HAL
open science

# Étude de quelques perturbations d'équations riches en symétries : résonances et stabilités

Joackim Bernier

▶ **To cite this version:**

Joackim Bernier. Étude de quelques perturbations d'équations riches en symétries : résonances et stabilités. Equations aux dérivées partielles [math.AP]. Université de Rennes, 2019. Français. NNT : 2019REN1S039 . tel-02397827

**HAL Id: tel-02397827**

**https://theses.hal.science/tel-02397827**

Submitted on 6 Dec 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# THÈSE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
**C**OMUE **U**NIVERSITE **B**RETAGNE **L**OIRE

Ecole Doctorale N°601
*Mathèmatique et Sciences et Technologies*
*de l'Information et de la Communication*
Spécialité : *Mathématiques et leurs interactions*

Par **Joackim BERNIER**

## Étude de quelques perturbations d'équations riches en symétries : résonances et stabilités

**Thèse présentée et soutenue à** RENNES , le 4 juillet 2019
**Unité de recherche : IRMAR, UMR CNRS 6625**

**Rapporteurs avant soutenance :**
Patrick GERARD, Professeur, Université Paris-Sud
Nader MASMOUDI, Professeur, New York University Abu Dhabi

**Composition du jury :**

| | |
|---|---|
| Examinateurs : | Manuela Valeria BANICA, Professeur, Sorbonne Université |
| | Lukas EINKEMMER, Professeur associé, University of Innsbruck |
| | Isabelle GALLAGHER, Professeur, Ecole Normale Supérieure |
| | Patrick GERARD, Professeur, Université Paris Sud |
| | Nader MASMOUDI, Professeur, New York University Abu Dhabi |
| | Miguel RODRIGUES, Professeur, Université de Rennes 1 |
| Dir. de thèse : | Erwan FAOU, Directeur de Recherches, INRIA |
| Co-dir. de thèse : | Nicolas CROUSEILLES, Chargé de Recherches, INRIA |

# REMERCIEMENTS

Tout d'abord, je tiens vraiment à vous remercier, Erwan et Nicolas, pour votre soutient, votre affection, votre confiance, votre savoir ainsi que votre enthousiasme ! Pour moi, vous êtes bien plus que mes directeurs de thèse !

Patrick Gérard et Nader Masmoudi, je vous remercie d'avoir rapporté cette thèse et de participer au jury de soutenance. Je suis très heureux et honoré que vous ainsi que Manuela Valeria Banica, Lukas Einkemmer, Isabelle Gallagher et Miguel Rodrigues fassiez partie de ce jury. Plus particulièrement, Miguel je tiens aussi à te remercier pour tous les précieux conseils et remarques éclairantes que tu as pu me donner tout au long de cette thèse.

Un grand merci à mes co-auteurs Benoît, Paul, Michel, Yingzhe, Fernando, Sever et Yann (ainsi que bien sûr Erwan et Nicolas) pour toutes les mathématiques que vous m'avez faites découvrir et tout les échanges passionnants que l'on a pu avoir durant cette thèse. J'espère bien avoir d'autres occasions de travailler avec vous ! Benoît, merci beaucoup de m'avoir fait découvrir une partie du Mexique et pour les conférences très enrichissantes auxquelles tu m'as permis d'aller ces derniers mois. Merci Karel pour tes relectures attentives et tes suggestions pour nos travaux avec Paul.

Je suis très reconnaissant envers tous les membres et personnels de l'IRMAR, de l'université de Rennes 1 et de l'ENS Rennes pour votre disponibilité, l'excellent cadre de travail, les discussions enrichissantes et l'environnement scientifique stimulant que vous nous offrez.

J'ai été très heureux de faire partie de l'équipe IPSO devenue MINGUS qui a été comme une famille mathématique pour moi. Merci pour tous ces bons moments, ces échanges, votre intérêt, les workshops où vous m'avez permis d'aller et toutes les opportunités que vous m'avez données de présenter mes travaux. Pierre je tiens à te remercier plus particulièrement pour toute l'aide que tu m'as apporté durant cette thèse.

Merci Didier pour tout le soutient et les précieux conseils que tu me donnes depuis tant d'années. C'est bien grâce à toi si j'ai fait cette thèse !

Merci Jean-François et les membres de l'ANR NABUCO pour la confiance que vous m'avez très tôt accordée en me permettant de poursuivre la recherche l'année prochaine avec vous.

Je tiens à remercier Pierre Germain, Michael Weinstein et les membres du Courant Institute pour m'avoir accueilli en avril et pour les échanges enrichissants que nous avons pu avoir.

Merci à ma famille, aux thésards de l'IRMAR, aux copains de promo, d'athlé ou de plus longue date pour votre soutien et tous les bons moments passés ensemble ces dernières années !

# SOMMAIRE

**Introduction** **9**
Optimalité et résonances des schémas aux différences finies compactes pour le problème de Dirichlet homogène sur un segment . . . 10
Existence et stabilité d'ondes progressives solitaires pour l'équation de Schrödinger non linéaire discrète . . . 13
Formes normales rationnelles pour les équations de Schrödinger non linéaires sur le tore de dimension 1 . . . 22
Quelques comportements non linéaires de solutions des équations de Vlasov-Poisson 34
Méthodes de splitting pour les rotations et leurs applications aux équations de Vlasov 39
Articles, prépublications et proceedings . . . 45

**1 Compact finite difference schemes** **47**
1.1 Introduction . . . 47
1.2 Formalism and main results . . . 48
1.2.1 Context . . . 48
1.2.2 Notions of consistency and stability . . . 50
1.2.3 Expression of the schemes . . . 51
1.2.4 Main results . . . 55
1.3 Polynomials and high order formulas . . . 58
1.3.1 Consistency for the polynomials . . . 59
1.3.2 The optimal case . . . 61
1.4 Stability . . . 63
1.4.1 Strong stability . . . 63
1.4.2 Relative stability . . . 66
1.5 Numerical experiments . . . 69
1.5.1 Efficiency of the optimal schemes . . . 69
1.5.2 Numerical resonances . . . 71
1.6 Appendix . . . 71
1.6.1 Proof of Proposition 1.2.2 . . . 71
1.6.2 Proof of Lemma 1.4.2 . . . 73
1.6.3 Proof of Theorem 1.4.3 . . . 73

**2 Discrete traveling waves for DNLS** **77**
2.1 Introduction . . . 77
2.1.1 Motivations and main results . . . 77
2.1.2 Notations . . . 83
2.2 Aliasing generating inhomogeneity . . . 84
2.2.1 Shannon's advection . . . 84
2.2.2 The aliasing error . . . 87
2.2.3 The flow of DNLS in the space of the Shannon interpolations . . . 88

2.3 Traveling waves of the homogeneous Hamiltonian . . . 89
2.3.1 Construction of the traveling waves . . . 90
2.3.2 Gevrey uniform regularity, Lyapunov stability and some adjustments . . . 102
2.4 Control of the instabilities and modulation . . . 110
2.5 Appendix . . . 117
2.5.1 Proof of Theorem 2.1.7 . . . 117
2.5.2 Proof of Lemma 2.4.1 . . . 120
2.5.3 Inverse function Theorem . . . 122
2.5.4 A result of coercivity . . . 123
2.5.5 Functional analysis lemmas . . . 125

**3 Bounds on the high Sobolev norms of DNLS 129**
3.1 Introduction . . . 129
3.2 Shannon interpolation . . . 132
3.3 Proof of Theorem 3.1.1 . . . 134
3.3.1 Construction of the modified energies . . . 134
3.3.2 Proof of Theorem 3.1.1 by induction . . . 142

**4 Rational normal forms for NLS 147**
4.1 Introduction . . . 147
4.2 Statement of the results and sketch of the proof . . . 150
4.2.1 Main results . . . 150
4.2.2 Sketch of proof . . . 152
4.3 General setting . . . 155
4.3.1 Hamiltonian formalism . . . 155
4.3.2 Hamiltonian flows . . . 157
4.3.3 Polynomial Hamiltonians . . . 159
4.4 Non-resonance conditions . . . 160
4.4.1 Small denominators . . . 161
4.4.2 Non resonant sets . . . 162
4.5 Probability estimates . . . 165
4.6 A class of rational Hamiltonians . . . 173
4.6.1 Construction of the class . . . 173
4.6.2 Structural lemmas . . . 176
4.7 Rational normal form . . . 181
4.7.1 Resonant normal form . . . 181
4.7.2 Elimination of the quintic term by the cubic . . . 186
4.7.3 Quintic normal form . . . 189
4.7.4 Proof of the rational normal form Theorem . . . 191
4.8 Dynamical consequences and probability estimates . . . 191
4.9 Appendix . . . 193
4.9.1 The case of (NLSP) . . . 193
4.9.2 Proof of Lemma 4.6.6 . . . 194

**5 Second order dispersion relations for Vlasov-Poisson 201**
5.1 Introduction . . . 201
5.2 Derivation of the dispersion relations . . . 208
5.2.1 Dispersion relations for first and second order . . . 208
5.2.2 A general linearized Vlasov-Poisson equation . . . 209
5.2.3 Proof of Propositions 5.2.1 and 5.2.2 . . . 211
5.3 Resolution and expansion of the linearized equation . . . 212
5.3.1 Introduction and statement of the result . . . 212
5.3.2 Definition of $N_k$ and theoretical tools . . . 213
5.3.3 Estimations for $D_k$ and $N_k$ . . . 214
5.3.4 A theoretical tool for the control of $\mathcal{L}^{-1}[N_k/D_k]$ . . . 216
5.3.5 Proof of Proposition 5.3.1 . . . 218
5.4 Resolution and expansion of the second order equation . . . 218
5.4.1 Introduction and statement of the result . . . 218
5.4.2 Definition of $\mathcal{N}_k^1$ and $\mathcal{N}_k^2$ (the right hand side) . . . 220
5.4.3 Estimates for $\mathcal{N}_k^1$ and $\mathcal{N}_k^2$ . . . 221
5.4.4 Proof of Lemma 5.4.2 in the non-resonant case. . . . 223
5.4.5 Proof of Lemma 5.4.2 in the resonant case . . . 224
5.4.6 Proof of Lemma 5.4.3 in the non resonant case . . . 225
5.4.7 Proof of Lemma 5.4.3 in the resonant case . . . 227
5.4.8 Proof of Proposition 5.4.1 . . . 231
5.5 Numerical results . . . 232
5.5.1 First example . . . 233
5.5.2 Another case where the Best frequency is almost dominant on a spatial mode . . . 234
5.5.3 A $2D$ case . . . 236
5.6 Appendix . . . 241
5.6.1 Some remarks about the space $\mathscr{E}(\mathbb{R}^d)$ . . . 241
5.6.2 An algebraic decomposition . . . 244
5.6.3 Computation of the zeros . . . 244

**6 Splitting methods for rotations : application to Vlasov equations 247**
6.1 Introduction . . . 247
6.2 Presentation of the method and its numerical analysis . . . 248
6.2.1 Numerical analysis . . . 250
6.2.2 Numerical illustrations . . . 261
6.3 Application to the Vlasov-Maxwell equations . . . 264
6.3.1 Reduced 1+1/2 Vlasov-Maxwell equations . . . 266
6.3.2 Splitting method . . . 266
6.3.3 Composition methods for systems separable into three parts . . . 267
6.4 Numerical results . . . 273
6.4.1 Vlasov-Maxwell system. . . . 273
6.4.2 Vlasov-HMF system. . . . 275
6.5 Conclusion . . . 279
6.6 Appendix . . . 279

6.6.1 Proof of Lemma 6.2.2 . . . . . 279
6.6.2 Proof of Lemma 6.2.3 . . . . . 280
6.6.3 Proof of Lemma 6.2.4 . . . . . 280
6.6.4 Proof of Lemma 6.2.7 . . . . . 281
6.6.5 Proof of Lemma 6.2.8 . . . . . 281

# INTRODUCTION

Les équations les plus élémentaires issues de la physique sont généralement des équations riches en symétries. Ces dernières imposent des contraintes sur les solutions pouvant permettre dans de nombreux cas une résolution explicite ou une description fine de certaines propriétés qualitatives. Cependant, pour bien des applications, on est amené à considérer des perturbations de ces équations ne partageant pas nécessairement ces symétries. On peut bien sûr penser à des perturbations associées à des modèles physiques plus réalistes ou plus précis mais aussi aux systèmes discrets introduits pour la résolution numérique des équations. En effet, pour la discrétisation en temps des problèmes d'évolution, *l'analyse rétrograde* permet d'identifier un système dynamique discret avec un système dynamique continu en des temps discrets. Dans le cas de la discrétisation en espace, on identifie une solution d'un problème discret avec celle d'un problème continu par l'intermédiaire d'un opérateur d'interpolation. C'est par exemple le principe de l'analyse de *Von-Neumann* ou des *éléments finis.*

Dans les cas les plus favorables, la perturbation peut simplement déformer les symétries. Lorsque l'on cherche à établir une méthode numérique ayant de bonnes propriétés qualitatives [1], on cherche naturellement à rentrer dans ce cadre. C'est par exemple tout l'intérêt des *méthodes symplectiques* pour l'intégration des systèmes hamiltoniens (voir [78]).

Cependant, que ce soit pour des raisons physiques ou des choix numériques, les perturbations peuvent briser les symétries de l'équation, ce qui peut changer radicalement le comportement des solutions (diminution du nombre d'invariants, perte de stabilité, ...). Par exemple, lorsque l'on perturbe un système *intégrable*, des phénomènes de *résonances* peuvent provoquer de larges variations des *actions* et ainsi changer radicalement la dynamique de l'équation (voir par exemple [51],[69]). Néanmoins, comme on le verra, dans de nombreuses situations, il est nécessaire d'attendre des *temps très longs* (relativement à la taille de la perturbation) pour pouvoir observer ces effets.

Ce manuscrit contient de nombreuses illustrations de ces phénomènes dans divers contextes tels que les *équations cinétiques*, *les problèmes elliptiques* et *les équations de Schrödinger non linéaires*. Il traite d'enjeux à la fois numériques, comme la conception et l'analyse de schémas ou la recherche de cas test pertinents permettant de valider l'implémentation des méthodes, mais aussi de problématiques plus théoriques orientées vers des questions de résonances et de stabilités. Avant de présenter en détail les résultats de cette thèse, on donne une rapide présentation de l'organisation du manuscrit, des motivations qui ont conduit ces travaux, des résultats obtenus et de certaines méthodes mises en œuvre.

Le premier chapitre est l'aboutissement d'un travail initié lors du stage de M2. Il est dévolu à l'étude d'une large classe de *schémas aux différences finies compactes* pour le problème de Dirichlet homogène sur le segment $[0, 1]$. Ces derniers jouissent de propriétés algébriques remarquables héritées d'une discrétisation de la dérivée seconde respectant ses symétries. Elles donnent des critères simples de consistance et de stabilité permettant de construire facilement des schémas convergeant à des ordres arbitrairement élevés. En mettant en œuvre des

1. souvent déterminantes, dans le cas des équations d'évolution, pour obtenir des comportements pertinents en temps longs.

méthodes *d'approximation de Padé*, on parvient à identifier les schémas les plus efficaces. A contrario, en étudiant certains phénomènes de *résonances* elliptiques, on montre que *presque tout schéma consistant est convergent*[2].

Les second et troisième chapitres de cette thèse portent sur l'équation de Schrödinger cubique focalisante discrète (DNLS). Il s'agit d'une semi-discrétisation par différences finies de l'équation de Schrödinger cubique focalisante présentant à la fois un intérêt numérique et un intérêt physique. Dans le second chapitre, qui est une collaboration avec Erwan Faou, on explique comment les termes *d'aliasing* engendrent une *inhomogénéité* dans DNLS. Malgré cette dernière, on parvient à construire des *ondes solitaires progressives* approchées orbitalement stables pendant des temps longs devant le paramètre de discrétisation en espace. On montre également que le défaut de stabilité est entièrement contrôlé par la croissance de normes de Sobolev discrètes d'indices élevés. Pour renforcer le résultat de stabilité et ainsi justifier l'existence de solitons sur des temps plus longs, dans le troisième chapitre, on établit, via la construction *d'énergies modifiées*, une borne polynomiale sur la croissance de ces normes.

Ces bornes étant effectives sur des temps d'autant plus longs que les normes sont initialement petites, il fut alors naturel de s'intéresser numériquement puis théoriquement à la dynamique des petites solutions de DNLS. Au vu des résultats des simulations et de la construction algébrique des énergies modifiées, il semblait alors envisageable de décrire leurs trajectoires à l'aide de méthodes de mise sous forme normale évitant d'avoir recours à des *paramètres externes*. L'analogue de telles constructions semblant être quasi-inexistant pour les équations de Schrödinger non linéaires continues, ces idées ont conduit à une collaboration avec Erwan Faou et Benoît Grébert présentée dans le quatrième chapitre de cette thèse. On y introduit une nouvelle construction de *formes normales* permettant de conjuguer le flot d'équations *résonnantes*, telles que les équations de *Schrödinger non linéaires* et de *Schrödinger-Poisson* sur le tore de dimension $1$, à une dynamique *intégrable*, sur des temps très longs et à des termes d'ordres arbitrairement élevés près, sur des ensembles contenant la plupart des petites fonctions régulières.

Le cinquième chapitre de cette thèse est le fruit d'une collaboration avec Michel Mehrenberger. En prolongeant au second ordre l'analyse linéaire classique des équations de *Vlasov-Poisson* au voisinage d'équilibres homogènes, on décrit certains phénomènes asymptotiques non-linéaires et multidimensionnels tels que les *ondes de Best*. Il s'agit de l'aboutissement d'un travail ayant débuté lors d'un projet CEMRACS[3] réalisé en 2016 (ayant donné lieu au proceeding [20]). A cette occasion, nous avions réalisé des calculs formels suggérant l'existence de certains de ces phénomènes, ce qui nous avait permis, grâce à leur observation, de valider l'implémentation de méthodes numériques pour les équations de *Vlasov-Poisson*.

Enfin dans le sixième chapitre, qui est une collaboration avec Fernando Casas et Nicolas Crouseilles, on réalise *l'analyse rétrograde* des méthodes de *splitting* pour résoudre l'équation de transport associée à une rotation dans le plan. Cela nous permet d'étudier et de corriger les pertes de symétries engendrées par ces méthodes. Après avoir réalisé l'analyse de convergence des méthodes *pseudo-spectrales* associées, on les met en œuvre pour obtenir des schémas précis et efficaces pour les équations de *Vlasov-Maxwell* et *Vlasov-HMF*.

2. l'ordre de convergence étant donné, à un facteur logarithmique près, par l'ordre de consistance.
3. Centre d'Eté Mathématique de Recherche Avancée en Calcul Scientifique.

## Optimalité et résonances des schémas aux différences finies compactes pour le problème de Dirichlet homogène sur un segment

Les schémas aux différences finies compactes sont des méthodes numériques visant à obtenir des approximations de solutions d'équations aux dérivées partielles. Ils sont étudiés depuis de très nombreuses années et la littérature les concernant est abondante. Cependant, pour les équations elliptiques, contrairement au cas des équations hyperboliques (voir [53]) et des équations différentielles ordinaires (voir [78]), il ne semble pas exister d'étude générale à leur propos. On trouve seulement beaucoup d'exemples de schémas efficaces construits à partir de considérations variées (énergie, monotonie, fonction de Green, ...)

Le travail présenté dans le premier chapitre de thèse a pour objectif d'être une première étape à une telle étude. Le sujet des équations elliptiques étant évidement trop vaste, on se restreint à étudier les schémas aux différences finies compactes pour un problème très académique : le problème de Dirichlet homogène en dimension $1$, i.e.

$$\begin{cases} -u''(x) = f(x), \ \forall x \in ]0,1[, \\ u(0) = u(1) = 0, \end{cases} \tag{DH}$$

où $f \in \mathcal{C}^\infty(\mathbb{R})$ est donnée et $u \in \mathcal{C}^\infty([0,1])$ est à déterminer. Bien que élémentaire, ce problème permet de mettre en évidence de nombreuses difficultés liées à l'existence d'un bord.

Dans ce contexte, un schéma au différences finies est un système linéaire de la forme

$$\mathbf{D}_N \mathbf{u}^N = h^2 \mathbf{S}_N \mathbf{f}^{N,ex},$$

où $\mathbf{f}^{N,ex} = f_{|h\mathbb{Z}}$ est la restriction du terme source sur la grille de pas $h = (N+1)^{-1}$, $\mathbf{D}_N \in \mathbb{C}^{N\times N}$ est une matrice carrée de taille $N$ et $\mathbf{S}_N \in \mathbb{C}^{(N\times\mathbb{Z})}$ est une matrice rectangulaire de hauteur $N$ et de largeur infinie à support fini. Enfin $\mathbf{u}^N \in \mathbb{C}^N$, l'inconnue de ce système, est une approximation de la solution $u$ de (DH) (i.e. on souhaite avoir $\mathbf{u}_j^N \simeq u(hj)$ pour $j = 1, \dots, N$).

Généralement, en excluant les premières et dernières lignes de $\mathbf{D}_N$ et $\mathbf{S}_N$, ces matrices sont creuses. Elles présentent seulement quelques diagonales non nulles sur lesquelles elles sont constantes. Dans cette classe très générale de schémas, beaucoup sont construits pour avoir une propriété de monotonie (voir par exemple [41],[100],[110]) car cette dernière permet d'obtenir facilement un critère de *stabilité* (voir [100]), i.e. (dans les cas simples)

$$\exists c > 0, \forall N \in \mathbb{N}^*, \forall \mathbf{v} \in \mathbb{C}^N, \ c\|\mathbf{v}\|_\infty \leqslant \|\mathbf{D}_N \mathbf{v}\|_\infty. \tag{1}$$

Dans la plupart des cas, pour obtenir cette propriété, on impose à $\mathbf{D}_N$ d'avoir ses coefficients diagonaux strictement positifs et ses coefficients extra-diagonaux négatifs. Cependant, ce critère est trop restrictif : il exclut de nombreux schémas convergents et rend difficile la conception de schémas d'ordres élevés, c'est-à-dire vérifiant (dans les cas simples)

$$\exists C > 0, \forall N \in \mathbb{N}^*, \forall f \in \mathcal{C}^\infty(\mathbb{R}), \ \|\mathbf{D}_N \mathbf{u}^{N,ex} - h^2 \mathbf{S}_N \mathbf{f}^{N,ex}\|_\infty \leqslant C h^{2n+2} \tag{2}$$

où $2n$ est l'ordre (de consistance) et $\mathbf{u}^{N,ex} = (u(hj))_{j=1,\dots,N}$, $u$ étant la solution de (DH).

L'étude proposée dans le premier chapitre de cette thèse porte sur une classe plus vaste de schémas. On considère des schémas pour lesquels $\mathbf{D}_N$ est un polynôme en $\mathbf{A}_N$, la discré-

tisation usuelle de la dérivée seconde, c'est-à-dire

$$\mathbf{A}_N = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{Z}^{N\times N}. \tag{3}$$

On peut tout d'abord remarquer qu'un tel choix n'est pas très restrictif car si on prend une formule de différences finies symétrique $d = (d_j)_{j\in\mathbb{Z}}$ qui approche la dérivée seconde, i.e. pour tout $u \in \mathcal{C}^\infty(\mathbb{R})$,

$$\sum_{j\in\mathbb{Z}} d_j u(hj) \simeq -h^2 u''(0),$$

alors on obtient[4] une matrice $\mathbf{D}_N$ qui est un polynôme en $\mathbf{A}_N$ et dont les termes centraux sont naturellement construits à partir de $d$, i.e.

$$1 \ll i \ll N \;\Rightarrow\; \forall j \in [\![1, N]\!],\ (\mathbf{D}_N)_{i,j} = d_{j-i}.$$

Grâce à ce choix, la dérivée seconde et sa discrétisation partagent les mêmes vecteurs propres. En effet, $\mathbf{D}_N$ a les même vecteurs propres que $\mathbf{A}_N$ et on peut facilement prouver que ceux-ci sont donnés par

$$\mathbf{A}_N \mathbf{e}_k^N = (2\sin(\pi k h/2))^2\, \mathbf{e}_k^N \text{ où } \mathbf{e}_k^N := (\sin(\pi k h j))_{j=1,\dots,N}.$$

Cette construction permet d'obtenir un critère simple et naturel de stabilité : si $\mathbf{D}_N = P(\mathbf{A}_N)$, il suffit que $P > 0$ sur $]0, 4]$, que $P(0) = 0$ et[5] $P'(0) \neq 0$. On parvient même à renforcer ce critère en permettant à $P$ de s'annuler sur $]0, 4]$ en imposant des conditions de nature *diophantienne* à ses zéros. En imposant aussi une structure naturelle à $\mathbf{S}_N$, de nature polynomiale, on parvient à établir une équivalence entre des estimations de consistance de type (2) et un problème *d'approximation de Padé* :

$$R(X) = C(X)Q(X) \mod X^n \tag{4}$$

où $P(X) = XR(X)$ est tel que $\mathbf{D}_N = P(\mathbf{A}_N)$, $\mathbf{S}_N$ est construite[6] à partir du polynôme $Q$ et $C$ est la série formelle

$$C(X) = 4\left(\frac{\arcsin(\frac{\sqrt{X}}{2})}{\sqrt{X}}\right)^2 = 2\sum_{n\in\mathbb{N}} \frac{X^n}{(n+1)^2 C_{2n+2}^{n+1}}.$$

Comme on le verra de façon détaillée, la plupart de ces bonnes propriétés sont en fait héritées des symétries de ces schémas. En effet, les polynômes en $\mathbf{A}_N$ sont exactement les matrices[7] des opérateurs de différences finies homogènes et symétriques restreints à l'espace (stable) des fonctions impaires en $0$ et en $1$. On peut donc dire que ces schémas ont les mêmes symétries que (DH).

Dans le premier chapitre de cette thèse, on cherche essentiellement à répondre à deux questions qualitatives :

4. par une formule de convolution et des choix naturels près du bord
5. ces deux dernières conditions sont naturelles pour obtenir des schémas étant au moins d'ordre $2$.
6. cette dernière étant plus technique, on ne la précise pas dans cette introduction.
7. dans la base canonique

- *En général,* ces schémas sont il stables ?
- Parmi tous ces schémas quels sont les plus *efficaces* ?

On parvient à répondre à la première question en démontrant que *presque tout*[8] *schéma consistant est convergent*, son ordre de convergence étant donné[9] par l'ordre de consistance. On obtient ce résultat en démontrant que la condition diophantienne garantissant la stabilité est vérifiée de façon *générique*. Il s'agit d'une démarche similaire à celle mise en œuvre pour établir la généricité de *conditions de non-résonance* pour les systèmes hamiltoniens (voir le quatrième chapitre de cette thèse).

La notion d'efficacité est plus délicate à définir car elle fait intervenir à la fois la précision du schéma et le temps de calcul requis pour résoudre le système linéaire. On répond donc à la seconde question en déterminant, à ordre de consistance fixé (i.e. $n$ dans (2)), quels sont les schémas ayant les matrices $\mathbf{D}_N$ et $\mathbf{S}_N$ les plus creuses. On montre que cela revient à déterminer les polynômes $R$ et $Q$ de degrés minimaux tels que $R(0) = 1$ et (4) soit satisfaite. En adoptant une terminologie classique, on cherche les *approximants de Padé* de $C$ (voir [11]). Cependant, pour que ceux-ci soient bien définis et que $R$ vérifie un critère garantissant la stabilité du schéma, il est nécessaire de connaitre des propriétés très fines sur $C$. Heureusement, de telles propriétés ont été établies récemment, par Karp et Prilepkina dans [86], pour une classe de *fonctions hypergéométriques généralisées* incluant $C$.

### Existence et stabilité d'ondes progressives solitaires pour l'équation de Schrödinger non linéaire discrète

On considère l'équation de Schrödinger sur $\mathbb{R}$ possédant une non linéarité cubique focalisante

$$\forall x \in \mathbb{R}, \qquad i\partial_t u(x) = \partial_x^2 u(x) + |u(x)|^2 u(x). \tag{NLS}$$

Il s'agit d'une équation autonome et réversible[10] qui est invariante par les transformations suivantes

$$\begin{cases} \text{Déphasage} & u(t,x) \mapsto e^{i\gamma} u(t,x) & \gamma \in \mathbb{T}, \\ \text{Advection} & u(t,x) \mapsto u(t, x - \mathrm{y}) & \mathrm{y} \in \mathbb{R}, \\ \text{Changement de référentiel galiléen} & u(t,x) \mapsto e^{iv(x-2vt)+iv^2 t} u(t, x-2vt) & v \in \mathbb{R}, \\ \text{Changement d'échelle} & u(t,x) \mapsto \lambda u(\lambda^2 t, \lambda x) & \lambda \in \mathbb{R}_+^*, \\ \text{Rotation} & u(t,x) \mapsto u(t,-x). \end{cases}$$

De plus, (NLS) est un système hamiltonien dans la mesure où elle s'écrit (formellement)

$$-i\partial_t u = \nabla H(u)$$

où le gradient est pris dans $L^2(\mathbb{R};\mathbb{C})$[11] et le hamiltonien $H$ est défini sur $H^1(\mathbb{R};\mathbb{C})$ par

$$H = \frac{1}{2}\| \bullet \|_{\dot{H}^1(\mathbb{R})}^2 - \frac{1}{4}\| \bullet \|_{L^4(\mathbb{R})}^4.$$

8. au sens de la mesure de Lebesgue sur les polynômes réels
9. à un facteur logarithmique près.
10. c'est-à-dire invariante par la transformation $u(t,x) \mapsto \bar{u}(-t,x)$.
11. Dans les chapitres $3$ et $4$ de cette thèse, on considère exclusivement $\mathbb{C}$ comme une $\mathbb{R}$-algèbre de dimension $2$ munie de la structure euclidienne (et non hermitienne) induite par module.

Par le théorème de Noether, on voit alors que les invariances par advection et déphasages sont associées à deux constantes du mouvement

$$\begin{cases} \text{La masse} & \|\bullet\|^2_{L^2(\mathbb{R})}, \\ \text{La quantité de mouvement} & \langle i\partial_x \bullet, \bullet\rangle_{L^2(\mathbb{R})}. \end{cases}$$

En dimension $1$ et munie d'une non linéarité cubique, l'équation de Schrödinger est une équation *intégrable*. On ne précise pas cette notion car elle n'a que peu d'importance dans cette partie. Cependant, on pourra simplement retenir que (NLS) possède une infinité de constantes du mouvement et de symétries non linéaires associées.

(NLS) possède une des *équilibres relatifs* intéressants appelés *solitons*. Il s'agit de solutions globales de (NLS) oscillant à la vitesse $\xi_1$ et se déplaçant à la vitesse $\xi_2$ où $\xi_1 > (\xi_2/2)^2$. Ils sont donnés explicitement par la formule

$$u(t,x) = e^{it\xi_1+\gamma}\psi_\xi(x - t\xi_2 - \mathrm{y})$$

où $\gamma \in \mathbb{T}$, $\mathrm{y} \in \mathbb{R}$ sont des degrés de liberté dus à l'invariance de (NLS) par déphasage et advection et $\psi_\xi$ est une fonction très régulière et localisée, explicitement donnée [12] par

$$\forall x \in \mathbb{R}, \quad \psi_\xi(x) = e^{\frac{1}{2}ix\xi_2}\frac{\sqrt{2}m_\xi}{\cosh(m_\xi x)} \quad \text{avec} \quad m_\xi = \sqrt{\xi_1 - \left(\frac{\xi_2}{2}\right)^2}.$$

Lorsque $\xi_2 = 0$ on parle *d'ondes stationnaires* tandis que lorsque $\xi_2 \neq 0$ on parle d'ondes *progressives* ou *voyageuses*.

Ces solitons sont généralement construits à partir des états de plus basses énergies ("*ground states*") de (NLS). Plus précisément, on peut montrer (voir par exemple [115]) que

$$\underset{\substack{u\in H^1(\mathbb{R}) \\ \|u\|_{L^2}=2}}{\arg\min} H(u) = \{e^{i\gamma}\psi_{(1,0)}(\bullet - \mathrm{y}),\ (\gamma,\mathrm{y}) \in \mathbb{T}\times\mathbb{R}\}.$$

En observant que ces minimiseurs sont des points critiques du hamiltonien sur la sphère de $L^2$, on montre naturellement que ce sont des ondes stationnaires. On obtient ensuite les autres solitons en ajustant la masse à l'aide de l'invariance par dilatation et la quantité de mouvement grâce à l'invariance galiléenne. L'intérêt principal de cette approche variationnelle est qu'elle permet de montrer la *stabilité orbitale* de ces équilibres relatifs. Elle fut démontrée pour la première fois en 1982 par Cazenave et Lions dans [46] en utilisant la méthode de *concentration-compacité*. Ce résultat de stabilité fut ensuite précisé en 1986 par Weinstein dans [115] grâce à l'introduction de fonctions de Lyapunov. Cette dernière méthode, désormais appelée *moment-énergie*, fut alors théorisée et étendue par Grillakis, Shatah et Strauss en 1987 dans [75], [76] (et récemment revisitée par De Bièvre, Genoud et Rota Nodari dans [58]). Le résultat que l'on obtient avec la méthode de *moment-énergie* peut alors s'énoncer sous la forme suivante.

**Théorème 1.** *Pour tout $\xi \in \mathbb{R}^2$ satisfaisant $\xi_1 > (\xi_2/2)^2$, on peut trouver une constante $c > 0$ telle que pour tout temps $t \in \mathbb{R}$ et toute solution $u$ de* (NLS) *dont la condition initiale vérifie $\|u(0) - \psi_\xi\|_{H^1(\mathbb{R})} < c$, il existe $\mathrm{y}, \gamma \in \mathbb{R}$ tels que*

$$c\|u(t) - e^{i\gamma}\psi_\xi(\bullet - \mathrm{y})\|_{H^1(\mathbb{R})} \leqslant \|u(0) - \psi_\xi\|_{H^1(\mathbb{R})}.$$

12. l'existence de telles formules est propre à la dimension $1$ et aux non linéarités de la forme $|u|^\nu u$. Elles ne sont pas indispensables pour notre étude mais permettent de simplifier de nombreuses preuves.

Ce résultat ne dit a priori rien sur l'évolution de $\mathrm{y}(t)$ et $\gamma(t)$. Cependant, ce défaut peut être comblé à l'aide de méthodes dites de *modulation* qui permettent de contrôler les variations de $|\dot{\gamma} - \xi_1|$ et $|\dot{\mathrm{y}} - \xi_2|$ par $\|u(0) - \psi_\xi\|_{H^1(\mathbb{R})}$ (voir par exemple [67], [81] ou [114]).

Dans les chapitres 2 et 3 de cette thèse, on s'intéresse à une semi-discrétisation de (NLS) sur une grille de pas $0 < h \ll 1$. Il s'agit d'une équation différentielle sur $\mathbb{C}^{h\mathbb{Z}}$ qui est définie par

$$\forall g \in h\mathbb{Z}, \quad i\partial_t \boldsymbol{u}_g = (\Delta_h \boldsymbol{u})_g + |\boldsymbol{u}_g|^2 \boldsymbol{u}_g, \tag{DNLS}$$

où $\Delta_h$, le laplacien discret, est défini sur $\mathbb{C}^{h\mathbb{Z}}$ par

$$(\Delta_h \boldsymbol{u})_g = \frac{\boldsymbol{u}_{g+h} - 2\boldsymbol{u}_g + \boldsymbol{u}_{g-h}}{h^2}.$$

Cette équation a bien moins de symétries que (NLS). En plus d'être autonome et réversible elle ne semble être invariante que par les transformations suivantes

$$\begin{cases} \text{Déphasage} & \boldsymbol{u}_g(t) \mapsto e^{i\gamma} u_g(t) \quad \gamma \in \mathbb{T}, \\ \text{Advection discrète} & \boldsymbol{u}_g(t) \mapsto \boldsymbol{u}_{g-a}(t) \quad a \in h\mathbb{Z}, \\ \text{Rotation} & \boldsymbol{u}_g(t) \mapsto \boldsymbol{u}_{-g}(t). \end{cases}$$

(DNLS) est elle aussi un système hamiltonien dans la mesure où elle s'écrit

$$-i\partial_t \boldsymbol{u} = \nabla H_h(\boldsymbol{u})$$

où le gradient est calculé pour la norme $L^2(h\mathbb{Z})$ et

$$H_h = \frac{1}{2}\| \bullet \|^2_{\dot{H}^1(h\mathbb{Z})} - \frac{1}{4}\| \bullet \|^4_{L^4(h\mathbb{Z})},$$

les normes de Lebesgue et de Sobolev homogènes discrètes étant naturellement définies pour $p \in [1, \infty)$, $n \in \mathbb{N}^*$ et $u \in \mathbb{C}^{h\mathbb{Z}}$ par

$$\|\boldsymbol{u}\|^p_{L^p(h\mathbb{Z})} = h \sum_{g \in h\mathbb{Z}} |\boldsymbol{u}_g|^p \quad \text{et} \quad \|\boldsymbol{u}\|^2_{\dot{H}^n(h\mathbb{Z})} = \langle (-\Delta_h)^n \boldsymbol{u}, \boldsymbol{u} \rangle_{L^2(h\mathbb{Z})}.$$

L'advection discrète n'étant pas associée à un *groupe de Lie continu*, elle n'a pas de raison d'être associée, par le théorème de Noether, à une constante du mouvement de (DNLS). Cependant, grâce à l'invariance par déphasage, la masse [13] continue d'être une constante du mouvement.

(DNLS) n'est pas la semi-discrétisation partageant le plus de symétries avec l'équation continue (NLS). On peut par exemple penser à la semi-discrétisation intégrable de (NLS) donnée par l'équation d'Ablowitz-Ladik (voir [2]) ou à des semi-discrétisations utilisant des méthodes de *désaliasing* qui seront introduites par la suite. Cependant, (DNLS) est probablement la plus élémentaire et la plus raisonnable en terme de coût de calcul. De plus, l'étude de (DNLS) est pertinente en elle même car elle intervient naturellement dans la modélisation de certains phénomènes physiques (voir par exemple [87]).

Dans le second chapitre de cette thèse, qui est un travail en commun avec Erwan Faou, on s'intéresse à l'existence et à la stabilité d'ondes progressives pour (DNLS), consistantes

13. i.e. définie par $\| \bullet \|^2_{L^2(h\mathbb{Z})}$.

avec celles de (NLS). Pour donner une idée du problème, on commence par présenter les résultats d'une expérience numérique (décrite dans [84]). On considère la solution de (DNLS), notée $\boldsymbol{u}$, dont la condition initiale est la restriction d'une onde voyageuse de (NLS) sur la grille $(\psi_\xi(g))_{g\in h\mathbb{Z}}$, avec $\xi_2 \neq 0$. Si (DNLS) possédait une onde progressive orbitalement stable et proche de celle de (NLS), cette solution devrait ressembler à $\psi_\xi$ pour tout temps et se déplacer globalement à la vitesse $\xi_2$. C'est exactement ce qui semble se passer dans un premier temps. Mais au bout d'un temps suffisamment long, l'onde semble ralentir puis s'arrêter. Cependant, ce phénomène devient quasiment impossible à observer pour des valeurs de $h$ suffisamment petites (par exemple $h = 1/10$ pour $\xi = (1,1)$). Dans la littérature physique, ce phénomène est connu sous le nom de *barrière de Peierls-Nabarro* (voir [84],[87], [96]). Il semble nécéssaire d'attendre des temps exponentiellement longs par rapport à $h^{-1}$ pour l'observer (voir [96]). Son existence ainsi que sa description semblent être un problème mathématique ouvert.

Ce phénomène est propre aux ondes voyageuses car pour les ondes stationnaires la situation est très différente. En 2010, Bambusi et Penati ont prouvé (dans [19]) l'existence d'ondes stationnaires consistantes avec celles de (NLS). En 2013, en étudiant une discrétisation totale[14] de (NLS), Bambusi, Faou et Grébert donnèrent des résultats sur leur stabilité orbitale dans [17]. Ces résultats se basent sur une interprétation de (DNLS) en termes d'éléments finis. Une autre construction, basée sur une interprétation pseudo-spectrale de (DNLS), a été réalisée en 2016 par Jenkinson et Weinstein (voir [84]). On résume en partie ces résultats dans le théorème suivant.

**Théorème 2.** *Pour tout $\xi_1 > 0$, il existe $h_0, C, c > 0$ tels que pour tout $h < h_0$, on puisse trouver une unique fonction $\boldsymbol{\phi}^h_{\xi_1} \in H^1(h\mathbb{Z};\mathbb{R})$ paire et un réel $\zeta_1 \in \mathbb{R}$, tels que*

- $e^{i\zeta_1 t}\,\boldsymbol{\phi}^h_{\xi_1}$ *est une solution de* (DNLS),
- $|\zeta_1 - \xi_1| + \|\,\boldsymbol{\phi}^h_{\xi_1} - \psi_{(\xi_1,0)\ |h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant Ch^2$,
- $\|\,\boldsymbol{\phi}^h_{\xi_1}\,\|_{L^2(h\mathbb{Z})} = \|\psi_{(\xi_1,0)}\|_{L^2(\mathbb{R})}$,
- *Si $\boldsymbol{u}$ est une solution de* (DNLS) *telle que $\boldsymbol{u}(0)$ est paire, $\|\,\boldsymbol{u}(0)\|_{L^2(h\mathbb{Z})} = \|\psi_{(\xi_1,0)}\|_{L^2(\mathbb{R})}$ et $\|\,\boldsymbol{u}(0) - \boldsymbol{\phi}^h_{\xi_1}\,\|_{H^1(h\mathbb{Z})} < c$ alors pour tout $t \in \mathbb{R}$, il existe $\gamma \in \mathbb{R}$ tel que*

$$\|\,\boldsymbol{u}(t) - e^{i\gamma}\,\boldsymbol{\phi}^h_{\xi_1}\,\|_{H^1(h\mathbb{Z})} \leqslant C\|\,\boldsymbol{u}(0) - \boldsymbol{\phi}^h_{\xi_1}\,\|_{H^1(h\mathbb{Z})}.$$

Ces résultats sont obtenus en se restreignant à l'ensemble des solutions paires de (DNLS) (qui, par invariance de (DNLS) par rotation, contient toutes les trajectoires issues de conditions initiales paires). Une solution paire de (DNLS) ne se déplaçant clairement pas, cette astuce permet d'éluder la question de l'instabilité que le déplacement engendré par une perturbation non paire pourrait générer mais elle exclut de fait l'étude des ondes progressives. De plus, puisque (DNLS) ne semble pas posséder de symétrie semblable au changement de référentiel galiléen de (NLS), on ne peut pas directement ramener l'étude de l'existence et de la stabilité des ondes voyageuses à celle des ondes stationnaires.

Le déplacement des ondes semblant être un facteur d'instabilité, pour l'étudier on a besoin de définir une advection sur $L^2(h\mathbb{Z})$ étendant continument l'advection discrète. Une façon naturelle de procéder consiste à transporter l'advection des fonctions sur $\mathbb{R}$ vers les fonctions sur la grille à l'aide d'un opérateur d'interpolation $\mathcal{I}_h : L^2(h\mathbb{Z}) \to L^2(\mathbb{R})$. Autrement dit, on définit

14. c'est-à-dire incluant la discrétisation en temps et les conditions aux limites.

l'advection de pas $a \in \mathbb{R}$ en faisant commuter le diagramme suivant

$$\begin{array}{ccc} L^2(h\mathbb{Z}) & \xrightarrow{\tau_a} & L^2(h\mathbb{Z}) \\ \mathcal{I}_h \downarrow & & \downarrow \mathcal{I}_h \\ L^2(\mathbb{R}) & \xrightarrow{u \mapsto u(\bullet - a)} & L^2(\mathbb{R}) \end{array}$$

Bien sûr, en général, cette approche ne fonctionne pas car l'advection d'une interpolation n'a aucune raison d'être une interpolation [15]. Cependant, il existe une interpolation pour laquelle on peut réaliser cette construction. Il s'agit de *l'interpolation de Shannon*. Elle consiste à prolonger une fonction de $L^2(h\mathbb{Z})$ en une fonction de $L^2(\mathbb{R})$ dont le support de la transformée de Fourier est compris entre $-\pi/h$ et $\pi/h$. On peut notamment la définir simplement par la formule

$$\mathcal{I}_h : \left\{ \begin{array}{lcl} L^2(h\mathbb{Z}) & \to & L^2(\mathbb{R}) \\ \boldsymbol{u} & \mapsto & h \displaystyle\sum_{g \in h\mathbb{Z}} \boldsymbol{u}_g \operatorname{sinc}\left(\frac{\pi}{h}(\bullet - g)\right), \end{array} \right.$$

où $\operatorname{sinc}(x) = x^{-1}\sin(x)$ est le sinus cardinal. Son image, notée $BL^2_h$ est constituée des fonctions de $L^2(\mathbb{R};\mathbb{C})$ dont le support de la transformée de Fourier est compris entre $-\pi/h$ et $\pi/h$.

Afin de mieux comprendre le défaut de symétrie de (DNLS) par rapport à cette nouvelle advection, on conjugue le flot de (DNLS) par l'interpolation $\mathcal{I}_h$ pour obtenir un système dynamique sur $BL^2_h$. Il s'agit encore d'un système hamiltonien, dont on montrera que ce dernier est donné par la formule

$$H_h \circ \mathcal{I}_h^{-1}(u) = \frac{1}{2}\int_{\mathbb{R}} \left|\frac{u(x+h)-u(x)}{h}\right|^2 \mathrm{d}x - \frac{1}{4}\int_{\mathbb{R}} \left(1 + 2\cos(2\pi x/h)\right)|u(x)|^4 \mathrm{d}x.$$

On observe que le terme quartique contient un terme inhomogène hautement oscillant. Il est dû au défaut de commutation entre la non linéarité et l'interpolation. On parle donc d'erreur *d'aliasing*. A cause de cette inhomogénéité, (DNLS) n'est pas invariant par rapport à *l'advection de Shannon*. On ne parvient donc pas à lui appliquer les méthodes classiques permettant d'obtenir des ondes progressives orbitalement stables.

Pour surmonter cette difficulté, on considère les solutions de (DNLS) comme des perturbations d'un système hamiltonien homogénéisé, appelé (DNLS/A) [16], et dont le hamiltonien est donné par

$$\widetilde{H}_h \circ \mathcal{I}_h^{-1}(u) = \frac{1}{2}\int_{\mathbb{R}} \left|\frac{u(x+h)-u(x)}{h}\right|^2 \mathrm{d}x - \frac{1}{4}\int_{\mathbb{R}} |u(x)|^4 \mathrm{d}x. \tag{DNLS/A}$$

Il s'agit d'une démarche semblable à celle mise en place en 2004 par Fröhlich, Gustafson, Jonsson et Sigal dans [67] et en 2008 par Holmer et Zworski dans [81] pour étudier l'impact d'un potentiel faiblement oscillant sur les solitons de (NLS).

Par construction, $\widetilde{H}_h$ est invariant par l'advection de Shannon et il reste une perturbation du hamiltonien de (NLS) invariante par déphasage [17]. Grâce à des méthodes perturbatives, on

15. par exemple dans le cas d'une interpolation par éléments finis.
16. Cette notation sert à rappeler que l'on a simplement supprimer les termes d'aliasing.
17. On peut cependant remarquer que (DNLS/A) est bien plus proche de (DNLS) que ne l'est (NLS). En terme de consistance, (DNLS/A) est une approximation de (DNLS) d'ordre infini tandis que (NLS) n'en est qu'une approximation d'ordre $2$.

parvient donc à construire des solitons pour (DNLS/A) consistants avec ceux de (NLS). De plus, en utilisant la méthode *moment-énergie* on construit des fonctions de Lyapunov garantissant leur stabilité orbitale. Pour faire simple, il s'agit de fonctions invariantes par le flot de $\widetilde{H}_h$ et qui contrôlent, localement dans $H^1(h\mathbb{Z})$, la distance aux solitons de (DNLS/A).

Étant donnée une solution $\boldsymbol{u}$ de (DNLS) initialement proche d'un des solitons de (DNLS/A) se déplaçant à la vitesse $\xi_2$, on parvient à montrer que la variation de ces fonctions de Lyapunov le long de $\boldsymbol{u}$ est essentiellement contrôlée par le produit entre la variation de la quantité de mouvement et la vitesse $\xi_2$

$$\xi_2\langle i\partial_x u(t), u(t)\rangle_{L^2(\mathbb{R})} - \xi_2\langle i\partial_x u(0), u(0)\rangle_{L^2(\mathbb{R})} = -\xi_2(2\pi/h)^{-1}\int_0^t\int_{\mathbb{R}} \sin(2\pi x/h)|u(s,x)|^4 \mathrm{d}x\mathrm{d}s, \tag{5}$$

où $u = \mathcal{I}_h\,\boldsymbol{u}$. Dans le cas des ondes stationnaires, $\xi_2 = 0$, cette variation est nulle. On retrouve donc un résultat de stabilité orbitale similaire à celui du Théorème 2. En ce qui concerne les ondes progressives, il y a deux approches pour contrôler (5).

La première consiste, pour chaque $s \in (0,t)$, à linéariser le terme de droite de (5) autour d'un soliton de (DNLS/A) proche de $\boldsymbol{u}(s)$. Les solitons de (DNLS/A) étant très réguliers, on verra que l'on peut négliger les deux premiers termes de cette linéarisation pour se focaliser sur le terme d'ordre $2$. On obtient alors essentiellement un contrôle de (5) par $th^{2n-1}\delta_n^2(t)$ où $\delta_n(t)$ est la distance, en norme $\dot{H}^n(h\mathbb{Z})$, entre $\boldsymbol{u}(s)$ et le soliton autour duquel on a linéarisé. Cependant, (5) ne contrôle cette distance qu'en norme $H^1(h\mathbb{Z})$. Pour pouvoir conclure avec un argument de type "*bootstrap*", on est alors contraint de choisir $n = 1$. Cela conduit à avoir un résultat de stabilité orbitale sur des temps de l'ordre de $h^{-1}$. Plus précisément, en suivant soigneusement la phase et la position des solutions à l'aide d'une méthode de modulation, on obtient le théorème suivant [18].

**Théorème 3.** *Soit $\Omega$ un ouvert relativement compact dans $\left\{\xi \in \mathbb{R}^2 \mid \xi_1 > (\xi_2/2)^2\right\}$.*
*Il existe $h_0, \kappa, r, \ell > 0$ tels que pour tout $h < h_0$, tout $\xi \in \Omega$, il existe une fonction très régulière $\eta_\xi^h \in H^\infty(\mathbb{R})$ et consistante à l'ordre $2$ avec $\psi_\xi$, i.e.*

$$\|\eta_\xi^h - \psi_\xi\|_{H^1(\mathbb{R})} \leqslant \kappa h^2,$$

*qui vérifie la propriété suivante. Si, à une advection et un déphasage près, $\boldsymbol{v} \in H^1(h\mathbb{Z})$ est proche de $\eta_\xi^h$, i.e.*

$$\exists \gamma_0, \mathrm{y}_0 \in \mathbb{R}, \quad \|\,\boldsymbol{v} - (e^{i\gamma_0}\eta_\xi^h(\bullet - \mathrm{y}_0))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant r,$$

*alors il existe $\gamma, \mathrm{y} \in C^1(\mathbb{R})$ avec $\gamma(0) = \gamma_0$ et $\mathrm{y}(0) = \mathrm{y}_0$ tels que la solution de* (DNLS) *valant initialement $\boldsymbol{v}$, notée $\boldsymbol{u}$, satisfait pour tout $t > 0$*

$$\delta(t) + |\dot{\gamma}(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant \kappa e^{\frac{h|\xi_2|t}{\ell^2}}(\delta(0) + e^{-\frac{\ell}{h}}). \tag{6}$$

*où*

$$\delta(t) := \|\,\boldsymbol{u}(t) - (e^{i\gamma(t)}\eta_\xi^h(\bullet - \mathrm{y}(t)))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})}.$$

**Remarque 1.** *Ce résultat appelle quelques remarques :*

18. Il s'agit du principal résultat du chapitre $3$ de cette thèse.

- *Si on omet les deux termes exponentiels dans* (6)*, il s'agit d'un résultat d'existence et de stabilité de solitons pour* (DNLS) *consistants avec ceux de* (NLS)*. Cependant, à cause du terme* $e^{-\ell/h}$*, il ne s'agit que de solitons approchés.*
- *Dans le cas stationnaire, on a* $\xi_2 = 0$*. Le premier terme exponentiel est donc constant. On retrouve donc, de façon approchée, l'existence d'ondes stationnaires pour* (DNLS) *donnée par le Théorème 2. Mais on étend le résultat de stabilité à des perturbations non symétriques.*
- *Dans le cas des ondes voyageuses, le premier terme exponentiel n'est contrôlé indépendamment de* $h$ *que sur des temps inférieurs à* $h^{-1}$*. Comme annoncé précédemment, il ne s'agit donc que d'un résultat de stabilité orbitale sur des temps de l'ordre de* $h^{-1}$*.*
- *Si on considère la solution de* (DNLS) *telle que* $\boldsymbol{u}(0) = (\eta_\xi^h(g))_{g\in\mathbb{Z}}$*, on a* $\delta(0) = 0$*. L'estimation* (6)*, nous garantie donc que* $\boldsymbol{u}$ *se comporte comme une onde voyageuse sur des temps de l'ordre de* $h^{-2}$*.*

La seconde approche pour contrôler (5) consiste à utiliser la régularité de $\boldsymbol{u}(s)$. En effet, grâce au terme hautement oscillant, on peut contrôler le terme de droite par

$$th^{2n-1} \max_{0\leqslant s\leqslant t} \| \boldsymbol{u}(s)\|^2_{\dot{H}^n(h\mathbb{Z})}.$$

Un contrôle *a priori* de la croissance des normes de Sobolev fournit donc un résultat de stabilité orbitale en temps long. Par exemple, une borne uniforme en temps sur la norme $\dot{H}^n(h\mathbb{Z})$ conduirait à un résultat de stabilité sur des temps de l'ordre de $h^{-2n+1}$. Pour NLS, l'existence d'une telle borne peut être obtenue grâce à *l'intégrabilité* (voir par exemple [104]). Cependant, pour (DNLS), elle ne semble pouvoir être obtenue que pour $n = 0$ et $n = 1$ grâce à la conservation de la masse et de l'énergie.

Le troisième chapitre de cette thèse est dévolu à l'obtention d'une borne polynomiale en temps sur la croissance des normes de Sobolev permettant de profiter de cette approche. Le résultat y étant établi est le suivant :

**Théorème 4.** *Pour tout* $n \in \mathbb{N}^*$*, il existe* $C > 0$*, tel que pour tout* $h > 0$*, si* $\boldsymbol{u}$ *est une solution de* (DNLS) *alors pour tout* $t \in \mathbb{R}$

$$\| \boldsymbol{u}(t)\|_{\dot{H}^n(h\mathbb{Z})} \leqslant C \left[ \| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})} + M_{\boldsymbol{u}(0)}^{\frac{2n+1}{3}} + |t|^{\frac{n-1}{2}} M_{\boldsymbol{u}(0)}^{\frac{4n-1}{3}} \right],$$

*où*

$$M_{\boldsymbol{u}(0)} = \| \boldsymbol{u}(0)\|_{\dot{H}^1(h\mathbb{Z})} + \| \boldsymbol{u}(0)\|^3_{L^2(h\mathbb{Z})}.$$

Grâce à ce théorème, on obtient alors un résultat de stabilité orbitale sur des temps de l'ordre de $h^{-2}$ pour des perturbations régulières des solitons.

**Théorème 5.** *Soit* $\Omega$ *un ouvert relativement compact dans* $\left\{\xi \in \mathbb{R}^2 \mid \xi_1 > (\xi_2/2)^2\right\}$ *et* $h_0, \kappa, r, \ell > 0$ *les constantes données par le théorème 3.*

*Pour tout* $\varepsilon, s > 0$*, il existe* $n \in \mathbb{N}^*$ *tel que pour tout* $\rho > 0$*, il existe* $C, T_0 > 0$ *avec*

$$T_0 = \infty \quad \textit{quand} \quad \xi_2 = 0 \qquad \textit{et} \qquad T_0 \to \infty \quad \textit{quand} \quad \xi_2 \to 0,$$

*et* $h_1 \in (0, h_0)$*, tel que pour tout* $h < h_1$*,* $\xi \in \Omega$ *et tout* $v \in H^n(\mathbb{R})$*, si*

$$\|v\|_{\dot{H}^n(\mathbb{R})} \leqslant \rho \quad \textit{et} \quad \|\psi_\xi - v\|_{H^1(\mathbb{R})} \leqslant \frac{r}{2(1+\kappa)}$$

*alors toute solution* $\boldsymbol{u}$ *de* (DNLS) *telle que*

$$\exists\, \mathrm{y}_0, \gamma_0 \in \mathbb{R}, \quad \forall g \in h\mathbb{Z}, \quad \boldsymbol{u}_g(0) = e^{i\gamma_0} v(g - \mathrm{y}_0)$$

*satisfait, pour tout* $t \geqslant 0$*,*

$$t \leqslant T_0 h^{-2+\varepsilon} \;\Rightarrow\; \|\, \boldsymbol{u}(t) - (e^{i\gamma(t)} \eta_\xi^h(\bullet - \mathrm{y}(t)))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant C \left( \|\eta_\xi^h - v\|_{H^1(\mathbb{R})} + h^s \right)$$

*où* $\gamma, \mathrm{y} \in C^1(\mathbb{R})$ *vérifie* $\gamma(0) = \gamma_0$*,* $\mathrm{y}(0) = \mathrm{y}_0$ *et*

$$t \leqslant T_0 h^{-2+\varepsilon} \;\Rightarrow\; |\dot{\gamma}(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant C \left( \|\eta_\xi^h - v\|_{H^1(\mathbb{R})} + h^s \right).$$

Dans de nombreux problèmes, un contrôle polynomial en temps des normes de Sobolev est obtenu en construisant des *énergies modifiées* et/ou en utilisant des arguments de *dispersion*. Originellement, ces méthodes furent introduites dans les années 90 par Bourgain (voir [37]) puis développées par Staffilani (voir [108]). Plus récemment, dans [104], Sohinger les a notamment mises en œuvre pour contrôler la croissance des normes $H^s$ par $t^{s^+/3}$ dans le cadre des équations de Schrödinger munies de non linéarités de type Hartree sur $\mathbb{R}$.

L'estimation donnée par le Théorème 3 est uniquement basée sur la construction d'énergies modifiées. Il est possible que cette estimation puisse être améliorée par l'utilisation d'estimations de dispersions (dans l'esprit de [104]). Cependant on ne s'attend pas à ce que le gain soit significatif, d'autant plus que la dispersion est connue pour être plus faible pour (DNLS) que pour (NLS) (voir [109]).

Afin de donner les grandes lignes de la construction qui est mise en œuvre dans le quatrième chapitre de cette thèse, on commence par remarquer que le résultat du Théorème 4 est homogène et qu'il suffit donc de l'établir pour $h = 1$. On introduit ensuite le *crochet de Poisson* sur $L^2(\mathbb{Z})$. Si $K_1$ et $K_2$ sont deux fonctions régulières sur $L^2(\mathbb{Z})$, le crochet de Poisson de $K_1$ et $K_2$ est défini par

$$\{K_1, K_2\} = \nabla K_2 \cdot i \nabla K_1,$$

où la relation $z_1 \cdot z_2 = \Re(z_1 \bar{z_2})$, $z_1, z_2 \in \mathbb{C}$ définit le produit scalaire sur $\mathbb{C}$.

Ensuite, on admet, dans un premier temps, que l'on peut construire des fonctions régulières sur $L^2(\mathbb{Z})$ à valeurs réelles, notées $\mathcal{E}_n^{(0)}$ et $\mathcal{E}_n^{(1)}$, telles que

(I) $\mathcal{E}_n^{(0)}$ commute [19] avec $\|\bullet\|^2_{\dot{H}^1(\mathbb{Z})}$ et $\mathcal{E}_n^{(1)}$ est solution de *l'équation homologique*

$$\{\|\bullet\|^2_{\dot{H}^1(\mathbb{Z})}, \mathcal{E}_n^{(1)}\} = \frac{1}{2}\{\|\bullet\|^4_{L^4(\mathbb{Z})}, \mathcal{E}_n^{(0)}\}, \tag{7}$$

(II) $\mathcal{E}_n^{(0)}$ soit équivalente au carré de la norme $\|\bullet\|_{\dot{H}^n(\mathbb{Z})}$

$$\exists c > 0,\; c\,\|\bullet\|^2_{\dot{H}^n(\mathbb{Z})} \leqslant \mathcal{E}_n^{(0)} \leqslant c^{-1}\|\bullet\|^2_{\dot{H}^n(\mathbb{Z})}.$$

19. c'est-à-dire qu'il vérifie $\{\|\bullet\|^2_{\dot{H}^1(\mathbb{Z})}, \mathcal{E}_n^{(0)}\} = 0$.

(III) $\mathcal{E}_n^{(1)}$ et $\{\| \bullet \|_{L^4(\mathbb{Z})}^4, \mathcal{E}_n^{(1)}\}$ soient chacun contrôlés par le produit du carré de la norme $\| \bullet \|_{\dot{H}^{n-1}(\mathbb{Z})}$ et d'une puissance de $M_\bullet$.

On peut facilement déduire de (I), que si $\boldsymbol{u}$ est une solution de (DNLS), alors

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathcal{E}_n^{(0)} + \mathcal{E}_n^{(1)}\right] \circ \boldsymbol{u} &= \frac{1}{2}\{\| \bullet \|_{\dot{H}^1(\mathbb{Z})}^2, \mathcal{E}_n^{(0)} + \mathcal{E}_n^{(1)}\} \circ \boldsymbol{u} - \frac{1}{4}\{\| \bullet \|_{L^4(\mathbb{Z})}^4, \mathcal{E}_n^{(0)} + \mathcal{E}_n^{(1)}\} \circ \boldsymbol{u} \\ &= -\frac{1}{4}\{\| \bullet \|_{L^4(\mathbb{Z})}^4, \mathcal{E}_n^{(1)}\} \circ \boldsymbol{u}\,.\end{aligned}$$

Or par conservation de l'énergie et de la masse, on peut contrôler $M_{\boldsymbol{u}(t)}$ par $M_{\boldsymbol{u}(0)}$. Ainsi, grâce à (II) et (III), on peut obtenir, par récurrence sur $n \geqslant 1$, l'estimation annoncée dans le Théorème 4.

Comme on vient de le voir, la preuve de ce théorème repose donc essentiellement sur la construction des fonctions $\mathcal{E}_n^{(0)}$ et $\mathcal{E}_n^{(1)}$. Afin d'expliquer comment ces fonctions sont obtenues, on introduit la transformée de Fourier discrète

$$\mathcal{F} : \begin{cases} L^2(\mathbb{T}) & \to & L^2(\mathbb{Z}) \\ u & \mapsto & \left(\displaystyle\int_{\mathbb{T}} u(\omega) e^{ig\omega} \mathrm{d}\omega\right)_{g \in \mathbb{Z}} \end{cases}.$$

On note alors $\widehat{\bullet} = \mathcal{F}^{-1}$ la transformée de Fourier inverse. On cherche $\mathcal{E}_n^{(0)}$ sous la forme

$$\mathcal{E}_n^{(0)}(\boldsymbol{u}) = \int_{\mathbb{T}} f_n(\omega) |\widehat{\boldsymbol{u}}(\omega)|^2 \mathrm{d}\omega$$

où $f_n$ est un fonction régulière sur $\mathbb{T}$ et $\mathcal{E}_n^{(1)}$ sous la forme

$$\mathcal{E}_n^{(1)} = \int_{w \in \mathcal{V}} g_n(w) \prod_{j=1}^{2} \widehat{\boldsymbol{u}}(w_j) \overline{\widehat{\boldsymbol{u}}(w_{-j})} \mathrm{d}w$$

où $g_n$ est une fonction régulière sur $\mathcal{V} = \{w \in \mathbb{T}^4,\ w_1 + w_2 = w_{-1} + w_{-2}\}$.

De simples calculs montrent alors que $\mathcal{E}_n^{(0)}$ commute avec $\| \bullet \|_{\dot{H}^1(\mathbb{Z})}^2$ et que l'équation homologique (7) peut être ramenée à l'équation suivante sur $f_n$ et $g_n$ :

$$\forall w \in \mathcal{V},\ \sum_{j=1}^{2} f_n(w_j) - f_n(w_{-j}) = \left(\sum_{j=1}^{2} \cos(w_j) - \cos(w_{-j})\right) g_n(w). \tag{8}$$

Il s'agit d'un problème de divisibilité que l'on parvient à résoudre grâce à une factorisation algébrique astucieuse[20]. En effet, pour tout $k \in \mathbb{N}$ impair et tout $w \in \mathbb{R}^4$ tel que $w_1 + w_2 = w_{-1} + w_{-2} + 2\pi j$ où $j \in \mathbb{Z}$, on a

$$\begin{aligned}\sum_{j=1}^{2} \cos(kw_j) - \cos(kw_{-j}) = 4\cos\left(k\frac{w_1 + w_2}{2}\right) \sin\left(k\frac{w_1 - w_2 - w_{-1} + w_{-2} + 2\pi j}{4}\right) \\ \times \sin\left(k\frac{w_1 - w_2 + w_{-1} - w_{-2} + 2\pi j}{4}\right).\end{aligned} \tag{9}$$

20. et difficilement généralisable.

En remarquant que lorsque $k$ est impair et $\omega \in \mathbb{T}$, on a

$$|\sin(k\omega)| \leqslant k|\sin(\omega)| \text{ et } |\cos(k\omega)| \leqslant k|\cos(\omega)|,$$

on en déduit que si $f_n$ est une somme de cosinus associés à des modes impairs alors on peut trouver une fonction $g_n$ régulière sur $\mathcal{V}$ telle que (8) soit vérifié. En déterminant explicitement une fonction $f_n$ satisfaisant cette condition et telle que l'hypothèse (II) soit vérifiée, on peut finalement montrer que $\mathcal{E}_n^{(1)}$ vérifie alors (III).

## Formes normales rationnelles pour les équations de Schrödinger non linéaires sur le tore de dimension 1

***Présentation du problème*** Dans le quatrième chapitre de cette thèse, qui est une collaboration avec Erwan Faou et Benoît Grébert, on étudie le comportement en temps longs des solutions initialement petites et régulières d'équations de Schrödinger non-linéaires. Plus précisément, on considère les équations de *Schrödinger non linéaires* suivantes

$$i\partial_t u = -\partial_x^2 u + \varphi(|u|^2)u, \ x \in \mathbb{T}, \ t \in \mathbb{R}, \tag{NLS}$$

où $\varphi$ est une fonction analytique réelle sur un voisinage de $0$ vérifiant $\varphi'(0) \neq 0$ ainsi que l'équation de *Schrödinger-Poisson*

$$\left\{ \begin{array}{rcl} i\partial_t u & = & -\partial_x^2 u + Wu, \\ -\partial_x^2 W & = & |u|^2 - \dfrac{1}{2\pi}\displaystyle\int_{\mathbb{T}} |u|^2 \mathrm{d}x \end{array} \right. , \ x \in \mathbb{T}, \ t \in \mathbb{R}. \tag{NLSP}$$

On est particulièrement intéressé par la dynamique des coefficients de Fourier de leurs solutions. Ces derniers étant définis pour les fonctions $u \in L^1(\mathbb{T})$ par

$$u(t,x) = \sum_{a \in \mathbb{Z}} u_a(t) e^{aix}.$$

Par la suite on identifiera toujours $u$ à sa suite de coefficients de Fourier.

Comme on vient de le voir dans la partie précédente (NLS) et (NLSP) sont des *systèmes hamiltoniens*. Cependant, ici et pour des raisons essentiellement calculatoires, il est préférable voir leurs solutions comme des solutions d'un système hamiltonien étendu de la forme

$$\forall j \in \mathbb{Z}, \ \left\{ \begin{array}{rl} i\partial_t u_j = & \partial_{v_j} H(u,v), \\ i\partial_t v_j = - & \partial_{u_j} H(u,v). \end{array} \right. \tag{10}$$

En effet, on peut facilement montrer que si $H(\bullet, \bar{\bullet})$ est à valeurs réelles (on dit alors que $H$ est réel) alors pour toute solution régulière $u, v$ de (10) vérifiant $v(t=0) = \overline{u(t=0)}$ on a $v(t) = \overline{u(t)}$ pour tout $t$. Puisque en pratique, on ne considère que de telles solutions pour des hamiltoniens réels, on identifie toujours $v$ à $\overline{u}$ et on note $z = (u, \overline{u})$. Il s'agit d'une technique très classique dans l'étude des systèmes hamiltoniens (voir par exemple [18]). Avec ce formalisme les hamiltoniens de (NLS) et (NLSP) sont

$$H_{\text{NLS}}(z) = \sum_{a \in \mathbb{Z}} (a^2 + \varphi(0))|u_a|^2 + \sum_{m \geqslant 2} \frac{\varphi^{(m-1)}(0)}{m!} \sum_{\sum_{j=1}^m k_j - \ell_j = 0} \prod_{j=1}^m u_{k_j} \overline{u_{\ell_j}},$$

et

$$H_{\text{NLSP}}(z) = \sum_{a\in\mathbb{Z}} a^2|u_a|^2 + \frac{1}{2}\sum_{\substack{k_1+k_2-\ell_1-\ell_2=0\\ k_1-\ell_1\neq 0}} \frac{u_{k_1}u_{k_2}\overline{u_{\ell_1}u_{\ell_2}}}{(k_1-\ell_1)^2}.$$

Pour pouvoir étudier le comportement de petites solutions régulières de (NLS) et (NLSP), on a besoin de pouvoir mesurer la taille et la régularité d'une fonction. On introduit donc la famille d'espace classiques suivants

$$\ell^1_s = \left\{ (u_a)_{a\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}},\ \|u\|_s := \sum_{a\in\mathbb{Z}} \langle a\rangle^s |u_a| < \infty \right\},$$

où $\langle a\rangle := \sqrt{1+a^2}$ désigne le crochet japonais. On utilise de tels espaces car ils simplifient significativement de nombreuses estimations mais il semble tout à fait possible d'adapter les résultats et constructions qui vont suivre dans des espaces de Sobolev (voir, par exemple, [18]).

La fonction constante égale à $0$ étant un équilibre de (NLS) et (NLSP), sur son voisinage ces équations peuvent être vues comme des perturbations de l'équation linéaire

$$i\partial_t u = (\mu - \partial_x^2)u,\ x\in\mathbb{T},\ t\in\mathbb{R}, \tag{LS}$$

où $\mu = \varphi(0) \in \mathbb{R}$ pour (NLS) et $\mu = 0$ pour (NLSP). L'équation de Schrödinger linéaire est elle aussi un système hamiltonien. Ce dernier est noté $Z_2$ et est donné par

$$Z_2(z) = \sum_{a\in\mathbb{Z}} (a^2+\mu)|u_a|^2,$$

il correspond naturellement à la partie quadratique de $H_{\text{NLS}}$ et $H_{\text{NLSP}}$. Comme on peut l'observer, $Z_2$ est une fonction dépendant de hamiltoniens réels, notés $I_a$ et définis par

$$I_a = |u_a|^2.$$

Pour un mode $a$ fixé, un tel hamiltonien engendre (par (10)) le flot

$$\Phi^t_{I_a} := u \mapsto (u_b e^{ti\delta_{a,b}})_{b\in\mathbb{Z}},\ t\in\mathbb{R}$$

où $\delta$ est le symbole de Kronecker. Il est clair que de tels flots commutent. Ils sont donc des symétries de $Z_2$ et ainsi de (LS). De plus par le théorème de Noether, les $I_a$ sont des constantes du mouvement pour (LS). La dynamique de (LS) peut même s'exprimer simplement à partir des $I_a$

$$\begin{cases} \partial_t I_a &= 0, \\ \partial_t \theta_a &= -\omega_a, \end{cases} \tag{11}$$

où $u_a = \sqrt{I_a}e^{i\theta_a}$ et les $\omega_a = a^2+\mu$ sont les *fréquences* du système linéaire. On dit donc que (LS) est *intégrable*[21] et on appelle les hamiltoniens $I_a$ des *actions*.

L'étude de perturbation de systèmes hamiltoniens intégrables remonte à Poincaré en 1892 dans [99]. Depuis, ce sujet est devenu classique et de nombreuses méthodes ont été développées. La plupart d'entre elles consistent à écrire le système dans de nouvelles variables

21. une définition rigoureuse de ce terme pour les systèmes hamiltoniens en dimension infinie serait particulièrement technique et peu utile pour ce qui va suivre. Dans le cas de la dimension finie, on peut se référer à [78].

(relativement proches des précédentes) dans lesquelles la dynamique est plus claire. Dans les cas les plus favorables, la dynamique des nouvelles variables est donnée par (11) (souvent à des termes de reste près d'ordre élevé) mais les fréquences sont des perturbations, dépendant uniquement des actions, des fréquences du système intégrable. Dans ce cas, on dit que l'on a mis le système sous *forme normale*. Pour avoir une telle dynamique, il suffit de pouvoir construire un changement de variables symplectique[22] $\tau$ relativement proche de l'identité tel que dans les nouvelles variables le hamiltonien soit une fonction des actions $(I_a)_{a\in\mathbb{Z}}$ (à des termes de reste près). Lorsque le changement de variables est défini sur un voisinage de $0$ et que les termes de reste peuvent être choisis d'ordre arbitrairement grand, on parle de *forme normale de Birkhoff*. Dans ce cas, on est en mesure de montrer que la variation relative des actions est petite sur des temps très longs.

Afin d'essayer de mettre (NLS) et (NLSP) sous forme normale de Birkhoff, on cherche le changement de variables sous la forme du flot au temps $t = 1$ d'un petit hamiltonien réel $\chi$, c'est-à-dire $\tau = \Phi^1_\chi$. Cela garantit, par construction, que le changement de variable est symplectique et proche de l'identité. Un simple calcul montre alors que si $\chi(z) = \mathcal{O}(z^4)$ on a

$$H \circ \tau^{-1} = Z_2(I) + P_4(z) - \{\chi, Z_2(I)\} + \mathcal{O}(z^6),$$

où $H = H_{\text{NLS}}$ ou $H = H_{\text{NLSP}}$, $P_4$ est sa partie quartique et $\{\bullet_1, \bullet_2\}$ est le crochet de Poisson qui est défini par

$$\{\chi_1, \chi_2\} = i \sum_{a\in\mathbb{Z}} \partial_{u_a}\chi_1 \partial_{\overline{u_a}}\chi_2 - \partial_{\overline{u_a}}\chi_1 \partial_{u_a}\chi_2.$$

Les termes de restes peuvent aussi être déterminés explicitement par un développement de Taylor[23] grâce à la formule $\partial_\alpha H \circ \Phi^\alpha_{-\chi} = -\{\chi, H \circ \Phi^\alpha_{-\chi}\}$.

On cherche alors à construire $\chi$ pour que $P_4(z) - \{\chi, Z_2(I)\}$ soit une fonction des actions. Or, un simple calcul montre que

$$\begin{aligned}\{u_{k_1}u_{k_2}\overline{u_{\ell_1}u_{\ell_2}}, Z_2(I)\} &= i(\omega_{k_1} + \omega_{k_2} - \omega_{\ell_1} - \omega_{\ell_2})u_{k_1}u_{k_2}\overline{u_{\ell_1}u_{\ell_2}}.\\ &= i(k_1^2 + k_2^2 - \ell_1^2 - \ell_2^2)u_{k_1}u_{k_2}\overline{u_{\ell_1}u_{\ell_2}}.\end{aligned}$$

Dans le cas de (NLS), pour simplifier autant de termes que possible, il est donc naturel de poser

$$\chi = \varphi'(0) \sum_{\substack{k_1+k_2=\ell_1+\ell_2\\ k_1^2+k_2^2\neq\ell_1^2+\ell_2^2}} \frac{u_{k_1}u_{k_2}\overline{u_{\ell_1}u_{\ell_2}}}{i(k_1^2 + k_2^2 - \ell_1^2 - \ell_2^2)}. \tag{12}$$

On obtient alors

$$H_{\text{NLS}} \circ \tau^{-1} = Z_2(I) + \varphi'(0) \sum_{\substack{k_1+k_2=\ell_1+\ell_2\\ k_1^2+k_2^2=\ell_1^2+\ell_2^2}} u_{k_1}u_{k_2}\overline{u_{\ell_1}u_{\ell_2}} + \mathcal{O}(z^6).$$

22. c'est-à-dire préservant la structure hamiltonienne.
23. ici en $\alpha = 0$

Il y a alors une propriété exceptionnelle, liée à la dimension $1$, à observer[24] :

$$\left.\begin{array}{rcl} k_1+k_2 & = & \ell_1+\ell_2 \\ k_1^2+k_2^2 & = & \ell_1^2+\ell_2^2 \end{array}\right\} \Rightarrow (k_1,k_2)=(\ell_1,\ell_2) \text{ ou } (k_1,k_2)=(\ell_2,\ell_1). \tag{13}$$

Ainsi, le terme quartique dans $H\circ\tau^{-1}$ est bien une fonction des actions, notée $Z_4$, et valant

$$Z_4(I)=2\varphi'(0)\left(\sum_{k\in\mathbb{Z}} I_k\right)^2-\varphi'(0)\sum_{k\in\mathbb{Z}} I_k^2. \tag{14}$$

En répétant la même opération avec les termes d'ordre $6$, on parvient facilement à construire un autre changement de variable $\tau$ tel que

$$H_{\text{NLS}}\circ\tau^{-1}=Z_2(I)+Z_4(I)+\sum_{\substack{k_1+k_2+k_3=\ell_1+\ell_2+\ell_3\\ k_1^2+k_2^2+k_3^2=\ell_1^2+\ell_2^2+\ell_3^2}} \alpha_{k,\ell}\, u_{k_1}u_{k_2}u_{k_3}\overline{u_{\ell_1}u_{\ell_2}u_{\ell_3}}+\mathcal{O}(z^8),$$

où les $\alpha_{k,\ell}\in\mathbb{C}$ sont uniformément bornés par rapport à $(k,\ell)$ et sont, a priori, non nuls. Cependant, la situation n'est pas aussi simple qu'à l'ordre $4$ car certains termes d'ordre $6$ ne sont pas des fonctions des actions : il y a des termes *résonnants*. En effet, par exemple, on a

$$-1+3+4=0+1+5 \text{ et } (-1)^2+3^2+4^2=0^2+1^2+5^2.$$

***Quelques résultats sur le sujet*** On aurait directement rencontré ce problème si on avait considéré (NLS) quintique (i.e. avec $\varphi(x)=\pm x^2$). Il s'agit d'un réel obstacle pour obtenir une forme normale de Birkhoff. Par exemple, en 2012, dans [73], Grébert et Thomann ont construit de petites solutions régulières de (NLS) quintique dont la variation relative des actions n'est pas petite sur des temps très longs. Un résultat similaire a été obtenu en 2012 par Gérard et Grellier dans [69](voir Corollary 1) pour une équation de "demi-onde" avec une non linéarité cubique (i.e. dans (NLS), il faut remplacer $-\partial_x^2$ par $|\partial_x|$ et prendre $\phi(x)=x$). Cela exclut l'existence d'une mise sous forme normale de Birkhoff pour ces équations. Des résultats semblables ont également été établis pour l'équation de Schrödinger cubique sur le tore de dimension $2$ (voir [43] ou [51]).

Dans ce cas, ce qui pose difficulté est l'existence de monômes $u_{k_1}\dots u_{k_m}\overline{u_{\ell_1}\dots u_{\ell_m}}$ n'étant pas des monômes en les actions et tels que $\omega_{k_1}+\dots+\omega_{k_m}=\omega_{\ell_1}+\dots+\omega_{\ell_m}$ et $k_1+\dots+k_m=\ell_1+\dots+\ell_m$. Dans cette situation, on dit que l'équation est *résonnante*. Pour parvenir à exclure ce genre de situations, une solution consiste à perturber les fréquences à l'aide de *paramètres externes*. Par exemple, il est classique de considérer une perturbation de (NLS) par un potentiel convolutif

$$i\partial_t u=-\Delta u+\varphi(|u|^2)u+W\star u,\ x\in\mathbb{T}^d,\ d\geqslant 1,\ t\in\mathbb{R}. \tag{NLS$\star$}$$

Dans ce cas, les fréquences sont les quantités $\omega_a=|a|^2+\varphi(0)+W_a$ pour $a\in\mathbb{Z}^d$. On peut démontrer que pour "la plupart"[25] des potentiels $W$, on a une minoration "raisonnable" de

24. il s'agit essentiellement du même type de "miracle" que la factorisation astucieuse (9) utilisée dans le travail sur la croissance des normes de Sobolev discrètes :

$$k_1+k_2=k_3+k_4 \Rightarrow k_1^2+k_2^2-k_3^2-k_4^2=2^{-1}(k_1-k_2-k_3+k_4)(k_1-k_2+k_3-k_4).$$

25. en un sens probabiliste.

$|\omega_{k_1} + \cdots + \omega_{k_m} - \omega_{\ell_1} - \cdots - \omega_{\ell_m}|$ par rapport à $(k, \ell)$ dès que $k_1 + \cdots + k_m = \ell_1 + \cdots + \ell_m$ et $u_{k_1} \dots u_{k_m} \overline{u_{\ell_1} \dots u_{\ell_m}}$ n'est pas un monôme en les actions. Il est alors possible de mener à bien une mise sous forme normale de Birkhoff en poursuivant la stratégie esquissée pour le terme quartique.

Une telle construction a été menée à bien par Bambusi et Grébert en 2006 dans [18]. Ils obtiennent ce résultat en prouvant un résultat de mise sous forme normale dans un cadre général. Ce dernier leur permet notamment de remplacer le terme $W \star u$ par $Wu$ (ce dernier étant plus physique) à condition de travailler en dimension $1$ avec des conditions de Dirichlet homogène. Cette approche a aussi permis, en 2013, dans [64], à Faou et Grébert de contrôler, pour la plupart des potentiels $W$, la variation relative des actions dans des espaces de Gevrey sur des temps exponentiellement longs. Enfin, en 2018, dans [31], Biasco, Massetti, et Procesi ont montré qu'il est possible, pour la plupart des potentiels $W$, de contrôler la variation relative des actions sur des temps exponentiellement longs mais dans des espaces de Sobolev.

Il existe très peu de résultats de formes normales pour les systèmes hamiltoniens résonnants pour lesquels on ne fait pas intervenir de paramètres externes (ce qui est naturel d'après les contre-exemples cités précédemment). Cependant dans le cas particulier de (NLS), en se plaçant bien en dimension $1$ et en ayant une perturbation cubique non triviale (i.e. $\varphi'(0) \neq 0$), on peut trouver quelques résultats. Tous s'appuient sur la propriété exceptionnelle (13) permettant de simplifier le terme d'ordre $4$. En effet, après avoir réalisé le changement de variables associé à cette première simplification, les fréquences (de la partie $Z_2 + Z_4$) s'écrivent $w_a = C(I) + a^2 + I_a$ où $C(I)$ est une fonction des actions et est indépendante de $a$. En se plaçant dans l'optique où l'on parviendrait à démontrer que les actions sont essentiellement des constantes du mouvement, ces fréquences sont les mêmes que pour (NLS$\star$) sauf que la condition initiale joue le rôle du potentiel convolutif. Cette observation a été faite pour la première fois en 1996, dans [89], par Kuksin et Pöschel. Ils s'en sont servi pour construire de nombreuses solutions quasi-périodiques à (NLS). Leur résultat s'appuie sur une autre construction de forme normale appelée *KAM*[26]. En 1999, dans [12], Bambusi l'a utilisée pour contrôler la variation relative des actions dans $H^1$ sur des temps exponentiellement longs (en se restreignant à des solutions satisfaisant des conditions de Dirichlet homogène). Enfin, Bourgain, en 2000, dans [39] s'en est servi pour contrôler sur des temps très longs la variation relative des actions pour "la plupart" des conditions initiales dans les espaces de Sobolev. Son résultat s'appuie sur une construction de forme normale très locale : pour une grande partie des fonctions $u^0$ dans la boule de taille $\varepsilon$ dans $H^s$, il parvient à conjuguer la dynamique de (NLS) à une dynamique intégrable[27], à des termes d'ordre $\varepsilon^{2r}$ près, sur un voisinage de taille $\varepsilon^r$ de $\{(e^{i\gamma_a} u_a^0)_{a\in\mathbb{Z}} \mid \gamma \in \mathbb{T}^{\mathbb{Z}}\}$ (ce résultat devant être compris pour des paramètres tels que $1 \ll r \ll s$). D'une certaine façon, il obtient une description très locale de la dynamique.

***Principaux résultats du chapitre*** Notre résultat est comparable à celui de Bourgain mais donne une description beaucoup plus globale de la dynamique : on construit le changement sur "presque toute" la boule de taille $\varepsilon$. Il nous suffit d'exclure quelques fonctions associées à des fréquences (i.e. $w_a = C(I) + a^2 + I_a$) résonnantes. Nos méthodes permettent aussi de mettre (NLSP) sous forme normale, qui comme on l'explique par la suite, peut simplement être vue comme une équation moins résonnante que (NLS).

26. en référence à Kolmogorov, Arnold et Moser.
27. i.e. semblable à (11).

Notre premier théorème est un résultat de mise sous forme normale à des ordres arbitrairement grands sur un ouvert de la boule de taille $4\varepsilon$. Sa "taille" est discutée par la suite.

**Théorème 6.** *Soit $H = H_{NLS}$ ou $H = H_{NLSP}$. Pour tout $r \geqslant 2$, il existe $s_0 \equiv s_0(r) = O(r^2)$ tel que pour tout $s \geqslant s_0$ il existe $\varepsilon_0 \equiv \varepsilon_0(r,s)$ tel que pour tout $\varepsilon < \varepsilon_0$, il existe des parties ouvertes $\mathcal{C}_{\varepsilon,r,s}$ et $\mathcal{O}_{\varepsilon,r,s}$ de $B_s(0,4\varepsilon)$, la boule de rayon $4\varepsilon$ de $\ell^1_s$, et un changement de variable canonique et analytique $\tau : \mathcal{O}_{\varepsilon,r,s} \to \mathcal{C}_{\varepsilon,r,s}$ vérifiant*

$$\|\tau(z) - z\|_s \leqslant \varepsilon^{3/2}, \quad \forall z \in \mathcal{O}_{\varepsilon,r,s}, \tag{15}$$

*et mettant $H$ sous forme normale à l'ordre $2r$ :*

$$H \circ \tau^{-1} = Z + R$$

*où*

- *$Z = Z(I)$ est une fonction régulière des actions ;*
- *le terme de reste, $R$, est d'ordre $\varepsilon^{2r+2}$ sur $\mathcal{C}_{\varepsilon,r,s}$, c'est-à-dire*

$$\|X_R(z)\|_s \leqslant \varepsilon^{2r+1}, \quad \forall z \in \mathcal{C}_{\varepsilon,r,s}. \tag{16}$$

*où $X_R$ est le champ de vecteur associé à $R$, c'est-à-dire $X_R = -i(\partial_{\bar{u}_a} R)_{a\in\mathbb{Z}}$.*

Le second résultat donne le principal corollaire dynamique que l'on est en droit d'attendre d'une telle mise sous forme normale : le flot engendre une faible variation relative des actions pendant des temps très longs, mais seulement sur une partie $\mathcal{V}_{\varepsilon,r,s}$ de l'ouvert sur lequel le changement de variable est défini (ce qui est naturel).

**Théorème 7.** *On considère $H$ et $\varepsilon_0(r,s)$ comme dans le Théorème 6 et $(u_a(t))_{a\in\mathbb{Z}}$ les coefficients de Fourier d'une solution $u$ du système Hamiltonien associée à $H$. Alors, pour tout $\varepsilon < \varepsilon_0$, il existe un ouvert $\mathcal{V}_{\varepsilon,r,s} \subset \mathcal{O}_{\varepsilon,r,s}$, tel que si $u(t=0) \in \mathcal{V}_{\varepsilon,r,s}$, on a*

$$|t| \leqslant \varepsilon^{-2r+1} \;\Rightarrow\; \begin{cases} \sup_{a\in\mathbb{Z}} \langle a\rangle^{2s} \left||u_a(t)|^2 - |u_a(0)|^2\right| \leqslant 3\varepsilon^{5/2}, \\ (u_a(t))_{a\in\mathbb{Z}} \in \mathcal{O}_{\varepsilon,r,s} \quad \textit{et en particulier} \quad \|u(t)\|_s \leqslant 4\varepsilon. \end{cases}$$

On discute enfin de la taille des ouverts sur lesquels ont lieu ces résultats. Le plus petit d'entre eux étant $\mathcal{V}_{\varepsilon,r,s}$, c'est ce dernier que l'on cherche à mesurer. Tout d'abord, on peut noter que $\mathcal{V}_{\varepsilon,r,s}$ est une sorte de cylindre dans le sens où

$$(u_a)_{a\in\mathbb{Z}} \in \mathcal{V}_{\varepsilon,r,s} \;\Rightarrow\; \forall \theta \in \mathbb{T}^{\mathbb{Z}},\ (e^{i\theta_a} u_a)_{a\in\mathbb{Z}} \in \mathcal{V}_{\varepsilon,r,s}.$$

Il suffit donc de mesurer la taille de l'ensemble des fonctions à coefficients réels appartenant à $\mathcal{V}_{\varepsilon,r,s}$. Pour faire cela, on définit une fonction aléatoire à valeurs dans la boule de taille $1/2$ de $\ell^1_s$

$$u^0(x) = c \sum_{a\in\mathbb{Z}} \sqrt{I_a} e^{iax}, \tag{17}$$

où $I = (I_a)_{a\in\mathbb{Z}}$ sont des variables aléatoires indépendantes distribuées dans $(0, \langle a\rangle^{-2s-4})$ et $c = (2\pi)^{-1} \tanh \pi$ est une constante de normalisation pour avoir presque sûrement $\|\varepsilon u^0\|_s < \varepsilon/2$. On estime alors la taille de $\mathcal{V}_{\varepsilon,r,s}$ en mesurant la probabilité que $\varepsilon u^0 \in \mathcal{V}_{\varepsilon,r,s}$ (auquel cas on dit que $\varepsilon u^0$ est *non résonant*). Il s'agit de la même démarche que celle mise en oeuvre par Bourgain dans [39] pour obtenir un résultat pour la plupart des petites fonctions de $H^s$.

**Théorème 8.** *Si pour tout $a \in \mathbb{Z}$, $I_a^2$ est uniformément distribuée dans $(0, \langle a \rangle^{-4s-8})$ alors il existe $\varepsilon_0(r,s) > 0$ tel que*

$$\forall \varepsilon \leqslant \varepsilon_0,\ \mathbb{P}(\varepsilon u^0 \in \mathcal{V}_{\varepsilon,r,s}) \geqslant 1 - \varepsilon^{1/3}.$$

L'obtention d'une estimation asymptotique de la mesure de l'ensemble non résonnant (i.e. $1 - \varepsilon^{1/3}$) est un des points forts de notre construction. Il permet d'affirmer d'une certaine façon que même lorsque l'on s'intéresse à des temps particulièrement grand (i.e. $r \gg 1$), on trouve toujours beaucoup de conditions initiales non résonantes. Avec la construction de Bourgain, cette borne se dégraderait vraisemblablement [28] avec $r$ (i.e. $1-\varepsilon^{\nu(r)}$ où $\nu(r) \to 0$ quand $r \to \infty$).

Pour analyser la stabilité des petites solutions régulières de (NLS) et (NLSP), il serait naturel (et tout particulièrement d'un point de vue numérique) de tirer aléatoirement $u^0$ et d'observer l'asymptotique des trajectoires issues de $\varepsilon u^0$ quand $\varepsilon$ tends vers $0$. Cependant, le Théorème 8 ne permet pas de prédire le résultat d'une telle expérience car il impose que $\varepsilon u^0$ soit tiré aléatoirement dans la boule de taille $\varepsilon/2$. Sur cette question, les résultats que nous obtenons sur (NLS) et (NLSP) diffèrent. Le premier résultat concerne (NLSP), il affirme que l'on observera bien une préservation relative des actions sur des temps toujours plus longs à mesure que $\varepsilon$ diminue.

**Théorème 9** (Cas (NLSP))**.** *Si pour tout $a \in \mathbb{Z}$, $I_a$ est uniformément distribuée dans $(0, \langle a \rangle^{-2s-4})$ alors pour $\varepsilon_0 < \varepsilon_0^{(0)}(r,s)$ on a*

$$\mathbb{P}(\forall \varepsilon \leqslant \varepsilon_0, \varepsilon u^0 \in \mathcal{V}_{\varepsilon,r,s}) \geqslant 1 - \varepsilon_0^{1/3}.$$

**Remarque 2.** *Pour* (NLSP)*, il est possible de tirer les actions selon une loi générant, avec une plus forte probabilité, des conditions initiales plus dégénérées (c'est-à-dire avec des coefficients de Fourier proches de $0$). Ce sont généralement des conditions initiales avec un nombre fini de coefficients de Fourier non nuls (i.e. des polynômes trigonométriques) qui permettent, pour les équations résonantes, de construire des solutions dont la variation relative des actions n'est pas petite en temps très longs (voir par exemple [43], [51],[73]).*

Pour (NLS), la situation semble plus complexe car notre construction laisse apparaitre un nouveau phénomène de résonances liant $\varepsilon$ à $u^0$ (on le discutera par la suite). Cependant on parvient à montrer que de telles résonances sont rares, dans le sens où en tirant aléatoirement $u^0$, on observe bien une préservation relative des actions sur des temps toujours plus longs à mesure que $\varepsilon$ diminue à condition de se restreindre à des valeurs de $\varepsilon$ choisies aléatoirement avec une décroissance au moins géométrique, c'est-à-dire $\varepsilon \in \{\varepsilon_n, n \in \mathbb{N}\}$ où

$$\varepsilon_n = 2^{-n+x_n},\ n \in \mathbb{N}$$

avec $(x_n)_{n\in\mathbb{N}}$ une famille de variables aléatoires uniformément distribuées dans $(0,1)$ et indépendantes de $I$.

**Théorème 10** (Cas (NLS))**.** *Si pour tout $a \in \mathbb{Z}$, $I_a^2$ est uniformément distribuée dans $(0, \langle a \rangle^{-4s-8})$ alors pour $n_0 \geqslant n_0^{(0)}(r,s)$, il y a une probabilité supérieure à $1 - 2^{-n_0/6}$ d'obtenir $u^0$ de telle sorte que la probabilité que $\varepsilon_n u^0$ soit non résonant pour tout $n \geqslant n_0$ est supérieure à $1 - 2^{-n_0/6}$. Plus formellement, on a*

$$\mathbb{P}\left(\mathbb{P}\left(\forall n \geqslant n_0,\ \varepsilon_n u^0 \in \mathcal{V}_{\varepsilon_n,r,s} \mid u^0\right) \geqslant 1 - 2^{-n_0/6}\right) \geqslant 1 - 2^{-n_0/6}.$$

28. dans son article, il ne cherche pas à lier la taille de son ensemble non résonant à $\varepsilon$.

**Remarque 3.** *Les $\varepsilon_n$ ne sont pas nécessairement indépendants. On peut tout à fait choisir $\varepsilon_n = 2^{-n}\varepsilon_0$.*

***Discussion sur la preuve de ces résultats*** La preuve de nos résultats repose sur une nouvelle théorie de formes normales rationnelles, elle est très différente de celle de Bourgain dans [39]. Elle s'apparente beaucoup plus à une mise sous forme normale de Birkhoff classique. Afin de l'expliquer, reprenons la construction que l'on avait débutée pour (NLS).

On avait construit un premier changement de variables symplectique proche de l'identité noté $\tau$ tel que

$$H_{\text{NLS}} \circ \tau^{-1}(z) = Z_2(I) + Z_4(I) + \mathcal{O}(z^6)$$

où $Z_4$ est donné par (14). Ce résultat avait été obtenu en prenant $\tau = \Phi^1_\chi$ où $\chi$ était obtenu en résolvant explicitement *l'équation homologique*

$$\{\chi, Z_2\} = \varphi'(0) \sum_{\substack{k_1+k_2=\ell_1+\ell_2 \\ k_1^2+k_2^2 \neq \ell_1^2+\ell_2^2}} u_{k_1} u_{k_2} \overline{u_{\ell_1} u_{\ell_2}}$$

et en observant que les termes quartiques restant étaient des monômes en les actions (voir (13)). En effectuant la même construction aux ordre supérieurs, on ne parvient pas à avoir des termes ne dépendant que des actions mais on peut néanmoins mettre $H_{\text{NLS}}$ sous *forme normale résonante*. C'est-à-dire construire $\tau$ tel que

$$H_{\text{NLS}} \circ \tau^{-1}(z) = Z_2(I) + Z_4(I) + \sum_{m=1}^{2r} \sum_{\substack{k_1+\cdots+k_m=\ell_1+\cdots+\ell_m \\ k_1^2+\cdots+k_m^2=\ell_1^2+\cdots+\ell_m^2}} \alpha_{k,\ell}\, u_{k_1} \dots u_{k_m} \overline{u_{\ell_1} \dots u_{\ell_m}} + \mathcal{O}(z^{2r+2})$$

où les $\alpha_{k,\ell} \in \mathbb{C}$ sont uniformément bornés par rapport à $(k, \ell)$ (cette construction est due à Kuksin et Pöschel dans [89]).

Avec cette construction tous les termes jusqu'à l'ordre $2r$ commutent avec $Z_2$. On ne peut donc plus espérer se servir de $Z_2$ pour les rendre intégrables (i.e. éliminer les termes n'étant pas des monômes en les actions). Il est donc naturel de chercher à utiliser $Z_4$. On cherche à faire un nouveau changement de variables donné par $\Phi^1_\chi$ où le hamiltonien $\chi$ est solution de l'équation homologique

$$\begin{aligned}\{\chi, Z_4\} &= \sum_{\substack{k_1+k_2+k_3=\ell_1+\ell_2+\ell_3 \\ k_1^2+k_2^2+k_3^2=\ell_1^2+\ell_2^2+\ell_3^2}} \alpha_{k,\ell}\, u_{k_1} u_{k_2} u_{k_3} \overline{u_{\ell_1} u_{\ell_2} u_{\ell_3}} - Z_6(I) \\ &= \sum_{\substack{k_1+k_2+k_3=\ell_1+\ell_2+\ell_3 \\ k_1^2+k_2^2+k_3^2=\ell_1^2+\ell_2^2+\ell_3^2 \\ \{k_1,k_2,k_3\}\cap\{\ell_1,\ell_2,\ell_3\}=\varnothing}} \alpha_{k,\ell}\, u_{k_1} u_{k_2} u_{k_3} \overline{u_{\ell_1} u_{\ell_2} u_{\ell_3}}\end{aligned}$$

où $Z_6$ est la partie du terme sixtique de $H_{\text{NLS}} \circ \tau^{-1}$ ne dépendant que des actions. Pour résoudre une telle équation, il suffit d'observer par un calcul très explicite que

$$\{u_{k_1} u_{k_2} u_{k_3} \overline{u_{\ell_1} u_{\ell_2} u_{\ell_3}}, Z_4\} = i(I_{k_1} + I_{k_2} + I_{k_3} - I_{\ell_1} - I_{\ell_2} - I_{\ell_3}) u_{k_1} u_{k_2} u_{k_3} \overline{u_{\ell_1} u_{\ell_2} u_{\ell_3}}.$$

Formellement, pour résoudre l'équation homologique, il suffirait donc de poser

$$\chi = \sum_{\substack{k_1+k_2+k_3=\ell_1+\ell_2+\ell_3 \\ k_1^2+k_2^2+k_3^2=\ell_1^2+\ell_2^2+\ell_3^2 \\ \{k_1,k_2,k_3\}\cap\{\ell_1,\ell_2,\ell_3\}=\varnothing}} \alpha_{k,\ell} \, \frac{u_{k_1}u_{k_2}u_{k_3}\overline{u_{\ell_1}u_{\ell_2}u_{\ell_3}}}{i(I_{k_1}+I_{k_2}+I_{k_3}-I_{\ell_1}-I_{\ell_2}-I_{\ell_3})}.$$

Il s'agit d'un hamiltonien plus complexe que ceux introduits précédemment. Il ne s'agit plus d'un polynôme mais d'une fraction rationnelle dont le dénominateur peut s'annuler. Pour définir son flot, il faut donc être en mesure de contrôler le dénominateur. C'est pour cela l'on ne parvient pas à faire une mise sous forme normale de Birkhoff sur toute une boule mais seulement sur un ouvert qu'elle contient.

Par des estimations de probabilités assez élémentaires, on peut montrer que si $u$ est tirée aléatoirement selon l'une des lois évoquées précédemment, on a, avec une probabilité supérieure à $1-c\gamma$ ($c$ étant une constante universelle)

$$\forall (k,\ell),\ |I_{k_1}+I_{k_2}+I_{k_3}-I_{\ell_1}-I_{\ell_2}-I_{\ell_3}| \geqslant \gamma\, \varepsilon^2 \left(\prod_{j=1}^{6} \langle \mu_j(k,\ell)\rangle^{-2}\right) \langle \mu_{\min}(k,\ell)\rangle^{-2s} \tag{18}$$

où $(\mu_j(k,\ell))_{j=1,\dots,6}$ est égal à $(k_1,k_2,k_3,\ell_1,\ell_2,\ell_3)$ à permutation près de sorte à avoir $|\mu_{\min}| = |\mu_6| \leqslant \cdots \leqslant |\mu_1|$.

Dans cette minoration, il y a quatre facteurs. Le premier, $\gamma$, peut être pensé [29] comme une constante. Le second facteur, $\varepsilon^2$ est un terme d'homogénéité [30], il ne pose aucun problème dans les estimations. Le troisième facteur (le produit) vient du fait que l'on a cherché à obtenir une estimation uniforme par rapport à $k$ et $\ell$. Le quatrième facteur est naturel car, dans la mesure où l'on considère des solutions régulières, les actions décroissent fortement.

A cause du troisième facteur l'estimation (18) n'est pas stable par une petite perturbation relative des actions dans $\ell^1_s$ (quitte à changer $\gamma$ en $\gamma/2$). Elle ne permet donc pas d'obtenir un ouvert sur lequel on pourrait définir le changement de variables. Il existe cependant une méthode classique [31] pour surmonter cette difficulté. Elle consiste à tronquer en fréquences les monômes par rapport au troisième plus grand indice.

En effet, considérons un monôme $M = u_{k_1}\dots u_{k_m}\overline{u_{\ell_1}\dots u_{\ell_m}}$ où $m \geqslant 2$ tel que $|\mu_3(k,\ell)| \geqslant N$ et $k_1+\cdots+k_m = \ell_1+\cdots+\ell_m$, alors son champ de vecteur $X_M = (i\partial_{\overline{u_a}}M)_{a\in\mathbb{Z}}$ vérifie l'estimation

$$\|X_M\|_s \leqslant m\langle \mu_1(k,\ell)\rangle^s \prod_{j=2}^{m} \langle \mu_j(k,\ell)\rangle^{-s} \|u\|_s^{2m-1} \leqslant m(2m-1)^s N^{-s}\|u\|_s^{2m-1},$$

car, grâce à la structure de convolution, on a $(2m-1)|\mu_2(k,\ell)| \geqslant |\mu_1(k,\ell)|$. Si on choisit

$$N \simeq \varepsilon^{-\frac{2r}{s}} \tag{19}$$

alors il s'agit d'un terme d'ordre $2r+2m-1$, c'est-à-dire un terme de reste qu'il n'y a donc pas besoin de résoudre.

29. même si à la fin de la preuve on le prend égal à $\varepsilon^{1/3}$, mais à ce niveau il s'agit surtout d'un détail technique.
30. les actions sont clairement des termes d'ordre $2$.
31. voir par exemple dans [64].

Il suffit donc de considérer le hamiltonien

$$\chi = \sum_{\substack{k_1+k_2+k_3=\ell_1+\ell_2+\ell_3 \\ k_1^2+k_2^2+k_3^2=\ell_1^2+\ell_2^2+\ell_3^2 \\ \{k_1,k_2,k_3\}\cap\{\ell_1,\ell_2,\ell_3\}=\varnothing \\ |\mu_3(k,\ell)|\leqslant N}} \alpha_{k,\ell} \, \frac{u_{k_1}u_{k_2}u_{k_3}\overline{u_{\ell_1}u_{\ell_2}u_{\ell_3}}}{i(I_{k_1}+I_{k_2}+I_{k_3}-I_{\ell_1}-I_{\ell_2}-I_{\ell_3})}. \tag{20}$$

Pour ce dernier, l'estimation du dénominateur (i.e. (18)) est plus favorable car si $(k,\ell)$ sont des indices de la somme précédente, on peut facilement vérifier que $|\mu_1(k,\ell)| \leqslant CN^2$ (où $C$ est une constante universelle). Ainsi, avec une probabilité supérieure à $1-c\gamma$ (où $c$ est une autre constante universelle), on a pour de tels indices

$$|I_{k_1}+I_{k_2}+I_{k_3}-I_{\ell_1}-I_{\ell_2}-I_{\ell_3}| \geqslant \gamma\, \varepsilon^2 N^{-8}\langle\mu_{\min}(k,\ell)\rangle^{-2s}. \tag{21}$$

Puisque $1 \ll r \ll s$, $\varepsilon^2 N^8$ est de la forme $\varepsilon^{2+}$ (voir (19)). L'estimation (21) est donc stable (quitte à changer $\gamma$ en $\gamma/2$) pour des perturbations relatives de $u$ dans $\ell^1_s$ de l'ordre de $\varepsilon^\nu$ pour n'importe quel constante universelle $\nu > 1$. Pour montrer que le flot engendré par $\chi$ est bien défini jusqu'au temps $1$ pour des fonctions vérifiant une estimation du type (21) et est proche de l'identité, il suffit de montrer son champ de vecteurs est petit devant $\varepsilon$. En omettant[32] les termes dus à une dérivation du dénominateur, on obtient naturellement une estimation du type :

$$\|X_\chi\|_s \lesssim \varepsilon^3 N^8 \langle\mu_1(k,\ell)\rangle^s \frac{\langle\mu_2(k,\ell)\rangle^{-s}\dots\langle\mu_6(k,\ell)\rangle^{-s}}{\langle\mu_{\min}(k,\ell)\rangle^{-2s}}\|u/\varepsilon\|_s^5.$$

Les premiers et derniers facteurs sont des termes d'homogénéité : si $u$ est dans la boule de taille $\varepsilon$, il s'agit d'un terme d'ordre $3^-$. Le facteur $\langle\mu_1(k,\ell)\rangle^s$ vient du contrôle du champ de vecteurs dans $\ell^1_s$, comme on l'a vu précédemment, il peut être facilement compensé par le terme $\langle\mu_2(k,\ell)\rangle^{-s}$ du numérateur. Enfin, puisque ici $\mu_6(k,\ell)=\mu_{\min}(k,\ell)$, le dénominateur peut être compensé par $\langle\mu_5(k,\ell)\rangle^{-s}\langle\mu_6(k,\ell)\rangle^{-s}$.

Par ce changement de variables, on a pu éliminer les termes d'ordre $6$ non intégrables. Le hamiltonien s'écrit donc[33]

$$H_{\mathsf{NLS}}\circ\tau^{-1} = Z_2(I)+Z_4(I)+Z_6(I)+\sum_{m=4}^{r} K_{2m}(z)+\mathcal{O}(\varepsilon^{2r+1}),$$

où le terme de reste a son champ de vecteurs contrôlé par $\varepsilon^{2r+1}$ et $K_{2m}$ est une fraction rationnelle homogène d'ordre $2m$, s'écrivant comme une somme de termes de la forme

$$u_{k_1}\dots u_{k_m}\overline{u_{\ell_1}\dots u_{\ell_m}}/P_{k,\ell}(I)$$

avec $P_{k,\ell}$ un polynôme[34] et $(k,\ell)$ vérifiant $k_1^\alpha+\dots+k_m^\alpha=\ell_1^\alpha+\dots+\ell_m^\alpha$ pour $\alpha=1,2$.

Grâce à la relation algébrique

$$\{\frac{1}{P_{k,\ell}(I)}u_{k_1}\dots u_{k_m}\overline{u_{\ell_1}\dots u_{\ell_m}}, Z_4(I)\} = \frac{1}{P_{k,\ell}(I)}\{u_{k_1}\dots u_{k_m}\overline{u_{\ell_1}\dots u_{\ell_m}}, Z_4(I)\},$$

32. ces derniers peuvent être contrôlés similairement.
33. pour un nouveau changement de variable toujours noté $\tau$.
34. dont le degré est contrôlé par rapport à $m$

il semble envisageable de supprimer les fractions rationnelles ne dépendant pas que des actions en faisant, comme pour les termes d'ordre $6$, un changement de variables s'appuyant sur $Z_4$. On pourrait d'ailleurs effectivement faire une telle construction pour les termes d'ordre $8$. Cependant, il existe, à partir de l'ordre $10$, des termes que l'on ne parvient pas à supprimer car le champ de vecteur du hamiltonien qu'il faudrait introduire n'est pas petit. Plus précisément, à l'ordre 10, il y a des termes de la forme

$$\frac{u_{k_1}u_{k_2}u_{k_3}\overline{u_{\ell_1}u_{\ell_2}u_{\ell_3}}I_0^4}{(I_{k_1}+I_{k_2}+I_{k_3}-I_{\ell_1}-I_{\ell_2}-I_{\ell_3})^2} \tag{22}$$

où $(k,\ell)$ vérifient exactement les mêmes hypothèses que les monômes d'ordre $6$ que l'on a résolu. Pour supprimer un tel terme, le hamiltonien que l'on devrait introduire contiendrait un terme de la forme

$$\chi = \frac{u_{k_1}u_{k_2}u_{k_3}\overline{u_{\ell_1}u_{\ell_2}u_{\ell_3}}I_0^4}{i(I_{k_1}+I_{k_2}+I_{k_3}-I_{\ell_1}-I_{\ell_2}-I_{\ell_3})^3}.$$

En omettant les termes dus à une dérivation du dénominateur, l'estimation naturelle du champ de vecteurs associé à cet hamiltonien est du type

$$\|X_\chi\|_s \lesssim \varepsilon^9 N^{24}\langle\mu_1(k,\ell)\rangle^s \frac{\langle\mu_2(k,\ell)\rangle^{-s}\dots\langle\mu_6(k,\ell)\rangle^{-s}}{\langle\mu_{\min}(k,\ell)\rangle^{-6s}}\|u/\varepsilon\|_s^{15}.$$

On voit apparaitre un problème : on ne parvient pas à compenser le dénominateur par des facteurs présents au numérateur. Pour s'en convaincre, on peut considérer un cas "critique" pour lequel $|\mu_1(k,\ell)| \simeq |\mu_6(k,\ell)| \simeq N$. Auquel cas, l'estimation devient $\|X_\chi\|_s \leqslant \varepsilon^9 N^{24+2s}\|u/\varepsilon\|_s^{15}$. Or $N$ avait été construit pour que $N^{-s}$ soit un terme d'ordre $2r$ en $\varepsilon$, on a donc $\varepsilon^9 N^{24+2s} \simeq \varepsilon^{(-4r+9)^-} \gg 1$.

Sur cet exemple, la principale difficulté provient du terme $I_0^4$ au numérateur. Dans la boule de taille $\varepsilon$ de $\ell^1_s$, il est typiquement de taille $\varepsilon$, ce qui ne permet pas de compenser les petits dénominateurs. Des termes problématiques de ce type apparaissent dans le développement du hamiltonien à partir de la résolution des termes d'ordre $6$. Plus précisément, si $\chi$ est le hamiltonien donné par (20), alors $\{\chi,\{\chi,Z_6\}\}$ contient naturellement un terme de la forme (22). Si $Z_6$ avait une forme similaire à $Z_4$ en étant un *polynôme symétrique en les actions*, de tels termes n'apparaitraient pas [35]. Cependant, la partie non symétrique de $Z_6$ est non nulle et est de la forme

$$\varphi'(0)^2\sum_{a\neq b}\frac{I_a^2 I_b}{(a-b)^2}.$$

Ce terme apparait à partir de la résolution des termes non résonants d'ordre $4$ comme une partie de $\{\chi,P_4\}$ où $\chi$ est donné par (12) et $P_4$ est la partie quartique de $H_{\text{NLS}}$. Il semble donc empêcher une mise sous forme normale au delà de l'ordre $10$. Cependant, en adaptant le processus de mise sous forme normale, ce terme peut s'avérer très utile. Pour mieux comprendre cette affirmation, retournons à (NLSP).

De même que pour (NLS), on peut facilement construire un premier de changement de variables symplectique $\tau$, proche de l'identité, tel que

$$H_{\text{NLSP}}\circ\tau^{-1} = Z_2(I)+Z_4(I)+\mathcal{O}(z^6)$$

35. il faudrait bien sûr faire le même genre d'hypothèse sur $Z_{2m}(I)$ avec $m\geqslant 4$ pour que des termes problématiques n'apparaissent pas à des ordres encore supérieurs.

où ici $Z_4(I)$ est très similaire à la partie non symétrique de $Z_6(I)$ (pour (NLS)) :

$$Z_4(I) = \sum_{a \neq b} \frac{I_a I_b}{(a-b)^2}.$$

Supposons que l'on veuille résoudre (avec $Z_4$) un terme de la forme

$$f(I) u_{k_1} \dots u_{k_m} \overline{u_{\ell_1} \dots u_{\ell_m}}$$

où $f(I)$ est une fonction ne dépendant que des actions, $k, \ell$ vérifient $|\mu_3(k,\ell)| \geqslant N$ et $k_1^\alpha + \dots + k_m^\alpha = \ell_1^\alpha + \dots + \ell_m^\alpha$ pour $\alpha = 1, 2$ et le numérateur ne dépend pas que des actions. On est naturellement amené à considérer le changement de variables induit par le hamiltonien

$$\chi = f(I) \frac{u_{k_1} \dots u_{k_m} \overline{u_{\ell_1} \dots u_{\ell_m}}}{i\omega_{k,\ell}(I)}$$

où

$$\omega_{k,\ell}(I) := \left( \sum_{j=1}^m \partial_{I_{k_j}} - \partial_{I_{\ell_j}} \right) Z_4(I) = \sum_{a \in \mathbb{Z} \setminus k,\ell} I_a \sum_{j=1}^m \frac{1}{(a-k_j)^2} - \frac{1}{(a-\ell_j)^2} + L_{k,\ell}(I_k, I_\ell)$$

avec $L_{k,\ell}(I_k, I_\ell)$ une fonction linéaire de $I_{k_1}, \dots, I_{k_m}, I_{\ell_1}, \dots, I_{\ell_m}$. On peut montrer qu'il existe toujours au moins un indice $a$ dans la somme de droite, inférieur à $3m$ et tel que le coefficient associé à $I_a$ soit non nul. On en déduit directement, qu'avec une grande probabilité, on a un contrôle de $\omega_{k,\ell}(I)$ du type

$$\omega_{k,\ell}(I) \gtrsim \varepsilon^2 N^{\alpha_m} \simeq \varepsilon^{2^+},$$

où $\alpha_m$ ne dépend pas de $s$ mais que de $m$. Contrairement au cas de (NLS), il n'y a pas le facteur $\langle \mu_{\min}(k,\ell) \rangle^{-2s}$. Ainsi grâce à la présence d'actions associées à des fréquences indépendantes de $k$ et $\ell$ au dénominateur, on parvient à résoudre des termes d'ordres arbitrairement élevés avec $Z_4$ (pour (NLSP)).

On voudrait donc procéder de même avec $Z_6$ pour (NLS). Cependant un simple argument de degré montre qu'une telle construction n'est pas possible, on n'obtiendrait pas des termes d'ordres de plus en plus élevés. En effet, si $\{\chi, Z_6\}$ permettait de simplifier un terme d'ordre $2m$ alors $\chi$ serait d'ordre $2m-4$. Or le changement changement de variables ferait aussi apparaitre un terme de la forme $\{\chi, Z_4\}$ qui est d'ordre $2m-2$ et qui n'a aucun raison de ne dépendre que des actions. Pour surmonter cette difficulté, on adopte une construction de forme normale inspirée de KAM : on résout avec le polynôme inhomogène $Z_4 + Z_6$. Cette dernière fait apparaitre des dénominateurs de la forme

$$\Omega_{k,\ell}(I) = \left( \sum_{j=1}^m \partial_{I_{k_j}} - \partial_{I_{\ell_j}} \right) (Z_4(I) + Z_6(I)).$$

Ces derniers peuvent typiquement être minorés par des quantités du type

$$N^{\alpha_m} \max(\varepsilon^2 \langle \mu_{\min}(k,\ell) \rangle^{-2s}, \varepsilon^4).$$

Autrement dit, ils vérifient à la fois l'estimation que l'obtiendrait pour $Z_4$ seul et celle pour $Z_6$. On peut donc soit voir le dénominateur comme un terme d'ordre $4$, soit comme un terme d'ordre

$2$ devenant dégénéré pour les hautes fréquences. Ce constat nous permet, via la construction d'une classe d'hamiltoniens sous forme de fractions rationnelles inhomogènes, de résoudre tous les termes d'ordres supérieurs ou égaux à $8$ pour (NLS). La construction et l'analyse de cette classe est la partie la plus technique de notre preuve. C'est l'inhomogénéité de ces fractions rationnelles qui rend possible l'existence de résonances liant $u^0$ et $\varepsilon$, expliquant ainsi les différences entre le Théorème 9 pour (NLSP) et le Théorème 10 pour (NLS).

### Quelques comportements non linéaires de solutions des équations de Vlasov-Poisson

Le cinquième chapitre de cette thèse est le fruit d'une collaboration avec Michel Mehrenberger traitant de la dynamique de certaines solutions des équations de Vlasov-Poisson

$$\begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x \phi \cdot \nabla_v f = 0 \\ \Delta_x \phi = n(f) - \displaystyle\int_{\mathbb{R}^d} f \mathrm{d}v \\ f(t=0) = f_0. \end{cases} \tag{VP}$$

où $f = f(t,x,v) : \mathbb{R} \times \mathbb{T}_d \times \mathbb{R}^d \to \mathbb{R}$ est une fonction de *distribution*, $\phi = \phi(t,x) : \mathbb{R} \times \mathbb{T}_d \to \mathbb{R}$ est un *potentiel électrique*, $\mathbb{T}_d = \mathbb{R}/L_1\mathbb{Z} \times \cdots \times \mathbb{R}/L_d\mathbb{Z}$ est un tore de dimension $d \geqslant 1$ et $n(f) = \iint_{\mathbb{T}_d \times \mathbb{R}^d} f \mathrm{d}x\mathrm{d}v$. D'un point de vue physique, il s'agit d'un modèle simple pouvant décrire l'évolution d'un *plasma*.

Ce travail trouve son origine dans un projet CEMRACS réalisé en 2016 avec Yann Barsamian, Sever Hirstoaga et Michel Mehrenberger et ayant donné lieu au proceeding [20]. Son objectif était l'implémentation, au sein de la bibliothèque SeLaLib [36], de méthodes *semi-lagrangiennes* et *particulaires* (Particules In Cell), pour la résolution de (VP) en dimension $d = 2$ (et aussi avec deux espèces). Ce projet comprenait donc aussi un travail de validation des résultats obtenus avec ces codes. On était donc amené à déterminer numériquement des approximations de solutions relativement bien connues de (VP). Pour en obtenir, une façon naturelle et classique [37] de procéder consiste à étudier les solutions des équations de Vlasov-Poisson linéarisées autour d'un état d'équilibre homogène $f^{eq} \equiv f^{eq}(v)$ [38]. Plus précisément, on cherche une solution de (VP) sous la forme

$$\begin{cases} f &= f^{eq} &+ \varepsilon g \\ \phi &= 0 &+ \varepsilon \psi. \end{cases}$$

En négligeant les termes d'ordre supérieur ou égal à $2$, on trouve alors que $g$ est formellement solution de

$$\begin{cases} \partial_t g + v \cdot \nabla_x g - \nabla_x \psi \cdot \nabla_v f^{eq} = 0, \\ \Delta_x \psi + \displaystyle\int_{\mathbb{R}^d} g \, \mathrm{d}v = 0, \\ g(t=0) = g_0. \end{cases} \tag{VPL}$$

36. `http://selalib.gforge.inria.fr/`
37. voir par exemple [106].
38. On peut remarquer que toute fonction de distribution ne dépendant que de $v$ est un état d'équilibre de (VP).

Il s'agit d'une équation linéaire et homogène. Il est donc naturel de chercher à la résoudre en faisant une transformation de Fourier selon dans la variable $x$. On obtient alors

$$\begin{cases} \partial_t \widehat{g} + i(v\cdot k)\widehat{g} - i\widehat{\psi}(k\cdot\nabla_v)f^{eq} = 0, \\ -|k|^2\widehat{\psi} + \displaystyle\int_{\mathbb{R}^d} \widehat{g}\,\mathrm{d}v = 0, \\ \widehat{g}(t=0) = \widehat{g}_0, \end{cases} \tag{VPLF}$$

où $k \in \widehat{\mathbb{T}}_d = 2\pi/L_1\mathbb{Z} \times \cdots \times 2\pi/L_d\mathbb{Z}$ et la transformée de Fourier en espace est définie pour $u \in L^1(\mathbb{T}_d)$ par

$$\widehat{u}(k) = \left(\prod_{j=1}^{d} L_j\right)^{-1} \int_{\mathbb{T}_d} u(x)e^{-ik\cdot x}\mathrm{d}x.$$

Il est important de remarquer sur (VPLF) qu'il n'y a pas *d'échange d'énergie* entre les modes au niveau linéaire. Autrement dit, si $g$ est une solution de (VPL) telle que $g_0 \equiv \widehat{g_0}(v)e^{ik\cdot x}$ alors elle est de la forme $g(t,x,v) = \widehat{g}(t,v)e^{ik\cdot x}$. Il s'agit d'une première limitation pour tester la capacité du code à reproduire des phénomènes multidimensionnels.

Puisque (VPLF) est une équation linéaire et autonome, il est naturel de chercher à la résoudre par transformation de Laplace. Cette dernière étant définie pour des fonctions $u : \mathbb{R}_+^* \to \mathbb{C}$ telles qu'il existe $\lambda \in \mathbb{R}$ satisfaisant $ue^{-\lambda t} \in L^\infty(\mathbb{R}_+^*)$ et pour des valeurs $z \in \mathbb{C}$ telles que $\Im z > \lambda$ par

$$\mathcal{L}[u](z) = \int_0^\infty u(t)e^{izt}\mathrm{d}t.$$

On peut alors prouver que les solutions de (VPLF) sont données par

$$g(t,x,v) = \sum_{k\in\widehat{\mathbb{T}}_d} e^{ik\cdot(x-vt)}\widehat{g_0}(k,v) + i\int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\psi}(s,k)k\cdot\nabla_v f^{eq}(v)\mathrm{d}s, \tag{23}$$

et pour $\Im z$ suffisamment grand

$$\mathcal{L}\left[\widehat{\psi}(t,k)\right](z) = \frac{N_k(z)}{D_k(z)} =: M_k(z). \tag{24}$$

où $N_k$ et $D_k$ sont des fonctions entières définies lorsque $\Im z$ est suffisamment grand par

$$N_k(z) = -\frac{i}{|k|^2}\int_{\mathbb{R}^d} \frac{\widehat{g_0}(k,v)}{v\cdot k - z}\mathrm{d}v \quad \text{et} \quad D_k(z) = 1 - \frac{1}{|k|^2}\int \frac{k\cdot\nabla_v f^{eq}(v)}{v\cdot k - z}\mathrm{d}v. \tag{25}$$

Ainsi, pour obtenir une solution $g$ de (VPL) par (23) il faut être en mesure de résoudre l'équation (24) (appelée *relation de dispersion*) en déterminant une transformée de Laplace inverse. Pour faire cela, moyennant des hypothèses fortes sur $f^{eq}$ et $g_0$, on montre que, pour tout $\lambda \in \mathbb{R}$, $D_k$ ne s'annule qu'en un nombre fini de points dont la partie imaginaire est plus grande que $\lambda$. Ainsi grâce à la formule

$$\mathcal{L}[t^m e^{-i\omega t}](z) = \frac{i^{m+1}m!}{(z-\omega)^{m+1}},\ \omega \in \mathbb{C}, \tag{26}$$

et à une estimation précise des termes de reste, on est en mesure de prouver que (24) admet une solution analytique $\widehat{\psi}$ dont le développement asymptotique est donné, pour tout $\lambda \in \mathbb{R}$, par

$$\widehat{\psi}(t,k) = \sum_{\substack{D_k(\omega)=0 \\ \Im\omega \geqslant \lambda}} P_{\omega,k}(t) e^{-i\omega t} + \mathcal{O}(e^{\lambda t}), \tag{27}$$

où $P_{\omega,k}$ est le polynôme tel que $M_k(z) \underset{z\to\omega}{=} \mathcal{L}[P_{\omega,k}(t)e^{-i\omega t}](z) + \mathcal{O}(1)$ soit le développement $M_k(z)$ en $\omega$.

Une telle analyse de l'équation linéarisée remonte à Landau en 1946 dans [90]. Elle a été rendue rigoureuse et généralisée en 1986 par Degond dans [59]. Elle donne notamment une première explication au phénomène *d'amortissement Landau*. Il correspond au cas où, pour tout $k \in \widehat{\mathbb{T}}_d$, les zéros de $D_k$ ont tous une partie imaginaire strictement négative. Dans ce cas, le potentiel électrique converge exponentiellement vite vers $0$. L'existence de ce phénomène a été prouvée pour les équations non linéaires (VP) en 2011 par Mouhot et Villani dans [93] puis reprise en $2016$ par Bedrossian, Masmoudi et Mouhot dans [22].

Comme on vient de le voir, la linéarisation des équations de (VP) ne permet pas d'expliquer de phénomènes vraiment multidimensionnels. De plus, évidement, elle ne saurait expliquer de phénomènes non linéaires. Afin de produire des *cas test* plus pertinents, on avait été amené dans [20] à considérer la dynamique du second terme dans le développement de $f$ en puissance de $\varepsilon$. Plus précisément, on cherche une solution à (VP) sous la forme

$$\left\{ \begin{array}{lllllllll} f & = & f^{eq} & + & \varepsilon g & + & \varepsilon^2 h & + & o(\varepsilon^2), \\ \phi & = & 0 & + & \varepsilon\psi & + & \varepsilon^2\mu & + & o(\varepsilon^2), \end{array} \right.$$

où $h(t=0) \equiv 0$. En négligeant les termes d'ordre $3$ en $\varepsilon$, on montre alors que $(h,v)$ est solution de

$$\left\{ \begin{array}{l} \partial_t h + v\cdot\nabla_x h - \nabla_x\mu\cdot\nabla_v f^{eq} = \nabla_x\psi\cdot\nabla_v g, \\ \Delta_x\mu + \displaystyle\int_{\mathbb{R}^d} h \,\mathrm{d}v = 0, \\ h(t=0) = 0. \end{array} \right. \tag{VPL2}$$

On reconnait les équations de Vlasov-Poisson linéarisées, avec condition initiale nulle mais un terme source les rendant non autonomes et non homogènes. On peut retrouver une étude de ces équations en dimension $d=1$ dans la littérature physique [101]. Dans le proceeding [20], une étude formelle de ces équations avait été réalisée grâce à une résolution par la formule de Duhamel. Elle avait mis en évidence certains comportements non linéaires et multidimensionnels permettant ainsi de réaliser des cas tests pertinents et d'expliquer les résultats de certaines simulations (notamment une présente dans [10]). Par exemple, on avait pu construire des solutions dont le champ électrique est amorti à l'ordre $1$ mais explose à l'ordre $2$. Cependant, cette analyse n'était pas complètement rigoureuse et ne nous avait pas permis d'expliquer un phénomène de résonance engendrant des ondes dont les fréquences sont appelées *fréquences de Best* (voir [101]).

Ici, comme dans le cas linéaire, on résout (VPL2) par transformation de Laplace en temps et transformation de Fourier en espace. Plus précisément, quelques calculs montrent que (VPL2)

est équivalent à

$$h(t,x,v) = \sum_{k\in\widehat{\mathbb{T}_d}} i\int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\mu}(s,k)k\cdot\nabla_v f^{eq}(v)\mathrm{d}s + \int_0^t \nabla_x\widehat{\psi\cdot\nabla}_v g(s,k,v)e^{ik\cdot(x-v(t-s))}\mathrm{d}s, \tag{28}$$

et pour $\Im z$ suffisamment grand

$$\mathcal{L}\left[\widehat{\mu}(t,k)\right](z) = \frac{\mathcal{N}_k(z)}{D_k(z)} =: \mathcal{M}_k(z), \tag{29}$$

où $D_k$ est donnée par (25) et $\mathcal{N}_k$ est une fonction méromorphe sur $\mathbb{C}$ explicitement connue.

Comme précédemment, il suffit donc de pouvoir inverser une transformation de Laplace pour résoudre (VPL2). Là encore, comme dans le cas linéaire, cette inversion passe par l'étude des pôles de $\mathcal{M}_k$, nous donnant par la même occasion un développement asymptotique de $\mu$ sous la forme :

$$\forall\lambda\in\mathbb{R},\ \widehat{\mu}(t,k) = \sum_{\substack{\mathcal{M}_k(\omega)=\infty\\ \Im\omega\geqslant\lambda}} Q_{\omega,k}(t)e^{-i\omega t} + \mathcal{O}(e^{\lambda t}). \tag{30}$$

où $Q_{\omega,k}$ est le polynôme tel que $\mathcal{M}_k(z) \underset{z\to\omega}{=} \mathcal{L}[Q_{\omega,k}(t)e^{-i\omega t}](z) + \mathcal{O}(1)$.

Les pôles de $\mathcal{M}_k$ peuvent soit être des zéros de $D_k$, donnant ainsi les même fréquences que dans le cas linéaires, soit être des pôles de $\mathcal{N}_k$. L'étude de ces pôles n'est pas simple car $\mathcal{N}_k$ s'exprime naturellement en fonction de la solution de (VPL). Cependant, à l'aide du développement asymptotique de $\psi$ (voir (27)), on est en mesure de décomposer $\mathcal{N}_k$ en une somme de termes dont on peut déterminer les pôles.

Afin de donner une intuition de ce que peuvent être ces pôles et dans quelle mesure il peuvent être associés à des phénomènes de résonances, on considère un terme particulièrement représentatif [39] de cette décomposition :

$$\mathcal{N}_k^{(rep)}(z) = \mathcal{L}\left[F_k^{(rep)}(t)\right](z)$$

où

$$F_k^{(rep)}(t) = e^{-i(\omega_1 t+\omega_2 t)} \iint\limits_{0\leqslant\tau\leqslant s\leqslant t} e^{i(\omega_1\tau+\omega_2 s)}\mathscr{F}[f^{eq}](\tau k_1 + sk_2)\,\mathrm{d}s\,\mathrm{d}\tau \tag{31}$$

avec $k_1, k_2 \in \widehat{\mathbb{T}_d}\backslash 0$ satisfaisant $k_1+k_2=k$, $\omega_1,\omega_2\in\mathbb{C}$ tels que $D_{k_1}(\omega_1) = D_{k_2}(\omega_2) = 0$ et $\mathscr{F}[f^{eq}]$ la transformée de Fourier de $f^{eq}$.

Puisque $\mathcal{N}_k^{(rep)}$ s'exprime comme la transformée de Laplace de $F_k^{(rep)}(t)$, on peut démontrer que ses pôles sont donnés par le développement asymptotique de $F_k^{(rep)}(t)$ grâce à la formule (26). Comme le suggère la formule (31), ce dernier ne fait pas apparaitre les mêmes termes selon si l'ensemble des $(\tau,s)$ tels que $\tau k_1 + sk_2 = 0$ est un segment (*cas résonant*) ou un point (*cas non-résonant*).

Dans le cas non-résonant, il existe une constante $c>0$ telle que

$$0\leqslant\tau\leqslant s\leqslant t,\ |\tau k_1 + sk_2| \geqslant cs.$$

39. bien que légèrement simplifié.

Ainsi, en supposant $f^{eq}$ suffisamment régulière pour que $\mathscr{F}[f^{eq}]$ décroisse plus vite que toute exponentielle en l'infini (par exemple comme une gaussienne), on peut montrer que l'intégrale double dans (31) converge plus vite que toute exponentielle quand $t$ tends vers $+\infty$. Cela permet d'obtenir une constante $a \in \mathbb{C}$ telle que

$$\forall \lambda \in \mathbb{R},\ F_k^{(rep)}(t) = ae^{-i(\omega_1 t + \omega_2 t)} + \mathcal{O}(e^{\lambda t}).$$

Dans le cas résonnant, il existe $\gamma \in (0,1)$ tel que

$$k_2 = -\gamma k_1.$$

En réalisant, dans (31), un changement de variable naturel, on obtient alors

$$F_k^{(rep)}(t) = \int_0^t \int_{-\gamma s}^{(1-\gamma)s} e^{-i(\omega_1(t-\tau-\gamma s)+\omega_2(t-s))} \mathscr{F}[f^{eq}](\tau k_1)\ \mathrm{d}\tau\ \mathrm{d}s.$$

Donc en supposant que $f^{eq}$ est suffisamment régulière pour que $\mathscr{F}[f^{eq}]$ décroisse plus vite que toute exponentielle, on a

$$\begin{aligned} F_k^{(rep)}(t) = &\left(\int_0^t e^{-i(\omega_1(t-\gamma s)+\omega_2(t-s))}\mathrm{d}s\right)\left(\int_{\mathbb{R}} e^{i\omega_1\tau}\mathscr{F}[f^{eq}](\tau k_1)\mathrm{d}\tau\right) \\ &- e^{-it(\omega_1+\omega_2)} \int_0^\infty \int_{\substack{\tau \geqslant (1-\gamma)s \\ \text{ou } \tau < -\gamma s}} e^{i(\omega_1(\tau+\gamma s)+\omega_2 s)} \mathscr{F}[f^{eq}](\tau k_1)\ \mathrm{d}\tau\ \mathrm{d}s \\ &+ e^{-it(\omega_1+\omega_2)} \int_t^\infty \int_{\substack{\tau \geqslant (1-\gamma)s \\ \text{ou } \tau < -\gamma s}} e^{i(\omega_1(\tau+\gamma s)+\omega_2 s)} \mathscr{F}[f^{eq}](\tau k_1)\ \mathrm{d}\tau\ \mathrm{d}s. \end{aligned}$$

Toujours sous cette même hypothèse, on peut montrer que le troisième terme décroit plus vite que toute exponentielle. Ainsi, cette décomposition permet d'obtenir le développement asymptotique suivant

$$\forall \lambda \in \mathbb{R},\ F_k^{(rep)}(t) = ae^{-it(\omega_1+\omega_2)} + be^{-it\omega_b} + \mathcal{O}(e^{\lambda t}),$$

où $a, b \in \mathbb{C}$ et $\omega_b = (1-\gamma)\omega_1 = (|k|/|k_1|)\omega_1$ est la *fréquence de Best*.

Comme le suggère cette esquisse de démonstration, on peut donc démontrer que $\mathcal{M}_k$ a trois types de pôles. Plus précisément, si $\omega$ est un pôle de $\mathcal{M}_k$ alors il vérifie l'un des conditions suivantes

(I) $\omega$ est un zéro de $D_k$

(II) $\omega = \omega_1 + \omega_2$ où $D_{k_1}(\omega_1) = D_{k_2}(\omega_2) = 0$ et $k_1 + k_2 = k$

(III) $\omega = (|k|/|k_1|)\omega_1$ où $D_{k_1}(\omega_1) = 0$ et il existe $\gamma \in (0,1)$ tel que $k = \gamma k_1$.

Ce sont ces pôles qui gouvernent le comportement asymptotique de $\widehat{\mu}(k)$ grâce à la formule (30).

Pour conclure la présentation de ce chapitre, on va donner le théorème précis qui y est démontré. Mais pour cela on a besoin d'introduire quelques notations.

- Si $k \in \widehat{\mathbb{T}}_d$, $n_{k,\omega}$ est la multiplicité de $\omega$ en tant que zéro de $D_k$

- $\mathscr{E}(\mathbb{R}^d)$ désigne l'espace des fonctions $f$ appartenant à l'espace de Schwartz $\mathscr{S}(\mathbb{R}^d)$ dont la transformée de Fourier, $\mathscr{F}f$, admet un prolongement holomorphe sur $\mathbb{C}^d$ et vérifie l'estimation suivante

$$\exists \alpha \in (0, \pi/2), \forall \beta \in (0, \alpha), \forall \lambda \in \mathbb{R}, \sup_{x \in \mathbb{R}^d} \sup_{\theta \in (-\beta, \beta)} e^{\lambda |x|} |\mathscr{F}f(e^{i\theta}x)| < \infty.$$

L'espace $\mathscr{E}(\mathbb{R}^d)$ contient de nombreuses[40] fonctions de distribution intervenant naturellement dans le cadre des équations de Vlasov-Poisson.

Cela nous permet donc d'obtenir le théorème suivant (qui est le principal résultat du cinquième chapitre de cette thèse)

**Théorème 11.** *Soit $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ et $g_0$ une fonction de la forme*

$$\forall x \in \mathbb{T}_d,\ g_0(x, v) = \sum_{k \in K} e^{ik \cdot x} \widehat{g_0}(k, v), \quad \textit{où } v \mapsto \widehat{g_0}(k, v) \in \mathscr{E}(\mathbb{R}^d)$$

*où $K$ est une partie finie de $\widehat{\mathbb{T}}_d \backslash \{0\}$. Il existe deux fonctions $\psi, \mu : \mathbb{R}_+^* \times \mathbb{T}_d \to \mathbb{R}$ de classe $C^\infty$ et deux fonctions continues $g, h : \mathbb{R}_+ \times \mathbb{T}_d \times \mathbb{R}^d \to \mathbb{R}$, $C^\infty$ sur $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$, telles que $(g, \psi, h, \mu)$ soit solution de* (VPL) *et* (VPL2)*.*

*De plus, si $\lambda \in \mathbb{R}$, $\psi$ est une combinaison linéaire de fonctions du type*

$$J(t, x) = t^m e^{i(k \cdot x - \omega t)} \textit{ et } R(t, x) = r(t) e^{i(k \cdot x - i\lambda t)}$$

*où $k \in K$, $D_k(\omega) = 0$, $\Im\omega \geqslant \lambda$, $0 \leqslant m < n_{k,\omega}$ et $r$ est une fonction analytique bornée sur $\mathbb{R}_+^*$.*

*De même, $\mu$ est une combinaison linéaire de fonctions du type*

$$\begin{array}{rclrcl} J(t,x) & = & t^m e^{i(k \cdot x - \omega t)} & I(t,x) & = & t^\ell e^{i(k \cdot x - (\omega_1 + \omega_2)t)} \\ B(t,x) & = & t^p e^{i(k \cdot x - \frac{|k|}{|k_1|}\omega_1 t)} d_{k_1}^{k_2} & R(t,x) & = & r(t) e^{i(k \cdot x - i\lambda t)} \end{array}$$

*où $k = k_1 + k_2$, $r$ est une fonction analytique bornée sur $\mathbb{R}_+^*$ et $k_1, k_2 \in K$ satisfont*

$$\begin{cases} D_k(\omega) = D_{k_1}(\omega_1) = D_{k_2}(\omega_2) = 0 \\ k \cdot k_1 \neq 0 \textit{ et } \left( d_{k_1}^{k_2} \neq 0 \iff \exists \gamma \in (0,1),\ k = \gamma k_1 \right) \\ \Im\omega \geqslant \lambda \textit{ et } \left( \Im\omega_1 + \Im\omega_2 \geqslant \lambda \textit{ ou } \frac{|k|}{|k_1|} \Im\omega_1 \geqslant \lambda \right) \\ m < n_{k,\omega},\ \ell < n_{k_1,\omega_1} + n_{k_2,\omega_2} - 1 + \sigma_{\omega_1,\omega_2}^{k_1,k_2},\ p < n_{k_1,\omega_1} + 1 + \nu_{\omega_1,\omega_2}^{k_1,k_2}, \end{cases}$$

*avec $\sigma_{\omega_1,\omega_2}^{k_1,k_2}, \nu_{\omega_1,\omega_2}^{k_1,k_2}$ des entiers positifs égaux à zéros dans les cas non dégénérés.*

Les cas dégénérés sont expliqués en détails dans le cinquième chapitre de cette thèse mais il semble qu'ils ne se produisent jamais en pratique. L'hypothèse consistant à prendre des conditions initiales étant des polynômes trigonométriques par rapport à $x$ est clairement trop forte. Cependant, elle est suffisante pour construire des cas tests intéressants et mettre en évidence le phénomène de résonance associé aux fréquences de Best. De plus, elle permet de ne pas alourdir la preuve en évitant de nombreux problèmes de convergence.

