COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

Thèse de doctorat de

l'Université de Rennes

École Doctorale N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Mathématiques et leurs Interactions*

Par
Yoann Le Hénaff

Modulated particle methods and high orders

Méthodes particulaires modulées et ordres élevés

Thèse présentée et soutenue à Rennes, le 12 juin 2024
Unité de recherche : IRMAR, UMR CNRS 6625

Rapporteurs avant soutenance :

Martin Campos Pinto	Directeur de recherche au Max Planck Institute for Plasma Physics
Bruno Després	Professeur au Laboratoire Jacques-Louis Lions, Université de la Sorbonne

Composition du Jury :

Présidente :	Virginie Ehrlacher	Professeur au CERMICS, Ecole des Ponts Paristech
Examinateurs :	Martin Campos Pinto	Directeur de recherche au Max Planck Institute for Plasma Physics
	Bruno Després	Professeur au Laboratoire Jacques-Louis Lions, Université de la Sorbonne
	Virginie Ehrlacher	Professeur au CERMICS, Ecole des Ponts Paristech
	Katharina Schratz	Professeure au Laboratoire Jacques-Louis Lions, Université de la Sorbonne
	Pierre Vernaz-Gris	Responsable du Pôle Simulations à Cailabs
Dir. de thèse :	Erwan Faou	Directeur de recherche à INRIA Bretagne Atlantique
Co-dir. de thèse :	Nicolas Crouseilles	Directeur de recherche à INRIA Bretagne Atlantique

Remerciements

Mes premiers et plus chaleureux remerciements vont à Erwan, qui a su me guider avec brio au cours des trois dernières années, tant au niveau scientifique que humain. Ta rigueur, ta détermination, tes idées et ta vision de la recherche sont devenus des modèles pour moi, et j'espère pouvoir faire preuve d'autant d'exemplarité scientifique par la suite. Ton encadrement a été composé en grande partie de ta bonne humeur et de ton soutien, qui m'ont souvent redonnés du baume au coeur. Des exemples parfaits de cela sont nos discussions non mathématiques à Cambridge et Angers qui ont été déterminantes pour moi. Nos nombreux échanges dans ton bureau ont aussi été très agréables, et j'en suis toujours ressorti avec de nouvelles idées, une nouvelle vision des choses, ou une détermination remise d'aplomb. Merci.

Nicolas, je ne t'oublie évidemment pas, ta complémentarité avec Erwan a été très bénéfique. Ta présence durant nos points avec lui m'a énormément rassuré au début, et ton point de vue numérique m'a beaucoup apporté, largement plus que ce que tu peux penser. Ça m'a aussi aidé à me rendre compte de ce que pouvait être un chercheur en num', et je t'en remercie. Si je continue dans cette voie, ce n'est pas étranger au fait de t'avoir eu comme directeur pendant trois ans.

Mes remerciements suivants ne peuvent aller à d'autres personnes que les membres du jury. Martin et Bruno, merci énormément d'avoir rapporté ce manuscrit qui contient plusieurs sujets différents et peut-être un peu trop de pages. Néanmoins vous avez fait l'exploit de le rapporter en à peine un mois ! Katarina et Virginie, merci d'avoir accepté de faire partie du jury malgré vos contraintes. Enfin, Pierre, je te remercie d'être présent malgré le fait que tu sois peut-être l'OVNI dans ce jury. Nos discussions ont donné lieu à ma partie préférée de la thèse.

Ce manuscrit a été réalisé presque entièrement dans les murs de l'IRMAR, et outre l'aspect scientifique évident, de nombreuses personnes ont permis sa complétion de façon plus indirecte. Marie-Aude, pour ta bonne humeur tous les jours, ton efficacité sans faille, ainsi que pour avoir été ma réponse habituelle aux questions *"On fait comment ? À qui faut-il demander ? Tu sais comment [quelque chose] fonctionne ?"*, merci. Sandra, Aude et Florian, merci également d'avoir toujours répondu à mes questions administratives et d'avoir été si efficaces.

L'aspect humain de cette aventure a été plus que déterminant, et essentiellement toutes les personnes de l'IRMAR y ont contribué de façon plus ou moins importante. Je tiens à remercier en particulier Léo pour l'ambiance qu'il a mis, et pour avoir instauré le FC IRMAR. Bon, t'es un coach en carton, donc merci surtout à Miguel et Florent d'avoir assumé cette responsabilité. Je n'oublie pas les quelques habitué.e.s du ballon rond : Bachir, Delphine, Stéphane et Épiphane, je sais que j'ai raté de nombreuses passes, tirs, et dribbles, mais j'ai tout de même beaucoup apprécié jouer avec vous.

Le cadre de travail au deuxième étage n'est heureusement pas uniquement l'oeuvre de Léo. Parmi les permanents reponsables de cela, je veux souligner l'impact de Miguel (encore lui, toujours dans les bons coups), Vincent, Louise, Zied et San. Parmi les non-permanents mention spéciale à Théo qui a organisé le Landau presque à lui tout seul, ainsi qu'aux anciens que j'ai pu côtoyer un ou deux ans et qui m'ont donné envie de continuer : Quentin, Grégoire, Alice, Pierre L.B., et Josselin.

Pour l'ambiance globale dans l'IRMAR et les discussions ou les cours que j'ai pu avoir avec eux, je tiens à remercier Benjamin, Nicolas S., et Pierre N. Plus généralement, je remercie tous les membres des équipes d'EDP et d'Analyse Numérique, et de manière encore plus globale toutes les personnes que j'ai pu côtoyer au sein des bâtiments 22 et 23.

Je ne peux pas écrire cette section sans remercier du fond du coeur les supers amis que je me suis fait au cours de ces trois ans. Non seulement j'ai pu faire le guignol avec eux, mais ils m'ont souvent encouragé[1] ! Hugo et François pour tout, mais en particulier les soirées à La Piste, au Fox ou au Tiff, ainsi que les blagues et la bonne humeur. La découverte du mexicain est aussi un souvenir mémorable. Antoine, Ketsia, Nathan, j'espère que cette année ne vous a pas dégouté des maths ou des matheux. Je vous dois, entre autres, mon introduction à l'escalade et une ambiance de bureau réanimée. Vous incarnez déjà l'esprit du bureau du fun. Il y aurait évidemment beaucoup plus de choses à dire mais je vais me cantonner à : Merci à vous cinq ! Je dois aussi remercier tous nos chats, qui sont une source intarissable de sujets de conversation : Umpa Lumpa le chat un peu trop gentil, Teapot et sa queue plumeau, Gramsci la chatastrophe, Noopy, Phoebe, ainsi que Fonzie et Lana.

Mérrriadec mérite sa petite ligne à lui tout seul, en particulier pour les sorties en plein air et les soirées burgers, mais aussi pour avoir fini par m'apprendre quelques petits trucs en LaTeX.

Marie, tu t'es beaucoup plus préoccupé de l'administratif que moi alors qu'on soutient à deux jours d'écart, et si je n'ai pas eu de gros retard c'est en partie grâce à toi. Ton

[1] Sans eux, probablement pas de blagues dans ce manuscrit…

implication, ainsi que celle de Mattia, dans la vie des doctorants se doit également d'être soulignée.

Merci également à mes amis de plus longue date, et tout particulièrement Lulu, Nada, et Clem'. Nos weekends escape-game et billard à Paris ou Rennes ont toujours été très cools, et j'espère pouvoir continuer ça même en étant à l'étranger ! Si je remonte un peu plus loin, je dois admettre que l'INSA a été plus agréable grâce à vous.

En parlant de l'INSA, je dois également mentionner quelques enseignants que j'ai eu et qui m'ont donné envie de continuer à faire des mathématiques. Je pense notamment à Loïc, Marc, James, Mounir, Olivier et Mohamed.

Mes derniers remerciements (mais pas des moindres !) vont à ma famille. Gwen, Kristen et Denia, merci d'avoir été là durant toutes mes études, vous m'avez souvent rappelé sans le savoir qu'il y avait autre chose que les maths.

Titi et Sam, vous m'avez donné plus que le nécessaire pour être un enfant puis un adulte heureux, et je vous en serai toujours reconnaissant. Je vous dois tout, et je vous remercie donc pour tout même si ce n'est pas assez. Il y a peut-être une part de moi inconsciente qui fait des maths grâce au jeu de multiplication de quand j'étais petit…

Abbreviations and Symbols

$\mathbb{N}$	The set of nonnegative integers $\{0, 1, \dots\}$.
$\mathbb{Z}$	The set of integers $\{\dots, -1, 0, 1, \dots\}$.
$\mathbb{R}$	The set of real numbers.
$\mathbb{C}$	The set of complex numbers.
(a, b)	The open interval, consisting of real numbers $a < x < b$, for $a, b \in \mathbb{R} \cup \{\pm\infty\}$.
$[a, b]$	The closed interval, consisting of real numbers $a \leq x \leq b$.
$[\![p, q]\!]$	The set of integers between p and q: $\{k \in \mathbb{Z} : p \leq k \leq q\}$.
$*$	The convolution operator, $(f * g)(x) = \int_{\mathbb{R}^d} f(y)g(x-y)dy$.
i	The imaginary unit such that $i^2 = -1$.
$\mathrm{Re}\,(z)$	The real part of a complex number z.
$\mathrm{Im}\,(z)$	The imaginary part of a complex number z.
$C^q(\Omega)$	The set of continuous functions that are q times differentiable over $\Omega \subset \mathbb{R}^d$, and such that $f^{(q)}$ is continuous on Ω. Here, $f^{(q)}$ denotes the q-th derivative of x.
$\mathbb{L}^p(\Omega)$	The set of functions f such that $\int_\Omega \|f(x)\|^p dx < \infty$, for $\Omega \subset \mathbb{R}^d$, $d \geq 1$.
$W^{s,p}(\Omega)$	For $s \in \mathbb{N}$, this is the set of functions f such that $D^\alpha f(x) \in \mathbb{L}^p(\Omega)$, for $\alpha \in \mathbb{N}^d$ and $\alpha_1 + \dots + \alpha_d < s$. Here, $D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \dots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}$, and the derivatives are taken in the distributional sense. When $s \in \mathbb{R} \setminus \mathbb{N}$, there is a well-defined meaning as well, but we will not use these spaces in this manuscript.
H^s	Other notation for $W^{s,2}$.
∇_z	The usual gradient with respect to the variable z. We may omit the subscript z if the variable is clear from the context.
$\mathcal{F}[f], \hat{f}$	The Fourier transform of $f \in \mathbb{L}^2(\mathbb{R}^d)$, see Section II-5.1.
$\mathcal{F}^{-1}[f]$	The inverse Fourier transform of $f \in \mathbb{L}^2(\mathbb{R}^d)$, see Section II-5.1.
$\mathrm{DFT}\left[\{x_k\}_{k=0}^{N-1}\right]$	The Discrete Fourier transform of a length-N sequence $\{x_k\}_{k=0}^{N-1}$, see Section II-5.2.
FFT	The Fast Fourier Transform, an efficient way of computing the Discrete Fourier Transform

DF	Abbreviation for Dirac-Frenkel.
(O|P)DE(s)	(Ordinary|Partial) Differential Equation(s)
(L|R)HS	(Left|Right)-Hand Side

Table of Contents

I

Introduction

Version Française

Il est généralement d'usage que la partie intitulée "Introduction" d'un manuscrit serve à introduire le sujet de ce dernier. Elle ne servira pas à cela ici.

Non pas par extravagance ou volonté de l'auteur de changer les codes, mais essentiellement car le contenu de ce manuscrit ne s'y prête guère. Il n'y a pas "un" sujet au coeur du manuscrit, mais trois (très) différents. Ces sujets ne sont d'ailleurs pas nécessairement en lien avec des méthodes particulaires modulées – le titre officiel de cette thèse – et c'est pourquoi un titre plus approprié pour ce manuscrit pourrait être

Quelques contributions en Analyse Numérique et Mathématiques Appliquées.

Le fil rouge est l'Analyse Numérique de façon générale. La thèse a été l'occasion de découvrir plusieurs domaines et plusieurs communautés, et cet intérêt pour différents sujets se traduit par des parties décorrélées.

Avant d'entamer le coeur de cette thèse, nous commencerons par faire quelques rappels généraux d'analyse numérique. C'est l'objet de la partie II – Some preliminaries of Numerical Analysis. Les différents chapitres de cette partie aborderons notamment la discrétisation d'un problème continu, le *splitting* en temps, les méthodes spectrales, les Transformées de Fourier (continue, discrète, et rapide), ainsi que le concept de complexité algorithmique. Tous ces éléments sont des parties essentielles de beaucoup de problèmes en analyse numérique. À ce titre, le contenu de chaque chapitre nécessiterait des livres entiers pour pouvoir en expliquer tous les détails. Le but dans cette partie n'est évidemment pas de parler de tout en détail, mais seulement de donner une idée (parfois grossière) des problématiques rencontrées et des solutions apportées. Cette partie II peut être vue comme une tentative de "vulgarisation" de l'analyse numérique. Le lecteur intéressé sera renvoyé aux références pour en apprendre plus.

Si la lecture de la partie II n'a pas effrayé le lecteur, ce dernier aura un choix à faire : lire les autres parties dans l'ordre, ou dans le désordre. En effet, les parties sont décorrélées et elles sont présentées dans ce manuscrit dans un ordre arbitraire [1].

La partie III – The Vlasov-Poisson system traitera du système de Vlasov-Poisson, et plus particulièrement d'une méthode particulaire qui a été développée au cours de la

[1] En réalité, il s'agit de l'ordre chronologique de réalisation de cette thèse, faute de mieux.

thèse. L'équation de Vlasov-Poisson est celle qui m'a donné envie d'essayer la recherche, malgré une connaissance très superficielle de celle-ci. L'idée initiale de la thèse consistait, comme son nom officiel l'indique, à regarder des *solutions modulées*. Par solutions modulées, nous entendons des fonctions dont l'expression est connue et qui dépendent de paramètres inconnus. Si la forme des fonctions est bien choisie, on peut espérer que ces fonctions soient des solutions approchées du problème dès lors que les paramètres réussissent à être identifiés. Il s'est avéré que la forme de ces fonctions modulées est loin d'être facile à identifier pour le système de Vlasov-Poisson, mais creuser la littérature associée a permis d'imaginer une nouvelle méthode numérique. Celle-ci fait en quelque sorte le pont entre les méthodes Semi-Lagrangiennes et particulaires (du type Particle-in-Cell et Cloud-in-Cell). En réalité, cette méthode avait déjà été introduite en 2011 par Barré, Olivetti et Yamaguchi sans preuve de convergence ou d'analyse poussée. On l'a donc étudiée et justifiée en détail. Des exemples numériques sont données pour vérifier son utilisation en pratique. Cet algorithme est simple, et se base sur des briques élémentaires bien connues : splitting en temps, quadratures, intégrateurs numériques simplectiques en temps, transformée de Fourier (non-uniforme), et bien évidemment quelques éléments de la théorie des espaces de Sobolev.

Le premier chapitre de la partie III sera dédié à la présentation de la littérature portant sur le système de Vlasov-Poisson. Nous commencerons par présenter l'origine physique des équations, ainsi que leur importance actuelle notamment dans le cadre du projet ITER. Quelques propriétés relativement basiques du système de Vlasov-Poisson seront ensuite discutées, avant de faire un tour d'horizon des travaux existants. Nous aborderons notamment les grandes familles de méthodes numériques utilisées pour la simulation de l'équation de Vlasov : les méthodes Semi-Lagrangiennes, les méthodes particulaires, et également les méthodes spectrales bien qu'elles soient moins utilisées aujourd'hui qu'il y a quelques décennies. Nous rentrerons ensuite dans le vif du sujet, en présentant dans le chapitre suivant les détails de la méthode numérique puis en obtenant un théorème de convergence. Il est à noter que cette estimation de l'erreur est "facilement" obtenue comme la somme des erreurs des différentes briques qui composent la méthode. Des résultats numériques sont ensuite présentés dans le cadre uni-dimensionnel, puis comparés aux résultats obtenus avec une méthode semi-Lagrangienne. Cette partie sur le système de Vlasov-Poisson s'achèvera sur la preuve du théorème de convergence, et sur une conclusion abordant les limitations de la méthode et les perspectives possibles.

Malheureusement, le but initial de l'étude du système de Vlasov-Poisson n'a pas été atteint. On rappelle que nous cherchions à obtenir des solutions modulées, c'est-à-dire des fonctions ayant une forme connue, et dépendant d'un certain nombre de paramètres que l'on met à jour afin que la fonction approche la solution du système. L'idée de ces solutions modulées a été en partie motivée par les travaux de Faou, Merle et Raphaël, qui construisent des solutions modulées pour l'équation de Schrödinger afin d'obtenir une explosion de la solution en temps infini, dans un cadre théorique. Dans le but de pouvoir mieux comprendre ces solutions modulées et tenter d'appliquer des idées similaires au système de Vlasov-Poisson, je me suis tourné vers ces travaux.

Il s'est avéré que, dans le cas de l'équation de Schrödinger linéaire avec potentiel quadratique – également appelée Oscillateur Harmonique Quantique – il est possible d'obtenir une description exacte et explicite des solutions modulées. Plus exactement, si la condition initiale peut s'écrire sous la forme d'une somme de Gaussiennes, alors on peut chercher la solution comme somme de fonctions Gaussiennes dépendant de certains paramètres. L'évolution des paramètres peut être connue de façon exacte et on obtient ainsi une solution exacte en tout temps. Ce résultat n'est pas nouveau, et est à la base des *Variational Gaussian wavepackets*, déjà beaucoup utilisés en pratique. Cependant, le résultat que nous avons obtenu avec les Gaussiennes s'étend naturellement à toutes les fonctions de Hermite-Gauss. Ainsi, il n'est pas nécessaire de supposer la condition initiale comme s'écrivant sous forme de somme de Gaussiennes. Cela est avantageux car la discrétisation d'une fonction arbitraire en somme de Gaussiennes est un problème difficile à traiter en pratique.

Le travail effectué dans ce cadre est décrit dans la partie IV – The Schrödinger equation. Tout comme pour la partie III, nous commençons par un tour d'horizon de la littérature concernant l'équation de Schrödinger. Nous résumons des résultats théoriques connus, et décrivons des méthodes numériques utilisées pour approcher la solution de l'équation. À la suite de cela, nous présentons la modulation de l'équation de Schrödinger linéaire, pour la base des fonctions de Hermite-Gauss. Cette modulation est *exacte*. Puis, nous considérons l'équation de Schrödinger non linéaire avec terme cubique, également appelée dans certains domaines équation de Gross-Pitaevskii. La nonlinéarité introduite met à mal l'intégration exacte des paramètres, et nous sommes réduits à utiliser une approximation numérique connue sous le nom de *Principe de Dirac-Frenkel*. Il s'agit d'une méthode essentiellement utilisée jusqu'à maintenant dans le cadre linéaire, qui présente des problèmes inhérents et auparavant identifiés. Nos exemples numériques montrent que les soucis identifiés dans le cadre linéaire apparaissent également dans le cadre non linéaire. Nous arrivons à construire des exemples qui évitent de faire apparaître lesdits problèmes, et pour ceux-ci l'approximation numérique est correcte. Nous construisons également des exemples numériques qui permettent à nouveau d'illustrer les soucis du principe de Dirac-Frenkel.

Le contenu de cette partie présente les travaux et résultats qui ont été faits en collaboration avec Erwan Faou et Pierre Raphaël, en partant de leurs travaux théoriques.

La liberté accordée durant ma thèse m'a également permis d'échanger avec des acteurs du secteur privé[2], et un problème de mathématiques appliquées s'est vite dégagé de ces échanges. La partie V – The spectral concentration problem en est le résultat, et peut être lu sans relation aucune avec les deux autres – si ce n'est via le prisme très général de l'analyse numérique et des mathématiques appliquées. Cette partie a été, pour moi, la plus intéréssante d'un point de vue humain et scientifique. Humainement, j'ai eu l'occasion de bloquer sur ce problème durant plusieurs mois[3], d'être amené à découvrir des mathématiques que je ne connaissais pas ou très peu, et de vraiment lutter contre un problème d'apparence très simple...Il a été parfois compliqué de trouver la motivation pour s'attaquer à un problème qui ne présente quasiment pas de prises et qui a été relativement peu étudié par le passé. Mais, et je ne pourrai pas le souligner assez, la motivation d'Erwan a fini par déteindre sur moi quand j'en manquais légèrement, et c'est grâce à lui que cette partie a pu voir le jour. En ce qui concerne le point de vue scientifique, il s'agit du contenu de la partie V.

