Université de Strasbourg

Mémoire d'Habilitation à Diriger des Recherches

Spécialité : **Mathématiques Appliquées**

préparée au laboratoire **Institut de Recherche Mathématique Avancée, UMR 7501**

dans le cadre de l'École Doctorale **Mathématiques, Sciences de l'Information et de l'Ingénieur**

présentée et soutenue publiquement
par

Michel Mehrenberger

le ?? ?? 2012

Titre:

Inégalités d'Ingham et schémas semi-Lagrangiens pour l'équation de Vlasov

Jury

M. Eric Sonnendrücker,	Garant
M. Christophe Besse,	Rapporteur
M. Francis Filbet,	Rapporteur
M. Thierry Gallouët,	Rapporteur
M. Vilmos Komornik,	Examinateur

Résumé

Dans une première partie, on rassemble plusieurs résultats en théorie du contrôle autour des inégalités d'Ingham, généralisations de l'égalité de Parseval, qui interviennent pour montrer l'observabilité, la contrôlabilité ou la stabilisation frontière ou interne de l'équation des ondes ou d'équations similaires dans certains cas particuliers. On s'intéresse dans un premier temps à l'optimalité de ce type d'inégalités en généralisant un résultat précédent au cas vectoriel. On développe ensuite un théorème de type Ingham adapté pour traiter le cas d'une géométrie cartésienne. Enfin, on donne des résultats d'observabilité dans le cas d'approximations numériques.

Dans une seconde partie, on présente les méthodes semi-Lagrangiennes qui sont composées essentiellement de deux ingrédients : calcul des caractéristiques le long desquelles la fonction de distribution est constante et étape d'interpolation. On analyse des schémas d'ordre élevé en temps pour le système de Vlasov-Poisson $1D \times 1D$, basés sur le splitting directionnel, qui est une succession d'étapes de transport linéaire. On étudie alors les méthodes semi-Lagrangiennes dans ce cas particulier et on fait le lien entre différentes formulations. On obtient également un théorème de convergence pour le système de Vlasov-Poisson dans ce cadre, qui reste valable pour des petits déplacements. On développe ensuite ce type de méthodes dans un cadre plus général, en se basant sur le splitting uni-dimensionnel conservatif, avec une variante de type Galerkin discontinu. Dans une dernière partie, on étudie l'opérateur de gyromoyenne qui intervient en physique des plasmas pour prendre en compte des corrections de rayon de Larmor fini. Enfin, on discute de la problématique de la divergence discrète nulle qui donne une compatibilité entre le calcul du champ et la méthode numérique de transport.

Mots-clés : physique des plasmas, méthodes semi-Lagrangiennes, équation de transport, gyromoyenne, inégalités d'Ingham, observabilité

Abstract

In the first part, we gather several results in the control theory around Ingham inequalities which are generalizations of Parseval's equality and appear for showing the internal or boundary observability, controlability or stabilization of the wave equation or similar equations in certain particular cases. We are interested at first in the optimality of such inequalities, by generalizing a previous result in the vectorial case. We then develop a Ingham type theorem adapted to treat the case of a cartesian geometry. Finally, we give some observability results in the case of numerical approximations.

In a second part, we present the semi-Lagrangian method which is composed by essentially two ingredients : the computation of the caracteristics along which the distribution function is constant et the interpolation step. We analyse high order schemes in time based on directional splitting, which are a succession of linear transport steps. We then study the semi-Lagrangian methods in this particular case and we make the link between different formulations. We also obtain a convergence theorem for the Vlasov-Poisson system in this framework, which remains valid in the case of small displacements. We then develop this type of methods in a more general framework, by using one dimensionnal conservative splitting. We also consider a discontinuous Galerkin variant of such schemes. In a last part, we study the gyroaverage operator which appears in plasma physics by taking care of finite Larmor radius corrections. Finally, we discuss the problematic of zero discrete divergence which gives a compatibility between field computations and the numerical method of transport.

Keywords : plasma physics, semi-Lagrangian methods, transport equation, gyroaverage, Ingham inequalities, observability

Table des matières

Introduction

1 Première partie

Dans une première partie, on rassemble plusieurs résultats en théorie du contrôle autour des inégalités d'Ingham. On rappelle ici le théorème d'Ingham [56].

Théorème 1.1. *Soit $(\omega_k)_{k\in\mathbb{Z}}$ une famille de nombres réels, satisfaisant la condition d'écart uniforme*

$$\gamma := \inf_{k\neq n} |\omega_k - \omega_n| > 0.$$

Si I est un intervalle borné de longueur $|I| > 2\pi/\gamma$, alors il existe deux constantes $c_1, c_2 > 0$

$$c_1 \sum_{k\in\mathbb{Z}} |x_k|^2 \leq \int_I |x(t)|^2 dt \leq c_2 \sum_{k\in\mathbb{Z}} |x_k|^2,$$

pour toutes les fonctions données par la somme

$$x(t) = \sum_{k\in\mathbb{Z}} x_k e^{i\omega_k t},$$

avec des coefficients complexes x_k de carré sommable.

Ce type de théorème intervient pour montrer l'observabilité, la contrôlabilité ou la stabilisation frontière ou interne de l'équation des ondes ou d'équations similaires dans certains cas particuliers (cf V. Komornik, P. Loreti, *Fourier Series in control theory*, 2005). Dans ce chapitre, on donne plusieurs résultats autour de ce type de technique. Ce type de méthode fonctionne bien pour des problèmes où le spectre de l'opérateur sous jacent est bien connu : pour l'équation des ondes en dimension 1, par exemple, et dans ce cas les inégalités d'Ingham se réduisent à l'égalité de Parseval. On cherche ici à développer et appliquer des théorèmes de type Ingham pour pouvoir traiter des situations plus complexes. On s'intéresse dans un premier temps à l'optimalité de ce type de théorème en généralisant un résultat précédent [73] au cas vectoriel [7]. On développe ensuite un théorème de type Ingham adapté pour traiter le cas d'une géométrie cartésienne [74]. Enfin, on donne des résultats d'observabilité dans le cas d'approximations numériques [63, 64] et on indique des perspectives de recherche.

2 Deuxième partie

En deuxième partie, on présente les méthodes semi-Lagrangiennes qui sont composées essentiellement de deux ingrédients : calcul des caractéristiques le long desquelles la fonction de distribution est constante et étape d'interpolation.

On étudie d'abord le cas de la discrétisation en temps par splitting directionnel pour l'équation de Vlasov-Poisson. On dérive les conditions d'ordre pour des schémas de splitting d'ordre quatre dans le cas du système de Vlasov-Poisson $1D \times 1D$ [27]. Les calculs pour obtenir de telles conditions sont motivés par la structure de Poisson spécifique du système de Vlasov-Poisson : la structure est similaire aux systèmes de Runge-Kutta-Nyström. On montre que les conditions obtenues sont les mêmes que les conditions RKN dérivées pour les EDO jusqu'à l'ordre quatre. Des résultats numériques sont effectués et montrent la pertinence d'utiliser des schémas de splitting dans ce contexte.
Grâce au splitting directionnel, une brique de base est le transport linéaire. On présente alors cette étape et on fait le lien entre diverses formulations : valeurs ponctuelles, valeurs moyennes, volumes finis, intégrateurs exponentiels. On trouve ainsi que certaines méthodes développées dans la littérature sont identiques ou semblables dans ce cas particulier de l'advection constante. On inclut aussi la description d'un schéma de Galerkin discontinu dans ce cadre qui a fait l'objet de récents développements.
On analyse ensuite les erreurs commises pour les méthodes semi-Lagrangiennes dans le cadre du système de Vlasov Poisson $1D \times 1D$ [21] avec splitting de Strang. On reprend et améliore les estimations du travail [9], en enlevant la singularité en $\frac{1}{\Delta t}$.
On développe ensuite les méthodes conservatives, de manière plus générale : on se place dans le cadre d'un champ à divergence nulle et on effectue un splitting directionnel. Chaque étape de splitting ne correspond alors plus forcément à un transport linéaire, comme c'était le cas pour le système de Vlasov-Poisson. Les méthodes conservatives présentent la possibilité d'ajouter des filtres adéquats pour assurer la positivité. Dans le cas de l'advection non constante, elles présentent aussi une alternative aux schémas semi-Lagrangiens classiques qui peuvent souffrir d'une mauvaise conservation de la masse, dans ce cadre du splitting directionnel. On développe aussi un schéma de Galerkin discontinu, en généralisant convenablement le cas précédemment traité du transport linéaire ; des résultats numériques sur le modèle centre-guide viennent confirmer la pertinence de l'approche utilisée.
Lors de la résolution de l'équation de Vlasov, on est aussi amené à résoudre le calcul de champs : Poisson, Maxwell... On regroupe alors dans ce dernier chapitre quelques résultats autour du calcul de champs. Dans un premier temps, on fait une étude de l'opérateur de gyromoyenne qui intervient en physique des plasmas pour prendre en compte des corrections de rayon de Larmor fini. Plusieurs méthodes sont proposées et comparées à la méthode spectrale de référence. On indique ensuite aussi une méthode numérique pour résoudre l'équation de Poisson dans le cadre de la méthode de Galerkin discontinu pour le transport. Enfin, on discute de la problématique de la divergence discrète nulle qui donne une compatibilité entre le calcul du champ et la méthode numérique de transport.
La partie se clôt avec une liste de travaux en cours et diverses perspectives de recherche.

3 Liste des publications

Une version des publications peut être téléchargée sur le site http ://www-irma.u-strasbg.fr/˜ mehrenbe/#publi. En gras sont indiqués les publications postérieures

aux travaux de thèse.

3.1 Thèse

[T] [72] M. Mehrenberger, *Inégalités d'observabilité et résolution adaptative de l'équation de Vlasov par éléments finis hiérarchiques*, Thèse de Doctorat, Université Louis Pasteur, décembre 2004.

3.2 Articles dans revues à comité de lecture

[**A11**] [3] K. Ammari, M. Mehrenberger, *Study of the nodal feedback stabilization of a string-beams network*, Journal of Applied Mathematics and Computing, Volume 36, Num. 1-2, pp. 441-458, (2011).
[**A10**] [1] F. Alauzet, M. Mehrenberger, *P1-conservative solution interpolation on unstructured triangular meshes*, Int. J. Numer. Meth. Engng, Vol. 84 (13), pages 1552–1588, 24 Décembre 2010.
[**A9**] [30] N. Crouseilles, M. Mehrenberger, H. Sellama, *Numerical solution of the gyroaverage operator for the finite gyroradius guiding-center model,* CiCP 8, pp. 484-510, (2010).
[**A8**] [31] N. Crouseilles, M. Mehrenberger, E. Sonnendrücker, *Conservative semi-Lagrangian schemes for Vlasov equations*, Journal of Computational Physics 229 (2010), 1927–1953.
[**A7**] [74] M. Mehrenberger, *An Ingham type proof for the boundary observability of a N-d wave equation*, C. R. Math. Acad. Sci. Paris 347 (2009), no. 1-2, 63–68.
[**A6**] [4] K. Ammari, M. Mehrenberger, *Stabilization of coupled systems,* Acta Math. Hungar. 123 (2009), no. 1-2, 1–10.
[**A5**] [2] K. Ammari, M. Jellouli, M. Mehrenberger, *Feedback stabilization of a coupled string-beam system.* Netw. Heterog. Media 4 (2009), no. 1, 19–34.
[**A4**] [63] P. Loreti, M. Mehrenberger, *An Ingham type proof for a two-grid observability theorem,* ESAIM Control Optim. Calc. Var. 14 (2008), no. 3, 604–631.
[A3] [18] M. Campos Pinto, M. Mehrenberger, *Convergence of an Adaptive Scheme for the one dimensional Vlasov-Poisson system,* Numer. Math. 108 (2008), no. 3, 407–444.
[A2] [9] N. Besse, M. Mehrenberger, *Convergence of classes of high order semi-Lagrangian schemes for the Vlasov Poisson system,* Math. of Comp. 77 (2008), 93–123.
[A1] [73] M. Mehrenberger, *Critical length for a Beurling type theorem,* Bollettino U.M.I. (8) 8-B (2005), 251–258.
[A0] [71] M. Mehrenberger, *Observability of coupled systems,* Acta. Mat. Hungar. 103 (4) (2004), 321–348.

3.3 Actes de colloques avec comité de lecture

[**P7**] [16] J. P. Braeunig, N. Crouseilles, M. Mehrenberger, E. Sonnendrücker, *Guiding-center simulations on curvilinear meshes*, Discrete and Continuous Dynamical Systems Series S, Volume 5, Number 3, June 2012.

[**P6**] [32] N. Crouseilles, M. Mehrenberger, F. Vecil, *Discontinuous Galerkin semi-Lagrangian method for Vlasov-Poisson*, ESAIM Proc, CEMRACS 2010, October 2011, Vol. 32, p. 211-230.
[**P5**] [7] A. Barhoumi, V. Komornik, M. Mehrenberger, *A vectorial Ingham-Beurling type theorem*, Ann. Univ. Sci. Budapest. Eötvös Sect. Math. 53 (2010), 17–32.
[**P4**] [64] P. Loreti, M. Mehrenberger, *"Observabilité uniforme de l'équation des ondes 1D, [Uniform observability of the 1D wave equation]* Paris-Sud Working Group on Modelling and Scientific Computing 2007–2008, 68–79, ESAIM Proc., 25, EDP Sci., Les Ulis, 2008.
[**P3**] [70] M. Mehrenberger, E. Violard, *A Hermite type adaptive semi-Lagrangian scheme,* Int. J. Appl. Math. Comput. Sci., 17 (3) (2007), 329–334.
[P2] [69] M. Mehrenberger, E. Violard, O. Hoenen, M. Campos Pinto, E. Sonnendrücker, *A Parallel Adaptive Vlasov Solver Based on Hierarchical Finite Element Interpolation,* Proceedings ICAP2004 St-Petersburg, Nuclear Inst. and Methods in Physics Research, A 558 (2006), 188–191.
[P1] [50] O. Hoenen, M. Mehrenberger, E. Violard, *Parallelization of an Adaptive Vlasov Solver,* ParSim04 proceedings, Lecture Notes in Computer Science, 3241 (2004), 430–435.
[P0] [17] M. Campos Pinto, M. Mehrenberger, *Adaptive numerical resolution of the Vlasov equation,* Numerical Methods for Hyperbolic and Kinetic Problems, CEMRACS 2003/IRMA Lectures in Mathematics and Theoretical Physics 7, 43–58.

3.4 Rapports/prépublications

[**S5**] [33] N. Crouseilles, M. Mehrenberger, F. Vecil, *A Discontinuous Galerkin semi-Lagrangian solver for the guiding-center problem*, hal-00717155, version 1, April 2012.
[**S4**] [28] N. Crouseilles, P. Glanc, M. Mehrenberger, C. Steiner *Finite Volume Schemes for Vlasov,* hal-00653038, version 1, December 2011.
[**S3**] [27] N. Crouseilles, E. Faou, M. Mehrenberger, *High order Runge-Kutta-Nyström splitting methods for the Vlasov- Poisson equation,* inria-00633934, version 1, October 2011.
[**S2**] [21] F. Charles, B. Després, M. Mehrenberger, *Enhanced convergence estimates for semi-Lagrangian schemes Application to the Vlasov-Poisson equation,* inria-00629081, version 1, October 2011.
[**S1**] [47] J. Guterl, J. P. Braeunig, N. Crouseilles, V. Grandgirard, G. Latu, M. Mehrenberger, E. Sonnendrücker, *Test of some numerical limiters for the conservative PSM scheme for 4D Drift-Kinetic simulations,* INRIA research report number 7467, November 2010.
[**S0**] [15] J. P. Braeunig, N. Crouseilles, V. Grandgirard, G. Latu, M. Mehrenberger, E. Sonnendrücker, *Some numerical aspects of the conservative PSM scheme in a 4D drift-kinetic code,* INRIA research report number 7109, November 2009.

Première partie

Chapitre 1

Inégalités d'Ingham et observabilité

1 Un théorème de type Ingham Beurling vectoriel

Baiocchi et al. ont généralisé il y a quelques années un théorème classique d'Ingham et Beurling en utilisant les différences divisées [5]. L'optimalité a été prouvée dans le cas scalaire dans [73]. On généralise ce résultat au cas vectoriel [7].

1.1 Introduction

Soit $\Omega := (\omega_k)_{k\in\mathbb{Z}}$ une famille de nombres réels satisfaisant la condition de gap

$$\gamma := \inf_{k\neq n} |\omega_k - \omega_n| > 0. \tag{1.1}$$

On note par $D^+ = D^+(\Omega)$ sa densité supérieure de Pólya, définie par la formule $D^+ := \lim_{r\to\infty} r^{-1}n^+(r)$, où $n^+(r) = n^+(\Omega, r)$ représente le plus grand nombre de termes de la suite $(\omega_k)_{k\in\mathbb{Z}}$ contenu dans un intervalle de longueur r.

Soit $(U_k)_{k\in\mathbb{Z}}$ une famille de vecteurs unitaires dans un espace de Hilbert complexe H et considérons les sommes

$$x(t) = \sum_{k\in\mathbb{Z}} x_k U_k e^{i\omega_k t}$$

avec des coefficients complexes de carré sommable x_k. On s'intéressé à la validité des estimations

$$\int_I |x(t)|_H^2 \, dt \asymp \sum_{k\in\mathbb{Z}} |x_k|^2 \tag{1.2}$$

où I est un intervalle borné de longueur $|I|$ et où l'on a écrit $A \asymp B$ s'il existe deux constantes strictement positives c_1, c_2 satisfaisant $c_1 A \leq B \leq c_2 A$.

Le résultat suivant généralise un théorème d'Ingham [56] ; pour $d = 1$ cela se réduit à un théorème de Beurling [11].

Théorème 1.1. *Soit $\Omega := (\omega_k)_{k\in\mathbb{Z}}$ une famille de nombres réels satisfaisant* (1.1).

(a) Si $|I| > 2\pi D^+$, alors les estimations (1.2) *sont satisfaites*

(b)Si les estimations (1.2)*sont satisfaites et H est de dimension finie d, alors $|I| \geq 2\pi D^+/d$.*

L'optimalité du Théorème 1.1 va se déduire du résultat suivant :

Théorème 1.2. *Soit Ω un ensemble de nombres réels de densité supérieure finie D^+ et soit $\alpha_1, \alpha_2, \dots$ une suite finie ou infinie de nombres dans $[0,1]$ satisfaisant $\alpha_1 + \alpha_2 + \cdots \geq 1$. Alors il existe une partition $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$ de Ω telle que la densité supérieure D_j^+ de Ω_j soit égale à $\alpha_j D^+$ pour chaque j.*

Remarque 1.3. *Il découle de la définition de la densité supérieure que si $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$ est une partition finie ou infinie de Ω, alors*

$$\max\{D^+(\Omega_1), D^+(\Omega_2), \dots\} \leq D^+(\Omega) \leq D^+(\Omega_1) + D^+(\Omega_2) + \cdots . \tag{1.3}$$

Ceci implique la nécessité des conditions $\alpha_j \leq 1$ et $\alpha_1 + \alpha_2 + \cdots \geq 1$ dans le théorème.

Maintenant, nous avons le corollaire suivant

Corollaire 1.4. *Soit $\Omega := (\omega_k)_{k\in\mathbb{Z}}$ une famille de nombre réels saitsfaisant (1.1), et H un espace de Hilbert de dimension finie. Etant donné un nombre réel arbitraire satisfaisant $\frac{1}{d} \leq \alpha \leq 1$ où $d = \dim H$, il existe une famille $(U_k)_{k\in\mathbb{Z}}$ de vecteurs unitaires dans H telle que les estimations (1.2) soient satisfaites si $|I| > 2\pi\alpha D^+$, et ne soient pas satisfaites si $|I| < 2\pi\alpha D^+$.*

On prouve le Théorème 1.1 dans la sous-section suivante et on l'étend dans le cas d'une condition d'écart affaibli dans la sous-section 1.3. Le Théorème 1.2 et le Corollaire 1.4 sont démontrés dans les sous-sections 1.4–1.6.

On réfère à [61] pour des applications variées au contrôle théorique de théorèmes de ce type.

1.2 Preuve du Théorème 1.1

La partie (a) découle du cas scalaire. En effet, en fixant une base orthonormée $(E_n)_{n\in N}$ de l'espace vectoriel fermé engendré par les $(U_k)_{k\in\mathbb{Z}}$ dans H et en développant les vecteurs U_k en série de Fourier : $U_k = \sum_{n\in N} u_{kn} E_n$, pour $|I| > 2\pi D^+$ on a

$$\begin{aligned}\int_I \Big\| \sum_{k\in\mathbb{Z}} x_k U_k e^{i\omega_k t} \Big\|_H^2 \, dt &= \sum_{n\in N} \int_I \Big| \sum_{k\in\mathbb{Z}} x_k u_{kn} e^{i\omega_k t} \Big|^2 \, dt \\ &\asymp \sum_{n\in N} \sum_{k\in\mathbb{Z}} |x_k u_{kn}|^2 \\ &= \sum_{k\in\mathbb{Z}} |x_k|^2.\end{aligned}$$

Pour la preuve de la partie (b) on adapte l'approche développée dans [46] et [73]. On pose $\gamma_k := 2\pi|I|^{-1}k$ pour simplifier. Etant donnés trois nombres réels y, r, R avec $r, R > 0$, on introduit les projections orthogonales

$$P_{y,r} : L^2(I,H) \to V_{y,r} \quad \text{and} \quad Q_{y,r+R} : L^2(I,H) \to W_{y,r+R}$$

sur les sous espaces vectoriels de dimension finie

$$V_{y,r} := \text{Vect}\left\{ U_k e^{i\omega_k t} \ : \ |\omega_k - y| < r \right\}$$

et

$$W_{y,r+R} := \mathrm{Vect}\left\{Ue^{i\gamma_k t} \ : \ |\gamma_k - y| < r+R \quad \text{and} \quad U \in H\right\}.$$

Notons que

$$n^+(2r) = \sup_y \dim V_{y,r} \tag{1.4}$$

et

$$(2r+2R)d\frac{|I|}{2\pi} \leq \dim W_{y,r+R} \leq (2r+2R+1)d\frac{|I|}{2\pi}. \tag{1.5}$$

En posant

$$S_{y,r,R} := P_{y,r} \circ Q_{y,r+R} \circ i$$

où i désigne l'injection $i : V_{y,r} \hookrightarrow L^2(I,H)$, on obtient une application linéaire de $V_{y,r}$ dans lui-même. On va étudier la trace de cette application.

Lemme 1.5. *On a*

$$|\mathrm{tr}(S_{y,r,R})| \leq \dim W_{y,r+R}.$$

Démonstration. On a

$$\|S_{y,r,R}\| \leq \|P_{y,r}\| \cdot \|Q_{y,r+R}\| \leq 1.$$

Ainsi les valeurs propres de $S_{y,r,R}$ ont leur module ≤ 1 et donc

$$|\mathrm{tr}(S_{y,r,R})| \leq \mathrm{rang}(S_{y,r,R}) \leq \dim(W_{y,r+R}). \qquad \square$$

Lemme 1.6. *En écrivant $e_k(t) := U_k e^{i\omega_k t}$ pour abréger, il existe $(\varphi_k)_{k\in\mathbb{Z}}$, une famille bornée biorthogonale à $(e_k)_{k\in\mathbb{Z}}$ dans $L^2(I,H)$ et on a*

$$\mathrm{tr}(S_{y,r,R}) = \dim V_{y,r} + \sum_{|\omega_k - y|<r} ((Q_{y,r+R} - \mathrm{Id})e_k, P_{y,r}\varphi_k)_H.$$

Démonstration. L'existence d'une famille biorthogonale bornée vient de (1.2) (voir [61] pour une preuve). On écrit ensuite $S_{y,r,R}e_k = \sum_{|\omega_j - y|<r} S_{k,j}e_j$. Puisque (φ_k) est biorthogonale, on a $(S_{y,r,R}e_k, \varphi_k)_{L^2(I,H)} = S_{k,k}$ et donc

$$\mathrm{tr}(S_{y,r,R}) = \sum_{|\omega_j - y|<r} S_{k,k} = \sum_{|\omega_k - y|<r} (S_{y,r,R}e_k, \varphi_k)_{L^2(I,H)},$$

de telle sorte que

$$\mathrm{tr}(S_{y,r,R}) = \sum_{|\omega_k - y|<r} (P_{y,r}e_k, \varphi_k)_{L^2(I,H)} + \sum_{|\omega_k - y|<r} (P_{y,r}(Q_{y,r+R} - \mathrm{Id})e_k, \varphi_k)_{L^2(I,H)}.$$

Puisque $P_{y,r}e_k = e_k$, on a $(P_{y,r}e_k, \varphi_k)_{L^2(I,H)} = 1$ et le résultat s'en déduit. $\square$

Lemme 1.7. *Pour $R \to \infty$ on a*

$$\|(Q_{y,r+R} - \mathrm{Id})e_k\| = O(1/\sqrt{R})$$

uniformément pour tout $y \in \mathbb{R}$, $r > 0$ et k satisfaisant $|\omega_k - y| < r$.

Démonstration. En fixant une base orthonormée $E_1, \ldots, E_d$ of H et en posant

$$f_{n,j}(t) := |I|^{-1/2} E_j e^{i\gamma_n t}$$

on a

$$e_k = \sum_{n\in\mathbb{Z}} \sum_{j=1}^{d} (e_k, f_{n,j})_{L^2(I,H)} f_{n,j}$$

et

$$Q_{y,r+R} e_k = \sum_{|\gamma_n - y| < r+R} \sum_{j=1}^{d} (e_k, f_{n,j})_{L^2(I,H)} f_{n,j}.$$

En appliquant l'égalité de Parseval, on obtient

$$\|(Q_{y,r+R} - \mathrm{Id})e_k\|^2 = \sum_{|\gamma_n - y| \geq r+R} \sum_{j=1}^{d} \left|(e_k, f_{n,j})_{L^2(I,H)}\right|^2.$$

Puisque

$$\left|(e_k, f_{n,j})_{L^2(I,H)}\right| = |I|^{-1/2} \left| \int_I (U_k, E_j)_H e^{i(\omega_k - \gamma_n)t}\, dt \right| \leq \frac{2|I|^{-1/2}}{|\omega_k - \gamma_n|}, \tag{1.6}$$

et $|\omega_k - y| < r$, on obtient alors

$$\|(Q_{y,r+R} - \mathrm{Id})e_k\|^2 \leq 4d|I|^{-1} \sum_{|\gamma_n - y| \geq r+R} \frac{1}{|\omega_k - \gamma_n|^2}$$

Notons qu'à partir de $|\gamma_n - y| \geq r + R$ et $|\omega_k - y| < r$, on obtient $|\gamma_n - \omega_k| > R$, et donc

$$\|(Q_{y,r+R} - \mathrm{Id})e_k\|^2 \leq 8d|I|^{-1} \sum_{n=0}^{\infty} \frac{1}{|2\pi|I|^{-1}n + R|^2}.$$

Puisque la dernière expression ne dépend pas de r, y et est un $O(1/R)$ quand $R \to \infty$, le lemme s'en déduit. □

Maintenant la preuve de la partie (b) du Théorème 1.1 peut être completée comme ceci. D'après les lemmes précédents, on a

$$\begin{aligned} \mathrm{dim} W_{y,r+R} &\geq |\mathrm{tr}(S)| \\ &= \left| \mathrm{dim} V_{y,r} + \sum_{|\omega_k - y| < r} ((Q_{y,r+R} - \mathrm{Id})e_k, P_{y,r}\varphi_k)_H \right| \\ &\geq \mathrm{dim} V_{y,r} - O(1/\sqrt{R}) \mathrm{dim} V_{y,r}. \end{aligned}$$

Donc

$$\dim V_{y,r} \leq (1 + O(1/\sqrt{R}))\dim W_{y,r+R}, \quad R \to \infty,$$

et en utilisant (1.4)–(1.5) on conclut que

$$n^+(2r) \leq (1 + O(1/\sqrt{R}))d\frac{|I|}{2\pi}(2r + 2R + 1), \quad R \to \infty.$$

On obtient alors

$$\begin{aligned} D^+ &= \lim_{r\to\infty} \frac{n^+(2r)}{2r} \\ &\leq (1 + O(1/\sqrt{R}))d\frac{|I|}{2\pi} \lim_{r\to\infty} \frac{2r + 2R + 1}{2r} \\ &= (1 + O(1/\sqrt{R}))d\frac{|I|}{2\pi} \end{aligned}$$

pour tout $R > 0$. En prenant $R \to \infty$, on conclut que $|I| \geq 2\pi D^+/d$.

1.3 Le cas des différences divisées

La condition d'écart (1.1) du théorème peut être affaiblie. En suivant [5], soit $(\omega_k)_{k\in\mathbb{Z}}$ une suite croissante de nombres réels satisfaisant pour un entier strictement positif M et un nombre réel strictement positif γ' la condition décart affaiblie

$$\omega_{k+M} - \omega_k \geq M\gamma' \quad \text{pour tout} \quad k \in \mathbb{Z}. \tag{1.7}$$

Cela implique que$D^+ < \infty$. On dit que $\omega_m, \ldots, \omega_{m+j-1}$ est une chaine d'exposants γ'-proche ($m \in \mathbb{Z}$, $j = 1, \ldots, M$) si

$$\begin{cases} \omega_m - \omega_{m-1} \geq \gamma', \\ \omega_k - \omega_{k-1} < \gamma' \quad \text{pour } k = m+1, \ldots, m+j-1, \\ \omega_{m+j} - \omega_{m+j-1} \geq \gamma'. \end{cases}$$

On définit ensuite les différences divisées $f_\ell = [\omega_m, \ldots, \omega_\ell]$ par la formule

$$[\omega_m](t) := \exp(i\omega_m t), \ [\omega_m, \omega_{m+1}](t) := it \int_0^1 \exp(i[s_m(\omega_{m+1} - \omega_m) + \omega_m]t)ds_m,$$

et pour $\ell = m+2, \ldots, m+j-1$,

$$\begin{aligned} [\omega_m, \ldots, \omega_\ell](t) \ := (it)^{\ell-m} \int_0^1 \int_0^{s_m} \cdots \int_0^{s_{\ell-2}} \\ \exp(i[s_{\ell-1}(\omega_\ell - \omega_{\ell-1}) + \cdots + s_m(\omega_{m+1} - \omega_m) + \omega_m)]t) \ ds_{\ell-1} \ldots \ ds_m. \end{aligned}$$

On peut maintenant établir une généralisation du Théorème 1.1 :

Théorème 1.8. *Théorème 1.1 est vrai si* (1.1) *est remplacé par* (1.7) *et* $e^{i\omega_k t}$ *est remplacé par* $f_k(t)$.

Démonstration. La plupart de la preuve du Théorème 1.1 peut être facilement adaptée. Pour la partie (b) on doit remplacer l'estimation (1.6) par :

$$\Big|\int_I (U_k, E_j)_H f_k(t) e^{-i\gamma_n t} dt\Big| \leq \Big|\int_I f_k(t) e^{-i\gamma_n t} dt\Big| \leq \frac{C}{|\omega_k - \gamma_n|}, \tag{1.8}$$

avec une constante C dependant seulement de γ', M et I. Cela se montre par un même argument que dans [73]. On a

$$A := \int_I f_k(t) e^{-i\gamma_n t} dt = \int_I g(t) e^{i\omega_k t} e^{-i\gamma_n t} dt$$

avec

$$g(t) = [\omega_m - \omega_k, \ldots, \omega_k - \omega_k](t).$$

En intégrant par parties sur $I = (a, b)$ on obtient que

$$A = \left[\frac{1}{i\omega_k - i\gamma_n} g(t) e^{i\omega_k t} e^{-i\gamma_n t}\right]_a^b - \int_I \frac{1}{i\omega_k - i\gamma_n} g'(t) e^{i\omega_k t} e^{-i\gamma_n t} dt.$$

Maintenant un calcul direct montre que pour tous nombres réels $\mu_1, \ldots, \mu_r$ les différences divisées satisfont l'inégalité

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

Ainsi, dans notre cas, grâce à la propriété d'exposants γ'-proches, on a

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

et cela donne (1.8). □

1.4 Preuve du Théorème 1.2 pour $\alpha_1 + \alpha_2 + \cdots = 1$

Supposons que le théorème est valable pour les ensembles Ω qui sont bornés inférieurement. Alors, en changeant Ω en $-\Omega$ on obtient que le théorème est aussi vrai si Ω est borné supérieurement. Finalement, si $\inf \Omega = -\infty$ et $\sup \Omega = \infty$, alors en appliquant ces deux spéciaux du théorème à

$$\Omega^- := \Omega \cap (-\infty, 0) \quad \text{et} \quad \Omega^+ := \Omega \cap [0, \infty)$$

on obtient deux partitions

$$\Omega^- = \Omega_1^- \cup \Omega_2^- \cup \cdots \quad \text{et} \quad \Omega^+ = \Omega_1^+ \cup \Omega_2^+ \cup \cdots$$

satisfaisant

$$D^+(\Omega_j^-) = \alpha_j D^+(\Omega^-) \quad \text{et} \quad D^+(\Omega_j^+) = \alpha_j D^+(\Omega^+)$$

pour tout j. Alors, en posant $\Omega_j := \Omega_j^- \cup \Omega_j^+$ on obtient une partition Ω avec les propriétés requises. Cela découle en appliquant le lemme suivant aux partitions $\Omega := \Omega^- \cup \Omega^+$ et $\Omega_j := \Omega_j^- \cup \Omega_j^+$.

Lemme 1.9. *Pour tout ensemble A de nombres réels, en posant $A^- := A \cap (-\infty, 0)$ et $A^+ := A \cap [0, \infty)$ on a*

$$D^+(A) = \max\{D^+(A^-), D^+(A^+)\}.$$

Démonstration. L'inégalité facile $\geq$ découle de (1.3). En posant

$$M := \max\{D^+(A^-), D^+(A^+)\}$$

pour abréger, pour l'inégalité inverse, il suffit de montrer que

$$\limsup_{r,s,\geq 0, r+s\to\infty} \frac{\mathrm{card}(A \cap [-r, s])}{r+s} \leq M.$$

Le cas $M = \infty$ est évident. Supposons donc que $M < \infty$. Pour chaque $\epsilon > 0$ fixé, on peut fixer deux nombres strictement positifs r_ϵ, s_ϵ satisfaisant

$$\mathrm{card}(A^- \cap [-r, 0]) \leq (D^+(A^-) + \epsilon)r \quad \text{for all} \quad r \geq r_\epsilon$$

and

$$\mathrm{card}(A^+ \cap [0, s]) \leq (D^+(A^+) + \epsilon)s \quad \text{for all} \quad s \geq s_\epsilon.$$

En ajoutant les deux inégalités et posant $K := \mathrm{card}(A \cap [-r_\epsilon, s_\epsilon])$ il s'ensuit que

$$\mathrm{card}(A \cap [-r, s]) \leq (M + \epsilon)(r + s) + K$$

pour tout $r, s \geq 0$. En divisant par $r + s$ et laissant $r + s \to \infty$ on conclut que

$$\limsup_{r,s,\geq 0, r+s\to\infty} \frac{\mathrm{card}(A \cap [-r, s])}{r+s} \leq M + \epsilon$$

pour tout $\epsilon > 0$, et le lemme en découle. □

Ainsi, on peut supposer que Ω est borné inférieurement. Si $D^+ = 0$, alors chaque partition de Ω a la propriété requise puisque toutes les densités supérieures sont égales à zéro. Ainsi, on peut aussi supposer que $0 < D^+ < \infty$; alors Ω est un ensemble non borné et on peut énumérer les éléments de Ω en une suite infinie strictement croissante $\omega_1 < \omega_2 < \cdots$. Finalement, en choisissant $\Omega_j = \emptyset$ chaque fois que $\alpha_j = 0$ on peut supposer sans perte de généralité que $0 < \alpha_j \leq 1$ pour tout j.

Afin d'expliquer l'idée de la preuve, on considère le cas spécial où la suite finie $\alpha_1, \ldots, \alpha_d$ est constituée de nombres rationnels. On fixe un entier positif N tel que $N\alpha_1, \ldots, N\alpha_d$ sont tous entiers, et on représente Ω comme union de blocs disjoins

$$B_n := \{\omega_k \in \Omega \ : \ k = nN + 1, nN + 2, \ldots, (n + 1)N\}, \quad n = 0, 1, \ldots. \tag{1.9}$$

Remarquons que chaque B_n a N éléments. Par conséquent, puisque $N\alpha_1 + \cdots + N\alpha_d = N$, on peut définir une partition $\Omega = \Omega_1 \cup \cdots \cup \Omega_d$ de Ω telle que

$$\mathrm{card}(B_n \cap \Omega_j) = N\alpha_j \quad \text{pour tout } n \text{ et } j. \tag{1.10}$$

On affirme que $D_j^+ \le \alpha_j D^+$ pour chaque j. C'est évident lorsque $D_j^+ = 0$. Si $D_j^+ > 0$ pour un j, alors on choisit une suite d'intervalles bornés (I_m^j) satisfaisant $|I_m^j| \to \infty$ et

$$\frac{\operatorname{card}(\Omega_j \cap I_m^j)}{|I_m^j|} \to D_j^+.$$

Puisque $D_j^+ > 0$, ainsi $\operatorname{card}(\Omega_j \cap I_m^j) \to \infty$ et donc $k_m \to \infty$ où k_m désigne le nombre de blocs consécutifs B_n contenus dans I_m^j. Puisque $I_m^j \cap \Omega$ est contenu dans l'union de $k_m + 2$ blocs B_n, en utilisant (1.10) il en découle que

$$\begin{aligned}\frac{\operatorname{card}(\Omega_j \cap I_m^j)}{|I_m^j|} &\le \frac{2N + k_m N\alpha_j}{|I_m^j|} \\ &\le \frac{2N + k_m N\alpha_j}{|I_m^j|} \cdot \frac{\operatorname{card}(\Omega \cap I_m^j)}{k_m N} \\ &\le \frac{2 + k_m\alpha_j}{k_m} \cdot \frac{n^+(\Omega, |I_m^j|)}{|I_m^j|}\end{aligned}$$

pour chaque m. En prenant $m \to \infty$ on conclut que $D_j^+ \le \alpha_j D^+$.

En fait $D_j^+ = \alpha_j D^+$ pour chaque j. En effet, si les inégalités $D_j^+ \le \alpha_j D^+$ n'étaient pas toutes des égalités, alors en utilisant (1.3) on obtiendrait une contradiction :

$$D^+ \le D_1^+ + \cdots + D_d^+ < \alpha_1 D^+ + \cdots + \alpha_d D^+ = D^+.$$

Maintenant on se tourne vers le cas général. On écrit $J = \{1, \ldots, d\}$ pour le cas fini $\alpha_1 + \cdots + \alpha_d = 1$ et $J = \{1, 2, \ldots\}$ pour le cas infini.

Soit

$$(k, j) : \{1, 2, \ldots\} \to \{1, 2, \ldots\} \times J$$

l'énumération lexicographique strictement croissante des paires

$$(k, j) \in \{1, 2, \ldots\} \times J \quad \text{satisfying} \quad [k\alpha_j] > [(k-1)\alpha_j],$$

où $[x]$ désigne la partie entière de x. Cette énumération est possible puisque pour chaque k seulement un nombre fini d'indices $j \in J$ peuvent satisfaire cette inégalité. En effet, par la convergence de la série $\sum_{j\in J} \alpha_j$ on a $k\alpha_j < 1$ et donc $[k\alpha_j] = [(k-1)\alpha_j] = 0$ pour tous les indices j suffisament grands.

Observons que

$$[k\alpha_j] - [(k-1)\alpha_j] = \begin{cases} 1 & \text{si } (k, j) \text{ appartient à la suite;} \\ 0 & \text{sinon.} \end{cases}$$

Pour chaque $j \in J$ on pose

$$\Omega_j := \{\omega_s \ : \ j(s) = j\}.$$

On affirme que $D_j^+ = \alpha_j D^+$ pour tout $j \in J$.

Tout d'abord, on montre que $D_j^+ \le \alpha_j D^+$ pour chaque $j \in J$. Le cas $D_j^+ = 0$ est évident. Soit $D_j^+ > 0$ et choisissons une suite (I_m^j) d'intervalles bornés tels que

$$|I_m^j| \to \infty \quad \text{and} \quad \frac{\operatorname{card}(\Omega_j \cap I_m^j)}{|I_m^j|} \to D_j^+.$$

En écrivant $\Omega \cap I_m^j = \{\omega_{s_m}, \dots, \omega_{t_m}\}$ on a

$$\ell_m := k(t_m) - k(s_m) \geq \operatorname{card}(\Omega_j \cap I_m^j) \to \infty;$$

la première inégalité découle de la définition de Ω_j, tandis que la seconde découle des hypothèses $D_j^+ > 0$ et $|I_m^j| \to \infty$.
Maintenant on a

$$\begin{aligned}\operatorname{card}(\Omega_j \cap I_m^j) &\leq \sum_{k=k(s_m)}^{k(t_m)} ([k\alpha_j] - [(k-1)\alpha_j]) \\ &= [k(t_m)\alpha_j] - [(k(s_m)-1)\alpha_j] \\ &\leq \ell_m \alpha_j + 1\end{aligned}$$

d'après la définition de Ω_j. De plus, on a

$$\begin{aligned}\operatorname{card}(\Omega \cap I_m^j) &\geq \sum_{n \leq \sqrt{\ell_m}} \sum_{k=k(s_m)+1}^{k(t_m)-1} ([k\alpha_n] - [(k-1)\alpha_n]) \\ &= \sum_{n \leq \sqrt{\ell_m}} ([(k(t_m)-1)\alpha_n] - [k(s_m)\alpha_n]) \\ &\geq \sum_{n \leq \sqrt{\ell_m}} (\ell_m \alpha_n - 2) \\ &\geq \ell_m \Big(\sum_{n \leq \sqrt{\ell_m}} \alpha_n \Big) - 2\sqrt{\ell_m}.\end{aligned}$$

Puisque $\ell_m \to \infty$ et $\sum \alpha_n = 1$, il découle des deux précédentes estimations que

$$\operatorname{card}(\Omega_j \cap I_m^j) \leq \ell_m \left(\alpha_j + o(1)\right)$$

et

$$\operatorname{card}(\Omega \cap I_m^j) \geq \ell_m \left(1 - o(1)\right)$$

quand $m \to \infty$. Ainsi, on a

$$\begin{aligned}\frac{\operatorname{card}(\Omega_j \cap I_m^j)}{|I_m^j|} &\leq \frac{\ell_m \left(\alpha_j + o(1)\right)}{|I_m^j|} \\ &\leq \frac{\ell_m \left(\alpha_j + o(1)\right)}{|I_m^j|} \cdot \frac{\operatorname{card}(\Omega \cap I_m^j)}{\ell_m \left(1 - o(1)\right)} \\ &= \frac{\alpha_j + o(1)}{1 - o(1)} \cdot \frac{\operatorname{card}(\Omega \cap I_m^j)}{|I_m^j|}.\end{aligned}$$

En posant $m \to \infty$ on conclut que

$$D_j^+ \leq \alpha_j \limsup \frac{\operatorname{card}(\Omega \cap I_m^j)}{|I_m^j|} \leq \alpha_j D^+.$$

Il reste à montrer qu'aucune des inégalités $D_j^+ \leq \alpha_j D^+$ est stricte. Si c'était le cas, en utilisant (1.3) on obtiendrait une contradiction

$$D^+ \leq D_1^+ + D_2^+ + \dots < \alpha_1 D^+ + \alpha_2 D^+ + \dots = D^+.$$

1.5 Preuve du Corollaire 1.4

Fixons $1/d \leq \alpha \leq 1$ arbitraire et choisissons ensuite $\alpha_1, \ldots, \alpha_d \geq 0$ tels que

$$\alpha_1 + \cdots + \alpha_d = 1 \quad \text{and} \quad \max\{\alpha_1, \ldots, \alpha_d\} = \alpha.$$

En appliquant la partie déjà prouvée du Théorème 1.2 on obtient une partition $\Omega = \Omega_1 \cup \cdots \cup \Omega_d$ de Ω telle que $D^+(\Omega_j) = \alpha_j D^+$ pour tout j. Fixons une base orthonormée $E_1, \ldots, E_d$ de H et posons $U_k = E_j$ si $\omega_k \in \Omega_j$. Alors, en utilisant l'identité

$$\int_I \Big\| \sum_{k\in\mathbb{Z}} x_k U_k e^{i\omega_k t} \Big\|_H^2 \, dt = \sum_{j=1}^{d} \int_I \Big| \sum_{\omega_k \in \Omega_j} x_k e^{i\omega_k t} \Big|^2 \, dt$$

et en appliquant le cas scalaire du théorème, on conclut que les estimations (1.2) sont valables si $|I| > 2\pi\alpha D^+$, et ne sont pas valables si $|I| < 2\pi\alpha D^+$.

1.6 Preuve du Théorème 1.2 pour $\alpha_1 + \alpha_2 + \cdots > 1$

En utilisant le même raisonnement qu'au début de la sous-section 1.6, on peut supposer que $0 < D^+ < \infty$, $0 < \alpha_j \leq 1$ pour tout j, et que les éléments de Ω forment une suite infinie strictement croissante $\omega_1 < \omega_2 < \cdots$.

Choisissons d'abord un entier strictement positif tel que

$$\sum_j \left([(n+1)N\alpha_j] - [nN\alpha_j] \right) \geq N \quad \text{pour tout} \quad n = 0, 1, \ldots. \tag{1.11}$$

Pour cela, choisissons un entier strictement positif k satisfaisant $\alpha_1 + \cdots + \alpha_k > 1$, et ensuite un entier strictement positif N tel que

$$\frac{k}{N} < \alpha_1 + \cdots + \alpha_k - 1.$$

Ensuite, en utilisant l'inégalité $[x+y] \geq [x] + [y]$ on obtient l'estimation suivante pour tout $n = 0, 1, \ldots$:

$$\begin{aligned} \frac{1}{N} \sum_j \left([(n+1)N\alpha_j] - [nN\alpha_j] \right) &\geq \frac{1}{N} \sum_{j=1}^{k} [N\alpha_j] \\ &> \frac{1}{N} \sum_{j=1}^{k} (N\alpha_j - 1) \\ &= \alpha_1 + \cdots + \alpha_k - \frac{k}{N} \\ &> 1. \end{aligned}$$

Notons que

$$0 \leq [(n+1)N\alpha_j] - [nN\alpha_j] \leq N \tag{1.12}$$

pour tout n et j puisqu'en utilisant la condition $\alpha_j \leq 1$ on a

$$[(n+1)N\alpha_j] = [nN\alpha_j + N\alpha_j] \leq [nN\alpha_j + N] = [nN\alpha_j] + N.$$

Ensuite, on représente Ω de nouveau comme une union de blocs disjoints B_n de N éléments comme dans (1.9). Puisque la densité supérieure de Ω ne change pas si on enlève un nombre fini de termes initiaux, on peut définir par récurrence une suite d'intervalles bornés (I_m) ayant les quatre propriétés suivantes :

$$\begin{aligned}
&\sup I_m < \inf I_{m+1} \quad \text{pour tout } m\,;\\
&\text{aucun bloc } B_n \text{ appartient à plus qu'un seul intervalle } I_m\,;\\
&|I_m| \to \infty;\\
&\frac{\operatorname{card}(\Omega \cap I_m)}{|I_m|} \to D^+.
\end{aligned} \tag{1.13}$$

Introduisons aussi une suite de nombres strictement positfs contenant chaque indice j une infinité de fois, par exemple

$$(b_m) := 1 \dots d\ 1 \dots d\ 1 \dots d\ \dots$$

dans le cas d'une suite finie $\alpha_1, \dots, \alpha_d$, et

$$(b_m) := 12\ 123\ 1234\ 12345\ 12 \dots$$

dans le cas d'une suite infinie $\alpha_1, \alpha_2, \dots$.