Dans le cinquième chapitre de cette thèse, on présente de nombreuses simulations numériques confirmant les résultats de ce théorème sur les ondes de Best. Les autres ondes avaient été observées et étudiées numériquement dans le proceeding [20].

40. il contient notamment les fonctions appartenant aux espaces de *Gelfand-Shilov*, $S_{1-\alpha}^{\alpha}(\mathbb{R}^d)$ pour $\alpha \in (0, 1)$.

## Méthodes de splitting pour les rotations et leurs applications aux équations de Vlasov

Le sixième chapitre de cette thèse est le fruit d'une collaboration avec Fernando Casas et Nicolas Crouseilles avec l'appui de Pierre Navaro. On y revisite le problème classique de la résolution numérique, par des méthodes de *splitting*, du problème de Cauchy pour l'équation de transport associée à une rotation du plan. Puis, on applique les nouvelles méthodes obtenues à la résolution des équations de *Vlasov-Maxwell* et *Vlasov-HMF*.

On considère le problème de Cauchy pour l'équation de transport associée à l'oscillateur harmonique sous la forme

$$\left\{ \begin{array}{lcl} \partial_t u(t,x) & = & (x_1\partial_{x_2} - x_2\partial_{x_1})u(t,x) \\ u(t=0,x) & = & u_0(x) \end{array} \right. , \ t \in \mathbb{R}, \ x = (x_1, x_2) \in \mathbb{R}^2 \tag{ROT}$$

où $u_0$ est une fonction régulière et bien localisée. Cette équation de transport est associée à un mouvement de rotation (à vitesse angulaire $-1$) dans la mesure où ses solutions s'écrivent

$$u(t) \equiv e^{tJx\cdot\nabla} u_0 = u_0 \circ e^{tJ} \tag{32}$$

où $J$ est la matrice de la forme symplectique dans la base canonique de $\mathbb{R}^2$, i.e.

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

Une façon naturelle de résoudre (ROT) consiste à utiliser une méthode dite *semi-lagrangienne*. Pour faire simple, on peut résumer la démarche de la façon suivante. On suppose que $u_0$ n'est connu que sur une grille et on voudrait connaitre aussi précisément que possible les valeurs de $u(t)$ sur cette même grille. Or, d'après la formule (32), la valeur de $u(t)$ en un point $g$ de la grille correspond à la valeur de $u_0$ au point $e^{-tJ}g$. On détermine donc une approximation de $u(t)$ en réalisant une interpolation de la valeur de $u_0$ en $e^{-tJ}g$ à partir de ses valeurs sur la grille.

Cette méthode donne des résultats précis mais est relativement lente car elle nécessite une interpolation en dimension $2$. Pour obtenir des méthodes ne nécessitant que des interpolations en dimension $1$, il est naturel de chercher à résoudre numériquement (ROT) à l'aide de méthodes de *splitting*. Les deux méthodes les plus élémentaires sont celles de Lie et de Strang. Elles consistent à approcher, sur un pas de temps $\delta_t \ll 1$, les solutions (ROT) par un *produit de cisaillements*, selon les formules

$$\left\{ \begin{array}{lcll} e^{\delta_t(x_1\partial_{x_2} - x_2\partial_{x_1})} & \simeq & e^{\delta_t x_1\partial_{x_2}} e^{-\delta_t x_2\partial_{x_1}} & \text{(LIE)} \\ e^{\delta_t(x_1\partial_{x_2} - x_2\partial_{x_1})} & \simeq & e^{-(\delta_t/2)x_2\partial_{x_1}} e^{\delta_t x_1\partial_{x_2}} e^{-(\delta_t/2)x_2\partial_{x_1}} & \text{(STRANG)} \end{array} \right.$$

Ces méthodes sont d'autant plus efficaces qu'elles sont naturellement parallélisables. Cependant elles n'ont pas de raison, a priori, de respecter les symétries de (ROT) car elles consistent à approcher le semi-groupe engendré par un opérateur isotrope par un produit de semi-groupes associés à des opérateurs ne l'étant pas. Afin de mesurer ce défaut, on réalise *l'analyse rétrograde* de ces méthodes. Plus précisément, on cherche à écrire exactement chacun de ces produits de semi-groupes comme un semi-groupe car l'étude des opérateurs associés permet alors une description fine de leurs dynamiques. Le théorème suivant fournit l'analyse rétrograde de méthodes de splitting incluant ceux de Lie et Strang.

**Théorème 12.** *Si $a, b \in \mathbb{R}$ vérifient $ab < 2$ alors*

$$e^{bx_1\partial_{x_2}} e^{-ax_2\partial_{x_1}} = e^{JL_{a,b}x\cdot\nabla}, \tag{33}$$

*et*

$$e^{-\frac{a}{2}x_2\partial_{x_1}} e^{bx_1\partial_{x_2}} e^{-\frac{a}{2}x_2\partial_{x_1}} = e^{JS_{a,b}x\cdot\nabla}, \tag{34}$$

*où*

$$L_{a,b} = \mu_{a,b}\begin{pmatrix} b & \frac{ab}{2} \\ \frac{ab}{2} & a \end{pmatrix} \qquad \textit{et} \qquad S_{a,b} = \mu_{a,b}\begin{pmatrix} b & 0 \\ 0 & a(1-\frac{ab}{4}) \end{pmatrix} \tag{35}$$

*avec $\mu_{a,b} = F(ab(1-ab/4))$, où $F : (-\infty, 1] \to \mathbb{R}$ est définie par*

$$F(x) = \begin{cases} \frac{\arcsin(\sqrt{x})}{\sqrt{x}} & \textit{si} \quad 0 < x \leqslant 1 \\ \frac{\textit{asinh}(\sqrt{-x})}{\sqrt{-x}} & \textit{si} \quad x < 0 \\ 1 & \textit{si} \quad x = 0. \end{cases}$$

**Remarque 4.** *Par la formule des caractéristiques, on prouve ce résultat non pas directement sur les semi-groupes mais sur les équations différentielles ordinaires associées. Ces dernières étant linéaires, il est naturel qu'il n'y ait pas de terme de reste. On peut aussi voir ce résultat comme un cas particulier de splitting de semi-groupe associé à un opérateur quadratique*[41]*. En effet, dans [4], qui est une collaboration*[42] *avec Paul Alphonse, on démontre notamment qu'un produit de tels semi-groupes est lui aussi un semi-groupe associé à un opérateur quadratique.*

Les méthodes de Lie et Strang sont des cas particuliers des méthodes analysées dans ce théorème car elles consistent à prendre $a = b = \delta_t$. On voit alors qu'elles sont équivalentes à des équations de transport dont le mouvement n'est plus une rotation classique. Pour le splitting de Lie, les trajectoires sont des ellipses "obliques" définies par

$$x_1^2 + \delta_t x_1 x_2 + x_2^2 = \text{cste},$$

tandis que pour celui de Strang, il s'agit d'ellipses définies par

$$x_1^2 + (1 - (\delta_t/2)^2)x_2^2 = \text{cste}.$$

Ils sont tous deux associés à un mouvement de rotation sur ces ellipses à la vitesse angulaires $-\delta_t^{-1}\arcsin(\delta_t\sqrt{1-(\delta_t/2)^2}) = -1 + \mathcal{O}(\delta_t^2)$. Les splittings de Lie et Strang génèrent donc deux types d'erreurs, la première concernant la forme des trajectoires et la seconde la vitesse de rotation.

On peut alors chercher à obtenir de meilleures méthodes de splitting en choisissant d'autres valeurs que $a = b = \delta_t$ pour celles étudiées dans le Théorème 12. On peut même chercher à avoir des méthodes exactes, c'est-à-dire trouver $a, b$ tels que $L_{a,b} = \delta_t I_2$ ou $S_{a,b} = \delta_t I_2$.

On voit tout d'abord qu'à cause des termes non diagonaux de $L$, il n'est pas possible d'adapter le splitting de Lie pour obtenir une méthode exacte. Cependant, cela est tout à fait possible pour le splitting de Strang pour lequel il suffit de prendre

$$a = 2\tan(\delta_t/2) \quad \text{et} \quad b = \sin(\delta_t)$$

41. i.e. un opérateur pseudo-différentiel dont le symbole est une forme quadratique.
42. non incluse dans de manuscrit.

pour avoir $S_{a,b} = \delta_t I_2$. Si on revient au niveau des caractéristiques ce splitting revient à décomposer une rotation en un produit de trois *transvections* selon la formule

$$\begin{pmatrix} 1 & -\tan(t/2) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \sin t & 1 \end{pmatrix} \begin{pmatrix} 1 & -\tan(t/2) \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = e^{tJ}. \tag{36}$$

Il s'agit d'un résultat classique en traitement de l'image (voir par exemple [77],[97]). Il a été utilisé dans de nombreux logiciels pour faire tourner les images. Cependant, du fait de l'augmentation de la puissance de calcul des ordinateurs, il semble délaissé au profit de méthodes utilisant de l'interpolation $2d$ car elles produisent[43] de meilleurs résultats sur les images très contrastées.

Le calcul scientifique n'est pas soumis au même contraintes car le coût de calcul reste souvent un facteur limitant et le nombre de rotations à effectuer peut être très important. De plus, pour de nombreuses applications on cherche à faire tourner des fonctions très régulières (à l'opposé des images très contrastées). Auquel cas, on parvient à démontrer que ces méthodes de splitting donnent de très bon résultats sur des temps très longs.

Afin de pouvoir préciser cette affirmation, on introduit des paramètres de discrétisations, des interpolations pseudo-spectrales et une échelle d'espaces mesurant la localisation et la régularité des fonctions.

On discrétise un carré de côté $R$, centré en $0$ avec une grille régulière à $N$ points par direction. On note $h = R/N$ son pas et on la note $\mathbb{G}^2$ où

$$\mathbb{G} = h \left[\!\!\left[ -\left\lfloor \frac{N-1}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor \right]\!\!\right].$$

Les normes de Lebesgue discrètes associées à cette grille sont définies pour $\boldsymbol{u} \in \mathbb{C}^{\mathbb{G}^2}$ par

$$\| \boldsymbol{u} \|^2_{L^2(\mathbb{G}^2)} = h^2 \sum_{g \in \mathbb{G}^2} | \boldsymbol{u}_g |^2 \qquad \text{et} \qquad \| \boldsymbol{u} \|_{L^\infty(\mathbb{G}^2)} = \max_{g \in \mathbb{G}^2} | \boldsymbol{u}_g |.$$

Les transformées de Fourier discrètes partielles sont définies par

$$\mathcal{F}_1 : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\widehat{\mathbb{G}} \times \mathbb{G}} \\ \boldsymbol{u} & \mapsto & h \sum_{g_1 \in \mathbb{G}} \boldsymbol{u}_{g_1, g_2} \, e^{-i g_1 \xi_1} \end{array} \right. \quad \text{et } \mathcal{F}_2 : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G} \times \widehat{\mathbb{G}}} \\ \boldsymbol{u} & \mapsto & h \sum_{g_2 \in \mathbb{G}} \boldsymbol{u}_{g_1, g_2} \, e^{-i g_2 \xi_2} \end{array} \right. ,$$

où $\widehat{\mathbb{G}} = \frac{2\pi}{R} \left[\!\!\left[ -\left\lfloor \frac{N-1}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor \right]\!\!\right]$ est le dual de $\mathbb{G}$.
On définit les cisaillements pseudo-spectraux de paramètres $\alpha \in \mathbb{R}$ par

$$\mathcal{S}_1^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}^2} \\ \boldsymbol{u} & \mapsto & \mathcal{F}_1^{-1} \left[ e^{i\alpha \xi_1 g_2} \mathcal{F}_1 \, \boldsymbol{u} \right] \end{array} \right. \quad \text{et } \mathcal{S}_2^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}^2} \\ \boldsymbol{u} & \mapsto & \mathcal{F}_2^{-1} \left[ e^{i\alpha \xi_2 g_1} \mathcal{F}_2 \, \boldsymbol{u} \right] \end{array} \right. .$$

La méthode pseudo-spectrale basée sur (36) permettant de calculer une rotation d'angle $-t$ est donc définie par

$$\mathcal{M}_{\delta_t} = \mathcal{S}_1^{-\tan(\delta_t/2)} \mathcal{S}_2^{\sin(\delta_t)} \mathcal{S}_1^{-\tan(\delta_t/2)}.$$

43. d'après [77].

Enfin, les espaces permettant de mesurer simplement la régularité et la localisation des fonctions sont notés $(X^s)_{s\geqslant 0}$ et sont définis par

$$X^s = \left\{ u \in L^2(\mathbb{R}^2),\ \|u\|_{X^s}^2 := \int_{\mathbb{R}^2} |x|^{2s}|u(x)|^2 \mathrm{d}x + \int_{\mathbb{R}^2} |\xi|^{2s}|\mathscr{F}u(\xi)|^2 \mathrm{d}\xi < \infty \right\}$$

où $\mathscr{F}u$ est la transformée de Fourier de $u$.

Le théorème suivant estime l'erreur de convergence associée à $\mathcal{M}_{\delta_t}$.

**Théorème 13.** *Pour tout $s, \nu > 0$ il existe $c > 0$ tel que pour tout $N \in \mathbb{N}^*$, $R > 0$, $u \in \mathscr{S}(\mathbb{R}^2)$, $n \in \mathbb{N}$ et $\delta_t \in \mathbb{R}$ satisfaisant $|\delta_t| < \pi - \nu$, en notant $t_n = n\delta_t$, on ait*

$$\|(\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}}, \tag{37}$$

*où $\boldsymbol{u} = u_{|\mathbb{G}^2}$.*

**Remarque 5.** *Ce résultat appelle quelques remarques :*

- *Cette méthode ne requiert pas de CFL, ce qui est naturel car elle est basée sur une approche semi-lagrangienne.*
- *L'erreur de convergence croit linéairement par rapport à $t_n$ et non pas exponentiellement. Cela s'explique par le fait que $\mathcal{M}_{\delta_t}$ est une isométrie pour la norme $\|\bullet\|_{L^2(\mathbb{G}^2)}$. Il s'agit d'ailleurs d'une qualité remarquable de cette méthode car $\exp(\delta_t Jx \cdot \nabla)$ préserve lui aussi les normes de Lebesgue.*
- *Des expériences numériques présentes dans le sixième chapitre de cette thèse semblent montrer que le facteur associé à l'erreur de discrétisation en espace est naturel : si $R$ est fixé suffisamment grand et que l'on considère des valeurs de $h$ de plus en plus petites, l'erreur commence par décroître très fortement puis se met à augmenter très doucement (comme $1/\sqrt{h}$).*
- *Puisque pour tout $\boldsymbol{u} \in L^2(\mathbb{G}^2)$ on a $\|\boldsymbol{u}\|_{L^\infty(\mathbb{G}^2)} \leqslant h^{-1}\|\boldsymbol{u}\|_{L^\infty(\mathbb{G}^2)}$,* (37) *donne aussi une estimation de l'erreur de convergence en norme $L^\infty$ discrète pendant des temps très longs :*

$$\|(\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx.\nabla} u\right)_{|\mathbb{G}^2}\|_{L^\infty(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{h^{3/2}} \|u\|_{X^{s+6}}.$$

On dispose donc d'une méthode pseudo-spectrale plus efficace que celles habituellement obtenues pour calculer des rotations (en utilisant par exemple un splitting de Lie ou de Strang). Or il s'agit d'une étape cruciale dans la résolution de certaines équations aux dérivées partielles non linéaires. On cherche donc à mettre à profit ce gain pour résoudre plus efficacement ces équations. On s'intéresse principalement à deux exemples : les équations de Vlasov-Maxwell et Vlasov-HMF.

En imposant certaines symétries [44] aux conditions initiales, les équations de Vlasov-Maxwell peuvent être réduites à une équation d'évolution donnée par

$$\partial_t \begin{pmatrix} f \\ E \\ B \end{pmatrix} = \begin{pmatrix} BJv \cdot \nabla_v f \\ \begin{pmatrix} 0 \\ -\partial_{x_1} B \end{pmatrix} \\ 0 \end{pmatrix} - \begin{pmatrix} E \cdot \nabla_v f \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \partial_{x_1} E_2 \end{pmatrix} - \begin{pmatrix} v_1 \partial_{x_1} f \\ \int_{\mathbb{R}^2} v f \mathrm{d}v - \overline{\mathcal{J}} \\ 0 \end{pmatrix} \tag{VMR}$$

44. précisée dans le sixième chapitre de cette thèse.

où $\overline{\mathcal{J}}(t) = 1/|L| \int_L \int_{\mathbb{R}^2} v f(t, x_1, v)\, \mathrm{d}x_1 \mathrm{d}v$ avec $L$ un tore de longueur $|L|$. Les inconnues sont la fonction de distribution $f : \mathbb{R}_+ \times L \times \mathbb{R}^2 \to \mathbb{R}$, le champ électrique $E : \mathbb{R}_+ \times L \to \mathbb{R}^2$ et le champ magnétique $B : \mathbb{R}_+ \times L \to \mathbb{R}$.

Dans l'équation (VMR), on a directement écrit le champ de vecteurs comme somme de trois champs de vecteurs plus élémentaires. Il s'agit d'une décomposition introduite en 2015 par Crouseilles, Einkemmer et Faou dans [55]. Elle permet de résoudre numériquement (VMR) par des méthodes de splitting car les flots de chacun de ces champs de vecteurs peuvent être calculer exactement. Cependant, le calcul du premier flot requiert la résolution numérique de (ROT) (car $B$ y est constant). Pour éviter d'avoir à effectuer une coûteuse interpolation en dimension $2$, il était donc approché par un splitting de Strang, ce qui induit un terme d'erreur d'ordre $2$ en temps.

La nouvelle méthode de splitting, permet de résoudre numériquement ce premier flot aussi rapidement qu'avec un splitting de Strang, tout en étant exact en temps. Grâce à cette méthode, on réalise des expériences numériques montrant que l'on parvient à résoudre plus précisément (VMR). De plus, elle rend possible une résolution de (VMR) par des méthodes de splitting d'ordre élevé requérant[45] une résolution exacte du flot.

L'équation de Vlasov-HMF est une équation de transport non linéaire s'écrivant

$$\partial_t f = -\{f, H[f]\} \tag{VHMF}$$

où $f : \mathbb{R} \times \mathbb{T} \times \mathbb{R} \to \mathbb{R}$, $\{f, g\} \equiv \partial_x f \partial_v g - \partial_v f \partial_x g$ est le crochet de Poisson, $\mathbb{T}$ est le tore de longueur $2\pi$ et le hamiltonien $H$ est défini par

$$H[f](t, x, v) = \frac{v^2}{2} - \Phi[f](t, x),$$

avec

$$\Phi[f](t, x) = \int_{\mathbb{T}\times\mathbb{R}} \cos(x - y) f(t, y, u)\, \mathrm{d}u\, \mathrm{d}y.$$

L'équation (VHMF) ressemble à un système hamiltonien dans la mesure où $\{f, \bullet\}$ est antisymétrique pour le produit scalaire $L^2$. Cependant, $\{f, \bullet\}$ admet un noyau et dépend elle même de l'inconnue $f$. On parle donc de *système de Poisson*. Puisque le hamiltonien $H$ est *séparable*[46], les méthodes de splitting sont particulièrement adaptées à sa résolution car elles sont explicites et préservent sa structure de système de Poisson (voir [78]). Plus précisément, en notant $\varphi_t^H$, le flot au temps $t$ d'une équation semblable à (VHMF), un splitting de Strang pour (VHMF) reviendrait à faire l'approximation

$$\varphi_{\delta_t}^H \simeq \varphi_{\delta_t/2}^{v^2/2} \circ \varphi_{\delta_t}^{-\Phi} \circ \varphi_{\delta_t/2}^{v^2/2}.$$

On s'intéresse à la dynamique en temps longs de (VHMF) autour d'états d'équilibres inhomogènes de la forme

$$f^{eq}(x, v) = \alpha e^{-\beta\left(\frac{v^2}{2} - M_0 \cos x\right)} \text{ avec } M_0 = \int_{\mathbb{T}\times\mathbb{R}} \cos(y) f^{eq}(y, u) \mathrm{d}y \mathrm{d}u.$$

45. comme on le verra dans le sixième chapitre, actuellement, les méthodes de splitting les plus efficaces sont basées sur des méthodes de composition pour lesquelles il n'est pas nécessaire que chaque étape soit résolue exactement. Cependant, il s'agit encore d'un sujet de recherche actif.

46. i.e. il s'écrit comme la somme d'une fonction de $v$ et d'une fonction de $x$

Dans [82], il est démontré que cette dernière est essentiellement gouvernée par la partie transport du linéarisé de (VHMF), c'est-à-dire $\{H[f^{eq}], \bullet\} = \{v^2/2 - M_0 \cos x, \bullet\}$. De plus, à cause d'effets de dispersion, les termes les plus significatifs sont localisés au voisinage de $x = 0$. Pour ces derniers, le terme de transport dominant (à une affinité près) est une rotation : $\{v^2/2 + M_0 x^2/2, \bullet\}$. On propose donc une méthode de splitting permettant d'approcher exactement ce terme

$$\varphi^H_{\delta_t} \simeq \varphi^{v^2/2}_{t_c \tan(\frac{\delta_t}{2t_c})} \circ \varphi^{-\Phi}_{t_c \sin(\frac{\delta_t}{t_c})} \circ \varphi^{v^2/2}_{t_c \tan(\frac{\delta_t}{2t_c})},$$

où $t_c = 1/\sqrt{M_0}$. Le sixième chapitre de cette thèse présente des expériences numériques montrant un net gain de précision dans la résolution de (VHMF) au voisinage de $f^{eq}$ grâce à cette nouvelle méthode.

### Articles, prépublications et proceedings

Les papiers suivants ont été réalisés durant cette thèse :

A) P. Alphonse and J. Bernier, *Smoothing Properties of fractional Ornstein-Uhlenbeck semi-groups and null-controllability*, arXiv: 1810.02629

B) P. Alphonse and J. Bernier, *Polar decomposition of semigroups generated by quadratic operators and regularizing effects*, to appear soon

C) Y. Barsamian, J. Bernier, S. Hirstoaga, et M. Mehrenberger, *Verification of 2D$\times$ 2D and two-species Vlasov-Poisson solvers*, ESAIM : ProcS, **63**(2018), pp. 78-108

D) J. Bernier, *Bounds on the growth of high discrete Sobolev norms for the cubic discrete nonlinear Schrödinger equations on $h\mathbb{Z}$*, Discrete and Continuous Dynamical Systems - Series A, 2019, **39** (6), pp. 3179-3195

E) J. Bernier, *Optimality and resonances in a class of compact finite difference schemes of high order*, *Calcolo* (2019) 56 : 12

F) J. Bernier, N. Crouseilles and F. Casas, *Splitting methods for rotations : application to Vlasov equations*, to appear soon

G) J. Bernier and E. Faou, *Existence and stability of traveling waves for discrete nonlinear Schroedinger equations over long times*, SIAM J. Math. Anal., 51(3), 1607-1656.

H) J. Bernier, E. Faou and B. Grébert, *Rational normal forms and stability of small solutions to nonlinear Schrödinger equations*, arXiv: 1812.11414

I) J. Bernier, E. Faou and B. Grébert, *Long time behavior of the solutions of NLW on the d-dimensional torus*, arXiv: 1906.05107

J) J. Bernier and M. Mehrenberger, *Long-time behavior of second order linearized Vlasov-Poisson equations near a homogeneous equilibrium*, arXiv : 1903.08374

Chaque chapitre de thèse correspond à un article. Les papiers A,B, C et I ne font pas partis de ce manuscrit. Le proceeding C est présenté dans l'introduction du cinquième chapitre. Les papiers A et B sont principalement consacrés à l'étude des effets régularisants de certaines familles de semi-groupes faisant interagir des phénomènes de transport et de diffusion. L'article I présente une nouvelle construction de formes normales permettant de contrôler la croissance

des normes de Sobolev d'indices élevés sur des temps très longs pour les petites solutions d'une large classe de systèmes hamiltoniens incluant l'équation des ondes non-linéaire sur le tore de dimension $d \geqslant 2$.

CHAPITRE 1

# OPTIMALITY AND RESONANCES IN A CLASS OF COMPACT FINITE DIFFERENCE SCHEMES OF HIGH ORDER

## 1.1 Introduction

Many decades ago, compact finite differences methods were widely studied. Nowadays, we still can find a huge literature about these methods that are very popular and very often used for the approximation of partial differential equations : KdV equation [85], hyperbolic equation [54], elliptic equation [41] or [110] for a more general overview. In particular, we can find a lot of examples of accurate schemes for elliptic problems and many classical mathematical arguments are proposed to prove their convergence (monotonicity, energy, green functions, ...). However, up to our knowledge, it seems that there is no general and algebraic study of compact finite difference schemes for elliptic problems, equivalent to what we can find, for example, for the Runge Kutta methods applied to Cauchy problems (general stability criteria, algebraic order conditions using Hopf algebras and trees as we can see in [79] or [78]).

This paper is a first step in the direction of such a general study. Consequently, it is devoted to compact finite difference schemes for one of the most elementary elliptic partial differential equation, the homogeneous Dirichlet problem in dimension $1$. We give a detailed derivation of a large and natural class of such schemes and we establish many of their qualitative properties. This derivation can be extended to multi-dimensional cartesian domains even if a study of qualitative properties of these schemes would involve more sophisticated algebra and would necessitate more investigations.

In our context, a compact finite difference scheme is a linear system of the form

$$\mathbf{D}_N \mathbf{u}^N = \mathbf{S}_N \mathbf{f}^{N,ex},$$

where $\mathbf{f}^{N,ex}$ is a discretization of the source term on a grid of stepsize $h = (N+1)^{-1}$, $\mathbf{D}_N$ and $\mathbf{S}_N$ are matrices and $\mathbf{u}^N$ is the approximation of the solution of the Dirichlet problem.

To study their convergence (i.e. the approximation of the exact solution by $\mathbf{u}^N$), we first introduce some specific and rigorous notions of consistency and stability taking into account the boundary conditions. Then, we describe precisely the class of schemes that we consider, namely when the matrix $\mathbf{D}_N$ is a polynomial in the usual discrete second derivative matrix $\mathbf{A}_N$

defined by

$$\mathbf{A}_N = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathscr{L}(\mathbb{C}^N). \tag{1.1}$$

This choice is made for two reasons. First it allows for a relatively simple stability analysis, and second it is in fact not so restrictive. Indeed, if we take a symmetric finite difference formula $d = (d_j)_{j\in\mathbb{Z}}$ that approximates the second derivative, i.e. for all smooth function $u$,

$$\sum_{j\in\mathbb{Z}} d_j u(hj) \simeq -h^2 u''(0),$$

then we get, from a convolution formula and for some specific and natural choice of the coefficients near the boundary, a matrix $\mathbf{D}_N$ that is a polynomial in $\mathbf{A}_N$.

In this paper, we give some general criterion of convergence for this family of schemes. Moreover, we address the following two questions :

- Are these schemes stable *in general* ?
- Amongst these schemes, what are the most *efficient* and are they stable ?

We will precise the two ambiguous terms *general* and *efficient* by introducing, on one hand, a Lebesgue measure on the set of schemes, and on the other hand, an optimization problem defining efficiency. The first main result of this paper will be to prove that almost all schemes are convergent at the same rate as its consistency order, up to some logarithmic correction. It is based on a careful analysis of small denominators appearing in the stability conditions, linked with diophantine approximation theory. The second main result of this paper is the design and construction of the most efficient schemes in the class considered, which turn to be stable, this latter property requiring the use of Padé approximant theory to be proved.

## 1.2 Formalism and main results

The goal of this section is to present the two main results of this paper. To this aim, we first define rigorously compact finite difference schemes for the homogeneous Dirichlet problem in dimension $1$. Then, we recall the usual concept of convergence, consistency and stability for these schemes. And finally, we define the particular set of schemes that we consider.

### 1.2.1 Context

We consider the homogeneous Dirichlet problem in dimension $1$, namely :

For a given $f : \mathbb{R} \to \mathbb{C}$, find $u : [0,1] \to \mathbb{C}$ such that

$$\begin{cases} -u''(x) = f(x), \ \forall x \in ]0,1[, \\ u(0) = u(1) = 0. \end{cases} \tag{1.2}$$

To design finite difference schemes, we will consider regular grids on $\mathbb{R}$. More precisely, we choose $N \in \mathbb{N}^*$ to be the number of grid points into $]0,1[$ (the number of unknowns) and we define $h$ as the stepsize of the grid. As a consequence, $h$ and $N$ are linked by the relation

$$h = \frac{1}{N+1}.$$

Let $x_j^N = jh$, $j \in \mathbb{Z}$ denote the grid points, see Figure 1.1.

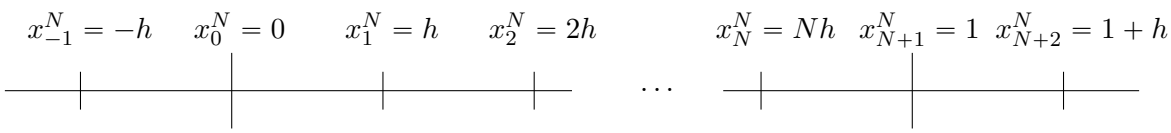


FIGURE 1.1 – Regular grid with $N$ points into $]0,1[$.

In this context, a finite difference scheme is a couple of sequences of matrices $((\mathbf{D}_N)_{N\in\mathbb{N}^*}, (\mathbf{S}_N)_{N\in\mathbb{N}^*})$ such that $\mathbf{D}_N \in \mathscr{L}(\mathbb{C}^N)$ is a square matrix of size $N$ and $\mathbf{S}_N \in \mathscr{L}(\mathbb{C}^{\mathbb{Z}};\mathbb{C}^N)$ is a rectangle matrix with $N$ rows and a finite number of columns.

If $\mathbf{D}_N$ is invertible for all $N$, such a scheme leads to an approximation of the solution $u$ of the Dirichlet problem (1.2). More precisely, we define $\mathbf{f}^{N,ex}$ and $\mathbf{u}^{N,ex}$ as the vectors of the values of $f$ and $u$ on the grid :

$$\mathbf{f}^{N,ex} = (f(x_j^N))_{j\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}} \text{ and } \mathbf{u}^{N,ex} = (u(x_j^N))_{j\in[\![1,N]\!]} \in \mathbb{C}^N. \tag{1.3}$$

Then, the scheme gives an approximation $\mathbf{u}^N$ of $\mathbf{u}^{N,ex}$ through the solution of the linear system

$$\mathbf{D}_N\mathbf{u}^N = h^2\mathbf{S}_N\mathbf{f}^{N,ex}. \tag{1.4}$$

It may seem unusual to use the values of $f$ outside $[0,1]$ but it is just a way to have more symmetric formulas. In practice, as we will explain in the subsection 1.2.3, we only use a finite number of values of $f$ outside $[0,1]$ independent of $N$ and at a distance of order $h$ of $[0,1]$. Consequently, as we will assume that $f$ is a regular function, these values could be extrapolated from those of $f$ in $[0,1]$ with some Newton series.

To estimate the accuracy of a scheme, we define a notion of rate of convergence and of order of convergence. Let $(\epsilon_N)_{N\in\mathbb{N}^*}$ be a sequence of positive numbers that tends to $0$, as $N$ goes to infinity. Then, a scheme $((\mathbf{D}_N)_{N\in\mathbb{N}^*}, (\mathbf{S}_N)_{N\in\mathbb{N}^*})$ is said to be convergent at the rate $(\epsilon_N)_{N\in\mathbb{N}^*}$, if $\mathbf{D}_N$ is invertible for all $N \in \mathbb{N}^*$ and if for all $f \in C^\infty(\mathbb{R})$ there exists a constant $c > 0$ such that for all $N \in \mathbb{N}^*$,

$$\sup_{j=1,\dots,N} |\mathbf{u}_j^N - \mathbf{u}_j^{N,ex}| =: \|\mathbf{u}^{N,ex} - \mathbf{u}^N\|_\infty \leqslant c\epsilon_N. \tag{1.5}$$

Furthermore, if $n$ is a positive integer and $\epsilon_N = h^n$, then a scheme that is convergent at the rate $(\epsilon_N)_{N\in\mathbb{N}^*}$ is said to be convergent of order $n$.

### 1.2.2 Notions of consistency and stability

In order to establish a convergence result of the form (1.5), we use introduce the notions of consistency and stability. Then, we give a Lax theorem to deduce the convergence from the consistency and the stability.

A scheme $((\mathbf{D}_N),(\mathbf{S}_N))$ is said to be consistent of order $n \in \mathbb{N}$, if for all $f \in C^\infty(\mathbb{R})$ there exists a constant $c > 0$ such that for all $N \in \mathbb{N}^*$, the vectors $\mathbf{u}^{N,ex}$ and $\mathbf{f}^{N,ex}$, defined by (1.3), verify

$$\|\mathbf{D}_N\mathbf{u}^{N,ex} - h^2\mathbf{S}_N\mathbf{f}^{N,ex}\|_\infty \leqslant ch^{n+2}. \tag{1.6}$$

In the context of the Dirichlet problem, it is usual to relax this notion of consistency near the boundary (see [41]). A scheme $((\mathbf{D}_N),(\mathbf{S}_N))$ is said to be consistent of order $n \in \mathbb{N}$ in the center and of order $n-2$ at a distance $l \in \mathbb{N}$ of the boundary, if for all $f \in C^\infty(\mathbb{R})$ there exists a constant $c > 0$ such that for all $N \in \mathbb{N}^*$, the vectors $\mathbf{u}^{N,ex}$ and $\mathbf{f}^{N,ex}$, defined by (1.3), verify for all $j = 1,\dots,N$,

$$\left|\left(\mathbf{D}_N\mathbf{u}^{N,ex}\right)_j - h^2\left(\mathbf{S}_N\mathbf{f}^{N,ex}\right)_j\right| \leqslant \begin{cases} ch^{n+2} & \text{if } l < j < N+1-l, \\ ch^n & \text{else.} \end{cases} \tag{1.7}$$

In this article, it is useful to distinguish some notions of stability. A scheme $((\mathbf{D}_N),(\mathbf{S}_N))$ or a sequence $(\mathbf{D}_N)_{N\in\mathbb{N}^*} \in \prod_{N\in\mathbb{N}^*} \mathscr{L}(\mathbb{C}^N)$ of matrices is said to be

- stable, if there exists a positive constant $c > 0$ such that for all $N \in \mathbb{N}^*$, we have

$$\forall \mathbf{v} \in \mathbb{C}^N,\ c\|\mathbf{v}\|_\infty \leqslant h^{-2}\|\mathbf{D}_N\mathbf{v}\|_\infty. \tag{1.8}$$

- strongly stable, if for all $l \in \mathbb{N}$, there exists a positive constant $c > 0$ such that for all $N \in \mathbb{N}^*$,

$$\forall \mathbf{v} \in \mathbb{C}^N,\ c\|\mathbf{v}\|_\infty \leqslant \sup_{j=1,\dots,N} \begin{cases} h^{-2}\left(\mathbf{D}_N\mathbf{v}\right)_j & \text{if } l < j < N+1-l, \\ \left(\mathbf{D}_N\mathbf{v}\right)_j & \text{else.} \end{cases}. \tag{1.9}$$

- stable relatively to a sequence $(\eta_N)_{N\in\mathbb{N}^*}$ of positive numbers, if there exists a positive constant $c > 0$ such that for all $N \in \mathbb{N}^*$, we have

$$\forall \mathbf{v} \in \mathbb{C}^N,\ c\|\mathbf{v}\|_\infty \leqslant \eta_N\|\mathbf{D}_N\mathbf{v}\|_\infty. \tag{1.10}$$

We remark that, if a scheme is strongly stable, then it is stable, and, if it is stable, then it is stable relatively to $\eta_N = (N+1)^2 = h^{-2}$.

To establish convergence from consistency and stability, we give a theorem due to Lax .

**Theorem 1.2.1.** *(Lax)*

- *A scheme that is strongly stable (see* (1.9)*) and consistent of order $n \geqslant 1$ in the center and of order $n-2$ at a distance $l \in \mathbb{N}$ of the boundary (see* (1.7)*) is convergent of order $n$.*
- *Let $(\eta_N)_{N\in\mathbb{N}^*}$ be a sequence of positive number and $n \in \mathbb{N}^*$ such that the sequence $(\eta_N h^{n+2})_{N\in\mathbb{N}^*}$ tends to zero as $N$ goes to infinity. Then a scheme that is stable relatively to the sequence $(\eta_N)_{N\in\mathbb{N}^*}$* (1.10) *and consistent of order $n$* (1.6) *is convergent at the rate $\epsilon_N = \eta_N h^{n+2}$* (1.5)*.*

*Proof.* The invertibility of $\mathbf{D}_N$ follows from the stability estimate. To prove the convergence estimate, it is enough to apply the stability estimate to the error of consistency

$$\mathbf{D}_N\mathbf{v} = \mathbf{D}_N\left(\mathbf{u}^{N,ex} - \mathbf{u}^N\right) = \mathbf{D}_N\mathbf{u}^{N,ex} - h^2\mathbf{S}_N\mathbf{f}^{N,ex}.$$

□

### 1.2.3 Expression of the schemes

Usually, to design a finite difference scheme $((\mathbf{D}_N),(\mathbf{S}_N))$, we need to introduce the notion of finite difference formulas. A finite difference formula is a sequence of complex numbers indexed by $\mathbb{Z}$ with finite support. We denote by $\mathbb{C}^{(\mathbb{Z})}$ their space. We say that a couple of finite difference formulas $(d,s) \in \left(\mathbb{C}^{(\mathbb{Z})}\right)^2$ is consistent of order $n$, if

$$\forall u \in C^\infty(\mathbb{R}),\ \sum_{j\in\mathbb{Z}} d_j u(x_j^N) + h^2 s_j u''(x_j^N) = \mathcal{O}(h^{n+2}). \tag{1.11}$$

For example, if we introduce the usual formula for the second derivative

$$a = 2\mathbb{1}_{\{0\}} - \mathbb{1}_{\{-1,1\}}, \tag{1.12}$$

then a Taylor expansion shows that $(a, \mathbb{1}_{\{0\}})$ is consistent of order $2$.

To preserve the classical properties of the second derivative, it is natural to assume that the sequences $d$ and $s$ are symmetric,

$$d, s \in \mathscr{S}_{\mathbb{C}} := \{b \in \mathbb{C}^{(\mathbb{Z})} \mid \forall j \in \mathbb{Z},\ b_j = b_{-j}\}, \tag{1.13}$$

and it is then natural to restrict the analysis to the case where $n$ is an even number. Sometimes, it is interesting and more effective –for instance using formal calculus– to consider finite difference formulas with coefficients in a smaller ring than $\mathbb{C}$. For example, the usual high order formulas have rational or integer coefficients. That is why, we introduce, the more general notation

$$\mathscr{S}_R := \{b \in R^{(\mathbb{Z})} \mid \forall j \in \mathbb{Z},\ b_j = b_{-j}\} \text{ with } R \text{ a ring such that } \mathbb{Z} \subset R \subset \mathbb{C}. \tag{1.14}$$

It is useful to associate to each finite difference formula the highest index associated to a non zero value. It is a measure of the stencil of a formula. More formally, if $b \in \mathscr{S}_{\mathbb{C}}$ is a symmetric formula then $\tau(b)$ is defined by

$$\tau(b) = \max\{j \in \mathbb{Z} \mid b_j \neq 0\}. \tag{1.15}$$

The following proposition explains that there is a simple way to get finite difference formulas $d, s \in \mathscr{S}_{\mathbb{C}}$ consistent of order $n$.

**Proposition 1.2.1.** *Let $n \in 2\mathbb{N}$ be an even integer and $d \in \mathscr{S}_{\mathbb{C}}$ be a symmetric formula with zero mean*

$$\sum_{j\in\mathbb{Z}} d_j = 0. \tag{1.16}$$

*Then there exists a unique* $s \in \mathscr{S}_{\mathbb{C}}$ *such that* $(d, s)$ *is consistent of order* $n$ (1.11) *and* $\tau(s) \leqslant \frac{n}{2} - 1$. *Furthermore,* $(\frac{s_0}{2}, s_1, \ldots, s_{\frac{n}{2}-1})$ *is the solution of the Vandermonde linear system*

$$(\frac{s_0}{2}, s_1, \ldots, s_{\frac{n}{2}-1})((i-1)^{2j-2})_{1 \leqslant i,j \leqslant \frac{n}{2}} = -\sum_{j>0} d_j \left( \frac{j^2}{2}, \ldots, \frac{j^n}{n(n-1)} \right). \tag{1.17}$$

*Proof.* If $1 \leqslant j \leqslant \frac{n}{2}$ is an integer and if we choose $u = x^{2j}$ in (1.11) then it comes

$$\sum_{i \in \mathbb{Z}} d_i (hi)^{2j} + s_j 2j(2j-1) h^{2j} i^{2(j-1)} = \mathcal{O}(h^{n+2}).$$

As $j \leqslant \frac{n}{2}$ and $h$ tends to $0$, we deduce that the remainder vanishes and we recognize the Vandermonde equation (1.17).
Conversely, since $d$ and $s$ are symmetric, if $u$ is an odd function then

$$\sum_{i \in \mathbb{Z}} d_i u(x_i^N) + h^2 s_i u''(x_i^N) = 0.$$

Furthermore, since $s$ is the solution of (1.17), this relation also holds if $u = x^{2j}$ with $1 \leqslant j \leqslant \frac{n}{2}$. As a consequence, it is enough to apply a Taylor Young expansion to prove (1.11). □

Then, to design the matrix $\mathbf{D}_N$ and $\mathbf{S}_N$ from the formulas $d$ and $s$, a natural choice would be the following :

$$(\mathbf{D}_N \mathbf{u})_i = \sum_{j \in \mathbb{Z}} d_{i-j} \mathbf{u}_j \text{ and } (\mathbf{S}_N \mathbf{f})_i = \sum_{j \in \mathbb{Z}} s_{i-j} \mathbf{f}_j. \tag{1.18}$$

However, $\mathbf{D}_N$ has to be square matrix. And, with such a definition, we use the values of $\mathbf{u}$ at the indexes $1-\tau(d), \ldots, 0$ and $N+1, \ldots, N+\tau(d)$. The usual way to solve this problem is to modify the formulas near the boundary (for $i \leqslant \tau(d)$ or $i \geqslant N+1-\tau(d)$). That is why, we introduce, for $i = 1, \ldots, \tau(d)$, some formulas $d^i \in \mathbb{C}^{(\mathbb{Z})}$ and $s^i \in \mathbb{C}^{(\mathbb{Z})}$ that satisfy a relation of consistency at a distance $i$ of the boundary

$$\forall u \in C^\infty(\mathbb{R}),\ u(0) = 0 \ \Rightarrow \ \sum_{j > -i} d_j^i u(x_{j+i}^N) + h^2 \sum_{j \in \mathbb{Z}} s_j^i u''(x_{j+i}^N) = \mathcal{O}(h^{\mu+2}), \tag{1.19}$$

here $\mu \in \{n-2, n\}$ is the desired order of consistency. We use symmetrically in $1$ these formulas to define, if $N$ is large enough, the following scheme $((\mathbf{D}_N), (\mathbf{S}_N))$, for $\mathbf{u} \in \mathbb{C}^N$ and $\mathbf{f} \in \mathbb{C}^{\mathbb{Z}}$, by

$$(\mathbf{D}_N \mathbf{u})_i := \begin{cases} \sum_{j>0} d_{j-i}^i \mathbf{u}_j & \text{if } 1 \leqslant i \leqslant \tau(d), \\ \sum_{j \in \mathbb{Z}} d_{j-i} \mathbf{u}_j & \text{if } \tau(d) < i < N+1-\tau(d), \\ \sum_{j<N+1} d_{-j+i}^{N+1-i} \mathbf{u}_j & \text{if } N+1-\tau(d) \leqslant i \leqslant N+1. \end{cases} \tag{1.20}$$

and

$$(\mathbf{S}_N \mathbf{f})_i := \begin{cases} \sum_{j \in \mathbb{Z}} s_{j-i}^i \mathbf{f}_j & \text{if } 1 \leqslant i \leqslant \tau(d), \\ \sum_{j \in \mathbb{Z}} s_{j-i} \mathbf{f}_j & \text{if } \tau(d) < i < N+1-\tau(d), \\ \sum_{j \in \mathbb{Z}} s_{-j+i}^{N+1-i} \mathbf{f}_j & \text{if } N+1-\tau(d) \leqslant i \leqslant N+1. \end{cases} \tag{1.21}$$

The following proposition enables to get the consistency of such a construction.

**Proposition 1.2.2.** *For $N$ large enough, let $((\mathbf{D}_N),(\mathbf{S}_N))$ be the scheme (defined by* (1.20) *and* (1.21)*), then*

- *if $\mu = n-2$, this scheme is consistent of order $n-2$ at a distance $\tau(d)$ of the boundary and of order $n$ in the center, see* (1.7)*.*
- *if $\mu = n$, this scheme is consistent of order $n$, see* (1.6)*.*

*Proof.* see Appendix 1.6.1. □

The main difficulty with such a construction is to get stability. There are at least two general ways for choosing the formulas near the boundary to ensure stability. A first principle is to rely on monotonicity arguments, as explained by Bramble and Hubbart [41] and Price [100]. The methods they consider to design the coefficients near the boundary are robust and lead, in general, to strong stability. However, the choice of formulas $d$ and $d^i$ is quite limited, as the conditions to ensure monotonicity are in general difficult to fulfil. Furthermore, it turns out that there exist very accurate high order schemes that do not satisfy any hypothesis of monotonicity.

A second natural way of obtaining the boundary coefficients is to start from *polynomial methods* that we consider below. For these methods, if we respect some algebraic structures, we can compute explicitly the eigenvalues and the eigenvectors of $\mathbf{D}_N$, and analyse directly the stability. This method is not very restrictive for the choice of the formulas $d$ and there is a natural choice for the formulas $d^i$ near the boundary.

The polynomial methods consists in studying schemes for which there exists a polynomial $P$ such that, for all $N \in \mathbb{N}$, $\mathbf{D}_N = P(\mathbf{A}_N)$ is a polynomial of $\mathbf{A}_N$ (the classical approximation of the second derivative, defined in (1.1)). The interest of this method is that the spectral decomposition of these matrices is well known. Indeed, we can verify by a straightforward calculation that

$$\mathbf{A}_N \mathbf{e}_k^N = 4\sin^2\left(\frac{\pi}{2}kh\right)\mathbf{e}_k^N, \text{ with } \mathbf{e}_k^N := (\sin(\pi khj))_{j=1,\dots,N}, \tag{1.22}$$

and deduce classically that

$$\mathbf{D}_N \mathbf{e}_k^N = P\left(4\sin^2\left(\frac{\pi}{2}kh\right)\right)\mathbf{e}_k^N. \tag{1.23}$$

Actually, it is not very restrictive to require for $\mathbf{D}_N$ to be a polynomial in $\mathbf{A}_N$. Indeed, for a given symmetric formulas $d$, there is a natural possible choice for the boundary formulas $d^i$, $i = 1,\dots,\tau(d)$ such that the matrix $\mathbf{D}_N$ defined by (1.20) is a polynomial in $\mathbf{A}_N$. This choice corresponds to extend all the vectors $\mathbf{u} \in \mathbb{C}^N$ in sequences defined on $\mathbb{Z}$ through the relations

$$\forall j \in \mathbb{Z},\ \mathbf{u}_j = -\mathbf{u}_{-j} \text{ and } \mathbf{u}_{N+1+j} = -\mathbf{u}_{N+1-j},$$

and use the natural convolution formula (1.18). In practice, when $N$ is large enough, this choice leads to

$$d_j^i = d_j - d_{2i+j}, \quad i = 1,\dots,\tau(d), \quad j \in \mathbb{Z} \tag{1.24}$$

In all this paper, we denote by $\mathbf{D}_N(d)$ the square matrix obtained from this construction (i.e. the matrix (1.20) and the boundary formulas (1.24)– see Definition 1.3.1 for a formal construction).

The following proposition shows that the previous construction is relevant : First, we prove that all the matrices $\mathbf{D}_N(d)$ are polynomials in $\mathbf{A}_N$, and second we can find formulas $s^i$, $i = 1,\dots,\tau(d)$ satisfying (1.19) for any given order of consistency $\mu$.

**Proposition 1.2.3.**

- *If $R$ is a ring such that $\mathbb{Z} \subset R \subset \mathbb{C}$ and if $d \in \mathscr{S}_R$ is a $R$ valued finite difference symmetric formula then there exists a polynomial $P \in R[X]$ such that*

$$\forall N \in \mathbb{N}^*,\ P(\mathbf{A}_N) = \mathbf{D}_N(d)$$

*and*

$$\deg P = \tau(d).$$

- *Let $n \in 2\mathbb{N}^*$ and $\mu = n$ or $\mu = n-2$. If there exists a finite difference formula $s \in \mathscr{S}_{\mathbb{C}}$ such that $(d, s)$ is consistent of order $\mu$ (see* (1.11)*) then for all $i = 1, \dots, \tau(d)$ there exists a unique symmetric formula $b^i \in \mathscr{S}_{\mathbb{C}}$ such that $\tau(b) \leqslant \frac{\mu}{2} - 1$ and*

$$s^i := s + (b^i_{i+j})_{j\in\mathbb{Z}} \text{ is consistent of order } \mu \text{ at a distance } i \text{ of the boundary, see (1.19).}$$

*Furthermore, $(\frac{b^i_0}{2}, b^i_1, \dots, b^i_{\frac{\mu}{2}-1})$ is the solution of the Vandermonde linear system*

$$(\frac{b^i_0}{2}, b^i_1, \dots, b^i_{\frac{\mu}{2}-1})((i-1)^{2j-2})_{1\leqslant i,j\leqslant \frac{\mu}{2}} = -\sum_{j>0} d_{i+j}\left(\frac{j^2}{2}, \dots, \frac{j^\mu}{\mu(\mu-1)}\right). \tag{1.25}$$

*Proof.* The first point will be proved in the next section as a direct consequence of Lemma 1.3.3 and Lemma 1.3.4. To prove the second point, let consider $u \in C^\infty(\mathbb{R})$ such that $u(0) = 0$. Then we have from (1.24), for $i = 1, \dots, \tau(d)$,

$$\begin{aligned}
\sum_{j>-i} d^i_j u(x^N_{j+i}) + h^2 \sum_{j\in\mathbb{Z}} s_j u''(x^N_{j+i}) &= \sum_{j>-i} (d_j - d_{j+2i})\, u(x^N_{j+i}) + h^2 \sum_{j\in\mathbb{Z}} s_j u''(x^N_{j+i}) \\
&= \sum_{j\in\mathbb{Z}} d_j u(x^N_{j+i}) + h^2 s_j u''(x^N_{j+i}) \\
&\quad - \sum_{j<-i} d_j u(x^N_{j+i}) - \sum_{j>-i} d_{j+2i} u(x^N_{j+i}) \\
&= -\sum_{j>0} d_{i+j}\left(u(x^N_{-j}) + u(x^N_j)\right) + \mathcal{O}(h^{\mu+2}) \\
&= -\sum_{j\in\mathbb{Z}} \widetilde{d}_j u(x^N_j) + \mathcal{O}(h^{\mu+2}),
\end{aligned}$$

with $\widetilde{d} \in \mathscr{S}_{\mathbb{C}}$ a symmetric finite difference formula with zero mean (1.16) defined by $\widetilde{d}_j = d_{i+j}$ if $j > 0$. Then applying Proposition 1.2.1 enables to conclude the proof. □

**Remark 1.2.2.** *The formula* (1.25) *implies in particular that $b^{\tau(d)} = 0$ because, for $i = 1, \dots, \tau(d)$, the right hand side term in* (1.25) *is zero by definition of $\tau(d)$.*

To conclude this part, explicit expressions of a class a high order schemes constructed using the previous principle are proposed. They will be used to give examples.

**Proposition 1.2.4.** *Let $(d, s) \in \mathscr{S}_{\mathbb{C}}$ be a couple of symmetric finite difference formulas that is consistent of order $n$ with $n \in 2\mathbb{N}^*$. Let $\mu \in \{n-2, n\}$ be an even integer. Define $l = \tau(d)-1$, $m = \tau(s)$ and for $i = 1, \dots, l$, $b^i$ as the solution of the system* (1.25)*. If we choose $s^i = s + (b^i_{i+j})_{j\in\mathbb{Z}}$*

*and* $d^i = d_j - d_{2i+j}$ *then the relations* (1.20) *and* (1.21) *define a scheme that is consistent of order* $n$ *in the center and of order* $\mu$ *at a distance* $l$ *of the boundary. More precisely, if* $N$ *is large enough, this scheme is given by the following band matrices :*

$$\mathbf{D}_N(d) = \begin{pmatrix} d_0 & \dots & d_{l+1} & & & \\ \vdots & \ddots & & \ddots & & \\ d_{l+1} & & \ddots & & \ddots & \\ & \ddots & & \ddots & & d_{l+1} \\ & & \ddots & & \ddots & \vdots \\ & & & d_{l+1} & \dots & d_0 \end{pmatrix} - \begin{pmatrix} d_2 & \dots & d_{l+1} & & & \\ \vdots & \iddots & & & & \\ d_{l+1} & & & & & \\ & & & & & d_{l+1} \\ & & & & \iddots & \vdots \\ & & & d_{l+1} & \dots & d_2 \end{pmatrix} \in \mathscr{L}(\mathbb{C}^N),$$

$$\mathbf{S}_N = \begin{pmatrix} s_m & \dots & s_0 & \dots & s_m & & & & \\ & \ddots & & \ddots & & \ddots & & & \\ & & \ddots & & \ddots & & \ddots & & \\ & & & \ddots & & \ddots & & \ddots & \\ & & & & \ddots & & \ddots & & \ddots \\ & & & & & s_m & \dots & s_0 & \dots & s_m \end{pmatrix} + \begin{pmatrix} \mathbf{B}_\mu^+ & & \\ & \mathbf{0}_{N-2l,N-\mu+2} & \\ & & \mathbf{B}_\mu^- \end{pmatrix}$$

*with* $\mathbf{S}_N \in \mathscr{L}(\mathbb{C}^{\mathbb{Z}}; \mathbb{C}^N)$, $\mathbf{0}_{N-2l,N-\mu+2}$ *the zero matrix of size* $(N-2l) \times (N-\mu+2)$,

$$\mathbf{B}_\mu^+ = \begin{pmatrix} b^1_{\frac{\mu}{2}-1} & \dots & \dots & b^1_0 & \dots & \dots & b^1_{\frac{\mu}{2}-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b^l_{\frac{\mu}{2}-1} & \dots & \dots & b^l_0 & \dots & \dots & b^l_{\frac{\mu}{2}-1} \end{pmatrix} \in \mathscr{L}(\mathbb{C}^{\mu-1}; \mathbb{C}^l)$$

*and*

$$\mathbf{B}_\mu^- = \begin{pmatrix} b^l_{\frac{\mu}{2}-1} & \dots & \dots & b^l_0 & \dots & \dots & b^l_{\frac{\mu}{2}-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b^1_{\frac{\mu}{2}-1} & \dots & \dots & b^1_0 & \dots & \dots & b^1_{\frac{\mu}{2}-1} \end{pmatrix} \in \mathscr{L}(\mathbb{C}^{\mu-1}; \mathbb{C}^l).$$

### 1.2.4 Main results

We will first define the notion of *efficiency* discussed in the introduction. If we design our schemes as in Proposition 1.2.4, and unless some more specific algebraic structure is given, the computation time for the approximation of one solution of the Dirichlet problem – that is the solution of the linear system (1.4)– grows *a priori* linearly with the size of the stencils $\tau(d)$ and $\tau(s)$. As a consequence, in general, the smaller $\tau(d)$ and $\tau(s)$ are, the larger the order of consistency of $(d, s)$ is, and hence the more *efficient* is our scheme. As a consequence, we will define schemes to be *the most efficient* for given $l, m \in \mathbb{N}$, those which are solutions to the following optimization problem :

$$\max_{\substack{(d,s)\in\mathscr{S}^2_{\mathbb{C}}\setminus\{(0,0)\} \\ \tau(d)\leqslant l+1,\ \tau(s)\leqslant m}} \operatorname{ord}(d, s), \tag{1.26}$$

where $\mathrm{ord}(d,s)$ is the exact order of consistency of $(d,s)$

$$\mathrm{ord}(d,s) = \sup\{n \in 2\mathbb{N} \mid (d,s) \text{ is consistent of order } n \text{ according to (1.11)}\}.$$

The following theorem proves that for any given stencil sizes $l$ and $m$ in $\mathbb{N}$, there exists a most efficient scheme solution of the previous optimization problem, and it is unique, up to a multiplication by a scalar.

**Theorem 1.2.3.** *For all $l, m \in \mathbb{N}$, there exists a couple of rational symmetric formulas $(d^{l,m}, s^{l,m}) \in \mathscr{S}^2_{\mathbb{Q}}$ such that*

$$\begin{cases} \tau(d^{l,m}) = l+1, \\ \tau(s^{l,m}) = m, \\ \displaystyle\sum_{j\in\mathbb{Z}} d^{l,m}_j j^2 = -2, \end{cases}$$

*that is solution of the problem of optimization*

$$\max_{\substack{(d,s)\in\mathscr{S}^2_{\mathbb{C}}\setminus\{(0,0)\} \\ \tau(d)\leqslant l+1,\ \tau(s)\leqslant m}} \mathrm{ord}(d,s) = \mathrm{ord}(d^{l,m}, s^{l,m}) = 2(l+m+1).$$

*Moreover if $(d,s) \in \mathscr{S}^2_{\mathbb{C}}$ is such that $\tau(d) \leqslant l+1$, $\tau(s) \leqslant m$ and $\mathrm{ord}(d,s) = 2(l+m+1)$ then there exists $\lambda \in \mathbb{C}$ such that $d = \lambda d^{l,m}$ and $s = \lambda s^{l,m}$.*

This theorem is the main result of the second section of this work (see Theorem 1.3.8). The proof relies on an interpretation of the optimization problem (1.26) as Padé approximant problem. Some of these optimal formulas are very classic. For example, the formulas $d^{0,0}, d^{1,0}, d^{2,0}$ and $d^{3,0}$ can be found explicitly in [66] whereas $(d^{0,1}, s^{0,1})$ is the classical Noumerov formula (see [95]). More generally, the optimal formulas of this theorem are effective because we can prove, with the property of uniqueness of Theorem 1.2.4, that they can be computed exactly as the solutions of these rational $(l+m+3)\times(l+m+3)$ linear systems

$$\begin{pmatrix} \mathbf{L}^0_{l+1} & \mathbf{0}_{1,m+1} \\ \mathbf{L}^2_{l+1} & \mathbf{0}_{1,m+1} \\ \mathbf{L}^2_{l+1} & 2\mathbf{L}^0_m \\ \vdots & \vdots \\ \mathbf{L}^{2(l+m+1)}_{l+1} & 2(m+l+1)(2(m+l)+1)\mathbf{L}^{2(l+m)}_m \end{pmatrix} \begin{pmatrix} d^{l,m}_0 \\ \vdots \\ d^{l,m}_{l+1} \\ s^{l,m}_0 \\ \vdots \\ s^{l,m}_m \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{1.27}$$

with

$$\mathbf{L}^k_n = \begin{pmatrix} \frac{0^k}{2} & 1^k & \dots & n^k \end{pmatrix} \text{ and } \mathbf{0}_{1,m+1} \text{ the zero row matrix of size } m+1.$$

The formulas $d^{l,m}$ being constructed as the solution of an optimization problem, there is *a priori* no reason that they generate stable schemes. However, the following theorem, which will be proved in the third section (see Application 2 of Theorem 1.4.1), precisely states that all these optimal schemes are indeed strongly stable.

**Theorem 1.2.4.** *For all $l \in \mathbb{N}^*$ and for all $m \in \mathbb{N}$, $(\mathbf{D}_N(d^{l,m}))_{N\in\mathbb{N}^*}$ is strongly stable, see* (1.9).

In particular, following Theorem 1.2.1, the schemes designed in Proposition 1.2.4, with $d = d^{l,m}$, $s = l, m$, $n = 2(l+m+1)$ and $\mu = 2(l+m)$, are convergent of order $2(l+m+1)$. The efficiency of these schemes is discussed in subsection 1.5.1.

As explained in the introduction, we now address the question of *generic* performance of the schemes that we have constructed above : are they stable and convergent *in general* once the algebraic order conditions are satisfied. To give a meaning to this question, we decide to use measure theory. Of course there exist formulas such that $(\mathbf{D}_N(d))_N$ can not be stable. It is the case, for example, when $\mathbf{D}_N(d)$ is not invertible for all $N$ which occurs for when the polynomial $P$ defining the scheme admit a root of the form $4\sin^2\left(\frac{\pi}{2}kh\right)$, see (1.23), which are eigenvalues of the matrix $\mathbf{A}_N$. But even if this is not the case, these eigenvalues can be very close to the roots of $P$, which induces small denominators in the stability estimates. Of course, these situations have to be avoided as well.

The following theorem gives an answer to these questions (see Application 1 of Theorem 1.4.1 and Application 2 of Theorem 1.4.3 for the proof).

**Theorem 1.2.5.** *Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ be a field and $l \in \mathbb{N}$ be an integer. Let $\mathscr{C}_{\mathbb{K},l}$ be the $\mathbb{K}$ finite dimensional vector space of symmetric formulas $d \in \mathscr{S}_{\mathbb{K}}$ with zero mean* (1.16) *and $\tau(d) \leqslant l+1$*

$$\mathscr{C}_{\mathbb{K},l} = \{d \in \mathscr{S}_{\mathbb{K}} \mid \sum_{j\in\mathbb{Z}} d_j = 0 \text{ and } \tau(d) \leqslant l+1\}.$$

*Then for any Lebesgue measure on $\mathscr{C}_{\mathbb{K},l}$, we have that :*

- *For almost all $d \in \mathscr{C}_{\mathbb{C},l}$, $(\mathbf{D}_N(d))_{N\in\mathbb{N}^*}$ is strongly stable* (1.9)*.*
- *For almost all $d \in \mathscr{C}_{\mathbb{R},l}$, $(\mathbf{D}_N(d))_{N\in\mathbb{N}^*}$ is stable relatively to any sequence $(\eta_N)_N$* (1.10) *such that*

$$\sum_{N\in\mathbb{N}^*} \frac{N+1}{\eta_N} < \infty \text{ and } \sup_{N\in\mathbb{N}^*} \frac{(N+1)^2}{\eta_N} < \infty.$$

As for Bertrand series, there is no optimal choice of sequence $(\eta_N)_{N\in\mathbb{N}}$ that satisfy this condition and we can not directly deduce stability in the sense of (1.8), but we can choose

$$\eta_N = ((N+1)\log(N+1))^2 = \left(\frac{\log h}{h}\right)^2.$$

As a consequence, we affirm that, up to some logarithmic corrections, almost all real symmetric formula generates stable schemes.

We use this theorem to deduce a convergence result.

**Proposition 1.2.5.** *With any given $d \in \mathscr{C}_{\mathbb{K},l}$, we can associate the scheme of Proposition 1.2.4, with $\mu = n-2$ if $\mathbb{K} = \mathbb{C}$ and $\mu = n$ if $\mathbb{K} = \mathbb{R}$, and the formula $s$ given by Proposition 1.2.1. Then, it follows from Theorem 1.2.1 that for all $l \in \mathbb{N}$,*

- *for almost all $d \in \mathscr{C}_{\mathbb{C},l}$, the associated scheme is convergent of order $n$.*
- *for almost all $d \in \mathscr{C}_{\mathbb{R},l}$, the associated scheme converges at the rate $h^n(\log(h))^2$.*

In the proof of Theorem 1.2.5 for real formulas, the logarithmic correction is due to the use of a diophantine control of some resonances. Experimentally, we can indeed evidence these quasi-resonances by plotting convergence curves for various schemes of Proposition 1.2.5 for randomly drawn formulas $d$ (see subsection 1.5.2).

## 1.3 Polynomials and high order formulas

The aim of this section is two fold. First, to explain why the matrices $\mathbf{D}_N(d)$ constructed in Proposition 1.2.4 are polynomials in $d$. Second, to give criteria of consistency on these polynomials to interpret Theorem 1.2.3 as a classical problem of Padé approximant.

To highlight the algebraic structure of the matrices $\mathbf{D}_N(d)$ of Proposition 1.2.4 we will now give a more formal definition of these matrices.

Let $d \in \mathscr{S}_{\mathbb{C}}$ be a symmetric formula. We introduce $T_d \in \mathscr{L}(\mathbb{C}^{\mathbb{Z}})$, the operator of convolution by $d$,

$$\forall w \in \mathbb{C}^{\mathbb{Z}},\ T_d(w) = d \star w = \left( \sum_{j \in \mathbb{Z}} d_j w_{i-j} \right)_{i \in \mathbb{Z}}. \tag{1.28}$$

Let $N \in \mathbb{N}^*$ and $\mathscr{E}_N$ be the space of the odd functions from $\mathbb{Z}$ to $\mathbb{C}$ that are odd in $0$ and in $N+1$

$$\mathscr{E}_N := \{ w \in \mathbb{C}^{\mathbb{Z}} \mid \forall j \in \mathbb{Z},\ w_{N+1+j} = -w_{N+1-j} \text{ and } w_{-j} = -w_j \}. \tag{1.29}$$

Let $\mathscr{B}_N$ be the canonical basis of $\mathscr{E}_N$

$$\mathscr{B}_N = (\mathbb{1}_{j+(2N+2)\mathbb{Z}} - \mathbb{1}_{-j+(2N+2)\mathbb{Z}})_{j=1,\dots,N}. \tag{1.30}$$

Since $d$ is symmetric, we verify that $\mathscr{E}_N$ is stable by $T_d$.

**Definition 1.3.1.** *With the previous construction, we define $\mathbf{D}_N(d)$ through the relation*

$$\mathbf{D}_N(d) = \mathrm{mat}_{\mathscr{B}_N} T_{d|\ \mathscr{E}_N}.$$

**Remark 1.3.2.** *Of course, this definition of $\mathbf{D}_N(d)$ gives the same matrices, when $N$ is large enough, as the matrix of Proposition 1.2.4.*

The space of symmetric formulas has a structure of free module on the ring of polynomial that is very useful to write efficient and accurate high order schemes. More precisely, we equip the set of formula $\mathbb{C}^{(\mathbb{Z})}$ of its structure of commutative algebra for the convolution

$$\forall d, s \in \mathbb{C}^{(\mathbb{Z})},\ d \star s = \left( \sum_{j \in \mathbb{Z}} d_j s_{i-j} \right)_{i \in \mathbb{Z}}.$$

Then, if $R$ is a ring such that $\mathbb{Z} \subset R \subset \mathbb{C}$, we consider $\mathscr{S}_R$ as a subalgebra of $\mathbb{C}^{(\mathbb{Z})}$.

On the one hand, this structure explains, through the following lemma, the importance of the formula $a$ (defined in (1.12) by $a = 2\mathbb{1}_{\{0\}} - \mathbb{1}_{\{-1,1\}}$).

**Lemma 1.3.3.** *If $R$ is a ring such that $\mathbb{Z} \subset R \subset \mathbb{C}$ then $\mathscr{S}_R$ is a free $R[X]$ module whose $a$ is a basis*

$$\forall d \in \mathscr{S}_R,\ \exists\,!\ P \in R[X],\ d = P(a).$$

*Furthermore, if $P \in \mathbb{C}[X]$ then*

$$\tau(P(a)) = \deg P.$$

*Proof.* If we consider $\mathbb{C}^{(\mathbb{Z})}$ as a subalgebra of $\mathbb{C}^{(\frac{\mathbb{Z}}{2})}$ then we remark that

$$a = -\left(\mathbb{1}_{\{\frac{1}{2}\}} - \mathbb{1}_{\{-\frac{1}{2}\}}\right)^{\star 2}.$$

Consequently, a binomial expansion gives

$$\forall n \in \mathbb{N},\ a^{\star n} = (-1)^n \left(\mathbb{1}_{\{\frac{1}{2}\}} - \mathbb{1}_{\{-\frac{1}{2}\}}\right)^{\star 2n} = \sum_{k=0}^{n} \frac{(2n)!}{(n+k)!(n-k)!}(-1)^k \mathbb{1}_{\{k,-k\}}.$$

The second point of the lemma is clearly a consequence of this expansion. Furthermore, since the term associated to the highest index of $a^{\star n}$ (i.e. $(-1)^n$) is invertible in $R$, the first point follows from an induction. □

On the other hand, this structure explains, why the matrices $\mathbf{D}_N(d)$ are polynomials in $\mathbf{A}_N$.

**Lemma 1.3.4.** *For all $N \in \mathbb{N}^*$, $d \mapsto \mathbf{D}_N(d)$ is a $\mathbb{C}[X]$ module morphism :*

$$\forall P \in \mathbb{C}[X], \forall d \in \mathscr{S}_{\mathbb{C}},\ \mathbf{D}_N(P(d)) = P(\mathbf{D}_N(d)).$$

*Proof.* It follows directly of Definition 1.3.1 of $\mathbf{D}_N(d)$ and of the associativity of the convolution. □

### 1.3.1 Consistency for the polynomials

We start with a lemma that we have used implicitly in the introduction (in Proposition (1.2.1) and Proposition (1.2.3)).

**Lemma 1.3.5.** *Let $n \in 2\mathbb{N}$. Then a couple of formulas $(d, s) \in \mathscr{S}_{\mathbb{C}}^2$ is consistent of order $n$* (1.11) *if and only if*

$$\forall p \in \mathbb{C}_{n+1}[X],\ \sum_{j\in\mathbb{Z}} d_j p(j) + s_j p''(j) = 0.$$

*Proof.* It is enough to choose $u(x) = x^i$ with $i \leqslant n+1$ is the definition of the consistency and then to simplify the powers of "$h$". Conversely, it is enough to do a Taylor expansion. □

In particular, if we choose $p = 1$, we find that the consistency of order $n = 0$ is nothing but the condition of zero mean (1.16) for $d$.

We introduce the formal Fourier transform $\mathscr{F}$ from the algebra of formulas $\mathbb{C}^{(\mathbb{Z})}$ to the algebra of formal series $\mathbb{C}[\![X]\!]$ defined by

$$\mathscr{F} : \left\{ \begin{array}{lcl} \mathbb{C}^{(\mathbb{Z})} & \to & \mathbb{C}[\![X]\!] \\ d & \mapsto & \displaystyle\sum_{j\in\mathbb{Z}} d_j e^{ijX}. \end{array} \right.$$

We give a characterization of consistency through this transform.

**Lemma 1.3.6.** *Let $n \in 2\mathbb{N}$. A couple of formulas $(d, s) \in \mathscr{S}_{\mathbb{C}}^2$ is consistent of order $n$* (1.11) *if and only if*

$$\mathscr{F}d = X^2 \mathscr{F}s \mod X^{n+2}.$$

*Proof.* Let $\partial_X \in \mathscr{L}(\mathbb{C}[X])$ be the formal derivative on the space of the polynomials $\mathbb{C}[X]$. As a consequence, since the Taylor expansion in $0$ of a polynomial is exact, if $p \in \mathbb{C}[X]$ and $x_0 \in \mathbb{C}$ then we have

$$p(x_0) = e^{x_0 \partial_X} p(0).$$

Consequently, following Lemma (1.3.5), $(d, s) \in \mathscr{S}^2_{\mathbb{C}}$ is consistent of order $n$ (1.11) if and only if

$$\forall p \in \mathbb{C}_{n+1}[X],\ \left(\mathscr{F}d - X^2 \mathscr{F}s\right)(i\partial_X)p(0) = 0.$$

We conclude the proof by considering the lowest power in the expansion of $\mathscr{F}d - X^2\mathscr{F}s$ in $0$. □

The formal Fourier transform is as usual an algebra morphism. As a consequence, if $P \in \mathbb{C}[X]$ and $d = P(a)$ then

$$\sum_{k \in \mathbb{N}^*} \sum_{j \in \mathbb{Z}} d_j \frac{(-1)^k j^{2k}}{(2k)!} X^{2k} = \mathscr{F}d = P(\mathscr{F}a) = P(2 - 2\cos(X)) = P\left(4\sin^2\left(\frac{X}{2}\right)\right). \tag{1.31}$$

The consistency and the stability of the formulas often involve moment of $d$ or $s$. In particular, this relation provides simple expressions for the first moments of $d$ in function of $P$

$$\sum_{j \in \mathbb{Z}} d_j = P(0) \text{ and } \sum_{j \in \mathbb{Z}} d_j j^2 = -2P'(0).$$

In fact, with the formula (1.31), we get a criterion of consistency directly on the polynomials.

**Lemma 1.3.7.** *Let $P, Q \in \mathbb{C}[X]$ and $n \in 2\mathbb{N}^*$. The couple of symmetric formulas $(P(a), Q(a))$ is consistent of order $n$* (1.11) *if and only if*

$$P(4X^2) = 4\left(\arcsin(X)\right)^2 Q(4X^2) \mod X^{n+2},$$

*where $\left(\arcsin(X)\right)^2$ is the square of the inverse sine function whose expansion is (for a reference, see, for example, [34])*

$$\left(\arcsin(X)\right)^2 = \sum_{n \in \mathbb{N}^*} \frac{2^{2n-1}}{n^2 C_{2n}^n} X^{2n}.$$

*Proof.* If we apply (1.31) to the criterion of consistency of Lemma 1.3.6 then it comes

$$P\left(4\sin^2\left(\frac{X}{2}\right)\right) = X^2 Q\left(4\sin^2\left(\frac{X}{2}\right)\right) \mod X^{n+2}.$$

To conclude the proof, it is enough to do the change of variable

$$X \leftarrow 2\arcsin(X).$$

□

### 1.3.2 The optimal case

In order to prove Theorem 1.2.3 we are going to explain the link between the problem of optimization (1.26) and the theory of Padé approximant. To see this link we introduce the usual valuation on $\mathbb{C}[\![X]\!]$ :

$$\forall C \in \mathbb{C}[\![X]\!],\ \mathrm{val}(C) = \min\{k \in \mathbb{N} \mid \forall 0 \leqslant j \leqslant k,\ C^{(j)}(0) = 0\}.$$

As a consequence, with this formalism, Lemma 1.3.7 can be written

$$\forall P, Q \in \mathbb{C}[X],\ \mathrm{ord}\,(P(a), Q(a)) = \mathrm{val}\left(P(4X^2) - 4\,(\arcsin(X))^2\,Q(4X^2)\right) - 2.$$

However, Lemma 1.3.3 proves that $(P, Q) \mapsto (P(a), Q(a))$ is a bijection from $\mathbb{C}_{l+1}[X] \times \mathbb{C}_m[X]$ to the space of the couples of symmetric formulas $(d, s)$ such that $\tau(d) \leqslant l+1$ and $\tau(s) \leqslant m$. As a consequence, the problem of optimization (1.26) is equivalent to the following

$$\max_{\substack{(P,Q)\in\mathbb{C}[X]^2\setminus\{(0,0)\}\\ \deg P\leqslant l+1,\ \deg Q\leqslant m}} \mathrm{val}\left(P(4X^2) - 4\,(\arcsin(X))^2\,Q(4X^2)\right).$$

Since if $P(0) \neq 0$ then $\mathrm{val}\left(P(4X^2) - 4\,(\arcsin(X))^2\,Q(4X^2)\right) = 0$, it is natural to study this problem of optimization for polynomials $P$ such that

$$P = XR \text{ where } R \in \mathbb{C}_l[X].$$

Consequently, it is enough to study the following problem of optimization

$$\max_{\substack{(R,Q)\in\mathbb{C}[X]^2\setminus\{(0,0)\}\\ \deg R\leqslant l,\ \deg Q\leqslant m}} \mathrm{val}\,(R - CQ)\,, \tag{1.32}$$

with (see [34] for the expansion)

$$C(X) := 4\left(\frac{\arcsin(\frac{\sqrt{X}}{2})}{\sqrt{X}}\right)^2 = 2\sum_{n\in\mathbb{N}} \frac{X^n}{(n+1)^2 C_{2n+2}^{n+1}}. \tag{1.33}$$

The theory of Padé approximants is a deep theory about approximation of formal series by rational ones. It has been extensively developed in the last decades (see [11] or [70] for an overview). Its aim is to give to each formal series $F \in \mathbb{C}[\![X]\!]$ a rational approximation $\frac{p_{l,m}}{q_{l,m}}$ (usually noted $[l/m]$) such that

$$F = \frac{p_{l,m}}{q_{l,m}} \mod X^{l+m+1}\text{ , with } p_{l,m} \in \mathbb{C}_l[X] \text{ and } q_{l,m} \in \mathbb{C}_m[X]. \tag{1.34}$$

A natural way to find such an approximation is to try to solve

$$p_{l,m} = F q_{l,m} \mod X^{l+m+1}\text{ , with } p_{l,m} \in \mathbb{C}_l[X] \text{ and } q_{l,m} \in \mathbb{C}_m[X]. \tag{1.35}$$

Indeed, if we get a solution $(p_{l,m}, q_{l,m})$ of (1.35) with $q_{l,m}(0) \neq 0$ then it is also a solution of (1.34). The second formulation (1.35) is interesting because it is a linear system of $l+m+1$

equations and $l+m+2$ unknowns. Consequently, it admits at least one non trivial solution. However, the question of its uniqueness (up to multiplication by a scalar) is generaly non trivial. In the classical Padé theory, if for a formal series $F$, the linear system (1.35) admits for all $l, m \in \mathbb{N}$, a unique non trivial solution (up to multiplication by a scalar), then it is said that the Padé table of $F$ is *normal*. Furthermore, if $F(0) \neq 0$ and if its Padé table is normal then a non trivial solution $(p_{l,m}, q_{l,m})$ of (1.35) satisfies $\deg p_{l,m} = l$, $\deg q_{l,m} = m$, $q_{l,m}(0) \neq 0$ and $\mathrm{val}(p_{l,m} - Fq_{l,m}) = l+m+1$ (see [11] or [70] for details).

What is crucial for us is that the Padé table of $C$ is normal. In fact, D. Karp and E. Prilepkina have proved in [86] that the Padé tables of many generalized hypergeometric functions are normals. To see that the Padé table of $C$ is normal, we just have to verify that $C$ is one of those generalized hypergeometric functions. The generalized hypergeometric functions are the formal series defined by

$$ {}_pF_q\left[\begin{matrix}\alpha_1 & \dots & \alpha_p\\ \beta_1 & \dots & \beta_q\end{matrix}; X\right] = \sum_{k\in\mathbb{N}} \frac{(\alpha_1)_k \dots (\alpha_p)_k}{(\beta)_1 \dots (\beta_q)_k} \frac{X^k}{k!} \text{ with } (\gamma)_k = \prod_{j=0}^{k-1} \gamma + j. \tag{1.36} $$

D. Karp and E. Prilepkina have proved in Theorem $9$ of [86] that if

$$ \begin{cases} p = q+1, \\ 0 < \alpha_{q+1} \leqslant 1, \\ 0 < \alpha_1 \leqslant \dots \leqslant \alpha_q, \\ 0 < \beta_1 \leqslant \dots \leqslant \beta_q, \\ \forall k \in [\![1, q]\!],\ \displaystyle\sum_{j=1}^{k} \alpha_j \leqslant \sum_{j=1}^{k} \beta_j \end{cases} $$

then the Padé table of ${}_pF_q\left[\begin{smallmatrix}\alpha_1 & \dots & \alpha_p\\ \beta_1 & \dots & \beta_q\end{smallmatrix}; X\right]$ is normal. However, $C$ is one of those generalized hypergeometric functions because

$$ C(X) = {}_3F_2\left[\begin{matrix}1 & 1 & 1\\ \frac{3}{2} & & 2\end{matrix}; \frac{X}{4}\right]. \tag{1.37} $$

We verify this assertion by the following elementary calculation

$$ \frac{(n+1)^2 C_{2n+2}^{n+1}}{(n+2)^2 C_{2n+4}^{n+2}} = \frac{(n+1)^2}{(2n+3)(2n+4)} = \frac{1}{4}\frac{(n+1)^2}{(n+\frac{3}{2})(n+2)}, $$

which shows by induction that the coefficients of $C(X)$ (see (1.33)) coincide with those of one of the generalized hypergeometric functions in (1.37), see (1.36).

Now, we just have to link these results of Padé approximation with our optimization problem (1.26). But if we denote by $(R_{l,m}, Q_{l,m})$ the solution of (1.35) such that $R_{l,m}(0) = 1$, then we have

$$ \mathrm{val}(R_{l,m} - CQ_{l,m}) = l+m+1. $$

Conversely, if $(R, Q)$ satisfies $\mathrm{val}(R - CQ) \geqslant l+m+1$ with $\deg R \leqslant l$ and $\deg Q \leqslant m$ then it is a solution of (1.35). But since the Padé table of $C$ is normal, $(R, Q)$ is equal to $(R_{l,m}, Q_{l,m})$, up to multiplication by a scalar.

Consequently, we have proved that the numerator and the denominator of the Padé approximant of $C$ are the solutions to the optimization problem (1.32), up to multiplication by a scalar. All the results of this analysis is summarized in the following theorem that is nothing but a version of Theorem 1.2.3 with polynomials.

**Theorem 1.3.8.** *For all $l, m \in \mathbb{N}$, there exists a couple of rational polynomial $(R_{l,m}, Q_{l,m}) \in \mathbb{Q}[X]^2$ such that*

$$\begin{cases} \deg R_{l,m} = l, \\ \deg Q_{l,m} = m, \\ R_{l,m}(0) = 1. \end{cases}$$

*Moreover $(R_{l,m}, Q_{l,m})$ is solution of the optimization problem*

$$\max_{\substack{(R,Q)\in\mathbb{C}[X]^2\setminus\{(0,0)\} \\ \deg R\leqslant l,\ \deg Q\leqslant m}} \operatorname{val}(R - CQ) = \operatorname{val}(R_{l,m} - CQ_{l,m}) = l + m + 1.$$

*Furthermore, this solution is essentially unique : if $(R, S) \in \mathbb{C}[X]^2$ is such that $\deg R \leqslant l$, $\deg Q \leqslant m$ and $\operatorname{val}(R_{l,m} - CQ_{l,m}) = l + m + 1$ then there exists $\lambda \in \mathbb{C}$ such that $R = \lambda R_{l,m}$ and $Q = \lambda Q_{l,m}$.*

There exists many very efficient methods to compute effectively Padé approximants (see for example [11] or [70]). Consequently, if the order of consistency is large enough, it is interesting to not compute the optimal formulas of Theorem 1.2.3 through the resolution of the linear system (1.27), but to compute them from the optimal polynomials of Theorem 1.3.8 through the relations

$$s^{l,m} = Q_{l,m}(a) \text{ and } d^{l,m} = P_{l,m}(a) \text{ with } P_{l,m}(X) = XR_{l,m}(X). \tag{1.38}$$

## 1.4 Stability

In this section we study criteria of stability for the sequences of matrices of the form $P(\mathbf{A}_N)$ with $P$ a polynomial. These conditions hold on the polynomial $P$. As a consequence, if we want to apply one of these criteria to a matrix of the form $\mathbf{D}_N(d)$, with $d$ a symmetric formula, we have to solve $P(a) = d$ (see Lemma 1.3.3 for details).

In the first part, we give a criterion of strong stability (1.9) and then we deduce Theorem 1.2.4 and the first part of Theorem 1.2.5 (when the formulas are complex). In the second part, we give a diophantine criterion of relative stability (1.10) that is enough to prove the second part of Theorem 1.2.5 (when the formulas are real).

### 1.4.1 Strong stability

**Theorem 1.4.1.** *Let $P \in \mathbb{C}[X]$ be a polynomial such that*

$$P(0) = 0,\ P'(0) \neq 0 \text{ and } \forall x \in ]0, 4],\ P(x) \neq 0. \tag{1.39}$$

*Then the sequence of matrices $(P(\mathbf{A}_N))_{N\in\mathbb{N}^*}$ is strongly stable* (1.9).

*Proof.* The assumptions (1.39) implies that there exists $\beta \neq 0$ a real number and a sequence $(\mu_k)_{k=1\dots d}$ of complex numbers such that

$$P(X) = \beta X \prod_{k=1}^{d} (X - \mu_k)\,.$$

On the one hand, a straightforward calculation shows that $\mathbf{A}_N$ is invertible and

$$\forall i, j \in [\![1, N]\!],\ (\mathbf{A}_N^{-1})_{i,j} = \min(j(1-hi), i(1-hj)).$$

On the other hand, since by assumption $\mu_k \notin [0,4]$, the following lemma (proved in Appendix 1.6.2) shows that $\mathbf{A}_N - \mu_k \mathbf{I}_N$ is invertible and that there exists a constant $c_{\mu_k}$ such that for all $N$, $\|(\mathbf{A}_N - \mu_k \mathbf{I}_N)^{-1}\|_\infty \leqslant c_{\mu_k}$.

**Lemma 1.4.2.** *If $\mu \in \mathbb{C}\backslash[0,4]$ then there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, $\mathbf{A}_N - \mu \mathbf{I}_N$ is invertible and for all $\mathbf{v} \in \mathbb{C}^N$*

$$\|\mathbf{v}\|_\infty \leqslant c \|\mathbf{A}_N \mathbf{v} - \mu \mathbf{v}\|_\infty.$$

As a consequence, $P(\mathbf{A}_N)$ is invertible and we have

$$\forall N \in \mathbb{N}^*, \forall \mathbf{v} \in \mathbb{R}^N,\ \|P(\mathbf{A}_N)^{-1}\mathbf{v}\|_\infty \leqslant |\beta|^{-1} \|\mathbf{A}_N^{-1}\mathbf{v}\|_\infty \prod_{k=1}^{d} c_{\mu_k}.$$

Hence, to prove Theorem 1.4.1, it is enough to prove that $(\mathbf{A}_N)_N$ is strongly stable (1.9). The estimation of strong stability of $(\mathbf{A}_N)_N$ is very explicit and is given for $l \in \mathbb{N}$ by

$$\begin{aligned}
&\|\mathbf{A}_N^{-1}\mathbf{v}\|_\infty \\
&\leqslant \sup_{i=1}^{N} \sum_{j=1}^{N} |\mathbf{v}_j| \min\{i(1-hj), j(1-hi)\} \\
&\leqslant \sup_{i=1}^{N} \sum_{j \in [\![l+1, N-l]\!]} |\mathbf{v}_j| \min\{i(1-hj), j(1-hi)\} + \sup_{i=1}^{N} \sum_{j \in [\![l+1, N-l]\!]^c} |\mathbf{v}_j| \min\{i(1-hj), j(1-hi)\} \\
&\leqslant \sup_{i=1}^{N} \sum_{j \in [\![l+1, N-l]\!]} |\mathbf{v}_j| \frac{4}{h} + \sup_{i=1}^{N} \sum_{j \in [\![l+1, N-l]\!]^c} |\mathbf{v}_j| l \\
&\leqslant \sup_{j=1}^{N} \begin{cases} 4h^{-2}|\mathbf{v}_j| & \text{if } l+1 \leqslant j \leqslant N-l, \\ 2l^2|\mathbf{v}_j| & \text{else.} \end{cases}
\end{aligned}$$

□

**Application 1 : Proof of the first part of Theorem 1.2.5** The more direct application of this criterion of strong stability is the first part of Theorem 1.2.5. Since we have proved in Lemma 1.3.3 that $P \mapsto P(a)$ induce an isomorphism of vector space between $X\mathbb{C}_l[X]$ and $\mathscr{C}_{\mathbb{C},l}$, it is enough to prove that almost all complex polynomials of degree smaller than $l+1$ do not have any zero point in $[0,4]$ to conclude with the criterion of stability of Theorem 1.4.1. In fact, we show that almost all complex polynomials of degree smaller than $l+1$ do not have any real zero point.

*Proof.* Since the null sets are the same for all the Lebesgue measures on $\mathbb{C}_l[X]$ it is enough to prove the result for one well chosen Lebesgue measure. As a consequence, we introduce $\lambda$ be a Lebesgue measure on $\mathbb{R}_l[X]$ and we consider $\lambda^{\otimes 2}$ as a Lebesgue measure on $\mathbb{C}_l[X]$ induced by the direct sum

$$\mathbb{C}_l[X] = \mathbb{R}_l[X] \oplus i\mathbb{R}_l[X].$$

Now, we remark that if a polynomial $P \in \mathbb{C}_l[X]$ admits the decomposition $P = P_1 + iP_2$ and has a real zero point $x \in \mathbb{R}$ then $x$ is a zero point of $P_1$ and of $P_2$. As a consequence, we conclude by the following calculation

$$\begin{aligned}
\lambda^{\otimes 2}\{P \in \mathbb{C}_l[X] \mid \exists x \in \mathbb{R},\ P(x) = 0\} &= \int_{\mathbb{R}_l[X]} \int_{\mathbb{R}_l[X]} \mathbb{1}_{\exists x \in \mathbb{R},\ (P_1+iP_2)(x)=0} \mathrm{d}\lambda(\mathrm{P}_1)\mathrm{d}\lambda(\mathrm{P}_2) \\
&= \int_{\mathbb{R}_l[X]} \int_{\mathbb{R}_l[X]} \mathbb{1}_{\exists x \in \mathbb{R},\ P_2(x)=P_1(x)=0} \mathrm{d}\lambda(\mathrm{P}_1)\mathrm{d}\lambda(\mathrm{P}_2) \\
&\leqslant \int_{\mathbb{R}_l[X]} \sum_{x \in \mathbb{R}, P_2(x)=0} \int_{\mathbb{R}_l[X]} \mathbb{1}_{P_1(x)=0} \mathrm{d}\lambda(\mathrm{P}_1)\mathrm{d}\lambda(\mathrm{P}_2) \\
&= 0.
\end{aligned}$$

The last equality is nothing but, since $\{P_1 \in \mathbb{R}_l[X] |\ P_1(x) = 0\}$ is an hyperplane of $\mathbb{R}_l[X]$, its Lebesgue measure is zero. □

**Application 2 : Proof of Theorem 1.2.4** The second application of the criterion of strong stability of Theorem 1.4.1 is the Theorem 1.2.4 about the strong stability of the most efficient schemes. In fact, to apply this criterion to the optimal formulas of Theorem 1.2.3, we exactly have to prove that the optimal polynomials $R_{l,m}$ of Theorem 1.3.8 do not have any zeros point in $[0, 4]$.

*Proof.* Let $l, m \in \mathbb{N}$ be some integers and $R_{l,m}$ the optimal polynomial given by Theorem 1.3.8. In the proof of this theorem, $R_{l,m}$ is built as the numerator of the Padé approximant of the function $C$ (1.33). Futhermore, as we have explained in the proof of Theorem 1.3.8, D. Karp and E. Prilepkina have proved in [86] that $C(-X)$ is a Stieltjes transform of a measure supported in $[0, 4]$. As a consequence, we can use the classical results about the localization of the zeros points and poles of the Padé approximants of such series.

On the one hand, it is enough to apply the point $(vii)$ of Theorem $3$ page $251$ of the book of J. Gilewicz [70] to prove that if $k \leqslant 0$ and $l \geqslant -k$ then all the zero points of $R_{l+k,l}$ are in $]4, \infty[$.

On the other hand, J. Gilewicz proves at the point $(iii)$ of this theorem that if $k \geqslant -1$, $l+k \geqslant 0$ and $l \geqslant 0$ then all the zero points of $Q_{l+k,l}$ (the denominator of the Padé approximant of $C$) are in $]4, \infty[$. Futhermore, page $264$ of his book [70], J. Gilewicz gives a theorem of Stieltjes and Wynn (point $(iii)$ of Theorem $5$) that implies that if $k \geqslant 0$ and $l \leqslant 0$ then

$$\forall x \in [0, 4],\ \frac{R_{l+k,l}(x)}{Q_{l+k,l}(x)} \leqslant \frac{R_{l+k+1,l+1}(x)}{Q_{l+k+1,l+1}(x)}.$$

Since $Q_{l+k,l}$ does not have any zero point on $[0, 4]$ and since by construction $Q_{l+k,l}(0) = R_{l+k,l}(0) = 1$ then it follows that for all $k \geqslant 0$ and all $l \geqslant 0$ we have

$$\forall x \in [0, 4],\ Q_{l+k,l}(x) > 0.$$

As a consequence, if $k \geqslant 0$ and $l \geqslant 0$ then, we have

$$\forall x \in [0,4],\ \frac{Q_{l+k+1,l+1}(x)}{Q_{l+k,l}(x)} R_{l+k,l}(x) \leqslant R_{l+k+1,l+1}(x).$$

Consequently, if for all $k \geqslant 0$, we prove that $R_{k,0}$ is positive on $[0,4]$, then we conclude by induction on $l \geqslant 0$, that $R_{l+k,l}$ is positive on $[0,4]$. Indeed, it is clear that $R_{k,0}$ is positive on $[0,4]$ because by construction (see Theorem 1.3.8), we have

$$R_{k,0} = 2\sum_{n=0}^{k} \frac{X^n}{(n+1)^2 C_{2n+2}^{n+1}} > 0 \text{ on } \mathbb{R}^+.$$

□

### 1.4.2 Relative stability

**Theorem 1.4.3.** *Let $P \in \mathbb{C}[X]$ be a polynomial and let $\Lambda$ be the set of the roots of $P$ in $[0,4]$ and assume that $P$ satisfies the following assumptions :*

*i)* $0 \in \Lambda$,

*ii)* $4 \notin \Lambda$,

*iii) the roots of $P$ in $[0,4]$ are simple,*

*iv)* $\exists \delta : \mathbb{N}^* \to \mathbb{R}_+^*$,

$$\forall \lambda \in \Lambda, \forall q \in \mathbb{N}^*, \forall 1 \leqslant p \leqslant q-1,\ \ 0 < \delta_q \leqslant \left|\lambda - 4\sin^2\left(\frac{\pi}{2}\frac{p}{q}\right)\right|. \tag{1.40}$$

*Then the sequence of finite difference matrices $(P(\mathbf{A}_N))_{N\in\mathbb{N}^*}$ is stable relatively to the sequence $\eta_N = \frac{1}{\delta_{N+1}}$ (1.10).*

*Proof.* see Appendix 1.6.3. □

**Application 1 : stability for second order algebraic zero points** The first application of this diophantine criteria is based on a classical result about approximation of algebraic numbers by rational ones. It gives a way to design sequences of matrices $\mathbf{D}_N$ that are stable (1.8), but such that $\mathbf{D}_N$ has not a positive or a negative spectrum for all $N$.

**Theorem 1.4.4.** *Liouville's Theorem. (from the book of Andrei B. Shidlovskii [102] page 23)*
*If $\alpha$ is a real algebraic number of degree $n$, $n \geqslant 1$, then there exists a constant $c = c(\alpha) > 0$ such that the following inequality holds for any $p \in \mathbb{Z}$ and $q \in \mathbb{N}^*$, $\frac{p}{q} \neq \alpha$ :*

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^n}.$$

**Corollary 1.4.1.** *If a polynomial $P \in \mathbb{C}[X]$ satisfies the three first hypothesis of Theorem 1.4.3 and if for all root $\lambda \in \Lambda\backslash\{0\}$ there exist an algebraic number of degree $2$, $\alpha$, such that $\lambda = 4\sin^2(\frac{\pi}{2}\alpha)$, then the sequence of finite difference matrices $(P(\mathbf{A}_N))_N$ is stable* (1.8).

**Application 2 : Proof of the second part of Theorem 1.2.5** The proof of the second part of Theorem 1.2.5 is an adaptation of a classical qualitative result about approximation of real numbers by rational ones.

**Theorem 1.4.5.** *A version of the Khinchin's Theorem. (see for example [102] page 17)*
*Let $(\nu_q)_q$ be a sequence of positive real numbers such that the series $\sum \nu_q$ converges. Then, for almost all $\alpha \in \mathbb{R}$, there exists a constant $c > 0$ such that for all $p, q \in \mathbb{Z} \times \mathbb{N}^*$, one has*

$$|\alpha - \frac{p}{q}| \geqslant c\frac{\nu_q}{q}.$$

More precisely, to prove the second part of Theorem 1.2.5 with the criterion of Theorem 1.4.3, it is enough to prove that the following set are null set for a Lebesgue measure on $\mathbb{R}_l[X]$ (they are the sets of the polynomials that do not satisfy $ii, iii$ or $iv$) :

$$E_1 = \{R \in \mathbb{R}_l[X] \mid R(4) = 0 \text{ or } R(0) = 0\},$$

$$E_2 = \{R \in \mathbb{R}_l[X] \mid \exists \lambda \in [0,4],\ R(\lambda) = R'(\lambda) = 0\},$$

$$E_3 = \left\{R \in \mathbb{R}_l[X] \mid \exists \lambda \in [0,4],\ R(\lambda) = 0 \text{ and } \liminf_{q\to\infty} \min_{p\in[\![1,q-1]\!]} \eta_{q-1} \left|\lambda - 4\sin^2\left(\frac{\pi}{2}\frac{p}{q}\right)\right| = 0\right\}.$$

Indeed, since we have proved in Lemma 1.3.3 that $R \mapsto (XR)(a)$ induce an isomorphism of vector space between $\mathbb{R}_l[X]$ and $\mathscr{C}_{\mathbb{R},l}$, the null sets for the Lebesgue measures on $\mathbb{R}_l[X]$ are associated to the null sets for the Lebesgue measures on $\mathscr{C}_{\mathbb{R},l}$.

It is quite clear that $E_1$ and $E_2$ are null sets. Indeed, $E_1$ is a null set because since $P \mapsto P(4)$ is linear, it is an hyperplane and $E_2$ is a null set because it is the set of the zero points of the discriminant $\Delta(R) = \mathrm{Res}(R, R')$ that is a non zero polynomial of $R$. However, to prove that $E_3$ is a null set, we have to adapt the proof of the Khinchin's Theorem 1.4.5.

In order to use the Borel Cantelli Theorem, we introduce a probability measure $\rho$ on $\mathbb{R}_l[X]$ with the same null set as a Lebesgue measure. More precisely, we introduce the Lebesgue measure $\mu$ on $\mathbb{R}_l[X]$ induced by the Hardy's scalar product $\langle ., . \rangle_{\mathcal{H}^2}$. This scalar product is defined by

$$\forall R_1, R_2 \in \mathbb{R}_l[X],\ \langle R_1, R_2 \rangle_{\mathcal{H}^2} := \sum_{k=0}^{l} \frac{R_1^{(k)}(0) R_2^{(k)}(0)}{k!^2}.$$

Then, we define $\rho$ through its density with respect to $\mu$

$$\frac{\mathrm{d}\rho}{\mathrm{d}\mu} = \frac{1}{\sqrt{2\pi}^{l+1}} e^{-\frac{1}{2}\|R\|^2_{\mathcal{H}^2}}.$$

As $\rho$ has a positive density with respect to $\mu$, $\rho$ and $\mu$ have the same null sets.
Hence, since $E_2$ is a null set, it is enough to prove that $E_3 \cap E_2^c$ is a null set. As a consequence, we can use the following inclusion

$$E_2^c \cap E_3 \subset E_2^c \cap \left\{R \in \mathbb{R}_l[X] \mid \liminf_{q\to\infty} \min_{p\in[\![1,q-1]\!]} \eta_{q-1} \left|R\left(4\sin^2\left(\frac{\pi}{2}\frac{p}{q}\right)\right)\right| = 0\right\}.$$

Then, we introduce the measurable sets

$$F_q := \left\{ R \in \mathbb{R}_l[X] \mid \min_{p \in [\![1, q-1]\!]} \left| R\left(4\sin^2\left(\frac{\pi}{2}\frac{p}{q}\right)\right)\right| \leqslant \frac{1}{\eta_{q-1}} \right\},$$

to get the inclusion

$$E_2^c \cap E_3 \subset E_2^c \cap \limsup_{q\to\infty} F_q.$$

Consequently, it is enough to prove that $\sum \rho(F_q) < \infty$ to conclude by the Theorem of Borel Cantelli that $E_3$ is a null set.

To control $\rho(F_q)$, we begin assuming the following lemma, that we will show at the end of this proof.

**Lemma 1.4.6.** *For all $\lambda \in \mathbb{R}$ and for all $\beta > 0$, we have*

$$\rho\left(\{R \in \mathbb{R}_l[X] \mid |R(\lambda)| \leqslant \beta\}\right) \leqslant \sqrt{\frac{2}{\pi}}\beta.$$

Consequently, we deduce from the last assumption of Theorem 1.2.5 that $\sum \rho(F_q) < \infty$,

$$\begin{aligned}\rho(F_q) &\leqslant \sum_{p=1}^{q-1} \rho\left\{R \in \mathbb{R}_l[X] \mid \left|R\left(4\sin^2\left(\frac{\pi}{2}\frac{p}{q}\right)\right)\right| \leqslant \frac{1}{\eta_{q-1}}\right\} \\ &\leqslant (q-1)\sqrt{\frac{2}{\pi}}\frac{1}{\eta_{q-1}} \in l^1(\mathbb{N}\backslash\{0,1\})\end{aligned}$$

To conclude this proof, we have to prove Lemma 1.4.6. We introduce the polynomial $R_\alpha \in \mathbb{R}_l[X]$ defined by

$$R_\alpha(X) = \sum_{k=0}^{l} (\alpha X)^k.$$

$R_\alpha$ is the Riesz representer of the evaluation in $\alpha$

$$\forall R \in \mathbb{R}_l[X],\ R(\alpha) = \langle R_\alpha, R\rangle_{\mathcal{H}^2}.$$

Consequently, since the Gaussian measure $\rho$ is isotropic, we have

$$\begin{aligned}\rho\left(\{R \in \mathbb{R}_l[X] \mid |R(\lambda)| \leqslant \beta\}\right) &= \rho\left(\{R \in \mathbb{R}_l[X] \mid |\langle R_\alpha, R\rangle_{\mathcal{H}^2}| \leqslant \beta\}\right) \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathbb{1}_{|y|\|R_\alpha\|^2_{\mathcal{H}^2}\leqslant\beta} e^{-\frac{y^2}{2}}\mathrm{dy} \\ &\leqslant \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathbb{1}_{|y|\|R_\alpha\|^2_{\mathcal{H}^2}\leqslant\beta}\mathrm{dy} \\ &= \sqrt{\frac{2}{\pi}}\beta\left(\sum_{k=0}^{l}\alpha^{2k}\right)^{-1} \leqslant \sqrt{\frac{2}{\pi}}\beta.\end{aligned}$$

## 1.5 Numerical experiments

In this section, in the light of a series of numerical experiments, we discuss the efficiency of the schemes and the optimality of our theorems.

### 1.5.1 Efficiency of the optimal schemes

In this subsection, we fix

$$u(x) = x(1-x)e^{4\cos(41x)} \text{ and } f = -u''.$$

This is a test case quite generic near the boundary : the derivatives of $f$ do not enjoy any algebraic cancellation law. Furthermore, the oscillations of the cosine term require the use of a fine grid to get an accurate approximation of the solution of the homogeneous Dirichlet problem. Thus, even with a high order method and relatively large values of $N$ (typically $N \simeq 10^2$), we are not limited by machine precision to compute convergence errors $E_N := \|\mathbf{u}^N - \mathbf{u}^{N,ex}\|_\infty$.

We consider $n, l, m$ some integers such that $2(l+m+1) = n$. Then, we consider the schemes $(\mathbf{D}_N^{l,m}, \mathbf{S}_N^{l,m})$ designed in Proposition 1.2.4 with $\mu = n-2$, $d = d_{l,m}$ and $s = s_{l,m}$. The optimal formulas $d_{l,m}$ and $s_{l,m}$ are determined by solving the linear system (1.27).

In Figure 1.2 (a), we plot their convergence curves for $n = 10$. As we have proven in Theorem 1.2.4, we observe that they are convergent of order $10$. We notice that the smaller $|m-l|$ is, the more accurate the scheme is.

Then to discuss their efficiency, we have to compare their convergence error with respect to the computational time required to solve the associated linear system : $\mathbf{D}_N^{l,m}\mathbf{u}^N = h^2\mathbf{S}_N^{l,m}\mathbf{f}^{N,ex}$. Of course, this time dependent on the linear solver used. Here, the simulations have been realized with Matlab version R2016a. Matrices $\mathbf{D}_N^{l,m}$ and $\mathbf{S}_N^{l,m}$ have been implemented as sparse matrices and the usual command $h^2\mathbf{D}_N^{l,m}\backslash(\mathbf{S}_N^{l,m} * \mathbf{f}^{N,ex})$ has been used to solve the linear system. Figure 1.2 (b) represents in $\log - \log$ scale the convergence error as a function of the time required to solve the system. The most classical scheme that is associated with $(l,m) = (4,0)$ is clearly less efficient than the others whereas the schemes associated with $(l,m) = (1,3), (2,2), (3,1)$ seem equally efficient. The scheme associated with $(l,m) = (0,4)$ is really more efficient than the others. This is due to the efficiency of Matlab to solve tridiagonal linear system. Indeed, this efficiency can be observe through the following numerical experiment.

We choose $N = 5001$ and we consider the sparse multi-diagonal matrices

$$\mathbf{M}_{i,j}^{(k)} = \begin{cases} 2k & \text{if } i = j \\ -1 & \text{else if } |i-j| \leqslant k \\ 0 & \text{else} \end{cases}$$

and the vector $\mathbf{f} = (1)_{k=1,\dots,N}$. Then plotting in Figure 1.2 (c) the time required to execute $\mathbf{M}^{(k)}\backslash\mathbf{f}$ as a function of $k$, we observe basically a linear function. However, its value in $k = 1$, corresponding to the resolution of a tridiagonal system, seems really lower than it should be if this executional time was an affine function of $k$. More precisely, realizing a linear regression excluding the tridiagonal case, we could expect a time of $0.4$ ms instead of the $0.12$ ms measured. So it is basically four time faster than what we could expect.

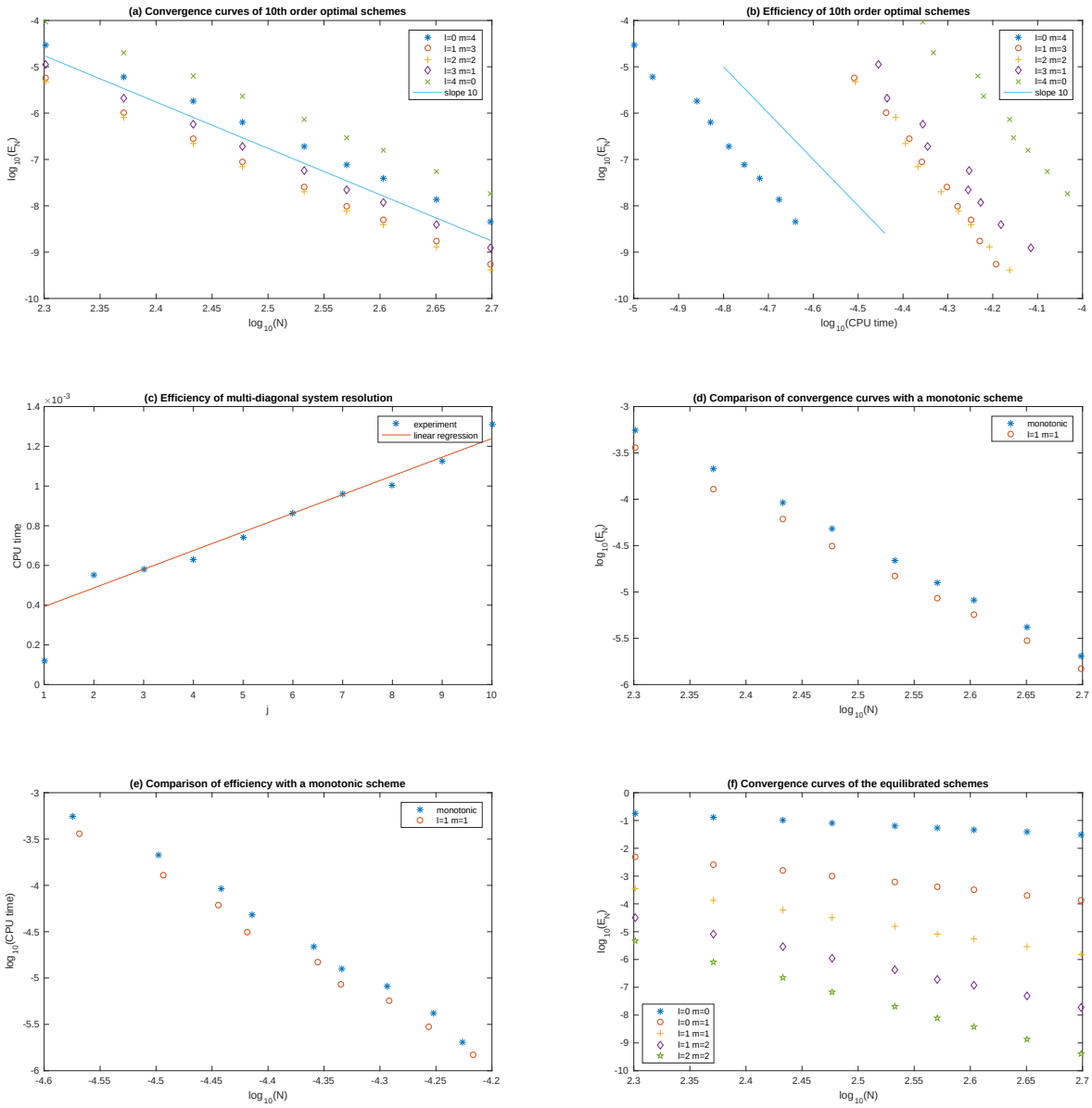


FIGURE 1.2 – Some numerical experiments about the efficiency and the accuracy of the optimal schemes designed in Proposition 1.2.4 with $\mu = n - 2$. To realize the figures (a),(b),(d),(e),(f), $N$ has been chosen in the set $\{200, 235, 271, 300, 341, 372, 401, 447, 500\}$.

In Figure 1.2 (f), we plot convergence curves of some equilibrated optimal schemes of diverse orders (i.e. schemes such that $m - \ell \geqslant 0$ is as small as possible). We observe that the high order schemes seem enjoying good convergence rate. More precisely, if $N$ is fixed, the higher the order of the scheme is, the lower the convergence error is.

Finally, we compare the optimal scheme $(\mathbf{D}_N^{1,1}, \mathbf{S}_N^{1,1})$ with a sixth order scheme denoted

$(\mathbf{D}_N^{mo}, \mathbf{S}_N^{mo})$ where

$$\mathbf{D}_N^{mo} = \begin{pmatrix} 120 & -60 & & & & & \\ -48 & 102 & -48 & -3 & & & \\ -3 & -48 & 102 & -48 & -3 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -3 & -48 & 102 & -48 & -3 \\ & & & -3 & -48 & 102 & -48 \\ & & & & & -60 & 120 \end{pmatrix}$$

and

$$\mathbf{S}_N^{mo} = \begin{pmatrix} 5 & 50 & 5 & & & & & \\ & 8 & 44 & 8 & & & & \\ & & 8 & 44 & 8 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & 8 & 44 & 8 & \\ & & & & & 8 & 44 & 8 \\ & & & & & & 5 & 50 & 5 \end{pmatrix}.$$

This scheme does not belong to the class of schemes designed in this paper. Its construction is based on monotonicity. A proof of its convergence of order $6$ can be found in [92]. We remark that, up to a scalar factor, it is designed with the optimal formulas $(d^{0,1}, s^{0,1})$ and $(d^{1,1}, s^{1,1})$. We can observe on Figure 1.2 (d) and (e) that on the test case designed in this subsection, the scheme $(\mathbf{D}_N^{mo}, \mathbf{S}_N^{mo})$ is less accurate and less efficient than the optimal scheme $(\mathbf{D}_N^{1,1}, \mathbf{S}_N^{1,1})$.

### 1.5.2 Numerical resonances

In the proof of Theorem 1.2.5 for real formulas, the logarithmic correction is due to the use of a diophantine control of some resonances. Experimentally, we aim at evidence these quasi-resonances by plotting convergence curves for various schemes of Proposition 1.2.5 for randomly drawn formulas $d$. Two typical kinds of behaviors for the convergence curves can be observed (see Figure 1.3 below). For random $d$, either we observe classical convergence curves (which are close to straight lines and correspond to non-resonant situations), or we obtain strange curves with a complex behavior corresponding to close to resonant situations. As suggested by Figure 1.3, it can be observed experimentally that the higher $n$ is, the lower the resonances are. However, we do not have any explanation of this phenomenon.

## 1.6 Appendix

### 1.6.1 Proof of Proposition 1.2.2

Let $f \in C^\infty(\mathbb{R})$ be the source function of the Dirichlet problem (1.2), and $u$ its solution. For each $N \in \mathbb{N}^*$ we consider $\mathbf{u}^{N,ex}$ and $\mathbf{f}^{N,ex}$ the discretizations of $u$ and $f$ defined by (1.3).

We begin proving the consistency of the scheme in the center. We introduce an integer $j$ such that $\tau(d) < j < N + 1 - \tau(d)$. Then, we do the estimation of consistency with a Taylor

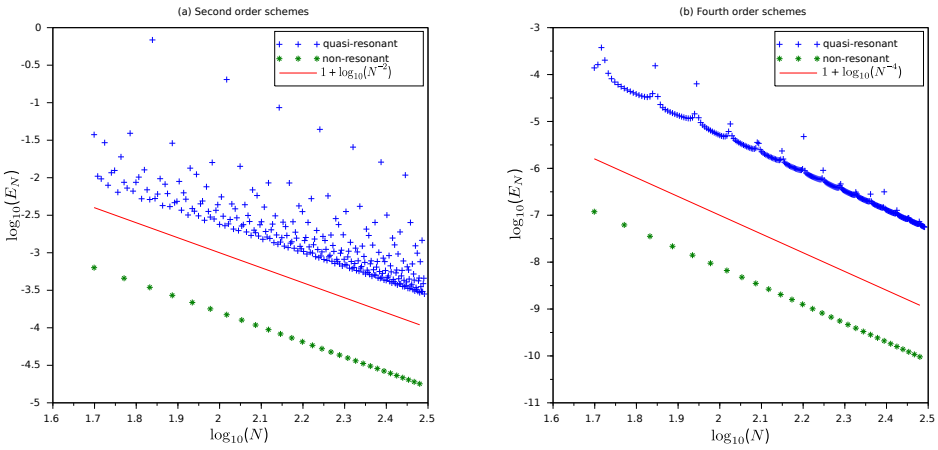


FIGURE 1.3 – Convergence curves with $u(x) = x(1-x)e^{2x}$ , $n = 2$ and $E_N := \|\mathbf{u}^N - \mathbf{u}^{N,ex}\|_\infty$. (a) For the non-resonant scheme $d = 2\mathbb{1}_{\{0\}} - \mathbb{1}_{\{-1,1\}}$ and for the quasi-resonant scheme $d = (2-6z)\mathbb{1}_{\{0\}} + (4z-1)\mathbb{1}_{\{-1,1\}} - z\mathbb{1}_{\{-2,2\}}$ with $z = 0.358946420670826$. (b) For the non-resonant scheme $d = 2\mathbb{1}_{\{0\}} - \mathbb{1}_{\{-1,1\}}$ and for the quasi-resonant scheme $d = (2-6z)\mathbb{1}_{\{0\}} + (4z-1)\mathbb{1}_{\{-1,1\}} - z\mathbb{1}_{\{-2,2\}}$ with $z = 32.12121212$

Lagrange formula

$$\begin{aligned}
&\left|\left(\mathbf{D}_N\mathbf{u}^{N,ex}\right)_j - h^2\left(\mathbf{S}_N\mathbf{f}^{N,ex}\right)_j\right| \\
&= \sum_{i\in\mathbb{Z}} d_i u(x^N_{i+j}) + h^2 s_i u''(x^N_{i+j}) \\
&= \sum_{i\in\mathbb{Z}} d_i p_j(x^N_i) + h^2 s_i p''_j(x^N_i) + \sum_{i\in\mathbb{Z}} d_i(\xi^{N,1}_{i,j} - x^N_i)^{n+2}\frac{u^{(n+2)}(\xi^{N,1}_{i,j})}{(n+2)!} + h^2 s_i(\xi^{N,2}_{i,j} - x^N_i)^n\frac{u^{(n)}(\xi^{N,2}_{i,j})}{n!},
\end{aligned}$$

with $\xi^{N,1}_{i,j}, \xi^{N,2}_{i,j} \in ]x^N_i, x^N_{i+j}[$ and

$$p_j(X) = \sum_{k=0}^{n+1}\frac{u^{(k)}(x^N_j)}{k!}X^k.$$

However, we have proved in Lemma 1.3.5, that the polynomial part of this sum is zero. Consequently, it is enough to estimate the second part. Finally, we get

$$\left|\left(\mathbf{D}_N\mathbf{u}^{N,ex}\right)_j - h^2\left(\mathbf{S}_N\mathbf{f}^{N,ex}\right)_j\right| \leqslant h^{n+2}\|u^{(n+2)}\|_{L^\infty(0,1)}\sum_{i\in\mathbb{Z}}|d_i|\frac{\tau(d)^{n+2}}{(n+2)!} + |s_i|\frac{\tau(s)^n}{n!}.$$

The same type of estimations holds near the boundary and we can prove similarly that, if $\tau(d) \geqslant j$ or $j \geqslant N+1-\tau(d)$ and if $\frac{\mu h}{2} \leqslant \gamma$ then

$$\left|\left(\mathbf{D}_N\mathbf{u}^{N,ex}\right)_j - h^2\left(\mathbf{S}_N\mathbf{f}^{N,ex}\right)_j\right| \leqslant h^{\mu+2}\|u^{(\mu+2)}\|_{L^\infty(-\gamma,1+\gamma)}\max_{1\leqslant k\leqslant\tau(d)}\sum_{i\in\mathbb{Z}}|d^k_i|\frac{\tau(d)^{\mu+2}}{(\mu+2)!} + |s^k_i|\frac{\tau(s^k)^\mu}{\mu!}.$$

### 1.6.2 Proof of Lemma 1.4.2

To prove this lemma, we need to use the notations introduced to define formally $\mathbf{D}_N(d)$ in Definition 1.3.1.

Now, for all $p \in \mathbb{Z}$ and for all $N \in \mathbb{N}^*$, we introduce an operator $O_{p,N}$ on $\mathscr{E}_N$ defined by

$$\forall w \in \mathscr{E}_N,\ O_{p,N} w = \frac{1}{2} T_{\mathbb{1}_{\{p,-p\}}} w = \left( \frac{w_{i+p} + w_{i-p}}{2} \right)_{i \in \mathbb{Z}}.$$

A straightforward calculation shows that the spectral decomposition of $O_{p,N}$ is

$$\forall k \in \mathbb{Z},\ O_{p,N} e_{k,N} = \cos(p\pi k h) e_{k,N} \text{ with } e_{k,N} = (\sin(k\pi h j))_{j \in \mathbb{Z}}. \tag{1.41}$$

Let $z \in \mathbb{C} \backslash [-1, 1]$ be a complex number. Since the periodic function $x \mapsto (\cos(x) - z)^{-1}$ is real analytic, its Fourier transform is summable. More precisely, there exists $(c_p(z)) \in l^1(\mathbb{N})$ such that

$$\forall x \in \mathbb{R},\ \frac{1}{\cos(x) - z} = \sum_{p \in \mathbb{N}} c_p(z) \cos(px).$$

Since $(c_p)$ is summable, it follows from (1.41) that $O_{1,N} - z I_{\mathscr{E}_N}$ is invertible and

$$(O_{1,N} - z I_{\mathscr{E}_N})^{-1} = \sum_{p \in \mathbb{N}} c_p(z) O_{p,N}.$$

Furthermore, if $w \in \mathscr{E}_N$ then for all $p \in \mathbb{N}$

$$\|O_{p,N} w\|_{l^\infty(\mathbb{Z})} \leqslant \|w\|_{l^\infty(\mathbb{Z})}.$$

As a consequence, we have

$$\|(O_{1,N} - z I_{\mathscr{E}_N})^{-1} w\|_{l^\infty(\mathbb{Z})} \leqslant \sum_{p \in \mathbb{N}} |c_p(z)| \|O_{p,N} w\|_{l^\infty(\mathbb{Z})} \leqslant \|(c_p(z))\|_{l^1(\mathbb{N})} \|w\|_{l^\infty(\mathbb{Z})}.$$

To finish the proof of Lemma 1.4.2, it is enough to see that

$$\operatorname{mat}_{\mathscr{B}_N} O_{1,N} = \mathbf{I}_N - \frac{1}{2} \mathbf{A}_N \text{ and } \forall w \in \mathscr{E}_N,\ \|\operatorname{mat}_{\mathscr{B}_N} w\|_\infty = \|w\|_{l^\infty(\mathbb{Z})}.$$

### 1.6.3 Proof of Theorem 1.4.3

It follows of the spectral decomposition of $\mathbf{A}_N$ (1.22), that $(\mathbf{e}_k^N)_{k=1\ldots N}$ is an orthogonal basis of $\mathbb{C}^N$. Furthermore, a straightforward calculation shows that, if $k \in [\![1, N]\!]$ then we have

$$\|\mathbf{e}_k^N\|_2 = \sum_{j=1}^N |(\mathbf{e}_k^N)_j|^2 = \sum_{j=1}^N \sin^2(\pi h k j) = \frac{1}{2} \sum_{j=1}^N 1 - \cos(2\pi h k j) = \frac{1}{2h}.$$

Consequently, if we take a vector $\mathbf{v} \in \mathbb{C}^N$, we get its discrete Fourier transform as

$$\mathbf{v} = 2h \sum_{k=1}^N \mathbf{e}_k^N \sum_{j=1}^N \mathbf{v}_j \sin(\pi h k j).$$

However, since the vectors $\mathbf{e}_k^N$ are eigenvectors of $\mathbf{A}_N$, there are eigenvectors of $P(\mathbf{A}_N)$ and their eigenvalues are $P(4\sin^2\left(\frac{\pi}{2}kh\right))$. Consequently, we know from assumption $(iv)$ that $P(\mathbf{A}_N)$ is invertible and that we have

$$P(\mathbf{A}_N)^{-1}\mathbf{v} = 2h\sum_{k=1}^{N}\frac{e_k^N}{P\left(4\sin^2\left(\frac{\pi}{2}kh\right)\right)}\sum_{j=1}^{N}\mathbf{v}_j\sin(\pi hkj).$$

Hence, if we do the estimation, $|\sin| \leqslant 1$, it comes

$$\|P(\mathbf{A}_N)^{-1}\mathbf{v}\|_\infty \leqslant 2h\sum_{k=1}^{N}\frac{1}{|P\left(4\sin^2\left(\frac{\pi}{2}kh\right)\right)|}\sum_{j=1}^{N}\|\mathbf{v}\|_\infty \leqslant \sum_{k=1}^{N}\frac{2}{|P\left(4\sin^2\left(\frac{\pi}{2}kh\right)\right)|}\|\mathbf{v}\|_\infty.$$

Consequently, to conclude the proof of Theorem 1.4.3, it is enough to proof that there exists a constant $c > 0$ such that

$$\forall N \in \mathbb{N}^*,\ \sum_{k=1}^{N}\frac{2}{|P\left(4\sin^2\left(\frac{\pi}{2}kh\right)\right)|} \leqslant \frac{c}{\delta_{N+1}}. \tag{1.42}$$

But from the assumption $(iii)$, we know that there exists a polynomial $Q \in \mathbb{R}[X]$ such that

$$P(X) = Q(X)\prod_{\lambda\in\Lambda}(X-\lambda) \text{ and } \forall x \in [0,4],\ Q(x) \neq 0. \tag{1.43}$$

Hence, we deduce from (1.43) that the following partial fraction decomposition holds

$$\frac{1}{P(X)} = \frac{1}{Q(X)}\sum_{\lambda\in\Lambda}\frac{Q(\lambda)}{P'(\lambda)}\frac{1}{X-\lambda}.$$

Consequently, to prove the estimation (1.42), it is enough to prove that

$$\forall\lambda \in \Lambda, \exists c > 0, \forall N \in \mathbb{N}^*,\ \sum_{k=1}^{N}\frac{1}{|4\sin^2\left(\frac{\pi}{2}kh\right)-\lambda|} \leqslant c\frac{c}{\delta_{N+1}}. \tag{1.44}$$

To prove (1.44), it is crucial to deduce, from the conditions $(i)$ and $(iv)$, that there exists a constant $c > 0$ such that

$$\forall q \in \mathbb{N}^*,\ \delta_q \leqslant \frac{c}{q^2}. \tag{1.45}$$

Then, it is enough, to distinguish the case $\lambda = 0$ from the case $\lambda \neq 0$. On the one hand, if $\lambda = 0$, using (1.45), we have

$$\sum_{k=1}^{N}\frac{1}{4\sin^2\left(\frac{\pi}{2}kh\right)} \leqslant \sum_{k=1}^{N}\frac{1}{4\,(kh)^2} \leqslant \frac{\pi^2}{6}\frac{1}{4h^2} \leqslant \frac{\pi^2}{24}\frac{c}{\delta_{N+1}}.$$

On the other hand, $x \mapsto 4\sin^2\left(\frac{\pi}{2}x\right)$ is a diffeomorphism from $]0,1[$ to $]0,4[$. Hence, if $\lambda \neq 0$, and since we know from assumption $(ii)$ that $\lambda \neq 4$, there exists a constant $\widetilde{c} > 0$ such that one has

$$\forall x \in [0,1],\ |x-\widetilde{\lambda}| \leqslant \widetilde{c}|4\sin^2\left(\frac{\pi}{2}x\right)-\lambda|,$$

where $\widetilde{\lambda} \in ]0,1[$ is defined by

$$4\sin^2\left(\frac{\pi}{2}\widetilde{\lambda}\right) = \lambda.$$

Since $\delta$ does not have any zero index, the assumption $(iv)$ provides $\widetilde{\lambda} \notin \mathbb{Q}$. Hence, we deduce that

$$\forall q \in \mathbb{N}^*, \exists! p_q \in [\![0,q]\!],\ |\widetilde{\lambda} - \frac{p_q}{q}| < \frac{1}{2q}.$$

As a consequence, with the estimation (1.45), we have

$$\begin{aligned}\sum_{k=1}^{N} \frac{1}{|4\sin^2\left(\frac{\pi}{2}kh\right) - \lambda|} &\leqslant \sum_{k\in[\![1,N]\!]\setminus\{p_{N+1}\}} \frac{\widetilde{c}}{|kh - \widetilde{\lambda}|} + \frac{1}{|4\sin^2\left(\frac{\pi}{2}p_{N+1}h\right) - \lambda|} \\ &\leqslant \sum_{k\in[\![1,N]\!]\setminus\{p_{N+1}\}} \frac{2\widetilde{c}}{h} + \frac{1}{\delta_{N+1}} \leqslant \frac{2\widetilde{c}}{h^2} + \frac{1}{\delta_{N+1}} \leqslant \frac{2\widetilde{c}c+1}{\delta_{N+1}}.\end{aligned}$$

CHAPITRE 2

# EXISTENCE AND STABILITY OF TRAVELING WAVES FOR DISCRETE NONLINEAR SCHRÖDINGER EQUATIONS OVER LONG TIMES.

## 2.1 Introduction

### 2.1.1 Motivations and main results

We study existence and stability of solitary traveling waves for the discrete nonlinear Schrödinger equation (DNLS) on a grid $h\mathbb{Z}$ of stepsize $h > 0$ and with a cubic focusing non linearity. This equation is a differential equation on $\mathbb{C}^{h\mathbb{Z}}$ defined by (see [87] for details about its derivation)

$$\forall g \in h\mathbb{Z}, \quad i\partial_t \boldsymbol{u}_g = \frac{\boldsymbol{u}_{g+h} - 2\,\boldsymbol{u}_g + \boldsymbol{u}_{g-h}}{h^2} + |\,\boldsymbol{u}_g\,|^2\,\boldsymbol{u}_g\,. \tag{2.1}$$

We focus on this equation near its continuum limit (as $h$ goes to $0$), called nonlinear Schrödinger equation (NLS), defined as the following partial differential equation

$$\forall x \in \mathbb{R}, \quad i\partial_t u(x) = \partial_x^2 u(x) + |u(x)|^2 u(x). \tag{2.2}$$

We study solutions of DNLS (2.1) with a behavior close to the continuous traveling waves of NLS (2.2). Such solitons $u$ are global solutions of NLS with speed of oscillation $\xi_1$ and speed of advection $\xi_2$, satisfying

$$\forall t_0 \in \mathbb{R}, \quad \forall t \in \mathbb{R}, \quad \forall x \in \mathbb{R}, \quad u(t_0 + t, x) = e^{i\xi_1 t} u(t_0, x - \xi_2 t). \tag{2.3}$$

The parameter $\xi = (\xi_1, \xi_2)$ characterizes travelling waves up to gauge transform $u(x) \mapsto e^{i\gamma}u(x)$ and advection $u(x) \mapsto u(x - \mathrm{y})$. For NLS they are given explicitly by their values at time $t = 0$

$$\forall x \in \mathbb{R}, \quad \psi_\xi(x) = e^{\frac{1}{2}ix\xi_2} \frac{\sqrt{2}m_\xi}{\cosh(m_\xi x)} \quad \text{with} \quad m_\xi = \sqrt{\xi_1 - \left(\frac{\xi_2}{2}\right)^2}. \tag{2.4}$$

for speed of oscillation $\xi_1$ and speed of advection $\xi_2$ satisfying

$$\xi_1 > \left(\frac{\xi_2}{2}\right)^2. \tag{2.5}$$

This chapter is a joint work with Erwan Faou realized in [26].

On a grid, the notion of traveling wave is not as clear as on a line, and we cannot define traveling waves for DNLS as easily as those of NLS by (2.3). The difficulty comes from the definition of the advection. Indeed, the canonical advection on a grid is only defined when the distance to cross is a multiple of the stepsize $h$. Of course, we could find some reasonable extensions of (2.3) in the discrete case. For example, a possible definition of discrete traveling waves could be for solution $\boldsymbol{u}$ to DNLS to satisfy

$$\forall t_0 \in \mathbb{R}, \quad \forall n \in \mathbb{Z}, \quad \forall g \in h\mathbb{Z}, \quad \boldsymbol{u}_g(t_0 + n\tau) = e^{i\xi_1 n\tau}\, \boldsymbol{u}_{g-nh}(t_0) \quad \text{with} \quad \xi_2 \tau = h, \tag{2.6}$$

for some speeds $\xi_1, \xi_2 \in \mathbb{R}$. Even if this definition seems to be the most natural, it is not the only one possible. For example, we could replace $h$ by $2h$ in this definition or to do things even more complicated, and no canonical choice appears obvious. There is at least one class of solutions that can be defined without ambiguity, the standing waves (i.e. when $\xi_2 = 0$) which are solutions of the form

$$\forall t_0 \in \mathbb{R}, \forall t \in \mathbb{R},\ \boldsymbol{u}(t_0 + t) = e^{i\xi_1 t}\, \boldsymbol{u}(t). \tag{2.7}$$

for some speed of oscillation $\xi_1 \in \mathbb{R}$.

We define the discrete $L^2$ and $H^1$ norms as follows : for $\boldsymbol{v} \in \mathbb{C}^{h\mathbb{Z}}$,

$$\| \boldsymbol{v} \|^2_{L^2(h\mathbb{Z})} = h \sum_{g \in h\mathbb{Z}} | \boldsymbol{v}_g |^2 \quad \text{and} \quad \| \boldsymbol{v} \|^2_{H^1(h\mathbb{Z})} = h \sum_{g \in h\mathbb{Z}} \left| \frac{\boldsymbol{v}_g - \boldsymbol{v}_{g-h}}{h} \right|^2 + \| \boldsymbol{v} \|^2_{L^2(h\mathbb{Z})}.$$

Of course, these norms are equivalents but not uniformly with respect to $h$. Since we focus on the continuous limit (i.e. when $h$ goes to $0$), uniformity with respect to $h$ is crucial.

The discrete $L^2$ norm, $\| \bullet \|^2_{L^2(h\mathbb{Z})}$ is a constant of the motion of DNLS associated, through the Noether Theorem (see, for example, [58] for details about this Theorem), to invariance under gauge transform action. As $L^2(h\mathbb{Z})$ is an algebra we can deduce by the Cauchy-Lipschitz Theorem that DNLS is globally well-posed in $L^2(h\mathbb{Z})$. Moreover, DNLS is a Hamiltonian system associated with the Hamiltonian

$$\mathrm{H}_{\mathrm{DNLS}}(\boldsymbol{u}) = \frac{h}{2} \sum_{g \in h\mathbb{Z}} \left| \frac{\boldsymbol{u}_{g+h} - \boldsymbol{u}_g}{h} \right|^2 - \frac{h}{4} \sum_{g \in h\mathbb{Z}} | \boldsymbol{u}_g |^4. \tag{2.8}$$

As we can guess from its expression, this Hamiltonian is very useful to establish some estimates of coercivity with the discrete $H^1$ norm, uniformly with respect to $h$.

The continuous traveling waves of NLS defined by (2.4) verify a property of stability called orbital stability. If for a given time a solution of NLS is close enough to a traveling wave, then it remains close to this traveling wave for all times, up to advection and gauge transformations. This property has been first proven by Cazenave and Lions in 1982 in [46] by a compactness method and in 1986 by Weinstein in [115] with what we call nowadays the *energy-momentum* method. This second method is more quantitative than the first one, and the estimates of stability we give in this article are all based on it. It has been developed by Grillakis, Shatah and Strauss in 1987 in [75] and [76] (see also [58] for a very clear presentation of this method).

**Theorem 2.1.1.** *Cazenave and Lions [46], Weinstein [115]*
*For each couple of speed $\xi \in \mathbb{R}^2$, such that $\xi_1 > \left(\frac{\xi_2}{2}\right)^2$, there exists a constant $c > 0$, such that for all solutions $u$ of NLS* (2.2) *with $\|u(0) - \psi_\xi\|_{H^1(\mathbb{R})} < c$, for all time $t \in \mathbb{R}$, there exist $\mathrm{y}, \gamma \in \mathbb{R}$ such that*

$$c\|u(t) - e^{i\gamma}\psi_\xi(\bullet - \mathrm{y})\|_{H^1(\mathbb{R})} \leqslant \|u(0) - \psi_\xi\|_{H^1(\mathbb{R})}.$$

This result does not give any information on the exact position of the solution. To remedy this problem, *modulational stability* methods have been developed, which allows to follow very precisely this solution (see [114], [67] or [81]).

If we try to apply the energy-momentum method to construct orbitally stable traveling waves for DNLS, the main difficulty comes from the definition of the advection on the grid. We discuss this issue in detail in section 2.2. However this problem is easily solved when considering standing waves (i.e. $\xi_2 = 0$) with symmetric perturbations for which the solution, remaining symmetric for all times, cannot move. In $2010$, Bambusi and Penati proved in [19] the existence of standing waves of DNLS looking like those of NLS. In fact, they constructed two kinds of standing waves. Each ones are real valued and symmetric but the first ones, called *Sievers-Takeno modes* or *onsite* , are centered in $0$ whereas the second ones, called *Page modes* or *off-site,* are centered in $\frac{h}{2}$. In $2013$, in [17], Bambusi, Faou and Grébert, studying fully discrete approximation in time and space of NLS standing waves, gave some results on their orbital stability. The construction of these standing waves is also realized in a $2016$ paper of Jenkinson and Weinstein (see [84]), with another kind of approximations. If we focus only on the *onsite* standing waves, we summarized a piece of these results in the following theorem.

**Theorem 2.1.2.** *Existence and orbital stability of standing waves*
*For all $\xi_1 > 0$, there exist $h_0, C, c > 0$ such that for all $h < h_0$, there exists a unique $\boldsymbol{\phi}^h_{\xi_1} \in H^1(h\mathbb{Z};\mathbb{R})$ symmetric, centered in $0$, and $\zeta_1 \in \mathbb{R}$, such that*

- $e^{i\zeta_1 t}\,\boldsymbol{\phi}^h_{\xi_1}$ *is a solution of DNLS,*
- $|\zeta_1 - \xi_1| + \|\,\boldsymbol{\phi}^h_{\xi_1} - \psi_{(\xi_1,0)\ |h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant Ch^2$,
- $\|\,\boldsymbol{\phi}^h_{\xi_1}\,\|_{L^2(h\mathbb{Z})} = \|\psi_{(\xi_1,0)}\|_{L^2(\mathbb{R})}$,
- *If $\boldsymbol{u}$ is a solution of DNLS such that $\boldsymbol{u}(0)$ is symmetric, centered in $0$, $\|\,\boldsymbol{u}(0)\|_{L^2(h\mathbb{Z})} = \|\psi_{(\xi_1,0)}\|_{L^2(\mathbb{R})}$ and $\|\,\boldsymbol{u}(0) - \boldsymbol{\phi}^h_{\xi_1}\,\|_{H^1(h\mathbb{Z})} < c$ then for all $t \in \mathbb{R}$, there exists $\gamma \in \mathbb{R}$ such that*

$$\|\,\boldsymbol{u}(t) - e^{i\gamma}\,\boldsymbol{\phi}^h_{\xi_1}\,\|_{H^1(h\mathbb{Z})} \leqslant C\|\,\boldsymbol{u}(0) - \boldsymbol{\phi}^h_{\xi_1}\,\|_{H^1(h\mathbb{Z})}.$$

Note that the same theorem holds, for the *off site* standing waves. We just need to write "symmetric, centered in $\frac{h}{2}$" instead of "symmetric, centered in $0$" and "$\psi_{(\xi_1,0)}(\bullet - \frac{h}{2})_{|h\mathbb{Z}}$" instead of "$\psi_{(\xi_1,0)\ |h\mathbb{Z}}$".

Usually, it is enough to prove existence and orbital stability of NLS standing waves to get some orbitally stable traveling wave. Indeed, NLS is invariant by *Galilean transformation* , defined by

$$u(t,x) \mapsto e^{i\frac{v}{2}(x-vt)+i\left(\frac{v}{2}\right)^2 t}u(t, x - vt).$$

However, it seems there is no such transformation for DNLS. So we cannot apply the same strategy.

The second reason why existence of orbitally stable traveling waves for DNLS seems very uncertain is more experimental. If we assume that DNLS admits a moving traveling wave (i.e. $\xi_2 \neq 0$) that is orbitally stable and looking like a continuous traveling wave, $\psi_\xi$, then the solution of DNLS generated by the discretization of $\psi_\xi$ on $h\mathbb{Z}$, should look like $\psi_\xi$ for all times, up to an advection and a gauge transform. But there are some reasonable numerical simulations for which it is not what is observed (see [84]). In fact, the speed of this solution seems going to $0$ as $t$ goes to infinity. In the literature, this phenomenon is usually called Peierls-Nabarro barrier (see [84], [87] and [96]). A rigorous proof of this phenomenon seems to be an open problem.

However, it is really difficult to observe when $h$ is small enough (in fact, stability for exponentially long times is expected, see [96]).

Before stating our main results, let us first formulate an easy corollary of them, showing that there exists quasi-traveling waves to DNLS close to the continuum limit, for times of order $\mathcal{O}(h^{-2})$, preventing the phenomena described above to appear before this time scale.