Comme parfois en mathématiques, les problématiques compliquées peuvent être formulées de façon simple[4] : étant donnés deux ensemble compacts $D_1, D_2 \subset \mathbb{R}^d$, il s'agit de chercher une fonction qui admet une transformée de Fourier à support dans D_2, et dont la norme $\mathbb{L}^2(D_1)$ est maximale par-rapport à sa norme $\mathbb{L}^2(\mathbb{R}^d)$. En plus d'avoir de nombreuses applications physiques, ce problème admet une particularité mathématique intéressante : une fois qu'une telle fonction est obtenue, on peut chercher une nouvelle fonction solution de ce problème de maximisation en imposant d'être orthogonale à la précédente. En procédant ainsi à l'infini, on obtient une base de $\mathbb{L}^2(\mathbb{R}^d)$, dont les éléments sont fonctions propres d'un certain opérateur, appelé *opérateur de concentration*. Ce sont les valeurs propres associées à ces fonctions propres qui admettent un comportement intéressant : avec la bonne normalisation, les premières valeurs propres sont très proches de 1 tandis que les suivantes sont très proches de 0. L'interprétation physique de ce phénomène est la suivante : si l'on cherche à décomposer une fonction qui admet une transformée de Fourier compacte et qui est concentrée en espace dans un autre ensemble compact, alors il suffit de décomposer cette fonction dans l'ensemble composé des fonctions propres de l'opérateur de concentration, associées aux valeurs propres proches de 1. Les fonctions propres associées aux valeurs propres proches de 0 peuvent être interprétées comme "ne contenant que très peu d'informations". Cela donne donc un seuil naturel et intuitif pour la troncature de cette base, contrairement à beaucoup d'autres bases utilisées en pratique (comme celles de Fourier ou Hermite-Gauss par exemple). Cependant, cette particularité rend la recherche de fonctions propres compliquée au niveau théorique. Qu'à

[2] L'entreprise souhaite rester anonyme pour des raisons de confidentialité.

[3] J'ai arrêté de compter après 6 mois...

[4] L'exemple le plus connu étant probablement le problème des $3n + 1$.

cela ne tienne, discrétisons notre opérateur de concentration et approchons les fonctions propres de celui-ci par des vecteurs propres de l'opérateur discrétisé. Enfer et damnation, le problème évoqué précédemment apparaît aussi au niveau numérique !

La partie V – The spectral concentration problem s'ouvre sur une présentation détaillée du problème, suivie d'une explication de la solution obtenue par Landau, Pollak et Slepian dans les années 1960 et 1970. Leur solution est extrêmement élégante, mais ne s'applique que dans un cadre restreint. De nouveau, nous ferons un tour d'horizon de la littérature concernant ce problème, et verrons notamment qu'il n'existe presque pas de solutions – ni même de procédés numériques – portant sur des cas qui s'éloignent du cadre "agréable" pour lequel une solution élégante est connue. Armés d'interprétation physique (en guise de couteau) et d'exemples numériques (en guise d'autre chose 🤡), nous exposerons les problèmes liés à la détermination numérique de ces vecteurs propres. Afin de pouvoir regarder des contextes qui s'éloignent du cadre connu, nous formaliserons la théorie nécessaire à l'étude de l'opérateur de concentration sur des domaines complètement arbitraires. L'intuition physique nous permet ensuite de construire un algorithme qui donne une approximation des vecteurs propres de l'opérateur de concentration discrétisé. Enfin, nous vérifions sur des exemples numériques non étudiés jusqu'à maintenant les résultats obtenus et leur cohérence physique.

La dernière partie de ce manuscrit sera une conclusion rappelant les résultats obtenus dans les parties III, IV et V. Elle sera rapide, puisqu'une conclusion détaillée sera donnée à la fin de chaque partie.

J'espère que ce manuscript sera agréable à lire.

Bonne lecture !

English version

It is generally accepted that the Part named “Introduction” in a manuscript is used to introduce the subject of said manuscript. It is not the case here.

It is neither because the author is crazy nor because he wants to change well-established conventions, but rather because it is not appropriate for this kind of manuscript. Indeed, there is not one unique subject treated in this manuscript, but three (very) different ones. A more appropriate title would be

Some contributions to Numerical Analysis and Applied Mathematics.

The common thread is Numerical Analysis in a global sense. During the PhD journey, I had the occasion of discovering several fields and communities, and my interest for various subjects is expressed via uncorrelated parts in this manuscript.

Before getting into the heart of the manuscript, we give some brief ideas, concepts, and results concerning numerical analysis. This is the topic of Part II – Some preliminaries of Numerical Analysis. The various chapters of this Part will cover the discretization of a continuous problem, *time-splitting*, spectral methods, Fourier transforms (continuous, discrete, and fast), as well as computational complexity. All of these notions are central to many problems in numerical analysis. As such, the content of each chapter would need several books in order to give all the details. The goal of this Part is obviously not to talk about everything in a complete sense, but rather to explain some issues one can face, and a few ideas about how to solve them. Part II can be understood as an attempt to explain numerical analysis in layman’s terms. We refer the interested reader to the references.

If the reading of Part II has not afraid the reader, they will have to make a choice: read all the parts in the order they are written, or in any other order. Indeed, since the parts are uncorrelated, they are presented in this manuscript in the chronological order of the thesis completion, but this is rather arbitrary.

Part III – The Vlasov-Poisson system will be treating the subject of the Vlasov-Poisson system, and more specifically a particle method which has been studied by the author of the manuscript. The Vlasov-Poisson system is perhaps the problem that gave me the will to pursue a PhD, even though I had a very poor knowledge about it. The initial idea of the thesis, as the official name states, was to look for *modulated solutions*. Here, “modulated solutions” means that we are looking for functions with a given form (known *a priori*), that depend on unknown parameters. The idea being, if one is able to know the time evolution of the parameters, we can get an approximate solution at any time by just plugging the value of the parameters into the function. If the form of the function is well chosen, we can hope that the functions depending on the parameters are good approximate solutions. We tried, in vain, to find modulated solution for the Vlasov-Poisson system. It turned out to be more difficult than we first hoped. In the meantime, I started reading the literature about the Vlasov-Poisson system, and an idea for a numerical method came up. It is a kind of link between the Semi-Lagrangian and particle methods (such as Particle-in-Cell and Cloud-in-Cell). To be honest, this method had already been introduced in 2011 by Barré, Olivetti and Yamaguchi, but they just described it briefly without performing any analysis on it. We studied a slight generalization of the method, and analyzed the error term. The algorithm is very simple, and relies on well-known elementary foundations: time-splitting, quadratures, simplectic numerical integrators in time, (nonuniform) Fourier transform, and obviously some elements of the theory of Sobolev spaces.

The first chapter of Part III is dedicated to a general presentation of the Vlasov-Poisson system. We start by presenting the physical and historical origins of the equations, as well as their current importance. This includes a very brief presentation of the ITER project. Some relatively basic properties of the Vlasov-Poisson system are then discussed, before giving an overview of the literature about this topic. We will also present the main families of numerical schemes used for the simulation of the Vlasov equation: Semi-Lagrangian methods, particle methods, and also spectral methods even though they are less used today than a few decades ago. We then proceed to the heart of Part III, by presenting the algorithm from Barré, Olivetti and Yamaguchin and stating a convergence result. The convergence result is intuitive (but cumbersome to show) once we understand that the error term is just the sum of the error terms of each elementary foundation. Some numerical results are then presented in the one-dimensional case, before being compared to the results obtained with a Semi-Lagrangian scheme. This Part about the Vlasov-Poisson system ends with the proof of the convergence result, and with a conclusion discussing the limitations of the methods and its perspectives.

Unfortunately, the initial goal of the study of the Vlasov-Poisson system has not been reached. We recall that we were looking for modulated solutions, or in other words, for functions with a known form and depending on a number of parameters that we can update to approximate the solution. Part of our motivation in looking for such functions lies in the works of Faou, Merle and Raphaël, who studied modulated solutions for the Schrödinger equation in order to obtain infinite-time blow-up of the solutions, in a theoretical framework. As an attempt to better understand their modulated solutions, I started studying their works in more details.

It turned out that, in the case of the linear Schrödinger equation with quadratic potential – also known as Quantum harmonic oscillator – it is possible to obtain an exact and explicit description of the modulated solutions. More specifically, if the initial condition can be written as a sum of Gaussian functions, then it is possible to look for the solution as a sum of Gaussian functions depending on some unknown parameters. Moreover, the time evolution of these parameters can be known exactly. This result is not new, and it is the starting point of the *Variational Gaussian wavepackets*, widely used today. However, the result we obtained with Gaussian functions naturally extends to all Hermite-Gauss functions. Thus, it is not necessary to assume that the initial condition can be written as a sum of Gaussian functions, which is a good thing: indeed, the discretization of an arbitrary function into a sum of Gaussian functions is still a difficult task in practice.

The work done in this framework is given in Part IV – The Schrödinger equation. As in Part III, we start with an overview of the literature treating the Schrödinger equation. We recall some known theoretical results, and also some numerical methods used to approximate the solution. After that, we present the modulation of the linear Schrödinger equation applied to the Hermite-Gauss basis. This modulation is *exact*. Then, we consider the nonlinear Schrödinger equation with cubic terms, also called in some fields the Gross-Pitaveskii equation. The nonlinearity introduced breaks the exact integration of parameters, and we have to resort to using a numerical approximation known as *Dirac-Frenkel principle*. It is a numerical method which has been used mostly in the linear case until now, and which has some known issues. Our numerical examples show that the issues identified in the linear case also appear when considering the nonlinear equation. Moreover, by understanding the cause of these issues, we can build some numerical examples that avoid the issues. For these examples, the numerical approximation is very satisfying. We also build some examples that illustrate situations where the Dirac-Frenkel principle breaks down.

The content of Part IV presents the work and results obtained in collaboration with Erwan Faou and Pierre Rapahël, with their common work as a starting point.

The freedom given to me during my PhD has also enabled me to talk with a company from the private sector[5], and during these discussions a numerical problem of applied mathematics has emerged. The Part V – The spectral concentration problem is the result, and bears no link with the two previous sections – except via the very general prism of numerical analysis and applied mathematics. This Part has been, to me, the most interesting from both the personal and scientific points of view. From the personal point of view, I have been stuck during ages on this problem[6], I worked on mathematics I

[5] The name of the company is not given for confidentiality reasons.

[6] I stopped counting after 6 months...

barely didn't know, and I struggled against a seemingly very simple problem. At times, it was difficult to find the motivation to work on a problem which just felt like a huge unclimbable wall. Fortunately, and I cannot stress this enough, Erwan's motivation always rub off on me, and it is thanks to him that this part of the manuscript is written today. The scientific side of this experience is the content of Part V.

It happens sometimes in mathematics that a difficult problem can be expressed in a very simple manner[7]. This is an example: given two compact sets $D_1, D_2 \subset \mathbb{R}^d$, we are looking for a function which has a compact Fourier transform with support in D_2, and whose $\mathbb{L}^2(D_1)$ norm is maximal with respect to its $\mathbb{L}^2(\mathbb{R}^d)$ norm. In addition to having several physical applications, this problem has an interesting property: once a function satisfying our criterion is obtained, we can look for a new solution with the additional constraint of being orthogonal to the previous one. By doing so an infinite number of times, we get a basis of $\mathbb{L}^2(\mathbb{R}^d)$, of which each element is an eigenfunction of a certain operator, called the *concentration operator*. The interesting property we mentioned is seen when looking at the associated eigenvalues: with the correct normalization, the first eigenvalues are very close to one, then quickly decrease to zero, and an infinite number of them are very close to zero. The physical interpretation of this phenomenon is the following: if we want to decompose a function with a compactly supported Fourier transform and which is concentrated in another compact set, then it suffices to decompose this function into the subset composed of the eigenfunctions of the concentration operator, associated to eigenvalues close to one. The eigenfunctions associated to eigenvalues close to zero can be interpreted as "being able to contain only a very small amount of information". This gives a natural and intuitive threshold on where to truncate the basis, unlike many other bases used in practice (for example, Fourier or Hermite-Gauss). However, this interesting property is also what makes the search for eigenfunctions difficult from the theoretical point of view. As any numerical analyst, if things don't work out theoretically, let's take a look at the numerical results! Damn, the previously mentioned issue also appears numerically in the discretized framework!

Part V – The spectral concentration problem starts with a detailed presentation of the problem, followed by an explanation of the solution obtained by Landau, Pollak and Slepian during the 1960s and 1970s. Their solution is extremely elegant, but only applies to a restricted framework. As in the previous parts, we then give an overview of the (relatively dry) literature about this problem, and we will see that there is almost no solution – nor numerical methods – treating situations very different from the "nice" framework studied by Slepian. Armed with physical interpretation and numerical examples, we explain the cause of the problems observed when trying to obtain eigenvectors

[7] One of the most famous examples being the $3n + 1$ problem.

of the discretized problem. In order to be able to look at situations that differ from the known framework, we formalize the theory needed to study the concentration operator on completely arbitrary domains. The physical intuition then allows us to give an algorithm which yields approximate eigenvectors of the discretized concentration operator. We then check on previously unstudied numerical examples the results obtained and their physical coherence.

The last Part of this manuscript is a conclusion recalling the main results obtained in parts III, IV and V. It is very brief, because a detailed conclusion will be given at the end of each of these parts.

I hope that this manuscript will be delightful to read.

Enjoy!

II

Some preliminaries of Numerical Analysis

This Part is devoted to general methods used in numerical analysis. It is self-contained, and can be read separately from the other chapters. The motivation is to get a grasp of some importants methods or ideas in numerical analysis, that have been used in the making of this thesis. We try to not go into too much detail for now, and only present essential ideas. If the reader is interested in studying these methods into more details, we refer them to the cited works and references therein.

Part II

CHAPTER 1

Representation of continuous problems on a computer

The usual problems we face in analysis and in the context of partial differential equations are of continuous nature: most of the physical problems are modelled as being continuous, and mathematics in the continuous setting are often easier to deal with. It is "easier to deal with" for us, humans, because we see the world as continuous. But even if we understand what these continuous problems mean, their solutions are, most of the time, very difficult to obtain exactly.

If it is not possible to obtain exact solutions, it should be easier to obtain approximate solutions, right? It turns out, it is also difficult to do so in general in the continuous setting for us humans.

If it is not possible to obtain approximate continuous solutions, is it easier to obtain approximate discrete solutions? *Technically* yes, in general. But it means that we, humans, would need to perform many, *many*, **many** (many!) computations by hand, usually the same boring ones. This is where computers come into play. They are wonderful in discrete settings, and love to do repeating operations[1].

Great! So when we have an equation, we just give it to the computer and it will give us a discrete approximate solution. Easy enough, right? … NO. There are several hidden layers of difficulties in the previous sentence.

II-1.1 Implicit hypotheses

II-1.1.1 Discretization of infinite domain

The first difficulty lies in that problems of interest are generally posed on the whole real line $\mathbb{R} = (-\infty, +\infty)$, or the d-dimensional real plane $\mathbb{R}^d = (-\infty, +\infty) \times \ldots \times (-\infty, +\infty)$. An infinite continuous domain must be discretized into an infinite discrete domain, but our computers don't have an infinite memory to hold an infinite number of points. Darn! Thus,

[1] At least mine never complained… 🤡

we need to reduce the number of points to a manageable one…When devising numerical schemes, we usually place ourselves in comfortable situations from the discretization point of view. Examples of "comfortable" situations are given below:

1. If the continuous problem is given on an unbounded domain (e.g. $\mathbb{R}^d$), we try to find the "most significant" finite-volume subset of the unbounded domain. For instance, for a time-dependent equation with a compactly supported initial condition, we can guess (and sometimes check theoretically) that after a small time the initial condition hasn't changed a lot. Hence, for a given time of simulation, we can make a *guesstimate* of how large the support of the solution will become. Depending on the equation at hand, the support of the solution can remain compact and thus one only needs to discretize the compact support. If no clear finite-volume subset can be found, we can try find a "good enough" subset. The typical example is $f(x) = e^{-\frac{x^2}{2}}$, where $|f(x)| < 10^{-16}$ for $|x| \geq 9$. For such a quickly decaying function, we can approximate it by a compactly supported function and use the previous ideas. Note that this value of 10^{-16} is not chosen at random, it is of the order of magnitude of the smallest nonzero value that a computer can represent in a usual floating-point format [2]. Hence, choosing $[-9, 9]$ instead of $\mathbb{R}$ is enough to represent $e^{-\frac{x^2}{2}}$ on a computer for its nonzero values, and then extend it by zero outside on $(-\infty, 9) \cup (9, +\infty)$.
2. If the continuous problem is given on a periodic domain, we discretize the periodic domain and use periodic boundary conditions to "wrap the solution around".

II-1.1.2 From continuous to discrete problem

The second issue is the discretization: how to go from a continuous problem to a discrete version? If I have a problem posed on $[-1, 1]$, and I want to discretize it using 3 points, I have infinitely many possibilities. I can choose my discretization points to be $\{-1, 0, 1\}$, or $\{-1, -0.5, 0\}$, or $\{-2/3, 0, 2/3\}$, or $\{-0.5, 0.99, 1\}$, …Depending on the discretization chosen, the results can greatly vary. The general intuition is that no region of the continuous domain is more interesting than an other, thus a uniform discretization where the points are equally spaced is generally best. There exist cases where the physical nature of the continuous problem imposes that some regions must be discretized in finer details than other regions, but this will be of no interest to us in this work.

[2] More precisely, a double-precision floating-point variable is represented by 64 bits: 1 for the sign, 11 for the exponent, and 52 for the significand. The *significand* is used for representing the decimal part, and $2^{-52} \approx 2.2 \times 10^{-16}$. This value is often called the *machine epsilon*, or *machine precision*.

II-1.1.3 How good is the discretization?

The third difficulty can only be understood once the two previous ones are clear, it is called *convergence*. If an algorithm is *convergent*, it is meant that, when the number of discretization points increases, the approximate solution gets closer to the exact solution evaluated at the discretization point. It may sound a little counter-intuitive, but knowing the exact solution is generally not needed for the study of convergence.

A major part of numerical analysis consists in converting continuous problems into discrete ones, and ensuring the discrete problems yield solutions that are close enough to the exact continuous solutions.

Part II

CHAPTER 2 Time-splitting

The method of *splitting*, sometimes called the *fractional step method*, consists in splitting an equation into simpler parts. The underlying idea is that, even though an equation is made of several simple parts, the full equation may be very difficult to solve. Some early records of fractional step methods can be found in [24] and the references therein for the USSR developments, and in [22] for the Western developments.

We refer to [18, 14, 16] for fairly recent guides on splitting methods.

Let us consider the following ordinary differential equation:

$$y' = Ay + By, \tag{II-2.1}$$

where A, B are some operators, and suppose that the solutions to

$$y' = Ay \tag{II-2.2}$$

and

$$y' = By \tag{II-2.3}$$

are known, or can be approximated easily.

The operators A, B may involve very different phenomena, and the solution to (II-2.1) has to approximate these very different phenomena. It can be difficult to obtain an approximate solution depending on the different phenomena at hand.

Moreover, it can happen that we wish to be able to use different numerical schemes for the A-part and the B-part, because some schemes are better suited than others for some phenomena.

It may seem intuitive that, if one solves (II-2.2) on a timestep Δt, and then (II-2.3) on another timestep Δt, the solution we obtain should be close to the solution of (II-2.1). But how close?

Well, this question can be answered using the Baker-Campbell-Hausdorff formula (see [2] and [14, Section III.4.2]):

$$\exp(tA)\exp(tB) = \exp(tC), \tag{II-2.4}$$

where

$$tC = t(A+B) + \frac{t^2}{2}[A,B] + \frac{t^3}{12}\left([A,[A,B]] + [B,[B,A]]\right) + \mathcal{O}(t^4), \tag{II-2.5}$$

and $[A,B] := AB - BA$ is the commutator of the operators A and B.