Maintenant, grâce à (1.11) et (1.12) on peut définir une partition $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$ ayant les propriétés suivantes, où on désigne par B'_m l'union des k_m blocs consecutifs $B_{n_m+1}, \dots, B_{n_m+k_m}$ contenus dans I_m :

$$\operatorname{card}(B_n \cap \Omega_j) \le [(n+1)N\alpha_j] - [nN\alpha_j] \quad \text{pour tout } n \text{ et } j\,; \tag{1.14}$$

$$\operatorname{card}(B_n \cap \Omega_j) = [(n+1)N\alpha_j] - [nN\alpha_j] \quad \text{si } B_n \subset B'_m \text{ et } j = b_m. \tag{1.15}$$

On affirme que $D_j^+ \le \alpha_j D^+$ pour chaque j. La preuve est similaire au cas de la partie (a). Le cas $D_j^+ = 0$ est évident. Si $D_j^+ > 0$ pour certains j, alors choisissons une suite d'intervalles bornés (I_m^j) satisfaisant $|I_m^j| \to \infty$ et

$$\frac{\operatorname{card}(\Omega_j \cap I_m^j)}{|I_m^j|} \to D_j^+.$$

Puisque $D_j^+ > 0$, ainsi $\operatorname{card}(\Omega_j \cap I_m^j) \to \infty$ et donc $k_m \to \infty$. Puisque $I_m^j \cap \Omega$ est contenu dans une union de $k_m + 2$ blocs $B_{n_m}, \dots, B_{n_m+k_m+1}$, en utilisant (1.14) et l'inégalité

$$[x+y] - [x] \le [y] + 1 \le y + 1 \le y + N$$

il s'en suit que

$$\begin{aligned}
\frac{\operatorname{card}(\Omega_j \cap I_m^j)}{|I_m^j|} &\le \frac{1}{|I_m^j|}\left(2N + \sum_{n=n_m+1}^{n_m+k_m} [(n+1)N\alpha_j] - [nN\alpha_j]\right)\\
&= \frac{2N + [(n_m+k_m+1)N\alpha_j] - [(n_m+1)N\alpha_j]}{|I_m^j|}\\
&\le \frac{3N + k_m N\alpha_j}{|I_m^j|}\\
&\le \frac{3N + k_m N\alpha_j}{|I_m^j|} \cdot \frac{\operatorname{card}(\Omega \cap I_m^j)}{k_m N}\\
&\le \frac{3 + k_m\alpha_j}{k_m} \cdot \frac{n^+(\Omega, |I_m^j|)}{|I_m^j|}
\end{aligned}$$

pour chaque m. En prenant $m \to \infty$ on conclut que $D_j^+ \leq \alpha_j D^+$.

On affirme que $D_j^+ \geq \alpha_j D^+$ et donc $D_j^+ = \alpha_j D^+$ pour chaque j. En effet, pour chaque j fixé, il existe des intervalles arbitrairement longs I_m tels que $b_m = j$. Pour ces intervalles,en utilisant (1.15) et l'inégalité

$$[x+y] - [x] \geq [y] > y - 1 \geq y - N$$

on obtient

$$\begin{aligned}
\operatorname{card}(\Omega_j \cap I_m) &\geq \operatorname{card}(B'_m \cap \Omega_j) \\
&= \sum_{n=n_m+1}^{n_m+k_m} [(n+1)N\alpha_j] - [nN\alpha_j] \\
&= [(n_m+k_m+1)N\alpha_j] - [(n_m+1)N\alpha_j] \\
&\geq k_m N\alpha_j - N
\end{aligned}$$

et donc

$$\begin{aligned}
\frac{\operatorname{card}(\Omega_j \cap I_m)}{|I_m|} &\geq \frac{k_m N\alpha_j - N}{|I_m|} \\
&\geq \frac{k_m N\alpha_j - N}{|I_m|} \cdot \frac{\operatorname{card}(\Omega \cap I_m)}{(k_m+2)N} \\
&= \frac{k_m\alpha_j - 1}{k_m+2} \cdot \frac{\operatorname{card}(\Omega \cap I_m)}{|I_m|}.
\end{aligned}$$

On en déduit que

$$\frac{n_j^+(\Omega_j, |I_m|)}{|I_m|} \geq \frac{k_m\alpha_j - 1}{k_m+2} \cdot \frac{\operatorname{card}(\Omega \cap I_m)}{|I_m|}.$$

Puisque $D^+ > 0$ et $|I_m| \to \infty$, par (1.13) on a $\operatorname{card}(\Omega \cap I_m) \to \infty$ et donc $k_m \to \infty$. Par conséquent, en prenant $|I_m| \to \infty$ on conclut que $D_j^+ \geq \alpha_j D^+$.

2 Une preuve de type Ingham pour l'observabilité frontière d'une équation des ondes $N - d$

L'observabilité frontière de l'équation des ondes a été étudiée par de nombreux auteurs. Une méthode de choix est d'utiliser la méthode des multiplicateurs (cf [60]). Récemment, dans [92], une première preuve basée sur les séries de Fourier a été donnée dans le cas où le domaine est un carré grâce à un test de type Hautus. On donne ici [74] une nouvelle preuve auto-contenue par une approche de type Ingham, dans le cas plus général où le domaine est un produit d'intervalles ; on obtient alors un temps et des constantes explicites, contrairement à la preuve de [92]. Cependant, on n'atteint pas le temps optimal, qui peut être obtenu pour ce problème par la méthode des multiplicateurs.

2.1 Introduction

Soit $N \in \mathbb{N}^*$, des réels $L_i > 0,\ i = 1, \dots, N$, $\Omega = \prod_{i=1}^N]0, L_i[$ et $u(t,x)$ solution de

$$\begin{cases} u'' = \Delta u, \quad 0 < t < T, \quad x \in \Omega \\ u = 0, \quad 0 < t < T, \quad x \in \partial\Omega \\ u(0,x) = u_0(x), \quad u'(0,x) = u_1(x), \quad x \in \Omega. \end{cases} \tag{2.1}$$

Le système est bien défini pour des données initiales satisfaisant $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$. On a la proposition classique suivante.

Proposition 2.1. *Soit* $\Gamma = \cup_{j=1}^N \left(\prod_{i=1}^{j-1} [0, L_i] \times \{0\} \times \prod_{i=j+1}^N [0, L_i] \right)$. *Il existe* $T_0 > 0$ *tel que pour* $T > T_0$, *le système (2.1) soit observable : il existe une constante* $c > 0$ *telle que l'on ait*

$$\|u_0\|^2_{H_0^1(\Omega)} + \|u_1\|^2_{L^2(\Omega)} \leq c \int_0^T \int_\Gamma |\partial_\nu u(t,x)|^2 \, \mathrm{d}\Gamma \mathrm{d}t, \tag{2.2}$$

pour tout $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$.

Une méthode de choix pour étudier une telle inégalité d'observabilité est d'utiliser la méthode des multiplicateurs (cf [60]). On peut ici effectivement appliquer cette méthode et obtenir (2.2) avec un temps optimal $T_0 = 2\sqrt{\sum_{i=1}^N L_i^2}$.
Dans certaines situations, on peut aussi essayer d'utiliser une inégalité d'Ingham (cf [56, 61]).
La solution de (2.1) est explicitement donnée en termes d'une série de Fourier :

$$u(t,x) = \sum_{k \in (\mathbb{N}^*)^N} \left(\alpha_k e^{i\omega_k t} + \alpha_{-k} e^{-i\omega_k t} \right) \prod_{j=1}^N \sin\left(\frac{\pi k_j x_j}{L_j} \right), \tag{2.3}$$

avec des coefficients α_k adéquats, et avec les notations

$$k = (k_1, \dots, k_N), \ x = (x_1, \dots, x_N), \ \omega_k = \sqrt{\sum_{j=1}^N \left(\frac{k_j \pi}{L_j} \right)^2}.$$

Une question naturelle est alors : peut-on obtenir (2.2) avec des arguments de type Ingham ?

La difficulté principale, dans le cadre multi-dimensionnel ($N > 1$) est que des valeurs propres en nombre arbitraire peuvent être arbitrairement proches, ce qui rend l'obtention de l'inégalité d'observabilité par une telle approche plus difficile. Notons que de tels cas ont déjà été considérés dans d'autres situations (voir [65], [63]). Certains théorèmes multi-dimensionnels de type Ingham ont été établis et ont permis de traiter par exemple du cas de l'équation des plaques avec contrôle interne [59]. Notons que (2.2) a été prouvé pour le cas du carré à l'aide de séries de Fourier, grâce à un test de type Hautus dans [92], sans temps explicite d'observation. Nous mentionnons que l'étude de l'inégalité d'observabilité peut aussi être effectuée par plusieurs autres techniques, telles que l'analyse micro-locale ou les estimations de Carleman. Le but de cette présente section est de fournir une preuve auto-contenue

de type Ingham de (2.2), dans le cas (déjà précisé) où le domaine est un produit d'intervalles. On utilise d'abord l'orthogonalité des vecteurs propres, puis on applique la première méthode d'Ingham. Grâce à une décomposition convenable, on peut se débarasser des termes où il n'y a pas d'écart spectral (comme dans [65]) et on peut donc obtenir l'inégalité d'observabilité désirée. Le temps d'observation, ainsi que les constantes impliquées dans l'inégalité peuvent être explicitées, mais nous n'avons pas pu atteindre le temps optimal par cette méthode. Notons aussi le travail [77] où une approche utilisant les séries de Fourier a également été utilisée pour le problème bidimensionnel ; les auteurs n'arrivent néanmoins pas non plus au temps optimal. Il semblerait que cette question reste ouverte.
Pour la suite, on va écrire $A \asymp B$ à la place de $c_1 A \leq B \leq c_2 A$ et $A \gtrsim B$ à la place de $A \geq c_1 B$ pour abréger, où $c_1, c_2 > 0$ sont des constantes.

2.2 Une nouvelle inégalité d'Ingham

On présente ici une nouvelle inégalité d'Ingham qui va nous permettre de prouver la Proposition 2.1 comme application, et qui peut avoir son intérêt propre. Notons d'ailleurs que le théorème suivant a été appliqué ou adapté dans [58, 79].

Théorème 2.2. *Soit $d \in \mathbb{N}^*$, des réels $(\lambda_k)_{k\in(\mathbb{N}^*)^d}$ et des complexes $(p_\ell)_{\ell\in\mathbb{N}^*}$.*
On suppose la condition d'écart suivante : pour $j = 1, \dots, d$, il existe $\gamma_j > 0$ tel que

$$\left|\lambda_{k_1,\dots,k_{j-1},k_j,k_{j+1},\dots,k_d} - \lambda_{k_1,\dots,k_{j-1},k'_j,k_{j+1},\dots,k_d}\right| \geq \gamma_j \left|k_j - k'_j\right|, \tag{2.4}$$

$$\left|\lambda_{k_1,\dots,k_{j-1},k_j,k_{j+1},\dots,k_d} + \lambda_{k_1,\dots,k_{j-1},k'_j,k_{j+1},\dots,k_d}\right| \geq \gamma_j \left|k_j + k'_j\right|, \tag{2.5}$$

pour tous les indices $k = (k_1, \dots, k_d) \in (\mathbb{N}^*)^d$ et $k'_j \in \mathbb{N}^*$ tels que les poids $(p_\ell)_{\ell\in\mathbb{N}^*}$ satisfassent

$$\max_{i=1,\dots,d,\ i\neq j} |p_{k_i}| \leq \max\left(\left|p_{k_j}\right|, \left|p_{k'_j}\right|\right). \tag{2.6}$$

Alors, pour $T > 2\pi\sqrt{\sum_{i=1}^d \frac{1}{\gamma_i^2}}$, il existe une constante $c_1 > 0$ telle que l'on ait,

$$\sum_{j=1}^{d} \sum_{k_1,\dots,k_{j-1},k_{j+1},\dots,k_d\in\mathbb{N}^*} \int_0^T \left|\sum_{k_j\in\mathbb{N}^*} p_{k_j}\left(\beta_k e^{i\lambda_k t} + \beta_{-k} e^{-i\lambda_k t}\right)\right|^2 \mathrm{d}t$$
$$\geq c_1 \sum_{k\in(\mathbb{N}^*)^d} \left(|\beta_k|^2 + |\beta_{-k}|^2\right)\left(\sum_{j=1}^{d} \left|p_{k_j}\right|^2\right),$$

pour tous les complexes $(\beta_k)_{k\in(\mathbb{N}^*)^d}$ et $(\beta_{-k})_{k\in(\mathbb{N}^*)^d}$, telles que les sommes impliquées soient finies.

La preuve consiste à utiliser la première méthode d'Ingham. On rappelle ici la définition et les principales propriétés qui vont être utilisées. On considère la fonction

$$f(t) = \begin{cases} \cos\frac{\pi t}{T} & \text{si } |t| \leq T/2 \\ 0 & \text{si } |t| > T/2. \end{cases}$$

Sa transformée de Fourier $\hat{k}$ satisfait : $\hat{f}(x) = \int_{-\infty}^{\infty} k(t)e^{ixt}\mathrm{d}t = -\frac{2T\pi\cos(xT/2)}{x^2T^2-\pi^2}$. On a $\hat{f}(0) = \frac{2T}{\pi}$, et pour $\gamma > \frac{2\pi}{T}$, $\ell \in \mathbb{N}^*$ et $|x| \geq \ell\gamma$,

$$|\hat{f}(x)| \leq \frac{2T}{\pi}\frac{1}{|\frac{x^2T^2}{\pi^2}-1|} = \frac{2T}{\pi}\big(\frac{2\pi}{\gamma T}\big)^2\frac{1}{|4\big(\frac{x}{\gamma}\big)^2-\big(\frac{2\pi}{\gamma T}\big)^2|} \leq \frac{2T}{\pi}\big(\frac{2\pi}{\gamma T}\big)^2\frac{1}{4\ell^2-1}. \quad (2.7)$$

On rappelle aussi que

$$\sum_{\ell=1}^{\infty}\frac{1}{4\ell^2-1} = \frac{1}{2}\sum_{\ell=1}^{\infty}\big(\frac{1}{2\ell-1}-\frac{1}{2\ell+1}\big) = \frac{1}{2}. \quad (2.8)$$

On pose $b_k(t) = \beta_k e^{i\lambda_k t} + \beta_{-k}e^{-i\lambda_k t}$. Maintenant, puisque $f(t) \leq 1_{[-T/2,T/2]}$, on a

$$A := \int_{-T/2}^{T/2}\left|\sum_{k_j\in\mathbb{N}^*} p_{k_j}b_k(t)\right|^2 \mathrm{d}t \geq \int_{\mathbb{R}} f(t)\left|\sum_{k_j\in\mathbb{N}^*} p_{k_j}b_k(t)\right|^2 \mathrm{d}t.$$

On note $\delta_j = \max_{i=1,\dots,d,\ i\neq j}|p_{k_i}|$ et on a alors la décomposition

$$\begin{aligned} A &\geq \int_{\mathbb{R}} k(t)\left|\sum_{k_j\in\mathbb{N}^*,\ \delta_j\leq|p_{k_j}|} p_{k_j}b_k(t) + \sum_{k_j\in\mathbb{N}^*,\ \delta_j>|p_{k_j}|} p_{k_j}b_k(t)\right|^2 \mathrm{d}t \\ &= \sum_{(k_j,k_j')\in\Lambda_1} + \sum_{(k_j,k_j')\in\Lambda_2} + \sum_{(k_j,k_j')\in\Lambda_3} B_k + \int_{\mathbb{R}} f(t)\left|\sum_{k_j\in\mathbb{N}^*,\ \delta_j>|p_{k_j}|} p_{k_j}b_k(t)\right|^2 \mathrm{d}t, \quad (2.9) \end{aligned}$$

en écrivant pour abréger $k' = (k_1,\dots,k_{j-1},k_j',k_{j+1},\dots,k_d)$,

$$\Lambda_1 := \left\{(k_j,k_j')\in\mathbb{N}^*,\ \delta_j\leq|p_{k_j}|, \delta_j\leq\left|p_{k_j'}\right|\right\},$$

$$\Lambda_2 := \left\{(k_j,k_j')\in\mathbb{N}^*,\ \delta_j\leq|p_{k_j}|,\ \delta_j>\left|p_{k_j'}\right|\right\},$$

$$\Lambda_3 := \left\{(k_j,k_j')\in\mathbb{N}^*,\ \delta_j\leq|p_{k_j}|,\ \delta_j>\left|p_{k_j'}\right|\right\},$$

et, puisque $\hat{f}(x) = \hat{f}(-x)$,

$$B_k = p_{k_j}\overline{p_{k_j'}}\left(\left(\beta_k\overline{\beta_{k'}}+\beta_{-k}\overline{\beta_{-k'}}\right)\hat{k}(\lambda_k-\lambda_{k'}) + \left(\beta_k\overline{\beta_{-k'}}+\beta_{-k}\overline{\beta_{k'}}\right)\hat{k}(\lambda_k+\lambda_{k'})\right).$$

Le point clé est que l'on peut se débarrasser du dernier terme de (2.9) où il n'y a pas de condition de gap grâce à la positivité de f. Cette astuce a déjà été utilisée dans [65] (voir aussi [63]).
Pour le premier terme de (2.9), en utilisant les hypothèses (2.4)-(2.5)-(2.6), le fait que $T > \frac{2\pi}{\gamma_j}$, la propriété (2.7), et puisque $\hat{f}(0) = \frac{2T}{\pi}$, on obtient, en écrivant $A_k = |\beta_k|^2 + |\beta_{-k}|^2$,

$$\sum_{(k_j,k_j')\in\Lambda_1} B_k \geq \frac{2T}{\pi}\left(\sum_{k_j\in\mathbb{N}^*,\ \delta_j\leq|p_{k_j}|}|p_{k_j}|^2 A_k - \sum_{(k_j,k_j')\in\Lambda_1,\ k_j\neq k_j'} C_k\right),$$

où

$$C_k = \frac{\left|p_{k_j}\right|^2 A_k + \left|p_{k'_j}\right|^2 A_{k'}}{2} \left(\frac{2\pi}{T\gamma_j}\right)^2 \left(\frac{1}{4(k_j-k'_j)^2-1} + \frac{1}{4(k_j+k'_j)^2-1}\right).$$

Pour les deuxièmes et troisièmes termes (2.9), on obtient de manière similaire

$$\sum_{(k_j,k'_j)\in\Lambda_2} B_k \geq -\frac{2T}{\pi} \sum_{(k_j,k'_j)\in\Lambda_2} C_k, \quad \sum_{(k_j,k'_j)\in\Lambda_3} B_k \geq -\frac{2T}{\pi} \sum_{(k_j,k'_j)\in\Lambda_3} C_k.$$

En rassemblant les estimations des quatre termes de (2.9) et en utilisant (2.8), on obtient donc

$$\begin{aligned}\frac{\pi}{2T}A \geq & \sum_{k_j\in\mathbb{N}^*,\ \delta_j\leq\left|p_{k_j}\right|} \left|p_{k_j}\right|^2 |\beta_k|^2 \\ & - \sum_{k_j,k'_j\in\mathbb{N}^*,\ k_j\neq k'_j} C_k \geq \sum_{k_j\in\mathbb{N}^*,\ \delta_j\leq\left|p_{k_j}\right|} \left|p_{k_j}\right|^2 |\beta_k|^2 - \sum_{k_j\in\mathbb{N}^*} \left|\beta_k p_{k_j}\right|^2 \left(\frac{2\pi}{T\gamma_j}\right)^2.\end{aligned}$$

On obtient alors, en sommant sur $j = 1, \ldots, d$ et $k \in (\mathbb{N}^*)^d$,

$$\begin{aligned}B := & \sum_{j=1}^{d} \sum_{k_1,\ldots,k_{j-1},k_{j+1},\ldots,k_d\in\mathbb{N}^*} \int_{-T/2}^{T/2} \left|\sum_{k_j\in\mathbb{N}^*} p_{k_j}\left(\beta_k e^{i\lambda_k t} + \beta_{-k} e^{-i\lambda_k t}\right)\right|^2 \mathrm{d}t \\ & \gtrsim \sum_{j=1}^{d} \sum_{k\in(\mathbb{N}^*)^d,\ \delta_j\leq\left|p_{k_j}\right|} A_k \left|p_{k_j}\right|^2 - \sum_{j=1}^{d} \left(\frac{2\pi}{T\gamma_j}\right)^2 \sum_{k\in(\mathbb{N}^*)^d} A_k \left|p_{k_j}\right|^2 \\ & = \sum_{k\in(\mathbb{N}^*)^d} A_k \sum_{j=1}^{d} \left(1_{\delta_j\leq\left|p_{k_j}\right|} - \left(\frac{2\pi}{T\gamma_j}\right)^2\right) \left|p_{k_j}\right|^2.\end{aligned}$$

Puisque l'on remarque que

$$\sum_{j=1}^{d} 1_{\delta_j\leq\left|p_{k_j}\right|} \left|p_{k_j}\right|^2 \geq \max_{j=1,\ldots,d} \left|p_{k_j}\right|^2 \geq \sum_{j=1}^{d} \left|\frac{p_{k_j}}{\gamma_j}\right|^2 / \sum_{j=1}^{d} \frac{1}{\gamma_j^2},$$

on obtient finalement

$$\begin{aligned}B & \gtrsim \left(1/\sum_{j=1}^{d} \frac{1}{\gamma_j^2} - \left(\frac{2\pi}{T}\right)^2\right) \sum_{k\in(\mathbb{N}^*)^d} A_k \left(\sum_{j=1}^{d} \left|\frac{p_{k_j}}{\gamma_j}\right|^2\right) \\ & \gtrsim \left(1/\sum_{j=1}^{d} \frac{1}{\gamma_j^2} - \left(\frac{2\pi}{T}\right)^2\right) \sum_{k\in(\mathbb{N}^*)^d} A_k \left(\sum_{j=1}^{d} \left|p_{k_j}\right|^2\right),\end{aligned}$$

ce qui termine la preuve du Théorème 2.2, à partir de l'hypothèse $T > 2\pi\sqrt{\sum_{i=1}^{d} \frac{1}{\gamma_i^2}}$ et pusique l'on peut changer l'intervalle $[-T/2, T/2]$ en $[0, T]$ par un argument classique de translation.

Application pour la preuve de la Proposition 2.1

Afin de prouver la Proposition 2.1 en utilisant le Théorème 2.2, on exprime d'abord l'inégalité (2.2) en termes de série de Fourier (2.3). On a d'abord

$$\int_0^T \int_\Gamma |\partial_\nu u(t,x)|^2 \,\mathrm{d}\Gamma\mathrm{d}t = \sum_{\mathrm{j}=1}^{\mathrm{N}} \int_0^{\mathrm{T}} \int_0^{\mathrm{L}_1} \dots \int_0^{\mathrm{L}_{\mathrm{j}-1}} \int_0^{\mathrm{L}_{\mathrm{j}+1}} \dots \int_0^{\mathrm{L}_{\mathrm{N}}}$$

$$\left| \sum_{k\in(\mathbb{N}^*)^N} \frac{\pi k_j}{L_j} \left(\alpha_k e^{i\omega_k t} + \alpha_{-k} e^{-i\omega_k t}\right) \prod_{\ell=1,\ell\neq j}^{N} \sin\left(\frac{\pi k_\ell x_\ell}{L_\ell}\right) \right|^2 \mathrm{d}x_1 \dots \mathrm{d}x_{j-1}\mathrm{d}x_{j+1}\mathrm{d}x_N\mathrm{d}t.$$

A cause de l'orthogonalité de la famille $\left(\sin\left(\frac{\pi k_\ell x_\ell}{L_\ell}\right)\right)_{k_\ell\geq 1}$ in $L^2(0,L_\ell)$, on obtient

$$\int_0^T \int_\Gamma |\partial_\nu u(t,x)|^2 \,\mathrm{d}\Gamma\mathrm{d}t$$

$$\asymp \sum_{j=1}^{N} \sum_{k_1,\dots,k_{j-1},k_{j+1},\dots,k_N\in\mathbb{N}^*} \int_0^T \left| \sum_{k_j\in\mathbb{N}^*} k_j \left(\alpha_k e^{i\omega_k t} + \alpha_{-k} e^{-i\omega_k t}\right) \right|^2 \mathrm{d}t. \quad (2.10)$$

D'autre part, on a

$$\|u_0\|^2_{H_0^1(\Omega)} + \|u_1\|^2_{L^2(\Omega)} \asymp \sum_{k\in(\mathbb{N}^*)^N} \left(\sum_{j=1}^{N} |k_j|^2\right) \left(|\alpha_k|^2 + |\alpha_{-k}|^2\right). \quad (2.11)$$

On applique ensuite le Théorème 2.2 : on prend $d = N$, $\lambda_k = \omega_k = \sqrt{\sum_{j=1}^N \left(\frac{k_j\pi}{L_j}\right)^2}$, pour $k \in (\mathbb{N}^*)^N$ et $p_\ell = \ell$, for $\ell \in \mathbb{N}^*$.
Afin de vérifier l'hypothèse d'écart (2.4)-(2.5)-(2.6), pour un $j \in \{1,\dots,N\}$ fixé, on considére donc $k = (k_1,\dots,k_N) \in (\mathbb{N}^*)^N$ et $k'_j \in \mathbb{N}^*$, s

$$0 < k_i \leq \max(k_j, k'_j),\ i = 1,\dots,N,\ i \neq j, \quad (2.12)$$

et on doit obtenir (2.4)-(2.5), pour un γ_j convenable. On a tout d'abord $|\omega_k + \omega_{k'}| \geq \left|k_j + k'_j\right| \frac{\pi}{L_j}$. On calcule ensuite

$$|\omega_k - \omega_{k'}| = \left|k_j - k'_j\right| \frac{\pi^2}{L_j^2} \frac{\left|k_j + k'_j\right|}{\sqrt{D + \left(\frac{k_j\pi}{L_j}\right)^2} + \sqrt{D + \left(\frac{k'_j\pi}{L_j}\right)^2}},$$

avec $D = \sum_{i=1,\ i\neq j}^N \left(\frac{k_i\pi}{L_i}\right)^2$. On peut supposer par exemple que $k_j < k'_j$ et on a donc, en utilisant (2.12)

$$\frac{|\omega_k - \omega_{k'}|}{\left|k_j - k'_j\right|} \geq \frac{\pi^2}{L_j^2} \frac{\left|1 + (k_j/k'_j)\right|}{\sqrt{\sum_{i=1,\ i\neq j}^d \left(\frac{\pi}{L_i}\right)^2 + (k_j/k'_j)^2 \frac{\pi^2}{L_j^2}} + \sqrt{\sum_{i=1}^d \left(\frac{\pi}{L_i}\right)^2}}$$

$$\geq \frac{\pi^2}{(L_j)^2} \frac{1}{2\sqrt{\sum_{i=1}^d \left(\frac{\pi}{L_i}\right)^2}}.$$

On obtient donc (2.4)-(2.5) en prenant $\gamma_j = \frac{\pi^2}{L_j^2}\frac{1}{2\sqrt{\sum_{i=1}^d \left(\frac{\pi}{L_i}\right)^2}}. < \frac{\pi}{L_j}$. A partir de (2.10) et (2.11), on obtient finalement (2.2) avec

$$T > 2\pi\sqrt{\sum_{j=1}^N \gamma_j^2} = 4\sqrt{\sum_{j=1}^N (L_j)^4}\sqrt{\sum_{j=1}^N L_j^{-2}},$$

ce qui termine la preuve de la Proposition 2.1. Notons que si on pouvait prendre $\gamma_j = \frac{\pi}{L_j}$, on obtiendrait le temps optimal $T_0 = 2\sqrt{\sum_{i=1}^N L_i^2}$.

3 Observabilité uniforme de l'équation des ondes

Dans ce travail [63, 64], on s'intéresse ici à l'observabilité uniforme de l'équation des ondes. Nous rassemblons d'abord, sous forme compacte et auto-contenue, plusieurs résultats obtenus à l'aide des multiplicateurs discrets. Puis, nous développons une approche de type Ingham en établissant de nouveaux théorèmes d'Ingham adaptés à cette situation. On s'intéresse enfin à l'optimalité du temps. A ce propos, nous établissons un exemple d'inégalité de type Ingham où la position de l'intervalle joue un rôle dans la détermination du temps optimal, ce qui n'est habituellement pas le cas.

3.1 Introduction

On considère l'équation des ondes $1D$

$$\begin{cases} u_{tt} - u_{xx} = 0, \quad 0 < x < 1, \quad 0 < t < T, \\ u(t,0) = 0, \quad u(t,1) = 0, \quad 0 < t < T \\ u(0,x) = u^0(x), \quad u_t(x,0) = u^1(x), \quad 0 < x < 1, \end{cases} \tag{3.1}$$

qui admet une unique solution $u \in C([0,T]; H_0^1(0,1)) \cap C^1([0,T]; L^2(0,1))$, pour $(u_0, u_1) \in H_0^1(0,1) \times L^2(0,1)$. L'énergie de la solution est donnée par

$$E(t) = \frac{1}{2}\int_0^1 |u_t(t,x)|^2 + |u_x(t,x)|^2 \, \mathrm{d}t,$$

et est conservée, c'est-à-dire que l'on a $E(t) = E(0), \quad 0 < t < T$. Il est bien connu que pour $T \geq 2$, on a l´*inégalité d'observabilité*

$$E(0) \leq C(T)\int_0^T |u_x(t,1)|^2 \, \mathrm{d}t, \tag{3.2}$$

pour chaque solution u de (3.1), avec une constante $C(T) > 0$ indépendante de la condition initiale (u_0, u_1). Cette inégalité signifie que l'énergie de la solution peut être estimée par l'énergie concentrée en l'extrémité $x = 1$ de la frontière et est également liée à la contrôlabilité frontière de l'équation des ondes (voir par exemple [60]).
Chaque solution u de (3.1) satisfait aussi une propriété supplémentaire de régularité

$$\int_0^T |u_x(t,1)|^2 \, \mathrm{d}t \leq C(T)E(0), \tag{3.3}$$

avec une autre constante $C(T) > 0$. Cette dernière inégalité est souvent appelée *inégalité directe*, alors que la première est aussi appelée *inégalité inverse* (cf [61]). L'inégalité directe est importante pour résoudre le problème frontière non homogène.

Le schéma aux différences finies On considère maintenant la semi-discrétisation classique de l'équation des ondes $1D$, avec $N \in \mathbb{N}^*$ impair et $h := 1/(N+1)$:

$$\begin{cases} u_j'' = \frac{1}{h^2}\left(u_{j+1} - 2u_j + u_{j-1}\right), & 0 < t < T, \quad j = 1, 2, \dots, N \\ u_0 = u_{N+1} = 0, & 0 < t < T \\ u_j(0) = u_j^0, \quad u_j'(0) = u_j^1, & j = 0, \dots, N+1. \end{cases} \tag{3.4}$$

Pour chaque condition initiale $(u_j^0, u_j^1)_{j=0}^{N+1}$ satisfaisant $u_0^0 = u_{N+1}^0 = u_0^1 = u_{N+1}^1 = 0$, le système (3.4) admet une unique solution, qui est explicitement donnée par

$$u_j(t) = \sum_{|k|=1}^{N} a_k e^{i\lambda_k^0 t} e_j^{|k|}, \quad e_j^{|k|} = \sin(j|k|\pi h), \quad \lambda_k^0 = \frac{2}{h}\sin(k\frac{\pi h}{2}), \tag{3.5}$$

où les coefficients $(a_k)_{|k|=1}^N$ sont uniquement déterminés par les relations

$$u_j^0 = \sum_{k=1}^{N}(a_k + a_{-k})e_j^k, \quad u_j^1 = \sum_{k=1}^{N} i\lambda_k^0(a_k - a_{-k})e_j^k, \quad j = 1, \dots, N. \tag{3.6}$$

L'énergie du système est donnée par :

$$E_h^0(t) = \frac{h}{2}\sum_{j=0}^{N}\left[|u_j'|^2 + \left|\frac{u_{j+1}(t) - u_j(t)}{h}\right|^2\right], \tag{3.7}$$

et est une discrétisation de l'énergie continue. Elle est aussi constante en temps : $E_h^0(t) = E_h^0(0), \quad 0 < t < T$. On cherche maintenant une version semi-discrète de (3.2), c'est-à-dire, une *inégalité d'observabilité uniforme* : a-t-on

$$C(T)E_h^0(0) \leq \int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t, \tag{3.8}$$

avec une constante $C(T) > 0$ indépendante des conditions initiales et de h ? Une telle inégalité est aussi liée à l'approximation numérique semi-discrète du contrôle frontière de l'équation des ondes en $1D$, qui a été intensément étudié ces derniers temps (cf [115]). La principale propriété et difficulté ici est que la constante $C(T)$ générique doit être indépendante du pas de discrétisation h. Comme dans le cas continu, on peut aussi chercher à obtenir une *inégalité directe* : a-t-on

$$\int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t \leq C(T)E_h^0(0) \tag{3.9}$$

avec une constante $C(T) > 0$ indépendante des conditions initiales et de h ?

Le θ-schéma Soit $0 \leq \theta \leq 1/4$. Le θ-schéma est une généralisation du schéma précédent et a été introduit dans [80]. Il est obtenu en remplaçant u_j'' avec $u_j'' + \theta(u_{j+1}'' - 2u_j'' + u_{j-1}'')$ dans (3.4) :

$$\begin{cases} u_j'' + \theta(u_{j+1}'' - 2u_j'' + u_{j-1}'') = \frac{1}{h^2}\left(u_{j+1} - 2u_j + u_{j-1}\right), & 0 < t < T, \quad j = 1, 2, \dots, N \\ u_0 = u_{N+1} = 0, & 0 < t < T \\ u_j(0) = u_j^0, \quad u_j'(0) = u_j^1, & j = 0, \dots, N+1. \end{cases} \tag{3.10}$$

Notons que le schéma aux différences finies correspond au cas $\theta = 0$. La valeur $\theta = 1/6$ correspond à une semi-discrétisation par éléments finis (voir par exemple [84]), et la valeur $\theta = 1/4$ peut aussi être dérivée à partir d'une méthode d'éléments finis, en discrétisant la position et la vitesse de manière différente, et est appelée méthode d'éléments finis mixtes (voir [19]).
La solution peut être exprimée en série de Fourier comme dans (3.5), en remplaçant λ_k^0 par λ_k^θ qui satisfait

$$-(\lambda_k^\theta)^2 + \theta h^2 (\lambda_k^\theta)^2 (\lambda_k^0)^2 = -(\lambda_k^0)^2. \tag{3.11}$$

L'énergie du système est donnée par

$$E_h^\theta(t) = \frac{h}{2} \sum_{j=1}^{N} \left|u_j'\right|^2 - \theta \left|u_{j+1}' - u_j'\right|^2 + \left|\frac{u_{j+1} - u_j}{h}\right|^2, \tag{3.12}$$

et satisfait $E_h^\theta(t) = E_h^\theta(0)$, pour $0 < t < T$.
L'*observabilité uniforme* est alors : a-t-on

$$C(T) E_h^\theta(0) \leq \int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t + \theta \int_0^T \left|u_N'(t)\right|^2 \mathrm{d}t, \tag{3.13}$$

avec une constante $C(T) > 0$ indépendante des conditions initiales et de h ?
L'*inégalité directe* est : a-t-on

$$\int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t + \theta \int_0^T \left|u_N'(t)\right|^2 \mathrm{d}t \leq C(T) E_h^\theta(0), \tag{3.14}$$

avec une constante $C(T) > 0$ indépendante des conditions initiales et de h ?
Il est maintenant bien connu que l'observabilité uniforme peut ne pas avoir lieu pour les semi-discrétisations classiques par l'effet de solutions numériques avec des modes de hautes fréquences parasites, et plusieurs méthodes ont été développées et analysées ces dernières années.

La méthode de filtrage Un remède est de filtrer les hautes fréquences, comme cela a été introduit dans [55]. Plus précisément, pour $0 < \alpha < 1$, on peut considérer le sous-espace des solutions de (3.4) ou (3.10) satisfaisant

$$a_k = 0, \quad |k| \geq \alpha N.$$

La méthode à deux grilles On peut aussi retrouver l'observabilité uniforme en modifiant les conditions initiales. Une méthode consiste à utiliser une grille fine et une grille grossière et de projeter les données initiales de la grille fine sur la grille grossière. Il s'agit de la méthode à deux grilles qui a été proposée par Glowinski [42] (dans le contexte des discrétisations complètes par éléments finis et différences finis en $2D$) et en premier analysée par Negreanu et Zuazua [85], avec une approche de type multiplicateurs discrets, comme nous allons le voir.
On suppose que $N \in \mathbb{N}^*$ est un nombre impair. Considérons donc des conditions initiales satisfaisant

$$u_{2k+1}^0 = \frac{u_{2k}^0 + u_{2k+2}^0}{2}, \quad u_{2k+1}^1 = \frac{u_{2k}^1 + u_{2k+2}^1}{2}, \quad k = 0, \dots \frac{N-1}{2}. \tag{3.15}$$

La méthode des éléments finis mixtes Il s'agit du θ-schéma, lorsque $\theta = 1/4$. Il a été analysé par [19].

Autres méthodes Il existe également d'autres moyens de retrouver l'observabilité uniforme : régularisation de Tychonoff [43], qui consiste à rajouter un terme agissant à l'intérieur du domaine et disparaissant lorsque la taille de maille tend vers zéro, l'ajout d'un terme de viscosité (voir [91] dans le cas de problèmes de stabilisation), l'utilisation de données initiales analytiques [76]. On réfère aussi le lecteur à [115], pour un aperçu des méthodes existantes.
L'observabilité uniforme, dans le contexte des semi-discrétisations de l'équation des ondes $1D$ a été étudiée dans plusieurs travaux récents : [55], [76], [19], [81], [85], [84]. Essentiellement deux sortes de méthodes ont été utilisées pour prouver l'observabilité uniforme.

La méthode des multiplicateurs Elle consiste à établir des identités pour les solutions considérées pour lesquelles l'observabilité uniforme est dérivée. Le moyen d'obtenir ces identités est d'utiliser l'équation, des intégrations par parties et des sommes téléscopiques dans le cas semi-discret (qui sont en fait un analogue de l'intégration par parties au niveau discret).

Approche de type Ingham Elle consiste à utiliser la solution sous forme de série de Fourier et d'utiliser ensuite un théorème d'Ingham [56] ou une variante.
Notons qu'il existe aussi d'autres méthodes pour prouver l'observabilité uniforme : voir par exemple [76], [54]. Toutes les questions mentionnées ici peuvent être posées pour d'autres équations et en dimension plus élevées, pour des discrétisations complètes, mais nous allons ici seulement traiter le cas de la semi-discrétisation de l'équation des ondes $1D$.
Pour la suite, nous allons rappeler la preuve de plusieurs résultats d'observabilité uniforme obtenus par la méthode des multiplicateurs. le but est de rassembler plusieurs résultats et preuves dans une forme compacte et auto contenue.
La méthode des multiplicateurs ne donne ici pas le temps optimal pour l'observabilité uniforme. Ainsi, par exemple, à l'aide des multiplicateurs discrets, on peut établir que (3.8) est vérifié pour des conditions initiales satisfaisant (3.15) pour tout temps $T > 4$. Ce temps peut être amélioré à $T > 2\sqrt{2}$ en utilisant une approche de type Ingham que l'on développe ensuite. La question de l'optimalité est enfin étudiée.

3.2 L'inégalité directe par la méthode des multiplicateurs

L'inégalité directe a été prouvée pour $\theta = 0$ dans [55] et [76], pour $\theta = 1/4$ dans [19], et pour $\theta = 1/6$ dans [84] (et est aussi contenue dans [55]).

Le schéma aux différences finies On utilise d'abord les conditions aux bords $u_0 = u_{N+1} = 0$ de (3.4). Les calculs algébriques suivants correspondent alors à une

intégration par parties discrète. On a tout d'abord

$$(N+1)u_N^2 = \sum_{j=0}^{N}(j+1)(u_{j+1}-u_j)^2 - \sum_{j=0}^{N-1}(j+1)(u_{j+1}-u_j)^2,$$

puis en découpant la première somme et avec un changement d'indice dans la deuxième somme, on a

$$(N+1)u_N^2 = \sum_{j=0}^{N}(u_{j+1}-u_j)^2 + \sum_{j=0}^{N} j(u_{j+1}-u_j)^2 - \sum_{j=1}^{N} j(u_j-u_{j-1})^2,$$

ce qui donne

$$(N+1)u_N^2 = \sum_{j=0}^{N}(u_{j+1}-u_j)^2 + \sum_{j=1}^{N} j(u_{j+1}-2u_j+u_{j-1})(u_{j+1}-u_{j-1})$$

On utilise alors la relation $u_j'' = \frac{1}{h^2}\left(u_{j+1}-2u_j+u_{j-1}\right)$ de (3.4) pour obtenir

$$(N+1)u_N^2 = \sum_{j=0}^{N}(u_{j+1}-u_j)^2 + \sum_{j=1}^{N} jh^2u_j''(u_{j+1}-u_{j-1}).$$

On fait ensuite une intégration par parties en temps :

$$\int_0^T ju_j''(u_{j+1}-u_{j-1})\mathrm{d}t = -\int_0^T ju_j'(u_{j+1}'-u_{j-1}')\mathrm{d}t + ju_j'(u_{j+1}-u_{j-1})\Big|_0^T,$$

De manière algébrique, comme précédemment, on obtient

$$\begin{aligned}-\sum_{j=0}^{N} u_j'j(u_{j+1}'-u_{j-1}') &= -\sum_{j=0}^{N} ju_j'u_{j+1}' + \sum_{j=0}^{N}(j+1)u_{j+1}'u_j' \\ &= \sum_{j=0}^{N} u_j'u_{j+1}' = \sum_{j=0}^{N}|u_j'|^2 - \frac{1}{2}\sum_{j=0}^{N}(u_j'-u_{j+1}')^2.\end{aligned}$$

En posant $X_h(t) = h\sum_{j=1}^{N} j\big(u_{j+1}-u_{j-1}\big)u_j'$, et en reprenant les 3 précédentes égalités, on a finalement l'identité

$$\int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t = 2TE_h(0) + X_h(t)\Big|_0^T - \frac{h}{2}\sum_{j=0}^{N}\int_0^T \left|u_{j+1}'-u_j'\right|^2 \mathrm{d}t$$

D'autre part, en utilisant l'inégalité triangulaire puis l'inégalité de Cauchy-Schwarz, on a l'estimation

$$\begin{aligned}|X_h(t)| &\le h\left(\sum_{j=0}^{N}\left|j(u_{j+1}-u_j)u_j'\right| + \left|j(u_j-u_{j-1})u_j'\right|\right) \\ &\le h\left(\sum_{j=0}^{N}\left|u_j'\right|^2 + \sum_{j=0}^{N}(1/h)^2\left|u_{j+1}-u_j\right|^2\right) = 2\left|E_h(0)\right|,\end{aligned}$$

de telle sorte que l'on obtient finalement (3.9) :

$$\int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t \le 2(T+2)E_h(0).$$

Le θ-schéma L'intégration par parties discrète en espace donne cette fois-ci avec (3.10)

$$(N+1)u_N^2 = \sum_{j=0}^{N}(u_{j+1}-u_j)^2 + \sum_{j=1}^{N} jh^2 u_j''(u_{j+1}-u_{j-1}) \\ + \theta\sum_{j=1}^{N} jh^2(u_{j+1}''-2u_j''+u_{j-1}'')(u_{j+1}-u_{j-1}),$$

tandis que l'intégration par parties en temps donne

$$\int_0^T (u_{j+1}''-2u_j''+u_{j-1}'')(u_{j+1}-u_{j-1})\mathrm{d}t \\ = -\int_0^T (u_{j+1}'-2u_j'+u_{j-1}')(u_{j+1}'-u_{j-1}') + (u_{j+1}'-2u_j'+u_{j-1}')(u_{j+1}-u_{j-1})|_0^T.$$

Pour étudier le premier terme, on obtient de manière algébrique, en utilisant les égalités $u_{j+1}'-2u_j'+u_{j-1}' = (u_{j+1}'-u_j')-(u_j'-u_{j-1}')$ et $u_{j+1}'-u_{j-1}' = (u_{j+1}'-u_j')+(u_j'-u_{j-1}')$ ainsi que $u_0 = u_{N+1} = u_0' = u_{N+1}' = 0$

$$\sum_{j=0}^{N} j(u_{j+1}'-2u_j'+u_{j-1}')(u_{j+1}'-u_{j-1}') \\ = \sum_{j=0}^{N} j(u_{j+1}'-u_j')(u_{j+1}'-u_{j-1}') - \sum_{j=0}^{N} j(u_j'-u_{j-1}')(u_{j+1}'-u_{j-1}') \\ = \sum_{j=0}^{N} j(u_{j+1}'-u_j')(u_{j+1}'-u_j') + \sum_{j=0}^{N} j(u_{j+1}'-u_j')(u_j'-u_{j-1}') \\ - \sum_{j=0}^{N} j(u_j'-u_{j-1}')(u_{j+1}'-u_j') - \sum_{j=0}^{N} j(u_j'-u_{j-1}')(u_j'-u_{j-1}') \\ = \sum_{j=0}^{N} j(u_{j+1}'-u_j')^2 - \sum_{j=0}^{N} j(u_j'-u_{j-1}')^2 \\ = \sum_{j=0}^{N} j(u_{j+1}'-u_j')^2 - \sum_{j=0}^{N-1}(j+1)(u_{j+1}'-u_j')^2 = (N+1)|u_N'|^2 - \sum_{j=0}^{N}(u_{j+1}'-u_j')^2.$$

0n obtient alors, en posant $Y_h(t) = h\sum_{j=1}^{N} j\big(u_{j+1}-u_{j-1}\big)(u_j'+\theta(u_{j+1}'-2u_j'+u_{j-1}'))$

$$\int_0^T \left|\frac{u_N}{h}\right|^2 + \theta\left|u_N'\right|^2 \mathrm{d}t - Y_h(t)|_0^T \\ = \int_0^T \sum_{j=0}^{N} h(\theta-1/2)(u_{j+1}'-u_j')^2 + h(\frac{u_{j+1}-u_j}{h})^2 + h\left|u_j'\right|^2 \mathrm{d}t \\ = 2TE_h^\theta(0) + (2\theta-1/2)h\int_0^T \sum_{j=0}^{N}(u_{j+1}'-u_j')^2\mathrm{d}t \quad (3.16)$$

D'autre part, on a

$$Y_h(t) = (1-4\theta)W_h(t) + \theta Z_h(t), \quad W_h(t) = h\sum_{j=1}^{N} j(u_{j+1}-u_j)u'_j,$$
$$Z_h(t) = h\sum_{j=1}^{N} j(u_{j+1}-u_j)(u'_{j+1}+2u'_j+u'_{j-1}),$$

et

$$|W_h(t)| \leq h\sum_{j=1}^{N} \left|u'_j\right|^2 + \left(\frac{u_{j+1}-u_j}{h}\right)^2,$$

$$\begin{aligned}|Z_h(t)| &\leq h\sum_{j=1}^{N} j\left(|u_{j+1}-u_j| + |u_j-u_{j-1}|\right)\left(\left|u'_{j+1}+u'_j\right| + \left|u'_j+u'_{j-1}\right|\right)\\ &\leq h\sum_{j=1}^{N} 2j^2\left|u_{j+1}-u_j\right|^2 + 2j^2\left|u_j-u_{j-1}\right|^2 + \frac{1}{2}\left|u'_{j+1}+u'_j\right|^2 + \frac{1}{2}\left|u'_j+u'_{j-1}\right|^2\\ &\leq h\sum_{j=1}^{N} 4\left(\frac{u_{j+1}-u_j}{h}\right)^2 + \left|u'_{j+1}+u'_j\right|^2,\end{aligned}$$

de telle sorte que

$$|Y_h(t)| \leq h\sum_{j=1}^{N}\left(\frac{u_{j+1}-u_j}{h}\right)^2 + h\sum_{j=1}^{N}(1-4\theta)\left|u'_j\right|^2 + \theta\left|u'_{j+1}+u'_j\right|^2 = 2E_h^\theta(0). \quad (3.17)$$

Pour chaque $0 \leq \theta \leq 1/4$ et $T > 0$, on obtient finalement (3.14) :

$$\int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t + \theta\int_0^T |u'_N(t)|^2\,\mathrm{d}t \leq 2(T+2)E_h^\theta.$$

3.3 L'observabilité uniforme par la méthode des multiplicateurs

L'observabilité uniforme a été obtenue pour la méthode des éléments finis mixtes [19], la méthode de filtrage [55] (pour $\theta = 0$ et $\theta = 1/6$) et la méthode à deux grilles [85] (pour $\theta = 0$) , [84] (pour $\theta = 0$ et $\theta = 1/6$).

