta v\max(|g(a)-g(b)|,|g(c)-g(b)|), \quad (7.7.27)$$

et dans les deux cas, on peut factoriser par $2r\Delta x = L\Delta v$, et donc le résultat en découle, puisqu'il s'agit du même calcul que précédemment. □

On donne alors enfin un lemme qui montre la stabilité de la projection et du transport, pour cette nouvelle norme $W^{1,\infty}_{L,r}$.

Lemme 7.7.8. *On a :*

$$|P^r_{\mathcal{M}}g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)} \leq |g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)}, \quad (7.7.28)$$

pour tout $g \in V$, et

$$|\mathcal{T}_x g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)} \leq (1+r\Delta t)|g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)}, \quad (7.7.29)$$

ainsi que

$$|\mathcal{T}_v g|_{W^{1,\infty}_{L,r+M\Delta t}(\Omega)} \leq \Big(1+\frac{\Delta t}{2r}(2M+|E|_{W^{1,\infty}([0,L])})\Big)|g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)}, \quad (7.7.30)$$

Preuve. La première inégalité découle directement du lemme 7.7.7.
Pour les autres inégalités, on cherche à procéder comme dans le lemme 7.7.1. Néanmoins, ici il faut reprendre les estimations locales du lemme 7.5.4 de manière plus précise, puisque les directions de dérivations sont maintenant affectées d'un poids.
0n reprend donc les calculs. On a :

$$T_x g(x,v) = g(x+\frac{\Delta t}{2}v, v) \quad T_v g(x,v) = g(x, v+\Delta t E(x)). \quad (7.7.31)$$

On en déduit :

$$L|\partial_x T_x g(x,v)|+2r|\partial_v T_x g(x,v)| = L|\partial_x g(x+\frac{\Delta t}{2}v,v)| \\ +2r|\frac{\Delta t}{2}\partial_x g(x+\frac{\Delta t}{2}v,v)+\partial_v g(x+\frac{\Delta t}{2}v,v)| \\ \leq L|\partial_x g(x+\frac{\Delta t}{2}v,v)|+2r|\partial_v g(x+\frac{\Delta t}{2}v,v)|+2r\frac{\Delta t}{2}|\partial_x g(x+\frac{\Delta t}{2}v,v)| \\ \leq |g|_{W^{1,\infty}(\mathbb{R}^2)}+\frac{r}{L}\Delta t|g|_{W^{1,\infty}(\mathbb{R}^2)}, \quad (7.7.32)$$

ce qui donne bien (7.7.29).
D'autre part, pour presque tous $(x,v) \in \Omega$, on a

$$\begin{aligned}
&L|\partial_x T_v g(x,v)| + 2(r+M\Delta t)|\partial_v T_v g(x,v)| = \\
&L|\partial_x g(x, v+\Delta t E(x)) + \Delta t E'(x)\partial_v g(x, v+\Delta t E(x))| + 2(r+M\Delta t)|\partial_v g(x, v+\Delta t E(x))| \\
&\qquad \le L|\partial_x g(x, v+\Delta t E(x))| + 2(r+M\Delta t)|\partial_v g(x, v+\Delta t E(x))| \\
&\qquad\qquad + \Delta t L|E'(x)|\ |\partial_v g(x, v+\Delta t E(x))| \\
&\qquad \le |g|_{W^{1,\infty}(\mathbb{R}^2)} + \Delta t(2M + |E|_{W^{1,\infty}([0,L])})\|\partial_v g\|_{L^\infty(\mathbb{R}^2)} \\
&\qquad\qquad \le |g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)} + \frac{\Delta t}{2r}(2M + |E|_{W^{1,\infty}([0,L])})|g|_{W^{1,\infty}_{L,r}(\mathbb{R}^2)} \qquad (7.7.33)
\end{aligned}$$

ce qui donne (7.7.30). □

Contrôle de π par projection On en déduit aussi que la projection reste bornée vis-à-vis de l'indicateur d'erreur π précédemment défini :

Lemme 7.7.9. *On a*

$$\pi_r(\mathcal{M}, P^r_{\mathcal{M}} g) \le \frac{\max(L,2r)}{\min(L,2r)} \pi_r(\mathcal{M}, g) \qquad (7.7.34)$$

pour tout maillage adaptatif $\mathcal{M}$ et tout $g \in V$.

Preuve. Soit α une maille de $\mathcal{M}$. Elle est composée de deux demi-mailles α^+ et α^-, notées α^*, si on ne distingue pas l'une par rapport à l'autre.
On a alors :

$$|P^r_{\mathcal{M}} g|_{W^{1,\infty}_{L,r}(\check{\alpha}^*_*)} \le |g|_{W^{1,\infty}_{L,r}(\check{\alpha}^*_*)}. \qquad (7.7.35)$$

Ainsi,

$$|P^r_{\mathcal{M}} g|_{W^{1,\infty}_{L,r}(\check{\alpha})} \le \sup_{*\in\{+,-\}} |g|_{W^{1,\infty}_{L,r}(\check{\alpha}^*_*)} \le \sup_{\beta} |g|_{W^{1,\infty}_{L,r}(\check{\beta})}, \qquad (7.7.36)$$

où β parcourt les deux cellules adjacentes de α, qui sont de même niveau (elles sont nécessaires dans la majoration, que lorsqu'il y a un noeud contraint).
On obtient alors :

$$|\alpha|\ |P^r_{\mathcal{M}} g|_{W^{1,\infty}_{L,r}(\check{\alpha})} \le \sup_{\beta} |\alpha|\ |g|_{W^{1,\infty}_{L,r}(\check{\beta})} = \sup_{\beta} |\beta||g|_{W^{1,\infty}_{L,r}(\check{\beta})}, \qquad (7.7.37)$$

puisque toutes les cellules β sont de même niveau (le même que celui de α).
On obtient alors :

$$\sup_{\alpha\in\mathcal{M}} |\alpha|\ |P^r_{\mathcal{M}} g|_{W^{1,\infty}_{L,r}(\check{\alpha})} \le \pi_{L,r}(\mathcal{M}, g) := \sup_{\alpha\in\mathcal{M}} |\alpha|\ |g|_{W^{1,\infty}_{L,r}(\check{\alpha})}, \qquad (7.7.38)$$

ce qui veut dire que

$$\pi_{L,r}(\mathcal{M}, P^r_{\mathcal{M}} g) \le \pi_{L,r}(\mathcal{M}, g). \qquad (7.7.39)$$

Comme les normes $W^{1,\infty}_{L,r}$ et $W^{1,\infty}$ sont équivalentes, on obtient bien le résultat ; grâce aux inégalités (7.7.23) et (7.7.24), on a plus précisément :

$$\pi_r(\mathcal{M}, P^r_{\mathcal{M}} g) \leq \frac{1}{\min(L,2r)} \pi_{L,r}(\mathcal{M}, P^r_{\mathcal{M}} g) \leq \frac{1}{\min(L,2r)} \pi_{L,r}(\mathcal{M}, g) \leq \frac{\max(L,2r)}{\min(L,2r)} \pi_{L,r}(\mathcal{M}, g). \quad (7.7.40)$$

□

Remarque 7.7.10. *On n'aurait pu en fait ici utiliser que le caractère borné de la projection pour la norme $W^{1,\infty}$. Le lemme indique que l'erreur est au plus multipliée par une constante lorsqu'on lui applique une projection, comme il en était de même pour le transport. On pourrait affiner le résultat, puisqu'une norme équivalente est stable (la norme $W^{1,\infty}_{L,r}$, cf (7.7.38)).*
Néanmoins, comme pour l'advection, on n'a que le caractère borné, ce n'est pas réellement une amélioration. Ce qu'il faudrait, c'est pouvoir contrôler le nombre de mailles du maillage prédit qui rend stable la quantité π par l'advection, en fonction de la régularité de la solution.
Pour la courbure, on peut appliquer la même remarque (mais trouver une quantité équivalente stable par la projection semble plus complexe). Notons que la courbure de la solution exacte, c'est-à-dire la norme $W^{2,1}$ reste elle bornée et s'exprime en fonction de la norme $W^{2,1}$ de la donnée initiale (au moins dans le cas où l'on suppose que $f_0 \in W^{2,\infty}(\Omega)$), cf lemme 5.7.12.

Contrôle de la courbure maximale par projection Concernant la courbure, on a le résultat suivant :

Lemme 7.7.11. *On a une estimation du type :*

$$\mu(\mathcal{M}, P_{\mathcal{M}} g) \leq C_{r/L} \mu(\mathcal{M}, g), \quad (7.7.41)$$

pour tout $g \in V$, et tout maillage adaptatif $\mathcal{M}$.
Comme indiqué, la constante $C_{r/L}$ ne dépend que du rapport r/L.

Avant de prouver ce résultat, on donne un lemme qui permet de simplifier les calculs :

Lemme 7.7.12. *On considère un rectangle ouvert ω, ayant pour sommets a, b, c et d, avec $\overrightarrow{ab} = \overrightarrow{cd}$.*
On suppose que ses côtés sont parallèles aux axes.
Alors, on a, pour tout $g \in V$

$$|g(d) - g(c) - g(b) + g(a)| \leq \int_\omega |\partial^2_{xv} g|. \quad (7.7.42)$$

De plus, si $\tilde{\omega}$ est un domaine image de ω, par une transformation affine, alors

$$|g(d) - g(c) - g(b) + g(a)| \leq C|g|_{W^{2,1,*}(\omega)}, \tag{7.7.43}$$

la constante C ne dépendant que de la transformation affine considérée.

Preuve. On ne restreint pas la généralité en supposant que (ab) est parallèle à l'axe des x. Les coordonnées des points sont alors : $a = (a_x, a_v)$, $b = (b_x, a_v)$, $c = (a_x, c_v)$ et $d = (b_x, c_v)$. On suppose d'abord que g est infiniment dérivable (en fait de classe C^2 suffit aussi). On a alors :

$$\begin{aligned} g(b) - g(a) - g(d) + g(c) = \int_{a_x}^{b_x} \partial_x g(x, a_v)\mathrm{d}x - \int_{a_x}^{b_x} \partial_x g(x, c_v)\mathrm{d}x \\ = \int_{a_x}^{b_x} \int_{a_v}^{c_v} \partial_v \partial_x g(x, v)\mathrm{d}x\mathrm{d}v, \end{aligned} \tag{7.7.44}$$

et donc (7.7.42) en découle. Maintenant si $g \in V$, on considère l'approximation $g_n = \rho_n \star g$, avec ρ_n, une suite d' approximation de l'unité (comme dans le lemme 7.4.3).
Comme g_n est de classe C^∞, g_n vérifie (7.7.42). En passant à la limite, le membre de droite tend vers $\int_{[0,1]^2} |\partial^2_{xv} g|$ (cf lemme 7.4.3).
D'autre part, comme g est continue, g_n converge uniformément sur tout compact de $\mathbb{R}^2$. En particulier, elle converge ponctuellement aux points a, b, c et d, et donc le membre de gauche dans (7.7.42) converge bien vers $|g(d) - g(c) - g(b) + g(a)|$, ce qui donne bien (7.7.42) pour g à la limite.
Le cas de ω' s'obtient à partir de (7.7.44), en remplacant g par $g \circ \tilde{\theta}$, où $\tilde{\theta}$ est l'application affine considérée qui envoie ω sur $\tilde{\omega}$ (on calcule d'abord les dérivées partielles, on effectue ensuite un changement variable et on conclut par densité de la même manière que précédemment). □

On donne maintenant la preuve du lemme 7.7.11.

Preuve du lemme 7.7.11. On considère donc une maille $\alpha \in \mathcal{M}$.
Soit α une maille du maillage $\mathcal{M}$.
On a :

$$|P^r_{\mathcal{M}} g|_{\star(\tilde{\alpha})} \leq |P^r_{\mathcal{M}} g|_{\star(\bar{\alpha^-})} + |P^r_{\mathcal{M}} g|_{\star(\bar{\alpha^+})}. \tag{7.7.45}$$

Il suffit de montrer que pour une demi-maille α^*, on a :

$$|P^r_{\mathcal{M}} g|_{\star(\bar{\alpha^*})} \leq C_{r/L}|g|_{\star\tilde{\omega}}, \tag{7.7.46}$$

où ω est un domaine qui peut être recouvert par un nombre fini fixé de mailles de $\mathcal{M}$.
On considère donc une telle demi-maille α^* de sommets a, b et c, perpendiculaire en b, avec (ab) parallèle à l'axe des v. On ne restreint pas la généralité en supposant que α est la maille en haut à gauche de la maille mère de α, et que $\alpha^* = \alpha^-$.

On note d le troisième sommet de la demi-maille β^* de la maille $\beta \in \mathcal{M}$ qui partage le côté bc.
Soit ensuite e le point symétrique à b par rapport à (ad), f le point symétrique de b par rapport à (ac) et enfin p le symétrique de a par rapport à (fc).
On voit alors que l'union ω des losanges ouverts $(afbe)$, $(bfcd)$ et $(bfpc)$ ainsi que du carré ouvert $(abcf)$ est contenu dans 4 mailles de $\mathcal{M}$ (voir figure 7.4)
Il s'agit maintenant de montrer qu'un tel ω satisfait bien (7.7.46), pour la maille α^* ici considérée.
Tout d'abord, on peut remarquer que les losanges et le carré en question peuvent être obtenus à partir d'un rectangle ayant ses côtés parallèles aux axes, par une transformation affine qui ne fait intervenir que le rapport r/L (dans le cas du carré, c'est l'identité et pour les losanges, on peut prendre une application du genre $x \to x + vL/r$, $v \to v$).
On est donc assuré, d'après le lemme précédent appliqué à ces 4 quadrilatères, que

$$g_{ebaf} + g_{dbcf} + g_{cbpf} + g_{abfc} \leq C|g|_{\star\bar{\omega}}, \tag{7.7.47}$$

où la notation g_{xyzt} signifie :

$$g_{xyzt} := |g(x) - g(y) - g(z) + g(t)|. \tag{7.7.48}$$

On calcule maintenant $|P^r_{\mathcal{M}} g|_{\star(\bar{\alpha^*})}$.
On note Δx et Δv les longueurs respectives du côté bc et du côté ab.
La contribution pour l'arête (ab) vaut alors :

$$\Delta v \Big| \frac{g(e) - P^r_{\mathcal{M}} g(b)}{\Delta x} - \frac{P^r_{\mathcal{M}} g(b) - g(c)}{\Delta x} \Big|. \tag{7.7.49}$$

Pour l'arête (bc), il s'agit de :

$$\Delta x \Big| \frac{g(a) - P^r_{\mathcal{M}} g(b)}{\Delta v} - \frac{P^r_{\mathcal{M}} g(b) - g(d)}{\Delta v} \Big|, \tag{7.7.50}$$

et enfin pour l'arête (ac), on trouve :

$$\sqrt{\Delta x^2 + \Delta v^2} \Big(\frac{1}{\Delta x} + \frac{1}{\Delta v} \Big) \Big| P^r_{\mathcal{M}} g(f) - g(a) - g(c) + g(b) \Big|. \tag{7.7.51}$$

Or on a :

$$\sqrt{\Delta x^2 + \Delta v^2} \leq \Delta x + \Delta v, \tag{7.7.52}$$

et

$$(\Delta x + \Delta v)\Big(\frac{1}{\Delta x} + \frac{1}{\Delta v} \Big) \leq 2 + \frac{L}{2r} + \frac{2r}{L} \leq (2 + 1/2 + 2)\sigma_{L,r}, \tag{7.7.53}$$

en rappelant que $\sigma_{L,r} = max(L/r, r/L) \geq 1$. On obtient alors :

$$\begin{aligned} |P_{\mathcal{M}} g|_{\star(\bar{\alpha^*})} \leq C\sigma_{L,r} \big(& |g(e) - 2P_{\mathcal{M}} g(b) + g(c)| \\ & + |g(a) - 2P_{\mathcal{M}} g(b) + g(d)| + |P_{\mathcal{M}} g(f) - g(a) - g(c) + g(b)|, \end{aligned} \tag{7.7.54}$$

et on peut remplacer $C\sigma_{L,r}$ par une nouvelle constante qui dépend cette fois (mais qui dépend que) de r/L.
D'autre part, on a toujours :

$$|P^r_{\mathcal{M}} g(b) - g(b)| \leq \frac{1}{2}|g(a) + g(d) - 2g(b)|, \tag{7.7.55}$$

en constatant que la quantité de gauche est nulle s'il y a un noeud contraint et vaut le terme de droite dans le cas contraire.
De même, on a :

$$|P_{\mathcal{M}} g(f) - g(f)| \leq \frac{1}{2}|g(a) + g(p) - 2g(f)| \tag{7.7.56}$$

On obtient alors, avec les notations précédentes, en utilisant (7.7.54),(7.7.55) et (7.7.56) :

$$|P_{\mathcal{M}} g|_{\star(\bar{\tilde{\alpha}^*})} \leq C_{L/r}(g_{ebbc} + g_{abbd} + g_{facb} + g_{abbd} + g_{affp}). \tag{7.7.57}$$

Il reste donc à vérifier que :

$$g_{ebbc} + g_{abbd} + g_{facb} + g_{affp} \leq C(g_{ebaf} + g_{dbcf} + g_{cbpf} + g_{abfc}), \tag{7.7.58}$$

d'après (7.7.47).
On a déjà :

$$g_{facb} = g_{abfc}, \tag{7.7.59}$$

et avec l'inégalité triangulaire, on obtient aussi en introduisant $g(b)$ et $g(c)$:

$$g_{affp} \leq g_{afbc} + g_{bcfp} = g_{abfc} + g_{cbpf}. \tag{7.7.60}$$

Or pour tous $x, y \in \mathbb{R}$, on a

$$|x| + |y| = |\frac{x+y}{2} + \frac{x-y}{2}| + |\frac{x+y}{2} - \frac{x-y}{2}| \leq |x+y| + |x-y| \tag{7.7.61}$$

Ainsi, on obtient :

$$g_{ebbc} + g_{abbd} \leq g_{eadc} + A \leq g_{ebaf} + g_{dbcf}, \tag{7.7.62}$$

avec

$$A := \big|g(e) - 2g(b) + g(c) + g(a) - 2g(b) + g(d)\big| \tag{7.7.63}$$

En ordonnant différemment les termes de A et en introduisant $g(f)$, on obtient

$$\begin{aligned} A = \big|&g(e) - g(b) - g(a) + g(f) + g(d) - g(b) - g(c) + g(f) \\ &+ 2(g(a) - g(b) - g(f) + g(c))\big| \leq g_{ebaf} + g_{dbcf} + 2g_{abfc}. \end{aligned} \tag{7.7.64}$$

Ainsi, (7.7.58) est bien vérifié, ce qui entraîne (7.7.46), grâce à (7.7.57) et (7.7.47). Finalement, on a (7.7.41). □

Remarque 7.7.13. *On a en fait montré plus précisément que pour toute maille $\alpha \in \mathcal{M}$, on a*

$$|P^r_{\mathcal{M}} g|_{\star(\bar{\tilde{\alpha}})} \leq C_{L/r}|g|_{\star(\tilde{\omega})}, \tag{7.7.65}$$

où ω est l'union de la maille α, ses deux soeurs qui partagent un côté, et les deux mailles extérieures à la maille mère de α qui partagent un côté avec α.