The notation $\exp(tA)$ is a shorthand for the evolution operator associated to the ODE (II-2.2). In other words, if (II-2.2) is supplied with the initial condition $y(t_0) = x$, we can write

$$y(t) = \exp((t-t_0)A)x.$$

Then, the LHS of (II-2.4) corresponds to solving (II-2.3) over a time t, and then using this as the new starting point for solving (II-2.2) over a time t.

What the BCH formula (II-2.4)–(II-2.5) states is that solving (II-2.3) over a time t and then solving (II-2.2) over a time t, is approximately the same as solving (II-2.1) over a time t. The error made is of order $\mathcal{O}(t^2)$.

The formula (II-2.4) is usually referred to as the Lie-Trotter splitting, which is of order 1. Another famous splitting method is the Strang splitting, sometimes also called Störmer-Verlet scheme:

$$\exp\left(\frac{t}{2}A\right)\exp tB \exp\left(\frac{t}{2}A\right).$$

It is this time of order 2, which means that this evolution operator is approximately the same as $\exp(t(A+B))$ up to some error term $\mathcal{O}(t^3)$.

The order of the splitting method can be increased, and the easiest way consists in composing a low-order splitting scheme with itself. However, some better properties can be obtained by choosing correctly the coefficients of the splitting. We refer to [8] for a work obtaining order-6 splitting coefficients in the context of solving the Vlasov-Poisson system.

Recently, exact splitting methods have been studied for a certain class of differential operators [1, 5], and their numerical efficiency has been assessed in [6].

Part II

CHAPTER 3 Spectral Methods

Spectral methods consists in expanding the unknown function f in a given basis $\{\psi_j\}_j$:

$$f = \sum_j f_j \psi_j$$

An equation on f can generally be reformulated so that it is now only about expansion coefficients f_j, and this new equation on the coefficients can sometimes be simpler than the original one.

It will be clearer with an example. We consider the one-dimensional linear Schrödinger equation with quadratic potential on $\mathbb{R}_+ \times \mathbb{R}$:

$$i\partial_t u(t,x) + \partial_x^2 u(t,x) - x^2 u(t,x) = 0. \tag{II-3.1}$$

We know that the Hermite functions $\{\psi_n\}_{n\in\mathbb{N}}$ form a basis of $\mathbb{L}^2(\mathbb{R})$, and they satisfy

$$\psi_n(x)'' + (2n+1-x^2)\psi_n(x) = 0.$$

Hence, at each time t, we can decompose the function $u(t,\cdot)$ into the Hermite basis, and we then get

$$u(t,x) = \sum_{n\in\mathbb{N}} c_n(t)\psi_n(x),$$

where c_n are complex coefficients depending on time. By plugging this expansion of u into (II-3.1) and using orthogonality of the basis functions, one obtains:

$$c_n'(t) = -i(2n+1)c_n(t),$$

which can be easily solved to get $c_n(t) = e^{-it(2n+1)}$, and thus

$$u(t,x) = \sum_{n\in\mathbb{N}} c_n(0)e^{-it(2n+1)}\psi_n(x).$$

By using an appropriate basis of functions, the partial differential equation (II-3.1) on u could be solved by solving only ordinary differential equations.

Another example is the following: consider the Poisson equation

$$\Delta_x \Phi(x) = \rho(x) \tag{II-3.2}$$

for $x \in \mathbb{T}^d := (\mathbb{R}/(2\pi\mathbb{Z}))^d$, where ρ is a given zero-average function, and Φ is the unknown. Since the domain is periodic, the Fourier family $\left\{e^{in\cdot x}\right\}_{n\in\mathbb{N}^d}$ is a basis of $\mathbb{L}^2(\mathbb{T}^d)$. Hence, we can write

$$\Phi(x) = \sum_{n\in\mathbb{N}^d} c_n e^{in\cdot x}$$

and

$$\rho(x) = \sum_{n\in\mathbb{N}^d} d_n e^{in\cdot x}.$$

Since $\int_{\mathbb{T}^d} \rho = 0$, we have $d_{(0,\ldots,0)} = 0$. We then get:

$$\Delta_x \Phi(x) = \sum_{n\in\mathbb{N}^d} -c_n |n|^2 e^{in\cdot x} = \rho(x) = \sum_{n\in\mathbb{N}^d} d_n e^{in\cdot x}.$$

By orthogonality of the basis functions, we obtain

$$c_n = \begin{cases} -\frac{d_n}{|n|^2} = -\frac{\int_{\mathbb{T}^d} \rho(z) e^{-in\cdot z} dz}{|n|^2}, & \text{if } n \neq 0, \\ 0, & \text{if } n = 0, \end{cases}$$

and thus

$$\Phi(x) = \sum_{n\in\mathbb{N}^d} -\frac{\int_{\mathbb{T}^d} \rho(z) e^{-in\cdot z} dz}{|n|^2} e^{in\cdot x}.$$

We refer to [21] for more details about spectral methods. Numerically, one has to be cautious because the basis is (generally) countably infinite and thus numerical simulations need to truncate the basis. Depending on the equation considered, the truncation may or may not have a huge numerical importance: for linear equations the truncation often does not cause too much trouble, but for nonlinear equations it generally does. This is due to the fact that nonlinear equations "create modes" and eventually modes that are beyond the truncation.

For instance, [17] used a Hermite expansion on the Vlasov equation and reported that truncating the Hermite basis leads to numerical unstabilities.

Part II

CHAPTER 4 Complexity

In this section we will focus on the computational complexity, and give some details about what is meant when an algorithm is "faster" than an other.

Of course, runnning a given algorithm on a modern-day NASA supercomputer will probably run faster than on the ENIAC[1]. This is not what is meant here.

In order to have a hardware-independent measure of the running-time of an algorithm, we count the number of elementary mathematical operations like additions, substractions, multiplications, and divisions. We assume that these four elementary mathematical operations are equally fast, though this is not true in general. Most of the time, the exact number of elementary operations is not known, so we give only an order of magnitude using the "big-O" notation. The "big-O" notation, generally denoted $\mathcal{O}$, means "of the order of magnitude of".

For example, let $T(N)$ be the total number of operations for an algorithm with a length-N input, and assume $T(N) = \mathcal{O}(C(N))$ for some function C. This means that there exist constants $k_2 > k_1 > 0$ that do not depend on N, such that

$$k_1 C(N) \leq T(N) \leq k_2 C(N)$$

Remark II.1

It is important to note that the complexity depends only on the algorithm used and the length of its input, and that it is not an intrinsic property of the problem considered.

In order to make this notation clearer, we give in the next Section two classical examples and estimate their complexity.

[1] Electronic Numerical Integrator and Computer, the first programmable, electronic, general-purpose digital computer, completed in 1945.

II-4.1 Examples

II-4.1.1 Sum of integers

The most straightforward way to compute the sum of the first n integers starting from 1 consists in going through each one of them and adding them. There are n integers, so it needs $n-1$ additions in total. Thus, its complexity is $\mathcal{O}(n-1) = \mathcal{O}(n)$.

On the other hand, if one uses the famous Gauss relation:

$$\sum_{k=1}^{n} k = \frac{n(n-1)}{2},$$

there are only three operations: a substraction, a multiplication, and a division. This number of three operations does not depend on n, thus its complexity is $\mathcal{O}(1)$.

II-4.1.2 Fibonacci sequence

The recurrence relation defininining the usual Fibonacci sequence is:

$$F_{n+2} = F_{n+1} + F_n, \quad F_0 = 0,\ F_1 = 1.$$

The most straightforward way to compute the n-th element consists in applying the recursive relation. If we let $T(N)$ be the number of additions required to compute the N-th Fibonacci number, we have the following relation:

$$T(N) = T(N-1) + T(N-2) + 1.$$

Indeed, to compute the N-th Fibonacci number we have to compute the two previous Fibonacci numbers and add them together. It is possible to show that

$$T(N) = \frac{1}{\sqrt{5}}\left[(-\varphi_-)\varphi_-^N + (-\varphi_+)\varphi_+^N - \sqrt{5}\right], \quad \varphi_+ = \frac{1+\sqrt{5}}{2},\ \varphi_- = \frac{1-\sqrt{5}}{2}.$$

Asymptotically, we get $T(N) = \mathcal{O}(\varphi_+^N)$.

On the other hand, one can note that, if F_n and F_{n+1} are stored in memory, then F_{n+2} can be computed with only one addition. This means that the total number of additions required to compute $F(N)$ is only $N-1$. In other words, $T(N) = \mathcal{O}(N)$. This example also shows that, sometimes, computational complexity can be improved by increasing memory complexity. However, here we always have to store only 2 values so that the memory complexity is $\mathcal{O}(1)$ (i.e. it does not increase as N increases).

In most algorithms, the total number of operations may depend on the input. We can then talk about a “worst case” complexity, “best case” complexity, and “average” complexity as well. The names are pretty self-explanatory. In general, the computational complexity refers to the asymptotic behavior as $N \to \infty$.

If an algorithm has a length-N input, it is said to be *constant* if its complexity is $\mathcal{O}(1)$, *linear* if its complexity is $\mathcal{O}(N)$, *quadratic* if the complexity is $\mathcal{O}(N^2)$, …

Part II

The Fourier transforms

5 CHAPTER

II-5.1 Continuous Fourier transform

The convention for the Fourier transform used in this manuscript is the following: for $f \in \mathbb{L}^2(\mathbb{R}^d)$, the Fourier transform $\mathcal{F} : \mathbb{L}^2(\mathbb{R}^d) \to \mathbb{L}^2(\mathbb{R}^d)$ is defined by

$$\mathcal{F}[f](\xi) := \int_{\mathbb{R}^d} f(x) e^{-i\xi \cdot x} dx. \tag{II-5.1}$$

We may use the shorthand $\hat{f}$ to denote $\mathcal{F}[f]$. For $g \in \mathbb{L}^2(\mathbb{R}^d)$, the inverse Fourier transform $\mathcal{F}^{-1} : \mathbb{L}^2(\mathbb{R}^d) \to \mathbb{L}^2(\mathbb{R}^d)$ is given by

$$\mathcal{F}^{-1}[g](x) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} g(\xi) e^{ix\cdot\xi} d\xi = \frac{1}{(2\pi)^d} \mathcal{F}[g](-x). \tag{II-5.2}$$

A fundamental theorem in the theory of Fourier integrals is the Plancherel theorem. We give below a statement of the theorem which we adapted to our Fourier convention. It can also be found in [19, Section IX.2] or [23, 9].

Theorem II.1: Plancherel

The Fourier transform $\mathcal{F} : \mathbb{L}^2(\mathbb{R}^d) \to \mathbb{L}^2(\mathbb{R}^d)$ defined by (II-5.1) is invertible, and its inverse $\mathcal{F}^{-1}$ is given by (II-5.2). We have,

$$\|\mathcal{F}[f]\|^2_{\mathbb{L}^2(\mathbb{R}^d)} = (2\pi)^d \|f\|^2_{\mathbb{L}^2(\mathbb{R}^d)},$$

and

$$\left\|\mathcal{F}^{-1}[f]\right\|^2_{\mathbb{L}^2(\mathbb{R}^d)} = (2\pi)^{-d} \|f\|^2_{\mathbb{L}^2(\mathbb{R}^d)}.$$

Moreover, for $f, g \in \mathbb{L}^2(\mathbb{R}^d)$,

$$(2\pi)^d \int_{\mathbb{R}^d} f(x)\overline{g(x)} dx = \int_{\mathbb{R}^d} \mathcal{F}[f](\xi)\overline{\mathcal{F}[g](\xi)} d\xi \tag{II-5.3}$$

Remark II.2

Relation (II-5.3) is sometimes called the *Parseval identity.*

Proof. For a proof of the first part of the theorem, see [23, 19, 9]. For the part about $\mathbb{L}^2$ inner products, it suffices to apply the polarization identity

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

to f and g in order to obtain

$$(2\pi)^d \mathrm{Re}\left(\int_{\mathbb{R}^d} f(x)\overline{g}(x)dx\right) = \mathrm{Re}\left(\int_{\mathbb{R}^d} \mathcal{F}[f](\xi)\overline{\mathcal{F}[g](\xi)}d\xi\right),$$

and to f and ig in order to obtain

$$(2\pi)^d \mathrm{Im}\left(\int_{\mathbb{R}^d} f(x)\overline{g}(x)dx\right) = \mathrm{Im}\left(\int_{\mathbb{R}^d} \mathcal{F}[f](\xi)\overline{\mathcal{F}[g](\xi)}d\xi\right).$$

Here, Re and Im denote respectively the real and imaginary parts. □

Moreover, [9] gives the following proposition:

Proposition II.1

Let $\phi, \psi \in \mathbb{L}^2(\mathbb{R}^d)$. The Fourier transform $\mathcal{F}$ defined by (II-5.1) enjoys the following properties:

1. $\int_{\mathbb{R}^d} \widehat{\phi}\psi = \int_{\mathbb{R}^d} \phi\widehat{\psi}$;
2. $\mathcal{F}(\phi * \psi) = \mathcal{F}(\phi)\mathcal{F}(\psi)$;
3. $(2\pi)^d\mathcal{F}(\phi\psi) = \mathcal{F}(\phi) * \mathcal{F}(\psi)$.

II-5.2 Discrete Fourier transform (DFT)

The discrete Fourier transform (DFT) of a sequence $\{x_n\}_{n=0}^{N-1}$ of N complex numbers is given by

$$\mathrm{DFT}\left[\{x_n\}_{n=0}^{N-1}\right](k) := \sum_{n=0}^{N-1} x_n e^{-2i\pi\frac{k}{N}n}.$$

The inverse Discrete Fourier transform (IDFT) of a sequence $\{X_k\}_{k=0}^{N-1}$ of N complex numbers is given by

$$\mathrm{IDFT}\left[\{X_k\}_{k=0}^{N-1}\right](n) := \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{2i\pi \frac{n}{N}k}.$$

Note that the DFT of a sequence $\{x_n\}_{n=0}^{N-1}$ can be represented via a matrix-vector product, with the matrix $\mathbf{W} \in M_N(\mathbb{C})$ being defined component-wise by

$$\mathbf{W}_{k,n} := e^{-2i\pi \frac{k}{N}n}$$

and the vector being naturally

$$x := \begin{pmatrix} x_0 \\ \vdots \\ x_{N-1} \end{pmatrix}.$$

Then,

$$\mathrm{DFT}\left[\{x_n\}_{n=0}^{N-1}\right](k) = (Wx)_k .$$

We refer to [7] for more details.

II-5.3 Link between the continuous and discrete Fourier transforms

Let us give some relations between the continuous and discrete Fourier transforms. Because of the conventions used, the relations make more sense when dealing with their inverses. This will simply allow us to get rid of some normalization constant.

We consider the one-dimensional framework for simplicity, but all the relations can be extended to an arbitrary dimension $d \geq 1$.

Let $f \in \mathbb{L}^2(\mathbb{R})$, and consider a uniform discretization $\{\xi_k := \frac{k}{N+1},\ k = 0, \dots, N\}$ of the interval $[0, 1]$, composed of $N + 1$ points. We assume N is odd for simplicity.

The continuous inverse Fourier transform of f is given by

$$\mathcal{F}^{-1}[f](x) = \frac{1}{2\pi}\int_{\mathbb{R}} f(\xi)e^{ix\xi}d\xi$$

If we assume f to be negligible outside $[-1/2, 1/2]$, we can consider $\tilde{f}$ the periodic extension of $f_{|[-\pi,\pi]}$ on $\mathbb{R}$ and get:

$$\begin{aligned}\mathcal{F}^{-1}[f](x) &\approx \frac{1}{2\pi}\int_{-1/2}^{1/2} f(\xi)e^{ix\xi}d\xi \\ &= \frac{1}{2\pi}\int_{0}^{1/2} f(\xi)e^{ix\xi}d\xi + \frac{1}{2\pi}\int_{-1/2}^{0} \tilde{f}(2\pi+\xi)e^{ix\xi}d\xi \\ &= \frac{1}{2\pi}\int_{0}^{1/2} f(\xi)e^{ix\xi}d\xi + \frac{1}{2\pi}\int_{1/2}^{1} \tilde{f}(\xi)e^{ix(\xi-1)}d\xi.\end{aligned}$$

For $n \in \mathbb{N}$,

$$\begin{aligned}\mathcal{F}^{-1}[f](2\pi n) &\approx \frac{1}{2\pi}\int_{0}^{1} \tilde{f}(\xi)e^{2i\pi n\xi}d\xi \\ &\approx \frac{1}{2\pi(N+1)}\sum_{k=0}^{N} \tilde{f}(\xi_k)\exp(2i\pi n\xi_k) \\ &\approx \frac{1}{2\pi(N+1)}\sum_{k=0}^{N} \tilde{f}(\xi_k)\exp\left(2i\pi n\frac{k}{N+1}\right). \end{aligned} \tag{II-5.4}$$

Let us now compute the IDFT of the sequence $\{\tilde{f}(\xi_k)\}_{k=0}^{N}$:

$$\text{IDFT}\left[\{\tilde{f}(\xi_k)\}_{k=0}^{N}\right](n) = \frac{1}{N+1}\sum_{k=0}^{N} f(\xi_k)\exp\left(2i\pi n\frac{k}{N+1}\right). \tag{II-5.5}$$

By comparing equations (II-5.4) and (II-5.5), we get that

$$\text{IDFT}\left[\{\tilde{f}(\xi_k)\}_{k=0}^{N}\right](n) \approx 2\pi\mathcal{F}^{-1}[f]\,(2\pi n)\,. \tag{II-5.6}$$

Letting $\xi_k = \frac{k}{N+1}$ for $k \in \mathbb{N}$, we obtain

$$\text{IDFT}\left[\{f(\xi_k)\}_{k=-\frac{N+1}{2}}^{\frac{N-1}{2}}\right](n) \approx 2\pi\mathcal{F}^{-1}[f]\,(2\pi n)\,. \tag{II-5.7}$$

This means that, for a periodic function f, the (inverse) Discrete Fourier transform converges to the (inverse) Continuous Fourier transform as $N \to \infty$.

II-5.3.1 Nonuniform discretization

There exist some occasions where the location of the points used for the discretization of the integral (II.1) are imposed and cannot be chosen evenly spaced. In this case, the FFT is not readily applicable, but some procedure were developped in order to adapt the FFT. The main idea consists in performing a Fourier interpolation of the nonuniform data, in order obtain approximate values at some uniform points, and then to apply the FFT to the uniform grid.

See [12] for the first detailed analysis of the Non-Uniform FFT, and [4, 3] for some improvements.

II-5.4 Fast Fourier Transform (FFT)

In order to compute the coefficients of the DFT, an efficient algorithm has been found. Quite surprisingly, the algorithm was already known and used by Gauss [15] but gained popularity primarily during the last sixty years.

The Fast Fourier Transform is one of the most important algorithms of the XX$^{\text{th}}$ century [10]. It is based on a Divide-and-Conquer approach in order to reduce the complexity of computing Fourier coefficients of a length-N signal, from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$. Even more impressive, the memory cost is minimal.

It was first introduced in its modern form by Cooley and Tukey in [11], and they heavily used the fact that N is a composite number. It is most efficient when $N = 2^m$ for some $m \in \mathbb{N}$. More modern formulations of the FFT algorithm allow prime N as well.

We give a quick presentation of the algorithm behind the FFT, following the presentation from [11].

The problem at hand is the computation of the complex Fourier coefficients X of a length-N complex input A. They are given by

$$X(j) = \sum_{k=0}^{N-1} A(k)\omega^{jk}, \quad j \in [\![0, N-1]\!] \tag{II-5.8}$$

where ω is the principal N-th root of unity:

$$\omega = e^{2\pi i/N}.$$

A straightforward computation of (II-5.8) would require N operations for each j, thus N^2 operations in total. By “operation” it is meant a complex multiplication followed by a complex addition.