**Theorem 2.1.3.** *For all $\varepsilon > 0$ and for all $\xi \in \mathbb{R}^2$ such that $\xi_1 > \left(\frac{\xi_2}{2}\right)^2$, there exist $h_0, C, T_0 > 0$ such that*

$$T_0 = \infty \quad \textit{when} \quad \xi_2 = 0 \qquad \textit{and} \qquad T_0 \to \infty \quad \textit{when the speed} \quad \xi_2 \to 0,$$

*and such that if $h < h_0$, $\mathrm{y}_0, \gamma_0 \in \mathbb{R}$ and $\boldsymbol{u}$ is the solution of DNLS such that*

$$\forall g \in h\mathbb{Z}, \qquad \boldsymbol{u}_g(0) = e^{i\gamma_0}\psi_\xi(g - \mathrm{y}_0),$$

*then, there exist $\gamma, \mathrm{y} \in C^1(\mathbb{R})$ satisfying $\gamma(0) = \gamma_0$ and $\mathrm{y}(0) = \mathrm{y}_0$ such that, for all $t \geqslant 0$,*

$$\forall t \leqslant T_0 h^{-2+\varepsilon}, \qquad \sup_{g \in h\mathbb{Z}} \left| \boldsymbol{u}_g(t) - e^{i\gamma(t)}\psi_\xi(g - \mathrm{y}(t)) \right| \leqslant Ch^2$$

*and*

$$\forall t \leqslant T_0 h^{-2+\varepsilon}, \qquad |\dot{\gamma}(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant Ch^2.$$

The proof of Theorem 2.1.3 is a straightforward application of Theorem 2.1.7 (or Theorem 2.1.4 if $\xi_2 = 0$). It would be possible to write the same result with the discrete $H^1$ norm instead of the $L^\infty$ norm.

To obtain this result, the strategy is to construct a function close to the continuous solitary wave $\psi_\xi$ for given parameters $\xi = (\xi_1, \xi_2)$, which define solitary waves of a modified version of DNLS essentially defined by removing the aliasing terms. This typically gives bound for time scales of order $\mathcal{O}(h^{-1})$ for orbital stability in $H^1(h\mathbb{Z})$. Moreover as the aliasing terms are small for regular functions, we can combine this analysis with a result of control of discrete Sobolev norms of DNLS to reach the time scale $\mathcal{O}(h^{-2})$. We give now the details of our results. The first one is a result of existence and stability in $H^1$ of discrete traveling waves for times of order $h^{-1}$.

**Theorem 2.1.4.** *Let $\Omega$ be a relatively compact open subset of $\left\{\xi \in \mathbb{R}^2 \mid \xi_1 > \left(\frac{\xi_2}{2}\right)^2\right\}$.*
*There exist $h_0, \kappa, r, \ell > 0$ such that for all $h < h_0$, for all $\xi \in \Omega$, there exists $\eta_\xi^h \in H^\infty(\mathbb{R})$ with*

$$\|\eta_\xi^h - \psi_\xi\|_{H^1(\mathbb{R})} \leqslant \kappa h^2, \tag{2.9}$$

*satisfying the following property. If $\boldsymbol{v} \in H^1(h\mathbb{Z})$ is an approximation of $\eta_\xi^h$ up to a gauge transform or an advection, i.e.*

$$\exists \gamma_0, \mathrm{y}_0 \in \mathbb{R}, \qquad \| \boldsymbol{v} - (e^{i\gamma_0}\eta_\xi^h(\bullet - \mathrm{y}_0))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant r,$$

*then there exist $\gamma, \mathrm{y} \in C^1(\mathbb{R})$ with $\gamma(0) = \gamma_0$ and $\mathrm{y}(0) = \mathrm{y}_0$ such that if $T > 0$ and $\boldsymbol{u}$, the solution of DNLS with $\boldsymbol{u}(0) = \boldsymbol{v}$, satisfy*

$$\forall t \in (0, T), \qquad \delta(t) := \| \boldsymbol{u}(t) - (e^{i\gamma(t)}\eta_\xi^h(\bullet - \mathrm{y}(t)))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant r, \tag{2.10}$$

*then we have for all* $t \in (0,T)$,

$$|\dot{\gamma}(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant \kappa\, (\delta(0) + \delta(t) + e^{-\frac{\ell}{h}}), \tag{2.11}$$

*and*

$$\delta(t) \leqslant \kappa\, e^{\frac{h|\xi_2|t}{\ell^2}} (\delta(0) + e^{-\frac{\ell}{h}}). \tag{2.12}$$

The functions $\eta_\xi^h$ are constructed in the third section and estimates (2.11) and (2.12) are proven in the fourth section. Now, we discuss this result. We focus on inequalities (2.11) and (2.12).

- If we remove the exponential terms, it is a result stronger than the classical inequality of orbital stability (see Theorem 2.1.1) as it includes a result of modulation.
- The exponential terms "$e^{-\frac{\ell}{h}}$" means that any discretization of $\eta_\xi^h$ is not exactly a traveling wave of DNLS.
- The time dependent exponential term means that the estimate of stability holds while $t|\xi_2|$ is smaller than $h^{-1}$. In particular, if we focus on standing waves (i.e. $\xi_2 = 0$), we get an estimate of stability for all times. Since our perturbation does not need to be symmetric, it is an extension of the previous results (see Theorem 2.1.2).
- If $\boldsymbol{u}(0)$ is a discretization of $\eta_\xi^h$ (i.e. if $\delta(0) = 0$) then the estimate of stability holds longer. Indeed, while $t|\xi_2|$ is smaller than $\frac{\ell^3}{h^2}$ (up to a multiplicative constant) , then the bootstrap (2.10) condition is satisfied. In particular, we deduce of the second inequality that at the end of this time, $\boldsymbol{u}$ completed a distance of order $\frac{\ell^3}{h^2}$, still looking like $\eta_\xi^h$.

Now, we discuss some consequences and applications of the proof of Theorem 2.1.4. These extensions are linked to the two relevant exponents of $h$ in this theorem.

First, there is a control of $\eta_\xi^h - \psi_\xi$ by $\mathcal{O}(h^2)$ (see (2.9)). This error is a consistency error. It is due to the approximation of the second derivative by a finite difference formula of order $2$. Such an estimate depends on the finite difference operator used to approximate second derivative in space. For example, if we consider the generalization of DNLS (2.1) called Discret Self-Trapping equation (DST, see [61])

$$\forall g \in h\mathbb{Z}, \quad i\partial_t\, \boldsymbol{u}_g = \frac{1}{h^2} \sum_{k\in\mathbb{Z}} a_k\, \boldsymbol{u}_{g-kh} + |\, \boldsymbol{u}_g\,|^2\, \boldsymbol{u}_g, \tag{2.13}$$

where $(a_k)_{k\in\mathbb{Z}} \in L^1(\mathbb{Z};\mathbb{R})$ is a symmetric sequence (i.e. $a_k = a_{-k}$ for all $k$), consistent of order $2n$, $n \in \mathbb{N}^*$,

$$\forall u \in H^\infty(\mathbb{R}), \quad \frac{1}{h^2} \sum_{k\in\mathbb{Z}} a_k u(hk) \underset{h\to 0}{=} \partial_x^2 u(0) + \mathcal{O}(h^{2n+2}), \tag{2.14}$$

and satisfying the estimate of stability

$$\exists \alpha > 0, \quad \forall \omega \in (0,\pi), \quad -\sum_{k\in\mathbb{Z}} a_k \cos(k\omega) \geqslant \alpha\omega^2 \tag{2.15}$$

then Theorem (2.1.4) holds for DST and we can replace (2.9) by $\|\eta_\xi^h - \psi_\xi\|_{H^1(\mathbb{R})} \leqslant \kappa h^{2n}$. In particular, this extension includes usual pseudo spectral methods and the usual high order discrete second derivatives (see [23] for details about these formulas) whose non-zero terms are given by

$$a_{\pm k} = \frac{2(-1)^{k+1}}{k^2} \frac{C_{2n}^{n-k}}{C_{2n}^{n}}, \quad \text{if} \quad 0 < k < n, \quad \text{and} \quad a_0 = -2\sum_{j=1}^{n} \frac{1}{j^2}.$$

Second, there is the right exponential term $e^{\frac{h|\xi_2|t}{\ell^2}}$ giving the stability estimates for times of order $h^{-1}$. As the error terms come mainly from aliasing effects, the control of stability for times larger than $h^{-1}$ essentially relies on a control of higher Sobolev norms for long times uniformly with respect to $h$. More precisely, we define the discrete homogeneous Sobolev norm $\| \bullet \|_{\dot{H}^n(h\mathbb{Z})}$ by

$$\| \boldsymbol{u} \|^2_{\dot{H}^n(h\mathbb{Z})} = \langle (-\Delta_h)^n \boldsymbol{u}, \boldsymbol{u} \rangle_{L^2(h\mathbb{Z})}, \quad \text{with} \quad (\Delta_h \boldsymbol{u})_g = \frac{\boldsymbol{u}_{g+h} - 2\boldsymbol{u}_g + \boldsymbol{u}_{g-h}}{h^2}, \tag{2.16}$$

and the Sobolev norm by

$$\| \boldsymbol{u} \|^2_{H^n(h\mathbb{Z})} = \sum_{k=0}^{n} \| \boldsymbol{u} \|^2_{\dot{H}^k(h\mathbb{Z})}.$$

Then we have the following version of Theorem 2.1.4 (see Remark 2.4.3 for its proof).

**Theorem 2.1.5.** *In Theorem 2.1.4, the inequality* (2.12) *can be replaced by*

$$\forall n \in \mathbb{N}^*,\ \delta(t) \leqslant \kappa \left( \delta(0) + e^{-\frac{\ell}{h}} + \sqrt{t|\xi_2|} h^{n-\frac{1}{2}} \sup_{0<s<t} \| \boldsymbol{u}(s) \|_{\dot{H}^n(h\mathbb{Z})} \right). \tag{2.17}$$

With such an estimate, we see that to obtain stability over exponentially long times, it would be enough to prove a control of the growth of the homogeneous Sobolev norm of the type $Ct^\alpha$, with $\alpha$ independent of $n$ and $h$ and $C$ independent of $h$. Note that for the continuous case, it is indeed the case for the solutions of NLS for which the $H^s$ norms are uniformly bounded in times by using integrability arguments (see for example [104]). Note that such bounds hold for linear Schrödinger equation with a smooth potential in $t$ and $x$ (see [38]).

For DNLS, it is possible to obtain polynomial control of the growth of Sobolev norms by using the *higher modified energy* method. The following result was obtained in [24] by the first author :

**Theorem 2.1.6** (Growth of discrete Sobolev norms, see [24])**.** *For all $n \in \mathbb{N}^*$, there exists $C > 0$, such that for all $h > 0$, if $\boldsymbol{u}$ is a solution of DNLS then for all $t \in \mathbb{R}$*

$$\| \boldsymbol{u}(t) \|_{\dot{H}^n(h\mathbb{Z})} \leqslant C \left[ \| \boldsymbol{u}(0) \|_{\dot{H}^n(h\mathbb{Z})} + M_{\boldsymbol{u}(0)}^{\frac{2n+1}{3}} + |t|^{\frac{n-1}{2}} M_{\boldsymbol{u}(0)}^{\frac{4n-1}{3}} \right], \tag{2.18}$$

*where*

$$M_{\boldsymbol{u}(0)} = \| \boldsymbol{u}(0) \|_{\dot{H}^1(h\mathbb{Z})} + \| \boldsymbol{u}(0) \|^3_{L^2(h\mathbb{Z})}.$$

The exponents of the $\boldsymbol{u}(0)$ norms are natural and correspond to an homogeneous estimate preserved by scaling with respect to $h$. As a corollary of Theorem 2.1.5 and Theorem 2.1.6, we get an extension of Theorem 2.1.3 for smooth perturbations of $\eta^h_\xi$. It is a result of stability for times of order $h^{-2}$ for such perturbations.

**Theorem 2.1.7.** *Let $\Omega$ be a relatively compact open subset of $\left\{ \xi \in \mathbb{R}^2 \mid \xi_1 > \left(\frac{\xi_2}{2}\right)^2 \right\}$ and $h_0, \kappa, r, \ell > 0$ be the constants given in Theorem 2.1.4.*

*For all $\varepsilon, s > 0$, there exists $n \in \mathbb{N}^*$ such that for all $\rho > 0$, there exist $C, T_0 > 0$ with*

$$T_0 = \infty \quad \textit{when} \quad \xi_2 = 0 \qquad \textit{and} \qquad T_0 \to \infty \quad \textit{when the speed} \quad \xi_2 \to 0, \tag{2.19}$$

*and* $h_1 \in (0, h_0)$*, such that for all* $h < h_1$*,* $\xi \in \Omega$ *and for all* $v \in H^n(\mathbb{R})$*, if*

$$\|v\|_{\dot{H}^n(\mathbb{R})} \leqslant \rho \quad \textit{and} \quad \|\psi_\xi - v\|_{H^1(\mathbb{R})} \leqslant \frac{r}{2(1+\kappa)}$$

*then any solution* $\boldsymbol{u}$ *of DNLS such that*

$$\exists\, \mathrm{y}_0, \gamma_0 \in \mathbb{R}, \quad \forall g \in h\mathbb{Z}, \quad \boldsymbol{u}_g(0) = e^{i\gamma_0} v(g - \mathrm{y}_0)$$

*satisfies, for all* $t \geqslant 0$*,*

$$t \leqslant T_0 h^{-2+\varepsilon} \;\Rightarrow\; \|\, \boldsymbol{u}(t) - (e^{i\gamma(t)} \eta_\xi^h(\bullet - \mathrm{y}(t)))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant C \left( \|\eta_\xi^h - v\|_{H^1(\mathbb{R})} + h^s \right) \tag{2.20}$$

*where* $\gamma, \mathrm{y} \in C^1(\mathbb{R})$ *satisfy* $\gamma(0) = \gamma_0$*,* $\mathrm{y}(0) = \mathrm{y}_0$ *and*

$$t \leqslant T_0 h^{-2+\varepsilon} \;\Rightarrow\; |\dot{\gamma}(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant C \left( \|\eta_\xi^h - v\|_{H^1(\mathbb{R})} + h^s \right). \tag{2.21}$$

This Theorem is proven in Appendix (see Section 2.5.1). Note that if we can prove a control on the growth of high Sobolev norms by $\mathcal{O}(t^{\alpha(n-1)})$ with $\alpha < \frac{1}{2}$, then we would adapt Theorem 2.1.7 to reach a stability time of order $h^{-1/\alpha+\varepsilon}$. Note that such a control of Sobolev norms holds for the continuous case with $\alpha = 0$ by using integrability results [118]. However, unlike the integrable discretization proposed in [1], DNLS is not integrable and a better control of Sobolev norms than in Theorem 2.1.6 and/or the optimality of such bounds remain open questions.

Let us emphasize that this question of regularity preservation is fundamental in the sense that our construction of discrete traveling waves is essentially of *infinite order for smooth functions*. However, with the variational techniques used to prove orbital stability, we cannot control higher Sobolev norms than the energy norm, and the time restriction is thus only driven by this question of regularity bounds over long times.

### 2.1.2 Notations

Sometimes some notations could be ambiguous, so, in this subsection, we clarify them.

- In all this paper, we consider $\mathbb{C}$ as an $\mathbb{R}$ Euclidian space of dimension $2$ equipped with the scalar product "·" defined by

  $$\forall z_1, z_2 \in \mathbb{C},\; z_1 \cdot z_2 = \Re(z_1 \overline{z_2}) = \Re z_1 \Re z_2 + \Im z_1 \Im z_2.$$

  Consequently, $L^2(\mathbb{R}; \mathbb{C})$ scalar product is defined by

  $$\forall u_1, u_2 \in L^2(\mathbb{R}; \mathbb{C}),\; \langle u_1, u_2 \rangle_{L^2(\mathbb{R})} = \int_{\mathbb{R}} u_1(x) \cdot u_2(x)\, \mathrm{dx}\,.$$

  In particular, we consider all the Fréchet differentials as $\mathbb{R}$ linear applications.
- If $u : \mathbb{R} \to \mathbb{C}$ is a real function and $h > 0$, we define the *discrete second derivative* of $u$ by

  $$\forall x \in \mathbb{R},\; \Delta_h u(x) = \frac{u(x+h) + u(x-h) - 2u(x)}{h^2}.$$

- We define the cardinal sine function on $\mathbb{R}$ by $\mathrm{sinc}(x) := x^{-1} \sin(x)$.

- As usual when we consider second derivatives, we identify the continuous bilinear forms with the operators from the space to its topological dual space. More precisely, if $E$ is a normed vector space and $b$ is a continuous bilinear form on $E$, we identify $b$ with the operator $\widetilde{b} : E \to E'$ defined by $b(x,y) = (\widetilde{b}(x))(y)$, $x, y \in E$. Consequently, it makes sense to try to invert $b$.
- If $M \in M_n(\mathbb{R})$ is a square matrix of length $n$ then $\|M\|_p$ is the matrix norm of $M$ associated to the $\ell^p$ norm on $\mathbb{R}^n$. Similarly, if $\xi \in \mathbb{R}^2$, $|\xi| := \sqrt{\xi_1^2 + \xi_2^2}$ is the $\ell^2$ norm of $\xi$.
- If $\mathscr{E}$ is a set then $\mathbb{1}_{\mathscr{E}}$ is the characteristic function of $\mathscr{E}$.

**Acknoledgements**

The authors are glad to thank Dario Bambusi, Benoît Grébert and Alberto Maspero for their helpful comments and discussions during the preparation of this work.

## 2.2 Aliasing generating inhomogeneity

In this section, we explain why DNLS can be interpreted as an inhomogeneous equation on $\mathbb{R}$ and why we cannot apply directly the *energy-momentum* method to get stable traveling waves. This section is also an introduction to most of the tools used in the this paper.

The *energy-momentum method* is a way to construct orbitally stable equilibria of a Hamiltonian system, relatively to a Lie group action. It has been used by Weinstein in [115] to prove the orbital stability of the traveling waves of NLS. Then it has been developed, in the general context of Hamiltonian systems by Grillakis, Shatah, Strauss in [75],[76]. A clear and rigorous presentation of the method and its formalism in a general setting is given in the paper [58] by De Bièvre, Genoud, and Rota Nodari.

A crucial part of this method is based on Noether Theorem, requiring to identify invariant Lie group actions with Hamiltonian flows. DNLS is invariant under two group actions : the gauge transform $\boldsymbol{u} \mapsto e^{i\gamma}\boldsymbol{u}$ and discrete advection $\boldsymbol{u} \mapsto (\boldsymbol{u}_{g+a})_{g\in h\mathbb{Z}}$. The gauge transform is clearly the flow of the Hamiltonian $\|\boldsymbol{u}\|^2_{L^2(h\mathbb{Z})}$ but the discrete advection is only defined for a countable set of values $a \in h\mathbb{Z}$ and cannot naturally be associated with a Hamiltonian.

First, we need to extend the advection for any values $a \in \mathbb{R}$ and then try to identify this extension with the flow of a Hamiltonian. Then we are going to see that the Hamiltonian of DNLS (see (2.8)) is not invariant by this advection, and that the error is driven by aliasing terms.

### 2.2.1 Shannon's advection

There are natural ways to define an advection, denoted by $\tau_a$, on the grid $h\mathbb{Z}$. For a given interpolation operator $\mathcal{I}_h : L^2(h\mathbb{Z}) \to L^2(\mathbb{R})$ we can carry the advection on $\mathbb{R}$ to the grid $h\mathbb{Z}$ by

making the following diagram commute

$$\begin{array}{ccc} L^2(h\mathbb{Z}) & \xrightarrow{\tau_a} & L^2(h\mathbb{Z}) \\ \mathcal{I}_h \big\downarrow & & \mathcal{I}_h \big\downarrow \\ L^2(\mathbb{R}) & \xrightarrow{u \mapsto u(\bullet - a)} & L^2(\mathbb{R}) \end{array} \tag{2.22}$$

In general, this construction does not work, as the advection of an interpolation is not necessarily an interpolation (see, for example with a finite element interpolation). However, there exists a classical interpolation called *Shannon interpolation* for which this construction can be applied. Let us define the *discrete Fourier transform* $\mathcal{F}_h$ and Fourier Plancherel transform $\mathscr{F}$

$$\mathcal{F}_h : \left\{ \begin{array}{lcl} L^2(h\mathbb{Z}) & \to & L^2(\mathbb{R}/\frac{2\pi}{h}\mathbb{Z}) \\ \boldsymbol{u} & \mapsto & \omega \mapsto h \sum_{g \in h\mathbb{Z}} \boldsymbol{u}_g \, e^{ig\omega} \end{array} \right. \quad \text{and} \quad \mathscr{F} : \left\{ \begin{array}{lcl} L^2(\mathbb{R}) & \to & L^2(\mathbb{R}) \\ u & \mapsto & \omega \mapsto \int_{\mathbb{R}} u(x) e^{ix\omega} \, \mathrm{dx} \end{array} \right. \tag{2.23}$$

where the last integral is defined by extending the operator defined on $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. We also use the notation $\widehat{u} = \mathscr{F} u$. The *Shannon interpolation* , denoted by $\mathcal{I}_h$, is defined through the following diagram

$$\begin{tikzcd} L^2(h\mathbb{Z}) \arrow[r, "\mathcal{F}_h"] \arrow[rrr, bend right, "\mathcal{I}_h"'] & L^2(\mathbb{R}/\frac{2\pi}{h}\mathbb{Z}) \arrow[r, "u \mapsto \mathbb{1}_{(-\frac{\pi}{h}, \frac{\pi}{h})} u"] & L^2(\mathbb{R}) \arrow[r, "\mathscr{F}^{-1}"] & L^2(\mathbb{R}) \end{tikzcd} . \tag{2.24}$$

With this construction, this interpolation clearly enjoys some useful properties.

**Proposition 2.2.1.** *$\mathcal{I}_h$ is an isometry between $L^2(h\mathbb{Z})$ and its image in $L^2(\mathbb{R})$. This image is denoted $BL^2_h$. It is the subspace of $L^2(\mathbb{R})$ whose Fourier transform support is a subset of $[-\frac{\pi}{h}, \frac{\pi}{h}]$, i.e.*

$$BL^2_h = \left\{ u \in L^2(\mathbb{R}) \mid \operatorname{Supp} \widehat{u} \subset [-\frac{\pi}{h}, \frac{\pi}{h}] \right\}.$$

*Moreover, the Shannon advection $\tau_a$* is well defined through (2.22).

*Proof.* We just need to verify that the advection of a Shannon interpolation is an interpolation. So let $u \in BL^2_h$. Since we have

$$\forall \omega \in \mathbb{R}, \qquad \widehat{u(\bullet - a)}(\omega) = e^{-i\omega a} \widehat{u}(\omega),$$

it is clear that $\operatorname{Supp} \widehat{u(\bullet - a)} = \operatorname{Supp} \widehat{u}$. Consequently, we have proven that $u(\bullet - a) \in BL^2_h$. □

Since the Fourier transform support of a Shannon interpolations is bounded, $BL^h_2$ functions are very regular functions (they are entire function). Consequently, when we deal with $BL^h_2$ functions we will not justify the algebraic calculations.

Now, we check that this advection is generated by a Hamiltonian flow. Introducing some formalism, since the Shannon interpolation is a $\mathbb{C}$ linear isometry, we prove in the following Lemma that it is a symplectomorphism between $\left(L^2(h\mathbb{Z}; \mathbb{C}), \langle i \ \bullet, \bullet \rangle_{L^2(h\mathbb{Z}; \mathbb{C})}\right)$ and $\left(BL^2_h, \langle i \ \bullet, \bullet \rangle_{L^2(\mathbb{R}; \mathbb{C})}\right)$ preserving the Hamiltonian structure.

**Lemma 2.2.1.** *Let $I$ be an open subset of $\mathbb{R}$, $\boldsymbol{u} \in C^1(I; L^2(h\mathbb{Z}; \mathbb{C}))$ and $H \in C^1(L^2(h\mathbb{Z}; \mathbb{C}); \mathbb{R})$. Defining $u = \mathcal{I}_h \boldsymbol{u}$, the following properties are equivalents*

$$\forall t \in I, \quad \forall \boldsymbol{v} \in L^2(h\mathbb{Z}; \mathbb{C}), \quad \langle i\partial_t \boldsymbol{u}(t), \boldsymbol{v}\rangle_{L^2(h\mathbb{Z})} = \mathrm{d}\, H(\boldsymbol{u}(t))(\boldsymbol{v}), \tag{2.25}$$

*and*

$$\forall t \in I, \quad \forall v \in BL^2_h, \quad \langle i\partial_t u(t), v\rangle_{L^2(\mathbb{R})} = \mathrm{d}(H \circ \mathcal{I}_h^{-1})(u(t))(v). \tag{2.26}$$

*Proof.* Assume (2.25) and $v \in BL^2_h$. Since $\mathcal{I}_h$ is bijective, there exists $\boldsymbol{v} \in L^2(h\mathbb{Z}; \mathbb{C})$ such that $v = \mathcal{I}_h \boldsymbol{v}$. So we have

$$\mathrm{d}(H \circ \mathcal{I}_h^{-1})(u(t))(v) = \mathrm{d}(H \circ \mathcal{I}_h^{-1})(u(t))(\mathcal{I}_h \boldsymbol{v}) = \mathrm{d}\, H(\boldsymbol{u}(t))(\boldsymbol{v}) = \langle i\partial_t \boldsymbol{u}(t), \boldsymbol{v}\rangle_{L^2(h\mathbb{Z})}.$$

However, we have

$$\langle i\partial_t u(t), v\rangle_{L^2(\mathbb{R})} = \langle \mathcal{I}_h^* i\mathcal{I}_h \partial_t \boldsymbol{u}(t), \boldsymbol{v}\rangle_{L^2(h\mathbb{Z})},$$

where $\mathcal{I}_h^*$ is the adjoint operator of $\mathcal{I}_h$. But $\mathcal{I}_h$ is $\mathbb{C}$ linear so we have

$$i\mathcal{I}_h \partial_t \boldsymbol{u}(t) = \mathcal{I}_h i\partial_t \boldsymbol{u}(t).$$

Furthermore, it is an isometry so we have $\mathcal{I}_h^* = \mathcal{I}_h^{-1}$. Consequently, we get

$$\langle i\partial_t u(t), v\rangle_{L^2(\mathbb{R})} = \langle i\partial_t \boldsymbol{u}(t), \boldsymbol{v}\rangle_{L^2(h\mathbb{Z})}.$$

So we have proven (2.26). Conversely, we can prove that (2.25) is a consequence of (2.26) using the same equalities. □

Applying Lemma 2.2.1 to the identify Shannon advection with a Hamiltonian flow, we just need to identify the canonical advection on $BL^2_h$.

**Lemma 2.2.2.** *Let $\mathcal{M} : BL^2_h \to \mathbb{R}$ be the momentum defined by*

$$\forall u \in BL^2_h, \quad \mathcal{M}(u) = \langle i\partial_x u, u\rangle_{L^2(\mathbb{R})}.$$

*If $u \in C^1(\mathbb{R}; BL^2_h)$ then the following properties are equivalent*

$$\forall t \in \mathbb{R}, \quad u(t, x) = u(0, x + 2t), \tag{2.27}$$

*and*

$$\forall t \in \mathbb{R}, \quad \forall v \in BL^2_h, \quad \langle i\partial_t u(t), v\rangle_{L^2(R)} = \mathrm{d}\, \mathcal{M}(u(t))(v). \tag{2.28}$$

*Proof.* Assume (2.28) and let $t \in \mathbb{R}$, $v \in BL^2_h$. We have

$$\langle \partial_t u(t), v\rangle_{L^2(R)} = \langle i\partial_t u(t), iv\rangle_{L^2(R)} = \mathrm{d}\, \mathcal{M}(u(t))(iv) = 2\langle i\partial_x u(t), iv\rangle_{L^2(R)} = 2\langle \partial_x u(t), v\rangle_{L^2(R)}.$$

So since $(BL^2_h, \| \bullet \|_{L^2(\mathbb{R})})$ is a Hilbert space, we have

$$\forall t, x \in \mathbb{R}, \quad \partial_t u(t, x) = 2\partial_x u(t, x).$$

Consequently, we have $u(t, x) = u(0, x + 2t)$. The converse is obvious. □

Applying Lemma 2.2.1 and Lemma 2.2.2, we deduce that Shannon's advection is the flow of the Hamiltonian $-\frac{1}{2}\mathcal{M} \circ I_h^{-1}$.

### 2.2.2 The aliasing error

In this subsection, we show that the DNLS Hamiltonian is not invariant by Shannon's advection. We recall some classical properties of Shannon's interpolation, see for example [98] for more details.

**Proposition 2.2.2.** *If $\boldsymbol{u} \in L^2(h\mathbb{Z})$ then $\mathcal{I}_h \boldsymbol{u}_{|\ h\mathbb{Z}} = \boldsymbol{u}$.*

This proposition is just a corollary of the following decomposition, where the series converges in $L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$,

$$\forall x \in \mathbb{R}, \quad \mathcal{I}_h \boldsymbol{u}(x) = \sum_{g \in h\mathbb{Z}} \boldsymbol{u}_g \operatorname{sinc}(\pi \frac{x-g}{h}).$$

**Corollary 2.2.1.** *The Shannon interpolation of $\boldsymbol{u}$ is the only function in $L^2(\mathbb{R})$ with Fourier transform support included in $[-\frac{\pi}{h}, \frac{\pi}{h}]$ and whose values on $h\mathbb{Z}$ are those of $\boldsymbol{u}$.*

Now, we detail a classical property of Shannon interpolation that is crucial in this paper.

**Proposition 2.2.3.** *If $u \in H^1(\mathbb{R})$ then $\boldsymbol{u} := u_{|\ h\mathbb{Z}} \in L^2(h\mathbb{Z})$ and for all $\omega \in (-\frac{\pi}{h}, \frac{\pi}{h})$ we have*

$$\widehat{\mathcal{I}_h \boldsymbol{u}}(\omega) = \sum_{k \in \mathbb{Z}} \widehat{u}(\omega + \frac{2\pi}{h}k). \tag{2.29}$$

*Proof.* First observe that the series (2.29) converges in $L^2(-\frac{\pi}{h}, \frac{\pi}{h})$. Indeed, using Cauchy Schwarz inequality, we have

$$\begin{aligned}\sum_{k \in \mathbb{Z}\setminus\{0\}} \|\widehat{u}(\omega + \frac{2\pi}{h}k)\|_{L^2(-\frac{\pi}{h},\frac{\pi}{h})} &\leqslant \sum_{k \in \mathbb{Z}\setminus\{0\}} \|\widehat{\partial_x u}(\omega + \frac{2\pi}{h}k)\|_{L^2(-\frac{\pi}{h},\frac{\pi}{h})} \frac{h}{|2k-1|\pi} \\ &\leqslant \sqrt{2\pi}\|\partial_x u\|_{L^2(\mathbb{R})} \sqrt{\sum_{k \in \mathbb{Z}\setminus\{0\}} \frac{h^2}{(2k-1)^2\pi^2}}.\end{aligned}$$

Now define $v \in BL^2_h$ through its Fourier transform

$$\widehat{v}(\omega) = \mathbb{1}_{(-\frac{\pi}{h},\frac{\pi}{h})} \sum_{k \in \mathbb{Z}} \widehat{u}(\omega + \frac{2\pi}{h}k).$$

If we prove that the values of $v$ on $h\mathbb{Z}$ are the same as the values of $u$ then we conclude the proof with Corollary 2.2.1. Using inverse Fourier transform formula and continuity of Fourier Plancherel transform, we get for $j \in \mathbb{Z}$,

$$\begin{aligned}v(hj) &= \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{v}(\omega)e^{-i\omega hj}\mathrm{d}\omega = \frac{1}{2\pi}\sum_{k \in \mathbb{Z}} \int_{-\frac{\pi}{h}-\frac{2\pi}{h}k}^{\frac{\pi}{h}-\frac{2\pi}{h}k} \widehat{u}(\omega)e^{-i(\omega-\frac{2\pi}{h}k)hj}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\sum_{k \in \mathbb{Z}} \int_{-\frac{\pi}{h}-\frac{2\pi}{h}k}^{\frac{\pi}{h}-\frac{2\pi}{h}k} \widehat{u}(\omega)e^{-i\omega hj}\mathrm{d}\omega = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{u}(\omega)e^{-i\omega hj}\mathrm{d}\omega = u(hj).\end{aligned}$$

□

We now express the DNLS Hamiltonian in terms of Shannon's interpolation :

**Lemma 2.2.3.** *For all $\boldsymbol{u} \in L^2(h\mathbb{Z})$, let $u = \mathcal{I}_h \boldsymbol{u}$, then we have*

$$\mathrm{H}_{\mathrm{DNLS}}(\boldsymbol{u}) = \frac{1}{2}\int_{\mathbb{R}} \left|\frac{u(x+h)-u(x)}{h}\right|^2 \mathrm{dx} - \frac{1}{4}\int_{\mathbb{R}} \left(1 + 2\cos(\frac{2\pi x}{h})\right) |u(x)|^4 \,\mathrm{dx}\,. \tag{2.30}$$

*Proof.* Since the Shannon interpolation $\mathcal{I}_h$ is an isometry between $L^2(h\mathbb{Z};\mathbb{C})$ and $L^2(\mathbb{R};\mathbb{C})$, we have

$$h\sum_{g\in h\mathbb{Z}} \left|\frac{\boldsymbol{u}_{g+h}-\boldsymbol{u}_g}{h}\right|^2 = \int_{\mathbb{R}} \left|\frac{u(x+h)-u(x)}{h}\right|^2 \mathrm{dx}\,.$$

Now we calculate the nonlinear part. First, we use the same argument of isometry to prove that

$$h\sum_{g\in h\mathbb{Z}} |\boldsymbol{u}_g|^4 = \langle \boldsymbol{u}, |\,\boldsymbol{u}\,|^2\,\boldsymbol{u}\rangle_{L^2(h\mathbb{Z})} = \langle u, \mathcal{I}_h(|\,\boldsymbol{u}\,|^2\,\boldsymbol{u})\rangle_{L^2(\mathbb{R})}. \tag{2.31}$$

But we deduce from Proposition 2.2.3 that for $\omega \in \mathbb{R}$

$$\mathscr{F}\mathcal{I}_h(|\,\boldsymbol{u}\,|^2\,\boldsymbol{u})(\omega) = \mathbb{1}_{(-\frac{\pi}{h},\frac{\pi}{h})}(\omega)\sum_{k\in\mathbb{Z}} \widehat{|u|^2 u}(\omega + \frac{2\pi}{h}k).$$

However, since $u \in BL^2_h$, we have

$$\operatorname{supp}\widehat{|u|^2 u} \subset \operatorname{supp}\widehat{u} + \operatorname{supp}\widehat{u} + \operatorname{supp}\widehat{\bar{u}} \subset \left[-\frac{3\pi}{h}, \frac{3\pi}{h}\right].$$

Consequently, if $k \notin \{-1,0,1\}$ the term in the sum is zero. Furthermore, it is clear that for any $v \in L^2(\mathbb{R})$, $\gamma \in \mathbb{R}$, $\widehat{v}(\bullet + \gamma) = \widehat{e^{i\gamma x} v}$. So we have

$$\mathscr{F}\mathcal{I}_h(|\,\boldsymbol{u}\,|^2\,\boldsymbol{u})(\omega) = \mathbb{1}_{(-\frac{\pi}{h},\frac{\pi}{h})}(\omega)\mathscr{F}\left[\left(1 + 2\cos(\frac{2\pi x}{h})\right)|u|^2 u\right](\omega).$$

We conclude by plugging this relation in (2.31). □

We this Lemma 2.2.3, we can observe that $\mathrm{H}_{\mathrm{DNLS}}$ is not invariant by advection. This default of invariance is due to an inhomogeneity generated by aliasing errors.

### 2.2.3 The flow of DNLS in the space of the Shannon interpolations

Thanks to Shannon's interpolation, we identify functions defined on a grid with functions of $BL^h_2$. Now, we are going to see that it is equivalent to consider the flow of DNLS on a grid, or to consider the Hamiltonian flow on $BL^h_2$ associated with the Hamiltonian $H^h_{\mathrm{DNLS}}$ defined by

$$\forall u \in BL^2_h,\ H^h_{\mathrm{DNLS}}(u) := \frac{1}{2}\int_{\mathbb{R}} \left|\frac{u(x+h)-u(x)}{h}\right|^2 \mathrm{dx} - \frac{1}{4}\int_{\mathbb{R}} \left(1 + 2\cos(\frac{2\pi x}{h})\right) |u(x)|^4 \,\mathrm{dx}\,. \tag{2.32}$$

Applying Lemma 2.2.1, we obtain :

**Lemma 2.2.4.** *Let $h > 0$, $\boldsymbol{u} \in C^1(\mathbb{R}; L^2(\mathbb{R}))$ and $u = \mathcal{I}_h(\boldsymbol{u})$. Then $\boldsymbol{u}$ is a solution of DNLS (see* (2.1)*) if and only if*

$$\forall t \in \mathbb{R}, \quad \forall v \in BL^2_h, \quad \langle i\partial_t u(t), v\rangle_{L^2(\mathbb{R})} = \mathrm{d}\, H^h_{\mathrm{DNLS}}(u(t))(v).$$

We conclude with the following result showing that discrete Sobolev norms are equivalent to continuous Sobolev norms on $BL^h_2$ :

**Lemma 2.2.5.** *Let $\boldsymbol{u} \in L^2(h\mathbb{Z})$ and $u = \mathcal{I}_h\, \boldsymbol{u} \in BL^2_h$. Then we have*

$$\frac{2}{\pi}\|u\|_{H^1(\mathbb{R})} \leqslant \|\, \boldsymbol{u}\, \|_{H^1(h\mathbb{Z})} \leqslant \|u\|_{H^1(\mathbb{R})}.$$

*Proof.* By construction, we know that $\|\, \boldsymbol{u}\, \|_{L^2(h\mathbb{Z})} = \|u\|_{L^2(\mathbb{R})}$. So we just need to focus on the other part of the $H^1(h\mathbb{Z})$ norm. Indeed, applying the Shannon isometry and the Fourier Plancherel isometry, we have

$$\begin{aligned}\|\, \boldsymbol{u}\, \|^2_{\dot{H}^1(h\mathbb{Z})} = \sum_{g\in h\mathbb{Z}} \left|\frac{\boldsymbol{u}_{g+h} - \boldsymbol{u}_g}{h}\right|^2 = \int_{\mathbb{R}} \left|\frac{u(x+h) - u(x)}{h}\right|^2 \mathrm{dx} &= \frac{1}{2\pi}\int_{\mathbb{R}} \frac{4}{h^2}\sin^2\left(\frac{\omega h}{2}\right)|\widehat{u}(\omega)|^2\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \mathrm{sinc}^2\left(\frac{\omega h}{2}\right)|\widehat{\partial_x u}(\omega)|^2\,\mathrm{d}\omega\,.\end{aligned}$$

Thus, the result follows from the bound

$$\mathrm{sinc}^2\left(\frac{\pi}{2}\right) \leqslant \mathrm{sinc}^2\left(\frac{\omega h}{2}\right) \leqslant 1.$$

□

Similarly, we can prove that for high order homogeneous Sobolev norms (see (2.16)), we have for all $\boldsymbol{u} \in L^2(h\mathbb{Z}; \mathbb{C})$ and $u = \mathcal{I}_h(\boldsymbol{u})$,

$$\left(\frac{2}{\pi}\right)^n \|u\|_{\dot{H}^n(\mathbb{R})} \leqslant \|\, \boldsymbol{u}\, \|_{\dot{H}^n(h\mathbb{Z})} \leqslant \|u\|_{\dot{H}^n(\mathbb{R})}. \tag{2.33}$$

## 2.3 Traveling waves of the homogeneous Hamiltonian

In the previous subsection, we have seen that the Hamiltonian of DNLS is not invariant by Shannon's advection. This default of invariance is due to an inhomogeneity generated by an aliasing error (the highly oscillatory terms in (2.32)), preventing a faire use of energy-momentum method to get stable traveling waves. Let us introduce the following perturbation of the DNLS Hamiltonian, obtained by removing these aliasing terms :

$$\forall u \in BL^2_h, \quad H_h(u) = \frac{1}{2}\int_{\mathbb{R}} \left|\frac{u(x+h) - u(x)}{h}\right|^2 \mathrm{dx} - \frac{1}{4}\|u\|^4_{L^4(\mathbb{R})}. \tag{2.34}$$

This new Hamiltonian is clearly invariant by gauge transform and advection, and we will be able to apply the energy-momentum method. Moreover, for smooth functions, it is very close to the DNLS Hamiltonian.

In the first subsection, we construct, with a perturbative method, critical points of Lagrange functions associated with (2.34). These critical points are the functions $\eta_\xi^h$ of Theorem 2.1.4. They are traveling waves for the dynamic associated to this homogeneous Hamiltonian. In the second subsection, we focus on their regularity and their orbital stability.

In all this section, we only consider speeds $\xi$ in $\Omega$, a relatively compact open subset of $\left\{\xi \in \mathbb{R}^2 \mid \xi_1 > \left(\frac{\xi_2}{2}\right)^2\right\}$.

### 2.3.1 Construction of the traveling waves

Let us introduce the Lagrange function $\mathscr{L}_\xi^h : BL_h^2 \to \mathbb{R}$ defined by

$$\forall u \in BL_h^2, \quad \mathscr{L}_\xi^h(u) = H_h(u) + \frac{\xi_1}{2}\|u\|_{L^2(\mathbb{R})}^2 + \frac{\xi_2}{2}\langle i\partial_x u, u\rangle_{L^2(\mathbb{R})}. \tag{2.35}$$

We prove in the following lemma that the traveling waves generated by $H_h$ are the critical points of $\mathscr{L}_\xi^h$.

**Proposition 2.3.1.** *Let $\xi \in \mathbb{R}^2$, $h > 0$ and $u \in C^1(\mathbb{R}; BL_h^2)$ be such that*

$$\forall t \in \mathbb{R}, \quad \forall x \in \mathbb{R}, \quad u(t,x) = e^{i\xi_1 t}u(0, x - \xi_2 t).$$

*Then the following properties are equivalents*

$$\forall t \in \mathbb{R}, \quad \forall v \in BL_h^2, \quad \langle i\partial_t u(t), v\rangle_{L^2(\mathbb{R})} = \mathrm{d}\, H_h(u(t))(v), \tag{2.36}$$

*and*

$$\mathrm{d}\,\mathscr{L}_\xi^h(u(0)) = 0. \tag{2.37}$$

*Proof.* By a straightforward calculation, we have, for all $t, x \in \mathbb{R}$,

$$\partial_t u(t,x) = i\xi_1 u(t,x) - \xi_2 \partial_x u(t,x).$$

Consequently, testing this relation against $v \in BL_h^2$, we get for all $t, x \in \mathbb{R}$,

$$\langle i\partial_t u(t), v\rangle_{L^2(\mathbb{R})} = -\,\mathrm{d}\left(\frac{\xi_1}{2}\|\bullet\|_{L^2(\mathbb{R})}^2 + \frac{\xi_2}{2}\langle i\partial_x \bullet, \bullet\rangle_{L^2(\mathbb{R})}\right)(u(t))(v).$$

So (2.36) is clearly equivalent to

$$\forall t \in \mathbb{R},\ \mathrm{d}\,\mathscr{L}_\xi^h(u(t)) = 0. \tag{2.38}$$

In particular (2.36) $\Rightarrow$ (2.37) is obvious.

Conversely, to prove (2.37) $\Rightarrow$ (2.36), we just need to prove that if $u_0 \in BL_h^2$ is a critical point of $\mathscr{L}_\xi^h$ and $\gamma, \mathrm{y} \in \mathbb{R}$ then $e^{i\gamma}u_0(\bullet - \mathrm{y})$ is also a critical point of $\mathscr{L}_\xi^h$. Define $T_{\gamma,\mathrm{y}} : BL_h^2 \to BL_h^2$ by

$$\forall v \in BL_h^2, \quad T_{\gamma,\mathrm{y}} v = e^{i\gamma} v(\bullet - \mathrm{y}).$$

Since $\mathscr{L}_\xi^h$ is invariant by gauge transform and advection, we have

$$\forall v \in BL_h^2, \quad \mathscr{L}_\xi^h(T_{\gamma,\mathrm{y}} v) = \mathscr{L}_\xi^h(v).$$

Calculating the derivative with respect to $v$ in $u_0$, we get

$$\forall v \in BL^2_h, \quad \mathrm{d}\,\mathscr{L}^h_\xi(T_{\gamma,\mathrm{y}}u_0)(T_{\gamma,\mathrm{y}}v) = \mathrm{d}\,\mathscr{L}^h_\xi(u_0)(v) = 0.$$

Since $T_{\gamma,\mathrm{y}}$ is an invertible operator on $BL^2_h$ (because $T^{-1}_{\gamma,\mathrm{y}} = T_{-\gamma,-\mathrm{y}}$), $T_{\gamma,\mathrm{y}}u_0$ is also a critical point of $\mathscr{L}^h_\xi$. □

In the following Theorem, we construct critical points of the Lagrange functions $\mathscr{L}^h_\xi$ as perturbations of the continuous traveling waves $\psi_\xi$ embedded in $BL^2_h$.

**Theorem 2.3.1.** *There exist $h_0, C, \rho, \alpha > 0$ such that for all $h < h_0$ and for all $\xi \in \Omega$, there exists $\eta^h_\xi \in BL^2_h$ satisfying*

*a)* $\mathrm{d}\,\mathscr{L}^h_\xi(\eta^h_\xi) = 0$,

*b)* $\|\eta^h_\xi - \psi_\xi\|_{H^1(\mathbb{R})} \leqslant Ch^2$,

*c)* $\forall x \in \mathbb{R},\ \overline{\eta^h_\xi}(-x) = \eta^h_\xi(x)$,

*d) if $u \in BL^2_h$ is such that $\|u - \eta^h_\xi\|_{H^1(\mathbb{R})} < \rho$, $\overline{u}(-x) = u(x)$ for all $x \in \mathbb{R}$ and $\mathrm{d}\,\mathscr{L}^h_\xi(u) = 0$ then $u = \eta^h_\xi$,*

*e) if $v \in BL^2_h \cap \mathrm{Span}(\eta^h_\xi, i\eta^h_\xi, \partial_x\eta^h_\xi)^{\perp_{L^2}}$, then we have*

$$\mathrm{d}^2\,\mathscr{L}^h_\xi(\eta^h_\xi)(v,v) \geqslant \alpha\|v\|^2_{H^1(\mathbb{R})}.$$

*Furthermore, $\xi \mapsto \eta^h_\xi$ is $C^1$ and for all $h < h_0$, for all $\xi \in \Omega$, we have*

$$\forall \zeta \in \mathbb{R}^2, \quad \|\,\mathrm{d}_\xi\,\eta^h_\xi(\xi)(\zeta) - \mathrm{d}_\xi\,\psi_\xi(\xi)(\zeta)\|_{H^1(\mathbb{R})} \leqslant C|\zeta|h^2.$$

The remainder of this section is devoted to the proof of this Theorem. It is divided in three steps. The idea of the proof is to apply, for each value of $\xi$, the inverse function Theorem to solve $\mathrm{d}\,\mathscr{L}^h_\xi(u) = 0$. We give an adapted version of this result, see Theorem 2.5.3, proven in Appendix. Moreover, we have to pay attention to symmetries and establish estimates uniform with respect to $\xi \in \Omega$ and $h$ small enough.

Step 1 : Identify the function to invert

First, we need a point around which apply the inverse function Theorem. To do this, we consider the orthogonal projection of the continuous traveling wave $\psi_\xi$ on $BL^2_h$ (for the $L^2(\mathbb{R})$ norm) denoted by $\psi^h_\xi$. Using Fourier Plancherel transform, we observe that $\psi^h_\xi$ and $\psi_\xi$ are linked by their Fourier transform through the relation

$$\widehat{\psi^h_\xi} = \mathbb{1}_{(-\frac{\pi}{h},\frac{\pi}{h})}\widehat{\psi_\xi}. \tag{2.39}$$

Sometimes it is useful to extend this notation for $h = 0$ with $\psi^0_\xi = \psi_\xi$.

Now, we have to take care about the symmetries of the problem. Indeed, since the set of the critical points of $\mathscr{L}^h_\xi$ is stable under advection and gauge transform, we expect that the differential of $\mathrm{d}\,\mathscr{L}^h_\xi$ is not invertible in this critical point. However, there is a classical trick to avoid the problem generated by these symmetries. To explain this trick we need to introduce an operator on $BL^2_h$

$$S_h : \left\{ \begin{array}{ccc} BL^2_h & \to & BL^2_h \\ u & \mapsto & (x \mapsto \overline{u(-x)}). \end{array} \right.$$

This symmetry is natural for our problem because $\mathscr{L}^h_\xi$ is invariant under its action.

**Lemma 2.3.2.** *For all $h > 0$, for all $\xi \in \mathbb{R}^2$, for all $u \in BL^2_h$, we have*

$$\mathscr{L}^h_\xi(S_h(u)) = \mathscr{L}^h_\xi(u).$$

*Proof.* It can be proven by a straightforward calculation. □

This operator induces a decomposition of $BL^2_h$ very well adapted to our problem

$$BL^2_h = \mathrm{Ker}(\mathrm{id} - S_h) \oplus \mathrm{Ker}(\mathrm{id} + S_h).$$

This decomposition is also a topological decomposition because these subspaces are closed for the $\| \bullet \|_{H^1(\mathbb{R})}$ norm. In all the paper, these spaces are always implicitly equipped with this norm.

The continuous traveling waves are invariant under this symmetry. Indeed, we can verify (see (2.4)) that

$$\forall x \in \mathbb{R}, \quad \overline{\psi_\xi(-x)} = \psi_\xi(x).$$

Consequently, we expect $\eta^h_\xi$ to be invariant under the action of $S_h$. The space $\mathrm{Ker}(\mathrm{id} - S_h)$ is not invariant under advection or gauge transform, so we avoid the previous difficulty. Moreover, we have the following result

**Lemma 2.3.3.** *For all $h > 0$, for all $\xi \in \mathbb{R}^2$, for all $u \in \mathrm{Ker}(\mathrm{id} - S_h)$, for all $v \in \mathrm{Ker}(\mathrm{id} + S_h)$, we have*

$$\mathrm{d}\,\mathscr{L}^h_\xi(u)(v) = 0.$$

*Proof.* Applying Lemma 2.3.2, we get

$$\mathscr{L}^h_\xi(u - v) = \mathscr{L}^h_\xi(u + v).$$

Then, if we compute the derivative with respect to $v \in \mathrm{Ker}(\mathrm{id} + S_h)$, we get

$$\mathrm{d}\,\mathscr{L}^h_\xi(u)(v) = -\,\mathrm{d}\,\mathscr{L}^h_\xi(u)(v).$$

□

With this lemma, we see that a critical point of $\mathrm{d}\,\mathscr{L}^h_{\xi\,|\mathrm{Ker}(\mathrm{id} - S_h)}$ is a critical point of $\mathscr{L}^h_\xi$. Hence we will apply the inverse function Theorem 2.5.3 in the point $\psi^h_\xi$ which is in $\mathrm{Ker}(\mathrm{id} - S_h)$ (it is a straightforward calculation), and to the function $\mathrm{d}\,\mathscr{L}^h_{\xi\,|\mathrm{Ker}(\mathrm{id} - S_h)}$.

Step 2 : Invertibility of the derivative

Now, we want to prove that $\mathrm{d}^2\,\mathscr{L}^h_{\xi\,|\mathrm{Ker}(\mathrm{id} - S_h)}(\psi^h_\xi)$ is invertible and to estimate the norm of its inverse uniformly with respect to $\xi \in \Omega$ and $h$ small enough. The strategy of the proof is to establish that $\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)$ is negative in the direction of $\psi^h_\xi$ and positive in the direction $L^2$-orthogonal to $\psi^h_\xi$ in $\mathrm{Ker}(\mathrm{id} - S_h)$. Then it will be possible to conclude using a classical lemma of functional analysis (see Lemma 2.5.7 ).

We are going to establish most of our estimates from the continuum limit. So we need to introduce the continuous Lagrange functions associated to NLS, defined on $H^1(\mathbb{R})$ by

$$\mathscr{L}_\xi(u) = \frac{1}{2}\|\partial_x u\|^2_{L^2(\mathbb{R})} - \frac{1}{4}\|u\|^4_{L^4(\mathbb{R})} + \frac{\xi_1}{2}\|u\|^2_{L^2(\mathbb{R})} + \frac{\xi_2}{2}\langle i\partial_x u, u\rangle_{L^2(\mathbb{R})}.$$

Of course, as expected, we can verify that $\psi_\xi$ is a critical point of $\mathscr{L}_\xi$. We will have to compare precisely $\psi^h_\xi$ and $\psi_\xi$. So we need a precise control of the regularity of $\psi_\xi$.

**Lemma 2.3.4.** *There exist $C > 0$ and $\varepsilon > 0$ such that for all $\xi \in \Omega$ and all $\omega \in \mathbb{R}$*

$$|\widehat{\psi_\xi}(\omega)| \leqslant Ce^{-\varepsilon|\omega|}.$$

*Proof.* It is a classical result of elliptic regularity. Here we can see it directly through formula (2.4). We also could prove it directly with the same ideas as in Theorem 2.3.13 below. □

First, we prove, through the following lemma, that $\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)$ is negative in the direction of $\psi^h_\xi$.

**Lemma 2.3.5.** *There exist $\alpha > 0$ and $h_0 > 0$ such that for all $h < h_0$ and all $\xi \in \Omega$ we have*

$$\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)(\psi^h_\xi, \psi^h_\xi) \leqslant -\alpha\|\psi^h_\xi\|^2_{H^1(\mathbb{R})}.$$

*Proof.* If $u \in H^1(\mathbb{R})$ we have

$$\mathrm{d}^2\,\mathscr{L}_\xi(u)(u,u) = \mathrm{d}\,\mathscr{L}_\xi(u)(u) - 2\|u\|^4_{L^4(\mathbb{R})}.$$

Consequently, since $\psi_\xi$ is a critical point of $\mathscr{L}_\xi$, we have

$$\mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(\psi_\xi, \psi_\xi) = -2\|\psi_\xi\|^4_{L^4(\mathbb{R})}.$$

However, $\xi \mapsto \|\psi_\xi\|^4_{L^4(\mathbb{R})}$ and $\xi \mapsto \|\psi_\xi\|^2_{H^1(\mathbb{R})}$ are continuous positive maps on $\overline{\Omega}$. So, there exists $\alpha > 0$ such that, for all $\xi \in \Omega$,

$$\mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(\psi_\xi, \psi_\xi) = -2\|\psi_\xi\|^4_{L^4(\mathbb{R})} \leqslant -\alpha\|\psi_\xi\|^2_{H^1(\mathbb{R})}.$$

Since $\|\psi^h_\xi\|^2_{H^1(\mathbb{R})} \leqslant \|\psi_\xi\|^2_{H^1(\mathbb{R})}$ (see (2.39)), to conclude this proof it is enough to prove that $\mathscr{L}^h_\xi(\psi^h_\xi)(\psi^h_\xi, \psi^h_\xi)$ goes to $\mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(\psi_\xi, \psi_\xi)$ when $h$ goes to $0$, uniformly with respect to $\xi \in \Omega$. We can write

$$\begin{aligned}\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)(\psi^h_\xi, \psi^h_\xi) = {} & \mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(\psi_\xi, \psi_\xi) + \int_{\mathbb{R}} \left|\frac{\psi^h_\xi(x+h) - \psi^h_\xi(x)}{h}\right|^2 - |\partial_x\psi^h_\xi|^2\,\mathrm{dx} \\ & + \mathrm{d}^2\,\mathscr{L}_\xi(\psi^h_\xi)(\psi^h_\xi, \psi^h_\xi) - \mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(\psi_\xi, \psi_\xi). \end{aligned} \tag{2.40}$$

First, with Fourier Plancherel isometry, we control by the classical estimate of consistency, the term generated by the discretization of the second derivative

$$\begin{aligned}\left|\int_{\mathbb{R}} |\partial_x\psi^h_\xi|^2 - \left|\frac{\psi^h_\xi(x+h) - \psi^h_\xi(x)}{h}\right|^2 \mathrm{dx}\right| &= \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \left[\omega^2 - \frac{4}{h}^2 \sin^2\left(\frac{\omega h}{2}\right)\right] |\widehat{\psi_\xi}(\omega)|^2\,\mathrm{d}\omega \\ &\leqslant \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \frac{1 - \mathrm{sinc}^2\left(\frac{\omega h}{2}\right)}{\omega^2}\omega^4|\widehat{\psi_\xi}(\omega)|^2\,\mathrm{d}\omega \\ &\leqslant \sup_{\omega\in\mathbb{R}} \frac{1 - \mathrm{sinc}^2\left(\frac{\omega h}{2}\right)}{\omega^2}\|\partial_x^2\psi_\xi\|^2_{L^2(\mathbb{R})} \\ &= \left(\frac{h}{2}\right)^2 \sup_{\omega\in\mathbb{R}} \frac{1 - \mathrm{sinc}^2(\omega)}{\omega^2}\|\partial_x^2\psi_\xi\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

Furthermore, we deduce from Lemma 2.3.4 that $\|\partial_x^2\psi_\xi\|^2_{L^2(\mathbb{R})}$ can be estimated uniformly with respect to $\xi \in \Omega$.

The convergence of the second term in (2.40) is easier. Indeed, we deduce from Lemma 2.3.4 that $\psi_\xi^h$ goes to $\psi_\xi$ when $h$ goes to $0$, uniformly with respect to $\xi \in \Omega$. We conclude because it is clear that the map $u \mapsto \mathrm{d}^2\,\mathscr{L}_\xi(u)(u,u)$ is Lipschitz on bounded subsets of $H^1(\mathbb{R})$, uniformly with respect to $\xi \in \Omega$. □

Now, we give the most important lemma of this proof, establishing the coercivity property of the discrete Lagrange functions uniformly with respect to the parameters.

**Lemma 2.3.6.** *There exist $\alpha > 0$ and $h_0 > 0$ such that for all $\xi \in \Omega$ and all $h < h_0$ we have*

$$\forall v \in BL_h^2 \cap \mathrm{Span}(i\psi_\xi^h, \partial_x\psi_\xi^h, \psi_\xi^h)^{\perp_{L^2}}, \quad \mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v,v) \geqslant \alpha\|v\|^2_{H^1(\mathbb{R})}. \tag{2.41}$$

*Proof.* We are going to establish this estimate by a perturbation of the continuum limit. Indeed, for the continuous Lagrangian functions this result has been proved by Weinstein in [115]. There exists $\alpha > 0$ such that for all $\xi \in \Omega$

$$\forall u \in H^1(\mathbb{R}) \cap \mathrm{Span}(i\psi_\xi, \partial_x\psi_\xi, \psi_\xi)^{\perp_{L^2}}, \quad \mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(v,v) \geqslant \alpha\|v\|^2_{H^1(\mathbb{R})}.$$

Literally, it is not exactly the result of Weinstein. We explain, in Lemma 2.5.4 of the Appendix how to get this estimate from the original result. Moreover, this result can be slightly extended to obtain the existence of two constants $c_1, c_2 > 0$ such that for all $\xi \in \Omega$,

$$\begin{aligned} &\text{if} \quad \|u - \psi_\xi\|_{H^1(\mathbb{R})} < c_1 \quad \text{and} \quad \max\big(|\langle\psi_\xi, v\rangle_{L^2(\mathbb{R})}|, |\langle i\psi_\xi, v\rangle_{L^2(\mathbb{R})}|, |\langle\partial_x\psi_\xi, v\rangle_{L^2(\mathbb{R})}|\big) < c_2\|v\|_{H^1} \\ &\qquad\qquad \text{then} \quad \mathrm{d}^2\,\mathscr{L}_\xi(u)(v,v) \geqslant \frac{\alpha}{8}\|v\|^2_{H^1(\mathbb{R})}. \end{aligned} \tag{2.42}$$

This result is a consequence of Lemma 2.5.6 given in Appendix. With its formalism we take $E = H^1(\mathbb{R})$, $b = \mathrm{d}^2\,\mathscr{L}_\xi$ and $X = \mathrm{Span}(i\psi_\xi, \partial_x\psi_\xi, \psi_\xi)$. This last family is free because $\psi_\xi$ is not a plane wave. Consequently, the associated Gram matrix is invertible. Finally, we just need to verify that the constants $c_1$ and $c_2$ given by the lemma can be controlled uniformly with respect to $\xi \in \overline{\Omega}$. But it is a direct consequence of the estimate proven in Lemma 2.5.6 since the Gram matrix is a continuous function of $\xi \in \overline{\Omega}$.

Now, we focus on estimate (2.41) of Lemma 2.3.6. Let $h_0 > 0$ be small enough to get that for all $h < h_0$ and all $\xi \in \Omega$, we have $\|\psi_\xi^h - \psi_\xi\|_{H^1(\mathbb{R})} < c_1$. Let us fix $h < h_0$, $\xi \in \Omega$ and consider a direction $v \in BL_h^2 \cap \mathrm{Span}(\psi_\xi^h, i\psi_\xi^h, \partial_x\psi_\xi^h)^{\perp_{L^2}}$. We decompose $v$ as

$$v = v_\ell + v_b \quad \text{with} \quad \widehat{v_\ell} = \mathbb{1}_{(-\omega_0,\omega_0)}\widehat{v} \quad \text{and} \quad \omega_0 = \frac{2\theta}{h_0}$$

where $\theta \in (0, \frac{\pi}{2})$ is a constant (independent of $h, \xi$ and $h_0$) that we will determine later. Consider the following decomposition

$$\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v,v) = \mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_\ell,v_\ell) + \mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_b,v_b) + 2\,\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_b,v_\ell). \tag{2.43}$$

We estimate separately each one of these terms as follows :

- For the first one, we deduce from Lemma 2.3.4 and the constraint on $v$ that there exists $\varepsilon, C > 0$ (independent of $\xi$) such that

$$\max\left(|\langle\psi_\xi, v_\ell\rangle_{L^2(\mathbb{R})}|, |\langle i\psi_\xi, v_\ell\rangle_{L^2(\mathbb{R})}|, |\langle\partial_x\psi_\xi, v_\ell\rangle_{L^2(\mathbb{R})}|\right) \leqslant Ce^{-\varepsilon\omega_0}\|v\|_{H^1(\mathbb{R})}.$$

Consequently, if $h_0$ is small enough to get $Ce^{-\varepsilon\omega_0} < c_2$, we can apply (2.42) to get

$$\mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi^h)(v_\ell, v_\ell) \geqslant \frac{\alpha}{8}\|v_\ell\|^2_{H^1(\mathbb{R})}.$$

Hence we have

$$\begin{aligned}
\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_\ell, v_\ell) &\geqslant \frac{\alpha}{8}\|v_\ell\|^2_{H^1(\mathbb{R})} + \mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_\ell, v_\ell) - \mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi^h)(v_\ell, v_\ell)\\
&= \frac{\alpha}{8}\|v_\ell\|^2_{H^1(\mathbb{R})} + \frac{1}{2\pi}\int_{\mathbb{R}}\left[\frac{4}{h^2}\sin^2\left(\frac{\omega h}{2}\right) - \omega^2\right]|\widehat{v_\ell}(\omega)|^2\,\mathrm{d}\omega\\
&= \frac{\alpha}{8}\|v_\ell\|^2_{H^1(\mathbb{R})} - \frac{1}{2\pi}\int_{|\omega|<\omega_0}\left[1 - \mathrm{sinc}^2(\frac{\omega h}{2})\right]|\omega\widehat{v_\ell}(\omega)|^2\,\mathrm{d}\omega\\
&\geqslant \frac{\alpha}{8}\|v_\ell\|^2_{H^1(\mathbb{R})} - \left[1 - \mathrm{sinc}^2(\theta)\right]\|v_\ell\|^2_{H^1(\mathbb{R})}
\end{aligned}$$

Choosing $\theta \in (0, \frac{\pi}{2})$ to have $1 - \mathrm{sinc}^2(\theta) < \frac{\alpha}{16}$, we get

$$\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_\ell, v_\ell) \geqslant \frac{\alpha}{16}\|v_\ell\|^2_{H^1(\mathbb{R})}.$$

- For the second term, we use the Fourier Plancherel isometry to get

$$\begin{aligned}
\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_b, v_b) \geqslant &\frac{1}{2\pi}\int_{\mathbb{R}}\mathrm{sinc}^2\left(\frac{\omega h}{2}\right)\omega^2|\widehat{v_b}(\omega)|^2\,\mathrm{d}\omega - 3\|\psi_\xi^h\|^2_{L^\infty(\mathbb{R})}\|v_b\|^2_{L^2(\mathbb{R})}\\
&- \frac{|\xi_2|}{2}\|\partial_x v_b\|_{L^2(\mathbb{R})}\|v_b\|_{L^2(\mathbb{R})}\\
\geqslant &\,\mathrm{sinc}^2(\theta)\|\partial_x v_b\|^2_{L^2(\mathbb{R})} - 3\|\psi_\xi^h\|^2_{L^\infty(\mathbb{R})}\|v_b\|^2_{L^2(\mathbb{R})} - \frac{|\xi_2|}{2}\|\partial_x v_b\|_{L^2(\mathbb{R})}\|v_b\|_{L^2(\mathbb{R})}.
\end{aligned}$$

However, applying the Fourier Plancherel isometry we get

$$\|v_b\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi}\int_{|\omega|>\omega_0}|\widehat{v}(\omega)|^2\,\mathrm{d}\omega \leqslant \frac{1}{\omega_0^2}\frac{1}{2\pi}\int_{|\omega|>\omega_0}|\omega\widehat{v}(\omega)|^2\,\mathrm{d}\omega = \frac{1}{\omega_0^2}\|\partial_x v_b\|^2_{L^2(\mathbb{R})}.$$

Consequently, we have

$$\begin{aligned}
\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_b, v_b) &\geqslant \left(\mathrm{sinc}^2(\theta) - \frac{3\|\psi_\xi^h\|^2_{L^\infty(\mathbb{R})}}{\omega_0^2} - \frac{|\xi_2|}{2\omega_0}\right)\|\partial_x v_b\|^2_{L^2(\mathbb{R})}\\
&\geqslant \left(\mathrm{sinc}^2(\theta) - \frac{3\|\psi_\xi^h\|^2_{L^\infty(\mathbb{R})}}{\omega_0^2} - \frac{|\xi_2|}{2\omega_0}\right)\frac{\omega_0^2}{1+\omega_0^2}\|v_b\|^2_{H^1(\mathbb{R})}
\end{aligned}$$

Since these quantities can be controlled uniformly with respect to $\xi \in \Omega$, if $h_0$ is small enough, we have for all $\xi \in \Omega$

$$\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v_b, v_b) \geqslant \frac{1}{2}\,\mathrm{sinc}^2(\theta)\|v_b\|^2_{H^1(\mathbb{R})}.$$

- For the third term, since the frequency supports of $v_\ell$ and $v_b$ are disjoint, we get

$$\begin{aligned}
\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)(v_b,v_\ell) &= \mathrm{d}^2\,\frac{\|\bullet\|^4_{L^4}}{4}(\psi^h_\xi)(v_b,v_l)\\
&\geqslant -3\|\psi^h_\xi\|^2_{L^\infty(\mathbb{R})}\|v_b\|_{L^2(\mathbb{R})}\|v_\ell\|_{L^2(\mathbb{R})}\\
&\geqslant -3\|\psi^h_\xi\|^2_{L^\infty(\mathbb{R})}\|v_\ell\|_{H^1(\mathbb{R})}\frac{\|v_b\|_{H^1(\mathbb{R})}}{\sqrt{1+\omega_0^2}}\\
&\geqslant -\frac{3\|\psi^h_\xi\|^2_{L^\infty(\mathbb{R})}}{2\sqrt{1+\omega_0^2}}\left(\|v_\ell\|^2_{H^1(\mathbb{R})}+\|v_b\|^2_{H^1(\mathbb{R})}\right).
\end{aligned}$$

  Controlling this quantity uniformly with respect to $\xi\in\Omega$, we deduce that if $h_0$ is small enough then

$$\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)(v_b,v_\ell)\geqslant -\frac{\beta}{2}\left(\|v_\ell\|^2_{H^1(\mathbb{R})}+\|v_b\|^2_{H^1(\mathbb{R})}\right),$$

  with $\beta=\min(\frac{1}{2}\operatorname{sinc}^2(\theta),\frac{\alpha}{16})$.

Applying these three estimates, we deduce that there exists an $h_0>0$ such that if $h<h_0$ and $\xi\in\Omega$ then for all $v\in BL^2_h\cap\operatorname{Span}(\psi^h_\xi,i\psi^h_\xi,\partial_x\psi^h_\xi)^{\perp_{L^2}}$, we have

$$\mathrm{d}^2\,\mathscr{L}^h_\xi(\psi^h_\xi)(v,v)\geqslant\frac{\beta}{2}\left(\|v_\ell\|^2_{H^1(\mathbb{R})}+\|v_b\|^2_{H^1(\mathbb{R})}\right)=\frac{\beta}{2}\|v\|^2_{H^1(\mathbb{R})}.$$

□

Before focusing on the invertibility of $\mathrm{d}^2\,\mathscr{L}^h_{\xi\,|\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)$, we give a small but useful lemma (particularly to control uniformly the norm of the inverse).

**Lemma 2.3.7.** *For all $r>0$, there exists $C>0$ such that for all $h>0$ and all $\xi\in\Omega$, we have for all $u,v,w\in BL^2_h$ with $\|w\|_{H^1(\mathbb{R})}<r$*

$$|\,\mathrm{d}^2\,\mathscr{L}^h_\xi(w)(u,v)|\leqslant C\|u\|_{H^1(\mathbb{R})}\|v\|_{H^1(\mathbb{R})}.$$

*Proof.* Since $|\sin(\omega)|\leqslant|\omega|$, we observe that, for all $u,v\in BL^2_h$

$$\begin{aligned}
|\,\mathrm{d}^2\,\mathscr{L}^h_\xi(w)| \leqslant&\|\partial_x u\|_{L^2(\mathbb{R})}\|\partial_x v\|_{L^2(\mathbb{R})}+3\|w\|^2_{L^\infty(\mathbb{R})}\|u\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})}\\
&+\xi_1\|u\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})}+|\xi_2|\|\partial_x u\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})}.
\end{aligned}$$

The result is thus a simple consequence of the classical Sobolev inequality,

$$\|w\|^2_{L^\infty(\mathbb{R})}\leqslant\|w\|_{L^2(\mathbb{R})}\|\partial_x w\|_{L^2(\mathbb{R})}.$$

□

In the following concluding Lemma, we prove the invertibility of $\mathrm{d}^2\,\mathscr{L}^h_{\xi\,|\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)$ and control the norm of its inverse uniformly with respect to $\xi\in\Omega$ and $h$ small enough.

**Lemma 2.3.8.** *There exist $h_0>0$ and $C>0$ such that for all $\xi\in\Omega$ and all $h<h_0$, $\mathrm{d}^2\,\mathscr{L}^h_{\xi\,|\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)$ is invertible and the norm of its inverse is smaller than $C$.*

*Proof.* We use Lemma 2.5.7 of the Appendix, by taking $E = \mathrm{Ker}(\mathrm{id} - S_h)$ (equipped with the $\| \bullet \|_{H^1(\mathbb{R})}$ norm), $T = \mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)$, $E_p = \mathrm{Span}(\psi^h_\xi)^{\perp_{L^2}} \cap \mathrm{Ker}(\mathrm{id} - S_h)$ and $E_m = \mathrm{Span}(\psi^h_\xi)$.

To get the coercivity estimate on $E_m$ we apply Lemma 2.3.5, while coercivity on $E_p$ is obtained from Lemma 2.3.6 after noticing that

$$\mathrm{Ker}(\mathrm{id} - S_h) \subset BL^2_h \cap \mathrm{Span}(\psi^h_\xi, i\psi^h_\xi, \partial_x \psi^h_\xi)^{\perp_{L^2}},$$

which is obvious since $i\psi^h_\xi, \partial_x \psi^h_\xi \in \mathrm{Ker}(\mathrm{id} + S_h) \subset \mathrm{Ker}(\mathrm{id} - S_h)^{\perp_{L^2}}$.

Applying Lemma 2.5.7, we obtain the invertibility of $\mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)$ and an explicit control of the norm of its inverse in terms of $\alpha_p$, $\alpha_m$ and $\|T\|$. However, with Lemma 2.3.5 and Lemma 2.3.6, we have a uniform control of $\alpha_p$ and $\alpha_m$ with respect to $\xi \in \Omega$ and $h$ small enough, the uniform control of $\|T\|$ being given by Lemma 2.3.7. □

Step 3 : The resolution and its consequences

Now, we want to apply the inverse function theorem 2.5.3 to $\mathrm{d}\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}$ in $\psi^h_\xi$. In the following Lemma, we focus on the last assumption required, i.e. $\mathrm{d}^2 \mathscr{L}^h_\xi$ is a Lipschitz function.

**Lemma 2.3.9.** *For all $R > 0$ there exists $k > 0$ such that for all $\xi \in \Omega$, $h > 0$, $u_1, u_2, v, w \in BL^2_h$, with $\|u_1\|_{H^1(\mathbb{R})} < R$ and $\|u_2\|_{H^1((\mathbb{R}))} < R$, we have*

$$\| \mathrm{d}^2 \mathscr{L}^h_\xi(u_1)(v,w) - \mathrm{d}^2 \mathscr{L}^h_\xi(u_2)(v,w)\| \leqslant k\|u_1 - u_2\|_{H^1(\mathbb{R})}\|v\|_{H^1(\mathbb{R})}\|w\|_{H^1(\mathbb{R})}.$$

*Proof.* We use mean value inequality. Indeed $\mathrm{d}^3 \mathscr{L}^h_\xi = -\frac{1}{4}\mathrm{d}^3 \| \bullet \|^4_{L^4(\mathbb{R})}$ is clearly a bounded function on bounded subsets of $H^1(\mathbb{R})$. □

Applying Lemma 2.3.9 and Lemma 2.3.8, we deduce that assumptions of the inverse function Theorem 2.5.3 are fulfilled. In the following Proposition, we give its conclusion.

**Proposition 2.3.2.** *There exist $h_0, r, \lambda, C > 0$ such that if $h < h_0$ and $\xi \in \Omega$ then*

- $\mathrm{d}\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}$ *is a $C^1$ diffeomorphism from $\{u \in \mathrm{Ker}(\mathrm{id} - S_h) \mid \|u - \psi^h_\xi\|_{H^1(\mathbb{R})} < r\}$ onto its image,*
- *if $u \in \mathrm{Ker}(\mathrm{id} - S_h)$ and $\|u - \psi^h_\xi\|_{H^1(\mathbb{R})} < r$ then*

$$\| \mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(u)^{-1}\|_{\mathscr{L}(\mathrm{Ker}(\mathrm{id}-S_h)';\mathrm{Ker}(\mathrm{id}-S_h))} \leqslant C,$$

- *if $\rho < r$ and $\Phi \in \mathrm{Ker}(\mathrm{id} - S_h)'$ with $\|\Phi - \mathrm{d}\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)\|_{\mathrm{Ker}(\mathrm{id}-S_h)'} < \lambda\rho$ then there exists $u \in \mathrm{Ker}(\mathrm{id} - S_h)$ such that $\|u - \psi^h_\xi\|_{H^1(\mathbb{R})} < \rho$ and*

$$\mathrm{d}\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(u) = \Phi.$$

To apply this result to $\Phi = 0$, we will show that the norm of $\mathrm{d}\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\psi^h_\xi)$ is small when $h \to 0$, uniformly in $\xi \in \Omega$. It is exactly, what we establish in the following Lemma, which also explains the error term "$h^2$" in Theorem 2.1.4.

**Lemma 2.3.10.** *For all $h_0 > 0$ there exists $M > 0$ such that if $h < h_0$ and $\xi \in \Omega$ then*

$$\forall v \in BL^2_h, \quad |\mathrm{d}\mathscr{L}^h_\xi(\psi^h_\xi)(v)| \leqslant Mh^2\|v\|_{H^1(\mathbb{R})}.$$

*Proof.* The arguments are very similar to the proof of Lemma 2.3.5. The key point is the estimate of the consistency error associated to the discretization of the second derivative by finite differences.

Since $\psi_\xi$ is a critical point of $\mathscr{L}_\xi$, we deduce from the definition of $\psi_\xi^h$ (see (2.39)) that

$$\begin{aligned}\mathrm{d}\,\mathscr{L}_\xi^h(\psi_\xi^h)(v) &= \mathrm{d}\,\mathscr{L}_\xi^h(\psi_\xi^h)(v) - \mathrm{d}\,\mathscr{L}_\xi(\psi_\xi)(v)\\ &= \langle(\partial_x^2-\Delta_h)\psi_\xi, v\rangle_{L^2(\mathbb{R})} + \mathrm{d}\,\frac{\|\bullet\|_{L^4(\mathbb{R})}^4}{4}(\psi_\xi)(v) - \mathrm{d}\,\frac{\|\bullet\|_{L^4(\mathbb{R})}^4}{4}(\psi_\xi^h)(v). \end{aligned} \tag{2.44}$$

To estimate the first term, we use Fourier Plancherel isometry to get

$$\begin{aligned}|\langle(\partial_x^2-\Delta_h)\psi_\xi, v\rangle_{L^2(\mathbb{R})}| &= \left|\frac{1}{2\pi}\int_{\mathbb{R}}\left[\frac{4}{h^2}\sin^2\left(\frac{\pi\omega h}{2}\right)-\omega^2\right]\widehat{\psi_\xi}(\omega)\cdot\widehat{v}(\omega)\,\mathrm{d}\omega\right|\\ &\leqslant \sup_{\omega\in\mathbb{R}}\left|\frac{\mathrm{sinc}^2\left(\frac{\omega h}{2}\right)-1}{\omega^2}\right|\|\partial_x^4\psi_\xi\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})} = \left(\frac{h}{2}\right)^2\sup_{\omega\in\mathbb{R}}\left|\frac{\mathrm{sinc}^2(\omega)-1}{\omega^2}\right|\|\partial_x^4\psi_\xi\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})}.\end{aligned}$$

As we can see from Lemma 2.3.4, $\|\partial_x^4\psi_\xi\|_{L^2(\mathbb{R})}$ is clearly bounded uniformly with respect to $\xi\in\Omega$.

To control the second term in (2.44), we use mean value inequality and Lemma 2.3.4 to get some constants $M, C>0$ independent of $h$ and $\xi\in\Omega$ such that

$$\left|\mathrm{d}\,\frac{\|\bullet\|_{L^4(\mathbb{R})}^4}{4}(\psi_\xi)(v) - \mathrm{d}\,\frac{\|\bullet\|_{L^4(\mathbb{R})}^4}{4}(\psi_\xi^h)(v)\right| \leqslant M\|\psi_\xi-\psi_\xi^h\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})} \leqslant Ce^{-\frac{\pi\varepsilon}{h}}\|v\|_{L^2(\mathbb{R})},$$

which shows the result, provided $h<h_0$ small enough. □

Applying Lemma 2.3.10, if $h_0$ is smaller than $\sqrt{\frac{\lambda r}{2M}}$ we can choose $\Phi=0$ in Proposition 2.3.2 and we denote by $\eta_\xi^h$ the corresponding critical point of $\mathscr{L}_{\xi\,|\mathrm{Ker}(\mathrm{id}-S_h)}^h$. As shown in the first step, $\eta_\xi^h$ is thus a critical point of $\mathscr{L}_\xi^h$, and with Proposition 2.3.2, we have proven the points $a)$ to $d)$ of Theorem 2.3.1. It remains to show the coercivity estimate $e)$ and the regularity with respect to $\xi$.

To obtain the coercivity estimate, we just have to perturb the estimate of Lemma 2.3.6 with Lemma 2.5.6 presented in Appendix. This is given by the following result

**Lemma 2.3.11.** *There exist $\alpha>0$, $h_0>0$ and $\rho>0$ such that for all $\xi\in\Omega$, $h<h_0$ and $u\in BL_h^2$ such that $\|u-\psi_\xi^h\|_{H^1(\mathbb{R})}<\rho$, we have*

$$\forall v\in BL_h^2\cap\mathrm{Span}(iu,\partial_x u,u)^{\perp_{L^2}},\quad \mathrm{d}^2\,\mathscr{L}_\xi^h(u)(v,v)\geqslant\alpha\|v\|_{H^1(\mathbb{R})}^2. \tag{2.45}$$

*Proof.* The proof is very similar to the first part of the proof of Lemma 2.3.6, but we need to track precisely the dependence of the constant with respect to $h$.

First, applying Lemma 2.3.6, we know that there exists $h_0>0$ and $\alpha>0$ such that for all $h<h_0$ and all $\xi\in\Omega$ we have

$$\forall v\in BL_h^2\cap\mathrm{Span}(i\psi_\xi^h,\partial_x\psi_\xi^h,\psi_\xi^h)^{\perp_{L^2}},\quad \mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)(v,v)\geqslant\alpha\|v\|_{H^1(\mathbb{R})}^2.$$

We want to apply Lemma 2.5.6 in $\psi_\xi^h$ in order to perturb this estimate and prove that there exist $h_0 > 0$, $c_1, c_2 > 0$ such that for all $\xi \in \Omega$ and all $h < h_0$, if

$$\|u - \psi_\xi^h\|_{H^1(\mathbb{R})} \leqslant c_1 \quad \text{and} \quad \max\Big(|\langle \psi_\xi^h, v\rangle_{L^2(\mathbb{R})}|, |\langle i\psi_\xi^h, v\rangle_{L^2(\mathbb{R})}|, |\langle \partial_x\psi_\xi^h, v\rangle_{L^2(\mathbb{R})}|\Big) \leqslant c_2\|v\|_{H^1(\mathbb{R})},$$

then

$$\mathrm{d}^2\,\mathscr{L}_\xi^h(u)(v,v) \geqslant \alpha\|v\|^2_{H^1(\mathbb{R})} \geqslant \frac{\alpha}{2}\|v\|^2_{H^1(\mathbb{R})}.$$

To do this, we apply Lemma 2.5.6 in $\psi_\xi^h$ with $E = BL^2_h$, $X = \mathrm{Span}(i\psi_\xi^h, \partial_x\psi_\xi^h, \psi_\xi^h)$ and $b = \mathrm{d}^2\,\mathscr{L}_\xi^h$. The Gram matrix is

$$G_\xi^h = \begin{pmatrix} \|\psi_\xi^h\|^2_{L^2(\mathbb{R})} & \langle i\psi_\xi^h, \partial_x\psi_\xi^h\rangle_{L^2(\mathbb{R})} & 0 \\ \langle i\psi_\xi^h, \partial_x\psi_\xi^h\rangle_{L^2(\mathbb{R})} & \|\partial_x\psi_\xi^h\|^2_{L^2(\mathbb{R})} & 0 \\ 0 & 0 & \|\psi_\xi^h\|^2_{L^2(\mathbb{R})} \end{pmatrix}.$$

To prove that the constants $c_1, c_2 > 0$ –explicitly given by Lemma 2.5.6– are independent of $\xi \in \Omega$ and $h$ small enough, we have to control uniformly the inverse of $G_\xi^h$, the norm of $\psi_\xi^h$ in $H^1(\mathbb{R})$, the norm of $\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)$ and prove that $\mathrm{d}^2\,\mathscr{L}_\xi^h$ is uniformly Lipschitz.

The control of $\psi_\xi^h$ in $H^1(\mathbb{R})$ is obvious, and Lemma 2.3.9 shows that $\mathrm{d}^2\,\mathscr{L}_\xi^h$ is uniformly Lipschitz. In Lemma 2.3.7, we have proven that the norm $\mathrm{d}^2\,\mathscr{L}_\xi^h(\psi_\xi^h)$ is uniformly bounded with respect to $h$ and $\xi \in \Omega$. So we just need to focus on the Gram matrix.

As explained in the proof of Lemma 2.3.6 $G_\xi^0 = G_\xi$ is invertible. Furthermore, $(h,\xi) \mapsto G_\xi^h$ is a continuous function on $\mathbb{R}_+ \times \overline{\Omega}$, so there exists $h_0 > 0$ and $M > 0$ such that for all $h < h_0$ and all $\xi \in \Omega$, $G_\xi^h$ is invertible and $\|(G_\xi^h)^{-1}\|_\infty \leqslant M$.

To prove (2.45), let us set $\rho = \min(c_1, c_2)$ and consider $h < h_0$ and $\xi \in \Omega$. Let $u, v \in BL^2_h$ be such that $\|u - \psi_\xi^h\|_{H^1(\mathbb{R})} < \rho$ and $v \in \mathrm{Span}(iu, \partial_x u, u)^{\perp_{L^2}}$. Then, we have

$$\max\Big(|\langle \psi_\xi^h, v\rangle_{L^2(\mathbb{R})}|, |\langle i\psi_\xi^h, v\rangle_{L^2(\mathbb{R})}|, |\langle \partial_x\psi_\xi^h, v\rangle_{L^2(\mathbb{R})}|\Big) \leqslant \|u - \psi_\xi^h\|_{H^1(\mathbb{R})}\|v\|_{H^1(\mathbb{R})} \leqslant c_2\|v\|_{H^1(\mathbb{R})}.$$

Consequently, we can apply the result of Lemma 2.5.6 to get

$$\mathrm{d}^2\,\mathscr{L}_\xi^h(u)(v,v) \geqslant \alpha\|v\|^2_{H^1(\mathbb{R})} \geqslant \frac{\alpha}{2}\|v\|^2_{H^1(\mathbb{R})}.$$

which shows the result. □

The following Lemma concludes the proof of Theorem 2.3.1. It shows that $\xi \mapsto \eta_\xi^h$ is $C^1$ and that its derivative with respect to $\xi$ is a good approximation of the derivative of $\psi_\xi$ with respect to $\xi$.

**Lemma 2.3.12.** *Let $h_0, r, \lambda, C > 0$ be the constants given in Proposition 2.3.2 and $M > 0$ be the constant associated with $h_0 > 0$ given in Lemma 2.3.10. Let $h_1 := \min(h_0, \sqrt{\frac{\lambda r}{2M}})$ and for any $h < h_1$ and $\xi \in \Omega$, let $\eta_\xi^h$ denotes the critical point of $\mathscr{L}_\xi^h$ at a distance smaller than $r$ from $\psi_\xi^h$. There exists $k > 0$ such that for all $h < h_1$, for all $\xi \in \Omega$, $\xi \mapsto \eta_\xi^h$ is $C^1$ and*

$$\|\,\mathrm{d}_\xi\,\psi_\xi(\zeta) - \mathrm{d}_\xi\,\eta_\xi^h(\zeta)\|_{H^1(\mathbb{R})} \leqslant k|\zeta|h^2.$$

*Proof.* Let $h < h_1$ and $\xi \in \Omega$. The function $(u, \zeta) \mapsto \mathrm{d}\,\mathscr{L}^h_{\zeta\ |\mathrm{Ker}(\mathrm{id}-S_h)}(u)$ is clearly a $C^1$ function. Applying Proposition 2.3.2, its derivative with respect to $u$ in $(\eta^h_\xi, \xi)$ is invertible. By construction, $(\eta^h_\xi, \xi)$ is a zero point of this function. So we can apply the implicit function theorem.

There exists $\rho > 0$ such that $B(\xi, \rho) \subset \Omega$ and $\Gamma \in C^1(B(\xi, \rho); \mathrm{Ker}(\mathrm{id}-S_h))$ such that

$$\forall \zeta \in B(\xi, \rho), \quad \mathrm{d}\,\mathscr{L}^h_{\zeta\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\Gamma(\zeta)) = 0.$$

To prove that $\Gamma(\zeta) = \eta^h_\zeta$, it is enough to prove that $\|\Gamma(\zeta) - \psi^h_\zeta\|_{H^1(\mathbb{R})} < r$. But by construction of $h_1$, we deduce of Proposition 2.3.2 that

$$\|\Gamma(\xi) - \psi^h_\xi\|_{H^1(\mathbb{R})} = \|\eta^h_\xi - \psi^h_\xi\|_{H^1(\mathbb{R})} \leqslant \frac{r}{2}.$$

Furthermore, $\zeta \mapsto \Gamma(\zeta)$ and $\zeta \mapsto \psi^h_\zeta$ are continuous functions. So there exists $\widetilde{\rho} < \rho$ such that,

$$\forall \zeta \in B(\xi, \widetilde{\rho}), \quad \|\Gamma(\xi) - \Gamma(\zeta)\|_{H^1(\mathbb{R})} + \|\psi^h_\zeta - \psi^h_\xi\|_{H^1(\mathbb{R})} \leqslant \frac{r}{4}.$$

Applying the triangle inequality for $\zeta \in B(\xi, \widetilde{\rho})$, we thus obtain

$$\|\Gamma(\zeta) - \psi^h_\zeta\|_{H^1(\mathbb{R})} \leqslant \frac{3}{4} r < r.$$

Since we have proven in Proposition 2.3.2 that $\mathrm{d}\,\mathscr{L}^h_{\zeta\ |\mathrm{Ker}(\mathrm{id}-S_h)}$ is invective on

$$\{u \in \mathrm{Ker}(\mathrm{id}-S_h) \mid \|u - \psi^h_\zeta\|_{H^1(\mathbb{R})} < r\},$$

we get $\Gamma(\zeta) = \eta^h_\zeta$ for all $\zeta \in B(\xi, \widetilde{\rho})$. Consequently, $\zeta \mapsto \eta^h_\zeta$ is $C^1$.

Now, we have to prove that $\mathrm{d}_\xi\, \eta^h_\xi$ is an approximation of $\mathrm{d}_\xi\, \psi_\xi$. First, we introduce some constants $c, \varepsilon > 0$ such that for all $\xi \in \Omega$ and all $\zeta \in \mathbb{R}^2$, we have

$$\forall \omega \in \mathbb{R}, \quad |\widehat{\mathrm{d}_\xi\, \psi_\xi(\zeta)}(\omega)| \leqslant c|\zeta| e^{-\varepsilon|\omega|}. \tag{2.46}$$

There are several ways to establish this property. The most direct is probably to deduce it from the explicit formula of $\psi_\xi$ (see (2.4)). But it can also be proven with elliptic regularity as in Theorem 2.3.13 below.

Then, we deduce from the definition of $\psi^h_\xi$ that for all $h > 0$, $\xi \mapsto \psi^h_\xi$ is $C^1$ and there exists $k > 0$ such that

$$\forall h > 0, \quad \forall \xi \in \Omega, \quad \forall \zeta \in \mathbb{R}^2, \quad \|\,\mathrm{d}_\xi\, \psi_\xi(\zeta) - \mathrm{d}_\xi\, \psi^h_\xi(\zeta)\|_{H^1(\mathbb{R})} \leqslant k|\zeta| e^{-\frac{\varepsilon\pi}{2h}}. \tag{2.47}$$

So we just need to prove that $\mathrm{d}_\xi\, \eta^h_\xi$ is an approximation of $\mathrm{d}_\xi\, \psi^h_\xi$ of order $2$ in $h$. To compare these quantities, we are going to prove that they are almost solutions of the same linear equation.

Since $\eta^h_\xi$ is a critical point of $\mathscr{L}^h_\xi$, it satisfies for all $v \in \mathrm{Ker}(\mathrm{id}-S_h)$, $\mathrm{d}\,\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\eta^h_\xi)(v) = 0$. So we can calculate the derivative with respect to $\xi$ to obtain that

$$\forall \zeta \in \mathbb{R}^2, \quad \forall v \in \mathrm{Ker}(\mathrm{id}-S_h), \quad \mathrm{d}^2\,\mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id}-S_h)}(\eta^h_\xi)(v, \mathrm{d}_\xi\, \eta^h_\xi(\zeta)) + b^h_\zeta[\eta^h_\xi](v) = 0,$$

where $b^h_\zeta[u] \in (\mathrm{Ker}(\mathrm{id}-S_h))'$ is defined for $u \in \mathrm{Ker}(\mathrm{id}-S_h)$ by

$$b^h_\zeta[u](v) := \zeta_1 \langle u, v\rangle_{L^2(\mathbb{R})} + \zeta_2 \langle i\partial_x u, v\rangle_{L^2(\mathbb{R})}.$$

Similarly, we define $E^h_{\xi,\zeta} \in \mathrm{Ker}(\mathrm{id} - S_h)'$ by

$$\forall \zeta \in \mathbb{R}^2, \quad \forall v \in \mathrm{Ker}(\mathrm{id} - S_h), \quad \mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\psi^h_\xi)(v, \mathrm{d}_\xi \psi^h_\xi(\zeta)) + b^h_\zeta[\psi^h_\xi](v) = E^h_{\xi,\zeta}(v).$$

Then, we get (in $\mathrm{Ker}(\mathrm{id} - S_h)'$), for all $\zeta \in \mathbb{R}^2$ ,

$$\begin{aligned}
&\mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\eta^h_\xi)(\mathrm{d}_\xi \psi^h_\xi(\zeta) - \mathrm{d}_\xi \eta^h_\xi(\zeta)) \\
&= \left[\mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\eta^h_\xi) - \mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\psi^h_\xi)\right](\mathrm{d}_\xi \psi^h_\xi(\zeta)) + b^h_\zeta[\eta^h_\xi - \psi^h_\xi] + E^h_{\xi,\zeta}(v).
\end{aligned}$$

However, we have proven in Proposition 2.3.2 that $\mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\eta^h_\xi)$ is invertible and that the norm of its invert is smaller than $C$. So we just need to control the three right terms of the last equality.

- Applying (2.46) and (2.47), for all $h > 0$ and all $\xi \in \Omega$, we have $\| \mathrm{d}_\xi \psi^h_\xi(\zeta)\|_{H^1(\mathbb{R})} \leqslant 2|\zeta|k$. So applying Lemma 2.3.9, there exists $\kappa > 0$, such that for all $h < h_1$, all $\xi \in \Omega$, all $\zeta \in \mathbb{R}^2$ and all $v \in \mathrm{Ker}(\mathrm{id} - S_h)$,

$$\begin{aligned}
&\left|\left[\mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\eta^h_\xi) - \mathrm{d}^2 \mathscr{L}^h_{\xi\ |\mathrm{Ker}(\mathrm{id} - S_h)}(\psi^h_\xi)\right](\mathrm{d}_\xi \psi^h_\xi(\zeta))(v)\right| \\
&\qquad \leqslant \kappa \|\eta^h_\xi - \psi^h_\xi\|_{H^1(\mathbb{R})}|\zeta|\|v\|_{H^1(\mathbb{R})} \leqslant \frac{M\kappa}{\lambda} h^2|\zeta|\|v\|_{H^1(\mathbb{R})}.
\end{aligned}$$

- The estimate of the second term is obvious. Indeed, for all $h < h_1$, all $\xi \in \Omega$, all $\zeta \in \mathbb{R}^2$ and all $v \in \mathrm{Ker}(\mathrm{id} - S_h)$ we have

$$\begin{aligned}
|b^h_\zeta[\eta^h_\xi - \psi^h_\xi](v)| &\leqslant |\zeta|(\|\eta^h_\xi - \psi^h_\xi\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})} + \|\partial_x(\eta^h_\xi - \psi^h_\xi)\|_{L^2(\mathbb{R})}\|\partial_x v\|_{L^2(\mathbb{R})}) \\
&\leqslant 2\frac{M}{\lambda} h^2|\zeta|\|v\|_{H^1(\mathbb{R})}.
\end{aligned}$$

- The bound on the term $E^h_{\xi,\zeta}$ is more difficult to obtain. First, we have to identify it. Since $\psi_\xi$ is a critical point of $\mathscr{L}_\xi$, it satisfies $\mathrm{d}\mathscr{L}_\xi(\psi_\xi)(v) = 0$ for all $v \in H^1(\mathbb{R})$. By calculating its derivative with respect to $\xi$, we get for $\zeta \in \mathbb{R}^2$,

$$\mathrm{d}^2 \mathscr{L}_\xi(\psi_\xi)(v, \mathrm{d}_\xi \psi_\xi(\zeta)) + \zeta_1\langle \psi_\xi, v\rangle_{L^2(\mathbb{R})} + \zeta_2\langle i\partial_x\psi_\xi, v\rangle_{L^2(\mathbb{R})} = 0.$$

  In particular, we can choose $v \in \mathrm{Ker}(\mathrm{id} - S_h)$. Consequently, we get

$$\begin{aligned}
&\mathrm{d}^2 \mathscr{L}^h_\xi(\psi^h_\xi)(v, \mathrm{d}_\xi \psi^h_\xi(\zeta)) + b^h_\zeta[\psi^h_\xi] + \langle(\Delta_h - \partial_x^2)\, \mathrm{d}_\xi \psi^h_\xi(\zeta), v\rangle_{L^2(\mathbb{R})} \\
&\qquad + \mathrm{d}^2 \frac{\|\bullet\|^4_{L^4(\mathbb{R})}}{4}(\psi^h_\xi)(v, \mathrm{d}_\xi \psi^h_\xi(\zeta)) - \mathrm{d}^2 \frac{\|\bullet\|^4_{L^4(\mathbb{R})}}{4}(\psi_\xi)(v, \mathrm{d}_\xi \psi_\xi(\zeta)) = 0.
\end{aligned}$$

  So we have

$$\begin{aligned}
E^h_{\xi,\zeta}(v) = \mathrm{d}^2 \frac{\|\bullet\|^4_{L^4(\mathbb{R})}}{4}(\psi_\xi)(v, \mathrm{d}_\xi \psi_\xi(\zeta)) - \mathrm{d}^2 \frac{\|\bullet\|^4_{L^4(\mathbb{R})}}{4}(\psi^h_\xi)(v, \mathrm{d}_\xi \psi^h_\xi(\zeta)) \\
+ \langle(\partial_x^2 - \Delta_h)\, \mathrm{d}_\xi \psi^h_\xi(\zeta), v\rangle_{L^2(\mathbb{R})}.
\end{aligned}$$

To estimate $\langle(\partial_x^2 - \Delta_h)\,\mathrm{d}_\xi\,\psi_\xi^h(\zeta), v\rangle_{L^2(\mathbb{R})}$ we use the same method as in Lemma 2.3.10 and we can find an universal constant $C_{\mathrm{univ}} > 0$ such that

$$\left|\langle(\partial_x^2 - \Delta_h)\,\mathrm{d}_\xi\,\psi_\xi^h(\zeta), v\rangle_{L^2(\mathbb{R})}\right| \leqslant C_{\mathrm{univ}} h^2 \|\,\mathrm{d}_\xi\,\psi_\xi^h(\zeta)\|_{H^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})}. \tag{2.48}$$

On the other hand, we have

$$\begin{aligned}\Big|\mathrm{d}^2\,\|\bullet\|^4_{L^4(\mathbb{R})}(\psi_\xi)(v, \mathrm{d}_\xi\,\psi_\xi(\zeta)) - \mathrm{d}^2\,\|\bullet\|_{L^4(\mathbb{R})}(\psi_\xi^h)(v, \mathrm{d}_\xi\,\psi_\xi^h(\zeta))\Big| \\ \leqslant 12\|\psi_\xi + \psi_\xi^h\|_{L^4(\mathbb{R})}\|\psi_\xi - \psi_\xi^h\|_{L^4(\mathbb{R})}\|v\|_{L^4(\mathbb{R})}\|\,\mathrm{d}_\xi\,\psi_\xi(\zeta)\|_{L^4(\mathbb{R})} \\ + 12\|\,\mathrm{d}_\xi\,\psi_\xi(\zeta) - \mathrm{d}_\xi\,\psi_\xi^h(\zeta)\|_{L^4(\mathbb{R})}\|\psi_\xi^h\|^2_{L^4(\mathbb{R})}\|v\|_{L^4(\mathbb{R})}.\end{aligned}$$

Applying Gagliardo-Nirenberg inequality, (2.46) and Lemma 2.3.4, it is clear that $\|\psi_\xi + \psi_\xi^h\|_{L^4(\mathbb{R})}$, $\|\psi_\xi^h\|^2_{L^4(\mathbb{R})}$, $|\zeta|^{-1}\|\,\mathrm{d}_\xi\,\psi_\xi(\zeta)\|_{L^4(\mathbb{R})}$ and $|\zeta|^{-1}\|\,\mathrm{d}_\xi\,\psi_\xi^h(\zeta)\|_{H^2(\mathbb{R})}$ are bounded uniformly with respect to $\xi \in \Omega$ and $h < h_1$.
Consequently, by using (2.47), there exist $\ell > 0$, $\kappa > 0$ such that for all $h < h_1$, all $\xi \in \Omega$ and all $\zeta \in \mathbb{R}^2$, we have

$$|E^h_{\xi,\zeta}(v)| \leqslant \kappa\left(h^2 + e^{-\frac{\ell}{h}}\right) \leqslant \kappa h^2\left(1 + \left(\frac{2}{e\ell}\right)^2\right),$$

which concludes the proof of the Lemma.

□

### 2.3.2 Gevrey uniform regularity, Lyapunov stability and some adjustments

The discrete traveling waves constructed in Theorem 2.3.1 enjoy most of the properties of the continuous traveling waves $\psi_\xi$. In this subsection, we analyse some of these properties useful to prove Theorem 2.1.4.

First, we study their regularity. Of course, since they belong to $BL^2_h$ they are entire functions but we can give a control of them in Gevrey norms uniformly with respect to $h$ and $\xi$.

**Theorem 2.3.13.** *There exists $h_0 > 0$ such that for all $M > 0$, there exist $C, \varepsilon > 0$ such that for all $h < h_0$ and all $\xi \in \Omega$, if $u \in BL^h_2$ satisfies $\|u\|_{H^1(\mathbb{R})} \leqslant M$ then*

$$\mathrm{d}\,\mathscr{L}^h_\xi(u) = 0 \;\Rightarrow\; \forall\omega \in \mathbb{R},\ |\widehat{u}(\omega)| < Ce^{-\varepsilon|\omega|}. \tag{2.49}$$

*Proof.* To get this result of elliptic regularity, we prove, in the following lemma, a result of coercivity.

**Lemma 2.3.14.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a function continuous in $0$ such that $f(0) = 1$. Assume that there exists $m > 0$ such that $f \geqslant m$ on $\mathbb{R}$. Then there exist $\alpha > 0$ and $h_0 > 0$ such that for all $\xi \in \Omega$ and $h < h_0$ we have*

$$\forall\omega \in \mathbb{R}, \quad \omega^2 f(h\omega) + \xi_2\omega + \xi_1 \geqslant \alpha\left(1 + \omega^2\right).$$

*Proof.* First, observe that we have

$$\omega^2 + \xi_2\omega + \xi_1 = \left(\omega + \frac{\xi_2}{2}\right)^2 + \xi_1 - \left(\frac{\xi_2}{2}\right)^2.$$

Consequently, there exists $\beta > 0$ such that for all $\xi \in \Omega$, we have

$$\omega^2 + \xi_2\omega + \xi_1 \geqslant \beta\left(1 + \omega^2\right).$$

Second observe that there exists $\omega_0 > 0$ such that, for all $|\omega| > \omega_0$ we have

$$m\omega^2 + \xi_2\omega + \xi_1 \geqslant \frac{m}{2}\left(1 + \omega^2\right).$$

Consequently, for such $\omega$ and for any $h > 0$, we have

$$\omega^2 f(h\omega) + \xi_2\omega + \xi_1 \geqslant \frac{m}{2}\left(1 + \omega^2\right).$$

Now, since $f$ is continuous in $0$, there exists $\delta > 0$ such that if $|\omega| < \delta$ then $|f(\omega) - 1| < \frac{\beta}{2}$. Consequently, if $|\omega| < \omega_0$ and $h < \frac{\delta}{\omega_0} =: h_0$ then we have

$$\omega^2 f(h\omega) + \xi_2\omega + \xi_1 = \omega^2 + \xi_2\omega + \xi_1 + \omega^2(f(h\omega) - 1) \geqslant \beta\left(1 + \omega^2\right) - \frac{\beta}{2}\omega^2 \geqslant \frac{\beta}{2}\left(1 + \omega^2\right).$$

□

We now prove the elliptic regularity result (2.49). Let us write the equation $\mathrm{d}\mathscr{L}^h_\xi(u) = 0$ in terms of the Fourier transform $\widehat{u}$. It is written

$$\forall \omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right), \qquad \left(\frac{4}{h^2}\sin^2\left(\frac{\omega h}{2}\right) - \xi_2\omega + \xi_1\right)\widehat{u}(\omega) = \widehat{u} * \widehat{\bar{u}} * \widehat{u}(\omega). \tag{2.50}$$

Applying Lemma 2.3.14 to $f(\omega) = \mathrm{sinc}^2\left(\frac{\omega}{2}\right) + \mathbb{1}_{(-\pi,\pi)^c}(\omega)$, for which $m = \frac{4}{\pi^2}$, there exist $h_0 > 0$ and $\alpha > 0$ such that if $\xi \in \Omega$ and $h < h_0$,

$$\forall \omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right), \qquad \frac{4}{h^2}\sin^2\left(\frac{\omega h}{2}\right) - \xi_2\omega + \xi_1 \geqslant \alpha\left(1 + \omega^2\right).$$

Hence, we have using (2.50)

$$\forall \omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right), \qquad \alpha\left(1 + \omega^2\right)|\widehat{u}(\omega)| \leqslant |\widehat{u}(\omega)| * |\widehat{\bar{u}}(\omega)| * |\widehat{u}(\omega)|. \tag{2.51}$$

Now, we prove by induction (on $n$) that there exists $C > 0$, that only depend of $\alpha$ and $M$ such that

$$\forall\ 1 \leqslant p \leqslant \infty, \qquad \|\omega^n \widehat{u}\|_{L^p(\mathbb{R})} \leqslant C^n n!. \tag{2.52}$$

First, we consider the cases $n = 1$ and $n = 0$. Since we have assumed that $\|u\|_{H^1(\mathbb{R})} \leqslant M$, we have

$$\|\widehat{u}\|_{L^1(\mathbb{R})} \leqslant \left\|\frac{1}{\sqrt{1+\omega^2}}\right\|_{L^2(\mathbb{R})} \|\sqrt{1+\omega^2}\widehat{u}(\omega)\|_{L^2(\mathbb{R})} = \sqrt{2}\pi\|u\|_{H^1(\mathbb{R})} \leqslant \sqrt{2}\pi M.$$

Then, we get from (2.51)

$$\|\omega\widehat{u}\|_{L^1(\mathbb{R})} \leqslant \|(1+\omega^2)\widehat{u}\|_{L^1(\mathbb{R})} \leqslant \frac{1}{\alpha}\|\widehat{u}\|^3_{L^1(\mathbb{R})} \leqslant \frac{\sqrt{8}\pi^3 M^3}{\alpha}.$$

Furthermore, we also get from (2.51),

$$\|(1+|\omega|)\widehat{u}\|_{L^\infty(\mathbb{R})} \leqslant \frac{1}{\alpha}\left\|\frac{1+|\omega|}{1+\omega^2}\right\|_{L^\infty(\mathbb{R})}\|\widehat{u}\|^3_{L^1(\mathbb{R})}.$$

We deduce (2.52) for $n=0$ and $1$ and for the other values of $p$ using Hölder inequality.

Now, we assume that (2.52) is proved for all $0 \leqslant n \leqslant N+1$. We deduce from (2.50) that for all $\omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right)$, we have

$$\begin{aligned}|\omega|^{N+2}|\widehat{u}(\omega)| &\leqslant |\omega|^N(1+\omega^2)|\widehat{u}(\omega)| \\ &\leqslant \frac{1}{\alpha}\left|\omega^N(\widehat{u}*\widehat{\bar{u}}*\widehat{u}(\omega))\right| = \frac{1}{\alpha}\left|\sum_{n_1+n_2+n_3=N}\frac{N!}{n_1!n_2!n_3!}(\omega^{n_1}\widehat{u})*(\omega^{n_2}\widehat{\bar{u}})*(\omega^{n_3}\widehat{u})(\omega)\right|.\end{aligned}$$

We deduce from Young convolution inequality that if $\frac{3}{q} = 2 + \frac{1}{p}$ then

$$\|\omega^{N+2}\widehat{u}\|_{L^p(\mathbb{R})} \leqslant \frac{1}{\alpha}\sum_{n_1+n_2+n_3=N}\frac{N!}{n_1!n_2!n_3!}\prod_{j=1}^{n}\|\omega^{n_j}\widehat{u}\|_{L^q(\mathbb{R})}.$$

Using the induction hypothesis, we obtain

$$\|\omega^{N+2}\widehat{u}\|_{L^p(\mathbb{R})} \leqslant \frac{1}{\alpha}C^N N!\,\#\{(n_1,n_2,n_3) \mid n_1+n_2+n_3=N\} = \frac{1}{2\alpha C^2}C^{N+2}(N+2)!.$$

So, if $C$ is chosen large enough to ensure $2\alpha C^2 \geqslant 1$, we obtain the result by induction.

Choosing $p=\infty$ in (2.52), we get

$$\forall n \in \mathbb{N}, \quad \forall \omega \in \mathbb{R}^*, \quad |\widehat{u}(\omega)| \leqslant \left(\frac{C}{|\omega|}\right)^n n!.$$

But using Stirling formula, we get an universal constant $c>0$ such that $n! \leqslant ce^{-\frac{2n}{3}}n^n$. Consequently, if $|\omega| \geqslant C$ and $n = \lfloor \frac{\omega}{C} \rfloor$, we have $|\widehat{u}(\omega)| \leqslant ce^{-\frac{|\omega|}{2C}}$, and this shows the result. □

In the following lemma, we prove that Lagrange functions are Lyapunov functions for the traveling waves of the homogeneous Hamiltonian. These uniform estimates are discrete versions of the continuous estimates, see for example Proposition 8.8 of [58]. They are the key estimates for applying the energy-momentum method.

**Lemma 2.3.15.** *Let $h_0, C, \rho, \alpha > 0$ be the constants given by Theorem 2.3.1. There exist $r, \beta, h_1 > 0$ such that for all $h < h_1$, all $\xi \in \Omega$, all $u \in BL^2_h \cap \mathrm{Span}(i\eta^h_\xi, \partial_x\eta^h_\xi)$, if $\|u-\eta^h_\xi\|_{H^1(\mathbb{R})} \leqslant r$ and $\|u\|^2_{L^2(\mathbb{R})} = \|\eta^h_\xi\|^2_{L^2(\mathbb{R})}$ then*

$$\beta\|u-\eta^h_\xi\|^2_{H^1(\mathbb{R})} \leqslant \mathscr{L}^h_\xi(u) - \mathscr{L}^h_\xi(\eta^h_\xi). \tag{2.53}$$

*Proof.* Let $h_1 < h_0$ and $\varepsilon > 0$ be such that

$$\forall h < h_1, \quad \forall \xi \in \Omega, \quad \|\eta_\xi^h\|_{L^2(\mathbb{R})}^2 \geqslant \frac{\|\psi_\xi\|_{L^2(\mathbb{R})}^2}{2} \geqslant \frac{\varepsilon}{2}.$$

Let $r \in (0,1)$ be a positive constant that will be determined later.

Since $\eta_\xi^h$ is bounded in $H^1(\mathbb{R})$, uniformly with respect to $\xi \in \Omega$ and $h < h_0$, there exists a constant $M > 0$ such that for all $\xi \in \Omega$, $h < h_0$, $w_1, w_2, w_3 \in BL_h^2$, we have $\|\eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant M$,

$$|\,\mathrm{d}^2\,\mathscr{L}_\xi^h(\eta_\xi^h)(w_1, w_2)| \leqslant M\|w_1\|_{H^1(\mathbb{R})}\|w_2\|_{H^1(\mathbb{R})}$$

and

$$\sup_{\|\eta_\xi^h - w\|_{H^1(\mathbb{R})} \leqslant 1} |\,\mathrm{d}^3\,\mathscr{L}_\xi^h(w)(w_1, w_2, w_3)| \leqslant M\|w_1\|_{H^1(\mathbb{R})}\|w_2\|_{H^1(\mathbb{R})}\|w_3\|_{H^1(\mathbb{R})}.$$

Indeed, the first estimate has been establish in Lemma 2.3.7 and the second is obvious since $\mathrm{d}^3\,\mathscr{L}_\xi^h = \mathrm{d}^3\,\frac{\|\bullet\|_{L^4(\mathbb{R})}^4}{4}$.

Consider $h < h_1$, $\xi \in \Omega$ and $u \in BL_h^2 \cap \mathrm{Span}(i\eta_\xi^h, \partial_x\eta_\xi^h)^{\perp_{L^2}}$ such that $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant r$ and $\|u\|_{L^2(\mathbb{R})}^2 = \|\eta_\xi^h\|_{L^2(\mathbb{R})}^2$. Then we define

$$v = \eta_\xi^h + \left[(u - \eta_\xi^h) - \frac{\eta_\xi^h}{\|\eta_\xi^h\|_{L^2(\mathbb{R})}}\langle u - \eta_\xi^h, \frac{\eta_\xi^h}{\|\eta_\xi^h\|_{L^2(\mathbb{R})}}\rangle_{L^2(\mathbb{R})}\right].$$

By construction, $v - \eta_\xi^h$ belongs to $\mathrm{Span}(i\eta_\xi^h, \partial_x\eta_\xi^h, \eta_\xi^h)^{\perp_{L^2}}$. Furthermore, $v - \eta_\xi^h$ is a second order perturbation of $u - \eta_\xi^h$ because, since $\|u\|_{L^2(\mathbb{R})}^2 = \|\eta_\xi^h\|_{L^2(\mathbb{R})}^2$, we have

$$\langle \eta_\xi^h, u - \eta_\xi^h\rangle_{L^2(\mathbb{R})} = -\frac{1}{2}\|u - \eta_\xi^h\|_{L^2(\mathbb{R})}^2.$$

So, we get

$$\|u - v\|_{H^1(\mathbb{R})} = \frac{\|\eta_\xi^h\|_{H^1(\mathbb{R})}}{2\|\eta_\xi^h\|_{L^2(\mathbb{R})}^2}\|u - \eta_\xi^h\|_{L^2(\mathbb{R})}^2 \leqslant \frac{2M}{\epsilon^2}\|u - \eta_\xi^h\|_{H^1(\mathbb{R})}^2.$$

Now, we can establish our estimate through a Taylor expansion of $\mathscr{L}_\xi^h(u)$ around $\eta_\xi^h$. The first order term vanishes since $\eta_\xi^h$ is a critical point of $\mathscr{L}_\xi^h$. The second order term is controlled

by applying the coercivity estimate of $\mathrm{d}^2\,\mathscr{L}_\xi^h$ (see (2.53)),

$$\begin{aligned}
&\mathscr{L}_\xi^h(u) - \mathscr{L}_\xi^h(\eta_\xi^h)\\
\geqslant&\, \mathrm{d}^2\,\mathscr{L}_\xi^h(\eta_\xi^h)(u-\eta_\xi^h, u-\eta_\xi^h) - M\|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^3\\
=&\, \mathrm{d}^2\,\mathscr{L}_\xi^h(\eta_\xi^h)(v-\eta_\xi^h, v-\eta_\xi^h) - \mathrm{d}^2\,\mathscr{L}_\xi^h(\eta_\xi^h)(u-v, u-v) + 2\,\mathrm{d}^2\,\mathscr{L}_\xi^h(\eta_\xi^h)(u-\eta_\xi^h, u-v)\\
&- M\|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^3\\
\geqslant&\, \alpha\|v-\eta_\xi^h\|_{H^1(\mathbb{R})}^2 - \|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^3\left(M\left(\frac{2M}{\epsilon^2}\right)^2\|u-\eta_\xi^h\|_{H^1(\mathbb{R})} + 2M\frac{2M}{\epsilon^2} + M\right)\\
=&\, \alpha\|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^2 + \alpha\|v-u\|_{H^1(\mathbb{R})}^2 - 2\alpha\langle v-u, u-\eta_\xi^h\rangle_{H^1(\mathbb{R})}\\
&- \|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^3\left(\left(\frac{2M^2}{\epsilon^2}\right)^2\|u-\eta_\xi^h\|_{H^1(\mathbb{R})} + \frac{4M^2}{\epsilon^2} + M\right)\\
\geqslant&\, \|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^2\left[\alpha - \|u-\eta_\xi^h\|_{H^1(\mathbb{R})}\left(2\alpha\frac{2M}{\epsilon^2} + \left(\frac{2M^2}{\epsilon^2}\right)^2\|u-\eta_\xi^h\|_{H^1(\mathbb{R})} + \frac{4M^2}{\epsilon^2} + M\right)\right]\\
\geqslant&\, \|u-\eta_\xi^h\|_{H^1(\mathbb{R})}^2\left[\alpha - r\left(\alpha\frac{4M}{\epsilon^2} + \left(\frac{2M^2}{\epsilon^2}\right)^2 + \frac{4M^2}{\epsilon^2} + M\right)\right].
\end{aligned}$$

Consequently, to prove the theorem, we just need to choose

$$r < \frac{\alpha}{2}\left(\alpha\frac{4M}{\epsilon^2} + \left(\frac{2M^2}{\epsilon^2}\right)^2 + \frac{4M^2}{\epsilon^2} + M\right)^{-1}.$$

□

The previous lemma provides a stability control for the solutions of the homogeneous Hamiltonian system. To apply it, two strong assumptions are required : $u \in \mathrm{Span}(i\eta_\xi^h, \partial_x\eta_\xi^h)$ and $\|u\|_{L^2(\mathbb{R})}^2 = \|\eta_\xi^h\|_{L^2(\mathbb{R})}^2$. If $u$ is close enough to $\eta_\xi^h$ there are two classical tricks to get these assumptions. To fulfill the first condition, the idea is to apply a small gauge transform and a small advection to $u$. We focus on this problem in the two following Lemmas. To satisfy the second assumption, the idea is to modify $\xi_1$. It is the object of the last Theorem of this section.

When $\eta_\xi^h$ is well defined through Theorem 2.3.1, for any $v \in BL_h^2$, we define the matrix $A_{\xi,h}[v]$ by

$$A_{\xi,h}[v] := \begin{pmatrix} \langle i\eta_\xi^h, iv\rangle_{L^2(\mathbb{R})} & -\langle i\eta_\xi^h, \partial_x v\rangle_{L^2(\mathbb{R})} \\ \langle \partial_x\eta_\xi^h, iv\rangle_{L^2(\mathbb{R})} & -\langle \partial_x\eta_\xi^h, \partial_x v\rangle_{L^2(\mathbb{R})} \end{pmatrix}. \tag{2.54}$$

We will explain later why this matrix is very useful, but first we give a technical Lemma.

**Lemma 2.3.16.** *Let $h_0, C, \rho, \alpha > 0$ be the constants given by Theorem 2.3.1. There exists $h_1 < h_0$, $M > 0$ and $\delta > 0$ such that for all $h < h_1$, all $\xi \in \Omega$ and all $v \in BL_h^2$ with $\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} < \delta$, $A_{\xi,h}[v]$ is invertible and $\|(A_{\xi,h}[v])^{-1}\|_\infty \leqslant M$.*

*Proof.* Let $h < h_0$, $\xi \in \Omega$ and $v \in BL_h^2$. Since $v \mapsto A_{\xi,h}[v]$ is a linear map we have

$$A_{\xi,h}[v] = A_{\xi,h}[\eta_\xi^h] + A_{\xi,h}[v - \eta_\xi^h]. \tag{2.55}$$

However, since $\|\eta_\xi^h - \psi_\xi\|_{H^1(\mathbb{R})} \leqslant Ch^2$, $A_{\xi,h}[\eta_\xi^h]$ converges to $G_\xi$, uniformly with respect to $\xi \in \Omega$, as $h$ goes to $0$, where

$$G_\xi = \begin{pmatrix} \|\psi_\xi\|^2_{L^2(\mathbb{R})} & -\langle i\psi_\xi, \partial_x\psi_\xi\rangle_{L^2(\mathbb{R})} \\ \langle i\psi_\xi, \partial_x\psi_\xi\rangle_{L^2(\mathbb{R})} & -\|\partial_x\psi_\xi\|^2_{L^2(\mathbb{R})} \end{pmatrix}.$$

Applying the Cauchy-Schwarz inequality, we have

$$\det G_\xi = \langle i\psi_\xi, \partial_x\psi_\xi\rangle^2_{L^2(\mathbb{R})} - \|\psi_\xi\|^2_{L^2(\mathbb{R})}\|\partial_x\psi_\xi\|^2_{L^2(\mathbb{R})} \leqslant 0.$$

But the case of equality is excluded since $\psi_\xi$ is not a plane wave (i.e. $\mathrm{Span}(i\psi_\xi, \partial_x\psi_\xi)$ is a free family). So $G_\xi$ is an invertible matrix. As $\xi \mapsto G_\xi$ is a continuous map on $\overline{\Omega}$, there exists $M > 0$ such that for all $\xi \in \Omega$

$$\|G_\xi^{-1}\|_\infty \leqslant \frac{M}{2}.$$

As $A_{\xi,h}[\eta_\xi^h]$ converges to $G_\xi$ when $h \to 0$, there exists $h_1 < h_0$ such that for all $h < h_1$ and $\xi \in \Omega$, $A_{\xi,h}[\eta_\xi^h]$ is invertible and

$$\|(A_{\xi,h}[\eta_\xi^h])^{-1}\|_\infty \leqslant M.$$

Applying the linear decomposition (2.55), we have

$$A_{\xi,h}[v] = A_{\xi,h}[\eta_\xi^h](I_2 + (A_{\xi,h}[\eta_\xi^h])^{-1}A_{\xi,h}[v - \eta_\xi^h]).$$

However, since $\eta_\xi^h$ is bounded in $H^1(\mathbb{R})$ uniformly with respect to $\xi$ and $h$, there exists $\delta > 0$ such that for all $\xi \in \Omega$ and all $h < h_1$, we have

$$\|(A_{\xi,h}[\eta_\xi^h])^{-1}A_{\xi,h}[v - \eta_\xi^h]\|_\infty < \frac{1}{2\delta}\|v - \eta_\xi^h\|_{H^1(\mathbb{R})}.$$

Consequently, if $\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant \delta$ then $A_{\xi,h}[v]$ is invertible and the norm of its inverse is bounded by $2M$. □

**Lemma 2.3.17.** *There exists $\lambda, \delta > 0$ and $h_1 < h_0$, such that for all $\xi \in \Omega$, $h < h_1$, $v \in BL^2_h$, if $\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} < \delta$ then there exists $\gamma, \mathrm{y} \in \mathbb{R}$ such that*

$$\max(|\gamma|, |\mathrm{y}|) \leqslant \lambda\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} \quad \textit{and} \quad e^{i\gamma}v(\bullet - \mathrm{y}) - \eta_\xi^h \in \mathrm{Span}(i\eta_\xi^h, \partial_x\eta_\xi^h)^{\perp_{L^2}}.$$

*Proof.* For this proof, we introduce a notation. If $\gamma, \mathrm{y} \in \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ then

$$T_{\gamma,\mathrm{y}}v := e^{i\gamma}v(\bullet - \mathrm{y})$$

Let $v \in BL^2_h$. We are going to apply the inverse function Theorem 2.5.3 to the following function

$$g^v_{\xi,h} : \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R}^2 \\ \begin{pmatrix}\gamma \\ \mathrm{y}\end{pmatrix} & \mapsto & \begin{pmatrix} \langle i\eta_\xi^h, T_{\gamma,\mathrm{y}}v - \eta_\xi^h\rangle_{L^2(\mathbb{R})} \\ \langle \partial_x\eta_\xi^h, T_{\gamma,\mathrm{y}}v - \eta_\xi^h\rangle_{L^2(\mathbb{R})} \end{pmatrix} \end{array} \right. .$$

$g^v_{\xi,h}$ is clearly a $C^1$ function whose Jacobian matrix is given by

$$\mathrm{J}g^v_{\xi,h}(\gamma, \mathrm{y}) = A_{\xi,h}[T_{\gamma,\mathrm{y}}v].$$

Applying Lemma 2.3.16, we can find $h_1 < h_0$, $\delta > 0$ and $M > 0$ such that if $h < h_1$ and $\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} < \delta$ then $\mathrm{J}g_{\xi,h}^v(0,0)$ is invertible and its norm is smaller than $M$. We want to prove that $\mathrm{J}g_{\xi,h}^v$ is Lipschitz uniformly with respect to $\xi, h, v$. In fact, since it is a $C^1$ function, we just need to control its derivative. Using integration by parts, there exists a constant $\kappa > 0$ such that for all $\mathrm{y}, \gamma \in \mathbb{R}$ we have

$$\|\,\mathrm{d}\,\mathrm{J}g_{\xi,h}^v(\gamma,\mathrm{y})\|_{\mathscr{L}(\mathbb{R}^2;M_2(\mathbb{R}^2))} \leqslant \kappa\|\eta_\xi^h\|_{H^2(\mathbb{R})}\|T_{\gamma,\mathrm{y}}v\|_{H^1(\mathbb{R})} = \kappa\|\eta_\xi^h\|_{H^2(\mathbb{R})}\|v\|_{H^1(\mathbb{R})}.$$

But, applying the result of elliptic regularity (Theorem 2.3.13), $\|\eta_\xi^h\|_{H^2(\mathbb{R})}$ is bounded in $H^2(\mathbb{R})$ uniformly with respect to $\xi \in \Omega$ and $h < h_0$. So, there exists $k > 0$ such that for all $\xi \in \Omega$, $h < h_0$ and $v \in BL_h^2$ with $\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} < \delta$, we have

$$\|\,\mathrm{d}\,\mathrm{J}g_{\xi,h}^v(\gamma,\mathrm{y})\|_{\mathscr{L}(\mathbb{R}^2;M_2(\mathbb{R}^2))} \leqslant k.$$

Now, we apply the inverse function theorem 2.5.3 to $g_{\xi,h}^v$ and we obtain some constants $\lambda > 0$ and $r > 0$, such that for all $h < h_1$, $\xi \in \Omega$ and $v \in BL_h^2$ with $\|v - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant R$,

$$\forall \nu \in \mathbb{R}^2, \quad |\nu| \leqslant r \;\Rightarrow\; \exists \gamma, \mathrm{y} \in \mathbb{R},\; g_{\xi,h}^v(\gamma,\mathrm{y}) = g_{\xi,h}^v(0,0) + \nu \quad \text{and} \quad \max(|\gamma|, |\,\mathrm{y}\,|) \leqslant \lambda|\nu|.$$

To prove the lemma, we would like to choose $\nu = -g_{\xi,h}^v(0,0)$ small enough. But since $\eta_\xi^h$ is uniformly bounded in $H^1(\mathbb{R})$, there exists a constant $K > 0$ such that for all $h < h_0$, $v \in BL_h^2$, $\xi \in \Omega$,

$$|g_{\xi,h}^v(0,0)| \leqslant K\|\eta_\xi^h - v\|_{H^1(\mathbb{R})}.$$

So, if $\|\eta_\xi^h - v\|_{H^1(\mathbb{R})} \leqslant \frac{r}{K}$, we can choose $\nu = -g_{\xi,h}^v(0,0)$ and the lemma is proven. □

In the following Theorem, we focus on a change of variable. Usually, NLS traveling waves are not indexed by $\xi$ but by their $L^2$ norm and their momentum. It would be possible to do the same here. Here, we prove that it is possible to index them by their $L^2$ norm and their speed of advection (i.e. $\xi_2$).

**Theorem 2.3.18.** *Let $h_0, C, \rho, \alpha > 0$ be the constants given by Theorem 2.3.1 and let $\widetilde{\Omega}$ be a relatively compact open subset of $\Omega$. Then there exist $h_1 < h_0, \delta > 0, k > 0$ such that for all $h < h_1$, for all $\xi \in \widetilde{\Omega}$ and for all $u \in BL_h^2$, if $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} < \delta$ then there exists $\zeta \in \Omega$ such that*

$$\left\{\begin{array}{lcl} \xi_2 & = & \zeta_2 \\ \|\eta_\zeta^h\|_{L^2(\mathbb{R})}^2 & = & \|u\|_{L^2(\mathbb{R})}^2 \end{array}\right. \qquad \textit{and} \qquad |\zeta - \xi| \leqslant k|\|\eta_\xi^h\|_{L^2(\mathbb{R})}^2 - \|u\|_{L^2(\mathbb{R})}^2|. \tag{2.56}$$

*Proof.* From the definition of $\psi_\xi$ (see (2.4)), we observe that for all $\xi \in \Omega$,

$$\|\psi_\xi\|_{L^2(\mathbb{R})}^2 = m_\xi\|\psi_{1,0}\|_{L^2(\mathbb{R})}^2 = 4m_\xi = 4\sqrt{\xi_1^2 - \left(\frac{\xi_2}{2}\right)^2}.$$

Consequently, there exists $\beta > 0$ such that for all $\xi \in \Omega$,

$$\partial_{\xi_1}\|\psi_\xi\|_{L^2(\mathbb{R})}^2 = \frac{2}{m_\xi} \geqslant 2\beta.$$

Let $h < h_0$. Applying Theorem 2.3.1, we know that $\xi \mapsto \eta_\xi^h$ is a $C^1$ approximation of $\xi \mapsto \psi_\xi$ up to an second order error term. Consequently, we have

$$\begin{aligned}
|\partial_{\xi_1}\|\psi_\xi\|^2_{L^2(\mathbb{R})} - \partial_{\xi_1}\|\eta_\xi^h\|^2_{L^2(\mathbb{R})}| &= 2|\langle \partial_{\xi_1}\psi_\xi - \partial_{\xi_1}\eta_\xi^h, \psi_\xi\rangle_{L^2(\mathbb{R})} + \langle \partial_{\xi_1}\eta_\xi^h, \psi_\xi - \eta_\xi^h\rangle_{L^2(\mathbb{R})}| \\
&\leqslant 2Ch^2\left(\|\psi_\xi\|_{L^2(\mathbb{R})} + \|\partial_{\xi_1}\eta_\xi^h\|_{L^2(\mathbb{R})}\right) \\
&\leqslant 2Ch^2\left(\|\psi_\xi\|_{L^2(\mathbb{R})} + Ch^2 + \|\partial_{\xi_1}\psi_\xi\|_{L^2(\mathbb{R})}\right) \\
&\leqslant 2Ch^2 \sup_{\xi\in\Omega}\left(\|\psi_\xi\|_{L^2(\mathbb{R})} + Ch_0^2 + \|\partial_{\xi_1}\psi_\xi\|_{L^2(\mathbb{R})}\right) \\
&=: Mh^2.
\end{aligned}$$

Let $h_1 = \min(h_0, \beta\sqrt{M})$. If $h < h_0$ and $\xi \in \Omega$, we have

$$\partial_{\xi_1}\|\eta_\xi^h\|^2_{L^2(\mathbb{R})} \geqslant \beta.$$

Since $\widetilde{\Omega}$ is relatively compact open subset of $\Omega$, there exists $r > 0$ such that

$$\widetilde{\Omega} + \overline{B_{\mathbb{R}^2}(0,r)} \subset \Omega.$$

Let $\xi \in \widetilde{\Omega}$, $h < h_1$ and let $g$ be the following function

$$g : \begin{cases} [\xi_1 - r, \xi_1 + r] & \to & \mathbb{R} \\ \zeta_1 & \mapsto & \|\eta^h_{\zeta_1,\xi_2}\|^2_{L^2(\mathbb{R})}. \end{cases}$$

Since $g$ is a continuous map, we have

$$[\|\eta^h_{\xi_1-r,\xi_2}\|^2_{L^2(\mathbb{R})}, \|\eta^h_{\xi_1+r,\xi_2}\|^2_{L^2(\mathbb{R})}] \subset g([\xi_1 - r, \xi_1 + r]). \tag{2.57}$$

But applying the mean value equality, we have

$$\|\eta^h_{\xi_1-r,\xi_2}\|^2_{L^2(\mathbb{R})} < \|\eta_\xi^h\|^2_{L^2(\mathbb{R})} - \beta r < \|\eta_\xi^h\|^2_{L^2(\mathbb{R})} + \beta r < \|\eta^h_{\xi_1+r,\xi_2}\|^2_{L^2(\mathbb{R})}.$$

Let $u \in BL^2_h$ be such that $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} < \delta$, where $\delta \in (0,1)$ is a positive constant that will be fixed later. Applying the triangle inequality, we get

$$\begin{aligned}
\left|\|u\|^2_{L^2(\mathbb{R})} - \|\eta_\xi^h\|^2_{L^2(\mathbb{R})}\right| &\leqslant \delta(\|u\|_{L^2(\mathbb{R})} + \|\eta_\xi^h\|_{L^2(\mathbb{R})}) \\
&\leqslant \delta(\delta + 2\|\eta_\xi^h\|_{L^2(\mathbb{R})}) \\
&\leqslant \delta(1 + 2\sup_{\xi\in\Omega,\ h<h_0}\|\eta_\xi^h\|_{L^2(\mathbb{R})}) \\
&=: \delta\kappa.
\end{aligned}$$

So, choosing $\delta = \frac{\beta r}{\kappa}$, we deduce from (2.57) that there exists $\zeta_1 \in [\xi_1 - r, \xi_1 + r]$ such that

$$\|u\|^2_{L^2(\mathbb{R})} = g(\zeta_1) = \|\eta_\zeta^h\|^2_{L^2(\mathbb{R})},$$

where $\zeta_2 := \xi_2$. Applying the mean value equality, we obtain

$$|\xi - \zeta| \leqslant \beta^{-1}\left|\|u\|^2_{L^2(\mathbb{R})} - \|\eta_\zeta^h\|^2_{L^2(\mathbb{R})}\right|.$$

which proves the result. □

## 2.4 Control of the instabilities and modulation

In the last section we have constructed approximate traveling waves $\eta_\xi^h$. In order to prove Theorem 2.1.4, now, we study the dynamics of DNLS around these approximate traveling waves.

We are going to use many results established in the previous section about $\eta_\xi^h$ and its properties. In a first paragraph, we summarize the results that will be useful and fix most of the constants.

Step 1 : variational properties around the equilibria

Let $\widetilde{\Omega}$ be a relatively compact open subset of $\left\{\xi \in \mathbb{R}^2 \mid \xi_1 > \left(\frac{\xi_2}{2}\right)^2\right\}$ and $\Omega$ a relatively compact open subset of $\widetilde{\Omega}$. In the previous section, we have proved that there exist some constants $h_0, \varepsilon, C, \rho > 0$ and, for all $\xi \in \widetilde{\Omega}$ and all $h < h_0$, a function $\eta_\xi^h \in BL_h^2$ satisfying the following properties.

- From Theorem 2.3.1, $\eta_\xi^h$ is a critical point of $\mathscr{L}_\xi^h$ and it is an approximation of $\psi_\xi$

$$\|\eta_\xi^h - \psi_\xi\|_{H^1(\mathbb{R})} \leqslant Ch^2.$$

- From Theorem 2.3.13, $\eta_\xi^h$ is regular function

$$\forall \omega \in \mathbb{R},\ |\widehat{\eta_\xi^h}(\omega)| \leqslant Ce^{-\varepsilon|\omega|}. \tag{2.58}$$

  Consequently, we also have $\|\widehat{\eta_\xi^h}\|_{H^3(\mathbb{R})} \leqslant C$.
- From Lemma 2.3.15, if $u \in BL_h^2 \cap \mathrm{Span}(i\eta_\xi^h, \partial_x\eta_\xi^h)^{\perp_{L^2}}$, $\|u\|_{L^2(\mathbb{R})}^2 = \|\eta_\xi^h\|_{L^2(\mathbb{R})}^2$ and $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant \rho$ then

$$\frac{1}{C}\|u - \eta_\xi^h\|_{H^1(\mathbb{R})}^2 \leqslant \mathscr{L}_\xi^h(u) - \mathscr{L}_\xi^h(\eta_\xi^h). \tag{2.59}$$

- From Theorem 2.3.18, if $u \in BL_h^2(\mathbb{R})$, $\xi \in \Omega$ and $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant \rho$ then there exists $\zeta \in \widetilde{\Omega}$ such that

$$\left\{\begin{array}{lcl} \xi_2 & = & \zeta_2 \\ \|\eta_\zeta^h\|_{L^2(\mathbb{R})}^2 & = & \|u\|_{L^2(\mathbb{R})}^2 \end{array}\right. \tag{2.60}$$

  and (using the regularity of $\xi \mapsto \eta_\xi^h$ uniformly with respect to $h$, see Theorem 2.3.1)

$$|\zeta - \xi| + \|u - \eta_\zeta^h\|_{H^1(\mathbb{R})} \leqslant C\|u - \eta_\xi^h\|_{H^1(\mathbb{R})}. \tag{2.61}$$

- From Lemma 2.3.17, for all $u \in BL_h^2(\mathbb{R})$, if $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant \rho$ then there exists $\gamma, \mathrm{y} \in \mathbb{R}$ such that

$$\max(|\gamma|, |\mathrm{y}|) \leqslant C\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} \quad \text{and} \quad e^{i\gamma}u(\bullet - \mathrm{y}) - \eta_\xi^h \in \mathrm{Span}(i\eta_\xi^h, \partial_x\eta_\xi^h)^{\perp_{L^2}}. \tag{2.62}$$

- From Lemma 2.3.16, for all $u \in BL_h^2(\mathbb{R})$, if $\|u - \eta_\xi^h\|_{H^1(\mathbb{R})} \leqslant \rho$ and $A_{h,\xi}[u]$ is the matrix defined in (2.54) then

$$A_{h,\xi}[u] \text{ is is invertible and } \|(A_{h,\xi}[u])^{-1}\|_1 \leqslant C. \tag{2.63}$$

- From Lemma 2.3.7 and Lemma 2.3.9, for all $u \in BL^2_h(\mathbb{R})$, if $\|u - \eta^h_\xi\|_{H^1(\mathbb{R})} \leqslant \rho$ then

$$\forall v, w \in BL^2_h, \quad \left|\mathrm{d}^2 \mathscr{L}^h_\xi(u)(v,w)\right| \leqslant C\|v\|_{H^1(\mathbb{R})}\|w\|_{H^1(\mathbb{R})}. \tag{2.64}$$

We finish this paragraph by a remark. In Theorem 2.1.4, we compare a solution $\boldsymbol{u}$ of DNLS with some discretizations of $\eta^h_\xi$ using discrete Sobolev norms. However, as we explain in Lemma 2.2.5, it is equivalent to compare directly the Shannon interpolation $u$ of the discrete solution with $\eta^h_\xi$ using continuous Sobolev norms.

Step 2 : Lyapunov estimation and modulation

Let $r > 0$ be a positive constant independent of $\xi$ and $h$ that will be determined at the end of this paragraph. Recall that for $v : \mathbb{R} \to \mathbb{R}$ we have

$$\forall x \in \mathbb{R}, \quad T_{\gamma,\mathrm{y}} v(x) := e^{i\gamma} v(x - \mathrm{y}),$$

and note that $T^{-1}_{\gamma,\mathrm{y}} = T_{-\gamma,-\mathrm{y}}$. Let $u_0 \in BL^2_h$ be such that $\delta(0) = \|u_0 - T_{\gamma_0,\mathrm{y}_0}\eta^h_\xi\|_{H^1(\mathbb{R})} < r$ where $\xi \in \Omega$, $\mathrm{y}_0, \gamma_0 \in \mathbb{R}$. Let $u$ be the solution of DNLS in $BL^2_h$ (see Lemma 2.2.4) such that $u(0) = u_0$.