La méthode des éléments finis mixtes Pour $\theta = 1/4$, qui correspond à la méthode des éléments finis mixtes, à partir de (3.16), on obtient

$$\int_0^T \left|\frac{u_N}{h}\right|^2 + 1/4\left|u'_N\right|^2\mathrm{d}t = 2TE_h^{1/4}(0) + Y_h(t)|_0^T \geq 2(T-2)E_h^{1/4}(0),$$

ce qui signifie que l'on a observabilité uniforme pour $T > 2$. Notons que l'on ne peut pas se débarasser du terme $\int_0^T \left|u'_N(t)\right|^2\mathrm{d}t$ (on peut le voir en prenant $a_N = 1$ et $a_k = 0$ sinon).

La méthode de filtrage On a d'après (3.5) et (3.11)

$$h^2\left|\lambda_k^\theta\right|^2=\frac{4\sin^2(k\pi h/2)}{1-4\theta\sin^2(k\pi h/2)}\leq\frac{4\sin^2(\alpha\pi h/2)}{1-4\theta\sin^2(\alpha\pi h/2)},\quad\text{pour}\quad|k|\leq\alpha N,$$

et donc, pour $0\leq\theta\leq 1/4$, on a, avec la solution explicite donnée par (3.5)

$$\begin{aligned}0&\leq(1/2-2\theta)h\int_0^T\sum_{j=0}^N(u'_{j+1}-u'_j)^2\mathrm{d}t\\&=(1/2-2\theta)\sum_{|k|\leq\alpha N}|a_k|^2h^2\left|\lambda_k^\theta\right|^2\left|\lambda_k^0\right|^2\leq\frac{2(1-4\theta)\sin^2(\alpha\pi h/2)}{1-4\theta\sin^2(\alpha\pi h/2)}E_h^\theta(0),\end{aligned}$$

de telle sorte que, de (3.16), on obtient

$$\int_0^T\left|\frac{u_N}{h}\right|^2+\theta\left|u'_N\right|^2\mathrm{d}t\geq(2T-4)E_h^\theta-T\frac{2(1-4\theta)\sin^2(\alpha\pi h/2)}{1-4\theta\sin^2(\alpha\pi h/2)}E_h^\theta,$$

ce qui signifie que l'observabilité uniforme est vraie pour

$$\begin{aligned}T&>2/(1-\frac{(1-4\theta)\sin^2(\alpha\pi h/2)}{1-4\theta\sin^2(\alpha\pi h/2)})\\&=2(1-4\theta\sin^2(\alpha\pi h/2))/(\cos^2(\alpha\pi h/2))=2(1+(1-4\theta)\tan^2(\alpha\pi h/2)).\end{aligned}$$

Notons que l'on obtient le même temps d'observation que dans [55] pour $\theta=0$ et $\theta=1/6$ (pour le dernier cas, le temps trouvé était $T>2/(1-\gamma/12)$, avec $|\lambda_k^\theta h|\leq\sqrt{\gamma}$, et puisque $\gamma=\frac{4\sin^2(\alpha\pi h/2)}{1-4\theta\sin^2(\alpha\pi h/2)}$, on peut vérifier qu'il s'agit du même temps).

La méthode à deux grilles En traduisant les relations (3.15) en série de Fourier, grâce à l'expression (3.6), on obtient pour $j=0,\dots,\frac{N-1}{2}$

$$\sum_{k=1}^N(a_k+a_{-k})\left(e_{2j+1}^k-\frac{e_{2j}^k+e_{2j+2}^k}{2}\right)=0,$$

ce qui donne, en rappelant que $e_j^k=\sin(jk\pi h)$

$$\begin{aligned}0&=\sum_{k=1}^N\left(\lambda_k^0\right)^2(a_k+a_{-k})\,e_{2j+1}^k\\&=\sum_{k=1}^{(N-1)/2}\left(\lambda_k^0\right)^2(a_k+a_{-k})\,e_{2j+1}^k+\sum_{k=1}^{(N-1)/2}\left(\lambda_{N+1-k}^0\right)^2(a_{N+1-k}+a_{k-N-1})\,e_{2j+1}^{N+1-k}\end{aligned}$$

On obtient alors par indépendance des $(e_{2j+1}^k)_{k=1}^{(N-1)/2}$ et en procédant de même pour la deuxième relation de (3.6) pour $k=0,\dots,\frac{N-1}{2}$:

$$\left\{\begin{array}{l}(\lambda_k^0)^2(a_k+a_{-k})=-(\lambda_{N+1-k}^0)^2(a_{N+1-k}+a_{-N-1+k}),\\(\lambda_k^0)^2\lambda_k^\theta(a_k-a_{-k})=-(\lambda_{N+1-k}^0)^2\lambda_{N+1-k}^\theta(a_{N+1-k}-a_{-N-1+k}).\end{array}\right.\tag{3.18}$$

En prenant le carré des relations et en sommant, puisque $\lambda_k^\theta \leq \lambda_{N+1-k}^\theta$ pour $k = 1, \dots, (N-1)/2$, on obtient

$$|a_{N+1-k}|^2 + |a_{-N-1+k}|^2 \leq (\nu_k)^4 \left(|a_k||^2 + |a_{-k}|^2 \right), \quad \nu_k = \frac{\lambda_k^0}{\lambda_{N+1-k}^0}, \quad k = 1, \dots, \frac{N-1}{2}. \tag{3.19}$$

On a déjà les estimations (3.16) et (3.17) dérivées de l'inégalité directe. On pose maintenant $C_\theta := \frac{1-4\theta}{1-2\theta}$, qui satisfait pour $k = 1, \dots, (N-1)/2$

$$(1/2 - 2\theta) \left| \lambda_{N+1-k}^\theta \right|^2 h^2 \geq C_\theta.$$

Donc, grâce au choix des conditions initiales, on obtient

$$\begin{aligned} &(2\theta - 1/2)\, h \sum_{j=0}^{N} \left| u'_{j+1} - u'_j \right|^2 - C_\theta E_h^\theta \\ &= \sum_{|k|=1}^{N} \left((2\theta - 1/2) \frac{h^2}{2} \left| \lambda_k^0 \right|^2 \left| \lambda_k^\theta \right|^2 - C_\theta / 2 \left| \lambda_k^0 \right|^2 \right) |a_k|^2 \leq \sum_{k=1}^{(N-1)/2} d_{|k|} \left| \lambda_k^0 \right|^2 / 2 \, |a_k|^2, \end{aligned}$$

où, pour $k = 1, \dots, (N-1)/2$ et $0 \leq \theta \leq 1/4$ on a

$$\begin{aligned} d_k &:= \left((1/2 - 2\theta) \left| \lambda_k^\theta \right|^2 h^2 - C_\theta \right) + \left| (1/2 - 2\theta) \left| \lambda_{N+1-k}^\theta \right|^2 h^2 - C_\theta \right| \cdot |\nu_k|^2 \\ &= C_\theta \left((1/2 - \theta) \left(|\lambda_k^\theta|^2 h^2 + |\lambda_{N+1-k}^\theta|^2 h^2 |\nu_k|^2 \right) - 1 - |\nu_k|^2 \right) \\ &= C_\theta \left(|\lambda_{N+1-k}^0|^2 h^2 \left(1 - \theta h^2 |\lambda_{N+1-k}^0|^2 \right) \left(1 - \theta h^2 |\lambda_k^0|^2 \right) \right)^{-1} e_k, \end{aligned}$$

avec

$$\begin{aligned} e_k &:= \left((1/2 - \theta) \left(|\lambda_k^0| |\lambda_{N+1-k}^0|^2 h^4 (2 - 4\theta) + |\nu_k|^2 \right) - 4 \left(1 - \theta h^2 |\lambda_{N+1-k}^0|^2 \right) \left(1 - \theta h^2 |\lambda_k^0|^2 \right) \right) \\ &= -4 + 16\theta + |\lambda_k^0|^2 |\lambda_{N+1-k}^0|^2 h^4 \left((-\theta + 1/2)(2 - 4\theta) - 4\theta^2 \right) \\ &= (-1 + 4\theta) \left(4 - |\lambda_k^0|^2 |\lambda_{N+1-k}^0|^2 h^4 \right) \leq 0. \end{aligned}$$

On obtient finalement

$$\int_0^T |\frac{u_N}{h}|^2 + \theta |u'_N|^2 \mathrm{d}t \geq \left((2 - C_\theta)\, T - 4 \right) E_h^\theta(0) = \left(T/(1 - 2\theta) - 4 \right) E_h^\theta(0),$$

ce qui signifie que l'observabilité uniforme a lieu pour $T > 4(1 - 2\theta)$. Remarquons que l'on obtient le temps $T > 4$, pour $\theta = 0$, comme dans [85] et [84]. Pour $\theta = 1/6$, on obtient l'estimation $T > 2 + 2/3$, qui est meilleure que $T > 4$ obtenue dans [84]. On peut aussi remarquer que l'on obtient le même temps que le temps obtenu par la méthode de filtrage de paramètre $\alpha = 1/2$.

3.4 Approche de type Ingham

Une autre façon classique d'étudier l'observabilité est d'utiliser une approche de type Ingham. Il s'agit d'utiliser la solution sous la forme de série de Fourier, et

d'utiliser ensuite un théorème de type Ingham ou une de ses variantes. On se place ici dans le cas des différences finies ($\theta = 0$) pour simplifier (voir [63] pour plus de détails).

Plus précisément, en introduisant le développement en série de Fourier des solutions et en calculant la dérivée normale pour une méthode à deux grilles, l'inégalité (3.8) prend la forme

$$C(T)\sum_{k=1}^{N}|\lambda_k^0|^2|a_k|^2 \leq \int_0^T \Big|\sum_{|k|=1}^{\frac{N-1}{2}} \frac{e_N^k}{h}(a_k e^{i\lambda_k^0 t} + b_k e^{i\mu_k^0 t})\Big|^2 \mathrm{d}t,$$

avec

$$\mu_{|k|}^0 = \lambda_{N+1-|k|}^0, \quad \mu_{-|k|}^0 = \lambda_{-N-1+|k|}^0,$$

et des suites $(a_k), (b_k)$ satisfaisant (3.18) (avec $\theta = 0$) où $b_{|k|} := a_{N+1-|k|}$ et $b_{-|k|} := a_{-N-1+|k|}$. De (3.19) (avec $\theta = 0$), on a en particulier

$$|b_k|^2 + |b_{-k}|^2 \leq (\nu_k)^4\big(|a_k|^2 + |a_{-k}|^2\big), \quad \nu_k = \tan(k\pi h/2), \quad k = 1, \ldots, (N-1)/2.$$

Ainsi, rappelons le théorème original d'Ingham [56], déjà signalé dans l'introduction :

Théorème 3.1. *Soit $\gamma > 0$ et soit (ν_k) une suite strictement croissante satisifaisant la* condition d'écart

$$\nu_{k+1} - \nu_k > \gamma, \quad \textit{pour} \quad k \in \mathbb{N}$$

Alors pour $T > \frac{2\pi}{\gamma}$, on a

$$c\sum_k |a_k|^2 \leq \int_0^T \Big|\sum_k a_k e^{i\nu_k t}\Big|^2 \mathrm{d}t \leq C\sum_k |a_k|^2,$$

avec des constantes $c, C > 0$ independantes de la suite (a_k).

Remarque 3.2. *L'inégalité est aussi valable sous la condition plus faible $\nu_{k+1} - \nu_k > \gamma, \quad for \quad k \geq k_0$, pour un certain entier k_0 (cf. [49]). Les constantes peuvent alors dépendre des premières fréquences correspondant à $k = 1, ..., k_0$ pour lesquelles la condtion décart n'est pas garantie (voir [78] par exemple).*

Difficultés pour une approche de type Ingham Chercher une preuve de type Ingham pour l'observabilité uniforme semble assez naturel dans le contexte de l'équation des ondes en dimension 1, où la solution est explicitement sous la forme de série de Fourier. Cette méthode a été appliquée avec succès pour certaines semi-discrétisations, comme la méthode de filtrage. [55] ou la méthode d'éléments finis mixtes [19]. Dans le cas de la méthode à deux grilles, la situation est plus subtile, on doit faire face à une infinité de valeurs propres qui peuvent être arbitrairement proches les unes des autres. En particulier, on ne peut pas appliquer le Théorème 3.1. La littérature dans ces cas est assez rare (voir [65], où une situation de ce type est considérée, qui est cependant différente de notre problème).

En regardant la Figure 1.1, on voit qu'il y a une compensation entre les écarts des suites (λ_k^0), (μ_k^0) et le coefficient ν_k. En effet, dans les régions où l'écart de (μ_k) est petit, le coefficient ν_k est aussi petit, et l'écart de (λ_k) est grand, de telle sorte que le terme $a_k e^{i\lambda_k^0 t}$ va dominer sur le terme $b_k e^{i\mu_k^0 t}$. D'autre part, quand le coefficient ν_k devient plus grand, l'écart (μ_k) devient aussi plus grand.

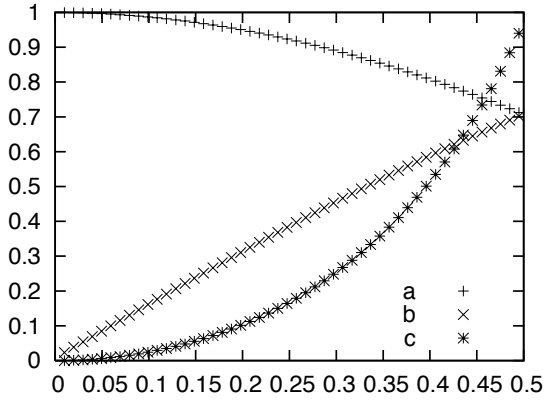

FIGURE 1.1 – Ecarts relatifs $(\lambda^0_{k+1}-\lambda^0_k)/\pi$, $(\mu^0_k-\mu^0_{k+1})/\pi$ et coefficient $\tan^2(k\pi h/2)$ vs k/N, pour $N=101$ et $k=1,\dots,(N-1)/2$; resp. a,b et c sur la légende.

Nouveaux théorèmes de type Ingham Afin de faire face à la situation précédente, on développe de nouveaux théorèmes de type Ingham, qui prennent en compte la sitation décrite juste avant.

On a d'abord le résultat suivant.

Théorème 3.3. *Soit $N\in\mathbb{N}^*$, $\gamma>0$, $\alpha>1/2$ et $M>0$. Soit $(\lambda_k)_{|k|=1}^N$ et $(\mu_k)_{|k|=1}^N$ des suites finies telles que*

$$\begin{gathered}\lambda_{k+1}-\lambda_k>\gamma, k=1,\dots,N-1,-2,\dots,-N,\\ \lambda_1-\lambda_{-1}>\gamma,\quad \mu_N-\lambda_N>\gamma,\quad \lambda_{-N}-\mu_{-N}>\gamma,\\ \mu_k-\mu_{k+1}>\gamma, |k|\geq N-N^\alpha,\\ \mu_k\geq\mu_{N-N^\alpha}, 1\leq k\leq N-N^\alpha,\quad \mu_k\leq\mu_{-N+N^\alpha}, -1\geq k\geq -N+N^\alpha.\end{gathered}$$

Alors, pour tout $T>\frac{2\pi}{\gamma}\sqrt{\max(1,1/2+M)}$, il existe une constante $C(T)>0$ telle que

$$C(T)\sum_{|k|=1}^N|a_k|^2\leq\int_0^T\Big|\sum_{|k|=1}^N a_ke^{i\lambda_kt}+b_ke^{i\mu_kt}\Big|^2\mathrm{d}t,$$

pour toutes les suites de coefficients $(a_k)_{|k|=1}^N$ et $(b_k)_{|k|=1}^N$ satisfaisant

$$|b_k|^2+|b_{-k}|^2\leq M^2(|a_k|^2+|a_{-k}|^2),\quad k=1,\dots,N. \tag{3.20}$$

La nouveauté dans ce Théorème est qu'il n'y a pas de condition d'écart pour les hautes fréquences, qui sont représentées par la suite $(\mu_k)_{|k|=1}^{N-N^\alpha}$. On peut remarquer que pour $M>1/2$, la borne inférieure $2\pi/\gamma(M+1/2)$ du temps T est toujours plus grande que la borne $2\pi/\gamma$ correspondant à la première suite $(\lambda_k)_{|k|=1}^N$.

En particulier, dans l'application à la méthode à deux grilles, la borne attendue $2\sqrt{2}$ de la suite $(\lambda^0_k)_{|k|=1}^{(N-1)/2}$ ne peut pas être atteinte par ce théorème. Afin de surmonter cette difficulté, nous avons développé une autre généralisation du théorème d'Ingham, qui est le résultat principal de cette section et qui va donner l'estimation précise $T>2\sqrt{2}$ pour la méthode à deux grilles.

Théorème 3.4. *Soit $N \in \mathbb{N}^*$ un nombre impair entier, $h := \frac{1}{N+1}$ et $f \in C^3([-1,1])$ une fonction impaire. Supposons que*

- *$f'(x) > 0$ pour $0 \le x < 1$,*
- *$f'(1) = 0$ et $f''(1) \neq 0$.*

On définit $\gamma > 0$ par

$$\gamma^2 = \min_{x\in[0,1/2]} \min\Big(\frac{f'(x)^2 + f'(1-x)^2}{2}, f'(x)^2\Big),$$

et on pose $\lambda_k := \frac{f(kh)}{h}$, $\gamma_k := |f'(kh)|$ for $|k| = 1, \dots, N$.
Alors, pour tout $T > 2\pi/\gamma$, il existe une constante $C(T) > 0$ independante de h, telle que l'on ait

$$\int_0^T \Big|\sum_{|k|=1}^{N} a_k e^{i\lambda_k t}\Big|^2 \ge C(T) \sum_{|k|=1}^{N} |a_k|^2,$$

pour chaque suite (a_k) satisfaisant

$$|a_{N+1-k}|^2 + |a_{-N-1+k}|^2 \le \Big(\frac{\gamma_{N+1-k}}{\gamma_k}\Big)^4 \big(|a_k|^2 + |a_{-k}|^2\big), \quad k = 1, \dots, (N-1)/2.$$

La principale nouveauté ici est que l'on peut mélanger l'écart $\gamma_{|k|}$ des basses fréquences λ_k avec l'écart $\gamma_{N+1-|k|}$ des hautes fréquences μ_k, comme si l'on avait un écart moyen $\sqrt{\frac{\gamma_{|k|}^2 + \gamma_{N+1-|k|}^2}{2}}$.

Application à la méthode à deux grilles Le résultat principal concernant l'application à la méthode à deux grilles peut être formulé ainsi :

Théorème 3.5. *Soit I un intervalle de longueur $|I| > 2\sqrt{2}$. Soit N un nombre impair et $h := \frac{1}{N+1}$.*
Alors il existe une constante $C_1(I)$ indépendante de h, telle que

$$E_h^0(0) \le C_1(I) \int_I \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t, \tag{3.21}$$

pour toutes les solutions de (3.4), écrites sous la forme (3.5), avec

$$|a_{N+1-k}|^2 + |a_{-N-1+k}|^2 \le \nu_k^4 \big(|a_k|^2 + |a_{-k}|^2\big),$$

où $\nu_k = \tan(k\pi h/2)$.

En particulier, la solution de la méthode à deux grilles satisfait les hypothèses (cf. (3.19)), de telle sorte que l'on obtient l'observabilité uniforme pour $|I| > 2\sqrt{2}$. On prouve ici le Théorème 3.5 en appliquant le Théorème 3.4.

Preuve du Théorème 3.5. En appliquant le Théorème 3.4 avec $f(x) = 2\sin(\pi x/2)$, on a $\frac{\gamma_{N+1-k}}{\gamma_k} = |\nu_k|$, et on obtient

$$\int_0^T \Big|\sum_{|k|=1}^{\frac{N-1}{2}} \frac{e_N^k}{h}(a_k e^{i\lambda_k^0 t} + b_k e^{i\mu_k^0 t})\Big|^2 \mathrm{d}t \ge C(T) \sum_{|k|=1}^{\frac{N-1}{2}} |\frac{e_N^k}{h}|^2 (|a_k|^2 + |b_k|^2),$$

pour toutes les suites $(a_k), (b_k)$ satisfaisant $|b_k|^2 + |b_{-k}|^2 \leq \nu_k^4\big(|a_k|^2 + |a_{-k}|^2\big)$ et pour $T > 2\sqrt{2}$. A partir de la dernière relation, et puisque $|e_N^k/h| = |\lambda_k^0 \cos(k\pi h/2)| \geq 1/2|\lambda_k^0|$, pour $k = 1, \ldots, (N-1)/2$, on obtient que

$$\int_0^T \Big|\sum_{|k|=1}^{\frac{N-1}{2}} \frac{e_N^k}{h}(a_k e^{i\lambda_k^0 t} + b_k e^{i\mu_k^0 t})\Big|^2 \mathrm{d}t \geq C(T) \sum_{|k|=1}^{\frac{N-1}{2}} |\lambda_k^0|^2 |a_k|^2 \geq C(T) \sum_{|k|=1}^{N} |\lambda_k^0|^2 |a_k|^2.$$

□

3.5 Inégalité directe par une approche de type Ingham

L'inégalité directe

On a déjà mentionné que l'inégalité directe est toujours valable en utilisant des multiplicateurs discrets. On peut se demander si on peut aussi obtenir ce résultat en utilisant les séries de Fourier. On a la proposition suivante :

Proposition 3.6. *Soit $N \in \mathbb{N}^*$, et une suite finie $(\lambda_k)_{k=1}^N$. Soit $(M_k)_{k=1}^N$ une suite finie positive telle qu'il existe deux constantes $M, \gamma > 0$ verifiant*

$$M_k \sum_{j, |\lambda_k - \lambda_j| < \gamma} M_j \leq M. \tag{3.22}$$

Alors pour tout $T > 0$ il existe une constante $C := C(T, \gamma, M) > 0$ telle que

$$\int_0^T \Big|\sum_{k=1}^{N} M_k a_k e^{i\lambda_k t}\Big|^2 \mathrm{d}t \leq C \sum_{k=1}^{N} |a_k|^2$$

pour toutes les suites de coefficients $(a_k)_{|k|=1}^N$.

Comme application, on peut obtenir une nouvelle preuve de (3.9). Une telle preuve a son intérêt propre, puisqu'elle pourrait certainement être appliquée à d'autres situations.

Proposition 3.7. *Soit $N \in \mathbb{N}^*$ et $h := \frac{1}{N+1}$. Alors, pour tout $T > 0$, il existe une constante $C_2(T)$ indépendante de h, telle que*

$$\int_0^T \left|\frac{u_N(t)}{h}\right|^2 \mathrm{d}t \leq C_2(T) E_h^0(0), \tag{3.23}$$

pour toutes les solutions de (3.4).

Preuve de la Proposition 3.6. On va utiliser cette fois-ci la seconde méthode d'Ingham (voir par exemple [61]). On utilise ici que $a \preceq b$, pour $a \leq cb$, avec un nombre c dépendant seulement de γ, T et M. On définit

$$H(x) = \begin{cases} \cos(\frac{\pi}{\gamma}x), & \text{if } |x| \leq \frac{\gamma}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Sa transformée de Fourier est donnée par

$$h(t) = \int_{-\infty}^{\infty} H(x)e^{-itx}\mathrm{d}x = \frac{2\gamma\pi\cos(\gamma t/2)}{\pi^2 - t^2\gamma^2}.$$

Soit g la transformée de Fourier du produit de convolution $G := H * H$. Il existe un intervalle $I_\gamma =]-r_\gamma, r_\gamma[$ tel que $1_{I_\gamma} \preceq g$, puisque $g(0) = 4\gamma^2/\pi^2 > 0$, $g = h^2$ est continue, positive et dépend seulement de γ. D'autre part, on a $|G| \preceq 1_{]-\gamma,\gamma[}$, puisque H est continue et s'annule en dehors de $]-\gamma/2, \gamma/2[$. On obtient alors

$$\begin{aligned}\int_{I_\gamma} |\sum_{k=1}^{N} a_k e^{i\lambda_k t}|^2\mathrm{dt} &\preceq \sum_{|\lambda_k-\lambda_j|\leq\gamma} M_k M_j |a_k||a_j| \preceq \sum_{|\lambda_k-\lambda_j|\leq\gamma} M_k M_j (|a_k|^2 + |a_j|^2) \\ &\preceq \sum_{|\lambda_k-\lambda_j|\leq\gamma} M_k M_j |a_k|^2 = \sum_{k=1}^{N} |a_k|^2 M_k \sum_{j,|\lambda_k-\lambda_j|\leq\gamma} M_j \preceq \sum_{k=1}^{N} |a_k|^2,\end{aligned}$$

ce qui donne le résultat puisque l'on peut remplacer I_γ par $[0, T]$ par un argument classique de translation. □

Preuve de la Proposition 3.7. La solution est sous la forme (3.5) et donc, il suffit de vérifier la condition (3.22) pour le second terme, avec $M_k = |kh| \succeq |\sin(k\pi h)|$. On fixe $\gamma = \pi/\sqrt{2}$, tel que l'on ait

$$|\mu_k - \mu_j| = |\frac{4}{h}\sin((k-j)\pi\frac{h}{4})\sin((k+j)\pi\frac{h}{4})| \succeq |k^2 - j^2|h,$$

pour $|k|, |j| = 1, \ldots, \frac{N-1}{2}$. Donc, la condition (3.22) peut être écrite comme

$$|k|h^2 \sum_{j,|k^2-j^2|<\delta/h} |j| \preceq 1,$$

avec une constante $\delta > 0$. Si $k^2 - \delta/h \geq 0$, on a

$$\begin{aligned}|k|h^2 \sum_{j,|k^2-j^2|<\delta/h} |j| &\preceq |k|h^2\sqrt{k^2+\delta/h}(\sqrt{k^2+\delta/h} - \sqrt{k^2-\delta/h}) \\ &\preceq |k|h^2\delta/h\frac{\sqrt{k^2+\delta/h}}{\sqrt{k^2+\delta/h} + \sqrt{k^2-\delta/h}} \preceq kh \preceq 1.\end{aligned}$$

D'autre part, si $k^2 - \delta/h < 0$, on obtient

$$|k|h^2 \sum_{j,|k^2-j^2|<\delta/h} |j| \preceq |k|h^2 \sum_{j,j^2\preceq\delta/h} |j| \preceq |k|h^2 h^{-1} \preceq 1,$$

ce qui donne le résultat. □

3.6 A propos de l'optimalité

On a vu dans le Théorème 3.5 que l'observabilité uniforme est valable pour $|I| > 2\sqrt{2}$ (qui est meilleur que le temps 4 obtenu par la méthode des multiplicateurs),

dans la classe des données à deux grilles. Notons que L. Ignat, dans sa thèse [54], montre que l'observabilité est valable pour $T > 2\sqrt{2}$ en utilisant d'autres méthodes. On peut donc se demander si ce temps est optimal. Nous allons voir que ce temps est optimal si on observe par exemple sur l'intervalle $[-T/2, T/2]$ au lieu de $[0, T]$. Nous allons aussi voir que dans certaines situations (pour une inégalité simplifiée), le temps optimal peut dépendre de la position de l'intervalle.
On cherche donc maintenant des exemples où la position de l'intervalle joue un rôle dans le temps optimal d'observation.
Soit $\alpha > 0$. Alors on a $T \geq \frac{\pi}{\alpha}$ si et seulement si

$$C(T)\sum_{k\geq 1}|a_k|^2 \leq \int_0^T \left|\sum_{k\geq 1} a_k(e^{ik\alpha t}+e^{-ik\alpha t})\right|^2 \mathrm{d}t$$

est vérifié pour tout (a_k).
D'autre part, on a $T \geq \frac{2\pi}{\alpha}$ si et seulement si

$$C(T)\sum_{k\geq 1}|a_k|^2 \leq \int_{-T/2}^{T/2} \left|\sum_{k\geq 1} a_k(e^{ik\alpha t}+e^{-ik\alpha t})\right|^2 \mathrm{d}t$$

est vérifié pour tout (a_k).
On voit donc sur cet exemple que le temps optimal pour avoir l'inégalité

$$C(I)\sum_{k\geq 1}|a_k|^2 \leq \int_I |\sum_{k\geq 1} a_k(e^{ik\alpha t}+e^{-ik\alpha t})|^2 \mathrm{d}t \tag{3.24}$$

dépend non seulement de la longueur de l'intervalle I mais aussi de sa position. Cet exemple a été mentionné dans [63].

On s'intéresse ici à savoir quand l'inégalité suivante est vérifiée :

$$C(I)\sum_{k\geq 1}|a_k|^2 \leq \int_I \left|\sum_{k\geq 1} a_k e^{ikat}\right|^2 \mathrm{d}t + \int_I \left|\sum_{k\geq 1} a_k e^{-ikbt}\right|^2 \mathrm{d}t, \tag{3.25}$$

où a, b sont des nombres strictement positifs.
Notons que si on a

$$C(I)\sum_{k\geq 1}|a_k|^2 \leq \int_I \left|\sum_{k\geq 1} a_k(e^{ikat}+e^{-ikbt})\right|^2 \mathrm{d}t, \tag{3.26}$$

alors (3.25) est vérifiée. L'inégalité (3.26) est une généralisation de (3.24), mais semble plus complexe à obtenir que (3.25), et c'est pour cela que l'on cherchera ici seulement à étudier sous quelle condition sur I l'inégalité (3.25) est vérifiée. Notons que par un changement de variable (3.26) revient à regarder la propriété de suite de Riesz sur l'union des 2 intervalles $aI \cup -bI$. On pourra consulter [66] sur le problème de trouver une base de Riesz sur une union d'intervalles et plus récemment [68] dans le cas multidimensionnel. On établit ici le théorème suivant :

Théorème 3.8. *Soit $a, b > 0$, $\alpha \in \mathbb{R}$ et $I = [\alpha T, \alpha T + T]$. On définit T_α par*
- *Si $\alpha = \ell + s$, avec $\ell \in \mathbb{N}$ et $0 < s < 1$, alors $T_\alpha = \min(\frac{2\pi}{\max(a,b)}, \frac{2\pi(\ell+2)}{(a+b)(\ell+1+s)})$.*
- *Si $\alpha = -\ell + s$, avec $\ell \in \mathbb{N}$, $\ell \geq 2$ et $0 < s < 1$, alors $T_\alpha = \min(\frac{2\pi}{\max(a,b)}, \frac{2\pi\ell}{(a+b)(\ell-s)})$.*
- *Si $\alpha = -1 + s$, avec $0 < s < 1$, alors $T_\alpha = \min(\frac{2\pi}{\max(a,b)}, \frac{2\pi}{(a+b)\max(1-s,s)})$.*
- *Si $\alpha \in \mathbb{Z}$ alors $T_\alpha = \min(\frac{2\pi}{\max(a,b)}, \frac{2\pi}{a+b}) = \frac{2\pi}{a+b}$.*

Alors, pour tout $T \geq T_\alpha$, (3.25) est satisfaite, avec une constante $C(I)$ indépendante des coefficients (a_k). Si $T < T_\alpha$, alors (3.25) ne peut pas être satisfaite avec une constante $C(I)$ indépendante des coefficients (a_k).

Démonstration. Soit $\alpha \in \mathbb{R}$ et $I = [\alpha T, \alpha T + T]$. On peut supposer que $aT < 2\pi$ et $bT < 2\pi$. En effet, si $T \geq 2\pi/a$ ou $T \geq 2\pi/b$, (3.25) sera vérifiée. On a

$$\int_I \left(\left| \sum_{k\geq 1} a_k e^{ikat} \right|^2 + \left| \sum_{k\geq 1} a_k e^{-ikbt} \right|^2 \right) \mathrm{d}t = \left(a \int_{\alpha a T}^{\alpha a T + aT} + b \int_{-\alpha b T - bT}^{-\alpha b T} \right) \left| \sum_{k\geq 1} a_k e^{ikt} \right|^2 \mathrm{d}t$$
$$= \left(a \int_0^{aT} + b \int_{-\alpha(a+b)T - bT}^{-\alpha(a+b)T} \right) \left| \sum_{k\geq 1} a_k e^{ikt} \right|^2 \mathrm{d}t. \quad (3.27)$$

On cherche ensuite $k \in \mathbb{Z}$ de telle sorte que

$$-\alpha(a+b)T - bT + 2\pi k \leq aT, \quad -\alpha(a+b)T + 2k\pi \geq 2\pi.$$

Si un tel k existe, on aura

$$\left(a \int_0^{aT} + b \int_{-\alpha(a+b)T - bT}^{-\alpha(a+b)T} \right) \left| \sum_{k\geq 1} a_k e^{ikt} \right|^2 \mathrm{d}t = \left(a \int_0^{aT} + b \int_{-\alpha(a+b)T - bT + 2\pi k}^{-\alpha(a+b)T + 2\pi k} \right) \left| \sum_{k\geq 1} a_k e^{ikt} \right|^2 \mathrm{d}t$$
$$\geq \left(a \int_0^{aT} + b \int_{aT}^{2\pi} \right) \left| \sum_{k\geq 1} a_k e^{ikt} \right|^2 \mathrm{d}t \geq \min(a,b) \int_0^{2\pi} \left| \sum_{k\geq 1} a_k e^{ikt} \right|^2 \mathrm{d}t$$

et on sera donc assuré d'avoir (3.25). On obtient donc

$$T(a+b)(\alpha+1) \geq 2k\pi, \quad \alpha T(a+b) \leq 2(k-1)\pi.$$

On distingue donc plusieurs cas.
- On suppose que $\alpha = \ell + s$, avec $\ell \in \mathbb{N}$ et $0 < s < 1$. On obtient alors

$$\frac{k}{\alpha+1} \leq \frac{T(a+b)}{2\pi} \leq \frac{k-1}{\alpha},$$

et donc $\alpha \leq k - 1$, ce qui implique que $k - 1 > \ell$, donc $k \geq \ell + 2$. Comme on doit avoir $T(a+b)(\alpha+1) \geq 2k\pi$, le temps le plus petit que l'on puisse obtenir est alors obtenu pour $k = \ell + 2$. Dans ce cas, on a $T(a+b) = 2\pi(\ell+2)/(\ell+1+s)$, et pour cette valeur de $T = T_\alpha$, on a donc (3.25). Si maintenant $T = T_\alpha - \varepsilon$, avec $\varepsilon > 0$ on aura

$$-\alpha(a+b)T - bT + 2\pi(\ell+1) - aT = (\alpha+1)\varepsilon > 0, \quad -\alpha(a+b)T + 2(\ell+1)\pi \geq 2\pi.$$

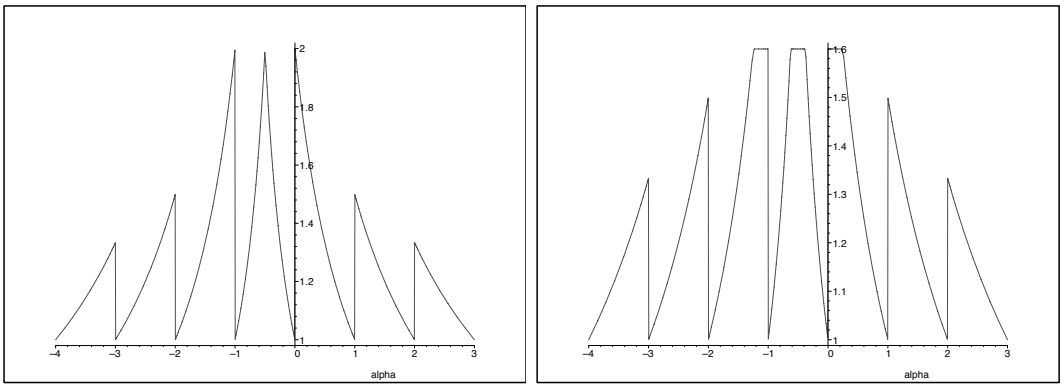

FIGURE 1.2 – $T_\alpha \frac{(a+b)}{2\pi}$ versus α, pour $a = b = 1$ (à gauche) et pour $a = 1$ et $b = 0.6$ (à droite).

D'autre part, on a $aT < 2\pi$, donc on peut choisir ε assez petit de telle sorte que $aT < -\alpha(a+b)T - bT + 2\pi(\ell+1) < 2\pi$. Dans ce cas, on doit avoir $bT \geq 2\pi$, ce qui n'est pas possible.

• On suppose maintenant que $\alpha = -\ell + s$, avec $\ell \in \mathbb{N}$, $\ell \geq 2$ et $0 < s < 1$. On obtient alors

$$\frac{k-1}{\alpha} \leq \frac{T(a+b)}{2\pi} \leq \frac{k}{\alpha+1},$$

et donc $\alpha \geq k-1$, ce qui implique que $k-1 \leq -\ell$. Comme on doit avoir $T(a+b)/(2\pi) \geq (k-1)/\alpha$, le temps le plus petit est obtenu pour $k-1 = -\ell$. On a alors $T = T_\alpha$ et (3.25). Enfin, si $T = T_\alpha - \varepsilon$, avec $\varepsilon > 0$, assez petit de telle sorte que l'on ait :

$$aT < -\alpha(a+b)T + 2(-\ell+1)\pi = 2\pi + \alpha(a+b)\varepsilon < 2\pi.$$

On doit alors avoir à nouveau $bT \geq 2\pi$, ce qui est impossible.

• On suppose maintenant que $\alpha = -1 + s$, avec $0 < s < 1$. On obtient alors

$$\frac{T(a+b)}{2\pi} \geq \max\left(\frac{k}{\alpha+1}, \frac{k-1}{\alpha}\right).$$

Le minimum de la quantité de droite est obtenu pour $k = 0$ ou $k = 1$, de telle sorte que l'on obtient (3.25) pour $T = T_\alpha$. Enfin, si $T = T_\alpha - \varepsilon$, comme $k/(\alpha+1)$ et $2(k-1)/\alpha$ sont distincts pour $k = 0, 1$ et $\alpha \notin \{0, 1\}$, on se retrouve dans l'un des cas des deux cas précédents et donc on ne peut pas avoir (3.25).

• On suppose que $\alpha \in \mathbb{Z}$. On a alors pour $T = T_\alpha$, en prenant le membre de droite de (3.27), $a\int_0^{\frac{2\pi a}{a+b}} + b\int_{\frac{-2\pi b}{a+b}}^0$, et donc on a un intervalle de longueur 2π, ce qui signifie que (3.25) est valide. Enfin, si $T < T_\alpha$, on a $aT + bT < 2\pi$ et donc (3.25) ne peut pas être satisfaite. □

A titre d'exemple, on a représenté sur le Figure 1.2 le temps T_α en fonction de α pour différentes valeurs de a et b. On déduit du Théorème 3.8 le corollaire suivant :

Corollaire 3.9. *Soit $I = [\varepsilon, \varepsilon + T]$, avec $0 < \varepsilon < T$. Alors, si (3.8) a lieu pour toutes les conditions initiales satisfaisant (3.15), on a $T \geq 2\sqrt{2} - \varepsilon$.*

Démonstration. Supposons que l'on ait montré que

$$\int_\varepsilon^{T+\varepsilon} \left| \sum_k a_k e^{i\pi/\sqrt{2}kt} \right|^2 \mathrm{d}t + \int_\varepsilon^{T+\varepsilon} \left| \sum_k a_k e^{-i\pi/\sqrt{2}kt} \right|^2 \mathrm{d}t \geq C(I) \sum_k |a_k|^2, \qquad (3.28)$$

Ceci implique donc que $T \geq T_\alpha$, avec $\alpha = \frac{\varepsilon}{T}$. Comme $0 < \varepsilon < T$, on a $0 < \alpha < 1$ et donc $\frac{a+b}{2\pi} T_\alpha = \frac{2}{1+\alpha}$, avec $a = b = \frac{\pi}{\sqrt{2}}$, ce qui implique que $T \geq 2\sqrt{2} - \varepsilon$, ce qui donne le résultat.
Il reste maintenant à montrer (3.28). On suppose que l'on a l'observabilité uniforme (3.8) pour $I = [\varepsilon, T+\varepsilon]$ pour des données initiales satisfaisant (3.15). On considère alors une sous classe de solutions satisfaisant

$$a_k = a_{-k}, \quad N^- := N/2 - N^{1/4} \leq k \leq N/2, \quad a_k = 0 \quad |k| < N^-.$$

On a alors en posant $\nu_k = \frac{\lambda_k^0}{\lambda_{N+1-k}^0}$

$$\int_I \left| \frac{u_N(t)}{h} \right|^2 \mathrm{d}t$$
$$\leq 2 \int_I \left| \sum_{k \geq N^-}^{N/2} \frac{e_N^k}{h} (a_k e^{i\lambda_k t} + a_k \nu_k^2 e^{i\lambda_{N+1-k} t}) \right|^2 + \left| \sum_{k \geq N^-}^{N/2} \frac{e_N^k}{h} (a_k e^{-i\lambda_k t} + a_k \nu_k^2 e^{-i\lambda_{N+1-k} t}) \right|^2 \mathrm{d}t,$$

et donc

$$C(T) \sum_{k=N^-}^{N/2} |\lambda_k|^2 |a_k|^2 \leq \int_I \left| \sum_{k \geq N^-}^{N/2} \frac{e_N^k}{h} a_k e^{i\lambda_k t} \right|^2 \mathrm{d}t + \int_I \left| \sum_{k \geq N^-}^{N/2} \nu_k^2 \frac{e_N^k}{h} a_k e^{i\lambda_{N+1-k} t} \right|^2 \mathrm{d}t,$$

ce qui donne (en utilisant le fait que $\left| \frac{e_N^k}{\lambda_k h} \right| \geq \cos(\pi/4)$, dès que $N/2 - \sqrt{N} - 1 \leq k \leq N/2$),

$$C(T) \sum_{k=N^-}^{N/2} |a_k|^2 \leq \int_I \left| \sum_{k \geq N^-}^{N/2} a_k e^{i\lambda_k t} \right|^2 \mathrm{d}t + \int_I \left| \sum_{k \geq N^-}^{N/2} \nu_k^2 a_k e^{i\lambda_{N+1-k} t} \right|^2 \mathrm{d}t.$$

En décomposant $\lambda_k^0 = 2/h \sin(\pi/4 + (k - 1/(2h))\pi h/2)$ et $\lambda_{N+1-k}^0 = 2/h \cos(\pi/4 + (k - 1/(2h))\pi h/2)$, on obtient

$$\int_\varepsilon^{T+\varepsilon} \left| \sum_k a_k e^{i\pi/\sqrt{2}kt} \right|^2 \mathrm{d}t + \int_\varepsilon^{T+\varepsilon} \left| \sum_k \nu_k^2 a_k e^{-i\pi/\sqrt{2}kt} \right|^2 \mathrm{d}t \geq C(T) \sum_k |a_k|^2.$$

Maintenant, en prenant une borne supérieure de $|\nu_k^2 - 1|$ qui tend vers zéro, l'inégalité se réduit à (3.28). □

Remarque 3.10. *On ne sait pas si le résultat reste valable pour $\varepsilon = 0$. On sait que si $T \geq 2\sqrt{2}$, alors (3.8) a lieu pour toutes les conditions initiales satisfaisant (3.15), indépendamment de la position de l'intervalle, cf [63]. Notons qu'en prenant $\varepsilon = 0$ dans (3.28), on obtient seulement $T \geq 2$, et le problème de connaître le temps optimal pour avoir (3.8) pour toutes les conditions initiales satisfaisant (3.15), dans le cas de l'intervalle $[0, T]$ reste ouvert.*

3.7 Preuve d'un théorème de type Ingham

On prouve ici le Théorème 3.3. On pourra trouver la démonstration du Théorème 3.4 (plus technique) dans [63].