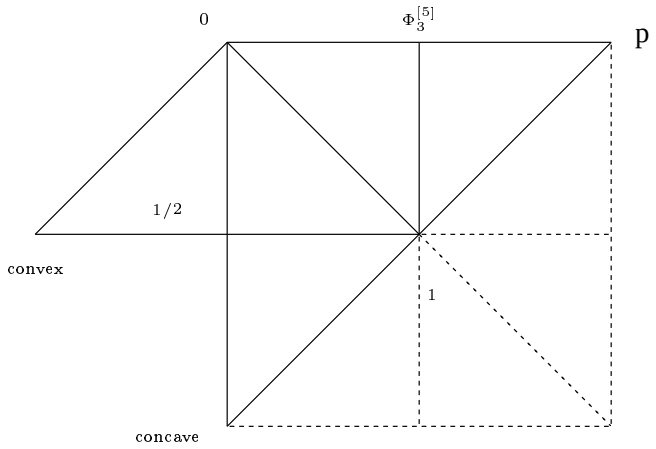


Fig. 7.4 – La demi-maille (abc) et le domaine $\omega = (afpcde)$

7.8 Le schéma numérique

On peut maintenant définir le schéma numérique adaptatif. Il consiste en une demi-advection en x suivi d'une advection en v, et enfin d'une demi-advection en x. Entre chaque étape, on projette sur un maillage adaptatif.
En se référant au chapitre 5 pour les notations, le schéma s'écrit alors

$$\begin{aligned} f_1^{n+1} &:= \tilde{\mathcal{I}}_x^{n+1/2} f^n, \\ f_2^{n+1} &:= \tilde{\mathcal{I}}_{v,n+1}^{E_h} f_1^{n+1}, \\ f^{n+1} &:= \tilde{\mathcal{I}}_x^{n+1} f_2^{n+1}. \end{aligned}$$

Il s'agit maintenant de préciser les opérateurs adaptatifs Π_1^{n+1}, Π_2^{n+1} et Π_1^{n+1}.
On reprend la suite r_n qui contrôle l'évolution du support en vitesse de la solution.
Pour simplifier, on choisit ici :

$$r_n := Q(t^n) = Q(0) + 2Lt^n. \tag{7.8.1}$$

Opérateurs de projection.
On définit

- $\Pi_1^{n+1} = P_{\mathcal{M}_1^{n+1}}^{r_n}$,

  avec $\mathcal{M}_1^{n+1} := \mathbb{T}(\tilde{\mathcal{M}}^n, \mathcal{I}_x, r_n)$ et $\tilde{\mathcal{M}}^n = \mathbb{A}(M^n, f^n, TOL_1^n, r_n)$
- $\Pi_2^{n+1} = P_{\mathcal{M}_2^{n+1}}^{r_{n+1}}$,

  avec $\mathcal{M}_2^{n+1} := \mathbb{T}(\tilde{\mathcal{M}}_1^{n+1}, \mathcal{I}_{v,n+1}^{E_h}, r_n)$,

  et $\tilde{\mathcal{M}}_1^{n+1} = \mathbb{A}(M_1^{n+1}, f_1^{n+1}, TOL_2^n, r_n)$
- $\Pi_3^{n+1} = P_{\mathcal{M}^{n+1}}^{r_{n+1}}$,

avec $\mathcal{M}^{n+1} := \mathbb{T}(\tilde{\mathcal{M}}_2^{n+1}, \mathcal{T}_x, r_{n+1})$,

et $\tilde{\mathcal{M}}_2^{n+1} = \mathbb{A}(M_2^{n+1}, f_2^{n+1}, TOL_3^n, r_{n+1})$,

les réels TOL_1^n, TOL_2^n et TOL_3^n seront choisis par la suite (assez petits).

Initialisation. A partir de la donnée initiale f_0, on construit une approximation

$$f^0 := \Pi_0 f_0, \tag{7.8.2}$$

où $\Pi_0 := P_{\mathcal{M}}^{r_0}$, pour un maillage (adaptatif gradué) que l'on va préciser.
On sait que $f_0 \in W^{2,1}(\Omega)$. En reprenant la proposition (7.4.1), on a à notre disposition la majoration de l'erreur d'interpolation suivante :

$$\|f_0 - \Pi f_0\|_{L^\infty(\Omega)} \leq C|f|_{W^{2,1}(\Omega)}. \tag{7.8.3}$$

L'opérateur Π_0 est choisi de telle manière à ce que pour chaque maille α, on ait :

$$|f|_{W^{2,1}(\check{\alpha})} \leq TOL_0, \tag{7.8.4}$$

Le réel TOL_0 sera aussi choisi par la suite assez petit.
Ceci implique alors :

$$\|f_0 - \Pi_0 f_0\|_{L^\infty(\Omega)} \leq CTOL_0. \tag{7.8.5}$$

Notons que l'on a aussi une majoration (comme pour les éléments de V, en regardant la démonstration des lemmes 7.7.3 et 7.7.5)

$$|\Pi_0 f_0|_{W^{1,\infty}} \leq C|f_0|_{W^{1,\infty}}. \tag{7.8.6}$$

Champs numériques de Poisson. En reprenant les notations du chapitre 5, on rappelle que l'on a défini le champ numérique

$$E^n(x) = \int_0^L K(x,y)\Big(\int_{\mathbb{R}} \tilde{T}_x^{n+1/2} f^n(y,v)\mathrm{d}v - 1\Big)\mathrm{d}y. \tag{7.8.7}$$

On spécifie maintenant le champ E_h^n. On remarque que E^n n'est pas L–périodique, puisque le schéma n'est pas conservatif.
On a plus précisément

$$E^n(L) - E^n(0) = \int_0^L \Big(\int_{\mathbb{R}} \tilde{T}_x^{n+1/2} f^n(x,v)\mathrm{d}v - 1\Big)\mathrm{d}x, \tag{7.8.8}$$

en utilisant l'identité $K(L,y) - K(0,y) = y/L - (y/L - 1) = 1$.
On définit ainsi d'abord le champ $\tilde{E}^n$ par

$$\tilde{E}^n(x) := E^n(x) + \frac{x}{L}(E^n(0) - E^n(L)), \tag{7.8.9}$$

pour $x \in [0, L]$ et on prolonge $\tilde{E}^n$ par périodicité sur $\mathbb{R}$.
Le champ E_h^n est alors défini comme l'approximation linéaire par morceaux de $\tilde{E}^n$, obtenue par interpolation affine de $\tilde{E}^n$ sur une discrétisation par des intervalles de taille h de l'intervalle $[0, L]$; le champ E_h^n se prolonge alors bien par périodicité en une fonction (continue) affine par morceaux sur $\mathbb{R}$.

Pas de temps. On suppose que le pas de temps est assez petit pour pouvoir appliquer le lemme 7.6.2, en prenant $G(v) = v\Delta t/2$, si $\mathcal{T} = \mathcal{T}^x$, et $G(x) = E_h^n(x)\Delta t$, si $\mathcal{T} = \mathcal{T}^v$, ce qui donne alors :

Lemme 7.8.1. *Sous la condition :*

$$r_{n+1}/r_n + \frac{L}{2r_n}\Delta t|E_h^n|_{W^{1,\infty}([0,L])} \leq 3/2, \tag{7.8.10}$$

et

$$1 + \frac{r_{n+1}}{L}\Delta t \leq 3/2, \tag{7.8.11}$$

on a :

$$\mu_{r_n}(\mathcal{M}_1^{n+1}, \mathcal{T}_x f^n) \leq 7(1+\Delta t/2)^2 \mu_{r_n}(\tilde{\mathcal{M}}^n, f^n), \tag{7.8.12}$$

$$\mu_{r_{n+1}}(\mathcal{M}_2^{n+1}, \mathcal{T}_v^{E_h^n} f_1^{n+1}) \leq 7\Big(1+\Delta t|E_h^n|_{W^{1,\infty}([0,L])}\Big)^2 \mu_{r_n}(\tilde{\mathcal{M}}_1^{n+1}, f_1^{n+1}) + 4\Delta t|\tilde{E}^n|_{W^{2,\infty}([0,L])}\pi_{r_n}(\tilde{\mathcal{M}}_1^{n+1}, f_1^{n+1}), \tag{7.8.13}$$

et

$$\mu_{r_{n+1}}(\mathcal{M}^{n+1}, \mathcal{T}_x f_2^{n+1}) \leq 7(1+\Delta t/2)^2 \mu_{r_{n+1}}(\tilde{\mathcal{M}}_2^{n+1}, f_2^{n+1}). \tag{7.8.14}$$

Et on a aussi :

$$\pi_{r_n}(\mathcal{M}_1^{n+1}, \mathcal{T}_x f^n) \leq 4(1+\Delta t/2)\pi_{r_n}(\tilde{\mathcal{M}}^n, f^n), \tag{7.8.15}$$

$$\pi_{r_{n+1}}(\mathcal{M}_2^{n+1}, \mathcal{T}_v^{E_h^n} f_1^{n+1}) \leq 4\Big(1+\Delta t|E_h^n|_{W^{1,\infty}([0,L])}\Big)\pi_{r_n}(\tilde{\mathcal{M}}^n, f_1^{n+1}), \tag{7.8.16}$$

et

$$\pi_{r_{n+1}}(\mathcal{M}^{n+1}, \mathcal{T}_x f_2^{n+1}) \leq 4(1+\Delta t/2)\pi_{r_{n+1}}(\tilde{\mathcal{M}}_2^{n+1}, f_2^{n+1}). \tag{7.8.17}$$

Erreur de périodicité du champ. Le lemme suivant permet de contrôler l'erreur donné par la modification précédente.

Lemme 7.8.2. *On a l'estimation :*

$$\big|E^n(L) - E^n(0)\big| \leq C_T(TOL_1^n + \|e_n\|_{L^\infty(\Omega)}). \tag{7.8.18}$$

Preuve. On remarque d'abord que

$$\int_0^L \int_{\mathbb{R}} \mathcal{T}_x f(t^n, x, v)\mathrm{d}x\mathrm{d}v = \int_0^L \int_{\mathbb{R}} f(t^n, x - v\Delta t/2, v)\mathrm{d}x\mathrm{d}v = \int_0^L \int_{\mathbb{R}} f(t^n, x, v)\mathrm{d}x\mathrm{d}v = L. \tag{7.8.19}$$

comme f est L–périodique en x, et comme f conserve la masse.
En injectant (7.8.19) dans (7.8.8), on obtient :

$$E^n(L) - E^n(0) = \int_0^L \int_{\mathbb{R}} \Big(\tilde{\mathcal{T}}_x^{n+1/2} f^n(x,v) - \mathcal{T}_x f(t^n, x, v)\Big)\mathrm{d}v\mathrm{d}x. \tag{7.8.20}$$

Or, on a

$$\tilde{\mathcal{T}}_x^{n+1/2} f^n(x,v) - \mathcal{T}_x f(t^n,x,v) = (I-\Pi_1^{n+1})\mathcal{T}_x f^n(x,v) + \Pi_1^{n+1}\mathcal{T}_x(f^n - f). \quad (7.8.21)$$

Donc, on en déduit que

$$\begin{aligned} |E^n(L) - E^n(0)| &\leq 2r_n L \|\tilde{\mathcal{T}}_x^{n+1/2} f^n(x,v) - \mathcal{T}_x f(t^n)\|_{L^\infty(\Omega)} \\ &\leq C_T(\|(I-\Pi_1^{n+1})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} \\ &\quad + \|\Pi_1^{n+1}\mathcal{T}_x(f^n - f(t^n))\|_{L^\infty(\Omega)}). \end{aligned} \quad (7.8.22)$$

On obtient alors, en appliquant successivement le corollaire 7.4.11, puis le lemme 7.8.1,

$$\begin{aligned} \|(I-\Pi_1^{n+1})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} &\leq C\sigma_{L,r_n}\mu(\mathcal{M}_1^{n+1}, \mathcal{T}_x f^n) \\ &\leq 7C\sigma_{L,r_n}(1+\Delta t/2)^2\mu(\tilde{\mathcal{M}}^n, f^n) \\ &\leq C_T TOL_1^n, \end{aligned} \quad (7.8.23)$$

la dernière inégalité provenant de la définition de $\tilde{\mathcal{M}}^n$.
D'autre part, on a également, en appliquant les lemmes 7.7.1 et 7.7.6

$$\|\Pi_1^{n+1}\mathcal{T}_x(f^n - f)\|_{L^\infty(\Omega)} \leq \|\mathcal{T}_x(f^n - f)\|_{L^\infty(\Omega)} \leq \|f^n - f\|_{L^\infty(\Omega)}, \quad (7.8.24)$$

ce qui donne finalement (7.8.18). □

Estimations sur E^n. En utilisant l'équation de Poisson, on obtient dans un premier temps des estimations sur le champ E^n.

Lemme 7.8.3. *L'approximation numérique E^n vérifie :*

$$E^n \in W^{2,\infty}(\Omega). \quad (7.8.25)$$

De plus, on a l'estimation :

$$\|E^n\|_{L^\infty([0,L])} \leq L(1 + 2r_n\|f_0\|_{L^\infty(\Omega)}) \leq C_T \quad (7.8.26)$$

$$|E^n|_{W^{1,\infty}([0,L])} \leq 2r_n\|f_0\|_{L^\infty(\Omega)} + 1 \leq C_T \quad (7.8.27)$$

et

$$|E^n|_{W^{2,\infty}([0,L])} \leq C_T |f_n|_{W^{1,\infty}_{L,r_n}(\Omega)}. \quad (7.8.28)$$

Preuve. De l'expression (7.8.7), on obtient :

$$|E^n(x)| \leq L\|K\|_{L^\infty[0,L]}(1+2r_n)\|f^n\|_{L^\infty(\Omega)}), \quad (7.8.29)$$

ce qui donne (7.8.26), puisque $\|K\|_{L^\infty([0,L])} \leq 1$.
Ensuite, on a :

$$\partial_x E^n(x) = \int_{-r_n}^{r_n} \tilde{T}_x^{n+1/2} f^n(x,v)\mathrm{d}v - 1, \tag{7.8.30}$$

et donc :

$$|\partial_x E^n(x)| \leq 2r_n \|\tilde{T}_x^{n+1/2} f^n\|_{L^\infty(\Omega)} + 1. \tag{7.8.31}$$

or, en appliquant les propriétés de stabilité L^∞ des opérateurs de transports et de projection des lemmes 7.7.1 et 7.7.6, on obtient :

$$\|\tilde{\mathcal{T}}_x^{n+1/2} f^n\|_{L^\infty(\Omega)} \leq \|f_0\|_\infty, \tag{7.8.32}$$

ce qui donne (7.8.27). Comme $\partial_x \tilde{\mathcal{T}}^{n+1/2} f^n(x,v)$ est constante sur chaque, en intégrant cette quantité suivant v, on obtient encore une fonction constante par mailles et ainsi $E^n \in W^{2,\infty}([0,L])$, et on a :

$$\begin{aligned}
|\partial_x^2 E^n(x)| = \Big|\int_{-r_n}^{r_n} \partial_x \tilde{\mathcal{T}}^{n+1/2} f^n(x,v)\mathrm{d}v\Big| & \\
&\leq \frac{2r_n}{L} |\tilde{\mathcal{T}}^{n+1/2} f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} \\
&\leq \frac{2r_n}{L} |\mathcal{T}^{n+1/2} f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} \\
&\leq \frac{2r_n}{L}(1 + r_n\Delta t)|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)},
\end{aligned} \tag{7.8.33}$$

en appliquant cette fois-ci les propriétés de stabilité $W^{1,\infty}$ des opérateurs de transport et de projection du lemme 7.7.8, et en utilisant le fait que

$$|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} = |f^n|_{W^{1,\infty}_{L,r_n}(\mathbb{R}^2)}, \tag{7.8.34}$$

comme la solution numérique f^n est périodique et à support dans $[-r_n, r_n]$.
Tout ceci donne alors bien (7.8.28). □

Estimations sur $\tilde{E}^n$ et E_h^n. On en déduit ensuite des estimations sur le champ $\tilde{E}^n$ $W^{2,\infty}$ et périodique, ainsi que sur son approximation affine par morceaux E_h^n également périodique.