Assume that $r < \rho$. Applying (2.60) and (2.61), there exists $\zeta \in \widetilde{\Omega}$ such that

$$\left\{ \begin{array}{lcl} \xi_2 & = & \zeta_2 \\ \|\eta^h_\zeta\|^2_{L^2(\mathbb{R})} & = & \|u_0\|^2_{L^2(\mathbb{R})} \end{array} \right. \quad \text{and} \quad |\zeta - \xi| + \|u_0 - T_{\gamma_0,\mathrm{y}_0}\eta^h_\xi\|_{H^1(\mathbb{R})} \leqslant C\delta(0).$$

Consequently, we have

$$\|\eta^h_\xi - \eta^h_\zeta\|_{H^1(\mathbb{R})} \leqslant (1 + C)\delta(0).$$

Now, assume that $Cr < \rho$, then applying (2.62), there exist $\delta_\gamma, \delta_\mathrm{y} \in \mathbb{R}$ such that

$$\left\{ \begin{array}{lcl} \theta_0 & = & \gamma_0 + \delta_\gamma \\ p_0 & = & \mathrm{y}_0 + \delta_\mathrm{y} \end{array} \right. \quad \text{with} \quad \max(|\delta_\gamma|, |\delta_\mathrm{y}|) \leqslant C^2\delta(0) \quad \text{and} \quad T^{-1}_{\theta_0,p_0} u_0 \in \mathrm{Span}(i\eta^h_\zeta, \partial_x \eta^h_\zeta)^{\perp_{L^2}}. \tag{2.65}$$

We would like to get some functions $\theta, p \in C^1(\mathbb{R}_+)$ such that as long as $u(t)$ is close to the orbit of $\eta^h_\zeta$ (up to gauge transform and advection), we have $T^{-1}_{\theta(t),p(t)} u(t) \in \mathrm{Span}(i\eta^h_\zeta, \partial_x \eta^h_\zeta)^{\perp_{L^2}}$. We are going to construct them by solving a differential equation. Taking a time derivative, if such functions exist they have to satisfy

$$A_{\zeta,h}[T^{-1}_{\theta(t),p(t)} u(t)] \begin{pmatrix} \dot\theta(t) \\ \dot p(t) \end{pmatrix} = \begin{pmatrix} \langle T^{-1}_{\theta(t),p(t)} \partial_t u(t), i\eta^h_\zeta \rangle_{L^2(\mathbb{R})} \\ \langle T^{-1}_{\theta(t),p(t)} \partial_t u(t), \partial_x \eta^h_\zeta \rangle_{L^2(\mathbb{R})} \end{pmatrix}. \tag{2.66}$$

We would like to solve the Cauchy problem associated with this ordinary differential equation with $\theta(0) = \theta_0$ and $p(0) = p_0$. Note that all the terms depend smoothly on $t, p(t), \theta(t)$, hence to get the existence of a local solution, we need to invert $A_{\zeta,h}[T^{-1}_{\theta(t),p(t)} u(t)]$. Using the regularity of $\eta^h_\zeta$ (see (2.58)), we have

$$\|u_0 - T_{\theta_0,p_0}\eta^h_\zeta\|_{H^1(\mathbb{R})} \leqslant C^3\delta(0).$$

Assuming that $C^3 r < \rho$, we get from (2.63) that $A_{\zeta,h}[T^{-1}_{\theta_0,p_0} u_0]$ is invertible and

$$\|(A_{\zeta,h}[T^{-1}_{\theta_0,p_0} u_0])^{-1}\|_1 \leqslant C.$$

Thus (applying, for example, Cauchy-Lipschitz Theorem or the implicit functions Theorem), there exist $T_{\max} \in (0, \infty]$ and a solution $\theta, p \in C^1([0, T_{\max}))$ of (2.66) on $[0, T_{\max})$ such that

- $\theta(0) = \theta_0$ and $p(0) = p_0$,
- for all $t \in [0, T_{\max})$, $A_{\zeta,h}[T^{-1}_{\theta(t),p(t)}u(t)]$ is invertible,
- $\lim\limits_{t\to T_{\max}} |\theta(t)| + |p(t)| + \|(A_{\zeta,h}[T^{-1}_{\theta(t),p(t)}u(t)])^{-1}\|_1 = \infty$

We would like to prove that while $\|u(t) - T_{\gamma(t),\mathrm{y}(t)}\eta^h_\xi\|_{H^1(\mathbb{R})} < r$, with $\gamma = \theta - \delta_\gamma$ and $\mathrm{y} = p - \delta_\mathrm{y}$ where $\delta_\gamma$ and $\delta_\mathrm{y}$ are given in (2.65), the last condition is not satisfied and so $\gamma(t)$ and $\mathrm{y}(t)$ are well defined. This is done by the following Lemma, whose proof is given in Section 2.5.2 of the Appendix.

**Lemma 2.4.1.** *There exist* $\gamma, \mathrm{y} \in C^1(\mathbb{R}_+)$ *such that* $\gamma(0) = \gamma_0$, $\mathrm{y}(0) = \mathrm{y}_0$ *and if* $T > 0$ *satisfies*

$$\forall t \in (0,T), \qquad \|u(t) - T_{\gamma(t),\mathrm{y}(t)}\eta^h_\xi\|_{H^1(\mathbb{R})} < r,$$

*then* $T < T_{\max}$ *and* $\gamma = \theta - \delta_\gamma$, $\mathrm{y} = p - \delta_\mathrm{y}$ *on* $(0,T)$*, where* $\delta_\gamma$ *and* $\delta_\mathrm{y}$ *are defined in* (2.65)*.*

From now on, we consider the functions $\gamma, \mathrm{y}$ given by Lemma 2.4.1 and $T > 0$ satisfying the bootstrap condition

$$\forall t \in (0,T), \qquad \delta(t) := \|u(t) - T_{\gamma(t),\mathrm{y}(t)}\eta^h_\xi\|_{H^1(\mathbb{R})} < r.$$

By construction, we have

$$\begin{aligned}\|u(t) - T_{\theta(t),p(t)}\eta^h_\zeta\|_{H^1(\mathbb{R})} &\leqslant \|u(t) - T_{\gamma(t),\mathrm{y}(t)}\eta^h_\xi\|_{H^1(\mathbb{R})} + \|\eta^h_\xi - T_{\delta_\gamma,\delta_\mathrm{y}}\eta^h_\xi\|_{H^1(\mathbb{R})} + \|\eta^h_\zeta - \eta^h_\xi\|_{H^1(\mathbb{R})}\\ &\leqslant \delta(t) + C^3\delta(0) + (1+C)\delta(0)\\ &< (2 + C + C^3)r.\end{aligned}$$

We assume that $(2 + C + C^3)r \leqslant \rho$. Since $\|u\|^2_{L^2(\mathbb{R})}$ is a constant of the motion, we have $\|u(t)\|^2_{L^2(\mathbb{R})} = \|\eta^h_\zeta\|^2_{L^2(\mathbb{R})}$. Furthermore, by construction $T^{-1}_{\theta(t),p(t)}u \in \mathrm{Span}(i\eta^h_\zeta, \partial_x\eta^h_\zeta)^{\perp_{L^2}}$, so we can apply (2.59) to get the Lyapunov control of the stability

$$\frac{1}{C}\|u(t) - T_{\theta(t),p(t)}\eta^h_\zeta\|^2_{H^1(\mathbb{R})} \leqslant \mathscr{L}^h_\zeta(u(t)) - \mathscr{L}^h_\zeta(\eta^h_\zeta). \tag{2.67}$$

To be rigorous, we can verify our assumptions on $r$ and observe that $r = \frac{\rho}{2+C+C^3}$ is a possible choice.

Step 3 : Estimation of $\delta(t)$

Usually, when we apply the energy-momentum method, the Lagrange function is a constant of the motion of DNLS. An estimate of the form (2.67) allows to control $\|u(t) - T_{\theta(t),p(t)}\eta^h_\zeta\|^2_{H^1(\mathbb{R})}$ by $\mathscr{L}^h_\zeta(u_0) - \mathscr{L}^h_\zeta(\eta^h_\zeta)$. This latter quantity can be controlled by using a Taylor expansion

$$\begin{aligned}\mathscr{L}^h_\zeta(u_0) - \mathscr{L}^h_\zeta(\eta^h_\zeta) &= \mathscr{L}^h_\zeta(T^{-1}_{\theta_0,p_0}u_0) - \mathscr{L}^h_\zeta(\eta^h_\zeta)\\ &\leqslant \frac{1}{2}\sup_{\|v-\eta^h_\zeta\|_{H^1(\mathbb{R})}\leqslant\rho}\left|\mathrm{d}^2\,\mathscr{L}^h_\zeta(v)(T^{-1}_{\theta_0,p_0}u_0 - \eta^h_\zeta)\right|\\ &\leqslant \frac{C}{2}\|u_0 - T_{\theta_0,p_0}\eta^h_\zeta\|^2_{H^1(\mathbb{R})},\end{aligned}$$

where the last estimate is given by (2.64).

In our case, because of the aliasing terms, $\mathscr{L}^h_\zeta(u(t))$ is not a constant of the motion. So we have to control its variations. Let $t < T$, since $H^h_{\mathrm{DNLS}}(u(t))$ and $\|u(t)\|^2_{L^2(\mathbb{R})}$ are constant of the motion, applying the formula of Lemma 2.30, we obtain the following decomposition

$$\begin{aligned}\mathscr{L}^h_\zeta(u(t)) - \mathscr{L}^h_\zeta(u(0)) =& H^h_{\mathrm{DNLS}}(u(t)) - H^h_{\mathrm{DNLS}}(u(0)) + \frac{\zeta_1}{2}\left(\|u(t)\|^2_{L^2(\mathbb{R})} - \|u(0)\|^2_{L^2(\mathbb{R})}\right)\\ &- \frac{1}{2}\int_{\mathbb{R}} \cos\left(\frac{2\pi x}{h}\right)(|u(t,x)|^4 - |u(0,x)|^4)\,\mathrm{dx}\\ &+ \frac{\zeta_2}{2}\left(\langle i\partial_x u(t), u(t)\rangle_{L^2(\mathbb{R})} - \langle i\partial_x u(0), u(0)\rangle_{L^2(\mathbb{R})}\right)\\ =& E_1(0) - E_1(t) + \frac{1}{2}E_2(t),\end{aligned} \tag{2.68}$$

where

$$E_1(t) = \frac{1}{2}\int_{\mathbb{R}} \cos\left(\frac{2\pi x}{h}\right)|u(t,x)|^4\,\mathrm{dx}$$

and

$$E_2(t) = \xi_2\left(\langle i\partial_x u(t), u(t)\rangle_{L^2(\mathbb{R})} - \langle i\partial_x u(0), u(0)\rangle_{L^2(\mathbb{R})}\right).$$

Note that we write $\xi_2$ instead of $\zeta_2$ as these two numbers are equal by construction (see (2.60)).

First, we explain how to bound $E_1(t)$. It can be decomposed as follow

$$E_1(t) = \frac{1}{4}\left(E_3(u(t)) + \overline{E_3(u(t))}\right), \quad \text{with} \quad E_3(v) = \int_{\mathbb{R}} e^{\frac{2i\pi}{h}}|u(t,x)|^4\,\mathrm{dx}.$$

Since $E_3$ is a $4-$homogeneous continuous function, its Taylor expansion is exact. So, we have

$$E_3(u(t)) = \sum_{j=0}^{4}\frac{1}{j!}\,\mathrm{d}^j\,E_3(T_{\theta(t),p(t)}\eta^h_\zeta)(\underbrace{u(t) - T_{\theta(t),p(t)}\eta_{\zeta,h}, \dots, u(t) - T_{\theta(t),p(t)}\eta_{\zeta,h}}_{j \text{ times}}). \tag{2.69}$$

To control these derivatives, we use the following lemma.

**Lemma 2.4.2.** *If* $u_1, u_2, u_3, u_4 \in BL^2_h$ *and*

$$M_h(u_1,u_2,u_3,u_4) = \int_{\mathbb{R}} e^{\frac{2i\pi x}{h}}u_1(x)u_2(x)u_3(x)u_4(x)\,\mathrm{dx},$$

*then we have*

$$|M_h(u_1,u_2,u_3,u_4)| \leqslant \frac{1}{4}\sum_{\sigma\in S_4}\|\widehat{u_{\sigma_1}}\mathbb{1}_{\omega\geqslant\frac{\pi}{3h}}\|_{L^2(\mathbb{R})}\|\widehat{u_{\sigma_2}}\mathbb{1}_{\omega\geqslant\frac{\pi}{3h}}\|_{L^2(\mathbb{R})}\|\widehat{u_{\sigma_3}}\|_{L^1(\mathbb{R})}\|\widehat{u_{\sigma_4}}\|_{L^1(\mathbb{R})}.$$

*Proof.* We identify $M_h$ with a convolution product

$$M_h(u_1,u_2,u_3,u_4) = \widehat{u_1} * \widehat{u_2} * \widehat{u_3} * \widehat{u_4}(\frac{2\pi}{h}).$$

But *if the sum of four numbers, all smaller than* $1$*, is equals to* $2$*, then at least* $2$ *of them are larger than* $\frac{1}{3}$. Consequently, since $\operatorname{supp}\widehat{u_j} \subset \left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$, it comes

$$|M_h(u_1,u_2,u_3,u_4)| \leqslant \frac{1}{4}\sum_{\sigma\in S_4}|\mathbb{1}_{\omega\geqslant\frac{\pi}{3h}}\widehat{u_{\sigma_1}}| * |\mathbb{1}_{\omega\geqslant\frac{\pi}{3h}}\widehat{u_{\sigma_2}}| * |\widehat{u_{\sigma_3}}| * |\widehat{u_{\sigma_4}}|(\frac{2\pi}{h}).$$

Then, we conclude the proof using Young convolution inequalities. □

Applying this Lemma to estimate the terms of (2.69) we obtain four types of contributions.

- Applying (2.58) and defining $\ell = \frac{\pi\varepsilon}{3}$, we have

$$\|\mathscr{F}[T_{\theta(t),p(t)}\eta_\zeta^h]\mathbb{1}_{\omega\geqslant\frac{\pi}{3h}}\|_{L^2(\mathbb{R})}^2 \leqslant C^2\int_{\omega\geqslant\frac{\pi}{3h}} e^{-2\varepsilon\omega}\,\mathrm{d}\omega = \frac{C^2}{\varepsilon}e^{-2\varepsilon\frac{\pi}{3h}} = \frac{C^2}{\varepsilon}e^{-\frac{2\ell}{h}}.$$

- Up to an universal constant $c>0$, we have

$$\|\mathscr{F}[T_{\theta(t),p(t)}\eta_\zeta^h]\|_{L^1(\mathbb{R})} \leqslant c\,C.$$

- Up to an universal constant $c>0$, we have

$$\begin{aligned}\|\mathscr{F}[u(t)-T_{\theta(t),p(t)}\eta_\zeta^h]\mathbb{1}_{\omega\geqslant\frac{\pi}{3h}}\|_{L^2(\mathbb{R})} &\leqslant \frac{3h}{\pi}\|\mathscr{F}[u(t)-T_{\theta(t),p(t)}\eta_\zeta^h]\;\omega\|_{L^2(\mathbb{R})}\\ &\leqslant ch\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}.\end{aligned}$$

- Up to an universal constant $c>0$, we have

$$\|\mathscr{F}[u(t)-T_{\theta(t),p(t)}\eta_\zeta^h]\|_{L^1(\mathbb{R})} \leqslant c\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}.$$

  Sometimes, it is also useful to control it by $c\rho$.

With these estimates, we get a constant $M>0$ (depending only of $\varepsilon, C, \rho, h_0$) such that

$$|E_3(u(t))| \leqslant 2Me^{-\frac{\ell}{h}} + 2Mh^2\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}^2. \tag{2.70}$$

So we deduce that

$$|E_1(t)| \leqslant Me^{-\frac{\ell}{h}} + h^2M\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}^2. \tag{2.71}$$

Now, we show how to control the term $E_2$ in (2.68). It is precisely the error generated by the default of invariance by advection. First, we give a more adapted expression of $E_2$ :

$$\begin{aligned}E_2(t) = \xi_2\int_0^t \partial_s\langle i\partial_x u(s), u(s)\rangle_{L^2(\mathbb{R})}\,\mathrm{ds} &= 2\xi_2\int_0^t\langle i\partial_x u(s), \partial_s u(s)\rangle_{L^2(\mathbb{R})}\,\mathrm{ds}\\ &= -4\xi_2\int_0^t\langle\partial_x u(s), \cos\left(\frac{2\pi x}{h}\right)|u(s)|^2u(s)\rangle_{L^2(\mathbb{R})}\,\mathrm{ds}\\ &= -\xi_2\frac{2\pi}{h}\int_0^t\int_{\mathbb{R}}\sin\left(\frac{2\pi x}{h}\right)|u(s,x)|^4\,\mathrm{dx}\,\mathrm{ds} = -\xi_2\frac{\pi}{h}\int_0^t E_3(u(s)) - \overline{E_3(u(s))}\,\mathrm{ds}\,.\end{aligned}$$

Applying Estimate of $E_3(u(s))$ (2.70), we obtain

$$|E_2(t)| \leqslant 4M\pi|\xi_2|h\int_0^t \frac{e^{-\frac{\ell}{h}}}{h^2} + \|u(s)-T_{\theta(s),p(s)}\eta_\zeta^h\|_{H^1(\mathbb{R})}^2\,\mathrm{ds}\,.$$

Finally, we apply estimate (2.67) and we get

$$\begin{aligned}
&\frac{1}{C}\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}\\
\leqslant&\mathscr{L}^h_\xi(u(t))-\mathscr{L}^h_\xi(\eta^h_\zeta)\\
=&\mathscr{L}^h_\xi(u(0))-\mathscr{L}^h_\xi(\eta^h_\zeta)+\mathscr{L}^h_\xi(u(t))-\mathscr{L}^h_\xi(u(0))\\
=&\mathscr{L}^h_\xi(u(0))-\mathscr{L}^h_\xi(\eta^h_\zeta)+E_1(0)-E_1(t)+E_2(t)\\
\leqslant&\frac{C}{2}\|u(0)-T_{\theta(0),p(0)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}+Me^{-\frac{\ell}{h}}+h^2M\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}\\
&+Me^{-\frac{\ell}{h}}+h^2M\|u(0)-T_{\theta(0),p(0)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}+4M\pi|\xi_2|h\int_0^t\frac{e^{-\frac{\ell}{h}}}{h^2}+\|u(s)-T_{\theta(s),p(s)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}\,\mathrm{ds}\,.
\end{aligned}$$

So there exist some constants $h_1<h_0$, $c>0$ and $\lambda>0$ (depending only of $\varepsilon, C, \rho, h_0$) such that, for all $h<h_1$, we have

$$\begin{aligned}
\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}\leqslant ce^{-\frac{\ell}{2h}}+c\|u(0)-T_{\theta(0),p(0)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}+2\lambda h|\xi_2|\int_0^t e^{-\frac{\ell}{2h}}\\
+\|u(s)-T_{\theta(s),p(s)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}\,\mathrm{ds}\,.
\end{aligned}$$

Applying Grönwall's lemma, we obtain the estimate

$$\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}+e^{-\frac{\ell}{2h}}\leqslant e^{2\lambda|\xi_2|ht}\left[(1+c)e^{-\frac{\ell}{2h}}+c\|u(0)-T_{\theta(0),p(0)}\eta_\zeta^h\|^2_{H^1(\mathbb{R})}\right].$$

Now applying Minkowski's inequality, we get

$$\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}\leqslant\sqrt{1+c}\,e^{\lambda|\xi_2|ht}\left[e^{-\frac{\ell}{4h}}+\|u(0)-T_{\theta(0),p(0)}\eta_\zeta^h\|_{H^1(\mathbb{R})}\right].$$

We want to deduce a bound on $\delta$ from this inequality. Applying the inequalities established in the previous paragraph, we have

$$\begin{aligned}
\|u(0)-T_{\theta(0),p(0)}\eta_\zeta^h\|_{H^1(\mathbb{R})}&\leqslant\|u(0)-T_{\gamma(0),\mathrm{y}(0)}\eta_\xi^h\|_{H^1(\mathbb{R})}+\|\eta_\xi^h-T_{\delta_\gamma,\delta_\mathrm{y}}\eta_\xi^h\|_{H^1(\mathbb{R})}+\|\eta_\zeta^h-\eta_\xi^h\|_{H^1(\mathbb{R})}\\
&\leqslant\delta(0)+C^3\delta(0)+(1+C)\delta(0).
\end{aligned}$$

On the other hand, applying the same inequalities, we have

$$\begin{aligned}
\delta(t)=\|u(t)-T_{\gamma(t),\mathrm{y}(t)}\eta_\xi^h\|_{H^1(\mathbb{R})}\leqslant&\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}+\|\eta_\xi^h-T_{\delta_\gamma,\delta_\mathrm{y}}\eta_\xi^h\|_{H^1(\mathbb{R})}\\
&+\|\eta_\zeta^h-\eta_\xi^h\|_{H^1(\mathbb{R})}\\
\leqslant&\|u(t)-T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})}+C^3\delta(0)+(1+C)\delta(0).
\end{aligned}$$

Consequently, we have proven our estimate :

$$\delta(t)\leqslant\sqrt{1+c}\,e^{\lambda|\xi_2|ht}\left[e^{-\frac{\ell}{4h}}+\delta(0)\left(2+C+C^3+\frac{1+C+C^3}{\sqrt{1+c}}\right)\right]. \tag{2.72}$$

**Remark 2.4.3.** *We could get another kind of estimate of $\delta(t)$ based on the high order Sobolev norms of $u(t)$. Indeed, if $n \in \mathbb{N}^*$, using Lemma 2.4.2, we have*

$$|E_3(u(t))| \lesssim \|\widehat{u}(\omega)\mathbb{1}_{|\omega|\geqslant\frac{\pi}{3h}}\|^2_{L^2(\mathbb{R})} \lesssim h^{2n}\|u(t)\|^2_{\dot{H}^n(\mathbb{R})}.$$

*Applying this inequality for $E_2$ and realizing the same proof without applying Grönwall's lemma, we get*

$$\delta(t) \lesssim \delta(0) + e^{-\frac{\ell}{h}} + \sqrt{t|\xi_2|}h^{n-\frac{1}{2}} \sup_{0<s<t} \|u(s)\|_{\dot{H}^n(\mathbb{R})}.$$

Step 4 : Control of $\dot{\gamma}$ and $\dot{\mathrm{y}}$

The idea to obtain the estimate (2.11) is that $\xi$ is the solution of a perturbed linear equation whose $(\dot{\gamma}, \dot{\mathrm{y}})$ is a solution (i.e. (2.66)). We work with a fixed $t < T$. To simplify the notation, we assume that $\theta(t) = p(t) = 0$. We introduce a notation : for $v \in BL^2_h$, we define

$$b_{\zeta,h}[v] := \begin{pmatrix} \langle \Delta_h v + |v|^2 v, \eta^h_\zeta \rangle_{L^2(\mathbb{R})} \\ -\langle \Delta_h v + |v|^2 v, i\partial_x \eta^h_\zeta \rangle_{L^2(\mathbb{R})} \end{pmatrix}. \tag{2.73}$$

With this formalism, equation (2.66) becomes (see Lemma 2.2.4)

$$A_{\zeta,h}[u(t)] \begin{pmatrix} \dot{\theta}(t) \\ \dot{p}(t) \end{pmatrix} = b_{\zeta,h}[u(t)] + 2E_4, \quad \text{where} \quad E_4 = \begin{pmatrix} \langle \cos\left(\frac{2\pi x}{h}\right) |v|^2 v, \eta^h_\zeta \rangle_{L^2(\mathbb{R})} \\ -\langle \cos\left(\frac{2\pi x}{h}\right) |v|^2 v, i\partial_x \eta^h_\zeta \rangle_{L^2(\mathbb{R})} \end{pmatrix}.$$

By construction $\eta^h_\zeta$ generates a traveling wave of the perturbation of DNLS whose speed is $\zeta$. It means we can apply Proposition 2.3.1 with $\mathfrak{u}(t,x) := e^{i\zeta_1}\eta^h_\zeta(x - \zeta_2 t)$. However, we have $e^{-i\zeta_1}\mathfrak{u}(t, \bullet + \xi_2 t) = \eta^h_\zeta \in \mathrm{Span}(i\eta^h_\zeta, \partial_x \eta^h_\zeta)^{\perp_{L^2}}$. So calculating $\partial_t \mathfrak{u}$ with Equation (2.36) of Proposition 2.3.1, we get

$$A_{\zeta,h}[\eta^h_\zeta]\zeta = b_{\zeta,h}[\eta^h_\zeta].$$

Consequently, we have

$$A_{\zeta,h}[u(t)] \begin{pmatrix} \dot{\theta}(t) - \zeta_1 \\ \dot{p}(t) - \zeta_2 \end{pmatrix} = \Big( b_{\zeta,h}[u(t)] - b_{\zeta,h}[\eta^h_\zeta] \Big) - A_{\zeta,h}[u(t) - \eta^h_\zeta]\zeta + 2E_4. \tag{2.74}$$

It is with this equation that we will obtain an estimate on $\dot{\theta}(t) - \zeta_1$ and $\dot{p}(t) - \zeta_2$. Indeed, as we have seen in the second step, since $t < T$, $A_{\zeta,h}[u(t)]$ is invertible and $\|A_{\zeta,h}[u(t)]^{-1}\|_1 \leqslant C$. So we just need to control the three terms in the right-hand side of the previous equation.

- First, we prove that $b_{\zeta,h}$ is a Lipschitz function on bounded subsets of $BL^2_h$, for the norm $\| \bullet \|_{H^1(\mathbb{R})}$, uniformly with respect to $\zeta$ and $h$. Considering the first coordinate (see (2.73)), we have

  $$(b_{\zeta,h}[v])_1 = \langle \Delta_h v + |v|^2 v, \eta^h_\zeta \rangle_{L^2(\mathbb{R})} = \langle v, \Delta_h \eta^h_\zeta \rangle_{L^2(\mathbb{R})} + \langle |v|^2 v, \eta^h_\zeta \rangle_{L^2(\mathbb{R})}.$$

  But $\|\Delta_h \eta^h_\zeta\|_{L^2(\mathbb{R})} \leqslant \|\partial^2_x \eta^h_\zeta\|_{L^2(\mathbb{R})} \leqslant C$ (see (2.58)) and $v \mapsto |v|^2 v$ is a Lipschitz function on bounded subsets of $H^1(\mathbb{R})$. So, since $\|\eta^h_\zeta\|_{H^1(\mathbb{R})} \leqslant C$ and $\|u(t) - \eta^h_\zeta\|_{H^1(\mathbb{R})} \leqslant \rho$, there exists a constant $k > 0$ (depending only of $C$ and $\rho$) such that

  $$|(b_{\zeta,h}[u(t)] - b_{\zeta,h}[\eta^h_\zeta])_1| \leqslant k\|u(t) - \eta^h_\zeta\|_{H^1(\mathbb{R})}.$$

  Since $\|\eta^h_\zeta\|_{H^3(\mathbb{R})} \leqslant C$, the second coordinate of $(b_{\zeta,h}[u(t)] - b_{\zeta,h}[\eta^h_\zeta])_1$ clearly enjoys the same estimate.

- Since $\|\eta_\zeta^h\|_{H^1(\mathbb{R})} \leqslant C$, it is obvious, from the definition of $A_{\zeta,h}$ (see (2.54)) that there exists an universal constant $c > 0$ such that

$$\|A_{\zeta,h}[u(t) - \eta_\zeta^h]\|_1 \leqslant cC\|u(t) - \eta_\zeta^h\|_{H^1(\mathbb{R})}.$$

- We can estimate $E_4$ as we have estimated $E_1(t)$ in the previous paragraph. Consequently, we get some constants $M, \ell$ independent of $h$ and $\zeta$ such that

$$|E_4| \leqslant Me^{-\frac{\ell}{h}} + M\|u(t) - \eta_\zeta^h\|_{H^1(\mathbb{R})}.$$

Applying these three estimates and the control of the norm of the invert of $A_{\zeta,h}[u(t)]$, we get from (2.74)

$$|\dot\theta(t) - \zeta_1| + |\dot p(t) - \zeta_2| \leqslant CMe^{-\frac{\ell}{h}} + C(M + k + cC)\|u(t) - \eta_\zeta^h\|_{H^1(\mathbb{R})}.$$

However, we have proven that $\|u(t) - \eta_\zeta^h\|_{H^1(\mathbb{R})} \leqslant \delta(t) + (1 + C + C^3)\delta(0)$ and $|\xi - \zeta| \leqslant C\delta(0)$. So, since $\dot\theta = \dot\gamma$ and $\dot p = \dot{\mathrm{y}}$, we have proven that

$$|\dot\gamma(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant K(e^{-\frac{\ell}{h}} + \delta(t) + \delta(0)),$$

where $K$ depends only of $C, M, c$ and $k$.

## 2.5 Appendix

### 2.5.1 Proof of Theorem 2.1.7

Let $s > 0$, $\varepsilon \in (0, 2)$ and $n \in \mathbb{N}^*$ be such that $n \geqslant n_0 \geqslant 2$ where $n_0 \in \mathbb{N}^*$ will be determined later to be large enough. Let $\rho > 0$ and $v \in H^n(\mathbb{R})$ be such that

$$\|v\|_{\dot H^n(\mathbb{R})} \leqslant \rho \quad \text{and} \quad \|\psi_\xi - v\|_{H^1(\mathbb{R})} \leqslant \frac{r}{2(1+\kappa)},$$

with $\xi \in \Omega$. Let $h_1 < h_0$ a constant that we will determine later.

Now consider $h < h_1$ and $\boldsymbol{u}$ a solution of DNLS such that

$$\exists\, \mathrm{y}_0, \gamma_0 \in \mathbb{R}, \quad \forall g \in h\mathbb{Z}, \quad \boldsymbol{u}_g(0) = e^{i\gamma_0} v(g - \mathrm{y}_0).$$

We denote by $u$ the Shannon interpolation of $\boldsymbol{u}$. Without loss of generality, since DNLS is invariant by gauge transform, we can assume that $\gamma_0 = 0$.

**Lemma 2.5.1.** *The following inequality holds :*

$$\|u_0 - \eta_\xi^h(\bullet - \mathrm{y}_0)\|_{H^1(\mathbb{R})} \leqslant \|v - \eta_\xi^h\|_{H^1(\mathbb{R})} + h^{n-1}\rho.$$

This lemma is a classical estimate of aliasing, it will be proven at the end of this subsection.

Since $u_0, \eta_\xi^h \in BL_h^2$, we can apply Lemma 2.2.5 to obtain

$$\delta(0) := \|\,\boldsymbol{u}(0) - \Big(\eta_\xi^h(\bullet - \mathrm{y}_0)\Big)_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant \|u_0 - \eta_\xi^h(\bullet - \mathrm{y}_0)\|_{H^1(\mathbb{R})} \leqslant \|v - \eta_\xi^h\|_{H^1(\mathbb{R})} + h^{n-1}\rho. \quad (2.75)$$

Applying the triangle inequality, we deduce of Theorem 2.1.4 that

$$\delta(0) \leqslant \|v - \psi_\xi\|_{H^1(\mathbb{R})} + \|\psi_\xi - \eta_\xi^h\|_{H^1(\mathbb{R})} + h^{n-1}\rho \leqslant \frac{r}{2(1+\kappa)} + \kappa h^2 + h^{n-1}\rho.$$

Consequently, if $h_1$ is small enough then $\delta(0) \leqslant \frac{r}{1+\kappa}$. So we can apply Theorem 2.1.4 and Theorem 2.1.5. In particular, we get functions $\gamma, \mathrm{y} \in C^1(\mathbb{R}_+)$ such that, if for all $t \in (0,T)$

$$\delta(t) := \| \boldsymbol{u}(t) - (e^{i\gamma(t)}\eta_\xi^h(\bullet - \mathrm{y}(t)))_{|h\mathbb{Z}}\|_{H^1(h\mathbb{Z})} \leqslant r, \tag{2.76}$$

then we have for all $t \in (0,T)$

$$\delta(t) \leqslant \kappa \left(\delta(0) + e^{-\frac{\ell}{h}} + \sqrt{t|\xi_2|}h^{n-\frac{3}{2}} \sup_{0<s<t} \| \boldsymbol{u}(s)\|_{\dot{H}^{n-1}(h\mathbb{Z})}\right), \tag{2.77}$$

and

$$|\dot{\gamma}(t) - \xi_1| + |\dot{\mathrm{y}}(t) - \xi_2| \leqslant \kappa\, (\delta(0) + \delta(t) + e^{-\frac{\ell}{h}}). \tag{2.78}$$

Applying Theorem 2.1.6, we deduce that if (2.77) is satisfied then

$$\delta(t) \leqslant \kappa \left(\delta(0) + e^{-\frac{\ell}{h}} + C\sqrt{|\xi_2|}t^{\frac{n}{2}}h^{n-\frac{1}{2}}M_{\boldsymbol{u}(0)}^{\frac{4n-1}{3}} + C\sqrt{|\xi_2|}\sqrt{t}h^{n-\frac{1}{2}}\left(\| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})} + M_{\boldsymbol{u}(0)}^{\frac{2n+1}{3}}\right)\right), \tag{2.79}$$

where

$$M_{\boldsymbol{u}(0)} = \| \boldsymbol{u}(0)\|_{\dot{H}^1(h\mathbb{Z})} + \| \boldsymbol{u}(0)\|_{L^2(h\mathbb{Z})}^3.$$

So, to use (2.79), we have to estimate $M_{\boldsymbol{u}(0)}$ and $\| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})}$ uniformly with respect to $h$ and $\xi$. We get these bounds in the following lemma that will be proven at the end of this subsection.

**Lemma 2.5.2.** *There exists a constant $K > 0$, depending only of $\Omega, \rho$ and $n$ such that for all $h < h_0$,*

$$\kappa C M_{\boldsymbol{u}(0)}^{\frac{4n-1}{3}} \leqslant K \quad \textit{and} \quad \kappa C\left(\| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})} + M_{\boldsymbol{u}(0)}^{\frac{2n+1}{3}}\right) \leqslant K.$$

With the estimate, (2.79) becomes

$$\delta(t) \leqslant \kappa\delta(0) + \kappa e^{-\frac{\ell}{h}} + K\sqrt{|\xi_2|}t^{\frac{n}{2}}h^{n-\frac{1}{2}} + K\sqrt{|\xi_2|}\sqrt{t}h^{n-\frac{1}{2}}. \tag{2.80}$$

Now, we overcome the bootstrap condition (2.76). Let $T_0 \in (0,\infty]$ be a function of $|\xi_2|$ that will be fixed later. Consider $t \in (0, T_0 h^{-2+\varepsilon})$ such that for all $\tau \leqslant t$, $\delta(\tau) \leqslant r$. We deduce from (2.80) that

$$\delta(t) \leqslant \kappa\delta(0) + \kappa e^{-\frac{\ell}{h}} + KT_0^{\frac{n}{2}}\sqrt{|\xi_2|}h^{\frac{n\varepsilon-1}{2}} + K\sqrt{T_0|\xi_2|}h^{n-\frac{3}{2}+\frac{\varepsilon}{2}}.$$

Assuming $n_0 \geqslant \max(2, \frac{1+2s}{\varepsilon}, \frac{s+3-\varepsilon}{2})$, $h_1 \leqslant 1$ and $T_0 = \min(|\xi_2|^{-1}, |\xi_2|^{-\frac{1}{n}})$, we deduce

$$\delta(t) \leqslant \kappa\delta(0) + \kappa e^{-\frac{\ell}{h}} + 2Kh^s \leqslant \kappa\delta(0) + \left(\kappa\left(\frac{s}{\ell e}\right)^s + 2K\right)h^s. \tag{2.81}$$

So assuming $h_1 < \left[\frac{r}{1+\kappa}\left(\kappa\left(\frac{s}{\ell e}\right)^s + 2K\right)^{-1}\right]^{-s}$, we get $\delta(t) < r$. Consequently, proceeding as usual by contradiction, we deduce that it was useless to assume that for all $\tau \leqslant t$, $\delta(\tau) \leqslant r$.

Finally, to conclude rigorously this proof, we have to explain how to get (2.20) and (2.21). On the one hand, to get (2.20), we just have to estimate $\delta(0)$ by (2.75) in (2.81) (and to assume that $n_0 - 1 \geqslant s$). On the other hand, we have to estimate the terms of (2.78). We control $\delta(0)$ as previously, $\delta(t)$ by (2.20) and $e^{-\frac{\ell}{h}}$ by $\left(\frac{hs}{\ell e}\right)^s$.

**Proof of Lemma 2.5.1.** Let $v_h$ be the $L^2$ orthogonal projection of $v$ on $BL^2_h$, $i.e.$

$$\widehat{v_h} = \mathbb{1}_{\left(-\frac{\pi}{h},\frac{\pi}{h}\right)}\widehat{v}.$$

We introduce $w_h = u_0 - v_h(\bullet - \mathrm{y}_0)$. Since the $H^1$ norm is invariant by advection, we have

$$\|u_0 - \eta^h_\xi(\bullet - \mathrm{y}_0)\|_{H^1(\mathbb{R})} \leqslant \|v_h - \eta^h_\xi\|_{H^1(\mathbb{R})} + \|w_h\|_{H^1(\mathbb{R})}.$$

Since $\eta^h_\xi \in BL^2_h$, $v - v_h$ is orthogonal to $\eta^h_\xi$ in $H^1(\mathbb{R})$. Consequently, we have $\|v_h - \eta^h_\xi\|_{H^1(\mathbb{R})} \leqslant \|v - \eta^h_\xi\|_{H^1(\mathbb{R})}$. So we just have to prove that $\|w_h\|_{H^1(\mathbb{R})} \leqslant \rho h^{n-1}$.

Applying Proposition 2.2.3, we have

$$\forall \omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right), \quad \widehat{w_h}(\omega) = \sum_{k\in\mathbb{Z}^*} e^{-i\left(\omega+\frac{2k\pi}{h}\right)\mathrm{y}_0}\widehat{v}(\omega + \frac{2k\pi}{h}).$$

Consequently, we have

$$\begin{aligned}\|w_h\|_{H^1(\mathbb{R})} &\leqslant \frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}^*} \|\widehat{v}(\omega + \frac{2k\pi}{h})\sqrt{1+\omega^2}\|_{L^2\left(-\frac{\pi}{h},\frac{\pi}{h}\right)} \\ &\leqslant \frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}^*} \left\|\widehat{\partial_x v}(\omega + \frac{2k\pi}{h})\frac{\sqrt{1+\omega^2}}{\omega + \frac{2k\pi}{h}}\right\|_{L^2\left(-\frac{\pi}{h},\frac{\pi}{h}\right)}.\end{aligned}$$

Assuming $h_1 \leqslant 2\pi$, we have $\left|\frac{\sqrt{1+\omega^2}}{\omega+\frac{2k\pi}{h}}\right| \leqslant \frac{2}{2|k|-1}$ for $\omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right)$. Consequently, applying the Cauchy-Schwarz inequality, we get

$$\|w_h\|_{H^1(\mathbb{R})} \leqslant \|\partial_x(v - v_h)\|_{L^2(\mathbb{R})}\sqrt{\sum_{k\in\mathbb{Z}^*} \frac{4}{(2|k|-1)^2}} = \frac{\pi}{\sqrt{2}}\|\partial_x(v - v_h)\|_{L^2(\mathbb{R})}.$$

Since the Fourier support of $v - v_h$ is localized outsize $\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$ and $n \geqslant 2$, we have

$$\|w_h\|_{H^1(\mathbb{R})} \leqslant \frac{\pi}{\sqrt{2}}\|\partial_x(v - v_h)\|_{L^2(\mathbb{R})} \leqslant \left(\frac{h}{\pi}\right)^{n-1} \frac{\pi}{\sqrt{2}}\|\partial^n_x(v - v_h)\|_{L^2(\mathbb{R})} \leqslant h^{n-1}\frac{\pi^{2-n}}{\sqrt{2}}\rho \leqslant h^{n-1}\rho.$$

**Proof of Lemma 2.5.2.** There are two quantities to control, $\|\boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})}$ and $M_{\boldsymbol{u}(0)}$. To control $\|\boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})}$, it is enough to prove that the restriction to $h\mathbb{Z}$ is a continuous map from $\dot{H}^n(\mathbb{R})$ to $\dot{H}^n(h\mathbb{Z})$, uniformly with respect to $h$. Indeed, denoting $w = v(\bullet - \mathrm{y}_0)$ and applying Proposition 2.2.3, we have, for all $\omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right)$,

$$\widehat{u_0}(\omega) = \sum_{k\in\mathbb{Z}^*} \widehat{w}(\omega + \frac{2k\pi}{h}).$$

Since for $k \neq 0$ and $\omega \in \left(-\frac{\pi}{h}, \frac{\pi}{h}\right)$, we have

$$\left|\frac{\omega}{\omega + \frac{2k\pi}{h}}\right| \leqslant \frac{1}{2|k|-1},$$

applying the Cauchy Schwarz inequality (and (2.33) ), we get

$$\begin{aligned}
\| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})} &\leqslant \|\omega^n \widehat{u_0}\|_{L^2\left(-\frac{\pi}{h},\frac{\pi}{h}\right)} \\
&\leqslant \|\omega^n \widehat{w}(\omega)\|_{L^2\left(-\frac{\pi}{h},\frac{\pi}{h}\right)} + \sum_{k\in\mathbb{Z}^*} \left\| \left(\frac{\omega}{\omega + \frac{2k\pi}{h}}\right)^n \widehat{\partial_x^n w}(\omega + \frac{2k\pi}{h})\right\|_{L^2\left(-\frac{\pi}{h},\frac{\pi}{h}\right)} \\
&\leqslant \|\partial_x^n w\|_{L^2(\mathbb{R})} + \sum_{k\in\mathbb{Z}^*} \|\widehat{\partial_x^n w}(\omega + \frac{2k\pi}{h})\|_{L^2\left(-\frac{\pi}{h},\frac{\pi}{h}\right)} \frac{1}{(2|k|-1)^n} \\
&\leqslant \|\partial_x^n w\|_{L^2(\mathbb{R})} + \|\partial_x^n w\|_{L^2(\mathbb{R})} \sqrt{\sum_{k\in\mathbb{Z}^*} \frac{1}{(2|k|-1)^{2n}}} \\
&= \left(1 + \sqrt{2\left(1 - \frac{1}{4^n}\right)\zeta(2n)}\right) \|\partial_x^n v\|_{L^2(\mathbb{R})} \leqslant \left(1 + \sqrt{2\left(1 - \frac{1}{4^n}\right)\zeta(2n)}\right) \rho,
\end{aligned}$$

where $\zeta$ is the Riemann zeta function.

Finally, to control $M_{\boldsymbol{u}(0)}$, we just have to control $\| \boldsymbol{u}(0)\|_{H^1(h\mathbb{Z})}$. But since we have proven that $\delta(0) \leqslant \frac{r}{1+\kappa}$, we just need to control $\|\eta_\xi^h\|_{H^1(\mathbb{R})}$ uniformly with respect to $\xi \in \Omega$ and $h < h_0$. Such an estimate can be obtained by using the bound $\|\eta_\xi^h - \psi_\xi^h\|_{H^1(\mathbb{R})} \leqslant \kappa h^2$ of Theorem 2.1.4.

### 2.5.2 Proof of Lemma 2.4.1

We would like to define the functions $\gamma$ and $\mathrm{y}$ from $\theta$ and $p$. So we introduce a new time : $T_{\mathrm{crit}}$. It is the largest time, smaller than $T_{\max}$, such that for all $t \in (0, T_{\mathrm{crit}})$, we have

$$\|(A_{\zeta,h}[T^{-1}_{\theta(t),p(t)}u(t)])^{-1}\| \leqslant 2C \tag{2.82}$$

and

$$|\theta(t) - \theta_0| + |p(t) - p_0| \leqslant 1 + c_2 t, \tag{2.83}$$

where $c_2 > 0$ is a real constant that will be determine later.

Now we define $\gamma$ and $\mathrm{y}$ as $C^1$ functions on $\mathbb{R}_+$ such that

$$\forall t \in (0, T_{\mathrm{crit}}),\ \gamma(t) = \theta(t) - \delta_\gamma \text{ and } \mathrm{y}(t) = p(t) - \delta_{\mathrm{y}}. \tag{2.84}$$

Let $T > 0$ be such that for all $t < T$, $\delta(t) = \|u(t) - T_{\gamma(t),\mathrm{y}(t)}\eta_\xi^h\|_{H^1(\mathbb{R})} < r$. To prove Lemma 2.4.1, it is enough to prove that $T \leqslant T_{\mathrm{crit}}$. We proceed by contradiction. Assume that $T_{\mathrm{crit}} < T$. So if $t < T_{\mathrm{crit}}$, we have

$$\|u(t) - T_{\theta(t),p(t)}\eta_\zeta^h\|_{H^1(\mathbb{R})} \leqslant (2 + C + C^3)r \leqslant \rho.$$

Applying (2.63), we know that

$$A_{\zeta,h}[T^{-1}_{\theta(t),p(t)}u(t)] \quad \text{is invertible and} \quad \|(A_{\zeta,h}[T^{-1}_{\theta(t),p(t)}u(t)])^{-1}\|_1 \leqslant C. \tag{2.85}$$

Furthermore, we can estimate $\langle T^{-1}_{\theta(t),p(t)}\partial_t u(t), i\eta^h_\zeta\rangle_{L^2(\mathbb{R})}$ and $\langle T^{-1}_{\theta(t),p(t)}\partial_t u(t), \partial_x\eta^h_\zeta\rangle_{L^2(\mathbb{R})}$. Indeed, since $u$ is a solution of DLNS in $BL^2_h$ (see Lemma 2.2.3), we have

$$\langle T^{-1}_{\theta(t),p(t)}\partial_t u(t), i\eta^h_\zeta\rangle_{L^2(\mathbb{R})} = -\left\langle \Delta_h u(t) + \left(1 + 2\cos\left(\frac{2\pi x}{h}\right)\right)|u(t)|^2 u(t), T^{-1}_{\theta(t),p(t)}\eta^h_\zeta\right\rangle_{L^2(\mathbb{R})}.$$

Since this operator is symmetric for the $L^2$ norm, we have

$$\begin{aligned}\langle T^{-1}_{\theta(t),p(t)}\partial_t u(t), i\eta^h_\zeta\rangle_{L^2(\mathbb{R})} = &-\langle u(t), T^{-1}_{\theta(t),p(t)}\Delta_h\eta^h_\zeta\rangle_{L^2(\mathbb{R})}\\ &-\left\langle \left(1 + 2\cos\left(\frac{2\pi x}{h}\right)\right)|u(t)|^2 u(t), T^{-1}_{\theta(t),p(t)}\eta^h_\zeta\right\rangle_{L^2(\mathbb{R})}.\end{aligned}$$

We are going to estimate these terms. Since $t < T$, we have $\|u(t) - T_{\gamma(t),\mathrm{y}(t)}\eta^h_\xi\|_{H^1(\mathbb{R})} < r$ by definition and so

$$\|u(t)\|_{H^1(\mathbb{R})} \leqslant r + C.$$

Consequently, we have

$$\||u(t)|^2 u(t)\|_{L^2(\mathbb{R})} \leqslant \|u(t)\|^2_{L^\infty}\|u(t)\|_{L^2(\mathbb{R})} \leqslant (r + C)^3.$$

Furthermore, we have seen in (2.33) that $\|\Delta_h\eta^h_\zeta\|_{L^2(\mathbb{R})} \leqslant \|\partial^2_x\eta^h_\zeta\|_{L^2(\mathbb{R})} \leqslant C$. Consequently, we have

$$|\langle T^{-1}_{\theta(t),p(t)}\partial_t u(t), i\eta^h_\zeta\rangle_{L^2(\mathbb{R})}| \leqslant C(r + C)^3 + C(r + C).$$

Similarly, we could prove that

$$|\langle T^{-1}_{\theta(t),p(t)}\partial_t u(t), \partial_x\eta^h_\zeta\rangle_{L^2(\mathbb{R})}| \leqslant C(r + C)^3 + C(r + C).$$

So, we have proved that

$$\max(|\dot\theta(t)|, |\dot p(t)|) \leqslant C^2(r + C)(1 + (r + C)^2).$$

Defining $c_2 = 2C^2(r + C)(1 + (r + C)^2)$, we have

$$|\theta(t) - \theta_0| + |p(t) - p_0| \leqslant c_2 t.$$

We can apply this inequality and (2.85) for $t = T_{\mathrm{crit}}$, so we have

$$\|(A_{\zeta,h}[T^{-1}_{\theta(T_{\mathrm{crit}}),p(T_{\mathrm{crit}})}u(t)])^{-1}\|_1 \leqslant C \text{ and } |\theta(T_{\mathrm{crit}}) - \theta_0| + |p(T_{\mathrm{crit}}) - p_0| \leqslant c_2 T_{\mathrm{crit}}.$$

But it is impossible because by definition of $T_{\mathrm{crit}}$ we should have

$$\|(A_{\zeta,h}[T^{-1}_{\theta(T_{\mathrm{crit}}),p(T_{\mathrm{crit}})}u(t)])^{-1}\|_1 = 2C \text{ or } |\theta(T_{\mathrm{crit}}) - \theta_0| + |p(T_{\mathrm{crit}}) - p_0| = 1 + c_2 T_{\mathrm{crit}}.$$

So, here is the contradiction and we have proven that $T \leqslant T_{\mathrm{crit}}$.

### 2.5.3 Inverse function Theorem

In this subsection, we give a version of the inverse function theorem.

**Theorem 2.5.3.** *Let $X, Y$ be some Banach spaces, $\Omega$ be an open convex subset of $X$ such that $0 \in \Omega$.*
*If $g : \Omega \to Y$ is a $C^1$ function such that*
- $\mathrm{d}\, g(0)$ *is invertible,*
- $\mathrm{d}\, g$ *is a $k$-Lipschitz function,*

*then, defining $\beta = \| \mathrm{d}\, g(0)^{-1}\|^{-1}$ and $r = \frac{\beta}{k}$, we have*
- *$g$ is a $C^1$ diffeomorphism from $B_X(0,r) \cap \Omega$ to $g(B_X(0,r) \cap \Omega)$,*
- *for all $x \in B_X(0,r) \cap \Omega$, $\| \mathrm{d}\, g(x)^{-1}\| \leqslant \frac{r}{\beta(r-\|x\|)}$,*
- *for all $0 < \rho \leqslant r$, if $B_X(0,\rho) \subset \Omega$ then $B_Y(g(0), \frac{\beta}{2}\rho) \subset g(B_X(0,\rho))$.*

*Proof.* First, we prove that $g$ is injective on $B_X(0,r) \cap \Omega$. Let $y \in B_X(0,r) \cap \Omega$. We introduce the application

$$\Phi_y : \begin{cases} B_X(0,r) \cap \Omega & \to & X \\ x & \mapsto & x - \mathrm{d}\, g(0)^{-1}(g(x) - g(y)). \end{cases}$$

It is enough to prove that $y$ is the only fix point of $\Phi_y$. But if $x \in B_X(0,r) \cap \Omega$ then

$$\| \mathrm{d}\, \Phi_y(x)\| = \|I_X - \mathrm{d}\, g(0)^{-1}\, \mathrm{d}\, g(x)\| \leqslant \| \mathrm{d}\, g(0)^{-1}\| \| \mathrm{d}\, g(0) - \mathrm{d}\, g(x)\| \leqslant \frac{\|x\|k}{\beta} < \frac{rk}{\beta} = 1. \quad (2.86)$$

Consequently, we deduce that if $x \neq y$ then $\|\Phi_y(x) - y\| < \|x - y\|$ and so $y$ is the only fix point of $\Phi_y$.

Then, we prove that $\mathrm{d}\, g(x)$ is invertible for any $x \in B_X(0,r) \cap \Omega$. Indeed, we have

$$\mathrm{d}\, g(x) = \mathrm{d}\, g(0) + \mathrm{d}\, g(x) - \mathrm{d}\, g(0) = \mathrm{d}\, g(0) \left[I_X + \mathrm{d}\, g(0)^{-1}(\mathrm{d}\, g(x) - \mathrm{d}\, g(0))\right]$$

with

$$\| \mathrm{d}\, g(0)^{-1}(\mathrm{d}\, g(x) - \mathrm{d}\, g(0))\| \leqslant \frac{k}{\beta}\|x\| < 1.$$

So we also deduce the second point of the theorem through the classical estimate of the Von Neumann series.

Now, applying the classical inverse function theorem, we have proven that $g$ is a $C^1$ diffeomorphism from $B_X(0,r) \cap \Omega$ to $g(B_X(0,r) \cap \Omega)$. Finally, we just need to prove the last assertion of the theorem. Let $\rho > 0$ be such that $0 < \rho \leqslant r$, $B_X(0,\rho) \subset \Omega$. We introduce $\delta \in (0,\rho)$ to prove that $\overline{B_Y(g(0), \frac{\beta}{2}\delta)} \subset g(\overline{B_X(0,\delta)})$. It is enough to prove the last point because

$$B_Y(g(0), \frac{\beta}{2}\rho) = \bigcup_{0<\delta<\rho} \overline{B_Y(g(0), \frac{\beta}{2}\delta)} \text{ and } g(B_X(0,\rho)) = \bigcup_{0<\delta<\rho} g(\overline{B_X(0,\delta)}).$$

Let $y \in \overline{B_Y(g(0), \frac{\beta}{2}\delta)}$, we want to solve $g(x) = y$. So, we introduce the application $\Psi = \Phi_{y\,|\overline{B_X(0,\delta)}}$. We want to apply the Banach fix point theorem. We have proven in (2.86) that $\Psi$ is

$\frac{\delta k}{\beta} < 1$ Lipschitz, so we just need to prove that it preserves $\overline{B_X(0,\delta)}$. Indeed, we have

$$\begin{aligned}\|\Psi(x)\| &\leqslant \|\Psi(0)\| + \|\Psi(x) - \Psi(0)\| \\ &\leqslant \frac{\beta}{2}\delta\|\,\mathrm{d}\,g(0)^{-1}\| + \|\,\mathrm{d}\,g(0)^{-1}\|\|g(x) - g(0) - \mathrm{d}\,g(0)x\| \\ &\leqslant \frac{\delta}{2} + \frac{1}{\beta}\left\|\int_0^1 \mathrm{d}\,g(sx)x\,\mathrm{d}s - \mathrm{d}\,g(0)x\right\| \leqslant \frac{\delta}{2} + \frac{k\delta}{\beta}\frac{\delta}{2} \leqslant \delta.\end{aligned}$$

□

### 2.5.4 A result of coercivity

**Lemma 2.5.4** (A reformulation of a Weinstein result in [115]). *If $\Omega$ is a relatively compact open subset of the set $\left\{\xi \in \mathbb{R}^2 \mid \xi_1 > \left(\frac{\xi_2}{2}\right)^2\right\}$ then there exists $c > 0$ such that for all $\xi \in \Omega$ we have*

$$\forall v \in H^1(\mathbb{R}) \cap \mathrm{Span}(\psi_\xi, i\psi_\xi, \partial_x\psi_\xi)^{\perp_{L^2}},\ \mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(v,v) \geqslant c\|v\|^2_{H^1(\mathbb{R})}. \tag{2.87}$$

*Proof.* Weinstein has proven in [115] that there exists $c > 0$ such that for all $v \in H^1(\mathbb{R})$,

$$v \in \mathrm{Span}(\psi_{(1,0)}, i\psi^3_{(1,0)}, \partial_x(\psi^3_{(1,0)}))^{\perp_{L^2}} \Rightarrow \mathrm{d}^2\,\mathscr{L}_{(1,0)}(\psi_{(1,0)})(v,v) \geqslant c\|v\|^2_{H^1}. \tag{2.88}$$

First, we will deduce from this estimate and Lemma 2.5.5 that (2.87) holds true for $\xi = (1,0)$. Then we will extend this result applying two transformations : *dilatation and boost.*

Step 1 : The case $\xi = (1,0)$. We apply Lemma 2.5.5 below, with the spaces

$$\begin{aligned} E &= H^1(\mathbb{R}) \cap \mathrm{Span}(\psi_{(1,0)})^{\perp_{L^2}} & G &= H^1(\mathbb{R}) \cap \mathrm{Span}(\psi_{(1,0)}, i\psi^3_{(1,0)}, \partial_x(\psi^3_{(1,0)}))^{\perp_{L^2}} \\ F &= H^1(\mathbb{R}) \cap \mathrm{Span}(\psi_{(1,0)}, i\psi_{(1,0)}, \partial_x\psi_{(1,0)})^{\perp_{L^2}} & H &= \mathrm{Span}(i\psi_{(1,0)}, \partial_x\psi_{(1,0)}). \end{aligned}$$

We equipped all these spaces with the $H^1(\mathbb{R})$ norm for which they are closed. By construction, $F$ and $H$ are obviously complementary spaces. However, we have to prove that $G$ and $H$ are complementary spaces.

First, we prove that $H \cap G = \{0\}$. If $g = \alpha i\psi_{(1,0)} + \beta\partial_x\psi_{(1,0)} \in G$ then $\langle g, i\psi^3_{(1,0)}\rangle_{L^2(\mathbb{R})} = \langle g, \partial_x(\psi^3_{(1,0)})\rangle_{L^2(\mathbb{R})} = 0$. However, since $\psi_{(1,0)}$ is a real valued function, we have

$$\langle \partial_x\psi_{(1,0)}, i\psi^3_{(1,0)}\rangle_{L^2(\mathbb{R})} = \langle \partial_x(\psi^3_{(1,0)}), i\psi_{(1,0)}\rangle_{L^2(\mathbb{R})} = 0. \tag{2.89}$$

Consequently, we deduce that $\alpha\|\psi_{(1,0)}\|^4_{L^4(\mathbb{R})} = \beta\langle \partial_x(\psi^3_{(1,0)}), \partial_x\psi_{(1,0)}\rangle_{L^2(\mathbb{R})} = 0$. So we just need to verify from (2.4) that $\langle \partial_x(\psi^3_{(1,0)}), \partial_x\psi_{(1,0)}\rangle_{L^2(\mathbb{R})} \neq 0$ which yields $\alpha = \beta = 0$.

Now, we prove that $H + G = E$. Since, by construction $G + \mathrm{Span}(i\psi^3_{(1,0)}, \partial_x(\psi^3_{(1,0)})) = E$, we just need to prove that $i\psi^3_{(1,0)}, \partial_x(\psi^3_{(1,0)}) \in H + G$. Since $i\psi^3_{(1,0)}$ and $\partial_x(\psi^3_{(1,0)})$ are orthogonal, we can decompose $i\psi_{(1,0)}$ and $\partial_x\psi_{(1,0)}$ through the decomposition $E = G + \mathrm{Span}(i\psi^3_{(1,0)}, \partial_x(\psi^3_{(1,0)}))$ to get (with (2.89))

$$\begin{cases} i\psi_{(1,0)}\|\psi_{(1,0)}\|^6_{L^6(\mathbb{R})} - \|\psi_{(1,0)}\|^4_{L^4(\mathbb{R})} i\psi^3_{(1,0)} \in G, \\ \partial_x\psi_{(1,0)}\|\partial_x(\psi^3_{(1,0)})\|^2_{L^2(\mathbb{R})} - \langle \partial_x(\psi^3_{(1,0)}), \partial_x\psi_{(1,0)}\rangle_{L^2(\mathbb{R})}\partial_x(\psi^3_{(1,0)}) \in G. \end{cases}$$

Since the coefficients associated with $i\psi_{(1,0)}^3$ and $\partial_x(\psi_{(1,0)}^3)$ are not zero, we deduce that

$$i\psi_{(1,0)}^3, \partial_x(\psi_{(1,0)}^3) \in H + G.$$

In order to apply Lemma 2.5.5, with $b = \mathrm{d}^2\,\mathscr{L}_{(1,0)}(\psi_{(1,0)})$ we have to prove that $\partial_x\psi_{(1,0)}$ and $i\psi_{(1,0)}$ belong to the kernel of $\mathrm{d}^2\,\mathscr{L}_{(1,0)}(\psi_{(1,0)})$. Indeed, since $\mathscr{L}_{(1,0)}(\psi_{(1,0)})$ is invariant by gauge transform and dilatation, the set of its critical points are also invariant by these transform, i.e.

$$\forall t \in \mathbb{R}, \forall v \in H^1(\mathbb{R}),\ \mathrm{d}\,\mathscr{L}_{(1,0)}(e^{it}\psi_{(1,0)})(v) = \mathrm{d}\,\mathscr{L}_{(1,0)}(\psi_{(1,0)}(\bullet - t))(v) = 0.$$

However, since $\psi_{(1,0)}$ is a very regular function (see Lemma 2.3.4 or directly (2.4)), we can compute the derivative in $t = 0$ to get

$$\forall t \in \mathbb{R}, \forall v \in H^1(\mathbb{R}),\ \mathrm{d}^2\,\mathscr{L}_{(1,0)}(\psi_{(1,0)})(i\psi_{(1,0)}, v) = \mathrm{d}^2\,\mathscr{L}_{(1,0)}(\psi_{(1,0)})(\partial_x\psi_{(1,0)}, v) = 0.$$

Now to apply Lemma 2.5.5, we observe that the required assumption of coercivity of $b$ on $G$ is the result of Weinstein (2.88), and we obtain the result.

Step 2 : Extension by dilatation and boost

Denote by $T$ the dilatation action defined by $T_m(u)(x) = mu(mx)$ for all $x \in \mathbb{R}$, $u \in H^1(\mathbb{R})$ and $m > 0$, and let $B$ be the boost action defined by $B_\nu u := e^{i\nu x}u$ for all $x \in \mathbb{R}$, $u \in H^1(\mathbb{R})$ and $\nu \in \mathbb{R}$. These transformations are useful because we have the following relations

$$\forall m, \mu > 0, \forall \nu \in \mathbb{R},\ \mathscr{L}_{(1,0)} \circ T_m = m^3\,\mathscr{L}_{(m^{-2},0)} \text{ and } \mathscr{L}_{(\mu,0)} \circ B_\nu = \mathscr{L}_{(\mu+\nu^2,-2\nu)}$$

With these relations a straightforward calculation shows that

$$\mathscr{L}_\xi = m_\xi^3\,\mathscr{L}_{(1,0)} \circ T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} \text{ with } m_\xi = \sqrt{\xi_1 - \left(\frac{\xi_2}{2}\right)^2}. \tag{2.90}$$

Furthermore, using the definition of $\psi_\xi$, we have

$$\psi_\xi = B_{\frac{\xi_2}{2}} \circ T_{m_\xi}\psi_{(1,0)}.$$

Consequently, we are able to transport the coercivity property from $\xi = (1,0)$ to any $\xi$, provided that $\xi_1 > \left(\frac{\xi_2}{2}\right)^2$. First, we observe that if $v \in H^1(\mathbb{R}) \cap \mathrm{Span}(\psi_\xi, i\psi_\xi, \psi_\xi')^{\perp_{L^2}}$ then

$$T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} v \in H^1(\mathbb{R}) \cap \mathrm{Span}(\psi_{(1,0)}, i\psi_{(1,0)}, \psi_{(1,0)}')^{\perp_{L^2}}.$$

Second, we calculate the derivative of the Lagrange function through the transport relation (2.90),

$$\mathrm{d}\,\mathscr{L}_\xi(\psi_\xi)(v) = m_\xi^3\,\mathrm{d}[\mathscr{L}_{(1,0)} \circ T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}}](\psi_\xi)(v) = m_\xi^3\,\mathrm{d}\,\mathscr{L}_{(1,0)}(\psi_{1,0})(T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} v) = 0.$$

Then we deduce a property of coercivity

$$\mathrm{d}^2\,\mathscr{L}_\xi(\psi_\xi)(v,v) = m_\xi^3\,\mathrm{d}^2\,\mathscr{L}_{(1,0)}(\psi_{1,0})(T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} v, T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} v) \geqslant c m_\xi^3 \left\| T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} v \right\|_{H^1}^2.$$

This inequality implies Estimate (2.87) because applying the Peetre inequality [1], we get

$$\left\|T_{m_\xi^{-1}} \circ B_{-\frac{\xi_2}{2}} v\right\|_{H^1}^2 = \left\|B_{-\frac{m_\xi^{-1}\xi_2}{2}} \circ T_{m_\xi^{-1}} v\right\|_{H^1}^2 \geqslant \frac{1}{2} \frac{\|T_{m_\xi^{-1}} v\|_{H^1}^2}{1+\left(\frac{m_\xi^{-1}\xi_2}{2}\right)^2} = \frac{1}{2} \frac{m_\xi^{-1}\|v\|_{L^2}^2 + m_\xi^{-3}\|\partial_x v\|_{L^2}^2}{1+\left(\frac{m_\xi^{-1}\xi_2}{2}\right)^2}.$$

□

### 2.5.5 Functional analysis lemmas

**Lemma 2.5.5.** *Let $F$, $G$ be two closed subspaces of a normed space $E$. If $F$ and $H$ admit a same finite dimensional complementary space $H$, denote by $\Pi$ the projection onto $G$ of kernel $H$. Then $\Pi_{|F}$ is a normed space vector isomorphism.*
*Furthermore, if $b$ is a bilinear symmetric form on $E$, $H$ is a subspace of its kernel and if there exists $\alpha > 0$ such that*

$$\forall x \in G,\ b(x,x) \geqslant \alpha \|x\|^2$$

*then there exists $\beta > 0$ such that*

$$\forall x \in F,\ b(x,x) \geqslant \beta \|x\|^2.$$

*Proof.* Let $P$ be the projection onto $F$ of kernel $H$. If $f \in F$ then $P\Pi f = f$. Indeed, if $f = g + h$ with $g \in G$ and $h \in H$ then $g = \Pi f = f - h$. Consequently, we would have $f = Pg = P\Pi f$. Similarly, we can prove that $\Pi P g = g$, for any $g \in G$. So, we have proven that $\Pi_{|F}^{-1} = P_{|G}$.
To prove the first part of the lemma, we just have to prove that $\Pi$ and $P$ are continuous to conclude this proof. This is a very classical exercise of normed space vector, whose proof is based on compactness.

The second part of the lemma is a straightforward calculation. Indeed, if $x \in F$ then

$$b(x,x) = b(\Pi x, \Pi x) \geqslant \alpha \|\Pi x\|^2 \geqslant \alpha \|\Pi_{|F}^{-1}\|^{-2} \|x\|^2.$$

□

**Lemma 2.5.6.** *Let $E$ be a real vector space whose $(x_j)_{j=1,\dots,n}$ is a free family. Define $X = \mathrm{Span}(x_j)_{j=1,\dots,n}$ the subspace generated by this family. Let $\langle \bullet, \bullet \rangle_1$,$\langle \bullet, \bullet \rangle_2$ be two scalar products on $E$ such that the induced norms satisfy $\| \bullet \|_1 \leqslant c \| \bullet \|_2$. Define $G \in M_n(\mathbb{R})$ the Gram matrix associated with $(x_j)_{j=1,\dots,n}$ for the scalar product $\langle \bullet, \bullet \rangle_1$, i.e.*

$$G = \begin{pmatrix} \langle x_1, x_1 \rangle_1 & \dots & \langle x_1, x_n \rangle_1 \\ \vdots & & \vdots \\ \langle x_n, x_1 \rangle_1 & \dots & \langle x_n, x_n \rangle_1 \end{pmatrix}.$$

*For any $u \in E$, let $b(u)$ be a bilinear symmetric form continuous for the $\| \bullet \|_2$ norm. Assume that $b$ is $k$ Lipschitz on a ball of radius $R > 0$, i.e.*

$$\forall u, v \in B_2(0,R), \quad \forall y, z \in E, \quad |b(u)(y,z) - b(v)(y,z)| \leqslant k \|u - v\|_2 \|y\|_2 \|z\|_2$$

1. If $x, y \in \mathbb{R}$ then $1 + (x-y)^2 \geqslant \frac{1}{2}(1+x^2)(1+y^2)^{-1}$.

*and that there exists* $\alpha > 0$ *such that*

$$\forall y \in X^{\perp_1}, \quad b(0)(y,y) \geqslant \alpha \|y\|_2^2.$$

*Define two constants* $c_1, c_2 > 0$ *by the explicit formulas*

$$c_1 = \max(R, \frac{\alpha}{8k}) \text{ and } c_2 = \frac{\alpha}{4}\left[\left(\sum_{j=1}^{n} \|x_j\|_2\right) \|G^{-1}\|_\infty \left(\frac{\alpha}{2} + \|b(0)\|_2 + \frac{2\|b(0)\|_2^2}{\alpha}\right)\right]^{-1}.$$

*If* $\|u\|_2 \leqslant c_1$ *and* $\sup_{j=1,\dots,n} |\langle x_j, y\rangle_1| \leqslant c_2 \|y\|_2$ *then*

$$b(u)(y,y) \geqslant \frac{\alpha}{8}\|y\|_2^2.$$

*Proof.* Let $y = y_\| + y_\perp$ be the decomposition of $y$ associated to the algebraic decomposition $E = X \oplus X^{\perp_1}$. So, we get

$$\begin{aligned}
b(0)(y,y) &= b(0)(y_\| + y_\perp, y_\| + y_\perp) \\
&= b(0)(y_\perp, y_\perp)b(0) + 2b(0)(y_\|, y_\perp) + (y_\|, y_\|) \\
&\geqslant \alpha\|y_\perp\|_2^2 - 2\|b(0)\|_2\|y_\|\|_2\|y_\perp\|_2 - |b(0)\|_2\|y_\|\|_2 \\
&\geqslant \frac{\alpha}{2}\|y_\perp\|_2^2 - \left(\|b(0)\|_2 + \frac{2\|b(0)\|_2^2}{\alpha}\right)\|y_\|\|_2^2 \\
&\geqslant \frac{\alpha}{2}\|y\|_2^2 - \left(\frac{\alpha}{2} + \|b(0)\|_2 + \frac{2\|b(0)\|_2^2}{\alpha}\right)\|y_\|\|_2^2.
\end{aligned}$$

Consequently, we just need to control $\|y_\|\|_2$ with $\|y\|_2$ to get the result when $u = 0$. However, using basis linear algebra we can prove that

$$y_\| = \sum_{j=1}^{n} a_j x_j \text{ with } (a_j)_{j=1,\dots,n} = G^{-1}(\langle x_j, y\rangle_1)_{j=1,\dots,n}.$$

So, we get

$$\|y_\|\|_2 \leqslant c_2 \left(\sum_{j=1}^{n} \|x_j\|_2\right) \|G^{-1}\|_\infty \|y\|_2.$$

Finally, by definition of $c_2$, we get $b(0)(y,y) \geqslant \frac{\alpha}{4}\|y\|_2^2$. Furthermore, since $b$ is k Lipschitz on $B(0,R)$, we deduce directly that if $\|u\|_1 \leqslant c_1$ then $b(u)(y,y) \geqslant \frac{\alpha}{8}\|y\|_2^2$.

□

**Lemma 2.5.7.** *Let* $E$ *be a Banach space of dual space* $E'$*. Consider a algebraic decomposition of* $E$*,* $E = E_p \oplus E_m$*, and a continuous linear application* $T : E \to E'$ *such that*

*i)* $\forall x, y \in E,\ \langle Tx, y\rangle_{E',E} = \langle Ty, x\rangle_{E',E}$,

*ii)* $\exists \alpha_p > 0,\ \forall x \in E_p,\ \langle Tx, x\rangle_{E',E} \geqslant \alpha_p \|x\|^2$,

*iii)* $\exists \alpha_m > 0,\ \forall x \in E_m,\ \langle Tx, x\rangle_{E',E} \leqslant -\alpha_m \|x\|^2$.

*Then $T$ is invertible and we have*

$$\|T^{-1}\| \leqslant \left( \frac{1}{\alpha_p} + \frac{1}{\alpha_m} + \frac{2\|T\|}{\alpha_m \alpha_p} + \frac{\|T\|^2}{\alpha_m (\alpha_p)^2} \right). \tag{2.91}$$

*Proof.* In the proof we omit the index $E', E$ for all the duality brackets. We define by restrictions $T_{\epsilon_1 \epsilon_2} \in \mathscr{L}(E_{\epsilon_2}; E_{\epsilon_1})$ for $\epsilon_1, \epsilon_2 \in \{p, m\}$. Then we use a direct corollary of Riesz Theorem to prove that $T_{pp}$ is invertible. This corollary is the following.

**Lemma 2.5.8.** *Let $E$ be a Banach space of dual $E'$. Consider a continuous linear application $T : E \to E'$ such that*

*i)* $\exists \alpha > 0, \forall x \in E,\ \langle Tx, x \rangle \geqslant \alpha \|x\|^2$,

*ii)* $\forall x, y \in E,\ \langle Tx, y \rangle = \langle Ty, x \rangle$,

*then $T$ is invertible and $\|T^{-1}\| \leqslant \alpha^{-1}$.*

Now, decomposing $x = x_p + x_m$ with $x_p \in E_p$ and $x_m \in E_m$, we introduce operators $P : E \to E_p$ and $S : E_m \to E'_m$ defined by

$$Px = x_p + T_{pp}^{-1} T_{pm} x_m \text{ and } S = T_{mm} - T_{mp} T_{pp}^{-1} T_{pm}.$$

Then we verify by symmetry of $T$ (with the same decomposition for $y$) that

$$\forall x, y \in E,\ \langle Tx, y \rangle = \langle T_{pp} Px, Py \rangle + \langle S x_m, y_m \rangle.$$

To prove the Lemma, we have to solve,

$$\forall y \in E,\ \langle Tx, y \rangle = \phi(y) \text{ with } \phi \in E'. \tag{2.92}$$

Let $z \in E_m$ and denote $y = z - T_{pp}^{-1} T_{pm} z$. First, we verify that $Py = 0$. Consequently, we deduce from (2.92) that

$$\phi(y) = \phi(z - T_{pp}^{-1} T_{pm} z) = \langle S x_m, z \rangle.$$

However, we verify that $-S$ satisfies assumptions of Lemma 2.5.8 with $\alpha = \alpha_m$. Consequently, $S$ is invertible and so we have

$$x_m = S^{-1} \phi_{|E_m} - S^{-1} \phi T_{pp} T_{pm}.$$

Now if we apply (2.92) for $y = y_p \in E_p$, we have

$$\phi(y) = \langle T_{pp} Px, y \rangle = \langle T_{pp} x_p, y \rangle + \langle T_{pm} x_m, y \rangle.$$

Consequently, we have

$$x_p = T_{pp}^{-1} \phi_{|E_p} - T_{pp}^{-1} T_{pm} x_m.$$

Finally, we have solved (2.92). So $T$ is bijective and we verify (2.91) using the estimate given by Lemma 2.5.8. □

CHAPITRE 3

# BOUNDS ON THE GROWTH OF HIGH DISCRETE SOBOLEV NORMS FOR THE CUBIC DISCRETE NONLINEAR SCHRÖDINGER EQUATIONS ON $h\mathbb{Z}$.

## 3.1 Introduction

We consider the cubic discrete nonlinear Schrödinger equation (called DNLS) on a grid $h\mathbb{Z}$ of stepsize $h > 0$. This equation is a differential equation on $\mathbb{C}^{h\mathbb{Z}}$ defined by (see [87] and the references therein for details about its derivation)

$$\forall g \in h\mathbb{Z},\ i\partial_t \boldsymbol{u}_g = (\Delta_h \boldsymbol{u})_g + \nu |\boldsymbol{u}_g|^2 \boldsymbol{u}_g, \tag{3.1}$$

where $\nu \in \{-1, 1\}$ is a parameter and $\Delta_h \boldsymbol{u}$ is the discrete second derivative of $\boldsymbol{u}$. It is defined by

$$\forall g \in h\mathbb{Z},\ (\Delta_h \boldsymbol{u})_g = \frac{\boldsymbol{u}_{g+h} - 2\boldsymbol{u}_g + \boldsymbol{u}_{g-h}}{h^2}.$$

We consider both the *focusing* and the *defocusing* equations. They correspond respectively to the choices $\nu = 1$ and $\nu = -1$.

DNLS is a popular model in numerical analysis for the spatial discretization of the cubic nonlinear Schrödinger equation (NLS), given by :

$$i\partial_t u = \partial_x^2 u + \nu |u|^2 u, \tag{3.2}$$

see, for example, [17],[19],[26],[83],[84],[87]. Motivated by the approximation properties of NLS by DNLS, we consider the discrete model near its continuous limit *i.e.* when $h$ goes to $0$. So, we introduce norms consistent with the usual continuous norms and we pay attention to establish estimates uniform with respect to $h$.

We introduce the discrete $L^2$ space. It is defined by

$$L^2(h\mathbb{Z}) = \left\{ \boldsymbol{u} \in \mathbb{C}^{h\mathbb{Z}},\ \|\boldsymbol{u}\|^2_{L^2(h\mathbb{Z})} = h \sum_{g \in h\mathbb{Z}} |\boldsymbol{u}_g|^2 < \infty \right\}.$$

This space is natural to solve DNLS. Indeed, as $L^2(h\mathbb{Z})$ is a Banach algebra (which is not the case in the continuous setting), Cauchy Lipschitz Theorem can be applied to get the local well posedness of DNLS in $L^2(h\mathbb{Z})$. Furthermore, since (3.1) is invariant by gauge transform, as

a consequence of the Noether Theorem the discrete $L^2$ norm is a constant of the motion of DNLS. Thus, DNLS is globally well posed in $L^2(h\mathbb{Z})$.

We introduce the homogeneous discrete Sobolev norms by analogy with respect to the continuous homogeneous Sobolev norms. If $n \in \mathbb{N}$ is an integer and $\boldsymbol{u} \in L^2(h\mathbb{Z})$, its discrete homogeneous Sobolev norm of order $n$ is defined by

$$\| \boldsymbol{u} \|^2_{\dot{H}^n(h\mathbb{Z})} = \langle (-\Delta_h)^n \boldsymbol{u}, \boldsymbol{u} \rangle_{L^2(h\mathbb{Z})}. \tag{3.3}$$

For example, if $\boldsymbol{u} \in L^2(h\mathbb{Z})$, its discrete homogeneous Sobolev norm of order $1$ is

$$\| \boldsymbol{u} \|_{\dot{H}^1(h\mathbb{Z})} = \sqrt{h \sum_{g \in h\mathbb{Z}} \left| \frac{\boldsymbol{u}_{g+h} - \boldsymbol{u}_g}{h} \right|^2}.$$

Naturally, we define as usual the non homogeneous discrete Sobolev norms by

$$\| \boldsymbol{u} \|^2_{H^n(h\mathbb{Z})} = \sum_{k=0}^{n} \| \boldsymbol{u} \|^2_{\dot{H}^k(h\mathbb{Z})}.$$

Applying the triangle inequality we can easily prove that all these norms are controlled by the discrete $L^2$ norm

$$\forall \boldsymbol{u} \in L^2(h\mathbb{Z}), \ \| \boldsymbol{u} \|_{\dot{H}^n(h\mathbb{Z})} \leqslant \left( \frac{2}{h} \right)^n \| \boldsymbol{u} \|_{L^2(h\mathbb{Z})}. \tag{3.4}$$

So, since the discrete $L^2$ norm is a constant of the motion of DNLS, any discrete Sobolev norm of a solution of DNLS is globally bounded. However, this bound is not uniform with respect to the stepsize $h$. Consequently, these estimates are trivial when we consider the continuous limit.

An uniform control of these norms with respect to $h$ may be crucial to establish aliasing [1] or consistency estimates. For example, in [26], the existence and the stability of traveling waves is studied near the continuous limit of the focusing DNLS. The discrete Sobolev norms are used to control an aliasing error generated by the variations of the momentum (see Theorem $1.5$ of [26]). It is proven that if for all $n \in \mathbb{N}$, the discrete Sobolev norm of order $n$ of the solutions of the focusing DNLS can be bounded by $t^{\alpha_n}$, uniformly with respect to $h$, then DNLS admits solutions whose behavior is similar to traveling waves for times of order $h^{-\beta}$, with $\beta = \limsup_n \frac{n}{\alpha_n}$.

There is a huge literature about the growth of the Sobolev norms for continuous Schrödinger equations. Since, we are focusing on the continuous limit of DNLS, it is natural to try to adapt the methods used for these equations. If we focus on the continuous Schrödinger equations on $\mathbb{R}$, it seems that there are three families of methods and results.

- First, there is the cubic nonlinear Schrödinger equation. This equation is known to be *completely integrable*. In particular, it admits a sequence of constants of the motion coercive in $H^n(\mathbb{R})$. Consequently, all the Sobolev norms are globally bounded (see, for example, [104]).

1. Aliasing usually refers to a default of commutation between an interpolation and a nonlinearity.

- Second, there is the linear Schrödinger equation with a potential smooth with respect to $t$ and $x$. In such case, for all $\varepsilon > 0$ there is a control of the growth by $t^{\varepsilon}$ (see [38]).
- Third, in the other cases, there are methods using dispersion and/or higher modified energy. They were first introduced by Bourgain [37] and continued in the work of Staffilani [108]. They provide a control of the growth of the $H^n$ norm by $t^{\alpha n+\beta}$ for some $\alpha, \beta \in \mathbb{R}$. More recently, applying these methods [104], Sohinger proves a control of the $H^s$ norm by $t^{\frac{1}{3}s+}$ for the nonlinear Schrödinger equation with an Hartree nonlinearity.

A priori, DNLS is not a completely integrable equation, so we can not control its Sobolev norms as for its continuous limit (for a completely integrable spatial discretization of NLS, we can refer to the Ablowitz-Ladik model, see [2]). In this paper, we adapt the last method to the discrete nonlinear Schrödinger equation. In [109], Stefanov and Kevrekidis proved that the dispersion is weaker for the linear discrete Schrödinger equation than for the continuous equation. They got a $L^\infty$ decay of the form $t^{-\frac{1}{2}} + (ht)^{-\frac{1}{3}}$ (see also [83]). Using dispersive arguments in our setting seems thus more difficult than in the continuous case and does not seem to strengthen significantly the results. However, the method of constructing *modified energies* can be applied and turns out to yield results comparable to the continuous case (i.e. a polynomial bound whose exponent is proportional to the index of the Sobolev norm detailed above). More precisely, with our construction, we get the following bound.

**Theorem 3.1.1.** *For all $n \in \mathbb{N}^*$, there exists $C > 0$, such that for all $h > 0$ and all $\nu \in \{-1, 1\}$, if $\boldsymbol{u} \in C^1(\mathbb{R}; L^2(h\mathbb{Z}))$ is a solution of DNLS then for all $t \in \mathbb{R}$*

$$\| \boldsymbol{u}(t)\|_{\dot{H}^n(h\mathbb{Z})} \leqslant C \left[ \| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})} + M_{\boldsymbol{u}(0)}^{\frac{2n+1}{3}} + |t|^{\frac{n-1}{2}} M_{\boldsymbol{u}(0)}^{\frac{4n-1}{3}} \right], \tag{3.5}$$

*where*

$$M_{\boldsymbol{u}(0)} = \| \boldsymbol{u}(0)\|_{\dot{H}^1(h\mathbb{Z})} + \| \boldsymbol{u}(0)\|^3_{L^2(h\mathbb{Z})}.$$

This theorem is the main result of this paper, it will be proven in the third section. The second section is devoted to the introduction of tools and notations useful to prove it.

We conclude this introduction with some remarks about estimate (3.5).

- If $n = 1$ then the discrete $H^1$ norm is globally bounded, uniformly with respect to $h$. It is a consequence of the conservation of the Hamiltonian of DNLS and its coercivity in $H^1(h\mathbb{Z})$. In the focusing case this argument is specific to the dimension $1$. It is based on a discrete Gagliardo-Nirenberg inequality. For the defocusing case, the coercivity is straighforward and can be extended to higher dimensions and with other nonlinearities.
- The factor associated to the growing term $t^{\frac{n-1}{2}}$ is $M_{\boldsymbol{u}(0)}^{\frac{4n-1}{3}}$. So the growth of the high Sobolev norms is controlled by the size of the initial condition with respect to the low Sobolev norms.
- The estimate (3.5) is homogeneous. More precisely, DNLS is invariant by dilatation in the sense that if $\boldsymbol{u}$ is a solution of DNLS then $(t, g) \mapsto \lambda\, \boldsymbol{u}_{\lambda g}(\lambda^2 t)$ is a solution of DNLS with stepsize $h\lambda^{-1}$. Estimate (3.5) is invariant by this transformation (as can be seen from the exponents of $M_{\boldsymbol{u}(0)}$). Consequently, to prove Theorem 3.1.1, we just have to prove it with $h = 1$.

- Here, the construction of higher modified energies relies essentially algebraic considerations. In particular, it does not use any dispersion effect. So it seems possible to realize almost the same proof to get estimate (3.5) with periodic boundary conditions.

## 3.2 Shannon interpolation

In order to use classical analysis tools, it is very useful to identify sequences of $L^2(\mathbb{Z})$ with functions defined on the real line through an interpolation method. Here, we choose the Shannon interpolation (this choice is quite natural, see [83] or [26]). More precisely, it is the usual interpolation we get extending a sequence into a real function whose Fourier transform is supported on $[-\pi, \pi]$.

In this section, we introduce this interpolation and we give some of its classical properties useful to prove Theorem 3.1.1. For details or proofs of these classical properties the reader can refer to [26] or [98].

First we need to define the *discrete Fourier transform*

$$\mathcal{F} : \begin{cases} L^2(\mathbb{Z}) & \to & L^2(\mathbb{R}/2\pi\mathbb{Z}) \\ \boldsymbol{u} & \mapsto & \omega \mapsto \sum_{g\in\mathbb{Z}} \boldsymbol{u}_g \, e^{ig\omega} \end{cases} , \tag{3.6}$$

and the *Fourier Plancherel transform*

$$\mathscr{F} : \begin{cases} L^2(\mathbb{R}) & \to & L^2(\mathbb{R}) \\ u & \mapsto & \omega \mapsto \int_{\mathbb{R}} u(x) e^{ix\omega} \, \mathrm{dx} \end{cases}$$

where the right integral is defined by extending the operator defined on $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. We also use the notation $\widehat{u} = \mathscr{F} u$.

Now, we define the *Shannon interpolation* , denoted $\mathcal{I}$, through the following diagram

$$\begin{tikzcd} L^2(\mathbb{Z}) \arrow[r, "\mathcal{F}"] \arrow[rrr, bend right, "\mathcal{I}"'] & L^2(\mathbb{R}/2\pi\mathbb{Z}) \arrow[r, "u\mapsto \mathbb{1}_{(-\pi,\pi)}u"] & L^2(\mathbb{R}) \arrow[r, "\mathscr{F}^{-1}"] & L^2(\mathbb{R}) \end{tikzcd} , \tag{3.7}$$

where $\mathbb{1}_{(-\pi,\pi)} : \mathbb{R} \to \mathbb{R}$ the characteristic function of $(-\pi, \pi)$ and $\mathscr{F}^{-1}$ is the inverse of the Fourier Plancherel transform.

It is possible to deduce a very explicit formula to determine $\mathcal{I}\,\boldsymbol{u}$ from $\boldsymbol{u}$. Indeed, for $\boldsymbol{u} \in L^2(\mathbb{Z})$ and $x \in \mathbb{R}$, we have

$$\mathcal{I}\,\boldsymbol{u}(x) = \sum_{g\in\mathbb{Z}} \boldsymbol{u}_g \,\mathrm{sinc}(\pi(x-g)),$$

where the sum converges in $L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and $\mathrm{sinc}(x) = \frac{\sin(x)}{x}$, denotes the cardinal sine function.

In the following proposition, we give some properties of this interpolation useful to prove Theorem 3.1.1 .

**Proposition 3.2.1.** *(see, for example Chapter $5.4$ in [98], for details)*

- *$\mathcal{I}$ is an isometry, i.e.*

$$\forall\, \boldsymbol{u} \in L^2(\mathbb{Z}),\ \sum_{g\in\mathbb{Z}} |\, \boldsymbol{u}_g \,|^2 = \int_{\mathbb{R}} |\mathcal{I}\, \boldsymbol{u}(x)|^2 \,\mathrm{dx}\,.$$

- *The image of $\mathcal{I}$ is the set of functions whose Fourier support is a subset of $[-\pi,\pi]$. It is denoted by*

$$BL^2 := \mathcal{I}(L^2(\mathbb{Z})) = \{u \in L^2(\mathbb{R}) \mid \mathrm{Supp}\ \widehat{u} \subset [-\pi,\pi]\}.$$

- *If $\boldsymbol{u} \in L^2(\mathbb{Z})$ then $\mathcal{I}\,\boldsymbol{u}$ is an entire function which $\boldsymbol{u}$ is the restriction on $\mathbb{Z}$, i.e.*

$$\forall g \in \mathbb{Z},\ (\mathcal{I}\,\boldsymbol{u})\,(g) = \boldsymbol{u}_g\,.$$

Now, we focus on properties more specific to the discrete Sobolev norms.

**Proposition 3.2.2.** *(see Proposition $2.6$ in [26]) Let $\boldsymbol{u} \in L^2(\mathbb{Z})$ be a sequence and let $u = \mathcal{I}\,\boldsymbol{u}$ denote its Shannon interpolation. Then we have for almost all $\omega \in (-\pi,\pi)$*

$$\widehat{\mathcal{I}\Delta_1\,\boldsymbol{u}}(\omega) = (2\cos(\omega) - 2)\widehat{u}(\omega) = -4\left(\sin\left(\frac{\omega}{2}\right)\right)^2 \widehat{u}(\omega)$$

*and*

$$\widehat{\mathcal{I}|\,\boldsymbol{u}\,|^2\,\boldsymbol{u}}(\omega) = \sum_{k\in\mathbb{Z}} \widehat{|u|^2 u}(\omega + 2k\pi) = \sum_{k=-1}^{1} \widehat{u} * \widehat{\bar{u}} * \widehat{u}(\omega + 2k\pi),$$

*where $*$ is the usual convolution product.*

We deduce two important direct corollaries of this proposition. In the first one we identify the differential equation satisfied by the Shannon interpolation of a solution of DNLS.

**Corollary 3.2.1.** *Let $\boldsymbol{u} \in C^1(\mathbb{R}; L^2(\mathbb{Z}))$ be a solution of DNLS and let $u = \mathcal{I}\,\boldsymbol{u} \in C^1(\mathbb{R}; BL^2)$ denote its Shannon interpolation, then for all $t \in \mathbb{R}$ and almost all $\omega \in (-\pi,\pi)$,*

$$i\partial_t \widehat{u}(t,\omega) = -4\left(\sin\left(\frac{\omega}{2}\right)\right)^2 \widehat{u}(\omega) + \nu \sum_{k=-1}^{1} \widehat{u} * \widehat{\bar{u}} * \widehat{u}(\omega + 2k\pi).$$

In the second corollary, we identify the discrete Sobolev norms.

**Corollary 3.2.2.** *Let $\boldsymbol{u} \in L^2(\mathbb{Z})$ be a sequence, let $u = \mathcal{I}\,\boldsymbol{u}$ denote its Shannon interpolation. If $n \in \mathbb{N}^*$ then*

$$\|\, \boldsymbol{u} \,\|^2_{\dot{H}^n(\mathbb{Z})} = \frac{1}{2\pi} \int 2^{2n} \left(\sin\left(\frac{\omega}{2}\right)\right)^{2n} |\widehat{u}(\omega)|^2 \,\mathrm{d}\omega\,.$$

*Consequently, the continuous and discrete homogeneous Sobolev norms are equivalents, i.e.*

$$\left(\frac{2}{\pi}\right)^n \|\partial_x^n u\|_{L^2(\mathbb{R})} \leqslant \|\, \boldsymbol{u} \,\|_{\dot{H}^n(\mathbb{Z})} \leqslant \|\partial_x^n u\|_{L^2(\mathbb{R})}.$$

## 3.3 Proof of Theorem 3.1.1

This section is devoted to the proof of Theorem 3.1.1. The idea is to construct some *higher modified energies* controlling $H^n(h\mathbb{Z})$ norms and whose growth can be controlled by $H^{n-1}(h\mathbb{Z})$ norms. The construction of *higher modified energies* to study growth of Sobolev norms is a well known method (see [50] or [104]).

As explained at the end of the introduction, since inequality (3.5) of Theorem 3.1.1 is homogeneous, without loss of generality, we just need to prove it when $h = 1$.

### 3.3.1 Construction of the modified energies

DNLS is a Hamiltonian differential equation (see [26]) whose Hamiltonian (i.e. its energy) is defined on $L^2(\mathbb{Z})$ by

$$\mathrm{H}_{\mathrm{DNLS}} = \frac{1}{2}\|\cdot\|^2_{\dot{H}^1(\mathbb{Z})} - \frac{\nu}{4}\|\cdot\|^4_{L^4(\mathbb{Z})}.$$

So $\mathrm{H}_{\mathrm{DNLS}}(\boldsymbol{u})$ is a constant of the motion (it can be proven directly computing the discrete $L^2$ inner product of (3.1) and $\boldsymbol{u}(t)$).

If $u \in BL^2$ is the Shannon interpolation of a sequence $\boldsymbol{u} \in L^2(\mathbb{Z})$ this Hamiltonian can be written as a function of $\widehat{u}$ (it is a consequence of Proposition 3.2.2)

$$2\pi\,\mathrm{H}_{\mathrm{DNLS}}(\boldsymbol{u}) = \frac{1}{2}\int\left(2\sin\frac{\omega}{2}\right)^2|\widehat{u}(\omega)|^2\,\mathrm{d}\omega \\ - \frac{\nu}{4}\int_{w_1+w_2=w_{-1}+w_{-2} \mod 2\pi}\widehat{u}(w_1)\overline{\widehat{u}(w_{-1})}\widehat{u}(w_2)\overline{\widehat{u}(w_{-2})}\mathrm{d}w_1\mathrm{d}w_2\mathrm{d}w_{-1}. \quad (3.8)$$

The principle of the construction of the modified energies is to change the weights of these integrals to get a control of high Sobolev norms. To explain this construction, we need to adopt more compact notations. Some of them are classical for NLS (see [104]).

First, if $m \in \mathbb{N}^*$, we define $\mathcal{V}_m$ by

$$\mathcal{V}_m := \left\{ w \in \mathbb{R}^{[\![-m,m]\!]\setminus\{0\}} \mid \sum_{j=1}^{m} w_j - w_{-j} = 0 \mod 2\pi \right\},$$

where $[\![-m,m[\![$ denotes the set $\{-m,\dots,m\}$, and we equip it with its natural measure, denoted $\mathrm{d}w$, induced by the canonical Lesbegue measure of $\mathbb{R}^{2m}$.

If $\mu \in L^\infty(\mathcal{V}_m)$ and if $v \in L^2(\mathbb{R})$ is supported on $[-\pi,\pi]$, we define $\Lambda_m(\mu,v)$ by

$$\Lambda_m(\mu,v) := \int_{\mathcal{V}_m}\mu(w)\prod_{j=1}^{m} v(w_j)\overline{v(w_{-j})}\mathrm{d}w.$$

To prove that $\Lambda_m$ is well defined, we just need to pay attention to the support of

$$w \mapsto \mu(w)\prod_{j=1}^{m} v(w_j)\overline{v(w_{-j})}$$

and to apply a convolution Young inequality (see Lemma 3.3.3 for details).

For example, with this notation, we have a more compact expression of (3.8) given by

$$2\pi\,\mathrm{H}_{\mathrm{DNLS}}(\boldsymbol{u}) = \frac{1}{2}\int\left(2\sin\frac{\omega}{2}\right)^2|\widehat{u}(\omega)|^2\,\mathrm{d}\omega - \frac{\nu}{4}\Lambda_2(\mathbb{1}_{\mathcal{V}_2},\widehat{u}). \tag{3.9}$$

Then, we define a transformation $S_m : L^\infty(\mathcal{V}_m) \to L^\infty(\mathcal{V}_{m+1})$ by

$$S_m\mu(w_{-m-1}, w, w_{m+1}) = \sum_{k=1}^{m}\sum_{\sigma\in\{-1,1\}}\sigma\mu(w + \sigma e_{\sigma k}(w_{m+1} - w_{-m-1})),$$

where $(e_k)_{k\in[\![-m,m]\!]\setminus\{0\}}$ is the canonical basis of $\mathbb{R}^{[\![-m,m]\!]\setminus\{0\}}$.

We define another transformation $D_m : L^\infty(\mathbb{R}) \to L^\infty(\mathcal{V}_m)$ by

$$D_m f(w) = \sum_{j=1}^{m} f(w_j) - f(w_{-j}).$$

We say that a function $\mu \in L^\infty(\mathcal{V}_m)$ is $2\pi$ periodic with respect to each one of its variables, and we denote it by $\mu \in L^\infty_{\mathrm{per}}(\mathcal{V}_m)$, if

$$\forall k \in [\![-m,m]\!]\setminus\{0\},\ \mu(w + 2\pi e_k) = \mu(w),\ w \text{ a.e.}$$

The following algebraic lemma explains why these notations are well suited to DNLS.

**Lemma 3.3.1.** *If $m \in \mathbb{N}^*$, $\mu \in L^\infty_{\mathrm{per}}(\mathcal{V}_m)$ and $\boldsymbol{u} \in C^1(\mathbb{R}; L^2(\mathbb{Z}))$ is a solution of DNLS whose Shannon interpolation is denoted $u$, then we have*

$$i\partial_t\Lambda_m(\mu,\widehat{u}) = 2\Lambda_m\left(\mu D_m\cos,\widehat{u}\right) + \nu\Lambda_{m+1}(S_m\mu,\widehat{u}). \tag{3.10}$$

*Proof.* By definition, the quantity to identify can expanded as follow

$$\begin{aligned}
&i\partial_t\Lambda_m(\mu,\widehat{u})\\
&= \sum_{k=1}^{m}\int_{\mathcal{V}_m}\mu(w)\left[\overline{\widehat{u}(w_{-k})}i\partial_t\widehat{u}(w_k) + \widehat{u}(w_k)i\partial_t\overline{\widehat{u}(w_{-k})}\right]\prod_{j\neq k}\widehat{u}(w_j)\overline{\widehat{u}(w_{-j})}\mathrm{d}w\\
&=: \sum_{k=1}^{m} I_k + I_{-k}.
\end{aligned}$$

Now, we have to expand $I_k$ and $I_{-k}$ using the definition of DNLS. Applying Proposition 3.2.2 we get

$$\forall w_k \in (-\pi,\pi),\ i\partial_t\widehat{u}(w_k) = (2\cos w_k - 2)\widehat{u}(w_k) + \nu\sum_{\ell\in\mathbb{Z}}\widehat{|u|^2u}(w_k + 2\pi\ell).$$

So, since $\mu$ is $2\pi$ periodic the direction $e_k$, we deduce

$$\begin{aligned}
&I_k - \Lambda_m((2\cos w_k - 2)\mu, \widehat{u}) \\
=&\nu \int_{\mathcal{V}_m} \mu(w)\overline{\widehat{u}(w_{-k})} \left( \mathbb{1}_{w_k \in (-\pi,\pi)} \sum_{\ell \in \mathbb{Z}} \widehat{|u|^2 u}(w_k + 2\pi\ell) \right) \prod_{j \neq k} \widehat{u}(w_j)\overline{\widehat{u}(w_{-j})} \mathrm{d}w \\
=&\nu \int_{\mathcal{V}_m} \mu(w)\overline{\widehat{u}(w_{-k})}\widehat{|u|^2 u}(w_k) \prod_{j \neq k} \widehat{u}(w_j)\overline{\widehat{u}(w_{-j})} \mathrm{d}w.
\end{aligned}$$

However, since for all $\omega \in \mathbb{R}$, $\overline{\widehat{u}}(\omega) = \widehat{\overline{u}}(-\omega)$, we have, for all $w_k \in \mathbb{R}$,

$$\widehat{|u|^2 u}(w_k) = \int_{w_{m+1} - w_{-m-1} + \widetilde{w}_k = w_k} \widehat{u}(w_{m+1})\widehat{u}(\widetilde{w}_k)\overline{\widehat{u}}(w_{-m-1}) \mathrm{d}w_{m+1} \mathrm{d}w_{-m-1}.$$

So, realizing the change of variable $w_k \leftarrow \widetilde{w}_k$, we get

$$\int_{\mathcal{V}_m} \mu(w)\overline{\widehat{u}(w_{-k})}\widehat{|u|^2 u}(w_k) \prod_{j \neq k} \widehat{u}(w_j)\overline{\widehat{u}(w_{-j})} \mathrm{d}w = \Lambda_{m+1}(\mu(w + e_k(w_{m+1} - w_{m+1})), \widehat{u}).$$

Similarly, we could prove that

$$I_{-k} = -\Lambda_m((2\cos w_{-k} - 2)\mu, \widehat{u}) - \nu\Lambda_{m+1}(\mu(w - e_{-k}(w_{m+1} - w_{m+1})), \widehat{u}).$$

So, finally, we get

$$\begin{aligned}
&i\partial_t \Lambda_m(\mu, \widehat{u}) = \sum_{k=1}^{m} I_k + I_{-k} \\
=&\Lambda_m \left( \sum_{k=1}^{m} \left[ (2\cos w_k - 2) - (2\cos w_{-k} - 2) \right] \mu, \widehat{u} \right) \\
&+ \nu\Lambda_{m+1} \left( \sum_{k=1}^{m} \mu(w + e_k(w_{m+1} - w_{m+1})) - \mu(w - e_{-k}(w_{m+1} - w_{m+1})), \widehat{u} \right) \\
=&2\Lambda_m \left( \mu D_m \cos, \widehat{u} \right) + \nu\Lambda_{m+1}(S_m\mu, \widehat{u}).
\end{aligned}$$

□

**Corollary 3.3.1.** *Let $f \in L^\infty(\mathbb{R})$, let $\boldsymbol{u} \in C^1(\mathbb{R}; L^2(\mathbb{Z}))$ be a solution of DNLS and let $u$ be its Shannon interpolation. Then, we have*

$$\partial_t \int f(\omega)|\widehat{u}(\omega)|^2 \, \mathrm{d}\omega = \nu \frac{i}{2} \Lambda_2(D_2 f, \widehat{u}).$$

*Proof.* The result only involves values of $f$ for $\omega \in (-\pi, \pi)$. So we can assume that $f$ is a $2\pi$ periodic function. Now, we observe that, by definition, we have

$$\int f(\omega)|\widehat{u}(\omega)|^2 \, \mathrm{d}\omega = \Lambda_1(f(w_1), \widehat{u}).$$

So, applying Lemma 3.3.1, we get

$$\partial_t \int f(\omega)|\widehat{u}(\omega)|^2 \,\mathrm{d}\omega = -2i\Lambda_1\left((D_1\cos)f(w_1),\widehat{u}\right) - i\nu\Lambda_2(S_2\left[f(w_1)\right],\widehat{u}).$$

Since $2\pi$ periodic functions clearly belong to the $D_1$ kernel, the first term is zero. So we just need to identify the second term. Indeed, paying attention to its symmetries and remembering that we have assumed that $f$ is $2\pi$ periodic function, we get

$$\begin{aligned}\Lambda_2(S_2\left[f(w_1)\right],\widehat{u}) &= \Lambda_2(f(w_1+w_2-w_{-2})-f(w_1),\widehat{u})\\ &= \Lambda_2(f(w_{-1})-f(w_1),\widehat{u})\\ &= -\frac{1}{2}\Lambda_2(f(w_1)+f(w_2)-f(w_{-1})-f(w_{-2}),\widehat{u}).\end{aligned}$$

□

With these notations and results we can explain more precisely the construction of our *higher modified energies* . But first, we explain why it is natural to introduce correction terms in the construction of our modified energy.

In order to control the discrete $\dot{H}^n$ norm, it would seem natural to control its derivative. Indeed, if $\boldsymbol{u}$ is a solution of DNLS and if $u$ is its Shannon interpolation, applying Corollary 3.3.1 (and Corollary 3.2.2), we have

$$\partial_t \| \boldsymbol{u} \|^2_{\dot{H}^n(\mathbb{Z})} = \nu\frac{i}{4\pi}\Lambda_2\left(D_2\left(2\sin\frac{\omega}{2}\right)^{2n},\widehat{u}\right). \tag{3.11}$$

So a direct estimation of this derivative would naturally lead to (see Lemma 3.3.3 for a proof of this estimate)

$$\left|\partial_t \| \boldsymbol{u} \|^2_{\dot{H}^n(\mathbb{Z})}\right| \leqslant C\| \boldsymbol{u} \|^2_{\dot{H}^n(\mathbb{Z})}\| \boldsymbol{u} \|_{\dot{H}^1(\mathbb{Z})}\| \boldsymbol{u} \|_{L^2(\mathbb{Z})},$$

where $C>0$ is an universal constant. Then assuming that the discrete homogeneous $H^1$ norm can be controlled uniformly on time by $M_{\boldsymbol{u}(0)}$ (see Theorem 3.5 for the definition of $M_{\boldsymbol{u}(0)}$ and the next subsection for a proof) and applying Grönwall's inequality, we would get an universal constant $C>0$ such that, for all $t\geqslant 0$,

$$\| \boldsymbol{u}(t)\|_{\dot{H}^n(\mathbb{Z})} \leqslant \| \boldsymbol{u}(0)\|_{\dot{H}^n(\mathbb{Z})} e^{CM_{\boldsymbol{u}(0)}^{\frac{4}{3}}t}.$$

If we proceed by homogeneity to get a result depending on the stepsize $h$, we would get

$$\| \boldsymbol{u}(t)\|_{\dot{H}^n(h\mathbb{Z})} \leqslant \| \boldsymbol{u}(0)\|_{\dot{H}^n(h\mathbb{Z})} e^{CM_{\boldsymbol{u}(0)}^{\frac{4}{3}}t}.$$

Such a control is better than the trivial estimate (3.4) only for times shorter than $-\frac{n}{C}\log(h)$. So it is quite weak, if we compare it with the estimate of Theorem 3.5 because this later gives a non trivial control of $\| \boldsymbol{u}(t)\|_{\dot{H}^n(\mathbb{Z})}$ for times shorter than $h^{-\frac{2n}{n-1}}$.

So to improve this exponential bound, the idea of modified energy is to add a corrector term to $\| \boldsymbol{u} \|^2_{\dot{H}^n(\mathbb{Z})}$ in order to cancel its time derivative (3.11). However, it turns out that there is an

algebraic obstruction to this construction as shown in Lemma 3.3.2 below. For this reason, we consider another functional $\int f_n(\omega)|\widehat{u}(\omega)|^2 \,\mathrm{d}\omega$ where $f_n$ is a real function and such that this last quantity is equivalent to the square of the $\dot{H}^n(\mathbb{Z})$ norm. More precisely, observing the formula of the Hamiltonian (see (3.9)), we consider a modified energy $E_n$ given by

$$E_n(u) = \int f_n(\omega)|\widehat{u}(\omega)|^2 \,\mathrm{d}\omega + \Lambda_2(\mu_n, \widehat{u}),$$

where $\mu_n \in L^\infty(\mathcal{V}_2)$ is a function.

Applying Lemma 3.3.1 and its Corollary 3.3.1, if we want the correction term to cancel the derivative of $\int f_n(\omega)|\widehat{u}(\omega)|^2 \,\mathrm{d}\omega$ then $\mu_n$ has to solve the equation

$$\nu D_2 f_n = 4\mu_n D_2 \cos. \tag{3.12}$$

Furthermore, if $\mu_n$ is a solution of (3.12), we would have

$$\partial_t E_n(u) = -i\nu\Lambda_3(S_2\mu_n, \widehat{u}).$$

With this construction, we will be able to prove Theorem 3.1.1 by induction because we will prove that $\Lambda_2(\mu_n, \widehat{u})$ and $\Lambda_3(S_2\mu_n, \widehat{u})$ are controlled by the square of the $\dot{H}^{n-1}(\mathbb{Z})$ norm.

Of course, we would like to iterate this process cancelling the derivative of $E_n(u)$ adding a new term to our modified energy. However, such a construction involve major algebraic issues and we do not know if it is possible (we should find some criteria of divisibility by $D_3 \cos$ on the ring of trigonometric polynomials on $\mathcal{V}_3$).

To realize this strategy, we need, on the one hand, to design a function $\mu_n$ satisfying (3.12) without any singularity and, on the other hand, we need to control $\Lambda_2(\mu_n, \widehat{u})$ and $\Lambda_3(S_2\mu_n, \widehat{u})$ by the square of the $\dot{H}^{n-1}(\mathbb{Z})$ norm. The two following lemmas treat each one of these issues.

**Lemma 3.3.2.** *If $f \in C^\infty(\mathbb{R})$ satisfies $f \underset{\omega\to 0}{=} \mathcal{O}(\omega^{2n})$ , where $n \in \mathbb{N}^*$, and if $f$ is an even function and $x \mapsto f(x - \frac{\pi}{2}) - f(\frac{\pi}{2})$ is an odd function then there exists $C > 0$ such that we have*

$$\forall w \in \mathcal{V}_2, |D_2 f(w)| \leqslant C|D_2 \cos(w)| \sum_{j\in\{\pm1,\pm2\}} w_j^{2n-2}.$$

*Proof.* Since $f$ is an even function and $x \mapsto f(x-\frac{\pi}{2})-f(\frac{\pi}{2})$ is an odd function, $f$ is a $2\pi$ periodic function whose Fourier series is

$$f(\omega) = f(\frac{\pi}{2}) + \sum_{k\in\mathbb{N}} \beta_k \cos((2k+1)\omega) \text{ with } (b_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}.$$

Furthermore, since $f$ is a $C^\infty$ function, for all $m \in \mathbb{N}^*$, there exists $C_m > 0$ such that

$$\sum_{k\in\mathbb{N}} |\beta_k|(2k+1)^m \leqslant C_m.$$

To get compact notations, we define the function $\cos_k$ (and similarly $\sin_k$) by

$$\forall \omega \in \mathbb{R},\ \cos_k \omega := \cos((2k+1)\omega).$$

If we assume that $w_1 + w_2 = w_{-1} + w_{-2} + 2\pi j$, with $j \in \mathbb{N}$ then we have

$$\begin{aligned}
&D_2 \cos_k w \\
=&2\cos_k\left(\frac{w_1+w_2}{2}\right)\cos_k\left(\frac{w_1-w_2}{2}\right) - 2\cos_k\left(\frac{w_{-1}+w_{-2}}{2}\right)\cos_k\left(\frac{w_{-1}-w_{-2}}{2}\right)\\
=&2\cos_k\left(\frac{w_1+w_2}{2}\right)\left[\cos_k\left(\frac{w_1-w_2}{2}\right) - (-1)^j\cos_k\left(\frac{w_{-1}-w_{-2}}{2}\right)\right].
\end{aligned}$$

But since $2k+1$ is odd, we have

$$(-1)^j\cos_k\left(\frac{w_{-1}-w_{-2}}{2}\right) = \cos_k\left(\frac{w_{-1}-w_{-2}}{2} + \pi j\right).$$

So, we get

$$\begin{aligned}
D_2\cos_k w = 2\cos_k\left(\frac{w_1+w_2}{2}\right)&\left[\cos_k\left(\frac{w_1-w_2}{2}\right) - \cos_k\left(\frac{w_{-1}-w_{-2}}{2}+\pi j\right)\right]\\
&= 4\cos_k\left(\frac{w_1+w_2}{2}\right)\sin_k\left(\frac{w_1-w_2-w_{-1}+w_{-2}+2\pi j}{4}\right)\\
&\qquad\sin_k\left(\frac{w_1-w_2+w_{-1}-w_{-2}+2\pi j}{4}\right).
\end{aligned}$$

However, we know that

$$\forall\omega\in\mathbb{R},\ |\sin_k\omega| \leqslant (2k+1)|\sin(\omega)|.$$

Consequently, we can prove the same relation for $\cos_k$. Indeed, since $2k+1$ is an odd number, for all $\omega \in \mathbb{R}$, we have

$$\left|\cos_k\left(\omega+\frac{\pi}{2}\right)\right| = |\sin_k\omega| \leqslant (2k+1)|\sin(\omega)| = (2k+1)\left|\cos\left(\omega+\frac{\pi}{2}\right)\right|.$$

So we deduce that for all $w \in \mathcal{V}_2$, we have

$$|D_2\cos_k(w)| \leqslant (2k+1)^3|D_2\cos(w)|.$$

Consequently, we have

$$|D_2 f(w)| \leqslant C_3|D_2\cos(w)|. \tag{3.13}$$

To conclude this proof, we just need to improve (3.13) when $w$ is small enough. In this case, we can forget the aliasing terms because if $\max_{j\in\{\pm1,\pm2\}}|w_j| < \frac{\pi}{2}$ then $w_1+w_2 = w_{-1}+w_{-2}$.

Now we realize the following change of variable

$$\begin{cases}
X &= \frac{w_1-w_2+w_3-w_4}{4},\\
Y &= \frac{w_1-w_2-w_3+w_4}{4},\\
Z &= \frac{w_1+w_2}{2},\\
H &= w_1+w_2-w_3-w_4.
\end{cases}$$

Then we define

$$F(X,Y,Z,H) = D_2 f(w).$$

Previously, we have proven that, for all $X, Y, Z \in \mathbb{R}$,

$$F(X,Y,Z,0) = \sum_{k\in\mathbb{N}} \beta_k \cos_k Z \sin_k X \sin_k Y.$$

Consequently, we have

$$F(0,Y,Z,0) = 0 \text{ and } \partial_X F(X,0,Z,0) = 0.$$

So applying a Taylor expansion, we get

$$\begin{aligned} F(X,Y,Z,0) &= F(0,Y,Z,0) + X\int_0^1 \partial_X F(\alpha X,Y,Z,0)\mathrm{d}\alpha \\ &= X\int_0^1\int_0^1 \partial_X F(\alpha X,0,Z,0) + Y\partial_X\partial_Y F(\alpha X,\beta Y,Z,0)\mathrm{d}\beta\mathrm{d}\alpha \\ &= XY\int_0^1\int_0^1 \partial_X\partial_Y F(\alpha X,\beta Y,Z,0)\mathrm{d}\beta\mathrm{d}\alpha. \end{aligned}$$

However, since $f \underset{\omega\to 0}{=} \mathcal{O}(\omega^{2n})$, all the derivatives of $f$ of order less than $2n$ vanish in $0$. Thus, we deduce that all the derivative of $F$ of order less than $2n$ also vanish in $0$. Consequently, realising a Taylor expansion, we get a constant $c > 0$ such that if $|X| + |Y| + |Z| + |H| < 1$ then

$$|\partial_X\partial_Y F(X,Y,Z,H)| \leqslant c(|X| + |Y| + |Z| + |H|)^{2n-2}.$$

So, if $|X| + |Y| + |Z| < 1$, we have

$$|F(X,Y,Z,0)| \leqslant cXY(|X| + |Y| + |Z|)^{2n-2}.$$

Then we get

$$|F(X,Y,Z,0)| \leqslant \frac{c}{\cos 1(\operatorname{sinc} 1)^2}\cos Z \sin X \sin Y(|X| + |Y| + |Z|)^{2n-2}.$$

We can write this inequality with the variables $w_1, w_2, w_{-1}, w_{-2}$. So, since the norms are equivalent on $\{w \in \mathbb{R}^{\{\pm1,\pm2\}},\ w_1 + w_2 = w_{-1} + w_{-2}\}$, there exists $\kappa > 0$ such that for all $w \in \{w \in \mathbb{R}^{\{\pm1,\pm2\}},\ w_1 + w_2 = w_{-1} + w_{-2}\}$, we have

$$\kappa^{-1}(|X| + |Y| + |Z|)^{2n-2} \leqslant \sum_{j\in\{\pm1,\pm2\}} |w_j|^{2n-2} \leqslant \kappa(|X| + |Y| + |Z|)^{2n-2}.$$

Thus, there exists $C \in (0, \frac{\pi}{2})$ such that if $w \in \mathcal{V}_2$ satisfies $\max_{j\in\{\pm1,\pm2\}} |w_j| < C^{-1}$ then

$$|D_2 f(w)| \leqslant C|D_2\cos(w)| \sum_{j\in\{\pm1,\pm2\}} |w_j|^{2n-2}. \tag{3.14}$$

Finally, to prove the lemma we just need to use (3.14) when $w$ is small enough and (3.13) when it is large. □

**Lemma 3.3.3.** *Let $n, m \in \mathbb{N}$, $m \geqslant 2$. There exists $K > 0$ such that for all $u \in BL^2$, we have*

$$\Lambda_m\left(\sum_{j=1}^{m} w_j^{2n} + w_{-j}^{2n}, |\widehat{u}|\right) \leqslant K\|\partial_x^n u\|_{L^2(\mathbb{R})}^2 \|\partial_x u\|_{L^2(\mathbb{R})}^{m-1} \|u\|_{L^2(\mathbb{R})}^{m-1}.$$

*Proof.* This lemma is somehow a discrete integration by parts. By linearity, we just need to prove that

$$\Lambda_m\left(|w_1|^{2n}, |\widehat{u}|\right) \leqslant C\|\partial_x^n u\|_{L^2(\mathbb{R})}^2 \|\partial_x u\|_{L^2(\mathbb{R})}^{m-1} \|u\|_{L^2(\mathbb{R})}^{m-1}.$$

Since, $\operatorname{supp}|\widehat{u}| \subset [-\pi, \pi]$ we have

$$\Lambda_m\left(|w_1|^{2n}, |\widehat{u}|\right) = \sum_{k=1-m}^{m-1} \int_{\mathcal{V}_{m,k}} w_1^{2n} \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})| \mathrm{d}w.$$

where

$$\mathcal{V}_{m,k} := \left\{ w \in \mathbb{R}^{[\![-m,m]\!]\setminus\{0\}} \mid \sum_{j=1}^{m} w_j - w_{-j} = 2k\pi \right\}.$$

So, applying Jensen's inequality to $x \mapsto x^n$, we get

$$\begin{aligned}
&\frac{1}{(2m)^{n-1}} \int_{\mathcal{V}_{m,k}} w_1^{2n} \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})| \mathrm{d}w \\
=&\frac{1}{(2m)^{n-1}} \int_{\mathcal{V}_{m,k}} |\omega_1|^n \left| w_{-1} + \sum_{j=2}^{m} w_j - w_{-j} - 2k\pi \right|^n \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})| \mathrm{d}w \\
\leqslant& \int_{\mathcal{V}_{m,k}} |w_1|^n \left( |w_{-1}|^n + \sum_{j=2}^{m} |w_j|^n + |w_{-j}|^n + |2k\pi|^n \right) \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})| \mathrm{d}w \\
=&(m-1)\left[ (|\omega|^n|\widehat{u}|)^{*2} * |\widehat{u}|^{*m-2} * |\widehat{\widehat{u}}|^{*m} \right](2k\pi) \\
&+ m\left[ (|\omega|^n|\widehat{u}|) * (|\omega|^n|\widehat{\widehat{u}}|) * |\widehat{u}|^{*m-1} * |\widehat{\widehat{u}}|^{*m-1} \right](2k\pi) \\
&+ \int_{\mathcal{V}_{m,k}} |w_1|^n |2k\pi|^n \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})| \mathrm{d}w.
\end{aligned}$$

The first term can be estimated by an elementary Young convolution inequality to get

$$\begin{aligned}
&(m-1)\left[ (|\omega|^n|\widehat{u}|)^{*2} * |\widehat{u}|^{*m-2} * |\widehat{\widehat{u}}|^{*m} \right](2k\pi) \\
&+ m\left[ (|\omega|^n|\widehat{u}|) * (|\omega|^n|\widehat{\widehat{u}}|) * |\widehat{u}|^{*m-1} * |\widehat{\widehat{u}}|^{*m-1} \right](2k\pi) \\
\leqslant&(2m-1)\|\omega^n \widehat{u}\|_{L^2(\mathbb{R})}^2 \|\widehat{u}\|_{L^1(\mathbb{R})}^{2m-2}.
\end{aligned}$$

The second term is an aliasing term. If $k = 0$, this term is $0$, so we can assume $k \neq 0$. Now observe that if the sum of $2m$ numbers, all smaller than $1$ is larger than $2$ then at least $2$ of them

are larger than $\frac{1}{2m-1}$. Consequently, applying the same Young convolution inequality, we have

$$\begin{aligned}
&\int_{\mathcal{V}_{m,k}} |w_1|^n\,|2k\pi|^n \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})|\mathrm{d}w \\
\leqslant &\int\limits_{\substack{w\in\mathcal{V}_{m,k}\\ |w_{-1}|\geqslant\frac{\pi}{2m-1}}} |w_1|^n\,|2k\pi|^n \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})|\mathrm{d}w \\
&+ \int\limits_{\substack{w\in\mathcal{V}_{m,k}\\ |w_2|\geqslant\frac{\pi}{2m-1}}} |\omega_1|^n\,|2k\pi|^n \prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})|\mathrm{d}w \\
\leqslant &(2|k|(2m-1))^n \int_{\mathcal{V}_{m,k}} |w_1|^n(|\omega_2|^n+|\omega_{-1}|^n)\prod_{j=1}^{m} |\widehat{u}(w_j)||\widehat{u}(w_{-j})|\mathrm{d}w \\
\leqslant &2(2|k|(2m-1))^n \|\omega^n\widehat{u}\|^2_{L^2(\mathbb{R})}\|\widehat{u}\|^{2m-2}_{L^1(\mathbb{R})}.
\end{aligned}$$

To conclude rigorously this proof, we just need to control classically $\|\widehat{u}\|^2_{L^1(\mathbb{R})}$ by the product of $\|u\|_{L^2(\mathbb{R})}$ and $\|\partial_x u\|_{L^2(\mathbb{R})}$. Indeed, if $v\in H^1(\mathbb{R})$, using Cauchy Schwarz inequality, we get

$$\|\widehat{v}\|_{L^1(\mathbb{R})} \leqslant \sqrt{2\pi}\|\sqrt{1+\omega^2}\widehat{v}\|_{L^2(\mathbb{R})} = 2\pi\|v\|_{H^1(\mathbb{R})}.$$

So, optimizing this inequality with respect to $\lambda$ through the transformation $v\to v(\lambda x)$, we get

$$\|\widehat{v}\|_{L^1(\mathbb{R})} \leqslant \sqrt{8}\pi\|v\|_{L^2(\mathbb{R})}\|\partial_x v\|_{L^2(\mathbb{R})}.$$

□

### 3.3.2 Proof of Theorem 3.1.1 by induction

With all these tools, now, we prove Theorem 3.1.1. As explained at the beginning of this section, we just need to focus on the case $h=1$. We are going to proceed by induction.

• We focus on the case $n=1$. Let $\boldsymbol{u}\in C^1(\mathbb{R};L^2(\mathbb{Z}))$ be a solution of DNLS. Since $\mathrm{H}_{\mathrm{DNLS}}$ is a constant of the motion of DNLS, for all $t\in\mathbb{R}$, we have

$$\|\,\boldsymbol{u}(t)\|^2_{\dot{H}^1(\mathbb{Z})} - \frac{\nu}{2}\|\,\boldsymbol{u}(t)\|^4_{L^4(\mathbb{Z})} = \|\,\boldsymbol{u}(0)\|^2_{\dot{H}^1(\mathbb{Z})} - \frac{\nu}{2}\|\,\boldsymbol{u}(0)\|^4_{L^4(\mathbb{Z})}. \tag{3.15}$$

Since $\|\,\boldsymbol{u}\,\|^2_{L^2(\mathbb{Z})}$ is also a constant of the motion, we have

$$\|\,\boldsymbol{u}(t)\|^4_{L^4(\mathbb{Z})} \leqslant \|\,\boldsymbol{u}(0)\|^2_{L^2(\mathbb{Z})}\|\,\boldsymbol{u}(t)\|^2_{L^\infty(\mathbb{Z})}.$$

Let $u$ be the Shannon interpolation of $\boldsymbol{u}$. Since $u_{|\mathbb{Z}}=\boldsymbol{u}$ (see Proposition 3.2.1), we have

$$\|\,\boldsymbol{u}(t)\|^2_{L^\infty(\mathbb{Z})} \leqslant \|u(t)\|^2_{L^\infty(R)} \leqslant c\|\partial_x u(t)\|_{L^2(\mathbb{R})}\|u(t)\|_{L^2(\mathbb{R})},$$

where $c$ is an universal constant associated to the classical Sobolev embedding. Since Shannon interpolation is an isometry we have proven that

$$\| \boldsymbol{u}(t)\|^4_{L^4(h\mathbb{Z})} \leqslant c\| \boldsymbol{u}(0)\|^3_{L^2(\mathbb{Z})}\|\partial_x u(t)\|_{L^2(\mathbb{R})}.$$

Now applying the estimate of Corollary (3.2.2), we get a *discrete Gagliardo-Nirenberg* inequality (for a sharper version of this inequality see Lemma $3.4$ in [68])

$$\| \boldsymbol{u}(t)\|^4_{L^4(\mathbb{Z})} \leqslant \frac{2c}{\pi}\| \boldsymbol{u}(0)\|^3_{L^2(\mathbb{Z})}\| \boldsymbol{u}(t)\|_{\dot{H}^1(\mathbb{Z})}.$$

Applying this inequality to (3.15), we get

$$\| \boldsymbol{u}(t)\|^2_{\dot{H}^1(\mathbb{Z})} - \frac{c}{\pi}\| \boldsymbol{u}(0)\|^3_{L^2(\mathbb{Z})}\| \boldsymbol{u}(t)\|_{\dot{H}^1(\mathbb{Z})} \leqslant \| \boldsymbol{u}(0)\|^2_{\dot{H}^1(\mathbb{Z})}.$$

Consequently, we have proven that

$$\begin{aligned}\| \boldsymbol{u}(t)\|_{\dot{H}^1(\mathbb{Z})} &\leqslant \frac{c}{2\pi}\| \boldsymbol{u}(0)\|^3_{L^2(\mathbb{Z})} + \frac{1}{2}\sqrt{\left(\frac{c}{\pi}\| \boldsymbol{u}(0)\|^3_{L^2(\mathbb{Z})}\right)^2 + 4\| \boldsymbol{u}(0)\|^2_{\dot{H}^1(\mathbb{Z})}}\\ &\leqslant C\left(\| \boldsymbol{u}(0)\|_{\dot{H}^1(\mathbb{Z})} + \| \boldsymbol{u}(0)\|^3_{L^2(\mathbb{Z})}\right),\end{aligned}$$

with $C = \max(1, \frac{c}{\pi})$.

• Let $n \geqslant 2$, let $\boldsymbol{u} \in C^1(\mathbb{R}; L^2(\mathbb{Z}))$ be a solution of DNLS satisfying for all $t \in \mathbb{R}$

$$\|\partial_x^{n-1}u(t)\|^2_{L^2(\mathbb{R})} \leqslant C\left(\|\partial_x^{n-1}u_0\|^2_{L^2(\mathbb{R})} + M_{u_0}^{\frac{4n-2}{3}} + |t|^{n-2}M_{u_0}^{\frac{8n-10}{3}}\right), \tag{3.16}$$

where $u$ is the Shannon interpolation of $u$ and

$$M_{u(0)} = \|\partial_x u_0\|_{L^2(\mathbb{R})} + \|u_0\|^3_{L^2(\mathbb{R})}.$$

Here, it is easier to work with an inequality on $u$ instead of $\boldsymbol{u}$ but applying the estimate of Corollary (3.2.2), (3.16) is equivalent to the inequality of Theorem 3.1.1.

First, we are going to construct our modified energy with Lemma 3.3.2. So we have to choose our function $f_n$. This function has to satisfy some criteria. First, we want $\int f_n|\widehat{u}(\omega)|^2\,\mathrm{d}\omega$ to be equivalent to square of the homogeneous $H^n$ norm of $u$. So we are looking for a regular function $f_n$ such that

$$\forall\omega \in (-\pi, \pi),\ \alpha\omega^{2n} \leqslant f_n(\omega) \leqslant \alpha^{-1}\omega^{2n}. \tag{3.17}$$

Second, we want $f_n - f_n(\frac{\pi}{2})$ to be even in $0$ and odd in $\frac{\pi}{2}$. So we cannot choose $f_n(\omega) = \omega^{2n}$ or $f_n(\omega) = \left(2\sin(\frac{\omega}{2})\right)^{2n}$. To satisfy these symmetries it is natural to look for $f_n$ as a trigonometric polynomial.

By performing an analysis involving elementary linear algebra, we find that $f_n$ defined by

$$f_n(\omega) := 1 - \cos(\omega)\sum_{k=0}^{n-1}\frac{C_{2k}^k}{4^k}(\sin\omega)^{2k}, \tag{3.18}$$

is the trigonometric polynomial of minimal degree (and such $f(\frac{\pi}{2}) = 1$) satisfying the previous hypothesis. Indeed, by construction, $f_n - 1$ is even in $0$ and odd in $\frac{\pi}{2}$. Furthermore, in $\mathbb{R}[\![X]\!]$ (i.e. formally), we have (see, for example, formula $3.6.9$ in [3])

$$\frac{1}{\sqrt{1-X^2}} = \sum_{k\in\mathbb{N}} \frac{C_{2k}^k}{4^k} X^{2k}.$$

Since, for all $\omega \in (-\frac{\pi}{2}, \frac{\pi}{2})$, $\cos\omega = \sqrt{1-(\sin\omega)^2}$, we deduce that

$$f_n(\omega) = \cos(\omega) \sum_{k\geqslant n} \frac{C_{2k}^k}{4^k} (\sin\omega)^{2k}.$$

Consequently, we get $f_n > 0$ on $\omega \in (0, \frac{\pi}{2})$ and $f_n(\omega) \underset{\omega\to 0}{\sim} \frac{C_{2n}^n}{4^n}\omega^{2n}$. So, using the symmetries of $f_n$, we deduce that there exists $\alpha > 0$ such that (3.17) is satisfied.

Then we define on $\mathcal{V}_2$ a function $\mu_n \in L^\infty(\mathcal{V}_2)$ by

$$\mu_n = \frac{\nu}{4} \frac{D_2 f_n}{D_2 \cos}.$$

In Lemma 3.3.2, we have proven that $\mu_n$ is well defined as a $L^\infty(\mathcal{V}_2)$ function (in fact, we could have proven that it is a regular function). Furthermore, we have proven that for all $w \in \mathcal{V}_2$, we have

$$|\mu_n(w)| \leqslant C_n \sum_{j\in\{\pm1,\pm2\}} |w_j|^{2n-2}, \tag{3.19}$$

where $C_n$ depends only of $n$.

Then we define our modified energy, for $v \in BL_1^2$ by

$$E_n(v) := \int_{\mathbb{R}} f_n(\omega)|\widehat{v}(\omega)|^2 \,\mathrm{d}\omega + \Lambda_2(\mu_n, \widehat{v}).$$

Applying (3.17), we get, for all $t \in \mathbb{R}$,

$$\begin{aligned} 2\pi\alpha\|\partial_x^n u(t)\|_{L^2(\mathbb{R})}^2 &= \alpha \int \omega^{2n}|\widehat{u}(t,\omega)|^2 \,\mathrm{d}\omega \\ &\leqslant \int f_n(\omega)|\widehat{u}(t,\omega)|^2 \,\mathrm{d}\omega \\ &\leqslant |E_n(u_0)| + |\Lambda_2(\mu_n, \widehat{u}(t))| + \int_0^t |\partial_s E_n(u(s))| \,\mathrm{d}s\,. \end{aligned}$$

To conclude the induction step we have to control each one of these terms.

▷ First, we focus on $\int_0^t |\partial_s E_n(u(s))| \,\mathrm{d}s\,.$

Applying Lemma 3.3.1, we get

$$\partial_t E_n(u(t)) = -i\nu\Lambda_3(S_2\mu_n, \widehat{u(t)}).$$

So, applying (3.19) and Jensen's inequality to $x \mapsto x^{2n-2}$, we get

$$|\partial_t E_n(u(t))| \leqslant C_n 3^{2n-3} 4\Lambda_3(\sum_{j=1}^{3} w_j^{2n-2} + w_{-j}^{2n-2}, |\widehat{u(t)}|).$$

Consequently, applying Lemma 3.3.3, we get a constant $K_n > 0$ such that

$$|\partial_t E_n(u(t))| \leqslant K_n \|\partial_x^{n-1} u(t)\|_{L^2(\mathbb{R})}^2 \|\partial_x u(t)\|_{L^2(\mathbb{R})}^2 \|u(t)\|_{L^2(\mathbb{R})}^2.$$

However, as we have proven at the initial step, there exists an universal constant $c > 0$ such that

$$\forall t \in \mathbb{R},\ \|\partial_x u(t)\|_{L^2(\mathbb{R})}^2 \|u(t)\|_{L^2(\mathbb{R})}^2 \leqslant c M_{u_0}^{\frac{8}{3}}.$$

So, from the induction hypothesis (see (3.16)), we get

$$|\partial_t E_n(u(t))| \leqslant \kappa \left( \|\partial_x^{n-1} u_0\|_{L^2(\mathbb{R})}^2 M_{u_0}^{\frac{8}{3}} + M_{u_0}^{\frac{4n+6}{3}} + |t|^{n-2} M_{u_0}^{\frac{8n-2}{3}} \right),$$

with $\kappa = cCK_n$. Consequently, we have

$$\left| \int_0^t |\partial_s E_n(u(s))| \, \mathrm{d}s \right| \leqslant \kappa \left( |t| \|\partial_x^{n-1} u_0\|_{L^2(\mathbb{R})}^2 M_{u_0}^{\frac{8}{3}} + |t| M_{u_0}^{\frac{4n+6}{3}} + \frac{|t|^{n-1}}{n-1} M_{u_0}^{\frac{8n-2}{3}} \right).$$

It is almost the required estimate of the induction. In fact, we just need to modify it using Young inequalities. Indeed, on the one hand we have

$$|t| M_{u_0}^{\frac{4n+6}{3}} \leqslant \frac{|t|^{n-1}}{n-1} M_{u_0}^{\frac{8n-2}{3}} + \frac{n-2}{n-1} M_{u_0}^{\frac{4n+2}{3}}.$$

On the other hand, since, by Hölder inequality,

$$\|\partial_x^{n-1} u_0\|_{L^2(\mathbb{R})}^2 \leqslant \|\partial_x^n u_0\|_{L^2(\mathbb{R})}^{2\frac{n-2}{n-1}} \|\partial_x u_0\|_{L^2(\mathbb{R})}^{\frac{2}{n-1}}, \tag{3.20}$$

we have

$$\begin{aligned} |t| \|\partial_x^{n-1} u_0\|_{L^2(\mathbb{R})}^2 M_{u_0}^{\frac{8}{3}} &\leqslant |t| \|\partial_x^n u_0\|_{L^2(\mathbb{R})}^{2\frac{n-2}{n-1}} M_{u_0}^{\frac{8}{3}+\frac{2}{n-1}} \\ &\leqslant \frac{n-2}{n-1} \|\partial_x^n u_0\|_{L^2(\mathbb{R})}^2 + \frac{|t|^{n-1}}{n-1} M_{u_0}^{\frac{8n-2}{3}}. \end{aligned}$$

▷ Second, we focus on $|\Lambda_2(\mu_n, \widehat{u}(t))|$.

Here, we just need to apply (3.19) to get

$$|\Lambda_2(\mu_n, \widehat{u}(t))| \leqslant C_n \Lambda_2(\sum_{j=1}^{2} |w_j|^{2n-2} + |w_{-j}|^{2n-2}, |\widehat{u}(t)|).$$

So, we deduce of Lemma 3.3.3, that there exists $\kappa_n > 0$ such that

$$|\Lambda_2(\mu_n, \widehat{u}(t))| \leqslant \kappa_n \|\partial_x^{n-1} u(t)\|_{L^2(\mathbb{R})}^2 \|\partial_x u(t)\|_{L^2(\mathbb{R})} \|u(t)\|_{L^2(\mathbb{R})}.$$

Consequently, applying the induction hypothesis (see (3.16)), and the initial step, we have

$$|\Lambda_2(\mu_n, \widehat{u}(t))| \leqslant K \left( \|\partial_x^{n-1} u_0\|_{L^2(\mathbb{R})}^2 M_{u_0}^{\frac{4}{3}} + M_{u_0}^{\frac{4n+2}{3}} + |t|^{n-2} M_{u_0}^{\frac{8n-6}{3}} \right),$$

with $K = C\kappa_n c$ where $c$ is an universal constant.
As previously, we need to apply some Young inequalities to modify this estimate to get the induction estimate. On the one hand, we have

$$|t|^{n-2} M_{u_0}^{\frac{8n-6}{3}} \leqslant \frac{n-2}{n-1} t^{n-1} M_{u_0}^{\frac{8n-2}{3}} + \frac{1}{n-1} M_{u_0}^{\frac{4n+2}{3}}.$$

On the other hand, applying (3.20), we get

$$\|\partial_x^{n-1} u_0\|_{L^2(\mathbb{R})}^2 M_{u_0}^{\frac{4}{3}} \leqslant \|\partial_x^n u_0\|_{L^2(\mathbb{R})}^{2\frac{n-2}{n-1}} M_{u(0)}^{\frac{4}{3}+\frac{2}{n-1}} \leqslant \frac{n-2}{n-1} \|\partial_x^n u_0\|_{L^2(\mathbb{R})}^2 + \frac{1}{n-1} M_{u_0}^{\frac{4n+2}{3}}.$$

▷ Finally, we focus on $|E_n(u_0)|$.

We apply the triangle inequality to get

$$|E_n(u_0)| \leqslant \int f_n(\omega) |\widehat{u_0}(\omega)|^2 \, \mathrm{d}\omega + |\Lambda_2(\mu_n, \widehat{u_0})|.$$

On the one hand, applying (3.17), we get

$$\int f_n(\omega) |\widehat{u_0}(\omega)|^2 \, \mathrm{d}\omega \leqslant \alpha^{-1} \int \omega^{2n} |\widehat{u}(\omega)|^2 \, \mathrm{d}\omega = 2\pi\alpha^{-1} \|\partial_x^{2n} u_0\|_{L^2(\mathbb{R})}^2.$$

On the other hand, applying the estimate of $\Lambda_2(\mu_n, \widehat{u}(t))$, when $t = 0$, we get

$$|\Lambda_2(\mu_n, \widehat{u_0})| \leqslant K \left( \frac{n-2}{n-1} \|\partial_x^n u_0\|_{L^2(\mathbb{R})}^2 + \left[ 1 + \frac{1}{n-1} \right] M_{u_0}^{\frac{4n+2}{3}} \right).$$

CHAPITRE 4

# RATIONAL NORMAL FORMS AND STABILITY OF SMALL SOLUTIONS TO NONLINEAR SCHRÖDINGER EQUATIONS

## 4.1 Introduction

In this paper we are interested in the long time behavior of small amplitude solutions of non-linear Hamiltonian partial differential equations on bounded domains. In this context, the competition between non-linear effects and energy conservation (typically the $H^1$ Sobolev norm) makes the problem intricate. One of the main issues is the control of higher order Sobolev norms of solutions for which typically a priori upper bounds are polynomials (see [36, 108, 40, 105, 52]).
Bourgain exhibited in [36] examples of growth of high order Sobolev norms for some solutions of a nonlinear wave equation in 1d with periodic boundary conditions. These examples were constructed by using as much as possible the totally resonant character of the equation (all the linear frequencies are integers and thus proportional).
On the contrary, Bambusi & Grébert have shown in [18] (see also [14, 40]) that, in a fairly general semi linear PDE framework, if an appropriate non-resonance condition is imposed on the linear part then the solution $u$ of the corresponding PDE satisfy a strong stability property :

$$\text{if } \|u(0)\|_{H^s} \leqslant \varepsilon \quad \text{then} \quad \|u(t)\|_{H^s} \leqslant 2\varepsilon \quad \text{for all} \quad t \leqslant \varepsilon^{-M}, \tag{4.1}$$

where $\|\cdot\|_{H^s}$ denotes the Sobolev norm of order $s$, $M$ can be chosen arbitrarily large and $\varepsilon$ is supposed to be small enough, $\varepsilon < \varepsilon_0(M, s)$. The method of proof is based on the construction of Birkhoff normal forms. To verify the appropriate non-resonance condition external parameters were used –such as a mass in the case of nonlinear wave equation– and the stability result were obtained for almost every value of these parameters. Then this technic was applied to prove almost global existence results for a lot of semi linear Hamiltonian PDEs (see [16, 15, 71, 72, 64]). However, the case of a non-linear perturbation of a fully resonant linear PDE was not achievable by this technique. Actually for the cubic nonlinear Schrödinger equation on the two dimensional torus it is proved in [51] that the high Sobolev norms may growth arbitrarily for some special initial data. Even in one dimension of space, it is proved in [73] that the quintic nonlinear Schrödinger equation on the circle does not satisfy (4.1) (but without arbitrary growth of the high Sobolev norms, see also [80] for a generalization or [43] for a two-dimensional example).