Now suppose that N is composite, i.e. $N = r_1 r_2$. We can write

$$\begin{aligned} j &= j_2 r_1 + j_1, \qquad && j_1 \in [\![0, r_1 - 1]\!], \quad j_2 \in [\![0, r_2 - 1]\!] \\ k &= k_1 r_2 + k_2, \qquad && k_1 \in [\![0, r_1 - 1]\!], \quad k_2 \in [\![0, r_2 - 1]\!], \end{aligned}$$

and then (II-5.8) becomes

$$X(j_2,j_1) = \sum_{k_1=0}^{r_1-1}\sum_{k_2=0}^{r_2-1} A(k_2,k_1)\omega^{jk_1r_2}\omega^{jk_2}.$$

Note the slight abuse of notation where A and X are now indexed by two indices:

$$A(k_2,k_1) = A(k_1r_2+k_2) \quad \text{and} \quad X(j_2,j_1) = X(j_2r_1+j_1).$$

Moreover, $\omega^{jk_1r_2} = \omega^{j_1k_1r_2}$ since

$$jk_1r_2 = j_2k_1r_1r_2 + j_1k_1r_2 = j_2k_1N + j_1k_1r_2,$$

and $(\omega^N)^{j_2k_1} = 1^{j_2k_1} = 1$.

Then, we can define

$$A_1(j_1,k_2) = \sum_{k_1=0}^{r_1-1} A(k_2,k_1)\omega^{j_1k_1r_2},$$

and obtain

$$X(j_2,j_1) = \sum_{k_2=0}^{r_2-1} A_1(j_1,k_2)\omega^{(j_2r_1+j_1)k_2}.$$

We have $A_1 \in \mathcal{M}_{r_1,r_2}(\mathbb{C})$, and the computation of each component of A_1 requires r_1 operations. This gives $r_1^2r_2 = Nr_1$ operations in total to compute A_1. On the other hand, given A_1, each component of $X \in \mathcal{M}_{r_2,r_1}(\mathbb{C})$ can be computed in r_2 operations. Hence $r_2^2r_1 = Nr_2$ operations are required in total to compute X from A_1.

Thus, the total number of required operations is

$$\underbrace{Nr_1}_{\text{To compute } A_1} + \underbrace{Nr_2}_{\text{To compute } X} = N(r_1+r_2).$$

One can see that, if $N = r_1 \cdots r_m$, one can derive a similar procedure which requires only $N(r_1+\cdots+r_d)$ operations. In this case, if $r_1 = \cdots = r_m =: r$, then the total number of operations is $rN\log_r N$. This is the most interesting when r is small, e.g. when $N = 2^m$.

Part II

CHAPTER 6

Useful results

II-6.1 Grönwall Lemma

Grönwall first stated his famous Lemma in the continuous setting in [13].

Lemma II.1: Grönwall

When, for $x_0 \leq x \leq x_0 + h$, the continuous function $z = z(x)$ satisfies the inequality

$$0 \leq z(x) \leq \int_{x_0}^{x} (Mz(y) + A) dy$$

where the constants M and A are nonnegative, then

$$0 \leq z(x) \leq Ahe^{Mh}, \quad x_0 \leq x \leq x_0 + h.$$

Proof. We start by writing

$$z(x) = e^{M(x-x_0)}\zeta(x),$$

and let the maximum of ζ on $[x_0, x_0 + h]$ occur at x_1. When $x = x_1$, we have

$$\begin{aligned} 0 \leq z(x_1) = e^{M(x_1-x_0)}\zeta(x_1) &\leq \int_{x_0}^{x_1} \left(Me^{M(y-x_0)}\zeta(y) + A\right) dy \\ &\leq \zeta(x_1) \int_{x_0}^{x_1} Me^{M(y-x_0)} dy + \int_{x_0}^{x_1} A dy \\ &\leq \zeta(x_1) \left(e^{M(x_1-x_0)} - 1\right) + A(x_1 - x_0). \end{aligned}$$

By substracting $\zeta(x_1) \left(e^{M(x_1-x_0)} - 1\right)$ from both sides, one obtains

$$\zeta(x_1) \leq Ah.$$

Recall that ζ is a nonnegative function (since z is nonnegative), hence

$$0 \leq \zeta(x_1) \leq Ah.$$

Finally, for $x \in [x_0, x_0 + h]$,

$$0 \leq z(x) \leq \max_{x \in [x_0, x_0+h]} \left(e^{M(x-x_0)} \zeta(x) \right) \leq e^{Mh} \zeta(x_1) \leq Ahe^{Mh}.$$

□

II-6.2 Cauchy-Lipschitz theorem

The following theorem is fundamental in the analysis of differential equations. The statement and his proof given below come from [20, p. 172]:

Theorem II.2: Cauchy-Lipschitz

Let $U \subset \mathbb{R} \times \mathbb{R}^m$ an open set, and $f \in C^1(U; \mathbb{R}^m)$. Then, for all initial data $(t_0, x) \in U$, the differential system

$$y' = f(t, y), \quad y(t_0) = x \tag{II-6.1}$$

has an unique maximal solution.

Proof. If y is solution to (II-6.1), by integrating we get

$$y(t) = x + \int_{t_0}^{t} f(s, y(s)) ds, \quad \forall t.$$

On the other hand, if y solves the above integral equation and f is continuous, then y is differentiable and solves (II-6.1).

Suppose the time interval I is a compact interval of length $l > 0$, then $y : I \to \mathbb{R}$ is bounded, thus $f(I, y(I)) \subset B(0, M)$ for some $M > 0$ large enough. We also have $f'(I, y(I)) \subset B(0, k)$ for some $k > 0$ large enough.

Let $\{y_n(t)\}_{n \in \mathbb{N}}$ the sequence of functions built by solving the following differential problems:

$$y_n'(t) = f(t, y_{n-1}(t)), \quad y_n(t_0) = x,$$

with y_0 a constant function equal to x. For $t \in I$, we have

$$\|y_1(t) - y_0(t)\| \leq \left| \int_{t_0}^{t} \|f(s, x)\| ds \right| \leq M|t - t_0|.$$

More generally, we have

$$\begin{aligned}\|y_n(t) - y_{n-1}(t)\| &\leq \left|\int_{t_0}^{t} \|f(s, y_{n-1}(s)) - f(s, y_{n-2}(s))\| ds\right| \\ &\leq k \left|\int_{t_0}^{t} \|y_{n-1}(s) - y_{n-2}(s)\| ds\right|,\end{aligned}$$

hence by induction,

$$\|y_n(t) - y_{n-1}(t)\| \leq M k^{n-1} \frac{|t - t_0|^n}{n!} \leq \frac{M}{k} \frac{(kl)^n}{n!}.$$

Thus,

$$\sum_{n \geq 1} \|y_n(t) - y_{n-1}(t)\| \leq \sum_{n \geq 1} \frac{M}{k} \frac{(kl)^n}{n!} = \frac{M}{k} \left(e^{kl} - 1\right),$$

i.e. the series $\sum_n (y_n(t) - y_{n-1}(t))$ converges normally for all $t \in I$. This implies the uniform convergence on I of the sequence of partial sums $\left\{\sum_{n \geq 1} (y_n - y_{n-1}) = y_n - x\right\}_{n \geq 1}$. We denote $y := \lim_{n \to \infty} y_n$. Using the fact that f' is continuous, we also have $f(s, y_n(s)) \to f(s, y(s))$. Hence, owing to the dominated convergence theorem,

$$\begin{aligned}y(t) = \lim_{n \to \infty} y_n(t) &= \lim_{n \to \infty} \left(x + \int_{t_0}^{t} f(s, y_n(s)) ds\right) \\ &= x + \lim_{n \to \infty} \int_{t_0}^{t} f(s, y_n(s)) ds = x + \int_{t_0}^{t} \lim_{n \to \infty} f(s, y_n(s)) ds \\ &= x + \int_{t_0}^{t} f(s, y(s)) ds.\end{aligned}$$

We deduce that y solves (II-6.1).

Let us now turn to the uniqueness: let y, z two solutions of (II-6.1), then

$$y(t) - z(t) = \int_{t_0}^{t} \left[f(s, y(s)) - f(s, z(s))\right] ds,$$

hence

$$\|y(t) - z(t)\| \leq k \left|\int_{t_0}^{t} \|y(s) - z(s)\| ds\right|,$$

and by induction

$$\|y(t) - z(t)\| \leq C k^n \frac{|t - t_0|^n}{n!},$$

where $C := \max_{t \in I} \|y(t) - z(t)\|$. By letting $n \to \infty$, we obtain $y(t) = z(t)$ for all $t \in I$.

Finally, let us extend the results to an arbitrary interval I. It can be written $I = \cup_j I_j$, with increasing compact intervals $I_0 \subset I_1 \subset I_2 \subset \ldots$, all containing the initial point t_0. We denote y_j the unique solution obtained on the compact time interval I_j.

If $y(t)$ is a solution of (II-6.1) on I, then y and y_j must coincide on I_j, according to the uniqueness on I_j. We define $y(t) := y_j(t)$ for all j such that $t \in I_j$, it is again unique and solution to (II-6.1).

The function y is then the unique solution on the arbitrary interval I. □

III

The Vlasov-Poisson system and its numerical simulation

Part III

CHAPTER 1 Introduction

Plasma is one of the four fundamental states of matter. Even though it is the main component of "ordinary matter" in our universe [1], it has been studied only recently. The first account of plasma dates back from 1879 and is due to Crookes [51]. It was then called "radiant matter", a term coined earlier by Faraday in 1816 during one of his thought experiments. It is only in 1928 that Langmuir [104] called this state of matter *plasma*. The term was used as an analogy with blood plasma, as explained by Tonks, a collaborator of Langmuir in 1928 [136]. A simple description of plasma is given in [42]:

> "A plasma is a quasineutral gas of charged and neutral particles which exhibits collective behavior."
>
> Francis F. Chen (1974)

Nowadays, plasmas are used daily, for instance in neon tubes. In the near future, it is also expected that plasma can be used to produce energy efficiently, via controlled fusion. Controlled fusion uses either inertial or magnetic confinement. The latter approach consists in confining low-density plasma using a magnetic field, but for a rather long time. Currently, the most promising configuration for a magnetic confinement device is the *tokamak* – a doughnut-shaped vacuum chamber, see an illustration on Figure III-1.1 – which was first developed by Soviet research in the late 1960s. Due to its promising properties, it is natural that the ITER project [2] also uses it.

[1] "It has often been said that 99% of the matter in the universe is in the plasma state [...]. This estimate may not be very accurate, but it is certainly a reasonable one", Francis Chen (1974, in [42]).

[2] More details are available on the dedicated website https://www.iter.org/.

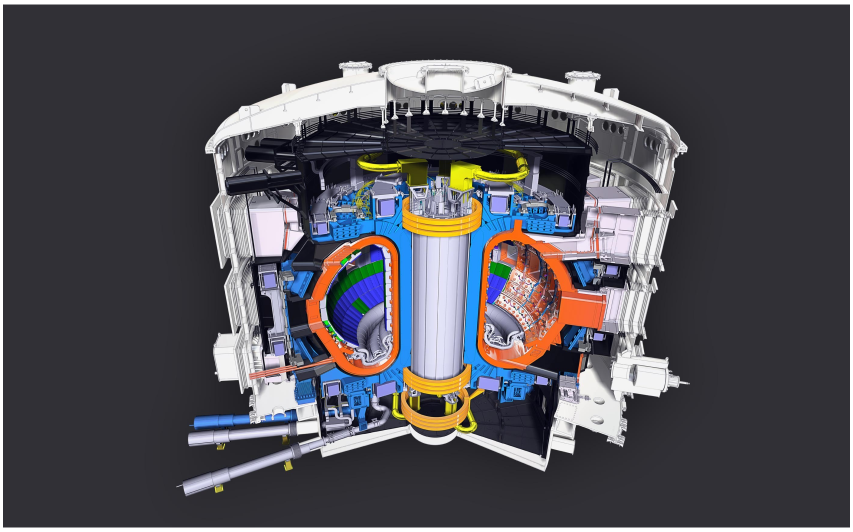


Figure III-1.1 – Cut-view of a tokamak, credits to https://www.iter.org/.

> The tokamak is an experimental machine designed to harness the energy of fusion. Inside a tokamak, the energy produced through the fusion of atoms is absorbed as heat in the walls of the vessel. Just like a conventional power plant, a fusion power plant will use this heat to produce steam and then electricity by way of turbines and generators. [...]
> The term "tokamak" comes to us from a Russian acronym that stands for "toroidal chamber with magnetic coils" [...] As a powerful electrical current is run through the vessel, the gas breaks down electrically, becomes ionized (electrons are stripped from the nuclei) and forms a plasma. As the plasma particles become energized and collide they also begin to heat up. Auxiliary heating methods help to bring the plasma to fusion temperatures (between 150 and 300 million °C). Particles "energized" to such a degree can overcome their natural electromagnetic repulsion on collision to fuse, releasing huge amounts of energy.
>
> Extract from https://www.iter.org/mach/tokamak

The long-term goal of ITER is to prove that producing energy through controlled fusion is possible.

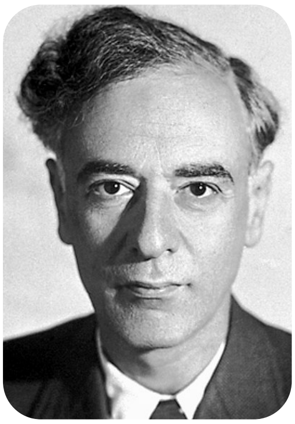
It is obvious that the ITER project could not have seen the light of day if it were not for decades full of countless works about plasma physics, time-consuming numerical simulations, and numerous mathematical theorems.

The theoretical study of plasma gained serious interest after the work of L. D. Landau [101], who studied the vibrations of a plasma under the influence of an electric field. This work gave the now famous name of *Landau damping*, which characterizes the damping and oscillations in plasma for certain initial conditions.

It is common in mathematics that, when studying physics problems, some simplifying assumptions are made in order to make the mathematics easier but still retain most of the physical meaning. Let us explain the framework used for the mathematical study of plasma physics, also described in [73].

From a physical point of view, we can represent any physical system using enough atoms and/or molecules. In the following, we will call these quantities "particles". Mathematically, it is not convenient to deal with too many particles, and a discrete model may be cumbersome. This is why the *kinetic theory* has been developed: to give a "fluid" representation of matter which is actually composed of particles. Here, the term "fluid" means a continuous setting, and it is not understood as a physical fluid (composed of molecules). Kinetic theory gives a statistical approach of the problem at hand. Instead of studying exactly where each particle is and how it evolves with time, we deal with a continuous function $f = f(t, x, v)$. This function depends on time $t \geq 0$, position $x \in \mathbb{R}^3$ and velocity $v \in \mathbb{R}^3$, and it is usually called the *particle density*, or *distribution function*. The probable number of particles in an infinitesimal volume of phase-space of size $dx \times dv$ and centered at (x, v) at time t is given by $f(t, x, v)dxdv$.

The three-dimensional spaces in position and velocity is due to the physical nature of the problems: we live in a three-dimensional world[Reference needed]🤡. Unfortunately, high-dimensional problems are often harder to deal with than low-dimensional problems. This is why one of the first simplifying assumption is generally to lower the dimension of space and/or velocity. Note that there are some physical situations where this reduction of dimension makes sense. The simplest framework consists in considering $x \in \mathbb{R}$ and $v \in \mathbb{R}$. A second simplifying assumption consists in restricting the position to "simple" domain, like the torus. In the following, we will treat the $(d_x + d_v)$-dimensional case, that is $v \in \mathbb{R}^{d_v}$, with either $x \in \mathbb{R}^{d_x}$ or $x \in \mathbb{T}^{d_x}$. Here, $\mathbb{T}^{d_x}$ denotes the d_x-dimensional torus, defined by $\mathbb{T}^{d_x} := (\mathbb{R}/(2\pi\mathbb{Z}))^{d_x}$. When the equality $d_x = d_v$ holds, we will define $d := d_x = d_v$ to make the notations a little lighter.



Lev D. Landau (1908-1968), in 1962. Credit to nobelprize.org.

	Vlasov	MHD
Time scale	Rapid	Slow
Temperature	High	Low
Density	Low	High
Collisions	Ignored	Very Important

Table III-1.1 – Main properties differences between the Vlasov theory and MHD, for completely ionized gases.

The next step consists in developing a theory to describe the completely ionized gases we are considering. Fortunately, there are essentially two distinct theories: the Vlasov theory, and MHD (Magnetohydrodynamics). Their main differences are summarized in Table III-1.1. We will focus in the following on the Vlasov theory, which concerns rapid time scales, high temperatures, low densities, and does not deal with collisions.

One of the main components of the Vlasov theory is the *Vlasov equation*, we detail now briefly how it is obtained. The Liouville equation [73, p. 2] is:

$$\frac{\mathcal{D}f}{\mathcal{D}t} = \text{“material derivative”} = \begin{array}{l}\text{rate of change due}\\ \text{to collisions.}\end{array}$$

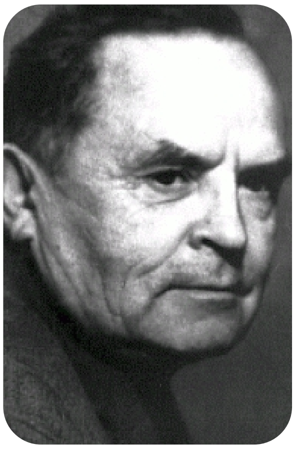
We neglect collisions, thus

$$\frac{\mathcal{D}f}{\mathcal{D}t} = \partial_t f + \frac{dx}{dt}\cdot\nabla_x f + \frac{dv}{dt}\cdot\nabla_v f = 0. \tag{III-1.1}$$

Newton's equations of motion are:

$$\begin{cases}\dfrac{dx}{dt} = v,\\ \dfrac{dv}{dt} = F,\end{cases}$$

where $F(t,x,v)$ is a force applied to the physical system (i.e. our plasma) at time t at position x and velocity v.



Anatoly Vlasov (1908–1975). Credit to ru.wikipedia.org.

Remark III.1

It may be easier conceptually to think of the particle distribution as a discrete set of particles instead of a continuum. In this case, one would consider N particles and apply Newton's equation of motion to each particle.

The theory of electromagnetism tells us that the force F to use in this context is the Lorentz force, given by:

$$F(t,x,v) = q\left(E(t,x,v) + \frac{v}{c} \times B(t,x,v)\right), \tag{III-1.2}$$

where c denotes the speed of light, and q is the electrical charge of the particle. We can compute the self-induced electromagnetic fields E and B within the plasma by using the system of Maxwell equations (see [73, 100], and [108] for the original paper where the equations are derived by Maxwell):

$$\begin{cases} \dfrac{\partial E}{\partial t} - \nabla \times B = -j, & \text{(III-1.3a)} \\ \dfrac{\partial B}{\partial t} + \nabla \times E = 0, & \text{(III-1.3b)} \\ \nabla \cdot E = \rho, & \text{(III-1.3c)} \\ \nabla \cdot B = 0, & \text{(III-1.3d)} \end{cases}$$

where

$$j := \int_{\mathbb{R}^d} fv dv \quad \text{and} \quad \rho := \int_{\mathbb{R}^d} f dv + \rho_0, \tag{III-1.4}$$

and ρ_0 is some constant to be defined later.

By plugging the Lorentz force (III-1.2) into (III-1.1), one gets the Vlasov equation, which describes the evolution of a single species of charged particles under self-consistent fields in the absence of collisions:

$$\begin{aligned} &\partial_t f + v \cdot \nabla_x f + q\,(E + v \times B) \cdot \nabla_v f = 0, && \text{(III-1.5a)} \\ &f(t=0) = f_0, && \text{(III-1.5b)} \end{aligned}$$

where f_0 is a given initial particle distribution at time $t = 0$.

The system of equations (III-1.3)-(III-1.5) is called the Vlasov-Maxwell system, and was first introduced by Anatoly Vlasov in 1938, in a Russian journal [3]. The English translation of one his works where the equation is derived is given in [139].

It is also possible to include external electric forces in E and external magnetic forces in B in the Vlasov-Maxwell system.

[3] It appears now that the earliest issues of this journal are nowhere to be found on the Internet (at least for a non-Russian speaking person), thus we can only date the results based on other works who reference Vlasov's work.

Remark III.2

The Vlasov equation can also be derived by taking the limit as $N \to \infty$ of a system of N particles with pair interactions, see for example [31, 133, 127]. This explains why Newton's equations of motion can still be used when the particle distribution is continuous instead of discrete.

The Vlasov equation could be formulated as well with two (or more!) distributions f_e and f_i, one corresponding to electrons and the other one to positive ions. However, the positive ions are much heavier than electrons, thus move much slower, and a common assumption is generally that they don't move. This allows to consider only one species of particles in the Vlasov equation, as is the case in Equation (III-1.5) .