Lemme 7.8.4. *Le champ $\tilde{E}^n$ appartient à $W^{2,\infty}([0,L])$, et on a :*

$$\|E_h^n\|_{L^\infty([0,L])} \leq \|\tilde{E}^n\|_{L^\infty([0,L])} \leq C_T(1 + TOL_1^n + \|e_n\|_{L^\infty(\Omega)}), \tag{7.8.35}$$

$$|E_h^n|_{W^{1,\infty}([0,L])} \leq |\tilde{E}^n|_{W^{1,\infty}([0,L])} \leq C_T(1 + TOL_1^n + \|e_n\|_{L^\infty(\Omega)}), \tag{7.8.36}$$

et

$$|\tilde{E}^n|_{W^{2,\infty}([0,L])} \leq C_T |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} \tag{7.8.37}$$

Preuve. Le champ E^n appartient à $W^{2,\infty}([0,L])$ et le champ $\tilde{E}^n$ et la somme de E^n et d'un champ affine sur $[0,L]$ (cf 7.8.38), et donc $\tilde{E}^n$ appartient également à $W^{2,\infty}([0,L])$. Les deux premières inégalités de gauche proviennent du fait que l'interpolation affine en dimension 1 est L^∞–stable.
Pour les autres inégalités, de la définition de $\tilde{E}^n$, on applique l'inégalité triangulaire à

$$\tilde{E}^n = E^n + (\frac{x}{L}(E^n(L) - E^n(0))) \tag{7.8.38}$$

et on applique les lemmes 7.8.3 et 7.8.2, en remarquant que la norme $W^{2,\infty}$ du deuxième terme est nulle. □

Estimation de l'erreur $E^n - E_h^n$ Le lemme suivant permet alors de donner une borne pour l'erreur entre l'approximation du champ numérique de Poisson E^n, et l'approximation utilisée E_h^n.

Lemme 7.8.5. *On a :*

$$\|E_h^n - E^n\|_{L^\infty([0,L])} \leq C_T(h^2|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} + TOL_1^n + \|e_n\|_{L^\infty(\Omega)}) \tag{7.8.39}$$

Preuve. On écrit :

$$E_h^n - E^n = (E_h^n - \tilde{E}^n) + (\tilde{E}^n - E^n). \tag{7.8.40}$$

Le second terme est une erreur d'interpolation, ainsi on a :

$$|\tilde{E}^n - E^n|_{L^\infty(\Omega)} \leq Ch^2|\tilde{E}^n|_{W^{2,\infty}([0,L])} \leq C_T h^2 |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}, \tag{7.8.41}$$

la dernière inégalité provenant du lemme 7.8.4, et pour le premier terme, on majore :

$$\begin{aligned} |E_h^n(x) - \tilde{E}^n(x)| = |\frac{x}{L}(E^n(L) - E^n(0))| &\leq |E^n(L) - E^n(0)| \\ &\leq C_T(TOL_1^n + \|e_n\|_{L^\infty(\Omega)}), \end{aligned} \tag{7.8.42}$$

en utilisant cette fois-ci le lemme 7.8.2, et tout ceci donne le résultat. □

Décomposition de l'erreur. On étudie l'erreur en norme L^∞, et on utilise la décomposition introduite dans le chapitre 5 :

$$e^{n+1} = e_1^{n+1} + \varepsilon_2^{n+1} + \varepsilon_3^{n+1} + \varepsilon_4^{n+1} + \varepsilon_5^{n+1} + \varepsilon_6^{n+1} \tag{7.8.43}$$

On a alors :

$$\|e_1^{n+1}\|_{L^\infty} = \|f(t^n) - \mathcal{T}_x\mathcal{T}_v\mathcal{T}_x f(t^n)\|_{L^\infty} \leq C_T \Delta t^3,$$

d'après le lemme 5.8.2.
On a ensuite :

$$\|\varepsilon_2^{n+1}\|_{L^\infty(\Omega)} = \|\mathcal{T}_x\mathcal{T}_{v,n+1}^E\mathcal{T}_x(f(t^n)-f^n)\|_{L^\infty(\Omega)} \leq \|f(t^n)-f^n\|_{L^\infty(\Omega)} = \|e^n\|_{L^\infty(\Omega)}. \quad (7.8.44)$$

Le terme suivant est l'erreur de couplage avec l'équation de Poisson.

$$\|\varepsilon_3^{n+1}\|_{L^\infty(\Omega)} = \|\mathcal{T}_x(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} \leq \|(\mathcal{T}_{v,n+1}^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} \quad (7.8.45)$$

Les derniers termes proviennent des erreurs de projections sur les différents maillages :

$$\|\varepsilon_4^{n+1}\|_{L^\infty(\Omega)} = \|\mathcal{T}_x\mathcal{T}_{v,n+1}^{E_h}(I-\Pi_1^{n+1})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} \leq \|(I-\Pi_1^{n+1})\mathcal{T}_x f^n\|_{L^\infty(\Omega)}, \quad (7.8.46)$$

$$\|\varepsilon_5^{n+1}\| = \|\mathcal{T}_x(I-\Pi_2^{n+1})\mathcal{T}_{v,n+1}^{E_h} f_1^{n+1}\|_{L^\infty(\Omega)} \leq \|(I-\Pi_2^{n+1})\mathcal{T}_{v,n+1}^{E_h} f_1^{n+1}\|_{L^\infty(\Omega)} \quad (7.8.47)$$

et

$$\|\varepsilon_6^{n+1}\|_{L^\infty(\Omega)} = \|(I-\Pi_3^{n+1})\mathcal{T}_x f_2^{n+1}\|_{L^\infty(\Omega)}. \quad (7.8.48)$$

Erreurs de projection. Le lemme suivant traite des erreurs de projection sur les différents maillages adaptatifs.

Lemme 7.8.6. *On a les estimations suivantes :*

$$\|(I-\Pi_1^{n+1})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} \leq C_T TOL_1^n \quad (7.8.49)$$

$$\|(I-\Pi_3^{n+1})\mathcal{T}_x f_2^{n+1}\|_{L^\infty(\Omega)} \leq C_T TOL_3^n. \quad (7.8.50)$$

et

$$\|(I-\Pi_2^{n+1})\mathcal{T}_{v,n+1}^{E_h} f_1^{n+1}\|_{L^\infty(\Omega)} \leq C_T TOL_2^n \Big(1+\Delta t(1+TOL_1^n+\|e_n\|_{L^\infty(\Omega)})^2 + \Delta t|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}\Big) \quad (7.8.51)$$

Preuve. La première estimation a déjà été montrée dans la preuve du lemme 7.8.2, et l'inégalité (7.8.50) se montre de la même manière.
Pour (7.8.51), on a d'abord

$$\|(I-\Pi_2^{n+1})\mathcal{T}_{v,n+1}^{E_h} f_1^{n+1}\|_{L^\infty(\Omega)} \leq C\mu_{r_{n+1}}(\mathcal{M}_2^{n+1}, \mathcal{T}_v^{E_h^n} f_1^{n+1}), \quad (7.8.52)$$

puis en utilisant le lemme 7.8.1, on a d'autre part

$$\mu_{r_{n+1}}(\mathcal{M}_2^{n+1}, \mathcal{T}_v^{E_h^n} f_1^{n+1}) \leq 7\Big(1+\Delta t|E_h^n|_{W^{1,\infty}([0,L])}\Big)^2 TOL_2^n + 4\Delta t|\tilde{E}^n|_{W^{2,\infty}([0,L])} TOL_2^n, \quad (7.8.53)$$

comme $\mu_{r_n}(\tilde{\mathcal{M}}_1^{n+1}, f_1^{n+1}) \leq TOL_2^n$ et $\pi_{r_n}(\tilde{\mathcal{M}}_1^{n+1}, f_1^{n+1}) \leq TOL_2^n$, par définition de $c\tilde{M}_1^n$. En utilisant alors le lemme 7.8.4 pour les termes $|\tilde{E}^n|_{W^{2,\infty}([0,L])}$ et E_h^n, on obtient bien le résultat. □

Couplage avec Poisson.

Lemme 7.8.7. *On a l'estimation :*

$$\|(\mathcal{T}_v^E - \mathcal{T}_{v,n+1}^{E_h})\mathcal{T}_x f^n\|_{L^\infty(\Omega)} \leq C_T |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} \Delta t \\ \big(\|e^n\|_{L^\infty(\Omega)} + TOL_1^n + h^2 |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}\big). \quad (7.8.54)$$

Preuve. En suivant toujours la décomposition du chapitre 5, on écrit d'abord :

$$\begin{aligned}
&\|E - E^n\|_{L^\infty([0,L])} \\
&\quad = \| \int_0^L K(x,y)([\int_{\mathbb{R}} T_x(f(t^n) - f^n) + (I - \Pi_1^{n+1})T_x f^n](y,v))\mathrm{d}v - 1)\mathrm{d}y\|_{L^\infty([0,L])} \\
&\qquad \leq \int_0^L \|K\|_{L^\infty} \big(\max(Q(\mathcal{T}_x f(t^n)), Q(\mathcal{T}_x f^n)) \\
&\qquad (\|\mathcal{T}_x(f(t^n) - f^n)\|_{L^\infty([0,L])} + \|(I - \Pi_1^{n+1})\mathcal{T}_x f(t^n))\big)\mathrm{d}y \\
&\qquad \leq L \max(Q(\mathcal{T}_x f(t^n)), Q(\mathcal{T}_x f^n)) \\
&\qquad (\|f(t^n) - f^n\|_{L^\infty(\Omega)} + \|(I - \Pi_1^{n+1})\mathcal{T}_x f^n\|_{L^\infty(\Omega)}) \quad (7.8.55)
\end{aligned}$$

Or, d'après le lemme 5.7.12, on a l'estimation suivante sur l'évolution du support en vitesse de la solution exacte :

$$Q(\mathcal{T}_x f(t^n)) \leq Q(0) + 2Lt^n \leq Q(0) + 2LT, \quad (7.8.56)$$

et on a aussi :

$$Q(\mathcal{T}_x f^n) \leq R. \quad (7.8.57)$$

Ainsi, on a :

$$\|E_{\mathcal{T}_x f(t^n)} - E^n\|_{L^\infty([0,L])} \leq C_T \|e^n\|_{L^\infty(\Omega)} + C_T TOL_1^n, \quad (7.8.58)$$

en utilisant au passage le lemme 7.8.6.
D'autre part, on a vu au lemme 7.8.5 que

$$\|E_h^n - E^n\|_{L^\infty([0,L])} \leq C_T(h^2 |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} + TOL_1^n + \|e_n\|_{L^\infty(\Omega)}). \quad (7.8.59)$$

En appliquant l'inégalité triangulaire, on obtient donc :

$$\|E_{\mathcal{T}_x f(t^n)} - E_h^n\|_{L^\infty([0,L])} \leq C_T(\|e^n\|_{L^\infty(\Omega)} + TOL_1^n + h^2 |f^n|_{W^{1,\infty}(\Omega)}). \quad (7.8.60)$$

0n obtient alors :

$$\begin{aligned}
|(\mathcal{T}_v^E - \mathcal{T}_v^{E_h^n})\mathcal{T}_x f^n(x,v)| &= |\ \mathcal{T}_x f^n(x, v - \Delta t E_{\mathcal{T}_x f(t^n)}(x)) - \mathcal{T}_x f^n(x, v - \Delta t E_h^n(x))| \\
&\leq |\mathcal{T}_x f^n|_{W^{1,\infty}_{L,r_n}(\Omega)} \frac{\Delta t}{2r_n} \big| E_{\mathcal{T}_x f(t^n)}(x) - E_h^n(x)\big|, \quad (7.8.61)
\end{aligned}$$

et comme $|\mathcal{T}_x f^n|_{W^{1,\infty}_{L,r_n}} \leq (1 + r_n \Delta t)|f^n|_{W^{1,\infty}_{L,r_n}}$ d'après le lemme (7.7.8), l'estimation (7.8.54) en découle. □

Estimation sur $|f^n|_{W^{1,\infty}(\Omega)}$. Les estimations précédentes font intervenir la norme $W^{1,\infty}_{L,r_n}$ de la solution numérique.
On n'arrive ici pas à avoir une borne directement indépendante de n, pour cette quantité ; néanmoins, le lemme suivant donne une information sur la stabilité de l'évolution de cette norme ; notons qu'à ce moment, la stabilité de l'opérateur de projection (et d'advection) du lemme 7.7.8 est cruciale.

Lemme 7.8.8. *On a :*

$$|f^{n+1}|_{W^{1,\infty}_{L,r_{n+1}}(\Omega)} \leq \Big(1 + C_T\Delta t(1 + \|e^n\|_{L^\infty(\Omega)} + TOL_1^n)\Big)|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}. \tag{7.8.62}$$

Preuve. On applique le lemme 7.7.8, et on obtient :

$$|f^{n+1}|_{W^{1,\infty}_{L,r_{n+1}}} \leq C_n|f^n|_{W^{1,\infty}_{L,r_n}}, \tag{7.8.63}$$

avec

$$C_n = (1 + r_{n+1}\Delta t)(1 + \frac{\Delta t}{2r_n}(4L\Delta t + |E_h^n|_{W^{1,\infty}([0,L])})(1 + r_n\Delta t). \tag{7.8.64}$$

Or, on sait d'après le lemme 7.8.4 que

$$|E_h^n|_{W^{1,\infty}([0,L])} \leq C_T(1 + TOL_1^n + \|e_n\|_{L^\infty(\Omega)}). \tag{7.8.65}$$

On en déduit donc que

$$C_n \leq C_T(1 + \|e^n\|_{L^\infty(\Omega)}TOL_1^n), \tag{7.8.66}$$

ce qui donne le résultat. □

Bilan. En rassemblant toutes les inégalités précédentes, on obtient :

$$\begin{aligned}\|e^{n+1}\|_{L^\infty(\Omega)} \leq\ & C_T\Delta t^3 \\ & + \|e^n\|_{L^\infty(\Omega)} \\ & + C_T(TOL_1^n + TOL_3^n) \\ & + C_T TOL_2^n\Big((1 + \Delta t[1 + TOL_1^n\|e^n\|_{L^\infty(\Omega)}])^2 + |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}\Big) \\ & + C_T|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}\Delta t(\|e^n\|_{L^\infty(\Omega)} + TOL_1^n + h^2|f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}).\end{aligned} \tag{7.8.67}$$

On note maintenant $u_n := \|e^n\|_{L^\infty(\Omega)}$, et $v_n := |f^n|_{W^{1,\infty}_{L,r_n}(\Omega)}$.
On suppose que TOL_1^n, TOL_2^n et TOL_3^n sont inférieurs à 1 et que $TOL_1^n + TOL_2^n + TOL_3^n + TOL_0 \leq TOL$.
Supposons maintenant que $u_i \leq 1$, pour $0 = 1, \ldots, n$.
En reprenant le lemme 7.8.8, on obtient

$$v_{n+1} \leq (1 + C_T\Delta t)v_n \leq (1 + C_T\Delta t)^n v_0. \tag{7.8.68}$$

Or, on a toujours $n\Delta t \leq T$, donc

$$v_{n+1} \leq (1 + C_T\Delta t)^{T/\Delta t}v_0 \leq C_T v_0, \tag{7.8.69}$$

car le membre du milieu admet une limite lorsque Δt tend vers zéro.
Or $v_0 \leq C|f_0|_{W^{1,\infty}(\Omega)}$, d'après (7.8.6).
On est donc assuré que

$$v_{n+1} \leq C_T, \tag{7.8.70}$$

et on a de même :

$$v_i \leq C_T, \quad \text{pour} \quad i = 0, \ldots, n. \tag{7.8.71}$$

La formule 7.8.67 se simplifie, alors, pour donner :

$$u_{n+1} \leq u_n(1 + C_T\Delta t) + C_T(\Delta t^3 + TOL + \Delta t h^2) \tag{7.8.72}$$

On obtient alors :

$$u_{n+1} \leq (1 + C_T\Delta t)^{n+1}(u_0 + C_T\frac{\Delta t^3 + TOL + \Delta t h^2}{\Delta t}) \\ \leq C_T(u_0 + \Delta t^2 + \frac{TOL}{\Delta t} + h^2) \tag{7.8.73}$$

comme à nouveau devant $(1 + C_T\Delta t)^{n+1} \leq (1 + C_\Delta t)^{T/\Delta t+1}$ est borné lorsque Δt tend vers 0. Pour le terme, initial, on a aussi, d'après (7.8.5)

$$u_0 \leq CTOL_0. \tag{7.8.74}$$

et on a donc

$$u_{n+1} \leq C_T(TOL_0 + \Delta t^2 + \frac{TOL}{\Delta t} + h^2) \tag{7.8.75}$$

Comme $TOL_0 \leq T/\Delta t TOL = C_T\frac{TOL}{\Delta t}$, on a finalement :

$$u_{n+1} \leq C_T(\Delta t^2 + \frac{TOL}{\Delta t} + h^2). \tag{7.8.76}$$

Maintenant, on choisit TOL et h assez petits tels que le membre de droite soit plus petit que 1 (et ceci indépendamment de n). Notons les TOL^0 et h^0.
Par récurrence, on a alors $u_n \leq 1$ pour tout n, et on a :

$$u_n \leq C_T(\Delta t^2 + \frac{TOL}{\Delta t} + h^2), \tag{7.8.77}$$

pour tous les réels $TOL \leq TOL^0$ et $h \leq h^0$.
Toutes ces estimations sont valables si le pas de temps est assez petit.
La condition s'écrit alors sous la forme

$$\Delta t \leq C_T, \tag{7.8.78}$$

où la constante C_T peut être explicitée en fonction de $Q(0)$,L, $\|f_0\|_{W^{1,\infty}(\Omega)}$.