Preuve du Théorème 3.3. On utilise la première méthode d'Ingham, en suivant d'abord [65]. On considère la fonction

$$G(t) = \begin{cases} \cos \frac{\pi t}{T} & \text{if } |t| \leq T/2 \\ 0 & \text{si } |t| > T/2. \end{cases}$$

Sa transformée de Fourier $\tilde{K}$ satisfait

$$\tilde{K}(\tau) = \int_{-\infty}^{\infty} G(t) e^{i\tau t} \mathrm{d}t = -\frac{2T\pi \cos(\tau T/2)}{\tau^2 T^2 - \pi^2} = \frac{2T}{\pi} K_T(\tau),$$

avec

$$K(\tau) = \frac{\cos(\frac{\pi}{2}\tau)}{1-\tau^2}, \quad K_T(\tau) = K(\frac{T}{\pi}\tau). \tag{3.29}$$

Donc, on a

$$\int_{-T/2}^{T/2} |\sum_{|k|=1}^{N} a_k e^{i\lambda_k t} + b_k e^{i\mu_k t}|^2 \mathrm{d}t \geq \frac{2T}{\pi} \sum_{|k|,|j|=1}^{N} K_T(\lambda_k - \lambda_j) a_k \overline{a_j} + b_k \overline{b_j} K_T(\mu_k - \mu_j) + a_k \overline{b_j} K_T(\lambda_k - \mu_j) + \overline{a_k} b_j K_T(\mu_k - \lambda_j),$$

car $0 \leq k \leq 1_{[-T/2,T/2]}$. D'autre part, puisque G est positif, on a aussi

$$\sum_{|k|,|j|=1}^{N} b_k \overline{b_j} K_T(\mu_k - \mu_j) = \int G(t) \Big| \sum_{|k|=1}^{N} b_k e^{i\mu_k t} \Big|^2 \mathrm{d}t \geq 0. \tag{3.30}$$

On peut donc se débarrasser de ces termes, comme cela avait été remarqué dans [65] ; notons que cet argument n'est pas valable si on utilise la deuxième méthode d'Ingham (et donc nous n'avons pas été capable de suivre la preuve plus courte du résultat de [65] dans [61]). On obtient donc, grâce à l'inégalité triangulaire

$$\int_{-T/2}^{T/2} |\sum_{|k|=1}^{N} a_k e^{i\lambda_k t} + b_k e^{i\mu_k t}|^2 \mathrm{d}t \geq \frac{2T}{\pi} [\sum_{|k|=1}^{N} K_T(0)|a_k|^2 - \sum_{|k|=1}^{N} \sum_{j \neq k} |a_k||a_j||K_T(\lambda_k - \lambda_j)| - 2 \sum_{|k|,|j|=1}^{N} |a_k||b_j||K_T(\lambda_k - \mu_j)|].$$

Nous devons maintenant procéder différemment que dans [65], afin de traiter les coefficients b_k dont la somme peut ne pas être bornée indépendamment de N.
Nous avons déjà à partir de l'hypothèse d'écart de la suite (λ_k) que $|\lambda_k - \lambda_j| \geq \gamma|k-j|$.
Il s'ensuit de l'hypothèse d'écart des termes mixtes que

$$|\lambda_k - \mu_j| = \mu_j - \mu_N + \mu_N - \lambda_N + \lambda_N - \lambda_k \geq (N-j)\gamma + \gamma + (N-k)\gamma = (2N+1-j-k)\gamma,$$

Si $k, j \geq N - N^\alpha$, et

$$|\lambda_k - \mu_j| = \lambda_k - \lambda_{-N} + \lambda_{-N} - \mu_{-N} + \mu_{-N} - \mu_j \geq (N+k)\gamma + \gamma + (N+j)\gamma = (2N+1+j+k)\gamma,$$

si $k, j \leq N^\alpha - N$, en utilisant les hypothèses

$$\begin{aligned}\mu_k \leq \mu_{N^\alpha - N} \leq \cdots \leq \mu_{-N} \leq \lambda_{-N} \leq \cdots \leq \lambda_{N^\alpha - N} \leq \cdots \leq \lambda_{-1} \\ \leq \lambda_1 \leq \cdots \leq \lambda_{N - N^\alpha} \leq \cdots \leq \lambda_N \leq \mu_N \leq \cdots \leq \mu_{N - N^\alpha} \leq \mu_\ell,\end{aligned}$$

pour $1 \leq \ell \leq N - N^\alpha$ and $-1 \geq k \geq N^\alpha - N$. On obtient aussi

$$\begin{aligned}&|\lambda_k - \mu_j| = \mu_j - \lambda_k \geq \mu_N - \lambda_{N-N^\alpha} \geq \lambda_N - \lambda_{N-N^\alpha} \geq CN^\alpha, \\ &\text{if } 1 \leq k \leq N - N^\alpha,\ j \geq 1, \\ &|\lambda_k - \mu_j| = \mu_j - \lambda_k \geq \mu_{N-N^\alpha} - \lambda_N \geq \mu_{N-N^\alpha} - \mu_N \geq CN^\alpha, \\ &\text{if } k \geq 1,\ 1 \leq j \leq N - N^\alpha, \\ &|\lambda_k - \mu_j| = \lambda_k - \mu_j \geq \lambda_{N^\alpha - N} - \mu_{-N} \geq \lambda_{N^\alpha - N} - \lambda_{-N} \geq CN^\alpha, \\ &\text{if } N^\alpha - N \leq k \leq -1,\ j \leq -1, \\ &|\lambda_k - \mu_j| = \lambda_k - \mu_j \geq \lambda_{-N} - \mu_{N^\alpha - N} \geq \mu_{-N} - \mu_{N^\alpha - N} \geq CN^\alpha, \\ &\text{if } k \leq -1,\ N^\alpha - N \leq j \leq -1.\end{aligned}$$

On a donc $|\lambda_k - \mu_j| \geq \gamma d_{k,j}$ avec une suite $d_{k,j}$ satisfaisant

$$\begin{aligned}d_{k,j} = 2N+1-k-j, \quad k, j \geq N - N^\alpha, \quad d_{k,j} = 2N+1-k-j, \quad k, j \leq -N + N^\alpha, \\ d_{k,j} \geq CN^\alpha, \quad \text{sinon.}\end{aligned}$$

On a donc

$$|K_T(\lambda_k - \mu_j)| \leq (\frac{2\pi}{T\gamma})^2 \frac{1}{4d_{k,j}^2 - 1}, \quad d_{k,j} = d_{j,k}, \quad |k|, |j| = 1, \ldots, N,$$

et

$$\begin{aligned}\int_{-T/2}^{T/2} |\sum_{|k|=1}^{N} a_k e^{i\lambda_k t} + b_k e^{i\mu_k t}|^2 \mathrm{d}t \geq \frac{2T}{\pi} [\sum_{|k|=1}^{N} K_T(0)|a_k|^2 \\ - (\frac{2\pi}{T\gamma})^2 (\sum_{|k|=1}^{N} \sum_{j \neq k} |a_k||a_j| \frac{1}{4(k-j)^2 - 1} - 2 \sum_{|k|,|j|=1}^{N} |a_k||b_j| \frac{1}{4d_{k,j}^2 - 1})].\end{aligned}$$

Pour les termes mixtes, on calcule

$$\begin{aligned}\sum_{|k|,|j|=1}^{N} |a_k||b_j| \frac{1}{4d_{k,j}^2 - 1} \leq \sum_{|k|,|j|=1}^{N} (\frac{M}{2}|a_k|^2 + \frac{1}{2M}|b_j|^2) \frac{1}{4d_{k,j}^2 - 1} \\ = \sum_{|k|=1}^{N} (\frac{M}{2}|a_k|^2 + \frac{1}{2M}|b_k|^2) \sum_{|j|=1}^{N} \frac{1}{4d_{k,j}^2 - 1} \\ = \sum_{k=1}^{N} (\frac{M}{2}(|a_k|^2 + |a_{-k}|^2 + \frac{1}{2M}(|b_k|^2 + |b_{-k}|^2) \sum_{|j|=1}^{N} \frac{1}{4d_{k,j}^2 - 1} \leq \sum_{k=1}^{N} M|a_k|^2 \sum_{|j|=1}^{N} \frac{1}{4d_{k,j}^2 - 1}.\end{aligned}$$

On a aussi classiquement

$$\sum_{|k|=1}^{N}\sum_{j\neq k}|a_k||a_j|\frac{1}{4(k-j)^2-1}\leq\sum_{k=1}^{N}|a_k|^2\sum_{j\neq k}^{N}\frac{1}{4(k-j)^2-1}.$$

On obtient alors

$$\int_{-T/2}^{T/2}|\sum_{|k|=1}^{N}a_ke^{i\lambda_kt}+b_ke^{i\mu_kt}|^2\mathrm{d}t\geq\frac{2T}{\pi}\sum_{|k|=1}^{N}(\frac{2\pi}{T\gamma})^2c_k|a_k|^2,$$

avec

$$c_k:=(\frac{T\gamma}{2\pi})^2K_T(0)-\sum_{|j|=1,j\neq k}^{N}\frac{1}{4(k-j)^2-1}-2M\sum_{j=1}^{N}\frac{1}{4d_{k,j}^2-1}. \tag{3.31}$$

On calcule, pour $k\geq N-N^\alpha$

$$\begin{aligned}c_k&=(\frac{T\gamma}{2\pi})^2-(\sum_{j=-N}^{k-1}+\sum_{j=k+1}^{N})\frac{1}{4(k-j)^2-1}-2M\sum_{j=N-N^\alpha}^{N}\frac{1}{4(2N+1-k-j)^2-1})+CN^{1-2\alpha}\\&=(\frac{T\gamma}{2\pi})^2-o(1)-(\sum_{\ell=1}^{N+k}+\sum_{\ell=1}^{N-k}+2M\sum_{\ell=N+1-k}^{N+1-k+N^\alpha})\frac{1}{2}(\frac{-1}{2\ell+1}+\frac{1}{2\ell-1})\\&\geq(\frac{T\gamma}{2\pi})^2-o(1)-(1+(M-\frac{1}{2})\frac{1}{2(N-k)+1})\geq(\frac{T\gamma}{2\pi})^2-o(1)-\max(1,M+\frac{1}{2}),\end{aligned}$$

et on obtient de manière similaire le même résultat pour $k\leq -N+N^\alpha$, de telle sorte que la preuve est faite, en prenant N suffisamment grand. □

4 Perspectives de recherche

Etude d'autres géométries On cherche à appliquer la méthode d'Ingham à d'autres géométries, où des informations suffisantes sur le spectre sont connues.

Etude de réseaux Dans la continuité de précédents travaux [2, 3], on cherche à appliquer des théorèmes de type Ingham. On s'intéresse aussi au taux de décroissance pour la stabilisation ainsi qu'à des développements numériques.

Deuxième partie

Chapitre 2

Introduction

1 Les schémas semi-Lagrangiens

Dans le cadre de la résolution des équations cinétiques (équation de Vlasov), il existe essentiellement deux types de méthodes. D'une part les méthodes PIC (Particle in Cell) et d'autre part les méthodes basées sur un maillage de l'espace des phases (méthodes eulériennes ou semi-Lagrangiennes).

Actuellement, beaucoup de codes utilisent encore les méthodes PIC et ses variantes. Mais la tendance change avec la montée en puissance des ordinateurs : de plus en plus de codes utilisent maintenant une grille de l'espace des phases. Il en est ainsi du code gyrocinétique semi-Lagrangien GYSELA (GYrokinetic SEmi LAgrangian).

Le principe d'une méthode semi-Lagrangienne est basé sur le fait que la fonction de distribution qui représente la probabilité de présence des particules est constante le long de courbes caractéristiques. Pour mettre à jour en temps la fonction de distribution connue sur un maillage, on doit alors préciser comment calculer numériquement les caractéristiques et comment interpoler. Il s'agit en quelque sorte d'un compromis entre les méthodes purement eulériennes (volumes finis/ différences finies) généralement contraintes à une condition CFL (typiquement, le déplacement doit être inférieur à une maille) et les méthodes PIC qui ne passent pas par une phase de projection dans l'espace des phases.

En ce qui concerne les équations cinétiques, ces méthodes ont été développées dès les années 70, avec un travail pionnier de Cheng et Knorr [22]. Elles ont été remises au goût du jour vers la fin des années 90 avec Eric Sonnendrücker [105]. On pourra consulter Shoucri [103] (et les nombreuses références) pour une vue d'ensemble historique de ce type de méthodes. Pour une revue plus récente et plus générale des méthodes numériques pour la fusion, on pourra consulter [44].

Notons aussi que les méthodes semi-Lagrangiennes ont été initialement développées et sont largement utilisées dans les simulations pour le climat (voir [108], pour une revue). Elles ont pris leur essor dans les années 80 (cf [96]). Parallèlement se sont aussi développées les méthodes de Galerkin charactéristiques, parfois aussi assimilées à des méthodes semi-Lagrangiennes (voir par exemple [86], pour des problèmes d'advection diffusion en mécanique des fluides). On mentionne aussi la conférence "Recent advances on theory and applications of Semi-Lagrangian methods", tenue à Rome, les 5-6 décembre 2011 (http ://www.mat.uniroma1.it/ricerca/convegni/2011/SL/.)

2 L'équation de Vlasov

2.1 La forme générale

L'équation de Vlasov s'écrit sous la forme

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

La fonction f qui dépend du temps t, de la position x et de la vitesse v est appelée fonction de distribution. La quantité $f(t,x,v)dxdv$ représente la probabilité de trouver des particules dans un élément de volume $dxdv$ au temps t et au point (x,v). Il s'agit d'une équation de transport. Cette équation est non linéaire à travers le champ F qui dépend de f par l'intermédiaire d'équations de Poisson ou Maxwell. Elle décrit la dynamique de particules chargées dans un plasma.

2.2 Le système de Vlasov-Poisson $1D \times 1D$

Un des premiers modèles étudiés dans le cadre de la simulation numérique des plasmas, qui rentre dans le cadre du transport linéaire, grâce à un splitting en temps, comme nous allons le voir par la suite, est le système de Vlasov-Poisson

$$\partial_t f(t,x,v) + v\partial_x f(t,x,v) + E(t,x)\partial_v f(t,x,v) = 0, \ (x,v) \in [0,L] \times \mathbb{R}. \tag{2.1}$$

Le champ E est solution de l'équation de Poisson

$$\partial_x E(t,x) = \int_{\mathbb{R}} f(t,x,v)dv - 1,$$

et on rajoute la condition d'intégrale nulle :

$$\int_0^L E(t,x)dx = 0.$$

En partant d'une donnée initiale régulière, la solution garde la régularité. Néanmoins, il se développe des fines structures. Typiquement, la fonction de distribution est constante le long de courbes appelées caractéristiques et celles-ci peuvent se rapprocher de plus en plus (on parle d'enroulement des caractéristiques), et au bout d'un moment le maillage ne peut plus résoudre ces fines structures.
On introduit la notation

$$E[f](x) = \partial_x^{-1}\left[\int_{\mathbb{R}} f(x,v)\mathrm{d}v - 1\right] = \int_0^L K(x,y)\left(\int_{\mathbb{R}} f(y,w)\mathrm{d}w - 1\right)\mathrm{d}y,$$

avec

$$K(x,y) = \frac{y}{L}, \text{ si } 0 \le y \le x, \ K(x,y) = \frac{y}{L} - 1, \text{ si } x \le y \le L.$$

Le Hamiltonien donné par

$$\begin{aligned} H[f] &= \frac{1}{2L}\int_0^L \int_{\mathbb{R}} v^2 f(x,v)\mathrm{d}x\mathrm{d}v + \frac{1}{2L}\int_0^L \big(E[f](x)\big)^2\mathrm{d}x \\ &=: T[f] + U[f] \end{aligned} \tag{2.2}$$

est préservé le long de la solution de (2.1).

2.3 Le modèle centre-guide

Un modèle plus complexe, qui ne rentre pas dans le cadre du transport linéaire, est le modèle centre-guide, dont l'inconnue $f(t,x,y)$ satisfait

$$\partial_t f + E_y(t,x,y)\partial_x f - E_x(t,x,y)\partial_y f = 0,\ -\Delta\Phi = f,\ E = -\nabla f. \tag{2.3}$$

On utilisera des conditions périodiques dans les deux directions, voir [102].

Chapitre 3

Discrétisation en temps pour Vlasov-Poisson

Ce chapitre est issu du travail [27].

1 Le splitting de Strang

Un des schémas en temps les plus populaires est le splitting de Strang, déjà utilisé par Cheng et Knorr [22]. Il consiste à résoudre successivement l'advection en x

$$\partial_t f + v\partial_x f = 0,$$

pendant un temps $\Delta t/2$, puis de mettre à jour le champ E par l'intermédiaire de l'équation de Poisson, de résoudre ensuite pendant un temps Δt l'advection en v

$$\partial_t f + E\partial_v f = 0,$$

et de recommencer avec l'advection en x pendant $\Delta t/2$. En partant d'une approximation de la solution à l'instant $t_n = t_0 + n\Delta t$, on obtient alors une approximation de l'erreur à l'instant t_{n+1}.
Par cette technique, on obtient un schéma en temps d'ordre 2 et on se ramène ici à une advection à coefficient constant pour la résolution spatiale. On peut économiser une advection, en factorisant le dernier splitting en x avec le premier splitting de l'étape suivante et ne pas faire cette factorisation lorsqu'on veut faire un diagnostique.

2 Schémas de splitting d'ordre élevé

S'il est souvent amplement satisfaisant d'utiliser le splitting de Strang, il parait naturel de chercher à augmenter l'ordre, notamment lorsque la discrétisation spatiale utilise elle-même un schéma d'ordre élevé (cf [39]). Quelques travaux existent sur le sujet dans le cadre de la résolution numérique de l'équation de Vlasov [88, 97, 112] et se basent généralement sur [113]. Notons aussi un résultat récent de Schaeffer [98] dans le cas linéaire. D'autre part, il existe maintenant une littérature bien développée sur la construction de schémas d'ordre élevé pour les EDO basés sur le splitting [12, 13, 48, 104].

Notons également un travail récent basé sur la méthode IDC (Integral Deferred Corrections) [87].

2.1 Introduction

Les méthodes de splitting que l'on considère ici sont basées sur la décomposition $H = T + U$ du Hamiltonien (2.2). Les équations hamiltoniennes associées à T et U sont tout simplement les équations

$$\partial_t f + v\partial_x f = 0 \quad \text{et} \quad \partial_t f + E[f]\partial_v f = 0 \tag{2.1}$$

respectivement. Ces deux équations peuvent être résolues explicitement en utilisant la formule des caractéristiques. Pour une donnée initiale f_0, la solution de la première est donnée par $f(t,x,v) = f_0(x - tv, v)$ et la deuxième par la relation $f(t,x,v) = f_0(x, v - tE[f_0])$ en notant que $E[f_0]$ est une constante du mouvement dans l'évolution du système Hamiltonien associé à U.
Ainsi, on est naturellement amené à étudier la classe suivante de méthodes : pour un nombre s donné $s \in \mathbb{N}^*$, des coefficients a_p, $p = 0, \ldots, 2s$, et un pas de temps $\Delta t > 0$, on définit le schéma de splitting avec $2s+1$ étapes par les relations $g_1(x,v) = f_0(x - a_0\Delta t v, v)$, et

$$\begin{aligned} g_{2j}(x,v) &= g_{2j-1}(x, v - a_{2j-1}E[g_{2j-1}](x)\Delta t), \\ g_{2j+1}(x,v) &= g_{2j}(x - a_{2j}\, v\, \Delta t, v), \end{aligned} \tag{2.2}$$

pour $j = 1, \ldots, 2s$. La quantité $g_{2s+1}(x,v)$ devrait être une approximation de $f(\Delta t, x, v)$. Notons que ces schémas préservent la masse

$$\int_0^L \int_{\mathbb{R}} f(t,x,v)dxdv = \int_0^L \int_{\mathbb{R}} f(0,x,v)dxdv,$$

et les normes L_p

$$\int_0^L \int_{\mathbb{R}} |f(t,x,v)|^p dxdv = \int_0^L \int_{\mathbb{R}} |f(0,x,v)|^p dxdv, \ p \in \mathbb{N}^*.$$

Malheureusement, cela ne reste généralement plus vrai, lorsque l'on introduit une discrétisation spatiale.
En ce qui concerne le cas de la dimension finie, il existe beaucoup de travaux concernant l'analyse des conditions d'ordre pour le splitting (voir en particulier [12, 13, 48] et références incluses). Dans ce cadre, rappelons que pour les systèmes d'équations différentielles de la forme

$$\dot{y}(t) = f_A(y(t)) + f_B(y(t)), \quad y(0) = y_0 \in \mathbb{R}^n, \tag{2.3}$$

avec $f_A, f_B : \mathbb{R}^n \to \mathbb{R}^n$, alors en notant by $\mathcal{L}_A$ and $\mathcal{L}_B$ les opérateurs de Lie associés à ces champs de vecteurs, une méthode de splitting d'ordre d est une composition de la forme

$$\prod_{i=1}^{2s+1} \exp(c_i \Delta t \mathcal{L}_A) \exp(d_i \Delta t \mathcal{L}_B) = \exp(\Delta t \mathcal{L}_{A+B}) + O((\Delta t)^{d+1}),$$

avec des coefficients appropriés c_i, d_i. Ici, $\exp(t\mathcal{L}_{A+B})$ est considéré comme le flot exact du système d'EDO (2.3) associé au champ de vecteur $f_A + f_B$.
Dans le cas particulier où f_A et f_B satisfont la relation

$$[[[\mathcal{L}_A, \mathcal{L}_B], \mathcal{L}_B], \mathcal{L}_B] = 0, \tag{2.4}$$

où $[\,\cdot\,,\,\cdot\,]$ est défini comme le crochet de Lie des deux opérateurs, alors les conditions d'ordre algébriques sur les coefficients c_i et d_i peuvent être simplifiés, et on parle de méthodes de Runge Kutta Nyström (RKN) (voir [12] pour une revue). Les systèmes différentiels du second ordre de la forme $\ddot{y}(t) = g(y(t)),\ g : \mathbb{R}^n \to \mathbb{R}^n$ en sont un cas particulier important.

2.2 Structure de Poisson

En ce qui concerne le système de Vlasov-Poisson, nous allons voir que les fonctionnelles T et U dans la décomposition (2.2) satisfont les relations de type RKN suivantes

$$\{\{\{T, U\}_f, U\}_f, U\}_f = 0, \tag{2.5}$$

où $\{\cdot,\cdot\}_f$ est le crochet de Poisson associé à la structure de Poisson de dimension infinie. Pour cela, on introduit d'abord le crochet de Poisson de dimension finie

$$\{f, g\} = \partial_x f \partial_v g - \partial_v f \partial_x g,$$

pour f et g des fonctions régulières de $(x, v) \in [0, L] \times \mathbb{R}$. On introduit ensuite le crochet de Poisson pour les fonctionnelles $H[f]$ et $G[f]$

$$\{H, G\}_f = \frac{1}{L}\int_0^L \int_{\mathbb{R}} \frac{\delta H}{\delta f}[f]\{\frac{\delta G}{\delta f}[f], f\}\mathrm{d}x\mathrm{d}v, \tag{2.6}$$

où $\dfrac{\delta H}{\delta f}[f]$ est la dérivée de Fréchet évaluée en f.
On a alors la proposition suivante

Proposition 2.1. *Les fonctionnelles $T[f]$ et $U[f]$ satisfont la relation*

$$\{\{T, U\}_f, U\}_f = 2U[f]. \tag{2.7}$$

En particulier, l'identité RKN (2.5) *est valable.*

Preuve. On a d'abord

$$\frac{\delta T}{\delta f}[f] = \frac{v^2}{2} \quad \text{et} \quad \frac{\delta U}{\delta f}[f] = \phi[f](x).$$

avec

$$\phi[f](x) = -\partial_{xx}^{-1}\left[\int_{\mathbb{R}} f(x, v)\mathrm{d}v - 1\right] = -\partial_x^{-1} E[f](x).$$

On calcule alors

$$\begin{aligned}\{T, U\}_f = \frac{1}{L}\int_0^L \int_{\mathbb{R}} \frac{\delta T}{\delta f}[f]\{\frac{\delta U}{\delta f}[f], f\}\mathrm{d}x\mathrm{d}v &= \frac{1}{L}\int_0^L \int_{\mathbb{R}} \frac{v^2}{2}\{\phi[f], f\}\mathrm{d}x\mathrm{d}v \\ &= -\frac{1}{L}\int_0^L \int_{\mathbb{R}} \phi[f]\{\frac{v^2}{2}, f\}\mathrm{d}x\mathrm{d}v,\end{aligned}$$

en utilisant les propriétés

$$\{f_1 f_2, f_3\} = f_1\{f_2, f_3\} + f_2\{f_1, f_3\}, \ \int_0^L \int_{\mathbb{R}} \{f_1, f_2\} dx dv = 0.$$

On peut donc écrire

$$\{T, U\}_f = \frac{1}{L} \int_0^L \int_{\mathbb{R}} \phi[f](x) v \partial_x f(x, v) \mathrm{d}x \mathrm{d}v.$$

Soit maintenant une fonction δf avec $\frac{1}{L}\int_0^L \int_{\mathbb{R}} \delta f = 0$. On a alors

$$\begin{aligned}\{T, U\}_{f+\delta f} = \{T, U\}_f &+ \frac{1}{L} \int_0^L \int_{\mathbb{R}} \phi[f](x) v \partial_x \delta f(x, v) \mathrm{d}x \mathrm{d}v \\ &- \frac{1}{L} \int_0^L \int_{\mathbb{R}} \Big(\partial_{xx}^{-1} \int_{\mathbb{R}} \delta f(x, w) \mathrm{d}w \Big) v \partial_x f(x, v) \mathrm{d}x \mathrm{d}v + \mathcal{O}(\delta f^2).\end{aligned}$$

On en déduit par intégrations par parties que

$$\begin{aligned}\frac{\delta\{T, U\}_f}{\delta f}[f] &= -v\partial_x \phi[f](x) - \int_{\mathbb{R}} v \partial_{xx}^{-1} (\partial_x f(x, v)) \mathrm{d}v \\ &=: vE[f](x) + Z[f](x),\end{aligned}$$

ce qui donne

$$\begin{aligned}\{\{T, U\}_f, U\}_f &= \frac{1}{L} \int_0^L \int_{\mathbb{R}} (Z[f](x) + vE[f](x)) \{\phi[f](x), f(x, v)\} dx dv \\ &= \frac{1}{L} \int_0^L \int_{\mathbb{R}} (Z[f](x) + vE[f](x)) \partial_x \phi[f](x) \partial_v f(x, v) dx dv.\end{aligned}$$

Comme $\int_{\mathbb{R}} \partial_v f(x, v) dv = 0$, on obtient

$$\begin{aligned}\{\{T, U\}_f, U\}_f &= -\frac{1}{L} \int_0^L \int_{\mathbb{R}} E[f](x)^2 v \partial_v f(x, v) \mathrm{d}x \mathrm{d}v \\ &= \frac{1}{L} \int_0^L E[f](x)^2 \Big(\int_{\mathbb{R}} f(x, v) \mathrm{d}v \Big) \mathrm{d}x.\end{aligned}$$

Puisque

$$\int_{\mathbb{R}} f(x, v) \mathrm{d}v = 1 + \partial_x E[f](x),$$

on a alors

$$\begin{aligned}\{\{T, U\}_f, U\}_f &= \frac{1}{L} \int_0^L E[f](x)^2 \mathrm{d}x + \frac{1}{3L} \int_0^L \partial_x \big(E[f](x)^3 \big) \mathrm{d}x \\ &= \frac{1}{L} \int_0^L E[f](x)^2 \mathrm{d}x = 2U[f],\end{aligned}$$

en utilisant l'expression de E et U. □

Ce type de relation permet de trouver les conditions d'ordre dans ce cadre abstrait. Notons en particulier que l'on peut espérer avoir un plus large ensemble de coefficients que ceux donnés par les relations RKN, puisque le système de Vlasov-Poisson vérifie la condition plus forte (2.7). Nous allons maintenant voir comment dériver les relations d'ordre de manière élémentaire, jusquà l'ordre 4, sans utiliser ce cadre abstrait qui demande de manipuler les opérateurs de Lie de dimension infinie agissant sur des espaces de Banach.

2.3 Conditions d'ordre ≤ 4 et caractéristiques

On choisit donc ici de travailler au niveau des charactéristiques. On introduit

$$X(t;h,x,v) = x + \int_h^t V(\sigma;h,x,v)d\sigma, \tag{2.8}$$

$$V(t;h,x,v) = v + \int_h^t E(\sigma, X(\sigma;h,x,v))d\sigma, \tag{2.9}$$

où le champ électrique est donné par

$$E(t,x) = \int_0^L K(x,y)\left(\int_{\mathbb{R}} f_0(X(0;t,y,w),V(0;t,y,w))\mathrm{d}w - 1\right)\mathrm{d}y, \tag{2.10}$$

avec f_0 la condition initiale. La solution de l'équation de Vlasov-Poisson est alors donnée par

$$f(t,x,v) = f_0(X(0;t,x,v),V(0;t,x,v)).$$

On définit maintenant les caractéristiques numériques en avant associées au schéma de splitting (2.2) par la relation

$$g_{2s+1}(X_s(\Delta t;0,x,v),V_s(\Delta t;0,x,v)) = f(\Delta t, X(\Delta t;0,x,v),V(\Delta t;0,x,v)). \tag{2.11}$$

La stratégie que l'on utilise est alors de dériver les conditions satisfaites par les coefficients $a_j,\ j=0,\dots,2s$, de telle sorte que les caractéristiques numériques en avant soient des approximations d'ordre quatre des caractéristiques (continues), c'est-à-dire

$$X_s(\Delta t;0,x,v) = X(\Delta t;0,x,v) + O(\Delta t^5),\ V_s(\Delta t;0,x,v) = V(\Delta t;0,x,v) + O(\Delta t^5).$$

On suppose ici assez de régularité sur la donnée initiale f_0 de telle sorte que le calcul des dérivées en temps puissent être effectuées.

Considérons les caractéristiques du système de Vlasov-Poisson donné par (2.8) et (2.9). Un développement de Taylor des caractéristiques en arrière autour de $h=0$ s'écrit

$$X(0;\Delta t,x,v) = \sum_{i=0}^{d} X_b^{[i]}\frac{\Delta t^i}{i!} + O(\Delta t^{d+1}),\ V(0;\Delta t,x,v) = \sum_{i=0}^{d} V_b^{[i]}\frac{\Delta t^i}{i!} + O(\Delta t^{d+1}),$$

avec

$$X_b^{[i]} = \partial_h^i X(0;0,x,v),\ V_b^{[i]} = \partial_h^i V(0;0,x,v),\ i=0,\dots,d.$$

De manière similaire, un développement de Taylor des caractéristiques en avant autour de $t=0$ s'écrit

$$X(\Delta t;0,x,v) = \sum_{i=0}^{d} X_f^{[i]}\frac{\Delta t^i}{i!} + O(\Delta t^{d+1}),\ V(0;\Delta t,x,v) = \sum_{i=0}^{d} V_f^{[i]}\frac{\Delta t^i}{i!} + O(\Delta t^{d+1}),$$

avec

$$X_f^{[i]} = \partial_t^i X(0;0,x,v),\ V_f^{[i]} = \partial_t^i V(0;0,x,v),\ i=0,\dots,d.$$

Notons que, puisque le système n'est pas autonome, on peut ne pas avoir

$$Z(0,\Delta t,x,v) = Z(-\Delta t,0,x,v),\ \text{for } Z\in\{X,Y\},$$

de telle sorte que les coefficients en avant $Z_f^{[i]}$ et les coefficients en arrière $Z_b^{[i]}$, avec $Z \in \{X, Y\}$, peuvent ne pas satisfaire $Z_b^{[i]} = (-1)^i Z_f^{[i]}$.
On définit les moments

$$I_k(x) = \int_{\mathbb{R}} v^k f_0(x,v)\mathrm{d}v, \quad k = 0, \dots, d, \tag{2.12}$$

et $\overline{I}_1 = \frac{1}{L}\int_0^L I_1(y)dy$.
On peut alors exprimer les dérivées en temps du champ électrique et des caractéristiques en termes des moments (2.12). Notons que certains de ces calculs ont été récemment effectués dans [94], [35], [97] pour dériver les schémas de Cauchy-Kovalevsky. Les expressions sont données dans les trois lemmes suivants.

Lemme 2.2. *Les trois premières dérivées du champ électrique sont données par*

$$\begin{aligned}
&\partial_x E(0,x) = I_0(x) - 1,\\
&\partial_t E(0,x) = -I_1(x) + \overline{I}_1,\\
&\partial_t^2 E(0,x) = \partial_x I_2(x) - E(0,x) I_0(x),\\
&\partial_t^3 E(0,x) = -\partial_x^2 I_3(x) + 3\partial_x(E(0,x) I_1(x)) + I_1(x) - \overline{I}_1 I_0(x).
\end{aligned}$$

Remarque 2.3. *Les expressions précédentes ont déjà été obtenues et utilisées dans [97] ; cela permet d'avoir une estimation d'ordre trois en temps du champ électrique dépendant du temps et d'obtenir un schéma d'ordre quatre en ayant besoin de calculer le champ électrique une seule fois par pas de temps.*

Lemme 2.4. *Les premiers coefficients pour les caractéristiques en avant sont données par*

$$\begin{aligned}
&X_f^{[0]} = x,\ X_f^{[j]} = V_f^{[j-1]},\ j = 1, \dots, d,\\
&V_f^{[0]} = v,\ V_f^{[1]} = E(0,x),\\
&V_f^{[2]} = v(I_0(x) - 1) - I_1(x) + \overline{I}_1,\\
&V_f^{[3]} = v^2 \partial_x I_0(x) - E(0,x) + \partial_x I_2(x) - 2v\partial_x I_1(x),\\
&V_f^{[4]} = -\partial_x^2 I_3(x) + 3v\partial_x^2 I_2(x) - 3v^2 \partial_x^2 I_1(x) + v^3 \partial_x^2 I_0(x),\\
&+ (I_0(x) - 1)(3(I_1(x) - vI_0(x)) + v(I_0(x) - 1) - \overline{I}_1).
\end{aligned}$$

Remarque 2.5. *Les formules pour les caractéristiques en avant ont été utilisées à l'ordre 3 pour dériver un schéma semi-Lagrangien en avant (FSL-CK3) dans [35, 94].*

Lemme 2.6. *Les premiers coefficients pour les caractéristiques en arrière sont don-*

nées par

$$
\begin{aligned}
&X_b^{[0]} = x,\ X_b^{[1]} = -v,\ X_b^{[2]} = E(0,x),\\
&X_b^{[3]} = -v(I_0(x)-1) + 2(-I_1(x) + \overline{I}_1),\\
&V_b^{[0]} = v,\ V_b^{[1]} = -E(0,x),\\
&V_b^{[2]} = v(I_0(x)-1) + I_1(x) - \overline{I}_1,\\
&V_b^{[3]} = -v^2\partial_x I_0(x) + E(0,x) - \partial_x I_2(x) - v\partial_x I_1(x),\\
&X_b^{[4]} = 3\partial_x I_2 - 3E(0,x)I_0 + 6v\partial_x I_1 + 3v^2\partial_x I_0 + (3E(0,x) - 2v)(I_0 - 1) - 6(I_1 - \overline{I}_1),\\
&V_b^{[4]} = -3(I_0-1)(-I_1 + \overline{I}_1) + v(I_0-1)^2 + v^3\partial_x^2 I_0 + v^2\partial_x^2 I_1\\
&- (-\partial_x^2 I_3 + 3E(0,x)\partial_x I_1 + 3(I_0-1)(I_1 - \overline{I}_1) + (I_1 - \overline{I}_1) + 2\overline{I}_1(I_0-1))\\
&+ E(0,x)\partial_x I_1 + 3vE(0,x)\partial_x I_0 + v(\partial_x^2 I_2 - (I_0-1)I_0 - E(0,x)\partial_x I_0).
\end{aligned}
$$

Remarque 2.7. *A partir de ces expressions, on peut obtenir les schémas CK3 et CK4 pour une méthode semi-l-Lagrangienne en arrière. On voit donc que dans ce contexte, on peut utiliser ce type de schémas aussi bien en avant qu'en arrière. Le coût de calcul est similaire, les formules sont juste légèrement modifiées. On a donné la formule pour la dérivée quatrième, mais celle-ci ne sera pas utile pour obtenir les conditions d'ordre* ≤ 4.

On considère ensuite un développement de Taylor des caractéristiques numériques en avant :

$$
X_s(\Delta t,0,x,v) = \sum_{i=0}^{d} X_{s,f}^{[i]}\frac{\Delta t^i}{i!} + O(\Delta t^{d+1}),\ V_s(\Delta t,0,x,v) = \sum_{i=0}^{d} V_{s,f}^{[i]}\frac{\Delta t^i}{i!} + O(\Delta t^{d+1}), \tag{2.13}
$$

avec

$$
X_{s,f}^{[i]} = \partial_t^i X_s(0;0,x,v),\ V_{s,f}^{[i]} = \partial_t^i V_s(0;0,x,v),\ i = 0,\dots,d.
$$

Le schéma de splitting sera alors d'ordre $\geq d$, si les caractéristiques numériques en avant coincident avec les caractéristiques (continues) jusqu'à l'ordre d, c'est-à-dire

$$
X_{s,f}^{[i]} = X_f^{[i]},\ V_{s,f}^{[i]} = V_f^{[i]},\ i = 0,\dots,d. \tag{2.14}
$$

On a déja donné les expressions pour le cas continu et il reste à donner les expressions pour le cas numérique. Pour cela, on calcule le développement des caractéristiques numériques en arrière qui sont utiles pour calculer les dérivées numériques en temps du champ électrique. Enfin, on peut calculer le développement des caractéristiques numériques en avant. On donne dans les lemmes suivants les expressions des dérivées en temps des champs électriques numériques puis des caractéristiques numériques en avant.

Lemme 2.8. *Pour* $p = 0,\dots,s-1$, *les premières dérivées du champ électrique pour*

le schéma de splitting (2.11) sont données par

$$\partial_t E[g_{2p+1}](0,x) = -\sum_{k=0}^{p} a_{2k}(I_1(x)-\overline{I}_1),$$
$$\partial_t^2 E[g_{2p+1}](0,x) = \left(\sum_{k=0}^{p} a_{2k}\right)^2 \partial_x I_2(x) - 2\sum_{k=0}^{p} a_{2k}\sum_{\ell=0}^{k-1} a_{2\ell+1}E_0(x)I_0(x),$$
$$\partial_t^3 E[g_{2p+1}](0,x) = -\left(\sum_{\ell=0}^{p} a_{2\ell}\right)^3 \partial_x^2 I_3(x)$$
$$+ 6\sum_{r=1}^{p} a_{2r-1}\sum_{m=r}^{p} a_{2m}\left(\sum_{\ell=0}^{p} a_{2\ell}\partial_x(E(0,x)I_1(x)) + \sum_{\ell=0}^{r-1} a_{2\ell}(I_1(x)-\overline{I}_1 I_0(x))\right).$$

où g_{2p+1} est la fonction définie dans (2.2).

Remarque 2.9. *La quantité $\partial_t E[g_{2p+1}](0,x)$ ne donne généralement pas une approximation d'ordre 3 des champs électriques intermédiaires $E(\sum_{j=0}^{p} a_{2j+1}\Delta t, x)$. Seulement après un pas de temps entier, on obtient la bonne approximation des caractéristiques. Cela contraste avec [97] où une approximation d'ordre 3 des champs électriques intermédiaires (qui peut être obtenue à partir du Lemme 2.2) est utilisée*

Lemme 2.10. *Les dérivées des caractéristiques numériques en avant sont données par*

$$X_{f,s}^{[0]} = x,\ V_{f,s}^{[0]} = v,\ X_{f,s}^{[1]} = \sum_{\ell=0}^{s} a_{2\ell}v, \quad \text{and} \quad V_{f,s}^{[1]} = \sum_{\ell=1}^{s} a_{2\ell-1}E(0,x),$$

pour les ordres zéro et un,

$$X_{f,s}^{[2]} = 2\sum_{\ell=0}^{s} a_{2\ell}\sum_{p=1}^{\ell} a_{2p-1}E(0,x),$$
$$V_{f,s}^{[2]} = 2\sum_{\ell=1}^{s} a_{2\ell-1}\sum_{r=0}^{\ell-1} a_{2r}(v(I_0(x)-1)+\overline{I}_1 - I_1(x)),$$

pour l'ordre deux,

$$X_{f,s}^{[3]} = 6\sum_{\ell=0}^{s} a_{2\ell}\sum_{p=1}^{\ell} a_{2p-1}\sum_{r=0}^{p-1} a_{2r}(\overline{I}_1 - I_1(x) + v(I_0(x)-1)),$$
$$V_{f,s}^{[3]} = 3\sum_{\ell=1}^{s} a_{2\ell-1}\times$$
$$\left(\left(\sum_{k=0}^{\ell-1} a_{2k}\right)^2 (\partial_x I_2(x) + v^2\partial_x I_0(x) - 2v\partial_x I_1(x)) - 2\sum_{p=0}^{\ell-1} a_{2p}\sum_{r=1}^{p} a_{2r-1}E(0,x)\right),$$

pour l'ordre trois, et

$$X_{f,s}^{[4]} = 12\sum_{\ell=0}^{s} a_{2\ell}\sum_{q=1}^{\ell} a_{2q-1}\left(\sum_{k=0}^{q-1} a_{2k}\right)^2 (\partial_x I_2(x) + v^2\partial_x I_0(x) - 2v\partial_x I_1(x)) \\ - 24\sum_{\ell=0}^{s} a_{2\ell}\sum_{q=1}^{\ell} a_{2q-1}\sum_{p=0}^{q-1} a_{2p}\sum_{r=1}^{p} a_{2r-1}E(0,x),$$

$$V_{f,s}^{[4]} = 4\sum_{\ell=1}^{s} a_{2\ell-1}\left(\sum_{k=0}^{\ell-1} a_{2k}\right)^3 \left(-\partial_x^2 I_3(x) + 3\partial_x^2 I_2(x)v - 3v^2\partial_x^2 I_1(x) + v^3\partial_x^2 I_0(x)\right) \\ + 24\sum_{\ell=1}^{s} a_{2\ell-1}\sum_{r=1}^{\ell-1} a_{2r-1}\sum_{m=r}^{\ell-1} a_{2m}\sum_{q=0}^{\ell-1} a_{2q}(I_0(x)-1)(I_1(x) - vI_0(x)) \\ + 24\sum_{\ell=1}^{s} a_{2\ell-1}\sum_{r=1}^{\ell-1} a_{2r-1}\sum_{m=r}^{\ell-1} a_{2m}\sum_{q=0}^{r-1} a_{2q}(I_0(x)-1)((I_0(x)-1)v - \overline{I}_1).$$

pour l'ordre quatre.

Pour calculer les coefficients $a_j, j = 0, \ldots, 2s+1$, on doit indentifier les dérivées en temps des caractéristiques numériques en avant. A partir des Lemmes 2.4 et 2.10, on voit que le développement des caractéristiques continues et numériques sont exprimées avec les mêmes termes (qui dépendent des moments $I_k, k = 0, \ldots, 3$ de f et de $E(0,x)$). Cela permet de dériver les conditions d'ordre.
Comme dans [12], on introduit les coefficients p_j qui permettent de faire le lien avec les méthodes PRK et RKN

$$p_0 = 0,\ p_{j+1} = a_j - p_j, \quad j = 0, \ldots, 2s,$$

et

$$B_1 = \sum_{j=1}^{2s} p_j, \qquad B_2 = \sum_{j=1}^{2s}(-1)^j p_j^2,$$
$$B_{4a} = \sum_{j=1}^{2s}(-1)^j p_j^4, \quad B_{4b} = \sum_{j=1}^{s}(p_{2j}^3 + p_{2j-1}^3)\sum_{k=1}^{2j-1} p_k,$$
$$B_{4c} = \sum_{j=1}^{s}(p_{2j}^2 - p_{2j-1}^2)\left(\sum_{k=1}^{j-1} p_{2k}\sum_{\ell=1}^{2k-1} p_\ell + \sum_{k=1}^{j} p_{2k-1}\sum_{\ell=1}^{2k-1} p_\ell\right).$$

Grâce à toutes ces relations précédentes, on obtient les conditions d'ordre pour l'ordre 4 sous une forme réduite.

Théorème 2.11. *Les conditions sur $p_j, j = 1, \ldots, 2s+1$ qui assurent que le splitting en temps avec $2s+1$ étapes est d'ordre 4 (i.e. (2.14)), peut être réécrit*

$$p_{2s+1} = 0,\ B_1 = 1, \tag{2.15}$$
$$B_2 = 0, \tag{2.16}$$
$$B_{3a} = B_{3b} = 0, \tag{2.17}$$
$$B_{4a} = -4B_{4b} = 4B_{4c}, \tag{2.18}$$

à condition que l'on suppose que les fonctions entre crochets suivantes soient indépendantes : $[vI_0(x) - 1, I_1(x) - \overline{I}_1]$, $[\partial_x I_2(x) + v^2\partial_x I_0(x) - 2v\partial_x I_1(x), E(0,x)]$ *et*

$$[-\partial_x^2 I_3(x) + 3\partial_x^2 I_2(x)v - 3v^2\partial_x^2 I_1(x) + v^3\partial_x^2 I_0(x), \\ (I_0(x)-1)(I_1(x) - vI_0(x)), (I_0(x)-1)((I_0(x)-1)v - \overline{I}_1)].$$

Remarque 2.12. *Plus précisément, (2.15) corresponds à l'ordre* 1*, (2.15)-(2.16) tà l'ordre* 2 *et (2.15)-(2.16)-(2.17) à l'ordre* 3*. On obtient donc les mêmes conditions que les méthodes PRK présentées dans [12] pour l'ordre* ≤ 3 *et que RKN pour l'ordre* 4 *(pour l'ordre* ≤ 3*, les coefficients pour les méthodes PRK et RKN coincident). Grâce à la relation (2.7), on conjecture que pour l'ordre* ≥ 5*, les conditions d'ordre seront différentes de RKN (il y aura plus de coefficients, pour le système de Vlasov-Poisson).*

2.4 Résultats numériques

On considère le cas test de l'amortissement Landau non linéaire. La condition initiale est donnée par

$$f_0(x,v) = \frac{1}{\sqrt{2\pi}}\exp(-v^2/2)(1 + \alpha\cos(kx)),$$

avec

$$N_x = N_v = 1024,\ \Delta t = 0.125,\ v_{\max} = 6,\ k = 0.5,\ \alpha = 0.5,$$

Pour la discrétisation en espace, on utilise une interpolation de Lagrange d'ordre 17 en espace ($x \in [0, 2\pi/k]$) et en vitesse ($v \in [-v_{\max}, v_{\max}]$), voir *e.g.* [9,21,31,37]. On reviendra sur la discrétisation en espace par la suite. Les coefficients que l'on utilise peuvent être trouvés dans la littérature (voir *e.g.* [13], [12]) ; les premiers digits sont donnés ici pour commodité. On va aussi considérer des schémas de splitting qui commencent avec une advection en v : $g_1(x,v) = f_0(x, v - b_0\, E[g_0](x)\,\Delta t)$, and

$$\begin{aligned} g_{2j}(x,v) &= g_{2j-1}(x - b_{2j-1}\, v\,\Delta t, v), \\ g_{2j+1}(x,v) &= g_{2j}(x, v - b_{2j}\, E[g_{2j}](x)\,\Delta t), \end{aligned} \tag{2.19}$$

pour $j = 1, \ldots, 2s$. Notons qu'un tel schéma peut être mis sous la forme 2.2), en prenant $a_0 = a_{2s+2} = 0,\ a_j = b_{j-1},\ j = 1, \ldots, 2s+1$.

• **Strang (s=1)** :
$[a_0, a_2] = [0.5, 0.5],\ a_1 = 1.$
• **Strang v-x (s=1)** :
$[b_0, b_2] = [0.5, 0.5],\ b_1 = 1.$
• **Triple jump (s=3)** :
$[a_0, a_2, a_4, a_6] = [0.676, -0.176, -0.176, 0.676],$
$[a_1, a_3, a_5] = [1.351, -1.70, 1.35].$
• **Order 4 (s=6)** :
$[b_0, b_2, \ldots, b_{12}] = [0.0830, 0.396, -0.0391, 0.120, -0.0391, 0.396, 0.0830],$
$[b_1, b_3, \ldots, b_{11}] = [0.245, 0.605, -0.350, -0.350, 0.605, 0.245].$
• **Order 6 (s=11)** :
$[b_0, b_2, \ldots, b_{22}] = [0.0415, 0.198, -0.04, 0.0753, -0.0115, 0.237,$
$0.237, -0.0115, 0.0753, -0.04, 0.198, 0.0415],$
$[b_1, b_3, \ldots, b_{21}] = [0.123, 0.291, -0.127, -0.246, 0.357, 0.205,$
$0.357, -0.246, -0.127, 0.291, 0.123].$

Pour ces différents splittings, on représente sur les Figures 3.1 et 3.2 l'évolution en temps de l'énergie électrique $\mathcal{E}_e(t) = \int_0^{2\pi/k} E(t,x)^2 dx$, l'énergie totale $\mathcal{E}(t)$ définie par

$$\mathcal{E}(t) = \int_{\mathbb{R}} \int_0^{2\pi/k} v^2 f(t,x,v) dx dv + \mathcal{E}_e(t),$$

et la norme $L^p, p = 1, 2$ de f. Rappelons que l'énergie totale et les normes L^p sont des quantités conservées du modèle. Le diagnostic de l'énergie électrique ne présente pas de différences significatives. Il est en de même pour les normes L^p, la conservation dépendant essentiellement de la discrétisation spatiale et en vitesse. Néanmoins, on voit clairement sur la Figure 3.1 l'avantage d'utiliser des schémas d'ordre élevé en temps pour la conservation de l'énergie ; cela a déjà été mis en valeur dans [112], mais les coefficients optimisés de Blanes et Moan [13] donnent encore de meilleurs résultats. Rappelons [13] qu'il ne faut pas toujours chercher à minimiser le nombre d'étapes, mais plutôt essayer d'éviter d'avoir des pas de temps négatifs trop importants (ceux-ci apparaissent automatiquement pour un ordre ≥ 3). Sur ces figures, on représente aussi le splitting de Strang en divisant le temps par 20, ce qui donne un temps de simulation similaire (même plus long) que le cas de l'ordre 6 avec $s = 11$. On voit alors que pour un temps de simulation donnée, le splitting d'ordre élevé avec les coefficients optimisés donne le meilleur résultat.