The particle distribution f changes over time, hence we need its state at initial time. We impose the following: $f(0, x, v) = f_0(x, v)$, where f_0 is called the *initial condition.*

When the magnetic part is neglected, only the electric field remains and the Maxwell equations yield a Poisson equation:

$$\Delta_x \Phi(t, x) = \rho(t, x), E \ \ = \nabla_x \Phi, \tag{III-1.6}$$

where Φ is a scalar potential and ρ is given by (III-1.4). The system composed of equations (III-1.5)-(III-1.6) is called the Vlasov-Poisson system, and writes

$$\begin{aligned}
&\partial_t f + v \cdot \nabla_x f + qE \cdot \nabla_v f = 0,\\
&E = \nabla_x \Phi,\\
&\Delta_x \Phi(t, x) = \rho(t, x),\\
&f(t = 0) = f_0.
\end{aligned} \tag{III-1.7}$$

Remark III.3

The Vlasov-Poisson system can also be used to study stellar dynamics. In fact, it was used by Jeans as early as 1915 [94]. In this case, the Vlasov equation with dimensionless variables is

$$\partial_t f + v \cdot \nabla_x f - E \cdot \nabla_v f = 0,$$

with $f(t = 0) = f_0$. Note the change of sign in front of the field E, compared to (III-1.7) for plasma physics. There is also no constant q here, but this only because we consider dimensionless variables. We refer the reader to [69, 127] for more details about the Vlasov-Poisson system for stellar dynamics.

The rest of this Part is organized as follows. In Chapter III-2 we give some useful results about the Vlasov-Poisson system. Chapter III-3 is dedicated to giving an overview of the literature concerning the Vlasov-Poisson equation, beginning with existence and uniqueness results, and then reviewing some numerical methods. The advantages and issues observed with numerical methods are discussed. Finally, in Chapter III-4, a grid-free particle method will be presented and analyzed. It was first described in [13], and its convergence proved in [105] constitutes the main contribution to this part of the thesis. The convergence result is Theorem III.2, and it is proven in Chapter III-5.

Part III

CHAPTER 2

Properties of the Vlasov-Poisson system

From now on, we consider *dimensionless* variables, which means that the one-species Vlasov equation (III-1.5) now writes

$$\partial_t f + v \cdot \nabla_x f + (E + v \times B) \cdot \nabla_v f = 0. \tag{III-2.1}$$

Moreover, we will focus on the periodic Vlasov-Poisson system, so that we are left with studying the following system of equations:

$$\begin{cases} \partial_t f(t,x,v) + v \cdot \nabla_x f(t,x,v) + E(t,x) \cdot \nabla_v f(t,x,v) = 0, & \text{(III-2.2a)} \\ E(t,x) = \nabla_x \Phi(t,x), \quad \Delta_x \Phi(t,x) = \rho(t,x,), & \text{(III-2.2b)} \\ f(0,x,v) = f_0(x,v), & \text{(III-2.2c)} \end{cases}$$

where $x \in \mathbb{T}_L^{d_x}$ and $\mathbb{T}_L^{d_x} := \mathbb{R}/(L_1\mathbb{Z}) \times \cdots \times \mathbb{R}/(L_{d_x}\mathbb{Z})$ for $L_i > 0, i = 1, \dots, d_x$. We let $v \in \mathbb{R}^{d_v}$. The time variable is $t \geq 0$, and the *charge density* ρ is defined by:

$$\rho(t,x) := \int_{\mathbb{R}^{d_v}} \left(f(t,x,v) - \frac{1}{|\mathbb{T}_L^{d_x}|} \int_{\mathbb{T}_L^{d_x}} f(t,y,v)dy \right) dv. \tag{III-2.3}$$

Usually, $d_x, d_v \in \{1,2,3\}$. In the following, we will simplify the presentation by letting $d := d_x = d_v$, but most of the work is applicable to the case $d_x \neq d_v$.

Remark III.4

In Equation (III-2.3), the quantity ρ_0 that was present in (III-1.4) has been chosen so that $\int \rho(x)dx = 0$. This condition is often called the “neutral background” condition. Physically, it means that we suppose there are as many positive particles as negative particles. In other words, the plasma is electrically neutral (meaning neither positive nor negative) and this makes the Poisson equation well-defined, for smooth f. When the one-species Vlasov equation is considered in plasma, it is generally to model the behavior of electrons. The fact that the plasma is considered electrically neutral means that there is a positively ionized background, namely the protons, and the substracted constant value represents the charge of this background, assumed to be uniform.

The first thing to note about the Vlasov equation (III-2.2a) is that it is a transport equation: it can be written

$$\partial_t f(t,U) + \begin{pmatrix} v \\ E(t,x) \end{pmatrix} \cdot \nabla_U f(t,U) = 0, \tag{III-2.4}$$

by letting $U := (x,v)$.

Definition – Lemma III.1: Characteristics of a transport equation

Consider the following transport equation

$$\partial_t f(t,u) + a(t,u) \cdot \nabla_u f(t,u) = 0, \quad t \in \mathbb{R}_+,\ u \in \mathbb{R}^d \tag{III-2.5}$$

where $a : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$, and with $f(t=0) = f_0$. We assume that a is L-Lipschitz with respect to its second variable, and $|a| \leq A$, $A \in \mathbb{R}_+$. Then the following differential system

$$U'(t) = a(t, U(t)), \quad U(0) = u_0 \tag{III-2.6}$$

is well-defined and has a unique solution $U(t)$ such that $U(0) = u_0$. The mapping $t \mapsto U(t)$ is called the *characteristics* of the transport equation (III-2.5). We shall denote it $U(t; 0, u_0)$ to emphasize the fact that $U(0) = u_0$. Moreover, there is a *time-reversal property*:

$$U(s;t,v) = w \iff U(t;s,w) = v. \tag{III-2.7}$$

The solution to (III-2.5) is given by

$$f(t,u) = f_0(U(0;t,u)). \tag{III-2.8}$$

Finally, if $\operatorname{div}_u(a) = 0$, then the mapping

$$w \mapsto U(s;t,w) \tag{III-2.9}$$

has unit Jacobian.

Proof. The first claim is simply an application of Theorem II.2 – Cauchy-Lipschitz.

For the time-reversal property, by uniqueness we get

$$U(t;s,U(s;r,v)) = U(t;r,v).$$

Thus, by taking $r = t$, we get

$$U(t;s,\cdot)^{-1} = U(s;t,\cdot),$$

which shows (III-2.7).

The characteristics (III-2.6) are now well-defined. Let us now consider the function f defined by

$$f(t,u) := f_0(U(0;t,u)) \qquad \Longleftrightarrow \qquad f(t,U(t;0,u)) = f_0(u),$$

and differentiate it with respect to time t:

$$\begin{aligned} \frac{d}{dt} f(t,U(t;0,u)) &= \frac{d}{dt} f_0(u) = 0 \\ \implies \partial_t f(t,U(t;0,u)) + \frac{d}{dt}U(t;0,u)\cdot\nabla_U f(t,U(t;0,u)) &= 0 \\ \implies \partial_t f(t,U(t;0,u)) + a(t,U(t;0,u))\cdot\nabla_U f(t,U(t;0,u)) &= 0. \end{aligned}$$

In other words,

$$f(t,U(t;0,u)) = f(0,U(0;0,u)) = f_0(u)$$

is a solution to the transport equation (III-2.5).

For the unit Jacobian property, we compute:

$$\frac{d}{dt}\det\frac{\partial U}{\partial u}(t;s,u) = \frac{d}{dt}\det\begin{pmatrix}\partial_1 U(t;s,u) & \partial_2 U(t;s,u) & \dots & \partial_d U(t;s,u)\end{pmatrix}.$$

Since the determinant of a matrix is a multilinear mapping of the columns of the matrix, we get

$$\frac{d}{dt}\det\frac{\partial U}{\partial u}(t;s,u) = \underbrace{\operatorname{div}_u a(t;s,u)}_{=0}\det\frac{\partial U}{\partial u} = 0.$$

Finally, by noting that $\det\frac{\partial U}{\partial u}(t;s,u) = I$ for $t = s$, we get that the Jacobian of $u \mapsto U(t;,s,u)$ is identically equal to one. □

The difficulty of the Vlasov-Poisson system (III-2.2) is that, unlike the transport equation (III-2.5), the electric field E depends on the solution. The method of characteristics could be applied directly if the field E was known, but additional work has to be done in order to prove existence to solutions of the Vlasov-Poisson system (III-2.2). We refer to the existence and uniqueness results given in Section III-3 – Review of the Vlasov-Poisson literature for more details.

Once we know that f_0 satisfies some smoothness conditions, and *a fortiori* the electric field E, we can apply the method of characterics and express the solution f using (III-2.8). A "physical" way of understanding (III-2.8) in the case of the Vlasov equation is given in [84]:

> "If we imagine that $x - v$ phase space is divided into a regular array of infinitesimal cells of volume $d\tau = dxdv$, where $d\tau$ is sufficiently small for not more than one electron to occupy it, then $f(x, v, t)d\tau$ gives the probability that the cell at (x, v) is occupied at time t. Given that there is an electron in the cell at time t, then it follows that there will be one in the cell at (x', v') at time t', where (x', v') are related to (x, v) by the electron equations of motion
>
> $$x' = x + \int_t^{t'} vds, \quad v' = v + \int_t^{t'} Edt.$$
>
> Pursuing this reasoning, we can show generally that
>
> $$f(x', v', t') = f(x, v, t).$$"
>
> Hockney and Eastwood (1988)

Let us come back to the Vlasov equation (III-2.4). Using the change of variables $U = (x, v)$, the characteristics of the Vlasov equation (III-2.2a) are the solutions to the following differential system:

$$\begin{cases} \dfrac{dX(t; s, x, v)}{dt} = V(t; s, x, v), & X(s; s, x, v) = x, \\ \dfrac{dV(t; s, x, v)}{dt} = E(t, X(t; s, x, v)), & V(s; s, x, v) = v. \end{cases} \tag{III-2.10}$$

The notation $X(t; s, x, v)$ (resp. $V(t; s, x, v)$) stands for the position (resp. velocity) component of the flow, starting from (x, v) at time s and evaluated at time t.

The solution to the Vlasov-Poisson system then writes

$$f(t, x, v) = f_0(X(0; t, x, v), V(0; t, x, v)). \tag{III-2.11}$$

III-2.1 Conserved Quantities

Just like many other physical systems, the plasma described by the Vlasov-Poisson system possesses some *conserved quantities*. A "conserved quantity" is a functional $\mathcal{A}$ depending on the solution f and such that $\frac{d}{dt}\mathcal{A}[f](t) = 0$. They are also often called *invariants*.

Some conserved quantities are given in the following Lemma:

Lemma III.1: Conservative quantities of the Vlasov-Poisson system (III-2.2)

Let f be the solution to the Vlasov-Poisson system (III-2.2) under suitable smoothness conditions on f_0. Then, the $\mathbb{L}^p$ norms,

$$\int_{D_x\times\mathbb{R}^{d_v}} |f(t,x,v)|^p dxdv, \quad p \geq 1,$$

the total energy

$$\mathcal{E} := \int_{D_x\times\mathbb{R}^{d_v}} |v|^2 f(t,x,v)dxdv + \int_{\mathbb{R}^d} |E(t,x)|^2 dx$$

and the momentum

$$\mathcal{M}(t) := \int_{D_x\times\mathbb{R}^{d_v}} vf(t,x,v)dxdv,$$

are constant with respect to time. We let D_x denote the x-domain, either $D_x = \mathbb{R}^{d_x}$ or $D_x = \mathbb{T}^{d_x}$.

Proof. For the $\mathbb{L}^p$ norms, we have

$$\begin{aligned}
&pf^{p-1}\partial_t f + pf^{p-1}v\cdot\nabla_x f + pf^{p-1}E\cdot\nabla_v f = 0\\
\Longrightarrow\ &\partial_t(f^p) + v\cdot\nabla_x(f^p) + E\cdot\nabla_v(f^p) = 0\\
\Longrightarrow\ &\int_{D_x\times\mathbb{R}^{d_v}} \partial_t(f^p)dxdv + \int_{\mathbb{R}^{d_v}}\left(v\cdot\int_{D_x}\nabla_x f^p dx\right)dv + \int_{D_x} E\cdot\left(\int_{\mathbb{R}^{d_v}}\nabla_v f^p dv\right)dx = 0,
\end{aligned}$$

where the last implication holds since E does not depend on v. Moreover,

$$\int_{D_x}\nabla_x(f^p) = 0 \quad \text{and} \quad \int_{\mathbb{R}^{d_v}}\nabla_v f^p dv = 0.$$

Thus,

$$\partial_t \int_{D_x \times \mathbb{R}^{d_v}} f^p = 0.$$

For the conservation of momentum, we make use of the momentum density $j(t,x)$ as defined in (III-1.4):

$$j(t,x) = \int_{\mathbb{R}^{d_v}} vf dv.$$

By differentiating with respect to time, we get

$$\partial_t j(t,x) = \int_{\mathbb{R}^{d_v}} v\partial_t f dv.$$

Recall that f is solution to the Vlasov equation, thus

$$\partial_t j(t,x) = \int_{\mathbb{R}^{d_v}} v \left(v \cdot \nabla_x f(t,x,v) + E(t,x) \cdot \nabla_v f(t,x,v) \right) dv.$$

Moreover, for $i = 1, \dots, d$,

$$\begin{aligned}
\int_{\mathbb{R}^{d_v}} v_i \left(E(t,x) \cdot \nabla_v f(t,x,v) \right) dv &= \sum_{j=1}^{d_v} E_j(t,x) \int_{\mathbb{R}^{d_v}} v_i \partial_{v_j} f(t,x,v) dv \\
&= - \sum_{j=1}^{d_v} E_j(t,x) \int_{\mathbb{R}^{d_v}} f(t,x,v) \partial_{v_j} v_i dv \\
&= -E_i(t,x) \int_{\mathbb{R}^{d_v}} f(t,x,v) dv = -E_i(t,x)\rho(t,x),
\end{aligned}$$

so that

$$\partial_t j(t,x) = \int_{\mathbb{R}^{d_v}} v \left(v \cdot \nabla_x f(t,x,v) \right) dv - E(t,x)\rho(t,x).$$

After integration with respect to $x \in D_x$, we obtain

$$\partial_t \mathcal{M}(t) = - \int_{D_x} E(t,x)\rho(t,x) dx.$$

Recall that

$$E(t,x) = \nabla_x \Phi(t,x) \quad \text{and} \quad \rho(t,x) = \Delta_x \Phi(t,x),$$

thus

$$\partial_t M(t) = - \int_{D_x} E(t,x)\rho(t,x) dx = - \int_{D_x} \nabla_x \Phi(t,x) \Delta \Phi(t,x) dx = \int_{D_x} \nabla_x \left(|\nabla_x \Phi|^2 \right) dx = 0,$$

since we assume Φ to be vanishing at infinity (when $D_x = \mathbb{R}^d$), or Φ is periodic (when

$D_x = \mathbb{T}^d$). This shows the conservation of momentum.

For the total energy, we copy the proof from [73, Sect. 4.2.3]. Start by multiplying the Vlasov equation by $|v|^2$ and integrate over $D_x \times \mathbb{R}^{d_v}$. Owing to the fact that $\int_{D_x} \nabla_x f = 0$, we get

$$\implies \partial_t \int_{D_x \times \mathbb{R}^{d_v}} |v|^2 f - 2 \int_{D_x} E \cdot \int_{\mathbb{R}^{d_v}} vf = 0.$$

Then

$$\partial_t \int_{D_x \times \mathbb{R}^{d_v}} |v|^2 f = 2 \int_{D_x} E \cdot j.$$

By integrating the Vlasov equation over v, one gets

$$\partial_t \rho + \nabla_x \cdot j = 0.$$

Moreover,

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_{D_x} |E|^2 dx &= \int_{D_x} E \cdot E_t = \int_{D_x} \nabla_x \Phi \cdot \nabla_x \partial_t \Phi \\ &= -\int_{D_x} \Phi \Delta \Phi = -\int_{D_x} \Phi \partial_t \rho = \int_{D_x} \Phi \nabla \cdot j \\ &= -\int_{D_x} j \cdot \nabla u = -\int_{D_x} j \cdot E. \end{aligned}$$

Finally,

$$\frac{d}{dt}\left(\int_{D_x \times \mathbb{R}^{d_v}} |v|^2 f dx dv + \int_{D_x \times \mathbb{R}^{d_v}} |E|^2 dx \right) = 0$$

□

Numerically, the exact conservation of quantities is challenging for many reasons: aliasing, finite-difference errors, finite-timestep errors, round-off errors...The numerical conservation of invariants is a key challenge that was identified early (see e.g. [84, Section 7.6]). In general, we cannot expect numerical methods to conserve exactly all the continuous invariants since numerical methods induce approximations in one way or another.

However, the size of the approximation is often a very good indicator of the accuracy of the numerical method. This is why the computation of conserved quantities is usually done to monitor the quality of solutions. In [28], Brackbill has underlined the importance of conserving both momentum and energy for the Vlasov-Poisson system.

For long-time simulations, it is sometimes argued that the conserved quantities are more important than the solution itself. Because of this, some authors have proposed schemes that specifically focus on conserving some invariants. One such example is the

Energy-conserving method proposed by Birdsall and Langdon in [22].

Some other authors have recently tried to focus on the geometrical structure of the equation, from which some invariants can be deduced. We can cite for instance GEMPIC, for GEometric Particle In Cell, see [100] and a modified version in [123]. We can also cite [52] who focuses on the Hamiltonian structure.

A comparison of two schemes is given in [86]: one scheme is based on standard finite elements and the other one on structure-preserving geometric finite elements. It is observed that the structure-preserving scheme yields better results.

Review of the Vlasov-Poisson literature

Part III

3 CHAPTER

The first result of global existence and uniqueness for the Vlasov-Poisson system in dimension one was proved by Iordanskii in 1964 [91]. In the multidimensional context, a global existence result for Vlasov-Poisson was stated by Caljub-Simon in 1973 [34].

The existence and uniqueness of classical solutions in the two-dimensional case is proved by Ukai and Okabe in 1978 [137].

The global existence result for weak solutions of the three-dimensional Vlasov equation is due to Arsen'ev [6]. Batt [14] showed global existence of classical solution of the Vlasov-Poisson system for spherically symmetric initial data. Horst [89] extended this result to cylindrically symmetric data. In [58], the global existence for the three-dimensional Vlasov-Poisson system is shown by Degond and Bardos for small initial data. The size restriction is lifted by Pfaffelmoser in [122]. Alternative proofs can be found in [129, 106, 88].

In dimension $d \geq 4$, the existence of global solutions for the Vlasov-Poisson system in stellar dynamics cannot happen. Indeed, Horst [90] constructed a counterexample. But there is at least local existence in dimension $d \geq 3$, as shown by Ukai and Okabe [137].

We have not stated any existence theorem from the cited references yet, because a more recent existence result in particular functional spaces will be given later in Section III-4.1.

III-3.1 Numerical methods

In addition to the theoretical studies performed on the Vlasov-Poisson or Vlasov-Maxwell systems, a large amount of works are dedicated to the numerical simulation of these systems. Numerous schemes have been developed since the 1960s, and still no general-purpose method is satisfactory today: some schemes are focused on long-time simulations, others on small-time simulations with a huge precision, some use the geometrical properties of the equations, some are dedicated to a high-dimensional phase-space, ... Moreover, nearly every one of these schemes have been improved throughout the decades, leading us to two broad families of schemes: the grid-based ones, and the particle ones.

A third type of schemes was initially considered, namely the spectral schemes[1], but the two former have been more widely used since. In addition to these "classical schemes", the search for new algorithms has very recently led people towards "newer approaches". Among them, we can find the machine-learning based methods. One such example is the PINN framework, for Physics-Informed Neural Networks. It was introduced by Raissi *et al.* in [125], and consists in using neural networks that are trained to respect given laws of physics. An application of PINN to the Vlasov-Poisson system can for instance be found in [144]. We refer to [54] for a review about PINN. Some authors have also used machine-learning based algorithms that focus on the Hamiltonian structure of the equations, see for instance [61]. Among the "newer approaches", one can also find quantum-based approaches, such as [142] where an algorithm generally used for quantum computation is applied to the Vlasov-Poisson system. The algorithm essentially consists in discretizing the Vlasov equation using finite differences, writing the distribution function under the form of a product of tensors, and applying a low-rank approximation to each of these tensors. See also [83] which presents a quantum algorithm to solve the Boltzmann-Maxwell six-dimensional equations. In this thesis we will not focus on any of these "newer approaches".