Conclusion. On obtient finalement le théorème suivant :

Théorème 7.8.1. *Il existe une constante C_T, des réels $\varepsilon_0, h_0, \Delta t_0$ ne dépendant que de $Q(0)$,L, $\|f_0\|_{W^{1,\infty}(\Omega)}$ et $|f_0|_{W^{2,1}(\Omega)}$ tels que pour tout $0 < \varepsilon \leq \varepsilon_0$, $0 < h \leq h_0$, et $0 < \Delta t \leq \Delta t_0$, en prenant $0 < TOL_0 + TOL_1 + TOL_2 + TOL_3 \leq \varepsilon$ on ait :*

$$\|f(t^n) - f^n\|_{L^\infty(\Omega)} \leq C_T(\Delta t^2 + \frac{\varepsilon + h^2}{\Delta t}). \tag{7.8.79}$$

et pour le champ E, on a aussi l'estimation :

$$\|E(t^n) - E_h^n\|_{L^\infty(\Omega)} \leq C_T(\Delta t^2 + \frac{\varepsilon + h^2}{\Delta t}). \tag{7.8.80}$$

Remarque 7.8.9. *On peut également obtenir la convergence en appliquant l'opérateur d'adaptation non pas à chaque étape, mais seulement au bout d'un nombre fini d'étapes (indépendant de Δt), grâce au lemme 7.7.11.*

Chapitre 8

Résolution numérique adaptative de l'équation de Vlasov basée sur des cellules

Collaboration Ce travail est issu d'une collaboration initiée au CEMRACS'03 avec Martin Campos Pinto. Les résultats sur la parallélisation ont été effectués par Eric Violard et Olivier Hoenen en 2004.

8.1 Introduction

Avec l'accroissement rapide de la puissance des ordinateurs ces dernières années, les simulations de plasma et de faisceaux de particules basées sur une résolution directe de l'équation de Vlasov dans l'espace des phases multi-dimensionnel deviennent attractives comme alternatives aux simulations PIC traditionnelles. Leur force vient essentiellement du fait qu'elles sont non bruitées et que toutes les parties de l'espace des phases, y compris la queue de distribution sont bien résolues. Leur principal défaut est que, pour des systèmes inhomogènes, beaucoup des points de la grille (là où les particules sont présentes) sont gaspillés. Cela est particulièrement le cas pour les simulations de faisceaux, lorsque le faisceau bouge rapidement dans l'espace des phases (par exemple à cause des forces focalisantes à gradient alterné). Afin de surmonter ce problème, des méthodes adaptatives sont utilisées. Dans ce chapitre, on s'intéresse à développer une méthode adaptative basée sur une décomposition dyadique de l'espace des phases en cellules. Un avantage à ce type de méthode est la parallélisation : chaque donnée dépend essentiellement de la donnée voisine au même niveau et donc il est plus aisé de distribuer les données. On développe plus particulièrement une méthode basée sur une interpolation par éléments finis biquadratiques hiérarchiques. Rappelons que des méthodes adaptatives ont été implémentées pour résoudre l'équation de Vlasov dans [68] and [14]-[42], en utilisant respectivement des

ondelettes de Deslaurier-Dubuc ou une grille mobile. Le plan de ce chapitre est alors le suivant. On commence par définir la forme générale d'un schéma semi-Lagrangien, puis on décrit plus particulièrement le cas de l'interpolation par éléments finis biquadratiques hiérarchiques, et on regarde comment ces éléments s'insèrent dans le cadre des bases hiérarchiques et de l'analyse multirésolution. Le schéma numérique est ensuite donné, et on présente des résultats numériques. Enfin, on donne des résultats sur la parallélisation et on donne les conclusions et les perspectives.

8.2 Une stratégie générale

Forme générale d'un schéma semi-Lagrangien Classiquement, un schéma semi-Lagrangien consiste à chercher les valeurs de la fonction de distribution en chaque point d'une grille uniforme pour un temps t donné, en utilisant les valeurs de la fonction de distribution au temps précédent, sur une même grille uniforme. La valeur en chaque point de l'espace des phases (x, v) est donnée de la manière suivante. On remonte la courbe caractéristique passant par (x, v) jusqu'au temps précédent $t - \Delta t$; on tombe alors sur un point (x^*, v^*), appelé origine de la caractéristique. On sait que la fonction de distribution est constante le long de cette caractéristique, donc la valeur de la fonction de distribution au temps t au point (x, v) est la même que la valeur de la fonction de distribution au temps $t - \Delta t$, au point (x^*, v^*). Cette valeur est alors donnée par une interpolation sur la grille uniforme du temps $t - \Delta t$.
De manière plus générale, un schéma semi-Lagrangien est composé principalement de deux opérateurs :

- un *opérateur d'advection* $\mathcal{A}_n$, calculé grâce aux courbes caractéristiques.

- un *opérateur de projection*, qui transforme la donnée initiale f_0 ou la solution transportée $\mathcal{A}_n f^{n-1}$ en une fonction plus simple f^n.

L'étape de projection entraîne des erreurs; néanmoins, si on voulait s'en passer, la solution numérique deviendrait extrêmement compliquée et en pratique incalculable, les courbes caractéristiques de l'équation de Vlasov se rapprochant de plus en plus (on parle d'*enroulement des caractéristiques*).

Paramètres de discrétisation On fixe $T > 0$ et un entier N, le pas de temps est alors donné par $\Delta t = T/(N + 1)$.
La fonction initiale f_0 est soit supposée périodique et à support compact en vitesse, soit à support compact en vitesse et en espace.
Si la fonction n'est pas à support compact, on peut néanmoins traiter le cas de fonctions à décroissance exponentielle, en prenant un domaine assez grand et en supposant que la

fonction est nulle en dehors de ce domaine. On sait que la fonction de distribution reste à support compact, si elle l'est au départ. Néanmoins le domaine peut s'agrandir en temps. On fixe ici un domaine assez grand

$$\Omega = [x_{min}, x_{max}] \times [v_{min}, v_{max}]$$

qui contient le support de la fonction de distribution sur l'intervalle $[0, T]$. Pour des méthodes uniformes, il est judicieux de prendre des domaines pas trop grands (mais pas trop petits dans le cas de fonctions à décroissance exponentielle). En ce qui concerne les méthodes adaptatives, on peut au contraire choisir de prendre des domaines assez grands. On pose $L_x = x_{max} - x_{min}$ et $L_v = v_{max} - v_{min}$.
La fonction de distribution $f(n\Delta t, \cdot, \cdot)$ va alors être approchée par

$$f^n : \Omega \to \mathbb{R}.$$

Grille dyadique On fixe deux entiers j_0 et J qui correspondent au niveau le plus grossier et le plus fin de discrétisation en espace.
On suppose que la fonction de distribution (numérique) f^n est polynomiale sur une partition de Ω en cellules définies par

$$\mathcal{M}_{j,k,\ell} = [x_{min} + \frac{k}{2^j} L_x, x_{min} + \frac{k+1}{2^j} L_x[\times [v_{min} + \frac{k}{2^j} L_v, v_{min} + \frac{k+1}{2^j} L_v[.$$

Les indices j, k et ℓ sont tirés de $j_0, \ldots, J$, $0, \ldots, 2^j - 1$ et $0, \ldots, 2^j - 1$ respectivement, de telle sorte que l'union de toutes les cellules parcourues par ces indices forment une partition de Ω.
On notera $I = (j, k, \ell)$ un tel indice ; j s'appelle le niveau.
Le *maillage* est alors la donnée de

$$\mathcal{M}_{\mathcal{I}} = (\mathcal{M}_I)_{I \in \mathcal{I}}, \tag{8.2.1}$$

$\mathcal{I}$ désigne l'ensemble de tous les indices permettant de définir la partition (de manière équivalente le maillage) ; on l'appelle *maillage d'indices*.

Remarque 8.2.1. *Au niveau de la place mémoire, cette information ne coûte pas trop cher puisqu'il s'agit d'entiers (néanmoins assez nombreux).*

On note aussi K l'ensemble de tous les indices possibles :

$$K = \{(j, k, \ell) : j \in \mathbb{N} : k, \ell = 0, \ldots, 2^j - 1\},$$

et K_j l'ensemble de tous les indices de niveau j :

$$K_j = \{(j, k, \ell) : k, \ell = 0, \ldots, 2^j - 1\}$$

Une fois les indices définis, on note P_I la fonction définie polynomialement sur $\mathcal{M}_I$ et prolongée en dehors de $\mathcal{M}_I$ par zéro.

Les fonctions P_I sont contraintes à appartenir à un espace $\mathcal{P}_I$ que l'on doit préciser et qui est décrit par un nombre fini de paramètres.
Une *représentation adaptative* est alors la donnée d'un couple

$$F = (\mathcal{I}, P_{\mathcal{I}})$$

où $\mathcal{I}$ est un maillage d'indices et $P_{\mathcal{I}}$ est la suite des P_I pour $I \in \mathcal{I}$.
Etant donnée une représentation adaptative F^n, on en déduit une fonction f^n définie sur Ω (par périodicité ; dans le cas de support compact, il est peut-être plus efficace de mettre la valeur directement à zéro sur le bord $\{x_{max}\} \times [v_{min}, v_{max}]$ et $[x_{min}, x_{max}] \times \{v_{max}\}$).
Notons que le contraire n'est pas vrai : plusieurs représentations peuvent donner la même fonction.
Nous verrons néanmoins plus tard un intérêt à avoir plusieurs représentations.

Remarque 8.2.2. *(i) La donnée d'une représentation adaptative est bien plus coûteuse, car elle fait intervenir des réels.*

(ii) La représentation que l'on a choisie ici est complètement locale, ce qui est intéressant pour le parallélisme.

(iii) Néanmoins, dans le cas classiques des éléments finis (et en particulier les éléments biquadratiques que nous étudierons), il semble plus intéressant de choisir une autre représentation, par les noeuds des cellules (voir section suivante).

Raffinement Etant donnée une représentation adaptative $F = (\mathcal{I}, P_{\mathcal{I}})$, on peut définir une représentation adaptative raffinée

$$R_I F,$$

pour $I = (j, k, \ell) \in \mathcal{I}$, tel que $j < J$. La *cellule mère* $\mathcal{M}_I$ est remplaceé par ses 4 *cellules filles* obtenues en coupant la cellule mère en quatre parts égales. Le nouveau maillage d'indices est obtenu en remplacant l'indice $I = (j, k, \ell) \in \mathcal{I}$ par les 4 indices des filles. $\mathcal{D}_1(I), \mathcal{D}_2(I), \mathcal{D}_3(I)$ et $\mathcal{D}_4(I)$.
Pour $i = 1, \ldots, 4$, les $P_{\mathcal{D}_i(I)}$ sont définis par $P_{\mathcal{D}_i(I)}(x, y) = P_I(x, v)$, pour $(x, v) \in \mathcal{M}_{\mathcal{D}_i(I)}$.

Structure d'arbre afin de pouvoir déraffiner la grille, le maillage d'indices $\mathcal{I}$ devrait satisfaire la *propriété de structure d'arbre* définie ci-après.
Pour chaque cellule $\mathcal{M}_I$ avec $I \in \mathcal{I}$ de niveau $j > j_0$, on peut définir *sa* cellule mère, comme l'unique cellule de niveau $j - 1$ qui contient $\mathcal{M}_I$.
Ensuite, l'arbre $\Lambda_{\mathcal{I}}$ est défini comme l'ensemble de toutes les cellules du maillage adaptatif $\mathcal{M}_{\mathcal{I}}$, ainsi que récursivement les mères de ces cellules. Une cellule du maillage adaptatif $\mathcal{M}_{\mathcal{I}}$ est dite *non raffinée* si elle est de niveau J ou si ses filles n'appartiennent pas à l'arbre.

Déraffinement étant donnée une représentation adaptative $F = (\mathcal{I}, P_{\mathcal{I}})$, on peut définir une représentation adaptative déraffinée

$$C_I F,$$

pour $I = (j, k, \ell)$, avec $j > j_0$, tel que les indices des filles $\mathcal{D}_i(I)$, $i = 1, \cdots, 4$ appartiennent à $\mathcal{I}$.
Les indices des filles $\mathcal{D}_i(I)$, $i = 1, \cdots, 4$ sont alors remplacés par I.
P_I est alors défini par une *projection interniveaux*

$$\Pi_I : \mathcal{P}_{\mathcal{D}_1(I)} \times \mathcal{P}_{\mathcal{D}_2(I)} \times \mathcal{P}_{\mathcal{D}_3(I)} \times \mathcal{P}_{\mathcal{D}_4(I)} \to \mathcal{P}_I, \tag{8.2.2}$$

On suppose que la projection satisfait :

$$C_I R_I F = F \tag{8.2.3}$$

pour une représentation adaptative $F = (\mathcal{I}, P_{\mathcal{I}})$, et un indice $I = (j, k, \ell) \in \mathcal{I}$, avec $j > j_0$.

Initialisation à partir d'une donnée initiale f_0 (souvent écrite sous forme analytique), on suppose que l'on a une procédure qui permet d'obtenir les P_I pour $I \in K$.
On commence avec une représentation adaptative F^0 vide.
Alors, pour chaque indice $I = (j, k, \ell)$, en commencant avec $j = J - 1$, si l'erreur de projection interniveaux obtenue entre P_I et ses filles $P_{\mathcal{D}(I)}$ n'est pas petit, on rajoute les $(I, P_{\mathcal{D}_i(I)})$, pour $i = 1, \ldots, 4$ à F^0 et peut -être encore d'autres éléments afin de respecter la structure d'arbre.
Ainsi, l'erreur entre le résultat f^0 représenté par F^0 est contrôlé à l'aide des erreurs de projection. On doit aussi vérifier que l'écart entre voisins P_{I_1} et P_{I_2} n'est pas trop grand, en raffinant F^0 si nécessaire.

Schéma en temps On *prédit* un nouveau maillage $\mathcal{M}_{n+1}$ obtenu comme une certaine approximation de l'advection en avant $A_n(\mathcal{M}_n)$ du maillage $\mathcal{M}_n$.
On calcule une approximation de l'advection en avant de la fonction de distribution f^n sur chaque cellule de $\mathcal{M}_{n+1}$, et on obtient ainsi F^{n+1}.
On raffine ensuite ou déraffine comme dans le cas de l'initialisation.
Ainsi on en déduit des nouveaux F^{n+1} et $\mathcal{M}_{n+1}$ et on continue à boucler sur n.

8.3 Interpolation biquadratique hiérarchique

Afin de spécifier le schéma numérique, on doit définir les espaces

$$\mathcal{P}_I : I \in K$$

que l'on considère, les projections et projections interniveaux.

On étudie ici le cas d'une interpolation biquadratique hiérarchique.
Pour $I \in K$, on introduit d'abord l'ensemble $\mathcal{Q}_\mathcal{I}$ qui consiste en tous les éléments P_I définis sur $\mathcal{M}_I$ par :

$$P_I(x,v) = \sum_{r,s=0}^{2} a^I_{r,s} x^r v^s \quad a_{r,s} \in \mathbb{R} \quad r,s = 0,1,2, \quad (x,v) \in \mathcal{M}_I, \tag{8.3.1}$$

et nuls en dehors de $\mathcal{M}_I$.

Noeuds pour chaque cellule $\mathcal{M}_I$, avec $I \in K$, on définit ses neuf *noeuds* : les quatre sommets de la cellule, les quatre points milieux des arêtes et le point milieu de la cellule. On indice les noeuds par $N^I_{r,s}$, pour $r,s = 0,1,2$, en notant que deux valeurs de I différentes peuvent donner le même noeud.
On note $\mathcal{N}_\mathcal{I}$ l'ensemble des noeuds relatifs à un maillage d'indices $\mathcal{I}$.

Base de Lagrange par morceaux On introduit sur $\mathbb{R}$ les fonctions :

$$\begin{aligned}\phi_0(x) &= 2(x-1/2)(x-1),\\ \phi_1(x) &= 4x(1-x),\\ \phi_2(x) &= 2x(x-1/2),\end{aligned}$$

Pour chaque $j = j_0, \ldots, J$, $k,\ell = 0, 2^j - 1$ et $r,s = 0,\ldots,2$, on note par $L^{r,s}_{j,k,\ell}$ les *fonctions élémentaires* de la cellule $\mathcal{M}_{j,k,\ell}$, au noeud $[r,s]$:

$$L^{r,s}_{j,k,\ell}(x,v) = \phi_r(\theta_x(x))\phi_s(\theta_v(v)), \tag{8.3.2}$$

où $\theta(x,v) = (\theta_x(x), \theta_v(v))$ est l'application affine de $\mathcal{M}_{j,k,\ell}$ dans $[0,1[^2$ définie par

$$\theta_x(j,k)(x) = \frac{2^j}{k}(x - x_{min}) \quad \theta_v(j,k)(v) = \frac{2^j}{k}(v - v_{min}). \tag{8.3.3}$$

L'élément $P_I \in \mathcal{Q}_I$ s'écrit alors :

$$P_I = \sum_{r,s=0}^{2} f^I_{r,s} L^{r,s}_I \chi_I \quad f^I_{r,s} \in \mathbb{R} \quad r,s = 0,1,2, \tag{8.3.4}$$

où χ_I est la fonction indicatrice de la cellule $\mathcal{M}_I$ (elle vaut 1 sur $\mathcal{M}_I$ et 0 ailleurs).
Le coefficient $f^I_{r,s}$ est la valeur au noeud $N^I_{r,s}$, pour la cellule $\mathcal{M}_I$.