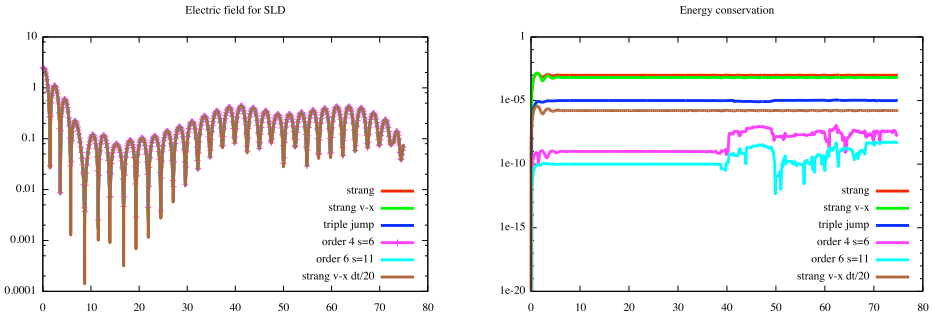

FIGURE 3.1 – Evolution en temps de l'énergie électrique (à gauche) et de l'énergie totale (à droite), pour différents algorithmes de splitting en temps

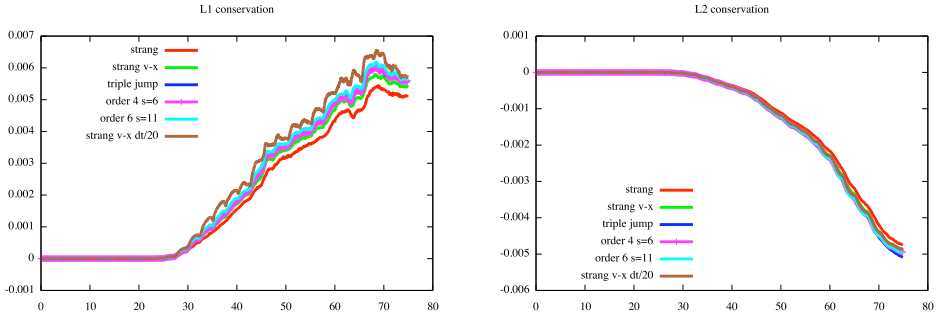

FIGURE 3.2 – Evolution en temps de la norme L^1 de f (à gauche) et de la norme L^2 de f (à gauche), pour différents algorithmes de splitting en temps.

Chapitre 4

Le transport linéaire

On regroupe des résultats issus de [21, 28, 31, 32]. Dans [31], il a été remarqué que dans le cadre du transport linéaire (l'advection constante) avec conditions limites périodiques, le schéma PFC sans limiteur [41] est équivalent à l'utilisation de l'interpolation de Lagrange de degré 3, et aussi que l'interpolation classique par splines cubiques a son analogue conservatif, la méthode PSM développée plus récemment dans [114]. Ce lien est exprimé ici dans une formulation abstraite [1]. On donne aussi des estimations utiles pour une analyse de convergence, qui sera l'objet du chapitre suivant. On fait également le lien avec d'autres types de méthodes (volumes finis, intégrateurs exponentiels). Des applications numériques à Vlasov-Poisson viennent illustrer le propos.

1 Introduction

Grâce à une discrétisation en temps par splitting directionnel, on se ramène pour le système de Vlasov-Poisson à devoir résoudre des équations d'advection linéaire (2.1). Cheng et Knorr [22] ont utilisé des splines cubiques qui offrent un bon compromis entre coût et précision. Les splines cubiques ont été remis au goût du jour dans un cadre général avec Eric Sonnendrücker [105]. Citons quelques travaux sur le sujet dans ce contexte : Fourier-Hermite [100], semi-Lagrangien (SL) avec interpolation d'Hermite [82, 83], la méthode SL PFC (Positive Flux Conservative) [41], SL sur maillage non structuré [10], SL en avant [34], SL WENO [89], SL en Galerkin Discontinu [32, 90, 97].
Soit $a \in \mathbb{R}$. On cherche à résoudre numériquement l'équation d'advection linéaire

$$\partial_t f(t,x) + a\partial_x f(t,x) = 0, x \in [0,1],\ t \in \mathbb{R},\ f(0,x) = f_0(x), x \in [0,1],$$

avec f_0 une donnée 1-périodique. La solution est donnée par

$$f(t,x) = f_0(x - at).$$

Maillage uniforme On considère un entier $N \in \mathbb{N}^*$ et un nombre $\Delta t > 0$. On définit

$$x_j = jh,\ h = 1/N,\ t_n = n\Delta t,\ j \in \mathbb{R},\ n \in \mathbb{R}.$$

1. En un certain sens, on rejoint les exigences d'une interface [20] pour une bibliothèque telle que SELALIB [101] : analyse de dépendances et de fonctionnalités pour la structuration des routines.

L'entier N permet de fixer une discrétisation en espace et le nombre réel Δt permet de fixer la discrétisation en temps.

Discrétisation de la fonction On introduit des valeurs ponctuelles

$$f_{n,j} \simeq f(t_n, x_j),\ j = 0, \dots, N-1$$

et des valeurs moyennes :

$$\bar{f}_{n,j} \simeq \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} f(t_n, x) dx,\ j = 0, \dots, N-1.$$

On peut aussi avoir plusieurs degrés de liberté dans chaque maille $[x_{j-1/2}, x_{j+1/2}]$, pour $j = 0, \dots, N-1$. On utilise la notation

$$f_{n,k,j} \simeq f(t_n, x_{k,j}),\ k = 0, \dots, N-1,\ j = 0, \dots, d.$$

avec

$$x_{k,j} = x_{k-1/2} + \alpha_j h,\ 0 \leq \alpha_0 < \cdots < \alpha_d < 1.$$

On notera aussi

$$f_{n,k} = (f_{n,k,0}, \dots, f_{n,k,d}) \in \mathbb{R}^{d+1}$$

Il se peut aussi que les valeurs $f_{n,k,j}$ ne correspondent pas forcément à des valeurs ponctuelles, mais plus généralement à des degrés de liberté.

2 Formulations de schémas semi-Lagragiens

2.1 Principe d'une méthode semi-Lagrangienne

Il s'agit de mettre à jour les degrés de liberté (valeurs ponctuelles/moyennes...). On introduit un opérateur de transport, pour $\alpha \in \mathbb{R}$,

$$\mathcal{T}_\alpha : (\mathbb{R}^{d+1})^{\mathbb{Z}} \to (\mathbb{R}^{d+1})^{\mathbb{Z}},$$

en considérant qu'il y a $d+1$ inconnues dans chaque maille. Comme on considère ici des données périodiques, on supposera en général que les suites $g \in (\mathbb{R}^{d+1})^{\mathbb{Z}}$ vérifient

$$g_{j+N} = g_j,\ j \in \mathbb{Z}.$$

Dans ce cadre général, une méthode semi-Lagrangienne s'écrit alors

$$(f_{n+1,j})_{j\in\mathbb{Z}} = \mathcal{T}_\alpha (f_{n,j})_{j\in\mathbb{Z}},$$

avec $\alpha = -\frac{a\Delta t}{h}$.
Notons que cette forme est valable, car on est dans le cadre d'un maillage uniforme et d'une advection constante.
Si α est un entier, on impose généralement que la transformation ne soit qu'un *shift* :

$$\mathcal{T}_\alpha (g_j)_{j\in\mathbb{Z}} = (g_{j+\alpha})_{j\in\mathbb{Z}}.$$

Il suffit alors de définir l'opérateur pour $0 < \alpha < 1$ (pour $\alpha = 0$, on prend l'application identité). En effet, on peut toujours écrire $\alpha = \alpha_0 + \tilde{\alpha}$, avec α_0 entier et $0 \leq \tilde{\alpha} < 1$ et

$$\mathcal{T}_\alpha(g_j)_{j\in\mathbb{Z}} = \mathcal{T}_{\tilde{\alpha}}(g_{j+\alpha_0})_{j\in\mathbb{Z}}.$$

On utilisera aussi la notation

$$(g_{j+\alpha})_{j\in\mathbb{Z}} = \mathcal{T}_\alpha(g_j)_{j\in\mathbb{Z}}.$$

Toutes ces définitions restent valables pour les suites périodiques (une fois qu'on a défini $g_{j+\alpha}$, pour $j = 0, \ldots, N-1$, on complète par périodicité).

2.2 Valeurs ponctuelles

On considère ici que les inconnues sont des valeurs ponctuelles. On cherche alors une reconstruction g 1-périodique telle que

$$g(x_j) = g_j,\ j = 0, \ldots, N-1.$$

A partir de cette reconstruction, on définit pour $0 < \alpha < 1$

$$(g_{j+\alpha})_{j\in\mathbb{Z}} = \mathcal{T}_\alpha(g_j)_{j\in\mathbb{Z}},$$

par

$$g_{j+\alpha} = g(x_{j+\alpha}),\ j \in \mathbb{Z}.$$

Remarquons que cette définition est consistante avec la périodicité et que l'on a un degré de liberté par maille $[x_{j-1/2}, x_{j+1/2}]$, $j = 0, \ldots, N-1$, c'est à dire que l'on considère le cas scalaire $d = 0$. On a défini ici l'opérateur de transport $\mathcal{T}_\alpha$ pour $0 < \alpha < 1$. Cette définition reste valable pour $\alpha \in \mathbb{R}$.
Remarquons aussi que si l'on connait un opérateur $\mathcal{T}_\alpha$ pour $0 < \alpha < 1$ agissant sur des données périodiques, alors on peut définir une reconstruction g 1-périodique, par la relation

$$g(x_{j+\alpha}) = g_{j+\alpha},\ j \in \mathbb{Z},\ 0 < \alpha < 1.$$

A partir de cette reconstruction, on retrouve le même opérateur $\mathcal{T}_\alpha$ qu'on avait au départ.

2.3 Valeurs moyennes

On considère ici que les inconnues sont des valeurs moyennes. On cherche alors une reconstruction g 1-périodique satisfaisant

$$\frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} g(x)dx = \overline{g}_j,\ j = 0, \ldots, N-1.$$

A partir de cette reconstruction, on définit pour $0 < \alpha < 1$

$$(\overline{g}_{j+\alpha})_{j\in\mathbb{Z}} = \overline{\mathcal{T}}_\alpha(\overline{g}_j)_{j\in\mathbb{Z}},$$

par

$$\overline{g}_{j+\alpha} = \frac{1}{h}\int_{x_{j+\alpha-1/2}}^{x_{j+\alpha+1/2}} g(x)dx,\ j \in \mathbb{Z}.$$

Remarquons que cette définition est consistante avec la périodicité et que l'on est toujours dans le cas scalaire. La méthode semi-Lagrangienne est bien conservative, au sens où

$$\sum_{j=0}^{N-1} \overline{g}_{j+\alpha} = \sum_{j=0}^{N-1} \overline{g}_j.$$

2.4 D'une reconstruction à l'autre

A partir d'une reconstruction par valeurs moyennes, on obtient un opérateur $(\overline{\mathcal{T}}_\alpha)_{0<\alpha<1}$, que l'on peut utiliser pour faire une reconstruction à partir de valeurs ponctuelles

$$(\mathcal{T}_\alpha)_{0<\alpha<1} = F^{M\to P}((\overline{\mathcal{T}}_\alpha)_{0<\alpha<1}).$$

Réciproquement, avec une reconstruction à partir de valeurs ponctuelles, donné par $\mathcal{T}_\alpha$, avec $0 < \alpha < 1$ qui agit sur les suite périodiques, on peut obtenir une reconstruction g pour les valeurs moyennes et donc un opérateur

$$(\overline{\mathcal{T}}_\alpha)_{0<\alpha<1} = F^{P\to M}((\mathcal{T}_\alpha)_{0<\alpha<1}).$$

Pour cela, on va reconstruire g à l'aide d'une primitive. En considérant

$$G_{j-1/2+\alpha} = \frac{1}{h}\int_{x_{-1/2}}^{x_{j-1/2+\alpha}} g(x)dx,$$

le schéma se réécrit

$$\overline{g}_{j+\alpha} = G_{j+1/2+\alpha} - G_{j-1/2+\alpha},\ j = 0,\dots,N-1,\ 0<\alpha<1,$$

et il suffit de définir les $G_{j-1/2+\alpha},\ j = 0,\dots,N$. Pour cela, on définit d'abord

$$M = \frac{1}{N}\sum_{k=0}^{N-1} \overline{g}_k$$

et

$$\tilde{g}_j = \overline{g}_j - M,\ j = 0,\dots,N-1,\ \tilde{g}_{j+N} = \tilde{g}_j,\ j \in \mathbb{Z}.$$

Notons que cette suite est construite pour que

$$\sum_{k=0}^{N-1} \tilde{g}_k = 0.$$

On définit alors la suite $(\tilde{G}_{j-1/2})_{j\in\mathbb{Z}}$ par

$$\tilde{G}_{-1/2} = 0,\ \tilde{G}_{j-1/2} = \sum_{k=0}^{j-1} \tilde{g}_k,\ j = 1,\dots N,\ \tilde{G}_{j-1/2+N} = \tilde{G}_{j-1/2},\ j \in \mathbb{Z}.$$

On utilise ensuite l'opérateur d'interpolation agissant sur la suite périodique $(\tilde{G}_{j-1/2})_{j\in\mathbb{Z}}$

$$\mathcal{T}_\alpha(\tilde{G}_{j-1/2})_{j\in\mathbb{Z}} = (\tilde{G}_{j-1/2+\alpha})_{j\in\mathbb{Z}},\ 0<\alpha<1,$$

On peut alors enfin définir

$$G_{j-1/2+\alpha} = \tilde{G}_{j-1/2+\alpha} + jM,\ j = 0,\dots,N,\ 0<\alpha<1.$$

ce qui donne en fait

$$\overline{g}_{j+\alpha} = \tilde{G}_{j+1/2+\alpha} - \tilde{G}_{j-1/2+\alpha} + M,\ j = 0,\dots,N-1,\ 0<\alpha<1.$$

2.5 Lien entre les deux reconstructions

A partir d'un opérateur $(\mathcal{T}_\alpha)_{0<\alpha<1}$ défini pour mettre à jour des valeurs ponctuelles. On peut définir un nouvel opérateur $(\overline{\mathcal{T}}_\alpha)_{0<\alpha<1}$ pour mettre à jour des valeurs moyennes. On peut alors réutiliser cet opérateur pour mettre à jour des valeurs ponctuelles. On peut donc se demander quand est-ce que l'on retombe sur le même opérateur que l'on a au départ, c'est-à-dire quand est-ce que

$$(\mathcal{T}_\alpha)_{0<\alpha<1} = F^{M\to P}F^{P\to M}((\mathcal{T}_\alpha)_{0<\alpha<1}). \tag{2.1}$$

En reprenant les notations précédentes, cela se traduit par

$$(\mathcal{T}_\alpha(g_j)_{j\in\mathbb{Z}})_k = (\mathcal{T}_\alpha(\tilde{G}_{j-1/2})_{j\in\mathbb{Z}})_{k+1} - (\mathcal{T}_\alpha(\tilde{G}_{j-1/2})_{j\in\mathbb{Z}})_k + M,\ k\in\mathbb{Z},$$

pour tout $0<\alpha<1$ et toute suite N-périodique $(g_j)_{j\in\mathbb{Z}}$. Les autres quantités sont définies par :

$$M = \frac{1}{N}\sum_{k=0}^{N-1} g_k$$

et

$$\tilde{G}_{-1/2} = 0,\ \tilde{G}_{j-1/2} = \sum_{k=0}^{j-1}\tilde{g}_k,\ j=1,\dots N,\ \tilde{G}_{j-1/2+N} = \tilde{G}_{j-1/2},\ j\in\mathbb{Z},$$

avec

$$\tilde{g}_j = g_j - M,\ j=0,\dots,N-1,\ \tilde{g}_{j+N} = \tilde{g}_j,\ j\in\mathbb{Z}.$$

En particulier, si l'on considère une suite $(G_{j+1/2})_{j\in\mathbb{Z}}$ N-périodique arbitraire, on doit avoir

$$(\mathcal{T}_\alpha(G_{j+1/2}-G_{j-1/2})_{j\in\mathbb{Z}})_k = (\mathcal{T}_\alpha(G_{j-1/2})_{j\in\mathbb{Z}})_{k+1} - (\mathcal{T}_\alpha(G_{j-1/2})_{j\in\mathbb{Z}})_k,\ k\in\mathbb{Z},$$

et la préservation des constantes :

$$\mathcal{T}_\alpha M = M,\ M\in\mathbb{R}.$$

2.6 Schémas homogènes

Un schéma homogène consiste à définir $\mathcal{T}_\alpha$ sous la forme

$$(\mathcal{T}_\alpha(f_j)_{j\in\mathbb{Z}})_j = f_{j+\alpha} = \sum_{k=0}^{N-1} L_{k,N}(\alpha)f_{j+k},\ j\in\mathbb{Z},\ 0<\alpha<1. \tag{2.2}$$

pour une suite $(f_j)_{j\in\mathbb{Z}}$ $N-$périodique. L'opérateur de transport agit ainsi sur les données à chaque endroit de la même manière. Si l'on suppose que les termes $L_{k,N}(\alpha),\ k=0,\dots,N-1$ vérifient

$$\sum_{k=0}^{N-1} L_{k,N} = 1,$$

on a bien la propriété d'équivalence (2.1). Soulignons aussi que ces schémas peuvent être implémentés en Fourier.

Remarque 2.1. *Les schémas (2.2) se généralisent au cas vectoriel en considérant* $f_j\in\mathbb{R}^{d+1}$ *et* $L_{k,N}(\alpha)\in\mathcal{M}_{d+1,d+1}(\mathbb{R})$.

3 Exemples de schémas

Par la suite, nous allons donner des exemples de schémas.

3.1 Lagrange LAG$(2d+1)$

On définit le schéma par

$$f_{j+\alpha} = \sum_{\ell=-d}^{d} L_\ell(\alpha) f_{j+\ell},\ L_\ell(\alpha) = \prod_{s=-d,\ s\neq\ell}^{d} \frac{\alpha - s}{\ell - s},\ 0 \leq \alpha < 1.$$

Pour $d = 0$ (sous CFL ≤ 1), il s'agit du schéma upwind.
Pour $d = 1$, il s'agit du schéma PFC [41] sans limiteur, aussi connu et développé dans [62].
De manière générale, il s'agit d'un schéma de Strang (sous CFL ≤ 1), voir [36, 37]. Sans la contrainte CFL, on peut parler de schémas de Strang shiftés. Ces schémas sont compacts et explicites.

3.2 Splines SPL(d)

On définit les fonctions B-splines par

$$B_j(x) = \int_{\mathbb{R}} B_{j-1}(t) B_0(x-t)dt,\ B_0(x) = 1_{[-1/2,1/2]}(x),\ j = 0, \ldots, d,$$

et les coefficients de splines $\eta_0, \ldots, \eta_{N-1}$ par

$$\sum_{\ell\in\mathbb{Z}} \eta_{\ell \mod N} B_d(j-\ell) = f_j,\ j = 0, \ldots, N-1.$$

On a alors

$$f_{j+\alpha} = \sum_{\ell\in\mathbb{Z}} \eta_{\ell \mod N} B_d(\alpha - \ell),\ 0 \leq \alpha < 1.$$

En général, on utilise $d = 3$, ce qui correspond aux splines cubiques classiques (formulation par point), mais aussi au schéma PSM [114] (formulation par volume). Notons que l'on peut utiliser une implémentation type FFT dans ce cadre de maillage uniforme avec des conditions périodiques. Cela est aussi valable pour l'interpolation de Lagrange d'ailleurs.
L'implémentation peut se faire en interpolant sur la primitive ; afin de garder le cas d'une interpolation périodique, on considère $f_j - \frac{1}{N}\sum_k f_k$ comme données.
Notons que l'on peut adapter des versions locales pour les splines [29], ce qui peut avoir son importance pour la parallélisation. On peut combiner aussi avec des techniques de quasi-interpolation, voir [93]. A titre d'exemple [2], voici une implémentation possible pour les splines cubiques. On suppose $0 \leq \alpha < 1$, pour donner les formules.

– Calcul d'une approximation locale des coefficients de splines

$$\omega_k = \sum_{\ell=-p/2}^{p/2-1} a_{\ell,p} f^{\text{old}}_{k+\ell},\ k = 0, \ldots, N-1$$

2. Un code Vlasov-Poisson $4D$ parallèle avec splitting dans chaque direction a été développée en collaboration avec E. Violard

– Calcul d'une première nouvelle valeur

$$f_k^{\text{new}} = \frac{(1-\alpha)^3}{6}\omega_{k-1} + \left(\frac{2}{3} - \alpha^2 + \frac{\alpha^3}{2}\right)\omega_k + \left(\frac{2}{3} - (1-\alpha)^2 + \frac{(1-\alpha)^3}{2}\right)\omega_{k+1} + \frac{\alpha^3}{6}\omega_{k+2}$$

– Etape de correction

$$f_{k+\alpha} = f_k^{\text{new}} + (1-\alpha)\left(f_k^{\text{old}} - \frac{\omega_{k-1} + 4f\omega_k + \omega_{k+1}}{6}\right) + \alpha\left(f_{k+1}^{\text{old}} - \frac{\omega_k + 4f\omega_{k+1} + \omega_{k+2}}{6}\right).$$

L'étape de correction (lift) permet de retrouver la condition d'interpolation $f_{k+0} = f_k^{\text{old}}$ qui n'est sinon pas vérifiée dans ce cadre de quasi-interpolation. Pour trouver les coefficients $a_{\ell,p}$, on résoud le système

$$\frac{a_{k-1,p} + 4a_{k,p} + a_{k+1,p}}{6} = \delta_k^0,\ k = 0, \dots, p-1,\ a_{\ell+p,p} = a_\ell,\ \ell \in \mathbb{Z},$$

ce qui assure que le système de splines est exactement résolu lorsque $N = p$, i.e. que l'on retombe sur SPL(3), et l'étape de correction est alors superflue dans ce cas. Notons que la méthode développée dans [93] correspond à choisir l'approximation suivante pour les coefficients de splines :

$$\omega_k = \frac{4}{3}f_k^{\text{old}} - \frac{1}{6}\left(f_{k-1}^{\text{old}} + f_{k+1}^{\text{old}}\right),\ k = 0, \dots, N-1.$$

Des résultats numériques sont donnés pour le cas test Landau non linéaire, avec $N_x = N_v = 128$, $\Delta t = 0.1$ et $v_{\max} = 6.5$ sur les Figures 4.1, 4.2 et 4.3. On remarque des différences pour $p = 16$ et $p = 18$ en temps long, lorsqu'on n'applique pas la procédure de correction : cela se voit surtout sur l'énergie totale. La quasi interpolation utilisée dans [93] donne une meilleure conservation de la positivité, mais au prix d'une diffusion bien plus importante (norme L^2). Pour $p = 32$, il n'est plus nécessaire de rajouter l'étape de correction. La masse est bien conservée, même si les erreurs d'arrondi semblent s'accumuler ici. Enfin, le choix de p doit être suffisament grand pour que le schéma ne soit pas trop diffusif.

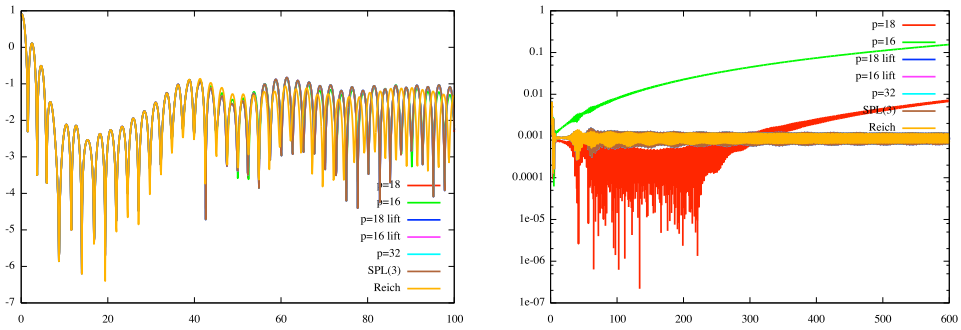

FIGURE 4.1 – Évolution en temps de l'énergie électrique (à gauche) et de l'énergie totale (à droite), pour $p = 16, 18$ avec ou sans correction (correction=lift), $p = 32$, méthode SPL(3) et quasi-interpolation de Reich avec correction

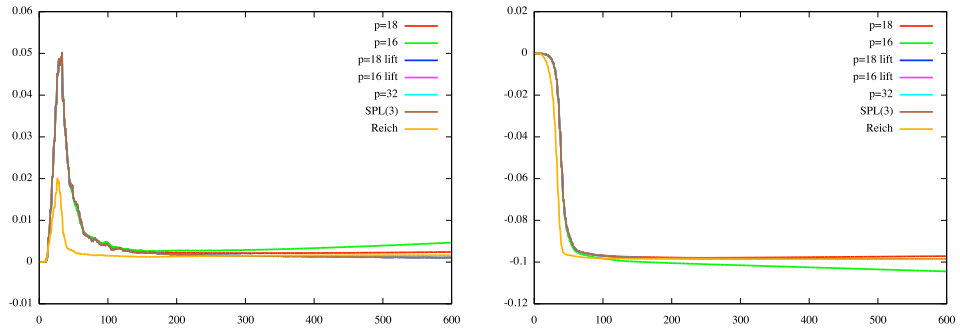

FIGURE 4.2 – Évolution en temps de la norme L^1 de f (à gauche) et de la norme L^2 de f (à gauche), pour $p = 16, 18$ avec ou sans correction (correction=lift), $p = 32$, méthode SPL(3) et quasi-interpolation de Reich avec correction

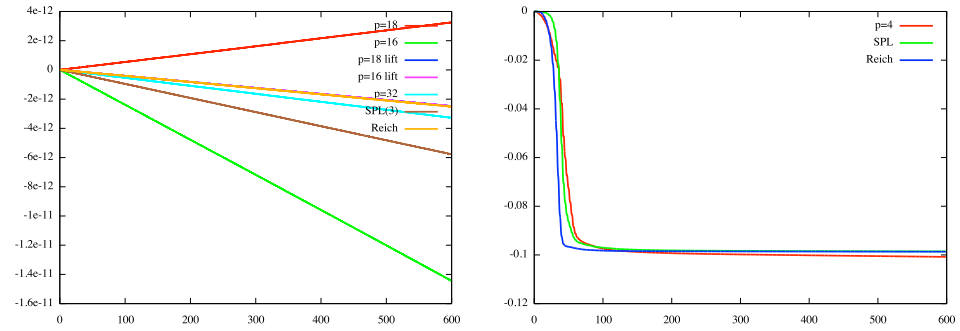

FIGURE 4.3 – Évolution en temps de la masse de f (à gauche) pour $p = 16, 18$ avec ou sans correction (correction=lift), $p = 32$, méthode SPL(3) et quasi-interpolation de Reich avec correction ; norme L^2 de f (à droite), pour $p = 4$ avec correction, SPL(3) et quasi-interpolation de Reich avec correction

3.3 Hermite

La reconstruction d'Hermite s'écrit

$$f_{j+\alpha} = f_j + f'_j\alpha + (f_{j+1} - f_j - f'_j)\alpha^2 + (f'_{j+1} + f'_j - 2(f_{j+1-f_j}))\alpha^2(\alpha - 1),\ 0 \leq \alpha < 1.$$

La dérivée doit être estimée. On aboutit à PPM0 (on parle aussi de splines de Catmull-Rom), PPM1 [24], PPM2 [25], en utilisant la formule d'ordre 2, 4 ou 6 :

$$f'_j \simeq \frac{f_{j+1} - f_{j-1}}{2h}, f'_j \simeq \frac{1}{12h}(f_{j-2} - f_{j+2} + 8(f_{j+1} - f_{j-1})),$$

$$f'_j \simeq \frac{1}{60h}(f_{j+3} - f_{j-3} + 9(f_{j-2} - f_{j+2}) + 45(f_{j+1} - f_{j-1})).$$

En différenciant la dérivée à gauche f'_{j^-} et à droite f'_{j^+}, on peut retrouver le schéma LAG(3) et aussi le schéma SPL(3) en imposant la continuité de la dérivée seconde, par résolution du système :

$$\frac{h}{3}\left(f'_{j-1} + 4f'_j + f'_{j+1}\right) = f_{j+1} - f_{j-1},$$

qui correspond à une approximation de Simpson de $\int_{x_{j-1}}^{x_{j+1}} f(x)dx$.
On peut aussi adapter ce type de formulation pour la parallélisation. Cette formulation présente d'ailleurs l'avantage fondamental de n'avoir qu'à communiquer la contribution des dérivées aux interfaces des processeurs, comme cela a été effectué dans [29] et appliqué avec succès sur un grand nombre de processeurs. Notons aussi dans la même veine, un algorithme local pour calculer les coefficients de splines implémenté dans SELALIB [101, 107], qui se base sur [111].

3.4 Galerkin discontinu GD(d+1)

On définit les $(d+1)$ points et poids de Gauss sur l'intervalle $[0,1]$: $\alpha_k, \omega_k,\ k = 0,\dots,d$. On a alors $f_{j,k} \simeq f(x_{j-1/2}+\alpha_k h)$ et $f_j = (f_{j,0},\dots,f_{j,d})$. Le schéma s'écrit alors pour $\ell = 0,\dots,d$,

$$\omega_\ell (f_{j+\alpha})_\ell = \sum_{\ell'=0}^{d} f_{j,\ell'} \int_\alpha^1 \phi_{\ell'}(s)\phi_\ell(s-\alpha)ds + \sum_{\ell'=0}^{d} f_{j+1,\ell'} \int_0^\alpha \phi_{\ell'}(s)\phi_\ell(s+1-\alpha)ds,$$

avec

$$\phi_\ell(x) = \prod_{j\neq\ell=0}^{d} \frac{x-\alpha_j}{\alpha_\ell - \alpha_j},\ \ell = 0,\dots,d. \tag{3.1}$$

Cette méthode a été développée dans [67, 90, 97] et [32].

4 Stabilité et convergence pour l'advection linéaire

4.1 Schémas de Lagrange

On considère une reconstruction de Lagrange LAG(2d+1) ; on utilise la définition (3.1) pour la norme discrète. On a alors le résultat suivant.

Lemme 4.1. *On considère l'advection linéaire $\partial_t f(t,x)+a\partial_x f(t,x) = 0$. En écrivant*

$$x_j - a\Delta t = x_{j+r} + \alpha\Delta x,\ 0 \le \alpha < 1,$$

l'erreur pour LAG(2d+1) avec n étapes, $n\Delta t \le T$ satisfait

$$\|(f(t_n,x_j) - f_j^n)_j\|_2 \le C_d T \frac{(1-\alpha)\alpha\Delta x^{2d+2}}{\Delta t} \|u_0^{(2d+2)}\|_{L^2}.$$

avec

$$C_d = O\left(\frac{(d+1)!d!}{(2d+2)!(2d+2)^{3/4}}\right) = O\left(\frac{1}{2^{2d}d^{1/4}}\right).$$

Remarque 4.2.
- *Pour la preuve, on consultera [21]. Une stratégie consiste à obtenir une estimation fine sur le noyau de Fourier. On peut aussi utiliser un résultat sur la norme maximum des B-splines [75] :*
  - *utilisation de l'erreur sous forme de noyau de B-splines*
  - *estimation fine grâce à l'uniformité de la grille*
- *Si l'on ne cherche pas à estimer finement la constante par rapport à d, la preuve est simple.*

- *Cette estimation n'est pas valide pour tous les schémas, par exemple, si on utilise le schéma de Lax Friedrichs, il n'y a pas de convergence pour Δt trop petit !*
- *On utilise la stabilité en norme L^2 du schéma, en effectuant une analyse de Von Neuman : on pourra consulter [9, 36–38, 40, 51, 52, 57, 109] sur cette question.*

L'estimation précédente implique alors que l'erreur vérifie

$$\text{erreur} \leq C \min(1, \frac{\Delta t}{\Delta x})\Delta x^{2d+2}.$$

4.2 Les splines cubiques

Si l'on considère une reconstruction par splines cubiques (PSM), on a toujours la stabilité L^2 (voir [9, 98]). On peut voir que l'erreur s'écrit, pour une représentation de type Hermite

$$\text{erreur} \leq (\Delta x)^4\alpha^2(1-\alpha)^2\frac{\max_\xi |f^{(4)}(\xi)|}{4!} + \Delta x \max|f_j' - f'(x_j)|\alpha(1-\alpha),$$

et pour les splines cubiques, on obtient

$$\max|f_j' - f'(x_j)| \leq C \max_\xi |f^{(5)}(\xi)|(\Delta x)^4,$$

ce qui donne

$$\text{erreur} \leq \min\left(1, \frac{\Delta t}{\Delta x}\right)^2 O\left(\frac{\Delta x^4}{\Delta t}\right) + \min\left(1, \frac{\Delta t}{\Delta x}\right) O\left(\frac{\Delta x^5}{\Delta t}\right).$$

Ainsi par exemple, pour $\Delta t = \Delta x^2$, on obtient erreur $\leq O(\Delta x^4)$, alors que pour LAG(3), on a seulement erreur $\leq O(\Delta x^3)$.

4.3 La stabilité pour les interpolettes

Les interpolettes ont été utilisés dans des schémas semi-Lagrangiens adaptatifs [8]. On note ici un résultat de stabilité en norme L^2, dans le cas d'une reconstruction uniforme. Avec

$$m(\omega) = \sum_n h_n e^{in\omega},\ \phi(2x) = \sum_n h_n \phi(2x-n),\ \phi_{j,k}(x) = \phi(2^{-j}x - k),$$

on définit

$$f_{k+\alpha} := \sum_\ell \phi_{J,\ell}(\alpha) f_{k+\ell},\ 0 \leq \alpha \leq 1.$$

Le lemme suivant permet alors d'obtenir la stabilité (voir aussi [40]).

Lemme 4.3.
Soit $J \in \mathbb{N}^*$, $\theta \in \mathbb{R}$.
Supposons que

$$m(\omega) + m(\omega+\pi) = 1,\ m(\omega) \geq 0, \omega \in \mathbb{R}.$$

Alors on a

$$\Big|\sum_{k\in\mathbb{Z}} \phi_{J,k}(\alpha)e^{ik\omega}\Big| \leq 1,\ \textit{pour } |\alpha| \leq 1.$$

Démonstration. On définit

$$\Phi(\alpha) = \sum_k e^{i\omega(k+\alpha)}\varphi(k+\alpha).$$

Il suffit alors de montrer que $|\Phi(\alpha)| \leq 1$, pour $\alpha = 0, 1/2, 1/4, 3/4, \ldots$.
Pour $n, p \in \mathbb{N}$, on calcule

$$\begin{aligned}
&\sum_{k=0}^{2^n-1} e^{i2\pi kp/2^n}\Phi(k/2^n) = \sum_{k=0}^{2^n-1}\sum_{\ell} e^{i(\omega+2p\pi)(\ell+k/2^n)}\varphi(\ell+k/2^n)\\
&= \sum_{k=0}^{2^n-1}\sum_{\ell,\ell_1,\ldots,\ell_n} h_{\ell_1}\ldots h_{\ell_n} e^{i(\omega+2p\pi)(\ell+k/2^n)}\varphi(2^n\ell + k - \sum_{s=1}^{n} 2^{n-s}\ell_s)\\
&= \sum_{k}\sum_{\ell_1,\ldots,\ell_n} h_{\ell_1}\ldots h_{\ell_n} e^{i(\omega+2p\pi)k/2^n}\delta^k_{\sum_{s=1}^n 2^{n-s}\ell_s} = 2^n\prod_{s=1}^{n} m((\omega+2p\pi)/2^s)
\end{aligned}$$

La transformée de Fourier inverse donne alors

$$\Phi(p/2^n) = \sum_{k=0}^{2^n-1} e^{-i2\pi kp/2^n}\prod_{s=1}^{n} m((\omega+2k\pi)/2^s),$$

et donc

$$|\Phi(p/2^n)| \leq \sum_{k=0}^{2^n-1}\prod_{s=1}^{n} m((\omega+2k\pi)/2^s) = \Phi(0) = 1,$$

puisque $m(\omega) \geq 0$, et cela donne le résultat.

□

Dans le cas des interpolettes d'ordre $2d+1$, avec $d \in \mathbb{N}$ le symbole $m(\omega)$ est défini par

$$m_d(\omega) = \cos^2(\omega/2)^N P_d(\sin^2(\omega/2)),$$

avec

$$P_d(x) = \sum_{n=0}^{d}\binom{2d+1}{d+n}x^n(1-x)^{d-n},$$

On note en particulier que l'on a bien $m_d(\omega) \geq 0$. L'autre condition est également satisfaite, puisque ϕ est interpolante, i. e. $\phi(k) = \delta^0_k$, voir [23].

4.4 Stabilité de schémas homogènes vectoriels

Comme on l'a vu, on analyse généralement la stabilité de schémas homogènes en utilisant l'analyse de Von Neuman. Cette analyse se généralise au cas vectoriel.
On écrit $f^{n+1} = Af^n$ où la matrice $A \in \mathcal{M}_{d+1}(\mathbb{R})$ est donnée par

$$A = \begin{pmatrix}
A_0 & A_1 & A_2 & \ldots & A_{N-1}\\
A_{N-1} & A_0 & A_1 & \ddots & \vdots\\
\vdots & \ddots & \ddots & \ddots & \\
A_2 & \ldots & A_{N-1} & A_0 & A_1\\
A_1 & A_2 & \ldots & & A_0
\end{pmatrix}.$$

On introduit alors la transformée de Fourier adaptée aux coefficients vectoriels de $\mathbb{R}^{d+1}$

$$\begin{array}{rcl} \left(\mathbb{R}^{d+1}\right)^N & \longrightarrow & \left(\mathbb{C}^{d+1}\right)^N \\ (f_0, f_1, ..., f_{N-1}) & \mapsto & (\hat{f}_0, \hat{f}_1, ..., \hat{f}_{N-1}), \end{array}$$

avec $\hat{f}_k = \sum_{j=0}^{N-1} e^{-2i\pi kj/N} f_j \in \mathbb{C}^{d+1}$. Similairement, on définit la transformée de Fourier $\hat{A}_k$ de la matrice A : $\hat{A}_k = \sum_{j=0}^{N-1} e^{-2i\pi kj/N} A_j \in \mathcal{M}_{d+1}(\mathbb{C})$. Avec ces notations, on peut exprimer l'évolution des modes de Fourier de la solution

$$\begin{aligned} \hat{f}_k^{n+1} &= \sum_{\ell=0}^{N-1} f_\ell^{n+1} e^{-2i\pi\ell k/N} = \sum_{\ell=0}^{N-1}\sum_{j=0}^{N-1} A_{\ell-j} f_j^n e^{-2i\pi\ell k/N} \\ &= \sum_{j=0}^{N-1} \left(\sum_{\ell=0}^{N-1} A_{\ell-j} e^{-2i\pi(\ell-j)k/N} \right) e^{-2i\pi jk/N} f_j^n = \hat{A}_k \hat{f}_k^n, \end{aligned}$$

où les indices sont pris modulo N puisque des conditions périodiques sont considérées. Maintenant, si on diagonalise la matrice $\hat{A}_k$ by $\hat{A}_k = P_k \Delta_k P_k^{-1}$, alors on a récursivement $\hat{f}_k^n = P_k(\Delta_k)^n P_k^{-1} \hat{f}_k^0$. Le facteur d'amplification est alors défini par $\rho(k/N) = \max_{0\le i\le d} |(\Delta_k)_{i,i}|$. Afin d'avoir une solution bornée, le facteur d'amplification doit satisfaire $\rho(\omega) \le 1$ pour $0 \le \omega \le 1$. On compare ici le facteur d'amplification pour les schémas de Galerkin Discontinu $DG0, DG1, DG2, DG3$, les schémas d'interpolation de Lagrange centrés de degré 9 et 17 ($LAG9$, $LAG17$) (voir [37,109]) et l'interpolation élément fini de degré $1, 2, 3$ et 4 ($Q1, Q2, Q3$ and $Q4$). Le facteur d'amplification $(1-\rho(\omega))$ de ces méthodes est tracé sur la Figure 4.4 pour des valeurs choisies de $\omega = k/N$, comme une fonction du déplacement $\alpha \in [0, 1]$. Remarquons que l'interpolation de type élément fini est instable pour un degré supérieur ou égal à 3 (voir $Q3$ et $Q4$).

5 Liens entre schémas semi-Lagrangiens et volumes finis

5.1 Le cadre du volume fini

Les inconnus sont $\bar{f}_i^n = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} f(t_n, x)dx$ et on écrit :

$$\bar{f}_i(t) = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} f(t, x)dx.$$

En intégrant l'équation de transport, on obtient

$$\frac{d\bar{f}_i(t)}{dt} = -\frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} a\partial_x f dx = -\frac{a}{h}[f(t, x_{i+1/2}) - f(t, x_{i-1/2})], \qquad (5.1)$$

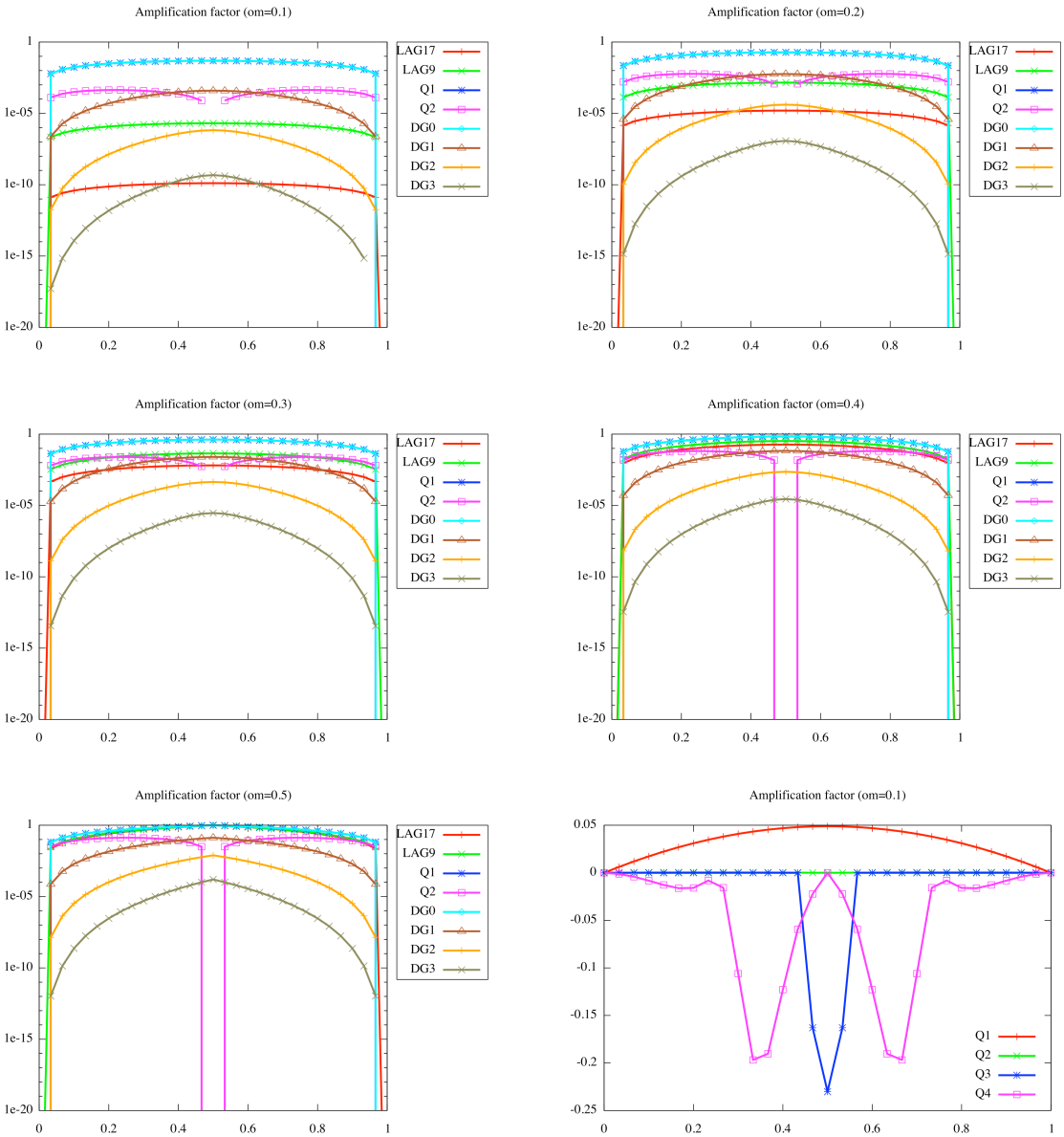

FIGURE 4.4 – **Facteur d'amplification** : $(1 - \rho(\omega))$ est tracé pour différentes valeurs de ω ($\omega = 0.1, 0.2, 0.3, 0.4, 0.5$ et de nouveau 0.1) pour différents schémas, en fonction du déplacement normalisé $\alpha \in [0, 1]$.

et il s'agit de donner un sens aux flux $f(t, x_{i\pm1/2})$ pour une suite donnée $(\bar{f}_i(t))_i$. Pour cela, on écrit classiquement

$$f(t, x_{i+1/2}) \approx \sum_{j=r}^{s} a_j \bar{f}_{i+j}(t), \tag{5.2}$$

Les conditions d'ordre permettent de déterminer les coefficients en résolvant le système

$$\frac{1}{2^k} = \sum_{j=r}^{s} a_j \int_{j-1/2}^{j+1/2} x^k dx, \ k = 0, \dots, s-r.$$

On a ainsi par exemple

$$\begin{aligned}
&CD2 : f(t, x_{i+1/2}) \approx \frac{1}{2}(\bar{f}_i(t) + \bar{f}_{i+1}(t)),\\
&CD4 : f(t, x_{i+1/2}) \approx \frac{7}{12}(\bar{f}_i(t) + \bar{f}_{i+1}(t)) - \frac{1}{12}(\bar{f}_{i-1}(t) + \bar{f}_{i+2}(t)),\\
&CD6 : f(t, x_{i+1/2}) \approx \frac{37(\bar{f}_i(t) + \bar{f}_{i+1}(t))}{60} - \frac{8(\bar{f}_{i-1}(t) + \bar{f}_{i+2}(t))}{60} + \frac{(\bar{f}_{i-2}(t) + \bar{f}_{i+3}(t))}{60},\\
&UP1 \ (a<0) : f(t, x_{i+1/2}) \approx \bar{f}_{i+1}(t),\\
&UP3 \ (a<0) : f(t, x_{i+1/2}) \approx \frac{1}{3}\bar{f}_i(t) + \frac{5}{6}\bar{f}_{i+1}(t) - \frac{1}{6}\bar{f}_{i+2}(t),\\
&UP5 \ (a<0) : f(t, x_{i+1/2}) \approx -\frac{1}{20}\bar{f}_{i-1}(t) + \frac{1}{9}\bar{f}_i(t) + \frac{47}{60}\bar{f}_{i+1}(t) - \frac{13}{60}\bar{f}_{i+2}(t) + \frac{1}{30}\bar{f}_{i+3}(t),\\
&UP1 \ (a>0) : f(t, x_{i+1/2}) \approx \bar{f}_i(t),\\
&UP3 \ (a>0) : f(t, x_{i+1/2}) \approx -\frac{1}{6}\bar{f}_{i-1}(t) + \frac{5}{6}\bar{f}_i(t) + \frac{1}{3}\bar{f}_{i+1}(t),\\
&UP5 \ (a>0) : f(t, x_{i+1/2}) \approx \frac{1}{30}\bar{f}_{i-2}(t) - \frac{13}{60}\bar{f}_{i-1}(t) + \frac{47}{60}\bar{f}_i(t) + \frac{1}{9}\bar{f}_{i+1}(t) - \frac{1}{20}\bar{f}_{i+2}(t).
\end{aligned}$$

5.2 Le flux semi-Lagrangien

On rappelle ici d'abord un lien entre les volumes finis et la forme semi-Lagrangienne du flux. Ce résultat, valide pour un champ général a déjà été prouvé dans [90], par exemple, en utilisant le théorème de la divergence. On donne ici une autre preuve basée sur un changement de variable.