This chapter is a joint work with Erwan Faou and Benoît Grébert realized in [27].

Now consider the nonlinear Schrödinger equation :

$$iu_t = -\Delta u + \varphi(|u|^2)u, \quad x \in \mathbb{T}, \quad t \in \mathbb{R}, \tag{NLS}$$

where $\varphi = \mathbb{R} \to \mathbb{R}$ is an analytic function on a neighborhood of the origin satisfying $\varphi(0) = m$ is the mass possibly 0 and $\varphi'(0) \neq 0$. Equation (NLS) is a Hamiltonian system associated with the Hamiltonian function

$$H_{\mathrm{NLS}}(u, \bar{u}) = \frac{1}{2\pi} \int_{\mathbb{T}} \left( |\nabla u|^2 + g(|u|^2) \right) \mathrm{d}x, \tag{4.2}$$

where $g(t) = \int_0^t \varphi$, and the complex symplectic structure $i\mathrm{d}u \wedge \mathrm{d}\bar{u}$.
This is an example of fully resonant Hamiltonian PDE, as the linear frequencies are $j^2 \in \mathbb{N}$ for $j \in \mathbb{Z}$. Nevertheless in [89] Kuksin & Pöschel proved for such equation the persistence of finite dimensional KAM tori, a result that requires a strong non resonant property on the unperturbed Hamiltonian. Actually they considered the cubic term as part of the unperturbed Hamiltonian to modulate the resonant linear frequencies and to avoid the problem of resonances. Roughly speaking the nonlinear term generates stability. Then Bourgain in [39] used the same idea to prove that for many random small initial data the solution of (NLS) satisfies (4.1). Although the method of proof is based on normal forms, the effective construction of the normal form depends on the initial datum in a very intricated way and actually the author does not obtain a Birkhoff normal form result for (NLS) but rather a way to break down the solution that allows him to obtain the property (4.1).
In this work we want to construct a new type of normal form, not based on polynomial functions but on rational functions (see Section 4.6), transforming the Hamiltonian of (NLS) into an integrable one up to a small remainder, over large open sets surrounding the origin. Then stability of higher order Sobolev norms during very long time is just one of the dynamical consequences. We stress out that since our rational normal form is built on open sets, the dynamical consequences remain stable with respect to the initial datum. In particular the property (4.1), although not verified on all a neighborhood of the origin, is locally uniform with respect to $u(0)$ in $H^s$ .
To describe our result let us introduce some notations. With a given function $u(x) \in L^2$, we associate the Fourier coefficients $(u_a)_{a\in\mathbb{Z}} \in \ell^2$ defined by

$$u_a = \frac{1}{2\pi} \int_{\mathbb{T}} u(x) e^{-iax} \mathrm{d}x.$$

In the remainder of the paper we identify the function with its sequence of Fourier coefficients $u = (u_a)_{a\in\mathbb{Z}}$, and as in [64, 63] we consider the spaces

$$\ell^1_s = \{ u = (u_a)_{a\in\mathbb{Z}} \in \ell^2 \quad | \quad \|u\|_s := \sum_{a\in\mathbb{Z}} \langle a \rangle^s |u_a| < +\infty \}, \tag{4.3}$$

where $\langle a \rangle = \sqrt{1 + a^2}$. Note that these spaces are linked with the classical Sobolev spaces by the relation $H^{s'} \subset \ell^1_s \subset H^s$ for $s' - s > 1/2$.

Our method also applies to equations with Hartree nonlinearity of the form (Schrödinger-Poisson equation)

$$\begin{aligned} iu_t &= -\Delta u + Wu, \quad x \in \mathbb{T}, \quad t \in \mathbb{R}, \\ -\Delta W &= |u|^2 - \frac{1}{2\pi} \int_{\mathbb{T}} |u|^2 \mathrm{d}x, \end{aligned} \tag{NLSP}$$

for which we have $W = V \star |u|^2$ where $V$ is the Green function of the operator $-\Delta$ with zero average on the torus. The Hamiltonian associated with this equation is

$$H_{\mathrm{NLSP}}(u, \bar{u}) = \frac{1}{2\pi} \int_{\mathbb{T}} \left( |\nabla u|^2 + \frac{1}{2}(V \star |u|^2)|u|^2 \right) \mathrm{d}x. \tag{4.4}$$

Different kind of convolution operators $V$ could also be considered, as well as higher order perturbations of (NLSP) (note however that unlike the (NLS) case the cubic (NLSP) equation is not integrable in dimension 1). As we will see, the probabilistic results obtained for (NLS) and (NLSP) differ significantly.

Our results are divided into two parts :

• **Abstract rational normal forms** (see Theorem 4.2.1). We construct a canonical transformation $\tau$ defined on an open set $\mathcal{V}_\varepsilon \subset \ell^1_s$ included in the ball of radius $\varepsilon$ centered at $0$ that puts the Hamiltonian of (NLS) (resp. (NLSP)) in normal form up to order $2r$ : $H \circ \tau = Z(I) + R$ where $Z$ depends only on the actions $I = (I_a)_{a\in\mathbb{Z}}$ with $I_a = |u_a|^2$, and $R = O(\varepsilon^{2r+1})$. The proof for this result is outlined in Section 4.2.2 and demonstrated in Section 4.7.
Of course the open set $\mathcal{V}_\varepsilon$ is defined in a rather complex way through non-resonant relationships between actions $|u_a|^2$ and $\varepsilon$ (see section 4.4). In particular it does not contain $u \equiv 0$ which is too resonant. Its construction relies on a *ultra-violet cut-off* as in classical KAM theory (particularly in [7]), here in an infinite dimensional setting. Moreover, these sets are invariant by angular rotation in the sense that

$$\forall\, (\theta_a)_{a\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}, \quad u = (u_a)_{a\in\mathbb{Z}} \in \mathcal{V}_\varepsilon \Longrightarrow (e^{i\theta_a} u_a)_{a\in\mathbb{Z}} \in \mathcal{V}_\varepsilon. \tag{4.5}$$

It is then necessary to show that the flow travels within these open sets. This is achieved in a second step.

• **Generic almost preservation of the actions over very long time** (see Theorems 4.2.3 and 4.2.4). For a given $\varepsilon > 0$, we set

$$u(0, x) = u^0_\varepsilon(x) = c\varepsilon \sum_{a\in\mathbb{Z}} \sqrt{I_a} e^{iax}, \tag{4.6}$$

where $I = (I_a)_{a\in\mathbb{Z}}$ are random variables with support included in the interval $(0, \langle a\rangle^{-2s+4})$, so that $u^0_\varepsilon$ belong to the space $\ell^1_s$ and $c = (2\pi)^{-1} \tanh \pi$ is a normalizing constant to ensure $\|u^0_\varepsilon\|_s < \varepsilon/2$ almost surely. We prove that under some assumptions on the law of $I_a$, then for essentially almost all couple $(\varepsilon, I)$ and $\varepsilon$ small enough, the initial values $u^0_\varepsilon$ of the form (4.6) are in the domain of definition of the normal forms, and thus have a dynamics that is essentially an integrable one over very long time. This implies the almost preservation of the actions $|u_a(t)|^2$ over times of order $\varepsilon^{-M}$ with $M$ arbitrary which in turn implies that the solution remains inside the open set $\mathcal{V}_\varepsilon$ . In particular we deduce the almost preservation of the Sobolev norm of the solution over times of order $\varepsilon^{-M}$, i.e. property (4.1). This second step is detailed in section 4.8.

We show a difference between (NLS) and (NLSP) for which we obtain a stronger result. Indeed, whereas possible resonances between $\varepsilon$ and the actions $I_a$ can appear in (NLS), this is not the case for (NLSP) which thus can be seen as a more robust equation than (NLS).

As previously mentioned, the possibility of obtaining normal forms without the help of external parameters was already known in the KAM theory (see [89] and also [62]). However these

normal forms were constructed around finite dimensional tori. The originality of our analysis is that we work with truly infinite dimensional objects.

The question of building full dimensional invariant tori by using our rational normal forms is under study.

It would also be nice to apply this new normal form technique to other PDEs, especially in higher dimension. Nevertheless, there is an important limitation : we use in an unavoidable way the fact that the dominant term of the non-linearity (the cubic term for (NLS) and (NLSP)) are completely integrable (they depend only on actions). This is no longer true for the quintic NLS equation (see [73]) or for (NLS) and (NLSP)) equations in higher dimension. It should be noted that in the case of the beam equation studied in [62], the cubic term is also not integrable and this does not prevent a KAM-type result from being obtained. But in this case, a finite number of symplectic transformations make it possible to get rid of the angles corresponding to the modes of the finite dimensional torus that is perturbed. In our case, we would need an infinite number of such transformations, which is not accessible because these transformations are not close to identity.

Finally let us mention two recent results that open new directions in the world of Birkhoff normal forms. In [29] Berti-Delors have considered recently Birkhoff normal forms for a quasi linear PDE, namely the capillarity-gravity water waves equation, and thus faced unbounded nonlinearity. In this paper capillarity plays the role of the external parameter. Also in [31] Biasco-Masseti-Procesi, considering a suitable Diophantine condition, prove exponential stability in Sobolev norm for parameter dependent NLS on the circle.

**Acknowledgments.** During the preparation of this work the three authors benefited from the support of the Centre Henri Lebesgue ANR-11-LABX- 0020-01 and B.G. was supported by ANR -15-CE40-0001-02 “BEKAM” and ANR-16-CE40-0013 “ISDEEC” of the Agence Nationale de la Recherche.

## 4.2 Statement of the results and sketch of the proof

### 4.2.1 Main results

First we introduce the Hamiltonians associated with (NLS) and (NLSP) written in Fourier variables :

$$H_{\mathrm{NLS}} = \sum_{a\in\mathbb{Z}} a^2|u_a|^2 + \frac{1}{2\pi}\int_{\mathbb{T}} g\Big(\sum_{a,b\in\mathbb{Z}} u_a\bar{u}_b e^{i(a-b)x}\Big)\mathrm{d}x, \quad \text{and} \tag{4.7}$$

$$H_{\mathrm{NLSP}} = \sum_{a\in\mathbb{Z}} a^2|u_a|^2 + \frac{1}{4\pi}\int_{\mathbb{T}} \Big(\sum_{a,b\in\mathbb{Z}} \hat{V}_{a-b}u_a\bar{u}_b e^{i(a-b)x}\Big)\Big(\sum_{a,b\in\mathbb{Z}} \hat{u}_a\bar{u}_b e^{i(a-b)x}\Big)\mathrm{d}x, \tag{4.8}$$

where the Fourier transform $\hat{V}_a = a^{-2}$, $a \neq 0$, $\hat{V}_0 = 0$ is associated with the Green function of the operator $-\Delta$ with zero average on $\mathbb{T}$.

**Theorem 4.2.1** ((NLS) and (NLSP) cases)**.** *Let $H$ equals $H_{\mathrm{NLS}}$ or $H_{\mathrm{NLSP}}$. For all $r \geqslant 2$, there exists $s_0 \equiv s_0(r) = O(r^2)$ such that for all $s \geqslant s_0$ the following holds :*
*There exists $\varepsilon_0 \equiv \varepsilon_0(r,s)$ such that for all $\varepsilon < \varepsilon_0$, there exist open sets $\mathcal{C}_{\varepsilon,r,s}$ and $\mathcal{O}_{\varepsilon,r,s}$ included*

*in $B_s(0,4\varepsilon)$ the ball of radius $4\varepsilon$ centered at the origin in $\ell^1_s$, and an analytic canonical and bijective transformation $\tau : \mathcal{C}_{\varepsilon,r,s} \mapsto \mathcal{O}_{\varepsilon,r,s}$ satisfying*

$$\|\tau(z) - z\|_s \leqslant \varepsilon^{\frac{3}{2}} \quad \forall z \in \mathcal{C}_{\varepsilon,r,s}, \tag{4.9}$$

*that puts $H$ in normal form up to order $2r$ :*

$$H \circ \tau = Z + R$$

*where*

- *$Z = Z(I)$ is a smooth function of the actions and thus is an integrable Hamiltonian ;*
- *the remainder $R$ is of order $\varepsilon^{2r+2}$ on $\mathcal{C}_{\varepsilon,r,s}$, precisely*

$$\|X_R(z)\|_s \leqslant \varepsilon^{2r+1} \quad \textit{for all } z \in \mathcal{C}_{\varepsilon,r,s}. \tag{4.10}$$

This theorem is proved in section 4.7.
In section 4.8, we prove that for all $\varepsilon > 0$, there exists a set of initial data included in $\mathcal{O}_{\varepsilon,r,s}$ on which (4.1) holds true :

**Theorem 4.2.2** ((NLS) and (NLSP) cases)**.** *Let $H$ and $\varepsilon_0(r,s)$ as in Theorem 4.2.1 and let $u_a(t)$ denotes the Fourier coefficients of the solution $u(t,x)$ of the Hamiltonian system associated with $H$. Then for all $\varepsilon < \varepsilon_0$ there exists an open set $\mathcal{V}_{\varepsilon,r,s} \subset \mathcal{O}_{\varepsilon,r,s}$ invariant by angle rotation in the sense of* (4.5)*, such that for all $(u_a(0))_{a\in\mathbb{Z}} \in \mathcal{V}_{\varepsilon,r,s}$ we have for all $t \leqslant \varepsilon^{-2r+1}$*

$$\sup_{a\in\mathbb{Z}} \langle a\rangle^{2s} \left||u_a(t)|^2 - |u_a(0)|^2\right| \leqslant 3\varepsilon^{\frac{5}{2}}, \tag{4.11}$$

$$(u_a(t))_{a\in\mathbb{Z}} \in \mathcal{O}_{\varepsilon,r,s} \quad \textit{and in particular} \quad \|u(t)\|_s \leqslant 4\varepsilon. \tag{4.12}$$

*Furthermore there exists a full dimensional torus $\mathcal{T}_0 \in \ell^1_s$ such that for all $r_1 + r_2 = 2r + 2$*

$$\textit{dist}_s(u(t), \mathcal{T}_0) \leqslant C\varepsilon^{r_1} \quad \textit{for } |t| \leqslant 1/\varepsilon^{r_2} \tag{4.13}$$

*where dist$_s$ denotes the distance on $\ell^1_s$ associated with the norm $\|\cdot\|_s$*

The next step is to describe the non resonant sets $\mathcal{V}_{\varepsilon,r,s}$ which, as we said, are open, invariant by rotation (see (4.5)) and included in the ball of $\ell^1_s$ centered at 0 and of radius $\varepsilon$ but does not contain the origin. The following Theorems, proved in section 4.4, show that in both cases these open sets contain *many* elements of the form (4.6) but is much larger in the (NLSP) case that in the (NLS) case.

The first result concerns the nonlinear Schrödinger case (NLS).

**Theorem 4.2.3** ((NLS) case)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let us assume that $I : \Omega \mapsto (\mathbb{R}_+)^{\mathbb{Z}}$ are random variables satisfying*

- *$(I_a)_{a\in\mathbb{Z}}$ are independent,*
- *for each $a \in \mathbb{Z}$, $I_a^2$ is uniformly distributed in $(0, \langle a\rangle^{-4s-8})$,*

*and let $u^0_\varepsilon$ be the familly of random variables defined by* (4.6)*.*
*Let $r$, $s \geqslant s_0(r)$ and $\varepsilon_0(r,s)$ as in Theorem*(4.2.1) *for* (NLS)*. Then*

- *for all $0 \leqslant \varepsilon < \varepsilon_0(r,s)$*

$$\mathbb{P}\left(u^0_\varepsilon \in \mathcal{V}_{\varepsilon,r,s}\right) \geqslant 1 - \varepsilon^{\frac{1}{3}}. \tag{4.14}$$

- *for all* $0 \leqslant \varepsilon < \varepsilon_0(r,s)$ *and for all sequence* $(x_n)_{n\in\mathbb{N}}$ *of random variables uniformly distributed in* $(0,1)$ *and independent of* $I_a$*, there is a probability larger than* $1-\varepsilon^{\frac{1}{6}}$ *to realize* $I$ *such that there is a probability larger than* $1-\varepsilon^{\frac{1}{6}}$ *to realize* $(\varepsilon_n)_{n\in\mathbb{N}} := (\varepsilon 2^{-(n+x_n)})_{n\in\mathbb{N}}$ *such that* $u^0_{\varepsilon_n}$ *is non-resonant for all* $n$ *(i.e.* $u^0_{\varepsilon_n} \in \mathcal{V}_{\varepsilon_n,r,s}$*). More formally, we have*

$$\mathbb{P}\left(\mathbb{P}\left(\forall n \in \mathbb{N},\ u^0_{\varepsilon_n} \in \mathcal{V}_{\varepsilon_n,r,s}\,|I\right) \geqslant 1-\varepsilon^{\frac{1}{6}}\right) \geqslant 1-\varepsilon^{\frac{1}{6}}, \tag{4.15}$$

*where* $\mathbb{P}(\,\cdot\,|I)$ *denote the probability conditionally to the distribution* $I = (I_a)_{a\in\mathbb{Z}}$

The first part of the statement corresponds to fixing an $\varepsilon$ and removing some resonant set of $I_a$ (depending on $\varepsilon$) the second part shows that for a given distribution of $I_a$, we can take a lot of arbitrarily small $\varepsilon$ fulfilling the assumptions of the Theorem. Moreover, as the set $\mathcal{V}_{\varepsilon,r,s}$ is invariant by angle rotation, then for a given $u^0_\varepsilon \in \mathcal{V}_{\varepsilon,r,s}$ of the form (4.6), all the rotated functions of the form (4.5) belong to $\mathcal{V}_{\varepsilon,r,s}$.

The authors would like to mention that (4.15) corresponds to many numerical experiments confirming the absence of drift over long times when $\varepsilon \to 0$ for a generic initial distribution of $I_a$ : in other words this statement correspond to what is generically numerically observed, confirming a sort of *generic behaviour* for solutions of (NLS) for which no energy exchanges is observed between the frequencies over very long times.

The corresponding analysis for the Schrödinger-Poisson case leads to a better result :

**Theorem 4.2.4** ((NLSP) case)**.** *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space, and let us assume that* $I : \Omega \mapsto (\mathbb{R}_+)^{\mathbb{Z}}$ *are random variables satisfying*

- $(I_a)_{a\in\mathbb{Z}}$ *are independent,*
- *for each* $a \in \mathbb{Z}$*,* $I_a$ *is uniformly distributed in* $(0, \langle a\rangle^{-2s-4})$*,*

*and let* $u^0_\varepsilon$ *be the familly of random variables defined by* (4.6)*.*
*Let* $r$*,* $s \geqslant s_0(r)$ *and* $\varepsilon_0(r,s)$ *as in Theorem*(4.2.1) *for* (NLSP)*. Then for all* $0 \leqslant \varepsilon < \varepsilon_0(r,s)$

$$\mathbb{P}\left(\forall \varepsilon' < \varepsilon, u^0_{\varepsilon'} \in \mathcal{V}_{\varepsilon',r,s}\right) \geqslant 1-\varepsilon^{\frac{1}{3}}. \tag{4.16}$$

This statement allows to take one distribution of the action $I_a$ fulfilling some generic non resonance condition, and then to take arbitrarily small $\varepsilon$ in the initial value (4.6) independently on the $I_a$. Thus the result for (NLSP) is much stronger from the point of view of phase space : stable initial distributions are much more likely for (NLSP) than for (NLS).

In the remainder of the paper, we will essentially focus on the (NLS) case. The proof of the (NLSP) case will be outlined in appendix 4.9.1, where we stress the difference with (NLS), which are mostly major simplifications.

### 4.2.2 Sketch of proof

In this section we explain the strategy of the proof and we describe the new mathematical objects needed. The starting point is to write (formally) the Hamiltonian (4.2) as

$$H = Z_2(I) + P_4 + \sum_{m\geqslant 3} P_{2m}$$

where $I$ denote the collection of $I_a = |u_a|^2$, $a \in \mathbb{Z}$, and where

$$Z_2 = \sum_{a\in\mathbb{Z}} (a^2 + \varphi(0)) I_a \tag{4.17}$$

is the Hamiltonian associated with the linear part of the equation. The Hamiltonians $P_{2m}$ are polynomials of order $2m$ in the variables $u_a$, and $P_4$ is explicitely given by

$$P_4 = \frac{1}{2} \sum_{a+b=c+d} \hat{V}_{a-c} u_a u_b \bar{u}_c \bar{u}_d$$

where $\hat{V}_a = \varphi'(0)$ in the case of (NLS) and $\hat{V}_a = a^{-2}$, $a \neq 0$, $\hat{V}_0 = 0$, in the case of (NLSP). The first step is to perform a first resonant normal form transform $\tau_2$ with respect to $Z_2$. This step is classic and can be found for the first time in [89]. After some iterations, the new Hamiltonian can be written

$$H \circ \tau_2 = Z_2(I) + Z_4(I) + Z_6(I) + R_6(u) + \sum_{m=4}^{r} K_{2m} + R$$

where $R$ is of order $2r+2$, $K_{2m}$ are polynomial of order $2m$, and $Z_4$ and $Z_6$ are polynomials of degre 4 and 6 containing only actions of the form $I_a$. Moreover, the polynomials $K_{2m}$ are resonant in the sense that they contain only monomials of the form $u_{a_1} \cdots u_{a_m} \bar{u}_{b_1} \cdots \bar{u}_{b_m}$ where the collection of indices satisfy the relation

$$a_1 + \cdots + a_m = b_1 + \cdots + b_m \quad \text{and} \quad a_1^2 + \cdots + a_m^2 = b_1^2 + \cdots + b_m^2. \tag{4.18}$$

Indeed, these monomials correspond to the kernel of the operator $\chi \mapsto \{Z_2, \chi\}$ when applied to polynomials, which is the engine of the construction of the normal form. Note that at this stage, no small divisor problem occur. Now natural idea consists in using the term $Z_4$ to eliminate iteratively the terms of order $2m$ that do not depend on the actions. The general strategy is the following :

*(i)* Truncate all polynomials $K_{2m}$ and remove all the monomials containing at least three indices of size greater than $N$. The remainder term, as already noticed, see for instance [39, 71, 18], is thus of order $\varepsilon^5 N^{-s}$. Moreover, taking into account the resonance condition (4.18), the remaining truncated monomials have *irreducible* part - meaning they do not contain actions - with indices bounded by $\mathcal{O}(N^2)$. Taking $N$ so that $N^{-s} = \mathcal{O}(\varepsilon^r)$ (so typically $N = \varepsilon^{-r/s}$) then ensures that this term will be small enough to control the dynamics over a time of order $\varepsilon^{-r}$ by using a bootstrap argument.

*(ii)* Construct iteratively normal form transformation to eliminate the remaining part of degree $2m$ that do not depend only of the action by using the integrable Hamiltonian $Z_4$ which is explicitly given. The engine underlying this construction is thus to solve iteratively homological equations of the form $\{Z_4, \chi\} = Z + K$ where $K$ is given and do not contains terms depending only on the actions. This step makes appear small denominators depending on $I$, so that such a construction makes naturally appear rational functions (see Section 4.6) and not only polynomials. However, the division by small denominators depending on $I$ also yields poles in the normal form transforms. To avoid them, we use generic non resonance conditions on the distribution $I$, and the resolution of the homological equation thus brings loss of derivatives. As the small denominators are generated by irreducible monomials whose modes are bounded by

$N^2$, in the estimates this step results in a loss of order $N^\alpha$ for some $\alpha > 0$ versus a gain of $\varepsilon$ at each step of the normal form construction.

*(iii)* Try to trade off a few powers of $\varepsilon$ to control the normal form construction. This is done by a condition of the form $\varepsilon N^\alpha < 1$. The heart of the analysis is thus to make $\alpha$ independent of $s$ so that for $s$ large enough, such condition can be satisfied and will be compatible with $N = \varepsilon^{-r/s}$. If this is the case, the Hamiltonian thus depends only on the actions and a small remainder term of order $\varepsilon^r$ which allows to conclude. The remaining difficulty is to handle the algebra of rational functionals, and the control of the non resonant sets after each normal form steps.

With this roadmap in hand, the expression of $Z_4$ is fundamental as it drives the small denominators. Here appears a drastic difference between (NLS) and (NLSP). Indeed, the first normal term corresponding to the Hamiltonian $P_4$ is of the form

$$Z_4 = \sum_{a \neq b \in \mathbb{Z}} \hat{V}_{a-b} I_a I_b + \frac{1}{2} \hat{V}_0 \sum_{a \in \mathbb{Z}} I_a^2 \tag{4.19}$$

The frequencies associated with this integrable Hamiltonian are of the form

$$\lambda_a(I) = \frac{\partial Z_4}{\partial I_a} = 2 \sum_{b \neq a \in \mathbb{Z}} \hat{V}_{a-b} I_b + \hat{V}_0 I_a.$$

We thus observe that for (NLSP) for which $\hat{V}_a = a^{-2}$, $a \neq 0$ and $\hat{V}_0 = 0$, if $u \in \ell^1_s$ with large $s$, these frequencies are essentially dominated by low modes $I_b$ at a scale $\langle a \rangle^{-2}$. Hence it is easy to prove that the small denominators associated with $Z_4$ which are linear combinations of the $\lambda_a(I)$ are generically non resonant, with a loss of derivative independent of $s$.

Hence for (NLSP), we can work through the previous programme, and the coefficients $\alpha$ in the condition $\varepsilon N^\alpha < 1$ will indeed be independent of $s$.

For (NLS) (for which $\hat{V}_a = \varphi'(0)$), the situation is much worse. Indeed the previous frequency degenerate to

$$\lambda_a(I) = \varphi'(0)\Big(2 \sum_{b \neq a \in \mathbb{Z}} I_b + I_a\Big) = \varphi'(0)\Big(2\|u\|^2_{L^2} - I_a\Big).$$

We thus see that the small denominators associated with $Z_4$ are of the form

$$\omega(I) = \varphi'(0)(I_{a_1} + \cdots + I_{a_m} - I_{b_1} - \cdots I_{b_m}), \tag{4.20}$$

with $a \cap b = \varnothing$ and as $u$ is in $\ell^1_s$, the $I_a$ are of order $\varepsilon^2 \langle a \rangle^{-2s}$. Hence the natural non resonant condition (that is generic in $I$) takes the form

$$|I_{a_1} + \cdots + I_{a_m} - I_{b_1} - \cdots I_{b_m}| \geqslant \gamma \varepsilon^2 \Big( \prod_{n=1}^m \langle a_n \rangle \langle b_n \rangle \Big)^{-2} \langle \mu_{\min} \rangle^{2s}, \tag{4.21}$$

where $\mu_{\min}$ denote the smallest index amongst the $a_n$ and $b_n$. Such a condition was used in [39]. We see that to run through the previous programme, we have to *distribute* the derivative of order $2s$ associated with the lowest index of the irreducible parts of the monomials, coming at each step of the normal form construction.

Unfortunately, such a distribution cannot be done straightforwardly. One of the reason is the presence of the remaining terms $Z(I)$ depending on the actions in the process. Indeed, take a

monomial of the form $f(I)\prod_{n=1}^{m} u_{a_n}\bar{u}_{b_n}$, where $f$ depends only on the actions associated with low modes, the remaining part being irreducible. These terms will enter into the normal form construction first as right-hand side of the homological equation, in which case they will be divided by $\omega(I)$ defined in (4.20), and then will contribute to the higher order terms by Poisson bracket with the other remaining terms. Now take some $Z$ previously constructed (for example $Z_6(I)$). New terms will enter into the next homological right hand side that are made of Poisson brackets between this term $Z$ and the constructed functional. Amongst the new term to solve, we will have terms of the form $f_1(I)\prod_{n=1}^{m}\omega(I)^{-1}u_{a_n}\bar{u}_{b_n}$ where $f_1$ depends again only on low modes. At this stage, it will be possible to distribute the derivative on the irreducible monomials, but by iterating, we see that at each resolution of the homological equation, we will have to divide by the *same* small denominator. After $p$ such iterations, we will end up with monomials of the form $f_p(I)\prod_{n=1}^{m}\omega(I)^{-p}u_{a_n}\bar{u}_{b_n}$ where $f_p(I)$ depends on low modes, and for $p > m$, we will not be able to control this terms independently of $s$. Hence the previous procedure cannot be applied.

To remedy this difficulty, a natural idea (coming from KAM strategy) is to include the term $Z_6$ in the normal form construction, that is to solve at each step the homological equation with $Z_4 + Z_6$. Nevertheless, this trick brings good and bad news :
-The bad news is that the frequencies associated with the Hamiltonian $Z_4 + Z_6$ are not perturbations of the frequencies of $Z_4$. We can even show that for a given generic distribution of the $I_a$, there are $\varepsilon$ producing resonances for the Hamiltonian $Z_4 + Z_6$ while $Z_4$ is non resonant.
-The good news is that the structure of $Z_6$ ressembles the structures of the Hamiltonian of (NLSP) with a similar convolution potential coming from the first resonant normal form done with the Laplace operator. In other words, $Z_6$ is much less resonant than $Z_4$, and has frequencies that satisfy generic non resonance conditions with loss of derivatives independent of $s$.

The strategy of proof is thus to apply the previous programme with $Z_4 + Z_6$ instead of $Z_4$ alone, after having taking care of the genericity condition on the initial data that have to link now the distribution of the $I_a$ and $\varepsilon$. This explain the major difference between the statement for (NLS) and (NLSP). The main drawback is that by doing so, we break the natural homogeneity in $\varepsilon$ which yields some specific technical difficulties, in particular in the definition of a class of rational functions, which must be stable by Poisson bracket and solution of homological equations, while preserving the asymptotic in $\varepsilon$.

## 4.3 General setting

### 4.3.1 Hamiltonian formalism

Let $\mathbb{U}_2 = \{\pm 1\}$. We identify a pair $(\xi, \eta) \in \mathbb{C}^{\mathbb{Z}} \times \mathbb{C}^{\mathbb{Z}}$ with $(z_j)_{j\in\mathbb{U}_2\times\mathbb{Z}} \in \mathbb{C}^{\mathbb{U}_2\times\mathbb{Z}}$ via the formula

$$j = (\delta, a) \in \mathbb{U}_2 \times \mathbb{Z} \Longrightarrow \begin{cases} z_j &= \xi_a \quad \text{if} \quad \delta = 1, \\ z_j &= \eta_a \quad \text{if} \quad \delta = -1. \end{cases} \tag{4.22}$$

We denote by $z = (\xi, \eta)$ such an element and we endow this set of sequences with the $\ell^1_s$ topology :

$$\ell^1_s = \ell^1_s(\mathbb{Z}, \mathbb{C})^2 = \{z \in \mathbb{C}^{\mathbb{U}_2\times\mathbb{Z}} \mid \|z\|_s < \infty\}$$

where [1]

$$\|z\|_s := \sum_{j\in\mathbb{U}_2\times\mathbb{Z}} \langle j\rangle^s |z_j| = \sum_{a\in\mathbb{Z}} \langle a\rangle^s |\xi_a| + \sum_{a\in\mathbb{Z}} \langle a\rangle^s |\eta_a|.$$

We associate with $z$ two complex functions on the torus $u$ and $v$ through the formulas

$$u(x) = \sum_{a\in\mathbb{Z}} \xi_a e^{iax} \quad \text{and} \quad v(x) = \sum_{a\in\mathbb{Z}} \eta_a e^{-iax}, \quad x \in \mathbb{T}. \tag{4.23}$$

We say that $z$ is *real* when $z_{\bar{j}} = \overline{z_j}$ for any $j \in \mathbb{U}_2 \times \mathbb{Z}$. In this case $v$ is the complex conjugate of $u : v(x) = \overline{u(x)},\ x \in \mathbb{T}$, and the definition of $\ell^1_s$ coincides with (4.3).

**Remark 4.3.1.** *The sequences spaces $\ell^1_s$, which are in fact Besov spaces, are not perfectly adapted to Fourier analysis : when $z \in \ell^1_s$ with $s \geqslant 0$ then the functions $u$ and $v$ belong to the Sobolev space $H^s(\mathbb{T})$ while when $u$ and $v$ belong to $H^s(\mathbb{T})$ then its sequence of Fourier coefficients $z$ belongs to $\ell^1_{s-\eta}$ only for $\eta > 1/2$. This lost of regularity would not happen in the Fourier space $\ell^2_s$ nevertheless we prefer $\ell^1_s$ because it leads to simpler estimates of the flows (see for instance Proposition 4.3.3). Anyway the results we obtain thus lead to control of Sobolev norms $H^{s-\frac{1}{2}^+}$ over long times.*

We endow $\ell^1_s$ with the symplectic structure

$$-i \sum_{a\in\mathbb{Z}} dz_{(+1,a)} \wedge dz_{(-1,a)} = -i \sum_{a\in\mathbb{Z}} d\xi_a \wedge d\eta_a. \tag{4.24}$$

For a function $F$ of $\mathcal{C}^1(\ell^1_s, \mathbb{C})$, we define its Hamiltonian vector field by $X_F = J\nabla F$ where $J$ is the symplectic operator induced by the symplectic form (4.24), $\nabla F(z) = \left(\frac{\partial F}{\partial z_j}\right)_{j\in\mathbb{U}\times\mathbb{Z}}$, and by definition we set for $j = (\delta, a) \in \mathbb{U}_2 \times \mathbb{Z}$,

$$\frac{\partial F}{\partial z_j} = \begin{cases} \dfrac{\partial F}{\partial \xi_a} & \text{if} \quad \delta = 1, \\ \dfrac{\partial F}{\partial \eta_a} & \text{if} \quad \delta = -1. \end{cases}$$

So $X_F = J\nabla F$ reads in coordinates

$$(X_F)_j = \begin{cases} i\dfrac{\partial F}{\partial \eta_a} & \text{if} \quad \delta = 1, \\ -i\dfrac{\partial F}{\partial \xi_a} & \text{if} \quad \delta = -1. \end{cases}$$

For two functions $F$ and $G$, the Poisson Bracket is (formally) defined as

$$\{F, G\} = \langle \nabla F, J\nabla G\rangle = i \sum_{a\in\mathbb{Z}} \frac{\partial F}{\partial \xi_a}\frac{\partial G}{\partial \eta_a} - \frac{\partial F}{\partial \eta_a}\frac{\partial G}{\partial \xi_a} \tag{4.25}$$

where $\langle\cdot,\cdot\rangle$ denotes the natural bilinear pairing : $\langle z, \zeta\rangle = \sum_{j\in\mathbb{U}_2\times\mathbb{Z}} z_j\zeta_j$.
We say that a Hamiltonian function $H$ is *real* if $H(z)$ is real for all real $z$.

1. Here for $j = (\delta, a)$ we set $\langle j\rangle = (1 + a^2)^{1/2} = \langle a\rangle$.

**Definition 4.3.2.** *For a given $s \geqslant 0$ and a given open set $\mathcal{U}$ in $\ell^1_s$, we denote by $\mathcal{H}_s(\mathcal{U})$ the space of real Hamiltonians $P$ satisfying*

$$P \in \mathcal{C}^1(\mathcal{U}, \mathbb{C}), \quad \textit{and} \quad X_P \in \mathcal{C}^1(\mathcal{U}, \ell^1_s).$$

We will use the shortcut $F \in \mathcal{H}_s$ to indicate that there exists an open set $\mathcal{U}$ in $\ell^1_s$ such that $F \in \mathcal{H}_s(\mathcal{U})$.
Notice that for $F$ and $G$ in $\mathcal{H}_s(\mathcal{U})$ the formula (4.25) is well defined.

### 4.3.2 Hamiltonian flows

With a given Hamiltonian function $H \in \mathcal{H}_s$, we associate the Hamiltonian system

$$\dot{z} = X_H(z) = J\nabla H(z),$$

which also reads in coordinates

$$\dot{\xi}_a = i\frac{\partial H}{\partial \eta_a} \quad \text{and} \quad \dot{\eta}_a = -i\frac{\partial H}{\partial \xi_a}, \quad a \in \mathbb{Z}. \tag{4.26}$$

Concerning the Hamiltonian flows we have

**Proposition 4.3.1.** *Let $s \geqslant 0$. Any Hamiltonian in $\mathcal{H}_s$ defines a local flow in $\ell^1_s$ which preserves the reality condition, i.e. if the initial condition $z = (\xi, \bar{\xi})$ is real, the flow $(\xi(t), \eta(t)) = \Phi^t_H(z)$ is also real, $\xi(t) = \overline{\eta(t)}$ for all $t$.*

*Proof.* The existence of the local flow is a consequence of the Cauchy-Lipschitz theorem.
Furthermore let us denote by $f$ the $\mathcal{C}^1$ function defined by $\ell^1_s(\mathbb{Z}, \mathbb{C}) \ni \xi \mapsto H(\xi, \bar{\xi}) - \overline{H(\xi, \bar{\xi})}$. Since $H$ is real we have $f \equiv 0$ and thus its differential at any point $\xi \in \ell^1_s(\mathbb{Z}, \mathbb{C})$ and in any direction $\zeta \in \ell^1_s(\mathbb{Z}, \mathbb{C})$ vanishes[2] :

$$\begin{aligned} 0 \equiv Df(\xi) \cdot \zeta &= \langle \nabla_\xi H(\xi, \bar{\xi}), \zeta \rangle + \langle \nabla_\eta H(\xi, \bar{\xi}), \bar{\zeta} \rangle \\ &- \overline{\langle \nabla_\xi H(\xi, \bar{\xi}), \zeta \rangle} - \overline{\langle \nabla_\eta H(\xi, \bar{\xi}), \bar{\zeta} \rangle} \\ &= \langle \nabla_\xi H(\xi, \bar{\xi}) - \overline{\nabla_\eta H(\xi, \bar{\xi})}, \zeta \rangle + \langle \nabla_\eta H(\xi, \bar{\xi}) - \overline{\nabla_\eta H(\xi, \bar{\xi})}, \bar{\zeta} \rangle. \end{aligned}$$

Therefore $\nabla_\xi H(\xi, \bar{\xi}) - \overline{\nabla_\eta H(\xi, \bar{\xi})}$ for all $\xi \in \ell^1_s(\mathbb{Z}, \mathbb{C})$ and the system (4.26) preserves the reality condition. □

In this setting Equations (NLS) and (NLSP) are equivalent to Hamiltonian systems associated with the real Hamiltonian function

$$H(\xi, \eta) = \sum_{a \in \mathbb{Z}} a^2 \xi_a \eta_a + P(z) \tag{4.27}$$

where

$$P(\xi, \eta) = \frac{1}{2\pi} \int_{\mathbb{T}} g\Big( \sum_{a,b \in \mathbb{Z}} \xi_a \eta_b e^{i(a-b)x} \Big) \mathrm{d}x, \tag{4.28}$$

2. Here by a slight abuse of notation $\langle \cdot, \cdot \rangle$ denotes the bilinear pairing in $\ell^1_s(\mathbb{Z}, \mathbb{C})$.

in the (NLS) case, and

$$P(\xi,\eta)=\frac{1}{4\pi}\int_{\mathbb{T}}\Big(\sum_{a,b\in\mathbb{Z}}\hat{\xi}_a\eta_b e^{i(a-b)x}\Big)\Big(\sum_{a,b\in\mathbb{Z}}\hat{V}_{a-b}\xi_a\eta_b e^{i(a-b)x}\Big)\mathrm{d}x, \tag{4.29}$$

in the (NLSP) case, where we recall that $\hat{V}_a=a^{-2}$ for $a\neq 0$ and $\hat{V}_0=0$. We first notice that in both cases, $P$ belongs to $\mathcal{H}_s$, in fact we have :

**Lemma 4.3.3.** *Let $s\geqslant 0$ and let $z\mapsto f(z)\in\mathbb{C}$ and $z\mapsto h(z)$ be two analytic functions defined on a neighborhood $\mathcal{V}$ of the origin in $\mathbb{C}$ that takes real values when $z$ is real. Then the formulas*

$$P(\xi,\eta)=\frac{1}{2\pi}\int_{\mathbb{T}}f\Big(\sum_{a,b\in\mathbb{Z}}\xi_a\eta_b e^{i(a-b)x}\Big)\mathrm{d}x, \tag{4.30}$$

$$Q(\xi,\eta)=\int_{\mathbb{T}}h\Big(\sum_{a,b\in\mathbb{Z}}\hat{V}_{a-b}\xi_a\eta_b e^{i(a-b)x}\Big)f\Big(\sum_{a,b\in\mathbb{Z}}\xi_a\eta_b e^{i(a-b)x}\Big)\mathrm{d}x \tag{4.31}$$

*define Hamiltonian $P$ and $Q$ belonging to $\mathcal{H}_s(\mathcal{U})$ where $\mathcal{U}$ is some neighborhood of the origin in $\ell^1_s$.*

*Proof.* First we verify that (4.30) and (4.31) define regular maps on $\ell^1_s$.
By definition, $\ell^1_s=\ell^1_s(\mathbb{Z},\mathbb{C})^2$ and the Fourier transform $\xi\mapsto\sum_{a\in\mathbb{Z}}\xi_a e^{iax}$ defines an isomorphism between $\ell^1_s(\mathbb{Z},\mathbb{C})$ and a subset of $L^2(\mathbb{R},\mathbb{C})$ that we still denote by $\ell^1_s(\mathbb{Z},\mathbb{C})$. Moreover, for $u,v\in\ell^1_s(\mathbb{Z},\mathbb{C})$, we have $\|uv\|_s\leqslant\|u\|_s\|v\|_s$ and thus the mapping $(u,v)\mapsto uv$ is analytic from $\ell^1_s(\mathbb{Z},\mathbb{C})^2$ to $\ell^1_s(\mathbb{Z},\mathbb{C})$. Extending this argument, if $h:\mathbb{C}\to\mathbb{C}$ is analytic in a neighborhood of the origin, the application $\xi\mapsto u(x)\mapsto h(u(x))$ is analytic from a neighborhood of the origin in $\ell^1_s(\mathbb{Z},\mathbb{C})$ into $\ell^1_s(\mathbb{Z},\mathbb{C})$.
Through the identification $(\xi,\eta)\mapsto(u(x),v(x))$, see (4.23) the Hamiltonian $Q$ reads

$$Q(\xi,\eta)=\int_{\mathbb{T}}h\Big(V\star(u(x)v(x))\Big)f\Big(u(x)v(x)\Big)\mathrm{d}x$$

Since the mapping $u\mapsto V\star u$ is analytic on $\ell_s(\mathbb{Z},\mathbb{C})$, we conclude that $Q$ is an analytic function from a neighborhood $\mathcal{U}$ of the origin in $\ell^1_s(\mathbb{Z},\mathbb{C})$ into $\mathbb{C}$. Similar arguments apply to $P$.
Next we verify that $X_P$ and $X_Q$ are still regular[3] function from $\mathcal{U}$ into $\ell^1_s(\mathbb{Z},\mathbb{C})$. We focus on $P$ but similar arguments apply to $Q$. We have

$$\frac{\partial Q}{\partial\xi_a}=\int_{\mathbb{T}}k(x)e^{iax}dx$$

with

$$k(x)=f'\Big(\sum_{a,b\in\mathbb{Z}}\xi_a\eta_b e^{i(a-b)x}\Big)\sum_{b\in\mathbb{Z}}\eta_b e^{-ibx}=f'(u(x)v(x))v(v).$$

Expanding $f'$ in entire series we rewrite $\frac{\partial Q}{\partial\xi_a}$ in a convergent sum of terms of the form

$$c_k\int_{\mathbb{T}}(u(x))^{k_1}(v(x))^{k_2}e^{iax}dx,$$

3. The analyticity of $P$ only insure that $X_P$ belongs to the dual of $\ell^1_s(\mathbb{Z},\mathbb{C})$.

i.e. the convolution product of $k = k_1 + k_2$ sequences in $\ell^1_s$. Then the conclusion follows from the fact that for any $s \geqslant 0$

$$\ell^1_s \star \ell^1_s \subset \ell^1_s.$$

□

On the contrary the quadratic part of $H$, $\sum_{a\in\mathbb{Z}} a^2\xi_a\eta_a$, corresponding to the linear part of (NLS) does not belong to $\mathcal{H}_s$. Nevertheless it generates a continuous flow which maps $\ell^1_s$ into $\ell^1_s$ explicitly given for all time $t$ and for all indices $a$ by $\xi_a(t) = e^{-ia^2t}\xi_a(0)$, $\eta_a(t) = e^{ia^2t}\eta_a(0)$. Furthermore this flow has the group property. By standard arguments (see for instance [45]), this is enough to define the local flow of $\dot{z} = X_H(z)$ in $\ell^1_s$ :

**Proposition 4.3.2.** *Let $s \geqslant 0$. Let $H$ be the NLS Hamiltonian defined by* (4.27) *and $z_0 \in \ell^1_s$ a sufficiently small initial datum. Then the Hamilton equation*

$$\dot{z}(t) = X_H(z(t)), \quad z(0) = z_0$$

*admits a local solution $t \mapsto z(t) \in \ell^1_s$ which is real if $z_0$ is real.*

The reality of the flow is proved as in the proof of Proposition 4.3.1.

### 4.3.3 Polynomial Hamiltonians

For $m \geqslant 1$ we define three nested subsets of $(\mathbb{U}_2\times\mathbb{Z})^m$ satisfying zero momentum conditions of increasing order :

$$\begin{aligned}
\mathcal{Z}_m &= \{\boldsymbol{j} = (\delta_\alpha, a_\alpha)_{\alpha=1}^m \mid \sum_{\alpha=1}^{2m} \delta_\alpha = 0\},\\
\mathcal{M}_m &= \{\boldsymbol{j} \in \mathcal{Z}_m \mid \sum_{\alpha=1}^{2m} \delta_\alpha a_\alpha = 0\},\\
\mathcal{R}_m &= \{\boldsymbol{j} \in \mathcal{M}_m \mid \sum_{\alpha=1}^{2m} \delta_\alpha a_\alpha^2 = 0\}.
\end{aligned} \tag{4.32}$$

We set $\mathcal{Z} = \bigsqcup_{m\geqslant 0} \mathcal{Z}_m$, $\mathcal{M} = \bigsqcup_{m\geqslant 0} \mathcal{M}_m$ and $\mathcal{R} = \bigsqcup_{m\geqslant 0} \mathcal{R}_m$.
For $\boldsymbol{j} \in \mathcal{Z}$, $\mathrm{Irr}(\boldsymbol{j})$ denotes the *irreductible* part of $\boldsymbol{j}$, i.e. a subsequence of maximal length $(j'_1, \dots, j'_{2p})$ containing no actions in the sense that if $j'_\alpha \neq \bar{j}'_\beta$ for all $\alpha, \beta = 1, \cdots, 2p$.
We set

$$\mathcal{I}rr = \mathrm{Irr}(\mathcal{R}) = \{\mathrm{Irr}(\boldsymbol{j}) \,|\, \boldsymbol{j} \in \mathcal{R}\}.$$

We will use indices **k** belonging to $\bigsqcup_{p\in\mathbb{N}} \mathcal{I}rr^p$, i.e. $\mathbf{k} = (\boldsymbol{k}_1, \dots, \boldsymbol{k}_p)$ for some $p \in \mathbb{N}$ with $\boldsymbol{k}_\alpha \in \mathcal{I}rr$.
We denote $\sharp\mathbf{k} = p$. We use the convention $\mathbf{k} = \varnothing$ for $p = 0$.
For $\boldsymbol{j} \in \mathcal{Z}_m$, we set $z_{\boldsymbol{j}} = z_{j_1} \cdots z_{j_{2m}}$. We also denote by $\bar{\boldsymbol{j}} = (\bar{j}_1, \dots, \bar{j}_{2m})$, and we notice that when $z$ is real, we have $\overline{z_{\boldsymbol{j}}} = z_{\bar{\boldsymbol{j}}}$.

**Definition 4.3.4.** *We say that $P(z)$ is a homogeneous polynomial of order $m$ if it can be written*

$$P(z) = P[c](z) = \sum_{\boldsymbol{j}\in\mathcal{M}_m} c_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad \textit{with} \quad c = (c_{\boldsymbol{j}})_{\boldsymbol{j}\in\mathcal{M}_m} \in \ell^\infty(\mathcal{M}_m), \tag{4.33}$$

*and such that the coefficients $c_{\boldsymbol{j}}$ satisfy $c_{\overline{\boldsymbol{j}}} = \overline{c_{\boldsymbol{j}}}$.*

Note that the last condition ensures that $P$ is real, as the set of indices are invariant by the application $\boldsymbol{j} \mapsto \overline{\boldsymbol{j}}$. Following [64] we easily get

**Proposition 4.3.3.** *Let $s \geqslant 0$. A homogeneous polynomial, $P[c]$, of degree $m \geqslant 2$ belongs to $\mathcal{H}_s(\ell^1_s)$ and we have*

$$\|X_{P[c]}(z)\|_s \leqslant 2m\|c\|_{\ell^\infty} \|z\|_s^{2m-1}. \tag{4.34}$$

*Furthermore for two homogeneous polynomials, $P[c]$ and $P[c']$, of degree respectively $m$ and $n$, the Poisson bracket is a homogeneous polynomial of degree $m+n-1$, $\{P[c], P[c']\} = P[c'']$ and we have the estimate*

$$\|c''\|_{\ell^\infty} \leqslant 2mn\|c\|_{\ell^\infty} \|c'\|_{\ell^\infty}. \tag{4.35}$$

For $j = (\delta, a) \in \mathbb{U}_2 \times \mathbb{Z}$, we set $I_j = I_a = \xi_a\eta_a = z_j z_{\overline{j}}$ the action of index $a$. $I = (I_a)_{a\in\mathbb{Z}}$ denote the set of all the actions.
We note that for a real $z$ in $\ell^1_s$ we have $\|z\|_s = \sum_a \langle a\rangle^s I_a^{1/2}$. Therefore, an integrable Hamiltonian, i.e. a Hamiltonian function depending only on the actions has a flow which leaves invariant each $\|\cdot\|_s$ norm.
We introduce $3$ integrable polynomials that will be used later (see Theorem 4.7.1) :

$$Z_2(I) = \sum_{a\in\mathbb{Z}} (a^2 + \varphi(0)) I_a,$$

$$Z_4(I) = \varphi'(0)\left(\sum_{a\in\mathbb{Z}} I_a\right)^2 - \frac{1}{2}\varphi'(0)\sum_{a\in\mathbb{Z}} I_a^2, \tag{4.36}$$

$$Z_6(I) = -\frac{1}{2}\varphi'(0)^2 \sum_{a\neq b\in\mathbb{Z}} \frac{1}{(a-b)^2} I_a^2 I_b \tag{4.37}$$

$$+ \frac{\varphi''(0)}{6}\left(6(\sum_{a\in\mathbb{Z}} I_a)^3 - 9(\sum_{a\in\mathbb{Z}} I_a^2)(\sum_{a\in\mathbb{Z}} I_a) + 4\sum_{a\in\mathbb{Z}} I_a^3\right).$$

The first one is the quadratic part of the NLS Hamiltonian, the second one is the quartic part and the third one contains the effective terms of the sextic part (see Theorem 4.7.1).

We note that $Z_4$ and $Z_6$ are polynomials of degree $2$ and $3$ in the sense of Definition 4.3.4, and thus define Hamiltonians in $\mathcal{H}_s$ for all $s \geqslant 0$.

## 4.4 Non-resonance conditions

In this section we discuss the control of small denominators corresponding to the the previous integrable Hamiltonians. We also give results allowing to control them, and show the probability estimates associated with non resonant sets.

### 4.4.1 Small denominators

For $\boldsymbol{j} = (j_1, \dots, j_{2m}) \in \mathbb{U}_2 \times \mathbb{Z}$, if $j_\alpha = (\delta_\alpha, a_\alpha)$ for $\alpha = 1, \dots, 2m$, we set

$$\Delta_{\boldsymbol{j}} = \sum_{\alpha=1}^{2m} \delta_\alpha a_\alpha^2, \quad \omega_{\boldsymbol{j}}(I) = \sum_{\alpha=1}^{2m} \delta_\alpha \frac{\partial Z_4}{\partial I_{a_\alpha}}(I)$$

and

$$\Omega_{\boldsymbol{j}}(I) = \omega_{\boldsymbol{j}}(I) + \sum_{\alpha=1}^{2m} \delta_\alpha \frac{\partial Z_6}{\partial I_{a_\alpha}}(I).$$

With these notations, we have for $\boldsymbol{j} \in \mathcal{Z}$, owing to the fact that $\omega_{\boldsymbol{j}} = \omega_{\mathrm{Irr}(\boldsymbol{j})}$ and $\Omega_{\boldsymbol{j}} = \Omega_{\mathrm{Irr}(\boldsymbol{j})}$,

$$\{Z_4, z_{\boldsymbol{j}}\} = i\omega_{\mathrm{Irr}(\boldsymbol{j})}(I) z_{\boldsymbol{j}} \quad \text{and} \quad \{Z_4 + Z_6, z_{\boldsymbol{j}}\} = i\Omega_{\mathrm{Irr}(\boldsymbol{j})}(I) z_{\boldsymbol{j}}. \tag{4.38}$$

Note also that

$$\{Z_2, z_{\boldsymbol{j}}\} = i\Delta_{\boldsymbol{j}} z_{\boldsymbol{j}}, \tag{4.39}$$

and that $|\Delta_{\boldsymbol{j}}| \geqslant 1$ except when $\boldsymbol{j} \in \mathcal{R}$ for which $\Delta_{\boldsymbol{j}} = 0$.

For $\boldsymbol{j} \in \mathcal{I}rr$, we have the expression

$$\omega_{\boldsymbol{j}}(I) = -\varphi'(0) \sum_{\alpha=1}^{2m} \delta_\alpha I_{a_\alpha}, \tag{4.40}$$

as the first term in (4.36) do not contribute using the relation $\sum_{\alpha=1}^{2m} \delta_\alpha = 0$.

We also introduce the following denominator :

$$\widetilde{\Omega}_{\boldsymbol{j}}(I) = -\varphi'(0) \sum_{\alpha=1}^{2m} \delta_\alpha I_{a_\alpha} - \frac{1}{2}\varphi'(0)^2 \sum_{\alpha=1}^{2m} \delta_\alpha \sum_{\substack{b \in \mathbb{Z} \\ b \neq a_\beta, \beta = 1, \dots, 2m}} \frac{I_b^2}{(a_\alpha - b)^2}.$$

The following lemma allows to control the evolution of the small denominators when moving the coordinates.

**Lemma 4.4.1.** *Let $r$ and $s$ be given. The exists a constant $C$ such that for all $\boldsymbol{j} \in \mathcal{I}rr$ with $\sharp \boldsymbol{j} \leqslant 2r$ and all $z \in \ell^1_s$, we have*

$$|\widetilde{\Omega}_{\boldsymbol{j}}(I) - \Omega_{\boldsymbol{j}}(I)| \leqslant C \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} \|z\|_s^4. \tag{4.41}$$

*Moreover, let $z, z' \in \ell^1_s$ be associated with the actions $I$ and $I'$, and let $h = \max(\|z\|_s, \|z'\|_s)$. Then we have*

$$|\omega_{\boldsymbol{j}}(I) - \omega_{\boldsymbol{j}}(I')| \leqslant C \|z - z'\|_s \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} h. \tag{4.42}$$

*and*

$$|\Omega_{\boldsymbol{j}}(I) - \Omega_{\boldsymbol{j}}(I')| \leqslant C \|z - z'\|_s \left( \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} h + h^3 \right) \tag{4.43}$$

*Proof.* Along the proof, $C$ will denote a constant depending on $r$, $s$ and derivatives of the function $\varphi$ at $0$. Let us denote $\boldsymbol{j} = (j_1, \dots, j_{2m}) \in \mathcal{Z}_m$ with $m \leqslant r$ and $j_\alpha = (\delta_\alpha, a_\alpha) \in \mathbb{U}_2 \times \mathbb{Z}$ for $\alpha = 1, \dots, 2m$. We calculate that

$$\Omega_{\boldsymbol{j}}(I) - \widetilde{\Omega}_{\boldsymbol{j}}(I) = -\frac{1}{2}\varphi'(0)^2 \sum_{\substack{\alpha,\beta=1\\ \beta\neq\alpha}}^{2m} \delta_\alpha \frac{I_{a_\beta}^2}{(a_\alpha - a_\beta)^2} - \varphi'(0)^2 \sum_{\alpha=1}^{2m} \sum_{b\neq a_\alpha} \delta_\alpha \frac{I_{a_\alpha} I_b}{(a_\alpha - b)^2}$$

$$+ \varphi''(0) \sum_{\alpha=1}^{2m} \delta_{a_\alpha} \left( -3 I_{a_\alpha} (\sum_{a\in\mathbb{Z}} I_a) + 2 I_{a_\alpha}^2 \right).$$

We see that the first term can be controlled by

$$C \sum_{\alpha=1}^{2m} I_{a_\alpha}^2 \leqslant C \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-4s} \sup_{\alpha=1,\dots,2m} \langle j_\alpha \rangle^{4s} I_{a_\alpha}^2 \leqslant C \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-4s} \|z\|_s^4,$$

and the second by

$$C \sum_{\alpha=1}^{2m} I_{a_\alpha} \sum_b I_b \leqslant C \|z\|_{L^2}^2 \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} \sup_\alpha \langle j_\alpha \rangle^{2s} I_{a_\alpha} \leqslant C \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} \|z\|_s^4.$$

The estimate for the third term is the same, which shows (4.41).

Now to prove (4.42), we have using the expression (4.40),

$$|\omega_{\boldsymbol{j}}(I) - \omega_{\boldsymbol{j}}(I')| \leqslant C \sum_{\alpha=1}^{2m} |I_{a_\alpha} - I'_{a_\alpha}| \leqslant C \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} \max_{\alpha=1}^{2m} \left( \langle j \rangle^{2s} |I_{a_\alpha} - I'_{a_\alpha}| \right).$$

We obtain (4.42) by noticing that

$$\langle j \rangle^{2s} |I_{a_\alpha} - I'_{a_\alpha}| = \langle j \rangle^s \left| \sqrt{I_{a_\alpha}} - \sqrt{I'_{a_\alpha}} \right| \langle j \rangle^s \left| \sqrt{I_{a_\alpha}} + \sqrt{I'_{a_\alpha}} \right|$$
$$\leqslant \| |z| - |z'| \|_s (\|z\|_s + \|z'\|_s) \leqslant 2 \|z - z'\|_s \max(\|z\|_s, \|z'\|_s).$$

The proof of (4.43) is then easily obtained by using the previous result, and explicit expressions of $\frac{\partial Z_6}{\partial I_{a_\alpha}}$ showing that this function is homogeneous of order $4$ and thus locally Lipschitz in $z$ with Lipschitz constant of order $h^3$ on balls of size $h$ in $\ell_s^1$ as can be seen by using estimates similar to the previous one. □

### 4.4.2 Non resonant sets

As usual we have to control the small divisors, this will be the case for $z$ belonging the following non resonant sets :

**Definition 4.4.2.** *Let $\varepsilon, \gamma > 0$, $r \geqslant 1$ and $s \geqslant 0$, we say that $z \in \ell_s^1$ belongs to the non resonant set $\mathcal{U}_{\gamma,\varepsilon,r,s}$, if for all $\boldsymbol{k} \in \mathcal{I}rr$ of length $\sharp\boldsymbol{k} \leqslant 2r$ we have*

$$|\omega_{\boldsymbol{k}}(I)| > \gamma \varepsilon^2 \left( \prod_{\alpha=1}^{\sharp\boldsymbol{k}} \langle k_\alpha \rangle^{-2} \right) \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s} \tag{4.44}$$

*and*

$$|\widetilde{\Omega}_{\boldsymbol{k}}(I)| > \gamma \left( \prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha \rangle^{-6} \right) \max(\varepsilon^2 \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s}, \varepsilon^4). \tag{4.45}$$

We also define the truncated non resonant set :

**Definition 4.4.3.** *Let $\varepsilon, \gamma > 0$, $N \geqslant 1$ $r \geqslant 1$ and $s \geqslant 0$, and let $\alpha_r = 24r$. We say that $z \in \ell^1_s$ belongs to the non resonant set $\mathcal{U}^N_{\gamma,\varepsilon,r,s}$, if for all $\boldsymbol{k} \in \mathcal{I}rr$ of length $\sharp \boldsymbol{k} \leqslant 7r$ such that $\langle \mu_1(\boldsymbol{k}) \rangle \leqslant N^2$, we have*

$$|\omega_{\boldsymbol{k}}(I)| > \gamma \varepsilon^2 N^{-\alpha_r} \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s} \tag{4.46}$$

*and*

$$|\Omega_{\boldsymbol{k}}(I)| > \gamma N^{-\alpha_r} \max(\varepsilon^2 \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s}, \varepsilon^4). \tag{4.47}$$

It turns out that for $N$ not too large depending on $\varepsilon$ and for $\gamma' < \gamma$ we have $\mathcal{U}_{\gamma,\varepsilon,r,s} \subset \mathcal{U}^N_{\varepsilon,\gamma',r,s}$. Precisely we have :

**Proposition 4.4.1.** *Let $r \geqslant 1$ and $s \geqslant 1$ be given. There exists $c$ such that for all $\varepsilon, \gamma > 0$, for all $N \geqslant 1$ and all $\gamma' < \gamma$ satisfying*

$$\varepsilon^2 < c N^{-\alpha_r} (\gamma - \gamma') \quad \textit{with } \alpha_r = 24r,$$

*we have that if $z \in \mathcal{U}_{\gamma,\varepsilon,r,s}$ and $\|z\|_s \leqslant 4\varepsilon$ then $z \in \mathcal{U}^N_{\varepsilon,\gamma',r,s}$.*

*Proof.* The hypothesis $z \in \mathcal{U}_{\gamma,\varepsilon,r,s}$ and $\langle \mu_1(\boldsymbol{k}) \rangle \leqslant N^2$ shows that if $z$ satisfies (4.44), we have

$$\begin{aligned} |\omega_{\boldsymbol{k}}(I)| &> \gamma \varepsilon^2 \left( \prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha \rangle^{-2} \right) \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s} \\ &\geqslant \gamma \varepsilon^2 N^{-4\sharp \boldsymbol{k}} \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s} \end{aligned}$$

which shows (4.46) for all $\gamma' \leqslant \gamma$, as $\sharp \boldsymbol{k} \leqslant 2r$ and hence $4\sharp \boldsymbol{k} \leqslant 24r = \alpha_r$. Similarly, we have

$$|\widetilde{\Omega}_{\boldsymbol{k}}(I)| > \gamma N^{-12\sharp \boldsymbol{k}} \max(\varepsilon^2 \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s}, \varepsilon^4),$$

which shows that $\widetilde{\Omega}_{\boldsymbol{k}}(I)$ satisfies (4.47) with $\alpha_r = 24r$.

To prove (4.47), we use the fact that using (4.41) we have

$$|\widetilde{\Omega}_{\boldsymbol{j}}(I) - \Omega_{\boldsymbol{j}}(I)| \leqslant C \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s} \varepsilon^4.$$

This shows that

$$|\Omega_{\boldsymbol{j}}(I)| \geqslant (\gamma N^{-\alpha_r} - C\varepsilon^2) \max(\varepsilon^2 \langle \mu_{\min}(\boldsymbol{j}) \rangle^{-2s}, \varepsilon^4),$$

and we deduce the result by choosing $c = 1/C$. □

We conclude this section with two stability results of the truncated resonant sets. The first one use the fact that the non resonance conditions depend only upon $I$ :

**Proposition 4.4.2.** *Let $r \geqslant 1$ and $s \geqslant 1$ be given. There exists $c$ such that the following holds : for $\varepsilon, \gamma > 0$ and $N \geqslant 1$, let $z \in \mathcal{U}^N_{\gamma,\varepsilon,r,s}$ such that $\|z\|_s \leqslant 4\varepsilon$, then for all $\gamma' > \gamma$ and for all $z' \in \ell^1_s$ such that*

$$\sup_{a \in \mathbb{Z}} |I'_a - I_a| \langle a \rangle^{2s} \leqslant c\varepsilon^2 N^{-\alpha_r} (\gamma' - \gamma) \tag{4.48}$$

*we have $z' \in \mathcal{U}^N_{\varepsilon,\gamma',r,s}$.*

*Proof.* We introduce the Banach space $\ell^\infty_s = \{(x_n)_{n\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}} \mid \sup\langle n\rangle^{2s}|x_n| < \infty\}$ that we endow with the norm $|x|_s := \sup\langle n\rangle^{2s}|x_n|$.
Let $\boldsymbol{j} \in \mathcal{I}rr$, we have using (4.40),

$$|\omega_{\boldsymbol{j}}(I) - \omega_{\boldsymbol{j}}(I')| \leqslant C\sum_{\alpha=1}^{2m} |I_{a_\alpha} - I'_{a_\alpha}| \leqslant C\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}|I - I'|_s.$$

Thus since $z \in \mathcal{U}^N_{\gamma,\varepsilon,r,s}$ we get using (4.48)

$$|\omega_{\boldsymbol{j}}(I')| \geqslant \big(\gamma + Cc(\gamma' - \gamma)\big)\varepsilon^2 N^{-\alpha_r}\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s} \geqslant \gamma'\varepsilon^2 N^{-\alpha_r}\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}$$

by choosing $c \leqslant \frac{1}{C}$.
On the other hand using that $I \mapsto \Omega_{\boldsymbol{j}}(I)$ is a homogeneous polynomial of order 3 on $\ell^\infty_s$. For $s > 1/2$ such polynomial (with bounded coefficients) is a $\mathcal{C}^\infty$ function and for $I, I'$ in a ball of $\ell^\infty_s$ of size $O(\varepsilon)$ centered at $0$ we have

$$|\Omega_{\boldsymbol{j}}(I) - \Omega_{\boldsymbol{j}}(I')| \leqslant C\varepsilon^2|I - I'|_s.$$

So we get

$$|\Omega_{\boldsymbol{j}}(I) - \Omega_{\boldsymbol{j}}(I')| \leqslant C|I - I'|_s \max(\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}, \varepsilon^2)$$

Thus since $z \in \mathcal{U}^N_{\gamma,\varepsilon,r,s}$ we get using (4.48)

$$\begin{aligned}|\Omega_{\boldsymbol{j}}(I')| &\geqslant \big(\gamma + Cc(\gamma' - \gamma)\big)\varepsilon^2 N^{-\alpha_r}\max(\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}, \varepsilon^2)\\ &\geqslant \gamma'\varepsilon^2 N^{-\alpha_r}\max(\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}, \varepsilon^2)\end{aligned}$$

by choosing $c \leqslant \frac{1}{C}$. □

The second stability result shows that the truncated resonant sets are stable by perturbation in $\ell^1_s$ up to change of constants.

**Proposition 4.4.3.** *Let $r \geqslant 1$ and $s \geqslant 1$ be given. There exists $c$ such that the following holds : for $\varepsilon, \gamma > 0$ and $N \geqslant 1$, let $z \in \mathcal{U}^N_{\gamma,\varepsilon,r,s}$ such that $\|z\|_s \leqslant 4\varepsilon$, then for all $\gamma' < \gamma$ and for all $z' \in \ell^1_s$ such that*

$$\|z - z'\|_s \leqslant c\varepsilon N^{-\alpha_r}(\gamma - \gamma'),$$

*we have $z' \in \mathcal{U}^N_{\varepsilon,\gamma',r,s}$.*

*Proof.* By using (4.42) with $h = 4\varepsilon$,

$$|\omega_{\boldsymbol{j}}(I) - \omega_{\boldsymbol{j}}(I')| \leqslant C\|z - z'\|_s\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}\varepsilon$$

We deduce that

$$|\omega_{\boldsymbol{j}}(I)| \leqslant |\omega_{\boldsymbol{j}}(I')| + C\|z - z'\|_s\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}\varepsilon,$$

and hence

$$|\omega_{\boldsymbol{j}}(I')| \geqslant \big(\gamma N^{-\alpha_r} - C\varepsilon^{-1}\|z - z'\|_s\big)\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}\varepsilon^2,$$

and we deduce the first part of the result by taking $c \leqslant 1/C$. Now using (4.43), we have

$$\begin{aligned}|\Omega_{\boldsymbol{j}}(I) - \Omega_{\boldsymbol{j}}(I')| &\leqslant C\|z - z'\|_s\Big(\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}\varepsilon + \varepsilon^3\Big)\\ &\leqslant 2C\varepsilon^{-1}\|z - z'\|_s\max(\varepsilon^2\langle\mu_{\min}(\boldsymbol{j})\rangle^{-2s}, \varepsilon^4)\end{aligned}$$

and we conclude as before by taking $c \geqslant 1/(2C)$. □

## 4.5 Probability estimates

In this section, we prove two genericity results for the (NLS) non-resonant sets (and give also some Lemmas that will be used in the (NLSP) case). We consider real $z$, we consider the actions $I_a \in \mathbb{R}_+$ as random variables. On the one hand, if $\varepsilon > 0$ we prove that typically $\varepsilon z$ is non-resonant. On the other hand, typically, up to some exceptional values of $\varepsilon$, we show that $\varepsilon z$ is non-resonant.

In this section $r > 0$ and $s > 1$ are fixed numbers, $\lambda = \frac{\mathbb{1}_{\varepsilon>0}}{\varepsilon}\mathrm{d}\varepsilon$ is the Haar measure of $\mathbb{R}_+^*$ and we consider $z = (\xi, \bar{\xi})$ as a function of the random variables $I = (I_a)_{a\in\mathbb{Z}}$, such that

- the actions $I_a$, $a \in \mathbb{Z}$, are independent variables,
- $I_a^2$ is uniformly distributed in $(0, \langle a\rangle^{-4s-8})$.

We note that this last assumption implies that $z \in \ell^1_s$.

The first proposition describes the case where $\varepsilon > 0$ is fixed.

**Proposition 4.5.1.** *There exists a constant $c > 0$ such that for all $\gamma \in (0,1)$ we have*

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\varepsilon z \in \mathcal{U}_{\gamma,\varepsilon,r,s}\right) \geqslant 1 - c\gamma.$$

The second proposition describes the case where $z$ is chosen randomly and the asymptotic of $\varepsilon z$ is considered as $\varepsilon$ goes to $0$.

**Proposition 4.5.2.** *There exists a constant $c > 0$ and $\nu_0 > 0$ such that for all $\gamma \in (0,1)$ and all $\nu \in (0,\nu_0)$ we have*

$$\mathbb{P}\left(\lambda\left(\varepsilon z \in \ell^1_s \backslash \mathcal{U}_{\gamma,\varepsilon,r,s}\right) < \nu\right) \geqslant 1 - \frac{c}{\nu}\gamma.$$

**Corollary 4.5.1.** *There exists a constant $c > 0$ and $\nu_0 > 0$ such that for all $\gamma \in (0,1)$, $\nu \in (0,\nu_0)$, all $\epsilon_\circ > 0$, all sequence $(x_n)_{n\in\mathbb{N}}$ of random variables uniformly distributed in $(0,1)$ and independent of $z$, there is a probability larger than $1 - c\gamma/\nu$ to realize $z$ such that that there is a probability larger than $1-\nu$ to realize $(\varepsilon_n)_{n\in\mathbb{N}} := (\epsilon_\circ 2^{-(n+x_n)})_{n\in\mathbb{N}}$ such that $\varepsilon_n z$ is non-resonant (i.e. $\varepsilon_n z \in \mathcal{U}_{\varepsilon_n,\gamma,r,s}$). More formally, we have*

$$\mathbb{P}\left(\mathbb{P}\left(\forall\, n,\ \varepsilon_n z \in \mathcal{U}_{\varepsilon_n,\gamma,r,s} \mid z\right) \geqslant 1-\nu\right) \geqslant 1 - c\frac{\gamma}{\nu}.$$

**Remark 4.5.1.** *The variables $x_n$ are not necessarily independent. For example, we could choose $x_n = x_0$ for all $n$.*

In order to prove these propositions and the corollary, we introduce some notations and elementary stochastic lemmas.

**Definition 4.5.2.** *If a random variable $I$ has a density with respect to the Lebesgue measure, we denote $f_I$ its density, i.e.*

$$\forall g \in C^0_b(\mathbb{R}), \quad \mathbb{E}\left[g(I)\right] = \int_{\mathbb{R}} g(x) f_I(x)\mathrm{dx}.$$

**Lemma 4.5.3.** *If $I$ is a random variable with density with values in $A \subset \mathbb{R}$ and $\phi : A \to \mathbb{R}$ satisfies $\phi' > 0$ then $\phi$ has a density and*

$$f_{\phi(I)} = \frac{f_I}{\phi'} \circ \phi^{-1}.$$

*Proof.* If $g \in C_b^0(\mathbb{R})$ then we have

$$\mathbb{E}\left[g \circ \phi(I)\right] = \int_I g \circ \phi(x) f_I(x) \mathrm{dx} = \int_{\mathrm{Im}\phi} g(y) \frac{f_I}{\phi'} \circ \phi^{-1}(y) \mathrm{dy}.$$

□

**Corollary 4.5.2.** *If $I^2$ is uniformly distributed on $(0,1)$ and $I \geqslant 0$ almost surely then $I$ has a density given by*

$$\forall x \in \mathbb{R},\ f_I(x) = 2x \mathbb{1}_{0<x<1}.$$

*Moreover, if $I$ has a density and $\varepsilon > 0$ then $\varepsilon I$ has a density and for all $x \in \mathbb{R}$*

$$\forall x \in \mathbb{R},\ f_{\varepsilon I}(x) = \frac{1}{\varepsilon} f_I(\frac{x}{\varepsilon}).$$

*In particular, we have for all $a \in \mathbb{Z}$ and $x \in \mathbb{R}$,*

$$f_{I_a^2}(x) = \langle a \rangle^{4s+8} \mathbb{1}_{0<x<\langle a \rangle^{-4s-8}} \quad \textit{and} \quad f_{I_a}(x) = 2x \langle a \rangle^{4s+8} \mathbb{1}_{0<x<\langle a \rangle^{-2s-4}}. \tag{4.49}$$

**Lemma 4.5.4.** *Let $I, J$ be some real independent random variables. If $I$ has a density, then for all $\gamma > 0$*

$$\mathbb{P}(|I+J| < \gamma) \leqslant 2\gamma \|f_I\|_{L^\infty}.$$

*Proof.* By Tonelli theorem, we have

$$\mathbb{P}(|I+J| < \gamma) = \mathbb{E}\left[\mathbb{1}_{|I+J|<\gamma}\right] = \mathbb{E}\left[\int_{J-\gamma}^{J+\gamma} f_I(x) \mathrm{dx}\right] \leqslant 2\gamma \|f_I\|_{L^\infty}.$$

□

**Lemma 4.5.5.** *If $a, b, \gamma \in \mathbb{R}^*$ are such that $0 < \gamma < |b|$ then for all $\sigma \in \mathbb{R}^*$*

$$\lambda(|a\varepsilon^\sigma + b| < \gamma) \leqslant \frac{2\gamma}{|\sigma|(|b| - \gamma)}.$$

*Proof.* Applying a natural change of coordinate, we get

$$\begin{aligned}\lambda(|a\varepsilon^\sigma + b| < \gamma) &= \int_{\mathbb{R}_+^*} \mathbb{1}_{|a\varepsilon^\sigma + b|<\gamma} \frac{\mathrm{d}\varepsilon}{\varepsilon} = \frac{1}{|\sigma|} \int_{\mathbb{R}_+^*} \mathbb{1}_{|a\varepsilon + b|<\gamma} \frac{\mathrm{d}\varepsilon}{\varepsilon} \\ &= \frac{1}{|\sigma|} \int_{\varepsilon \in \left(\frac{-b-\gamma}{a}; \frac{-b+\gamma}{a}\right) \cap \mathbb{R}_+^*} \frac{\mathrm{d}\varepsilon}{\varepsilon} \leqslant \frac{2\gamma}{|\sigma|(|b| - \gamma)}.\end{aligned}$$

□

A first application of these lemmas is the genericity of the non-resonance assumption (4.44).

**Lemma 4.5.6.** *There exists a constant $c > 0$ such that for all $\gamma \in (0,1)$ we have*

$$\mathbb{P}\left(\forall \boldsymbol{k} \in \mathcal{I}rr,\ \sharp k \leqslant 2r \Rightarrow |\omega_{\boldsymbol{k}}(I)| \geqslant \gamma \left(\prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha \rangle^{-2}\right) \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s}\right) \geqslant 1 - c\gamma.$$

*Proof.* We are going to bound the probability of the complementary event by $c\gamma$. For each $\boldsymbol{k} \in \mathrm{Irr}$ of length smaller than or equal to $2r$, we have to estimate $\mathbb{P}\left(|\omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right)$, where $\gamma_{\boldsymbol{k}} > 0$ will be judiciously chosen.
We recall that by definition, if $\boldsymbol{k} = (k_1, \ldots, k_{\sharp \boldsymbol{k}})$ and $k_\alpha = (\delta_\alpha, a_\alpha)$ then we have

$$\omega_{\boldsymbol{k}}(I) = -\varphi'(0) \sum_{\alpha=1}^{\sharp \boldsymbol{k}} \delta_\alpha I_{a_\alpha}.$$

So paying attention to the multiplicity, this sum writes

$$\omega_{\boldsymbol{k}}(I) = \sum_{\beta=1}^{m} p_\beta I_{a_{\alpha_\beta}},$$

where $m \leqslant \sharp \boldsymbol{k}$ is an integer, $p \in \mathbb{R}^m$ satisfies $|p_\beta| \geqslant |\varphi'(0)|$ and $(a_{\alpha_\beta})_\beta$ is a subsequence of $a$. Thus, using Lemma 4.5.4, we have

$$\mathbb{P}\left(|\omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right) \leqslant 2\gamma_{\boldsymbol{k}} \min_{\beta=1,\ldots,m} \frac{\|f_{I_{a_{\alpha_\beta}}}\|_{L^\infty}}{|p_\beta|} \leqslant \frac{2\gamma_{\boldsymbol{k}}}{|\varphi'(0)|} \min_{\alpha=1,\ldots,\sharp \boldsymbol{k}} \|f_{I_{a_\alpha}}\|_{L^\infty} \leqslant \frac{4\gamma_{\boldsymbol{k}}}{|\varphi'(0)|} \langle \mu_{\min}(\boldsymbol{k}) \rangle^{2s-4},$$

by using (4.49). Consequently, if we take

$$\gamma_{\boldsymbol{k}} = \gamma \left( \prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha \rangle^{-2} \right) \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s}$$

we get

$$\mathbb{P}\left(|\omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right) \leqslant \frac{4\gamma}{|\varphi'(0)|} \prod_{\alpha=1}^{\sharp \boldsymbol{k}-2} \langle \mu_\alpha(\boldsymbol{k}) \rangle^{-2}.$$

Using the fact that $\boldsymbol{k} \in \mathcal{R}$ and the zero momenta conditions (4.32), we see that $\mu_{\sharp \boldsymbol{k}}(\boldsymbol{k})$ and $\mu_{\sharp \boldsymbol{k}-1}(\boldsymbol{k})$ can be expressed as functions of $\mu_{\sharp \boldsymbol{k}-2}(\boldsymbol{k}), \ldots, \mu_1(\boldsymbol{k})$, so this last product is summable on $\{\boldsymbol{k} \in \mathrm{Irr} \mid \sharp \boldsymbol{k} \leqslant 2r\}$. Consequently, there exists a constant $c > 0$ such that

$$\mathbb{P}\left(\exists \boldsymbol{k} \in \mathcal{I}rr,\ \sharp k \leqslant 2r \text{ and } |\omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right) \leqslant \sum_{\substack{\boldsymbol{k} \in \mathrm{Irr} \\ \sharp \boldsymbol{k} \leqslant 2r}} \mathbb{P}\left(|\omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right) \leqslant c\gamma.$$

□

A second application is the proof of Corollary 4.5.1 of Proposition 4.5.2.

**Proof of Corollary 4.5.1**.

We denote $\mathscr{E}_\lambda$ the event defined by

$$\mathscr{E}_\lambda = \{\lambda\left(\varepsilon z \in \ell^1_s \backslash \mathcal{U}_{\gamma,\varepsilon,r,s}\right) < \nu\}.$$

Applying Proposition (4.5.2), there exists a constant $c > 0$ such that for all $\gamma \in (0,1)$ we have $\mathbb{P}(\mathscr{E}_\lambda) \geqslant 1 - c\gamma/\nu$. Thus, we will conclude this proof showing that

$$\mathbb{P}\left(\forall\, n,\ \varepsilon_n z \in \mathcal{U}_{\varepsilon_n,\gamma,r,s} \mid \mathscr{E}_\lambda\right) \geqslant 1 - \nu.$$

To show this, we just have to prove that

$$\sum_{n\in\mathbb{N}} \mathbb{P}\left(\varepsilon_n z \in \ell^1_s \backslash \mathcal{U}_{\varepsilon_n,\gamma,r,s} \mid \mathscr{E}_\lambda\right) < \nu.$$

By a natural change of variable (see Lemma 4.5.3), $\varepsilon_n$ has a density given by

$$f_{\varepsilon_n} = \frac{1}{\log 2}\begin{cases} \epsilon_\circ^{-1}\varepsilon^{-1} & \text{if } \epsilon_\circ 2^{-n-1} < \varepsilon < \epsilon_\circ 2^{-n}, \\ 0 & \text{else.} \end{cases}$$

Consequently, since $(\varepsilon_n)$ is independent of $z$, applying Chasles formula, we get

$$\begin{aligned}\sum_{n\in\mathbb{N}} \mathbb{P}\left(\varepsilon_n z \in \ell^1_s \backslash \mathcal{U}_{\varepsilon_n,\gamma,r,s} \mid \mathscr{E}_\lambda\right) &= \sum_{n\in\mathbb{N}} \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\varepsilon z\in\ell^1_s\backslash\mathcal{U}_{\gamma,\varepsilon,r,s}} f_{\varepsilon_n}(\varepsilon)\mathrm{d}\varepsilon \mid \mathscr{E}_\lambda\right] \\ &= \frac{1}{\log 2}\mathbb{E}\left[\sum_{n\in\mathbb{N}}\int_{2^{-n-1}}^{2^{-n}} \mathbb{1}_{\varepsilon z\in\ell^1_s\backslash\mathcal{U}_{\gamma,\varepsilon,r,s}} \frac{\mathrm{d}\varepsilon}{\varepsilon} \mid \mathscr{E}_\lambda\right] \leqslant \frac{1}{\log 2}\mathbb{E}\left[\nu \mid \mathscr{E}_\lambda\right] = \frac{\nu}{\log 2},\end{aligned}$$

and we easily obtain the result after a scaling in $\nu$. □

To take into account the terms induced by $Z_6$ in the proof Proposition 4.5.1 and Proposition 4.5.2, we are going to need an useful algebraic lemma.

**Lemma 4.5.7.** *If $\boldsymbol{k} = (k_1, \dots, k_{2m}) \in \mathcal{I}rr$ with $k_\alpha = (\delta_\alpha, a_\alpha)$, there exists $a^* \in ]\!]-3m, 3m[\![\backslash\{a_1, \dots, a_{2m}\}$ such that*

$$\left|\sum_{\alpha=1}^{2m} \frac{\delta_\alpha}{(a^* - a_\alpha)^2}\right| \geqslant (6m)^{-4m} \prod_{\alpha=1}^{2m} \langle a_\alpha\rangle^{-2}. \tag{4.50}$$

*Proof.* First, we observe that there exists $P \in \mathbb{Z}[X]$ of degree smaller than or equal to $4m-2$ such that

$$\sum_{\alpha=1}^{2m} \frac{\delta_\alpha}{(X - a_\alpha)^2} = P(X) \prod_{\alpha=1}^{2m} \frac{1}{(X - a_\alpha)^2}. \tag{4.51}$$

Since $\boldsymbol{k}$ is irreductible, we deduce of the uniqueness of partial fraction decomposition that $P \neq 0$. Hence, $P$ vanishes in, at most, $4m-2$ points. But there are, at least, $4m-1$ points into $]\!]-3m, 3m[\![\backslash\{a_1, \dots, a_{2m}\}$. So we can find $a^*$ in this set such that $P(a^*) \neq 0$.

Then, since $P \in \mathbb{Z}[X]$ and $a^* \in \mathbb{Z}$, we deduce that $|P(a^*)| \geqslant 1$. Thus, to prove (4.50), we just have to bound each factor of the denominator in (4.51) by

$$|a_\alpha - a^*| \leqslant 6m\langle a_\alpha\rangle.$$

To get this estimate we just have to observe that if $|a_\alpha| < 3m$ then

$$|a_\alpha - a^*| \leqslant |a_\alpha| + |a^*| \leqslant 6m \leqslant 6m\langle a_\alpha\rangle,$$

whereas, if $|a_\alpha| \geqslant 3m$ then $|a^*| \leqslant |a_\alpha|$ and so

$$|a_\alpha - a^*| \leqslant |a_\alpha| + |a^*| \leqslant 2|a_\alpha| \leqslant 6m\langle a_\alpha\rangle.$$

□

**Proof of Proposition 4.5.2.** Let $\nu \in (0, \frac{1}{2})$ and $\gamma \in (0,1)$. We introduce three events

$$\mathscr{E}_4 = \left\{ \forall \boldsymbol{k} \in \mathcal{I}rr,\ \sharp k \leqslant 2r \Rightarrow |\omega_{\boldsymbol{k}}(I)| \geqslant \frac{\gamma}{\nu} \left( \prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha \rangle^{-2} \right) \langle \mu_{\min}(\boldsymbol{k}) \rangle^{-2s} \right\},$$

$$\mathscr{E}_6 = \left\{ \forall \boldsymbol{k} \in \mathcal{I}rr,\ \sharp k \leqslant 2r \Rightarrow |\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I)| \geqslant \frac{\gamma}{\nu} \left( \prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha \rangle^{-4} \right) \right\},$$

and

$$\mathscr{E}_\lambda = \left\{ \lambda \left( \varepsilon z \in \ell^1_s \backslash \mathcal{U}_{\gamma,\varepsilon,r,s} \right) < c_\lambda \nu \right\},$$

where $c_\lambda$ is a positive constant that will be determine later.

We have proven in Lemma 4.5.6 that there exists a constant $c_4 > 0$ such that

$$\mathbb{P}\left(\mathscr{E}_4\right) \geqslant 1 - c_4 \frac{\gamma}{\nu}.$$

We are going to prove, on the one hand, that there exists a constant $c_6 > 0$ (independent of $\gamma$ and $\nu$) such that

$$\mathbb{P}\left(\mathscr{E}_6\right) \geqslant 1 - c_6 \frac{\gamma}{\nu}. \tag{4.52}$$

On the other hand, we will prove that

$$\mathscr{E}_4 \cap \mathscr{E}_6 \subset \mathscr{E}_\lambda. \tag{4.53}$$

Assuming (4.52) and (4.53), and up to a natural rescaling with respect to $\nu$, Proposition 4.5.2 becomes a straightforward estimate :

$$\begin{aligned} \mathbb{P}(\mathscr{E}_\lambda) \geqslant \mathbb{P}(\mathscr{E}_4 \cap \mathscr{E}_6) = 1 - \mathbb{P}(\mathscr{E}_4^c \cup \mathscr{E}_6^c) &\geqslant 1 - (\mathbb{P}(\mathscr{E}_4^c) + \mathbb{P}(\mathscr{E}_6^c)) \\ &= \mathbb{P}(\mathscr{E}_4) + \mathbb{P}(\mathscr{E}_6) - 1 \geqslant 1 - \frac{c_4 + c_6}{\nu}\gamma. \end{aligned}$$

First, we focus on the proof of (4.52), which is similar to the proof of Lemma 4.5.6. We are going to bound the probability of the complementary event by $c_6 \frac{\gamma}{\nu}$. For each $\boldsymbol{k} \in \mathrm{Irr}$ of length smaller than or equal to $2r$, we have to estimate $\mathbb{P}\left(|\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right)$, where $\gamma_{\boldsymbol{k}} > 0$ will be judiciously chosen.

We recall that by definition, if $\boldsymbol{k} = (k_1, \ldots, k_{\sharp k})$ and $k_\alpha = (\delta_\alpha, a_\alpha)$ then we have

$$\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I) = -\frac{1}{2}\varphi'(0)^2 \sum_{\substack{b \in \mathbb{Z} \\ b \notin \{a_1, \ldots, a_{2m}\}}} I_b^2 \sum_{\alpha=1}^{2m} \frac{\delta_\alpha}{(a_\alpha - b)^2}. \tag{4.54}$$

Thus, using Lemma 4.5.4 and Corollary (4.5.2) we have

$$\mathbb{P}\left(|\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I)| < \gamma_{\boldsymbol{k}}\right) \leqslant 2\gamma_{\boldsymbol{k}} \inf_{b \in \mathbb{Z}} \left| \frac{1}{2}\varphi'(0)^2 \langle b \rangle^{-4s-8} \sum_{\alpha=1}^{2m} \frac{\delta_\alpha}{(a_\alpha - b)^2} \right|^{-1}.$$

Applying Lemma 4.5.7 to estimate this infimum, we get

$$\mathbb{P}\left(|\widetilde{\Omega}_{\boldsymbol{k}}(I)-\omega_{\boldsymbol{k}}(I)|<\gamma_{\boldsymbol{k}}\right)\leqslant 4\gamma_{\boldsymbol{k}}(\varphi'(0))^{-2}(6r)^{4r}(3r)^{4s+8}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle a_\alpha\rangle^2.$$

Consequently, choosing $\gamma_{\boldsymbol{k}}=\frac{\gamma}{\nu}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-4}$, we get $\mathbb{P}(\mathscr{E}_6)\geqslant 1-c_6\frac{\gamma}{\nu}$ with

$$c_6=4\varphi'(0)^{-2}(6r)^{4r}(3r)^{4s+8}\sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}.$$

Now, we focus on the proof of (4.53). So, we consider a realization of the actions $I=(I_a)_{a\in\mathbb{Z}}$ where the lower bounds characterising $\mathscr{E}_4$ and $\mathscr{E}_6$ are satisfied, *i.e.* for all $\boldsymbol{k}$ irreductible of length smaller than or equal to $2r$ we have

$$|\omega_{\boldsymbol{k}}(I)|\geqslant\frac{\gamma}{\nu}\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s},\tag{4.55}$$

and

$$|\widetilde{\Omega}_{\boldsymbol{k}}(I)-\omega_{\boldsymbol{k}}(I)|\geqslant\frac{\gamma}{\nu}\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-4}\right).\tag{4.56}$$

We have to estimate $\lambda\left(\varepsilon z\in\ell^1_s\backslash\mathcal{U}_{\gamma,\varepsilon,r,s}\right)$ for such a realization $I$. Thus, we decompose naturally the set we are estimating :

$$\left\{\varepsilon>0\mid\varepsilon z\in\ell^1_s\backslash\mathcal{U}_{\gamma,\varepsilon,r,s}\right\}=\Sigma_1\cup\Sigma_2\cup\Sigma_3,$$

with

$$\Sigma_1=\left\{\varepsilon>0\mid\exists\boldsymbol{k}\in\mathcal{I}rr,\ \sharp\boldsymbol{k}\leqslant 2r\ \text{ and }|\omega_{\boldsymbol{k}}(\varepsilon^2 I)|<\gamma\varepsilon^2\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s}\right\},$$

$$\Sigma_2=\left\{\varepsilon>0\mid\exists\boldsymbol{k}\in\mathcal{I}rr,\ \sharp\boldsymbol{k}\leqslant 2r\ \text{ and }|\widetilde{\Omega}_{\boldsymbol{k}}(\varepsilon^2 I)|<\gamma\varepsilon^2\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s}\right\},$$

and

$$\Sigma_3=\left\{\varepsilon>0\mid\exists\boldsymbol{k}\in\mathcal{I}rr,\ \sharp\boldsymbol{k}\leqslant 2r\text{ and }|\widetilde{\Omega}_{\boldsymbol{k}}(\varepsilon^2 I)|<\gamma\varepsilon^4\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\right\}.$$

In fact, since $\omega_{\boldsymbol{k}}$ is linear, $I$ satisfies (4.55) and $\nu<1$, $\Sigma_1$ is the empty set. Thus, we have

$$\lambda\left(\varepsilon z\in\ell^1_s\backslash\mathcal{U}_{\gamma,\varepsilon,r,s}\right)\leqslant\lambda(\Sigma_2)+\lambda(\Sigma_3).$$

So, we have to estimate $\lambda(\Sigma_2)$ and $\lambda(\Sigma_3)$. Observing that $\widetilde{\Omega}_{\boldsymbol{k}}-\omega_{\boldsymbol{k}}$ is quadratic, we have

$$\widetilde{\Omega}_{\boldsymbol{k}}(\varepsilon^2 I)=\varepsilon^2\omega_{\boldsymbol{k}}(I)+\varepsilon^4\left(\widetilde{\Omega}_{\boldsymbol{k}}(I)-\omega_{\boldsymbol{k}}(I)\right).$$

Thus, we are going to estimate $\lambda(\Sigma_2)$ and $\lambda(\Sigma_3)$ with Lemma 4.5.5. To apply, this Lemma, we observe that from (4.55) and (4.56) we have

$$|\omega_{\boldsymbol{k}}(I)| - \gamma\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s} \geqslant \gamma\left(\frac{1}{\nu}-1\right)\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s},$$

and

$$|\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I)| - \gamma\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right) \geqslant \gamma\left(\frac{1}{\nu}-1\right)\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-4}\right).$$

Consequently, applying Lemma 4.5.5 with $\gamma_{\boldsymbol{k},s} := \gamma\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s}$, we get

$$\begin{aligned}\lambda(\Sigma_2) \leqslant \sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}} \lambda\left(|\omega_{\boldsymbol{k}}(I) + \varepsilon^2\left(\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I)\right)| < \gamma_{\boldsymbol{k},s}\right) &\leqslant \sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}} \frac{\gamma_{\boldsymbol{k},s}}{|\omega_{\boldsymbol{k}}(I)| - \gamma_{\boldsymbol{k},s}}\\ &\leqslant \left(\sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-4}\right)\frac{\nu}{1-\nu},\end{aligned}$$

as the first previous estimate can be recast as $|\omega_{\boldsymbol{k}}(I)| - \gamma_{\boldsymbol{k},s} \geqslant \gamma_{\boldsymbol{k},s}(\frac{1}{\nu}-1)\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^4$. Similarly, we obtain

$$\begin{aligned}\lambda(\Sigma_3) \leqslant \sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}} \lambda\left(|\varepsilon^{-2}\omega_{\boldsymbol{k}}(I) + \left(\widetilde{\Omega}_{\boldsymbol{k}}(I) - \omega_{\boldsymbol{k}}(I)\right)| < \gamma\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\right)\\ \leqslant \left(\sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}\right)\frac{\nu}{1-\nu}.\end{aligned}$$

Hence, since these sum are clearly convergent, we have proven that $\mathscr{E}_4 \cap \mathscr{E}_6 \subset \mathscr{E}_\lambda$ for a convenient choice of $c_\lambda$. □

**Proof of Proposition 4.5.1.** Let $\varepsilon > 0$ be a fixed positive real number. By definition of $\mathcal{U}_{\gamma,\varepsilon,r,s}$, we decompose $\{\varepsilon z \in \mathcal{U}_{\gamma,\varepsilon,r,s}\}$ into

$$\{\varepsilon z \in \mathcal{U}_{\gamma,\varepsilon,r,s}\} = \mathscr{E}_4 \cap \mathscr{E}_{46}$$

where

$$\mathscr{E}_4 = \left\{\forall \boldsymbol{k}\in\mathrm{Irr},\ \sharp k \leqslant 2r \Rightarrow |\omega_{\boldsymbol{k}}(\varepsilon^2 I)| \geqslant \gamma\varepsilon^2\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}\right)\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s}\right\}$$

and

$$\mathscr{E}_{46} = \left\{\forall \boldsymbol{k}\in\mathrm{Irr},\ \sharp k \leqslant 2r \Rightarrow |\widetilde{\Omega}_{\boldsymbol{k}}(\varepsilon^2 I)| \geqslant \gamma\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\max\left(\varepsilon^2\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s}, \varepsilon^4\right)\right\}.$$

Since $\omega_{\boldsymbol{k}}$ is linear, $\mathscr{E}_4$ does not depend of $\varepsilon$. Consequently, applying Lemma 4.5.6, we get a constant $c_4 > 0$ such that $\mathbb{P}(\mathscr{E}_4) \geqslant 1 - c_4\gamma$. So assuming that there exists a constant $c_{46} > 0$

such that $\mathbb{P}(\mathscr{E}_{46}) \geqslant 1 - c_{46}\gamma$, we could conclude the proof of Proposition 4.5.1 by the following estimate

$$\mathbb{P}\left(\varepsilon z \in \mathcal{U}_{\gamma,\varepsilon,r,s}\right) \;=\; 1 - \mathbb{P}(\mathscr{E}_4^c \cup \mathscr{E}_{46}^c) \;\geqslant\; \mathbb{P}(\mathscr{E}_4) + \mathbb{P}(\mathscr{E}_{46}) - 1 \;\geqslant\; 1 - (c_4 + c_{46})\gamma.$$

Thus, we just have to focus on the proof of the existence of $c_{46}$. So, we recall that by definition, if $\boldsymbol{k} = (k_1, \dots, k_{\sharp\boldsymbol{k}})$ and $k_\alpha = (\delta_\alpha, a_\alpha)$ then we have with $\sharp\boldsymbol{k} = 2m$,

$$\widetilde{\Omega}_{\boldsymbol{k}}(\varepsilon^2 I) = -\varepsilon^2\varphi'(0)\sum_{\alpha=1}^{2m}\delta_\alpha I_{a_\alpha} - \frac{\varepsilon^4}{2}\varphi'(0)^2 \sum_{\substack{b\in\mathbb{Z}\\ b\notin\{a_1,\dots,a_{2m}\}}} I_b^2 \sum_{\alpha=1}^{2m}\frac{\delta_\alpha}{(a_\alpha - b)^2}.$$

By construction, it is a sum of independent random variable, thus applying Lemma 4.5.4, we have the estimate on the complement

$$\mathbb{P}(\mathscr{E}_{46}^c) \leqslant \sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}} 2\gamma_{\boldsymbol{k},\varepsilon}\min\left(\min_{\alpha=1,\dots,2m}\|F_\alpha\|_{L^\infty}, \inf_{b\in\mathbb{Z}}\|G_b\|_{L^\infty}\right),$$

where $F_\alpha$ is the probability density function of the part depending on $I_\alpha$ in $\widetilde{\Omega}_{\boldsymbol{k}}(\varepsilon^2 I)$ and $G_b$ the probability density function of the part depending on the variable $I_b^2$, and where

$$\gamma_{\boldsymbol{k},\varepsilon} = \gamma\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-6}\right)\max\left(\varepsilon^2\langle\mu_{\min}(\boldsymbol{k})\rangle^{-2s}, \varepsilon^4\right).$$

By using (4.49), we have that

$$\|F_\alpha\|_{L^\infty} \leqslant \frac{\|f_{I_\alpha}\|_{L^\infty}}{|\varepsilon^2\varphi'(0)\mathrm{Card}\{\beta \mid a_\beta = a_\alpha\}|} \leqslant \frac{\langle a_\alpha\rangle^{2s+4}}{|\varepsilon^2\varphi'(0)|}$$

and thus

$$\min_{\alpha=1,\dots,2m}\|F_\alpha\|_{L^\infty} \leqslant \frac{\langle\mu_{\min}(\boldsymbol{k})\rangle^{2s+4}}{|\varepsilon^2\varphi'(0)|}.$$

Similarly using Lemma 4.5.7, we obtain

$$\begin{aligned}\inf_{b\in\mathbb{Z}}\|G_b\|_{L^\infty} &\leqslant \min_{b\in]\!]-3m,3m[\![\setminus\{a_1,\dots,a_{2m}\}} \frac{\|f_{I_b^2}\|_{L^\infty}}{\left|\frac{\varepsilon^4}{2}\varphi'(0)^2\sum_{\alpha=1}^{2m}\frac{\delta_\alpha}{(a_\alpha-b)^2}\right|}\\ &\leqslant \frac{4\langle 3m\rangle^{2s+4}(6m)^{4m}}{\varepsilon^4\varphi'(0)^2}\prod_{\alpha=1}^{2m}\langle a_\alpha\rangle^2 \leqslant \frac{4\langle 3r\rangle^{2s+4}(6r)^{4r}}{\varepsilon^4\varphi'(0)^2}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^2.\end{aligned}$$

Hence there exists a constant $c_{r,s}$ depending on $r, s$ such that

$$\mathbb{P}(\mathscr{E}_{46}^c) \leqslant \sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}} 2\gamma_{\boldsymbol{k},\varepsilon}\min\left(\frac{\langle\mu_{\min}(\boldsymbol{k})\rangle^{2s+4}}{|\varepsilon^2\varphi'(0)|}, \frac{c_{r,s}}{\varepsilon^4\varphi'(0)^2}\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^2\right) \leqslant \gamma c\sum_{\substack{\boldsymbol{k}\in\mathrm{Irr}\\ \sharp\boldsymbol{k}\leqslant 2r}}\left(\prod_{\alpha=1}^{\sharp\boldsymbol{k}}\langle k_\alpha\rangle^{-2}\right) \leqslant c_{46}\gamma,$$

for some constants $c$ and $c_{46}$, and this shows the result. □

## 4.6 A class of rational Hamiltonians

The set of Hamiltonian functions constructed in the normal form process arise from solving homological equation associated with small denominators $\omega_{\boldsymbol{k}}(I)$ and $\Omega_{\boldsymbol{k}}(I)$ (see (4.38)). Then a natural class of Hamiltonians should be

$$F(z) = \sum_{\boldsymbol{j}\in\mathcal{R}} f_{\boldsymbol{j}}(I) z_{\boldsymbol{j}} \tag{4.57}$$

for some functions $f_{\boldsymbol{j}}$ which are inverse of products of small denominators $\omega_{\boldsymbol{k}}(I)$ and $\Omega_{\boldsymbol{k}}(I)$ associated with multi-indices depending on the construction process. Note that we sum over $\boldsymbol{j} \in \mathcal{R}$ since the non resonant part will be killed beforehand by a standard resonant normal form procedure involving polynomial Hamiltonians (see section 4.7.1). Each term of (4.57) will be controlled by the non resonance conditions (4.44) and (4.45), provided we can compensate the loss of derivative arising in the small denominator by terms in the numerator $z_{\boldsymbol{j}}$. For a given $\boldsymbol{j}$, several terms can appear in $f_{\boldsymbol{j}}$ that are associated with different small denominators. To take into account the specificity of each term arising in the normal form process described in section 4.7 we will introduce four sub-classes of rational Hamiltonians.
Notice that in the case of (NLSP), the situation is simpler and only two sub-classes are needed (see Appendix 4.9.1).

### 4.6.1 Construction of the class

First we introduce the following set of indices, encoding the structure of the possible terms $f_{\boldsymbol{j}}(I)$ arising in (4.57).

For $r \in \mathbb{N}$, let $\mathscr{H}_r$ be a set of multi-indices valued functions

$$(\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) : \mathbb{Z}^* \to \mathcal{R} \times \bigsqcup_{p\in\mathbb{N}} \mathcal{I}rr^p \times \bigsqcup_{q\in\mathbb{N}} \mathcal{I}rr^q \times \mathbb{Z}^*. \tag{4.58}$$

For a given $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ and $\ell \in \mathbb{Z}^*$, we associated $\pi_\ell \in \mathcal{R}_{m_\ell}$ for some $m_\ell \in \mathbb{N}$, $\mathbf{k}_\ell = (\boldsymbol{k}_{\ell,1}, \dots, \boldsymbol{k}_{\ell,p_\ell})$ and $\mathbf{h}_\ell = (\boldsymbol{h}_{\ell,1}, \dots, \boldsymbol{h}_{\ell,q_\ell})$ for some $p_\ell$ and $q_\ell$ in $\mathbb{N}$. For some given set of coefficients $c = (c_\ell)_{\ell\in\mathbb{Z}^*} \in \mathbb{C}^{\mathbb{Z}^*}$ we will define the Hamiltonian function

$$Q_\Gamma[c](z) = \sum_{\ell\in\mathbb{Z}^*} c_\ell (-i)^{p_\ell+q_\ell} \frac{z_{\boldsymbol{\pi}_\ell}}{\prod_{\alpha=1}^{n_\ell} \omega_{\boldsymbol{k}_{\ell,\alpha}} \prod_{\alpha=n_\ell+1}^{p_\ell} \Omega_{\boldsymbol{k}_{\ell,\alpha}} \prod_{\alpha=1}^{q_\ell} \Omega_{\boldsymbol{h}_{\ell,\alpha}}}. \tag{4.59}$$

Note that such Hamiltonian can be recast under the form (4.57) by setting

$$f_{\boldsymbol{j}}(I) = \sum_{\ell\in\boldsymbol{\pi}^{-1}(\boldsymbol{j})} c_\ell (-i)^{p_\ell+q_\ell} \frac{1}{\prod_{\alpha=1}^{n_\ell} \omega_{\boldsymbol{k}_{\ell,\alpha}} \prod_{\alpha=n_\ell+1}^{p_\ell} \Omega_{\boldsymbol{k}_{\ell,\alpha}} \prod_{\alpha=1}^{q_\ell} \Omega_{\boldsymbol{h}_{\ell,\alpha}}}. \tag{4.60}$$

Roughly speaking, the structure of the class can be explained as follows : Each time a homological equation for $Z_4$ or $Z_4 + Z_6$ is solved, the term $f_{\boldsymbol{j}}(I)$ is divided by $\omega_{\boldsymbol{j}}$ or $\Omega_{\boldsymbol{j}}$ so the functionals are naturally under the previous form. The fact that we decompose into two parts the contribution of $\Omega_{\boldsymbol{j}}$ in the denominator is explained by the $2r$ order condition **(ii)** just below.

To ensure that the Hamiltonians $Q_\Gamma[c]$ are well defined and that their vectorfield can be controlled in $\ell^1_s$, we impose several restrictions on the set $\mathscr{H}_r : \Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ if the following conditions are satisfied

**(i) Reality**. The functional is real, *i.e.* $Q_\Gamma[c](z) \in \mathbb{R}$ for real $z$. This condition is satisfied by imposing for all $\ell \in \mathbb{Z}^*$,

$$\boldsymbol{\pi}_{-\ell} = \overline{\boldsymbol{\pi}_\ell}, \quad \overline{\mathbf{k}_\ell} = \mathbf{k}_{-\ell}, \quad n_{-\ell} = n_\ell, \quad \text{and} \quad \overline{\mathbf{h}_\ell} = \mathbf{h}_{-\ell}$$

after noticing that $\omega_{\overline{\boldsymbol{j}}} = -\omega_{\boldsymbol{j}}$ and $\Omega_{\overline{\boldsymbol{j}}} = -\Omega_{\boldsymbol{j}}$. Note that this implies that $p_\ell, q_\ell$ and $m_\ell$ are even functions of $\ell$.

**(ii) $2r$ order**. The link between the terms of the class and the order $2r$ is given by the relation

$$\forall \ell \in \mathbb{Z}^*, \quad r = m_\ell - p_\ell - 2q_\ell. \tag{4.61}$$

This definition corresponds to the fact that while the numerator $z_{\pi_\ell}$ is of order $\varepsilon^{2m_\ell}$ and the homogeneity of the non resonance condition of $\omega_{\boldsymbol{k}_{\ell,\alpha}}$ is $\varepsilon^2$, the homogeneity of the small denominator $\Omega_{\boldsymbol{k}}$ can be of order $\varepsilon^2$ or $\varepsilon^4$ depending on the non resonance condition we use in (4.47). The previous notation then specifies that $\Omega_{\boldsymbol{k}_{\ell,\alpha}}$ will be controlled by the non resonance condition homogeneous to $\varepsilon^2$ while the others, $\Omega_{\boldsymbol{h}_{\ell,\alpha}}$, will be controlled by the non resonance condition homogeneous to $\varepsilon^4$.

**(iii) Consistency**. We assume that $\forall \ell \in \mathbb{Z}^*,\ 0 \leqslant n_\ell \leqslant p_\ell$.

**(iv) Finite numerator and denominator degrees**. We assume that

$$\sup_{\ell \in \mathbb{N}} m_\ell < \infty, \tag{4.62}$$

and

$$\forall\, \ell \in \mathbb{Z}^*, \quad \forall\, \alpha, \quad \sharp \boldsymbol{k}_{\ell,\alpha} \leqslant m_\ell \quad \text{and} \quad \sharp \boldsymbol{h}_{\ell,\alpha} \leqslant m_\ell. \tag{4.63}$$

**(v) Finite multiplicity**. We assume that $\sup_{j \in \mathcal{R}} \operatorname{Card} \boldsymbol{\pi}^{-1}(j) < \infty$. This condition ensures that the number of terms defining $f_j(I)$ in (4.60) is finite.

**(vi) Distribution of the derivatives**. There exists a positive constant $C > 0$ such that for all $\ell \in \mathbb{Z}^*$, there exists $\iota : [\![1, 2p_\ell]\!] \to [\![3, 2m_\ell]\!]$, an injective function satisfying

$$\begin{cases} \displaystyle\max_{1 \leqslant \alpha \leqslant p_\ell} \frac{\langle \mu_{\min}(\boldsymbol{k}_{\ell,\alpha}) \rangle}{\langle \mu_{\iota_{2\alpha-1}}(\boldsymbol{\pi}_\ell) \rangle} \leqslant C \quad \text{and} \\ \displaystyle\max_{1 \leqslant \alpha \leqslant p_\ell} \frac{\langle \mu_{\min}(\boldsymbol{k}_{\ell,\alpha}) \rangle}{\langle \mu_{\iota_{2\alpha}}(\boldsymbol{\pi}_\ell) \rangle} \leqslant C. \end{cases} \tag{4.64}$$

This condition ensures that terms of the form $\langle \mu_{\min}(\boldsymbol{k}_{\ell,\alpha}) \rangle^{-2s}$ arising in the denominators when using (4.44) or (4.45) can be compensated by modes in the numerators $z_{\boldsymbol{\pi}_\ell}$ smaller than the third largest. In other words, the first and second largest indices in $\boldsymbol{\pi}_\ell$, $\mu_1(\boldsymbol{\pi}_\ell)$ and $\mu_2(\boldsymbol{\pi}_\ell)$ will be free and not required to control the small denominators $\omega_{\boldsymbol{k}_{\ell,\alpha}}$ and $\Omega_{\boldsymbol{k}_{\ell,\beta}}$. This will ensure a global control of the vector field associated with $Q_\Gamma[c]$ after truncation independently of $s$.

**(vii) Global control of the structure**. The following condition ensures that the structure has a kind of memory of the zero-momentum condition.

$$\sup_{\ell \in \mathbb{Z}^*} \max_{1 \leqslant \alpha \leqslant q_\ell} \frac{\langle \mu_{\min}(\boldsymbol{h}_{\ell,\alpha}) \rangle}{\langle \mu_2(\boldsymbol{\pi}_\ell) \rangle} < \infty. \tag{4.65}$$

For $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ and $c : \mathbb{Z}^* \to \mathbb{C}$, we define the weight of $c$ relatively to $\Gamma$ by

$$\mathcal{N}_\Gamma(c) = \sup_{\substack{\ell \in \mathbb{Z}^* \\ c_\ell \neq 0}} \max_{\substack{1 \leqslant \alpha \leqslant p_\ell \\ 1 \leqslant \beta \leqslant q_\ell}} \big( \langle \mu_1(\mathrm{Irr}(\boldsymbol{\pi}_\ell) \rangle, \langle \mu_1(\boldsymbol{k}_{\ell,\alpha} \rangle, \langle \mu_1(\boldsymbol{h}_{\ell,\beta} \rangle \big). \tag{4.66}$$

Then we introduce the space

$$\ell^\infty_\Gamma(\mathbb{Z}^*) = \{ c \in \ell^\infty(\mathbb{Z}^*) \mid \quad \forall\, \ell \in \mathbb{Z}^*, \quad c_{-\ell} = \overline{c_\ell} \quad \text{and} \quad \mathcal{N}_\Gamma(c) < \infty \}.$$

Note that the Hamiltonian defined by (4.59) is clearly an analytic function on an open subset of $\ell^1_s$ avoiding the zeros of the denominators. Further we note that $\mathcal{N}_\Gamma(c)$ is the maximal size of indices we have at the denominator and thus the control that we will have on this denominator when $z$ belongs to the non resonant set (see Definition 4.4.2) will only depend on $\mathcal{N}_\Gamma(c)$.

In order to stick as closely as possible to the rational Hamiltonians we are going to build in the next sections, we introduce four subclasses of $\mathscr{H}_r$ denoted by $\mathscr{H}_{r,\omega}, \mathscr{H}_{r,\Omega}, \mathscr{H}^*_{r,\omega}$ and $\mathscr{H}^*_{r,\Omega}$ respectively. This technical refinement, not really indispensable, will allow us to control $m_\ell$ (see Remark 4.6.2) which in turn will allow us to obtain better constants in our Theorems [4]. We first give the four definitions and then comment on them.

- $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ belongs to $\mathscr{H}_{r,\omega}$ if

$$\forall \ell \in \mathbb{Z}^*,\ n_\ell = p_\ell \text{ and } q_\ell = 0 \text{ and } n_\ell \leqslant 2r - 6.$$

- $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ belongs to $\mathscr{H}^*_{r,\omega}$ if

$$\forall \ell \in \mathbb{Z}^*,\ n_\ell = p_\ell \text{ and } q_\ell = 0 \text{ and } n_\ell \leqslant 2(r+1) - 5.$$

- $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ belongs to $\mathscr{H}_{r,\Omega}$ if

$$n_\ell = \alpha_1 + \alpha_2 \text{ and } p_\ell = n_\ell + \alpha_3 \text{ and } q_\ell = \alpha_4 + \alpha_5 \tag{4.67}$$

  where $\alpha \in (\mathbb{N})^5$ satisfies

$$\alpha_1 \leqslant 2r - 6 \text{ and } \alpha_2 + \alpha_3 + \alpha_4 \leqslant \alpha_5 \text{ and } \alpha_5 \leqslant r - 4. \tag{4.68}$$

- $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}_r$ belongs to $\mathscr{H}^*_{r,\Omega}$ if

$$n_\ell = \alpha_1 + \alpha_2 \text{ and } p_\ell = n_\ell + \alpha_3 \text{ and } q_\ell = \alpha_4 + \alpha_5 + 1$$

  where $\alpha \in (\mathbb{N})^5$ satisfies

$$\alpha_1 \leqslant 2(r+2) - 6 \text{ and } \alpha_2 + \alpha_3 + \alpha_4 \leqslant \alpha_5 \text{ and } \alpha_5 \leqslant (r+2) - 4. \tag{4.69}$$

Some comments to clarify the meaning of these definitions :

- $\mathscr{H}_{r,\omega}$ and $\mathscr{H}_{r,\Omega}$ will be used to describe the Hamiltonians arising in our normal forms. $\mathscr{H}^*_{r,\omega}$ and $\mathscr{H}^*_{r,\Omega}$ will be used to describe the Hamiltonians obtained after solving a Homological equation (see Lemmas 4.6.4 and 4.6.5), and thus, that govern our canonical changes of variables.

4. Without tracking the form of our rational normal forms we will obtain in the right hand side of (4.14) $1 - \varepsilon^\nu$ for some constant $\nu$ depending on $r$ and $s$, instead of $1 - \varepsilon^{1/3}$.

- $\mathscr{H}_{r,\omega} \subset \mathscr{H}_{r,\Omega}$ by taking $\alpha_i = 0$ for $i = 2, \dots, 5$. Nevertheless we prefer to introduce the class $\mathscr{H}_{r,\omega}$ since it plays a special role in our construction. Actually in our second step of normal form (see section 4.7.2) we only use the class $\mathscr{H}_{r,\omega}$ while in the third step of normal form (see section 4.7.3) we only use the class $\mathscr{H}_{r,\Omega}$.
- the $\alpha_i$ give some informations about the history that generated the term $\Gamma$ : $\alpha_1$ counts the number of homological equations we solved with $Z_4$ in the second normal form process (section 4.7.2) ; $\alpha_2$ increases when in a Poisson bracket, some $\omega_{\boldsymbol{k}}(I)$ is involved (see (4.113)) in the third normal form process (section 4.7.3) ; $\alpha_5$ control the number of homological equations we solved with $Z_4 + Z_6$ in the third normal form process (section 4.7.3) ; $\alpha_3$ increases when in a Poisson bracket, some $\Omega_{\boldsymbol{k}}(I)$ is involved and we apply the derivative on the part of $\Omega_{\boldsymbol{k}}(I)$ that comes from $Z_4$ ; $\alpha_4$ increases when in a Poisson bracket, some $\Omega_{\boldsymbol{k}}(I)$ is involved and we apply the derivative on the part of $\Omega_{\boldsymbol{k}}(I)$ that comes from $Z_6$.
- the precise numerology is dictated by the experience of calculating the first terms and by the need for the overall structure to be stable by Poisson bracket (see Lemma 4.6.6 which underlies the whole construction).

We eventually define the set of functionals associated with a structure in $\mathscr{H}_r$,

$$\mathscr{F}_r := \{F = Q_\Gamma[c], \quad \Gamma \in \mathscr{H}_r, \quad c \in \ell^\infty_\Gamma(\mathbb{Z}^*)\}.$$

Then, we define naturally its subsets $\mathscr{F}_{r,\omega}, \mathscr{F}_{r,\Omega}, \mathscr{F}^*_{r,\omega}$ and $\mathscr{F}^*_{r,\Omega}$.

**Remark 4.6.1.** *Note that all polynomials of the form* (4.33) *can be written under the form* $Q_\Gamma[c]$ *for some* $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n)$ *with* $\mathbf{k} = \mathbf{h} = \varnothing$ *and the convention* $n = 0$*. More precisely, if* $P$ *is a polynomial of order* $2m$*, then it can be written under the previous form, with* $\Gamma \in \mathscr{H}_{m,\omega} \subset \mathscr{H}_{m,\Omega}$*.*

**Remark 4.6.2.** *The uniform bound on the numerator in condition* (4.62) *can be specified on the subclasses. More precisely, using* (4.61) *we deduce that if* $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n)$ *belongs to* $\mathscr{H}_{r,\Omega}$ *or* $\mathscr{H}^*_{r-2,\Omega}$ *and* $r \geqslant 4$ *then*

$$m_\ell \leqslant 7r - 22. \tag{4.70}$$

*Similarly, if* $\Gamma$ *belongs to* $\mathscr{H}_{r,\omega}$ *or* $\mathscr{H}^*_{r-1,\omega}$ *and* $r \geqslant 3$ *then*

$$m_\ell \leqslant 3r - 6. \tag{4.71}$$

### 4.6.2 Structural lemmas

In this section we verify that our class allows to define flows and that this class is stable by resolution of homological equations and by Poisson bracket.

#### Control of the vector fields

First, we have to verify that the vector field associated with Hamiltonian belonging to the class defined above are under control in $\ell^1_s$ in such way it defines a regular flow. In other words we would like to prove that such Hamiltonian are regular in the sense of Definition 4.3.2. Actually, we will control the vector field of Hamiltonian of the form $Q_\Gamma[c]$ for which $\mathcal{N}_\Gamma(c) \leqslant N^2$ for a given $N$, a property that is stable by Poisson bracket and solution of homological equation according to Lemmas 4.6.4 and 4.6.6.

**Lemma 4.6.3.** *Let $r \geqslant 2$, $\alpha_r = 24r$, and $s$ be given. For all $\Gamma \in \mathscr{H}_{r,\omega}, \mathscr{H}^*_{r,\omega}, \mathscr{H}_{r,\Omega}$ or $\mathscr{H}^*_{r,\Omega}$ there exists a constant $C > 0$ such that for all $\varepsilon, \gamma < 1$, all $c \in \ell^\infty_\Gamma(\mathbb{Z}^*)$ and all $N \geqslant 1$ such that $\mathcal{N}_\Gamma(c) \leqslant N^2$, then $Q_\Gamma[c] \in \mathcal{C}^\infty(\mathcal{U}^N_{\gamma,\varepsilon,r,s})$ is a regular Hamiltonian in the sense of Definition 4.3.2 and for all $z \in B_s(0, 4\varepsilon) \cap \mathcal{U}^N_{\gamma,\varepsilon,r,s}$*

$$\|X_{Q_\Gamma[c]}(z)\|_s \leqslant C\varepsilon^{2r-1}\|c\|_{\ell^\infty}\left(\frac{N^{\alpha_r}}{\gamma}\right)^{\beta_r}. \tag{4.72}$$

*with*

$$\beta_r = \begin{cases} 2r-5 & \textit{for} \quad \Gamma \in \mathscr{H}_{r,\omega}, \\ 2r-2 & \textit{for} \quad \Gamma \in \mathscr{H}^*_{r,\omega}, \end{cases} \quad \textit{and} \quad \beta_r = \begin{cases} 4r-13 & \textit{for} \quad \Gamma \in \mathscr{H}_{r,\Omega}, \\ 4r-5 & \textit{for} \quad \Gamma \in \mathscr{H}^*_{r,\Omega}. \end{cases}$$

*Proof.* Let $\rho = \sup_{\ell\in\mathbb{N}} m_\ell$. We have seen in (4.70) and (4.71) that this quantity is bounded by $7r$. The functional $Q_\Gamma[c]$ can be written under the form

$$Q_\Gamma[c](z) = \sum_{m=1}^{\rho}\sum_{\boldsymbol{j}\in\mathcal{R}_m} f_{\boldsymbol{j},m}(I) z_{\boldsymbol{j}},$$

where the coefficients $f_{\boldsymbol{j},m}(I)$, which depend on $\Gamma$, are given by (4.60).
Let $j_0 \in \mathbb{U}_2 \times \mathbb{Z}$ be fixed, the component of the vector field $(X_{Q_\Gamma[c]})_{j_0}(z)$ is given by

$$(X_{Q_\Gamma[c]})_{j_0}(z) = \frac{\partial}{\partial z_{\overline{j_0}}} Q_\Gamma[c](z) = \sum_{m=1}^{\rho}\sum_{\boldsymbol{j}\in\mathcal{R}_m} f_{\boldsymbol{j},m}(I)\frac{\partial}{\partial z_{\overline{j_0}}}(z_{\boldsymbol{j}}) + z_{\boldsymbol{j}}\frac{\partial}{\partial z_{\overline{j_0}}}(f_{\boldsymbol{j},m}(I)) \tag{4.73}$$

Let us examine the contributions coming from the first type of terms in the right-hand side.

Let $\ell \in \boldsymbol{\pi}^{-1}(\boldsymbol{j})$ with $m_\ell = m$ be given, and $p_\ell$, $q_\ell$ and $n_\ell$ the integers associated with one term in the decomposition (4.60). To control the denominators, as $z \in \mathcal{U}^N_{\gamma,\varepsilon,r,s}$ we will use the estimates (4.46) and (4.47). More precisely, as $\sharp\boldsymbol{k}_{\ell,\alpha} \leqslant m_\alpha \leqslant 7r$ (see (4.63)), we have

$$|\omega_{\boldsymbol{k}_{\ell,\alpha}}(I)| > \gamma\varepsilon^2 N^{-\alpha_r}\langle\mu_{\min}(\boldsymbol{k}_{\ell,\alpha})\rangle^{-2s}$$

by definition of the weight and using the fact that $\langle\mu_1(\boldsymbol{k}_{\ell,\alpha})\rangle \leqslant \mathcal{N}_\Gamma(c) \leqslant N^2$. Similarly, we will use

$$|\Omega_{\boldsymbol{k}_{\ell,\alpha}}(I)| > \gamma\varepsilon^2 N^{-\alpha_r}\langle\mu_{\min}(\boldsymbol{k}_{\ell,\alpha})\rangle^{-2s}$$

and

$$|\Omega_{\boldsymbol{h}_{\ell,\alpha}}(I)| > \gamma\varepsilon^4 N^{-\alpha_r}.$$

After using these bounds, we can conclude that for $z \in \mathcal{U}^N_{\gamma,\varepsilon,r,s}$, there exists $C$ depending only on $r$ and $s$ such that

$$\begin{aligned} |f_{\boldsymbol{j},m}(I)| &\leqslant C\|c\|_{\ell^\infty}\sum_{\substack{\ell\in\boldsymbol{\pi}^{-1}(\boldsymbol{j})\\ m_\ell=m}}\frac{N^{\alpha_r(p_\ell+q_\ell)}}{\gamma^{p_\ell+q_\ell}\varepsilon^{2p_\ell+4q_\ell}}\prod_{\alpha=1}^{p_\ell}\langle\mu_{\min}(\boldsymbol{k}_{\ell,\alpha})\rangle^{2s} \\ &\leqslant C\|c\|_{\ell^\infty}\left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r}\varepsilon^{2r-2m}\prod_{\alpha=1}^{p_\ell}\langle\mu_{\min}(\boldsymbol{k}_{\ell,\alpha})\rangle^{2s} \end{aligned}$$

where we verify that $b_r = 2r-6$ for $\Gamma \in \mathscr{H}_{r,\omega}$, and $b_r = 2r-3$ for $\mathscr{H}^*_{r,\omega}$, $b_r = 2r-14$ for $\Gamma \in \mathscr{H}_{r,\Omega}$ and $b_r = 4r-6$ for $\Gamma \in \mathscr{H}^*_{r,\Omega}$. Indeed, we used that in all those cases we always have $p_\ell + 2q_\ell = m_\ell - r$ by using (4.61). In the other hand if $\Gamma \in \mathscr{H}_{r,\omega}$ or $\mathscr{H}^*_{r,\omega}$, we have $q_\ell = 0$ and $p_\ell \leqslant 2r-6$ or $2r-3$ for $\mathscr{H}_{r,\omega}$ and $\mathscr{H}^*_{r,\omega}$ respectively. Hence the value of $b_r$ in these both cases. Now if $\Gamma \in \mathscr{H}_{r,\Omega}$, we have with the notations (4.68)-(4.69), $p_\ell + q_\ell \leqslant \alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5 \leqslant \alpha_1 + 2\alpha_5$, infering the value of $b_r$. The case or $\Gamma \in \mathscr{H}^*_{r,\Omega}$ is treated similarly.
Up to a combinatorial factor, we can assume that $j_1 = \overline{j_0}$, and hence $\partial_{z_{\overline{j_0}}}(z_{\boldsymbol{j}}) = z_{j_2}\cdots z_{j_{2m}}$, and moreover we can also assume that $j_2$ is the largest index amongst $(j_2,\dots,j_{2m})$. Hence $j_2 = \mu_1(\boldsymbol{j})$ or $j_2 = \mu_2(\boldsymbol{j})$ depending if $j_1 = \mu_1(\boldsymbol{j})$ or not. Furthermore using our Hypothesis (vi) on the repartition of the derivatives (see (4.64)) we have

$$\prod_{\alpha=1}^{p_\ell} \langle \mu_{\min}(\boldsymbol{k}_{\ell,\alpha})\rangle^{2s} \leqslant \prod_{\alpha=3}^{m} \langle \mu_\alpha(\boldsymbol{j})\rangle^{s}. \tag{4.74}$$

With these choices and this estimate, we get

$$\begin{aligned}
&\sum_{j_0\in\mathbb{U}_2\times\mathbb{Z}} \langle j_0\rangle^s \left| \sum_{\boldsymbol{j}\in\mathcal{R}_m} f_{\boldsymbol{j},m}(I) \frac{\partial}{\partial z_{\overline{j_0}}}(z_{\boldsymbol{j}}) \right| \\
&\qquad \leqslant C\|c\|_{\ell^\infty} \left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r} \varepsilon^{2r-2m} \sum_{\boldsymbol{j}=(j_1,\dots j_{2m})\in\mathcal{R}_m} \prod_{\alpha=3}^{m} \langle \mu_\alpha(\boldsymbol{j})\rangle^s \langle j_1\rangle^s |z_{j_2}\cdots z_{j_{2m}}| \\
&\qquad\qquad \leqslant C\|c\|_{\ell^\infty} \left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r} \varepsilon^{2r-2m} \sum_{\boldsymbol{j}=(j_1,\dots j_{2m})\in\mathcal{R}_m} \langle j_1\rangle^s |z_{j_2}||v_{j_3}|\cdots|v_{j_{2m}}|
\end{aligned}$$

where $v_k = \langle k\rangle z_k$ is in $\ell^1$ and of norm smaller than $\varepsilon$ by assumption. Since $\boldsymbol{j}\in\mathcal{R}_m$ it satisfies the zero-momentum condition and thus $\langle j_1\rangle \leqslant (2m-1)\langle j_2\rangle$ . Hence the last sum is bounded by

$$\sum_{(j_2,\dots j_{2m})} |v_{j_2}||v_{j_3}|\cdots|v_{j_{2m}}| \leqslant C\varepsilon^{2m-1}.$$

By summing with respect to $m$, we get that the first contribution of the right-hand side of (4.73) for the estimate of $\|X_{Q_\Gamma[c]}(z)\|_s = \sum_j \langle j\rangle^s |(X_{Q_\Gamma[c]})_j(z)|$ satisfies the bound (4.72).

Now we study the second contribution in the equation (4.73). To this aim, let us write

$$f_{\boldsymbol{j},m}(I) = \sum_{\substack{\ell\in\boldsymbol{\pi}^{-1}(\boldsymbol{j})\\ m_\ell = m}} c_\ell f^\ell_{\boldsymbol{j},m}(I)$$

where $f^\ell_{\boldsymbol{j},m}(I)$ correspond to the decomposition (4.60). In view of the structure of $f^\ell_{\boldsymbol{j},m}(I)$, we have

$$z_{\boldsymbol{j}} \frac{\partial}{\partial z_{\overline{j_0}}}(f^\ell_{\boldsymbol{j},m}(I)) = -z_{\boldsymbol{j}} z_{j_0} f^\ell_{\boldsymbol{j},m}(I) \left( \sum_{\alpha=1}^{n_\ell} \frac{\partial_{I_j}\omega_{\boldsymbol{k}_{\ell,\alpha}}}{\omega_{\boldsymbol{k}_{\ell,\alpha}}} + \sum_{\alpha=n_\ell+1}^{p_\ell} \frac{\partial_{I_j}\Omega_{\boldsymbol{k}_{\ell,\alpha}}}{\Omega_{\boldsymbol{k}_{\ell,\alpha}}} + \sum_{\beta=n_1}^{q_\ell} \frac{\partial_{I_j}\Omega_{\boldsymbol{h}_{\ell,\beta}}}{\Omega_{\boldsymbol{h}_{\ell,\beta}}} \right). \tag{4.75}$$

Let us assume that $j_1 = \mu_1(\boldsymbol{j})$ and $j_2 = \mu_2(\boldsymbol{j})$. We have with the previous notation and using again (4.74)

$$\langle j_0\rangle^s |z_{\boldsymbol{j}} z_{j_0} f^\ell_{\boldsymbol{j},m}(I)| \leqslant C\|c\|_{\ell^\infty} \left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r} \varepsilon^{2r-2m} |z_{j_1}||z_{j_2}||v_j||v_{j_3}|\cdots|v_{j_{2m}}|.$$

Now as $\partial_{I_j}\omega_{\boldsymbol{k}_{\ell,\alpha}} = \pm 1$, we have by using (4.64) and the fact that $n_\ell \leqslant m - r$,

$$\left|\frac{\partial_{I_j}\omega_{\boldsymbol{k}_{\ell,\alpha}}}{\omega_{\boldsymbol{k}_{\ell,\alpha}}}\right| \leqslant \frac{C}{\gamma\varepsilon^2} N^{\alpha_r}\langle\mu_{\min}(\boldsymbol{k}_{\ell,\alpha})\rangle^{2s} \leqslant \frac{C}{\gamma\varepsilon^2} N^{\alpha_r}\langle\mu_3(\boldsymbol{j})\rangle^{2s},$$

and the contribution corresponding to this term in the expression

$$\sum_{j_0}\langle j_0\rangle^s \left|\sum_{\boldsymbol{j}\in\mathcal{R}_m} z_{\boldsymbol{j}}\frac{\partial}{\partial z_{\bar{j}_0}}(f_{\boldsymbol{j},m}(I))\right| \leqslant C\sum_{j_0}\sum_{\substack{\ell\in\boldsymbol{\pi}^{-1}(\boldsymbol{j})\\ m_\ell = m}}\langle j_0\rangle^s \left|z_{\boldsymbol{j}}\frac{\partial}{\partial z_{\bar{j}_0}}(f^\ell_{\boldsymbol{j},m}(I))\right|$$

is thus bounded by

$$C\|c\|_{\ell^\infty}\left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r+1}\varepsilon^{2r-2m-2}\sum_{j_0,\boldsymbol{j}}\langle\mu_3(\boldsymbol{j})\rangle^{2s}|z_{j_1}||z_{j_2}||v_{j_0}||v_{j_3}|\cdots|v_{j_{2m}}|$$
$$\leqslant C\|c\|_{\ell^\infty}\left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r+1}\varepsilon^{2r-1},$$

as $j_1$ and $j_2$ are larger than the third largest index in $\boldsymbol{j}$. By summing with respect to $m$, the global contribution of these terms satisfies the estimate (4.72).

We obtain similar estimates for the terms in (4.75) associated with the part of $\Omega_{\boldsymbol{k}_{\ell,\alpha}}$ and $\Omega_{\boldsymbol{h}_{\ell,\alpha}}$ coming from $Z_4$. It remains to estimates the part coming from $Z_6$ in (4.75). Typically a term of the form $\frac{\partial_{I_j}\Omega_{\boldsymbol{k}_{\ell,\alpha}}}{\Omega_{\boldsymbol{k}_{\ell,\alpha}}}$ will yield a contribution of the form

$$\sum_p \alpha_p \frac{I_p}{\Omega_{\boldsymbol{k}_{\ell,\beta}}}$$

where $\alpha_p$ are uniformly bounded in $p$. The global contribution of these term, by estimating $\Omega_{\boldsymbol{k}_{\ell,\alpha}}$ and $\Omega_{\boldsymbol{h}_{\ell,\beta}}$ by $\gamma\varepsilon^4 N^{\alpha_r}$ will be

$$C\|c\|_{\ell^\infty}\left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r+1}\varepsilon^{2r-2m-4}\sum_{j_0,p,\boldsymbol{j}}|z_{j_1}||z_{j_2}||v_{j_0}||v_{j_3}|\cdots|v_{j_{2m}}||z_p|^2 \leqslant C\|c\|_{\ell^\infty}\left(\frac{N^{\alpha_r}}{\gamma}\right)^{b_r+1}\varepsilon^{2r-1}.$$

This shows the result with $\beta_r = b_r + 1$. □

### Homological equations

In this section we will see that our class is particularly well adapted to the solution homological equations, the central step in the construction of normal forms. Actually, this class was constructed precisely to be invariant by Poisson bracket and by solution of the homological equation with $Z_4$ or $Z_4 + Z_6$.

We define the set $\mathscr{A}_r$ as the subset of elements $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n)$ of $\mathscr{H}_r$ for which $Q_\Gamma[c]$ depends only on the actions. This means that for all $\ell \in \mathbb{Z}^*$, $\mathrm{Irr}(\boldsymbol{\pi}_\ell) = \varnothing$.

Then we define $\mathscr{R}_r$ the subset of elements $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n)$ of $\mathscr{H}_r$ such that for all $\ell \in \mathbb{Z}^*$,

$\mathrm{Irr}(\boldsymbol{\pi}_\ell) \neq \varnothing$. Note that for all $\Gamma \in \mathscr{H}_r$ there exists $A \in \mathscr{A}_r$ and $R \in \mathscr{R}_r$ such that for all $c \in \ell^\infty_\Gamma$, $c \in \ell^\infty_A \cap \ell^\infty_R$, and

$$Q_\Gamma[c] = Q_A[c] + Q_R[c]. \tag{4.76}$$

We also naturally define the corresponding subsets of $\mathscr{F}_r$

$$\mathscr{F}_r^A := \{F = Q_\Gamma[c], \quad \Gamma \in \mathscr{A}_r, \quad c \in \ell^\infty_\Gamma(\mathbb{Z}^*)\}. \tag{4.77}$$

the functionals of order $r$ depending only on the actions, and

$$\mathscr{F}_r^R := \{F = Q_\Gamma[c], \quad \Gamma \in \mathscr{R}_r, \quad c \in \ell^\infty_\Gamma(\mathbb{Z}^*)\}. \tag{4.78}$$

Naturally, we define $\mathscr{R}_{r,\omega}, \mathscr{R}_{r,\Omega}, \mathscr{R}^*_{r,\omega}, \mathscr{R}^*_{r,\Omega}$ as the restrictions of $\mathscr{R}_r$ to $\mathscr{H}_{r,\omega}$, $\mathscr{H}_{r,\Omega}$, $\mathscr{H}^*_{r,\omega}$, $\mathscr{H}^*_{r,\Omega}$.

With this formalism, the resolution of the homological equation is trivial, after noticing that $Z_4$ and $Z_6$ commute with terms depending only of the actions and by using the relations (4.38).

**Lemma 4.6.4.** *Let $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{R}_{r,\Omega}$. Defining $\Gamma' = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}', n) \in \mathscr{H}^*_{r-2,\Omega}$ with*

$$\forall \ell \in \mathbb{Z}^*, \quad \mathbf{h}'_\ell = (\mathbf{h}_\ell, \mathrm{Irr}(\boldsymbol{\pi}_\ell)),$$

*Then for all $c \in \ell^\infty_\Gamma(\mathbb{Z}^*)$, $Q'_\Gamma[c]$ is solution of the homological equation*

$$\{Z_4 + Z_6, Q_{\Gamma'}[c]\} = Q_\Gamma[c], \tag{4.79}$$

*and we have*

$$\mathcal{N}_{\Gamma'}(c) = \mathcal{N}_\Gamma(c).$$

We will also need to solve a homological equation associated with $Z_4$ :

**Lemma 4.6.5.** *Let $\Gamma = (\boldsymbol{\pi}, \varnothing, \varnothing, 0) \in \mathscr{R}_{3,\omega}$. Defining $\Gamma' = (\boldsymbol{\pi}, \mathbf{k}', \varnothing, n') \in \mathscr{H}^*_{2,\omega}$ with*

$$\forall \ell \in \mathbb{Z}^*, \quad \mathbf{k}'_\ell = (\boldsymbol{\pi}_\ell) \text{ and } n'_\ell = 1,$$

*Then for all $c \in \ell^\infty_\Gamma(\mathbb{Z}^*)$, $Q'_\Gamma[c]$ is solution of the homological equation*

$$\{Z_4, Q_{\Gamma'}[c]\} = Q_\Gamma[c],$$

*and we have*

$$\mathcal{N}_{\Gamma'}(c) = \mathcal{N}_\Gamma(c).$$

### Stability by Poisson bracket

Now comes the main technical result of this paper : the stability of our classes by Poisson bracket.

**Lemma 4.6.6.** *Let $W \in \{\omega, \Omega\}$, let $\Gamma \in \mathscr{H}^*_{r,W}$ and let $\Gamma' \in \mathscr{H}_{r',W}$.*
*There exists $\Gamma'' \in \mathscr{H}_{r'',W}$, where*

$$r'' = r + r' - 1$$

*and there exists a bilinear continuous application*

$$g : \ell^\infty_\Gamma(\mathbb{Z}^*) \times \ell^\infty_{\Gamma'}(\mathbb{Z}^*) \to \ell^\infty_{\Gamma''}(\mathbb{Z}^*)$$

*such that for all* $c \in \ell^\infty_\Gamma(\mathbb{Z}^*)$, $c' \in \ell^\infty_{\Gamma'}(\mathbb{Z}^*)$

$$\{Q_\Gamma[c], Q_{\Gamma'}[c']\} = Q_{\Gamma''}[g(c,c')]$$

*and*

$$\mathcal{N}_{\Gamma''}(g(c,c')) \leqslant \max(\mathcal{N}_\Gamma(c), \mathcal{N}_{\Gamma'}(c')). \tag{4.80}$$

*Proof.* We postpone the proof to appendix 4.9.2 □

## 4.7 Rational normal form

In this section we prove Theorem 4.2.1 for (NLS). As announced in section 4.2.2 this is achieved in three steps : First we kill the non resonant monomials in the Hamiltonian $P$ by using $Z_2$ as normal form (Section 4.7.1), then we kill the remaining non integrable terms ($K_6$) of order 6 by including the resonant part of order 4, namely $Z_4$ (which is integrable), in the normal form (Section 4.7.2), finally we kill all the non integrable terms up to order $r$ by including the integrable part of order 6, namely $Z_6$, in the normal form (Section 4.7.3).