Interpolation continue imposer la continuité entre les cellules pour f consiste à restreindre l'ensemble $\mathcal{Q}$ pour la définition des $\mathcal{P}_I$.
L'ensemble $\mathcal{P}_I$ peut alors être directement décrit par les noeuds (et non aux noeuds relativement à une cellule), avec une attention spéciale pour les *noeuds contraints* : lorsqu'un noeud d'une cellule α est un point de l'adhérence d'une autre (unique) cellule β, on impose à la valeur au noeud d'être égale à la valeur obtenue par interpolation sur la cellule β.

Projection ici, la projection (et la projection inter-niveaux) consiste simplement en une interpolation sur les noeuds de la cellule sur laquelle on projette. On a alors, si l'on note $C_I f$ la fonction projetée, pour un indice de maille I de niveau $j > j_0$:

$$f = C_I f + \sum (f(a) - C_I f(a))\phi_a \tag{8.3.5}$$

où la somme est portée sur tous les 16 noeuds

$$a = N^{r,s}_{\mathcal{D}_i(I)} \quad (r,s) = (0,1),(1,0),(1,1),(1,2) \quad \text{or} \quad (2,1),$$

des cellules filles $\mathcal{M}_{\mathcal{D}_i(I)}$, $i = 1,\ldots,4$ qui ne sont pas des noeuds de $\mathcal{M}_I$.
La valeur $C_I f(a)$ est obtenue par interpolation sur la cellule $\mathcal{M}_I$, avec les 9 valeurs aux noeuds de celle-ci.
Les fonctions ϕ_a sont définies pour tout $a \in \mathcal{N}_K$ de niveau j comme valant 1 au point a, centre d'un rectangle de taille $2^j L_x \times 2^j L_v$, quadratique à l'intérieur strict de ce rectangle et nul ailleurs.
On a alors :

$$\|f - C_I f\| \leq \sum |f(a) - C_I f(a)| \|\phi_a\|, \tag{8.3.6}$$

et donc pour s'assurer que l'erreur de projection n'est pas trop grande, il suffit de contrôler la quantité de droite dans (8.3.6).
Remarquons que cette quantité est liée à la régularité locale de la fonction que f approche (cela se voit en introduisant la fonction que l'on approche et en utilisant l'inégalité triangulaire).

8.4 Bases hiérarchiques

On rappelle ici la définition de bases hiérarchiques, on donne d'abord l'exemple des bases de Faber-Schauder, pour fixer les idées, puis on donne une définition générale. Enfin, on regarde comment les éléments biquadratiques qui sont utilisés dans l'algorithme s'insèrent dans le cadre des bases hiérarchiques.
Pour faciliter l'exposé, on regarde essentiellement le cas de la dimension un (les éléments quadratiques donc).
On indique à la fin les modifications apportées pour le cas des dimensions supérieures.

Historique les bases hiérarchiques ont été développées par Yserentant en 1986 pour résoudre de manière efficace des systèmes linéaires provenant de problèmes elliptiques du second ordre. Néanmoins, l'utilisation de ces bases est bien plus ancienne. Faber [32] en 1909 avait déjà l'idée de représenter les fonctions de manière hiérarchique.

Suite de Faber-Schauder On considère l'intervalle $\mathbb{R}$ et on définit

$$I_{j,k} = [k2^{-j}, (k+1)2^{-j}[\quad k \in \mathbb{Z}.$$

On définit ensuite l'espace d'approximation

$$V_j = \{f \in C^0(\mathbb{R}) : f \text{ affine sur } I_{j,k}, \quad k \in \mathbb{Z}\}. \tag{8.4.1}$$

On a alors la propriété des espaces emboités : $V_j \subset V_{j+1}$ pour tout $j \in \mathbb{Z}$.
D'autre part, une fonction dans V_j est complètement déterminée par ses noeuds : $\Gamma_j := \{k2^{-j} \quad k \in \mathbb{Z}\}$.
On note également $\Gamma := \cup_{j\in\mathbb{Z}}\Gamma_j$.
On considère ensuite un opérateur d'interpolation :

$$P_j : C^0(\mathbb{R}) \to V_j \quad P_j u(x) = u(x) \quad x \in \Gamma_j. \tag{8.4.2}$$

L'idée de Faber a été alors de représenter la fonction u en termes de P_0 et des différences

$$P_{j+1} - P_j.$$

Ces différences sont des fonctions dans V_j et sont nulles en tous les points de Γ_{j-1}. Ainsi, elles peuvent être représentées par leur valeurs aux noeuds $x \in \nabla_j := \Gamma_{j+1}\backslash\Gamma_j$.
Une base naturelle pour V_j est donnée par les fonctions $\phi_{j,k}$ définies par

$$\phi(x) = \max\{0, 1 - |x|\} \quad \phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k) \quad k \in \mathbb{Z}. \tag{8.4.3}$$

La suite de Faber-Schauder est alors donnée par

$$\{\phi_{0,k}\}_{k\in\mathbb{Z}} \cup_{j\geq 1} \{\phi_{j,k}\}_{x=2^{-j}k\in\nabla_j} \tag{8.4.4}$$

On note que pour chaque j, l'espace W_j engendré par les

$$\{\phi_{j,k}\}_{x=2^{-j}k\in\nabla_j} \tag{8.4.5}$$

est supplémentaire de V_j dans V_{j+1} :

$$V_{j+1} = V_j \oplus W_j, \tag{8.4.6}$$

et pour une fonction $u \in C^0(\mathbb{R})$, le terme $P_{j+1}u - P_j u$ appartient à l'espace W_j.
Ainsi, avoir une décomposition d'une fonction u sous la forme

$$u = P_0 u + \sum (P_{j+1} - P_j)u, \tag{8.4.7}$$

revient à décomposer u dans l'espace :

$$V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_j \oplus \ldots, \tag{8.4.8}$$

et pour tout j, l'espace d'approximation V_j se décompose sous la forme :

$$V_j = V_0 \oplus W_1 \oplus \ldots \oplus W_{j-1}. \tag{8.4.9}$$

Ainsi, pour avoir une approximation d'une fonction u, on regarde successivement ses approximations dans les espaces V_0, puis $V_0 \oplus W_0$ et ainsi de suite ; suivant la construction hiérarchique du système de Faber-Schauder, étant donnée une certaine approximation d'une fonction u dans V_j, pour avoir une approximation plus "précise" dans V_{j+1}, il suffit de conserver l'approximation dans V_j et de rajouter la contribution des éléments de base additionnels de W_j (au lieu de réexprimer la nouvelle approximation de u dans une nouvelle base, on tire ici profit de conserver l'ancienne base).

Erreur d'interpolation. L'erreur d'interpolation dans $C(\mathbb{R})$ peut être estimée par le module de continuité $\omega(u,\delta) := \omega_0(u,\delta)_\infty$ défini par 5.5.13 :

$$\|u - P_j u\|_{L^\infty(\mathbb{R})} \leq \omega(u, 2^{-j}). \tag{8.4.10}$$

Comme le module de continuité tend vers zéro lorsque δ tend vers zéro, on obtient :

$$u = P_0 u + \sum_{k=1}^{\infty} (P_j u - P_{j-1} u), \tag{8.4.11}$$

avec convergence dans $L^\infty(\mathbb{R})$. Cette estimation n'est en faite pas optimale ; une majoration optimale est donnée en prenant le deuxième module de régularité dans L^∞ :

$$\|u - P_j u\|_{L^\infty(\mathbb{R})} \leq \omega_2(u, 2^{-j})_\infty \tag{8.4.12}$$

Remarquons alors que les différences deviennent rapidement très petites.
Pour $u \in C^1(\mathbb{R})$, on obtient :

$$\|P_j u - P_{j-1} u\|_{L^\infty(\mathbb{R})} \leq 2^{-j} \|u'\|_{L^\infty(\mathbb{R})} \tag{8.4.13}$$

Pour $u \in C^2(\mathbb{R})$, on obtient :

$$\|P_j u - P_{j-1} u\|_{L^\infty(\mathbb{R})} \leq 4^{-j}/2 \|u''\|_{L^\infty} \tag{8.4.14}$$

D'autre part, on peut remplacer $\mathbb{R}$ par un intervalle quelconque I. Ainsi, si u est une fonction très régulière dans un voisinage de I les termes de la série vont rapidement tendre vers zéro, et donc on pourra en garder un petit nombre.
On reprend maintenant la définition utilisée dans [74] pour définir les bases hiérarchiques.
Soit ω un domaine borné Soit X un sous espace de $C(\bar{\omega})$, on suppose que l'inclusion est linéaire et bornée.
On note σ une fonction de $\mathbb{N}$ dans $\mathbb{N}^*$.

Une grille hiérarchique de points. On définit un ensemble

$$\Delta = \{\Delta_k\}_{k=0}^{\infty}, \tag{8.4.15}$$

avec :

$$\Delta_0 = \{z_{j,0} \quad j = 1, \ldots, \sigma(0)\} \tag{8.4.16}$$

et

$$\Delta_k = \Delta_{k-1} \cup \{z_{j,k} \quad j = 1, \ldots, \sigma(k)\}. \tag{8.4.17}$$

Les éléments $z_{j,k}$ appartiennent à $\bar{\omega}$ et vérifient :

$$z_{j,k} \notin \Delta_{k-1}, \quad k = 1, 2, \ldots \tag{8.4.18}$$

et

$$\overline{\cup \Delta_k} = \bar{\Omega}. \tag{8.4.19}$$

La suite Δ est appelée *multigrille* de $\bar{\omega}$, et Δ_k est appelé un k-niveau de Δ, et les éléments $z_{j,k}$ sont appelés *noeuds* de Δ.

Système hiérarchique. Soit

$$H = \{h_{j,k} \quad k \geq 1, \quad j = 1, \ldots, \sigma(k)\}, \tag{8.4.20}$$

un sous ensemble dénombrable de X.
H est alors un système Δ–hiérarchique, si

$$h_{j,k}(z_{\ell,m}) = \delta_j^\ell \delta_k^m \quad m = 0, \ldots, k \quad \ell = 1, \ldots, \sigma(k), \tag{8.4.21}$$

où δ_j^ℓ est le symbole de Kronecker (δ_j^ℓ vaut 1 si $j = \ell$ et 0 sinon).
il est alors clair que tout sous-ensemble fini de H est linéairement indépendant.
0n notera aussi ϕ_a, l'élément $h_{j,k} \in H$, pour $a = z_{j,k} \in \Delta$.

Définition des bases hiérarchiques. On rappelle qu'une suite (e_n) est une base (de Schauder) d'un espace de Banach E si tout élément de E se décompose de manière unique sous la forme

$$x = \sum x_k e_k \quad x_k \in \mathbb{R}, \tag{8.4.22}$$

qui converge dans E.
Si H est une base de X et un système Δ–hiérarchique, on dit alors que H est une *base hiérarchique*.
Reprenons maintenant l'exemple du système de Faber-Schauder. On a la propriété suivante (cf [74]) :

Proposition 8.4.1. *Le système de Faber-Schauder, en restreignant les éléments à l'intervalle* $[0,1]$ *et en prenant* $\Delta := \Gamma \cap [0,1]$ *est une base hiérarchique simultanément pour* $C^0([0,1])$ *et* $W^{1,2}([0,1])$.

Système quadratique hiérarchique en dimension 1**.** On considère cette fois-ci pour espace V_j :

$$V_j := \{f \in C^0(\mathbb{R}) \quad f \text{ quadratique sur } I_{j,k}\}, \tag{8.4.23}$$

c'est-à-dire que f s'écrit sous forme d'un polynôme de degré inférieur ou égal à 2 sur chaque intervalle $I_{j,k}$, avec continuité aux bords de chaque intervalle.
On a toujours la propriété d'espaces emboîtés, $V_j \subset V_{j+1}$.
On introduit alors les noeuds : $\Gamma_0 := \{2^{-1}k, \quad k \in \mathbb{Z}\}$ et

$$\Gamma_j := \{2^{-j-1}(2k+1), \quad k \in \mathbb{Z}\}, \tag{8.4.24}$$

pour $j \geq 1$. On définit maintenant deux fonctions de bases. Soit d'abord ϕ^1 la fonction définie par les conditions

$$\phi^1(k) = \delta_0^k, k \in \mathbb{Z} + 1/2\mathbb{Z}, \quad \text{ et } \phi^1 \in V_0,$$

qui s'écrit explicitement

$$\phi^1(x) = (1 - |x|)(1 - 2|x|), \text{ si } x \in [-1,1], \quad \phi^1(x) = 0, \text{ si } |x| \geq 1,$$

qui est la fonction de base dans V_0 associée au noeud entier 0.
Soit d'autre part ϕ^2, la fonction définie par les conditions

$$\phi^2(k) = \delta^k_{(1/2)}, k \in \mathbb{Z} + 1/2\mathbb{Z}, \quad \text{et } \phi^2 \in V_0,$$

qui est la fonction de base dans V_0 associée au noeud demi-entier $1/2$, et qui s'écrit aussi explicitement

$$\phi^2(x) = 2x(1-x), \text{ si } x \in [0,1], \quad \phi^2(x) = 0, \text{ si } x \in \mathbb{R}\backslash[0,1].$$

On introduit alors les éléments $\phi^1_{j,k}$ et $\phi^2_{j,k}$ définis par

$$\phi^s_{j,k}(x) = \phi^s(2^j x - k), \quad j,k \in \mathbb{Z}, \quad s = 1,2. \tag{8.4.25}$$

Le système quadratique hiérarchique est alors donné par :

$$\{\phi_a = \phi^1_{0,k}, \quad a = k \in \Gamma_0\} \cup \{\phi_a = \phi^2_{0,k} \quad a = 2^{-1}(2k+1) \in \Gamma_0\} \\ \cup_{j\geq 1} \{\phi_a = \phi^2_{j,k}, \quad a = 2^{-j-1}(2k+1) \in \nabla_j\}, \tag{8.4.26}$$

avec toujours la notation $\nabla_j := \Gamma_{j+1}\backslash\Gamma_j$.
On définit de la même manière l'opérateur de projection P_j.
On a alors, en utilisant la propriété (8.4.10) qui est valable aussi ici :

Proposition 8.4.2. *Le système quadratique forme une base hiérarchique pour* $C^0([0,1])$.

Ondelettes. Les bases hiérarchiques sont à priori définies indépendamment des ondelettes ; on cherche ici à voir comment le système quadratique s'interprète dans le cadre de l'analyse multirésolution.

Analyse multirésolution en dimension 1 L'analyse multirésolution (définition introduite par Mallat [55]) est une suite $\{V_j\}_{j\in\mathbb{Z}}$ de sous-espaces vectoriels fermés de $L^2(\mathbb{R})$, vérifiant les propriétés suivantes :

$$V_j \subset V_{j+1} \tag{8.4.27}$$
$$f(x) \in V_j \leftrightarrow f(2x) \in V_{j+1} \tag{8.4.28}$$
$$\cap_{j\in\mathbb{Z}} V_j = 0 \tag{8.4.29}$$
$$\cup_{j\in\mathbb{Z}} V_j \text{ est dense dans } L^2(\mathbb{R}) \tag{8.4.30}$$
$$\text{Il existe une fonction } \phi(x) \text{ dans } V_0 \text{ telle que } \{\phi(x-k)\}_{k\in\mathbb{Z}} \text{ soit une base de Riesz de } V_0. \tag{8.4.31}$$

Le système de Faber-Schauder définit une analyse multirésolution (cf [24]), et on peut prendre pour ϕ la fonction ϕ du système de Faber-Schauder ; ϕ est appelée *fonction d'échelle*.

En ce qui concerne le système hiérarchique, les propriétés (8.4.27)–(8.4.30) sont bien vérifiées, mais la propriété (8.4.31) n'est pas satisfaite du moins pour une fonction de base ϕ_a pour $a \in \Gamma$.
L'espace V_0 est ici engendré par deux *fonctions d'échelles* : ϕ^1 et ϕ^2, correspondant aux noeuds entiers et aux noeuds demi-entiers respectivement.

Proposition 8.4.3. *Les fonctions*

$$\phi^1(\cdot - k), \quad \phi^2(\cdot - k), \quad k \in \mathbb{Z} \tag{8.4.32}$$

forment une base de Riesz de V_0.