Proposition 5.1. *On a*

$$\int_{t_n}^{t_{n+1}} a(t, x_{i+1/2}) f(t, x_{i+1/2}) dt = \int_{x^*_{i+1/2}}^{x_{i+1/2}} f(t_n, y) dy,$$

où

$$\partial_t f(t,x) + \partial_x(a(t,x)f(t,x)) = 0, \ X'(t) = a(t, X(t)), \ X(t_{n+1}) = x_{i+1/2}, \ X(t_n) = x^*_{i+1/2}.$$

Preuve. On écrit $X(t,s,x)$ la caractéristique satisfaisant $\partial_t X(t,s,x) = a(t, X(t,s,x)), \ X(s,s,x) = x$. On a tout d'abord, en suivant [41]

$$\int_{t_n}^{t_{n+1}} a(t, x_{i+1/2}) f(t, x_{i+1/2}) dt = \int_{t_n}^{t_{n+1}} a(t, x_{i+1/2}) f(t_n, X(t_n, t, x_{i+1/2})) \partial_x X(t_n, t, x_{i+1/2}) dt.$$

On fait ensuite le changement de variable $y = X(t_n, t, x_{i+1/2})$, afin de passer de l'intégrale en temps à l'intégrale en espace. Notons que l'on a

$$X(t_n, t', X(t', t, x_{i+1/2})) = X(t_n, t, x_{i+1/2}),\ \forall t',$$

ce qui signifie que cette quantité ne dépend pas de t'. La dérivée par rapport à t' est donc nulle, ce qui signifie que

$$\partial_s X(t_n, t', X(t', t, x_{i+1/2})) + \partial_t X(t', t, x_{i+1/2}) \partial_x X(t_n, t', X(t', t, x_{i+1/2})) = 0,$$

c'est-à-dire

$$\partial_s X(t_n, t', X(t', t, x_{i+1/2})) = -a(t, X(t', t, x_{i+1/2})) \partial_x X(t_n, t', X(t', t, x_{i+1/2})).$$

En prenant $t' = t$, on obtient

$$\partial_s X(t_n, t, x_{i+1/2}) = -a(t, x_{i+1/2}) \partial_x X(t_n, t, x_{i+1/2}),$$

et donc $dy = -a(t, x_{i+1/2}) \partial_x X(t_n, t, x_{i+1/2}) dt$. Comme on a $X(t_n, t_n, x_{i+1/2}) = x_{i+1/2}$ et $X(t_n, t_{n+1}, x_{i+1/2}) = x^*_{i+1/2}$, on obtient

$$\int_{t_n}^{t_{n+1}} a(t, x_{i+1/2}) f(t, x_{i+1/2}) dt = -\int_{x_{i+1/2}}^{x^*_{i+1/2}} f(t_n, y) dy,$$

ce qui donne le résultat. On réfère à [14] pour de tels calculs sur les caractéristiques. □

5.3 Intégrateur exponentiel

On fait maintenant un lien entre le système d'EDO qui vient de la formulation volumes finis (5.1)-(5.2) et les schémas semi-Lagrangiens pour une discrétisation en espace donnée.

Proposition 5.2. *Considérons le schéma semi-Lagrangien avec une reconstruction LAG-$2d+1$ appliquée M fois avec le pas de temps $\Delta t/M$ qui peut être écrite sous la forme*

$$(f_j^{n+1,M})_{j=0,\dots,N-1} = \prod_{k=1}^{M} \mathcal{T}_{\Delta t/M} (f_j^{n,M})_{j=0,\dots,N-1}.$$

On a a alors

$$\lim_{M\to\infty} f_j^{n,M} = \bar{f}_j(t_n),\ j = 0, \dots, N-1,$$

où $(\bar{f}_j)_{j=0,\dots,N-1}$ résoud (5.1) en prenant l'approximation upwind UP-$2d+1$ (5.2) avec $s = -r = d$ (pour $a > 0$).

Preuve. En considérant d'abord le système semi-discret de la méthode des volumes finis, on a

$$\frac{d\bar{f}_i}{dt} = -a(f_{i+1/2} - f_{i-1/2}),\ a > 0,$$

où les flux sont approchés par un schéma upwind $f(t, x_{i+1/2}) \approx \sum_{j=-d}^{d} a_j \bar{f}_{i+j}(t)$, les coefficients satisfaisant

$$\frac{1}{2^k} = \sum_{j=-d}^{d} a_j \int_{j-1/2}^{j+1/2} x^k dx, \ k = 0, \ldots, 2d.$$

La solution du système d'EDO (l'intégrateur exponentiel) peut être vu comme une approximation d'Euler en temps, en utilisant le pas de temps $\Delta t/M$ et en cherchant la limite $M \to +\infty$. Cela peut-être calculé facilement en utilisant l'analyse de Von Neumann. En effet, avec $f_{i+1/2} = \sum_{j=-d}^{d} c_j \bar{f}_{i+j}^n$, on a alors $(\widehat{f_{i+1/2}})_k = \sum_{j=-d}^{d} c_j \widehat{\bar{f}}_k^n e^{ikjh}$ et

$$(\widehat{\bar{f}}^{n+1})_k = \widehat{\bar{f}}_k^n \left(1 - \nu \left(\sum_{j=-d}^{d} c_j (e^{ijk\Delta x} - e^{i(j-1)k\Delta x})\right)\right) = \widehat{\bar{f}}_k^n (1 - \nu h(k)),$$

où $\nu = a\Delta t/(Mh)$ et $h(k)$ désigne la transformée de Fourier des flux.
On a alors

$$\begin{aligned} \lim_{M\to+\infty} (1 - \nu h(k))^M &= \lim_{M\to+\infty} \exp(M \ln(1 - \nu h(k))) \\ &= \lim_{M\to+\infty} \exp(-M\nu h(k)) \\ &= \exp(-a\Delta t/h \ h(k)). \end{aligned}$$

On observe ainsi que l'intégrateur est donné par l'exponentielle de $-a\Delta t/\Delta x$ multiplié par la transformée de Fourier des flux.
D'autre part, on considère la méthode semi-Lagrangienne

$$\bar{f}_i^{n+1} = \frac{1}{h} \int_{x_{i-1/2}-a\Delta t/M}^{x_{i+1/2}-a\Delta t/M} f(t^n, x) dx,$$

où $f(t^n, x)$ est reconstruit par une fonction polynômiale p_i de degré $\leq 2d$ satisfaisant les conditions

$$\frac{1}{h} \int_{x_{i-1/2+j}}^{x_{i+1/2+j}} p_i(x) dx = \bar{f}_{i+j}^n = \frac{1}{h}(P_i(x_{i+1/2+j}) - P_i(x_{i-1/2+j})), \ j = -d, \ldots, d,$$

où P_i désigne une primitive de p_i. Ainsi, le schéma numérique s'écrit aussi en termes des P_i

$$\bar{f}_i^{n+1} = \frac{1}{h}(P_i(x_{i+1/2} - a\Delta t/M) - P_i(x_{i-1/2} - a\Delta t/M)),$$

où l'on a supposé $a > 0$ et $a\Delta t/M < h$. Avec un développement de Taylor de $P_i(x_{i+1/2} - a\Delta t/M)$, on obtient

$$\bar{f}_i^{n+1} = \bar{f}_i^n - \nu(p_i(x_{i+1/2}) - p_i(x_{i-1/2})) + O(\nu^2),$$

avec $\nu = a\Delta t/(M\Delta x)$. Une analyse de Von Neuman conduit à $(\widehat{\bar{f}}^{n+1})_k = (\widehat{\bar{f}}^n)_k(1 - \nu h(k) + O(\nu^2))$ où $h(k)\widehat{\bar{f}}_k^n$ désigne la transformée de Fourier de $[p_i(x_{i+1/2}) - p_i(x_{i-1/2})]$. Ensuite, en prenant la limite $M \to +\infty$, on obtient

$$\lim_{M\to+\infty} (1 - \nu h(k) + O(\nu^2))^M = \lim_{M\to+\infty} \exp(M \ln(1 - \nu h(k) + O(\nu^2))) = \exp(-a\Delta t/\Delta x \ h(k)).$$

Cela suffit à montrer que $p_i(x_{i+1/2})$ (pour la méthode conservative) est égal à l'approximation de $f(x_{i+1/2})$ (pour la méthode de volumes finis). La valeur $p_i(x_{i+1/2})$ peut être écrite comme $p_i(x_{i+1/2}) = \sum_{j=-d}^{d} a_j \bar{f}^n_{i+j}$ où a_j satisfait le système de Vandermonde, qui correspond bien avec l'approximation des flux $f(x_{i+1/2})$ obtenus par la méthode des volumes finis. □

Remarque 5.3. *Une correspondance similaire peut être établie pour les schémas aux différences finies centrées (CD). En particulier, les analogues de CD2, CD4 et CD6 sont PPM0, PPM1, PPM2.*

Remarque 5.4. *Pour les schémas semi-Lagrangiens, on peut aussi utiliser des approximations upwind dans la reconstruction au lieu de PPM :*

$$P_i(x_{i-1/2} + \alpha\Delta x) =$$
$$(3\alpha^2 - 4\alpha + 1) f^n_{(i-1/2)^+} + (3\alpha^2 - 2\alpha) f^n_{(i+1/2)^-} + (6\alpha - 6\alpha^2)\bar{f}^n_i \quad \text{avec } \alpha \in [0,1],$$

et $f^n_{(i+1/2)^+}$ (resp. $f^n_{(i+1/2)^-}$) est reconstruit en utilisant (5.2) avec $s = d+1, r = -d+1$ (resp. $s = -r = d$). Dans le cas $d = 0, 1$, ce schéma coincide avec LAG-$2d+1$. Pour d plus grand, il ne coincide pas avec LAG-$2d+1$ (puisque la reconstruction est toujours de degré trois et n'a donc pas le même degré que LAG-$2d+1$). Mais la limite "intégrateur exponentiel" (comme définie dans la Proposition 5.2) le sera. En particulier, on peut gagner un ordre plus grand de précision à la limite.

Remarque 5.5. *On peut vérifier que les schémas CD préservent exactement la norme L^2 discrète $\sum_{j=0}^{N-1} |f_j(t)|^2$. D'autre part, les schémas upwind font décroître la norme L^2 : on peut vérifier que*

$$\sum_{j=-d}^{d} a_j(\cos(j\omega) - \cos((j-1)\omega)) \geq 0, 0 \leq \omega \leq 2\pi,$$

pour $d = 3$ par exemple, et cette relation reste vraie grâce à la stabilité du schéma LAG-$2d+1$, pour tout $d \in \mathbb{N}$. La conservation de la norme L^2 qui est a priori une bonne propriété n'est pas très satisfaisante, puisqu'elle génère généralement des oscillations parasites. Au contraire, un peu de dissipation, obtenue avec une approximation upwind d'ordre élevé semble être meilleure dans le régime limite. Voir aussi [37], pour une discussion sur les schémas d'ordre pair et impair. Notons que c'est un point crucial dans [6] ; un schéma non linéaire est dérivé : une approximation centrée est utilisée là où la solution est régulière et une approximation d'un ordre de moins est utilisée là où la solution n'est pas régulière. On peut alors remarquer que, lorsque des pas de temps pas trop petits sont utilisés, la norme L^2 décroit généralement pour un schéma semi-Lagrangien avec une reconstruction centrée des dérivées (comme les splines cubiques, PPM) et cela peut empêcher les oscillations parasites qui sont observées dans le cas des volumes finis.

Remarque 5.6. *Revenons sur le fait que l'on utilise l'approximation point milieu au temps initial*

$$\bar{f}^0_j = f(0, x_j),$$

Un choix plus naturel serait d'utiliser

$$\bar{f}^0_j = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} f(0,x)dx.$$

Cependant, avec ce choix, on perd l'approximation d'ordre élevé (sic !), puisque dans ce contexte semi-Lagrangien, le schéma est alors équivalent au schéma semi-Lagrangien ponctuel. D'ailleurs, dans [89], les auteurs classifient ce type de méthodes parmi les schémas semi-Lagrangiens de type différences finies (et non volumes finis) et présente ce type de schémas en introduisant une fonction H satisfaisant

$$f(t_n, x_j) \simeq \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} H(t_n, x) dx,$$

qui est mise à jour d'une façon volumes finis. Voir [89] pour des détails et d'autres reconstructions similaires dans le cadre WENO.

5.4 Résultats numériques

Comme illustration numérique, on considère le cas test de l'instabilité double -faisceau. La condition initiale est donnée par

$$f_0(x, y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) (1 + 0.05 \cos(0.5x)), \quad (x, y) \in [0, L] \times [-9, 9],$$

avec $L = 4\pi$.
On présente les diagnostiques 2D de la fonction de distribution sur les Figures 4.5 et 4.6.
On remarque que CD4 présente beaucoup d'oscillations, ce qui donne une mauvaise qualité de la solution (cf Remarque 5.5). On voit aussi bien le lien entre up5 et LAG5 et entre CD4 et PPM1 pour un (très) petit pas de temps. Enfin, on voit que PPM1 se comporte correctement lorsque le pas de temps n'est pas trop grand, ce qui est possible dans le cadre d'un schéma semi-Lagrangien.

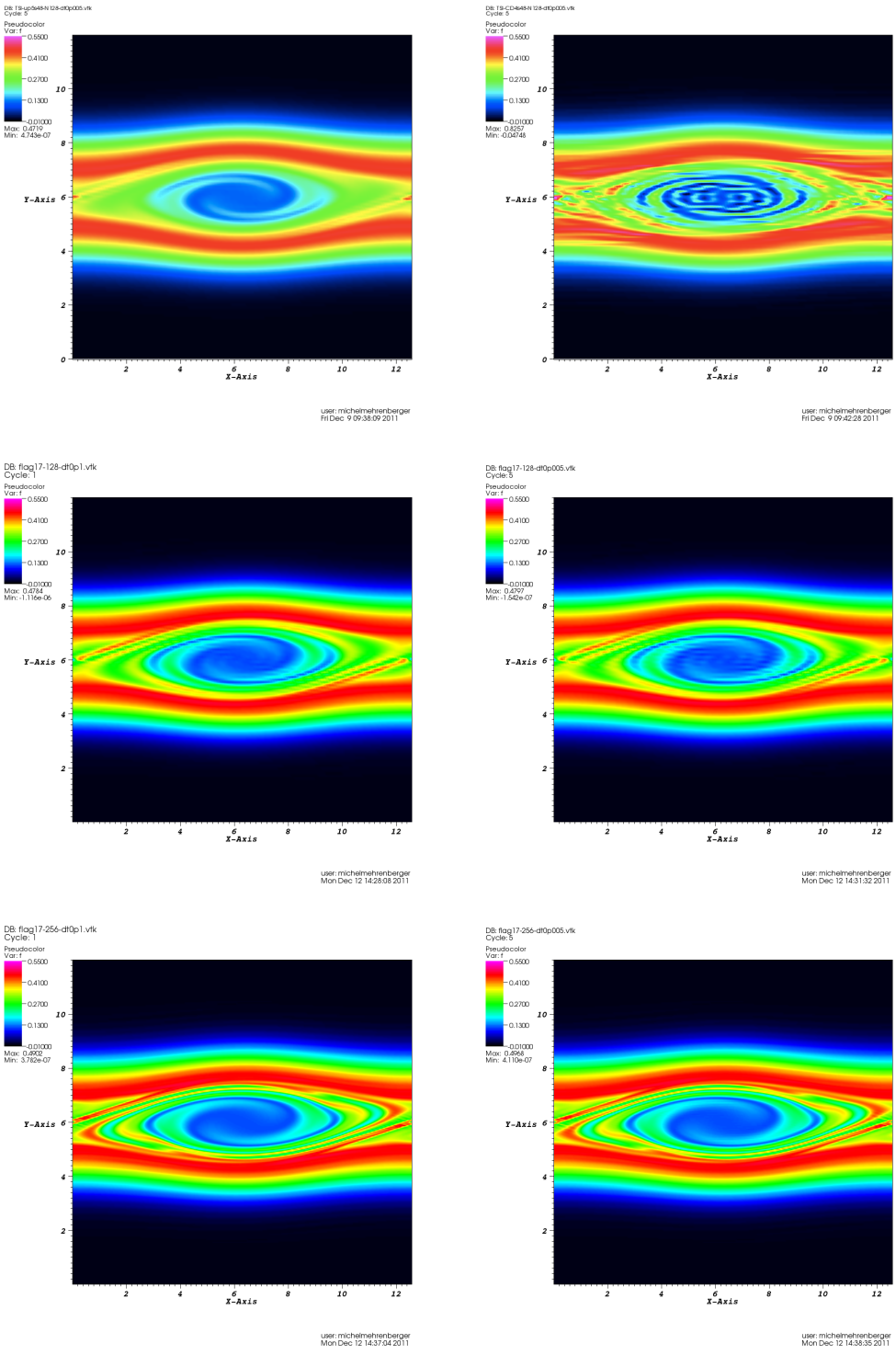

FIGURE 4.5 – Test de l'instabilité double-faisceau : fonction de distribution en fonction de x et v au temps $t = 53$ pour (en haut) up5 (a gauche) CD4 (à droite) $N_x = N_v = 128$, $\Delta t = 0.005$; puis (comme référence) pour une méthode semi-Lagrangienne avec une reconstruction de Lagrange d'ordre 17 et $\Delta t = 0.1$ (à gauche), $\Delta t = 0.005$ (à droite), et $N_x = N_v = 128$ (au milieu), $N_x = N_v = 256$ (en bas).

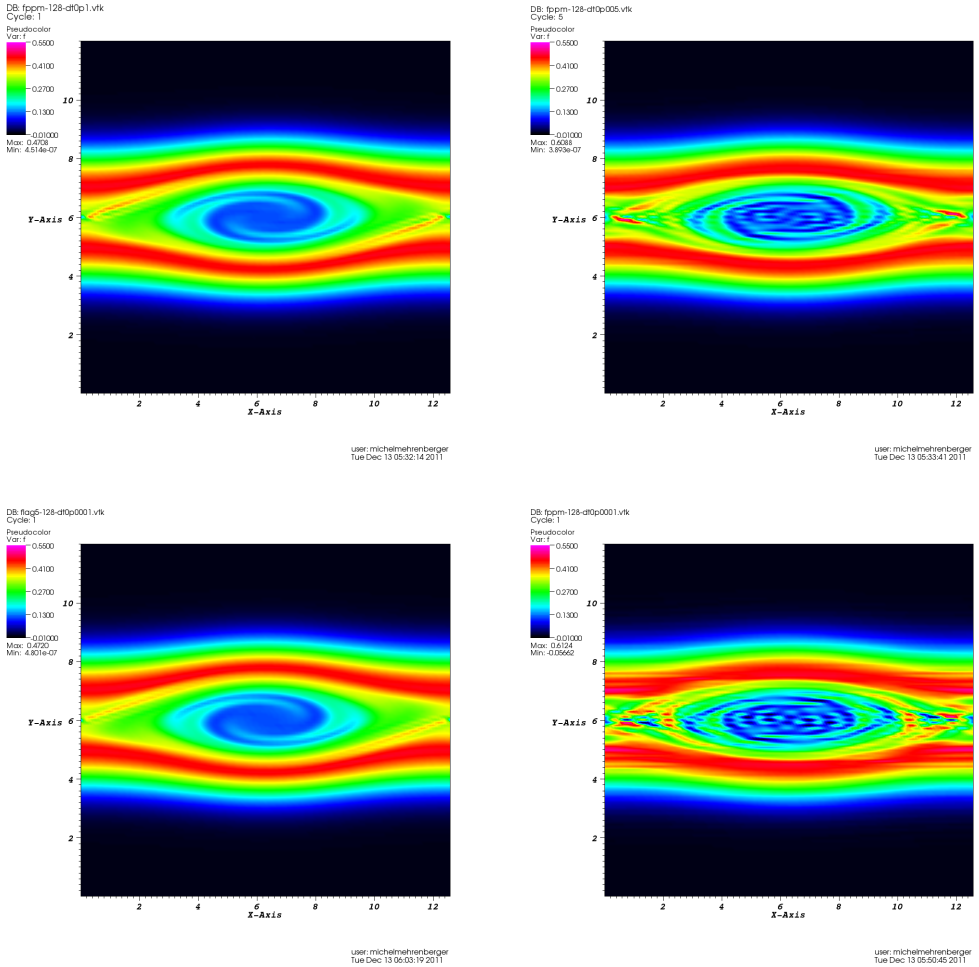

FIGURE 4.6 – Test de l'instabilité double-faisceau : fonction de distribution en fonction de x et v au temps $t = 53$ pour une méthode semi-Lagrangienne avec $N_x = N_v = 128$ et une reconstruction PPM1 avec $\Delta t = 0.1$ (en haut à gauche), $\Delta t = 0.005$ (en haut à droite), $\Delta t = 0.0001$ (en bas à droite), et une reconstruction de Lagrange d'ordre 5 avec $\Delta t = 0.0001$ (en bas à gauche).

Chapitre 5

Un résultat de convergence pour le système de Vlasov-Poisson

Ce chapitre est issu du travail [21].

1 Introduction

L'étude du transport linéaire est une brique fondamentale dans le cadre de la résolution numérique du système de Vlasov-Poisson par splitting. On donne ici un résultat de convergence dans le cadre d'un splitting de Strang. Une attention particulière est portée pour éviter la singularité en $1/\Delta t$ d'estimations précédentes [9], qui arrive lorsque le pas de temps devient très petit par rapport au pas d'espace. On donne ici des conditions sur l'opérateur d'interpolation (sous section 4) pour avoir l'estimation de convergence (7.2) qui n'est plus singulière. En particulier, en prenant une reconstruction de Lagrange LAG (2d+1), pour le transport et la reconstruction $\mathcal{R}_x \otimes \mathcal{R}_v$, on est dans ce cadre ($p = 2d + 1$), en utilisant le Lemme 4.1 (voir [21] et [9] aussi). Cette hypothèse relaxée peut permettre aussi de montrer la convergence de schémas de type intégrateur exponentiel, vus comme limites de schémas semi-Lagrangiens.

2 Algorithme

On fixe $N_x, N_v \in \mathbb{N}^*$, $L, v_{\max} > 0$. On définit la grille

$$x_i = i\Delta x, v_j = -v_{\max} + j\Delta v, v_{\min} = -v_{\max}, \Delta x = L/N_x,\ \Delta v = 2v_{\max}/N_v,$$

du domaine de simulation $\Omega = [0, L] \times [-v_{\max}, v_{\max}]$; on prend ici des conditions périodiques en x et en v. On définit une approximation numérique $f^{(n)} \in \mathbb{R}^{N_x \times N_v}$, par

$$f_{i,j}^{(n)},\ i = 0, \ldots, N_x - 1,\ j = 0, \ldots, N_v - 1,\ n \in \mathbb{N}.$$

L'algorithme de Vlasov-Poisson avec splitting de Strang s'écrit sous la forme

$$f^{(n+1)} = \mathcal{T}_{x,\Delta t/2}\mathcal{T}_{v,E_h^n,\Delta t}\mathcal{T}_{x,\Delta t/2}f^{(n)},\ n \in \mathbb{N},\ f^{(0)} = \Pi f(0).$$

Pour une fonction continue $g \in \mathcal{C}(\Omega)$, on définit sa fonction discrète projetée associée $\Pi g \in \mathbb{R}^{N_x \times N_v}$, par

$$(\Pi g)_{i,j} = g(x_i, v_j),\ i = 0, \dots, N_x - 1,\ j = 0, \dots, N_v - 1.$$

L'advection en x $\mathcal{T}_{x,\Delta t/2}$ agit sur $g_{0,j}, \dots, g_{N_x-1,j}$, avec la vitesse $\alpha_j = v_j \Delta t/2$:

$$\mathcal{T}_{x,\Delta t/2} g = \mathcal{T}_{\alpha_j}(g_{0,j}, \dots, g_{N_x-1,j}),\ g \in \mathbb{R}^{N_x \times N_v}.$$

L'advection en v $\mathcal{T}_{v,E_h^n,\Delta t}$ agit sur $g_{i,0}, \dots, g_{i,N_v-1}$, avec la vitesse $\alpha_i = E_h^n(x_i)\Delta t$:

$$\mathcal{T}_{v,E_h^n,\Delta t} g = \mathcal{T}_{\alpha_i}(g_{i,0}, \dots, g_{i,N_v-1}),\ g \in \mathbb{R}^{N_x \times N_v}.$$

Il reste alors à définir le champ $E_h^n(x)$. Pour cela, on peut prendre

$$E_h^n(x) = \int_0^L K(x,y) \left(\int_{-v_{\max}}^{v_{\max}} \left(\mathcal{R}_x \otimes \mathcal{R}_v \mathcal{T}_{x,\Delta t/2} f^{(n)} \right)(y,v) dv - 1 \right) dy, \tag{2.1}$$

où $\mathcal{R}_x \otimes \mathcal{R}_v : \mathbb{R}^{N_x \times N_v} \to L^\infty([0,L]) \times L^\infty([-v_{\max}, v_{\max}])$ est un opérateur de reconstruction.

3 Décomposition de l'erreur

L'erreur numérique est par définition $e^{(n)} \in \mathbb{R}^{N_x \times N_v}$

$$e^{(n)} = \Pi f(t_n) - f^{(n)}.$$

On a alors la décomposition de l'erreur

$$e^{(n+1)} = \varepsilon_1 + \varepsilon_2 + \mathcal{T}_{x,\Delta t/2}\varepsilon_3 + \mathcal{T}_{x,\Delta t/2}\varepsilon_4 + \mathcal{T}_{x,\Delta t/2}\mathcal{T}_{v,E_h^n,\Delta t}(\varepsilon_5 + \varepsilon_6),$$

avec

$$\begin{aligned}
\varepsilon_1 &= \Pi f(t_{n+1}) - \Pi \widetilde{\mathcal{T}_{x,\Delta t/2}} \widetilde{\mathcal{T}_{v,E[\cdot],\Delta t}} \widetilde{\mathcal{T}_{x,\Delta t/2}} f(t_n), \\
\varepsilon_2 &= \left(\Pi \widetilde{\mathcal{T}_{x,\Delta t/2}} - \mathcal{T}_{x,\Delta t/2}\Pi \right) \widetilde{\mathcal{T}_{v,E[\cdot],\Delta t}} \widetilde{\mathcal{T}_{x,\Delta t/2}} f(t_n). \\
\varepsilon_3 &= \Pi \left(\widetilde{\mathcal{T}_{v,E[\cdot],\Delta t}} - \widetilde{\mathcal{T}_{v,E_h^n,\Delta t}} \right) \widetilde{\mathcal{T}_{x,\Delta t/2}} f(t_n), \\
\varepsilon_4 &= \left(\Pi \widetilde{\mathcal{T}_{v,E_h^n,\Delta t}} - \mathcal{T}_{v,E_h^n,\Delta t}\Pi \right) \widetilde{\mathcal{T}_{x,\Delta t/2}} f(t_n), \\
\varepsilon_5 &= \left(\Pi \widetilde{\mathcal{T}_{x,\Delta t/2}} - \mathcal{T}_{x,\Delta t/2}\Pi \right) f(t_n) \\
\varepsilon_6 &= \mathcal{T}_{x,\Delta t/2} e^{(n)}.
\end{aligned}$$

Pour cela, on utilise le transport exact en x : $\widetilde{\mathcal{T}_{x,\Delta t/2}} : L^\infty([0,L]) \times L^\infty([-v_{\max}, v_{\max}]) \to L^\infty([0,L]) \times L^\infty([-v_{\max}, v_{\max}])$

$$\widetilde{\mathcal{T}_{x,\Delta t/2}} g = g(x - v\Delta t/2, v),$$

et le transport exact en v $\widetilde{\mathcal{T}_{v,E[\cdot],\Delta t}} : L^\infty([0,L]) \times L^\infty([-v_{\max}, v_{\max}]) \to L^\infty([0,L]) \times L^\infty([-v_{\max}, v_{\max}])$

$$\widetilde{\mathcal{T}_{v,E[\cdot],\Delta t}} g = g(x, v - E[g](x)\Delta t),\ E[g](x) = \int_0^L K(x,y)(\int_{-v_{\max}}^{v_{\max}} g(y,v) dv - 1) dy.$$

On définit de la même manière $\widetilde{\mathcal{T}_{v,E_h^n,\Delta t}}$, en remplaçant $E[\cdot]$ par E_h^n. La norme L^2 d'une fonction discrète est définie par (3.1) [1].

4 Hypothèses sur les reconstructions

Pour pouvoir traiter simultanément x et v, on considère la variable z. On considère ainsi

$$z_k = z_{\min} + k\Delta z,\ \Delta z = (z_{\max} - z_{\min})/N_z,$$

et $(\Pi_z g)_k = g(z_k),\ k = 0, \ldots, N_z - 1$.

On introduit aussi

$$h = \max(\Delta x, \Delta v).$$

On suppose que l'on a la stabilité en norme L^2 du transport linéaire :

$$\|\mathcal{T}_\alpha(g_0, \ldots, g_{N_z-1})\|_2 \leq \|(g_0, \ldots, g_{N_z-1})\|_2. \tag{4.1}$$

On suppose l'approximation suivante pour $g \in \mathbb{R}^{N_z}$

$$\|(\Pi_z\widetilde{\mathcal{T}}_\alpha - \mathcal{T}_\alpha\Pi_z)g)\|_2 \leq C\alpha(1-\alpha)\Delta z^{p+1}\|\partial_z^{p+1}g\|_{L^2([z_{\min},z_{\max}])}. \tag{4.2}$$

On a ici $\widetilde{T}_\alpha$, l'opérateur de transport exact : $\widetilde{T}_\alpha(g)(x) = g(x + \alpha\Delta z)$.
On suppose aussi une condition de stabilité entre l'approximation discrète et continue :

$$c\|g\|_2 \leq \|\mathbb{R}_x \otimes \mathbb{R}_v g\|_{L_2(\Omega)} \leq C\|g\|_2, \tag{4.3}$$

avec des constantes $c, C > 0$ indépendantes de $\Delta x, \Delta v$, ainsi qu'une propriété d'ordre d'approximation :

$$\|g - \mathbb{R}_x \otimes \mathbb{R}_v \Pi g\|_{L_2(\Omega)} \leq Ch^{p+1}. \tag{4.4}$$

1. On va utiliser les notations naturelles pour la norme de fonctions discrètes sur $\mathbb{R}$ discretisées par une longueur de maille $\Delta x > 0$. Par exemple la norme L^p d'une fonction discrète $w = (w_i)_{i\in\mathbb{Z}}$ est

$$\|w\|_p = \left(\Delta x \sum_i |w_i|^p\right)^{\frac{1}{p}} \quad 1 \leq p < \infty, \quad \text{et } \|w\|_\infty = \sup_i |w_i|.$$

Si le domaine est fini, par exemple $\Omega =]0,1[^2$, alors $N\Delta x = 1$ est demandé pour $N \in \mathbb{N}$: dans ce cas, la fonction discrète est $w = (w_{ij})_{1\leq i,j\leq N}$ avec des normes definies par (3.1).

$$\|w\|_p = \left(\Delta x \sum_{1\leq i,j\leq N} |w_{ij}|^p\right)^{\frac{1}{p}} \quad 1 \leq p < \infty, \quad \text{et } \|w\|_\infty = \sup_{1\leq i,j\leq N} |w_{ij}|. \tag{3.1}$$

Ces notations sont compatibles avec la définition standard de la norme L^p d'une fonction

$$\|z\|_{L^p(\Omega)} = \left(\int_\Omega |z(x)|^p dx\right)^{\frac{1}{p}} \quad 1 \leq p < \infty, \quad \text{et } \|z\|_{L^\infty(\Omega)} \sup_{x\in\Omega} |z(x)|.$$

5 Erreur en temps du splitting de Strang

On rappelle le résultat suivant (cf [18] par exemple, voir aussi la partie sur la discrétisation en temps pour Vlasov-Poisson)

Lemme 5.1. *Si $f(t_n)$ est borné dans $W^{1,\infty}(\Omega)$, alors*

$$\|f(t_{n+1}) - \widetilde{\mathcal{T}_{x,\Delta t/2}}\widetilde{\mathcal{T}_{v,E[\cdot],\Delta t}}\widetilde{\mathcal{T}_{x,\Delta t/2}}f(t_n)\|_{L^\infty(\Omega)} \leq C_T\Delta t^3.$$

On obtient alors

$$\|\varepsilon_1\|_2 \leq C_1\Delta t^3, \tag{5.1}$$

puisque le domaine est borné, les conditions sont périodiques et la donnée "initiale" $f(t_n)$ est à support compact dans Ω.

6 Estimation pour le champ électrique

Le troisième terme dans la décomposition est

$$\begin{aligned}(\varepsilon_3)_{i,j} =& f\left(t_n, x_i - (v_j - E^n(x_i)\Delta t)\Delta t/2, v_j - E^n(x_i)\Delta t\right)\\ &- f\left(t_n, x_i - (v_j - E_h^n(x_i)\Delta t)\Delta t/2, v_j - E_h^n(x_i)\Delta t\right),\end{aligned}$$

où l'on dénote par E^n le champ électrique calculé à partir de $\widetilde{\mathcal{T}_{x,\Delta t/2}}f(t_n)$

$$E^n(x) := E[\widetilde{\mathcal{T}_{x,\Delta t/2}}f(t_n)](x) = \int_0^L K(x,y)\left(\int_{\mathbb{R}}\left(\widetilde{\mathcal{T}_{x,\Delta t/2}}f(t_n)\right)(y,v)dv - 1\right)dy$$

On a donc à estimer la quantité suivante

$$E^n(x_i) - E_h^n(x_i) = \int_0^L K(x_i,y)\int_{-v_{\max}}^{v_{\max}} g(y,v)dvdy,$$

avec

$$\begin{aligned}g =& \widetilde{\mathcal{T}_{x,\Delta t/2}}f(t_n) - \mathcal{R}_x\otimes\mathcal{R}_v\Pi\widetilde{\mathcal{T}_{x,\Delta t/2}}f(t_n)\\ &+ \mathcal{R}_x\otimes\mathcal{R}_v(\Pi\widetilde{\mathcal{T}_{x,\Delta t/2}} - \mathcal{T}_{x,\Delta t/2}\Pi)f(t_n)\\ &+ \mathcal{R}_x\otimes\mathcal{R}_v\mathcal{T}_{x,\Delta t/2}(\Pi f(t_n) - f^{(n)}).\end{aligned}$$

En utilisant les hypothèses (4.1),(4.2),(4.3), on obtient l'estimation a priori suivante

$$\max_{i\in\{0,\dots,N_x-1\}}|E^n(x_i) - E_h^n(x_i)| \leq C\left(\max(\Delta x,\Delta v)^{p+1} + \|e^{(n)}\|_2\right). \tag{6.1}$$

De plus, puisque $E^n \in \mathrm{L}^\infty(0,L)$, on obtient

$$\begin{aligned}\max_{i\in\{0,\dots,N_x-1\}}|E_h^n(x_i)| &\leq C'(1 + \|e^{(n)}\|_2 + h^{p+1})\\ &\leq C''(1 + \|f^{(n)}\|_2 + \|\Pi f(t_n)\|_2)\end{aligned} \tag{6.2}$$

et alors

$$\sup_{n\leq\frac{T}{\Delta t}}\max_{i\in\{0,\dots,N_x-1\}}|E_h^n(x_i)| < +\infty \tag{6.3}$$

en utilisant le fait que

$$\|\Pi f(t_n)\|_2 \leq 2v_{\max}L\|f(t_n)\|_{L^\infty(\Omega)} \leq 2v_{\max}L\|f(0)\|_{L^\infty(\Omega)},$$

et, l'estimation (4.1)

$$\|f^{(n)}\|_2 \leq \|f^{(0)}\|_2 \leq 2v_{\max}L\|f(0)\|_{L^\infty(\Omega)}.$$

7 Estimation de convergence

L'inégalité triangulaire et (4.1) impliquent que

$$\|e^{(n+1)}\|_2 \leq \|\varepsilon_1\|_2 + \|\varepsilon_2\|_2 + \|\varepsilon_3\|_2 + \|\varepsilon_4\|_2 + \|\varepsilon_5\|_2 + \|e^{(n)}\|_2.$$

On a en utilisant (4.2)

$$\|\varepsilon_2\|_2 \leq C_2\tau_x\Delta x^{p+1},\ \|\varepsilon_4\|_2 \leq C_4\tau_v\Delta v^{p+1}, \|\varepsilon_5\|_2 \leq C_5\tau_x\Delta x^{p+1},$$

avec

$$\tau_x = \min\left(v_{\max}\frac{\Delta t}{\Delta x}, 1\right),\ \tau_v = \min\left(\sup_{n\leq\frac{T}{\Delta t}}\max_{i\in\{0,\ldots,N_x-1\}}|E_h^n(x_i)|\frac{\Delta t}{\Delta v}, 1\right), \tag{7.1}$$

ce qui donne en utilisant (6.3)

$$\|\varepsilon_2\|_2 + \|\varepsilon_4\|_2 + \|\varepsilon_5\|_2 \leq C_{245}\min\left(\frac{\Delta t}{h}, 1\right)h^{p+1}$$

En utilisant (5.1) et (6.1), on a aussi

$$\|\varepsilon_1\|_2 \leq C_1\Delta t^3, \|\varepsilon_3\|_2 \leq C_3\Delta t\left(\max(\Delta x, \Delta v)^{p+1} + \|e^{(n)}\|_2\right).$$

On obtient alors

$$\begin{aligned}\|e^{(n+1)}\|_2 &\leq \|e^{(n)}\|_2 + c\left(\min\left(\frac{\Delta t}{h}, 1\right)h^{p+1} + \Delta t\|e^{(n)}\|_2 + \Delta t^3\right)\\ &\leq e^{c\Delta t}\|e^{(n)}\|_2 + c\left(\min\left(\frac{\Delta t}{\Delta x}, 1\right)\Delta x^{p+1} + \Delta t^3\right)\end{aligned}$$

puisque $1 + c\Delta t \leq e^{c\Delta t}$. Ainsi, on a après sommation

$$\begin{aligned}\|e^{(n)}\|_2 \leq& e^{cn\Delta t}\|e^{(0)}\|_2\\ &+ c\left(\min\left(\frac{\Delta t}{h}, 1\right)h^{p+1} + \Delta t^3\right)\left(1 + e^{c\Delta t} + e^{c2\Delta t} + \cdots + e^{c(n-1)\Delta t}\right).\end{aligned}$$

Puisque $e^{(0)} = 0$ et $n\Delta t \leq T$, on obtient l'estimation finale

$$\|e^{(n)}\|_2 \leq \left(\min\left(\frac{\Delta t}{h}, 1\right)h^{p+1} + \Delta t^3\right)\frac{C}{\Delta t},\quad C > 0. \tag{7.2}$$

Chapitre 6

Le cas non constant

Jusqu'à présent, on s'est focalisé sur le cas de l'advection à coefficient constant (transport linéaire). On considère maintenant la généralisation des méthodes semi-Lagrangiennes au cas d'une advection générale à divergence nulle.

1 Introduction

On cherche à résoudre cette fois-ci

$$\partial_t f(t,x,y) + a_1(t,x,y)\partial_x f(t,x,y) + a_2(t,x,y)\partial_y f(t,x,y) = 0,$$

sous la condition de divergence nulle :

$$\partial_x a_1(t,x,y) + \partial_y a_2(t,x,y) = 0.$$

Par contre, on ne suppose plus cette fois-ci que

$$\partial_x a_1(t,x,y) = 0.$$

Dans ce cas, on peut considérer une advection $2D$ avec recherche des caractéristiques en $2D$: méthode BSL ou FSL (Backward or Forward Semi Lagrangian, voir [34] et [105]). On peut aussi se ramener à un dimension, en faisant du splitting sur la forme conservative qui est généralement préférée à la forme advective (pour pouvoir avoir une conservation de la masse). On se ramène alors à considérer l'équation $1D$

$$\partial_t g(t,x) + \partial_x(a(t,x)g(t,x)) = 0. \tag{1.1}$$

On utilise alors la propriété

$$g(t,x) = g(s, X(s;t,x))\partial_x X(s;t,x)$$

pour passer d'un instant s à un instant t, en utilisant une reconstruction de la fonction à l'instant s à partir des degrés de liberté qui doivent être mis à jour à l'instant t ; comme on l'a vu dans le cas de l'advection à coefficient constant. On a aussi besoin d'approcher les caractéristiques $X(\tau) = X(\tau;t,x)$, solutions de

$$X'(\tau) = a(\tau, X(\tau)),\ X(s) = x.$$

En faisant ce splitting, on perd néanmoins la propriété du principe du maximum.
Comme exemple de cas test, on va ici considérer le modèle centre-guide (2.3).
On peut aussi considérer l'équation de Vlasov-Poisson, lorsque l'on ne fait pas de splitting.

2 Méthodes semi-Lagrangiennes conservatives splittées

Cette section est issue du travail [31]. On généralise l'approche de [41].
On se base sur la résolution de l'équation (1.1).

Principe de la méthode

Pour $N \in \mathbb{N}^*$, on définit les points de grille

$$x_i = x_{\min} + i\Delta x,\ i \in \frac{1}{2}\mathbb{Z},\ \text{avec } \Delta x = (x_{\max} - x_{\min})/N \text{ and } I = [x_{\min}, x_{\max}].$$

On considère la quantité moyennée pour un temps donné s

$$\bar{g}_i(s) = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} g(s,x)dx,\ i = 0,\dots,N-1. \tag{2.1}$$

Maintenant, pour un autre temps t, grâce à la conservation du volume, on peut écrire l'égalité suivante

$$\int_{x_{i-1/2}}^{x_{i+1/2}} g(t,x)dx = \int_{x_{i-1/2}(s)}^{x_{i+1/2}(s)} g(s,x)dx, \tag{2.2}$$

où $x_{i-1/2}$ and $x_{i-1/2}(s)$ appartiennent à la même courbe caractéristique définie par

$$\frac{dX(\tau)}{d\tau} = a(\tau, X(\tau)),\quad X(t) = x_{i-1/2},\ X(s) = x_{i-1/2}(s),\ \ i = 0,..,N. \tag{2.3}$$

En supposant que les valeurs $\bar{g}_i(s),\ i = 0,\dots,N-1$ sont connues, on peut reconstruire la fonction primitive $G(s,x) = \frac{1}{\Delta x}\int_{x_{-1/2}}^{x} g(s,y)dy$ sur les points de grille comme fonction cumulative

$$G(s, x_{i-1/2}) = \sum_{k=0}^{i-1} \bar{g}_k(s),\ \ i = 1,\dots,N,\ G(s,x_{-1/2}) = 0. \tag{2.4}$$

En utilisant (2.2), on a alors

$$\bar{g}_i(t) = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} g(t,x)dx = \frac{1}{\Delta x}\int_{x_{i-1/2}(s)}^{x_{i+1/2}(s)} g(s,x)dx = G(s,x_{i+1/2}(s)) - G(s,x_{i-1/2}(s)). \tag{2.5}$$

Grâce à cette égalité, pour aller du temps s au temps t, on a besoin de
- calculer au moins numériquement les valeurs $x_{i-1/2}(s),\ i = 0,\dots,N$,
- reconstruire numériquement une fonction primitive (satisfaisant les contraintes d'interpolation (2.4)) sur $x_{i-1/2}(s),\ i = 0,\dots,N$.

Ainsi, comme dans le cas d'une méthode semi-Lagrangienne ponctuelle, l'algorithme de méthodes conservatives est composé de deux principales étapes : calcul des courbes caractéristiques et étape de reconstruction

2.1 Calcul des courbes caractéristiques

Dans le cas précédemment étudié où a est constant, l'intégration de (2.3) est directe. Notons que dans le cas général, on n'a pas d'information sur a a un temps donné. Typiquement, a peut dépendre de g à travers une équation de Poisson. Afin de surmonter cette difficulté, on peut utiliser un schéma à deux pas de temps comme dans [105] qui est d'ordre 2, ou utilise run schéma prédicteur-correcteur.
On considère une discrétisation en temps

$$t^n = n\Delta t,\ n \in \mathbb{N},\ \Delta t > 0,$$

et introduisons $\bar{g}_i^n \approx \bar{g}_i(t^n)$ défini par (2.1). Si on suppose que $\bar{g}_i^{n-1}$ et $\bar{g}_i^n$ sont connus pour $i = 0, \ldots, N-1$, on reconstruit le terme d'advection $a^n(x) \approx a(t^n, x)$, qui dépend généralement de $\bar{g}_i^n,\ i = 0, \ldots, N-1$. La valeur $x_{i+1/2}(t^{n-1})$ est alors approchée par $X_{i+1/2}^{n-1} \approx X(t^{n-1})$ qui est donné par

$$\frac{dX(t)}{dt} = a^n(X(t)),\ X(t^{n+1}) = x_{i+1/2},\ t \in [t^{n-1}, t^{n+1}] \tag{2.6}$$

et $\bar{g}_i^{n+1}$ peut alors être calculé par la formule (2.5), avec $t = t^{n+1}$ et $s = t^{n-1}$:

$$\bar{g}_i^{n+1} = G^{n-1}(X_{i+1/2}^{n-1}) - G^{n-1}(X_{i-1/2}^{n-1}), \tag{2.7}$$

avec $G^{n-1} \approx G(t^{n-1}, \cdot)$ calculé avec les valeurs $\bar{g}_i^{n-1},\ i = 0, \ldots, N-1$.
Pour calculer $X_{i+1/2}^{n-1}$, on peut soit calculer directement les pieds de la caractéristique se terminant aux interfaces $x_{i+1/2}$ comme suggéré par (2.6). On peut aussi résoudre la même équation avec condition finale $X(t^{n+1}) = x_i$ pour avoir X_i^{n-1} et après interpoler pour avoir $X_{i+1/2}^{n-1}$. En pratique, on utilise cette dernière approche avec l'approximation $X_{i+1/2}^{n-1} = (X_i^{n-1} + X_{i+1}^{n-1})/2$, qui reste d'ordre deux. Par la suite, on va présenter des moyens numériques de calculer la solution approchée X_i^{n-1}.

Formule du point milieu Comme dans [45, 105], une formule du point milieu peut être employée :

$$x_i - X_i^{n-1} = 2\Delta t a^n \left(\frac{x_i + X_i^{n-1}}{2} \right).$$

En écrivant $X_i^{n-1} = x_i - 2\alpha_i$, le déplacement α_i peut être calculé au second ordre en résolvant le point fixe uni-dimensionnel

$$\alpha_i = \Delta t\ a^n(x_i - \alpha_i). \tag{2.8}$$

Dans [105], un algorithme de Newton est utilisé. On mentionne aussi [45] dans lequel un développement de Taylor du membre de droite est utilisé, ce qui revient à utiliser l'agorithme de Newton avec deux itérations. Cependant, l'inconvénient de ces algorithmes est qu'ils demandent d'évaluer la matrice jacobienne de a^n. Un algorithme de point fixe peut-être alors aussi implémenté. Mais, si on suppose que l'on a une reconstruction linéaire par morceaux du terme d'advection, comme cel est supposé dans [45, 105], on peut rendre l'algorithme explicite. Les principales étapes de ce nouvel algorithme sont détaillées ci-après.