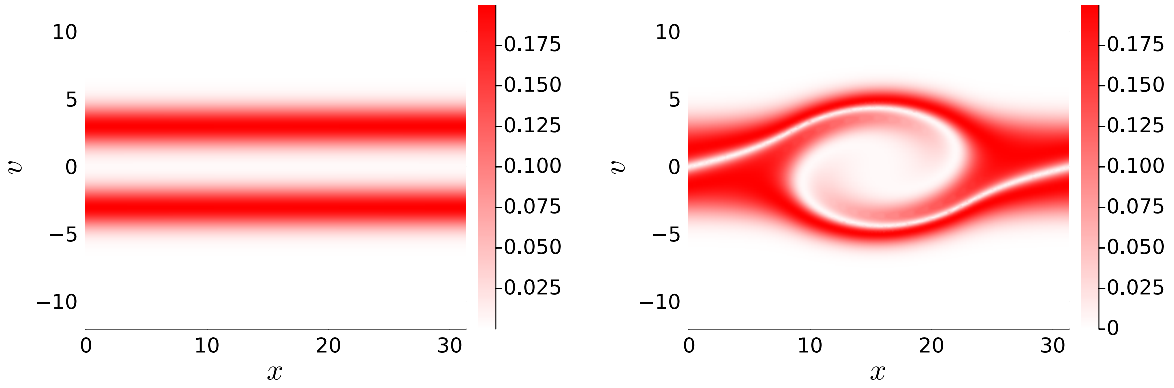
It has been observed very early – as soon as the 1960s and 1970s, see [10, 99, 5, 118, 43, 97] – that one of the main difficulties arising in the numerical simulation of the Vlasov-Poisson system (III-2.2) are the steep gradients appearing in the distribution function. They are called *filaments*, and the phenomenon is called *filamentation*. It is illustrated in Figure III-3.1: starting from two straight beams of electrons (the red parts) in a periodic x-domain at time $t = 0$, the solution after some time $t = T$ has "mixed" the two beams of electrons and a vortex is created.

III-3.1.1 Spectral methods

Spectral methods were studied in detail at the beginning of computer simulations of the Vlasov equation. Early studies, starting from the late 1960s, were concerned only about the one-dimensional case. This is partly due to the computational power available at that time, but also because it is easier to perform numerical analysis (theoretical as well as numerical).

Among the earliest works, Knorr [99] used a Fourier transform in space, and Armstrong [5] considered a Hermite basis in velocity, in addition to a Fourier expansion in space. The Hermite basis had been studied previously for the Boltzmann equation [75]. In these bases, the Vlasov-Poisson system can be expressed as coupled nonlinear ordinary differential equations on the expansion coefficients. Numerically, infinite bases have to be

[1] See Chapter II-3 – Spectral Methods for more details about spectral methods.



(a) Initial condition f_0.

(b) Approximate solution after a time $T = 30$.

Figure III-3.1 – Illustration of filamentation on the Two-Stream Instability example, detailed in Section III-4.4.4.

truncated because we can only deal with a finite number of modes. This approximation is investigated numerically on the Landau damping example[2], and it is verified that the Fourier series converges rapidly so that the truncation is justified. However, the truncation of the Hermite basis was inadequate since the Vlasov-Poisson system creates filamentation with time, i.e. steep velocity gradients, and thus more and more Hermites modes are needed. The Hermite and Fourier bases were also investigated in [95]. The Fourier basis is appropriate if only a few Fourier modes are necessary, and it has also been reported that the truncated Hermite expansion is numerically unstable. The authors of [95] provide a cut-off procedure of the bases, which is reported to work well for the linear Vlasov equation. For the nonlinear Vlasov equation, the quality of the cut-off depends on the case studied.

In [118, p. 1058] , J. Nührenberg sums up quite well the restrictions the above works suffer from:

> “... the difficulties related to the truncation of the Hermite expansion appear to be more serious than those which arise from using a finite interval in y-space. Therefore this method has so far been restricted to a one-dimensional velocity space.”
>
> J. Nührenberg (1971)

The work [118] was developed as a way of bypassing the limitations of [5, 95] related to the use of Fourier transform in x – which can only be used for linear or weakly nonlinear

[2] More details in Section III-4.4 – Numerical Simulations.

problem – and Hermite basis in v. After performing a Fourier transform in v, a second-order scheme explicit in time is used. The derived scheme is stable under the condition $\frac{\Delta t}{\Delta x \Delta v} < \frac{1}{2}$, where Δt is the timestep, and Δx (resp. Δv) denotes the space (resp. velocity) grid stepsize.

In [98], Klimas studies the Fourier-Fourier basis on a modified Vlasov-Poisson system, following an early study of the Fourier-Fourier basis by Knorr [99]. The system is modified in order to take into account the displacement of the space average of the electric field, which is a phenomenon observed in physical situations.

A Fourier-transformed velocity space is investigated by Eliasson in [63, 64, 62] in the context of the Vlasov-Maxwell system, and he uses outflow boundary conditions for the Fourier-transformed velocity space in order to create dissipation and reduce the so-called "recurrence phenomenon" that usually appears in velocity space.

In order to tackle the issues related to the Hermite basis, Holloway [87] proposed using an asymmetrically-weighted Hermite basis for the velocity expansion, and reported a much better stability and better conservation results. In particular, a "usual" (i.e. symmetrically-weighted) Hermite expansion prevents the simultaneous conservation of mass and momentum, but conserves the $\mathbb{L}^2$ norm. An asymmetrically-weighted expansion allows for the conservation of mass, momentum and total energy, but does not conserve the $\mathbb{L}^2$ norm. This can be explained as follows: with a symmetrically-weighted Hermite expansion, the time dependence of the n-th mode is coupled to the $(n+1)$-th mode, which yields the truncation issues we mentioned previously. On the other hand, with an asymmetrically-weighted Hermite expansion, the time dependence of the n-th mode only depends on the $(n-1)$ lower modes, hence the exact time dependence of each Hermite mode can be recovered. However, since the $\mathbb{L}^2$ norm is not conserved with an asymmetrically-weighted Hermite basis, a scheme based on this expansion is not numerically stable. Some numerical tests involving this asymmetrically-weighted Hermite basis are presented in [59, 39]. Recently, Bessemoulin-Chatard and Filbet [21] introduced a weighted $\mathbb{L}^2$ space in which the asymmetrically-weighted Hermite expansion is stable. A convergence result by the same authors is then given in [20]. We can also cite [24] which considers renormalized Hermite coefficients in order to be able to use the classical unweighted $\mathbb{L}^2$ space, and [25] which uses the same idea in the context of the Vlasov-Poisson-Fokker-Planck system.

The transition between spectral and grid-based methods is done in some sense by Galerkine-type methods. They are methods for which we consider a discretization of the phase-space into cells, and in each cell we look for an approximate solution belonging to some functional space or with some given form (often, under the form of a piecewise polynomial). In most applications, the discontinuous Galerkine method is used, and its early applications to Vlasov equations can be found in [80, 44]. The first step of these methods consists in discretizing the phase space in order to obtain an equation of the form $\frac{d}{dt}G_h = R(G_h)$ (following notations of [44]). The second step consists in applying a time integrator (e.g. a Runge-Kutta scheme) in order to get an approximate solution at the next time step. This general procedure is sometimes called *method of lines*. For more details, see [45, 44, 21].

III-3.1.2 Grid-based methods

As their name indicates, these methods are based on a grid discretization of the phase-space (x, v). Perhaps the easiest schemes to think of are the difference schemes, which were proposed by Kellog in [96] and by Nührenberg in [118]. See also [135].

Soon after, in 1976, Cheng and Knorr [43] split the position and velocity part of the equation, using Strang splitting[3] (second order in time). This is possibly the first account of what will be later called a *Backward Semi-Lagrangian* scheme.

> “We have thus reduced the integration of the Vlasov equation to two successive interpolation problems.”
>
> Cheng & Knorr (1976) in [43, p.332]

More details about this scheme can be found at the end of this section. The Backward Semi-Lagrangian scheme from [43] makes use of Fourier interpolation to avoid dissipation from the linear and cubic splines. Then, incremental studies followed [93, 71, 92], studying variations of [43]. Cubic Hermite interpolation has also been used by Nakamura and Yabe [115]. A modern presentation of the semi-Lagrangian scheme is given in [132]. Among other grid-based methods, we can cite the work of Filbet [68], who proposed a finite volume scheme and showed its convergence under a CFL condition.

Unfortunately, the common denominator of all grid-based methods is that they are not suited for long-term simulations. Indeed, filamentation, which often occurs in nonlinear Vlasov simulations, cannot be resolved by grid methods if the filaments are finer than the grid. This makes long-time simulations of small phenomena very hard to do, as underlined in [97, 67].

It has to be noted that the Backward Semi-Lagrangian scheme is still widely used

[3] More details about splitting can be found in Section II-2 – Time-splitting.

today, with improvements that appeared over the years. It has proved its relative efficiency for high-dimensional problems because it is possible – through some splitting – to only solve a sequence of one-dimensional problems, though it remains computationally very expensive to simulate high-dimensional problem. These methods are also particularly interesting from a convergence point of view, as can be seen from the error estimate of Besse and Mehrenberger [19], improved later by Charles, Després and Mehrenberger in [40]. The improved version of the error estimate is:

$$\mathcal{O}\left(\min\left(\frac{\Delta x}{\Delta t}, 1\right)\Delta x^p + \Delta t^2\right).$$

In this case, the error estimate quantifies the absolute difference between exact and approximate solutions, evaluated at grid points. In order to obtain this error estimate in the semi-Lagrangian framework, they considered a Lagrangian interpolation of order $p+1$. We also refer to [18] for an earlier convergence estimate, as well as [36] for a convergence estimate in the case of an adaptive mesh.

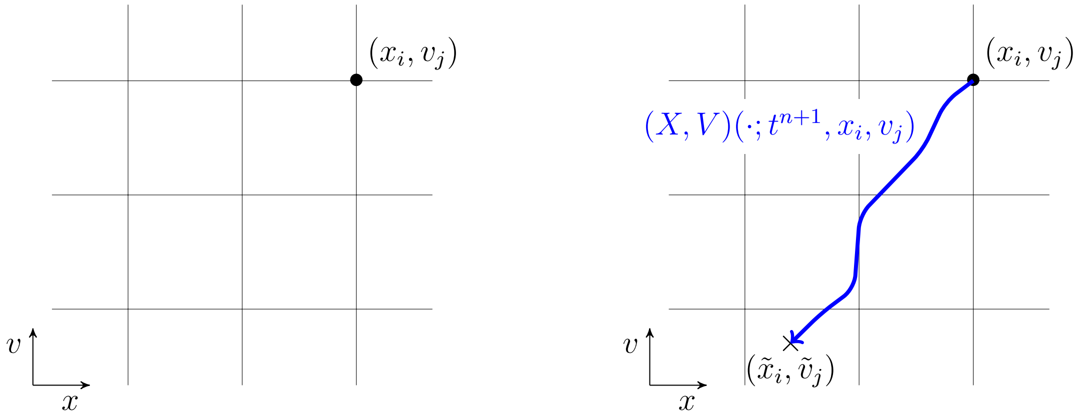
For the good convergence properties, its wide usage, and also because we will use it later in Section III-4, let us give details on how the Backward Semi-Lagrangian scheme works.

Backward Semi-Lagrangian scheme The main idea behind this method is that the unknown function f is constant along the characteristics. More precisely, using the previously defined notations for the characteristics, we make important use of the following relation:

$$f(t_1, x, v) = f(t_0, X(t_0; t_1, x, v), V(t_0; t_1, x, v)). \tag{III-3.1}$$

This relation is the core of Semi-Lagrangian schemes. If $t_0 < t_1$, we talk about a Backward Semi-Lagrangian scheme (BSL), and if $t_1 < t_0$ we call the scheme Forward Semi-Lagrangian (FSL). The forward scheme was introduced by Crouseilles, Respaud and Sonnendrücker [53] in 2009. Let us focus on the case $t_1 > t_0$ since it is the most widely used of the two.

We assume a time-discretization $\{t^n\}_{n=0}^{N_T}$, $N_T \in \mathbb{N}^*$. It is not required that this time discretization is uniform. Suppose that one knows exactly the values of the unknown function f at time t^n on the whole domain $\mathbb{T}^d \times \mathbb{R}^d$, and wishes to get an approximation of the function f at time t^{n+1}. The Backward Semi-Lagrangian scheme consists, for every point (x, v) of the phase-space, in solving the characteristics backward in time. This means solving (III-2.10) when $t < s$. Using (III-3.1), the value of the unknown function f at (x, v) at time t^{n+1} is the same as the value at the point $(X(t^n; t^{n+1}, x, v), V(t^n; t^{n+1}, x, v))$ at


(x_i, v_j)
v
x
(x_i, v_j)
$(X, V)(\cdot; t^{n+1}, x_i, v_j)$
$(\tilde{x}_i, \tilde{v}_j)$
v
x

(a) Discrete phase-space.

(b) Following the characteristics backward in time.

Figure III-3.2 – Grid discretization of the phase space.

time t^n. In other words,

$$f(t^n, X(t^n; t^{n+1}, x, v), V(t^n; t^{n+1}, x, v)) = f(t^{n+1}, x, v).$$

Since the function f at time t^n is supposed to be known exactly, we only need to solve the characteristics in order to get the function f at time t^{n+1} on the whole phase-space.

This simple idea holds for a continuous phase-space, and needs to be adapted to a phase-space discretization. The essential change is that the phase-space is not anymore $\mathbb{T}^d \times \mathbb{R}^d$ but only a finite number of points, the *grid*.

For this *grid discretization*, the true phase-space $\mathbb{T}^d \times \mathbb{R}^d$ is first truncated to one of finite volume, and then the finite-volume phase space is represented using only *points* (or *nodes*). The points in this discrete phase-space can be labelled (x_i, v_j) for $i \in J_x$, $j \in J_v$, where J_x and J_v are finite subsets of $\mathbb{N}^d$. The two-dimensional situation is illustrated in Figure III-3.2a.

Once the phase-space has been discretized, we can apply the ideas mentioned previously. If one knows the characteristics $s \mapsto (X(s; t^{n+1}, x_i, v_j), V(s; t^{n+1}, x_i, v_j))$, they just have to evaluate $f(t^n, X(t^n; t^{n+1}, x_i, v_j), V(t^n; t^{n+1}, x_i, v_j))$ in order to know $f(t^{n+1}, x_i, v_j)$. However, there are in practice two issues: the first one is that we don't know the characteristics exactly, and we can only resort to numerical integration of (III-2.10). The other issue is that we generally don't know the function f at time t^n for the whole continuous phase-space, but only at nodes of the discretized phase-space. This is illustrated in Figure III-3.2b.

To solve the first issue, we can simply suppose that the characteristics are obtained with a sufficient precision so that the error is negligible, or can at least be estimated.

They can even be obtained efficiently if one uses a splitting method[4] to solve for X and V successively in (III-2.10). For the second issue, we generally interpolate the value of $f(t^n, X(t^n; t^{n+1}, x_i, v_j), V(t^n; t^{n+1}, x_i, v_j))$: if $(X(t^n; t^{n+1}, x_i, v_j), V(t^n; t^{n+1}, x_i, v_j))$ falls strictly inside a cell of the discretized phase-space, then we can interpolate the value of

$$f(t^n, X(t^n; t^{n+1}, x_i, v_j), V(t^n; t^{n+1}, x_i, v_j))$$

using the value of the corners of the cell. This interpolation step may seem costly, but it can actually be a series of one-dimensional steps and hence remain relatively efficient even in high-dimensions.

This achieves the detailed presentation of Semi-Lagrangian schemes. Let us now focus on the last main family of numerical methods, namely the particle methods.

III-3.1.3 Particle methods

Let us use an analogy in order to explain the idea behind particle methods. Suppose you want to know how the water in a swimming pool behaves. It is too hard to track every single molecule (which a continuous treatment of the solution would allow you to do), thus you take the water out of the swimming pool and fill it with plastic balls. Besides the incredible fun you'll have in your new ball pit 🤡 , you are now able to track every ball when you swim in it. If, when you increase the number of balls they also get smaller, at one point you'll get a pool filled with almost-liquid plastic. The behavior of this "quasi-fluid" is almost the same as that of the water which is of interest, but it requires many, many, plastic balls. There is then a tradeoff between how many balls you are able to track, and how accurately you want the balls to behave like a liquid.

A particle method is exactly this. The continuous representation of the solution may be too costly from a computational point of view, so it is discretized into small chunks, the *particles*. It is clear that it suffices to discretize the initial condition f_0 to have a particle solution at all times. When the size of the chunks get smaller and smaller, the particle representation of the plasma get closer and closer to the continuous plasma. If there are sufficiently many particles then one is able to get a good grasp of the behavior of the solution by looking at the particles. This is in essence related to the kinetic nature of the Vlasov-Poisson system (III-2.2): the initial condition is assumed to be composed of small indivisible, distinct, particles, instead of a fluid.

This avoids one of the main problems of the previous schemes, namely filamentation, because only the characteristics are used and they do not depend on the velocity gradient of the solution. However, particles methods present other issues. In particular, they are

[4] See Section II-2 – Time-splitting for more details.

generally very noisy and require lots of particle to yield satisfying results.

Early records of particle methods date back from the 1960s, see [32, 55] , though they considered situations simpler than the Vlasov-Poisson system (III-2.2). In these works and a few others – e.g. [56] – the Poisson equation was not solved but instead the Coulomb interaction was used to model interactions between particles. This means that if N particles were used, the complexity [5] was $\mathcal{O}(N^2)$. At that time, the number of particles was limited to about 10^2 to 10^3 due to the computing power available.

The need for faster algorithms has quickly been recognized, and particle-mesh methods were introduced. They allow having a particle method and in the same time computing efficiently the Poisson equation on a grid. The Poisson equation can be efficiently solved using a Discrete Fourier transform, but it requires points to be unformly distributed. One issue in the particle method is that we do not know ρ at equally spaced points. The transition from particle to grid points is often called the *deposition step* , because the particles are deposited onto the grid. Several deposition steps exist, among which:

- Nearest-Grid Point (NGP) [85]: for each x-cell, all the particles within the cell give their charge to the cell center. Then the electric field is supposed constant on each cell.
- Particle-In-Cell (PIC) [79, 114]: invented by Harlow in 1962, it uses the same deposition step as NGP, but the electric field is then interpolated linearly between two grid centers. This reduces fluctuations from NGP, as well as improve energy conservation.
- Cloud-In-Cell (CIC) [23, 103]: each particle is assumed to have a given shape, the particle charge is then deposited on nearby cells according to this shape.

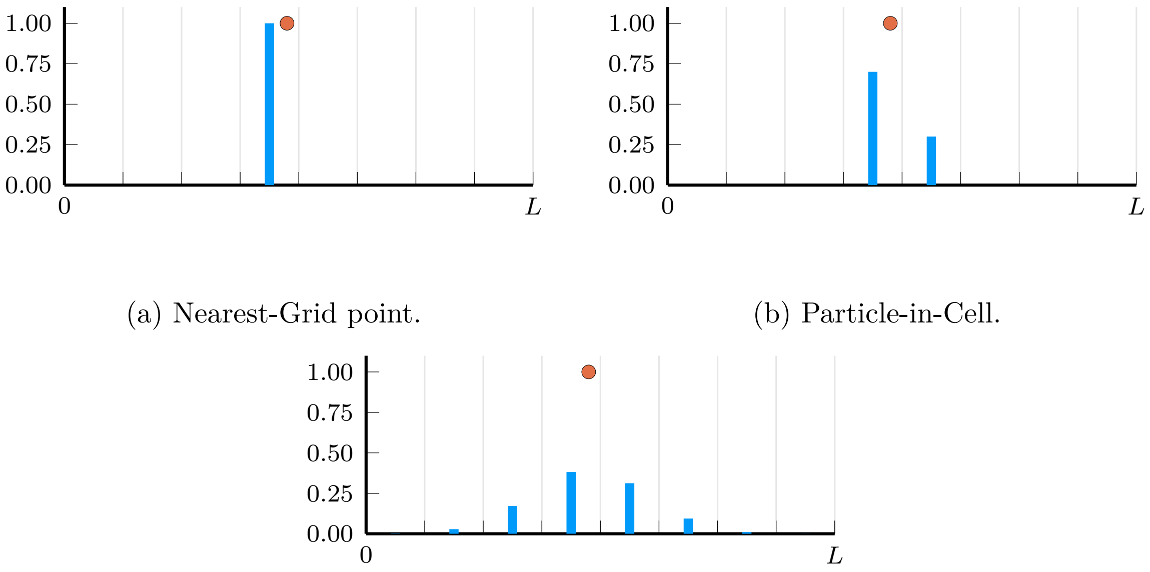
These deposition steps are illustrated in Figure III-3.3.