Preuve. On utilise la décomposition :

$$\int_{\mathbb{R}} |\sum_\ell a_\ell \phi^1(x-\ell) + b_\ell \phi^2(x-\ell)|^2 \mathrm{d}x = \sum_{k\in\mathbb{Z}} \int_{I_{0,k}} |\sum_\ell a_\ell \phi^1(x-\ell) + b_\ell \phi^2(x-\ell)|^2 \mathrm{d}x$$
$$= \sum_{k\in\mathbb{Z}} \int_{I_{0,0}} |\sum_\ell a_{\ell+k} \phi^1(x-\ell) + b_{\ell+k} \phi^2(x-\ell)|^2 \mathrm{d}x,$$

et on procéde comme dans le lemme 6.4.2, pour les polynômes de Lagrange. □

Remarque 8.4.4. *On peut dire aussi que les espaces V_j forment une analyse multi-résolution de multiplicité 2 : si les propriétés (8.4.27)–(8.4.30) sont vérifiées, et s'il existe plusieurs fonctions $\phi_1, \ldots, \phi_r$ telles que*

$$\phi_q(x-k) \quad k \in \mathbb{Z} \quad q = 1, \ldots, r \tag{8.4.33}$$

forment une base de Riesz de V_0, on parle d'analyse multi-résolution de multiplicité r (cf [24]).

Détails Dans un intervalle $I_{j,k}$ donné, considérons deux niveaux d'approximations d'une fonction u donnée.
Le premier est donné grâce aux valeurs aux noeuds extrémités de l'intervalle $[a,b] := I_{j,k}$ et du noeud milieu m de l'intervalle. L'approximation dans V_j sur $I_{j,k}$ consiste alors à interpoler par un polynôme de degré 2 sur $I_{j,k}$.
Pour l'approximation dans V_{j+1}, on a besoin des valeurs additionnelles aux deux noeuds milieux m_1 et m_2 de chacun des sous-intervalles $[a,m]$ et $[m,b]$.
La différence entre les deux approximations $P_{j+1}u - P_j u$ est alors donnée par :

$$(u(m_1) - P_j u(m_1))\phi_{m_1} + (u(m_2) - P_j u(m_2))\phi_{m_2}. \tag{8.4.34}$$

La figure 8.1 représente dans ce cas les différentes fonctions de bases ϕ_a, ϕ_b, ϕ_m et ϕ_{m_1}, ϕ_{m_2}.
On voit alors que sur la maille $I_{j,k}$, l'erreur de l'approximation $P_j u$ par rapport à l'approximation $P_{j+1}u$ s'exprime en fonction des détails $d_{m_1} := u(m_1) - P_j u(m_1)$ et $d_{m_2} := u(m_2) - P_j u(m_2)$.

Compression et stockage On considère toujours une fonction u que l'on approche à l'aide des éléments hiérarchiques. Supposons ainsi que l'on ait un système Δ_u–hiérarchique, où Δ_u est un ensemble fini, "adapté" à la fonction u, qui est obtenu par exemple en éliminant les détails inférieurs à un certain seuil pour tous les niveaux jusqu'à un certain niveau d'approximation J. Notons P_{Δ_u} cette approximation. Le meilleur choix serait de considérer, pour une norme $\|\cdot\|$ donnée (le choix dépend de la norme), toutes les approximations possibles, qui sont en nombre fini, et de prendre ensuite une qui donne l'erreur $\|u - P_{\Delta_u}u\|$ la plus petite possible ; bien sûr un tel choix n'est pas envisageable et le *seuillage*, précédement décrit peut être considéré comme une approximation *presque optimale* (voir [24], pour plus de détails).
On cherche maintenant à connaître la valeur en un point donné arbitraire, ce qui est une étape importante, dans un algorithme de type semi-lagrangien.
Une possibilité est de faire une *décomposition en ondelettes*. Cette étape consiste à calculer la donnée P_{Δ_u} sur tous les points de Γ_J. Il s'agit d'une étape de *décompression* et qui demande de la place mémoire, au moment où elle est faite.
La valeur aux points de Γ_J peut s'obtenir en parcourant tous les niveaux j, en partant du niveau 0 (plus généralement d'un niveau grossier j_0) et de calculer la valeur $P_{\Delta_u}(a)$, pour $a \in \nabla_{j-1}$ (en prenant pour convention ici, $\nabla_{-1} := \Gamma_0$).
Cette valeur est soit égale à $u(a)$, si $a \in \Delta_u$ (ce qui est le cas pour le niveau grossier), soit il s'agit d'une valeur interpolée avec les 3 points de niveau inférieur (les valeurs en ces points sont donc déjà calculés).
C'est ainsi que l'on peut faire dès que l'on a un opérateur de *prédiction*.
Néanmoins, pour calculer la valeur en un unique point x, il n'est pas nécessaire de connaître la valeur de P_{Δ_u} en tous les points de Γ_J. Si on a juste pour information les points de Δ_u, on peut tester si chacun des 3 noeuds du J–niveau utiles pour l'interpolation en x existent, si aucun existe, il n'est pas nécessaire de considérer le $J-$niveau et on recommence avec les noeuds du $J-1$–niveau (et ainsi de suite récursivement), s'il y en a au moins un qui existe, cela veut dire que les noeuds du J–niveau sont nécessaires, et donc on calcule la valeur aux autres noeuds, de la même manière que x récursivement. Cette dernière solution est la plus *adaptive* possible, au sens où l'on conserve vraiment que l'information strictement nécessaire ; néanmoins, elle semble plus coûteuse lors de l'évaluation en un point, puisqu'il faut tester si chaque point existe ou non (i. e., appartient ou non à Δ_u).
On peut alors choisir de stocker en plus le niveau de régions ; avec cette information, pour un point donné, il suffira de commencer à chercher les points du niveau de la région laquelle il appartient, il y en a alors au moins un qui appartient à Δ_u, la valeur aux autres points est alors calculée comme précédemment.
La manière que l'on a considéré est de stocker non seulement le niveau, mais également la valeur de u aux 3 points nécessaires à l'interpolation. Il en résulte que l'on stocke encore plus de données ; néanmoins, de cette manière, on est assuré de rester local, et de ne pas avoir besoin de chercher des points de niveau différent pour interpoler (ce qui facilite la parallélisation).

Remarque 8.4.5. *Pour une décomposition en ondelettes qui ne provient pas nécessairement d'un système hiérarchique, pour calculer la valeur en un point, on a besoin de la contribution de toutes les ondelettes de tous les niveaux qui sont dans le support, tandis*

que pour ce système quadratique, la valeur ne dépend que de 3 *paramètres.*

Système hiérarchique biquadratique en dimension 2**.** En dimension 2 (et en dimension quelconque de manière générale), le système hiérarchique se construit par produit tensoriel.
L'espace V_j est donné par :

$$V_j := \{f \in C^0(\mathbb{R}^2) \quad f \text{ biquadratique sur } I_{j,k} \times I_{j,\ell}\}, \tag{8.4.35}$$

c'est-à-dire que f s'écrit sous la forme d'un polynôme de degré inférieur ou égal à 2 en chacune des deux variables :

$$f(x,v) = \sum_{i,j=0}^{2} a_{i,j} x^i v^j, \quad a_{i,j} \in \mathbb{R} \quad i = 0,1,2, \quad j = 0,1,2, \tag{8.4.36}$$

sur chaque carré $I_{j,k,\ell} := I_{j,k} \times I_{j,\ell}$ avec continuité aux bords.
On introduit les noeuds dyadiques

$$\Gamma_0 := \{2^{-1}(k,\ell), \quad k,\ell \in \mathbb{Z}\}, \tag{8.4.37}$$

et

$$\Gamma_j := \{2^{-j-1}(2k+1, 2\ell), \quad k,\ell \in \mathbb{Z}\} \cup \{2^{-j-1}(2k, 2\ell+1), \quad k,\ell \in \mathbb{Z}\}. \tag{8.4.38}$$

Le système biquadratique en dimension 2 est alors obtenu par produit tensoriel du système quadratique en dimension 1 et définit une base hiérarchique de $C^0([0,1]^2)$; on peut définir aussi l'analyse multirésolution de $L^2(\mathbb{R}^2)$ par produit tensoriel, et les propriétés précédentes se transmettent.

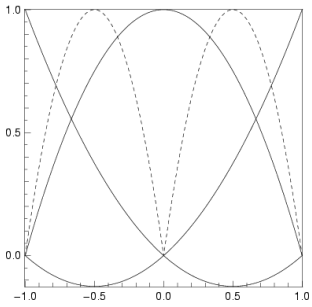


Fig. 8.1 – fonctions de bases quadratiques hiérarchiques. Fonctions de bases sur la cellule grossière (trait plein). Fonctions de bases additionnelles pour les cellules fines (en pointillé).

8.5 Le schéma numérique

Résolution en temps L'advection est classiquement effectuée par un splitting en temps (voir e.g [59]). Une méthode générale à deux pas de temps a été introduite dans [67] pour résoudre les caractéristiques, par un calcul direct en $2D$ (voir aussi [68] pour une nouvelle méthode efficace). On a utilisé ici l'algorithme complètement local suivant. Connaissant la position finale $a = (X^{n+1}, V^{n+1})$ au temps $(n+1)\Delta t$, on calcule une approximation (X^n, V^n) d'ordre deux de la position arrière $(X, V)(n\Delta t; a, (n+1)\Delta t)$.
Ainsi, avec les notations $t^n = n\Delta t$, $t^{n+\frac{1}{2}} = (n+\frac{1}{2})\Delta t$ et $X^{n+\frac{1}{2}} = \frac{X^n + X^{n+1}}{2}$, on peut écrire à l'ordre deux en temps :

$$\frac{X^{n+1} - X^n}{\Delta t} = \frac{V^n + V^{n+1}}{2} + O(\Delta t^2).$$

D'autre part, à nouveau à l'ordre deux, on obtient :

$$\frac{V^{n+1} - V^n}{\Delta t} = E(t^{n+\frac{1}{2}}, X^{n+\frac{1}{2}}) + O(\Delta t^2) \tag{8.5.1}$$

$$\tag{8.5.2}$$

et

$$E(t^n, X^{n+\frac{1}{2}}) = \frac{1}{2}[E(t^{n+\frac{1}{2}}, X^{n+\frac{1}{2}}) + E(t^{n-\frac{1}{2}}, X^{n+\frac{1}{2}})] + O(\Delta t^2), \tag{8.5.3}$$

ainsi que

$$E(t^{n-\frac{1}{2}}, X^{n+\frac{1}{2}}) = \frac{1}{2}[E(t^{n-1}, X^n) + E(t^n, X^{n+1})] + O(\Delta t^2). \tag{8.5.4}$$

On obtient alors :

$$\begin{aligned} E(t^{n+\frac{1}{2}}, X^{n+\frac{1}{2}}) &= 2E(t^n, X^{n+\frac{1}{2}}) - E(t^{n-\frac{1}{2}}, X^{n+\frac{1}{2}}) + O(\Delta t^2) \\ &= 2E(t^n, X^{n+\frac{1}{2}}) - \frac{1}{2}[E(t^{n-1}, X^n) + E(t^n, X^{n+1})] + O(\Delta t^2). \end{aligned}$$

Ainsi, on obtient la formule d'ordre deux suivante :

$$\frac{V^{n+1} - V^n}{\Delta t} = 2E(t^n, \frac{X^n + X^{n+1}}{2}) - \frac{E(t^{n-1}, X^n) + E(t^n, X^{n+1})}{2} + O(\Delta t^2). \tag{8.5.5}$$

Champ électrique de Poisson : A partir d'une représentation adaptative F^n, on peut en déduire la densité de charge $\rho^n = \int f^n(x, v)\mathrm{d}v$ et obtenir le champ électrique E^n, à partir de l'équation de Poisson :

$$-\partial_x E^n = \rho^n - 1$$

(dans le cas axisymétrique, on prendra $-\frac{1}{r}\partial_r(rE_r) = \rho$).

Advection avant/arrière : on écrit $B_n^{wd}(a)$ la position advectée arrière d'un noeud a. On l'obtient ici par (8.5.5) avec $a = (X^{n+1}, V^{n+1})$ et $B_n^{wd}(a) = (X^n, V^n)$. Ce système peut être résolu itérativement en l'inconnue V^n. D'autre part $F_n^{wd}(a)$ désigne la position advectée avant et est également donnée par (8.5.5), en écrivant $a = (X^n, V^n)$ et $F_n^{wd}(a) = (X^{n+1}, V^{n+1})$.

L'algorithme est maintenant guidé par deux quantités : le **niveau de résolution le plus fin** J et un **seuil de tolérance** ε .

Algorithme. La solution au temps t^n est donnée par un maillage adaptatif $\mathcal{M}_n$ qui consiste en une partition dyadique de cellules. Passer du temps t^n à t^{n+1} consiste en trois étapes :

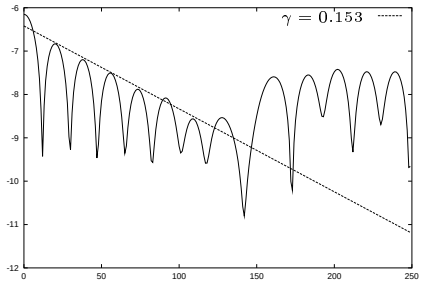
1. **Prédiction** de $\mathcal{M}^{n+1}$: Pour chaque cellule $\alpha \in \mathcal{M}^n$, calculer le centre c_α et le point advecté en avant, $\mathcal{A}(c_\alpha)$ en suivant les caractéristiques de l'équation de Vlasov. Rajouter ensuite à $\mathcal{M}^{n+1}$ l'unique cellule $\bar{\alpha}$ de niveau j, qui convient à la place de $\mathcal{M}^{n+1}$ et rajouter les cellules nécessaires de telle sorte que $\mathcal{M}^{n+1}$ est un maillage adaptatif dyadique. Enfin, si $j < J$, raffiner $\bar{\alpha}$ d'un niveau, c'est-à-dire, remplacer $\bar{\alpha}$ par ses 4 cellules de niveau $j+1$ qui couvrent la même surface.
2. **Evaluation** : Pour chaque noeud a de $\mathcal{M}^{n+1}$, calculer le point advecté en arrière $\mathcal{A}^{-1}(a)$ et mettre $f^{n+1}(a)$ à $f^n(\mathcal{A}^{-1}(a))$: l'évaluation $f^n(c)$ de la solution en chaque point $c \in [0,1[^2$ est obtenu en cherchant l'unique cellule α du maillage adaptatif $\mathcal{M}^n$ où le point est localisé, en utilisant les valeurs aux noeuds de cette cellule et en calculant l'interpolation biquadratique sur cette cellule, disons $I(c, \alpha, f^n(c))$.
3. **Compression** de $\mathcal{M}^{n+1}$: En partant de $j = J-1$ à j_0, remplacer 4 cellules de niveau $j+1$ par une cellule α de niveau j (faire le contraire du raffinement de α) où la norme de la différence $f^{n+1}(a) - I(a, \alpha, f^n(a))$, pour chaque noeud a of α, est assez petit.

8.6 Cas tests

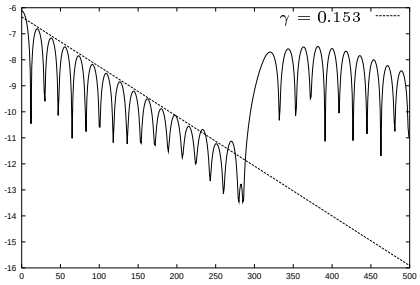
L'amortissement linéaire de Landau Afin de tester la précision numérique du schéma, l'amortissement linéaire de Landau est très classique. La condition initiale est donnée par

$$f(0,x,v) = \frac{1}{\sqrt{2\pi}} e^{-v^2/2}(1 + \alpha\cos(kx))$$

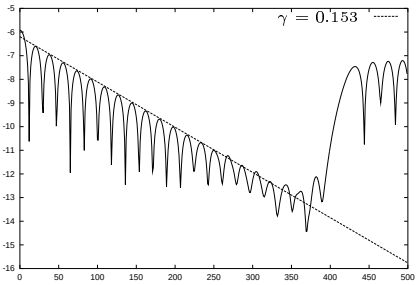
avec $\alpha = 0.01$, la période vaut $L = 4\pi$ et $k = 0.5$. Pour la discrétisation en temps, on choisit $\Delta t = 1/8$. On restreint le domaine de calcul en vitesse en un intervalle $[-v_{max}, v_{max}]$, avec un nombre v_{max} suffisamment grand. Le champ électrique décroit exponentiellement avec un taux théorique de $\gamma = 0.1533$ en norme L_2 (voir e.g. [59]). En fait, la solution numérique ne peut pas décroitre tout le temps et la solution devrait restituer son énergie



(a) $\Delta v = 0.281$



(b) $\Delta v = 0.148$

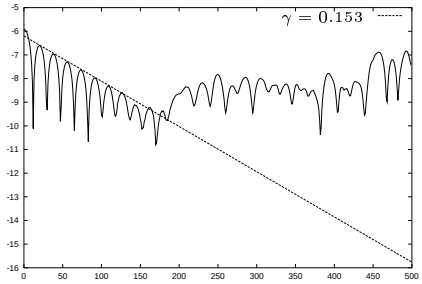


(c) $\Delta v = 0.112$

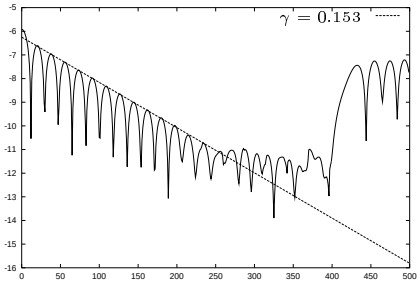
Fig. 8.2 – Evolution de $\log(\int E(x,t)^2 dx)$ en function du nombre d'itérations ($\Delta t = 0.125$) dans le cas uniforme avec un niveau de résolution $J = 4, 5$ et 6 (c'est-à-dire, respectivement 256, 1024 et 4096 cellules) pour l'amortissement linéaire de Landau.

en un "temps de récurrence"(see e.g [59]). Cependant, en prenant plus de points dans la direction des vitesses, on peut repousser ce phénomène et donc avoir une meilleure description du champ électrique pour des temps plus longs.