En partant de (2.8) et en notant $[x_j, x_{j+1}]$ la maille dans lequelle $(x_i - \alpha_i)$ tombe, la reconstruction linéaire de a^n s'écrit

$$\alpha_i = \Delta t \ [(1-\beta)a^n(x_j) + \beta a^n(x_{j+1})], \tag{2.9}$$

où β est tel que

$$x_i - \alpha_i = x_j + \beta, \ \ x_j = x_0 + j\Delta x, \ \ x_i = x_0 + i\Delta x. \tag{2.10}$$

L'injection de l'expression de α_i dans (2.9) conduit à

$$\beta\left[\Delta x + \Delta t\left(a^n(x_{j+1}) - a^n(x_j)\right)\right] = (i-j)\Delta x - \Delta t\, a^n(x_j), \tag{2.11}$$

où l'on peut déduire une expression de β

$$\beta = \left[(i-j)\Delta x - \Delta t\, a^n(x_j)\right] / \left[\Delta x + \Delta t(a^n(x_{j+1}) - a^n(x_j))\right]. \tag{2.12}$$

Maintenant, il reste à déterminer l'indice j. Pour faire cela, on doit remarquer que β donné par (2.12) est dans l'intervalle $[0, 1[$. Ainsi, de (2.11), on peut déduire une expression de $x_i = i\Delta x$

$$i\Delta x = j\Delta x + \Delta t \ a^n(x_j) + \beta\left[\Delta x + \Delta t(a^n(x_{j+1}) - a^n(x_j))\right].$$

En utilisant le fait que $\beta \in [0,1]$, et en remarquant que $[\Delta x + \Delta t(a^n(x_{j+1}) - a^n(x_j))] > 0$ pourvu que Δt soit suffisamment petit, on en déduit que

$$i\Delta x \in [M_j, M_{j+1}], \ \ \text{with } M_j = x_j + \Delta t\, a^n(x_j).$$

Sous l'hypothèse que Δt est assez petit, la suite $(M_j)_{j=0,..,N-1}$ est croissante et forme donc un maillage non uniforme pour lequel la position de x_i peut être trouvé facilement. L'algorithme est alors le suivant pour $i = 0, .., N-1$:

- détermination de j tel que $x_i \in [M_j, M_{j+1}]$
- détermination de β avec (2.12)
- détermination de α_i avec (2.10)

Méthodes de Runge-Kutta On peut aussi utiliser des techniques classiques comme les méthodes de Runge-Kutta (RK) pour l'intégration de (2.6). Notons que mêm si on utilise des méthodes de Runge Kutta d'ordre plus élevé, on ne peut pas avoir plus que de l'ordre deux en temps, puisque l'on résoud (2.6) à la place de (2.3). Cependant, on observe un meilleur comportement pour l'ordre 4 au leiu de l'ordre 2. Comme exemple, une méthode de Runge Kutta d'ordre 2 peut être définie ainsi :

$$k_1 = a^n(x_i), \quad k_2 = a^n(x_i - 2\Delta t\, k_1),$$

puis

$$X_i^{n-1} = x_i - \Delta t(k_1 + k_2).$$

Dans nos tests, une interpolation avec splines cubiques a été utilisée pour évaluer le champ a^n qui est connu aux points de grille $x_i, i = 0, \ldots, N-1$.

2.2 Étape de reconstruction

Pour l'étape de reconstruction, on a besoin d'interpoler la primitive en chacun des points $x_{i-1/2}(s)$. Pour cela, on peut utiliser les mêmes opérateurs d'interpolation présentés dans le cas de l'advection à coefficient constant.
Dans le cas d'une interpolation de type Hermite de la primitive, avec reconstruction des dérivées (LAG3, PPM, PSM...) de la primitive aux interfaces $g_{(k+1/2)^{\pm}}$, on peut implémenter la mise à jour des degrés de liberté de la manière suivante : pour $i = 0, \ldots, N$, on calcule

$$G_{i-1/2} = \beta(1-\beta)^2 g_{(j_i-1/2)^+} + \beta^2(\beta-1) g_{(j_i+1/2)^-} + \beta^2(3-2\beta)\bar{g}_{j_i}^{\text{old}}, \tag{2.13}$$

avec $x_{i-1/2}(s) = x_{j_i-1/2} + \beta\Delta x,\ 0 \leq \beta < 1$, et les nouvelles valeurs sont données par

$$\bar{g}_i^{\text{new}} = \sum_{k=j_i}^{j_{i+1}-1} \bar{g}_k^{\text{old}} + (G_{i+1/2} - G_{i-1/2}),\ i = 0, \ldots, N-1.$$

Dans le cadre de la reconstruction, on peut rajouter des limiteurs de pente qui consistent à modifier les valeurs aux extrémités $g_{(k+1/2)^{\pm}}$.

2.3 Limiteurs de pente

On s'intéresse ici à la description de différents filtres qui peuvent être adaptés à la reconstruction précédente. Il est bien connu que des schémas d'ordre élevé peuvent générer des nouveaux extrema, violer la monotonie et développer des oscillations numériques. Afin d'éviter ou de réduire ces problèmes, des filtres ont été introduits. Ce point a été étudié par beaucoup d'auteurs et reste le sujet de récents développements (voir par exemple [25, 26, 41, 110, 114] et références incluses).
Une première contrainte physique, qui est le principal objectif est la conservation de la positivité. Notons que cette propriété est globale et bien définie. Une propriété plus générale est la conservation du principe du maximum ; on doit distinguer ici les extrema globaux et locaux. Le maximum et le minimum sont bien définis pour la fonction initiale (qui est généralement donnée par une formule) et sont des bons candidats pour des extrema globaux pendant toute la simulation, comme c'est fait dans [41]. L'utilisation de bornes locales est plus ambiguë. En effet, même pour des bornes globales calculées par la solution courante, la valeur maximale en un temps donné peut diminuer à cause de la diffusion numérique. Donc, des filtres basés sur de telles valeurs peuvent forcer artificiellement la baisse du maximum et donc accélerer la diffusion d'autant plus. En fait, on devrait essayer de garder les extrema existants, de ne pas en générer d'autres et faire attention à ce qu'on ne dégrade pas l'ordre de convergence du schéma dans les régions où la solution est régulière. Dans [110], une reconstruction linéaire avec les mailles les plus proches permet de déterminer un extremum local. D'autres stratégies consistent à limiter les dérivées de la fonction reconstruite (qui peuvent être grandes lorsque la solution n'a pas de comportement régulier) comme dans [25, 53].
On a testé plusieurs filtres. Au lieu de les présenter tous, on va traiter seulement certains qui nous paraissent pertinents. Un filtre va avoir ici trois ingrédients ; tout d'abord la définition des extrema, ensuite la procédure de limitation qui oblige la

reconstruction à ne pas violer la définition des extrema et enfin une procédure de limitation des extrema.

On considère ici une reconstruction (2.13) et on va modifier les valeurs $g_{(j-1/2)+}$ et $g_{(j+1/2)-}$ sur la maille $[x_{j-1/2}, x_{j+1/2}]$.

Définition des extrema On doit d'abord définir les bornes $g_{\min}, g_{\max}$ où l'on veut garder la solution. Pour cela, on va considérer :

- *extrema positifs :* $g_{\min} = 0, g_{\max} = \infty$,
- *extrema globaux :* $g_{\min} = \min g^0(x), g_{\max} = \max g^0(x)$, où g^0 est la fonction initiale qui va être advectée,
- *extrema d'Umeda :* les extrema locaux qui sont définis comme dans [110] :

$$g_{\max} = \min(\max g^0(x), \max(g_{\max 1}, g_{\max 2})),\ g_{\min} = \max(\min g^0(x), \min(g_{\min 1}, g_{\min 2})),$$

avec

$$\begin{aligned}
g_{\max 1} &= \max(\max(g_{i-1}^{\text{old}}, g_i^{\text{old}}), \min(2g_{i-1}^{\text{old}} - g_{i-2}^{\text{old}}, 2g_i^{\text{old}} - g_{i+1}^{\text{old}})) \\
g_{\max 2} &= \max(\max(g_{i+1}^{\text{old}}, g_i^{\text{old}}), \min(2g_{i+1}^{\text{old}} - g_{i+2}^{\text{old}}, 2g_i^{\text{old}} - g_{i-1}^{\text{old}})) \\
g_{\min 1} &= \min(\min(g_{i-1}^{\text{old}}, g_i^{\text{old}}), \max(2g_{i-1}^{\text{old}} - g_{i-2}^{\text{old}}, 2g_i^{\text{old}} - g_{i+1}^{\text{old}})) \\
g_{\min 2} &= \min(\min(g_{i+1}^{\text{old}}, g_i^{\text{old}}), \max(2g_{i+1}^{\text{old}} - g_{i+2}^{\text{old}}, 2g_i^{\text{old}} - g_{i-1}^{\text{old}}))
\end{aligned}$$

Dans le cas non constant, les extrema positifs peuvent être définis. Cependant, dans le cas général, les contractions de volume peuvent amener la valeur g_i en dehors des bornes (cf par exemple le cas d'une condition initiale constante). Ainsi, on propose de relaxer la définition des extrema comme suit : on remplace $g_{\min}$ par $\min(g_i^{\text{old}}, g_{\min})$ et $g_{\max}$ par $\max(g_i^{\text{old}}, g_{\max})$. Notons que cette procédure n'a pas d'effet dans le cas de l'advection à coefficient constant.

Limitation des extrema Le *filtre d'Hyman* est donné par l'algorithme suivant

$$\begin{aligned}
g &= \max(g, \max(g_{\min}, -2g_{\max} + 3g_{j_i}^{\text{old}}); \\
g &= \min(g, \min(g_{\max}, 3g_{j_i}^{\text{old}} - 2g_{\min}));
\end{aligned}$$

où g prend successivement la valeur $g_{(j-1/2)+}$ et $g_{(j+1/2)-}$. Ce filtre assure que les fonctions $x \to G_i(x) - xg_{\min}$ et $x \to xg_{\max} - G_i(x)$ sont croissantes sur $[0, 1]$, si $g_{\min} \leq g_{j_i}^{\text{old}} \leq g_{\max}$. Donc les extrema positifs sont préservés et dans le cas de l'advection constante, les extrema globaux sont aussi préservés.

On pourrait aussi utiliser le limiteur PFC (voir [41]), qui a été destiné à la reconstruction LAG. Une autre possibilité est de modifier les valeurs $g_{(j-1/2)+}, g_{(j+1/2)-}$ le moins possible de telle sorte que la contrainte $g_{\min} \leq G_i'(x) \leq g_{\max}$, pour tout $0 \leq x \leq 1$ soit satisfaite. Comme exemple, on peut résoudre le problème de minimisation pour $|(g'_{j+})^{\text{new}} - g'_{j+}| + |(g'_{(j+1)-})^{\text{new}} - g'_{(j+1)-}|$. Cependant, cela peut ne pas être toujours une bonne idée ou utile de rester le plus proche de la première reconstruction, qui peut être parfois une mauvaise approximation.

Limitation des oscillations On a vu que la dérivée de la primitive peut être calculée par la reconstruction PSM ou LAG (supposé plus diffusif). On peut même

utiliser la reconstruction PPM0 $g_{j-1/2,m} = (\bar{g}_j^{\rm old} + \bar{g}_{j-1}^{\rm old})/2$, qui donne une reconstruciton encore plus diffusive. Ona alors rajouté le filtre suivant qui tend à privilégier la dérivée de la reconstruction la plus diffusive, si l'erreur entre les deux reconstructions est trop grande. Le but est d'atténuer les oscillations parasites, qui sont détectées quand l'erreur est grande. Une telle approche a été effectuée dans [25]. On considère ici une stratégie similaire. On compare la reconstruction LAG à PSM afin de corriger au mieux la formule PPM0 $g_{j-1/2,m} = \frac{\bar{g}_{j-1}^{\rm old} + \bar{g}_j^{\rm old}}{2}$.
En d'autres termes, si $(g_{(j-1/2)+,LAG} - g_{j-1/2,m})(g_{j-1/2,PSM} - g_{j-1/2,m}) < 0$, on prend $g_{(j-1/2)+} = g_{j-1/2,m}$, sinon

$$g_{(j-1/2)+} = g_{j-1/2,m} + s\,\min(C|g_{(j-1/2)+,LAG} - g_{j-1/2,m}|, |g_{j-1/2,PSM} - g_{j-1/2,m}|), \tag{2.14}$$

avec $s = sign(g_{j-1/2,PSM} - g_{j-1/2,m})$, et où $C > 1$. On modifie de manière similaire $g_{(j+1)-}$. Ce schéma dépend du choix de la constante C. On a pris ici $C = 2.5$ dans nos tests numériques. En augmentant C, le minimum dans (2.14) sera $g_{j-1/2,PSM}$, de telle sorte que le filtre n'aura plus d'effet. En baissant la valeur de C, la méthode aura un comportement plus diffusif.

3 Méthodes semi-Lagrangiennes Galerkin Discontinu

Cette section est issu du travail [33].

3.1 Principe de la méthode

On généralise le travail précédent [32] (ou [67,90,97]) concernant l'advection à coefficient constant au cas de la résolution de l'équation conservative (1.1). Notons qu'il y a déjà eu des développements de schémas semi-Lagrangiens Galerkin Discontinu, avec applications en météorologie [95]. Pour l'advection constante, le calcul des caractéristiques pouvait se faire de manière exacte. On va considérer ici calculer les pieds des caractéristiques des points de Gauss et faire évoluer les autres points de manière linéaire par morceaux. Modulo cette approximation, le calcul des intégrales se fait alors aussi de manière exacte, comme dans le cas de l'advection constante.
On part à nouveau de la discrétisation de l'équation d'advection non linéaire 1d (1.1). On définit les $(d+1)$ points et poids de Gauss sur l'intervalle $[0,1]$: $\gamma_k, \omega_k,\ k = 0,\ldots,d$. Les degrés de liberté sont alors $g_{j,\ell}^n \simeq g(t_n, x_{j,\ell}),\ x_{j,\ell} = x_{j-1/2} + \gamma_\ell h$.
On définit les fonctions de base

$$\varphi_{j,\ell}(x) \;=\; \begin{cases} \prod_{0\le\ell'\le d,\ell'\neq\ell} \frac{x - x_{j,\ell'}}{x_{j,\ell} - x_{j,\ell'}} & x \in [x_{j-1/2}, x_{j+1/2}] \\ \\ 0 & \text{sinon} \end{cases}.$$

L'approximation à l'instant t_n s'écrit alors

$$g_n(x) = \sum_{j=0}^{N-1} \sum_{\ell=0}^{d} g_{j,\ell}^n \varphi_{j,\ell}(x),$$

et on a, en utilisant la propriété des points de Gauss

$$\int_{x_{j-1/2}}^{x_{j+1/2}} g_n(x)\varphi_{j,\ell}(x)dx = \omega_\ell \Delta x g_{j,\ell}^n.$$

Afin de mettre à jour les degrés de liberté, on utilise les caractéristiques :

$$\int_{x_{j-1/2}}^{x_{j+1/2}} g(t_{n+1},x)\varphi_{j,\ell}(x)dx = \int_{x_{j-1/2}}^{x_{j+1/2}} g(t_n, X(t_n;t_{n+1},x))\varphi_{j,\ell}(x)\partial_x X(t_n,t_{n+1},x)dx.$$

On calcule alors numériquement les pieds des caractéristiques aux points de Gauss $x_{j,\ell}^b \approx X(t_n;t_{n+1},x_{j,\ell}), \ell = 0,\dots,d$, en utilisant par exemple une méthode de Runge Kutta (cf section précédente). Maintenant, pour $x \in [x_{j-1/2}, x_{j+1/2}[$, on peut trouver $u \in [0,1[$ et $\ell \in \{-1,\dots,d\}$ tels que $x = (1-u)x_{j,\ell} + ux_{j,\ell+1}$, en utilisant la convention $x_{j,-1} = x_{j-1,d+1} = \frac{1}{2}(x_{j-1,d} + x_{j,0})$. On fait alors l'approximation $x^b \approx X(t_n;t_{n+1},x)$ suivante

$$x^b = (1-u)x_{j,\ell}^b + ux_{j,\ell+1}^b,$$

avec à nouveau $x_{j,-1}^b = x_{j-1,d+1}^b = \frac{1}{2}(x_{j-1,d}^b + x_{j,0}^b)$.
Le schéma s'écrit alors

$$\begin{aligned} g_{j,\ell}^{n+1}\omega_\ell\Delta x &= \sum_{i',j'} g_{j',\ell'}^n \sum_{m=-1}^{d}\int_0^1 \varphi_{j',\ell'}((1-u)x_{j,m}^b + ux_{j,m+1}^b) \\ &\times \varphi_{j,\ell}((1-u)x_{j,m} + ux_{j,m+1})\mathrm{d}u(x_{j,m+1}^b - x_{j,m}^b). \end{aligned} \tag{3.1}$$

Les intégrales sont calculées de manière exacte, en séparant les parties où chacun des termes est un polynôme (par exemple si $x_{k-1} < x_{j,m}^b < x_k < x_{j,m+1}^b < x_{k+1}$, on doit séparer en deux : la partie $(1-u)x_{j,m}^b + ux_{j,m+1}^b \in [x_{j,m}^b, x_k]$ et la partie $(1-u)x_{j,m}^b + ux_{j,m+1}^b \in [x_k, x_{j,m+1}^b]$).

3.2 Simulations du modèle centre-guide

Afin de valider le code, on le compare avec la méthode BSL (Backward Semi-Lagrangian) (BSL) [105]. On utilise pour le code DG $\Delta t = 0.01$ afin d'éviter au code qu'il réduise lui-même le pas de temps (sinon, le code réduit le pas de temps à $\Delta t \approx 0.02$; le pas de temps est calculé à chaque pas de temps pour ne pas se déplacer de plus d'une maille et un schéma prédicteur-correcteur est utilisé). Le code BSL permet d'avoir des pas de temps plus grands. Sur la Figure 6.1 et la Figure 6.2, on dessine la densité $\rho(t,x_1,x_2)$ aux temps $t = 30$ et $t = 60$ et l'évolution de certaines quantités scalaires pour les deux méthodes. La méthode DG donne des résultats qualitatifs et quantitatifs comparables.

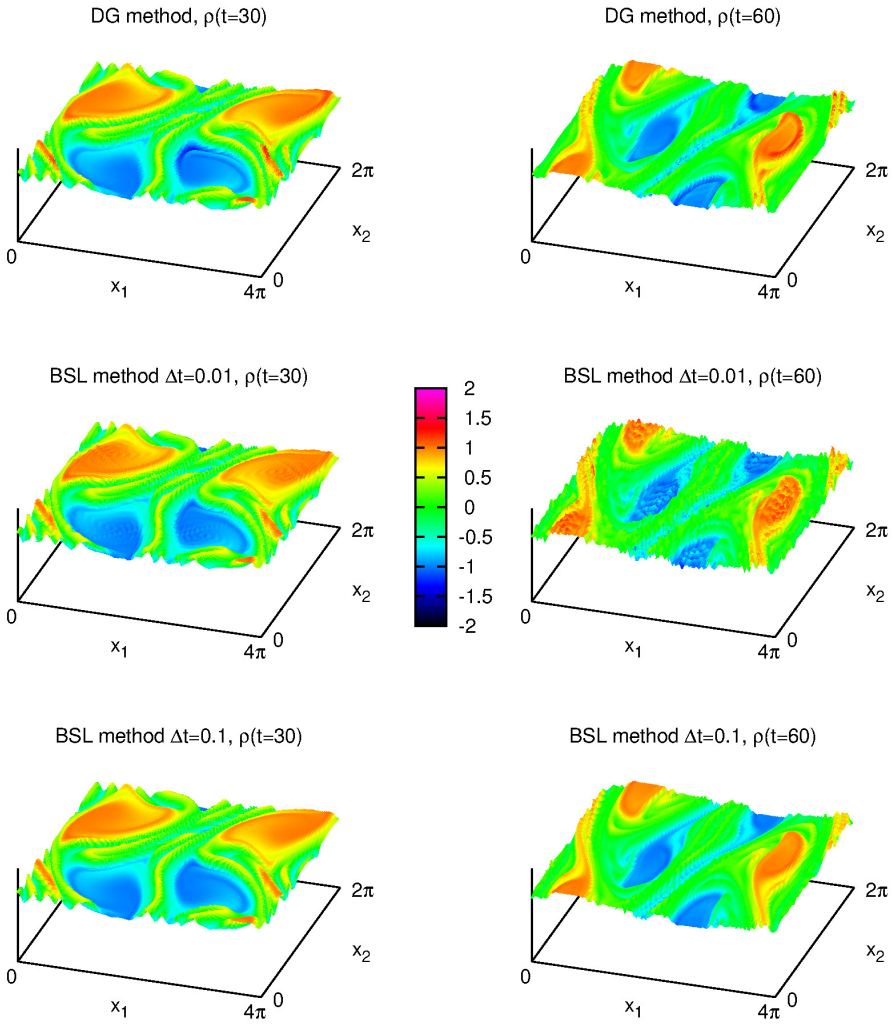

FIGURE 6.1 – **Modèle centre-guide.** Evolution de la densité $\rho(x_1, x_2)$, dessinée aux temps $t = 30$ et $t = 60$. Le pas de temps est $\Delta t = 0.01$ pour DG. Les maillages sont $(N_x, N_y, d) = (32, 32, 3)$ pour DG, $(N_x, N_y) = (128, 128)$ pour BSL.

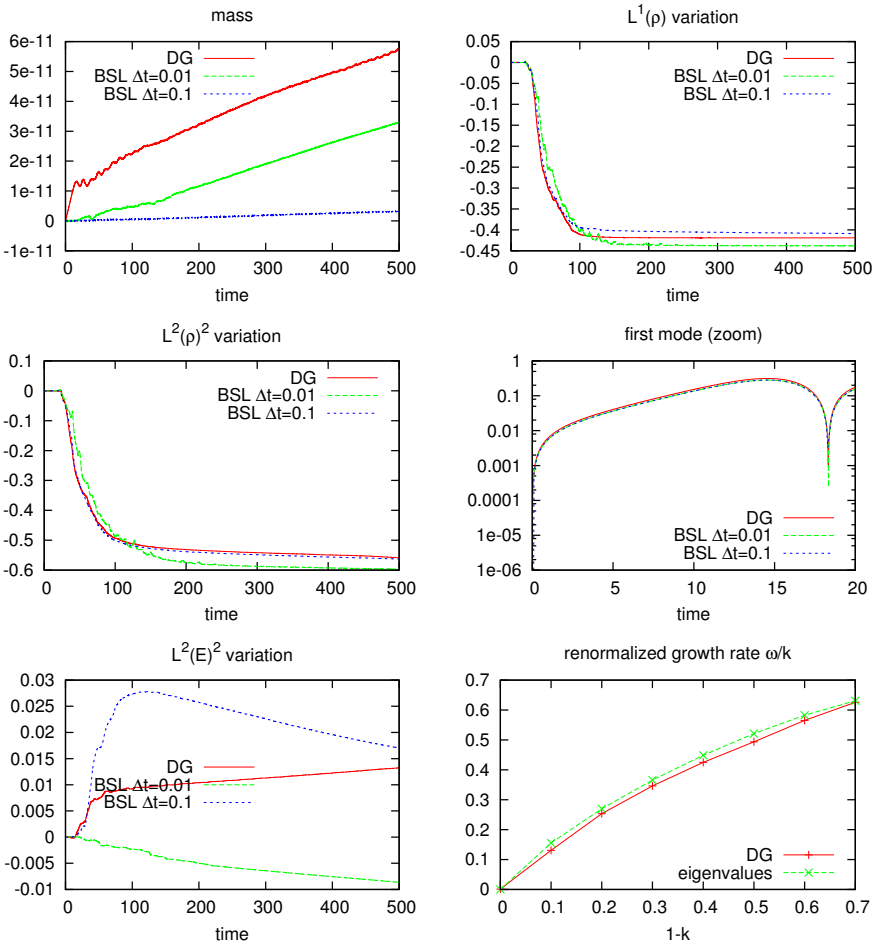

FIGURE 6.2 – **Modèle centre-guide.** Comportement de quelques quantités scalaires. En haut à gauche : évolution en temps de la masse totale. En haut à droite : évolution en temps de la variation de $\|\rho(t)\|_{L^1}$. Au centre à gauche : évolution en temps de la variation de $\|\rho(t)\|_{L^2}$ (enstrophie). Au centre à droite : évolution en temps du premier mode de Fourier $\|\phi_{k=0.5}(t)\|_{L^2}$ En bas à gauche : évolution en temps de la variation de $\|E(t)\|_{L^2}$. En bas à droite : taux de croissance ω/k en fonction de $k-1$. Le pas de temps est $\Delta t = 0.01$ pour DG. Les maillages sont $(N_x, N_y, d) = (32, 32, 3)$ pour DG, $(N_x, N_y) = (128, 128)$ pour BSL.

Chapitre 7

Le calcul de champs

Dans le cadre de la résolution numérique de l'équation de Vlasov, la méthode de transport est généralement couplée à un calcul de champ : résolution de l'équation de Poisson, des équations de Maxwell. On utilise aussi parfois l'opérateur de gyromoyenne.

1 Etude de l'opérateur de gyromoyenne

On se base sur le travail [30].
L'opérateur de gyromoyenne est défini par

$$J(f)(x,y) = \frac{1}{2\pi}\int_0^{2\pi} f(x+\rho\cos(\theta), y+\rho\sin(\theta))d\theta.$$

Dans un champ magnétique uniforme, les particules décrivent une trajectoire hélicoïdale et la projection sur le plan perpendiculaire est un cercle. L'opérateur de gyromoyenne traduit alors, dans la théorie gyrocinétique, l'idée de moyenner la fonction de distribution des particules autour d'un cercle d'un rayon caractéristique (le rayon de Larmor ρ) représentant le mouvement de gyration très rapide des particules autour des lignes de champs. On s'intéresse ici à la résolution numérique de cet opérateur en présentant et comparant différentes méthodes numériques.
On suppose f 2π périodique en x et en y. On définit une grille cartésienne de taille $N_x \times N_y$.

1.1 Décomposition dans une base

On écrit

$$f(x,y) = \sum_{j,k} \eta_{j,k} B_{j,k}(x,y),$$

Le calcul de la gyromoyenne se réduit au calcul de la gyromoyenne sur les éléments de base

$$\mathcal{J}(f)(x,y) = \sum_{j,k} \eta_{j,k} \mathcal{J}(B_{j,k})(x,y).$$

1.2 Expression en Fourier

En prenant la base de Fourier

$$B_{j,k}(x,y) = \exp(ijx)\exp(iky),$$

on obtient

$$\mathcal{J}(B_{j,k}) = \mathcal{J}(B_{j,k})(0,0)B_{j,k} = J_0(\sqrt{j^2+k^2}\rho)B_{j,k},$$

où J_0 est la fonction de Bessel.
Sur la grille cartésienne, on obtient la méthode de Bessel, qui est exacte sur la grille pour les fonctions $B_{j,k}$. Il s'agit de la méthode de référence.

1.3 Approximation de Padé et autres variantes

On considère l'approximation

$$J_0(\sqrt{j^2+k^2}\rho) \simeq \left(1+\rho^2\frac{j^2+k^2}{4}\right)^{-1},$$

qui se traduit par

$$\left(1-\frac{\rho^2}{4}\Delta\right)\mathcal{J}(f) = f.$$

Sur la grille cartésienne, on obtient la méthode de Padé, notée PADE1. On définit de manière similaire PADE2 par

$$J_0(\sqrt{j^2+k^2}\rho) \simeq \left(1+\rho^2\frac{j^2+k^2}{4}+\rho^4\frac{(j^2+k^2)^2}{64}\right)^{-1},$$

TAYLOR1

$$J_0(\sqrt{j^2+k^2}\rho) \simeq 1-\rho^2\frac{j^2+k^2}{4},$$

et TAYLOR2

$$J_0(\sqrt{j^2+k^2}\rho) \simeq 1-\rho^2\frac{j^2+k^2}{4}+\rho^4\frac{(j^2+k^2)^2}{64}.$$

1.4 Approximation linéaire et par splines cubiques

On prend les fonctions de base linéaire (LIN) ou splines cubiques (SPL) sur la grille cartésienne.
Pour la quadrature des fonctions de bases, on discrétise $[0, 2\pi[$ de manière uniforme, pour obtenir les méthodes

$$\text{LIN4},\ \text{LIN8},\ \text{LIN16},\ \text{SPL4},\ \text{SPL8},\ \text{SPL16},$$

le numéro correspondant au nombre de points pris, ou de manière adaptative suivant le rayon (3 points de Gauss utilisés sur chaque arc d'intersection avec le maillage), pour obtenir les méthodes

$$\text{IM}-\text{LIN},\ \text{IM}-\text{SPL}$$

1.5 Comparaison des méthodes

Connaissant f sur la grille, chacune des méthodes permet de calculer $\mathcal{J}(f)$ sur la grille :

$$(f_{j,k}) \in \mathbb{R}^{N_x \times N_y} \to (\mathcal{J}(f)_{j,k}) \in \mathbb{R}^{N_x \times N_y}$$

En notant $\widehat{g}_{j,k}$ la transformation de Fourier discrète de $g_{j,k}$, pour la méthode de Bessel, on a

$$\frac{\widehat{J(f)}_{j,k}}{\widehat{f}_{j,k}} = J_0(\rho\sqrt{j^2+k^2}). \tag{1.1}$$

Les autres méthodes fournissent alors par le membre de gauche de (1.1) une approximation de la fonction de Bessel. Plus précisément, pour chaque méthode, on peut écrire

$$\mathcal{J}(f)_{r,s} = \sum_{j,k} a_{j,k} f_{j+r,k+s}.$$

Le terme $a_{j,k}$ correspond à la contribution du point d'indice (j,k) de la gyromoyenne au point $(0,0)$. On obtient alors

$$\widehat{\mathcal{J}(f)}_{j,k} = \overline{\widehat{a}}_{j,k}\widehat{f}_{j,k}.$$

1.6 Résultats numériques

On a ici une présentation d'un cadre général englobant la méthode de Bessel et les méthodes de quadrature. On compare ces méthodes dans le cas périodique, dans l'espace réel et de Fourier. Les comparaisons des différentes méthodes sont données sur les Figures 7.1, 7.2, 7.3,7.4, 7.5 et 7.6.
Les méthodes d'intégration montrent clairement leur avantage ; néanmoins, dans le cas où le rayon devient grand, le nombre de points de quadrature doit augmenter pour garder une bonne approximation, ce qui engendre a priori un surcoût, mais en même temps, la contribution des grands rayons est en général faible... Pour ces méthodes d'intégration, on introduit une formulation matricielle qui ramène le calcul de la gyromoyenne à celui des éléments de base de l'espace d'approximation choisi, ce qui peut être fait une fois pour toutes. Ainsi, dans le cas d'une géométrie cartésienne périodique, cette approche a un coût similaire à la méthode de Bessel, la matrice ayant dans ce cas là le bon goût d'être circulante.

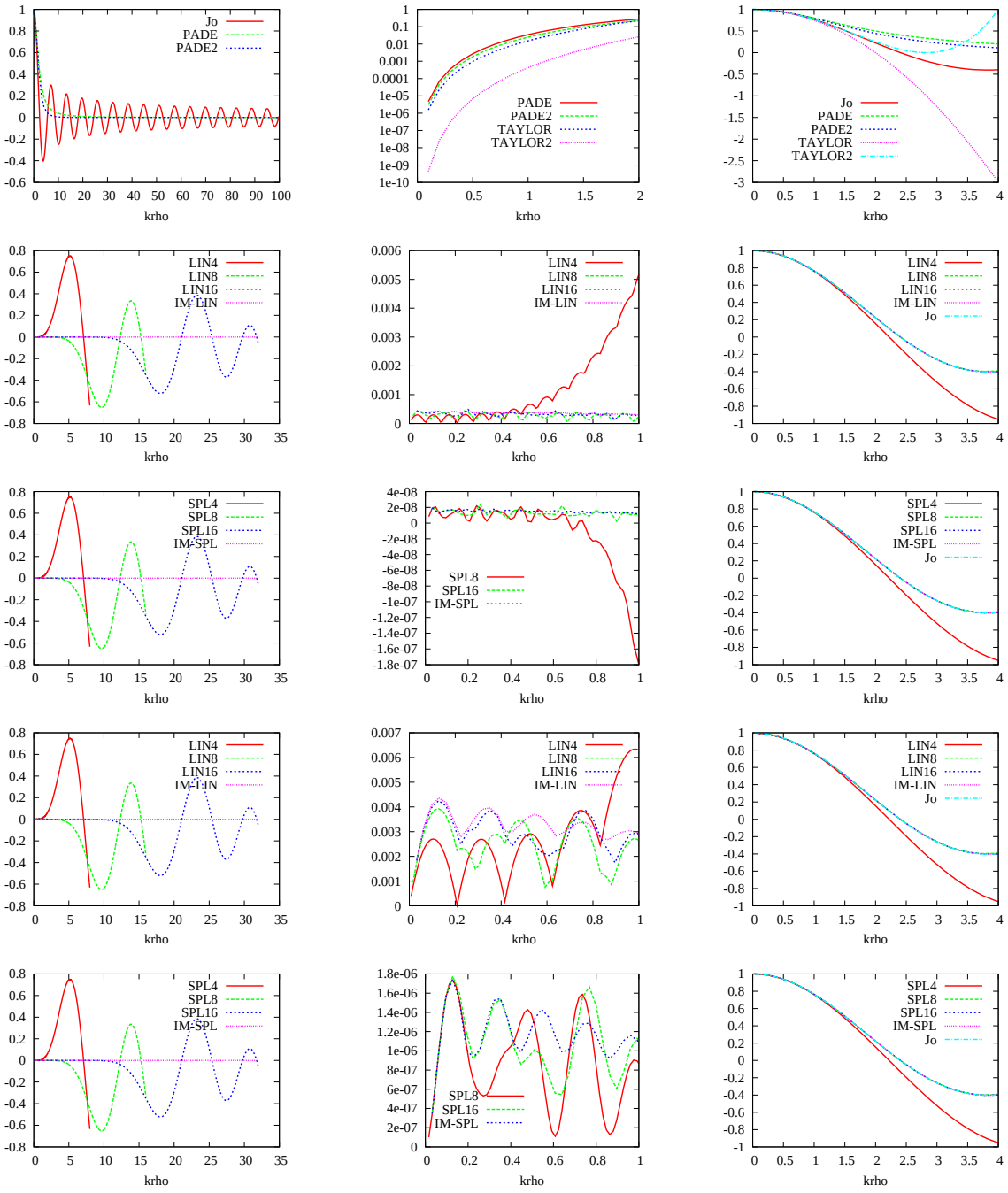

FIGURE 7.1 – Approximation de la fonction de Bessel et erreur pour différentes méthodes $N_x = N_y = 128$, $(j,k) = (1,1),(3,3)$.

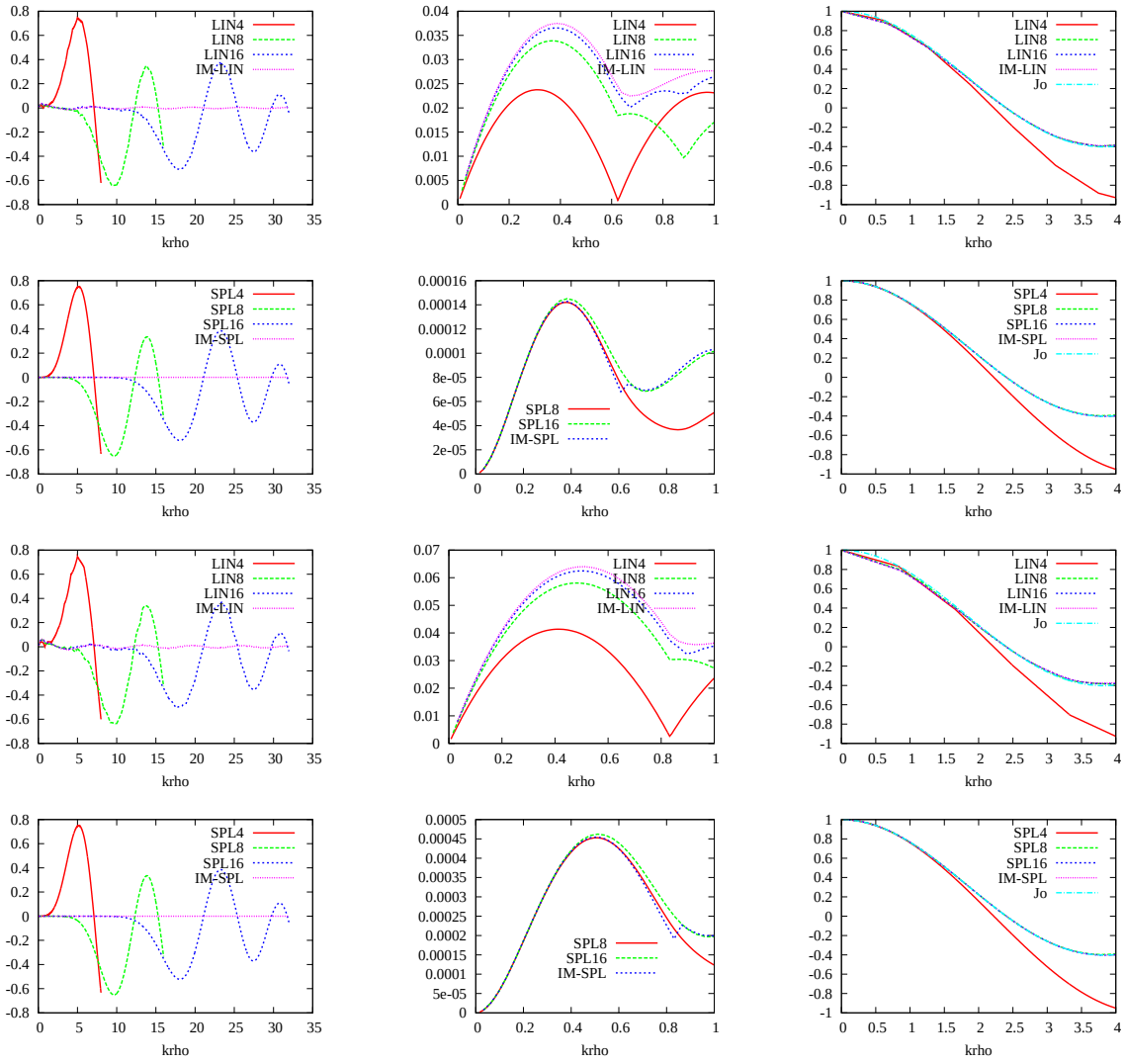

FIGURE 7.2 – Erreur pour différentes méthodes $N_x = N_y = 128$, $(j,k) = (9,9),(12,12)$.

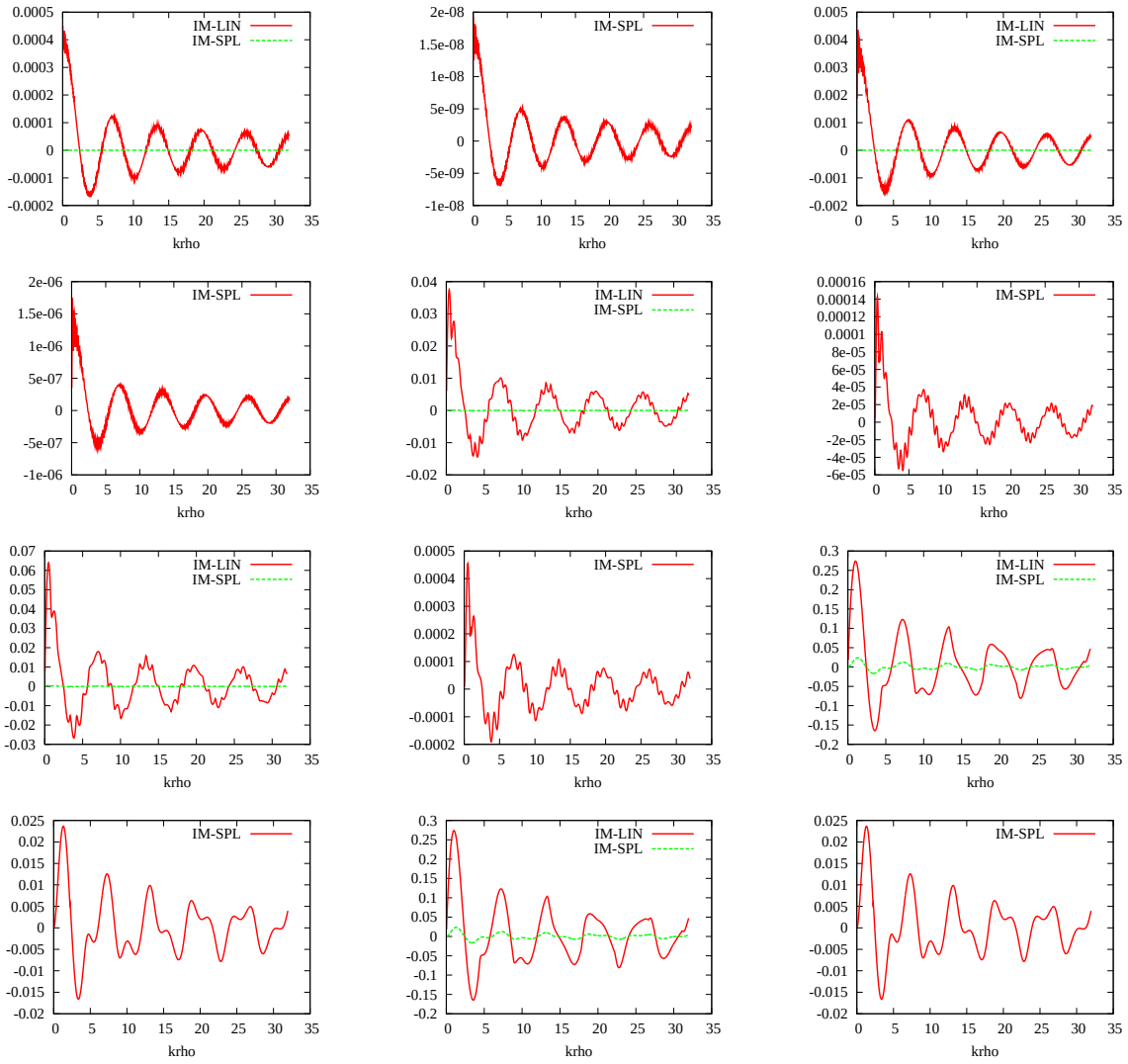

FIGURE 7.3 – Erreur avec la fonction de Bessel pour IM-LIN et IM-SPL $N_x = N_y = 128$, $(j,k) = (1,1),(3,3),(9,9),(12,12),(32,32),(64,64)$.

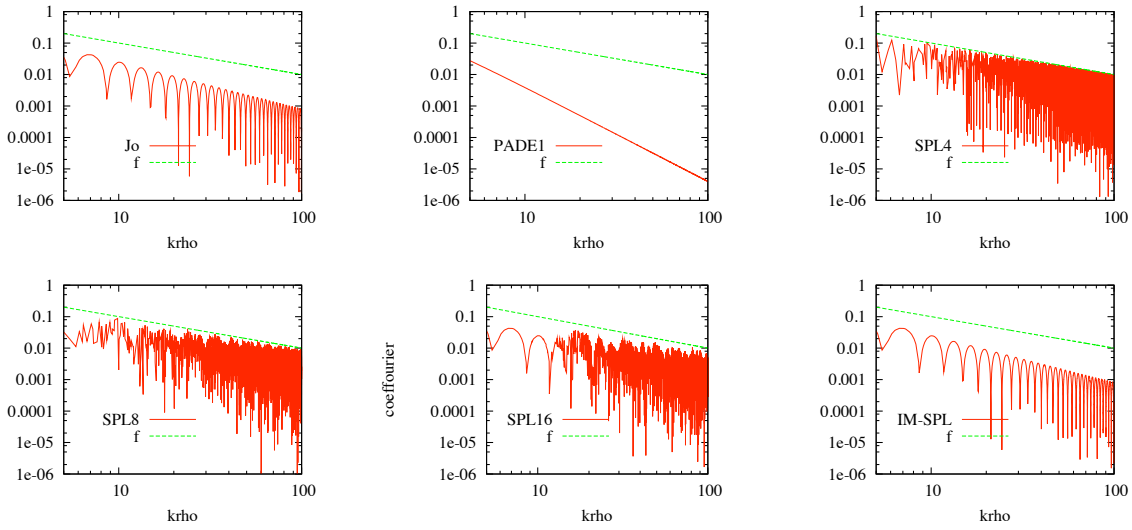

FIGURE 7.4 – Comparaison de la gyromoyenne et de la fonction dans l'espace de Fourier pour différentes méthodes $\rho = 1$.

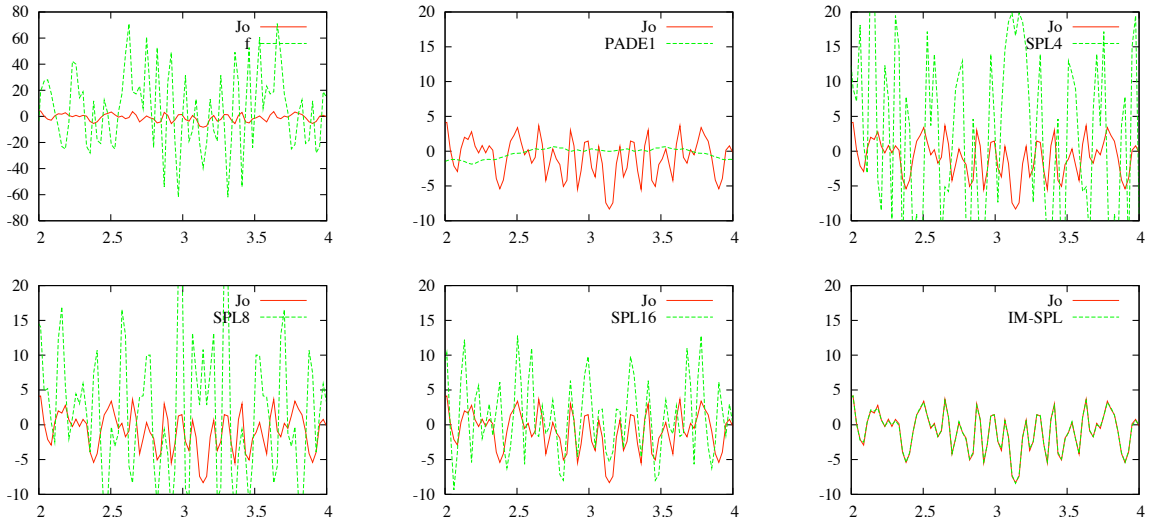

FIGURE 7.5 – Gyromoyenne en fonction de x, pour $y = \pi/2$. Comparaison de différentes méthodes. $\rho = 0.5$.

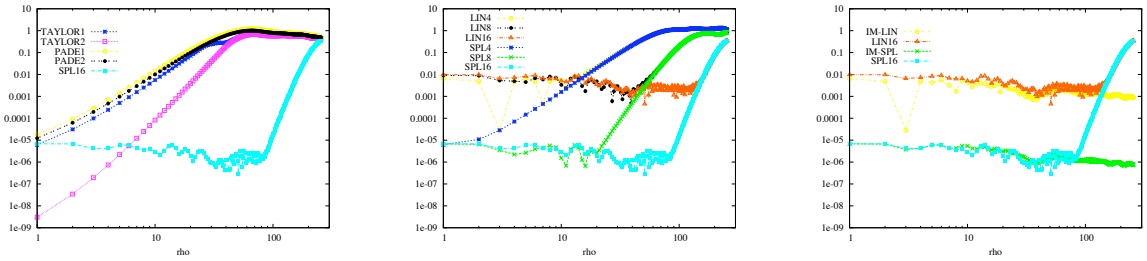

FIGURE 7.6 – Erreur L^1 (par rapport à la méthode de Bessel) en fonction de ρ (échelle log).