For methods assuming a given shape function for the particles instead of a Dirac mass, we talk about “finite-size particles”, as opposed to “zero-size particles” which would be the Dirac masses. Reviews of particle methods and their deposition steps are given in [119] and [138]. In [102], the effect of spatial grid and its influence on plasma behavior are studied: it is illustrated numerically that – for some deposition methods – some nonphysical instabilities appear when the grid size is too large compared to the Debye length [6]. We now turn to the description of the Particle-In-Cell approach.

Particle-in-Cell scheme It consists in following the evolution of some point particles. More precisely, given an initial distribution f_0, we approximate it by a sum of P Dirac

[5] See Section II-4 – Complexity for the notation.

[6] Roughly stated, the Debye length is the minimal distance between two electrons in a plasma to be able to distinguish one from the other. Intuitively, the nonphysical instabilities occur because a coarse grid cannot take into account electron interactions which occur within a grid cell.



(a) Nearest-Grid point.

(b) Particle-in-Cell.

(c) Cloud-in-Cell, with a Gaussian shape function.

Figure III-3.3 – Illustrations of different step depositions: Nearest-Grid Point (top left), PIC interpolation (linear, top right), and CIC interpolation (Gaussian shape function, bottom). The particle (orange circle) has a weight of 1, and the value of ρ on the grid (made of 8 cells, delimited by grey vertical thin lines) according to each deposition step is given in blue.

masses, where $P \in \mathbb{N}^*$ is supposedly large:

$$f_0(x,v) \approx \sum_{p=1}^{P} \beta_p \delta(x-x_p)\delta(v-v_p) =: \tilde{f}_0(x,v). \tag{III-3.2}$$

The variables (x_p, v_p) are called the initial coordinates of the particle p in the phase-space, $p \in [\![1, P]\!]$. Here $\delta(\cdot)$ denotes the usual Dirac mass. The quantity β_p is the weight of the particle labelled p, and the weights are usually chosen uniform in Particle-In-Cell methods. In (III-3.2), each Dirac mass represents a collection of sub-particles who are defined only by a point in the phase-space. Moreover it is implicitely assumed that all sub-particles from the same collection remain "close" for all times to the Dirac mass representing the collection. These Dirac masses are usually called *meta-particles*, because each one of them is treated numerically as one particle but may represent physically many sub-particles. Some schemes, such as those presented in [82, 37, 35, 70], allow the meta-particles to be deformed, and even to split or recombine.

Variants of the PIC algorithm have also been used for fluid equations, where additional care is paid to thermodynamic variables. We can cite for instance GAP [107], PAL [109], SOAP [117], and FLIP [30, 29]. The PIC algorithm has also been used in magnetohydrodynamics. See also [134] for use cases where electrons and ions are treated differently.

In the aforementioned works, the Poisson equation (III-1.6) is solved on a grid using a difference scheme (see for example [84, 22]). Later, the FFT [7] has been more widely used. Starting from the late 1970s, simulations with several millions of particles were doable (see [84, Section 9.1.3]).

However, the use of the PIC algorithms or its variants still cause problems. The main ones are instabilities – "a fundamental property of particle-in-cell codes" [27] – and the nonconservation of energy or momentum [28]. The grid used in PIC methods is at the heart of most PIC-related issues. The SPH method [72, 113] is one of the rare grid-free particle methods, and relies on the representation of the solution via an interpolant kernel. It has found numerous applications in astrophysics [112] but, depending on the interpolant kernel, it can have a complexity of order $\mathcal{O}(N^2)$. The use of compactly supported kernels, as suggested in [112], reduces complexity at the expense of a rougher approximation of the electric field: only the nearby particles contribute to the force applied to a given particle, instead of every particle as Coulomb interaction tells us.

Another grid-free approach has been proposed in [46]. It is based on the *Boundary Integral/Treecode (BIT)* method, and the key idea is to replace the particle-particle interactions by particle-cluster interactions. The Fast Multipole Method [8] can be used [76, 41].

Recently, an algorithm based on the Non Uniform Fourier Transform has been proposed [65], and an efficient version [111] has been made using Non Uniform Fast Fourier Transform.

A big argument in favor of particle methods is the good scaling with respect to dimension. In order to compute the charge density ρ as well as some other physical quantities (e.g. total energy, electrical energy, momentum, …), a numerical integration is required. In high-dimensional cases, this quickly becomes a difficult problem. The Monte-Carlo integration [9] allows to obtain approximate values for these integrals with an error scaling as $\mathcal{O}(N^{-1/2})$, where N is the total number of particles used. On the other hand, if a grid-based method was used with M discretization points in each dimension, the number of points would be $\mathcal{O}(M^d)$ while the error is $\mathcal{O}(1/M^\gamma)$, where γ is the order of the numerical integration (generally, $\gamma \approx 1, 2, 3$).

Brackbill mentions in [26] the huge link between computational fluid dynamics and plasma modelling. In particular, the PIC method and its variants in the context of plasma physics are very similar to the Vortex methods for Euler fluid equations.

[7] See Section II-5 – The Fourier transforms for more details on FFT.

[8] Considered one the top 10 algorithms of the XX$^{\text{th}}$ century, according to [47].

Finally, we refer to [126, 48] for a study of particle methods and to [130] for a discussion of physical situations.

The Weighted Particle method

Part III

CHAPTER 4

This chapter is based on the paper [105], published in *Numerische Mathematik*.

In this Chapter we present a particle scheme introduced in [13], and give a convergence result. The proof of the convergence result is detailed in the next Chapter. The scheme was first used to study the magnetization of the Hamiltonian Mean-Field model. Unfortunately, the authors only gave a brief description of the algorithm with no convergence proof, even though the scheme is promising. Our goal is to detail thoroughly the method, and to prove its convergence. The approach presented is different from the Particle-in-Fourier method [111], mainly in the way the charge density is computed, and how the approximate solution is represented. The main advantages of the approach considered is that it allows to obtain high-order estimates by combining well-studied high-order methods, such as integral quadratures and time integration schemes. The convergence estimate shows that with smooth enough initial data, the Fourier truncation error becomes negligible, so that we don't need many Fourier modes in practice.

A by-product of this approach is that all the error terms are decoupled, yielding a relatively easy proof of convergence. This method is named "Weighted Particle method".

We start Section III-4.1 by giving a particular existence result in Sobolev regularity from [38]. Then we will discuss several ways of computing the electric field in the Vlasov-Poisson system. We then finish with a presentation of the Fourier approach to solve the Poisson equation, which will be at the core of the method presented and will allow the definition of a truncated Fourier kernel to the Vlasov-Poisson equation. This truncated Fourier kernel can be seen as an approximation to the exact Fourier kernel which involves infinitely many modes. In Section III-4.2 we explain how the Weighted Particle method is obtained naturally from the truncated Fourier kernel. The building blocks of this scheme are integral quadratures and time integration schemes, allowing a high-order method. Starting from the quadratures, we deduce the particle representation of the approximate solution in a natural way. Moreover, the method presented is totally grid-free since the particles don't require to be deposited onto some grid as it is done, for example, in the Particle-In-Cell method. Section III-4.3 is dedicated to the Weighted Particle method. We start by discussing how this method differs from others in the literature, and then present the main result: the convergence of the approximate characteristics obtained through the Weighted Particle method towards the true characteristics of the Vlasov-Poisson system. One-dimensional numerical results are presented in Section III-4.4 to illustrate the accuracy one can obtain with relatively few particles. The proof of the convergence result will be given in the next Chapter.

III-4.1 Preliminaries

For a given multi-index $p = (p_1, \dots, p_d) \in \mathbb{N}^d$, we denote by ∂_x^p the multi-derivative $\partial_{x_1}^{p_1} \dots \partial_{x_d}^{p_d}$. Similarly, we set $v^m = v_1^{m_1} \dots v_d^{m_d}$ for $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ and $m = (m_1, \dots, m_d) \in \mathbb{N}^d$. We let $| \cdot |$ the usual Euclidian norm on $\mathbb{R}^d$. As the functional framework, we will consider the spaces $\mathcal{H}_\nu^r(U \times V)$ equipped with the norms

$$||f||_{\mathcal{H}_\nu^r}^2 = \sum_{\substack{(m,p,q)\in(\mathbb{N}^d)^3 \\ |p|+|q|\leq r \\ |m|\leq \nu}} \int_V \int_U \left|v^m \partial_x^p \partial_v^q f(x,v)\right|^2 dxdv. \tag{III-4.1}$$

We will mostly talk about the space $\mathcal{H}_\nu^r(\mathbb{T}^d \times \mathbb{R}^d)$, and for sake of clarity we will simply denote this space $\mathcal{H}_\nu^r$. These weighted Sobolev spaces were already considered in [57]. We have the following existence result from [38]:

Theorem III.1

Let $\nu > d/2$, $r \geq 3\nu$. There exist constants $C_{r,\nu}$ and $L_{r,\nu}$ such that for all given $B > 0$ and $f_0 \in \mathcal{H}_\nu^{r+2\nu+1}$ such that $||f_0||_{\mathcal{H}_\nu^{r+2\nu+1}} \leq B$, then for all $\alpha, \beta \in [0,1]$, there exists a solution $f(t,x,v)$ of the Vlasov-Poisson equation

$$\partial_t f + \alpha v \cdot \nabla_x f + \beta \nabla_x \Phi \cdot \nabla_v f = 0, \tag{III-4.2}$$

with initial value $f(0,x,v) = f_0(x,v)$ on the interval $[0,T]$ with

$$T := \frac{C_{r,\nu}}{1+B}, \tag{III-4.3}$$

and we have the estimate

$$\forall t \in [0,T], \quad ||f(t)||_{\mathcal{H}_\nu^{r+2\nu+1}} \leq \min\left(2B, e^{L_{r,\nu}(1+B)t}\right) ||f_0||_{\mathcal{H}_\nu^{r+2\nu+1}}. \tag{III-4.4}$$

Moreover, for two initial conditions f_0 and g_0 satisfying the previous hypothesis, we have

$$\forall t \in [0,T], \quad ||f(t) - g(t)||_{\mathcal{H}^r_\nu} \leq e^{L_{r,\nu}(1+B)t} \, ||f_0 - g_0||_{\mathcal{H}^r_\nu} . \tag{III-4.5}$$

This result holds in the functional space $\mathcal{H}^{r+2\nu+1}_\nu$ which is a subspace of the usual Sobolev space $H^{r+2\nu+1}(\mathbb{T}^d_L \times \mathbb{R}^d) = W^{r+2\nu+1,\,2}(\mathbb{T}^d_L \times \mathbb{R}^d)$.

III-4.1.1 Particle methods

From (III-2.11), the approximate solution to the Vlasov-Poisson system with initial condition $\tilde{f}_0$ can be reconstructed at time t if we know the characteristics at time t. The approximate solution at time t writes

$$f(t,x,v) \approx \tilde{f}_0(X(0;t,x,v), V(0;t,x,v)) = \sum_{p=1}^{P} \beta_p \delta(X(0;t,x,v) - x_p)\delta(V(0;t,x,v) - v_p). \tag{III-4.6}$$

The product of Dirac masses in the sum is non zero if and only if

$$\begin{cases} X(0;t,x,v) = x_p \\ V(0;t,x,v) = v_p \end{cases} \tag{III-4.7}$$

From (III-2.9) this is equivalent to

$$\begin{cases} x = X(t;0,x_p,v_p) \\ v = V(t;0,x_p,v_p) \end{cases} \tag{III-4.8}$$

Therefore, the approximate solution to the Vlasov-Poisson system with initial condition $\tilde{f}_0$ can be written as

$$f(t,x,v) \approx \sum_{p=1}^{P} \beta_p \delta(x - X(t;0,x_p,v_p))\delta(v - V(t;0,x_p,v_p)). \tag{III-4.9}$$

Hence, it is sufficient to follow the characteristics forward in time in order to be able to reconstruct the approximate solution for all times. The main problem with this approach is that, after a time t, the particles are completely disorganized in the phase-space, and hence one needs a “pre-processing” step before being able to compute the electric field $E(t,\cdot)$ which is obtained as $E(t,\cdot) = \nabla_x \Phi(t,\cdot)$, where $\Phi(t,\cdot)$ is the solution to (III-2.2b).

III-4.1.2 Electric field

Kernel-based computation In order to solve the Poisson equation (III-2.2b), one may want to use a Green kernel G to compute E exactly:

$$E(t,x) = \int_{\mathbb{T}^d} \mathcal{K}(x,y) \cdot \left(\rho(t,y) - \frac{1}{|\mathbb{T}_L^d|} \int_{\mathbb{T}^d} \rho(t,\tilde{x}) d\tilde{x} \right) dy, \tag{III-4.10}$$

where

$$\mathcal{K}(x,y) = -\nabla_x G(x,y), \quad -\Delta_x G(x,y) = \delta_0(x-y).$$

This approach can be found in [141, 18, 128], and because it introduces a discontinuity in the kernel $\mathcal{K}$ along the line $\{x = y\}$, there have been some attempts at smoothing it, see e.g. [140].

However the way the electric field is smoothed depends on the authors, and it may seem arbitrary to choose one way or another. In the case of initial particles nonuniformly spaced, the authors of [141] conclude that a mollified version of the kernel G, depending on some mollification parameter, may be preferable to the unmollified version.

This Green kernel-based approach has also been used for numerical computations of fluid dynamics (e.g. Euler equations) in the so-called Vortex and Vortex Blob methods (see [3, 121, 7]). These methods face the same issues, but the convergence of the former methods seems to be have treated more thoroughly (see e.g. [78, 16, 15, 49, 74, 50]). In particular, the authors of these papers have also faced the question of whether or not to mollify the Green kernel, and the overwhelming opinion is that the kernel has to be mollified in order to obtain realistic physical results. Because of the similarities between particle and vortex methods, we can assume this conclusion also holds for particle methods applied to plasma situations. We can also cite [81], where the authors obtain a smooth, high-order kernel approximating the Green kernel G.

The mollification of the Green kernel involved in plasma or fluid dynamic simulations depends on some mollification parameter which is chosen rather arbitrarily in the cited papers. Hence it may not be satisfactory to rely on mollifying the Green kernel, even though its regularized version yields more physical results.

Fourier approach

It has been mentioned previously, the Poisson equation on an uniform grid is easier and faster to solve than the full Coulomb interactions. This explains why the Poisson formulation has been more studied for particle methods in the last decades. The Fourier approach detailed thereafter consists in solving the Poisson equation with periodic boundary conditions by making use the Fourier transform. The Fourier transform can be approximated numerically in a very efficient manner using the Fast Fourier Transform [1], thus the Fourier approach is a good direction in which to look for a fast and efficient particle algorithm.

The issues observed in practice are generally linked to the deposition step and not to the solving of the Poisson equation [2]. We detail below a mesh-free particle method, which is in particular aimed at getting rid of the deposition step.

Let $L := (L_1, \cdots, L_d)$, and for $z \in \mathbb{R}^d$ define

$$\frac{z}{L} := \left(\frac{z_1}{L_1}, \ldots, \frac{z_d}{L_d} \right).$$

We use common notations: $|z|$ for the ℓ^2 norm of a vector $z \in \mathbb{R}^d$, $z \cdot w$ for the ℓ^2 inner-product of two vectors $z, w \in \mathbb{R}^d$, and $|[0, L_1] \times [0, L_d]| = \prod_{i=1}^d L_i$. Moreover, for a multi-index $p \in \mathbb{N}^d$, define

$$z^p := (z_1, \cdots, z_d)^{(p_1, \cdots, p_d)} := z_1^{p_1} \cdots z_d^{p_d}.$$

The convention we use in this Chapter for the Fourier transform $\hat{g}$ of a periodic function $g \in \mathbb{L}^2(\mathbb{T}_L^d)$ is the following:

$$\hat{g}(k) = \frac{1}{|\mathbb{T}_L^d|} \int_{\mathbb{T}_L^d} g(x) e^{-2i\pi k \cdot \frac{x}{L}} dx, \quad k \in \mathbb{Z}^d. \tag{III-4.11}$$

The solution Φ of the Poisson equation (III-2.2b) can be obtained via straightforward computations:

$$\Phi(t,x) = \frac{-1}{|\mathbb{T}_L^d|} \sum_{k \in (\mathbb{Z}^d)^*} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2} \int_{\mathbb{T}_L^d \times \mathbb{R}^d} e^{2i\pi k \cdot \frac{x-y}{L}} f(t,y,v) dy dv. \tag{III-4.12}$$

Moreover, since Φ is a real quantity, the imaginary part of the right-hand side is equal to zero, so that

$$\Phi(t,x) = \frac{-1}{|\mathbb{T}_L^d|} \sum_{k \in (\mathbb{Z}^d)^*} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2} \left[\cos\left(2\pi k \cdot \frac{x}{L}\right) C_k(t) + \sin\left(2\pi k \cdot \frac{x}{L}\right) S_k(t) \right], \tag{III-4.13}$$

[1] See Section II-5 – The Fourier transforms for more details.

[2] See Section III-3 – Review of the Vlasov-Poisson literature for more details.

where

$$C_k(t) := \int_{\mathbb{T}_L^d \times \mathbb{R}^d} \cos\left(2\pi k \cdot \frac{y}{L}\right) f(t,y,v)dydv,$$
$$S_k(t) := \int_{\mathbb{T}_L^d \times \mathbb{R}^d} \sin\left(2\pi k \cdot \frac{y}{L}\right) f(t,y,v)dydv.$$

We easily obtain the electrical field E:

$$\begin{aligned} E(t,x) &= \nabla_x \Phi(t,x) && \text{(III-4.14)} \\ &= \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{k \in (\mathbb{Z}^d)^*} \frac{1}{2\pi \left|\frac{k}{L}\right|^2} \frac{k}{L} \left[\sin\left(2\pi k \cdot \frac{x}{L}\right) C_k(t) - \cos\left(2\pi k \cdot \frac{x}{L}\right) S_k(t)\right] && \text{(III-4.15)} \end{aligned}$$

The formula here, with a series over $k \in (\mathbb{Z}^d)^*$, corresponds to the Poisson framework. However, any truncation in the sum over k can be done in order to approximate E. It is intuitive to consider only a finite number of Fourier modes, and we choose to keep only the modes $\left\{k \in (\mathbb{Z}^d)^* : |k| \leq K\right\}$ where $K \in \mathbb{N}^*$ is some parameter (think of it as *user-input*).