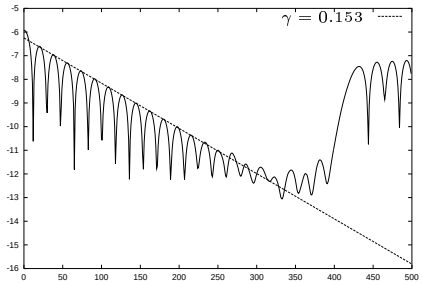
Dans le cas adaptatif (voir figure 8.3), on prend $J = 6$ et $v_{max} = 7.15$ puisque ces paramètres these paramètres donnent de bons résultats dans le cas uniforme (voir figure 8.2) : on obtient environ 20 periodes d'oscillations. Comme prévu, la précision de l'amortissement augmente lorsque la tolérance baisse, et on atteint la précision du cas uniforme sous-jacent avec $\varepsilon = 10^{-8}$. Les normes L_1 et L_2 sont bien conservées : la norme L_2 perd moins de $4 \cdot 10^{-6}$ norme relative pour $\varepsilon = 10^{-6}$ et l'erreur de masse totale relative est inférieur à 10^{-6} (cependant, si la tolérance est trop grande, les résultats sont plus mauvais : si $\varepsilon = 10^{-5}$, la norme L_2 augmente de 10^{-4} après 500 itérations).
D'un autre côté, la proportion de cellules économisée est d'à peu près 1/4 (pour $\varepsilon = 10^{-8}$, environ 3100 cellules sont utilisées au départ et 2900 après les 100 premières itérations ; pour $\varepsilon = 10^{-6}$, on passe de 2400 à 2200 cellules) : la compression se fait là où la solution est presque nulle.



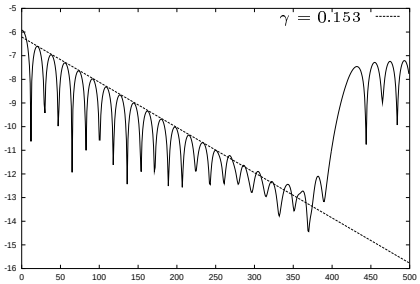
(a) $\simeq 2000 - 1400$ mailles



(b) $\simeq 2400 - 2200$ mailles



(c) $\simeq 2700 - 2650$ mailles



(d) $\simeq 3100 - 2900$ mailles

Fig. 8.3 – Evolution de $\log(\int E(x,t)^2 dx)$ en fonction du nombre d'itérations ($\Delta t = 0.125$) avec $\Delta v = 0.112$ dans le cas adaptatif (avec un niveau de résolution fin $J = 6$ et une tolérance $\varepsilon = 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}$) pour l'amortissement de Landau linéaire.

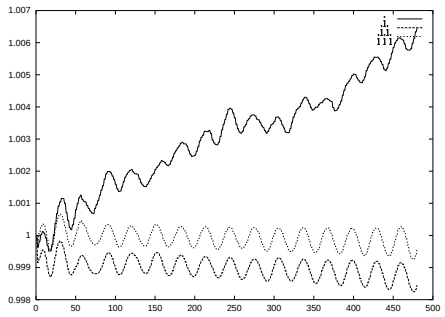
Ainsi, si l'utilisation d'un tel maillage adaptatif ne semble pas être le moyen le plus naturel pour traiter ce cas test puisque la solution ne développe pas de petites échelles, nous sommes maintenant convaincus, d'après cet exemple que le schéma adaptatif peut aller jusqu'à la précision de la solution uniforme si on choisit une tolérance assez petite. Notons également que l'on peut augmenter le domaine de calcul avec coût additionnel minime.

Faisceau semi-Gaussien On considère maintenant un faisceau semi-Gaussien défini par la condition initiale

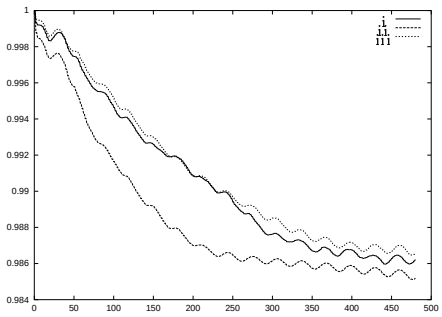
$$f_0(r,v) = \frac{1}{\pi a^2 \sqrt{2\pi} b} e^{-\frac{1}{2}(v^2/b^2)}, \quad \text{if} \quad r < a,$$

et $f_0(r,v) = 0$ ailleurs. Ici $a = 4/\sqrt{15}$ et $b = 1/(2\sqrt{15})$. Le pas de temps est $\Delta t = \pi/16$, ce qui correspond à $1/32$-ième de période. On effectue 480 itérations, c'est-à-dire 15 périodes.

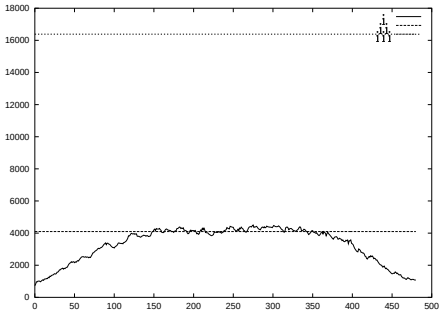
La solution numérique développe des petites échelles qui disparaissent après un certain temps, par diffusion. Dans notre simulation, on a pris $\varepsilon = 10^{-3}$ et $J = 7$ pour le cas adaptatif.
On compare notre solution adaptative à une solution grossière uniforme (avec une résolution plus grossière $J = 6$), et à une uniforme plus fine (avec la même résolution $J = 7$). On voit sur les figures 8.5, 8.6 et 8.7 que la grille suit le développement des petites échelles, tandis que le ratio $\#(\mathcal{M}^n)/\#(\mathcal{M}_6)$ passe en 6 périodes de $1/4$ à 1 (voir figure 8.4(c)). Après 3 périodes (figure 8.6), la solution adaptative semble mieux cerner les effets non linéaires que la solution uniforme grossière, elle reste en fait aussi précise que la solution uniforme fine. Sur la figure 8.7, on approche la zone de pleine filamentation ; la solution adaptative reste meilleure que la plus grossière, avec toujours le même ordre de grandeur de cellules (voir aussi 8.4(c)), mais cette fois-ci, on n'atteint plus la même précision que la grille fine uniforme. Après 15 périodes, la diffusion apparait et le processus de déraffinement apparait : on se retrouve alors avec le même nombre de points qu'initialement (voir figure 8.8 et 8.4(c)).
En utilisant une interpolation biquadratique, le schéma est assez diffusif, cependant l'adaptivité n'accélère pas ce phénomène : la norme L_2 norm des solutions fines unfiormes et adaptatives sont assez similaires (voir figure 8.4(b)). D'autre part, la masse n'est pas bien conservée dans le cas adaptatif (figure 8.4(a)) : elle augmente et cela semble être dû au gain/défaut de masse dans l'étape de compression et le manque de conservation dans le cas de l'interpolation sur un maillage non-uniforme pour un schéma semi-Lagrangien.



(a) masse totale



(b) norme L_2



(c) nombre de mailles

Fig. 8.4 – Evolution de la **masse totale relative**, de la **norme** L_2 de la fonction de distribution f pour le faisceau semi-Gaussien et du **nombre de cellules** en fonction du nombre d'itérations (32 itérations = 1 période) dans le cas adaptatif avec $J = 7$ et $\varepsilon = 10^{-3}$ (i), dans le cas uniforme avec $J = 6$ (ii) et avec $J = 7$ (iii).

8.7 Implémentation parallèle

Cette partie a été établie grâce à Eric Violard et Olivier Hoenen (voir [45], [44] pour plus de détails).

Régions le domaine de calcul est subdivisé en *régions*. Une région est une surface du domaine de calcul qui est définie par une union de cellules. Les régions sont allouées à des processeurs de telle sorte que chaque processeur a connaissance de ses propres cellules et noeuds, qui sont inclus dans sa propre région. Lorsque le maillage s'adapte à l'évolution en temps de la physique, le nombre de cellules contenu dans une région change et il est donc nécessaire d'inclure un mécanisme d'équilbrage de charge, qui consiste à redéfinir les régions sur chaque processeur.

Compression Afin de minimiser les communications, on n'applique la compression que dans la limite de sa région. Ainsi, la phase de compression ne requiert aucune communication. On perd un peu en adaptativité, puisque l'on élimine moins de cellules que la méthode originale.

Structure de données chaque processeur connait une représentation locale du maillage. Celui-ci est représenté par deux tables de hachage (outil classique utilisé pour stocker des données irrégulièrement réparties) :

- la table de hachage des cellules qui garde en mémoire un ensemble de cellules qui forme une partition du domaine de calcul, et qui associe à chaque cellule l'identité de son processeur.

- la table de hachage des noeuds : elle stocke les noeuds du maillage ainsi que la région d'appartenance pour chaque noeud.

Cette représentation permet d'accéder aux cellules et aux noeuds en temps constant tout en minimisant l'espace mémoire.

Equilibrage de charges : pour chaque région, le nombre de cellules doit être approximativement le même et chaque région doit avoir une bonne forme pour améliorer la compression. De plus, chaque région doit être connexe afin de réduire le volume de communications ; les courbes de Hilbert sont alors utilisées [52].

Résultats numériques On considère toujours le cas test du faisceau semi-Gaussien, pour une simulation de 5 périodes (160 itérations). Les résultats sont reportés sur les figures 8.9, 8.10 et la table 8.1.

Sur la figure 6.3, le niveau le plus fin est $J = 8$, ce qui correspond à une grille fine de 512×512 points. On observe que la grille adaptative suit bien les structures fines.

La figure 8.10 montre l'interprétation graphique de l'accélération (speed-up) pour une échelle logarithmique.
On observe que l'accélération est approximativement constant lorsque le niveau des détails augmente.

La table 8.1 montre que le temps continue à décroitre lorsque le nombre de processeurs augmente jusqu'à 64 processeurs.

8.8 Conclusion

Nous avons présenté un solveur de Vlasov basé sur une interpolation par éléments finis hiérarchiques. Les résultats numériques obtenus montrent que le code fonctionne bien dans le cas de l'espace des phases $2D$, pour des simulations de faisceaux.
la méthode se généralise directement à une dimension quelconque ; pour qu'elle reste efficace en $4D$, il semble indispensable de la paralléliser.
L'inhérente bonne localisation des cellules a déjà permis de donner une implémentation parallèle efficace en $2D$; un code $4D$ parallèle est en cours de développement.
Néanmoins, il subsiste des points à améliorer.

Prédiction de la grille. Pour prédire la grille, nous avons utilisé une advection en avant suivie d'un raffinement automatique d'un niveau.

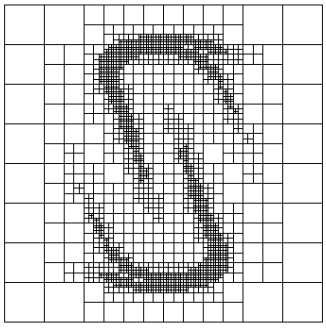
• Il est en fait aussi possible de faire une prédiction en arrière de la grille (comme celle utilisée dans la preuve de convergence du chapitre 7).
Ainsi, au temps t^{n+1}, on considère les mailles de niveau $j = j_0$, on advecte an arrière le centre de chacune de ces mailles et on recommence avec les filles, si le niveau sur lequel le point advecté en arrière tombe sur une maille de niveau strictement supérieur à j.
L'avantage est que l'on ne calcule plus des points qui servent juste à prédire la grille.

• D'autre part, on raffine ici d'un niveau automatiquement (ce qui semble suffisant dans la pratique) ; cette procédure a l'avantage de rester locale, mais prédit un maillage souvent bien plus gros que le maillage qui est vraiment utile. Il pourrait donc être intéressant de remplacer cette technique en utilisant des indicateurs d'erreurs (un analogue de la courbure qui était utilisée au chapitre 7). Par cette nouvelle manière de prédire le maillage, on a besoin de l'information des cellules voisines. Si celles-ci appartiennent à un autre processeur, cela entraîne des communications supplémentaires ; une idée serait alors de combiner ces deux techniques : indicateurs d'erreurs, là où l'information est disponible sans communication, et raffinement d'un niveau aux autres endroits.

Ordre élevé et conservation. La méthode implémentée a le principal défaut de ne pas être d'ordre élevé et surtout de ne pas être conservative. En ce qui concerne l'ordre, on a été limité à l'ordre 2, puisqu'au delà, l'algorithme présente un défaut de stabilité (voir chapitre 6). L'utilisation de maillages adaptatifs décalés pour rester dans la zone de stabilité semble assez complexe, il semblerait préférable d'utiliser d'autres reconstructions (qui font intervenir les gradients par exemple, ou des reconstructions non interpolantes, que l'on peut chercher aussi de telle sorte que l'on ait d'autres propriétés telles que la conservation de la masse).
Pour ce qui est de la conservation, l'équation de Vlasov a la propriété de conserver beaucoup de quantités, il est généralement difficile d'en conserver beaucoup numériquement. Les plus importantes sont la masse, le moment d'ordre un et l'énergie du système. Il est

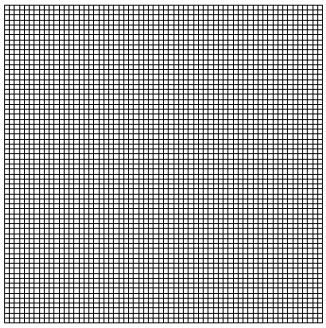
aussi parfois souhaitable de garder la positivité.
Dans notre méthode, l'erreur de masse est à chaque pas de temps de l'ordre du détail. Néanmoins, pour des simulations en temps longs (telles que l'instabilité double-faisceaux), l'accumulation des erreurs entraîne des phénomènes non physiques.
Signalons qu'il existe une manière de recouvrir la masse (et les autres quantités), dans le cas d'un maillage non structuré, par une méthode a posteriori (cf [15]). Néanmoins cette méthode détériore l'ordre de précision ; l'interpolation se base sur une combinaison linéaire d' interpolation de degré 1 et du degré initial et les coefficients sont solutions d'un problème d'optimisation.
Dans le cas des ondelettes, une technique de lifting est souvent utilisée (voir [71]).

8.9 Figures



(a) $J = 7$, $\varepsilon = 10^{-3}$

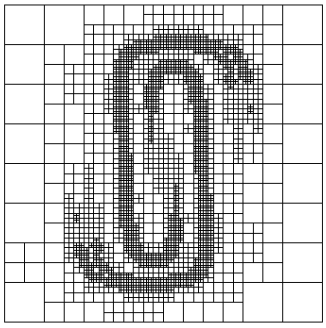
(b) $J = 7$, $\varepsilon = 10^{-3}$



(c) $J = 6$ uniform

(d) $J = 6$ uniform

FIG. 8.5 – grille adaptative et fonction de distribution f dans l'espace des phases (v, x) après 1.5 périodes (48 itérations).



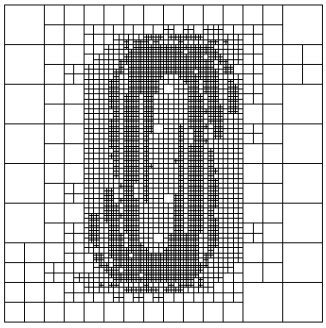
(a) $J = 7$, $\varepsilon = 10^{-3}$

(b) $J = 7$, $\varepsilon = 10^{-3}$

(c) $J = 6$ uniform

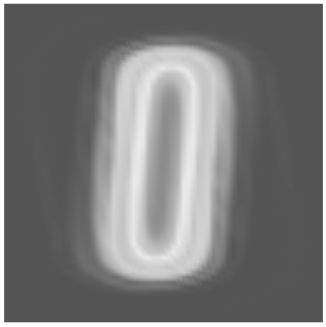
(d) $J = 7$ uniform

Fig. 8.6 – grille adaptative et fonction de distribution f dans l'espace des phases (v, x) après 3 périodes (96 itérations).

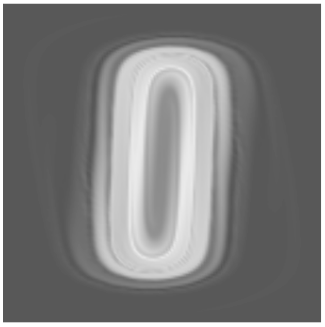


(a) $J = 7$, $\varepsilon = 10^{-3}$

(b) $J = 7$, $\varepsilon = 10^{-3}$

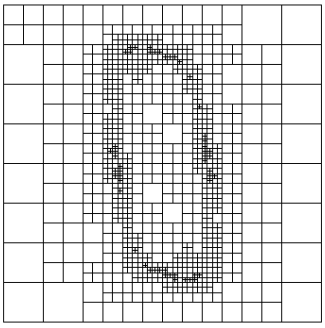


(c) $J = 6$ uniform



(d) $J = 7$ uniform

FIG. 8.7 – grille adaptative et fonction de distribution f dans l'espace des phases (v, x) après 6 périodes (192 itérations).



(a) $J = 7$, $\varepsilon = 10^{-3}$

(b) $J = 7$, $\varepsilon = 10^{-3}$

(c) $J = 6$ uniform

(d) $J = 7$ uniform

Fig. 8.8 – grille adaptative et fonction de distribution f dans l'espace des phases (v, x) après 15 périodes (480 itérations).