### 4.7.1 Resonant normal form

In this section we apply a Birkhoff normal form procedure to kill iteratively the non resonant monomials up to order $r$ of the Hamiltonian $P$.

**Theorem 4.7.1.** *For all* $r \geqslant 4$ *and* $s \geqslant 0$*, there exists* $\tau_2$ *a* $\mathcal{C}^1$ *symplectomorphism in a neighborhood of the origin in* $\ell^s_1$ *close to the identity :*

$$\|\tau_2(z) - z\|_s \leqslant C\|z\|_s^3$$

*which puts* $H$ *in normal form up to order 6 :*

$$H \circ \tau_2(z) = Z_2(I) + Z_4(I) + Z_6(I) + K_6(z) + \sum_{m=4}^{r} K_{2m}(z) + R(z) \tag{4.81}$$

*where for all* $m = 4, \dots, r$, $K_{2m}$ *is a homogeneous resonant polynomial of order* $m$

$$K_{2m}(z) = P[c^{(m)}](z) = \sum_{\boldsymbol{j} \in \mathcal{R}_m} c^{(m)}_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad \textit{with} \quad c^{(m)} \in \ell^\infty(\mathcal{R}_m), \tag{4.82}$$

*and where* $K_6(z)$ *contains only irreducible monomials*

$$K_6(z) = \sum_{\boldsymbol{j} \in \mathcal{R}_3 \cap \mathcal{I}rr} c_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad \textit{with} \quad c \in \ell^\infty(\mathcal{R}_3). \tag{4.83}$$

*Moreover,* $R$ *is smooth in a neighborhood of the origin and satisfies*

$$\|X_R(z)\|_s \leqslant C\|z\|_s^{2r+1}, \tag{4.84}$$

*for* $z$ *small enough in* $\ell^1_s$.

*Proof.* The proof is standard (it first appears in [89]) except for the calculation of the resonant terms of order six. For convenience of the reader we give the details.
We have $H = \sum_{a\in\mathbb{Z}} a^2\xi_a\eta_a + P$ where $P$ is given by (4.28) and we write

$$P = \sum_{m=1}^{r} P_{2m} + R_{2r+2}$$

where

$$\begin{aligned} P_{2m} &= \frac{\varphi^{(m-1)}(0)}{m!}\frac{1}{2\pi}\int_{\mathbb{T}}(\sum_{a\in\mathbb{Z}}\xi_a e^{iax})(\sum_{b\in\mathbb{Z}}\eta_b e^{-ibx})dx \\ &= \frac{\varphi^{(m-1)}(0)}{m!}\sum_{a_1+\cdots+a_m=b_1+\cdots b_m}\xi_{a_1}\cdots\xi_{a_m}\eta_{b_1}\cdots\eta_{b_m} \\ &= \frac{m!\,\varphi^{(m-1)}(0)}{(2m)!}\sum_{\boldsymbol{j}\in\mathcal{M}_m} z_{\boldsymbol{j}} \end{aligned} \tag{4.85}$$

and $R_{2r+2}$ is a remainder of order $2r+2$ *i.e.* $R_{2r+2}\in\mathcal{H}_s(\ell^1_s)$ and $\|X_{R_{2r+2}}(z)\|_s \leqslant C\|z\|_s^{2r+1}$. We note that the integrable Hamiltonian given by (4.17) reads

$$Z_2 = \sum_{a\in\mathbb{Z}} a^2\xi_a\eta_a + P_2.$$

First we kill the non resonant monomials of order 4 by a change of variables $\Psi_4$. We search for $\Psi_4 = \Phi^1_{\chi_4}$, the time one flow of $\chi_4$ of a polynomial Hamiltonian homogeneous of order 4 :

$$\chi_4 = \sum_{j\in\mathcal{M}_2} a_{\boldsymbol{j}}\, z_{\boldsymbol{j}}.$$

For any $F\in\mathcal{H}_s$, the Taylor expansion of $F\circ\Phi^t_\chi$ between $t=0$ and $t=1$ gives

$$F\circ\Phi^1_\chi = F + \{F,\chi\} + \frac{1}{2}\int_0^1 (1-t)\{\{F,\chi\},\chi\}\circ\Phi^t_\chi \mathrm{d}t.$$

Applying this formula to $H = Z_2 + P_4 + R_6$ we get

$$H\circ\Psi_4 = Z_2 + (P-P_2) + \{Z_2,\chi\} + \{(P-P_2),\chi\} + \frac{1}{2}\int_0^1 (1-t)\{\{H,\chi\},\chi\}\circ\Phi^t_\chi \mathrm{d}t.$$

In this formule the homogeneous part of order 4 is $P_4 + \{Z_2,\chi_4\}$. Then we set

$$\chi_4 := \frac{1}{12}\varphi'(0)\sum_{\boldsymbol{j}\in\mathcal{M}_2\setminus\mathcal{R}_2}\frac{1}{i\Delta_{\boldsymbol{j}}}z_{\boldsymbol{j}}, \quad Z_4 = \frac{1}{12}\varphi'(0)\sum_{\boldsymbol{j}\in\mathcal{R}_2} z_{\boldsymbol{j}}.$$

We note that at this stage there are no small divisors problem since $|\Delta_{\boldsymbol{j}}|\geqslant 1$ except when $\boldsymbol{j}\in\mathcal{R}$ in which case $\Delta_{\boldsymbol{j}}=0$. So $\chi_4$ and $Z_4$ are well defined homogeneous polynomials of order 4 and, using (4.39) they solve the homological equation

$$Z_4 = P_4 + \{Z_2,\chi_4\}. \tag{4.86}$$

Further $H \circ \Psi_4 = Z_2 + Z_4 + R_6$ with

$$R_6 = (P - P_2 - P_4) + \{P - P_2, \chi_4\} + \frac{1}{2}\int_0^1 (1-t)\{\{H,\chi\},\chi\} \circ \Phi_\chi^t \mathrm{d}t \tag{4.87}$$

is a smooth Hamiltonian beginning at order 6 *i.e.* $R_6 = 0(z^6)$.
We can iterate this procedure to kill successively the non resonant monomials of order $6, \cdots, 2r$. Then we get the existence of a symplectomorphism $\Psi$ close to the identity and defined on a neighborhood of the origin in $\ell_1^s$ such that

$$H \circ \Psi = Z_2 + Z_4 + Z_6' + \sum_{m=4}^{r} K_{2m}(z) + R(z), \tag{4.88}$$

where $K_{2m}$ are resonant polynomials of the form (4.82), $R$ is a smooth remainder satisfying (4.84) on a neighborhood of the origin and $Z_6'$ is a resonant monomial of order 6. It remains to compute $Z_4$ and $Z_6'$.
Concerning $Z_4$ we have

$$Z_4 = \frac{1}{2}\varphi'(0) \sum_{\substack{a_1+a_2=b_1+b_2 \\ a_1^2+a_2^2=b_1^2+b_2^2}} \xi_{a_1}\xi_{a_2}\eta_{b_1}\eta_{b_2}$$

but

$$a_1 + a_2 = b_1 + b_2 \quad \text{and} \quad a_1^2 + a_2^2 = b_1^2 + b_2^2$$

leads to

$$\{a_1, a_2\} = \{b_1, b_2\}.$$

Therefore we get as anounced in (4.36)

$$\begin{aligned} Z_4 = Z_4(I) =& \frac{1}{2}\varphi'(0) \sum_{a,b\in\mathbb{Z}} I_a I_b (2 - \delta_{ab}) \\ =& \varphi'(0)(\sum_{a\in\mathbb{Z}} I_a)^2 - \frac{1}{2}\varphi'(0) \sum_{a\in\mathbb{Z}} I_a^2. \end{aligned}$$

After the first two Birkhoff procedures we get[5] $Z_6' = Z_{6,1} + Z_{6,2}$ where $Z_{6,1}$ is the resonant part of $\{P_4, \chi_4\} + \frac{1}{2}\{\{Z_2, \chi_4\}, \chi_4\}$ and $Z_{6,2}$ is the resonant part of $P_6$.
Let us start with the latter, following (4.85) we have

$$Z_{6,2} = \frac{\varphi''(0)}{6} \sum_{\substack{a_1+a_2+a_3=b_1+b_2+b_3 \\ a_1^2+a_2^2+a_3^2=b_1^2+b_2^2+b_3^2}} \xi_{a_1}\xi_{a_2}\xi_{a_3}\eta_{b_1}\eta_{b_2}\eta_{b_3}.$$

If $\{a_1, a_2, a_3\} \cap \{b_1, b_2, b_3\} \neq \emptyset$ then, assuming for instance $a_3 = b_3$, we get

$$((1, a_1), (1, a_2), (-1, b_1), (-1, b_2)) \in \mathcal{R}_2$$

5. Recall that the Poisson bracket of a Polynomial of order $m$ with a polynomial of order $m$ is a polynomial of order $m + n - 2$.

which leads as before to $\{a_1, a_2\} = \{b_1, b_2\}$. So either

$$((1, a_1), (1, a_2), (1, a_3), (-1, b_1), (-1, b_2), (1, b_3)) \in \mathcal{I}rr \text{ or } \{a_1, a_2, a_3\} = \{b_1, b_2, b_3\}$$

, i.e.

$$\begin{aligned} Z_{6,2} =& K_6'(z) + \frac{\varphi''(0)}{6} \sum_{\{a_1,a_2,a_3\}=\{b_1,b_2,b_3\}} \xi_{a_1}\xi_{a_2}\xi_{a_3}\eta_{b_1}\eta_{b_2}\eta_{b_3} \\ =& K_6(z) + \frac{\varphi''(0)}{6} \left( \sum_{\substack{a\neq b,\ a\neq c \\ b\neq c}} 6 I_a I_b I_c + \sum_{a\neq b} 9 I_a^2 I_b + \sum_a I_a^3 \right) \\ =& K_6'(z) + \frac{\varphi''(0)}{6} \left( 6(\sum_{a\in\mathbb{Z}} I_a)^3 - 9(\sum_{a\in\mathbb{Z}} I_a^2)(\sum_{a\in\mathbb{Z}} I_a) + 4\sum_{a\in\mathbb{Z}} I_a^3 \right) \end{aligned} \tag{4.89}$$

where $K_6'$ is of the form (4.83).
It remains to compute $Z_{6,1}$. First we notice that using the homological equation (4.86) we get

$$\{P_4, \chi_4\} + \frac{1}{2}\{\{Z_2, \chi_4\}, \chi_4\} = \{P_4, \chi_4\} + \frac{1}{2}\{Z_4 - P_4, \chi_4\} = \{Z_4, \chi_4\} + \frac{1}{2}\{Q_4, \chi_4\}$$

where $Q_4$ denotes the non resonant part of $P_4$ :

$$Q_4 = P_4 - Z_4 = Z_4 = \frac{1}{12}\varphi'(0) \sum_{\boldsymbol{j}\in\mathcal{M}_2\setminus\mathcal{R}_2} z_{\boldsymbol{j}}.$$

We easily verify that the Poisson bracket of a resonant monomial with a non resonant monomial cannot be resonant. Therefore $Z_{6,1}$ is the resonant part of

$$\frac{1}{2}\{Q_4, \chi_4\} = \frac{\varphi'(0)^2}{288} \left\{ \sum_{\boldsymbol{j}\in\mathcal{M}_2\setminus\mathcal{R}_2} z_{\boldsymbol{j}}, \sum_{\boldsymbol{k}\in\mathcal{M}_2\setminus\mathcal{R}_2} \frac{1}{i\Delta_{\boldsymbol{k}}} z_{\boldsymbol{k}} \right\} \tag{4.90}$$

$$= \frac{\varphi'(0)^2}{8} \left\{ \sum_{\substack{a_1+a_2=b_1+b_2 \\ a_1^2+a_2^2\neq b_1^2+b_2^2}} \xi_{a_1}\xi_{a_2}\eta_{b_1}\eta_{b_2}, \sum_{\substack{a_1+a_2=b_1+b_2 \\ a_1^2+a_2^2\neq b_1^2+b_2^2}} \frac{\xi_{a_1}\xi_{a_2}\eta_{b_1}\eta_{b_2}}{i(a_1^2+a_2^2-b_1^2-b_2^2)} \right\}. \tag{4.91}$$

Then we proceed as for $Z_{6,2}$ to conclude that

$$Z_{6,1} = K_6'' + Z_6'(I)$$

where $K_6''$ is of the form[6] (4.83) and $Z_6(I)$ is the part of $\frac{1}{2}\{Q_4, \chi_4\}$ depending only on the actions. So we can write

$$Z_6'(I) = \sum_{a\in\mathbb{Z}} \alpha_a\, I_a^3 + \sum_{a\neq b\in\mathbb{Z}} \beta_{ab}\, I_a^2 I_b + \sum_{\substack{a\neq b,\ a\neq c \\ b\neq c}} \gamma_{abc}\, I_a I_b I_c$$

6. In fact a long but straightforward computation leads to $K_6'' = 0$ which means that, up to order 6, the Birkhoff normal form of the cubic NLS depends only on the actions. A sort of reminiscence of the complete integrability. Nevertheless this result is not needed in this paper and the calculation is long...

where the values of $\alpha_a$, $\beta_{ab}$, $\gamma_{abc}$ are compute in Lemma 4.7.2 below.
Thus we get

$$Z_{6,1} = K_6'' - \frac{1}{2}\varphi'(0)^2 \sum_{a\neq b\in\mathbb{Z}} \frac{1}{(a-b)^2}\, I_a^2 I_b$$

and using (4.89)

$$Z_6'(z) = Z_{6,1} + Z_{6,2} = K_6(z) + Z_6(I)$$

where $K_6 = K_6' + K_6''$ is of the form (4.83) and $Z_6$ is given by (4.37) as expected. □

**Lemma 4.7.2.** *The coefficients of the term $Z_6'(I)$ satisfy*

*(i)* $\alpha_a = 0$ *for all* $a \in \mathbb{Z}$,
*(i)* $\gamma_{abc} = 0$ *for all* $a \neq b$, $a \neq b$ *and* $b \neq c \in \mathbb{Z}$,
*(i)* $\beta_{ab} = \frac{-\varphi'(0)^2}{2(a-b)^2}$ *for all* $a \neq b \in \mathbb{Z}$.

*Proof.* We use formulas (4.90) and (4.91) to identify the terms of $\frac{1}{2}\{Q_4, \chi_4\}$ depending only on actions.
(i) If $I_a^3 = \frac{\partial z_{\boldsymbol{j}}}{\partial \xi_b}\frac{\partial z_{\boldsymbol{k}}}{\partial \eta_b}$ with $z_{\boldsymbol{j}}$ a monomial from $Q_4$ and $z_{\boldsymbol{k}}$ a monomial from $\chi_4$ then necessarily $z_{\boldsymbol{j}} = \xi_b\xi_a\eta_a^2$ and $z_{\boldsymbol{k}} = \xi_a^2\eta_a\eta_b$. But since $\boldsymbol{j}, \boldsymbol{k} \in \mathcal{M}_2$ we get $a = b$ and thus $\boldsymbol{j} = \boldsymbol{k}$ is resonant which is not possible.

(ii) Assume $I_aI_bI_c = \frac{\partial z_{\boldsymbol{j}}}{\partial \xi_d}\frac{\partial z_{\boldsymbol{k}}}{\partial \eta_d}$ with $\boldsymbol{j}, \boldsymbol{k} \in \mathcal{M}_2\backslash\mathcal{R}_2$. We consider two different cases :
• $z_{\boldsymbol{j}} = \xi_a\xi_d\eta_a\eta_b$ and $z_{\boldsymbol{k}} = \xi_c\xi_b\eta_d\eta_c$. Since $\boldsymbol{j} \in \mathcal{M}_2$ we get $b = d$ which is incompatible with $\boldsymbol{j} \notin \mathcal{R}_2$. All similar cases obtained by permutation of $a, b, c$ lead to the same incompatibility.
• $z_{\boldsymbol{j}} = \xi_a\xi_d\eta_c\eta_b$ and $z_{\boldsymbol{k}} = \xi_c\xi_b\eta_d\eta_a$. Since $\boldsymbol{j} \in \mathcal{M}_2$ we get $d = c + b - a$ and then we calculate $\Delta_k = -2(a-b)(a-c)$. By permutation we get up to an irrelevant constant $c$

$$c\gamma_{abc} = \frac{1}{(a-b)(a-c)} + \frac{1}{(b-a)(b-c)} + \frac{1}{(c-a)(c-b)} = 0.$$

(iii) Assume $I_a^2I_b = \frac{\partial z_{\boldsymbol{j}}}{\partial \xi_c}\frac{\partial z_{\boldsymbol{k}}}{\partial \eta_c}$ with $\boldsymbol{j}, \boldsymbol{k} \in \mathcal{M}_2\backslash\mathcal{R}_2$. We consider different cases :
• $z_{\boldsymbol{j}} = \xi_a\eta_a^2\xi_c$ and $z_{\boldsymbol{k}} = \xi_a\xi_b\eta_b\eta_c$. Since $\boldsymbol{j} \in \mathcal{M}_2$ we get $a = c$ which is incompatible with $\boldsymbol{j} \notin \mathcal{R}_2$.
• $z_{\boldsymbol{j}} = \xi_b\eta_a\eta_b\xi_c$ and $z_{\boldsymbol{k}} = \xi_a^2\eta_a\eta_c$. We get again using the zero momentum condition that $a = c$ which is incompatible with $\boldsymbol{j} \notin \mathcal{R}_2$.
• $z_{\boldsymbol{j}} = \xi_b\eta_a^2\xi_c$ and $z_{\boldsymbol{k}} = \xi_a^2\eta_b\eta_c$. The zero momentum leads to $c = 2a-b$ and we get $\Delta_{\boldsymbol{k}} = -\Delta_{\boldsymbol{j}} = 2a^2 - b^2 - (2a-b)^2 = -2(a-b)^2$. So $\boldsymbol{j}, \boldsymbol{k} \in \mathcal{M}_2\backslash\mathcal{R}_2$.
It remains to calculate the number of occurrences of this configuration in $D(z)$ : we can exchange $a$ and $b$ in $z_{\boldsymbol{j}}$ and in $z_{\boldsymbol{k}}$ and we can exchange $\boldsymbol{j}$ and $\boldsymbol{k}$. So 8 occurrences in (4.91) and thus

$$\beta_{ab} = \frac{\varphi'(0)^2}{8} 8 \frac{1}{-2(a-b)^2} = \frac{-\varphi'(0)^2}{2(a-b)^2}.$$

□

### 4.7.2 Elimination of the quintic term by the cubic

*It's mercy, compassion, and forgiveness I lack. Not rationality.* Beatrix Kiddo in "Kill Bill : Volume 1" (Q. Tarentino, 2003).

In this section we will truncate the new Hamiltonian $H \circ \tau_2$ and eliminate the resonant term $K_6$ with the help of $Z_4$. Moreover, we will show that the new Hamiltonian admits a development with terms of the form $Q_\Gamma[c]$ with $\Gamma$ in the class $\mathscr{H}_{r,\omega}$.

**Proposition 4.7.1.** *Let $r \geqslant 4$ and $s \geqslant 0$ be given. There exist $N_0$, $\Gamma_{2m} \in \mathscr{H}_{m,\omega}$ for $4 \leqslant m \leqslant r$, and a constant $C$ , such that for all $0 < \varepsilon \leqslant 1, \gamma > 0$, $N \geqslant 1$ with*

$$C\varepsilon^{\frac{3}{2}} \left(\frac{N^{\alpha_r}}{\gamma}\right)^3 < 1, \tag{4.92}$$

*there exist*

- $c_8, \dots, c_{2r} \in \ell^\infty_{\Gamma_8} \times \cdots \times \ell^\infty_{\Gamma_{2r}}$
- $\tau_4 : B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma/2,\varepsilon,r,s} \to \ell^1_s$ *a $\mathcal{C}^1$ injective symplectomorphism,*
- $R^{\text{high}}, R^{\text{ord}} \in \mathcal{C}^1(B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma/2,\varepsilon,r,s})$

*such that*

$$H \circ \tau_2 \circ \tau_4 = Z_2 + Z_4 + Z_6 + \sum_{m=4}^{r} Q_{\Gamma_{2m}}[c_{2m}] + R^{\text{high}} + R^{\text{ord}}, \tag{4.93}$$

*and we have the following bounds*

- *for all $m = 4 \dots r$, $\mathcal{N}_{\Gamma_{2m}}(c_{2m}) \leqslant N^2$*
- *for all $m = 4 \dots r$, $\|c_{2m}\|_{\ell^\infty} \leqslant C_r$*
- *For all $z \in B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma/2,\varepsilon,r,s}$, we have*

$$\|X_{R^{\text{high}}}(z)\|_{\ell^1_s} \leqslant C\varepsilon^5 N^{-s} \quad \textit{and} \quad \|X_{R^{\text{ord}}}(z)\|_{\ell^1_s} \leqslant C\varepsilon^{2r+1} \left(\frac{N^{\alpha_{r+1}}}{\gamma}\right)^{2r-3}$$

- *$\tau_4$ takes values in $z \in B_s(0, 3\varepsilon) \cap \mathcal{U}^N_{\gamma/4,\varepsilon,r,s}$ and satisfies the estimates*

$$\|\tau_4(z) - z\|_s \leqslant C\varepsilon^{\frac{3}{2}} \gamma N^{-\alpha_r}. \tag{4.94}$$

*Proof.* The proof is divided into two steps. First we introduce a cut-off in frequency allowing to work only with rational functionals whose irreducible monomials have their largest index bounded by $N^2$. Then we will define the change of variable $\tau_4$ and express the Hamiltonian in the new variable.

**First step : Truncation**. For all $4 \leqslant m \leqslant r$, we decompose $K_{2m}(z)$ of (4.82) into $K^N_{2m} + R^N_{2m}$, with

$$K^N_{2m}(z) = \sum_{\substack{\boldsymbol{j} \in \mathcal{R}_m \\ \langle \mu_3(\boldsymbol{j}) \rangle \leqslant \nu_m N}} b_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad \text{and} \quad R^N_{2m}(z) = \sum_{\substack{\boldsymbol{j} \in \mathcal{R}_m \\ \langle \mu_3(\boldsymbol{j}) \rangle > \nu_m N}} b_{\boldsymbol{j}} z_{\boldsymbol{j}},$$

where $\nu_m$ will be constant that will be specified later. Let $j_0 \in \mathbb{U}_2 \times \mathbb{Z}$ be a given index. Up to a combinatorial factor, we have

$$|X_{R_{2m}}(z)_{j_0}| \leqslant C_m \sum_{\substack{\boldsymbol{j} = (j_1, \dots, j_{2m}) \in \mathcal{R}_m \\ j_1 = \overline{j_0}}} |z_{j_2} \cdots z_{j_{2m}}|.$$

Let us assume that $j_2$ is the highest index in the monomial in the right-hand side, and $j_3$ the second highest. We thus have at least $\langle j_3\rangle > \nu_m N$. By using the zero momentum condition, we have

$$\|X_{R_{2m}}(z)\|_s \leqslant CN^{-s} \sum_{j_2,\dots,j_{2m}} \langle j_2\rangle^s |z_{j_2}| \langle j_3\rangle^s ||z_{j_3}\cdots z_{j_{2m}}|$$

where the constant $C$ depends on $m$ and $s$. It is thus easy to verify that when $\|z\|_s \leqslant \varepsilon$, we have

$$\|X_{R_{2m}}(z)\|_s \leqslant C\varepsilon^{2m-1}N^{-s}.$$

If we define $R^{\text{high}} = \sum_{m=3}^{r-1} R_{2m}$, we easily verify that it satisfies the hypothesis of the Proposition.

Now let us consider $K_{2m}^N$. By Remark 4.6.1, there exists $\Lambda_{2m} \in \mathscr{H}_{m,\omega}$ and $b_{2m} : \mathbb{Z}^* \to \mathbb{C}$ with $\|b\|_{\ell^\infty_{\Lambda_{2m}}} \leqslant C_m$ and such that

$$K_{2m}^N = Q_{\Lambda_{2m}}[b_{2m}].$$

Let us prove that, up to a choice of $\nu_m$, $\mathcal{N}_{\Lambda_{2m}}(b_{2m}) \leqslant N^2$,

For a given monomial $\boldsymbol{j} = (j_1,\dots,j_{2m}) = \boldsymbol{\pi}_\ell$ up to a combinatorial factor, let us assume that $j_1$ and $j_2$ correspond to the first and second largest indexes. We thus have $\langle j_p\rangle \leqslant \nu_m N$ for $p = 3,\dots,2m$. Let us denote by $j_p = (\delta_p, a_p) \in \mathbb{U}_2 \times \mathbb{Z}$. We have by definition of $\mathcal{R}_m$,

$$|\delta_1 a_1 + \delta_2 a_2| \leqslant (2m-2)\nu_m N \quad \text{and} \quad |\delta_1 a_1^2 + \delta_2 a_2^2| \leqslant (2m-2)\nu_m^2 N^2. \tag{4.95}$$

If $\delta_1$ and $\delta_2$ are of the same sign, this implies that $|a_1|$ and $|a_2|$ are smaller than $(2m-2)^{1/2}\nu_m N \leqslant N^2$ for $\nu_m$ small enough. If $\delta_1$ and $\delta_2$ are opposite signs, then two cases can occur.

- $a_1 = a_2$. In this case, $j_1 = \bar{j}_2$ and the product $z_{j_1}z_{j_2} = I_{a_1}$ is an action. In this situation, $\mathrm{Irr}(\boldsymbol{j}) \subset (j_3,\cdots,j_{2m})$ and hence $\langle \mu_1(\mathrm{Irr}(\boldsymbol{j}))\rangle \leqslant \langle j_3\rangle \leqslant N \leqslant N^2$.
- $a_1 = -a_2$. In this case, the first equation in (4.95) yields if $a_1 - a_2 \neq 0$, then we have

  $$|2a_1| \leqslant (2m-2)\nu_m N$$

  showing that $\langle j_1\rangle \leqslant N^2$ for $\nu_m$ small enough. As necessarily, $j_1 = \mu_1(\mathrm{Irr}(\boldsymbol{j}))$ we conclude that $\mathcal{N}_{\Lambda_{2m}}(b_{2m}) \leqslant N^2$.
- In any other situation, we have

  $$|a_1 + a_2||a_1 - a_2| \leqslant (2m-2)\nu_m^2 N^2$$

  with $|a_1 + a_2|$ and $|a_1 - a_2| \geqslant 1$. This shows that $|a_1 \pm a_2| \leqslant (2m-2)\nu_m^2 N^2$ and hence $\langle j_1\rangle$ and $\langle j_2\rangle$ smaller than $N^2$, for a good choice of $\nu_m$. We conclude as in the previous case that $\mathcal{N}_{\Lambda_{2m}}(b_{2m}) \leqslant N^2$.

**Second step : Construction of** $\tau_4$. As we have seen, $K_6^N$ can be written under the form $Q_{[\Lambda_6]}[b_6]$ for some $\Lambda_6 \in \mathscr{H}_{3,\omega}$ and weight $\mathcal{N}_{[\Lambda_6]}(b_6) \leqslant N^2$. Furthermore by Theorem 4.7.1, $K_6$ contains only irreducible monomials so $\Lambda_6 \in \mathscr{R}_{3,\omega}$. By using Lemma 4.6.5, there exists $\Lambda_6' \in \mathscr{H}_{2,\omega}^*$ such that $\chi := Q_{\Lambda_6'}[b_6]$ is solution of the homological equation

$$\{Z_4, \chi\} = \left\{Z_4, Q_{\Lambda_6'}[b_3]\right\} = Q_{\Lambda_6}[b_3] = K_6^N.$$

Moreover, $\mathcal{N}_{\Lambda_6'}(b_6) = \mathcal{N}_{\Lambda_6}(b_6) \leqslant N^2$. Hence by using (4.72), we immediately obtain the estimate

$$\|X_\chi(z)\|_s \leqslant C\varepsilon^3 \left(\frac{N^{\alpha_r}}{\gamma}\right)^2 \tag{4.96}$$

for $z \in B_s(0, 2\varepsilon) \cap \mathcal{U}_{\gamma/2,\varepsilon,r,s}$.

We then define $\tau_4 = \Phi_\chi^1$ the flow at time $1$ associated with the Hamiltonian $\chi$. Now let $z \in B_s(0, \varepsilon) \cap \mathcal{U}^N_{\gamma,\varepsilon,r,s}$. We have to prove that the flow at time $1$ of the Hamiltonian $\chi_4$ remain in the set $B_s(0, 3\varepsilon) \cap \mathcal{U}^N_{\gamma/4,\varepsilon,r,s}$. To prove this result, we use a bootstrap argument. Let us assume that this is the case.
By using (4.96), we easily obtain that

$$\|\Phi_\chi^1(z) - z\|_s \leqslant C\varepsilon^3 \left(\frac{N^{\alpha_r}}{\gamma}\right)^2 \leqslant \varepsilon^{3/2} \left(\frac{N^{\alpha_r}}{\gamma}\right)^{-1},$$

for some constant $C$ that we choose to be the one of assumption (4.92) . So we have in particular $C\varepsilon^3 \left(\frac{N^{\alpha_r}}{\gamma}\right)^2 \leqslant \varepsilon^{\frac{3}{2}}$, and hence $\|\tau_4(z)\|_s \leqslant 3\varepsilon$ provided $\varepsilon < 1$. Moreover, using Proposition 4.4.3, with $z' = \tau_4(z)$ and $\gamma' = \gamma/2$, we have

$$\|z - \tau_4(z)\|_s \leqslant \frac{1}{2} c\varepsilon N^{-\alpha_r}\gamma$$

where $c$ is the constant of Proposition 4.4.3. As a consequence we have $\tau_4(z) \in \mathcal{U}^N_{\varepsilon,\gamma/2,r,s}$ which concludes the bootstarp argument. Estimate (4.94) then easily follows. Note that $\tau_4$ is injective by definition of the flow.

Now we apply $\tau_4$ to (4.81), taking into account, $\sum_{m=3} K^r_{2m} = \sum_{m=3}^r K^N_{2m} + R^{\text{high}}$, we get

$$H \circ \tau_2 \circ \tau_4(z) = \big(Z_2 + Z_4(I) + Z_6 + K^N_6 + \sum_{m=4}^r K^N_{2m} + R^{\text{high}} + R\big) \circ \tau_4. \tag{4.97}$$

First we notice that $Z_2 \circ \tau_4 = Z_2$. Indeed, $\chi_4 = \sum_{\boldsymbol{j}\in\mathcal{R}_3} f_j(I) z_{\boldsymbol{j}}$, and $\boldsymbol{j}$ is a resonant monomial in $\mathcal{R}_3$ thus we have $\{Z_2(I), z_{\boldsymbol{j}}\} = \Delta_{\boldsymbol{j}} z_{\boldsymbol{j}} = 0$ as well as $\{Z_2(I), f_{\boldsymbol{j}}(I)\} = 0$. Hence $\{Z_2, \chi_4\} = 0$.
On the other hand we have, using the notation $\mathrm{ad}_\chi G = \{G, \chi\}$,

$$Z_4(I) \circ \tau_4 = Z_4 + \{Z_4, \chi\} + \sum_{\alpha=2}^M \frac{1}{\alpha!}\mathrm{ad}^\alpha_\chi Z_4 + \int_0^1 \frac{(t-s)^{M+1}}{(M)!}\mathrm{ad}^M_\chi Z_4 \circ \Phi^s_\chi(z)\mathrm{d}s,$$

and a similar formula for all the terms of (4.97), in particular

$$K^N_{2m} \circ \tau_4 = K^N_{2m} + \sum_{\alpha=1}^M \frac{1}{\alpha!}\mathrm{ad}^\alpha_\chi K^N_{2m} + \int_0^1 \frac{(t-s)^{M+1}}{(M)!}\mathrm{ad}^M_\chi K^N_{2m} \circ \Phi^s_\chi(z)\mathrm{d}s. \tag{4.98}$$

Note that by definition of $\chi$, the term $\{Z_4, \chi\}$ and $K^N_6$ cancel. The first three terms in the asymptotics are thus $Z_2 + Z_4 + Z_6$. Now let us look at the other terms generated. As $\chi := Q_{\Lambda_6'}[b_3]$ with $\Lambda_6' \in \mathscr{H}^*_{2,\omega}$, and as $Z_4(I) \in \mathscr{H}_{2,\omega}$, Lemma 4.6.6 shows that $\mathrm{ad}^\alpha_\chi Z_4 \in \mathscr{F}_{2+\alpha,\omega}$. Similarly, we have $\mathrm{ad}^\alpha_\chi K^N_{2m} \in \mathscr{F}_{m+\alpha,\omega}$ By collecting the elements of same degree, we obtain the claimed decomposition (4.93) where $R^{\text{ord}}$ is a sum of terms of order greater than $2r + 2$ and where by

a slight abuse of notation we still denote by $R^{\text{high}}$ its composition by $\tau_4$ (which is closed to the identity).
The estimates on the remainder are then consequences of the previous estimates on $\tau_4$, upon using the condition (4.92). □

**Remark 4.7.3.** *We have for* $z \in B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma/2,\varepsilon,r,s}$

$$\|z - \tau_2 \circ \tau_4(z)\|_s \leqslant \|z - \tau_2(z)\|_s + \|\tau_2(z) - \tau_2 \circ \tau_4(z)\|_s$$
$$\leqslant C\varepsilon^3 + C\|z - \tau_4(z)\|_s \leqslant C\varepsilon^3 \left(\frac{N^{\alpha_r}}{\gamma}\right)^2$$

*as* $\tau_2$ *is* $\mathcal{C}^1$ *in a neighborhood of the origin and up to some change of constant* $C$*. Hence by the same argument as in the proof, we have that the application* $\tau_2 \circ \tau_4$ *maps* $B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma/2,\varepsilon,r,s}$ *into* $B_s(0, 4\varepsilon) \cap \mathcal{U}^N_{\gamma/4,\varepsilon,r,s}$.

### 4.7.3 Quintic normal form

*You and I have unfinished business.* Beatrix Kiddo in "Kill Bill : Volume 2" (Q. Tarentino, 2004).

Recall that $\mathscr{F}^A_{r,\Omega}$ is the set of rational functions that depend only on the actions and can be written $Q_\Gamma$ with $\Gamma \in \mathscr{A}_{r,\Omega}$. By solving iteratively homological equations with the normal form term $Z_4 + Z_6$, we obtain the following proposition :

**Proposition 4.7.2.** *Let* $r \geqslant 4$ *be given. For all* $s \geqslant 0$*, there exist* $A_{2m} \in \mathscr{A}_{m,\Omega}$*, for* $4 \leqslant m \leqslant r$*, and a constant* $C > 0$*, such that for all* $0 < \varepsilon < 1$*,* $\gamma > 0$*,* $N \geqslant 1$ *satisfying*

$$C\varepsilon^{\frac{3}{2}} \left(\frac{N^{\alpha_r}}{\gamma}\right)^3 < 1, \tag{4.99}$$

*there exist*

- $e_8, \dots, e_{2r} \in \ell^\infty_{A_8} \times \cdots \times \ell^\infty_{A_{2r}}$
- $\tau_6 : B_s(0, \frac{3}{2}\varepsilon) \cap \mathcal{U}^N_{\gamma,\varepsilon,r,s} \to \ell^1_s$ *a* $\mathcal{C}^1$ *injective symplectomorphism,*
- $R^{\text{high}}, R^{\text{ord}} \in \mathcal{C}^1(B_s(0, \frac{3}{2}\varepsilon) \cap \mathcal{U}^N_{\gamma,\varepsilon,r,s})$

*such that*

$$H \circ \tau_2 \circ \tau_4 \circ \tau_6 = Z_2 + Z_4 + Z_6 + \sum_{m=4}^{r} Z_{2m} + R^{\text{high}} + R^{\text{ord}}, \tag{4.100}$$

*where* $Z_{2m} = Q_{A_{2m}}[e_{2m}] \in \mathscr{F}^A_{m,\Omega}$ *depends only on the actions. Furthermore we have the following bounds*

- *for all* $m = 4 \dots r$*,* $\mathcal{N}_{\Gamma_{2m}}(e_{2m}) \leqslant N^2$
- *for all* $m = 4 \dots r$*,* $\|c_{2m}\|_{\ell^\infty} \leqslant C_r$
- *For all* $z \in B_s(0, \frac{3}{2}\varepsilon) \cap \mathcal{U}^N_{\gamma,\varepsilon,r,s}$*, we have*

$$\|X_{R^{\text{high}}}(z)\|_{\ell^1_s} \leqslant C\varepsilon^5 N^{-s} \quad \textit{and} \quad \|X_{R^{\text{ord}}}(z)\|_{\ell^1_s} \leqslant C\varepsilon^{2r+1} \left(\frac{N^{\alpha_{r+1}}}{\gamma}\right)^{4r-9} \tag{4.101}$$

- $\tau_6$ *takes values in* $z \in B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma/2,\varepsilon,r,s}$ *and satisfies the estimates*

$$\|\tau_6(z) - z\|_s \leqslant C\varepsilon^{\frac{3}{2}}\gamma N^{-\alpha_r}. \tag{4.102}$$

*Proof.* We construct $\tau_6$ by induction. Note that in the Hamiltonian (4.93), the terms are in $\mathscr{F}_{m,\omega} \subset \mathscr{F}_{m,\Omega}$. Starting with this Hamiltonian, we define $A_8$ and $R_8$ according to the decomposition (see (4.76))

$$Q_{\Gamma_8}[c_8] = Q_{A_8}[c_8] + Q_{R_8}[c_8],$$

where $R_8 \in \mathscr{R}_{4,\Omega}$. Then Lemma 4.6.4 gives us $\Lambda_8 \in \mathscr{R}^*_{2,\Omega}$ such that

$$\{Z_4 + Z_6, Q_{\Lambda_8}[c_8]\} = R_8[c_8],$$

We define $e_8 = c_8$, and we easily verity that $A_8$ and $e_8$ satisfy the hypothesis of the proposition. Setting $\chi_8 = Q_{\Lambda_8}[c_8]$, and using (4.72), the application $\Phi^1_{\chi_8}$ satisfies an estimate under the form

$$\|\Phi^1_{\chi_8}(z) - z\|_{\ell^1_s} \leqslant C\varepsilon^{2p-1}\Big(\frac{N^{\alpha_r}}{\gamma}\Big)^{4p-5} \quad \text{with} \quad p = 8.$$

Thus using (4.99) we conclude

$$\|\Phi^1_{\chi_8}(z) - z\|_{\ell^1_s} \leqslant C\varepsilon^{\frac{3}{2}}\gamma N^{-\alpha_r}.$$

As in the previous Proposition, we verify by using Proposition 4.4.3 that if $z \in B_s(0, \frac{3}{2}\varepsilon) \cap \mathcal{U}^N_{\gamma,\varepsilon,r,s}$, then $\Phi^s_{\chi_8}(z) \in B_s(0, 2\varepsilon) \cap \mathcal{U}^N_{\gamma',\varepsilon,r,s}$ for all $s \in (0, 1)$ where we take $\gamma' = \frac{\gamma}{2}(2 - \frac{1}{r})$.

By using formulas similar to (4.98) and shrinking $\gamma'$ up to $\gamma/2$, we see that

$$H \circ \tau_2 \circ \tau_4 \circ \Phi^1_{\chi_8} = Z_2 + Z_4 + Z_6 + Z_8 + \sum_{m=5}^{r} Q_{\Gamma'_{2m}}[c'_m] + R^{\text{high}} + R^{\text{ord}},$$

where $\Gamma'_{2m}$, $c'_{2m}$, $R^{\text{high}}$ and $R^{\text{ord}}$ satisfy the condition of the Theorem.
By induction, for a given $p$, let us assume that the Hamiltonian is put on normal form up to order $2p$,

$$\sum_{m=2}^{p-1} Z_{2m} + \sum_{m=p}^{r-1} K_{2m} + R^{\text{high}} + R^{\text{ord}},$$

with remainder terms $R^{\text{high}}$, $R^{\text{ord}}$ satisfying (4.101), $Z_{2m} \in \mathscr{F}^A_{m,\Omega}$ and $K_{2m} \in \mathscr{F}^R_{m,\Omega}$. Let us decompose $K_{2p} = Z_{2p} + R_{2p}$ where $Z_{2m} \in \mathscr{F}^A_{p,\Omega}$ and $R_{2p} \in \mathscr{F}^R_{p,\Omega}$. Then to eliminate $R_{2p}$ we construct $\chi_{2p} \in \mathscr{H}^*_{p-2,\Omega}$ by solving the homological equation of Lemma (4.6.4). We have $\chi_{2p} = Q_{\Lambda_{2p}}[c_{2p}]$ with $\Lambda_{2p} \in \mathscr{R}_{p,\Omega}$ and by Lemma 4.6.3 and under the assumption (4.92)

$$\|\Phi^1_{\chi_{2p}}(z) - z\|_{\ell^1_s} \leqslant C\varepsilon^{2p-1}\Big(\frac{N^{\alpha_r}}{\gamma}\Big)^{4p-5} \leqslant C\varepsilon^{\frac{3}{2}}\gamma N^{-\alpha_r}.$$

We then easily verify that the transformation $\tau_6 = \Phi^1_{\chi_8} \circ \cdots \circ \Phi^1_{\chi_{2r}}$ satisfies the conditions of the Theorem. □

### 4.7.4 Proof of the rational normal form Theorem

*Proof of Theorem 4.2.1.* To prove Theorem 4.2.1 it suffices to apply Proposition 4.7.2 at order $r' = 6r$ and to choose

$$\alpha_{r'} = 24r', \quad N_\varepsilon = \varepsilon^{-\frac{2r-2}{s}}, \quad \gamma_\varepsilon = \varepsilon^{\frac{1}{3}+\frac{1}{12}}, \quad s \geqslant s_0(r) = 48 \times 24 \times 6(2r-2). \tag{4.103}$$

With this choice of $N = N_\varepsilon$, we have

$$\varepsilon^5 N_\varepsilon^{-s} = \varepsilon^{2r+3} \quad \text{and} \quad \varepsilon^{2r'+1}\left(\frac{N^{\alpha_{r'+1}}}{\gamma}\right)^{4r'-9} \leqslant \varepsilon^{2r+3}$$

so that the estimate (4.10) is satisfied for $R = R^{\text{high}} + R^{\text{ord}}$ in (4.101) for $\varepsilon$ small enough (we reach the order $2r+3$ instead of $2r+1$ to normalize the constant to $1$).

Now condition (4.99) can be written

$$\varepsilon^{\frac{1}{2}-\frac{1}{3}-\frac{1}{12}-24r'\frac{2r-2}{s}} \leqslant C^{-\frac{1}{3}}.$$

Choosing $\varepsilon < \varepsilon_0(r,s) = C^{-16/3}$ and using the definition of $s_0$, this condition is satisfied :

$$\varepsilon^{\frac{1}{2}-\frac{1}{3}-\frac{1}{12}-24r'\frac{2r-2}{s}} \leqslant \varepsilon^{\frac{3}{48}} \leqslant C^{-\frac{1}{3}}.$$

With these choices, Theorem 4.2.1 holds true with

$$\mathcal{C}_{\varepsilon,r,s} = \mathcal{U}^{N_\varepsilon}_{\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\tfrac{3}{2}\varepsilon), \qquad \tau = \tau_2 \circ \tau_4 \circ \tau_6, \quad \text{and} \quad \mathcal{O}_{\varepsilon,r,s} = \tau(\mathcal{C}_{\varepsilon,r,s}), \tag{4.104}$$

as $\tau$ is injective on $\mathcal{C}_{\varepsilon,r,s}$. Moreover, by Remark (4.7.3) and the previous estimates, we have $\mathcal{O}_{\varepsilon,r,s} \subset \mathcal{U}^{N_\varepsilon}_{\gamma_\varepsilon/4,\varepsilon,r',s} \cap B_s(0,4\varepsilon)$, and

$$\|\tau(z) - z\|_s \leqslant C\varepsilon^{\frac{3}{2}}\gamma N^{-\alpha'_r} \leqslant \varepsilon^{\frac{3}{2}}. \tag{4.105}$$

□

## 4.8 Dynamical consequences and probability estimates

We are now in position to prove Corollary 4.2.2 and Theorem (4.2.3). First, we define the sets

$$\mathcal{V}_{\varepsilon,r,s} = \mathcal{U}^{N_\varepsilon}_{4\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\tfrac{1}{2}\varepsilon) \subset \mathcal{O}_{\varepsilon,r,s} \quad \text{and} \quad \mathcal{W}_{\varepsilon,r,s} = \tau^{-1}(\mathcal{V}_{\varepsilon,r,s}) \subset \mathcal{C}_{\varepsilon,r,s},$$

where as previously $r' = 6r$. With this definition, Estimate (4.14) is a consequence of Proposition 4.5.1 and Proposition (4.4.1). Note that the condition required in this proposition is ensured (with $\gamma' = \gamma/2 = 4\gamma_\varepsilon$) under the condition (4.99). Note moreover that we use the term $\varepsilon^{\frac{1}{12}}$ in the definition of $\gamma$ to fix the constant to $1$ in the final probability estimate and obtain $\varepsilon^{\frac{1}{3}}$.

Similarly, (4.15) is obtained from Corollary (4.5.1) with $\nu = \varepsilon^{\frac{1}{6}}$. This proves Theorem (4.2.3).

To prove the dynamical consequences, we note that the open set $\mathcal{W}_{\varepsilon,r,s}$ contains the initial value in the new variable. Let $z \in \mathcal{V}_\varepsilon$ and $z' = \tau^{-1}(z) \in \mathcal{W}_{\varepsilon,r,s}$. By using (4.105) we have $\|z' - z\|_s = \|z' - \tau(z')\|_s \leqslant C\varepsilon^{\frac{3}{2}}\gamma N^{-\alpha_r}$. Hence by using Proposition 4.4.3, we deduce that

$$\mathcal{W}_{\varepsilon,r,s} \subset \mathcal{U}^{N_\varepsilon}_{2\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\varepsilon).$$

The goal of the analysis is thus to prove that the dynamics starting in $\mathcal{U}^{N^\varepsilon}_{2\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\varepsilon)$ remains in the set $\mathcal{C}_{\varepsilon,r,s} = \mathcal{U}^{N^\varepsilon}_{\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\frac{3}{2}\varepsilon)$ for a time $T \leqslant \varepsilon^{-2r+1}$. To prove this, we first recall a *small* lemma proved in [64] :

**Lemma 4.8.1.** *let $f : \mathbb{R} \to \mathbb{R}_+$ a continuous function, and $y : \mathbb{R} \to \mathbb{R}_+$ a differentiable function satisfying the inequality*

$$\forall\, t \in \mathbb{R}, \quad \frac{\mathrm{d}}{\mathrm{d}t} y(t) \leqslant 2 f(t) \sqrt{y(t)}.$$

*Then we have the estimate*

$$\forall\, t \in \mathbb{R}, \quad \sqrt{y(t)} \leqslant \sqrt{y(0)} + \int_0^t f(s)\,\mathrm{d}s.$$

*Proof of Corollary 4.2.2.* We use a bootstrap argument. Let us fix $r \geqslant 2$, $s \geqslant s_0(r)$ and $\varepsilon < \varepsilon_0(r,s)$ as in Theorem 4.2.1. Let $U(0) = (u_a(0))_{a\in\mathbb{Z}} \in \mathcal{V}_{\varepsilon,r,s}$ and $V(0) = \tau^{-1}(U(0)) = (v_a(0))_{a\in\mathbb{Z}} \in \mathcal{W}_{\varepsilon,r,s}$. By definition, we have

$$V(0) \in \mathcal{W}_{\varepsilon,r,s} \subset \mathcal{U}^{N^\varepsilon}_{2\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\varepsilon)$$

and let

$$T = \sup\{t > 0 \mid V(t') \in \mathcal{C}_{\varepsilon,r,s} \text{ for all } 0 \leqslant t' < t\}.$$

Note that for $t < T$, we have $U(t) = \tau(V(t)) \in \mathcal{O}_{\varepsilon,r,s}$ which coincides with the solution governed by the Hamiltonian $H_{\mathrm{NLS}}$ by uniqueness of the solution. We are going to prove that if $t \leqslant \min(T, \varepsilon^{-2r+1})$ then $V(t) \in \mathcal{U}^{N^\varepsilon}_{\gamma_\varepsilon,\varepsilon,r',s} \cap B_s(0,\frac{3}{2}\varepsilon) = \mathcal{C}_{\varepsilon,r,s}$ and then conclude to $T \geqslant \varepsilon^{-2r}$ by a continuity argument. To prove this we have, in view of (4.104), to control the small divisors (4.46) and (4.47) and the norm $\|V(t)\|_s$.

Let $J_a(t) = |v_a(t)|^2$ denote the actions of $V(t)$. For $t < T$ we can use Theorem 4.2.1 to conclude that

$$\begin{aligned} \Big|\frac{\mathrm{d}}{\mathrm{d}t} J_a(t)\Big| &\leqslant (J_a(t))^{1/2} \left( \left|\frac{\partial R}{\partial \xi_a}(V(t))\right| + \left|\frac{\partial R}{\partial \xi_a}(V(t))\right| \right) && (4.106)\\ &\leqslant (J_a(t))^{1/2} \|X_R(V(t))\|_s \langle a\rangle^{-s} \leqslant C\varepsilon^{2r+2}\langle a\rangle^{-2s}. && (4.107) \end{aligned}$$

Therefore for $t < T$, we have

$$|J_a(t) - J_a(0)| \leqslant CT\varepsilon^{2r+2}\langle a\rangle^{-2s} \quad \forall a \in \mathbb{Z}. \tag{4.108}$$

Together with Proposition 4.4.2, this equation shows that for $T \leqslant \varepsilon^{-2r+1}$ and under the condition (4.99) fulfilled by $N_\varepsilon$ and $\gamma_\varepsilon$, we have $V(t) \in \mathcal{U}^{N^\varepsilon}_{\gamma_\varepsilon,\varepsilon,r',s}$.

To control the norm of $V(t)$, we note that since $\|V(t)\|_s = \sum_{a\in\mathbb{Z}} \langle a\rangle^s |v_a(t)| = \sum_{a\in\mathbb{Z}} \langle a\rangle^s |J_a(t)|^{1/2}$ we get using (4.106) and Lemma 4.8.1

$$\|V(t)\|_s \leqslant \|V(0)\|_s + \int_0^t \|X_R(V(s))\|_s ds \leqslant \|V(0)\|_s + t\varepsilon^{2r+1}.$$

Using (4.9) we get for $t \leqslant \min(T, \varepsilon^{-2r+1})$ and $\varepsilon$ small enough

$$\|V(t)\|_s \leqslant \|V(0)\|_s + t\varepsilon^{2r+1} \leqslant \varepsilon + \varepsilon^2 \leqslant \tfrac{3}{2}\varepsilon, \tag{4.109}$$

hence $V(t) \in \mathcal{C}_{\varepsilon,r,s}$ for $T \leqslant \varepsilon^{-2r+1}$. This shows in particular that $U(t) \in \mathcal{O}_{\varepsilon,r,s}$ on this time horizon and conclude our bootstrap argument.

Finally it remains to prove (4.11). Let $I_a(t) = |u_a(t)|$, by (4.9) we get that $|J_a(t) - I_a(t)| \leqslant C\varepsilon^{\frac{5}{2}+\frac{1}{24}}\langle a\rangle^{-2s}$. We then deduce that for $t \leqslant \min(T, \varepsilon^{-2r+2})$

$$\begin{aligned}|I_a(t) - I_a(0)| &\leqslant |I_a(t) - J_a(t)| + |\tilde{I}_a(t) - J_a(0)| + |J_a(0) - I_a(0)| \\ &\leqslant 2\varepsilon^{\frac{5}{2}}\langle a\rangle^{-2s} + T\varepsilon^{2r+2}\langle a\rangle^{-2s} \leqslant 3\varepsilon^{\frac{5}{2}}\langle a\rangle^{-2s},\end{aligned} \tag{4.110}$$

which shows (4.11) and conclude the proof of the Corollary. □

## 4.9 Appendix

### 4.9.1 The case of (NLSP)

As explain in section 4.2.2, the main difference between (NLS) and (NLSP) appears when we calculate $Z_4$. Indeed, the resonant normal form procedure used in section 4.7.1 leads, in the (NLSP) case, to the following formula (see (4.19) with $\hat{V}_a = a^2$, $a \neq 0$ and $\hat{V}_0 = 0$)

$$Z_4 = \varphi'(0) \sum_{a \neq b \in \mathbb{Z}} \frac{1}{(a-b)^2} I_a I_b.$$

Thus the frequencies associated with this integrable Hamiltonian are

$$\lambda_a(I) = \frac{\partial Z_4}{\partial I_a} = 2\varphi'(0) \sum_{b \neq a \in \mathbb{Z}} \frac{1}{(a-b)^2} I_b.$$

For these frequencies we obtain a much better control of the small denominators that the one obtained for (NLS), in particular, contrary to the (NLS) case (see (4.21)), the loss of derivative is independent of $s$.
For $\boldsymbol{j} = (j_1, \dots, j_{2m}) \in \mathbb{U}_2 \times \mathbb{Z}$, if $j_\alpha = (\delta_\alpha, a_\alpha)$ for $\alpha = 1, \dots, 2m$, the small denominators in the (NLSP) case are given by

$$\omega_{\boldsymbol{j}}(I) = \sum_{\alpha=1}^{2m} \delta_\alpha \frac{\partial Z_4}{\partial I_{a_\alpha}}(I) = 2\varphi'(0) \sum_{\alpha=1}^{2m} \delta_\alpha \sum_{b \neq a_\alpha \in \mathbb{Z}} \frac{1}{(a_\alpha - b)^2} I_b.$$

Let us remark that $\omega_{\boldsymbol{j}}(I)$ has the same structure of the small denominator associated with $Z_6$ used to obtain non resonance estimates, except that $I_b^2$ is replaced by $I_b$ as it can be easily seen by comparing the previous formula with (4.54). By proceeding as in Section 4.5, with the crucial use of Lemma 4.5.7, we obtain the following result whose proof whose proof is left to the reader.

**Lemma 4.9.1.** *Assume that $I_a$, $a \in \mathbb{Z}$ are independent random variable with $I_a$ uniformly distributed in $(0, \langle a\rangle^{-2s-4})$, then there exists a constant $c > 0$ such that for all $\gamma \in (0,1)$ we have*

$$\mathbb{P}\left(\forall \boldsymbol{k} \in \mathcal{I}rr,\ \sharp k \leqslant 2r \Rightarrow |\omega_{\boldsymbol{k}}(I)| \geqslant \gamma \left(\prod_{\alpha=1}^{\sharp \boldsymbol{k}} \langle k_\alpha\rangle^{-4}\right)\right) \geqslant 1 - c\gamma.$$

The major difference with the (NLS) case is that now the small denominator do not depend on $s$ (compare with Lemma 4.5.6). Hence, the construction can be performed without having to distribute the derivative and we can apply a normal form procedure using only $Z_4$ (and not $Z_4 + Z_6$ as in the (NLS) case).

Following the general strategy, for $\varepsilon, \gamma > 0$, $r \geqslant 1$, $N \geqslant 1$ and $s \geqslant 0$, we say that $z \in \ell^1_s$ belongs to the non resonant set $\mathcal{U}_{\gamma,\varepsilon,r,s}$, if for all $\boldsymbol{k} \in \mathcal{I}rr$ of length $\sharp\boldsymbol{k} \leqslant 2r$ we have

$$|\omega_{\boldsymbol{k}}(I)| > \gamma\varepsilon^2 \left( \prod_{\alpha=1}^{\sharp\boldsymbol{k}} \langle k_\alpha \rangle^{-4} \right);$$

and the that $z \in \ell^1_s$ belongs to the truncated non resonant set $\mathcal{U}^N_{\gamma,\varepsilon,r,s}$, if for all $\boldsymbol{k} \in \mathcal{I}rr$ of length $\sharp\boldsymbol{k} \leqslant 2r$ such that $\langle \mu_1(\boldsymbol{k}) \rangle \leqslant N^2$, we have

$$|\omega_{\boldsymbol{k}}(I)| > \gamma\varepsilon^2 N^{-16r}. \tag{4.111}$$

An adapted Proposition 4.4.1 remains valid, namely : for $N$ large enough depending on $\varepsilon$ and on $\gamma' < \gamma$ we have $\mathcal{U}_{\gamma,\varepsilon,r,s} \subset \mathcal{U}^N_{\varepsilon,\gamma',r,s}$. Moreover, by using the previous Lemma, if $z \in \ell^1_s$ depends on random actions $I_a$ independent and uniformly distributed in $(0, \langle a \rangle^{-2s-4})$, there exists a constant $c > 0$ such that for all $\gamma \in (0,1)$ we have

$$\mathbb{P}\left(\forall \varepsilon > 0,\ \varepsilon z \in \mathcal{U}_{\gamma,\varepsilon,r,s}\right) \geqslant 1 - c\gamma. \tag{4.112}$$

Note that the difference with Proposition 4.5.1 is that for one choice of non resonant actions, the non resonance condition holds for all $\varepsilon$. In other words, the phenomenon of resonances between $\varepsilon$ and $I$ cannot occur in the (NLSP) case.

The class of rational Hamiltonians we need is also simpler : we only need to consider $\mathscr{H}_\omega$ and $\mathscr{H}^*_\omega$ defined in Section (4.6), *i.e.* functionals of the form

$$Q_\Gamma[c](z) = \sum_{\ell \in \mathbb{Z}^*} c_\ell (-1)^{n_\ell} \frac{z_{\boldsymbol{\pi}_\ell}}{\prod_{\alpha=1}^{p_\ell} \omega_{\boldsymbol{k}_{\ell,\alpha}}}.$$

with the same condition as in the (NLS) case, but *without* the restrictive condition **(vi)** on the distribution of derivatives, making the proof of the Poisson bracket estimate considerably much simpler, as can be seen in the next Appendix.

By using the estimate (4.111), we can prove an equivalent of Lemma (4.6.3) for this class of functional (with $\alpha_r = 16r$) and the steps of the rational normal form construction can be then followed as in Section 4.7 under the same condition (4.99). The optimization process in $N$ and $\gamma$ can then be done in the same way.
In the end, the probability estimate (4.112) gives Theorem (4.2.4).

### 4.9.2 Proof of Lemma 4.6.6

This section is devoted to the proof of Lemma 4.6.6. As in the statement of the Lemma, let $W = \omega$ or $W = \Omega$ and let $\Gamma = (\boldsymbol{\pi}, \mathbf{k}, \mathbf{h}, n) \in \mathscr{H}^*_{r,W}$ and $\Gamma = (\boldsymbol{\pi}', \mathbf{k}', \mathbf{h}', n') \in \mathscr{H}_{r',W}$.

To compute the poisson bracket between $Q_\Gamma[c]$ and $Q_{\Gamma'}[c']$, we only need to calculate the poisson brackets of the summands (see the expression (4.59)). Applying the Leibniz's rule we

see that, up to combinatorial factors and finite linear combinations depending on $r$, four kind of terms appear depending on which part of the Hamiltonians the Poisson bracket applies to :

***Type I.*** The first type of terms we consider are those where the derivatives apply only on the numerators. They are of the form

$$\frac{c_\ell c_{\ell'}(-i)^{p_\ell+q_\ell+p_{\ell'}+q_{\ell'}}}{\prod_{\alpha=1}^{n_\ell}\omega_{\boldsymbol{k}_{\ell,\alpha}}\prod_{\alpha=n_\ell+1}^{p_\ell}\Omega_{\boldsymbol{k}_{\ell,\alpha}}\prod_{\alpha=1}^{q_\ell}\Omega_{\boldsymbol{h}_{\ell,\alpha}}\prod_{\alpha=1}^{n_{\ell'}}\omega_{\boldsymbol{k}'_{\ell',\alpha}}\prod_{\alpha=n_{\ell'}+1}^{p_{\ell'}}\Omega_{\boldsymbol{k}'_{\ell',\alpha}}\prod_{\alpha=1}^{q_{\ell'}}\Omega_{\boldsymbol{h}_{\ell',\alpha}}}\{z_{\boldsymbol{\pi}_\ell}, z_{\boldsymbol{\pi}'_{\ell'}}\}$$

for some $\ell$ and $\ell'$ in $\mathbb{Z}^*$. Let us set $\boldsymbol{j} = \boldsymbol{\pi}_\ell$ and $\boldsymbol{j}' = \boldsymbol{\pi}'_{\ell'}$, i.e. $z_{\boldsymbol{\pi}_\ell} = z_{j_1}\cdots z_{j_{2m}}$ and $z_{\boldsymbol{\pi}'_{\ell'}} = z_{j'_1}\cdots z_{j'_{2m'}}$. The product $\{z_{\boldsymbol{\pi}_\ell}, z_{\boldsymbol{\pi}_{\ell'}}\}$ is a linear combination of terms of the form $z_{\boldsymbol{j}''}$ with $\boldsymbol{j}'' \in \mathcal{R}_{m_\ell+m_{\ell'}-1}$ .

Up to a combinatorial factor, linear combinations and renumbering to define the application $\boldsymbol{\pi}''$, we can concentrate on terms $z_{\boldsymbol{j}''}$ with $\boldsymbol{j}'' = \boldsymbol{\pi}''_{\ell''}$ of the form

$$z_{j_2}\cdots z_{j_{2m}} z_{j'_2}\cdots z_{j'_{2m'}},$$

provided $\bar{j}_1 = j'_1$. Clearly, the produced term is of the good form with $r'' = r + r' - 1$, $n_{\ell'} = n_\ell + n'_\ell$, $q_{\ell'} = q_\ell + q_{\ell'}$ and $p_{\ell''} = p_\ell + p_{\ell'}$. In particular the reality condition is easily verified by considering the terms corresponding to $-\ell$ and $-\ell'$ and imposing $\overline{\pi''_{\ell''}} = \pi''_{-\ell''}$, and the conditions $(\mathbf{i}) - (\mathbf{v})$ of the definition of the class are trivially satisfied. We can also verify that these terms fulfill the conditions defining the subclass $\mathscr{H}_{r'',W}$. Indeed, in the case when $W = \omega$, we have $q_{\ell''} = q_\ell + q_{\ell'} = 0$ and $n_{\ell''} = n_\ell + n_{\ell'} \leqslant 2(r+1) - 5 + 2r' - 6 = 2(r+r') - 9 \leqslant 2r'' - 7$. In the case $W = \Omega$, we can set $\alpha''_i = \alpha_i + \alpha'_i$ for $i = 1, \ldots, 4$ and $\alpha''_5 = \alpha_5 + \alpha'_5 + 1$, and we can easily check that the relations (4.67) are satisfied for $\Gamma''$. Moreover, using (4.68) and (4.69), we check that $\alpha''_5 = \alpha_5 + \alpha'_5 + 1 \leqslant (r+2) - 4 + r' - 4 + 1 \leqslant r'' - 4$, and similarly that the three conditions in (4.68) are satisfied.

It remains to prove the conditions $(\mathbf{vi})$ and $(\mathbf{vii})$ that are the most delicate. We analyze different cases according to which are the largest indices among $\boldsymbol{j}$, $\boldsymbol{j}'$ and $\boldsymbol{j}''$. The three main case are $\langle j_1\rangle \leqslant \langle\mu_3(\boldsymbol{j})\rangle$, $j_1 = \mu_2(\boldsymbol{j})$ and $j_1 = \mu_1(\boldsymbol{j})$, and by symmetry, we are left to the following cases to be studied :

| | | | | |
|---|---|---|---|---|
| $\langle j_1\rangle \leqslant \langle\mu_3(\boldsymbol{j})\rangle$ | $\langle j'_1\rangle \leqslant \langle\mu_3(\boldsymbol{j}')\rangle$ | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$ | 3.3.a |
| | | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ | 3.3.b |
| | $j'_1 = \mu_2(\boldsymbol{j}')$ | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$ | $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | 3.2.a |
| | | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ | 3.2.b |
| | | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$ | 3.2.c |
| | $j'_1 = \mu_1(\boldsymbol{j}')$ | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ | 3.1 |
| $j_1 = \mu_2(\boldsymbol{j})$ | $j'_1 = \mu_2(\boldsymbol{j}')$ | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$ | 2.2 |
| | $j'_1 = \mu_1(\boldsymbol{j}')$ | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_3(\boldsymbol{j})$ | 2.1.a |
| | | $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j}')$ | 2.1.b |
| $j_1 = \mu_1(\boldsymbol{j})$ | $j'_1 = \mu_1(\boldsymbol{j}')$ | $\mu_1(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j}')$ | 1.1.a |
| | | $\mu_1(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ | $\mu_2(\boldsymbol{j}'') = \mu_3(\boldsymbol{j})$ | 1.1.b |

These cases are summarized in Figure 4.1 where we try to visualize the different configurations.

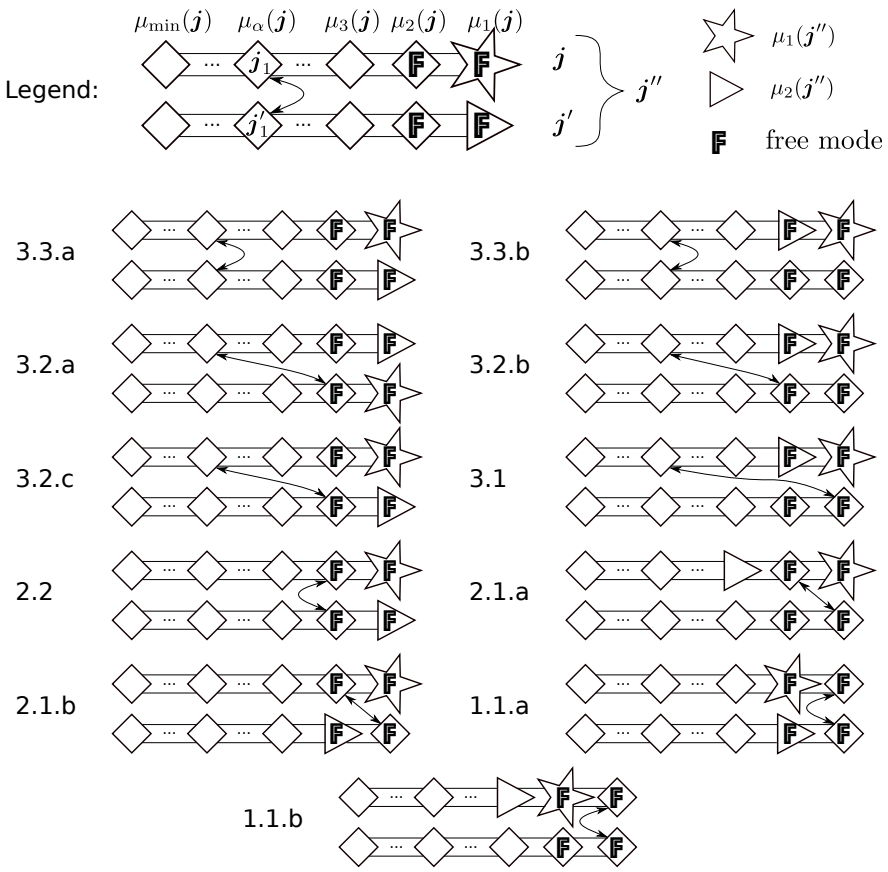


FIGURE 4.1 – Possible configurations arising from the calculation of $\{z_j, z_{j'}\}$.

**Cases 3.3**. In these cases, we have $\langle j_1 \rangle \leqslant \langle \mu_3(\boldsymbol{j}) \rangle$ and $\langle j_1' \rangle \leqslant \langle \mu_3(\boldsymbol{j}') \rangle$ and $(\mathbf{vii})$ is automatically satisfied as $\mu_2(\boldsymbol{j}'')$ is always greater than $\mu_2(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}')$.

To prove $(\mathbf{vi})$, we must contruct a fontion $\iota''$ that distributes the derivatives in $\boldsymbol{j}''$ from the functions $\iota$ and $\iota'$ distributing the derivatives in $\boldsymbol{j}$ and $\boldsymbol{j}'$. Note that by induction hypothesis and the definition of the condition (4.64), the modes $\mu_1(\boldsymbol{j})$, $\mu_2(\boldsymbol{j})$, $\mu_1(\boldsymbol{j}')$ and $\mu_2(\boldsymbol{j}')$ are *free* in the sense that $1$ and $2$ are not in the image of $\iota$ and $\iota'$.

We see that we can build $\iota''$ from $\iota$ and $\iota'$ easily if $j_1$ or $j_1'$ do not correspond to some $\mu_{\iota_\alpha}(\boldsymbol{j})$ or $\mu_{\iota'_\alpha}(\boldsymbol{j}')$, as $j_1$ and $j_1'$ do not appear in $\boldsymbol{j}''$. We thus see that the issue is to control $\langle j_1 \rangle = \langle j_1' \rangle$ by two free modes *and* by letting the two highest modes $\mu_1(\boldsymbol{j}'')$ and $\mu_2(\boldsymbol{j}'')$ free. Indeed, in such a case, up to a reconfiguration of $\iota''$, the relation (4.64) will hold again for $\boldsymbol{j}''$, by using the induction hypothesis on $\boldsymbol{j}$ and $\boldsymbol{j}'$. By symmetry, we thus are led to distinguish two cases :

*Case 3.3.a*. $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$. In this case, $\langle j_1 \rangle \leqslant \langle \mu_2(\boldsymbol{j}) \rangle$, $\langle j_1' \rangle \leqslant \langle \mu_2(\boldsymbol{j}') \rangle$ and we can distribute the derivative to the free modes $\mu_2(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}')$ by letting the two highest modes of $\boldsymbol{j}''$ free.

*Case 3.3.b*. $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$. In this case, we use the fact that $\langle j_1 \rangle = \langle j_1' \rangle$ to control $\langle j_1 \rangle$ by $\langle \mu_2(\boldsymbol{j}') \rangle$ and $\langle j_1' \rangle$ by $\langle \mu_1(\boldsymbol{j}') \rangle$ which are modes always smaller that $\langle \mu_3(\boldsymbol{j}'') \rangle$.

**Cases 3.2**. $\langle j_1 \rangle \leqslant \langle \mu_3(\boldsymbol{j}) \rangle$ and $j_1' = \mu_2(\boldsymbol{j}')$. The main difference with the previous case is that condition **(vii)** is not automatically satisfied. To prove it, we need a control of $\langle \mu_2(\boldsymbol{j}) \rangle$ and $\langle \mu_2(\boldsymbol{j}') \rangle$ by $\langle \mu_2(\boldsymbol{j}'') \rangle$. But on the other hand, we only need to control $\langle j_1 \rangle$ by one mode, as $j_1'$ was not used in the distribution of the derivative (condition (4.64)) for $\boldsymbol{j}'$. As necessarily the first to highest modes of $\boldsymbol{j}''$ are in the set $\{\mu_1(\boldsymbol{j}), \mu_2(\boldsymbol{j}), \mu_1(\boldsymbol{j}')\}$ we are thus led to the following three cases :

*Case 3.2.a*. $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$ and $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$. In this situation, we can easily control $\langle j_1 \rangle$ by $\langle \mu_2(\boldsymbol{j}) \rangle$ which is free, and fulfill condition **(vi)**. Moreover, we have $\langle \mu_2(\boldsymbol{j}) \rangle \leqslant \langle \mu_1(\boldsymbol{j}) \rangle = \langle \mu_2(\boldsymbol{j}'') \rangle$ and $\langle \mu_2(\boldsymbol{j}') \rangle = \langle j_1 \rangle \leqslant \langle \mu_2(\boldsymbol{j}) \rangle$ and hence condition (4.65) for $\boldsymbol{j}''$ is inherited from condition **(vii)** for $\boldsymbol{j}$ and $\boldsymbol{j}'$.

*Case 3.2.b*. $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$. Here, we can control $\langle j_1 \rangle = \langle \mu_2(\boldsymbol{j}') \rangle$ by $\langle \mu_1(\boldsymbol{j}') \rangle$ which is free and smaller than $\mu_2(\boldsymbol{j}'')$ which shows $(\mathbf{vi})$. Moreover, in this situation, we have $\mu_2(\boldsymbol{j}) = \mu_2(\boldsymbol{j}'')$ and $\langle \mu_2(\boldsymbol{j}') \rangle = \langle j_1 \rangle \leqslant \langle \mu_2(\boldsymbol{j}) \rangle = \langle \mu_2(\boldsymbol{j}'') \rangle$ so that $(\mathbf{vii})$ holds true for $\boldsymbol{j}''$.

*Case 3.2.c*. $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$. In this situation $(\mathbf{vi})$ can be easily shown as $\langle j_1 \rangle \leqslant \langle \mu_2(\boldsymbol{j}) \rangle$ which free and smaller than $\langle \mu_2(\boldsymbol{j}'') \rangle$. To prove $(\mathbf{vii})$, we notice that $\mu_2(\boldsymbol{j}) = \mu_2(\boldsymbol{j}'')$ and $\langle \mu_2(\boldsymbol{j}') \rangle = \langle j_1 \rangle \leqslant \langle \mu_2(\boldsymbol{j}) \rangle$.

**Case 3.1**. $\langle j_1 \rangle \leqslant \langle \mu_3(\boldsymbol{j}) \rangle$ and $j_1' = \mu_1(\boldsymbol{j}')$. In this situation we have $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$. As in the previous case, we only have to distribute derivative in one free mode, *i.e.* control $\langle j_1 \rangle$ by $\langle \mu_2(\boldsymbol{j}') \rangle$. This is done by using the zero momentum condition : we have $\langle j_1 \rangle = \langle \mu_1(\boldsymbol{j}') \rangle \leqslant C_{r'} \langle \mu_2(\boldsymbol{j}') \rangle$ where $C_{r'}$ depends only on $r'$. This shows $(\mathbf{vi})$ and $(\mathbf{vii})$ is proved by noticing that $\mu_2(\boldsymbol{j}) = \mu_2(\boldsymbol{j}'')$ and $\langle \mu_2(\boldsymbol{j}') \rangle \leqslant \langle \mu_1(\boldsymbol{j}') \rangle = \langle j_1 \rangle \leqslant \langle \mu_2(\boldsymbol{j}) \rangle$.

**Case 2.2**. $j_1 = \mu_2(\boldsymbol{j})$ and $j_1' = \mu_2(\boldsymbol{j}')$ and by symmetry we can assume $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_1(\boldsymbol{j}')$. In this case, $(\mathbf{vi})$ for $\boldsymbol{j}''$ is directly inherited from the condition for $\boldsymbol{j}$ and $\boldsymbol{j}'$ as $j_1$ and $j_1'$ were not involved in them. To prove $(\mathbf{vii})$, we notice that $\langle \mu_2(\boldsymbol{j}) \rangle = \langle \mu_2(\boldsymbol{j}') \rangle \leqslant \langle \mu_1(\boldsymbol{j}') \rangle = \langle \mu_2(\boldsymbol{j}'') \rangle$.

**Cases 2.1**. $j_1 = \mu_2(\boldsymbol{j})$ and $j_1' = \mu_1(\boldsymbol{j}')$. In this necessarily, we have $\mu_1(\boldsymbol{j}'') = \mu_1(\boldsymbol{j})$. As in the previous case, $(\mathbf{vi})$ is easily obtained. To prove $(\mathbf{vii})$ we have to distinguish two cases :

*Case 2.1.a*. $\mu_2(\boldsymbol{j}'') = \mu_3(\boldsymbol{j})$, which means in particular that $\langle\mu_2(\boldsymbol{j}')\rangle \leqslant \langle\mu_3(\boldsymbol{j})\rangle = \langle\mu_2(\boldsymbol{j}'')\rangle$. Moreover, by using the zero-momentum condition, we have $\langle\mu_2(\boldsymbol{j})\rangle = \langle\mu_1(\boldsymbol{j}') \leqslant C_{r'}\langle\mu_2(\boldsymbol{j}')\rangle \leqslant C_{r'}\langle\mu_2(\boldsymbol{j}'')\rangle$ and this shows the result.

*Case 2.1.a*. $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j}')$. In this situation we just need to prove that $\langle\mu_2(\boldsymbol{j})\rangle$ is controlled by $\langle\mu_2(\boldsymbol{j}'')\rangle$ which is ensured by the fact that $\langle\mu_2(\boldsymbol{j})\rangle = \langle\mu_1(\boldsymbol{j}')\rangle \leqslant C_{r'}\langle\mu_2(\boldsymbol{j}')\rangle$ by using the zero momentum condition.

**Cases 1.1**. $j_1 = \mu_1(\boldsymbol{j})$ and $j_1' = \mu_1(\boldsymbol{j}')$. As before, $(\mathbf{vi})$ is easily obtained. To verify $(\mathbf{vii})$, by symmetry, we have only two cases to examine :

*Case 1.1.a*. $\mu_1(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_2(\boldsymbol{j}')$. In this situation, we have by using the zero momentum condition $\langle\mu_2(\boldsymbol{j})\rangle = \langle\mu_1(\boldsymbol{j}'')\rangle \leqslant C_{r''}\langle\mu_2(\boldsymbol{j}'')\rangle$ which shows $(\mathbf{vii})$.

*Case 1.1.b*. $\mu_1(\boldsymbol{j}'') = \mu_2(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}'') = \mu_3(\boldsymbol{j})$. In this case we have necessarily $\langle\mu_2(\boldsymbol{j}')\rangle \leqslant \langle\mu_3(\boldsymbol{j})\rangle \leqslant \langle\mu_2(\boldsymbol{j})\rangle = \langle\mu_1(\boldsymbol{j}'')\rangle$ and we conclude by using the zero momentum condition as in the previous case.

To conclude the analysis of this type, we just observe that (4.80) is a consequence of the fact that in all the previous cases, we have $\langle\mu_1(\boldsymbol{j}'')\rangle \leqslant \max(\langle\mu_1(\boldsymbol{j}'')\rangle, \langle\mu_1(\boldsymbol{j}'')\rangle)$ and the definition (4.66) of $\mathcal{N}_\Gamma(c)$.

***Type II.*** The second type of terms we consider are those where one $\omega_{\boldsymbol{k}_{\ell,\alpha}}$ appears in the Poisson bracket. They are of the form

$$\frac{c_\ell c_{\ell'}(-i)^{p_\ell+q_\ell+p_{\ell'}+q_{\ell'}} z_{\boldsymbol{\pi}_\ell}}{\prod_{\alpha=1}^{n_\ell-1}\omega_{\boldsymbol{k}_{\ell,\alpha}} \prod_{\alpha=n_\ell+1}^{p_\ell}\Omega_{\boldsymbol{k}_{\ell,\alpha}} \prod_{\alpha=1}^{q_\ell}\Omega_{\boldsymbol{h}_{\ell,\alpha}} \prod_{\alpha=1}^{n_{\ell'}}\omega_{\boldsymbol{k}'_{\ell',\alpha}} \prod_{\alpha=n_{\ell'}+1}^{p_{\ell'}}\Omega_{\boldsymbol{k}'_{\ell',\alpha}} \prod_{\alpha=1}^{q_{\ell'}}\Omega_{\boldsymbol{h}_{\ell',\alpha}}}\{\frac{1}{\omega_{\boldsymbol{k}_{\ell,n_\ell}}}, z_{\boldsymbol{\pi}'_{\ell'}}\}$$

Let us set $\boldsymbol{j}^* = \boldsymbol{k}_{\ell,n_\ell} = (j_1^*, \dots, j^*_{\sharp\boldsymbol{k}_{\ell,n_\ell}})$. The Poisson bracket above is in general zero, except if one of the index of $\boldsymbol{j}^*$ is conjugated to one of the index of $\boldsymbol{j}' = \boldsymbol{\pi}'_{\ell'}$. We can assume here that $\bar{j}_1^* = j_1'$. In this case, we have

$$\{\frac{1}{\omega_{\boldsymbol{j}^*}}, z_{\boldsymbol{j}'}\} = \pm i\frac{1}{\omega^2_{\boldsymbol{j}^*}} z_{\boldsymbol{j}'} \tag{4.113}$$

So the new term is of the good form with $\boldsymbol{j}'' = \boldsymbol{j} \cup \boldsymbol{j}'$ and up to a combinatorial factor, linear combinations and renumbering we can define the application $\boldsymbol{\pi}''$ in such a way that $\boldsymbol{j}'' = \boldsymbol{\pi}''_{\ell''}$. The term in the denominator has one more factor repeating the index $\boldsymbol{k}_{\ell,n_\ell}$. Hence we have $m_{\ell''} = m_\ell + m_{\ell'}$, $n_{\ell''} = n_\ell + n_{\ell'} + 1$, $p_{\ell''} = p_\ell + p_{\ell'} + 1$, $q_{\ell''} = q_\ell + q_{\ell'}$ and $r'' = r + r' - 1$. As in the Type I case, we can fulfill the reality condition by considering the terms corresponding to $-\ell$ and $-\ell'$ and imposing $\overline{\pi''_{\ell''}} = \pi''_{-\ell''}$, and the conditions $(\mathbf{i}) - (\mathbf{v})$ of the definition of the class are hence satisfied. Moreover, we can verify that these terms fulfill the conditions associated with the subclass $\mathscr{H}_{r'',W}$. In the case when $W = \omega$, we have $q_{\ell''} = q_\ell + q_{\ell'} = 0$ and $n_{\ell''} = n_\ell + n_{\ell'} + 1 \leqslant 2(r+1) - 5 + 2r' - 5 = 2(r+r') - 8 \leqslant 2r'' - 6$. Moreover, in the case $W = \Omega$, we can set $\alpha''_i = \alpha_i + \alpha'_i$ for $i \in \{1,3,4\}$ and $\alpha''_i = \alpha_i + \alpha'_i + 1$ for $i \in \{2,5\}$ and check that the relations (4.67) and (4.68) are satisfied for $\Gamma''$.

In this case $\mu_2(\boldsymbol{j}'')$ is necessarily greater than $\mu_2(\boldsymbol{j})$ and $\mu_2(\boldsymbol{j}')$, so that $(\mathbf{vii})$ is easily proved.

To prove $(\mathbf{vi})$, we observe that the functions $\iota$ and $\iota'$ distribute the indices $\boldsymbol{k}_{\ell,\alpha}$ and $\boldsymbol{k}'_{\ell',\alpha}$ to some indices in $\boldsymbol{j}$ and $\boldsymbol{j}'$ respectively that are always lower than the third ones. Hence we have four free indices, and two new derivatives to distribute coming from the presence of the new term $\omega_{\boldsymbol{j}^*}$. We can distinguish two cases :

- $\langle j'_1\rangle \leqslant \langle\mu_2(\boldsymbol{j}')\rangle$. In this situation, we use $(\mathbf{vi})$ saying that $\langle\mu_{\min}(\boldsymbol{j}^*)\rangle \leqslant C\langle\mu_2(\boldsymbol{j})\rangle$. Hence as $\langle\mu_{\min}(\boldsymbol{j}^*)\rangle \leqslant \langle j_1^*\rangle = \langle j'_1\rangle \leqslant \langle\mu_2(\boldsymbol{j}')\rangle$, we can construct $\iota''$ from $\iota$ and $\iota'$ and by making $\boldsymbol{j}^*$ correspond to the third and fourth largest indices amongst $\mu_1(\boldsymbol{j}), \mu_2(\boldsymbol{j}), \mu_1(\boldsymbol{j}')$ and $\mu_2(\boldsymbol{j}')$.
- $j'_1 = \mu_1(\boldsymbol{j}')$. We still have by $(\mathbf{vi})$ that $\langle\mu_{\min}(\boldsymbol{j}^*)\rangle \leqslant C\langle\mu_2(\boldsymbol{j})\rangle$. Moreover by zero momentum condition, we have $\langle\mu_{\min}(\boldsymbol{j}^*)\rangle \leqslant \langle j_1^*\rangle = \langle j'_1\rangle = \langle\mu_1(\boldsymbol{j}')\rangle \leqslant C_{r'}\langle\mu_2(\boldsymbol{j}')\rangle$ and we are back the the previous case.

***Type III.*** Now we consider terms where one $\Omega_{\boldsymbol{k}_{\ell,\alpha}}$ appears in the Poisson bracket. They are of the form

$$\frac{c_\ell c_{\ell'}(-i)^{p_\ell+q_\ell+p_{\ell'}+q_{\ell'}} z_{\boldsymbol{\pi}_\ell}}{\prod_{\alpha=1}^{n_\ell}\omega_{\boldsymbol{k}_{\ell,\alpha}}\prod_{\alpha=n_\ell+1}^{p_\ell-1}\Omega_{\boldsymbol{k}_{\ell,\alpha}}\prod_{\alpha=1}^{q_\ell}\Omega_{\boldsymbol{h}_{\ell,\alpha}}\prod_{\alpha=1}^{n_{\ell'}}\omega_{\boldsymbol{k}'_{\ell',\alpha}}\prod_{\alpha=n_{\ell'}+1}^{p_{\ell'}}\Omega_{\boldsymbol{k}'_{\ell',\alpha}}\prod_{\alpha=1}^{q_{\ell'}}\Omega_{\boldsymbol{h}_{\ell',\alpha}}}\{\frac{1}{\Omega_{\boldsymbol{k}_{\ell,p_\ell}}}, z_{\boldsymbol{\pi}'_{\ell'}}\}$$

Let us set $\boldsymbol{j}^* = \boldsymbol{k}_{\ell,p_\ell} = (j_1^*,\dots,j^*_{\sharp\boldsymbol{k}_{\ell,p_\ell}})$, $\boldsymbol{j} = \boldsymbol{\pi}_\ell$ and $\boldsymbol{j}' = \boldsymbol{\pi}'_{\ell'}$ as before. To compute the Poisson bracket there two case to examine.

- First, if $\overline{\boldsymbol{j}^*}\cap\boldsymbol{j}' = \varnothing$ then
$$\{\frac{1}{\Omega_{\boldsymbol{j}^*}}, z_{\boldsymbol{j}'}\} = \pm iP(I)\frac{z_{\boldsymbol{j}'}}{\Omega^2_{\boldsymbol{j}^*}},$$
where, in view of the form $Z_6$ (see (4.37)) , $P$ is a polynomial of degree $1$ with real coefficients. Up to a combinatorial factor, linear combinations and renumbering we can define the application $\boldsymbol{\pi}''$ satisfying the reality condition, and we can set $m_{\ell''} = m_\ell + m_{\ell'} + 1$, $n_{\ell''} = n_\ell + n_{\ell'}$, $p_{\ell''} = p_\ell + p_{\ell'}$ and $q_{\ell''} = q_\ell + q_{\ell'} + 1$. The conditions $(\mathbf{i})-(\mathbf{v})$ of the definition of the class are hence satisfied. Moreover, we can set $\alpha''_i = \alpha_i + \alpha'_i$ for $i \in \{1,2,3\}$, $\alpha''_i = \alpha_i + \alpha'_i + 1$ for $i \in \{4,5\}$ and check that the relations (4.67) and (4.68) are satisfied for $\Gamma''$. Moreover, $(\mathbf{vii})$ is satisfied as $\langle\mu_2(\boldsymbol{j})\rangle \leqslant \langle\mu_2(\boldsymbol{j}'')\rangle$ and $\langle\mu_2(\boldsymbol{j}')\rangle \leqslant \langle\mu_2(\boldsymbol{j}'')\rangle$, and $(\mathbf{vi})$ is also satisfied as there is no new derivative to distribute.
- If one of the index of $\boldsymbol{j}^*$ is conjugate of one of the index of $\boldsymbol{j}'$, then we get
$$\{\frac{1}{\Omega_{\boldsymbol{j}^*}}, z_{\boldsymbol{j}'}\} = \pm i\frac{z_{\boldsymbol{j}'}}{\Omega^2_{\boldsymbol{j}^*}} + \pm iP(I)\frac{z_{\boldsymbol{j}'}}{\Omega^2_{\boldsymbol{j}^*}}, \tag{4.114}$$
where $P$ is a polynomial of degree $1$ with real coefficients. We thus treat the second term as previously. To treat the first term, we use the same analysis than the one in type II with $n_{\ell'} = n_\ell + n_{\ell'}$, $p_{\ell''} = p_\ell + p_{\ell'} + 1$ $q_{\ell''} = q_\ell + q_{\ell'}$. The only difference is that we set $\alpha''_i = \alpha_i + \alpha'_i$ for $i \in \{1,2,4\}$ and $\alpha''_i = \alpha_i + \alpha'_i + 1$ for $i \in \{3,5\}$ but the distribution of derivatives is achieved in a similar way.

***Type IV.*** Finally we consider terms where one $\Omega_{\boldsymbol{h}_{\ell,\alpha}}$ appears in the Poisson bracket. They are of the form

$$\frac{c_\ell c_{\ell'}(-i)^{p_\ell+q_\ell+p_{\ell'}+q_{\ell'}} z_{\boldsymbol{\pi}_\ell}}{\prod_{\alpha=1}^{n_\ell}\omega_{\boldsymbol{k}_{\ell,\alpha}}\prod_{\alpha=n_\ell+1}^{p_\ell}\Omega_{\boldsymbol{k}_{\ell,\alpha}}\prod_{\alpha=1}^{q_\ell-1}\Omega_{\boldsymbol{h}_{\ell,\alpha}}\prod_{\alpha=1}^{n_{\ell'}}\omega_{\boldsymbol{k}'_{\ell',\alpha}}\prod_{\alpha=n_{\ell'}+1}^{p_{\ell'}}\Omega_{\boldsymbol{k}'_{\ell',\alpha}}\prod_{\alpha=1}^{q_{\ell'}}\Omega_{\boldsymbol{h}_{\ell',\alpha}}}\{\frac{1}{\Omega_{\boldsymbol{h}_{\ell,q_\ell}}}, z_{\boldsymbol{\pi}'_{\ell'}}\}$$

It is almost the same as type III except that to deal with the first term in the right-hand side of (4.114) we count one $\Omega_{\boldsymbol{j}*}$ in the denominator as $\Omega_{\boldsymbol{h}_{\ell'',q_{\ell''}}}$ with $q_{\ell''} = q_\ell + q_{\ell'} + 1$ and the other is counted as $\Omega_{\boldsymbol{k}_{\ell'',p_{\ell''}}}$ with $p_{\ell''} = p_\ell + p_{\ell'}$. The analysis is then the same as in Type II for the distribution of derivatives.

CHAPITRE 5

# LONG-TIME BEHAVIOR OF SECOND ORDER LINEARIZED VLASOV-POISSON EQUATIONS NEAR A HOMOGENEOUS EQUILIBRIUM.

## 5.1 Introduction

We consider distribution functions $f = f(t,x,v) : \mathbb{R} \times \mathbb{T}_d \times \mathbb{R}^d \to \mathbb{R}$ and potentials $\phi = \phi(t,x) : \mathbb{R} \times \mathbb{T}_d \to \mathbb{R}$ satisfying the Vlasov-Poisson system

$$\begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x \phi \cdot \nabla_v f = 0 \\ \Delta_x \phi = n(f) - \int_{\mathbb{R}^d} f \mathrm{d}v \\ f(t=0) = f_0. \end{cases} \tag{VP}$$

Here periodic boundary conditions being used, $\mathbb{T}_d$ is a $d$ dimensional torus : there exist $L_1, \dots, L_d > 0$ such that $\mathbb{T}_d = (\mathbb{R}/L_1\mathbb{Z}) \times \cdots \times (\mathbb{R}/L_d\mathbb{Z})$. Furthermore, doing an assumption of neutrality, we only consider solutions of (VP) such that

$$n(f) = \iint_{\mathbb{T}_d \times \mathbb{R}^d} f \mathrm{d}x \mathrm{d}v.$$

In this paper, we aim at exhibiting nonlinear and multidimensional phenomena of solutions of (VP), pursuing a first preliminary work [20] on this subject. Beyond their physical interest, these phenomena can be relevant to evaluate the performances and the qualitative properties of numerical methods.

Since the very first developments of numerical methods for solving VP (we refer to [106], for a review ; the litterature is particularly huge in $1D \times 1D$ and we can mention [103], as one of the earliest works in $2D \times 2D$), the numerical solutions are compared to the solutions of the Vlasov-Poisson system linearized around a homogeneous equilibria $f^{eq} \equiv f^{eq}(v)$ (it can be noticed that every function depending only on $v$ is an equilibrium of (VP)). It consists in looking for solutions of (VP) of the type

$$\begin{cases} f &= f^{eq} &+ \varepsilon g \\ \phi &= 0 &+ \varepsilon \psi. \end{cases}$$

This chapter is a joint work with Michel Mehrenberger realized in [28].

Neglecting second order terms, $g$ is formally a solution of

$$\begin{cases} \partial_t g + v \cdot \nabla_x g - \nabla_x \psi \cdot \nabla_v f^{eq} = 0, \\ \Delta_x \psi + \displaystyle\int_{\mathbb{R}^d} g \, \mathrm{d}v = 0, \\ g(t=0) = g_0. \end{cases} \tag{VPL}$$

This equation being linear and homogeneous, it is natural to try to solve it realizing a Fourier transform we respect to the variable $x$. Thus, we get

$$\begin{cases} \partial_t \widehat{g} + i(v \cdot k)\widehat{g} - i\widehat{\psi}(k \cdot \nabla_v) f^{eq} = 0, \\ -|k|^2 \widehat{\psi} + \displaystyle\int_{\mathbb{R}^d} \widehat{g} \, \mathrm{d}v = 0, \\ \widehat{g}(t=0) = \widehat{g}_0, \end{cases} \tag{VPLF}$$

where $k \in \widehat{\mathbb{T}}_d = (2\pi/L_1)\mathbb{Z} \times \cdots \times (2\pi/L_d)\mathbb{Z}$ and the Fourier transform with respect to the space variable is defined for $u \in L^1(\mathbb{T}_d)$ and $k \in \widehat{\mathbb{T}}_d$ by

$$\widehat{u}(k) = \left( \prod_{j=1}^{d} L_j \right)^{-1} \int_{\mathbb{T}_d} u(x) e^{-ik \cdot x} \mathrm{d}x.$$

It is relevant to notice on (VPLF) that there is no *energy exchange* between space modes at the linear level. In other words, if $g$ is a solution of (VPL) such that $g_0 \equiv \widehat{g_0}(v) e^{ik \cdot x}$ then it is of the form $g(t,x,v) = \widehat{g}(t,v) e^{ik \cdot x}$. As a consequence, the linear analysis is not well suited to describe multidimensional phenomena that could be confronted with numerical simulations.

Since (VPLF) is linear and autonomous, it is natural to solve it with the *Laplace transform*. This transform is defined for functions $u : \mathbb{R}_+^* \to \mathbb{C}$ such that there exists $\lambda \in \mathbb{R}$ satisfying $ue^{-\lambda t} \in L^\infty(\mathbb{R}_+^*)$ and $z \in \mathbb{C}$ such that $\Im z > \lambda$ by

$$\mathcal{L}[u](z) = \int_0^\infty u(t) e^{izt} \mathrm{d}t.$$

Thus, it can be proven that solutions of (VPLF) are given by

$$g(t,x,v) = \sum_{k \in \widehat{\mathbb{T}_d}} e^{ik \cdot (x - vt)} \widehat{g_0}(k,v) + i \int_0^t e^{ik \cdot (x - v(t-s))} \widehat{\psi}(s,k) k \cdot \nabla_v f^{eq}(v) \mathrm{d}s, \tag{5.1}$$

and for $\Im z$ large enough

$$\mathcal{L}\left[\widehat{\psi}(t,k)\right](z) = \frac{N_k(z)}{D_k(z)} =: M_k(z). \tag{5.2}$$

where $N_k$ and $D_k$ are holomorphic functions defined when $\Im z$ is large enough by

$$N_k(z) = -\frac{i}{|k|^2} \int_{\mathbb{R}^d} \frac{\widehat{g_0}(k,v)}{v \cdot k - z} \mathrm{d}v \quad \text{and} \quad D_k(z) = 1 - \frac{1}{|k|^2} \int \frac{k \cdot \nabla_v f^{eq}(v)}{v \cdot k - z} \mathrm{d}v. \tag{5.3}$$

Thus to get a solution $g$ of (VPL) by (5.1) we just have to solve the equation (5.2) (called *dispersion relation*) determining an inverse Laplace transform.

Up to some strong assumptions on $f^{eq}$ and $\widehat{g_0}(k)$ (precised later), it can be proven that $N_k$ and $D_k$ are entire functions and that, for all $\lambda \in \mathbb{R}$, the number of zeros of $D_k$ with an imaginary part larger than $\lambda$ is finite (see Remark 5.1.3). Thus, using the formula

$$\mathcal{L}[t^m e^{-i\omega t}](z) = \frac{i^{m+1} m!}{(z-\omega)^{m+1}},\ \omega \in \mathbb{C},\ m \in \mathbb{N}, \tag{5.4}$$

and realizing precise estimates of remainder terms, we can prove that (5.2) has an analytic solution $\widehat{\psi}$ whose analytic expansion is given, for all $\lambda \in \mathbb{R}$, by

$$\widehat{\psi}(t,k) = \sum_{\substack{D_k(\omega)=0 \\ \Im\omega \geqslant \lambda}} P_{\omega,k}(t) e^{-i\omega t} + \mathcal{O}(e^{\lambda t}), \tag{5.5}$$

where $P_{\omega,k}$ is the polynomial such that $M_k(z) \underset{z\to\omega}{=} \mathcal{L}[P_{\omega,k}(t)e^{-i\omega t}](z) + \mathcal{O}(1)$ is the expansion of $M_k(z)$ in $\omega$.

Such an analysis was first realized by Landau [90], in 1946. It has been done rigorously and generalized in 1986 by Degond [59]. It gave a partial explanation to the phenomenon of *Landau damping*. This latter corresponds to the dynamic of (VP) when for all $k \in \widehat{\mathbb{T}_d}$, $D_k(z)$ does not vanish if $\Im z \geqslant 0$. In this case, the electric potential goes exponentially fast to zero as $t$ goes to $+\infty$. In 2011, Mouhot and Villani proved the existence of this phenomenon for the nonlinear Vlasov-Poisson equation (VP) in [93] and continued in the work of Bedrossian, Masmoudi and Mouhot in [22].

As we have just seen, due to the absence of energy exchange between the spaces modes at the linear level, the linearization is not relevant to explain really multidimensional phenomena. Furthermore, of course, it can not explain nonlinear phenomena. This motivates thus the study of the dynamic of the *second order term* in the expansion of $f$ as powers of $\varepsilon$. More precisely, we look for a solution of (VP) under the form

$$\left\{ \begin{array}{lllllllll} f & = & f^{eq} & + & \varepsilon g & + & \varepsilon^2 h & + & o(\varepsilon^2), \\ \phi & = & 0 & + & \varepsilon\psi & + & \varepsilon^2\mu & + & o(\varepsilon^2), \end{array} \right.$$

where $h(t=0) \equiv 0$. Neglecting the third order terms, it can be proven formally that $(h,v)$ is a solution of

$$\left\{ \begin{array}{l} \partial_t h + v\cdot\nabla_x h - \nabla_x\mu\cdot\nabla_v f^{eq} = \nabla_x\psi\cdot\nabla_v g, \\ \Delta_x\mu + \displaystyle\int_{\mathbb{R}^d} h \,\mathrm{d}v = 0, \\ h(t=0) = 0. \end{array} \right. \tag{VPL2}$$

We recognize the linearized Vlasov-Poisson equations, with an initial condition equal to zero but with a source term. In that case, we refer to Denavit [60], for one of the first works on the subject, in 1965. Different second order oscillations appear and have been studied by physicists (see for example [101] and references therein ; there are many references especially from the 1960s and 1970s). Our aim is to make here a rigorous mathematical study of the asymptotical behavior of the solutions of these equations, which has, to the best of our knowledge, not already been performed.

Here, as for the linear case, we solve (VPL2) using a Laplace transform for the time variable and a Fourier transform for the space variable. More precisely, some calculations prove that

(VPL2) is equivalent to

$$h(t,x,v) = \sum_{k\in\widehat{\mathbb{T}_d}} i\int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\mu}(s,k)k\cdot\nabla_v f^{eq}(v)\mathrm{d}s + \int_0^t \widehat{\nabla_x\psi\cdot\nabla_v g(s,k,v)}e^{ik\cdot(x-v(t-s))}\mathrm{d}s, \tag{5.6}$$

and when $\Im z$ is large enough

$$\mathcal{L}\left[\widehat{\mu}(t,k)\right](z) = \frac{\mathcal{N}_k(z)}{D_k(z)} =: \mathcal{M}_k(z), \tag{5.7}$$

where $D_k$ is given by (5.3) and $\mathcal{N}_k$ is a meromorphic function on $\mathbb{C}$ explicitly known.

As previously, there is just to invert a Laplace transform to solve (VPL2). As for the linear case, a precise study of $\mathcal{M}_k$ and its poles gives a solution $\mu$ of (5.7) and an asymptotic expansion of the form

$$\forall\lambda\in\mathbb{R},\ \widehat{\mu}(t,k) = \sum_{\substack{\mathcal{M}_k(\omega)=\infty\\ \Im\omega\geqslant\lambda}} Q_{\omega,k}(t)e^{-i\omega t} + \mathcal{O}(e^{\lambda t}). \tag{5.8}$$

where $Q_{\omega,k}$ is the polynomial such that $\mathcal{M}_k(z) \underset{z\to\omega}{=} \mathcal{L}[Q_{\omega,k}(t)e^{-i\omega t}](z) + \mathcal{O}(1)$.

The poles of $\mathcal{M}_k$ are of two kinds : they can be zeros of $D_k$ (generating the same frequencies as at the first order) or poles of $\mathcal{N}_k$. The study of the poles is technical because $\mathcal{N}_k$ is defined from the solution of (VPL). However, the asymptotic expansion of $\psi$ (see (5.5)) enables a decomposition of $\mathcal{N}_k$ in more elementary terms whose poles can be determined.

In order to give an intuition of these poles, we consider a term that is very representative [1] of this decomposition :

$$\mathcal{N}_k^{(rep)}(z) = \mathcal{L}\left[F_k^{(rep)}(t)\right](z)$$

where

$$F_k^{(rep)}(t) = e^{-i(\omega_1 t+\omega_2 t)} \iint\limits_{0\leqslant\tau\leqslant s\leqslant t} e^{i(\omega_1\tau+\omega_2 s)}\mathscr{F}[f^{eq}](\tau k_1 + sk_2)\,\mathrm{d}s\,\mathrm{d}\tau \tag{5.9}$$

with $k_1,k_2\in\widehat{\mathbb{T}_d}\backslash 0$ satisfy $k_1+k_2=k$, $\omega_1,\omega_2\in\mathbb{C}$ are such that $D_{k_1}(\omega_1)=D_{k_2}(\omega_2)=0$ and $\mathscr{F}[f^{eq}]$ is the Fourier transform of $f^{eq}$. The later being defined for $u\in L^1(\mathbb{R}^d)$ and $\xi\in\mathbb{R}^d$ by

$$\mathscr{F}[u](\xi) = \int_{\mathbb{R}^d} u(v)e^{-iv\cdot\xi}\mathrm{d}v.$$

Since $\mathcal{N}_k^{(rep)}$ is the Laplace transform of $F_k^{(rep)}(t)$, it can be proven that its poles are given by the asymptotic expansion of $F_k^{(rep)}(t)$ with the formula (5.4). As it is suggested by the formula (5.9), the behavior of this later is quite different if the set of the points $(\tau,s)$ such that $\tau k_1+sk_2=0$ is a line segment (*resonant case*) or a point (*non-resonant case*).

In the non-resonant case, there exists a constant $c>0$ such that

$$0\leqslant\tau\leqslant s\leqslant t,\ |\tau k_1+sk_2|\geqslant cs.$$

So, assuming that $f^{eq}$ is regular enough so that $\mathscr{F}[f^{eq}](\xi)$ decreases faster than any exponential as $|\xi|$ goes to $+\infty$ (for example like a Gaussian), we can prove that the integral in (5.9)

1. but slightly simplified.

converges faster than any exponential as $t$ goes to $+\infty$. As a consequence, we get a constant $a \in \mathbb{C}$ such that

$$\forall \lambda \in \mathbb{R},\ F_k^{(rep)}(t) = a e^{-i(\omega_1 t + \omega_2 t)} + \mathcal{O}(e^{\lambda t}).$$

In the resonant case, there exists $\gamma \in (0,1)$ such that

$$k_2 = -\gamma k_1.$$

Realizing a natural change of coordinates in (5.9), we get

$$F_k^{(rep)}(t) = \int_0^t \int_{-\gamma s}^{(1-\gamma)s} e^{-i(\omega_1(t-\tau-\gamma s)+\omega_2(t-s))} \mathscr{F}[f^{eq}](\tau k_1)\ \mathrm{d}\tau\ \mathrm{d}s.$$

Thus, assuming that $f^{eq}$ is regular enough so that $\mathscr{F}[f^{eq}](\xi)$ decrease faster than any exponential as $|\xi|$ goes to $+\infty$, we have

$$\begin{aligned} F_k^{(rep)}(t) = &\left(\int_0^t e^{-i(\omega_1(t-\gamma s)+\omega_2(t-s))}\mathrm{d}s\right)\left(\int_{\mathbb{R}} e^{i\omega_1\tau}\mathscr{F}[f^{eq}](\tau k_1)\mathrm{d}\tau\right) \\ &- e^{-it(\omega_1+\omega_2)} \int_0^\infty \int_{\substack{\tau \geqslant (1-\gamma)s \\ \text{or } \tau < -\gamma s}} e^{i(\omega_1(\tau+\gamma s)+\omega_2 s)} \mathscr{F}[f^{eq}](\tau k_1)\ \mathrm{d}\tau\ \mathrm{d}s \\ &+ e^{-it(\omega_1+\omega_2)} \int_t^\infty \int_{\substack{\tau \geqslant (1-\gamma)s \\ \text{or } \tau < -\gamma s}} e^{i(\omega_1(\tau+\gamma s)+\omega_2 s)} \mathscr{F}[f^{eq}](\tau k_1)\ \mathrm{d}\tau\ \mathrm{d}s, \end{aligned}$$

and we can prove that the third term decreases faster than any exponential. Thus, this decomposition provides the following asymptotic expansion

$$\forall \lambda \in \mathbb{R},\ F_k^{(rep)}(t) = a e^{-it(\omega_1+\omega_2)} + b e^{-it\omega_b} + \mathcal{O}(e^{\lambda t}),$$

where $a, b \in \mathbb{C}$ and $\omega_b = (1-\gamma)\omega_1 = (|k|/|k_1|)\omega_1$ is the *Best frequency* (according to [101]).

As suggested by this sketch of proof, we can prove that $\mathcal{M}_k$ have three kinds of poles. More precisely, if $\omega$ is a pole of $\mathcal{M}_k$ it satisfies one of the following conditions

(I) $\omega$ is a zero of $D_k$,

(II) $\omega = \omega_1 + \omega_2$ where $D_{k_1}(\omega_1) = D_{k_2}(\omega_2) = 0$ and $k_1 + k_2 = k$,

(III) $\omega = (|k|/|k_1|)\omega_1$ where $D_{k_1}(\omega_1) = 0$ and there exists $\gamma \in (0,1)$ such that $k = \gamma k_1$.

We recall that these poles drive the asymptotic behavior of $\widehat{\mu}(k)$ through formula (5.8). The frequencies (I) and (II) have already been identified in our preliminary work on this subject [20], but not the frequency (III). We emphasize that all the three type of frequencies are listed in [101], which makes our analysis coherent with the physics litterature.

To conclude this presentation, we are going to state a precise theorem giving the asymptotic behavior of the solutions of (VPL2). To this end, we need to introduce some notations.

**Definition 5.1.1.** *Let $\mathscr{E}(\mathbb{R}^d)$ be the subspace of the Schwartz space $\mathscr{S}(\mathbb{R}^d)$, of functions $f$, whose Fourier transform, $\mathscr{F} f$, extends to an entire function on $\mathbb{C}^d$ and such that*

$$\exists \alpha \in (0, \frac{\pi}{2}), \forall \beta \in (0,\alpha), \forall \lambda \in \mathbb{R},\ \sup_{x \in \mathbb{R}^d} \sup_{\theta \in (-\beta,\beta)} e^{\lambda|x|} |\mathscr{F} f(e^{i\theta}x)| < \infty, \tag{5.10}$$

*where $|\cdot|$ denotes the canonical Hermitian norm of $\mathbb{C}^d$.*

**Remark 5.1.2.** *Most of our results require that $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $v \mapsto \widehat{g_0}(k,v) \in \mathscr{E}(\mathbb{R}^d)$. This assumption is probably not optimal but the space $\mathscr{E}(\mathbb{R}^d)$ contains most of the usual functions used in Vlasov-Poisson simulations. For example Maxwellian functions belong to this space. Appendix 5.6.1 provides many examples and details about this space.*

**Remark 5.1.3.** *Assuming that $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $v \mapsto \widehat{g_0}(k,v) \in \mathscr{E}(\mathbb{R}^d)$, $D_k$ and $N_k$ are entire functions and for all $\lambda \in \mathbb{R}$, the number of zeros of $D_k$ with an imaginary part larger than $\lambda$ is finite (proof will be given in Corollary 5.3.3 and Proposition 5.3.3). Appendix 5.6.3 provides an algorithm to computate the zeros of $D_k$.*

**Definition 5.1.4.** *If $k \in \widehat{\mathbb{T}_d}$, $n_{k,\omega}$ denotes the multiplicity of $\omega$ as zero of $D_k$, i.e.*

$$n_{k,\omega} = \max\{m \in \mathbb{N} \mid \forall \ell < m,\ D_k^{(\ell)}(\omega) = 0\}.$$

Most of the result of this paper will require that $g_0$ is supported on a finite number of spatial modes whose set is denoted $K \subset \widehat{\mathbb{T}_d}\backslash\{0\}$. More precisely, they require the following assumtion

**Assumption 1.** *There exists $K$, a finite part of $\widehat{\mathbb{T}_d}\backslash\{0\}$ such that*

$$\forall x \in \mathbb{T}_d,\ g_0(x,v) = \sum_{k\in K} e^{ik\cdot x}\widehat{g_0}(k,v), \quad \textit{with } v \mapsto \widehat{g_0}(k,v) \in \mathscr{E}(\mathbb{R}^d).$$

This assumption seems clearly not optimal but it is general enough to exhibit the relevant phenomena and it corresponds to the usual initial data used for numerical simulations. Furthermore, it simplifies most of the proof avoiding several problems of convergences.

We can now state the main result of this paper : the following theorem proves the existence of smooth solutions of (VPL) and (VPL2) and describes their asymptotic behavior.

**Theorem 5.1.5.** *Let $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $g_0$ be a function satisfying Assumption 1. Then there exist two $C^\infty$ functions $\psi, \mu : \mathbb{R}_+^* \times \mathbb{T}_d \to \mathbb{R}$ and two continuous functions $g, h : \mathbb{R}_+ \times \mathbb{T}_d \times \mathbb{R}^d \to \mathbb{R}$, $C^\infty$ on $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$, such that $(g,\psi,h,\mu)$ is solution of* (VPL) *and* (VPL2)*.*

*Furthermore, if $\lambda \in \mathbb{R}$, $\psi$ is a linear combination of functions of the two types*

$$J(t,x) = t^m e^{i(k\cdot x - \omega t)} \textit{ and } R(t,x) = r(t)e^{i(k\cdot x - i\lambda t)}$$

*where $k \in K$, $D_k(\omega) = 0$, $\Im\omega \geqslant \lambda$ and $0 \leqslant m < n_{k,\omega}$ and $r$ is a bounded analytic function on $\mathbb{R}_+^*$.*

*Similarly, $\mu$ is a linear combination of functions of the four types*

$$\begin{array}{rclrcl} J(t,x) & = & t^m e^{i(k\cdot x - \omega t)} & I(t,x) & = & t^\ell e^{i(k\cdot x - (\omega_1+\omega_2)t)} \\ B(t,x) & = & t^p e^{i(k\cdot x - \frac{|k|}{|k_1|}\omega_1 t)} d_{k_1}^{k_2} & R(t,x) & = & r(t)e^{i(k\cdot x - i\lambda t)} \end{array}$$

*where $k = k_1 + k_2$, $r$ is a bounded analytic function on $\mathbb{R}_+^*$ and $k_1, k_2 \in K$ satisfy*

$$\begin{cases} D_k(\omega) = D_{k_1}(\omega_1) = D_{k_2}(\omega_2) = 0 \\ k\cdot k_1 \neq 0 \textit{ and } \left(d_{k_1}^{k_2} \neq 0 \iff \exists \gamma \in (0,1),\ k = \gamma k_1\right) \\ \Im\omega \geqslant \lambda \textit{ and } \left(\Im\omega_1 + \Im\omega_2 \geqslant \lambda \textit{ or } \frac{|k|}{|k_1|}\Im\omega_1 \geqslant \lambda\right) \\ m < n_{k,\omega},\ \ell < n_{k_1,\omega_1} + n_{k_2,\omega_2} - 1 + \sigma_{\omega_1,\omega_2}^{k_1,k_2},\ p < n_{k_1,\omega_1} + 1 + \nu_{\omega_1,\omega_2}^{k_1,k_2}, \end{cases}$$

*with $\sigma_{\omega_1,\omega_2}^{k_1,k_2}, \nu_{\omega_1,\omega_2}^{k_1,k_2}$ some non negative integers equal to zero in the non degenerate cases (see Remark 5.1.8 for details).*

**Remark 5.1.6.** *We have $e^{-i\omega t} = e^{\Im(\omega)t - i\Re(\omega)t}$, so that all the terms of the sum except maybe the remainder $R$ are of the form $e^{ik\cdot x}P(t)e^{\tilde{\lambda}t+i\beta t}$, with $P(t) \in \mathbb{C}[t]$. If one of this term satisfies $\tilde{\lambda} < \lambda$, it can be put in the remainder term $R$.*

**Remark 5.1.7.** *Taking $\lambda$ decreasing to $-\infty$ makes the sum larger, but it always remain finite, for a fixed $\lambda$, since $K, K+K$ are finite together with the zeros (see Remark 5.1.3). We warn the reader that, a priori, the expansion does not converge as $\lambda$ goes to to $-\infty$.*

**Remark 5.1.8.** *It may exist some degenerate cases for which the four types of functions introduced in the second part of Theorem 5.1.5 are non distinct. In such a case, the numbers $\sigma^{k_1,k_2}_{\omega_1,\omega_2}$ and $\nu^{k_1,k_2}_{\omega_1,\omega_2}$ do not vanish and we have*

$$\sigma^{k_1,k_2}_{\omega_1,\omega_2} = (n_{k,\omega_1+\omega_2} - 1)\mathbb{1}_{D_k(\omega_1+\omega_2)=0} + 2\cdot\mathbb{1}_{d^{k_2}_{k_1}\neq 0 \text{ and } \omega_1+\omega_2=\frac{|k|}{|k_1|}\omega_1}$$

*and*

$$\nu^{k_1,k_2}_{\omega_1,\omega_2} = (n_{k,\frac{|k|}{|k_1|}\omega_1} - 1)\mathbb{1}_{D_k(\frac{|k|}{|k_1|}\omega_1)=0}$$

*where $\mathbb{1}_P$ denotes the characteristic function of the property $P$.*

**Remark 5.1.9.** *In the case where $d^{k_2}_{k_1} \neq 0$, which we will call* resonant case*, where the* Best frequency*, that is the term $B$ appears, $p$ can a priori be $\geqslant 1$. For the $J$ and $I$ terms, the multiplicity can be equal to one, corresponding to $m = \ell = 0$.*

The fifth section of this paper is devoted to some numerical experiments. They principally aim at highlighting the Best's waves because most of the other phenomena associated with second order terms have been studied numerically in the proceedings [20]. Unlike the linear case, it seems that there is no elementary way to determine *a priori* the coefficients associated with the asymptotic expansion of $\mu$. Indeed, they depend non trivially on the solution of (VPL) (and not only on its asymptotic expansion). Consequently, we use here least squares procedures, which permit to have a simple and quick way to find these coefficients.

There are some difficulty arising of these computations because, as we compare the solution of the second order expansion to the solution of (VP), this gives a constraint on $\varepsilon$ and the final time that should be small enough. As we have seen, the final time should also not be too small in order to be in the asymptotic regime, and this is also true for $\varepsilon$ (which is here put to the square, as we consider second order expansion) due to the limits imposed by machine precision.

We admit that for the numerical checking of codes, second order terms have not gained much popularity, maybe as the linear terms generally already give the main phenomena. We emphasize that we are here able to identify the contributions of the different frequencies, and thus do an effective comparison with, as already told, multidimensional and nonlinear features.

***Some remarks about the notations*** In order to keep proofs as readable as possible, we do some classical abuses of notation for integral transforms. For example, the Fourier transform on $\mathbb{R}^d$ is always associated with the variable $v$, it means that if $u \in L^1(\mathbb{R}^d)$ then $\mathscr{F}[u]$ and $\mathscr{F}[u(v)]$ denotes the same functions. Similarly, if $u$ is a function of $t, x, v$ then $\mathscr{F}[u](t,x,\xi)$ denotes $\mathscr{F}[v \mapsto u(t,x,v)](t,x,\xi)$. Similarly, $t$ is associated with $\mathcal{L}$, $z$ with $\mathcal{L}^{-1}$, $\xi$ with $\mathscr{F}^{-1}$, $x$ with $u \mapsto \widehat{u}$ and $k$ with $(u \mapsto \widehat{u})^{-1}$.

***Outline of the work*** In Section 5.2, we derive some integral equations (called dispersion relations) satisfied by solutions $\psi, \mu$ of (VPL) and (VPL2). Then we prove that it is enough to solve these dispersion relations to get solutions for (VPL) and (VPL2). The next two sections are devoted to the resolution of these dispersion relations and to the asymptotic expansions of their solutions : Section 5.3 is for the first order expansion and Section 5.4 is for the second order expansion. Finally in Section 5.5, we give some numerical results.

## 5.2 Derivation of the dispersion relations

### 5.2.1 Dispersion relations for first and second order

In the following propositions, we give the *dispersion relations*, that are obtained through Fourier and Laplace transforms. Note that we have an expression for both the electric potentials $\psi$ resp. $\mu$ and the distribution function $g$ resp. $h$ of the first resp. second order dispersion relations.

**Proposition 5.2.1.** *Assume $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $g_0$ satisfies Assumption 1. Assume there exists a $C^\infty$ function on $\mathbb{R}_+^* \times \mathbb{T}_d$, denoted $\psi$, and there exists $\lambda_0 > 0$ such that $e^{-\lambda_0 t}\psi(t,x)$ is bounded on $\mathbb{R}_+^* \times \mathbb{T}_d$. Furthermore, assume that, for all $k \in \widehat{\mathbb{T}_d}\backslash\{0\}$, $\mathcal{L}\left[\widehat{\psi}(t,k)\right](z)$ is a solution of*

$$\mathcal{L}\left[\widehat{\psi}(t,k)\right](z)D_k(z) = -\frac{i}{|k|^2}\int_{\mathbb{R}^d}\frac{\widehat{g_0}(k,v)}{v\cdot k - z}\mathrm{d}v. \tag{5.11}$$

*for $\Im z > \lambda_0$.*
*If we define $g$ by*

$$g(t,x,v) = \sum_{k\in K} e^{ik\cdot(x-vt)}\widehat{g_0}(k,v) + i\int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\psi}(s,k)k\cdot\nabla_v f^{eq}(v)\mathrm{d}s, \tag{5.12}$$

*then $g$ is a $C^\infty$ function on $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$, continuous on $\mathbb{R}_+ \times \mathbb{T}_d \times \mathbb{R}^d$ and $(g,\psi)$ is a solution of* (VPL)*.*

**Proposition 5.2.2.** *Assume $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $g_0$ satisfies Assumption 1. Assume there exists a solution of* (VPL) *as in Proposition* (5.2.1)*. Assume there exists a $C^\infty$ function on $\mathbb{R}_+^* \times \mathbb{T}_d$, denoted $\mu$, and there exists $\lambda_1 > 2\lambda_0$ such that $e^{-\lambda_1 t}\psi(t)$ is bounded on $\mathbb{R}_+^* \times \mathbb{T}_d$. Furthermore, assume that, for all $k \in \widehat{\mathbb{T}_d}\backslash\{0\}$, $\mathcal{L}\left[\widehat{\mu}(t,k)\right](z)$ is a solution of*

$$\mathcal{L}\left[\widehat{\mu}(t,k)\right](z)D_k(z) = -\frac{i}{|k|^2}\int_{\mathbb{R}^d}\frac{\mathcal{L}\left[\widehat{\nabla_x\psi\cdot\nabla_v g(t,k,v)}\right](z)}{v\cdot k - z}\mathrm{d}v. \tag{5.13}$$

*for $\Im z > \lambda_1$.*
*If we define $h$ by*

$$h(t,x,v) = \sum_{k\in K+K} i\int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\mu}(s,k)k\cdot\nabla_v f^{eq}(v)\mathrm{d}s + \int_0^t \widehat{\nabla_x\psi\cdot\nabla_v g(s,k,v)}e^{ik\cdot(x-v(t-s))}\mathrm{d}s, \tag{5.14}$$

*then $h$ is a $C^\infty$ function on $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$, continuous on $\mathbb{R}_+ \times \mathbb{T}_d \times \mathbb{R}^d$ and $(h,\mu)$ is a solution of* (VPL2)*.*

### 5.2.2 A general linearized Vlasov-Poisson equation

In order to prove Propositions 5.2.1 and 5.2.2, as (VPL) and (VPL2) share the same structure, we focus on a general linearized Vlasov Poisson equation

$$\begin{cases} \partial_t \mathfrak{g}(t,x,v) + v\cdot\nabla_x \mathfrak{g}(t,x,v) - \nabla_x \mathfrak{u}(t,x)\cdot\nabla_v f^{eq}(v) = \mathfrak{S}(t,x,v), \\ \Delta_x \mathfrak{u}(t,x) + \int \mathfrak{g}(t,x,v)\mathrm{d}v = 0, \\ \mathfrak{g}(0,x,v) = \mathfrak{g}_0(x,v). \end{cases} \tag{VPLG}$$

In the following proposition, we derive a general dispersion relation satisfied by $\mathfrak{u}$. We first do not consider the coupling with the Poisson equation.

**Proposition 5.2.3.** *Assume $\mathfrak{g}_0 \in C^1(\mathbb{T}_d\times\mathbb{R}^d)$, $f^{eq}\in C^2(\mathbb{R}^d)$ and $\mathfrak{S}(t,x,v)\in C^1(\mathbb{R}_+^*\times\mathbb{T}_d\times\mathbb{R}^d)$ and there exist $\lambda_0>0$, $\mathfrak{d}\in C^0(\mathbb{R}^d)\cap L^1(\mathbb{R}^d)$ satisfying*

$$\forall (t,k,v)\in\mathbb{R}_+^*\times\widehat{\mathbb{T}_d}\times\mathbb{R}^d,\ e^{-\lambda_0 t}|\widehat{\mathfrak{S}}(t,k,v)| + |\widehat{\mathfrak{g}_0}(k,v)| + |\nabla_v f^{eq}(v)| \leqslant \mathfrak{d}(v).$$

*Assume there exists $\mathfrak{u}\in C^1(\mathbb{R}_+^*\times\mathbb{T}_d)$ such that $e^{-\lambda_0 t}\mathfrak{u}(t)$ is bounded on $\mathbb{R}_+^*\times\mathbb{T}_d$. Assume there exists a continuous function $\mathfrak{g}\in C^1(\mathbb{R}_+^*\times\mathbb{T}_d\times\mathbb{R}^d)$, continuous on $\mathbb{R}_+\times\mathbb{T}_d\times\mathbb{R}^d$ such that $\mathfrak{g}$ is solution of the Vlasov equation*

$$\forall (t,x,v)\in\mathbb{R}_+^*\times\mathbb{T}_d\times\mathbb{R}^d,\ \begin{cases} \partial_t \mathfrak{g}(t,x,v) + v\cdot\nabla_x \mathfrak{g}(t,x,v) - \nabla_x \mathfrak{u}(t,x)\cdot\nabla_v f^{eq}(v) = \mathfrak{S}(t,x,v), \\ \mathfrak{g}(0,x,v) = \mathfrak{g}_0(x,v). \end{cases} \tag{5.15}$$

*If $\lambda>\lambda_0$ then for all $k\in\widehat{\mathbb{T}_d}$, there exists $C>0$,*

$$\forall v\in\mathbb{R}^d,\ \sup_{t\in\mathbb{R}_+}|e^{-\lambda t}\widehat{\mathfrak{g}}(t,k,v)| \leqslant C\mathfrak{d}(v). \tag{5.16}$$

*Furthermore, for all $z\in\mathbb{C}$ with $\Im(z)>\lambda_0$ we have*

$$\mathcal{L}\left[\int_{\mathbb{R}^d}\widehat{\mathfrak{g}}(t,k,v)\mathrm{d}v\right](z) = -i\int_{\mathbb{R}^d}\frac{\widehat{\mathfrak{g}_0}(k,v)}{v\cdot k - z}\mathrm{d}v + \mathcal{L}[\widehat{\mathfrak{u}}(t,k)](z)\int_{\mathbb{R}^d}\frac{k\cdot\nabla_v f^{eq}(v)}{v\cdot k - z}\mathrm{d}v - i\int_{\mathbb{R}^d}\frac{\mathcal{L}\left[\widehat{\mathfrak{S}}(t,k,v)\right](z)}{v\cdot k - z}\mathrm{d}v. \tag{5.17}$$

*Proof.* First, applying a space Fourier transform to (5.15), we get for all $(t,k,v)\in\mathbb{R}_+^*\times\widehat{\mathbb{T}_d}\times\mathbb{R}^d$

$$\partial_t\widehat{\mathfrak{g}}(t,k,v) + iv\cdot k\,\widehat{\mathfrak{g}}(t,k,v) - i\widehat{\mathfrak{u}}(t,k)k\cdot\nabla_v f^{eq}(v) = \widehat{\mathfrak{S}}(t,k,v). \tag{5.18}$$

Consequently, applying Duhamel formula, we get for all $(t,k,v)\in\mathbb{R}_+\times\widehat{\mathbb{T}_d}\times\mathbb{R}^d$

$$\widehat{\mathfrak{g}}(t,k,v) = e^{-ik\cdot vt}\widehat{\mathfrak{g}_0}(k,v) + i\int_0^t e^{-ik\cdot v(t-s)}\widehat{\mathfrak{u}}(s,k)k\cdot\nabla f^{eq}(v)\mathrm{d}s + \int_0^t e^{-ik\cdot v(t-s)}\widehat{\mathfrak{S}}(s,k,v)\mathrm{d}s. \tag{5.19}$$

So, we deduce, there exist $M, C > 0$ such that, if $\lambda > \lambda_0$ then

$$\begin{aligned}|\widehat{\mathfrak{g}}(t,k,v)| &\leqslant |\widehat{\mathfrak{g}_0}(k,v)| + \int_0^t e^{\lambda_0 s} M|k||\nabla_v f^{eq}(v)|\mathrm{d}s + \int_0^t e^{\lambda_0 s} e^{-\lambda_0 s}|\widehat{\mathfrak{S}}(s,k,v)|\mathrm{d}s,\\ &\leqslant \mathfrak{d}(v) + te^{\lambda_0 t}(1+|k|M)\mathfrak{d}(v),\\ &\leqslant Ce^{\lambda t}\mathfrak{d}(v).\end{aligned}$$

We deduce of this last estimation, that for any fixed $v \in \mathbb{R}^d$ and for any $\lambda > \lambda_0$, $e^{-\lambda t}\widehat{g}(t,k,v)$ is continuous and bounded on $\mathbb{R}_+$. Consequently, we can apply a Laplace transform on (5.18) and get for all $z \in \mathbb{C}$ such that $\Im z > \lambda_0$ and $v \in \mathbb{R}^d$,

$$\begin{aligned}-iz\,\mathcal{L}[\widehat{\mathfrak{g}}(t,k,v)](z) - \widehat{\mathfrak{g}_0}(k,v) + iv\cdot k\,\mathcal{L}[\widehat{\mathfrak{g}}(t,k,v)](z) - i\,\mathcal{L}[\widehat{\mathfrak{u}}(t,k)](z)k\cdot\nabla_v f^{eq}(v)\\ = \mathcal{L}\left[\widehat{\mathfrak{S}}(t,k,v)\right](z).\end{aligned}$$

Since $\Im z > 0$, this relation can be divided by $i(v\cdot k - z)$ to get for all $v \in \mathbb{R}^d$

$$\mathcal{L}[\widehat{\mathfrak{g}}(t,k,v)](z) = -i\frac{\widehat{\mathfrak{g}_0}(k,v)}{v\cdot k - z} + \mathcal{L}[\widehat{\mathfrak{u}}(t,k)](z)\frac{k\cdot\nabla_v f^{eq}(v)}{v\cdot k - z} - i\frac{\mathcal{L}\left[\widehat{\mathfrak{S}}(t,k,v)\right](z)}{v\cdot k - z}.$$

Finally we conclude this proof integrating with respect to $v$ and applying Fubini Theorem (with the control (5.16)) to get for all $z \in \mathbb{C}$ with $\Im z > \lambda_0$

$$\mathcal{L}\left[\int_{\mathbb{R}^d}\widehat{\mathfrak{g}}(t,k,v)\mathrm{d}v\right](z) = \int_{\mathbb{R}^d}\mathcal{L}\left[\widehat{\mathfrak{g}}(t,k,v)\right](z)\mathrm{d}v.$$

□

If we want to get a closed equation on $\mathfrak{u}$, we have to use Poisson equation

$$\Delta_x\mathfrak{u}(t,x) = -\int_{\mathbb{R}^d}\mathfrak{g}(t,x,v)\mathrm{d}v. \tag{5.20}$$

Formally, applying a space Fourier transform and a Laplace transform we would get

$$|k|^2\,\mathcal{L}\left[\widehat{\mathfrak{u}}(t,k)\right] = \mathcal{L}\left[\int_{\mathbb{R}^d}\widehat{\mathfrak{g}}(t,k,v)\mathrm{d}v\right].$$

Consequently, applying (5.17), we should get the following dispersion relation

$$\mathcal{L}\left[\widehat{\mathfrak{u}}(t,k)\right](z)D_k(z) = -\frac{i}{|k|^2}\int_{\mathbb{R}^d}\frac{\widehat{\mathfrak{g}_0}(k,v)}{v\cdot k - z}\mathrm{d}v - \frac{i}{|k|^2}\int_{\mathbb{R}^d}\frac{\mathcal{L}\left[\widehat{\mathfrak{S}}(t,k,v)\right](z)}{v\cdot k - z}\mathrm{d}v, \tag{5.21}$$

where $D_k$ is defined by (5.3).

**Proposition 5.2.4.** *Assume* $\mathfrak{g}_0 \in C^1(\mathbb{T}_d\times\mathbb{R}^d)$, $f^{eq} \in C^2(\mathbb{R}^d)$ *and* $\mathfrak{S}(t,x,v) \in C^1(\mathbb{R}_+^*\times\mathbb{T}_d\times\mathbb{R}^d)$ *and there exist* $\lambda_0 > 0$, $\mathfrak{d} \in C^0(\mathbb{R}^d)\cap L^1(\mathbb{R}^d)$ *satisfying*

$$\forall(t,k,v) \in \mathbb{R}_+^*\times\widehat{\mathbb{T}_d}\times\mathbb{R}^d,\ e^{-\lambda_0 t}|\widehat{\mathfrak{S}}(t,k,v)| + |\widehat{\mathfrak{g}_0}(k,v)| + |\nabla_v f^{eq}(v)| \leqslant \mathfrak{d}(v).$$

*Assume there exists* $\mathfrak{u} \in C^1(\mathbb{R}_+^* \times \mathbb{T}_d)$ *such that* $e^{-\lambda_0 t}\mathfrak{u}(t)$ *is bounded on* $\mathbb{R}_+^* \times \mathbb{T}_d$*. Furthermore, assume that, for all* $k \in \widehat{\mathbb{T}_d}\backslash\{0\}$*,* $\mathcal{L}\left[\widehat{\mathfrak{u}}(t,k)\right](z)$ *is a solution of* (5.21) *for* $\Im z > \lambda_0$*. Assume there exists a finite set* $\mathfrak{K} \subset \widehat{\mathbb{T}_d}$ *such that*

$$\forall t \in \mathbb{R}, \forall v \in \mathbb{R}^d,\ k \in \widehat{\mathbb{T}_d}\backslash\mathfrak{K} \ \Rightarrow\ \widehat{\mathfrak{g}_0}(k,v) = \widehat{\mathfrak{S}}(t,k,v) = \widehat{\mathfrak{u}}(t,k) = 0.$$

*If we define* $\mathfrak{g}$ *by*

$$\mathfrak{g}(t,x,v) = \sum_{k\in\mathfrak{K}} e^{ik\cdot(x-vt)}\widehat{\mathfrak{g}_0}(k,v) + i\int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\mathfrak{u}}(s,k)k\cdot\nabla f^{eq}(v)\mathrm{d}s + \int_0^t e^{ik\cdot(x-v(t-s))}\widehat{\mathfrak{S}}(s,k,v)\mathrm{d}s, \quad (5.22)$$

*then* $\mathfrak{g} \in C^1(\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d)$ *is continuous on* $\mathbb{R}_+ \times \mathbb{T}_d \times \mathbb{R}^d$ *and* $(\mathfrak{g},\mathfrak{u})$ *is a solution of* (VPLG)*.*

*Proof.* By construction of $\mathfrak{g}$ through Duhamel formula (5.22), $\mathfrak{g}$ is obviously a continuous function on $\mathbb{R}_+ \times \mathbb{T}_d \times \mathbb{R}^d$ and $C^1$ on $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$. Furthermore, we may verify by a straightforward calculation that $\mathfrak{g}$ is solution of the Vlasov equation (5.15). Consequently, we just have to prove that $\mathfrak{g},\mathfrak{u}$ is solution of Poisson equation (5.20).

However $\mathfrak{u}$ and $\mathfrak{g}$ satisfy assumptions of Proposition 5.2.3, so we can apply it. Consequently, we know that if $\lambda > \lambda_0$ then $e^{-\lambda t}\int\widehat{\mathfrak{g}}(t,k,v)\mathrm{d}v$ is continuous and bounded and that its Laplace transform satisfies (5.17). But since $\mathcal{L}[\widehat{\mathfrak{u}}(t,k)]$ is a solution of the dispersion relation (5.11), we deduce that for all $z \in \mathbb{C}$ such that $\Im z > \lambda_0$ we have

$$|k|^2\,\mathcal{L}\left[\widehat{\mathfrak{u}}(t,k)\right](z) = \mathcal{L}\left[\int_{\mathbb{R}^d}\widehat{g}(t,k,v)\mathrm{d}v\right](z).$$

But it is well known that Laplace transform is injective on continuous functions with an exponential order (i.e. bounded by an exponential function), see Theorem $1.7.3$ in [6]. Consequently, we have for all $t > 0$

$$|k|^2\widehat{\mathfrak{u}}(t,k) = \int_{\mathbb{R}^d}\widehat{\mathfrak{g}}(t,k,v)\mathrm{d}v.$$

Since space Fourier transform is also injective on regular functions, we have proven that $\mathfrak{u},\mathfrak{g}$ is a solution of Poisson equation (5.20). □

### 5.2.3 Proof of Propositions 5.2.1 and 5.2.2

We now apply Proposition 5.2.4 for the proof of Propositions 5.2.1 and 5.2.2.

*Proof of Proposition 5.2.1.* First, observe that if $k \in \widehat{\mathbb{T}_d}\backslash(\{K\}\cup\{0\})$ then for any $t > 0$, we have $\widehat{\psi}(t,k) = 0$. Indeed, since $\mathcal{L}\left[\widehat{\psi}(t,k)\right](z)$ is a solution of (5.11), we have

$$D_k(z)\,\mathcal{L}\left[\widehat{\psi}(t,k)\right](z) = 0.$$

But, we have proven in Lemma 5.3.2 that $D_k(z) \neq 0$ if $\Im z$ is large enough. Consequently, $\mathcal{L}\left[\widehat{\psi}(t,k)\right](z) = 0$ if $\Im z$ is large enough. So we deduce by a classical criterion about Laplace transform (see Theorem $1.7.3$ in [6]) that $\widehat{\psi}(t,k) = 0$.

We observe on (5.12) that $g$ is clearly a $C^\infty$ function on $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$. Finally, we just need to apply Proposition 5.2.4 to prove that $(g,\psi)$ is a solution of (VPL). □

*Proof of Proposition 5.2.2.* Let $\mathfrak{S}$ be defined by

$$\mathfrak{S}(t,x,v) = \nabla_x \psi(t,x) \cdot \nabla_v g(t,x,v).$$

By construction, it is a $C^\infty$ function on $\mathbb{R}_+^* \times \mathbb{T}_d \times \mathbb{R}^d$. Since, space Fourier transform of $\psi$ is supported by $K$ (see proof of Proposition 5.2.1), its space Fourier transform is supported by $K+K$. Furthermore, since $\lambda_1 > 2\lambda_0$, we can construct, by a straightforward estimation, a continuous function $\mathfrak{d} \in C^0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ such that

$$\forall v \in \mathbb{R}^d,\ e^{-\lambda_1 t}|\widehat{\mathfrak{S}}(t,k,v)| \leqslant \mathfrak{d}(v).$$

In particular, this estimation proves that the right member of (5.13) is well defined if $\Im z \geqslant \lambda_1$.

Now, as in Proposition 5.2.1, we can first prove that space Fourier transform of $\mu$ is supported by $(K+K) \cup \{0\}$, then observe that $h$ is a $C^\infty$ function and conclude that $(h,\mu)$ is a solution of (VPL) by Proposition 5.2.4. □

## 5.3 Resolution and expansion of the linearized equation

### 5.3.1 Introduction and statement of the result

In Proposition 5.2.1, we have proved that it is enough to solve dispersion relation (5.11) to get a solution $(g,\psi)$ to linearized Vlasov-Poisson equation (VPL). So the aim of this section is to solve this dispersion relation introducing most of the theoretical tools useful in the resolution of the second order relation (5.13). In particular, many of them deal with analytic function defined on *open sectors*, denoted $\Sigma_\alpha$, with $\alpha \in (0,\pi)$, and defined by

$$\Sigma_\alpha = \{re^{i\beta} \mid -\alpha < \beta < \alpha \text{ and } r > 0\}.$$

The result we are going to establish in this section is the following.

**Proposition 5.3.1.** *Assume $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $g_0$ satisfies Assumption 1. For all $\lambda \in \mathbb{R}$, for all $k \in K$, for all zero point $\omega$ of $D_k$ there exists a polynomial, denoted $P_{k,\omega}$, whose degree is strictly smaller than the multiplicity of $\omega$, $\alpha \in (0,\frac{\pi}{2})$ and there exists $r_{k,\lambda}$ an analytic and bounded function on $\Sigma_\alpha$ such that the following expansion defines a solution of the dispersion relation* (5.11)

$$\forall t \in \mathbb{R}_+^*, \forall x \in \mathbb{T}_d,\ \psi(t,x) = \sum_{k\in K} \sum_{\substack{D_k(\omega)=0 \\ \Im\omega \geqslant \lambda}} P_{k,\omega}(t) e^{i(k\cdot x - \omega t)} + e^{ik\cdot x} e^{\lambda t} r_{k,\lambda}(t).$$

This proposition will be proven at the end of this section. First, we introduce some notations and many useful theoretical tools.