2 Résolution spectrale de l'équation de Poisson pour Galerkin Discontinu

On explique ici (cf [33]) comment calculer, avec précision spectrale le potentiel électrique Φ et le champ électrique $E = -\nabla_{x_1,x_2}\Phi$ obtenu en résolvant l'équation de Poisson

$$-\Delta_{x_1,x_2}\Phi = f.$$

La particularité ici est que f est représenté aux points de Gauss

$$f_{k_1,j_1,k_2,j_2} \simeq f(x_{k_1,j_1}, y_{k_2,j_2}).$$

On a ainsi la représentation DG de f sous la forme

$$f(x,y) = \sum_{k_1=0}^{N_x-1}\sum_{j_1=0}^{d}\sum_{k_2=0}^{N_y-1}\sum_{j_2=0}^{d} f_{k_1,j_1,k_2,j_2}\phi_{k_1,j_1}(x)\psi_{k_2,j_2}(y). \tag{2.1}$$

On représente maintenant f dans l'espace de Fourier par

$$f(x,y) = \sum_{(p,q)\in\mathbb{Z}\times\mathbb{Z}} a_{p,q} e^{ipx\frac{2\pi}{L_x}} e^{iqy\frac{2\pi}{L_y}}, \tag{2.2}$$

avec

$$a_{p,q} = \frac{1}{L_x L_y}\int_{x_{\min}}^{x_{\max}}\int_{y_{\min}}^{y_{\max}} f(x,y) e^{-ipx\frac{2\pi}{L_x}} e^{-iqy\frac{2\pi}{L_y}}\,\mathrm{d}x\mathrm{d}y,$$

$L_x = x_{\max} - x_{\min}$ est la longueur du domaine en x, et $L_y = y_{\max} - y_{\min}$ est la longueur du domaine en y. Après injection de la représentation (2.1), on obtient

$$a_{p,q} = \sum_{k_1,j_1,k_2,j_2} f_{k_1,j_1,k_2,j_2}\frac{1}{L_x L_y}\int_{x_{\min}}^{x_{\max}} \varphi_{k_1,j_1}(x) e^{-ipx\frac{2\pi}{L_x}}\,\mathrm{d}x \int_{y_{\min}}^{y_{\max}} \psi_{k_2,j_2}(y) e^{-iqy\frac{2\pi}{L_y}}\,\mathrm{d}y.$$

On peut alors tirer profit de la structure circulante de l'opération, en calculant certaines quantités en pré-calcul.
On introduit d'abord la transformée de Fourier discrète

$$\begin{aligned}
\mathcal{F}[f]_{p,j_1,q,j_2} &= \sum_{k_1'=0}^{N_x-1}\sum_{k_2'=0}^{N_y-1} f_{k_1',j_1,k_2',j_2} e^{-\frac{2\pi}{N_x}ipk_1'} e^{-\frac{2\pi}{N_y}iqk_2'} \\
\mathcal{F}^{-1}[f]_{k_1,j_1,k_2,j_2} &= \sum_{p=0}^{N_x-1}\sum_{q=0}^{N_y-1} f_{p,j_1,q,j_2} e^{\frac{2\pi}{N_x}ik_1p} e^{\frac{2\pi}{N_y}ik_2q}.
\end{aligned}$$

On peut alors écrire (avec $\star \in \{x,y\}$ et $E = (E^x, E^y)$)

$$\begin{aligned}
\tilde{\rho}^{\star}_{p'',j_1,q'',j_2} &:= \sum_{j_1'=0}^{d}\sum_{j_2'=0}^{d} \mathcal{F}[f]_{p'',j_1',q'',j_2'} S^{\star}_{p'',j_1',j_1,q'',j_2',j_2}, \\
E^{\star}_{k_1,j_1,k_2,j_2} &= \frac{-i}{\omega_{j_1}\omega_{j_2}N_xN_y}\mathcal{F}^{-1}[\tilde{\rho}^{\star}]_{k_1,j_1,k_2,j_2}., \qquad (2.3)
\end{aligned}$$

et les quantités à précalculer sont

$$S^{\{x,y\}}_{p'',j'_1,j_1,q'',j'_2,j_2} := \sum_{(p',q')\in\mathbb{Z}\times\mathbb{Z}} \begin{cases} 0 & \text{if } (p',p'',q',q'') = (0,0,0,0) \\ \frac{\left\{(p'N_x+p'')\frac{2\pi}{L_x},(q'N_y+q'')\frac{2\pi}{L_y}\right\}}{\left((p'N_x+p'')\frac{2\pi}{L_x}\right)^2+\left((q'N_y+q'')\frac{2\pi}{L_y}\right)^2} & \\ \times I_{j_1,(p'N_x+p'')\frac{2\pi}{N_x}} I_{j'_1,-(p'N_x+p'')\frac{2\pi}{N_x}} & \\ \times I_{j_2,(q'N_y+q'')\frac{2\pi}{N_y}} I_{j'_2,-(q'N_y+q'')\frac{2\pi}{N_y}} & \text{sinon,} \end{cases}.$$

avec

$$I_{j,\alpha} := \int_0^1 \phi^j(s) e^{i\alpha s} \mathrm{d}s, \text{ où } \phi^j \text{ est défini par (3.1).}$$

3 Compatibilité entre le champ et l'advection

On dérive ici une condition du premier ordre pour avoir une meilleur compatibilité entre le calcul du champ et l'advection. Notons le travail [15] où le fait d'imposer une condition de divergence nulle discrète s'est avéré crucial dans le cadre d'une simulation de Vlasov par une méthode semi-Lagrangienne conservative. On se place dans le cadre du modèle centre guide 2D.

$$\partial_t f(t,x,y) - \partial_x(E_y f) + \partial_y(E_x f) = 0,\ E = -\nabla\Phi,\ -\Delta\Phi = f. \tag{3.1}$$

On introduit les caractéristiques $X(t;s,x,y) = X(t), Y(t;,s,x,y) = Y(t)$, avec

$$X'(t) = -E_y(X(t),Y(t)),\ Y'(t) = E_x(X(t),Y(t)),\ X(s) = x,\ Y(s) = y. \tag{3.2}$$

Au niveau continu, on sait que le champ est à divergence nulle :

$$\partial_x E_y - \partial_x E_x = 0,$$

de telle sorte que l'équation (3.1) est équivalente à (3.3) :

$$\partial_t f(t,x,y) - E_y\partial_x f + E_x\partial_y f = 0,\ E = -\nabla\Phi,\ -\Delta\Phi = f. \tag{3.3}$$

Grâce à la propriété de divergence nulle, les caractéristiques satisfont la propriété de conservation de la masse :

$$\int g(s,x,y)dxdy = \int g(t,X(t;s,x,y),Y(t;s,x,y))dxdy,\ \forall s,t\in\mathbb{R},$$

et aussi la *préservation des constantes* :

$$g(s,x,y) = 1, \forall x,y \Rightarrow g(t,x,y) = g(s,X(s;t,x,y),Y(s;t,x,y)) = 1.$$

Une méthode semi-Lagrangienne peut être vue comme une succession d'interpolations (ou projections) et de transport (en utilisant les caractéristiques). Étant donné une fonction de distribution constante au temps t_n, en utilisant les caractéristiques du schéma, la fonction de distribution devrait rester constante au temps t_{n+1}. Ceci est l'analogue de la préservation des constantes au niveau discret. D'autre part,

étant donné une fonction de distribution arbitraire au temps t_n, cette fonction devrait avoir la même masse au temps t_{n+1} ; ceci est l'analogue de la conservation de la masse au niveau discret. Généralement, les schémas semi-Lagrangiens ne conservent pas les deux conditions. Plus précisément, souvent une des conditions est satisfaite, mais l'autre n'est vraie qu'approximativement. On va ici dériver une condition de divergence discrète sur le champ, dépendant de la méthode semi-Lagrangienne qui est utilisée, de telle sorte que cette dernière soit satisfaite à l'ordre 1 en temps. Par la suite, on va considérer plusieurs types de schémas semi-Lagrangiens et expliquer comment dériver cette condition.

3.1 Méthodes semi-Lagrangiennes 2D en arrière

On fixe une grille uniforme $(x_i, y_j)_{i=0,\dots,N_x, j=0,\dots,N_y}$, du domaine périodique $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$. On suppose qu'au temps t_n, la fonction de distribution s'écrit

$$f_n(x,y) = \sum_{i,j} \gamma_{n,i,j} \phi_{i,j}(x,y),$$

et $f_n(x_k, y_\ell) \simeq f(t_n, x_k, y_\ell)$. La fonction de distributions est alors mise à jour de la manière suivante :

$$f_{n+1}(x_k, y_\ell) = f_n(X_h(t_n; t_{n+1}, x_k, y_\ell), Y_h(t_n; t_{n+1}, x_k, y_\ell)),$$

où (X_h, Y_h) sont les caracteristiques qui résolvent numériquement (3.2). Finalement f_{n+1} est obtenu en résolvant le système suivant (on suppose qu'il y a unique solution)

$$f_{n+1}(x_k, y_\ell) = \sum_{i,j} \gamma_{n+1,i,j} \phi_{i,j}(x_k, y_\ell), \;\; k = 0, \dots, N_x - 1, \;\; \ell = 0, \dots, N_y - 1.$$

Préservation des constantes On peut vérifier que la préservation des constantes est satisfaite : on prend une fonction qui est constante sur la grille au temps t_n

$$g_n(x_k, y_\ell) = 1 \;\; k = 0, \dots, N_x - 1, \;\; \ell = 0, \dots, N_y - 1.$$

On calcule ensuite $\gamma_{n,i,j}$ avec

$$\sum_{i,j} \gamma_{n,i,j} \phi_{i,j}(x_k, y_\ell) = 1, \;\; k = 0, \dots, N_x - 1, \;\; \ell = 0, \dots, N_y - 1.$$

On a alors

$$g_n(x,y) = \sum_{i,j} \gamma_{n,i,j} \phi_{i,j}(x,y).$$

Au temps t_{n+1}, on obtient

$$g_{n+1}(x_k, y_\ell) = g_n(X_h(t_n; t_{n+1}, x_k, y_\ell), Y_h(t_n; t_{n+1}, x_k, y_\ell)).$$

et donc les constantes sont préservées, si on a

$$g_n(x,y) = 1, \;\; x, y \in \Omega.$$

Pour cela, il suffit que la fonction 1 soit générée par les fonctions de bases $\phi_{i,j}$, ce qui est généralement le cas pour une telle méthode semi-Lagrangienne.

Conservation de la masse On considère maintenant une fonction arbitraire sur la grille au temps t_n. Sans perte de généralité, on peut prendre une fonction de base $\phi_{i0,j0}$. La masse de la fonciton de base au temps t_n sur une grille uniforme est

$$M(t_n) = \sum_{i,j} \phi_{i0,j0}(x_i, y_j)$$

Au temps t_{n+1}, la masse devient

$$M(t_{n+1}) = \sum_{i,j} \phi_{i0,j0}(X_h(t_n; t_{n+1}, x_i, y_j), Y_h(t_n; t_{n+1}, x_i, y_j)).$$

Cette fois-ci, on n'a pas généralement $M(t_n) = M(t_{n+1})$, cette valeur dépend grandement de la façon dont sont approchées les caractéristiques. Cependant, on peut imposer une condition de conservation de masse du *premier ordre* en imposant

$$M_n'(0) = 0, \text{ with } M_n(s) = M(t_n + s),$$

ce qui donne

$$\sum_{i,j} E_{y,i,j} \partial_x \phi_{i0,j0}(x_i, y_j) - \sum_{i,j} E_{x,i,j} \partial_y \phi_{i0,j0}(x_i, y_j) = 0. \tag{3.4}$$

Un calcul de champ à divergence nulle Le champ électrique peut ne pas satisfaire la condition de divergence nulle (3.4). On cherche ici un calcul qui satisfait naturellement le condition de divergence nulle. On définit

$$E_x(x, y) = \sum_{i,j} E_{x,i,j} \phi_{i,j}(x, y), \; E_y(x, y) = \sum_{i,j} E_{y,i,j} \phi_{i,j}(x, y),$$

tel que

$$E_x(x, y) = -\partial_x \Phi(x, y), \; E_y(x, y) = -\partial_y \Phi(x, y).$$

Pour cela, on suppose que l'on a

$$\Phi(x, y) = \sum_{i,j} \Phi_{i,j} \phi_{i,j}(x, y),$$

ce qui donne

$$E_x(x, y) = -\sum_{i,j} \Phi_{i,j} \partial_x \phi_{i,j}(x, y), \; E_y(x, y) = -\sum_{i,j} \Phi_{i,j} \partial_y \phi_{i,j}(x, y).$$

On obtient alors

$$\begin{aligned} &\sum_{i,j} E_{y,i,j} \partial_x \phi_{i0,j0}(x_i, y_j) - \sum_{i,j} E_{x,i,j} \partial_y \phi_{i0,j0}(x_i, y_j) \\ &= -\sum_{i,j} \sum_{k,\ell} \Phi_{k,\ell} \partial_y \phi_{k,\ell}(x_i, y_j) \partial_x \phi_{i0,j0}(x_i, y_j) + \sum_{i,j} \sum_{k,\ell} \Phi_{k,\ell} \partial_x \phi_{k,\ell}(x_i, y_j) \partial_y \phi_{i0,j0}(x_i, y_j). \end{aligned}$$

Afin que la somme soit nulle, on doit avoir

$$\sum_{i,j} \partial_y \phi_{k,\ell}(x_i, y_j) \partial_x \phi_{i0,j0}(x_i, y_j) - \sum_{i,j} \partial_x \phi_{k,\ell}(x_i, y_j) \partial_y \phi_{i0,j0}(x_i, y_j) = 0.$$

Si l'on suppose que $\phi_{k,\ell}(x,y) = \phi_k(x)\psi_\ell(y)$, et que $\phi_k(x_i) = \psi_k(y_i) = \delta_{i,k}$ et que $\phi_k(x) = \phi(\frac{x-x_k}{\Delta x})$ et $\psi_\ell(y) = \phi(\frac{y-y_\ell}{\Delta y})$, on obtient

$$\sum_{i,j} \partial_y\phi_{k,\ell}(x_i,y_j)\partial_x\phi_{i0,j0}(x_i,y_j) - \sum_{i,j} \partial_x\phi_{k,\ell}(x_i,y_j)\partial_y\phi_{i0,j0}(x_i,y_j)$$
$$= \psi'_\ell(y_{j0})\phi'_{i0}(x_k) - \phi'_k(x_{i0})\psi'_{j0}(y_\ell) = \phi'(j0-\ell)\phi'(k-i0) - \phi'(-j0+\ell)\phi'(-k+i0).$$

Si $\phi(x) = \phi(-x)$, la somme est nulle.

3.2 Méthode semi-Lagrangiennes conservatives splittées

On suppose connaitre

$$f_{n,i,j} \simeq \frac{1}{\Delta x\Delta y}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} f(t_n,x,y)dxdy.$$

On a une représentation continue

$$f_n(x,y) = \sum_{i,j} f_{n,i,j}\phi_{i,j}(x,y).$$

Afin de calculer la solution au temps $t_{n+1} = t_n + s$, on calcule d'abord

$$\tilde{f}_{n,i,j} = \frac{1}{\Delta x\Delta y}\int_{X_{i-1/2,j}(s)}^{X_{i+1/2,j}(s)}\int_{y_{j-1/2}}^{y_{j+1/2}} f_n(x,y)dydx,$$

avec

$$X'_{i-1/2,j}(s) = -E_y(X_{i-1/2,j}(s),y_j),\ X_{i-1/2,j}(0) = x_{i-1/2}.$$

Ce dernier système est obtenu en faisant le changement de variable $X(s) = \tilde{X}(t_{n+1}-s)$, où $\tilde{X}$ satisfait

$$\tilde{X}'(t) = E_y(\tilde{X}(t)),\ \tilde{X}(t_{n+1}) = x_{i-1/2}.$$

On calcule alors

$$f_{n+1,i,j} = \frac{1}{\Delta x\Delta y}\int_{Y_{i,j-1/2}(s)}^{Y_{i,j+1/2}(s)}\int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}_n(x,y)dxdy,$$

avec

$$Y'_{i,j-1/2}(s) = E_x(x_i,Y_{i,j-1/2}(s)),\ Y_{i,j-1/2}(0) = y_{j-1/2}.$$

On a considéré ici un cas particulier de splitting : Δt en x et Δt en y, mais on aurait pu considérer des stratégies plus générales, comme par exemple $\Delta t/2$ pour x, Δt pour y et de nouveau $\Delta t/2$ pour x.
Le schéma conserve par définition la masse ; on va donc regarder ce qu'il en est pour la préservation des constantes.

Préservation des constantes On a

$$M_{i,j}^n = \frac{1}{\Delta x \Delta y} \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{y_{j-1/2}}^{y_{j+1/2}} f(x,y) dx dy.$$

Si l'on prend une fonction constante $f = 1$, on obtient alors

$$M_{i,j}^n = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} dx = \frac{1}{\Delta x} \int_{X_{i-1/2,j}(s)}^{X_{i+1/2,j}(s)} dx,$$

c'est-à-dire

$$M_{i,j}^n = \frac{X_{i+1/2,j}(s) - X_{i-1/2,j}(s)}{\Delta x}.$$

Ainsi, la fonction après le splitting devient

$$f_i(y) = \sum_k M_{i,k} \varphi_k(y),$$

où $M_{i,j}^n = (1/\Delta y) \int_{Y_{i,j-1/2}(s)}^{Y_{i,j+1/2}(s)} f_i(y) dy$.
On a alors

$$\begin{aligned} M^{n+1}(s) &= \frac{1}{\Delta y} \int_{Y_{i,j-1/2}(s)}^{Y_{i,j+1/2}(s)} f_i(y) dy \\ &= \frac{1}{\Delta y} \sum_k M_{i,k} \int_{Y_{i,j-1/2}(s)}^{Y_{i,j+1/2}(s)} \varphi_k(y) dy, \end{aligned}$$

de telle sorte que

$$M_{i,j}^{n+1}(s) = \sum_k \left(\frac{X_{i+1/2,k}(s) - X_{i-1/2,k}(s)}{\Delta x} \right) \times \frac{1}{\Delta y} \int_{Y_{i,j-1/2}(s)}^{Y_{i,j+1/2}(s)} \varphi_k(y) dy,$$

et

$$\begin{aligned} -\frac{d}{ds} M_{i,j}^{n+1}(0) &= \sum_k \frac{E_{i+1/2,k}^y - E_{i-1/2,k}^y}{\Delta x} \times \delta_{j,k} \\ &\quad + \sum_k 1 \times \frac{1}{\Delta y} \left(\varphi_k(y_{j+1/2})(-E_{i,j-1/2}^x) - \varphi_k(y_{j-1/2})(-E_{i,j-1/2}^x) \right) \\ &= \frac{E_{i+1/2,j}^y - E_{i-1/2,j}^y}{\Delta x} - \frac{E_{i,j+1/2}^x - E_{i,j-1/2}^x}{\Delta y} \end{aligned}$$

La condition de divergence nulle devient alors

$$\frac{E_{i+1/2,j}^y - E_{i-1/2,j}^y}{\Delta x} - \frac{E_{i,j+1/2}^x - E_{i,j-1/2}^x}{\Delta y} = 0. \tag{3.5}$$

Un calcul de champ à divergence nulle On suppose que le potentiel s'écrit

$$\Phi(x,y) = \sum_{i,j} \Phi_{i,j}\phi_{i,j}(x,y),$$

et on calcule le champ électrique de la manière suivante :

$$E_y(x,y_j) = -\frac{1}{\Delta y}\int_{y_{j-1/2}}^{y_{j+1/2}} \partial_y\Phi(x,y)dy = -\frac{1}{\Delta y}\left(\Phi(x,y_{j+1/2}) - \Phi(x,y_{j-1/2})\right) \quad (3.6)$$

et

$$E_x(x_i,y) = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \partial_x\Phi(x,y)dx = \frac{1}{\Delta x}\left(\Phi(x_{i+1/2},y) - \Phi(x_{i-1/2},y)\right). \quad (3.7)$$

On obtient alors

$$\begin{aligned}
&\Delta x\Delta y\left(\frac{E^y_{i+1/2,j} - E^y_{i-1/2,j}}{\Delta x} - \frac{E^x_{i,j+1/2} - E^x_{i,j-1/2}}{\Delta y}\right)\\
&\quad = \left(\Phi(x_{i+1/2},y_{j+1/2}) - \Phi(x_{i+1/2},y_{j-1/2}) - \Phi(x_{i-1/2},y_{j+1/2}) + \Phi(x_{i-1/2},y_{j-1/2})\right)\\
&\quad - \left(\Phi(x_{i+1/2},y_{j+1/2}) - \Phi(x_{i-1/2},y_{j+1/2}) - \Phi(x_{i+1/2},y_{j-1/2}) + \Phi(x_{i-1/2},y_{j-1/2})\right)\\
&\quad = 0.
\end{aligned}$$

Reconstruction PSM Pour l'exemple de la méthode PSM, les valeurs $\Phi_{i,j}$ sont calculées à partir de $f_{n,i,j}$ et donc $\Phi_{i,j}$ represente une valeur moyenne sur la maille $[x_{i-1/2},x_{i+1/2}]\times[y_{j-1/2},y_{j+1/2}]$:

$$\Phi_{i,j} = \frac{1}{\Delta x\Delta y}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} \Phi(x,y)dxdy$$

On calcule d'abord les valeurs moyennes dans la direction x

$$\Phi_{i,j+1/2} = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \Phi(x,y_{j+1/2})dx,$$

en résolvant le système suivant

$$\Phi_{i,j-1/2} + 4\Phi_{i,j+1/2} + \Phi_{i,j+3/2} = 3(\Phi_{i,j} + \Phi_{i,j+1}).$$

On calcule ensuite les valeurs ponctuelles $\Phi_{i+1/2,j+1/2} \simeq \Phi(x_{i+1/2},y_{j+1/2})$

$$\Phi_{i-1/2,j+1/2} + 4\Phi_{i+1/2,j+1/2} + \Phi_{i+3/2,j+1/2} = 3(\Phi_{i,j+1/2} + \Phi_{i+1,j+1/2}).$$

On définit ensuite

$$E_{y,i+1/2,j} = -\frac{1}{\Delta y}(\Phi_{i+1/2,j+1/2} - \Phi_{i+1/2,j-1/2}),$$

et

$$E_{y,i,j} = -\frac{1}{\Delta y}(\Phi_{i,j+1/2} - \Phi_{i,j-1/2}),$$

de telle sorte que l'on obtient une reconstruction explicite de

$$\begin{aligned}
E_y(x,y_j) &= E_{y,i-1/2,j} + \alpha(E_{y,i+1/2,j} - E_{y,i-1/2,j})\\
&\quad + \alpha(1-\alpha)(6E_{y,i,j} - 3E_{y,i+1/2,j} - 3E_{y,i-1/2,j}),\ x = x_{i-1/2} + \alpha\Delta x,\ \alpha\in[0,1].
\end{aligned}$$

Un calcul similaire peut se faire pour E_x.

3.3 Illustration numérique

On a considéré le modèle centre guide et testé la méthode BSL avec reconstruction du champ par Fourier et reconstruction du champ par splines (à partir du potentiel en Fourier). On remarque que la conservation de la masse est bien liée à cette conditon de divergence nulle. En particulier, a priori, on aurait pu supposer que le champ serait mieux calculé en utilisant la méthode de Fourier (précision spectrale) ; mais on voit ici qu'il est préférable d'utiliser les splines (ce qui est d'ailleurs habituellement fait) qui sont plus compatibles, la relation de divergence discrète étant satsfaite.

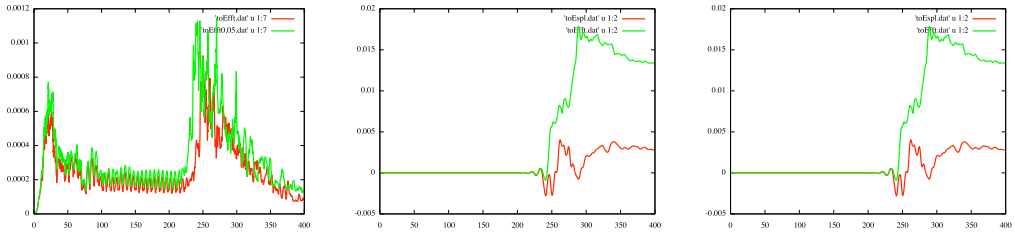

FIGURE 7.7 – Divergence discrète pour le champ électrique calculé par Fourier avec $dt = 0.1$ (en vert) et $dt = 0.05$ (en rouge) (à gauche) ; conservation de la masse pour Fourier (en vert) et les splines (en rouge) $dt = 0.1$ (au milieu) et $dt = 0.05$ (à droite) en fonction du temps $Nx = Ny = 64$, pour une simulation du modèle centre guide

Travaux en cours/Perspectives

Participation à l'encadrement doctoral, post-doctoral ou de stages de recherche

Gyromoyenne L'opérateur de gyromoyenne a été étudié dans le cadre d'une géométrie cartésienne. Des adaptations doivent être faites pour d'autres géométries (thèse de C. Steiner, débutée en septembre 2011).

Méthodes conservatives multi-dimensionnelles Les méthodes conservatives ont été développées dans le cadre mono-dimensionnel avec splitting directionnel. Le splitting casse certaines propriétés (principe du maximum, conservation du volume...). Il est alors intéressant de développer des méthodes conservatives multi-dimensionnelles (thèse de P. Glanc, débutée en octobre 2010).

Méthodes semi-Lagrangiennes en avant (FSL) et sur maillage curviligne, maillage adapté aux invariants L'étude des méthodes semi-Lagrangiennes dans le cadre d'un maillage curviligne présentent quelques complications et nécessitent donc d'être revisitées. On s'intéresse aussi au cas le maillage suit les isolignes du hamiltonien, ce qui peut-être utile lorsque l'on recherche des instabilités proches d'un équilibre (postdoc M. Bergot, H. Sellama, 2011-2012).

Equations équivalentes pour le transport L'étude d'équations équivalentes permet d'aider à comprendre et comparer les opérateurs d'interpolation pour le transport linéaire (stage de M2, C. Steiner, 2011).

Développements de méthodes semi-Lagrangiennes en géométrie polaire On s'intéresse au développement de cas tests 2D en géométrie circulaire pour pouvoir tester les méthodes numériques plus facilement avant intégration éventuelle dans le code GYSELA (stage de fin de deuxième année d'école d'ingénieur, E. Madaule, 2012).

Autres axes de recherche

Convergence en temps et en espace pour Vlasov-Poisson Dans le manuscript, on s'est limité au cas d'un splitting de Strang. On attend un résultat plus général, mêlant l'ordre en espace et l'ordre en temps

Séparation échelle rapide/lente Il s'agit de développer des méthodes qui permettent de séparer une dynamique rapide linéaire d'une dynamique non linéaire lente. Autour de ces questions en lien avec le maillage curviligne, se posent les questions de splitting, intégration géométrique, de propriété de divergence nulle, conservation de la masse et d'autres quantités.

Etude et développement de méthode semi-Lagrangiennes Il reste de nombreuses questions sur la convergence des méthodes semi-Lagrangiennes : cas de l'advection non constante, étude de la stabilité de certaines reconstructions, développements de nouvelles reconstructions, maillages uniformes, non uniformes, non structurés, adaptatifs, curvilignes...

Développement logiciel Participation à SELALIB [101], bibliothèque semi Lagrangienne pour la simulation gyrocinétique, en lien avec le code GYSELA.

Bibliographie

[1] F. Alauzet, M. Mehrenberger, *P1-conservative solution interpolation on unstructured triangular meshes*, Int. J. Numer. Meth. Engng (2010, online).

[2] K. Ammari, M. Jellouli, M. Mehrenberger, *Feedback stabilization of a coupled string-beam system.* Netw. Heterog. Media 4 (2009), no. 1, 19–34.

[3] K. Ammari, M. Mehrenberger, *Study of the nodal feedback stabilization of a string-beams network*, Journal of Applied Mathematics and Computing (16 June 2010, online).

[4] K. Ammari, M. Mehrenberger, *Stabilization of coupled systems,* Acta Math. Hungar. 123 (2009), no. 1-2, 1–10.

[5] C. Baiocchi, V. Komornik, P. Loreti, *Ingham–Beurling type theorems with weakened gap conditions*, Acta Math. Hungar. 97 (1–2) (2002), 55–95.

[6] J. W. Banks, J. A. F. Hittinger, *A new class of nonlinear finite-volume methods for Vlasov simulation*, IEEE Trans. Plasma Sc **38**, (2010).

[7] A. Barhoumi, V. Komornik, M. Mehrenberger, *A vectorial Ingham-Beurling type theorem*, Ann. Univ. Sci. Budapest. Eötvös Sect. Math. 53 (2010), 17–32.

[8] N. Besse, F. Filbet, M. Gutnic, I. Paun, E. Sonnendrücker, *Adaptive numerical method for the Vlasov equation based on a multiresolution analysis,* In F. Brezzi, A. Buffa, S. Escorsaro, and A. Murli editors, Numerical Mathematics and Advanced Applications ENUMATH 01, 437-446, Springer 2001.

[9] N. Besse, M. Mehrenberger, *Convergence of classes of high order semi-Lagrangian schemes for the Vlasov equation,* Math. of Comp. 77 (2008), 93-123.

[10] N. Besse, E. Sonnendrücker, *Semi-Lagrangian schemes for the Vlasov equation on an unstructured mesh of phase space*, J. Comput. Phys., 191 (2003), 341-376.,

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

[12] S. Blanes, F. Casas, A. Murua, *Splitting and composition methods in the numerical integration of differential equations*, Bol. Soc. Esp. Mat. Apl. 45 (2008), 89-145.

[13] S. Blanes, P. C. Moan, *Practical Symplectic Partitioned Runge-Kutta and Runge-Kutta-Nystrom Methods*, J. Comput. Appl. Math. 142 (2002), 313-330.

[14] A. S. Bonnet-BenDhia, S. Fliss, P. Joly, P. Moireau, *Introduction aux équations aux dérivées partielles et à leur approximation numérique*, polycopié, cours ENSTA, 2011.

[15] J. P. Braeunig, N. Crouseilles, V. Grandgirard, G. Latu, M. Mehrenberger, E. Sonnendrücker, *Some numerical aspects of the conservative PSM scheme in a 4D drift-kinetic code,* INRIA research report number 7109, November 2009.

[16] J. P. Braeunig, N. Crouseilles, M. Mehrenberger, E. Sonnendrücker, *Guiding-center simulations on curvilinear meshes*, Discrete and Continuous Dynamical Systems Series S, Volume 5, Number 3, June 2012.

[17] M. Campos Pinto, M. Mehrenberger, *Adaptive numerical resolution of the Vlasov equation,* Numerical Methods for Hyperbolic and Kinetic Problems, CEMRACS 2003/IRMA Lectures in Mathematics and Theoretical Physics 7, 43–58.

[18] M. Campos Pinto, M. Mehrenberger, *Convergence of an adaptive semi-Lagrangian scheme for the Vlasov-Poisson system,* Numerische Mathematik 108 (2008), no. 3, pp. 407-444.

[19] C. Castro, S. Micu, *Boundary controllability of a linear semi-discrete 1D wave equation derived from a mixed finite element method*, Numer. Math. 102 (3) (2006), 413–462.

[20] Edwin Chacon-Golcher, *Selalib Coding Guidelines*, documents pour selalib, https ://gforge.inria.fr/

[21] F. Charles, B. Després, M. Mehrenberger, *Enhanced convergence estimates for semi-Lagrangian schemes Application to the Vlasov-Poisson equation,* inria-00629081, version 1, October 2011.

[22] C.Z. Cheng, G. Knorr, *The integration of the Vlasov equation in configuration space*, J. Comput. Phys. 22, pp. 330-351, (1976).

[23] A. Cohen, *Numerical analysis of wavelet methods*, studies in mathematics and its applications, 32, North-Holland, Elsevier, 2003.

[24] P. Colella, P. R. Woodward, *The piecewise parabolic method (PPM) for gas-dynamical simulations*, J. Comput. Phys., **54**, 174-201, (1984).

[25] P. Colella, M. D. Sekora, *A limiter for PPM that preserves accuracy at smooth extrema*, J. Comput. Phys., **227**, 7069-7076, (2008).

[26] F. Coquel, Ph. Helluy, J. Schneider, *Second order entropy diminishing scheme for the Euler equations*, Intern. J. Numer. Meth. in Fluids, **50**, pp. 1029-1061, (2006).

[27] N. Crouseilles, E. Faou, M. Mehrenberger, *High order Runge-Kutta-Nyström splitting methods for the Vlasov- Poisson equation,* inria-00633934, version 1, October 2011.

[28] N. Crouseilles, P. Glanc, M. Mehrenberger, C. Steiner *Finite Volume Schemes for Vlasov,* hal-00653038, version 1, December 2011.

[29] N. Crouseilles, G. Latu, E. Sonnendrücker *A Vlasov solver based on local cubic spline interpolation on patches,* J. Comput. Phys., 228, pp. 1429-1446, (2009).

[30] N. Crouseilles, M. Mehrenberger, H. Sellama, *Numerical solution of the gyroaverage operator for the finite gyroradius guiding-center model,* CiCP 8, pp. 484-510, (2010).

[31] N. Crouseilles, M. Mehrenberger, E. Sonnendrücker, *Conservative semi-Lagrangian schemes for Vlasov equations*, J. Comput. Phys. 229 (2010), 1927-1953.

[32] N. Crouseilles, M. Mehrenberger, F. Vecil, *Discontinuous Galerkin semi-Lagrangian method for Vlasov-Poisson*, ESAIM Proc, CEMRACS 2010, to appear.

[33] N. Crouseilles, M. Mehrenberger, F. Vecil, *A Discontinuous Galerkin semi-Lagrangian solver for the guiding-center problem*, submitted.

[34] N. Crouseilles, T. Respaud, E. Sonnendrücker, *A forward semi-Lagrangian scheme for the numerical solution of the Vlasov equation*, Comput. Phys. Comm. **180**, pp. 1730-1745 (2009).

[35] N. Crouseilles, T. Respaud, *A charge preserving scheme for the numerical resolution of the Vlasov-Ampere equations*, Commun. Comput. Phys. **10** (2011), pp. 1001-1026.

[36] B. Després, *Finite Volume Transport Schemes,* Numer. Math. 108 (2008), no. 4, 529-556.

[37] B. Després, *Uniform asymptotic stability of Strang's explicit schemes*, SIAM J. Numer. Anal. 47, pp. 3956-3976, (2009).

[38] M. Falcone, R. Ferretti *Convergence Analysis for a Class of High-Order Semi-Lagrangian Advection Schemes*, SIAM Journal on Numerical Analysis, Volume 35 Issue 3, June 1998.

[39] M. Falcone, R. Ferretti, and T. Manfroni, *Optimal discretization steps in semi- Lagrangian approximation of first order PDEs,* in Numerical Methods for Viscosity Solutions and Applications (Heraklion, 1999), Ser. Adv. Math. Appl. Sci. **59**, M. Falcone and C. Makridakis, eds., World Scientific, River Edge, NJ, 2001, pp. 95-117.

[40] R. Ferretti *Equivalence of semi-Lagrangian and Lagrange-Galerkin schemes under constant advection speed*, Journal of Computational Mathematics, 2010,V28(4) : 461-473.

[41] F. Filbet, E. Sonnendrücker, P. Bertrand, *Conservative numerical schemes for the Vlasov equation*, J. Comput. Phys., **172**, pp. 166-187, (2001).

[42] R. Glowinski, C. H. Li and J. L. Lions, *A numerical approach to the exact boundary controllability of the wave equation (I), Dirichlet Controls : Description of the numerical methods* Japan. J. Appl. Math., 7 (1990), 1–76.

[43] R. Glowinski, *Ensuring well posedness by analogy : Stokes problem and boundary control of the wave equation,* J. Comput. Physics, 103(2) (1992), 189–221.

[44] V. Grandgirard, *Numerical methods for magnetic confinement fusion*, http ://smai.emath.fr/cemracs/cemracs10/fr_courses.html

[45] V. Grandgirard, M. Brunetti, P. Bertrand, N. Besse, X. Garbet, P. Ghendrih, G. Manfredi, Y. Sarrazin, O. Sauter, E. Sonnendrücker, J. Vaclavik, L. Villard, *A drift-kinetic semi-Lagrangian 4D code for ion turbulence simulation*, J. Comput. Phys., **217**, pp. 395-423, (2006).

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

[47] J. Guterl, J. P. Braeunig, N. Crouseilles, V. Grandgirard, G. Latu, M. Mehrenberger, E. Sonnendrücker, *Test of some numerical limiters for the conservative PSM scheme for 4D Drift-Kinetic simulations,* INRIA research report number 7467, November 2010.

[48] E. Hairer, C. Lubich, G. Wanner, *Geometric numerical integration : Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics, 2006.

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

[50] O. Hoenen, M. Mehrenberger, E. Violard, *Parallelization of an Adaptive Vlasov Solver,* ParSim04 proceedings, Lecture Notes in Computer Science, 3241 (2004), 430–435.

[51] H. Hong, S. Steinberg, *Accuracy and stability of polynomial interpolation schemes for advection equations,* preprint, 2001 (http ://wendouree.org/˜stanly/prints/Publications.html)

[52] W. Hundsdorfer, J. Verwer, *Numerical solution of time-dependent advection-diffusion-reaction equations,* volume 33 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2003.

[53] James M. Hyman, *Accurate monotonicity preserving cubic interpolation*, SIAM J. Sci. Stat. Comput. **4** pp. 645-654.RA

[54] L. Ignat, *Propiedades cualitativas de esquemas numéricos de aproximción de ecuaciones de difusión y de dispersión,* Thesis, Universidad Autónoma de Madrid (2006).

[55] J. A. Infante, E. Zuazua, *Boundary observability for the space discretization of the* $1D$ *wave equation*, M^2AN, 33(2) (1999), 407–438.

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

[57] A. Iserles, G. Strang, *The optimal accuracy of difference schemes,* Trans. of the AMS, Vol. 277, 2, 198, 779–803, 1983.

[58] F.A. Khodja, K. Mauffrey, A. Münch, *Exact controllability of a system of mixed order with essential spectrum*, SIAM Control and Optimization (2011).

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

[60] V. Komornik, *Exact Controllability and Stabilization-The Multiplier Method*, John Wiley and Masson, Chicester and Paris, 1994.

[61] V. Komornik, P. Loreti, *Fourier Series in Control Theory*, Springer Monographs in Mathematics, Springer-Verlag, New York, 2005.

[62] J. Laprise, A. Plante *A class of semi-Lagrangian integrated-mass (SLIM) numerical transport algorithms*, Mon. Wea. Rev. **123**, pp. 553-565, (1995).

[63] P. Loreti, M. Mehrenberger, *An Ingham type proof for a two-grid observability theorem,* ESAIM Control Optim. Calc. Var. 14 (2008), no. 3, 604–631.

[64] P. Loreti, M. Mehrenberger, *"Observabilité uniforme de l'équation des ondes 1D, [Uniform observability of the 1D wave equation]* Paris-Sud Working

Group on Modelling and Scientific Computing 2007–2008, 68–79, ESAIM Proc., 25, EDP Sci., Les Ulis, 2008.

[65] P. Loreti, V. Valente, *Partial exact controllability for spherical membranes*, SIAM J. Control Optim. 35 (1997), 641–653.

[66] Y. I. Lyubarskii, K. Seip, *Sampling and Intepolating Sequences for Multiband-Limited Functions end Exponential Bases on Disconnected Sets,* J. Fourier Anal. Appl., 3, 5, 597-615, 1997.

[67] A. Mangeney, F. Califano, C. Cavazzoni, P. Travnicek, *A Numerical Scheme for the Integration of the Vlasov-Maxwell System of Equations*, J. Comput. Phys. 179, 495–538 (2002).

[68] J. Marzo, *Riesz basis of exponentials for a union of cubes in* $\mathbb{R}^d$, http ://arxiv.org/aabs/math.FA/0601288, 2005.

[69] M. Mehrenberger, E. Violard, O. Hoenen, M. Campos Pinto, E. Sonnendrücker, *A Parallel Adaptive Vlasov Solver Based on Hierarchical Finite Element Interpolation,* Proceedings ICAP2004 St-Petersburg, Nuclear Inst. and Methods in Physics Research, A 558 (2006), 188–191.

[70] M. Mehrenberger, E. Violard, *A Hermite type adaptive semi-Lagrangian scheme,* Int. J. Appl. Math. Comput. Sci., 17 (3) (2007), 329–334.

[71] M. Mehrenberger, *Observability of coupled systems,* Acta. Mat. Hungar. 103 (4) (2004), 321–348.

[72] M. Mehrenberger, *Inégalités d'observabilité et résolution adaptative de l'équation de Vlasov par éléments finis hiérarchiques*, Thèse de Doctorat, Université Louis Pasteur, décembre 2004.

[73] M. Mehrenberger, *Critical length for a Beurling type theorem*, Bol. Un. Mat. Ital. B (8), 8-B (2005), 251–258.

[74] M. Mehrenberger, *An Ingham type proof for the boundary observability of a N-d wave equation*, C. R. Math. Acad. Sci. Paris 347 (2009), no. 1-2, 63–68.

[75] G. Meinardus, H. Morsche, G. Walz, *On the Chebyshev Norm of Polynomial B-Splines,* Journal of Approximation Theory, 82, 99–122 (1995).

[76] S. Micu, *Uniform boundary controllability of a semi-discrete* 1*D wave equation,* Numer. Math. 91 (2002), 723–766.

[77] S. Micu, Luz de Teresa, *A spectral study of the boundary controllability of the linear 2-D wave equation in a rectangle*, Asympt. Anal. 66 (2010), no. 3-4, 139–160.

[78] S. Micu, E. Zuazua, *Boundary controllability of a linear hybrid system arising in the control of noise,* SIAM J. Cont. Optim., 35(5)(1997), 1614–1638.

[79] A. Moulahi, S. Nouira, *Stabilisation polynomiale et analytique de l'équation des ondes sur un rectangle*, Annales mathématiques Blaise Pascal, 2010, vol. 17, no 2, 401–424.

[80] A. Münch, *Family of implicit and controllable schemes for the 1D wave equation* , C. R. Acad. Sci. Paris Série I, 339(10) (2004), 733–738.

[81] A. Münch, *A uniformly controllable and implicit scheme for the 1D wave equation*, Mathematical Modelling and numerical analysis 39(2) (2005), 1–42.

[82] T. Nakamura, R. Tanaka, T. Yabe, K. Takizawa, *Exactly conservative semi-Lagrangian scheme for multi-dimensional hyperbolic equations with directional splitting technique*, J. Comput. Phys., **174**, pp. 171-207, (2001).

[83] T.Nakamura, T.Yabe, *Cubic Interpolated Propagation Scheme for Solving the Hyper-Dimensional Vlasov-Poisson Equation in Phase Space*, Comput. Phys. Comm., **120**, pp.122-154 (1999).

[84] M. Negreanu, *Numerical methods for the analysis of the propagation, observation and control of waves,* Thesis, Universidad Complutense Madrid (2003). Available at http ://www.uam.es/proyectosinv/cen/indocumentos.html

[85] M. Negreanu, E. Zuazua, *Convergence of a multigrid method for the controllability of a 1D wave equation,* C. R. Acad. Sci. Paris, Série I, 338 (5), (2004), 413-418.

[86] O. Pironneau,*On the transport-diffusion algorithm and its applications to the Navier Stokes equations.*, Numer. Math. 38 (1982), 309

[87] J. Qiu, *Integral deferred correction methods : introduction and application*, seminaire Novembre 2011, IRMA, Université de Strasbourg.

[88] E. Pohn, M. Shoucri, G. Kamelander, *Eulerian Vlasov codes*, Comput. Phys. Comm. 166 (2005), 81-93.

[89] J.M. Qiu, C. W. Shu, *Conservative semi-Lagrangian finite difference WENO formulations with applications to the Vlasov equation*, Comm. Comput. Phys. **10** (2011), pp.979-1000.

[90] J.M. Qiu, C. W. Shu, *Positivity preserving semi-Lagrangian discontinuous Galerkin formulation : theoretical analysis and application to the Vlasov-Poisson system*, J. Comput. Phys., Volume 230, Issue 23 (2011), pp. 8386-8409.

[91] K. Ramdani, T. Takahashi, M. Tucsnak, *Uniformly exponentially stable approximations for a class of second order evolution equations. Application to LQR optimization problems*, ESAIM COCV, 13 (3), (2007), 503–527.

[92] K. Ramdani, T. Takahashi, G. Tenenbaum, M. Tucsnak, *A spectral approach for the exact observability of infinite-dimensional systems with skew-adjoint generator*, J. Funct. Anal. 226 (1) (2005), 193–229.

[93] S. Reich, *An explicit and conservative remapping strategy for semi-Lagrangian advection*, Atmospheric Science Letters, vol. 8, issue 2, pp. 58-63.

[94] T. Respaud, E. Sonnendrücker, *Analysis of a new class of Forward semi-Lagrangian schemes for the 1D Vlasov-Poisson Equations*, Numer. Math., **118** (2011), pp. 329-366.

[95] M. Restelli, L. Bonaventura, R. Sacco, *A semi-Lagrangian discontinuous galerkin method for scalar advection by incompressible flows,* Journal of Computational Physics 216 (2006) 195–215.

[96] A. Robert, *A stable numerical integration scheme for the primitive meteorological equations,* Atmos. Ocean., 19 (1981), pp. 35–46.

[97] J. A. Rossmanith, D. C. Seal, *A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov-Poisson equations*, J. Comp. Phys. 230, pp. 6203-6232, (2011).

[98] J. Schaeffer *Higher order time splitting for the linear Vlasov equation*, SIAM, J. Numer Anal. 47 (2009), 2203-2223.

[99] L. Schumaker, *Spline functions : basic theory*, Cambridge University Press, 2007.

[100] Joseph W. Schumer, James Paul Holloway, *Vlasov Simulations Using Velocity-Scaled Hermite Representations*, Journal of Computational Physics Volume 144, Issue 2, 10 August 1998, Pages 626-66.

[101] SELALIB, http ://selalib.gforge.inria.fr/

[102] M. Shoucri, *A two-level implicit scheme for the numerical solution of the linearized vorticity equation*, Int. J. Numer. Meth. Eng. **17**, p. 1525, (1981).

[103] M. Shoucri *Eulerian codes for the numerical solution of the Vlasov equation.* Commun. Nonlinear Sci. Numer. Simul. 13 (2008), no. 1, 174–182.

[104] M. Sofroniou, G. Spaletta, *Derivation of symmetric composition constants for symmetric integrators*, Optimization Methods and Software 20 (2005), 597-613.

[105] E. Sonnendrücker, J. Roche, P. Bertrand, A. Ghizzo *The semi-Lagrangian method for the numerical resolution of the Vlasov equation,* J. Comput. Phys., **149**, pp. 201-220, (1999).

[106] E. Sonnendrücker, *Approximations numérique des équations Vlasov-Maxwell*, (Notes du Cours M2), (2010).

[107] E. Sonnendrücker, *A possibly faster algorithm for cubic splines on a uniform grid*, unpublished.

[108] A. Staniforth, J. Côté, *Semi-Lagrangian integration schemes for atmospheric models–A review,* Mon. Weather Rev. 119 (1991).

[109] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. 5, pp. 506-517, (1968).

[110] T. Umeda, *A conservative and non-oscillatory scheme for Vlasov code simulations*, Earth Planets Space, **60**, pp. 773-779, (2008).

[111] M. Unser, *Fast B-Spline Transforms for Continuous Image Representation and Interpolations,* IEEE Trans on Pattern Analysis and Machine Intelligence, vol 13 (3), 1991.

[112] T. H. Watanabe, H. Sugama, *Vlasov and Drift Kinetic Simulation Methods Based on the Symplectic Integrator*, National Institute for Fusion Science 792 (2004).

[113] H. Yoshida, *Construction of higher order symplectic integrators*, Physics Letters A 150 (1990).

[114] M. Zerroukat, N. Wood, A. Staniforth, *The Parabolic Spline Method (PSM) for conservative transport problems*, Int. J. Numer. Meth. Fluids, **51**, pp. 1297-1318, (2006).

[115] E. Zuazua, *Propagation, Observation, Control and Numerical Approximation of Waves approximated by finite difference methods,* SIAM Review, 47 (2) (2005), 197-243.