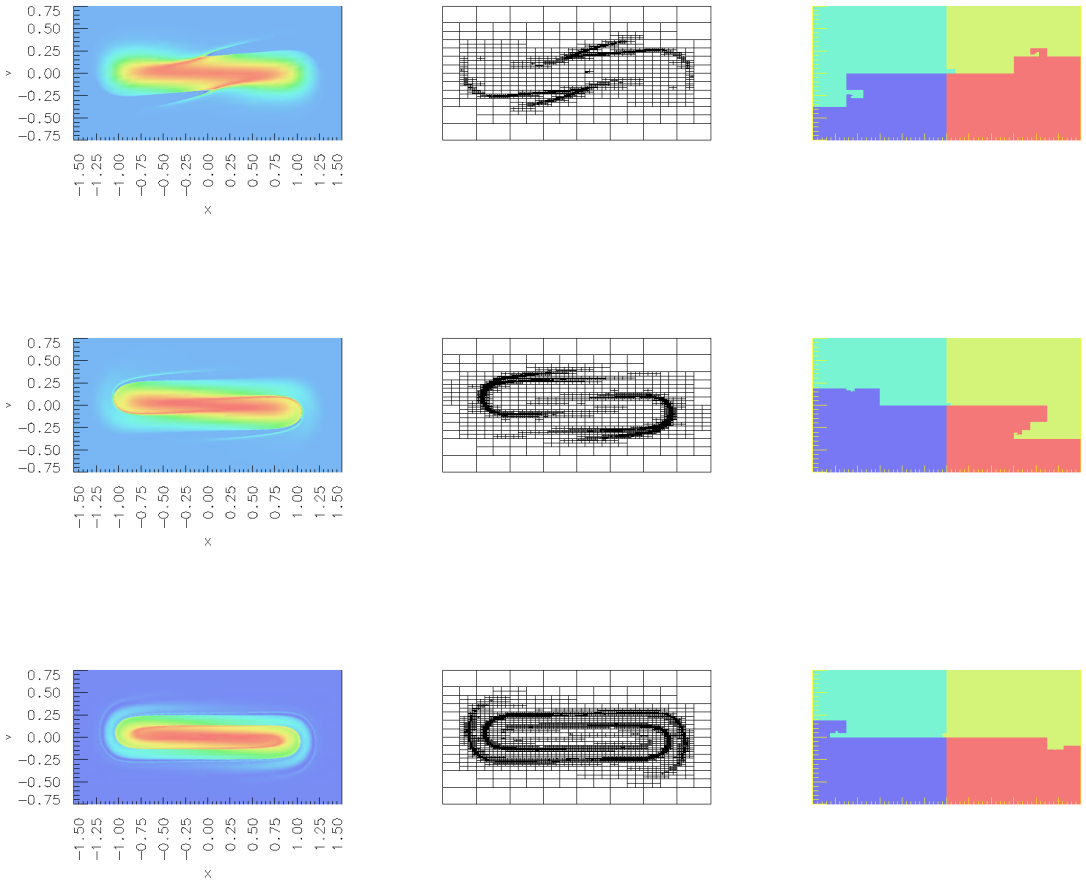


FIG. 8.9 – Evolution de la fonction de distribution ainsi que la grille et la distribution des régions après une demi, une et deux périodes (resp. 32, 64 et 128 itérations)

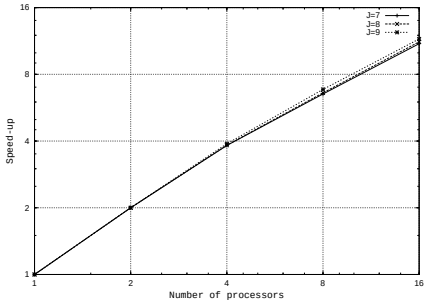


FIG. 8.10 – speed-up sur un HP cluster

# procs	$J = 8$
1	3074
2	1514
4	797
8	459
16	268
32	177
64	153

Tab. 8.1 – temps écoulé en secondes sur un SGI O3800

Bibliographie

[1] R. Adams, *Sobolev spaces*, Pure and Applied Mathematics, A series of monographs and textbooks, Academic press, 1975.

[2] F. Alauzet, P. L. George, B. Mohammadi, P. Frey, H. Borouchaki, *Transient fixed point-based unstructured mesh adaptation,* ECCOMAS Computational Fluid Dynamics Conference, Part II (Swansea, 2001), Internat. J. Numer. Methods Fluids 43 (6-7) (2003), 729–745.

[3] B. J. C. Baxter, N. Sivakumar, *On shifted cardinal interpolation by Gaussians and multiquadrics*, J. Approx. Theory 87 (1996), 36–59.

[4] I. Babushka et W. C. RHeinboldt, *Error estimates for adaptive finite element computations,* SIAM, J. Numer. Anal. 15 (1978), 736–754.

[5] J. Batt, *Global symmetric solutions of the initial value problem of stellar dynamics*, J. Differ. Equations (1977), 342–364.

[6] R. Becker, R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods,* Acta Numerica 10 (2001), 1–102.

[7] M. L. Begue, A. Ghizzo et P. Bertrand, *Two-dimensional Vlasov simulation of Raman scattering and plasma beatwave acceleration on parallel computers,* J. Comput. Phys. 151 (1999), 458–478.

[8] M. Berger et J. Oliger, *Adaptive mesh refinement for hyperbolic partial differential equations,* J. Comput. Phys. 53 (1984), 482–512.

[9] M. Berger et P. Collela, *Local adaptive mesh refinement for hyperbolic partial differential equations* J. Comput. Phys. 82 (1984), 64–84.

[10] R. Bermejo, *Analysis of an algorithm for the Galerkin-characteristic method,* Numer. Math. 60 (1991), 163–194.

[11] R. Bermejo, *Analysis of a class of quasi-monotone and conservative semi-Lagrangian advection schemes,* Numer. Math. 87 (2001), 597–623.

[12] S. Bertoluzza, *Adaptive wavelet collocation for the solution of steady state equation,* SPIE Proc. Wavelet Appl. II. 2491 (1995).

[13] S. Bertoluzza, *An adaptive collocation method based on interpolating wavelets,* W. Dahmen, A. J. Kurdila et P. Oswald Eds. (1997), Multiscale Wavelet Methods for PDE's, Academic Press, New York.

[14] N. Besse, F. Filbet, M. Gutnic, I. Paun et E. Sonnendrücker, *An adaptive numerical method for the Vlasov equation based on a multiresolution analysis*, Numerical Mathematics and Advanced Applications ENUMATH 2001, Eds F. Brezzi, A. Buffa, S. Escorsaro, A. Murli, Springer (2001), 437–446.

[15] N. Besse, *Etude mathématique et numérique de l'équation de Vlasov non linéaire sur des maillages non structurés de l'espace des phases*, Thèse, ULP 2003.

[16] N. Besse, *Convergence of a semi-Lagrangian scheme for the one-dimensional Vlasov-Poisson system* , SIAM J. Numer. Anal. 42 (2004), 350–382.

[17] N. Besse, *Convergence of a high-order semi-Lagrangian scheme with propagation of gradients for the Vlasov-Poisson system*, submitted.

[18] N. Besse et E. Sonnendrücker, *Semi-Lagrangian schemes for the Vlasov equation on an unstructured mesh of phase space*, J. Comput. Phys. 191 (2) (2003), 341–376.

[19] C. K. Birdshall et A.B. Langdon, *Plasma physics via computer simulation,* McGraw-Hill, 1985.

[20] F. Bouchut, F. Golse et M. Pulvirenti, *Kinetic equations and asymptotic theory,* Series in Applied Mathematics, P.G. Ciarlet and P.-L. Lions (Eds.) Gauthier-villars (2000).

[21] H. Brezis, *Analyse fonctionnelle, théorie et applications*, Mathématiques appliquées pour la maîtrise, Masson, 1992.

[22] Carl de Boor, *On the cardinal spline interpolant to* e^{int}, SIAM J. Math. Anal 7 (6) (1976).

[23] P. G. Ciarlet, *Basic Error Estimates for Elliptic Problems*, Handbook of Numerical Analysis, Vol. II, Finite Element Methods (Part 1), 1991.

[24] A. Cohen, *Numerical analysis of wavelet methods*, studies in mathematics and its applications, 32, North-Holland, Elsevier, 2003.

[25] A. Cohen, S. M. Kaber, S. Müller et M. Postel, *Fully adaptive multiresolution finite volume schemes for conservation laws,* Math. Comp. 72 (241) (2003), 183–225.

[26] C.Z. Cheng et G. Knorr, *The integration of the Vlasov equation in configuration space,* J. Comput. Phys. 22 (1976), 330–351.

[27] J. Cooper et A. Klimas, *Boundary Value Problems for the Vlasov-Maxwell Equation in One dimension,* J. of Math. Anal. and Appl. 75 (1980), 306–329.

[28] G.-H. Cottet, P.-A. Raviart, *Particle methods for the one-dimensional Vlasov-Poisson equations,* SIAM, J. Numer. Anal. 21 (1) (1984), 52–76.

[29] R. DeVore, *Nonlinear approximation,* Acta Numerica, Cambridge Univ. Press 7 (1998), 51–150.

[30] W. Doerfler, *A convergent adaptive algorithm for Poisson's equation,* SIAM, J. Numer. Anal. 33 (1996), 1106–1124.

[31] K. Eriksson, D. Estep, P. Hansbo et C. Johnson, *Introduction to adaptive methods for differential equations* Acta Numerica, Cambridge Univ. Press (1995), 105–158.

[32] G. Faber, *Über stetige Funktionen,* Mathematische Annalen 66 (1909), 81–94.

[33] Maurizio Falcone et Roberto Ferretti, *Convergence analysis for a class of high-order semi-lagrangian advection schemes,* SIAM J. Numer Anal. 35 (3) (1998), 909–940.

[34] F. Filbet, *Contribution à l'analyse et à la simulation numérique de l'équation de Vlasov,* Thèse de l'université de Nancy, (2001).

[35] F. Filbet, *Convergence of a finite volume scheme for the Vlasov-Poisson system,* SIAM, J. of Numer. Anal. 39 (4) (2001), 1146–1169.

[36] F. Filbet, E. Sonnendrücker *Comparison of Eulerian Vlasov solvers*, cComput. Phys. Comm. 150 (2003), 247–266.

[37] F. Filbet, *Numerical Methods for the Vlasov equation* ENUMATH'01 Proceedings.

[38] F. Filbet, E. Sonnendrücker et P. Bertrand, *Conservative Numerical schemes for the Vlasov equation*, J. Comput. Phys. 172 (1) (2000), 166–187.

[39] E. Forest et J. Bengtsson, *Application of the Yoshida-Ruth techniques to implicit integration and multi-map explicit integration,* Phys. Let. A 158 (1991), 99–101.

[40] D. Goldman et T. J. Kaper, *Nth-order operator splitting schemes and non reversible systems,* SIAM J. Numer. Anal. 33 (1996), 349–367.

[41] R. T. Glassey, *The Cauchy problem in kinetic theory,* Society for Industrial and Applied Mathematics (SIAM), Philadephia, PA, 1996.

[42] M. Gutnic, M. Haefele, Ioana Paun et E. Sonnendrücker, *Vlasov simulations on an adaptive phase-space grid* to appear in Comput. Phys. Comm.

[43] A. Harten, *Multiresolution algorithms for the numerical solution of hyperbolic conservation laws,* Comm. Pure Appl. Math. 48 (1993), 1305–1342.

[44] O. Hoenen, M. Mehrenberger et E. Violard, *Parallelization of an Adaptive Vlasov Solver,* to appear in ParSim04 (Special Session of EuroPVM/MPI 2004).

[45] 0. Hoenen, *Mécanisme de régulation dynamique de la charge pour un solveur de Vlasov semi-Lagrangien,* rapport de stage de DEA d'informatique 2003/2004 Image et Calcul Parallèle Scientifique.

[46] H. Hong et S. Steinberg, *Accuracy and stability of polynomial interpolation schemes for advection equations*, Preprint.

[47] E. Horst, *On classical solutions of the initial value problem for the unmodified nonlinear Vlasov equation,* Math. Meth. Appl. Sci. 3 (1981), 229–248.

[48] E. Horst, *On the asymptotic growth of the solutions of the Vlasov-Poisson system,* Math. Meth. in the Appl. Sci. 16 (1993), 75–85.

[49] S. V. Iordanskii, *The Cauchy problem for the kinetic equation of plasma,* Amer. Math. Soc. Transl. 35 (2) (1964), 351–363.

[50] A. Klimas et W. M. Farell, *A splitting algorithm for the Vlasov simulation with filamentation filtration,* J. Comput. Phys. 68 (1987), 202–226.

[51] K. Jetter, S. D. Riemenschneider et N. Sivakumar, *Schoenberg's exponential Euler spline curves*, Proceedings of the Royal Society of Edinburgh 118A (1991), 21–33.

[52] J. K. Lawder, P. J. H. King, *Using Space-Filling Curves for Multi-dimensional Indexing*, Lecture Notes in Computer Science 1832 (2000).

[53] P. L. Lions et B. Perthame, *Propagation of moments and regularity for the 3-dimensional Vlasov-system*, Ivent. Math. 10 (1991), 415–430.

[54] Y. Maday, V. Perrier et J. C. Ravel, *Adaptativité dynamique sur bases d'ondelettes pour l'approximation d'équations aux dérivées partielles,* CRAS Paris, Série I. 1 (1991), 405–410.

[55] S. Mallat, *Multiresolution approxiamtion and wavelet orthonormal bases of* $L^2(\mathbb{R})$ Trans. Amer. Math. Soc. 315 (1989), 69–88.

[56] S. Mallat, *A wavelet tour of signal processing.* Academic Press, 1998.

[57] P. Morin, R. Nocetto et K. Siebert, *Data oscillations and convergence of adaptive FEM* SIAM J. Numer. Anal. 38 (2000), 466–488.

[58] Ramachandran D. Nair, *Extension of a Conservative Cascade Scheme on the Sphere to Large Courant Numbers,* Mounthly Weather review 132 (2004), 390–395.

[59] T. Nakamura et T. Yabe, *Cubic interpolated propagation scheme for solving the hyper-dimensional Vlasov-Poisson equation in phase space*, Comput. Phys. Comm. 120 (1999), 122–154.

[60] F. J. Narcowitch, N. Sivakumar et J. D. Ward, *Stability results for scattered-data interpolation on euclidean spheres*, Adv. Comput. Math. 8 (3) (1998), 137–163.

[61] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Éditeurs, Paris ; Academia, Éditeurs, Prague 1967.

[62] P. M. Prenter, *Splines and variational methods*, Wiley Classics Library, 1975.

[63] K. Pfaffelmoser, *Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data,* J. of Diff. Eqs. 95 (1992), 281–303.

[64] P. A. Raviart, *An analysis of particle methods*, Numerical methods in fluid dynamics, Lecture Notes in Math. Vol 1127, 343–324, Springer, Berlin 1985.

[65] J. Schaeffer, *Global existence of smooth solutions to the Vlasov-Poisson system in three dimensions,* Commun. Part. Diff. Eq. 16 (1991), 1313–1335.

[66] I. J. Schoenberg, *Cardinal interpolation and spline functions*, Journal of Approximation Theory, 2 (1969), 167–20 6.

[67] E. Sonnendrücker, J. Roche, P. Bertrand et A. Ghizzo, *The Semi-Lagrangian Method for the Numerical Resolution of Vlasov Equations,* J. Comput. Phys. 149 (1999), 201–220.

[68] E. Sonnendrücker, F. Filbet, A. Friedman, E. Oudet et J. L. Vay, *Vlasov simulation of beams on a moving phase-space grid,* to appear in Comput. Phys. Comm.

[69] A. Staniforth, J. Cote, *Semi-Lagrangian integration schemes for atmospheric models-a review,* Monthly Weather Review 119 (1991), 2206–2223.

[70] G. Strang, *Trigonometric polynomials and difference methods of maximum accuracy,* J. Math. Phys. 41 (2) (1962), 147–154.

[71] W. Sweldens, *The lifting scheme : construction of Second Generation Wavelets*, SIAM, Journal on Mathematical Analysis, 29 (2) (1998), 511–546.

[72] H. Triebel, *Theory of functions spaces,* Birkhauser, Basel, 1983.

[73] S. Ukai, T. Okabe, *On classical solutions in the large in time of two dimensional Vlasov's equation,* Osaka J. Math. 15 (1978), 245–261.

[74] Vladimir L. Vaskevitch, *Best approximation and hierarchical bases*, Selcuk J. appl. Math. 2 (2) (2001), 83–106.

[75] R. Verfurth, *A posteriori error estimation and adaptive mesh refinement techniques,* J. Comp. Appl. Math. 50 (1994), 67–83.

[76] S. Wollman et E. Ozizmir, *Numerical approximation of the one-dimensional Vlasov-Poisson system with periodic boundary conditions,* SIAM J. Numer. Anal. 33 (1996), 1377–1409.

[77] S. Wollman, *On the approximation of the Vlasov-Poisson system by particle methods,* SIAM J. Numer. Anal. 37 (4) (2000), 1369–1398.

[78] H. Yoshida, *Construction of higher order sympletic integrators,* Phys. Let. A 150 (1990), 262–268.

[79] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numerische Mathematik (1986), 379–412.

[80] H. Yserentant, *Hierarchical bases,* ICIAM 91 (R. E. O'Malley, ed.), SIAM (Philadelphia).

[81] H. Yserentant, *Old and new convergence proof for multigrid methods*, Acta Numerica, Cambridge University Press (1993), 285–326.