### 5.3.2 Definition of $N_k$ and theoretical tools

The right member of the dispersion relation (5.11) is very important in our study. We denote it $N_k(z)$. More precisely, it is an analytic function defined, when $\Im z > 0$ by

$$N_k(z) = -\frac{i}{|k|^2}\int \frac{\widehat{g_0}(k,v)}{v\cdot k - z}\mathrm{d}v.$$

In the first part of this proof we study the regularity and the behavior of $D_k$ and $N_k$. However, we need to introduce some classical results on Laplace transform.

First, consider the following Theorem that is very useful to invert Laplace transforms and to control it.

**Theorem 5.3.1.** *(Analytic representation)*
*Let $\alpha \in (0, \frac{\pi}{2})$, $\lambda_0 \in \mathbb{R}$ and $q : i(\lambda_0, \infty) \to \mathbb{C}$. The following assertions are equivalent :*

*(i) There exists a holomorphic function $f : \Sigma_\alpha \to \mathbb{C}$ such that*

$$\forall 0 < \beta < \alpha, \sup_{z\in\Sigma_\beta} |e^{-\lambda_0 z} f(z)| < \infty \text{ and } \forall \lambda > \lambda_0,\ q(i\lambda) = \mathcal{L}[f](i\lambda).$$

*(ii) The function $q$ has a holomorphic extension $\widetilde{q} : i\lambda_0 + i\Sigma_{\alpha+\frac{\pi}{2}} \to \mathbb{C}$ such that*

$$\forall 0 < \gamma < \alpha, \sup_{z\in i\lambda_0 + i\Sigma_{\gamma+\frac{\pi}{2}}} |(z - i\lambda_0)\widetilde{q}(z)| < \infty.$$

*Proof.* See Theorem $2.6.1$ in [6] page $87$. □

**Remark 5.3.2.** *Note that if $e^{-\lambda_0 t} f(t)$ is bounded on $\mathbb{R}_+^*$ then for $\lambda > \lambda_0$, $e^{-\lambda t} f(t) \in L^1(\mathbb{R}_+)$ and so $\mathcal{L}[f]$ is well defined for $\Im(z) > \lambda_0$, which is the set $i\lambda_0 + i\Sigma_{\frac{\pi}{2}}$.*

There is a direct corollary of the proof of Theorem 5.3.1 that is useful in our study.

**Corollary 5.3.1.** *Assume that conclusion of Theorem 5.3.1 holds. Then for all $0 < \gamma < \beta < \alpha$, we have*

$$\sup_{z\in i\lambda_0 + i\Sigma_{\gamma+\frac{\pi}{2}}} |(z - i\lambda_0)\widetilde{q}(z)| \leqslant \frac{1}{\sin(\beta-\gamma)} \sup_{z\in\Sigma_\beta} |e^{-\lambda_0 z} f(z)|.$$

Then, we observe that $D_k$ and $N_k$ are defined through a integral operator whose kernel is $\frac{1}{v\cdot k - z}$. The following lemma links this operator to more classical ones.

**Lemma 5.3.3.** *Let $f \in L^1(\mathbb{R}^d)$ and $k \in \mathbb{R}^d\backslash\{0\}$ then for all $z \in \mathbb{C}$ with $\Im z > 0$*

$$\int_{\mathbb{R}^d} \frac{f(v)}{k\cdot v - z}\mathrm{d}v = i\,\mathcal{L}\left[\mathscr{F}[f](kt)\right](z).$$

*Proof.* First, remark that since $f \in L^1(\mathbb{R}^d)$, $t \to \mathscr{F}[f](kt) = \int_{\mathbb{R}^d} f(v)e^{-itv\cdot k}\mathrm{d}v$ is a continuous and bounded function, so its Laplace transform is well defined if $\Im z > 0$. Now, consider the following function

$$F(t) = \int_{\mathbb{R}^d} \frac{f(v)}{k\cdot v - z} e^{-i(k\cdot v - z)t}\mathrm{d}v.$$

Since $f \in L^1(\mathbb{R}^d)$, it is a regular function and we have

$$F'(t) = -i \int_{\mathbb{R}^d} f(v) e^{-i(k\cdot v - z)t} \mathrm{d}v = -i\, \mathscr{F}[f](kt) e^{izt}.$$

But, since $\Im z > 0$ we observe that $F(t)$ goes to $0$ when $t$ goes to $+\infty$. Consequently, we get

$$F(0) = -\int_0^\infty F'(t)\mathrm{d}t = i \int_0^\infty \mathscr{F}[f](kt) e^{izt} \mathrm{d}t = i\, \mathcal{L}\left[\mathscr{F}[f](kt)\right](z).$$

□

### 5.3.3 Estimations for $D_k$ and $N_k$

With Lemma 5.3.3, we can write $D_k$ and $N_k$ as Laplace transforms. So, in the following proposition, we can prove their analyticity using the analytic representation theorem (Theorem 5.3.1). In particular, we prove and extend Remark 5.1.3.

**Proposition 5.3.2.** *If $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ then*

- *for all $k \in \mathbb{R}^d \backslash \{0\}$, $D_k$ is an entire function,*
- *there exists $\alpha \in (0, \frac{\pi}{2})$ such that for all $0 < \gamma < \alpha$ and for all $\lambda_0 \in \mathbb{R}$ there exists $C > 0$ satisfying*

$$\forall k \in \mathbb{R}^d \backslash \{0\}, \forall z \in i|k|\lambda_0 + i\Sigma_{\gamma + \frac{\pi}{2}},\ |D_k(z) - 1| \leqslant \frac{C}{|k||z - i|k|\lambda_0|}.$$

*Proof.* Since $f \in \mathscr{S}(\mathbb{R}^d)$, we have $k \cdot \nabla_v f^{eq} \in L^1(\mathbb{R}^d)$. Consequently, $D_k$ is well defined as an analytic function on $\{z \in \mathbb{C} \mid \Im z > 0\}$. Furthermore, we can apply Lemma 5.3.3 to get for $\Im z > 0$

$$D_k(z) = 1 - \frac{i}{|k|^2} \mathcal{L}\left[\mathscr{F}\left[k \cdot \nabla_v f^{eq}\right](kt)\right](z).$$

Then we define $e_k = \frac{k}{|k|}$ and we get by the change of variable $t' = |k|t$

$$\begin{aligned}
&\mathcal{L}\left[\mathscr{F}\left[k \cdot \nabla_v f^{eq}\right](kt)\right](z) \\
&= \int_0^\infty \mathscr{F}\left[|k| e_k \cdot \nabla_v f^{eq}\right](|k| e_k t) e^{i\frac{z}{|k|}|k|t} \mathrm{d}t \\
&= \int_0^\infty \mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k t') e^{i\frac{z}{|k|}t'} \mathrm{d}t',
\end{aligned}$$

so that

$$D_k(z) = 1 - \frac{i}{|k|^2} \mathcal{L}\left[\mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k t)\right]\left(\frac{z}{|k|}\right). \tag{5.23}$$

Now, using Theorem 5.3.1, we are going to prove this Laplace transform defines an entire function and we are going to control it.

Since $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$, it extends to an analytic function and there exists $\alpha \in (0, \frac{\pi}{2})$ such that for all $\beta \in (0, \alpha)$ and for all $\lambda \in \mathbb{R}$, there exist $C > 0$ such that

$$\forall z \in \Sigma_\beta \mathbb{R}^d,\ |\mathscr{F}[f^{eq}](z)| \leqslant C|e^{-\lambda z}|.$$

Consequently, we get

$$\forall z \in \Sigma_\beta,\ |\mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k z)| = |iz\,\mathscr{F}[f^{eq}](z)| \leqslant C|z|e^{-\lambda \Re z}.$$

Finally, we have proven that for all $\lambda_0 \in \mathbb{R}$, there exists a constant $M > 0$ (independent of $e_k$) such that

$$\forall z \in \Sigma_\beta,\ |\mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k z)| \leqslant M|e^{-\lambda_0 z}|.$$

Applying Theorem 5.3.1 and its corollary, we have proven that $\mathcal{L}[\mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k t)]$ is an entire function and that for all $\gamma \in (0, \alpha)$ and all $\lambda_0 \in \mathbb{R}$, there exists $M > 0$ (associated to $\beta = \frac{\alpha+\gamma}{2}$) such that

$$\forall z \in i\lambda_0 + i\Sigma_{\gamma+\frac{\pi}{2}},\ |\mathcal{L}\left[\mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k t)\right](z)| \leqslant \frac{M}{\sin(\frac{\alpha-\gamma}{2})} \frac{1}{|z - i\lambda_0|}.$$

Finally, we deduce directly the result from formula (5.23) :

$$\begin{aligned} |D_k(z) - 1| &= \left| \frac{i}{|k|^2} \mathcal{L}\left[\mathscr{F}\left[e_k \cdot \nabla_v f^{eq}\right](e_k t)\right]\left(\frac{z}{|k|}\right) \right| \\ &\leqslant \frac{M}{|k|^2 \sin(\frac{\alpha-\gamma}{2})} \frac{1}{|\frac{z}{|k|} - i\lambda_0|} = \frac{M}{|k| \sin(\frac{\alpha-\gamma}{2})} \frac{1}{|z - i|k|\lambda_0|}. \end{aligned}$$

□

**Corollary 5.3.2.** *If $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ then there exists $\alpha \in (0, \frac{\pi}{2})$ such that for all $\lambda_0 \in \mathbb{R}$ and $\gamma \in (0, \alpha)$, there exists $c > 0$ such that for all $k \in \mathbb{R}^d \backslash \{0\}$ we have*

$$\{z \in \mathbb{C} \mid D_k(z) = 0\} \subset i|k|\lambda_0 - \left(\mathbb{D}(0, \frac{c}{|k|}) \cup \overline{i\Sigma_{\frac{\pi}{2}-\gamma}}\right).$$

*Proof.* Indeed, we have either $z \in i|k|\lambda_0 + i\Sigma_{\gamma+\frac{\pi}{2}}$, so that $1 = |D_k(z) - 1| \leqslant \frac{C}{|k||z - i|k|\lambda_0|}$ and thus $z \in i|k|\lambda_0 - \mathbb{D}(0, \frac{C}{|k|})$. Otherwise, we have $z \in \mathbb{C} \backslash \left\{i|k|\lambda_0 + i\Sigma_{\gamma+\frac{\pi}{2}}\right\}$, that is $z = i|k|\lambda_0 + ire^{i\delta}$, $\delta \in [\frac{\pi}{2} + \gamma, 2\pi - \frac{\pi}{2} - \gamma]$, and thus $\pi - \delta \in [-\frac{\pi}{2} + \gamma, -\gamma + \frac{\pi}{2}]$, which leads to $z = i|k|\lambda_0 - ire^{-i\tilde{\delta}}$, with $\tilde{\delta} = \pi - \delta \in [-\frac{\pi}{2} + \gamma, -\gamma + \frac{\pi}{2}]$. □

**Corollary 5.3.3.** *If $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ then for all $k \in \widehat{\mathbb{T}_d} \backslash \{0\}$, $D_k$ is an entire function and for all $\lambda \in \mathbb{R}$, $\{\omega \in \mathbb{C} \mid D_k(\omega) = 0$ and $\Im\omega \geqslant \lambda\}$ is a finite set.*

*Proof.* In Proposition 5.3.2, we have proved that $D_k$ is an entire function and it can be directly deduced from its Corollary 5.3.2 that $\{z \in \mathbb{C} \mid D_k(z) = 0$ and $\Im\omega \geqslant \lambda\}$ is bounded. Consequently, since zero points of $D_k$ are isolated, it is a finite set. □

It is very natural to adapt this result to $N_k$. More precisely, we deduce the following proposition.

**Proposition 5.3.3.** *For all $k \in \widehat{\mathbb{T}_d}\backslash\{0\}$, $N_k$ is an entire function and there exists $\alpha \in (0, \frac{\pi}{2})$ such that for all $\lambda_0 \in \mathbb{R}$ and for all $\beta \in (0, \alpha)$, we have*

$$\sup_{z \in i\lambda_0 + i\Sigma_{\beta+\frac{\pi}{2}}} |N_k(z)||z - i\lambda_0| < \infty.$$

### 5.3.4 A theoretical tool for the control of $\mathcal{L}^{-1}[N_k/D_k]$

Now, we introduce a general criterion to invert Laplace transform and get an asymptotic expansion.

**Lemma 5.3.4.** *Let $R \in \mathcal{H}(\mathbb{C})$ be an entire function and $N$ be a meromorphic function defined on $\mathbb{C}$. If there exists $\alpha \in (0, \frac{\pi}{2})$ such that*

$$\exists C > 0, \sup_{z \in i\Sigma_{\alpha+\frac{\pi}{2}}} |zR(z)| + |zN(z)| < C,$$

*then for any $\lambda \in \mathbb{R}$, there exist $\beta \in (0, \alpha)$ and a function $r \in \mathcal{H}(\Sigma_\beta)$ analytic and bounded on $\Sigma_\beta$ such that if $\Im z$ is large enough then*

$$\frac{N(z)}{1 - R(z)} = \mathcal{L}\left[\sum_{\substack{\omega \in Z \\ \Im\omega \geqslant \lambda}} P_\omega(t) e^{-i\omega t} + e^{\lambda t} r(t)\right](z),$$

*where $Z$ is the set of poles of $\frac{N(z)}{1-R(z)}$,*

$$P_\omega = \sum_{k=0}^{n_\omega - 1} \frac{a_{k+1,\omega}(-i)^{k+1}}{k!} X^k$$

*is the polynomial whose coefficients are defined by the expansion of $\frac{N}{1-R}$ in $z = \omega$*

$$\frac{N(z)}{1 - R(z)} \underset{z \to \omega}{=} \sum_{j=1}^{n_\omega} \frac{a_{j,\omega}}{(z - \omega)^j} + \mathcal{O}(1).$$

**Remark 5.3.5.** *In the application, for the proof of Proposition 5.3.1, $N$ will be entire (thus meromorphic), but for the second order case, in the next section, we will really need that $N$ is meromorphic.*

*Proof of Lemma 5.3.4.* Many geometrical objects are going to be introduced in this proof. The reader can refer to Figure 5.1 to an illustration of these constructions.
First observe that to prove the lemma, we can assume that $\lambda$ is negative enough. In particular we assume that $\lambda < -2C$.
By construction, if $|z| \geqslant 2C$ and $z \in \overline{i\Sigma_{\alpha+\frac{\pi}{2}}}$ then $|1 - R(z)| \geqslant 1 - |R(z)| > 1 - \frac{C}{|z|} \geqslant \frac{1}{2}$.

Consequently, all zero points of $(1-R)$ belong to $\mathbb{D}(0,2C)\cup -i\Sigma_{\frac{\pi}{2}-\alpha}$ (note that $-i\Sigma_{\frac{\pi}{2}-\alpha}\backslash\{0\} = \left(\overline{i\Sigma_{\alpha+\frac{\pi}{2}}}\right)^c$). Since the poles of $N$ lie on $-i\Sigma_{\frac{\pi}{2}-\alpha}$, the poles of $\frac{N(z)}{1-R(z)}$ lie on $\mathbb{D}(0,2C)\cup -i\Sigma_{\frac{\pi}{2}-\alpha}$. In particular, the set of its poles with an imaginary part larger than or equal to $\lambda$ is finite (see Figure 5.1).
Now consider the following rational fraction

$$Q(z) = \sum_{\substack{\omega\in Z\\ \Im\omega\geqslant\lambda}} \sum_{j=1}^{n_\omega} \frac{a_{j,\omega}}{(z-\omega)^j}.$$

We introduce $r_1>0$ such that we have

$$\mathbb{D}(0,2C)\cup\left(\{z\in\mathbb{C}\mid \Im z\geqslant\lambda\}\cap -i\Sigma_{\frac{\pi}{2}-\alpha}\right)\subset\mathbb{D}(i\lambda,r_1).$$

Now, we observe that there exists $\beta\in(0,\alpha)$ such that $\frac{N}{1-R}-Q$ is a continuous function on $\overline{i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}}$. Indeed, it is a meromorphic function whose poles lie on $\{z\in\mathbb{C}\mid\Im z<\lambda\}\cap -i\Sigma_{\frac{\pi}{2}-\alpha}$ and are isolated, and thus we can choose such $\beta$ (small enough). Consequently, there exists $M>0$ such that

$$\forall z\in\mathbb{D}(i\lambda,r_1)\cap\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right),\ \left|\frac{N(z)}{1-R(z)}-Q(z)\right|<M\leqslant\frac{Mr_1}{|z-i\lambda|}.$$

Furthermore since $zN(z)$ is bounded on $i\Sigma_{\alpha+\frac{\pi}{2}}$, there exists $M_1>0$ such that

$$\forall z\in i\lambda+i\Sigma_{\beta+\frac{\pi}{2}},\ |N(z)|\leqslant\frac{M_1}{|z-i\lambda|}.$$

Indeed, we distinguish the case $z\in i\Sigma_{\alpha+\frac{\pi}{2}}\cap\mathbb{D}^c(0,2|\lambda|)$, for which there exists $C>0$ such that

$$|N(z)|\leqslant\frac{C}{|z|}\leqslant\frac{C}{|z-i\lambda|}\frac{|z-i\lambda|}{|z|}\leqslant\frac{C}{|z-i\lambda|}\frac{3|\lambda|}{2|\lambda|}\leqslant\frac{M_1}{|z-i\lambda|},$$

and the case $z\in\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right)\cap\left((i\Sigma_{\alpha+\frac{\pi}{2}})^c\cup\mathbb{D}(0,2|\lambda|)\right)$ which is a bounded set ensuring

$$|z-i\lambda||N(z)|\leqslant M_1.$$

Consequently, by construction of $r_1$, we get $|1-R(z)|\geqslant\frac{1}{2}$, when $|z|\geqslant 2C$ and $z\in i\Sigma_{\alpha+\frac{\pi}{2}}$ [so, in particular when $z\in i\Sigma_{\alpha+\frac{\pi}{2}}\cap\mathbb{D}^c(i\lambda,r_1)\cap\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right)$] and $|1-R(z)|\geqslant c$, with $c>0$, when $z\in\left(i\Sigma_{\alpha+\frac{\pi}{2}}\right)^c\cap\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right)$ (which is a bounded set) [so, in particular when $z\in\left(i\Sigma_{\alpha+\frac{\pi}{2}}\right)^c\cap\mathbb{D}^c(i\lambda,r_1)\cap\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right)$] and thus

$$\forall z\in\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right)\cap\mathbb{D}^c(i\lambda,r_1),\ \left|\frac{N(z)}{1-R(z)}\right|\leqslant\frac{\max(2,1/c)M_1}{|z-i\lambda|}.$$

Finally, since $Q$ is a rational fraction whose poles lie on $\mathbb{D}(i\lambda,r_1)$ and vanishing as $z$ goes to $\infty$, the function $z\to(z-i\lambda)Q(z)$ is bounded on $\mathbb{D}^c(i\lambda,r_1)$ and thus there exists $M_2>0$ such that

$$\forall z\in\left(i\lambda+i\Sigma_{\beta+\frac{\pi}{2}}\right)\cap\mathbb{D}^c(i\lambda,r_1),\ |Q(z)|\leqslant\frac{M_2}{|z-i\lambda|}.$$

Then we get a constant $M_3 > 0$ such that

$$\forall z \in i\lambda + i\Sigma_{\beta+\frac{\pi}{2}}, \ \left| \frac{N(z)}{1-R(z)} - Q(z) \right| < \frac{M_3}{|z - i\lambda|}.$$

Applying Theorem 5.3.1 to $\frac{N}{1-R} - Q$, we get an analytic and bounded function $t \to e^{-\lambda t} e^{\lambda t} r(t) = r(t)$ on $\Sigma_\gamma$ (with $\gamma = \frac{\beta}{2}$), such that

$$\mathcal{L}\left[e^{\lambda t} r(t)\right](z) = \frac{N(z)}{1-R(z)} - Q(z).$$

To conclude the proof of Lemma 5.3.4, we just need to determine the invert Laplace transform of $Q$. But we get, by straightforward calculation,

$$Q(z) = \mathcal{L}\left[\sum_{\substack{\omega \in Z \\ \Im\omega \geqslant \lambda}} P_\omega(t) e^{-i\omega t}\right](z).$$

□

### 5.3.5 Proof of Proposition 5.3.1

Finally we can prove the result stated at the beginning of this section.

*Proof of Proposition 5.3.1.* In Proposition 5.3.2 and 5.3.3 we have proven that we can apply Lemma 5.3.4 with $D_k = 1 - R$ and $N = N_k$, taking $\lambda_0 = 0$. But the result of this lemma is exactly the expansion of Proposition 5.3.1. □

**Remark 5.3.6.** *As we use only $\lambda_0 = 0$ in Proposition 5.3.2 and 5.3.3 for the proof of Proposition 5.3.1, we may wonder of we could use a weaker assumption on $f^{eq}$ for getting the estimate on $D_k$ for example. Indeed, that estimate derives from Theorem 5.3.1 for $\lambda_0 = 0$ and so the weaker assumption could be the hypothesis of (i) in Theorem 5.3.1 for $\lambda_0 = 0$. However, we also need to have that $D_k$ is entire, and there we have used Theorem 5.3.1* for all $\lambda_0 \in \mathbb{R}$.

## 5.4 Resolution and expansion of the second order equation

### 5.4.1 Introduction and statement of the result

In Proposition 5.2.2, we have proven that it is enough to solve dispersion relation (5.13) to get a solution $(h, \mu)$ to second order linearized Vlasov-Poisson equation (VPL2). So this section is devoted to the resolution of this second order dispersion relation, following the strategy established for the first order dispersion relation in the previous section, by proving the following proposition, which permits to complete the proof of our main result, Theorem 5.1.5.

**Proposition 5.4.1.** *Assume $f^{eq} \in \mathscr{E}(\mathbb{R}^d)$ and $g_0$ satisfies Assumption 1. Consider the solution $(g, \psi)$ of* (VPL) *given by Proposition 5.3.1 and Proposition 5.2.1. Then there exists a solution $\mu$ of the dispersion relation* (5.13) *whose expansion is detailed in Theorem 5.1.5.*

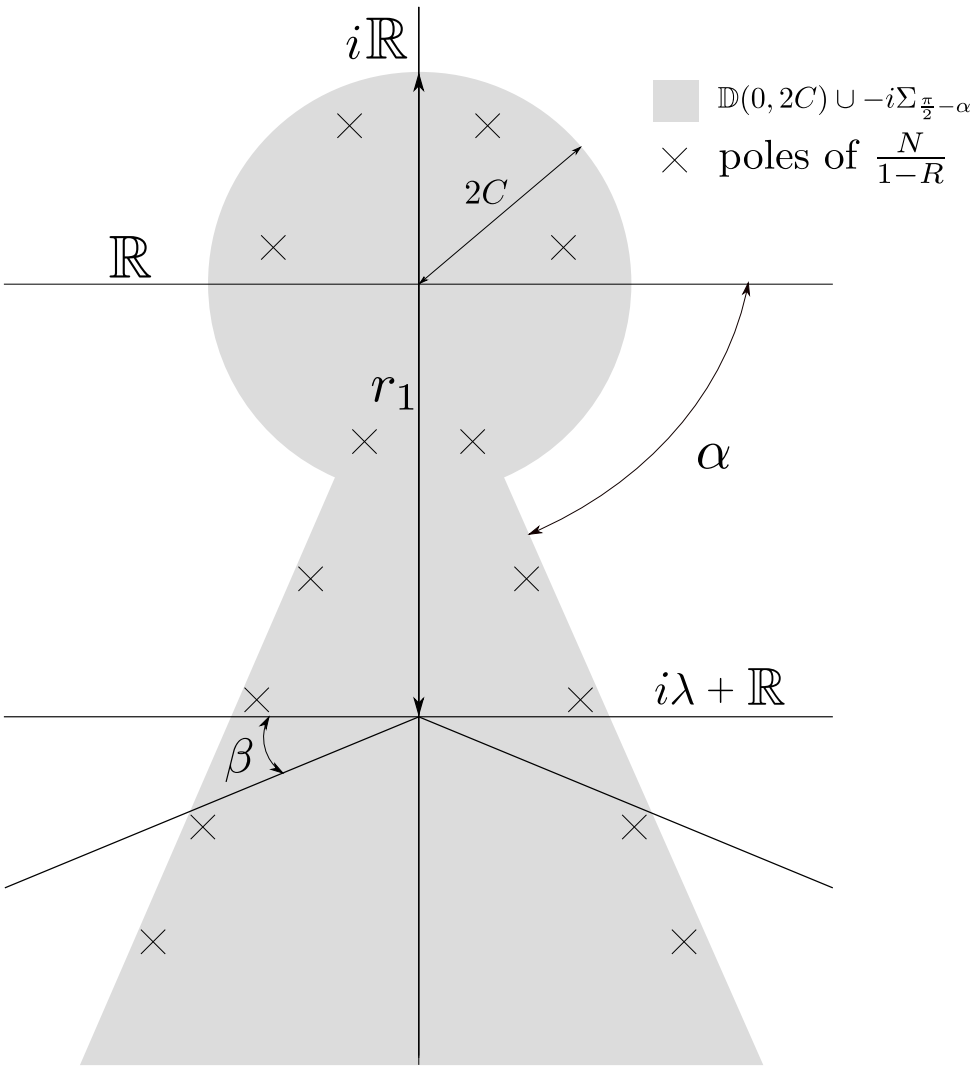


FIGURE 5.1 – An illustration of the geometrical constructions introduced in the proof of Lemma 5.3.4.

### 5.4.2 Definition of $\mathcal{N}_k^1$ and $\mathcal{N}_k^2$ (the right hand side)

We will first look for the right hand side of the second order dispersion relation, that was $N_k$ in the first order case.

Let $k \in (K+K)\backslash\{0\}$. By looking at the second order dispersion relation (5.13), we can assume, without loss of generality, that there exist $k_1, k_2 \in K$ such that $k_1 + k_2 = k$ and

$$\nabla_x \widehat{\psi \cdot \nabla_v} g(t,k,v) = i\widehat{\psi}(t,k_1)k_1 \cdot \nabla_v \widehat{g}(t,k_2,v).$$

Consequently, we can determine more precisely the right member of (5.13). However, to be rigorous we need to prove that our integrals are convergent. Indeed, as in Proposition 5.2.2, there exists $\lambda_0 \in \mathbb{R}$ such that for any $\lambda > \lambda_0$, we can construct a continuous and integrable function $\mathfrak{d} \in C^0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ such that

$$\forall (t,v) \in \mathbb{R}_+^* \times \mathbb{R}^d,\ |\widehat{\psi}(t,k_1)|e^{-\lambda t} + e^{-\lambda t}|k_1 \cdot \nabla_v \widehat{g}(t,k_2,v)| \leqslant \mathfrak{d}(v).$$

Consequently, if $\Im z > 2\lambda_0$, we can apply Lemma 5.3.3 to prove that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{\mathcal{L}\left[i\widehat{\psi}(t,k_1)k_1 \cdot \nabla_v \widehat{g}(t,k_2,v)\right](z)}{v \cdot k - z}\,\mathrm{d}v \\
&= i\,\mathcal{L}\left[\mathscr{F}\left[\mathcal{L}\left[i\widehat{\psi}(t,k_1)k_1 \cdot \nabla_v \widehat{g}(t,k_2,v)\right](z)\right](kt)\right](z) \\
&= i\int_0^\infty \int_{\mathbb{R}^d} \int_0^\infty i\widehat{\psi}(t,k_1)k_1 \cdot \nabla_v \widehat{g}(t,k_2,v) e^{izt}\mathrm{d}t\, e^{-i(k\tau)\cdot v}\,\mathrm{d}v\, e^{iz\tau}\,\mathrm{d}\tau \\
&= -i\int_0^\infty \int_0^\infty \widehat{\psi}(t,k_1)(k_1 \cdot k)\tau\, \mathscr{F}\,\widehat{g}(t,k_2,k\tau)\, e^{iz(\tau+t)}\mathrm{d}t\,\mathrm{d}\tau \\
&= -i(k_1 \cdot k)\int_0^\infty \int_0^s \widehat{\psi}(t,k_1)(s-t)\,\mathscr{F}\,\widehat{g}(t,k_2,k(s-t))\mathrm{d}t\, e^{izs}\mathrm{d}s \\
&= -i(k_1 \cdot k)\,\mathcal{L}\left[\int_0^t \widehat{\psi}(\tau,k_1)(t-\tau)\,\mathscr{F}\,\widehat{g}(\tau,k_2,k(t-\tau))\mathrm{d}\tau\right](z),
\end{aligned}$$

where we have used the change of variable $\tau + t \leftarrow s$ and the notation

$$\mathscr{F}\,\widehat{g}(t,k,\xi) = \mathscr{F}[\widehat{g}(t,k,v)](\xi).$$

Furthermore, using definition of $g$ (see (5.12)), we can precise $\mathscr{F}\,\widehat{g}(\tau,k_2,k(t-\tau))$. Indeed, we start from

$$\hat{g}(t,k,v) = e^{-ik\cdot vt}\widehat{g_0}(k,v) + i\int_0^t e^{-ik\cdot v(t-s)}\widehat{\psi}(s,k)k \cdot \nabla_v f^{eq}(v)\mathrm{d}s,$$

so that we have

$$\mathscr{F}\,\widehat{g}(t,k_2,\xi) = \mathscr{F}[\widehat{g_0}(k_2,v)](\xi + k_2 t) + i\int_0^t \widehat{\psi}(s,k_2)\,\mathscr{F}\left[k_2.\nabla_v f^{eq}\right](\xi + k_2(t-s))\mathrm{d}s.$$

Consequently, we get

$$\mathscr{F}\,\widehat{g}(\tau,k_2,k(t-\tau)) = \mathscr{F}[\widehat{g_0}(k_2,v)](kt+(k_2-k)\tau) \\ - i\int_0^\tau \widehat{\psi}(s,k_2)\,\mathscr{F}\left[k_2\cdot\nabla_v f^{eq}\right](k(t-\tau)+k_2(\tau-s))\mathrm{d}s.$$

So we have two numerators to study for this dispersion relation. On the one hand, we have

$$\mathcal{N}_k^1(z) := -\frac{k_1\cdot k}{|k|^2}\,\mathcal{L}\left[F_k^1(t)\right](z), \tag{5.24}$$

with

$$F_k^1(t) := \int_0^t \widehat{\psi}(\tau,k_1)(t-\tau)\,\mathscr{F}[\widehat{g_0}(k_2,v)](kt+(k_2-k)\tau)\mathrm{d}\tau.$$

On the other hand, we have

$$\mathcal{N}_k^2(z) := i\frac{k_1\cdot k}{|k|^2}\,\mathcal{L}\left[F_k^2(t)\right](z), \tag{5.25}$$

with

$$F_k^2(t) = \int_0^t \widehat{\psi}(\tau,k_1)(t-\tau)\int_0^\tau \widehat{\psi}(s,k_2)\,\mathscr{F}\left[k_2\cdot\nabla_v f^{eq}\right](k(t-\tau)+k_2(\tau-s))\mathrm{d}s\;\mathrm{d}\tau.$$

So with these notations,the dispersion relation (5.4.1) may be written as, for all $z$ such that $\Im z > 2\lambda_0$,

$$\mathcal{L}\left[\widehat{\mu}(t,k)\right](z)D_k(z) = \mathcal{N}_k^1(z)+\mathcal{N}_k^2(z). \tag{5.26}$$

### 5.4.3 Estimates for $\mathcal{N}_k^1$ and $\mathcal{N}_k^2$

We are going to apply the same strategy as for the resolution of the first order dispersion relation. It will be solve using Lemma 5.3.4. The denominator has been studied in Proposition 5.3.2. The following lemma describes the regularity and the behavior of the numerators $\mathcal{N}_k^1$ and $\mathcal{N}_k^2$.

**Lemma 5.4.1.** *The function $\mathcal{N}_k^1$, $\mathcal{N}_k^2$ have a meromorphic continuation and there exist $\widetilde{\lambda}\in\mathbb{R}$ and $\beta\in(0,\frac{\pi}{2})$ such that*

$$\sup_{z\in i\widetilde{\lambda}+i\Sigma_{\beta+\frac{\pi}{2}}} |z-i\widetilde{\lambda}|\left[|\mathcal{N}_k^1(z)|+|\mathcal{N}_k^2(z)|\right] < \infty.$$

• *If there exists $\gamma\in(0,1)$ such $k_2=-\gamma k_1$ then the poles of $\mathcal{N}_k^1$ are the points $\omega\in\mathbb{C}$ such that $\omega=\omega_1\frac{|k|}{|k_1|}$ where $D_{k_1}(\omega_1)=0$ and its multiplicity is smaller than or equal to $n_{k_1,\omega_1}+1$. $\mathcal{N}_k^2$ has two kinds of poles. On the one hand, there are the points $\omega\in\mathbb{C}$ such that $\omega=\omega_1+\omega_2$ where $D_{k_1}(\omega_1)=D_{k_2}(\omega_2)=0$. On the other hand, there are the points $\omega\in\mathbb{C}$ such that $\omega=\omega_1\frac{|k|}{|k_1|}$ where $D_{k_1}(\omega_1)=0$. The multiplicity of a pole belonging to the two families is smaller than or*

*equal to* $n_{k_1,\omega_1}+n_{k_2,\omega_2}+1$. *Else the multiplicity of a pole of the first kind is smaller than or equal to* $n_{k_1,\omega_1}+n_{k_2,\omega_2}-1$ *and the multiplicity of a pole of the second kind is smaller than or equal to* $n_{k_1,\omega_1}+1$.
• *Else* $\mathcal{N}_k^1$ *is an entire function and the poles of* $\mathcal{N}_k^2$ *are the points* $\omega\in\mathbb{C}$ *such that* $\omega=\omega_1+\omega_2$ *where* $D_{k_1}(\omega_1)=D_{k_2}(\omega_2)=0$ *and its multiplicity is smaller than or equal to* $n_{k_1,\omega_1}+n_{k_2,\omega_2}-1$.

Now, we focus on proving Lemma 5.4.1. However, using analytic representation Theorem 5.3.1, it is directly deduced of the two following lemmas (Lemma 5.4.2 and Lemma 5.4.3) involving properties of $F_k^1$ and $F_k^2$.

**Lemma 5.4.2.** *For all* $\lambda\in\mathbb{R}$ *there exist* $\beta\in(0,\frac{\pi}{2})$ *and an analytic and bounded function on* $\Sigma_\beta$ *denoted* $r$ *such that*
• *if* $k_2=-\gamma k_1,\ \gamma\in(0,1)$, *then for all* $t>0$

$$F_k^1(t)=\sum_{\substack{D_{k_1}(\omega_1)=0\\ \Im\omega_1\geqslant\lambda}} R_{\omega_1}(t)e^{-i\omega_1\frac{|k|}{|k_1|}t}+e^{\lambda t}r(t), \tag{5.27}$$

*with* $R_{\omega_1}$ *a polynomial of degree smaller than or equal to* $n_{k_1,\omega_1}$,
• *else, for all* $t>0$

$$F_k^1(t)=e^{\lambda t}r(t). \tag{5.28}$$

**Lemma 5.4.3.** *For all* $\lambda\in\mathbb{R}$ *there exist* $\beta\in(0,\frac{\pi}{2})$ *and an analytic and bounded function on* $\Sigma_\beta$ *denoted* $r$ *such that*
• *if* $k_2=-\gamma k_1,\ \gamma\in(0,1)$, *then for all* $t>0$

$$F_k^2(t)=\sum_{\substack{D_{k_1}(\omega_1)=0\\ \Im\omega_1\geqslant\lambda}}\left[R_{k_1,k_2}^{\omega_1}(t)e^{-i\omega_1\frac{|k|}{|k_1|}t}+\sum_{\substack{D_{k_2}(\omega_2)=0\\ \Im\omega_2\geqslant\lambda}}Q_{\omega_1,\omega_2}^{k_1,k_2}(t)e^{-i(\omega_1+\omega_2)t}\right]+e^{\lambda t}r(t), \tag{5.29}$$

*with* $Q_{\omega_1,\omega_2}^{k_1,k_2}$ *a polynomial of degree smaller than or equal to* $n_{k_1,\omega_1}+n_{k_2,\omega_2}-2$ *and* $R_{k_1,k_2}^{\omega_1}$ *a polynomial of degree smaller than or equal to* $n_{k_1,\omega_1}$ *(if there exist* $\omega_1,\omega_2$ *such that* $\omega_1+\omega_2=\omega_1\frac{|k|}{|k_1|}$ *the maximal possible degree of* $Q_{\omega_1,\omega_2}^{k_1,k_2}$ *and* $R_{k_1,k_2}^{\omega_1}$ *is* $n_{k_1,\omega_1}+n_{k_2,\omega_2}$ *)*
• *else, for all* $t>0$

$$F_k^2(t)=\sum_{\substack{D_{k_1}(\omega_1)=0\\ \Im\omega_1\geqslant\lambda}}\sum_{\substack{D_{k_2}(\omega_2)=0\\ \Im\omega_2\geqslant\lambda}}Q_{\omega_1,\omega_2}^{k_1,k_2}(t)e^{-i(\omega_1+\omega_2)t}+e^{\lambda t}r(t), \tag{5.30}$$

*with* $Q_{\omega_1,\omega_2}^{k_1,k_2}$ *a polynomial of degree smaller than or equal to* $n_{k_1,\omega_1}+n_{k_2,\omega_2}-2$.

**Remark 5.4.4.** *In Lemma 5.4.1, we need that the inequality is true for a given* $\tilde\lambda$, *in order to apply Lemma 5.3.4. However, applying Theorem 5.3.1, we deduce from Lemmae 5.4.2 and 5.4.3 that the inequality is true* for all $\tilde\lambda\in\mathbb{R}$. *On the other hand, we have needed that Lemma 5.4.2 and 5.4.3 are true for all* $\lambda\in\mathbb{R}$, *in order to prove that* $\mathcal{N}_k^1$ *and* $\mathcal{N}_k^2$ *are meromorphic.*

We are going to prove these lemmas distinguishing the non resonant case from the resonant case (when there exists $\gamma \in (0,1)$ such that $k_2 = -\gamma k_1$). In order to get proofs as clear as possible we do not prove that the remainder term can be extended on complex cones and we only control them on $\mathbb{R}_+^*$. Indeed, there are no real issues to extend them and the arguments to control them on $\Sigma_\alpha$ or $\mathbb{R}_+^*$ are the same. Furthermore, the notations induced for the complex extensions are quite heavy and so do not help to understand the ideas. However, in the first proof, to give an example, we prove the analytic extension and we really estimate it.

### 5.4.4 Proof of Lemma 5.4.2 in the non-resonant case.

Since we are studying the non-resonant case, there exists $\delta > 0$ such that

$$\forall \theta \in (0,1),\ \delta \leqslant |(1-\theta)k + \theta k_2|.$$

Indeed, in the resonant case there exists $\gamma \in (0,1)$ such that $k_2 = -\gamma k_1 = \gamma(k_2 - k)$, so that $(1-\gamma)k_2 + \gamma k = 0$. We have proven in Proposition 5.3.1 that there exists $\alpha \in (0, \frac{\pi}{2})$ such that $\widehat{\psi}(t, k_1)$ extends to an analytic function on $\Sigma_\alpha$ and that there exists $\lambda_0 \in \mathbb{R}$ and $M > 0$ such that

$$\forall z \in \Sigma_\alpha, |\widehat{\psi}(z, k_1)| \leqslant M e^{\lambda_0 \Re z}.$$

Furthermore, since $v \mapsto \widehat{g_0}(k_2, v) \in \mathscr{E}(\mathbb{R}^d)$, its Fourier transform extends to an entire function on $\mathbb{C}^d$ and we can assume (choosing $\alpha$ small enough) that for all $\lambda_2 \in \mathbb{R}$ there exists $C_{\lambda_2} > 0$ such that

$$\forall z \in \Sigma_\alpha \mathbb{R}^d,\ |\mathscr{F}\left[\widehat{g_0}(k_2, v)\right](z)| \leqslant C_{\lambda_2} e^{\lambda_2 |\Re z|}. \tag{5.31}$$

Now observe that by a change of variable, $F_k^1(t)$ can be written as

$$F_k^1(t) = t^2 \int_0^1 (1-\theta)\widehat{\psi}(\theta t, k_1)\, \mathscr{F}[\widehat{g_0}(k_2, v)](t\left((1-\theta)k + \theta k_2\right))\mathrm{d}\theta.$$

Consequently, $F_k^1(t)$ naturally extends to an analytic function on $\Sigma_\alpha$. Now, we have to control $F_k^1(z)e^{-\lambda z}$ on $\Sigma_\alpha$ for any $\lambda \in \mathbb{R}$. Indeed, we have, for $z \in \Sigma_\alpha$, as we can assume $\lambda_2 \leqslant 0$,

$$\begin{aligned}
|F_k^1(z)e^{-\lambda z}| &\leqslant |z|^2 e^{-\lambda \Re z} \int_0^1 |\widehat{\psi}(\theta z, k_1)|(1-\theta)\, |\mathscr{F}[\widehat{g_0}(k_2, v)](z\left((1-\theta)k + \theta k_2\right))|\, \mathrm{d}\theta \\
&\leqslant C_{\lambda_2} M |z|^2 \int_0^1 e^{(\lambda_0 \theta - \lambda)\Re z}(1-\theta) e^{\lambda_2 \Re z |(1-\theta)k + \theta k_2|} \mathrm{d}\theta \\
&\leqslant C_{\lambda_2} M |z|^2 e^{(|\lambda_0| - \lambda + \delta\lambda_2)\Re z} \\
&\leqslant C_{\lambda_2} M \left(\frac{\Re z}{\cos\alpha}\right)^2 e^{(|\lambda_0| - \lambda + \delta\lambda_2)\Re z}.
\end{aligned}$$

So this quantity is bounded uniformly with respect to $z \in \Sigma_\alpha$ if $\lambda_2 < \frac{\lambda - |\lambda_0|}{\delta}$.

### 5.4.5 Proof of Lemma 5.4.2 in the resonant case

As explained before, from now, we do not pay attention to the analytic extension anymore. First, we use the resonance to give a more adapted expression of $F_k^1$

$$\begin{aligned} F_k^1(t) &= \int_0^t \widehat{\psi}(\tau, k_1)(t-\tau)\, \mathscr{F}[\widehat{g_0}(k_2, v)](k_1\left((1-\gamma)t-\tau\right))\mathrm{d}\tau \\ &= \int_{-\gamma t}^{(1-\gamma)t} \widehat{\psi}((1-\gamma)t-s, k_1)(\gamma t+s)\, \mathscr{F}\, \widehat{g_0}(k_2, k_1 s)\mathrm{d}s, \end{aligned}$$

making the change of variable $s \leftarrow (1-\gamma)t-\tau$. We want to expand $\psi$, so we introduce the dependency of $F_k^1$ with respect to $t \mapsto \widehat{\psi}(t, k_1)$ by denoting $F_k^1[\widehat{\psi}(t, k_1)](t)$. Consequently, using the expansion of $\psi$ of Proposition 5.3.1, for any $\lambda_1 \in \mathbb{R}$, we get

$$F_k^1[\widehat{\psi}(t, k_1)](t) = \sum_{\substack{D_{k_1}(\omega_1)=0 \\ \Im\omega_1 \geqslant \lambda_1}} F_k^1[P_{k_1,\omega_1}(t)e^{-i\omega_1 t}](t) + F_k^1[e^{\lambda_1 t} r_{k_1,\lambda_1}(t)](t),$$

where $r_{k_1,\lambda_1}$ is a bounded function on $\mathbb{R}_+^*$.

First, we are going to control the remainder term $F_k^1[e^{\lambda_1 t} r_{k_1,\lambda_1}(t)](t)$. Using the same control of the Fourier transform as previously (see (5.31)), we have, as we can assume $\lambda_1, \lambda_2 \leqslant 0$,

$$\begin{aligned} &e^{-\lambda t}|F_k^1[e^{\lambda_1 t} r_{k_1,\lambda_1}(t)](t)| \\ &\qquad \leqslant \|r_{k_1,\lambda_1}\|_{L^\infty} e^{-\lambda t} \int_{-\gamma t}^{(1-\gamma)t} e^{\lambda_1[(1-\gamma)t-s]}(\gamma t+s)\, |\mathscr{F}[\widehat{g_0}(k_2, v)](k_1 s)|\, \mathrm{d}s \\ &\qquad \leqslant \|r_{k_1,\lambda_1}\|_{L^\infty} C_{\lambda_2} e^{-\lambda t} \int_{-\gamma t}^{(1-\gamma)t} e^{\lambda_1[(1-\gamma)t-s]}(\gamma t+s) e^{\lambda_2|k_1||s|}\mathrm{d}s \\ &\qquad \leqslant \|r_{k_1,\lambda_1}\|_{L^\infty} C_{\lambda_2} e^{[(1-\gamma)\lambda_1-\lambda]t} \int_{\mathbb{R}} (\gamma t+s) e^{(\lambda_2|k_1|-\lambda_1)|s|}\mathrm{d}s. \end{aligned}$$

So this quantity is bounded uniformly with respect to $t \in \mathbb{R}_+^*$ if $(1-\gamma)\lambda_1 < \lambda$ and $\lambda_2|k_1| < \lambda_1$.

Now, we are going to study one leading term of the type $F_k^1[t^n e^{-i\omega_1 t}](t)$. So, we are doing a new expansion.

$$\begin{aligned} F_k^1[t^n e^{-i\omega_1 t}](t) &= e^{-i\omega_1(1-\gamma)t} \int_{-\gamma t}^{(1-\gamma)t} ((1-\gamma)t-s)^n e^{i\omega_1 s}(\gamma t+s)\, \mathscr{F}[\widehat{g_0}(k_2, v)](k_1 s)\mathrm{d}s \\ &= \sum_{j=0}^{n+1} b_j t^j e^{-i\omega_1(1-\gamma)t} \int_{-\gamma t}^{(1-\gamma)t} s^{n+1-j} e^{i\omega_1 s}\, \mathscr{F}[\widehat{g_0}(k_2, v)](k_1 s)\mathrm{d}s, \end{aligned}$$

where $b_0, \ldots, b_{n+1}$ are real numbers. Here we recognise the leading terms of (5.27) since, by construction, $1-\gamma = \frac{|k|}{|k_1|}$.

Then, observe that since the right integral is convergent (see (5.31)), there exists $A \in \mathbb{C}$ such that for any $t \in \mathbb{R}_+^*$, we have

$$\int_{-\gamma t}^{(1-\gamma)t} s^{n+1-j} e^{i\omega_1 s} \mathscr{F}[\widehat{g_0}(k_2,v)](k_1 s)\mathrm{d}s = A - \int_{-\infty}^{-\gamma t} s^{n+1-j} e^{i\omega_1 s} \mathscr{F}[\widehat{g_0}(k_2,v)](k_1 s)\mathrm{d}s - \int_{(1-\gamma)t}^{+\infty} s^{n+1-j} e^{i\omega_1 s} \mathscr{F}[\widehat{g_0}(k_2,v)](k_1 s)\mathrm{d}s.$$

The complex number $A$ is the leading term of this integral whereas the other ones are remainder terms. So we just have to control them. Indeed, we have

$$\begin{aligned}\left| e^{-\lambda t} t^j e^{-i\omega_1(1-\gamma)t} \int_{-\infty}^{-\gamma t} s^{n+1-j} e^{i\omega_1 s} \mathscr{F}[\widehat{g_0}(k_2,v)](k_1 s)\mathrm{d}s \right| \\ \leqslant C_{\lambda_2} e^{-\lambda t} t^j e^{\Im\omega_1(1-\gamma)t} \int_{\gamma t}^{\infty} s^{n+1-j} e^{\Im\omega_1 s} e^{\lambda_2|k_1|s}\mathrm{d}s \\ \leqslant C_{\lambda_2} \int_0^{\infty} e^{-\lambda\frac{s}{\gamma}} \left(\frac{s}{\gamma}\right)^j e^{\Im\omega_1(1-\gamma)\frac{s}{\gamma}} s^{n+1-j} e^{\Im\omega_1 s} e^{\lambda_2|k_1|s}\mathrm{d}s,\end{aligned}$$

as we can assume $\lambda \leqslant 0$ and since $t \leqslant \frac{s}{\gamma}$. Consequently, it is bounded uniformly with respect to $t$ if $|k_1|\gamma\lambda_2 < \lambda - \Im\omega_1$. The estimation of the third integral can be realized with the same ideas.

As we have a term in $t^{n+1}$, we see that $R_{\omega_1}$ is of degree $\leqslant n_{k_1,\omega_1}$, since $P_{k_1,\omega_1}$ is of degree $\leqslant n_{k_1,\omega_1} - 1$.

### 5.4.6 Proof of Lemma 5.4.3 in the non resonant case

First, operating the change of variable $\tau' = t - \tau$, $s' = t - s$, we can write $F_k^2$ as

$$F_k^2(t) = \int_{0\leqslant\tau'\leqslant s'\leqslant t} \widehat{\psi}(t-\tau',k_1)\widehat{\psi}(t-s',k_2)\tau' \mathscr{F}\left[k_2\cdot\nabla_v f^{eq}\right](k\tau' + k_2(s'-\tau'))\mathrm{d}s'\mathrm{d}\tau',$$

since, if $0 \leqslant s \leqslant \tau \leqslant t$ we get $0 \leqslant t-\tau \leqslant t-s$ and $t-\tau \leqslant t-s \leqslant t$, that is $0 \leqslant \tau' \leqslant s' \leqslant t$. In order to get notations general enough but compact, we denote $u = \mathscr{F}\left[k_2.\nabla_v f^{eq}\right]$. Since $u \in \mathscr{E}(\mathbb{R}^d)$, for all $\lambda_3 \in \mathbb{R}$ there exists a constant $C_{\lambda_3} > 0$ such that

$$\forall\xi \in \mathbb{R}^d, \forall t > 0,\ |u(t\xi)| \leqslant C_{\lambda_3} e^{\lambda_3 t|\xi|}. \tag{5.32}$$

We define, for continuous functions $\phi_1, \phi_2$ with an exponential order, a bilinear operator $q$ by

$$q[\phi_1,\phi_2](t) = q[\phi_1(t),\phi_2(t)](t) = \int_{0\leqslant\tau\leqslant s\leqslant t} \phi_1(t-\tau)\phi_2(t-s)\tau u(k\tau + k_2(s-\tau))\mathrm{d}s\ \mathrm{d}\tau.$$

With these notations, we have

$$F_k^2(t) = q[\widehat{\psi}(t,k_1),\widehat{\psi}(t,k_2)](t).$$

Consequently, using the expansions of $\widehat{\psi}(t,k_1)$ and $\widehat{\psi}(t,k_2)$ established in Proposition 5.3.1, we get[2] for $\lambda_1,\lambda_2\in\mathbb{R}$,

$$F_k^2=\sum_{\substack{D_{k_1}(\omega_1)=0\\ \Im\omega_1\geqslant\lambda_1}}\sum_{\substack{D_{k_2}(\omega_2)=0\\ \Im\omega_2\geqslant\lambda_2}} q[P_{k_1,\omega_1}e^{-i\omega_1 t},P_{k_2,\omega_2}e^{-i\omega_2 t}]+q[e^{\lambda_1 t}r_{k_1,\lambda_1},\widehat{\psi}(t,k_2)]$$
$$+q[\widehat{\psi}(tk_1),e^{\lambda_2 t}r_{k_2,\lambda_2}]-q[e^{\lambda_1 t}r_{k_1,\lambda_1},e^{\lambda_2 t}r_{k_2,\lambda_2}] \quad (5.33)$$

where $r_{k_1,\lambda_1}$ and $r_{k_2,\lambda_2}$ are respectively bounded by constants $C_{\lambda_1}$ and $C_{\lambda_2}$.

Furthermore, we can also assume that there exists $\lambda_0\in\mathbb{R}$ and $M>0$ such that

$$\forall t>0,\ |\widehat{\psi}(t,k_1)|+|\widehat{\psi}(t,k_2)|\leqslant Me^{\lambda_0 t}.$$

Finally, since we are treating the non-resonant case, we may assume that there exists $\delta>0$ such that

$$\forall 0\leqslant\tau\leqslant s,\ \delta s\leqslant|\tau k+(s-\tau)k_2|. \quad (5.34)$$

So first, we are going to control the remainder terms of (5.33). For example, we consider $q[e^{\lambda_1 t}r_{k_1,\lambda_1},\widehat{\psi}(t,k_2)]$. So, if $t>0$, $\lambda_3<0$, $\lambda_1<0$, we have

$$\begin{aligned}
&e^{-\lambda t}|q[e^{\lambda_1 t}r_{k_1,\lambda_1},\widehat{\psi}(t,k_2)](t)|\\
&\qquad\leqslant C_{\lambda_1}MC_{\lambda_3}e^{(\lambda_1+\lambda_0-\lambda)t}\int_{0\leqslant\tau\leqslant s\leqslant t}e^{-\lambda_1\tau-\lambda_0 s}\tau e^{\lambda_3|k\tau+k_2(s-\tau)|}\mathrm{d}s\mathrm{d}\tau\\
&\qquad\leqslant C_{\lambda_1}MC_{\lambda_3}e^{(\lambda_1+\lambda_0-\lambda)t}\int_{0\leqslant\tau\leqslant s\leqslant t}e^{-\lambda_1\tau-\lambda_0 s+\lambda_3\delta s}\tau\mathrm{d}s\mathrm{d}\tau\\
&\qquad\qquad\leqslant C_{\lambda_1}MC_{\lambda_3}t^2e^{(\lambda_1+\lambda_0-\lambda)t}\int_{s>0}e^{-\lambda_1 s-\lambda_0 s+\lambda_3\delta s}\mathrm{d}s.
\end{aligned}$$

So, this quantity is bounded uniformly with respect to $t>0$ if $\lambda_1<\lambda-\lambda_0$ and $\lambda_3<\frac{\lambda_1+\lambda_0}{\delta}<\frac{\lambda}{\delta}$. Similarly, we could prove that if $\lambda_2$ is chosen negative enough then we could control, uniformly with respect to $t$, $q[\widehat{\psi}(t,k_1),e^{\lambda_2 t}r_{k_2,\lambda_2}](t)e^{-\lambda t}$, and also $q[e^{\lambda_1 t}r_{k_1,\lambda_1},e^{\lambda_2 t}r_{k_2,\lambda_2}]$.

Now, we consider a generic leading terms of (5.33) of the type $q[t^{n_1}e^{-i\omega_1 t},t^{n_2}e^{-i\omega_2 t}]$. So first, we can expand it

$$\begin{aligned}
&q[t^{n_1}e^{-i\omega_1 t},t^{n_2}e^{-i\omega_2 t}](t)\\
&=\int_{0\leqslant\tau\leqslant s\leqslant t}(t-\tau)^{n_1}e^{-i\omega_1(t-\tau)}(t-s)^{n_2}e^{-i\omega_2(t-s)}\tau u(k\tau+k_2(s-\tau))\mathrm{d}s\mathrm{d}\tau\\
&=\sum_{j_1=0}^{n_1}\sum_{j_2=0}^{n_2}b_{j_1,j_2}e^{-i(\omega_1+\omega_2)t}t^{n_1-j_1+n_2-j_2}\int_{0\leqslant\tau\leqslant s\leqslant t}\tau^{j_1+1}s^{j_2}e^{i\omega_1\tau+i\omega_2 s}u(k\tau+k_2(s-\tau))\mathrm{d}s\mathrm{d}\tau,
\end{aligned}$$

2. Realizing a decomposition of the form

$$q[a_1+b_1,a_2+b_2]=q[a_1,a_2]+q[b_1,a_2+b_2]+q[a_1+b_1,b_2]-q[b_1,b_2].$$

where $b \in \mathbb{R}^{[\![0,n_1]\!]\times[\![0,n_2]\!]}$ are some real coefficients.

We observe that this last integral converge when $t$ goes to $+\infty$. Indeed, we have

$$\left|\int_0^s \tau^{j_1+1}s^{j_2}e^{i\omega_1\tau+i\omega_2 s}u(k\tau+k_2(s-\tau))\mathrm{d}\tau\right| \leqslant C_{\lambda_3}s^{j_1+j_2+2}e^{(|\omega_1|+|\omega_2|)s}e^{\lambda_3\delta s}$$
$$\in L^1(\mathbb{R}_+), \quad \text{if } \delta\lambda_3 < -|\omega_1|-|\omega_2|.$$

Consequently, there exists a complex constant $A \in \mathbb{C}$ such that

$$\int_{0\leqslant\tau\leqslant s\leqslant t} \tau^{j_1+1}s^{j_2}e^{i\omega_1\tau+i\omega_2 s}u(k\tau+k_2(s-\tau))\mathrm{d}s\mathrm{d}\tau$$
$$= A - \int_{\substack{0\leqslant\tau\leqslant s\\ t\leqslant s}} \tau^{j_1+1}s^{j_2}e^{i\omega_1\tau+i\omega_2 s}u(k\tau+k_2(s-\tau))\mathrm{d}s\mathrm{d}\tau.$$

This complex number $A$ generates the term of frequency $\omega_1+\omega_2$ in (5.30). So we just need to prove that the other term is a remainder term controlling it. Indeed, we have

$$\begin{aligned}
&e^{-\lambda t}\left|e^{-i(\omega_1+\omega_2)t}t^{n_1-j_1+n_2-j_2}\int_{\substack{0\leqslant\tau\leqslant s\\ t\leqslant s}} \tau^{j_1+1}s^{j_2}e^{i\omega_1\tau+i\omega_2 s}u(k\tau+k_2(s-\tau))\mathrm{d}s\mathrm{d}\tau\right|\\
\leqslant& C_{\lambda_3}e^{-\lambda t}e^{\Im(\omega_1+\omega_2)t}t^{n_1-j_1+n_2-j_2}\int_{\substack{0\leqslant\tau\leqslant s\\ t\leqslant s}} s^{j_1+j_2+1}e^{-\Im\omega_1\tau-\Im\omega_2 s}e^{\lambda_3\delta s}\mathrm{d}s\mathrm{d}\tau\\
\leqslant& C_{\lambda_3}e^{-\lambda t}e^{\Im(\omega_1+\omega_2)t}t^{n_1-j_1+n_2-j_2}\int_{t\leqslant s} s^{j_1+j_2+2}e^{|\Im\omega_1|s-\Im\omega_2 s}e^{\lambda_3\delta s}\mathrm{d}s\\
\leqslant& C_{\lambda_3}\int_{s>0} e^{|\Im(\omega_1+\omega_2)|s-\lambda s}s^{n_1+2+n_2}e^{|\Im\omega_1|s-\Im\omega_2 s}e^{\lambda_3\delta s}\mathrm{d}s,
\end{aligned}$$

as $\lambda$ can be supposed $\leqslant 0$, and this last quantity is finite if $\lambda_3$ is negative enough ($\lambda_3 < \frac{\lambda-|\Im(\omega_1+\omega_2)|-|\Im\omega_1|+\Im\omega_2}{\delta}$).

Concerning the degree, we see that it is $\leqslant n_{k_1,\omega_1}-1+n_{k_2,\omega_2}-1$, since $n_1 \leqslant n_{k_1,\omega_1}-1$ and $n_2 \leqslant n_{k_2,\omega_2}-1$, which corresponds to what is expected.

### 5.4.7 Proof of Lemma 5.4.3 in the resonant case

We consider now the last case, which is the most complex. We keep the notations of the previous subsection but we need a new expression of $q$ adapted to the resonance :

$$\begin{aligned}
q[\phi_1,\phi_2](t) &= \int_{0\leqslant\tau\leqslant s\leqslant t} \phi_1(t-\tau)\phi_2(t-s)\tau u(k_1\left[(1-\gamma)\tau-\gamma(s-\tau)\right])\mathrm{d}s\mathrm{d}\tau,\\
&= \int_0^t\int_0^s \phi_1(t-\tau)\phi_2(t-s)\tau u(k_1\left[\tau-\gamma s\right])\mathrm{d}\tau\mathrm{d}s,\\
&= \int_0^t\int_{-\gamma s}^{(1-\gamma)s} \phi_1(t-\tau-\gamma s)\phi_2(t-s)(\tau+\gamma s)u(k_1\tau)\mathrm{d}\tau\mathrm{d}s.
\end{aligned}$$

The term $\tau + \gamma s$ is quite heavy for our estimations, so we introduce a last notation

$$q_l^m[\phi_1, \phi_2](t) = \int_0^t \int_{-\gamma s}^{(1-\gamma)s} \phi_1(t - \tau - \gamma s)\phi_2(t - s)\tau^l s^m u(k_1\tau)\mathrm{d}\tau\mathrm{d}s.$$

Consequently, we can expand $q[\phi_1, \phi_2]$ as follow

$$q[\phi_1, \phi_2] = q_1^0[\phi_1, \phi_2] + \gamma q_0^1[\phi_1, \phi_2].$$

We also introduce [3] a new expansion of $F_k^2$ more adapted to the resonance

$$F_k^2 = \sum_{\substack{D_{k_1}(\omega_1)=0 \\ \Im\omega_1 \geqslant \lambda_1}} \sum_{\substack{D_{k_2}(\omega_2)=0 \\ \Im\omega_2 \geqslant \lambda_2}} q[P_{k_1,\omega_1} e^{-i\omega_1 t}, P_{k_2,\omega_2} e^{-i\omega_2 t}]$$
$$+ \sum_{\substack{D_{k_1}(\omega_1)=0 \\ \Im\omega_1 \geqslant \lambda_1}} q[P_{k_1,\omega_1} e^{-i\omega_1 t}, e^{\lambda_2 t} r_{k_2,\lambda_2}] + q[e^{\lambda_1 t} r_{k_1,\lambda_1}, \widehat{\psi}(t, k_2)].$$

Now, we are going to study each one of the terms of this expansion.

**Last term**

First, we control the last remainder term, $q[e^{\lambda_1 t} r_{k_1,\lambda_1}, \widehat{\psi}(t, k_2)]$. Indeed, if $t > 0$ we have

$$\begin{aligned}
&e^{-\lambda t}|q_l^m[e^{\lambda_1 t} r_{k_1,\lambda_1}, \widehat{\psi}(t, k_2)](z)| \\
\leqslant &C_{\lambda_1} M C_{\lambda_3} e^{-\lambda t} t^{l+m} \int_0^t \int_{-\gamma s}^{(1-\gamma)s} e^{\lambda_1[t-\tau-\gamma s]} e^{\lambda_0(t-s)} e^{\lambda_3|k_1||\tau|} \mathrm{d}\tau\mathrm{d}s \\
\leqslant &C_{\lambda_1} M C_{\lambda_3} \left(\int_{\mathbb{R}} e^{-\lambda_1\tau + \lambda_3|k_1||\tau|}\mathrm{d}\tau\right) t^{l+m} e^{(-\lambda+\lambda_0+\lambda_1)t} \left(\int_0^t e^{-\gamma s\lambda_1 - \lambda_0 s}\mathrm{d}s\right) \\
\leqslant &C_{\lambda_1} M C_{\lambda_3} \left(\int_{\mathbb{R}} e^{-\lambda_1\tau + \lambda_3|k_1||\tau|}\mathrm{d}\tau\right) t^{l+m+1} e^{(-\lambda+\lambda_0+\lambda_1)t} e^{-\gamma t\lambda_1 - \lambda_0 t} \\
\leqslant &C_{\lambda_1} M C_{\lambda_3} \left(\int_{\mathbb{R}} e^{(-\lambda_1 + \lambda_3|k_1|)|\tau|}\mathrm{d}\tau\right) t^{l+m+1} e^{(-\lambda+\lambda_1(1-\gamma))t}
\end{aligned}$$

So this last quantity is bounded uniformly with respect to $t$ if $\lambda_1$ and $\lambda_3$ are chosen negative enough. More precisely, we need $(1-\gamma)\lambda_1 < \lambda$ and $\lambda_3|k_1| < \lambda_1$.

**Second term**

Now, we study the behavior of the second kind of term in the expansion of $F_k^2$.

Expanding $P_{k_1,\omega_1}(t-\tau-\gamma s)$, we can write $q[P_{k_1,\omega_1} e^{-i\omega_1 t}, e^{\lambda_2 t} r_{k_2,\lambda_2}](t)$ as a linear combination of term of the type

$$t^j q_{l+1}^m[e^{-i\omega_1 t}, e^{\lambda_2 t} r_{k_2,\lambda_2}](t) \text{ and } t^j q_l^{m+1}[e^{-i\omega_1 t}, e^{\lambda_2 t} r_{k_2,\lambda_2}](t),$$

with $j + l + m \leqslant \deg P_{k_1,\omega_1}$.

3. Realizing a decomposition of the form : $q[a_1 + b_1, a_2 + b_2] = q[a_1, a_2] + q[a_1, b_2] + q[b_1, a_2 + b_2]$.

Let $t > 0$, then we have

$$q_l^m[e^{-i\omega_1 t}, e^{\lambda_2 t} r_{k_2,\lambda_2}](t) = e^{-i\omega_1 t} \int_0^t \int_{-\gamma s}^{(1-\gamma)s} e^{i\omega_1 \tau} e^{i\omega_1 \gamma s} e^{\lambda_2 (t-s)} r_{k_2,\lambda_2}(t-s) \tau^l s^m u(k_1 \tau) \mathrm{d}\tau \mathrm{d}s.$$

So, using (5.32),we introduce

$$\mathfrak{R}_-(s) = \int_{-\infty}^{-\gamma s} e^{i\omega_1 \tau} \tau^l u(k_1 \tau) \mathrm{d}\tau \text{ and } \mathfrak{R}_+(s) = \int_{(1-\gamma)s}^{\infty} e^{i\omega_1 \tau} \tau^l u(k_1 \tau) \mathrm{d}\tau$$

and

$$A = \int_{\mathbb{R}} e^{i\omega_1 \tau} \tau^l u(k_1 \tau) \mathrm{d}\tau \text{ and } B_{\lambda_2,p} = \int_0^{\infty} e^{-i\omega_1 \gamma s} s^p e^{\lambda_2 s} r_{k_2,\lambda_2}(s) \mathrm{d}s, \tag{5.35}$$

where $B_{\lambda_2,p}$ is well defined if $\lambda_2$ is negative enough (i.e. $\lambda_2 < -\gamma|\lambda_0|$). Consequently, we get (since $1-\gamma = \frac{|k|}{|k_1|}$)

$$\begin{aligned}
& q_l^m[e^{-i\omega_1 t}, e^{\lambda_2 t} r_{k_2,\lambda_2}](t) \\
=& A e^{-i\omega_1 t} \int_0^t e^{i\omega_1 \gamma s} s^m e^{\lambda_2 (t-s)} r_{k_2,\lambda_2}(t-s) \mathrm{d}s \\
& + \int_0^t e^{i\omega_1 \gamma s} e^{\lambda_2 (t-s)} s^m r_{k_2,\lambda_2}(t-s) \left(\mathfrak{R}_-(s) + \mathfrak{R}_+(s)\right) \mathrm{d}s \\
=& A e^{-i\omega_1 \frac{|k|}{|k_1|} t} \int_0^t e^{-i\omega_1 \gamma s} (t-s)^m e^{\lambda_2 s} r_{k_2,\lambda_2}(s) \mathrm{d}s \\
& + \int_0^t e^{i\omega_1 \gamma s} e^{\lambda_2 (t-s)} s^m r_{k_2,\lambda_2}(t-s) \left(\mathfrak{R}_-(s) + \mathfrak{R}_+(s)\right) \mathrm{d}s \\
=& \sum_{p=0}^m C_m^p t^{m-p} A B_{\lambda_2,p} e^{-i\omega_1 \frac{|k|}{|k_1|} t} \\
& + \sum_{p=0}^m C_m^p t^{m-p} A e^{-i\omega_1 \frac{|k|}{|k_1|} t} \int_t^{\infty} e^{-i\omega_1 \gamma s} s^p e^{\lambda_2 s} r_{k_2,\lambda_2}(s) \mathrm{d}s \\
& + \int_0^t e^{i\omega_1 \gamma s} e^{\lambda_2 (t-s)} s^m r_{k_2,\lambda_2}(t-s) \left(\mathfrak{R}_-(s) + \mathfrak{R}_+(s)\right) \mathrm{d}s,
\end{aligned}$$

where $C_m^p = \binom{m}{p}$ is a binomial coefficient. Here there are three kinds of terms. The first one is one of expected leading term. The two others are remainder terms. So we have to control them.

First, we control the second kind of term. If $t > 0$ then

$$\begin{aligned}
& \left| e^{-\lambda t} e^{-i\omega_1 \frac{|k|}{|k_1|} t} \int_t^{\infty} e^{-i\omega_1 \gamma s} s^p e^{\lambda_2 s} r_{k_2,\lambda_2}(s) \mathrm{d}s \right| \\
& \qquad \leqslant C_{\lambda_2} e^{-\lambda t} e^{\Im\omega_1 \frac{|k|}{|k_1|} t} \int_t^{\infty} e^{\Im\omega_1 \gamma s} s^p e^{\lambda_2 s} \mathrm{d}s \\
& \qquad\qquad \leqslant C_{\lambda_2} \int_{s>0} s^p e^{[|\Im\omega_1| + |\lambda| + \lambda_2] s} s^p e^{\lambda_2 s} \mathrm{d}s.
\end{aligned}$$

So this last quantity is finite if $\lambda_2$ is negative enough.

Then we control the last kind of term. If $t > 0$ then

$$\begin{aligned}\left|e^{-\lambda t}\int_0^t e^{i\omega_1\gamma s}e^{\lambda_2(t-s)}s^m r_{k_2,\lambda_2}(t-s)\mathfrak{R}_-(s)\mathrm{d}s\right| \\ \leqslant C_{\lambda_2}C_{\lambda_3}e^{-\lambda t}\int_0^t e^{-\Im\omega_1\gamma s}e^{\lambda_2(t-s)}s^m\int_{\gamma s}^{\infty}e^{-\Im\omega_1\tau}\tau^l e^{\lambda_3|k_1|\tau}\mathrm{d}\tau\mathrm{d}s \\ \leqslant C_{\lambda_2}C_{\lambda_3}t^m e^{(\lambda_2-\lambda+|\Im\omega_1|\gamma)t}\int_{\tau>0}\tau^l e^{\left(\lambda_3|k_1|-\Im\omega_1+\frac{|\lambda_2|}{\gamma}\right)\tau}\mathrm{d}\tau.\end{aligned}$$

So this last quantity is bounded uniformly with respect to $t$ if $\lambda_2 < \lambda - |\Im\omega_1|\gamma$ and $\lambda_3|k_1| < \Im\omega_1) - \frac{|\lambda_2|}{\gamma}$. Of course, we could control the other remainder term (with $\mathfrak{R}_+$) in a similar way.

Concerning the degree, it is smaller or equal than the degree of $P_{k_1,\omega_1}$, that is $\leqslant n_{k_1,\omega_1} - 1$, as $j + m \leqslant \deg P_{k_1,\omega_1}$. This is for the moment one degree less than what is expected in the Lemma 5.4.3.

**Remark 5.4.5.** *Note the term $B_{\lambda_2,p}$ in* (5.35) *is not explicit, as it relies on a remainder term of the first order dispersion relation. It is worth mentioning that this term contributes to the second order expansion, and not as a remainder term.*

**First term**

Finally we study the first kind of terms in the expansion of $F_k^2$. These terms are of the type $q[P_{k_1,\omega_1}e^{-i\omega_1 t}, P_{k_2,\omega_2}e^{-i\omega_2 t}]$. By a straightforward calculation, as in the previous case, it can be extended as a linear combination of terms of the type $t^j q_{l+1}^m[e^{-i\omega_1 t}, t^n e^{-i\omega_2 t}]$ and of the type $t^j q_l^{m+1}[e^{-i\omega_1 t}, t^n e^{-i\omega_2 t}]$ with $j + l + m = \deg P_{k_1,\omega_1}$ and $n \leqslant \deg P_{k_2,\omega_2}$.

In order to pursue the proof for this first kind of terms, in the following elementary lemma, we introduce a useful algebraic decomposition. It is proven in Appendix 5.6.2.

**Lemma 5.4.6.** *For all $n, m \in \mathbb{N}$, for all $\omega \in \mathbb{C}$, there exists $Q_{m,n,\omega}, R_{m,n,\omega} \in \mathbb{C}[X]$ such that*

$$\forall t > 0,\ \int_0^t e^{i\omega s}s^m(t-s)^n\mathrm{d}s = Q_{m,n,\omega}(t)e^{i\omega t} + R_{m,n,\omega}(t).$$

*If $\omega \neq 0$ then $\deg Q_{m,n,\omega} = m$ and $\deg R_{m,n,\omega} = n$. If $\omega = 0$ then $Q_{m,n,\omega} = 0$ and $\deg R_{m,n,\omega} = m + n + 1$.*

**Remark 5.4.7.** *The fact that the degree of $R_{m,n,\omega}$ can change contains the discussion on the multiplicity. Indeed, it will be applied for $\omega = \gamma\omega_1 + \omega_2$ which is equal to zero when $\omega_1 + \omega_2 = \frac{|k|}{|k_1|}\omega_1$, since $\frac{|k|}{|k_1|} = \frac{|k_1+k_2|}{|k_1|} = (1-\gamma)$.*

Furthermore, using the previous constructions, we introduce

$$B(t) = \int_{s>0} e^{i\omega_1\gamma s}e^{i\omega_2 s}(t-s)^n s^m\left(\mathfrak{R}_-(s) + \mathfrak{R}_+(s)\right)\mathrm{d}s \in \mathbb{C}_n[t].$$

Now, if $t > 0$, we have

$$\begin{aligned}
&q_l^m[e^{-i\omega_1 t}, t^n e^{-i\omega_2 t}](t)\\
=&\int_0^t \int_{-\gamma s}^{(1-\gamma)s} e^{-i\omega_1(t-\tau-\gamma s)} e^{-i\omega_2(t-s)}(t-s)^n \tau^l s^m u(k_1\tau)\mathrm{d}\tau\mathrm{d}s\\
=&A\int_0^t e^{-i\omega_1(t-\gamma s)} e^{-i\omega_2(t-s)}(t-s)^n s^m \mathrm{d}s\\
&+\int_0^t e^{-i\omega_1(t-\gamma s)} e^{-i\omega_2(t-s)}(t-s)^n s^m \left(\mathfrak{R}_-(s)+\mathfrak{R}_+(s)\right)\mathrm{d}s\\
=&Ae^{-i(\omega_1+\omega_2)t}\left[Q_{m,n,\gamma\omega_1+\omega_2}(t)e^{i(\gamma\omega_1+\omega_2)t}+R_{m,n,\gamma\omega_1+\omega_2}(t)\right]\\
&+B(t)e^{-i(\omega_1+\omega_2)t}-\int_t^\infty e^{-i\omega_1(t-\gamma s)} e^{-i\omega_2(t-s)} s^m \left(\mathfrak{R}_-(s)+\mathfrak{R}_+(s)\right)\mathrm{d}s\\
=&(AR_{m,n,\gamma\omega_1+\omega_2}+B(t))e^{-i(\omega_1+\omega_2)t}+AQ_{m,n,\gamma\omega_1+\omega_2}(t)e^{-i\omega_1\frac{|k|}{|k_1|}t}\\
&-\int_t^\infty e^{-i\omega_1(t-\gamma s)} e^{-i\omega_2(t-s)}(t-s)^n s^m \left(\mathfrak{R}_-(s)+\mathfrak{R}_+(s)\right)\mathrm{d}s.
\end{aligned}$$

Finally we just have to prove that this last integral is a remainder term. Indeed, we have

$$\begin{aligned}
&\left|e^{-\lambda t}\int_t^\infty e^{-i\omega_1(t-\gamma s)} e^{-i\omega_2(t-s)}(t-s)^n s^m \mathfrak{R}_-(s)\mathrm{d}s\right|\\
\leqslant& C_{\lambda_3} t^n e^{(-\lambda+\Im\omega_1+\Im\omega_2)t}\int_t^\infty e^{-(\gamma\Im\omega_1+\Im\omega_2)s} s^m \int_{\gamma s}^\infty e^{\Im\omega_1\tau}\tau^l e^{\lambda_3|k_1|\tau}\mathrm{d}\tau\mathrm{d}s\\
\leqslant& C_{\lambda_3}\int_t^\infty e^{-\gamma s}\int_{\gamma s}^\infty \frac{\tau^{n+l+m}}{\gamma^{n+m}} e^{(1+\Im\omega_1+\frac{|-\lambda+\Im\omega_1+\Im\omega_2|+|\gamma\Im\omega_1+\Im\omega_2|}{\gamma}+\lambda_3|k_1|)\tau}\mathrm{d}\tau\mathrm{d}s\\
\leqslant& C_{\lambda_3}\int_{s>0} e^{-\gamma s}\mathrm{d}s\int_{\tau>0}\frac{\tau^{n+l+m}}{\gamma^{n+m}} e^{(1+\Im\omega_1+\frac{|-\lambda+\Im\omega_1+\Im\omega_2|+|\gamma\Im\omega_1+\Im\omega_2|}{\gamma}+\lambda_3|k_1|)\tau}\mathrm{d}\tau.
\end{aligned}$$

So this last quantity is finite if $\lambda_3$ is negative enough.

Concerning the degree, we consider first the case $\gamma\omega_1+\omega_2 \neq 0$. As $B$ is of degree $\leqslant n$ and $R_{m,n,\gamma\omega_1+\omega_2}$ is of degree $\leqslant n$. So we get, as $j \leqslant \deg P_{k_1,\omega_1}$ and $n \leqslant \deg P_{k_2,\omega_2}$, that $Q^{k_1,k_2}_{\omega_1,\omega_2}$ is of degree $\leqslant n_{k_1,\omega_1}-1+n_{k_2,\omega_2}-1$, which is the expected value. Now, as we can have a $q_l^{m+1}$ term, leading to $Q_{m+1,n,\gamma\omega_1+\omega_2}$ which is of degree $\leqslant m+1$ and as $m$ can be chosen $\leqslant \deg P_{k_1,\omega_1}$, $R^{\omega_1}_{k_1,k_2}$ is of degree $\leqslant n_{k_1,\omega_1}$, which is now the expected value. We consider finally the case $\gamma\omega_1+\omega_2=0$, so that the terms $e^{-i(\omega_1+\omega_2)t}$ and $e^{-i\omega_1\frac{|k|}{|k_1|}t}$ are the same. The terms of highest degree is then $R_{m+1,n,\gamma\omega_1+\omega_2}$ which is here of degree $\leqslant m+n+2$, that is $\leqslant n_{k_1,\omega_1}-1+n_{k_2,\omega_2}-1+2$. All the values found are thus those that are expected.

### 5.4.8 Proof of Proposition 5.4.1

*Proof of Proposition 5.4.1.* In Proposition 5.3.2 and 5.4.1 we have proven that we can apply Lemma 5.3.4 with $1-R(z)=D_k(z+i\widetilde{\lambda})$ and $N(z)=\mathcal{N}_k^1(z+i\widetilde{\lambda})+\mathcal{N}_k^2(z+i\widetilde{\lambda})$, taking $\lambda_0=\widetilde{\lambda}/|k|$

in Proposition 5.3.2 : we get from Proposition 5.3.2 and Lemma 5.4.1

$$\forall z \in i\Sigma_{\gamma+\frac{\pi}{2}},\ |zR(z)| \leqslant \frac{C}{|k|},\ \sup_{z\in i\Sigma_{\beta+\frac{\pi}{2}}} |zN(z)| < \infty.$$

But the result of this lemma is that for all $\lambda \in \mathbb{R}$, we have

$$\frac{N(z)}{1-R(z)} = \mathcal{L}\Bigg[\sum_{\substack{\omega \text{ pole of } \frac{N}{1-R} \\ \Im\omega \geqslant \lambda}} P_\omega(t)e^{-i\omega t} + e^{\lambda t} r(t)\Bigg](z),$$

with a function $r \in \mathcal{H}(\Sigma_{\tilde\beta})$ analytic and bounded on $\Sigma_{\tilde\beta}$, for $\Im z$ large enough, with some $\tilde\beta$ satisfying $0 < \tilde\beta < \gamma < \beta$ and $P_\omega$ is the polynomial such that

$$\frac{N(z)}{1-R(z)} \underset{z\to\omega}{=} \mathcal{L}[P_\omega(t)e^{-i\omega t}] + \mathcal{O}(1).$$

Thus, we have

$$\begin{aligned}
\frac{\mathcal{N}_k^1(z)+\mathcal{N}_k^2(z)}{D_k(z)} &= \mathcal{L}\Bigg[\sum_{\substack{\omega \text{ pole of } \frac{N}{1-R} \\ \Im\omega \geqslant \lambda}} P_\omega(t)e^{-i\omega t} + e^{\lambda t} r(t)\Bigg](z-i\widetilde{\lambda}) \\
&= \mathcal{L}\Bigg[\sum_{\substack{\omega \text{ pole of } \frac{N}{1-R} \\ \Im\omega \geqslant \lambda}} P_\omega(t)e^{-i(\omega+i\widetilde{\lambda})t} + e^{(\lambda+\widetilde{\lambda})t} r(t)\Bigg](z) \\
&= \mathcal{L}\Bigg[\sum_{\substack{\omega \text{ pole of } \frac{\mathcal{N}_k^1+\mathcal{N}_k^2}{D_k} \\ \Im(\omega) \geqslant \lambda+\widetilde{\lambda}}} P_{\omega-i\widetilde{\lambda}}(t)e^{-i\omega t} + e^{(\lambda+\widetilde{\lambda})t} r(t)\Bigg](z)
\end{aligned}$$

So, defining $\mu$ by

$$\widehat{\mu}(t,k) = \sum_{\substack{D(\omega)=1 \\ \Im(\omega) \geqslant \lambda+\widetilde{\lambda}}} P_{\omega-i\widetilde{\lambda}}(t)e^{-i\omega t} + e^{(\lambda+\widetilde{\lambda})t} r(t),$$

we get (5.26), which is (5.13). We finally have the expansion of Theorem 5.1.5. Concerning the multiplicity, if one pole is common to $\mathcal{N}_k^1 + \mathcal{N}_k^2$ and $D_k^{-1}$ we have to sum up the multiplicity, leading to add $n_{k,\omega_1+\omega_2} - 1$ to the range for $\ell$ and $n_{k,\frac{|k|}{|k_1|}\omega_1} - 1$ to the range for $p$. The other concerns about the multiplicity follow from Lemmae 5.4.2 and 5.4.3, and the condition $k \cdot k_1 \neq 0$ directly follows from the factor $k \cdot k_1$ in front of (5.24) and (5.25). Note also that $\mathbb{R}_+^* \subset \Sigma_{\widetilde\beta}$, so that $r$ is bounded on $\mathbb{R}_+^*$ as stated in Theorem 5.1.5. □

## 5.5 Numerical results

Simulations have already been performed for multi-dimensional and multi-species simulations in [20], highlighting the relevance of second order expansion. We focus here more specifically on exhibiting a case where the Best frequency, that corresponds to the terms $B$ in Theorem 5.1.5, appears.

### 5.5.1 First example

We consider the one dimensional case ($d = 1$ and $L_1 = 2\pi$) and solve numerically (VP) with a Semi-Lagrangian scheme and an adapted $6$-th order splitting [44]. $1D$ periodic centered Lagrange interpolation of degree $17$ is used in both $x$ and $v$ directions and the periodic Poisson solver is solved with fast Fourier transform.

Initial condition is $f_0(x,v) = f^{eq}(v) + \varepsilon g_0(x,v)$, with

$$f^{eq}(v) = e^{-v^2/2},\ g_0(x,v) = \cos(2x)e^{-v^2/(2\sigma_2^2)} + \cos(3x)e^{-v^2/(2\sigma_3^2)}$$

and $\sigma_2 = 2^{1/4},\ \sigma_3 = \sqrt{\pi}/2$ and $\varepsilon = 0.001$.

We take $v \in [-v_{\max}, v_{\max}]$, with $v_{\max} = 10$. Numerical parameters are : the number of uniform cells in $x$ (resp. $v$) that are $N_x$ (resp. $N_v$) and the time step $\Delta t \in \mathbb{R}_+^*$, leading to a grid which will be referred as $N_x \times N_v \times \Delta t$ grid.

The first Fourier mode $\widehat{E}_{1,num}(t)$ of the electric field $E := -\nabla\Phi$ is computed from the simulation at each time step $t = t_n = n\Delta t$, using a discrete Fourier transform.

We first compute the zeros of $D_k = D_{-k}$ (see Remark 5.6.3), for $|k| = 1,2,3$ with greatest imaginary part that are

$$\begin{aligned}
\omega_{1,\pm} &\simeq \pm 2.511728081 - 0.4796966410i,\\
\omega_{2,\pm} &\simeq \pm 3.734976684 - 2.087460944i,\\
\omega_{3,\pm} &\simeq \pm 4.866872949 - 4.113005968i.
\end{aligned}$$

The second frequency of the mode $1$ is $\omega_{1,\pm}^{(2)} \simeq \pm 3.498058625 - 2.374303389i$. Such zeros can be computing with a symbolic calculus software. An example using Maple is provided in the Appendix. Here the modes that are initialized are $k_1, k_2 \in \{\pm 2, \pm 3\}$. The main term is for $k = k_1 + k_2 = \pm 1$, with $k_1 = \mp 2$ and $k_2 = \pm 3$, as $\omega_{\pm 1}$ has the greatest imaginary part among the $\omega_{k_1+k_2}$, with $k_1, k_2 \in \{\pm 2, \pm 3\}$. For having $k_2 = -\gamma k_1$, with $\gamma \in (0,1)$, we have to take $k_2 = \pm 2$ and $k_1 = \mp 3$, so that the Best frequencies $\omega_{b,\pm}$ of greatest imaginary part are defined by

$$\omega_{b,\pm} = \frac{|k_1 + k_2|}{|k_1|}\omega_{3,\pm} = \frac{\omega_{3,\pm}}{3}.$$

In order to see such term, we have to remove the main part coming from $\omega_{\pm 1}$. The procedure is detailed as follows. From Theorem 5.1.5, we look here for

$$\Re(\widehat{E}_{1,num})(t) \simeq \Re\left(ze^{-i\omega_1 t} + (z_1 + tz_2)e^{-i\omega_b t}\right), \text{with } z, z_1, z_2 \in \mathbb{C},$$

with $\omega_1 = \omega_{1,+}$ or $\omega_1 = \omega_{1,-}$, as it leads to the same value, and similarly for $\omega_b$. We estimate $z$ by using a least square procedure : we first define

$$\chi^2(y) = \sum_{t_{\min} \leqslant t_j \leqslant t_{\max}} \left(\Re\left(ye^{-i\omega_1 t_j}\right) - \Re(\widehat{E}_{1,num})(t_j)\right)^2$$

and then define $z$ by minimizing this quantity, that is, $\chi^2(z) = \min_{y\in\mathbb{C}} \chi^2(y)$, which is explicitely given by as solution of

$$A^T A \begin{bmatrix} \Re(z) \\ \Im(z) \end{bmatrix} = A^T b,\ A = [\Re(e^{-i\omega_1 t_j})_j; -\Im(e^{-i\omega_1 t_j})_j],\ b = \Re(\widehat{E}_{1,num})(t_j)_j,$$

with $A$ a matrix given by its 2 columns and $b$ a vector, all the three vectors being indexed by $j$ that goes through all the values such that $t_{\min} \leqslant t_j \leqslant t_{\max}$.

Once $z$ is found, we estimate $z_1$ and $z_2$ using again a least square procedure on the remainder : defining this time

$$\tilde{\chi}^2(y_1, y_2) = \sum_{\tilde{t}_{\min} \leqslant t_j \leqslant \tilde{t}_{\max}} \left( \Re\left((y_1 + t_j y_2) e^{-i\omega_b t_j}\right) - \Re\left(\widehat{E}_{1,num}(t_j) - z e^{-i\omega_1 t}\right)\right)^2,$$

$z_1$ and $z_2$ are obtained by minimizing this quantity, that is,

$$\tilde{\chi}^2(z_1, z_2) = \min_{y_1, y_2 \in \mathbb{C}} \tilde{\chi}^2(y_1, y_2).$$

Again the solution is explicitely given, the matrix $A$ being here

$$A = [\Re(e^{-i\omega_b t_j})_j; -\Im(e^{-i\omega_b t_j})_j; \Re(t_j e^{-i\omega_b t_j})_j; -\Im(t_j e^{-i\omega_b t_j})_j].$$

On Figure 5.2, we represent the time evolution of the real part of the first Fourier mode $|\Re(\widehat{E}_{1,num})(t)|$ in absolute value, together with $|\Re(\widehat{E}_{1,num})(t) - \Re\left(z e^{-i\omega_1 t}\right)|$, that is the quantity where we have removed the main part (it is a term $J$ in Theorem 5.1.5) ; the latter is compared to $|\Re\left((z_1 + t_j z_2) e^{-i\omega_b t_j}\right)|$ that corresponds to the Best term. The parameters $t_{\min}$, $t_{\max}$, $\tilde{t}_{\min}$ and $\tilde{t}_{\max}$ are chosen properly so that, in the corresponding interval, the approximation is valid. Note that a too low value is not good, as the expansion is only asymptotic and we consider only one term which is the main term asymptotically. A too high value is also not good, as we have to face with the round off or numerical error and the nonlinear behavior (note that we do not solve here the second linearized equation but the full nonlinear equation). We observe a well agreement, which is even better, by refining the grid, so that we can claim that we have exhibited the Best frequency in the numerical results, which is fully coherent with the theoretical results.

### 5.5.2 Another case where the Best frequency is almost dominant on a spatial mode

Now we consider again $d = 1$ (dimension $1$), but we change the spatial length of the domain $L_1 = 20\pi$, and take

$$f^{eq}(v) = e^{-v^2/2},\ g_0(x, v) = \cos(x) e^{-v^2/(2\sigma_2^2)} + \cos(0.1x) e^{-v^2/(2\sigma_3^2)}$$

and $\sigma_2 = 2^{1/4}$, $\sigma_3 = \sqrt{\pi}/2$ and $\varepsilon = 0.001$. Now the modes that are initialized are $k_1, k_2 \in \{\pm 1, \pm 0.1\}$. We now need to know (we already have the value of $\omega_{1,\pm}$ from the previous subsection)

$$\begin{aligned}
\omega_{0.1,\pm} &\simeq \pm 1.592755970 + 3.218848582 \cdot 10^{-52} i, \\
\omega_{0.2,\pm} &\simeq \pm 1.621955006 - 2.569883158 \cdot 10^{-12} i, \\
\omega_{0.9,\pm} &\simeq \pm 2.382548194 - 0.3594880484 i, \\
\omega_{1.1,\pm} &\simeq \pm 2.639613224 - 0.6100786528 i.
\end{aligned}$$

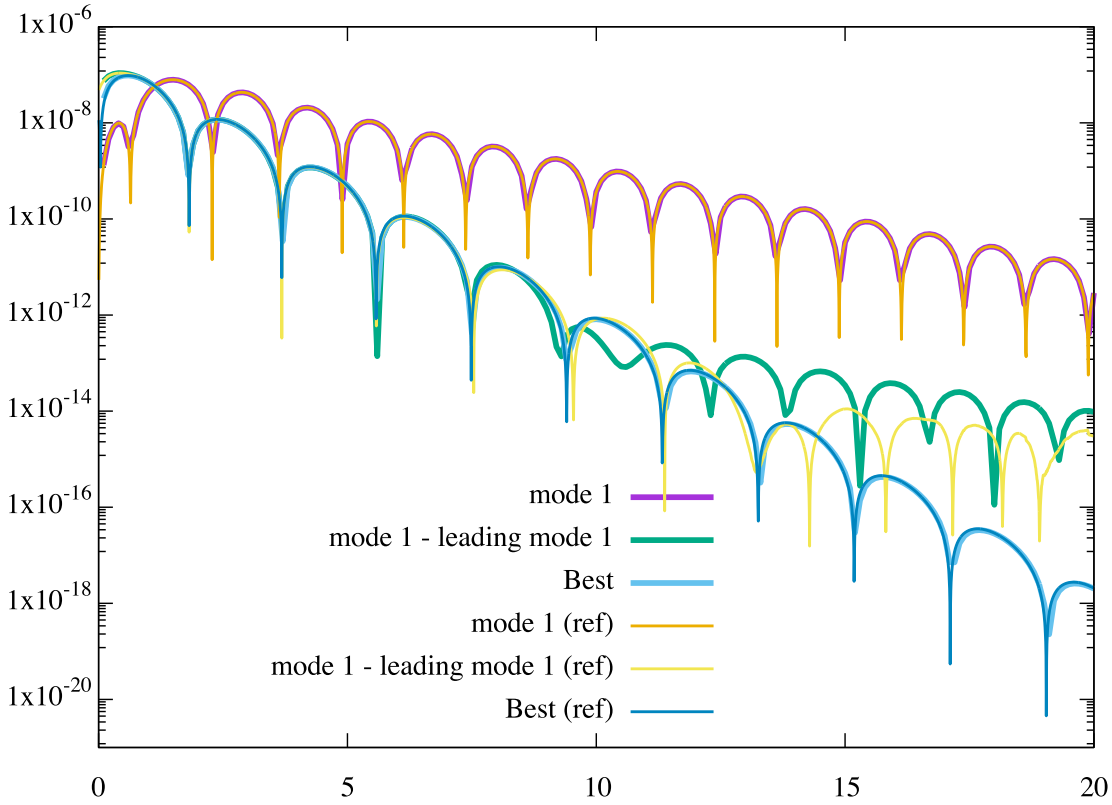


FIGURE 5.2 – Time evolution of $|\Re(\widehat{E}_{1,num})(t)|$ (mode 1), $|\Re\left(\widehat{E}_{1,num}(t) - ze^{-i\omega_1 t}\right)|$ (mode 1 - leading mode 1) and $|\Re\left((z_1 + t_j z_2)e^{-i\omega_b t_j}\right)|$ (Best), for coarse $128 \times 256 \times 0.1$ and refined $2048 \times 4096 \times 0.00625$ grids, the latter being referred as (ref) in the legend. The parameters $[t_{\min}, t_{\max}] = [17.5, 35]$ and $[\tilde{t}_{\min}, \tilde{t}_{\max}] = [1.75, 17.5]$ are used for the least square procedures.

The second frequency of the mode $0.9$ is $\omega^{(2)}_{0.9,\pm} \simeq \pm 3.181466437 - 2.102684847i$. The possible values of $k = k_1 + k_2$ are in the set $\{\pm 0.2, \pm 0.9, \pm 1.1, \pm 2\}$. The first order expansion already gives a term that is not damped (the imaginary part is almost equal to zero). We also have terms on the second order expansion that are not damped (for $k = \pm 0.2$). Nevertheless, if one consider the mode $k = \pm 0.9$, one can look at $|\Re(\widehat{E}_{0.9,num})(t)|$. From Theorem 5.1.5, we look thus here for an approximation of $\varepsilon^{-2}\Re(\widehat{E}_{0.9,num})(t)$ in the form

$$\mathcal{E}(t,z) = \Re\left(z_1 e^{-i\omega_{0.9}t} + (z_2 t + z_3)e^{-i0.9\omega_1 t} + z_4 e^{-i(\omega_1+\omega_{0.1,-})t} + z_5 e^{-i(\omega_1+\omega_{0.1,+})t}\right),$$

with $z = (z_1, z_2, z_3, z_4, z_5) \in \mathbb{C}^5$, using again $\omega_\ell = \omega_{\ell,+}$ or $\omega_\ell = \omega_{\ell,-}$, for $\ell \in \mathbb{R}$, as it leads to the same result. In order to estimate $z$, we compute

$$\min_{y\in\mathbb{C}^5} \sum_{t_{\min}\leqslant t_j \leqslant t_{\max}} \left(e^{\lambda t_j}\Re\left(\mathcal{E}(t_j,y) - \varepsilon^{-2}\widehat{E}_{0.9,num}(t_j)\right)\right)^2,$$

that is attained for $y = z$, by using the least square method as previously. Note that we add here the weight $e^{\lambda t}$, with $\lambda = 0.48$ and then we look for all the coefficients in one step. The choice of the value of $\lambda$ is coherent with the fact that from Theorem 5.1.5, the function $e^{\lambda t}\left(\varepsilon^{-2}\Re(\widehat{E}_{0.9,num})(t) - \mathcal{E}(t,z)\right)$ should be bounded. Numerical results are shown on Figure 5.3 and Figure 5.4. We use $t_{\min} = 0$ and $t_{\max} = 30$ for the coarse grid and have increased $t_{\max}$ to $35$ for the fine grid (for the fine grid, we could even increase this value, which was not possible for the coarse grid : the results were worse, as the solution is not precise enough for the coarse grid on late times, as shown on on Figure 5.3 and Figure 5.4). For the fine grid, we could also not really increase further than around $t_{\max} = 50$, as we are limited, with non-linear effects, convergence and/or machine precision ; we have also preferred not to go until $t_{\max} = 50$, as it leads to a worser matching, since the least square procedure tends to match for values around $50$, where the matching is less good. We could also change the initial time, but it has not so much impact, as it was the case for the previous subsection, since we have added here a weight function in the least square procedure. We emphasize that we can again exhibit the Best frequency and also the two other types of frequencies, which are all in the same range, for this example. In order to get this results, we note that we had to adapt he strategy concerning the least square method that was presented for the first example ; this is due to the fact the several modes are in a similar range, and it was not easy to use the first procedure (used for the first example) to catch the different frequencies.

The values of $z$ are given here for coarse and fine mesh :

$$\begin{aligned}
z_{1,\text{coarse}} &\simeq 1.2463 - 11.578i, & z_{1,\text{fine}} &\simeq 1.2183 - 11.548i,\\
z_{2,\text{coarse}} &\simeq 0.21502 + 0.28932i, & z_{2,\text{fine}} &\simeq 0.23103 + 0.31652i,\\
z_{3,\text{coarse}} &\simeq -4.2852 + 9.2615i, & z_{3,\text{fine}} &\simeq -4.5484 + 8.9068i,\\
z_{4,\text{coarse}} &\simeq 2.3853 + 1.2186i, & z_{4,\text{fine}} &\simeq 2.7369 + 1.1629i,\\
z_{5,\text{coarse}} &\simeq 1.5556 + 1.2385i, & z_{5,\text{fine}} &\simeq 1.5611 + 1.1445i.
\end{aligned}$$

### 5.5.3 A $2D$ case

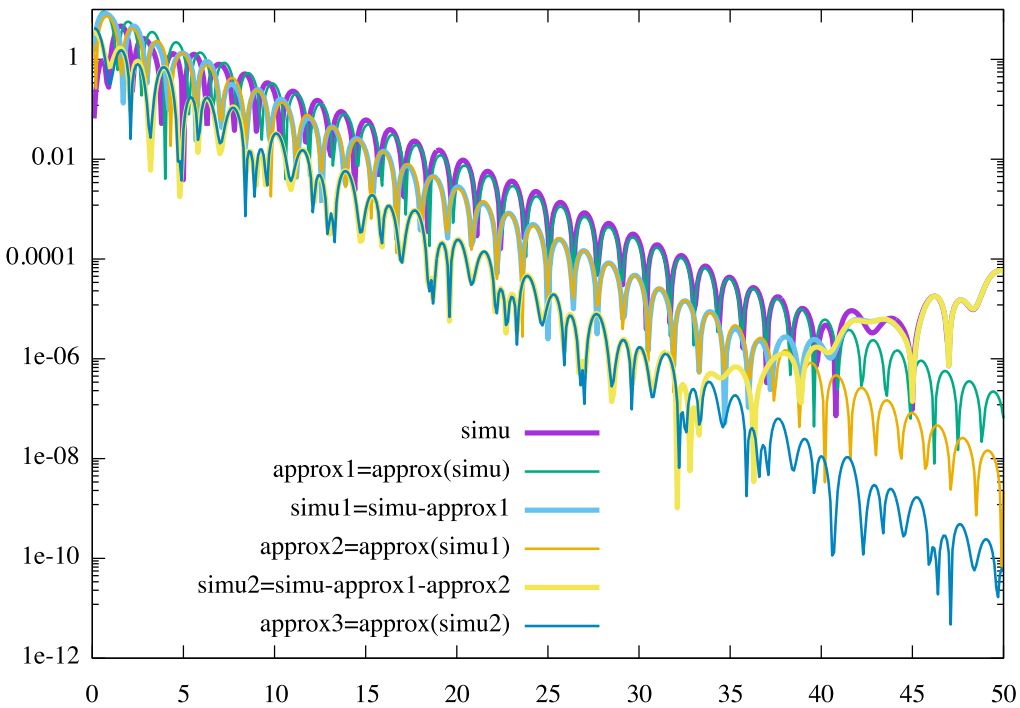


FIGURE 5.3 – Time evolution of

- $|\Re(\varepsilon^{-2}\widehat{E}_{1,num})(t)|$ (simu) vs $|z_1 e^{-i\omega_{0.9}t}|$ (approx1),
- $|\Re\left(\varepsilon^{-2}\widehat{E}_{1,num}(t) - z_1 e^{-i\omega_{0.9}t}\right)|$ (simu1) vs $|\Re\left((z_2 t + z_3)e^{-i0.9\omega_1 t}\right)|$ (approx2),
- $|\Re\left(\varepsilon^{-2}\widehat{E}_{1,num}(t) - z_1 e^{-i\omega_{0.9}t} - (z_2 t + z_3)e^{-i0.9\omega_1 t}\right)|$ (simu2) vs $|\Re\left(z_4 e^{-i(\omega_1+\omega_{0.1,-})t} + z_5 e^{-i(\omega_1+\omega_{0.1,+})t}\right)|$ (approx3),

for coarse $128 \times 256 \times 0.1$ grid. The parameters for the least square procedure is $[t_{\min}, t_{\max}] = [0, 30]$.

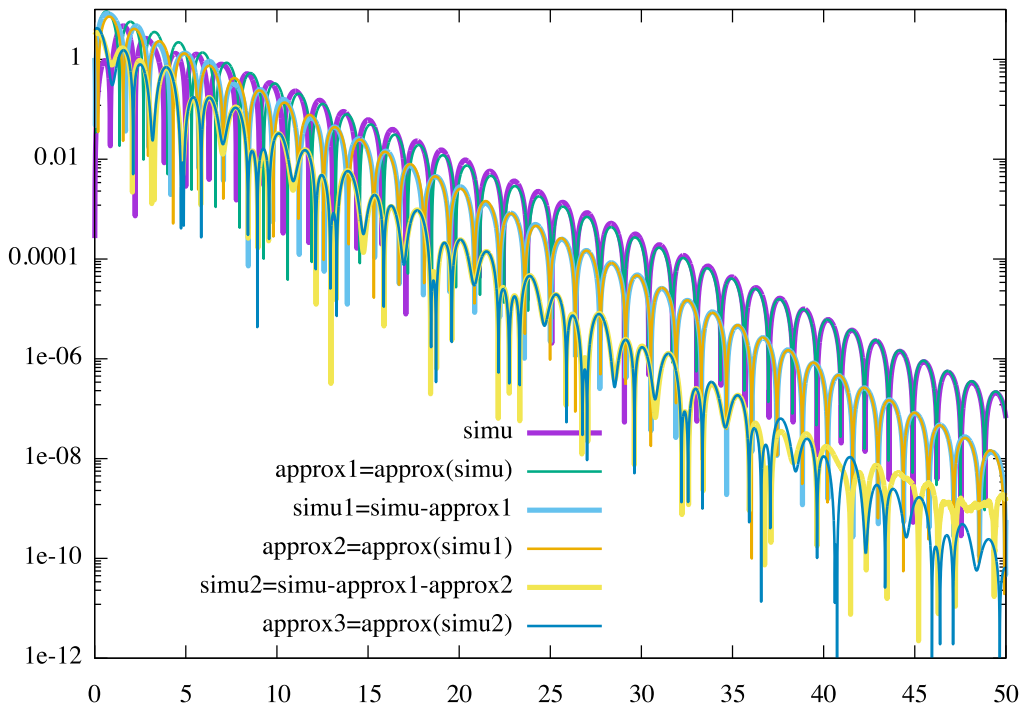


FIGURE 5.4 – Time evolution of

- $|\Re(\varepsilon^{-2}\widehat{E}_{1,num})(t)|$ (simu) vs $|z_1 e^{-i\omega_{0.9}t}|$ (approx1),
- $|\Re\left(\varepsilon^{-2}\widehat{E}_{1,num}(t) - z_1 e^{-i\omega_{0.9}t}\right)|$ (simu1)
  vs $|\Re\left((z_2 t + z_3)e^{-i0.9\omega_1 t}\right)|$ (approx2),
- $|\Re\left(\varepsilon^{-2}\widehat{E}_{1,num}(t) - z_1 e^{-i\omega_{0.9}t} - (z_2 t + z_3)e^{-i0.9\omega_1 t}\right)|$
  (simu2) vs $|\Re\left(z_4 e^{-i(\omega_1+\omega_{0.1,-})t} + z_5 e^{-i(\omega_1+\omega_{0.1,+})t}\right)|$ (approx3),

for refined $2048 \times 4096 \times 0.00625$ grid. The parameters for the least square procedure is $[t_{\min}, t_{\max}] = [0, 35]$.

### Looking for Best frequencies in 2D

Finally, we focus on a $2D$ case. Here we can write $k = k_1 + k_2$ with

$$k_j = \left(m_j \frac{2\pi}{L_1}, n_j \frac{2\pi}{L_2}\right),\ j = 1, 2,\ m_j, n_j \in \mathbb{Z}.$$

Now if $k = \gamma k_1$, with $\gamma \in (0, 1)$, we get :

$$m_1 + m_2 = \gamma m_1,\ n_1 + n_2 = \gamma n_1,$$

which leads to

$$1 - \gamma = -\frac{m_2}{m_1} = -\frac{n_2}{n_1},$$

if $m_1 \neq 0$ and $n_1 \neq 0$. If $m_1$ or $m_2 = 0$, we get $m_1 = m_2 = 0$, and similarly for $n_1$ and $n_2$. In order to have a "real" $2D$ case, we can suppose that $m_1 \neq 0$ and $n_1 \neq 0$. We have

$$\frac{-m_2}{m_1} = \frac{-n_2}{n_1} = 1 - \gamma = \frac{p}{q},\ p, q \in \mathbb{N},\ p < q,\ p \wedge q = 1.$$

So we obtain $-m_2 q = p m_1$, and thus $m_1 = \ell q$, $\ell \in \mathbb{Z}^*$ and $-m_2 = \ell p$ together with $n_1 = \tilde{\ell} q$, $\tilde{\ell} \in \mathbb{Z}^*$ and $-n_2 = \tilde{\ell} p$. Note that we then have $k \cdot k_1 = \gamma |k_1|^2 \neq 0$.

### A $2D$ test case with Best frequency

We choose here $L_1 = L_2 = L$, $m_1 = n_1 = 3,\ m_2 = n_2 = -2$, so that

$$k_1 = (3, 3)\frac{2\pi}{L},\ k_2 = (-2, -2)\frac{2\pi}{L},\ k_1 + k_2 = (1, 1)\frac{2\pi}{L},$$

and

$$|k_1| = 3\sqrt{2}\frac{2\pi}{L},\ |k_2| = 2\sqrt{2}\frac{2\pi}{L},\ |k_1 + k_2| = \sqrt{2}\frac{2\pi}{L}.$$

We will write $\omega_\ell$, instead of $\omega_{\ell,+}$ or $\omega_{\ell,-}$, when we can either use $\omega_{\ell,+}$ or $\omega_{\ell,-}$. We will need for this subsection and the next one, the following values (note that the values are here not the same as in the one dimensional case, since the dispersion relation is not the same, as we have considered here a normalized Maxwellian) :

$$\begin{aligned}
\omega_{\sqrt{2}/10,\pm} &\simeq \pm 1.030839024 - 6.410202539 \cdot 10^{-10} i,\\
\omega_{2\sqrt{2}/10,\pm} &\simeq \pm 1.140206800 - 0.007780445579 i,\\
\omega_{3\sqrt{2}/10,\pm} &\simeq \pm 1.316627173 - 0.08467369148 i,\\
\omega_{\sqrt{5}/10,\pm} &\simeq \pm 1.081943401 - 0.0004485284614 i,\\
\omega_{\sqrt{13}/10,\pm} &\simeq \pm 1.234323666 - 0.04025247555 i.
\end{aligned}$$

Also, the second frequency of the mode $\sqrt{2}/10$ is $\omega^{(2)}_{\sqrt{2}/10,\pm} \simeq \pm 0.5196579915 - 0.2520173386 i$.

The main frequencies that intervene on the spatial mode $(1,1)\frac{2\pi}{L}$ are $\omega_{\sqrt{2}\frac{2\pi}{L},\pm}$, and $\frac{\omega_{3\sqrt{2}\frac{2\pi}{L},\pm}}{3}$ (the last one is the Best frequency). In that case, we expect a similar behavior as the test case of the first subsection.

We use here $f^{eq}(v) = \frac{1}{2\pi}e^{-v_1^2/2-v_2^2/2}$, with $x = (x_1, x_2),\ v = (v_1, v_2)$, together with

$$g_0(x,v) = (\cos(0.3x_1 + 0.3x_2) + \cos(0.2x_1 + 0.2x_2))\, f^{eq}(v),$$

taking $L = 20\pi$. We solve again numerically (VP) with a Semi-Lagrangian scheme and an adapted $6$-th order splitting [44] (here $d = 2$). The parameter $\varepsilon$ is always fixed to $\varepsilon = 10^{-3}$.

We take $v_1, v_2 \in [-6, 6]$, $32$ cells in $x_1$ and $x_2$ directions, $64$ cells in $v_1$ and $v_2$ directions ; time step is fixed to $\Delta t = 0.1$, leading to a $32 \times 32 \times 64 \times 64 \times 0.1$ grid. The diagnostics are here obtained form the charge density $\rho(t, x_1, x_2) = \int_{\mathbb{R}^2} f \mathrm{d}v$ (computed from trapezoidal rule) : we define $\widehat{\rho}_{\ell_1,\ell_2,num}(t)$ the Discrete Fourier Transform of the charge density at time $t = t_n = n\Delta t$. Results are given on Figure 5.5. The least square procedure is here applied to minimize :

$$\min_{y\in\mathbb{C}^3} \sum_{t_{\min}\leqslant t_j\leqslant t_{\max}} \left(e^{\lambda t_j}\Re\left(\mathcal{E}(t_j, y) - \varepsilon^{-2}\widehat{\rho}_{1,1,num}(t_j)\right)\right)^2,$$

with

$$\mathcal{E}(t,y) = \Re\left(y_1 e^{-i\omega_{\sqrt{2}/10}t} + (y_2 t + y_3)e^{-i\frac{1}{3}\omega_{3\sqrt{2}/10}t}\right),$$

and is attained for $y_j = z_j,\ j = 1, 2, 3$, where the $z_j$ are given in Figure 5.5. We clearly see on Figure 5.5, that the Best frequency $\frac{\omega_{3\sqrt{2}\frac{2\pi}{L}}}{3}$ is needed : with the combination of the main frequency $\omega_{\sqrt{2}\frac{2\pi}{L}}$ the simulated mode $\widehat{\rho}_{1,1,num}$ is accurately asymptotically described. In that case, we see both frequencies are useful ; the main frequency is not enough as we can see it on Figure 5.5. Indeed both modes (main and Best) are shown (they are shifted towards bottom of the Figure in order to see them better), and we see that the *combination* of the modes is needed to describe the simulated mode.

### A $2D$ test case without Best frequency

Now, if we change and take $n_1 = 2,\ n_2 = -1$, we have no more Best frequency, and the main frequencies that intervene on the same spatial mode $(1,1)\frac{2\pi}{L}$ are $\omega_{\sqrt{2}\frac{2\pi}{L},\pm}$ and $\omega_{\sqrt{13}\frac{2\pi}{L},\pm} + \omega_{\sqrt{5}\frac{2\pi}{L},\pm}$ (thesse frequencies were defined in the previous subsection), as we have this time

$$k_1 = (3,2)\frac{2\pi}{L},\ k_2 = (-2,-1)\frac{2\pi}{L},\ k_1 + k_2 = (1,1)\frac{2\pi}{L},$$

and

$$|k_1| = \sqrt{13}\frac{2\pi}{L},\ |k_2| = \sqrt{5}\frac{2\pi}{L},\ |k_1 + k_2| = \sqrt{2}\frac{2\pi}{L},$$

the initial data being changed to

$$g_0(x,v) = (\cos(0.3x_1 + 0.2x_2) + \cos(0.2x_1 + 0.1x_2))\, f^{eq}(v),$$

and we have still $L = 20\pi$. For the least square procedure, we consider the minimization problem

$$\min_{y\in\mathbb{C}^3} \sum_{t_{\min}\leqslant t_j\leqslant t_{\max}} \left(e^{\lambda t_j}\Re\left(\mathcal{E}(t_j,y)-\varepsilon^{-2}\widehat{\rho}_{1,1,num}(t_j)\right)\right)^2,$$

with

$$\mathcal{E}(t,y)=\Re\left(y_1e^{-i\omega_{\sqrt{2}/10}t}+y_2e^{-i(\omega_{\sqrt{5}/10,+}+\omega_{\sqrt{13}/10,-})t}+y_3e^{-i(\omega_{\sqrt{5}/10,+}+\omega_{\sqrt{13}/10,+})t}\right),$$

attained for $y_j = z_j,\ j = 1,2,3$, where the $z_j$ are given in Figure 5.6. We remark here that we have some unexpected frequency at the beggining which might be interpreted as a Best frequency (the simu-first approx curve), but such one is damped and we get the right asymptotic behavior, which shows that we cannot get a Best frequency in the asymptotic limit, which is fully consistant with Theorem 5.1.5.

## 5.6 Appendix

### 5.6.1 Some remarks about the space $\mathscr{E}(\mathbb{R}^d)$

The aim of this subsection is to present some tools to construct explicit examples of functions of $\mathscr{E}(\mathbb{R}^d)$ (characterized by (5.10)).

The *Gelfand-Shilov* spaces $S^\beta_\alpha(\mathbb{R}^d)$ provide many useful examples of functions of $\mathscr{E}(\mathbb{R}^d)$. Their usual definition is the following $\alpha,\beta>0$,

$$S^\beta_\alpha(\mathbb{R}^d) := \{f\in\mathscr{S}(\mathbb{R}^d)\mid \exists\varepsilon,C>0,\forall v,\xi\in\mathbb{R}^d,\ |f(v)|\leqslant Ce^{-\epsilon|v|^{\frac{1}{\alpha}}} \text{ and } |\mathscr{F}f(\xi)|\leqslant Ce^{-\epsilon|\xi|^{\frac{1}{\beta}}}\}.$$

Many details about these spaces can be found in [94], in particular these spaces are stable by multiplication by a polynomial or a trigonometric polynomial, derivation and the natural action of the affine group of $\mathbb{R}^d$. Furthermore, we obviously have $\mathscr{F}S^\beta_\alpha(\mathbb{R}^d)=S^\alpha_\beta(\mathbb{R}^d)$.

**Proposition 5.6.1.** *If $\nu\in(0,1)$, then $S^{1-\nu}_\nu(\mathbb{R}^d)\subset\mathscr{E}(\mathbb{R}^d)$.*

*Proof.* It is a direct corollary of Proposition $6.1.8$ of [94]. □

**Example 5.6.1.**

- $|v|^2\cos(v_1-v_2)e^{-v_1^2-(v_1+v_2)^2}\in S^{\frac{1}{2}}_{\frac{1}{2}}(\mathbb{R}^2)\subset\mathscr{E}(\mathbb{R}^2)$,
- *If $k\in\mathbb{N}^*$ then $e^{-v^{2k}}\in S^{1-\frac{1}{2k}}_{\frac{1}{2k}}(\mathbb{R})\subset\mathscr{E}(\mathbb{R})$ (see [94]).*

To get some other example, we remark that $\mathscr{E}(\mathbb{R}^d)$ is clearly stable by multiplication by a trigonometric polynomial, derivation and the natural action of the affine group of $\mathbb{R}^d$. Furthermore, it enjoys the following tensor product property.

**Proposition 5.6.2.** *If $d_1,d_2\in\mathbb{N}^*$ then $\mathscr{E}(\mathbb{R}^{d_1})\otimes\mathscr{E}(\mathbb{R}^{d_2})\subset\mathscr{E}(\mathbb{R}^{d_1+d_2})$.*

**Example 5.6.2.** $\partial_{v_1}e^{-v_1^4-(v_1-3v_2)^2}\in\mathscr{E}(\mathbb{R}^2)$.

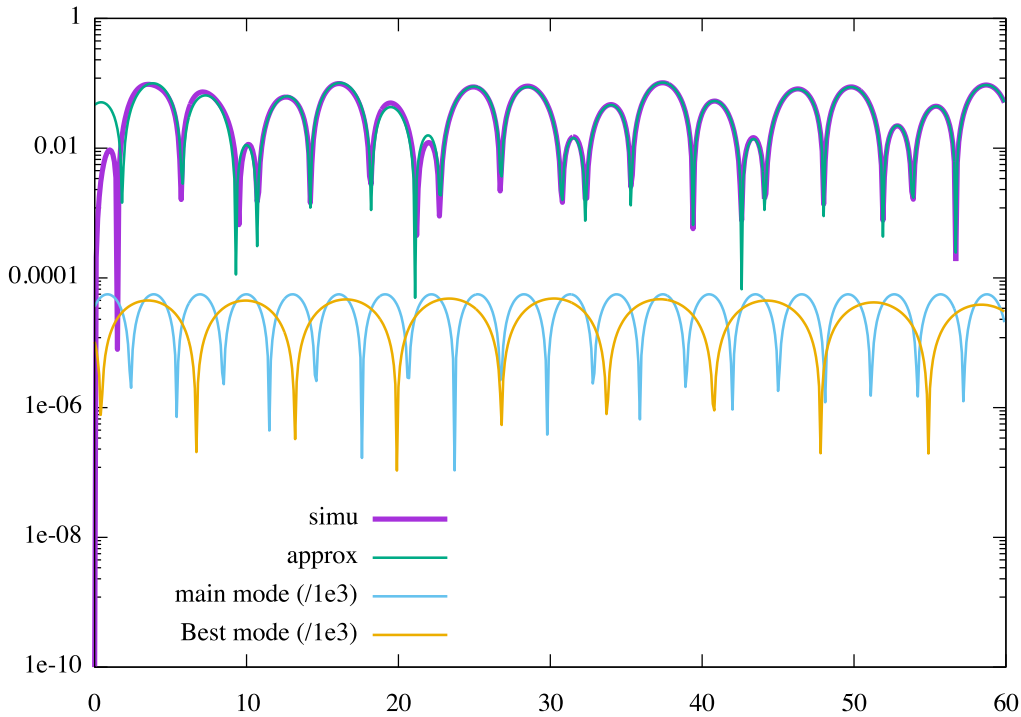


FIGURE 5.5 – A $2D$-case with Best frequency : time evolution of

- $|\Re(\varepsilon^{-2}\widehat{\rho}_{1,1,num})(t)|$ (simu)
- $|\Re\left(z_1 e^{-i\omega_{\sqrt{2}/10}t} + (z_2 t + z_3)e^{-i\frac{1}{3}\omega_{3\sqrt{2}/10}t}\right)|$ (approx)
- $10^{-3}|\Re\left(z_1 e^{-i\omega_{\sqrt{2}/10}t}\right)|$ (main mode /1e3)
- $10^{-3}|\Re\left((z_2 t + z_3)e^{-i\frac{1}{3}\omega_{3\sqrt{2}/10}t}\right)|$ (Best mode /1e3)

The parameters $\lambda = 0.09$ and $[t_{\min}, t_{\max}] = [0, 60]$ are used for the least square procedure to fit (simu) by (approx) and leads to $z_1 \simeq 0.036159 + 0.042602i$, $z_2 \simeq -0.0031761 - 0.00089598i$ and $z_3 \simeq 0.010351 - 0.046355i$.

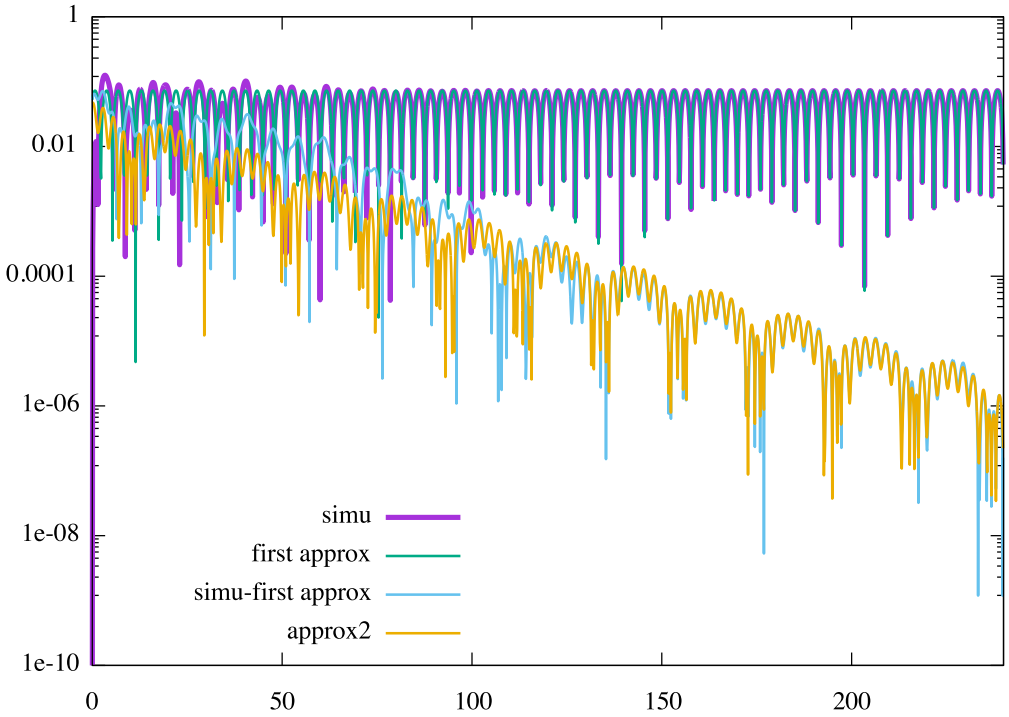


FIGURE 5.6 – A $2D$-case without Best frequency : time evolution of

- $|\Re(\varepsilon^{-2}\widehat{\rho}_{1,1,num})(t)|$ (simu)
- $|\Re\left(z_1 e^{-i\omega_{\sqrt{2}/10}t}\right)|$ (first approx)
- $|\Re\left(\varepsilon^{-2}\widehat{\rho}_{1,1,num}(t) - z_1 e^{-i\omega_{\sqrt{2}/10}t}\right)|$ (simu - first approx)
- $|\Re\left(z_2 e^{-i(\omega_{\sqrt{5}/10,+}+\omega_{\sqrt{13}/10,-})t} + z_3 e^{-i(\omega_{\sqrt{5}/10,+}+\omega_{\sqrt{13}/10,+})t}\right)|$ (approx2)

The parameters $\lambda = 0.05$ and $[t_{\min}, t_{\max}] = [0, 240]$ are used for the least square procedure leads to $z_1 \simeq 0.052836 + 0.049810i$, $z_2 \simeq -0.032921 - 0.0010657i$ and $z_3 \simeq -0.013703 - 0.0050901i$.

### 5.6.2 An algebraic decomposition

The aim of this subsection is to prove Lemma 5.4.6.

*Proof of Lemma 5.4.6.* If $\omega = 0$, we get the result by expanding the polynomial $(t-s)^n$. So we suppose now that $\omega \neq 0$. Since we recognize a convolution product, we apply a Laplace transform. So we get

$$\mathcal{L}\left[\int_0^t e^{i\omega s} s^m (t-s)^n \mathrm{d}s\right](z) = \mathcal{L}\left[t^m\right](z+\omega)\,\mathcal{L}\left[t^n\right](z) = \frac{n!m!(-i)^{n+m+2}}{(z+\omega)^{m+1}z^{n+1}}.$$

We can apply a partial fraction decomposition to get some complex coefficients $(a_j)_{j=0,\dots,n}$ and $(b_j)_{j=0,\dots,m}$ such that

$$\frac{n!m!(-i)^{n+m+2}}{(z+\omega)^{m+1}z^{n+1}} = \sum_{j=0}^{n} \frac{a_j}{z^{j+1}} + \sum_{j=0}^{m} \frac{b_j}{(z+\omega)^{j+1}}.$$

Consequently, we have

$$\mathcal{L}\left[\int_0^t e^{i\omega s} s^m (t-s)^n \mathrm{d}s\right](z) = \mathcal{L}\left[\sum_{j=0}^{n} \frac{a_j i^{j+1}}{j!} t^j + \sum_{j=0}^{m} \frac{b_j i^{j+1}}{j!} t^j e^{i\omega t}\right](z),$$

Since the Laplace transform characterizes the continuous functions with an exponential order (see Theorem $1.7.3$ in [6]), we have proved the lemma. □

### 5.6.3 Computation of the zeros

We have used the following Maple code to compute the zeros of $D_k$. We recall from Lemma 5.3.3 that for $\Im(z) > 0$

$$\begin{aligned} D_k(z) &= 1 - \frac{1}{|k|^2}\int \frac{k\cdot\nabla_v f^{eq}(v)}{v\cdot k - z}\mathrm{d}v = 1 - \frac{i}{|k|^2}\,\mathcal{L}\left[\mathscr{F}[k\cdot\nabla_v f^{eq}(v)](kt)\right](z) \\ &= 1 - \frac{i}{|k|^2}\,\mathcal{L}\left[ik\cdot(kt)\,\mathscr{F}[f^{eq}(v)](kt)\right](z) = 1 + \mathcal{L}\left[t\,\mathscr{F}[f^{eq}(v)](kt)\right](z) \\ &= 1 + \int_0^\infty t\,\mathscr{F}[f^{eq}(v)](kt)e^{izt}\mathrm{d}t = 1 + \int_0^\infty t\int_{\mathbb{R}^d} f^{eq}(v)e^{itk\cdot v}\mathrm{d}v e^{izt}\mathrm{d}t. \end{aligned}$$

We can write $v = v_\parallel + v_\perp$, with $v_\parallel$ the component of $v$ along $k$ and $v_\perp$ perpendicular to $k$,(when $d \geqslant 2$), so that

$$D_k(z) = 1 + \int_0^\infty t\int_{(\mathbb{R}^d)_\parallel}\left(\int_{(\mathbb{R}^d)_\perp} f^{eq}(v_\parallel + v_\perp)\mathrm{d}v_\perp\right)e^{itk\cdot v_\parallel}\mathrm{d}v_\parallel e^{izt}\mathrm{d}t.$$

**Remark 5.6.3.** *Note that $D_{-k}(z) = D_k(z)$, if $\mathscr{F}[f^{eq}(v)]$ is an even function.*

```
with(inttrans):
with(RootFinding):
Digits:=20:
feq:=exp(-((v)^2)/2);
#the space mode
k:=1.;
#Fourier transform of the equilibrium
Tfeq:=fourier(feq,v,t):
#the analytic function
Dk:=1+int(t*subs(t=k*t,Tfeq)*exp(I*om*t),t=0..infinity):
#the time modes
l:=sort([Analytic(Dk,om,re=-8..8,im=-8..8)],(a,b)->Im(a)>Im(b));
```

CHAPITRE 6

# SPLITTING METHODS FOR ROTATIONS : APPLICATION TO VLASOV EQUATIONS

## 6.1 Introduction

The main goal of this work is to introduce a splitting strategy to deal with rotations motions and to apply it to construct efficient high order time integrators for Vlasov type equations. The splitting is based on the fact that a rotation of angle $\theta$ can be decomposed into a product of three shear transformations

$$\begin{pmatrix} 1 & -\tan\theta/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \sin\theta & 1 \end{pmatrix} \begin{pmatrix} 1 & -\tan\theta/2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = e^{\theta J}, \tag{6.1}$$

for $\theta \neq k\pi, k \in \mathbb{Z}^\star$ and where $J$ is the $2$x$2$ following matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{6.2}$$

Note that this decomposition into shear matrices can be derived using formal computations and has been already introduced in the image processing community (see [97, 112, 5, 116, 47]), in which several approaches have been developed to rotate an image on a computer screen. Moreover, this approach has also been used to design numerical methods for Gross-Pitaevskii equations (see [49] and [9], Lemma II.2) in which this underlying splitting is used to solve exactly the harmonic oscillator.

To make the link between (6.1) and the underlying partial differential equation, we introduce the following two-dimensional transport equation

$$\partial_t u = Jx \cdot \nabla_x u, \quad x \in \mathbb{R}^2, \tag{6.3}$$

with the initial condition $u(t=0,x) = u^{in}(x)$. The exact solution of (6.3) at time $t$ writes $u(t,x) = u^{in}(e^{tJ}x)$ which is nothing but the rotation of angle $t$ of the initial condition $u^{in}$. When the initial condition is not known analytically or when equation (6.3) is a part of a more complicated model, then one only has access to a discrete information of the initial condition and a numerical method is required to approximate (6.3). Our goal in this work is to introduce a directional splitting inspired by (6.1) which is exact with respect to the time variable.

Obviously, standard finite differences or finite volumes based methods can be used to approximate the spatial direction $x$ and coupled to Runge-Kutta strategies in time. However, this

This chapter is a joint work with Nicolas Crouseilles and Fernando Casas realized in [25].

leads to methods which usually suffer from strong CFL condition on the time step. Then, semi-Lagrangian methods are preferred, since they are free from stability condition still keeping eulerian accuracy (see [107, 65, 119, 57]). For (6.3), the feet of the characteristics can be computed exactly and a two-dimensional interpolation has to be performed to update the numerical unknown. However, high-dimensional interpolation is known to be non conservative and it is obviously more demanding in term of complexity and time. Then, splitting methods are very competitive since they reduce the problem into very simple one-dimensional linear transport equations which can be solved with efficiently with semi-Lagrangian methods (using high-order or even spectral interpolation). Moreover, in a splitting procedure, the variable that does not appear in the derivative is just a parameter so that a very simple parallelization can be performed by distributing the computation on the processors according to the values of this parameter.

For rotation dynamics however, the standard splitting strategy (like Strang or Lie splitting for example) can induce some error since it involves a wrong rotational velocity (see [32]). Here, we propose a new splitting which enables to solve (6.3) exactly in time (like in [49, 9]). Moreover, when this splitting is coupled with spectral methods (and under some assumptions detailed in the sequel), the so-obtained method is able to capture to a very high accuracy the exact solution (spectral accuracy in practice). A complete proof of convergence of the fully discretized numerical method is performed. We will see that this strategy and some simple extensions turn out to be very efficient compared to standard methods when applied to the following problems. First, it enables us to design high order (in time) methods for the Vlasov-Maxwell system. Second, when applied to close to equilibrium of the Vlasov-HMF model as in [82], this splitting turns out to be more accurate than the Strang one, at the same cost.

Concerning the Vlasov-Maxwell solvers, our goal was to improve the method introduced in [55] in which a splitting into three parts has been proposed. Among these three parts, two were solved exactly in time whereas for the magnetic part, a standard directional splitting was performed. Here, the new method enables us to also solve this part exactly in time. This is then very helpful to design high order splitting methods for the full Vlasov-Maxwell system. The resulting schemes are fourth order accurate in time and preserves the Gauss condition exactly. We also use the new splitting to approximate the solution of the Vlasov-HMF system, for which the close to equilibrium dynamic is driven by the linearized Hamiltonian part (see [82]). For such Hamiltonian, the new splitting has a good behavior (see [21]) and we compare its efficiency to standard Strang splitting by studying perturbation of a non homogeneous equilibrium state.

The rest of the paper is organized as follows. First, the method is presented in the context of the numerical approximation of transport equation of the form (6.3) and a complete proof of convergence is performed with some numerical illustrations. Then, the Vlasov-Maxwell system is presented and we explain how the new method is used to design high-order Vlasov-Maxwell solver. Finally, some numerical results are given to show the benefit of the new method in the Vlasov context.

## 6.2 Presentation of the method and its numerical analysis

In this section, we focus on the following two-dimensional equation

$$\partial_t u = Jx \cdot \nabla u, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{6.4}$$

supplemented with an initial condition $u(t = 0, x) = u^{in}(x)$.

We intend to analyse the convergence of a splitting in time based numerical scheme coupled with a spectral method in space (ie in the $x$-direction). More precisely, we want to solve (6.4) on $[t_n, t_{n+1}]$ ; then we want to compute $u^{n+1}(x)$ an approximation of $u(t^{n+1}, x_1, x_2)$ the solution at time $t_{n+1} = t_n + \delta_t$ ($\delta_t > 0$ being the time step and $n \in \mathbb{N}$) of (6.4) with the initial condition $u^{in}(x_1, x_2) = u(t^n, x_1, x_2)$ at time $t^n = n\delta_t, n \in \mathbb{N}$. To do so, we propose a new splitting in which each step is a shear transformation.

Let us introduce some notations. For a given 2x2 matrix $A$, we denote by $\exp(\delta_t Ax \cdot \nabla)u^n$ the solution at time $t_{n+1}$ of

$$\left\{ \begin{array}{lcl} \partial_t u(t,x) & = & Ax \cdot \nabla u(t,x), \quad x \in \mathbb{R}^2 \\ u^{in}(x) & = & u^n(x) \end{array} \right. , \tag{6.5}$$

Then, inspired by (6.1), we search for $a, b \in \mathbb{R}$ so that the following relation holds true

$$e^{-\frac{a}{2}x_2\partial_{x_1}} e^{bx_1\partial_{x_2}} e^{-\frac{a}{2}x_2\partial_{x_1}} u^n = e^{\delta_t Jx\cdot\nabla} u^n, \tag{6.6}$$

which can be written equivalently by

$$e^{A_1 x\cdot\nabla} e^{A_2 x\cdot\nabla} e^{A_1 x\cdot\nabla} u^n = e^{\delta_t Jx\cdot\nabla} u^n, \tag{6.7}$$

with

$$A_1 = \begin{pmatrix} 0 & -a/2 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix}. \tag{6.8}$$

Using the method of the characteristics, we have for (6.5)

$$e^{\delta_t Ax\cdot\nabla u^n} = u^n \circ e^{\delta_t A}, \quad \delta_t \geqslant 0,$$

so that (6.7) is nothing but

$$u^n(e^{A_1} e^{A_2} e^{A_1} x) = u^n(e^{\delta_t J} x).$$

Since the exponential of matrices can be computed easily,

$$e^{A_1} = \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix}, \quad e^{A_2} = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix},$$

it comes from (6.1) that the choices $a = 2\tan(\delta_t/2)$, and $b = \sin(\delta_t)$, leads to an exact splitting in time so that the scheme then writes $u^{n+1}(x) = u^n(e^{A_1} e^{A_2} e^{A_1} x)$ with $A_1$ and $A_2$ given by (6.8). Let us remark that the usual Strang splitting corresponds to $a = b = \delta_t$.

Then, now we have to solve shear transformations which is nothing but one-dimensional linear advections. We consider here to use a pseudo-spectral method. To do so, we discretize a square, of size $R$, centered in 0, (i.e. $\left[-\frac{R}{2}, \frac{R}{2}\right]^2$) with a regular grid with $N \in \mathbb{N}^*$ points per direction. Its stepsize is $h = R/N$. We denote this grid $\mathbb{G}^2$ with

$$\mathbb{G} = h \left[\!\left[ -\left\lfloor \frac{N-1}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor \right]\!\right]. \tag{6.9}$$

Then, we define the discrete partial Fourier transforms

$$\mathcal{F}_1 : \left\{ \begin{array}{lcl} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\widehat{\mathbb{G}}\times\mathbb{G}} \\ \boldsymbol{u} & \mapsto & h \sum_{g_1\in\mathbb{G}} \boldsymbol{u}_{g_1,g_2}\, e^{-ig_1\xi_1} \end{array} \right. \text{ and } \mathcal{F}_2 : \left\{ \begin{array}{lcl} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}\times\widehat{\mathbb{G}}} \\ \boldsymbol{u} & \mapsto & h \sum_{g_2\in\mathbb{G}} \boldsymbol{u}_{g_1,g_2}\, e^{-ig_2\xi_2} \end{array} \right. ,$$

where $\widehat{\mathbb{G}} = \eta \left[\!\left[ -\left\lfloor \frac{N-1}{2} \right\rfloor, \left\lfloor \frac{N}{2} \right\rfloor \right]\!\right]$ stands for the set of discrete frequencies with $\eta = 2\pi/R$.

Now, we want to solve the continuous shear transformations ($\alpha \in \mathbb{R}$) :

$$\begin{aligned} \partial_t u &= \alpha x_2 \partial_{x_1} u, \\ \partial_t u &= \alpha x_1 \partial_{x_2} u, \end{aligned} \tag{6.10}$$

which are the basic building blocks of the splitting presented above. These shear transformations are particularly simple to solve and we shall use pseudo-spectral method. Then, for any parameter $\alpha \in \mathbb{R}$, we introduce two pseudo-spectral shear transformations,

$$\mathcal{S}_1^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}^2} \\ \boldsymbol{u} & \mapsto & \mathcal{F}_1^{-1}\left[ e^{i\alpha\xi_1 g_2} \mathcal{F}_1 \boldsymbol{u} \right] \end{array} \right. \tag{6.11}$$

and

$$\mathcal{S}_2^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}^2} & \to & \mathbb{C}^{\mathbb{G}^2} \\ \boldsymbol{u} & \mapsto & \mathcal{F}_2^{-1}\left[ e^{i\alpha\xi_2 g_1} \mathcal{F}_2 \boldsymbol{u} \right] \end{array} \right. \tag{6.12}$$

**Remark 6.2.1.** *If $N$ is even, we have to pay attention to the mode $\frac{N}{2}$ associated to the frequency $\frac{\eta N}{2}$. Indeed, we can easily verify that $\mathcal{S}_i^\alpha \mathbb{R}^{\mathbb{G}^2} \subset \mathbb{R}^{\mathbb{G}^2}$ (for $i = 1, 2$) if and only if $N$ is odd or $\alpha \in \mathbb{Z}$.*

Finally, the numerical solution $(\boldsymbol{u}^n)_{n\in\mathbb{N}}$ of the numerical schemes we consider are defined by (for $\delta_t \neq k\pi, k \in \mathbb{Z}^\star$)

$$\begin{aligned} \boldsymbol{u}^n &= (\mathcal{L}_{\delta_t})^n \, u^{in}_{|\mathbb{G}^2} := (\mathcal{S}_2^{\delta_t} \mathcal{S}_1^{-\delta_t})^n \, u^{in}_{|\mathbb{G}^2}, && \text{(Lie)} \\ \boldsymbol{u}^n &= (\mathcal{T}_{\delta_t})^n \, u^{in}_{|\mathbb{G}^2} := (\mathcal{S}_1^{-\delta_t/2} \mathcal{S}_2^{\delta_t} \mathcal{S}_1^{-\delta_t/2})^n \, u^{in}_{|\mathbb{G}^2}, && \text{(Strang)} \\ \boldsymbol{u}^n &= (\mathcal{M}_{\delta_t})^n \, u^{in}_{|\mathbb{G}^2} := (\mathcal{S}_1^{-\tan(\delta_t/2)} \mathcal{S}_2^{\sin(\delta_t)} \mathcal{S}_1^{-\tan(\delta_t/2)})^n \, u^{in}_{|\mathbb{G}^2}, && \text{(New)} \end{aligned} \tag{6.13}$$

where $u^{in}_{|\mathbb{G}^2}$ is the evaluation of the initial condition $u^{in}$ on the grid $\mathbb{G}^2$.
The main goal of this section is now to perform a complete numerical analysis of these splittings defined in (6.13).

### 6.2.1 Numerical analysis

We define some associated discrete Lebesgue norms. They are defined for $\boldsymbol{u} \in \mathbb{C}^{\mathbb{G}^2}$ by

$$\| \boldsymbol{u} \|^2_{L^2(\mathbb{G}^2)} = h^2 \sum_{g\in\mathbb{G}^2} | \boldsymbol{u}_g |^2 \qquad \text{and} \qquad \| \boldsymbol{u} \|_{L^\infty(\mathbb{G}^2)} = \max_{g\in\mathbb{G}^2} | \boldsymbol{u}_g |.$$

We also introduce a scale of spaces, denoted $(X^s)_{s\geqslant 0}$, defined by

$$X^s = \left\{ u \in L^2(\mathbb{R}^2),\ \|u\|^2_{X^s} := \int |x|^{2s} |u(x)|^2 \mathrm{d}x + \int |\xi|^{2s} |\mathscr{F}u(\xi)|^2 \mathrm{d}\xi < \infty \right\}$$

where $\mathscr{F}u$ denotes the Fourier transform of $u$.

### Consistency

First, we prove that the pseudo-spectral shear transformations (6.11) and (6.12) are consistent with the continuous ones (6.10). Let us remark that in addition to the analysis of the spectral consistency, we will also pay attention to the truncation $R$.

**Proposition 6.2.1.** *For all $s > 1$ and for all $M > 0$, there exists $c > 0$ such that for all $u \in \mathscr{S}(\mathbb{R}^2)$, $\alpha \in (-M, M)$, $R > 0$ and $N \in \mathbb{N}^*$ we have*

$$\|\mathcal{S}_1^\alpha \boldsymbol{u} - \boldsymbol{v}\|_{L^2(\mathbb{G}^2)} \leqslant c\,|\alpha| \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}}.$$

*where $\boldsymbol{u} = u_{|\mathbb{G}^2}$ and $\boldsymbol{v} = v_{|\mathbb{G}^2}$ with $v(x) = u(x_1 + \alpha x_2, x_2)$.*

*Proof.* Applying the discrete Fourier-Plancherel isometry, we get

$$h^2 \sum_{g \in \mathbb{G}^2} |\,(\mathcal{S}_1^\alpha \boldsymbol{u})_g - \boldsymbol{v}_g\,|^2 = \frac{h\eta}{2\pi} \sum_{(\xi_1, g_2) \in \widehat{\mathbb{G}} \times \mathbb{G}} |\,(\mathcal{F}_1 \mathcal{S}_1^\alpha \boldsymbol{u})_{\xi_1, g_2} - (\mathcal{F}_1 \boldsymbol{v})_{\xi_1, g_2}\,|^2. \tag{6.14}$$

Thus, we are going to expand $\mathcal{F}_1 \boldsymbol{v}$ and $\mathcal{F}_1 \boldsymbol{u}$ with respect to $u$. More precisely, we apply the Poisson formula to get

$$\begin{aligned}
\mathcal{F}_1 \boldsymbol{u} &= h \sum_{g_1 \in h\mathbb{Z}} u(g_1, g_2) e^{-i\xi_1 g_1} - h \sum_{g_1 \in \mathbb{G}^c} u(g_1, g_2) e^{-i\xi_1 g_1} \\
&= \mathscr{F}_1 u(\xi_1, g_2) + \sum_{k \in \mathbb{Z}^*} \mathscr{F}_1 u(\xi_1 + \frac{2k\pi}{h}, g_2) - h \sum_{g_1 \in \mathbb{G}^c} u(g_1, g_2) e^{-i\xi_1 g_1},
\end{aligned}$$

where $\mathscr{F}_1 u(\xi_1, x_2) = \displaystyle\int u(x) e^{-i\xi_1 x_1} \mathrm{d}x_1$ is the continuous Fourier transform of $u$ along the first direction and $\mathbb{G}^c = h\mathbb{Z} \backslash \mathbb{G}$. Consequently, since $\mathscr{F}_1 v(\xi_1, x_2) = e^{i\alpha\xi_1 x_2} \mathscr{F}_1 u$, we decompose the consistency error into three terms

$$\begin{aligned}
(\mathcal{F}_1 \mathcal{S}_1^\alpha \boldsymbol{u})_{\xi_1, g_2} - (\mathcal{F}_1 \boldsymbol{v})_{\xi_1, g_2} = &\sum_{k \in \mathbb{Z}^*} \left(1 - e^{i\alpha \frac{2k\pi}{h} g_2}\right) e^{i\alpha\xi_1 g_2} \mathscr{F}_1 u(\xi_1 + \frac{2k\pi}{h}, g_2) && (\boldsymbol{\varepsilon}^1_{\xi_1, g_2}) \\
&+ h \sum_{g_1 \in \mathbb{G}^c} \left(1 - e^{i\alpha\xi_1 g_2}\right) u(g_1, g_2) e^{-i\xi_1 g_1} && (\boldsymbol{\varepsilon}^2_{\xi_1, g_2}) \\
&+ h \sum_{g_1 \in \mathbb{G}^c} \left[u(g_1 + \alpha g_2, g_2) - u(g_1, g_2)\right] e^{-i\xi_1 g_1}. && (\boldsymbol{\varepsilon}^3_{\xi_1, g_2})
\end{aligned}$$

Now we bound each one of these consistency errors.
<u>Estimation of $\boldsymbol{\varepsilon}^1$</u> :

First, we have

$$
\begin{aligned}
|\,\varepsilon^1_{\xi_1,g_2}\,| &\leqslant \sum_{k\in\mathbb{Z}^*}\left|\alpha\frac{2k\pi}{h}g_2\right|\left|\mathscr{F}_1 u(\xi_1+\frac{2k\pi}{h},g_2)\right| \\
&= \sum_{k\in\mathbb{Z}^*}\left|\alpha\frac{2k\pi}{h}g_2\right|\left|\xi_1+\frac{2k\pi}{h}\right|^{-s-1}\left|\mathscr{F}_1(|\partial_{x_1}|^{s+1}u)(\xi_1+\frac{2k\pi}{h},g_2)\right| \\
&\leqslant \sum_{k\in\mathbb{Z}^*}\left|\alpha\frac{2k\pi}{h}g_2\right|\left(\frac{(2|k|-1)\pi}{h}\right)^{-s-1}\left|\mathscr{F}_1(|\partial_{x_1}|^{s+1}u)(\xi_1+\frac{2k\pi}{h},g_2)\right| \\
&\leqslant \frac{4|\alpha|\zeta(s+1)}{\sqrt{1+g_2^2}}\left(\frac{\pi}{h}\right)^s\sup_{x_2\in\mathbb{R}}\int_{\mathbb{R}}(1+x_2^2)|(|\partial_{x_1}|^s u)(x_1,x_2)|\mathrm{d}x_1,
\end{aligned}
$$

where $\zeta$ denotes the Riemann function. This estimate involves a norm of $u$ that is neither usual nor isotropic. Furthermore, the estimates of $\varepsilon^2$ and $\varepsilon^3$ will lead to some other norms of this kind. Consequently, in order to get an estimate as readable as possible, we control these norms by the $X^{s+6}$ norm. Such a control can be realized with classical techniques of pseudo-differential calculus. As these estimates are technical but not crucial here, we omit details (the interested reader could refer for example to [94]).

Now, we observe that by monotonicity we have

$$
h\sum_{g_2\in\mathbb{G}\setminus\{0\}}\frac{1}{1+g_2^2}\leqslant\int_{\mathbb{R}}\frac{1}{1+y^2}\mathrm{d}y\leqslant\pi. \tag{6.15}
$$

Thus, since $\varepsilon^1_{\xi_1,0}=0$, there exists constant $c>0$, depending only on $s$, such that

$$
\frac{h\eta}{2\pi}\sum_{(\xi_1,g_2)\in\widehat{\mathbb{G}}\times\mathbb{G}}|\,\varepsilon^1_{\xi_1,g_2}\,|^2\leqslant c|\alpha|^2R^{-1}h^{2s}(\#\widehat{\mathbb{G}})\|u\|_{X^{s+6}}\leqslant c|\alpha|^2h^{2s-1}\|u\|^2_{X^{s+6}}.
$$

<u>Estimation of $\varepsilon^2$</u> :
First, naturally, we control $\varepsilon^2$ by

$$
|\,\varepsilon^2_{\xi_1,g_2}\,|\leqslant\alpha|\xi_1||g_2|\left|h\sum_{g_1\in\mathbb{G}^c}u(g_1,g_2)e^{-i\xi_1g_1}\right|. \tag{6.16}
$$

In order, to absorb the factor $\xi_1$ on the left, we realize a discrete integration by part. So, we assume that $\xi_1\neq 0$, we denote $\xi_1=k_1\eta$ and $g_1=g_1=n_1h$, where $k_1\in[\![-\lfloor\frac{N-1}{2}\rfloor,\lfloor\frac{N}{2}\rfloor]\!]$ and $n_1\in\mathbb{Z}\setminus[\![-\lfloor\frac{N-1}{2}\rfloor,\lfloor\frac{N}{2}\rfloor]\!]$.

Then we introduce $N^+=1+\lfloor N/2\rfloor$ and $N^-=-1-\lfloor(N-1)/2\rfloor$. Consequently, we have

$$
\begin{aligned}
h\sum_{g_1\in\mathbb{G}^c}u(g_1,g_2)e^{-i\xi_1g_1} =& h\sum_{n_1\geqslant N^+}u(g_1,g_2)e^{-\frac{2i\pi n_1k_1}{N}}+h\sum_{n_1\leqslant N^-}u(g_1,g_2)e^{-\frac{2i\pi n_1k_1}{N}} \\
=& h\sum_{n_1\geqslant N^+}\frac{u(g_1,g_2)-u(g_1+h,g_2)}{h}h\sum_{n=N^+}^{n_1}e^{-\frac{2i\pi nk_1}{N}} \qquad (E_+) \\
&+h\sum_{n_1\leqslant N^-}\frac{u(g_1,g_2)-u(g_1-h,g_2)}{h}h\sum_{n=n_1}^{N^-}e^{-\frac{2i\pi nk_1}{N}}. \qquad (E_-)
\end{aligned}
$$

To control $(E_+)$, first we observe that since $0 \leqslant |k_1| \leqslant N/2$, we have

$$\left| h \sum_{n=N^+}^{n_1} e^{-\frac{2i\pi n k_1}{N}} \right| \leqslant \frac{2h}{|1 - e^{-\frac{2i\pi k_1}{N}}|} \leqslant c \frac{hN}{2\pi |k_1|} = \frac{c}{|\xi_1|},$$

where $c$ is an universal constant.

Then, applying the mean value theorem, we get

$$\begin{aligned} |E_+| &\leqslant \frac{c}{|\xi_1|(1+g_2^2)} \left( \sup_{x \in \mathbb{R}^2} (1+x_2^2)|x_1|^{s+1} |\partial_{x_1} u(x)| \right) h \sum_{n_1 \geqslant N^+} g_1^{s+1} \\ &\leqslant \frac{c_s}{|\xi_1|(1+g_2^2)} \|u\|_{X^{s+6}} R^{-s} \frac{1}{N} \sum_{n_1 \geqslant 0} \left( \frac{N^+ + n_1}{N} \right)^{s+1} \\ &\leqslant \frac{c_s}{|\xi_1|(1+g_2^2)} \|u\|_{X^{s+6}} R^{-s} \frac{1}{N} \sum_{n_1 \geqslant 0} \left( \frac{1}{2} + \frac{n_1}{N} \right)^{s+1} \end{aligned} \tag{6.17}$$

where $c_s > 0$ is a constant depending only on $s$. We recognize a Riemann sum, so we have

$$\frac{1}{N} \sum_{n_1 \geqslant 0} \left( \frac{1}{2} + \frac{n_1}{N} \right)^{s+1} \xrightarrow[N \to \infty]{} 2^{-s-2} \int_1^\infty y^{-s-1} \mathrm{d}y = \frac{2^{-s-2}}{s}.$$

In particular, since this sequence converges, it is bounded by a constant depending only on $s$. Thus, we obtain the following bound for $|E_+|$

$$|E_+| \leqslant \frac{c_s}{|\xi_1|(1+g_2^2)} \|u\|_{X^{s+6}} R^{-s},$$

where $c_s$ is another constant depending only on $s$. Note that, by symmetry, the same control holds for $E^-$.

Finally, coming back to (6.16) and using (6.15), we have another constant, denoted $c_s$, depending only on $s$ such that

$$\frac{h\eta}{2\pi} \sum_{(\xi_1, g_2) \in \widehat{\mathbb{G}} \times \mathbb{G}} |\, \varepsilon^2_{\xi_1, g_2} \,|^2 \leqslant c_s |\alpha|^2 R^{-1} R^{-2s} (\#\widehat{\mathbb{G}}) \|u\|^2_{X^{s+6}} \leqslant c_s |\alpha|^2 h^{-1} R^{-2s} \|u\|^2_{X^{s+6}}. \tag{6.18}$$

<u>Estimation of $\varepsilon^3$</u> :
Let us introduce a small technical lemma whose proof is postponed to the Appendix.

**Lemma 6.2.2.** *If $y_1, y_2, y_3, \lambda \in \mathbb{R}$ are such that $y_3 \in [y_1; y_1 + \lambda y_2]$ then we have*

$$\left| \begin{pmatrix} y_3 \\ y_2 \end{pmatrix} \right| \geqslant \frac{|y_1|}{\sqrt{1+\lambda^2}}.$$

Then applying the mean value theorem, for any $g_1, g_2, \alpha \in \mathbb{R}$, there exists $m_{g_1, g_2, \alpha}$ in $[g_1; g_1 + \alpha g_2]$ such that

$$u(g_1 + \alpha g_2, g_2) - u(g_1, g_2) = \alpha g_2 \partial_{x_1} u(m_{g_1, g_2, \alpha}, g_2).$$

Since $|\alpha| \leqslant M$, applying Lemma 6.2.2, we get

$$\begin{aligned}
|\,\boldsymbol{\varepsilon}^3_{\xi_1,g_2}\,| &\leqslant h \sum_{g_1\in\mathbb{G}^c} |\alpha||g_2||\partial_{x_1} u(m_{g_1,g_2,\alpha}, g_2)| \\
&\leqslant h \sum_{g_1\in\mathbb{G}^c} \frac{|\alpha|}{\sqrt{1+g_2^2}} \left|\begin{pmatrix} m_{g_1,g_2,\alpha} \\ g_2 \end{pmatrix}\right|^{-s-1} \sup_{x\in\mathbb{R}^2} (1+x_2^2)|x|^{s+1}|\partial_{x_1} u(x)| \\
&\leqslant c_s \frac{|\alpha|}{\sqrt{1+g_2^2}} \left|\frac{R}{2\sqrt{1+M^2}}\right|^{-s-1} \|u\|_{X^{s+6}} \left( h \sum_{g_1\in\mathbb{G}^c} \left|\frac{2g_1}{R}\right|^{-s-1} \right),
\end{aligned}$$

where $c_s$ is a constant depending only on $s$.

Then, realizing the same estimate as in (6.17), we get another constant $c_s$ depending only on $s$ such that

$$\left( h \sum_{g_1\in\mathbb{G}^c} \left|\frac{2g_1}{R}\right|^{-s-1} \right) \leqslant c_s R.$$

Thus we have the estimate

$$|\,\boldsymbol{\varepsilon}^3_{\xi_1,g_2}\,| \leqslant c_{s,M} \frac{|\alpha|}{\sqrt{1+g_2^2}} R^{-s} \|u\|_{X^{s+6}}$$

where $c_{s,M}$ is a constant depending only on $s$ and $M$. Consequently, we can realize the same estimate as (6.18) to get

$$\frac{h\eta}{2\pi} \sum_{(\xi_1,g_2)\in\widehat{\mathbb{G}}\times\mathbb{G}} |\,\boldsymbol{\varepsilon}^3_{\xi_1,g_2}\,|^2 \leqslant c_{s,M} |\alpha|^2 h^{-1} R^{-2s} \|u\|^2_{X^{s+6}}.$$

where $c_{s,M}$ is another constant depending only on $s$ and $M$.

□

### Backward error analysis

We aim at describing the long time behavior of the splitting methods. So, we perform a general backward error analysis [1] for a large class of methods including Lie and Strang splittings but also the new splitting. Note that since we deal with a linear problem the expansions are convergent.

**Proposition 6.2.2.** *If $a, b \in \mathbb{R}$ satisfy $ab < 2$ then*

$$e^{bx_1\partial_{x_2}} e^{-ax_2\partial_{x_1}} = e^{JL_{a,b}x\cdot\nabla}, \tag{6.19}$$

*and*

$$e^{-\frac{a}{2}x_2\partial_{x_1}} e^{bx_1\partial_{x_2}} e^{-\frac{a}{2}x_2\partial_{x_1}} = e^{JS_{a,b}x\cdot\nabla}, \tag{6.20}$$

*where*

$$L_{a,b} = \mu_{a,b} \begin{pmatrix} b & \frac{ab}{2} \\ \frac{ab}{2} & a \end{pmatrix} \qquad \textit{and} \qquad S_{a,b} = \mu_{a,b} \begin{pmatrix} b & 0 \\ 0 & a(1-\frac{ab}{4}) \end{pmatrix} \tag{6.21}$$

1. The reader can refer to [78] for an overview on backward error analysis.

*with* $\mu_{a,b} = F(ab(1-ab/4))$*, where* $F$ *is the continuous function on* $(-\infty, 1]$ *defined by*

$$F(x) = \begin{cases} \frac{\arcsin(\sqrt{x})}{\sqrt{x}} & \textit{if} \quad 0 < x \leqslant 1 \\ \frac{\operatorname{asinh}(\sqrt{-x})}{\sqrt{-x}} & \textit{if} \quad x < 0 \\ 1 & \textit{if} \quad x = 0. \end{cases}$$

*Proof.* Considering the transport equation (6.5) which can be solved with the method of the characteristics, we have

$$e^{tAx\cdot\nabla} u_0 = u^{in} \circ e^{tA}.$$

Thus (6.19) is equivalent to

$$\exp\begin{pmatrix} 0 & -a \\ 0 & 0 \end{pmatrix} \exp\begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix} = e^{JL_{a,b}},$$

with $J$ given by (6.2). These exponentials of matrices can be written as shear transforms. So (6.19) is equivalent to

$$P_{a,b} := \begin{pmatrix} 1 & -a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} = \begin{pmatrix} 1-ab & -a \\ b & 1 \end{pmatrix} = e^{JL_{a,b}}. \tag{6.22}$$

Similarly, (6.20) is equivalent to

$$\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = e^{JS_{a,b}}. \tag{6.23}$$

First, we prove that if (6.22) holds with $L_{a,b}$ given by (6.21) then (6.22) also holds with $S_{a,b}$ given by (6.21). Indeed, observing that a Lie splitting is always conjugated to a Strang splitting we have

$$\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & a/2 \\ 0 & 1 \end{pmatrix} P_{a,b} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & a/2 \\ 0 & 1 \end{pmatrix} e^{JL_{a,b}} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix}.$$

But $\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix}$ is symplectic, i.e.

$${}^{\mathrm{t}}\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} J \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = J.$$

Thus, we have

$$\begin{pmatrix} 1 & a/2 \\ 0 & 1 \end{pmatrix} e^{JL_{a,b}} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} = \exp\left( J\, {}^{\mathrm{t}}\begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} L_{a,b} \begin{pmatrix} 1 & -a/2 \\ 0 & 1 \end{pmatrix} \right) = e^{JS_{a,b}},$$

where $S_{a,b}$ is given by (6.21).

So, now we aim at proving (6.22). The existence of such a $L_{a,b}$ is ensured by the following lemma (an elementary proof is given in Appendix).

**Lemma 6.2.3.** *If* $a, b$ *are small enough, there exists a symmetric matrix* $L_{a,b}$ *such that* $L_{a,b}$ *goes to* $0$ *as* $(a, b)$ *goes to* $0$ *and* $P_{a,b} = \exp(JL_{a,b})$*.*

Then, we have to determine a formula for $L_{a,b}$. Since $L_{a,b}$ is a symmetric matrix, $e^{JL_{a,b}}$ is an Hamiltonian flow at time $1$. A fortiori, $L_{a,b}$ is a constant of the motion. So we have

$$^{\mathrm{t}}(e^{JL_{a,b}})L_{a,b}e^{JL_{a,b}} = L_{a,b}.$$

But, by construction, we have $e^{JL_{a,b}} = P_{a,b}$, so $L_{a,b}$ is an eigenvector associated with the eigenvalue $1$ of the following linear application

$$R_{a,b} : \left\{ \begin{array}{lcl} S_2(\mathbb{R}) & \to & S_2(\mathbb{R}) \\ Q & \mapsto & {}^{\mathrm{t}}P_{a,b}QP_{a,b}. \end{array} \right.$$

By a straightforward calculation, we observe that

$$Q_{a,b} = \begin{pmatrix} b & \frac{ab}{2} \\ \frac{ab}{2} & a \end{pmatrix} \text{ satisfies } R_{a,b}(Q_{a,b}) = Q_{a,b}. \tag{6.24}$$

Then we deduce of the following lemma (proven in Appendix) that if $0 < ab < 4$ then there exists $\mu_{a,b} \in \mathbb{R}$ such that

$$L_{a,b} = \mu_{a,b}Q_{a,b}. \tag{6.25}$$

**Lemma 6.2.4.** *If $0 < ab < 4$ the eigenspace of $R_{a,b}$ associated with the eigenvalue $1$ is of dimension $1$.*

Now, we just have to determined $\mu_{a,b}$. Since $L_{a,b}$ is symmetric, it is diagonalizable in an orthonormal basis, i.e.

$$\exists \lambda \in \mathbb{R}^2, \exists \Omega \in O_2(\mathbb{R}),\ L_{a,b} = \Omega^{-1} \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \Omega = \Omega^{-1} D \Omega.$$

So, since $J$ and $\Omega$ commute, we have

$$P_{a,b} = \Omega^{-1} e^{JD} \Omega.$$

Since we assume that $0 < ab < 4$, we deduce from (6.25) that $L_{a,b}$ is either positive or negative. In particular we have $\lambda_1\lambda_2 > 0$. Thus we can define the symplectic matrix

$$K = \begin{pmatrix} \sqrt[4]{\lambda_1/\lambda_2} & \\ & \sqrt[4]{\lambda_2/\lambda_1} \end{pmatrix}.$$

This matrix satisfies $\sqrt{\lambda_1\lambda_2}\,{}^{\mathrm{t}}KK = D$ and $J\,{}^{\mathrm{t}}K = K^{-1}J$ . Thus, we have

$$P_{a,b} = (K\Omega)^{-1} e^{\sqrt{\lambda_1\lambda_2}J} (K\Omega) = (K\Omega)^{-1} \begin{pmatrix} \cos(\sqrt{\lambda_1\lambda_2}) & -\sin(\sqrt{\lambda_1\lambda_2}) \\ \sin(\sqrt{\lambda_1\lambda_2}) & \cos(\sqrt{\lambda_1\lambda_2}) \end{pmatrix} (K\Omega). \tag{6.26}$$

In particular, we have

$$\operatorname{Tr} P_{a,b} = 2\cos(\sqrt{\lambda_1\lambda_2}) = 2\cos(\sqrt{\det L_{a,b}}) = 2\cos(\mu_{a,b}\sqrt{\det Q_{a,b}}).$$

As a consequence, since $\sqrt{\det L_{a,b}}$ goes to zero when $(a,b)$ goes to $0$, we deduce of a straightforward calculation that if $ab$ is small enough then

$$\mu_{a,b} = \pm F(ab(1 - ab/4)).$$

Finally, we have to determine the sign of $\mu_{a,b}$. First, observe that by continuity, we have either $\mu_{a,b} > 0$ for all $a, b$ small enough satisfying $ab > 0$ or $\mu_{a,b} < 0$ for all $a, b$ small enough satisfying $ab > 0$. This second case is excluded because when $a$ goes to zero we have

$$e^{-F(a^2(1-a^2/4))JQ_{a,a}} = e^{-aJ+\mathcal{O}(a^2)} = P_{-a,-a} + \mathcal{O}(a^2) \neq P_{a,a} + \mathcal{O}(a^2).$$

To conclude, we have proved that if $ab > 0$ and $(a, b)$ is small enough then

$$P_{a,b} = e^{F(ab(1-ab/4))JQ_{a,b}}.$$

Furthermore, this relation is analytic with respect to $a$ and $b$, so it can be extended to all $a, b \in \mathbb{R}$ such that $ab < 2$. Indeed, under this assumption we have $ab(1 - ab/4) \in (-\infty, 1)$ which is the domain of analyticity of $F$.

□

The classical splitting formulas of Lie and Strang correspond to the choice $a = b = \delta_t$ in (6.19) and (6.20). However, these choices are not necessarily the best. For the Strang like splittings, a straigthforward calculation proves that there exists an optimal choice for which the splitting is exact. This choice can be obtained by direct formal calculations by assuming a decomposition of the rotation matrix. Note that this splitting has been introduced in the imaging community (see [97, 112, 5]) and also in the PDE context (see [49, 9]).

**Lemma 6.2.5.** *If $\delta_t \in (-\pi, \pi)$ then we have*

$$S_{2\tan(\delta_t/2),\sin(\delta_t)} = \delta_t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that due to the non-diagonal terms of $L_{a,b}$, it is impossible to design an exact splitting based on the Lie splitting.

### Convergence

We now consider the convergence of the pseudo-spectral splittings (6.13) to approximate our problem (6.3). Then, for a discrete initial condition $\boldsymbol{u} = u^{in}_{|\mathbb{G}^2}$, the numerical solution at time $t_n = n\delta_t$ $(n \in \mathbb{N})$ is given by $n$ compositions of the operators defined in (6.13). For instance for the standard Strang splitting, the numerical solution at time $t_n$ is $(\mathcal{T}_{\delta_t})^n \boldsymbol{u}$. In the following theorem, we give, up to a spectral spatial error, the dynamic generated by the Strang pseudo-spectral method $\mathcal{T}_{\delta_t}$ and by the Lie pseudo-spectral method $\mathcal{L}_{\delta_t}$ over very long times.

**Theorem 6.2.6.** *For all $s > 0$ there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$ and all $\delta_t \in [-1, 1]$ satisfying, denoting $t_n = n\delta_t$, we have*

$$\|(\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}},$$

*and*

$$\|(\mathcal{T}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n J S^{\mathcal{T}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}},$$

*where $\boldsymbol{u} = u_{|\mathbb{G}^2}$, $S^{\mathcal{L}}_{\delta_t} := L_{\delta_t,\delta_t}/\delta_t = I_2 + \mathcal{O}(\delta_t)$ and $S^{\mathcal{T}}_{\delta_t} := S_{\delta_t,\delta_t}/\delta_t = I_2 + \mathcal{O}(\delta_t^2)$. The definitions of $S_{a,b}$ and $L_{a,b}$ are given by* (6.21) *in Proposition 6.2.2, whereas $\mathcal{L}_{\delta_t}$ and $\mathcal{T}_{\delta_t}$ are given by* (6.13).

*Proof.* We focus only on proving the convergence estimate for the Lie splitting. The same proof could be applied to prove the estimate for the Strang splitting.

Let $\varepsilon_n \in L^2(\mathbb{G}^2)$ be the consistency error at time $t_n$. It is defined by

$$\boldsymbol{\varepsilon}_n = \mathcal{L}_{\delta_t}\left(e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2} - \left(e^{t_{n+1} JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2}.$$

As usual, for linear schemes, the convergence error is given by

$$(\mathcal{L}_{\delta_t})^n\, \boldsymbol{u} - \left(e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2} = \sum_{k=0}^{n-1} \mathcal{L}_{\delta_t}^{n-1-k}\, \boldsymbol{\varepsilon}_k\,.$$

Here, $\mathcal{L}_{\delta_t}$ is an isometry of $L^2(\mathbb{G}^2)$, so we have

$$\|(\mathcal{L}_{\delta_t})^n\, \boldsymbol{u} - \left(e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant \sum_{k=0}^{n-1} \|\,\boldsymbol{\varepsilon}_k\,\|_{L^2(\mathbb{G}^2)} \leqslant n \sup_{k\in\mathbb{N}} \|\,\boldsymbol{\varepsilon}_k\,\|_{L^2(\mathbb{G}^2)}.$$

Thus, we just have to bound $\varepsilon_k$. Using formulas of Proposition 6.2.2, we decompose $\varepsilon_k$ into two consistency errors for the pseudo-spectral shear transformations

$$\begin{aligned}\boldsymbol{\varepsilon}_k = \mathcal{S}_2^{\delta_t}\left(e^{-\delta_t x_2\partial_{x_1}} e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2} - \left(e^{\delta_t x_1\partial_{x_2}} e^{-\delta_t x_2\partial_{x_1}} e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2} \\ + \mathcal{S}_2^{\delta_t}\left[\mathcal{S}_1^{-\delta_t}\left(e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2} - \left(e^{-\delta_t x_2\partial_{x_1}} e^{t_n JS^{\mathcal{L}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2}\right].\end{aligned}$$

Then applying Proposition 6.2.1, we get a constant $c > 0$, depending only on $s > 0$ such that

$$\|\,\boldsymbol{\varepsilon}_k\,\|_{L^2(\mathbb{G}^2)} \leqslant c\,|\delta_t| \frac{R^{-s} + h^s}{\sqrt{h}^s}\left(\|e^{t_n JS^{\mathcal{L}}_{\delta_t} x.\nabla} u\|_{X^{s+6}} + \|e^{-\delta_t x_2\partial_{x_1}} e^{t_n JS^{\mathcal{L}}_{\delta_t} x.\nabla} u\|_{X^{s+6}}\right).$$

Now, we introduce a lemma to control these norms, whose proof is available in Appendix.

**Lemma 6.2.7.** *For all $\kappa > 0$ and all $s > 0$ there exists a constant $c > 0$ such that if $\tau \in GL_2(\mathbb{R})$ satisfies*

$$\forall x \in \mathbb{R}^2,\ \kappa^{-1}|x| \leqslant |\tau(x)| \leqslant \kappa|x| \tag{6.27}$$

*then for all $u \in \mathscr{S}(\mathbb{R}^2)$ we have*

$$\|u \circ \tau\|_{X^s} \leqslant c\,\|u\|_{X^s}.$$

We recall that if $A \in M_2(\mathbb{R})$ then $e^{(Ax\cdot\nabla)}u = u^{in} \circ e^A$. Thus we just have to get estimates of the form (6.27) for $\tau = \exp(tJS^{\mathcal{L}}_{\delta_t})$ and $\tau = \begin{pmatrix} 1 & -\delta_t \\ 0 & 1 \end{pmatrix}$, uniformly with respect to $t \in \mathbb{R}$ and $\delta_t$ satisfying $|\delta_t| \leqslant 1$.

Since $\begin{pmatrix} 1 & -\delta_t \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & \delta_t \\ 0 & 1 \end{pmatrix}$ and $\delta_t \in [-1, 1]$ which is compact, by continuity, we get $\kappa > 0$ such that

$$\forall \delta_t \in [-1,1], \forall x \in \mathbb{R}^2,\ \kappa^{-1}|x| \leqslant \left|\begin{pmatrix} 1 & \delta_t \\ 0 & 1 \end{pmatrix} x\right| \leqslant \kappa|x|.$$

For the other estimate, we observe that the quadratic form associated $S^{\mathcal{L}}_{\delta_t}$ is a constant of the motion of $\exp(tJS^{\mathcal{L}}_{\delta_t})$ : for all $t \in \mathbb{R}$ and all $\delta_t \in [-1,1]$ we have

$$\forall x \in \mathbb{R}^2, \ {}^{\mathrm{t}}\left(e^{tJS^{\mathcal{L}}_{\delta_t}}x\right) S^{\mathcal{L}}_{\delta_t} e^{tJS^{\mathcal{L}}_{\delta_t}}x = {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x. \tag{6.28}$$

Furthermore, for all $\delta_t \in [-1,1]$, we have

$$\det S^{\mathcal{L}}_{\delta_t} = \delta_t^{-2} \arcsin^2\left(\sqrt{\delta_t^2(1-\delta_t^2/4)}\right) > 0.$$

So, $S^{\mathcal{L}}_{\delta_t}$ is either a positive or negative. Thus, since $(S^{\mathcal{L}}_{\delta_t})_{1,1} > 0$, it is positive. Then, since $\delta_t \mapsto S^{\mathcal{L}}_{\delta_t}$ is a continuous map, $S^{\mathcal{L}}_{\delta_t}$ and $(S^{\mathcal{L}}_{\delta_t})^{-1}$ are bounded uniformly with respect to $\delta_t \in [-1,1]$. Consequently, there exists $\kappa > 0$ such that for all $\delta_t \in [-1,1]$ and all $x \in \mathbb{R}^2$ we have

$$\kappa^{-1} \ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x \leqslant \kappa^{-1} |S^{\mathcal{L}}_{\delta_t}||x|^2 \leqslant |x|^2 \leqslant \kappa |(S^{\mathcal{L}}_{\delta_t})^{-1}|^{-1}|x|^2 \leqslant \kappa \ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x.$$

Thus we deduce of (6.28) that for all $t \in \mathbb{R}$, all $\delta_t \in [-1,1]$ and all $x \in \mathbb{R}^2$ we have

$$|e^{tJS^{\mathcal{L}}_{\delta_t}}x|^2 \leqslant \kappa \ {}^{\mathrm{t}}\left(e^{tJS^{\mathcal{L}}_{\delta_t}}x\right) S^{\mathcal{L}}_{\delta_t} e^{tJS^{\mathcal{L}}_{\delta_t}}x = \kappa \ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x \leqslant \kappa^2 |x|^2.$$

and

$$|e^{tJS^{\mathcal{L}}_{\delta_t}}x|^2 \geqslant \kappa^{-1} \ {}^{\mathrm{t}}\left(e^{tJS^{\mathcal{L}}_{\delta_t}}x\right) S^{\mathcal{L}}_{\delta_t} e^{tJS^{\mathcal{L}}_{\delta_t}}x = \kappa^{-1} \ {}^{\mathrm{t}}x S^{\mathcal{L}}_{\delta_t} x \geqslant \kappa^{-2} |x|^2.$$

□

As a corollary, we deduce the convergence error of these methods.

**Corollary 6.2.1.** *For all $s > 0$ and all $h_0 > 0$, there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$ and all $\delta_t \in [-1,1]$ and $h = R/N \leqslant h_0$, denoting $t_n = n\delta_t$, we have*

$$\|(\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s}+h^s}{\sqrt{h}} \|u\|_{X^{s+6}} + c\, |e^{t_n J} - e^{t_n JS^{\mathcal{L}}_{\delta_t}}| \|u\|_{X^4},$$

*and*

$$\|(\mathcal{T}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s}+h^s}{\sqrt{h}} \|u\|_{X^{s+6}} + c\, |e^{t_n J} - e^{t_n JS^{\mathcal{T}}_{\delta_t}}| \|u\|_{X^4},$$

*where $\boldsymbol{u} = u_{|\mathbb{G}^2}$, $\mathcal{L}_{\delta_t}$ and $\mathcal{T}_{\delta_t}$ are given by* (6.13)

*Proof of Corollary 6.2.1.* We only focus on proving the convergence estimate for the Lie slitting, the case of the Strang splitting being similar. Applying Theroem 6.2.6 and the triangle inequality, we have

$$\|(\mathcal{L}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s}+h^s}{\sqrt{h}} \|u\|_{X^{s+6}} + E_{u,\delta_t,n,\mathbb{G}},$$

with

$$E_{u,\delta_t,n,\mathbb{G}} = \| \left(e^{t_n JS^{\mathcal{L}}_{\delta_t}x\cdot\nabla}u - e^{t_n Jx\cdot\nabla}u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)}.$$

Consequently, to prove the corollary, we just have to bound $E_{u,\delta_t,n,\mathbb{G}}$ by $|e^{t_n J} - e^{t_n JS^{\mathcal{L}}_{\delta_t}}| \|u\|_{X^4}$. First, we introduce a technical lemma that will be proved in Appendix.

**Lemma 6.2.8.** *There exists an universal constant $c > 0$ such that for all $v \in H^2(\mathbb{R}^2)$, all $R > 0$ and all $N \in \mathbb{N}^*$ we have*

$$\|v_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant \|u\|_{L^2(\mathbb{R}^2)} + c\, h^2 \|\Delta u\|_{L^2(\mathbb{R}^2)}.$$

Since $h \leqslant h_0$, applying this lemma we get a constant $c > 0$, depending only on $h_0 > 0$, such that

$$E_{u,\delta_t,n,\mathbb{G}} \leqslant c\, \|(1-\Delta)\left(e^{t_n J S^{\mathcal{L}}_{\delta_t} x.\nabla} u - e^{t_n J x.\nabla} u\right)\|_{L^2(\mathbb{R}^2)}.$$

Then applying the Fourier Plancherel isometry, we get

$$E_{u,\delta_t,n,\mathbb{G}} \leqslant \frac{c}{2\pi} \|(1+|\xi|^2)\left(\mathscr{F}u \circ {}^{\mathrm{t}}\left(e^{-t_n J S^{\mathcal{L}}_{\delta_t}}\right) - \mathscr{F}u \circ {}^{\mathrm{t}}\left(e^{-t_n J}\right)\right)\|_{L^2(\mathbb{R}^2)}.$$

Then introducing a Taylor remainder under its integral form, it comes

$$\begin{aligned} E_{u,\delta_t,n,\mathbb{G}} &\leqslant \|(1+|\xi|^2)\int_0^1 \nabla_\xi \mathscr{F}u(y_{\alpha,\xi,n,\delta_t}) \cdot {}^{\mathrm{t}}\left(e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}\right)\xi \, \mathrm{d}\alpha\|_{L^2(\mathbb{R}^2)} \\ &\leqslant |e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}| \max_{\alpha\in(0,1)} \|(1+|\xi|^2)^{3/2}\nabla_\xi \mathscr{F}u(y_{\alpha,\xi,n,\delta_t})\|_{L^2(\mathbb{R}^2)} \end{aligned} \tag{6.29}$$

where $y_{\alpha,\xi,n,\delta_t} = {}^{\mathrm{t}}M_{\alpha,n,\delta_t}\xi$ and $M_{\alpha,n,\delta_t} = I_2 - \alpha\left(e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}\right)$.

Now, we distinguish two cases. If $|e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}| \leqslant 1/2$ then we deduce that we have $|M_{\alpha,n,\delta_t} - I_2| \leqslant \frac{1}{2}$. Consequently, we have

$$|\det M_{\alpha,n,\delta_t}| \geqslant \kappa \text{ and } |M^{-1}_{\alpha,n,\delta_t}| \leqslant 2,$$

where $\kappa$ is an universal constant.

Thus, realizing a natural change of coordinates, we get

$$\begin{aligned} E_{u,\delta_t,n,\mathbb{G}} &\leqslant \frac{|e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|}{\kappa} \|(1+|\,{}^{\mathrm{t}}M_{\alpha,n,\delta_t}{}^{-1}\xi|^2)^{3/2}\nabla_\xi \mathscr{F}u\|_{L^2(\mathbb{R}^2)} \\ &\leqslant 8\frac{|e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|}{\kappa} \|(1+|\xi|^2)^{3/2}\nabla_\xi \mathscr{F}u\|_{L^2(\mathbb{R}^2)} \leqslant c|e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|\|u\|_{X^4} \end{aligned}$$

where $c > 0$ is an universal constant.

Finally, we have to consider the case where $|e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}| \geqslant 1/2$. Applying Lemma 6.2.8 and the Fourier-Plancherel isometry we get two constant $c, \kappa > 0$ depending only on $h_0$ such that

$$\begin{aligned} E_{u,\delta_t,n,\mathbb{G}} &\leqslant \|\,\boldsymbol{u}\,\|_{L^2(\mathbb{G}^2)} + \|\left(e^{t_n J x\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\|(1-\Delta)u\|_{L^2(\mathbb{R})} \\ &\leqslant \kappa|e^{-t_n J S^{\mathcal{L}}_{\delta_t}} - e^{-t_n J}|\|u\|_{X^4}. \end{aligned}$$

□

Finally, we focus on the new splitting $\mathcal{M}_{\delta_t}$. We give a theorem proving that its dynamic corresponds, up to a spectral spatial error, to the rotation with the exact speed, for very long times.

**Theorem 6.2.9.** *For all $s, \nu > 0$ there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$ and all $\delta_t \in \mathbb{R}$ satisfying $|\delta_t| < \pi - \nu$, denoting $t_n = n\delta_t$, we have*

$$\|(\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}}, \tag{6.30}$$

*where $\boldsymbol{u} = u_{|\mathbb{G}^2}$ , and $\mathcal{M}_{\delta_t}$ is given by* (6.13)*.*

*Proof.* Realizing the same proof as in Theorem 6.2.6, we could easily prove that for all $s, \nu > 0$ there exists $c > 0$ such that for all $N \in \mathbb{N}^*$, all $R > 0$, all $u \in \mathscr{S}(\mathbb{R}^2)$, all $n \in \mathbb{N}$ and all $\delta_t \in \mathbb{R}$ satisfying $|\delta_t| < \pi - \nu$, denoting $t_n = n\delta_t$, we have

$$\|(\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n JS^{\mathcal{M}}_{\delta_t} x\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{\sqrt{h}} \|u\|_{X^{s+6}}$$

where $\boldsymbol{u} = u_{|\mathbb{G}^2}$ and $S^{\mathcal{M}}_{\delta_t} := S_{2\tan(\delta_t/2),\sin(\delta_t)}/\delta_t$ where $S_{a,b}$ is given by (6.21).

Thus, to conclude this proof, we just have to observe that by Lemma 6.2.5 we have $S_{2\tan(\delta_t/2),\sin(\delta_t)} = \delta_t I_2$. □

**Remark 6.2.10.** *For all $\boldsymbol{u} \in L^2(\mathbb{G}^2)$ we have $\|\boldsymbol{u}\|_{L^\infty(\mathbb{G}^2)} \leqslant h^{-1}\|\boldsymbol{u}\|_{L^2(\mathbb{G}^2)}$, thus* (6.30) *gives a control of convergence error with the discrete $L^\infty$ norm for very long times :*

$$\|(\mathcal{M}_{\delta_t})^n \boldsymbol{u} - \left(e^{t_n Jx\cdot\nabla} u\right)_{|\mathbb{G}^2}\|_{L^\infty(\mathbb{G}^2)} \leqslant c\, t_n \frac{R^{-s} + h^s}{h^{3/2}} \|u\|_{X^{s+6}}.$$

### 6.2.2 Numerical illustrations

In this subsection, we intend to illustrate the different results obtained previously, namely the spatial accuracy of pseudo-spectral method and the time accuracy of the time splitting.

**Spatial accuracy**

First, we present some numerical results to illustrate the estimates obtained in Proposition 6.2.1. To do so, we consider the following function

$$u(x) = \exp\left(-\frac{|x|^2}{2}\right), \quad x = (x_1, x_2) \in \mathbb{R}^2,$$

which is shifted by $\alpha = 0.01$. We denote $v_{|\mathbb{G}^2}$ where $v(x) = u(x_1 + \alpha x_2, x_2)$ the exact shifted solution, and we compute the (discrete) $L^2$ norm of the difference between $\mathcal{S}_1^\alpha u_{|\mathbb{G}^2}$ and $v_{|\mathbb{G}^2}$. The spatial grid $\mathbb{G}^2$ is defined by (6.9) where $h = R/N$, $R = 15$ and different values of $N$ are considered to check the spatial accuracy. The results are displayed in Figures 6.1 and 6.2. One can observe that for large $h$ (or small $N$), the term $R^{-s}$ is negligible and the term $h^s$ gives the exponential decreasing of the error which is the typical behavior of spectral methods. On the contrary, for very small values of $h$ (or large values of $N$), the term $R^{-s}/h^{-1/2}$ becomes prominent even if the error is quite small (around $10^{-11}$).

**Time accuracy**

In this part, we give some numerical illustrations of the efficiency of the new splitting. To do so, we consider the following equation

$$\partial_t u = Jx \cdot \nabla_x u, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{6.31}$$

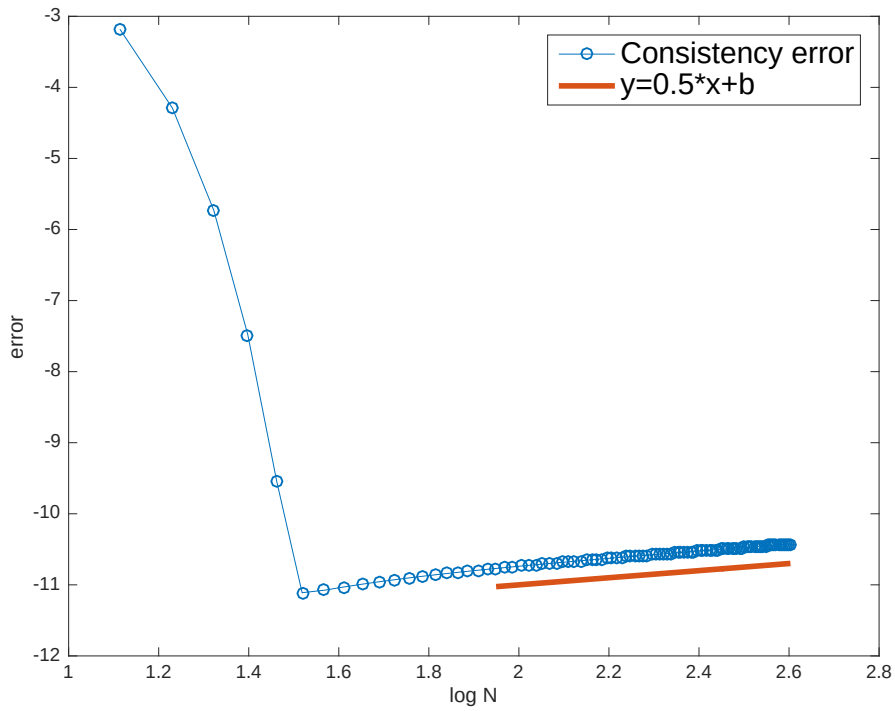


FIGURE 6.1 – Spatial error ($\log - \log$ scale) as a function of the number of points $N$ between the exact shifted solution and the approximation computed using fast Fourier transform.

with the initial condition

$$u^{in}(x) = \frac{1}{2\pi\beta}\left[\exp\left(-\frac{(x_1-0.3)^2}{\beta}\right) + \exp\left(-\frac{(x_1+0.3)^2}{\beta}\right)\right]\exp\left(-\frac{x_2^2}{\beta}\right),$$

with $\beta = 0.01$. The spatial truncated domain $[-2,2]^2$ is discretized with the grid $\mathbb{G}^2$ defined by (6.9) with $R = 4$ and a space step $h = R/N$, $N = 243 = 3^5$. The time step is $h \approx 0.139$ and the final time is $T = 10^5$ (the number of iterations is $71888$). In the next figures, some results are displayed where we compare the exact solution, the solution given by $(\mathcal{T}_{\delta_t})^n\ u^{in}_{|\mathbb{G}^2}$ (Strang splitting and spectral interpolation), the solution given $(\mathcal{L}_{\delta_t})^n\ u^{in}_{|\mathbb{G}^2}$ (Lie splitting and spectral interpolation) and the solution given by the new method $(\mathcal{M}_{\delta_t})^n\ u^{in}_{|\mathbb{G}^2}$ (see (6.13)). First, in Figures 6.3, the three solution are plotted at the final time. We can observe that the exact solution and the solution obtained by the new method are very close whereas the solution obtained by the Strang splitting is not good due to the fact that the angular velocity of the Strang method is not exact. To precise these observations, we plot on Figure 6.4 (Figure 6.5 is a zoom) the relative $L^\infty$ error of the new method, the Strang and the Lie methods. The error produced by the new method is close to $10^{-13}$ which is the spectral error. On the contrary, the Strang and Lie methods periodically produce an error of order one. This is due to its wrong angular velocity : the solution move away from the exact solution producing large error and at some times, the Strang method recover the exact solution so that the error becomes very small.

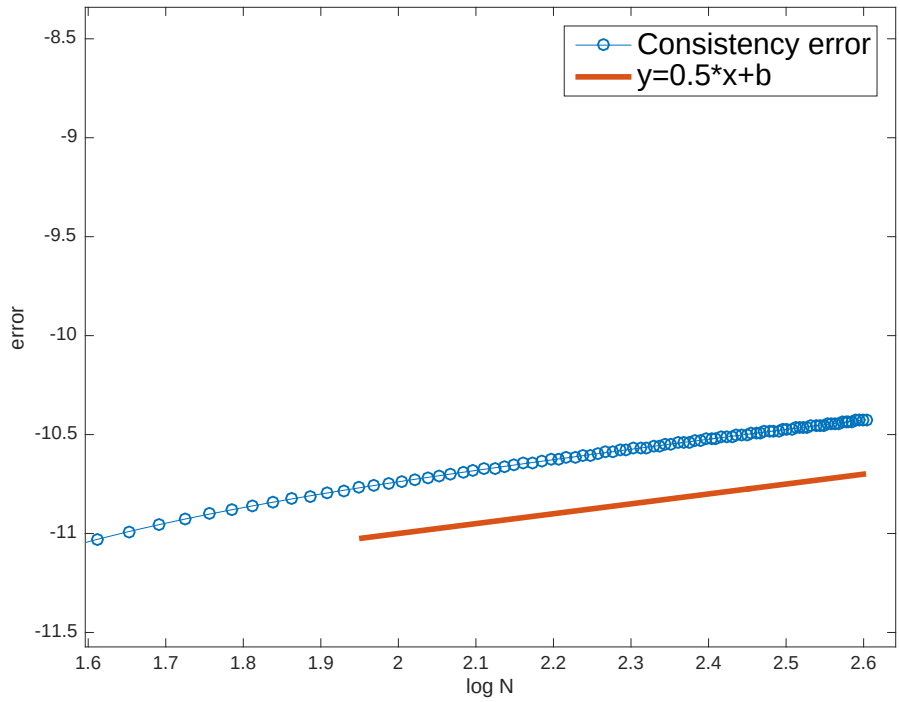


FIGURE 6.2 – Zoom of Figure 6.1.

These times can be computed from the above analysis. Indeed, from the rotation speed of the Strang splitting $\omega_{\delta_t} = \frac{\arcsin(\sqrt{\delta_t^2(1-\delta_t^2/4)})}{\delta_t}$, we deduce that the exact solution (which rotates with a speed $\omega_{ex} = 1$) and the numerical solution obtained by the Strang method will coincide every times $\bar{T}$ such that $t^n + \omega_{ex}\bar{T} = t^n + \omega_{\delta_t}\bar{T}\ [\pi]$ (the factor $\pi$ (instead of a factor $2\pi$) is due to our choice of a symmetric initial condition). Then, we have $\bar{T} = \pi/(\omega_{\delta_t} - 1)$ which gives with our choice of time step $\delta_t \approx 0.139$, $\bar{T} \approx 3888$. We can observe a very good agreement on Figures 6.4 and 6.5 and also on Figure 6.6 for which $\delta_t = \pi/4$ and then $\bar{T} \approx 113$.

Finally, we study the performance of the new method. Indeed, we compare the new splitting and a direct two-dimensional solving of (6.31). The direct resolution is done by a semi-Lagrangian type strategy : at each time step, we first compute exactly the feet of the characteristics equations and we then use a two-dimensional spectral interpolation by means of the non uniform fast Fourier transform (the so-called nufft procedure introduced in [74]). We checked that this approach also leads to spectral accuracy, and we want here to compare the two spectral methods in terms of CPU time with respect to the total number of points $N^2$ ($N$ being the number of points per direction). The results are displayed in Figure 6.7 : the time execution (for $10$ iterations) for both methods (new splitting and nufft) as a function of $N^2$ (for $N = 2^5, \ldots, 2^{11}$), in $\log - \log$ scale. Even if the method have the same complexity $\mathcal{O}(N^2 \log(N))$, the new approach clearly has a smaller constant (around ten times smaller). Moreover, in such a splitting procedure, a simple and efficient parallelization can be performed since the variable that does not appear in the derivative is just a parameter.

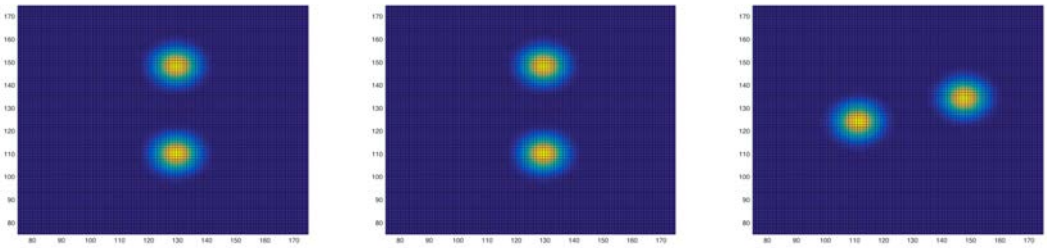

FIGURE 6.3 – Solution $u(T,x)$ of (6.31). Left : Exact solution $u(T,x)$. Middle : Numerical solution obtained by the new splitting. Right : Numerical solution obtained by the Strang splitting.

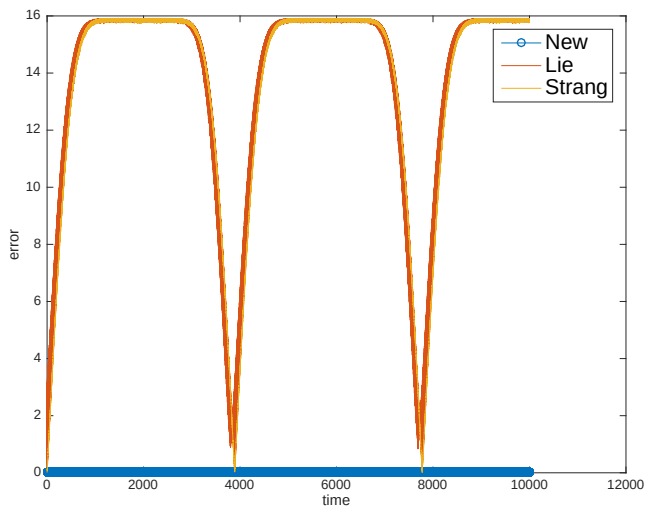


FIGURE 6.4 – Time history of the relative errors between the exact solution of (6.31) and the numerical solution obtained by the new splitting ('New'), the Lie splitting ('Lie') and the Strang splitting ('Strang').

## 6.3 Application to the Vlasov-Maxwell equations

In this section, we intend to apply the above splitting to the context of the 1+1/2 Vlasov-Maxwell system. Indeed, the time discretization of this system is based on a time splitting, and one of this step (the so-called magnetic part) corresponds to a rotation in the velocity direction due to the presence of the self-consistent electromagnetic field. Then, instead of using a Strang splitting like in [55], we shall use the exact splitting presented in the previous section, so that this magnetic part will be solved exactly in time and with a spectral accuracy in the velocity directions. This is very helpful to design high order methods for the full Vlasov-Maxwell system. After introducing the 1+1/2 Vlasov-Maxwell system we intend to solve, the splitting method introduced in [55] is recalled and then high order methods dedicated to systems split into three parts are introduced.

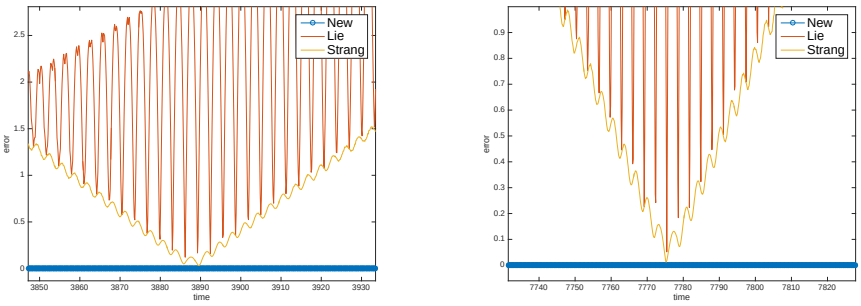


FIGURE 6.5 – Time history of the relative errors (zoom of Figure 6.4 around $\bar{T} \approx 3188$ (left) and $\bar{T} \approx 2 \times 3188$ (right).

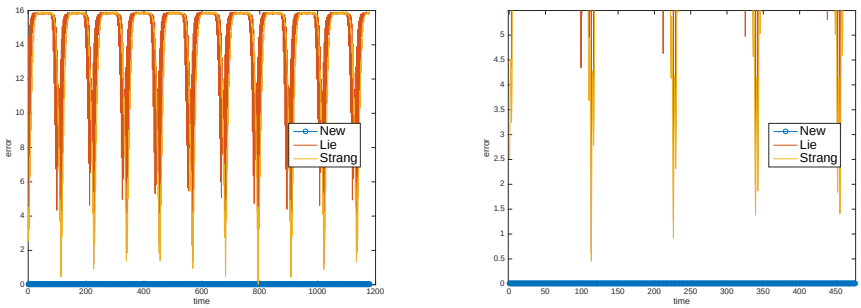


FIGURE 6.6 – Time history of the relative errors between the exact solution of (6.31) and the numerical solution obtained by the new splitting ('New'), the Lie splitting ('Lie') and the Strang splitting ('Strang'), with $\delta_t = \pi/4$. The right figure is a zoom of the left one around $k\bar{T}$ with $\bar{T} \approx 113, k = 1, 2, 3, 4$.

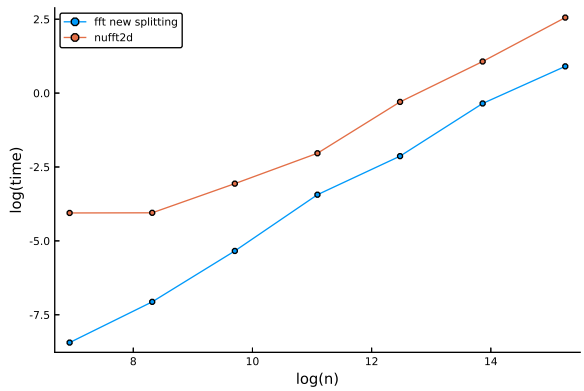


FIGURE 6.7 – Time execution as a function of the total number of points ($\log - \log$ scale). Blue : new method (new splitting and one-dimensional fast Fourier transform). Red : exact computation of the feet of the characteristics and two-dimensional non uniform fast Fourier transform.

### 6.3.1 Reduced 1+1/2 Vlasov-Maxwell equations

We consider the phase space $(x_1, v_1, v_2) \in L \times \mathbb{R}^2$, where $L = \mathbb{R}/2\pi\mathbb{Z}$ is a one-dimensional torus, and the unknown functions $f(t, x_1, v_1, v_2)$, $B(t, x_1)$ and $E(t, x_1) = (E_1, E_2)(t, x_1)$ which are determined by solving the following system of evolution equations

$$\begin{aligned}
&\partial_t f + v_1 \partial_{x_1} f + E \cdot \nabla_v f - BJv \cdot \nabla_v f = 0,\\
&\partial_t B = -\partial_{x_1} E_2,\\
&\partial_t E_2 = -\partial_{x_1} B - \int_{\mathbb{R}^2} v_2 f(t, x_1, v)\mathrm{d}v + \overline{\mathcal{J}}_2(t),\\
&\partial_t E_1 = -\int_{\mathbb{R}^2} v_1 f(t, x_1, v)\mathrm{d}v + \overline{\mathcal{J}}_1(t),
\end{aligned} \tag{6.32}$$

where $v = (v_1, v_2)$, $\overline{\mathcal{J}}_i(t) = 1/|L| \int_L \int_{\mathbb{R}^2} v_i f(t, x_1, v)\mathrm{d}x_1\mathrm{d}v, i = 1, 2$ ($|L|$ denotes the measure of $L$) and $J$ denotes the symplectic matrix

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

This reduced system, which has been considered in several former studies (see [42, 55, 48]), has to be supplemented with the Gauss condition

$$\partial_{x_1} E_1(t, x_1) = \int_{\mathbb{R}^2} f(t, x_1, v)\mathrm{d}v - 1, \forall t \geqslant 0, \tag{6.33}$$

and with initial conditions $f(t = 0, x_1, v) = f^{in}(x_1, v), E_2(t = 0, x_1) = E_2^{in}(x_1)$ and $B(t = 0, x_1) = B^{in}(x_1)$. Note that $E_1^{in}(x_1)$ is implied by the Gauss condition (6.33) at the initial time.

### 6.3.2 Splitting method

Here we propose to use the splitting method introduced in [55]. By reformulating the Vlasov-Maxwell system into

$$\frac{dF}{dt} = \mathcal{H}_E(F) + \mathcal{H}_f(F) + \mathcal{H}_B(F), \quad F(0) = F^{in},$$

where the fields $\mathcal{H}_E(F), \mathcal{H}_f(F)$ and $\mathcal{H}_B(F)$ will be written below. We denote by the solution of the Vlasov-Maxwell system (6.32) by $F(\delta_t) = (f, E_1, E_2, B)(\delta_t)$. This solution can be formally written as $F(\delta_t) = \varphi_{\delta_t}(F^{in}) := \exp((\mathcal{H}_E + \mathcal{H}_f + \mathcal{H}_B)\delta_t)F^{in}$, where $F^{in} = (f^{in}, E_1^{in}, E_2^{in}, B^{in})$ denotes the initial condition.

Now, we want to use a splitting method to approximate the system (6.32). To do so, we shall use the splitting introduced in [55, 56] based on a decomposition into three parts corresponding respectively to the fields $\mathcal{H}_E(F), \mathcal{H}_f(F)$ and $\mathcal{H}_B(F)$. Then, a first order Lie method based on this decomposition into three parts writes $\chi_{\delta_t}(F^{in}) = \varphi_{\delta_t}(F^{in}) + \mathcal{O}({\delta_t}^2)$ with

$$\chi_{\delta_t} = \varphi_{\delta_t}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t}^{[\mathcal{H}_B]} \tag{6.34}$$

where $\varphi_{\delta_t}^{[\mathcal{H}_E]}$, $\varphi_{\delta_t}^{[\mathcal{H}_f]}$, $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ denotes the exact solutions corresponding to the fields $\mathcal{H}_E$, $\mathcal{H}_f$ and $\mathcal{H}_B$. Using these notations, the adjoint of the Lie method $\chi_t^\star$ writes

$$\chi_{\delta_t}^\star = \varphi_{\delta_t}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t}^{[\mathcal{H}_E]}. \tag{6.35}$$

In the following we write down the equations associated to the fields $\mathcal{H}_E, \mathcal{H}_f$ and $\mathcal{H}_B$

$$\varphi_{\delta_t}^{[\mathcal{H}_E]} : \partial_t f + E \cdot \nabla_v f = 0, \partial_t E = 0, \partial_t B = -\partial_{x_1} E_2,$$

$$\varphi_{\delta_t}^{[\mathcal{H}_f]} : \partial_t f + v_1 \partial_{x_1} f = 0, \partial_t E = -\int_{\mathbb{R}^2} v f dv + \overline{\mathcal{J}}, \partial_t B = 0,$$

$$\varphi_{\delta_t}^{[\mathcal{H}_B]} : \partial_t f - BJv \cdot \nabla_v f = 0, \partial_t E_1 = 0, \partial_t E_2 = -\partial_{x_1} B, \partial_t B = 0.$$

Then, as mentioned in [55], $\varphi_{\delta_t}^{[\mathcal{H}_E]}$ and $\varphi_{\delta_t}^{[\mathcal{H}_f]}$ can be computed exactly in time and efficiently in phase space using spectral methods. However, the computation of $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ was performed using a Strang splitting. Instead, we shall use the new splitting $\mathcal{M}_{\delta_t}$ introduced above to compute exactly in time $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ and efficiently in phase space using spectral methods. Let us remark that the application of the new splitting to the $\mathcal{H}_B$ part requires a slight modification. Indeed, to solve $\partial_t f - BJv \cdot \nabla_v f = 0$ (with $B$ constant in time during this part) on one time step $\delta_t$ from an initial condition $f^{in}$ (defined on the velocity grid), we will use the new splitting with a modified time step $B\delta_t$ to capture the right rotation speed, i.e. $(\mathcal{M}_{B\delta_t})$ with $(\mathcal{M}_{\delta_t})$ defined by (6.13).

Based on the fact that each step can be computed exactly in time, we now look for efficient integration methods for systems separable into three parts which enable us to design efficient high order methods in time. A simple and efficient way to achieve this goal is to consider compositions of a first-order method with its adjoint computed at fractional step sizes. This is the main subject of the next part.

### 6.3.3 Composition methods for systems separable into three parts

To simplify the presentation, we restrict ourselves to the ODE context. The so-obtained composition methods will then be used within the Vlasov-Maxwell framework.

Let us consider the following ODE

$$\dot{x}(t) \equiv \frac{dx}{dt}(t) = u(x(t)), \qquad x(0) = x^{in} \in \mathbb{R}^D, \tag{6.36}$$

with $D \in \mathbb{N}^\star$, whose exact solution at time $t = \delta_t$ will be denoted by $x(\delta_t) = \varphi_{\delta_t}(x^{in})$. We are interested in problems where $u$ in (6.36) can be split into three parts,

$$u(x) = u_a(x) + u_b(x) + u_c(x)$$

in such a way that the exact flows $\varphi_{\delta_t}^{[a]}$, $\varphi_{\delta_t}^{[b]}$, $\varphi_{\delta_t}^{[c]}$, corresponding to $u_a$, $u_b$, $u_c$, respectively, can be computed exactly. One might consider then splitting methods of the form

$$\varphi_{a_s\delta_t}^{[a]} \circ \varphi_{b_s\delta_t}^{[b]} \circ \varphi_{c_s\delta_t}^{[c]} \circ \cdots \varphi_{a_1\delta_t}^{[a]} \circ \varphi_{b_1\delta_t}^{[b]} \circ \varphi_{c_1\delta_t}^{[c]} \tag{6.37}$$

and fix the coefficients $a_i, b_i, c_i$, $i = 1, \ldots, s$ so that it provides an approximation of order, say, $p$. It turns out, however, that the number of order conditions to be satisfied by these parameters grows very rapidly with the order. Thus, time-symmetric schemes of order $p = 4$ (resp. $p = 6$) require solving $11$ (resp. $56$) conditions. A more convenient way consists in considering compositions of $\chi_{\delta_t}$ and its adjoint $\chi_{\delta_t}^\star$, with

$$\chi_{\delta_t} = \varphi_{\delta_t}^{[a]} \circ \varphi_{\delta_t}^{[b]} \circ \varphi_{\delta_t}^{[c]} \qquad \text{and} \qquad \chi_{\delta_t}^\star = \varphi_{\delta_t}^{[c]} \circ \varphi_{\delta_t}^{[b]} \circ \varphi_{\delta_t}^{[a]}. \tag{6.38}$$

More specifically, we construct integrators within the family

$$\mathcal{G}_1 \equiv \left\{\psi_{\delta_t} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \cdots \circ \chi_{\alpha_{2s-1}\delta_t} \circ \chi^\star_{\alpha_{2s}\delta_t} :\ s \geqslant 1,\ (\alpha_j)_{1\leqslant j\leqslant 2s} \in \mathbb{R}^{2s}\right\}, \tag{6.39}$$

where $\chi_{\delta_t}$ and $\chi^\star_{\delta_t}$ are given by (6.38), so that

$$\chi_{\delta_t}(x^{in}) = \varphi_{\delta_t}(x^{in}) + \mathcal{O}(\delta_t^2), \tag{6.40}$$

and an analogous relation for $\chi^\star_{\delta_t}$. Composition integrators $\psi_{\delta_t} \in \mathcal{G}_1$ are time-symmetric whenever they have left-right palindromic sequences of coefficients $\alpha_i$, i.e. if $\alpha_{2s+1-i} = \alpha_i,\quad i = 1,\ldots,s$.

Notice that one could achieve methods of order $p$ within this family even if only first-order approximations to the flows $\varphi^{[a]}_{\delta_t}$, $\varphi^{[b]}_{\delta_t}$, and $\varphi^{[c]}_{\delta_t}$ are available, as long as one is able to construct the corresponding adjoint $\chi^\star_{\delta_t}$.

**Remark 6.3.1.** *Another well-known class $\mathcal{G}_2$ of integrators is formed by compositions*

$$\mathcal{G}_2 = \{\psi_{\delta_t} = \phi_{\alpha_1\delta_t} \circ \cdots \circ \phi_{\alpha_s\delta_t} :\ s \geqslant 1,\ (\alpha_1,\ldots,\alpha_s) \in \mathbb{R}^s\}, \tag{6.41}$$

*where $\phi_{\delta_t} : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is any second-order self-adjoint integrator. Notice that, if $\phi_{\delta_t}$ is chosen as $\phi_{\delta_t} = \chi_{\delta_t/2} \circ \chi^\star_{\delta_t/2}$, then $\mathcal{G}_2$ is contained in $\mathcal{G}_1$. These integrators also enjoy the time-symmetric property if $\alpha_{s+1-i} = \alpha_i,\ i = 1,\ldots,s$.*

### Analysis of the methods

For the analysis, it is convenient to introduce the graded Lie algebra associated with the vector field defining the ODE (6.36) and its corresponding exact flow $\varphi_{\delta_t}$. As is well known, for each infinitely differentiable map $g : \mathbb{R}^D \longrightarrow \mathbb{R}$, $g(\varphi_{\delta_t}(x))$ admits an expansion of the form

$$g(\varphi_{\delta_t}(x)) = e^{\delta_t F}[g](x) = g(x) + \sum_{k\geqslant 1} \frac{{\delta_t}^k}{k!} F^k[g](x), \qquad x \in \mathbb{R}^D,$$

where $F$ is the vector field associated with $u$,

$$F = \sum_{i=1}^{D} u_i(x)\,\frac{\partial}{\partial x_i}. \tag{6.42}$$

Similarly, for the basic first-order method $\chi_{\delta_t}$ defined by (6.40), one has $g(\chi_{\delta_t}(x)) = e^{Y_{\delta_t}}[g](x)$ with $Y_{\delta_t} = \sum_{k\geqslant 1} {\delta_t}^k Y_k$ and for its adjoint $\chi^\star_{\delta_t}$ also defined in (6.40), one has $g(\chi^\star_{\delta_t}(x)) = e^{-Y_{-\delta_t}}[g](x)$. Then, one can formally compute the operator series associated to any integrator $\psi_{\delta_t} \in \mathcal{G}_1$ defined by (6.39)

$$\Psi_{\delta_t} \quad = \quad \exp(Y_{\delta_t\alpha_1})\exp(-Y_{-\delta_t\alpha_2})\cdots\exp(Y_{\delta_t\alpha_{2s-1}})\exp(-Y_{-\delta_t\alpha_{2s}}).$$

By repeated application of the Baker–Campbell–Hausdorff formula we can express formally $\Psi_{\delta_t}$ as the exponential of an operator $F_{\delta_t}$,

$$\Psi_{\delta_t} = e^{F_{\delta_t}}, \qquad \text{with} \qquad F_{\delta_t} = \sum_{k\geqslant 1} {\delta_t}^k F_k,$$

${\delta_t}^k F_k \in \mathcal{L}_k$ for each $k \geqslant 1$ and $\mathcal{L} = \bigoplus_{k\geqslant 1} \mathcal{L}_k$ is the graded Lie algebra generated by the vector fields $\{\delta_t Y_1, \delta_t^2 Y_2, \delta_t^3 Y_3, \ldots\}$ where, by consistency, $Y_1 = F$.

Notice that

$$\begin{aligned} Y_{\delta_t \alpha_i} &= \delta_t \alpha_i Y_1 + (\delta_t \alpha_i)^2 Y_2 + (\delta_t \alpha_i)^3 Y_3 + \cdots \\ -Y_{-\delta_t \alpha_i} &= \delta_t \alpha_i Y_1 - (\delta_t \alpha_i)^2 Y_2 + (\delta_t \alpha_i)^3 Y_3 - \cdots \end{aligned}$$

so that

$$\begin{aligned} \Psi_{\delta_t} &= \exp\big(\delta_t w_1 Y_1 + {\delta_t}^2 w_2 Y_2 + {\delta_t}^3 (w_3 Y_3 + w_{12}[Y_1, Y_2]) \\ &\quad + {\delta_t}^4 (w_4 Y_4 + w_{13}[Y_1, Y_3] + w_{112}[Y_1, [Y_1, Y_2]]) + \mathcal{O}({\delta_t}^5)\big), \end{aligned}$$

where $w_1, w_2, \ldots$ are polynomials in the coefficients $\alpha_i$. In particular,

$$\begin{aligned} w_1 &= \sum_{i=1}^{2s} \alpha_i, \qquad w_2 = \sum_{i=1}^{2s} (-1)^i \alpha_i^2, \\ w_3 &= \sum_{i=1}^{2s} \alpha_i^3, \qquad w_4 = \sum_{i=1}^{2s} (-1)^i \alpha_i^4, \\ w_{12} &= \frac{1}{2} \left( \sum_{i=1}^{2s-1} (-1)^{i+1} \alpha_i^2 \sum_{j=i+1}^{2s} \alpha_j + \sum_{i=1}^{2s-1} \alpha_i \sum_{j=i+1}^{2s} (-1)^j \alpha_j^2 \right) \end{aligned}$$

Then, a time-symmetric 4th-order method has to satisfy only consistency ($w_1 = 1$) and the order conditions at order three, that is $w_3 = w_{12} = 0$. Let us remark that conditions at even order ($w_2 = w_4 = 0$) are automatically verified by symmetry. Notice, then, that the minimum number of maps to be considered in $\psi_{\delta_t} \in \mathcal{G}_1$ is $s = 3$.

### Methods of order $4$

It turns out, however, that methods involving the minimum number of maps (or *stages*) do not usually provide the best efficiency. In other words, considering additional stages (and thus some free parameters) leads to more efficient schemes, even when the computational cost per step is also higher. The difficulty then lies in the way the free parameters are fixed. In this respect, several objective functions have been considered in the literature. In particular we mention the following (let us recall that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s}$)

$$\mathcal{E}_1(\boldsymbol{\alpha}) = \sum_{i=1}^{2s} |\alpha_i| \qquad \text{and} \qquad \mathcal{E}_2(\boldsymbol{\alpha}) = 2s \left| \sum_{i=1}^{2s} \alpha_i^5 \right|^{1/4}. \tag{6.43}$$

The quantity $\mathcal{E}_2$ is usually the dominant error term for a number of problems. The criterion we follow here will be to look for symmetric methods with small values of $\mathcal{E}_1$ which, in addition, have also small values of $\mathcal{E}_2$. In the sequel, we consider composition methods in the class $\mathcal{G}_1$ with $s = 3, 4, 5, 6$ (see (6.39)) which have been designed by optimizing the both functions $\mathcal{E}_1$ and $\mathcal{E}_2$.

**Case** $s = 3$**.** The integrator reads

$$\psi_{\delta_t}^{[3]} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^\star_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^\star_{\alpha_1\delta_t} \tag{6.44}$$

and the unique (real) solution to the order conditions $w_1 = 1$, $w_3 = w_{12} = 0$ is given by

$$\alpha_1 = \alpha_2 = \frac{1}{2(2-2^{1/3})}, \qquad \alpha_3 = \frac{1}{2} - 2\alpha_1.$$

If $\chi_{\delta_t} = \varphi_{\delta_t}^{[a]} \circ \varphi_{\delta_t}^{[b]} \circ \varphi_{\delta_t}^{[c]}$, then it involves $13$ maps (the minimum number). For future reference, the values of the objective functions are

$$\mathcal{E}_1(\boldsymbol{\alpha}) = 4.40483, \qquad \mathcal{E}_2(\boldsymbol{\alpha}) = 4.55004.$$

**Remark 6.3.2.** *Notice that this corresponds to the familiar scheme of Yoshida [117]*

$$\psi_{\delta_t} = \phi_{\gamma\delta_t/2} \circ \phi_{\beta\delta_t} \circ \phi_{\gamma\delta_t/2}$$

*in* $\mathcal{G}_2$ *with* $\gamma = 1/(2-2^{1/3})$*. Moreover, this method is also recovered in [88] when considering splitting methods of the form (6.37).*

**Case** $s = 4$**.** The composition is

$$\psi_{\delta_t}^{[4]} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^\star_{\alpha_4\delta_t} \circ \chi_{\alpha_4\delta_t} \circ \chi^\star_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^\star_{\alpha_1\delta_t}, \tag{6.45}$$

involving $17$ maps. Now we have a free parameter, which we take as $\alpha_1$. The minima of both $\mathcal{E}_1$ and $\mathcal{E}_2$ are achieved at approximately $\alpha_1 = 0.358$, and so the coefficients are

$$\begin{aligned} \alpha_1 &= 0.358 & \alpha_2 &= -0.47710242361717810834 \\ \alpha_3 &= 0.35230499471528197958 & \alpha_4 &= 0.26679742890189612876 \end{aligned}$$

with

$$\mathcal{E}_1(\boldsymbol{\alpha}) = 2.9084, \qquad \mathcal{E}_2(\boldsymbol{\alpha}) = 3.1527.$$

**Case** $s = 5$**.** Now the composition

$$\psi_{\delta_t}^{[5]} = \chi_{\alpha_1\delta_t} \circ \chi^\star_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^\star_{\alpha_4\delta_t} \circ \chi_{\alpha_5\delta_t} \circ \chi^\star_{\alpha_5\delta_t} \circ \chi_{\alpha_4\delta_t} \circ \chi^\star_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^\star_{\alpha_1\delta_t} \tag{6.46}$$

involves $21$ maps when applied to a system separable into three parts. By carrying out a similar analysis we conclude that the best solution according with the criterion adopted is achieved when

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2(4-4^{1/3})}, \qquad \alpha_5 = \frac{1}{2} - 4\alpha_1,$$

which give $\mathcal{E}_1(\boldsymbol{\alpha}) = 2.3159, \mathcal{E}_2(\boldsymbol{\alpha}) = 2.6111.$

**Remark 6.3.3.** *This method also belongs to* $\mathcal{G}_2$ *since it can be written as*

$$\psi_{\delta_t} = \phi_{\gamma\delta_t} \circ \phi_{\gamma\delta_t} \circ \phi_{\beta\delta_t} \circ \phi_{\gamma\delta_t} \circ \phi_{\gamma\delta_t}$$

*belonging to* $\mathcal{G}_2$ *with coefficients*

$$\gamma = 2\alpha_1, \qquad \beta = 2\alpha_5.$$

*This method was proposed by Suzuki in [111].*

**Case $s = 6$.** Analogously we have considered a composition involving $3$ free parameters and $25$ maps :

$$\psi_{\delta_t}^{[6]} = \chi_{\alpha_1\delta_t} \circ \chi^{\star}_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^{\star}_{\alpha_4\delta_t} \circ \chi_{\alpha_5\delta_t} \circ \chi^{\star}_{\alpha_6\delta_t} \circ \chi_{\alpha_6\delta_t} \circ \chi^{\star}_{\alpha_5\delta_t} \circ \chi_{\alpha_4\delta_t} \circ \chi^{\star}_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^{\star}_{\alpha_1\delta_t}. \quad (6.47)$$

A solution leading to small values of $\mathcal{E}_1$ and $\mathcal{E}_2$ is

$$\begin{aligned} &\alpha_1 = \alpha_2 = \frac{3}{20} && \alpha_3 = \frac{17}{100} \\ &\alpha_4 = -0.2628463256938681137 && \alpha_5 = 0.16217658484020533783 \\ &\alpha_6 = 0.13066974085366277593 \end{aligned}$$

with

$$\mathcal{E}_1(\boldsymbol{\alpha}) = 2.0513, \qquad \mathcal{E}_2(\boldsymbol{\alpha}) = 2.4078.$$

**Remark 6.3.4.** *Although the optimization criterion we have adopted here usually leads to good methods, one can find schemes in the literature with larger values of $\mathcal{E}_1$ and $\mathcal{E}_2$ which are very efficient in practice. Thus, in particular, we mention the fourth-order splitting method designed in [33] which, once written as a method in $\mathcal{G}_1$, also involves $s = 6$ stages.*

**Remark 6.3.5.** *When a method in $\mathcal{G}_1$ is applied to the problem* (6.36) *and the basic first-order method is the composition $\chi_{\delta_t} = \varphi^{[a]}_{\delta_t} \circ \varphi^{[b]}_{\delta_t} \circ \varphi^{[c]}_{\delta_t}$, the corresponding algorithm can be implemented as (here we take as example method (6.47) with $s = 6$)*

$$\begin{aligned} &y = x_n \\ &\textit{do } j = 1:6 \\ &\quad y = \varphi^{[a]}_{\alpha_{2j-1}h} y \\ &\quad y = \varphi^{[b]}_{\alpha_{2j-1}h} y \\ &\quad \bar{\alpha} = \alpha_{2j-1} + \alpha_{2j} \\ &\quad y = \varphi^{[c]}_{\bar{\alpha}h} y \\ &\quad y = \varphi^{[b]}_{\alpha_{2j}h} y \\ &\quad y = \varphi^{[a]}_{\alpha_{2j}h} y \\ &\textit{end} \\ &x_{n+1} = y \end{aligned}$$

*where $x_n$ is the approximation of the solution at time $t^n = n\delta_t$.*

**An ODE example.**

Before we consider the application of the preceding methods in the context of the Vlasov–Maxwell system, we end this section by presenting some numerical results illustrating their relative efficiency on a simple ODE system, namely the so-called ABC-flow, with equations

$$\dot{x} = \mathcal{B}\cos y + \mathcal{C}\sin z, \qquad \dot{y} = \mathcal{C}\cos z + \mathcal{A}\sin x, \qquad \dot{z} = \mathcal{A}\cos x + \mathcal{B}\sin y, \quad (6.48)$$

which has been studied as a model volume-preserving three-dimensional flow satisfied by $x(t), y(t), z(t)$ (see [78]). The vector field $u$ is separable into three solvable parts, namely

$$u = u_a + u_b + u_c = \mathcal{A}(0, \sin x, \cos x) + \mathcal{B}(\cos y, 0, \sin y) + \mathcal{C}(\sin z, \cos z, 0).$$

For $\mathcal{B} = \mathcal{C} = 1$, $\mathcal{A} = 0.3$ we take as initial condition $(x^{in}, y^{in}, z^{in}) = (3.14, 2.77, 0)$, integrate until $t = 30$ and measure the error in phase space at the final time obtained with several values of $\delta_t$. The resulting efficiency diagram is shown in Figure 6.8. The error is measured as the number of evaluations of the different maps.

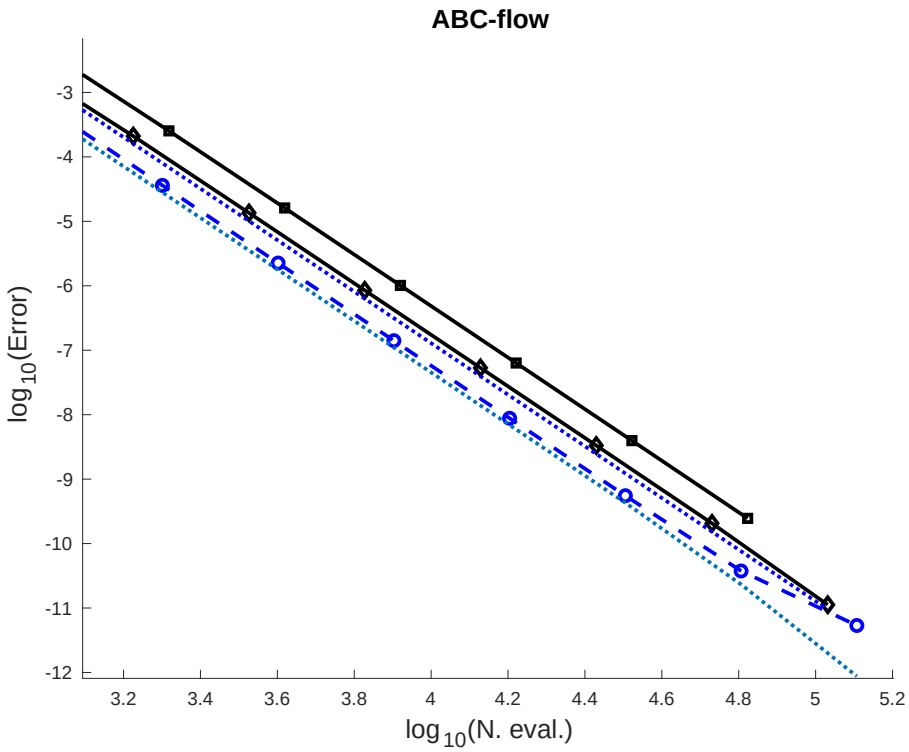


FIGURE 6.8 – ABC-flow. Efficiency diagrams obtained by different composition and splitting methods

Lines are encoded as follows : the black solid line with squares corresponds to Yoshida's method ($s = 3$) ; the black solid line with diamond is obtained with the splitting method AK 11-4 presented in [8] for systems separated into three parts, involving 21 maps ; the blue dotted line corresponds to Suzuki's method ($s = 5$) ; the blue dashed line with circles corresponds to the Blanes & Moan $S_6$ splitting method (Table 2 in [33], rewritten as a composition of the form (6.39)), and the blue dotted line has been obtained with the new composition method with $s = 6$.

For other values of the constants $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ we get different results, but essentially the same pattern holds : Suzuki's method is always more efficient than Yoshida's, and both the new composition with $s = 6$ and the splitting method of Blanes & Moan are always more efficient. For all values of the parameters we have checked, the new composition works quite

well, sometimes better than the splitting method of Blanes & Moan, sometimes slightly less efficient.

## 6.4 Numerical results

In this section, we show some numerical results to illustrate the efficiency and performance of the methods derived below. We focus on Vlasov applications by considering the Vlasov-HMF and the Vlasov-Maxwell system for which the different methods presented above are applied (new splitting and composition methods).

### 6.4.1 Vlasov-Maxwell system.

The composition methods introduced in the previous sections can then be used to derive a global $4$th order method for the Vlasov-Maxwell equation. As an example, the Yoshida method ($s = 3$) in the Vlasov-Maxwell context writes

$$\psi_{\delta_t}^{[3]} = \chi_{\alpha_1\delta_t} \circ \chi^{\star}_{\alpha_2\delta_t} \circ \chi_{\alpha_3\delta_t} \circ \chi^{\star}_{\alpha_3\delta_t} \circ \chi_{\alpha_2\delta_t} \circ \chi^{\star}_{\alpha_1\delta_t}$$

with $\alpha_1 = \alpha_2 = \frac{1}{2(2-2^{1/3})}, \quad \alpha_3 = \frac{1}{2} - 2\alpha_1$, and where $\chi_{\delta t}$ and $\chi^{\star}_{\delta t}$ are given by (6.34) and (6.35). Then, if we denote by $F^n$ an approximation at time $t^n = n\delta_t, n \in \mathbb{N}$ of the Vlasov-Maxwell solution $F(t^n)$, we have

$$F^n = \left(\psi_{\delta_t}^{[3]}\right)^n (F^{in}),$$

and $F^n$ is a $4$th order approximation of $F(t^n)$. The other $4$th order methods $\psi_{\delta_t}^{[s]}, s = 4, 5, 6$ are defined by (6.45), (6.46) and (6.47) in Subsection 6.3.3. We also define the standard Strang splitting $\psi_{\delta t}^{[2]}$ which, with our notations writes

$$\begin{aligned}
\psi_{\delta t}^{[2]} &= \chi_{\delta_t/2} \circ \chi^{\star}_{\delta_t/2} \\
&= \varphi_{\delta_t/2}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_E]} \\
&= \varphi_{\delta_t/2}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t}^{[\mathcal{H}_B]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_f]} \circ \varphi_{\delta_t/2}^{[\mathcal{H}_E]}.
\end{aligned}$$

The Strang splitting for a decomposition into three parts involves $5$ maps since, as usual, two maps can be concatenated.

We present some numerical results to illustrate the efficiency of the different methods. First of all, we used the methods $\psi_{\delta t}^{[s]}, s = 2, 3, 4, 5, 6$. In this context, one goal is to compare the new exact splitting for the rotation applied to the field $\mathcal{H}_B$ to a standard Strang method. In the methods $\psi_{\delta t}^{[s]}$, the flow $\varphi_{\delta_t}^{[\mathcal{H}_B]}$ is then approximated by the Strang splitting $\mathcal{T}_{\delta_t}^{[\mathcal{H}_B]}$ given by (6.13). This means that in this method, $\chi_{\delta_t}$ is now replaced by $\tilde{\chi}_{\delta_t}$ defined by

$$\tilde{\chi}_{\delta_t} = \varphi_{\delta_t}^{[\mathcal{H}_E]} \circ \varphi_{\delta_t}^{[\mathcal{H}_f]} \circ \mathcal{T}_{\delta_t}^{[\mathcal{H}_B]}.$$

The global Strang splitting is then defined by $\tilde{\psi}_{\delta_t}^{[2]}$

$$\tilde{\psi}_{\delta_t}^{[2]} = \tilde{\chi}_{\delta_t/2} \circ \tilde{\chi}^{\star}_{\delta_t/2},$$

and the definition of $\tilde{\psi}_{\delta_t}^{[s]}$ for $s = 3, 4, 5, 6$ follows directly. Let us remark that even if the magnetic part $\mathcal{H}_B$ is not solved exactly in time, the global method $\tilde{\psi}_{\delta t}^{[s]}$ still have the same order as $\psi_{\delta t}^{[s]}$ (i.e. of order $2$ for $s = 2$ or of order $4$ for $s = 3, 4, 5, 6$). We then want to investigate the impact of this approximation on the global error of the so-obtained splitting.

To do so, we consider the following initial condition for (6.32)

$$f^{in}(x_1, v_1, v_2) = \frac{1}{\pi v_{\text{th}^2}\sqrt{T_r}} e^{-(v_1^2+v_2^2/T_r)/v_{\text{th}}}(1 + \alpha\cos(kx_1)),$$

and $B^{in}(x_1) = 10 + 3\cos(kx_1), E_2^{in}(x_1) = 0$. We consider $\alpha = 10^{-4}, k = 0.4, v_{th} = 0.02, k = 0.4$ and $T_r = 12$. The phase space domain is $(x_1, v_1, v_2) \in [0, 2\pi/k] \times [-1, 1]^2$ and the number of points is $N_x = 8$ in space and $N_v = 513$ per direction in velocity. The runs are performed up to a final time $T = 2$ and different values of the time step $\delta_t$ are considered between $10^{-3}$ to $0.4$. The results are given in Figures 6.9 and 6.10 where we have plotted the $L^\infty$ error on the total energy with respect to $\delta t/M$ where $M$ is the number of maps. The total energy (which is conserved with time at the continuous level) is defined by

$$\mathcal{H}(t) = \frac{1}{2}\int_L |E|^2 dx + \frac{1}{2}\int_L |B|^2 \mathrm{d}x + \frac{1}{2}\int_{L\times\mathbb{R}^2} |v|^2 f dv dx \tag{6.49}$$

with $L = [0, 2\pi/k]$, and the error we consider is

$$\text{err} := \max_{t\in[0,T]} \Big|\frac{\mathcal{H}(t) - \mathcal{H}(0)}{\mathcal{H}(0)}\Big|.$$

First, one can see that the order of convergence are well recovered for all the methods but some fourth order methods present some better efficiency. For instance, the two methods corresponding to $s = 5$ and $s = 6$ are clearly the best, and are much more efficient that the triple jump method ($s = 3$) or the Strang one ($s = 2$) even if they involve a larger number of maps (as exemplified in the ODE context in Subsection 6.3.3). Second, we can observe than the error produced by the methods $\psi_{\delta t}^{[s]}$ (i.e. when the exact splitting is used for the part $\mathcal{H}_B$) is smaller than the error performed by the methods $\tilde{\psi}_{\delta t}^{[s]}$ (i.e. when a Strang splitting is used for the part $\mathcal{H}_B$). Note that on Figures 6.9 and 6.10, the lines indicating the order are kept fixed. Moreover, on Figure 6.10, in addition to the error produced by the methods $\tilde{\psi}_{\delta t}^{[s]}$, the error curves produced by the methods $\psi_{\delta t}^{[2]}$ and $\psi_{\delta t}^{[5]}$ (labelled by $s = 2$ new (5) and $s = 5$ new (21) in the legend) have been displayed to ease the comparisons. Note that we have chosen to plot the $\psi_{\delta t}^{[5]}$ method but very similar conclusions arise with the choice $\psi_{\delta t}^{[6]}$ since the two methods $\psi_{\delta t}^{[5]}$ and $\psi_{\delta t}^{[6]}$ have a very close efficiency in our context. For the global Strang method the ratio between the error produced by $\tilde{\psi}_{\delta t}^{[2]}$ and $\psi_{\delta t}^{[2]}$ is about $2.5$ whereas the ratio between the error produced by $\tilde{\psi}_{\delta t}^{[5]}$ and $\psi_{\delta t}^{[5]}$ is about $6$ (the same ratio is observed between $\tilde{\psi}_{\delta t}^{[6]}$ and $\psi_{\delta t}^{[6]}$). Let us remark that, for a given method (i.e. a given $s$), the cost of a $\tilde{\psi}_{\delta t}^{[s]}$ method is the same as a the one $\psi_{\delta t}^{[s]}$ method.

We end this subsection by considering other splitting methods from the literature, namely the splitting methods from [8] which assume that each subpart are solved exactly, which is our case when the exact splitting is used for the magnetic part. The results are displayed in Figure 6.11 where we have tested second order methods (AK 3-2 and AK 5-2 involve $9$ maps), a fourth order method (AK 11-4 involves $21$ maps) and even a sixth order method (AY 15-6

involves $29$ maps). We refer to see [8] for more details on these methods. As previously we also added $\psi_{\delta t}^{[2]}$ (second order) and $\psi_{\delta t}^{[5]}$ (fourth order) for comparison, whereas the slope $2$ and $4$ are the same as in Figures 6.9 and 6.10. First, we observe that AK 3-2 is the best second order method. The third order PP method is not very attractive in this context compared to second order methods. Second, among the two fourth order methods (AK 11-4 and $\psi_{\delta t}^{[5]}$), the method $\psi_{\delta t}^{[5]}$ offers a better efficiency since the error is about $5$ times smaller, as already noticed in the ODE framework (see Subsection 6.3.3). Finally, the method AY 15-6 offers sixth order accuracy and turns out to be the best method here when the time step is small enough.

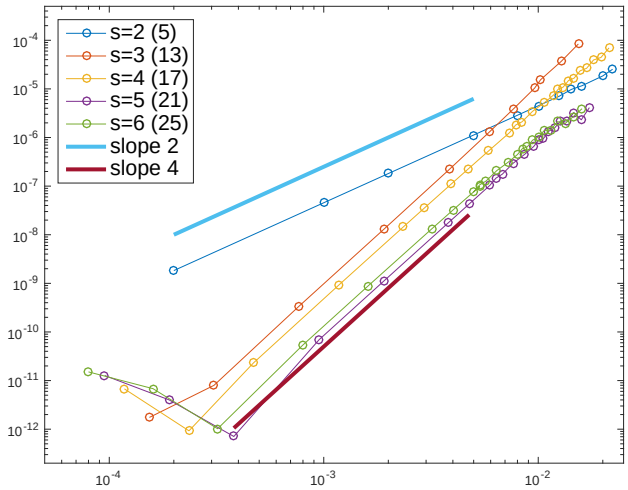


FIGURE 6.9 – Efficiency diagrams obtained by different composition methods $\psi_{\delta_t}^{[s]}, s = 2, 3, 4, 5, 6$ for the Vlasov-Maxwell system. The number of maps for each method is indicated into parenthesis.

### 6.4.2 Vlasov-HMF system.

Our goal is to solve numerically the Vlasov-HMF model satisfied by $f(t, x, v)$, $(x, v) \in L \times \mathbb{R}$, with $L = \mathbb{R}/2\pi\mathbb{Z}$

$$\partial_t f + \{f, H[f]\} = 0, \tag{6.50}$$

where $\{f, g\} = \partial_x f \partial_v g - \partial_v f \partial_x g$ and $H[f]$ is given by

$$H[f] = \frac{v^2}{2} - \Phi[f](x).$$

Finally, the potential is defined by

$$\Phi[f](x) = \cos x \int_{L\times\mathbb{R}} \cos(y) f(y, u) \mathrm{dydu} + \sin x \int_{L\times\mathbb{R}} \sin(y) f(y, u) \mathrm{dydu}. \tag{6.51}$$

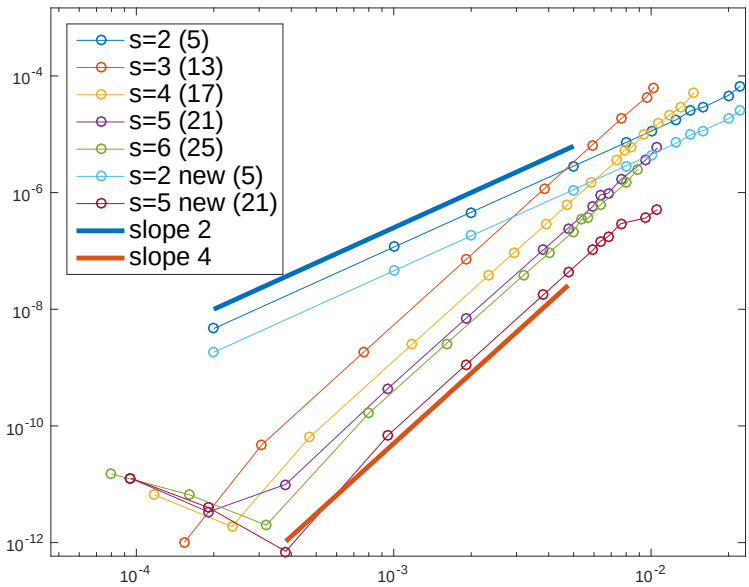


FIGURE 6.10 – Efficiency diagrams obtained by different composition methods $\tilde{\psi}_{\delta_t}^{[s]}, s = 2, 3, 4, 5, 6$ and $\psi_{\delta_t}^{[2]}$, $\psi_{\delta_t}^{[5]}$ for the Vlasov-Maxwell system. The order lines 'slope 2' and 'slope 4' are the same as in Figure 6.9. The number of maps for each method is indicated into parenthesis.

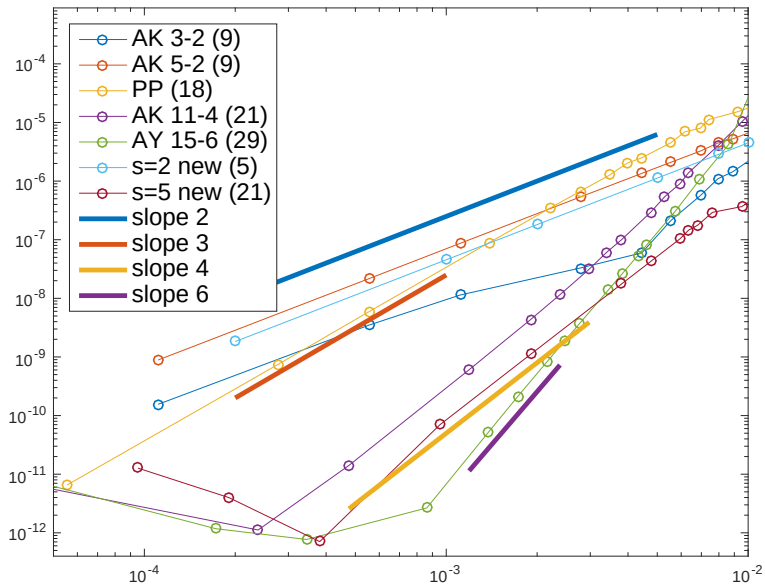


FIGURE 6.11 – Efficiency diagrams obtained by different methods from [8] and $\psi_{\delta_t}^{[2]}$, $\psi_{\delta_t}^{[5]}$ for the Vlasov-Maxwell system. The order lines 'slope 2' and 'slope 4' are the same as in Figure 6.9. The number of maps for each method is indicated into parenthesis.

We consider the following stationary solution (see [82] for more details)

$$f^{eq}(x,v) = \gamma e^{-\beta\left(\frac{v^2}{2} - M_0 \cos x\right)} \text{ with } M_0 = \int_{L\times\mathbb{R}} \cos(y) f^{eq}(y,u)\mathrm{dydu}, \tag{6.52}$$

where $\gamma, \beta, M_0 \in \mathbb{R}$ will be explicitly given below. Following [82], the long time behavior of (6.50) is driven by the linearized Hamiltonian part, i.e. $\partial_t f + \{f, H[f^{eq}]\} = 0$, with $H[f^{eq}] = \frac{v^2}{2} - M_0 \cos(x)$. We recognize the pendulum Hamiltonian for which a slight modification of the new splitting is able to capture the rotation phenomena with high accuracy compare to standard Strang splitting (see [21]). In this HMF context, the material introduced before have to be slightly modified.

First, let us introduce the discretization of the phase space $L\times[-v_{\max}, v_{\max}]$, with $v_{\max} > 0$ a truncation of the velocity direction. We consider $\mathbb{G}_x := h_x [\![0, N_x - 1]\!]$ the space grid (with $h_x = L/N_x$ the stepsize and $N_x \in \mathbb{N}^*$ the number of points) and $\mathbb{G}_v := h_v [\![-\lfloor (N_v-1)/2\rfloor, \lfloor N_v/2\rfloor]\!]$ the speed grid (with $h_v = 2v_{\max}/N_v$ the stepsize and $N_v \in \mathbb{N}^*$ the number of points). We also introduce the set of discrete frequencies : $\widehat{\mathbb{G}}_x = \eta_x [\![-\lfloor (N_x-1)/2\rfloor, \lfloor N_x/2\rfloor]\!]$ and $\widehat{\mathbb{G}}_v = \eta_v [\![-\lfloor (N_v-1)/2\rfloor, \lfloor N_v/2\rfloor]\!]$ with $\eta_x = 2\pi/L$ and $\eta_v = \pi/v_{\max}$. Then, we define the discrete partial Fourier transforms

$$\mathcal{F}_1 : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}_x\times\mathbb{G}_v} & \to & \mathbb{C}^{\widehat{\mathbb{G}}_x\times\mathbb{G}_v} \\ \boldsymbol{u} & \mapsto & h_x \sum_{g_1\in\mathbb{G}_x} \boldsymbol{u}_{g_1,g_2}\, e^{-ig_1\xi_1} \end{array} \right.$$

and

$$\mathcal{F}_2 : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v} & \to & \mathbb{C}^{\mathbb{G}_x \times \widehat{\mathbb{G}}_v} \\ \boldsymbol{u} & \mapsto & h_v \sum_{g_2 \in \mathbb{G}_v} \boldsymbol{u}_{g_1,g_2}\, e^{-ig_2\xi_2} \end{array} \right. ,$$

whereas the shears are now defined by

$$\mathcal{S}_1^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v} & \to & \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v} \\ \boldsymbol{u} & \mapsto & \mathcal{F}_1^{-1}\left[e^{i\alpha\xi_1 g_2}\mathcal{F}_1\,\boldsymbol{u}\right] \end{array} \right.$$

and

$$\widetilde{\mathcal{S}}_2^\alpha : \left\{ \begin{array}{ccc} \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v} & \to & \mathbb{C}^{\mathbb{G}_x \times \mathbb{G}_v} \\ \boldsymbol{u} & \mapsto & \mathcal{F}_2^{-1}\left[e^{i\alpha\xi_2 E[\boldsymbol{u}]_{g_1}}\mathcal{F}_2\boldsymbol{u}\right] \end{array} \right. ,$$

where $E[\boldsymbol{u}]_{g_1}$ is deduced from the relation $E[\boldsymbol{u}](x) = -\partial_x\Phi[\boldsymbol{u}](x)$ and (6.51)

$$\begin{aligned} E[\boldsymbol{u}]_{g_1} &= \sin(g_1h_x)h_xh_v \sum_{(g_1,g_2)\in\mathbb{G}_x\times\mathbb{G}_v} \cos(g_1h_x)\,\boldsymbol{u}_{g_1,g_2} \\ &\quad - \cos(g_1h_x)h_xh_v \sum_{(g_1,g_2)\in\mathbb{G}_x\times\mathbb{G}_v} \sin(g_1h_x)\,\boldsymbol{u}_{g_1,g_2}\,. \end{aligned} \tag{6.53}$$

Then, at time $t^n = n\delta_t$, we denote by $f^n$ an approximation of the solution $f(t^n)$ on the phase space grid computed by the Strang splitting $\widetilde{\mathcal{T}}_{\delta_t}$ and the new splitting $\widetilde{\mathcal{M}}_{\delta_t}$ which are defined by

$$\begin{aligned} f^{n+1} &= \widetilde{\mathcal{T}}_{\delta_t}\, f^n := \mathcal{S}_1^{-\delta_t/2}\widetilde{\mathcal{S}}_2^{\delta_t}\mathcal{S}_1^{-\delta_t/2}\, f^n, && \text{(Strang)} \\ f^{n+1} &= \widetilde{\mathcal{M}}_{\delta_t}\, f^n \\ &:= \mathcal{S}_1^{-t_c\tan(\delta_t/(2t_c))}\widetilde{\mathcal{S}}_2^{t_c\sin(\delta_t/t_c)}\mathcal{S}_1^{-t_c\tan(\delta_t/(2t_c))}\, f^n, && \text{(New)} \end{aligned} \tag{6.54}$$

where $f^0 := f^{in}$, and $t_c = \frac{1}{\sqrt{M_0}}$ is the characteristic time of the Vlasov-HMF model which has been introduced to capture the correct angular velocity. Let us remark that the electric field $E[f]$ has to be solved using (6.53) before the shear $\widetilde{\mathcal{S}}_2^\alpha$ in the splittings (6.54).

To evaluate the performance of the new splitting compare to the Strang one, we consider an initial condition $f^{in}$ as a perturbation of the equilibrium solution (6.52) (with $\beta = 10, M_0 = 0.9455421864232981$ and $\alpha = 0.0001194365987897421$)

$$f^{in}(x,v) = f^{eq}(x,v)(1+\varepsilon\cos(x)), \quad (x,v)\in[-\pi,\pi]\times\mathbb{R},$$

with $\varepsilon = 10^{-3}$. We consider a truncated velocity domain of $[-8,8]$, the number of points in the spatial direction is $N_x = 128$ whereas we considered $N_v = 256$ points in the velocity direction, and the final time is $T = 25$. Note that the splitting can also be coupled to a semi-Lagrangian method ; the shears $\mathcal{S}_1$ and $\widetilde{\mathcal{S}}_2$ have to be modified accordingly (see [30] for instance).

We look at the $L^\infty$ error between a reference distribution function (obtained with the new splitting with a small time step $\delta_t = T/1000$) and the one obtained by Strang or new splitting given by (6.54) (with $t_c = 1.0283940255$) for different time steps $\delta_t \in \{T/50, T/100, T/150, T/200, T/250\}$. The results are displayed in Figure 6.12 in $\log$-$\log$ scale. First we observe that, as expected, the two methods are second order accurate in time. But, one can remark that the error produced by the new splitting is much more smaller that the error produced by the Strang splitting, for a same cost (the number of maps is the same for the two methods).

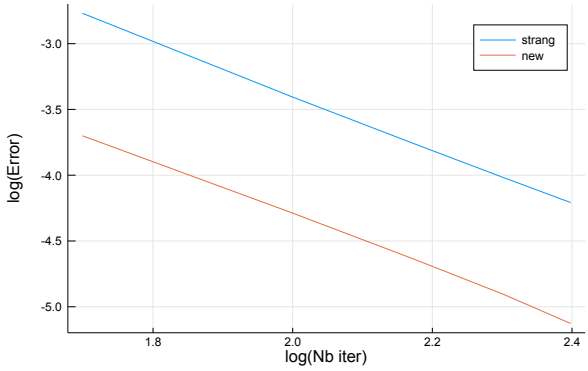


FIGURE 6.12 – Error as a function of the number of iterations for the HMF-Poisson system. Comparison of the Strang splitting ('strang') and the New splitting ('new').

## 6.5 Conclusion

In this work, we study a directional splitting which preserves exactly the rotations and apply to the PDE context. A careful numerical analysis of this splitting coupled with spectral interpolation techniques has been performed. These results are illustrated by some numerical experiments.

Then, this step serves as a building block of a splitting for the Vlasov-Maxwell system. Indeed, this system can be split into three parts which, thanks to this new splitting, can all be solved exactly. New high order composition methods are then designed to accurately and efficiently solve the full Vlasov-Maxwell system. Numerical results show the good behavior of these methods. Finally, a direct application of the new splitting for close to equilibrium simulations of the Vlasov-HMF model also shows very good results compared to the standard Strang splitting, with no additional cost.

The extension to the relativistic Vlasov-Maxwell equations in two or three dimensions in the velocity space are planned. The approach should be even more attractive in this context.

## 6.6 Appendix

In this Appendix, we gather the proofs of the different lemmas occuring in the proof of convergence of the pseudo-spectral splitting methods.

### 6.6.1 Proof of Lemma 6.2.2

If $0 \in [y_1; y_1 + \lambda y_2]$ then we have $|y_1| \leqslant \lambda |y_2|$ and so we get

$$\left| \begin{pmatrix} y_3 \\ y_2 \end{pmatrix} \right| \geqslant |y_2| \geqslant \frac{|y_1|}{|\lambda|} \geqslant \frac{|y_1|}{\sqrt{1+\lambda^2}}.$$

Else we have $|y_3| = |y_1|$ or $|y_3| = |y_1 + \lambda y_2|$. If $|y_3| = |y_1|$ then we have

$$\left|\begin{pmatrix} y_3 \\ y_2 \end{pmatrix}\right| \geqslant |y_3| = |y_1| \geqslant \frac{|y_1|}{\sqrt{1+\lambda^2}}.$$

Else if $|y_3| = |y_1 + \lambda y_2|$, we have

$$\left|\begin{pmatrix} y_3 \\ y_2 \end{pmatrix}\right|^2 = y_2^2 + (y_1 + \lambda y_2)^2.$$

This last quantity is a second order polynomial with respect to $y_2$. Thus its infimum can be determined explicitly. More precisely, we have

$$y_2^2 + (y_1 + \lambda y_2)^2 \geqslant \frac{|y_1|^2}{1+\lambda^2}.$$

### 6.6.2 Proof of Lemma 6.2.3

However, if $a$ and $b$ are small enough, $P_{a,b}$ is closed to the identity. Consequently, it admits a logarithm $M_{a,b} \in M_2(\mathbb{R})$, defined by

$$M_{a,b} = \sum_{n\in\mathbb{N}} \frac{(-1)^n}{n+1}(P_{a,b} - I_2)^n,$$

and satisfying

$$e^{M_{a,b}} = P_{a,b}.$$

A fortiori, we have $\exp(\operatorname{Tr} M_{a,b}) = \det P_{a,b} = 1$. Hence we have $\operatorname{Tr} M_{a,b} = 0$. Furthermore, the following application define an isomorphism of vector space (it is an injection between two spaces of dimension $3$)

$$\begin{cases} S_2(\mathbb{R}) & \to & \mathfrak{sl}_2(\mathbb{R}) \\ L & \mapsto & JL \end{cases}.$$

where $\mathfrak{sl}_2(\mathbb{R}) = \{M \in M_2(\mathbb{R}) \mid \operatorname{Tr} M = 0\}$. As a consequence, there exists a symmetric matrix $L_{a,b} \in S_2(\mathbb{R})$ such that

$$M_{a,b} = JL_{a,b}.$$

### 6.6.3 Proof of Lemma 6.2.4

Since $0 < ab < 4$, $Q_{a,b}$ is either positive or negative, and, as a consequence, the following Euclidian norm is well defined on $S_2(\mathbb{R})$

$$\forall K \in S_2(\mathbb{R}),\ \|K\|_{a,b}^2 := \int_{\mathbb{R}^2} ({}^{\mathrm{t}}xKx)^2 e^{-|{}^{\mathrm{t}}xQ_{a,b}x|}\mathrm{d}x.$$

Since $\det P_{a,b} = 1$, computing $\|R_{a,b}^{-1}K\|_{a,b}$, we deduce of a change of variables and of (6.24) that

$$\forall K \in S_2(\mathbb{R}),\ \|R_{a,b}K\|_{a,b} = \|K\|_{a,b}.$$

This relation means that $R_{a,b}$ is an isometry for the Euclidian norm $\|\cdot\|_{a,b}$. A fortiori, we have $\det R_{a,b} = \pm 1$. But, since $R_{0,0} = I_2$ and $(a,b) \mapsto \det R_{a,b}$ is a continuous map, we deduce that $\det R_{a,b} = 1$. Consequently, $R_{a,b}$ is a rotation in a space of dimension $3$. So, there are only two possibilities : either $R_{a,b}$ is the identity or the eigenspace of $R_{a,b}$ associated with the eigenvalue $1$ is of dimension $1$.

To conclude, we just have to verify that $P_{a,b}$ is not the identity. First, we observe that $P_{a,b}$ is not a scalar matrix, so there exists $x \in \mathbb{R}^2$ such that $x$ is not an eigenvector of $P_{a,b}$. Then, we consider a vector $y \in \mathbb{R}^2 \backslash \{0\}$ such that $x$ and $y$ are orthogonal. By construction, we have

$$ {}^{\mathrm{t}}y P_{a,b} x \neq 0. $$

Consequently, if $K = y\, {}^{\mathrm{t}}y \in S_2(\mathbb{R})$, we have

$$ {}^{\mathrm{t}}x R_{a,b}(K) x = ({}^{\mathrm{t}}y P_{a,b} x)^2 \neq 0 = ({}^{\mathrm{t}}y x)^2 = {}^{\mathrm{t}}x K x. $$

Thus, we have $R_{a,b}(K) \neq K$.

### 6.6.4 Proof of Lemma 6.2.7

We have to bound $\||x|^s (u \circ \tau)\|_{L^2(\mathbb{R}^2)}$ and $\||\xi|^s \mathscr{F}(u \circ \tau)\|_{L^2(\mathbb{R}^2)}$. However, a straightforward calculation shows that we have

$$ \mathscr{F}(u \circ \tau) = |\det \tau|^{-1} (\mathscr{F} u) \circ^t \tau^{-1}, $$

and equation (6.27) is clearly equivalent to

$$ |\tau| \leqslant \kappa \text{ and } |\tau^{-1}| \leqslant \kappa. $$

Thus, since $|\tau| = |{}^t\tau|$, if we get a bound on $\||x|^s (u \circ \tau)\|_{L^2(\mathbb{R}^2)}$, uniform with respect to $\tau$, we also a a bound on $\||\xi|^s \mathscr{F}(u \circ \tau)\|_{L^2(\mathbb{R}^2)}$ uniform with respect to $\tau$.

Finally, to bound $\||x|^s (u \circ \tau)\|_{L^2(\mathbb{R}^2)}$, we just have to apply a change of coordinates :

$$ \||x|^s (u \circ \tau)\|_{L^2(\mathbb{R}^2)} = \sqrt{|\det \tau|^{-1}} \||\tau(x)|^s u\|_{L^2(\mathbb{R}^2)} \leqslant \sqrt{|\det \tau^{-1}|} |\tau|^s \|u\|_{X^s} \leqslant \kappa^{s+1} \|u\|_{X^s}. $$

### 6.6.5 Proof of Lemma 6.2.8

First, we apply the Poisson formula and the discrete Fourier Plancherel isometry to get

$$ \|v_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)} \leqslant \|v_{|h\mathbb{Z}^2}\|_{L^2(h\mathbb{Z}^2)} = \frac{1}{2\pi} \| \sum_{k \in \mathbb{Z}^2} \mathscr{F} v(\cdot + \frac{2k\pi}{h}) \|_{L^2((-\frac{\pi}{h}, \frac{\pi}{h})^2)}. $$

Then we observe that if $k \in \mathbb{Z}^2 \backslash \{0\}$ and $\xi \in (-\frac{\pi}{h}, \frac{\pi}{h})^2$ then we have

$$ |\xi + \frac{2k\pi}{h}| \geqslant \frac{\pi}{h}(2|k| - \sqrt{2}). $$

Thus, we control $\|v_{|\mathbb{G}^2}\|_{L^2(\mathbb{G}^2)}$ by

$$ \frac{1}{2\pi} \|\mathscr{F} v\|_{L^2((-\frac{\pi}{h}, \frac{\pi}{h})^2)} + \frac{h^2}{2\pi^3} \sum_{k \in \mathbb{Z}^2 \backslash \{0\}} \frac{1}{(2|k| - \sqrt{2})^2} \|(|\xi|^2 \mathscr{F} v)(\cdot + \frac{2k\pi}{h})\|_{L^2((-\frac{\pi}{h}, \frac{\pi}{h})^2)}. $$

Finally, applying the Cauchy Schwarz inequality and the Chasles relation, we control the second term by

$$\frac{h^2}{2\pi^3}\||\xi|^2\mathscr{F}v\|_{L^2(\mathbb{R}^2)}\sqrt{\sum_{k\in\mathbb{Z}^2\setminus\{0\}}\frac{1}{(2|k|-\sqrt{2})^4}}.$$

# BIBLIOGRAPHIE


[1] M. J. ABLOWITZ AND J. F. LADIK, *A nonlinear difference scheme and inverse scattering*, Studies in Appl. Math., 55 (1976), pp. 213–229.

[2] M. J. ABLOWITZ, B. PRINARI, AND A. D. TRUBATCH, *Discrete and continuous nonlinear Schrödinger systems*, vol. 302 of London Mathematical Society Lecture Note Series, Cambridge University Press, Cambridge, 2004.

[3] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55 of National Bureau of Standards Applied Mathematics Series, For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[4] P. ALPHONSE AND J. BERNIER, *Polar decomposition of semigroups generated by quadratic operators and regularizing effects*. preprint May 2019.

[5] E. ANDRES, *The quasi-shear rotation*, In : Miguet S., Montanvert A., Ubéda S. (eds) Discrete Geometry for Computer Imagery., 1176 (1996).

[6] W. ARENDT, C. J. K. BATTY, M. HIEBER, AND F. NEUBRANDER, *Vector-valued Laplace transforms and Cauchy problems*, vol. 96 of Monographs in Mathematics, Birkhäuser/Springer Basel AG, Basel, second ed., 2011.

[7] V. I. ARNOLD, *Small denominators and problems of stability of motion in classical and celestial mechanics*, Russian Math. Surveys, 18 (1963), pp. 24–34.

[8] W. AUZINGER, H. HOFSTÄTTER, D. KETCHESON, AND O. KOCH, *Practical splitting methods for the adaptive integration of nonlinear evolution equations. part i : Construction of optimized schemes and pairs of schemes*, BIT Numer. Math., 57 (2017), pp. 55–74.

[9] P. BADER AND S. BLANES, *Fourier methods for the perturbed harmonic oscillator in linear and nonlinear schrödinger equations*, Phys. Rev. E, 83 (2011), p. 046711.

[10] M. BADSI AND M. HERDA, *Modelling and simulating a multispecies plasma*, ESAIM : ProcS, 53 (2016), pp. 27–37.

[11] G. A. BAKER, JR. AND P. GRAVES-MORRIS, *Padé approximants*, vol. 59 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, second ed., 1996.

[12] D. BAMBUSI, *Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equations*, Math. Z., 130 (1999), pp. 345–387.

[13] ——, *On long time stability in Hamiltonian perturbations of non-resonant linear PDEs*, Nonlinearity, 12 (1999), pp. 823–850.

[14] ——, *Birkhoff normal form for some nonlinear PDEs*, Comm. Math. Physics, 234 (2003), pp. 253–283.

[15] ——, *A Birkhoff normal form theorem for some semilinear PDEs*, in Hamiltonian dynamical systems and applications, NATO Sci. Peace Secur. Ser. B Phys. Biophys., Springer, Dordrecht, 2008, pp. 213–247.

[16] D. BAMBUSI, J.-M. DELORT, B. GRÉBERT, AND J. SZEFTEL, *Almost global existence for Hamiltonian semilinear Klein-Gordon equations with small Cauchy data on Zoll manifolds*, Comm. Pure Appl. Math., 60 (2007), pp. 1665–1690.

[17] D. BAMBUSI, E. FAOU, AND B. GRÉBERT, *Existence and stability of ground states for fully discrete approximations of the nonlinear Schrödinger equation*, Numer. Math., 123 (2013), pp. 461–492.

[18] D. BAMBUSI AND B. GRÉBERT, *Birkhoff normal form for PDE's with tame modulus*, Duke Math. J., 135 (2006), pp. 507–567.

[19] D. BAMBUSI AND T. PENATI, *Continuous approximation of breathers in one- and two-dimensional DNLS lattices*, Nonlinearity, 23 (2010), pp. 143–157.

[20] Y. BARSAMIAN, J. BERNIER, S. A. HIRSTOAGA, AND M. MEHRENBERGER, *Verification of 2D$\times$ 2D and two-species vlasov-poisson solvers*, ESAIM : ProcS, 63 (2018), pp. 78–108.

[21] K. BEAUCHARD AND F. MARBACH, *Personnal communication*.

[22] J. BEDROSSIAN, N. MASMOUDI, AND C. MOUHOT, *Landau damping : paraproducts and Gevrey regularity*, Ann. PDE, 2 (2016), pp. Art. 4, 71.

[23] J. BERNIER, *Optimality and resonances in a class of compact finite difference schemes of high order*. preprint, Oct. 2017.

[24] ——, *Bounds on the growth of high discrete Sobolev norms for the cubic discrete non-linear Schödinger equations on $h\mathbb{Z}$*, Discrete and Continuous Dynamical Systems-A, 39 (2019), pp. 3179–3195.

[25] J. BERNIER, N. CROUSEILLES, AND F. CASAS, *Splitting methods for rotations : application to vlasov equations*. to appear soon.

[26] J. BERNIER AND E. FAOU, *Existence and stability of traveling waves for discrete nonlinear Schrödinger equations over long times*. preprint, May 2018.

[27] J. BERNIER, E. FAOU, AND B. GRÉBERT, *Rational normal forms and stability of small solutions to nonlinear schrödinger equations*. arXiv : 1812.11414.

[28] J. BERNIER AND M. MEHRENBERGER, *Long-time behavior of second order linearized vlasov-poisson equations near a homogeneous equilibrium*. arXiv : 1903.08374.

[29] M. BERTI AND J. DELORT, *Almost global existence of solutions for capillarity-gravity water waves equations with periodic spatial boundary conditions*. arXiv :1702.04674.

[30] N. BESSE AND M. MEHRENBERGER, *Convergence of classes of high-order semi-lagrangian schemes for the vlasov-poisson system*, Math. Comp., 77 (2008), pp. 93–123.

[31] L. BIASCO, J. MASSETTI, AND M. PROCESI, *Exponential stability estimates for the 1D NLS*. arXiv :1810.06440.

[32] S. BLANES AND P. MOAN, *Relativistic plasma simulation-optimization of a hybrid code*, Proceedings of the Fourth Conference on Numerical Simulations of Plasmas held at the Naval Research Laboratory, Washington DC„ (1970).

[33] ——, *Practical symplectic partitioned runge–kutta and runge–kutta–nyström methods*, J. Comput. Appl. Math., 142 (2002), pp. 313–330.

[34] J. M. Borwein and M. Chamberland, *Integer powers of arcsin*, Int. J. Math. Math. Sci., (2007), pp. Art. ID 19381, 10.

[35] J. Bourgain, *Construction of approximative and almost-periodic solutions of perturbed linear schrödinger and wave equations.*, Geom. Funct. Anal., 6 (1996), pp. 201–230.

[36] ——, *On the growth in time of higher order sobolev norms of smooth solutions of hamiltonian pde.*, International Mathematics Research Notices, 6 (1996), pp. 277–304.

[37] ——, *On the growth in time of higher Sobolev norms of smooth solutions of Hamiltonian PDE*, Internat. Math. Res. Notices, (1996), pp. 277–304.

[38] ——, *On growth of Sobolev norms in linear Schrödinger equations with smooth time dependent potential*, J. Anal. Math., 77 (1999), pp. 315–348.

[39] ——, *On diffusion in high-dimensional Hamiltonian systems and PDE.*, J. Anal. Math., 80 (2000), pp. 1–35.

[40] ——, *Remarks on stability and diffusion in high-dimensional hamiltonian systems and partial differential equations.*, Ergod. Th. & Dynam. Sys., 24 (2003), pp. 1331–1357.

[41] J. H. Bramble and B. E. Hubbard, *New monotone type approximations for elliptic problems*, Math. Comp., 18 (1964), pp. 349–367.

[42] F. Califano, F. Pegoraro, S. V. Bulanov, and A. Mangeney, *Kinetic saturation of the weibel instability in a collisionless plasma*, Phys. Rev. E, 57 (1998), pp. 7048–7059.

[43] R. Carles and E. Faou, *Energy cascades for NLS on the torus*, Discrete Contin. Dyn. Syst., 32 (2012), pp. 2063–2077.

[44] F. Casas, N. Crouseilles, E. Faou, and M. Mehrenberger, *High-order hamiltonian splitting for the Vlasov-Poisson equations*, Numerische Mathematik, 135 (2017), pp. 769–801.

[45] T. Cazenave, *Semilinear Schrödinger equations*, vol. 10 of Courant Lecture Notes in Mathematics, New York University, Courant Institute of Mathematical Sciences, New York ; American Mathematical Society, Providence, RI, 2003.

[46] T. Cazenave and P.-L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[47] B. Chen and A. Kaufman, *3d volume rotation using shear transformation*, Graphical Models, 62 (2000), pp. 308–322.

[48] Y. Cheng, I. Gamba, F. Lie, and P. Morrison, *Discontinuous galerkin methods for vlasov-maxwell equations*, SIAM J. Numer. Anal., 52 (2014), pp. 1017–1049.

[49] S. A. Chin and E. Krotscheck, *Fourth-order algorithms for solving the imaginary-time gross-pitaevskii equation in a rotating anisotropic trap*, Phys. Rev. E, 72 (2005), p. 036705.

[50] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Polynomial upper bounds for the orbital instability of the 1D cubic NLS below the energy norm*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 31–54.

[51] ——, *Transfer of energy to high frequencies in the cubic defocusing nonlinear schrödinger equation*, Invent. Math., 181 (2010), pp. 39–113.

[52] J. COLLIANDER, S. KWON, AND T. OH, *A remark on normal forms and the "upside-down" I-method for periodic NLS : Growth of higher sobolev norms*, Journal d'Analyse Mathématique, 118 (2012), pp. 55–82.

[53] J.-F. COULOMBEL, *Stability of finite difference schemes for hyperbolic initial boundary value problems*, in HCDTE lecture notes. Part I. Nonlinear hyperbolic PDEs, dispersive and transport equations, vol. 6 of AIMS Ser. Appl. Math., Am. Inst. Math. Sci. (AIMS), Springfield, MO, 2013, p. 146.

[54] ——, *On the strong stability of finite difference schemes for hyperbolic systems in two space dimensions*, Calcolo, 51 (2014), pp. 97–108.

[55] N. CROUSEILLES, L. EINKEMMER, AND E. FAOU, *Hamiltonian splitting for the vlasov-maxwell equations*, J. Comput. Phys., 283 (2015), pp. 224–240.

[56] ——, *An asymptotic preserving scheme for the relativistic vlasov-maxwell equations in the classical limit*, Comput. Phys. Comm., 209 (2016), pp. 13–26.

[57] N. CROUSEILLES, M. MEHRENBERGER, AND E. SONNENDRÜCKER, *Conservative semi-lagrangian schemes for vlasov equations*, Journal of Computational Physics, 229 (2010), pp. 1927–1953.

[58] S. DE BIÈVRE, F. GENOUD, AND S. ROTA NODARI, *Orbital stability : analysis meets geometry*, in Nonlinear optical and atomic systems, vol. 2146 of Lecture Notes in Math., Springer, Cham, 2015, pp. 147–273.

[59] P. DEGOND, *Spectral theory of the linearized Vlasov-Poisson equation*, Trans. Amer. Math. Soc., 294 (1986), pp. 435–453.

[60] J. DENAVIT, *First and second order landau damping in maxwellian plasmas*, Physics of Fluids, 8 (1965), pp. 471–478.

[61] J. C. EILBECK, P. S. LOMDAHL, AND A. C. SCOTT, *The discrete self-trapping equation*, Phys. D, 16 (1985), pp. 318–338.

[62] H. ELIASSON, B. GRÉBERT, AND S. B. KUKSIN, *Kam for non-linear beam equation*, Geometric And Functional Analysis, 26 (2016), pp. 1588–1715.

[63] E. FAOU, *Geometric numerical integration and Schrödinger equations*, European Math. Soc., 2012.

[64] E. FAOU AND B. GRÉBERT, *A Nekhoroshev-type theorem for the nonlinear schrödinger equation on the torus*, Analysis & PDE, 6 (2013), pp. 1243–1262.

[65] F. FILBET, E. SONNENDRÜCKER, AND P. BERTRAND, *Conservative numerical schemes for the vlasov equation*, Journal of Computational Physics, 172 (2001), pp. 166–187.

[66] B. FORNBERG, *Generation of finite difference formulas on arbitrarily spaced grids*, Math. Comp., 51 (1988), pp. 699–706.

[67] J. FRÖHLICH, S. GUSTAFSON, B. L. G. JONSSON, AND I. M. SIGAL, *Solitary wave dynamics in an external potential*, Comm. Math. Phys., 250 (2004), pp. 613–642.

[68] D. FURIHATA AND T. MATSUO, *Discrete variational derivative method—a structure preserving numerical method for partial differential equations*, Sūgaku, 66 (2014), pp. 135–156.

[69] P. GÉRARD AND S. GRELLIER, *Effective integrable dynamics for a certain nonlinear wave equation*, Analysis and PDE, 5 (2012), p. 1139–1155.

[70] J. GILEWICZ, *Approximants de Padé*, vol. 667 of Lecture Notes in Mathematics, Springer, Berlin, 1978.

[71] B. GRÉBERT, *Birkhoff normal form and hamiltonian pdes*, Séminaires et Congrès, 15 (2007), pp. 1–46.

[72] B. GRÉBERT, R. IMEKRAZ, AND E. PATUREL, *Normal forms for semilinear quantum harmonic oscillators*, Commun. Math. Phys., 291 (2009), pp. 763–798.

[73] B. GRÉBERT AND L. THOMANN, *Resonant dynamics for the quintic non linear schrödinger equation*, Ann. I. H. Poincaré - AN, 29 (2012), pp. 455–477.

[74] L. GREENGARD AND J. LEE, *Accelerating the nonuniform fast fourier transform*, SIAM Review, 46 (2004), p. 443.

[75] M. GRILLAKIS, J. SHATAH, AND W. STRAUSS, *Stability theory of solitary waves in the presence of symmetry. I*, J. Funct. Anal., 74 (1987), pp. 160–197.

[76] ——, *Stability theory of solitary waves in the presence of symmetry. II*, J. Funct. Anal., 94 (1990), pp. 308–348.

[77] P.-A. GUIHÉNEUF, *Rotations discrètes*. http ://images.math.cnrs.fr/Rotations-discretes.html, 5 Mai 2018.

[78] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2010. Structure-preserving algorithms for ordinary differential equations, Reprint of the second edition (2006).

[79] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving ordinary differential equations. I*, vol. 8 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 1993. Nonstiff problems.

[80] E. HAUS AND M. PROCESI, *KAM for beating solutions of the quintic NLS*, Comm. Math. Phys., 354 (2017), pp. 1101–1132.

[81] J. HOLMER AND M. ZWORSKI, *Soliton Interaction with Slowly Varying Potentials*, International Mathematics Research Notices, (2008).

[82] R. HORSIN, *Comportement en temps long d'équations de type Vlasov : études mathématiques et numériques*, PhD thesis, Rennes 1, 2017.

[83] L. I. IGNAT AND E. ZUAZUA, *Dispersive properties of numerical schemes for nonlinear Schrödinger equations*, in Foundations of computational mathematics, Santander 2005, vol. 331 of London Math. Soc. Lecture Note Ser., Cambridge Univ. Press, Cambridge, 2006, pp. 181–207.

[84] M. JENKINSON AND M. I. WEINSTEIN, *Onsite and offsite bound states of the discrete nonlinear Schrödinger equation and the Peierls-Nabarro barrier*, Nonlinearity, 29 (2016), pp. 27–86.

[85] H. KANAZAWA, T. MATSUO, AND T. YAGUCHI, *A conservative compact finite difference scheme for the KdV equation*, JSIAM Lett., 4 (2012), pp. 5–8.

[86] D. KARP AND E. PRILEPKINA, *Hypergeometric functions as generalized Stieltjes transforms*, Journal of Mathematical Analysis and Applications, 393 (2012), pp. 348 – 359.

[87] P. G. KEVREKIDIS, *The discrete nonlinear Schrödinger equation*, vol. 232 of Springer Tracts in Modern Physics, Springer-Verlag, Berlin, 2009. Mathematical analysis, numerical computations and physical perspectives, Edited by Kevrekidis and with contributions by Ricardo Carretero-González, Alan R. Champneys, Jesús Cuevas, Sergey V. Dmitriev, Dimitri J. Frantzeskakis, Ying-Ji He, Q. Enam Hoq, Avinash Khare, Kody J. H. Law, Boris A. Malomed, Thomas R. O. Melvin, Faustino Palmero, Mason A. Porter, Vassilis M. Rothos, Atanas Stefanov and Hadi Susanto.

[88] P. KOSELEFF, *Exhaustive search of symplectic integrators using computer algebra*, Fields Inst. Commun, 10 (1996), pp. 103–120.

[89] S. B. KUKSIN AND J. PÖSCHEL, *Invariant cantor manifolds of quasi-periodic oscillations for a nonlinear schrödinger equation*, Annals of Math., 143 (1996), pp. 149–179.

[90] L. LANDAU, *On the vibrations of the electronic plasma*, J.Phys.(USSR), 10 (1946), pp. 25–34.

[91] M. MALO, *Modèles mathématiques de type HMF : stabilité méthodes numériques autour d'états stationnaires*, PhD thesis, ENS Rennes, 2018.

[92] J. MERRIEN, *Approximation de problèmes faiblement non linéaires par des schémas de différences finies superconvergents*, PhD thesis, Rennes 1, 1985.

[93] C. MOUHOT AND C. VILLANI, *On Landau damping*, Acta Math., 207 (2011), pp. 29–201.

[94] F. NICOLA AND L. RODINO, *Global pseudo-differential calculus on Euclidean spaces*, vol. 4 of Pseudo-Differential Operators. Theory and Applications, Birkhäuser Verlag, Basel, 2010.

[95] B. NOUMEROV, *A method of extrapolation of perturbations*, Monthly Notices of the Royal Astronomical Society, 84 (1924), p. 592.

[96] O. F. OXTOBY AND I. V. BARASHENKOV, *Moving solitons in the discrete nonlinear Schrödinger equation*, Phys. Rev. E (3), 76 (2007), pp. 036603, 18.

[97] A. W. PAETH, *A fast algorithm for general raster rotation*, in Graphics Gems, A. S. Glassner, ed., Academic Press Professional, Inc., San Diego, CA, USA, 1990, pp. 179–195.

[98] J. PEYRIÈRE, *Convolution, séries et intégrales de Fourier*, Références Sciences, Ellipses, Paris, 2012.

[99] H. POINCARÉ, *Les Méthodes Nouvelles de la Mécanique Céleste*, Tome I, Gauthier-Villars, Paris, 1892.

[100] H. S. PRICE, *Monotone and oscillation matrices applied to finite difference approximations*, Math. Comp., 22 (1968), pp. 489–516.

[101] Z. SEDLÁČEK AND L. NOCERA, *Second-order oscillations of a Vlasov–Poisson plasma in Fourier-transformed velocity space*, Journal of Plasma Physics, 48(3) (1992), pp. 367–389.

[102] A. B. SHIDLOVSKII, *Transcendental numbers*, vol. 12 of De Gruyter Studies in Mathematics, Walter de Gruyter & Co., Berlin, 1989. Translated from the Russian by Neal Koblitz, With a foreword by W. Dale Brownawell.

[103] M. M. SHOUCRI AND R. R. GAGNÉ, *A multistep technique for the numerical solution of a two-dimensional Vlasov equation*, Journal of Computational Physics, 23 (1977), pp. 243–262.

[104] V. SOHINGER, *Bounds on the growth of high Sobolev norms of solutions to nonlinear Schrödinger equations on* $\mathbb{R}$, Indiana Univ. Math. J., 60 (2011), pp. 1487–1516.

[105] ——, *Bounds on the growth of high sobolev norms of solutions to nonlinear schrödinger equations on* $\mathbb{S}^1$, Differential and Integral Equations, 24 (2011), pp. 653–718.

[106] E. SONNENDRÜCKER, *Numerical Methods for the Vlasov-Maxwell equations*. book in preparation, 2016.

[107] E. SONNENDRÜCKER, J. ROCHE, AND A. GHIZZO, *The semi- lagrangian method for the numerical resolution of the vlasov equation*, Journal of Computational Physics, 149 (1999), pp. 201–220.

[108] G. STAFFILANI, *On the growth of high Sobolev norms of solutions for KdV and Schrödinger equations*, Duke Math. J., 86 (1997), pp. 109–142.

[109] A. STEFANOV AND P. G. KEVREKIDIS, *Asymptotic behaviour of small solutions for the discrete nonlinear Schrödinger and Klein-Gordon equations*, Nonlinearity, 18 (2005), pp. 1841–1857.

[110] J. C. STRIKWERDA, *Finite difference schemes and partial differential equations*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2004.

[111] M. SUZUKI, *Fractal decomposition of exponential operators with applications to many-body theories and monte carlo simulations*, Phys. Lett. A., 146 (1990), pp. 319–323.

[112] A. TANAKA, *A rotation method for raster image using skew transformation*, Proc. IEEE Conf. Corn.put. Vision and Pattern Rec, (1986), pp. 272–277.

[113] E. WEIBEL, *Spontaneously growing transverse waves in a plasma due to an anisotropic velocity distribution*, Phys. Rev. Lett. 2, 83 (1959).

[114] M. I. WEINSTEIN, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[115] ——, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–67.

[116] J. WELLING, W. EDDY, AND T. YOUNG, *Rotation of 3d volumes by fourier-interpolated shears*, Graphical Models, 68 (2006), pp. 356–370.

[117] H. YOSHIDA, *Construction of higher order symplectic integrators*, Physics Letters A, 150 (1990), pp. 262–268.

[118] V. E. ZAKHAROV AND A. B. SHABAT, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Journal of Experimental and Theoretical Physics, 34 (1972), pp. 62–69.

[119] M. ZERROUKAT, N. WOOD, AND A. STANIFORTH, *The parabolic spline method (psm) for conservative transport problems*, International Journal for Numerical Methods in Fluids, 51 (2006), pp. 1297–1318.

**Titre : Étude de quelques perturbations d'équations riches en symétries : résonances et stabilités**

**Mot clés :** équations de Schrödinger non linéaires, formes normales rationnelles, ondes progressives discrètes, fréquences de Best, schémas aux différences finies compacts

**Résumé :** Cette thèse est un recueil de constructions et de résultats variés autour de problèmes de résonances et de stabilités. Premièrement, on s'intéresse à la conception et à l'analyse de méthodes numériques pour des problèmes académiques tels que le problème de Dirichlet sur un segment ou l'équation de transport associée à une rotation du plan. Ensuite, on étend l'analyse linéaire classique des équations de Vlasov-Poisson autour d'états d'équilibre homogènes pour décrire des phénomènes multidimensionnels et non linéaires. Enfin, une large partie est consacrée à l'étude d'équations de Schrödinger non linéaires en dimension $1$. D'une part, on étudie l'impact d'une semi-discrétisation naturelle sur les ondes solitaires progressives et la croissance des normes de Sobolev. D'autre part, on développe une nouvelle famille de formes normales permettant de décrire la dynamique des petites solutions régulières pendant des temps très longs.

**Title : Study of some pertubation of equations with many symetries : resonances and stabilities**

**Keywords :** nonlinear Schrödinger equations, rational normal forms, discrete traveling waves, Best frequency, compact finite difference schemes

**Abstract :** This manuscript deals with many problems about resonance and stability. First, we design and analyse numerical methods for academic problems like the Dirichlet problem on a segment line or the transport equation associated with a two dimensional rotation. Then, we extend the classical linear analysis of Vlasov-Poisson equations near homogeneous equilibria to describe nonlinear and multidimensional phenomena. Finally, a large part of this thesis is devoted to nonlinear Schrödinger equations in dimension $1$. On the one hand, we study the impact of a natural semi-discretisation on the solitary traveling waves and on the growth of the high order Sobolev norms. On the other hand, we develop a new family of normal forms to describe the dynamic of small and smooth solutions for very long times.