The approximation of the field E for a given K is given by:

$$\begin{aligned} E^K(t,x) &= \nabla_x \Phi^K[f^K](t,x) && \text{(III-4.16)} \\ &= \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} \frac{1}{2\pi \left|\frac{k}{L}\right|^2} \frac{k}{L} \left[\sin\left(2\pi k \cdot \frac{x}{L}\right) C_k^K(t) - \cos\left(2\pi k \cdot \frac{x}{L}\right) S_k^K(t)\right], \\ & && \text{(III-4.17)} \end{aligned}$$

where

$$\begin{aligned} C_k^K(t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi k \cdot \frac{y}{L}\right) f^K(t,y,v)dydv \\ S_k^K(t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} \sin\left(2\pi k \cdot \frac{y}{L}\right) f^K(t,y,v)dydv, \end{aligned} \qquad \text{(III-4.18)}$$

and where the function f^K is solution to the Vlasov-Poisson equation with a truncated kernel:

$$\partial_t f^K(t,x,v) + v \cdot \nabla_x f^K(t,x,v) + E^K(t,x) \cdot \nabla_v f^K(t,x,v) = 0 \qquad \text{(III-4.19a)}$$
$$f^K(0,x,v) = f_0(x,v) \qquad \text{(III-4.19b)}$$

We can define for a given $K \in \mathbb{N}^*$ the characteristics of (III-4.19a) in the following way:

$$\begin{cases} \dfrac{dX^K(t;s,x,v)}{dt} = V^K(t;s,x,v), & X^K(s;s,x,v) = x \\ \dfrac{dV^K(t;s,x,v)}{dt} = E^K(t,X^K(t;s,x,v)), & V^K(s;s,x,v) = v \end{cases} \tag{III-4.20}$$

These characteristics exhibit the same properties as those given in Section III-4.1.1, in particular they are measure-preserving. Thus, for all $k \in (\mathbb{Z}^d)^*$ such that $|k| \leq K$, we have

$$\begin{aligned} C_k^K(t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi k \cdot \frac{X^K(t;0,y,v)}{L}\right) f_0(y,v)dydv, \\ S_k^K(t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} \sin\left(2\pi k \cdot \frac{X^K(t;0,y,v)}{L}\right) f_0(y,v)dydv. \end{aligned} \tag{III-4.21}$$

Remark III.5

Our electric field E^K is presented here as an approximation to the exact E, however one could also understand (III-4.19a) as an intermediate system “between” Vlasov-HMF (in which case $K = 1$) and Vlasov-Poisson (in which case $K \to \infty$).

III-4.2 Building blocks of the Weighted Particle method

The difficulty in the computations of (III-4.21) lies in the fact that we cannot know in practice the characteristics $X^K(t;0,y,v)$ and $V^K(t;0,y,v)$ for all starting points $(y,v) \in \mathbb{T}_L^d \times \mathbb{R}^d$. Hence, it is natural to look at quadrature approximations, which would only involve the characteristics for a finite number of starting points.

III-4.2.1 Quadratures

Denote by $z = (x_1, \cdots, x_d, v_1, \cdots, v_d) \in \mathbb{R}^{2d}$ a variable of the phase-space, and suppose along the dimension i of the phase space we have a quadrature rule of order q_i over a closed interval I_i. The quadrature is defined by some nodes $\left\{z_i^j\right\}_j$, $z_i^j \in I_i$, and nonnegative weights $\left\{w_i^j\right\}_j$. We suppose the nodes are equispaced with step Δz_i, i.e. $z_i^{j_i} = z_i^0 + j_i \Delta z_i$ for some $\Delta z_i > 0$ and $z_i^0 \in I_i$. Under these conditions, the variable j_i belongs to some finite set $J_i := \{0, 1, \cdots, N_i\}$, where $N_i \in \mathbb{N}^*$ and $N_i \leq \left\lfloor \frac{|I_i|}{\Delta z_i} \right\rfloor$.

The error of the quadrature along dimension i is characterized as follows: there exists a constant $C > 0$ such that for all $g \in C^{q_i+1}(I_i)$ we have

$$\left| \int_{I_i} g(\zeta_i) d\zeta_i - \sum_{j_i \in J_i} w_i^{j_i} g(z_i^{j_i}) \right| \leq C \left\| \partial_{\zeta_i}^{q_i+1} g(\zeta_i) \right\|_{\mathbb{L}^\infty(I_i)}.$$

Examples of quadratures satisfying these conditions are the rectangle rule and Newton-Cotes formulae of low order (high orders may involve negative weights).

Remark III.6

We consider uniform quadratures nodes with nonnegative weights for simplicity, in order to obtain a convergence result. However it is also possible to consider in practice non-uniform quadratures (e.g. Gauss-Legendre or Gauss-Hermite quadratures) or negative weights (e.g. high-order Newton-Cotes formulae).

Our notations for the one-dimensional case have been set so that a generalization to the multidimensional case is straightforward. Let $j \in J := J_1 \times \dots \times J_{2d}$ the label of the node $z^j = (z_1^{j_1}, \dots, z_{2d}^{j_{2d}})$ in the multidimensional quadrature over $I_1 \times \dots \times I_{2d}$. The weight of the node z^j is $w^j = w_1^{j_1} \dots w_{2d}^{j_{2d}}$. The multidimensional quadrature over $I_1 \times \dots \times I_{2d}$ is simply a cartesian product of one-dimensional quadratures over $I_1, \dots, I_{2d}$.

In order to understand how (III-4.21) is approximated using this multidimensional integral, suppose for now that the initial condition f_0 has a compact support in velocity: this is only for the sake of understanding, and we will not use this hypothesis later. Under this assumption, let $I_v = I_d \times \dots \times I_{2d}$ a cartesian product of finite intervals $I_d, \dots, I_{2d}$, such that $\operatorname{supp} f_0 \subset \mathbb{T}_L^d \times I_v$. Then, the integrals of (III-4.21) are integrals over $\mathbb{T}_L^d \times I_v$ and we are able to apply quadrature rules as described above to each dimension of the phase-space. We obtain, for all $k \in (\mathbb{Z}^d)^*$ such that $|k| \leq K$,

$$\begin{aligned} C_k^{K,h}(t) &= \sum_{j=(j_1,\dots,j_{2d}) \in J} \cos\left(2\pi k \cdot \frac{X^K(t;0,z^j)}{L}\right) f_0(z^j) w^j, \\ S_k^{K,h}(t) &= \sum_{j=(j_1,\dots,j_{2d}) \in J} \sin\left(2\pi k \cdot \frac{X^K(t;0,z^j)}{L}\right) f_0(z^j) w^j. \end{aligned} \tag{III-4.22}$$

We give later in Proposition III.5 an estimate on the approximation errors

$$\left| C_k^{K,h}(t) - C_k^K(t) \right| \quad \text{and} \quad \left| S_k^{K,h}(t) - S_k^K(t) \right|,$$

depending on the order q_i of the quadratures and the quadrature steps Δz_i.

From the coefficients $C_k^{K,h}$ and $S_k^{K,h}$, one gets the following approximation to the electric field E^K:

$$E^{K,h}(t,x) := \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{2\pi \left|\frac{k}{L}\right|^2} \frac{k}{L} \left[\sin\left(2\pi k\cdot \frac{x}{L}\right) C_k^{K,h}(t) - \cos\left(2\pi k \cdot \frac{x}{L}\right) S_k^{K,h}(t)\right]. \tag{III-4.23}$$

In our notations, the exponent h denotes a phase-space discretization. With this electric field $E^{K,h}$, one can define an approximation to the equation (III-4.19a), which reads

$$\partial_t f^{K,h}(t,x,v) + v\cdot\nabla_x f^{K,h}(t,x,v) + E^{K,h}(t,x)\cdot\nabla_v f^{K,h}(t,x,v) = 0, \tag{VP^{K,h}}$$
$$f^{K,h}(0,x,v) = f_0(x,v). \tag{III-4.24a}$$

Bearing in mind that we are trying to obtain a particle method, the sums in (III-4.22) suggest to have a particle corresponding to each j. We then have $P = |J| = |J_1| \times \dots \times |J_{2d}|$ particles in total. For each $p = 1,\dots,P$, we can find a unique index $j \in J$ such that $(x_p, v_p) := z^j$. The name “Weighted Particle method” stems from the fact that we can understand $f_0(z^j)w^j$ in (III-4.22) as the weight β_p of the particle numbered j (or equivalently, the particle labelled p). Finally, we can define the characteristics of equation (VP^{K,h}) as:

$$\begin{cases} \dfrac{dX_p^K(t)}{dt} = V_p^K(t), & X_p^K(0) = x_p \\ \dfrac{dV_p^K(t)}{dt} = E^{K,h}(t, X_p^K(t)), & V_p^K(0) = v_p \end{cases} \quad p = 1,\dots,P. \tag{III-4.25}$$

The notations for these characteristics are deliberately distinct from those defined in (III-4.20) in order to distinguish them easily.

III-4.2.2 Time integration

We now have only a finite number of particles to follow, and their time evolution is defined by (III-4.25) which is an Ordinary Differential Equation (ODE). Therefore, integrating the ODE over $[0,t]$ gives the characteristics at time t. The problem of integrating numerically an ODE has been thoroughly studied and many numerical schemes exist.

Let $N_t \in \mathbb{N}$, we consider a uniform time-discretization $t^n = n\Delta t$, $0 \leq n \leq N_t$, of stepsize $\Delta t > 0$. We let $T := N_t\Delta t$. The ODE (III-4.25) is written as a first-order ODE, but it can be easily rewritten as a second-order ODE. Therefore, in order to integrate numerically (III-4.25), one can choose a time integration scheme to solve either first-order or second-order ODEs. We suppose the time integration scheme is globally of order γ. As an example, we could take the explicit Euler method which is of order 1, or Runge-Kutta methods whose order depend on the coefficients. It would also be possible to use splitting methods in order to integrate (III-4.25).

Note that (III-4.25) exhibits a Hamiltonian structure since $E^{K,h} = \nabla_x \Phi^{K,h}[f^{K,h}]$ where $f^{K,h}$ is the solution to (VP^{K,h}) and where

$$\Phi^{K,h}[f^{K,h}](t,x) := \frac{-1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{4\pi^2\left|\frac{k}{L}\right|^2} \left[\cos\left(2\pi k\cdot\frac{x}{L}\right) C_k^{K,h}(t) + \sin\left(2\pi k\cdot\frac{x}{L}\right) S_k^{K,h}(t)\right].$$

Therefore, we may benefit from using a symplectic time integrator. Such time integration schemes have also been studied thoroughly, we can cite for instance [124, 77, 66].

For the numerical results that we will present in Section III-4.4, we have chosen a symplectic, 3-stage, explicit, Runge-Kutta-Nyström scheme of order 4. Its Butcher tableau is given in [66, p. 327]. For higher-order symplectic integrators, we refer to [143] or more recently to [38].

Once we have applied our favorite time integration scheme to the particle numbered $p \in \{1, \ldots, P\}$, we obtain an approximation to the solution $(X_p^K(t^n), V_p^K(t^n))_{p=1,\ldots,P}$ of (III-4.25) . We will denote this approximation by

$$X_p^{K,n}, V_p^{K,n}.$$

These are the approximate characteristics that we will compute in practice. Finally, our method can be summed up via Algorithm 1.

III-4.3 Summary of the Weighted Particle method

The Weighted Particle method simply consists in applying the ideas discussed above in Section III-4.2. That is, for a given k, we have to compute the approximate coefficients $C_k^{K,h}$ and $S_k^{K,h}$ via quadratures as written in (III-4.22). This has a complexity $\mathcal{O}(P)$ where P is the total number of particles. Then we have to do this for all $k \in (\mathbb{Z}^d)^*$ such that $|k| \leq K$, in order to compute the approximate electric field $E^{K,h}$ as defined by (III-4.23). This amounts to computing $\mathcal{O}(K^d)$ times the coefficients $C_k^{K,h}, S_k^{K,h}$, once for each k. When this is done and the coefficients $C_k^{K,h}, S_k^{K,h}$ are stored in memory, the computation of the electric field $E^{K,h}$ given by (III-4.23) can be done in $\mathcal{O}(K^d)$ for each particle. In order to update the position and velocity of all particles, one needs to compute the electric field for each one of them. This yields a complexity of $\mathcal{O}(PK^d)$ for each timestep. Then, we can compute the approximate characteristics via a time integration scheme.

Algorithm 1 Weighted Particle Method

```
Require:
  — $f_0$: initial distribution
  — The compact intervals $I_{d+1}, \dots, I_{2d}$.
  — time integration scheme (specifying the timestep $\Delta t$ and the number of timesteps $N_t$)
  — Quadrature rule for each dimension (specifying, for each dimension $i = 1, \dots, 2d$, the number of nodes $N_i+1$, their locations $\{z_i^j\}_{j=0,\dots,N_i}$, and their weights $\{w_i^j\}_{j=0,\dots,N_i}$)
  — $K$: the truncation parameter
  $P = (N_1 + 1) \times \cdots \times (N_{2d} + 1)$.   (Total number of particles)
  x[p], v[p], $\beta$[p] $\leftarrow (x_p, v_p, \beta_p)$, $p = 1, \dots, P$.   (Initial positions, velocities, and weights)
  for $n = 0, \dots, N_t$ do
      $t^n = n\Delta t$
      for all stages of the time integration over a timestep do
          Use NUFFT to compute approximate coefficients $C_k^{K,h}, S_k^{K,h}$ for $|k| \leq K$
          Update x, v with (III-4.25) by using (III-4.23).
          if Last stage of timestep then
              Compute Observables (e.g. electrical energy, momentum, total energy).
          end if
      end for
  end for
```

However, the complexity of order $\mathcal{O}(PK^d)$ may not be satisfying with many dimensions, even with K small. To reduce this, we can use the Non-Uniform Fast Fourier Transform (NUFFT). Roughly put, the NUFFT is an FFT on an upscaled grid and the non-uniform data is interpolated to this grid. The idea is to notice that Equations (III-4.22) and (III-4.23) are (inverse) Fourier transforms. Leveraging the power of the usual FFT, the cost to compute the Fourier transform corresponding to (III-4.22) can be $\mathcal{O}(P + K^d \log K^d)$. The cost to compute the inverse Fourier transform corresponding to (III-4.23) can be $\mathcal{O}(K^d + P \log P)$. Finally, using NUFFT, the global cost for each update of all positions and velocities is $\mathcal{O}(P \log P + K^d \log K^d)$. See [11] and references therein for details about the complexity reductions of NUFFT.

Moreover, it is also possible to specify a desired numerical tolerance and to choose NUFFT parameters so that the relative error for each Fourier coefficient does not exceeed the numerical tolerance. For all the numerical results presented here, this numerical tolerance has been set to 10^{-12}. Following the notation of [11], let ε_∞ the maximal relative error for the (inverse) NUFFT. Then $C_k^{K,h}$ is known up to an error of the order $P\varepsilon_\infty$, and the electric field is known up to an error of the order $PK^d\varepsilon_\infty$. By choosing a small numerical tolerance, ε_∞ get smaller than the given tolerance, and the error $PK^d\varepsilon_\infty$ can get negligible when compared to other error terms.

The basic idea of this scheme had already been given in [13]. However the algorithm proposed in the referenced paper, named "Weighted Particle code", imposes a regular lattice, does not consider Fourier modes other than $k = \pm 1$, imposes a normalization condition on the particle weights, and is only used to study the magnetization of the N-body simulation in the Hamiltonian Mean-Field framework. Finally, no proof of convergence of the algorithm is given, and the time integration scheme is not discussed. We do not have such restrictions here. Our proposed algorithm thus appears to be an extension of the "Weighted Particle code" from [13], and it is guaranteed to converge by Theorem III.2.

It can also be seen as an improvement of the grid-free method presented in [141]: in that work the authors use a smoothed Green kernel, and the rectangle rule to approximate integrals. We allow other types of quadratures here.

Finally, it can be seen as an application to the Vlasov-Poisson system of the method presented in [57], where the authors use the Weighted Particle method to approximate the solution to convection-diffusion equations. Our method could also be understood as a Vortex method with a Fourier regularization of the Green kernel.

We can find such approach to the Vlasov equations via the Fourier kernel mentioned in papers related to the Vlasov-HMF models – such as [33, 4] – but no link to the general Poisson framework is discussed. A similar idea has been proposed in [120] to approximate the collision operator of the Boltzmann equation, called the Fourier-Galerkin spectral method.

The approach presented here is closely related to the Particle-In-Fourier method (PIF), see [111]. In the PIF method the charge density ρ is approximated as a sum of shape functions, which is similar to what is done in the Cloud-In-Cell method. The authors proposed Gaussian shapes as a natural choice, but one could argue that this is pretty arbitrary. Our Weighted Particle method does not require shape functions, and can compute ρ exactly up to the quadrature error. The PIF method also makes use of the Non-Uniform Fast Fourier Transform, so our method is not computationally worse than PIC or PIF. Finally, some ideas leading to the Weighted Particle method are very different from the PIC or PIF approach. In particular we do not seek an approximate solution as a sum of Dirac masses or shape functions, which is a simplifying assumption in PIC and PIF methods: in WPM this representation of the solution is simply a consequence of the quadrature rules, hence the Dirac masses appear naturally as a consequence of the quadrature dirscretization. In our numerical examples, we use the library `FINUFFT.jl`, described in [12, 11].

To be coherent with the paper [13] which first proposed the basic ideas presented here, we name our method "Weighted Particle method" (abbreviated WPM).

III-4.3.1 Convergence of the Weighted Particle method

The following result gives an estimate on how the numerical approximations of the characteristics of (III-4.19a) – with our notations, $X_p^{K,n}$ and $V_p^{K,n}$ – approach the true characteristics of the Vlasov-Poisson equation (III-2.2a) – with our notations, $X(t^n; 0, x_p, v_p)$ and $V(t^n; 0, x_p, v_p)$. We recall that the quantity γ is the global order of the time-integration scheme used, and q_i is the order of the quadrature rule along dimension i.

Theorem III.2: Convergence of the Weighted Particle method

Let $j \in \mathbb{N}$ such that $j \geq 1 + \max_i q_i$, and $\nu, r, \alpha \in \mathbb{N}$ such that $\nu + j > d/2$, $r \geq \max(3(\nu + j), (j-1)(d+1))$, $\alpha \geq 2(r+d)$. Let $K \in \mathbb{N}$, and assume $f_0 \in \mathcal{H}^{r+\alpha}_{\nu+j}$. Then there exists a constant $C > 0$ such that the following holds: for $\delta \geq 0$, define finite intervals $I_{d+1} := [a_1, b_1], \dots, I_{2d} = [a_d, b_d]$ and $I_v := I_{d+1} \times \dots \times I_{2d}$ such that

$$||f_0||_{\mathcal{H}^0_\nu(\mathbb{T}^d_L \times (\mathbb{R}^d \setminus I_v))} \leq \delta.$$

Then for all $K \in \mathbb{N}^*$, and $n = 1, \dots, N_t$

$$\begin{aligned} &\max_{p=1,\dots,P} \left(\left|X_p^{K,n} - X(t^n; 0, x_p, v_p)\right| + \left|V_p^{K,n} - V(t^n; 0, x_p, v_p)\right| \right) \\ &\leq C \left(K^d \left[\delta + K^{\gamma+1} \Delta t^\gamma + \sum_{i=1}^{2d} K^{q_i} \Delta z_i^{q_i} \right] + \frac{1}{(1+K)^{\frac{\alpha+1}{2} - d}} \right) \end{aligned} \tag{III-4.26}$$

where C is independent of $n, \Delta t, \Delta z_i, K$.

The proof of this result relies on the following inequality:

$$\begin{aligned} &|X_p^{K,n} - X(t^n; 0, x_p, v_p)| + |V_p^{K,n} - V(t^n; 0, x_p, v_p)| \\ &\quad \leq |X_p^{K,n} - X_p^K(t^n)| + |V_p^{K,n} - V_p^K(t^n)| \\ &\quad + |X_p^K(t^n) - X^K(t^n; 0, x_p, v_p)| + |V_p^K(t^n) - V^K(t^n; 0, x_p, v_p)| \\ &\quad + |X^K(t^n; 0, x_p, v_p) - X(t^n; 0, x_p, v_p)| + |V^K(t^n; 0, x_p, v_p) - V(t^n; 0, x_p, v_p)| \end{aligned} \tag{III-4.27}$$

Remark III.7

The condition $||f_0||_{\mathcal{H}^0_\nu(\mathbb{T}^d_L \times (\mathbb{R}^d \setminus I_v))} \leq \delta$ means that, for a given $\delta > 0$, we choose I_v large enough so that most of the weighted $\mathbb{L}^2$ mass of f_0 is inside the domain $\mathbb{T}^d \times I_v$. The motivation behind this condition can be roughly stated as: "the quadrature rules are not set on domains where f_0 is negligible (up to an error of order δ)".

Some comments are in order about the error estimate (III-4.26). Intuitively, we would like to have K large so that the system (III-4.19) approximates well the system (III-2.2). However, the error estimate "explodes" as $K \to +\infty$ if Δt and Δz_i are fixed. This creates a CFL-like condition, not between Δt and Δz_i as a usual CFL condition would, but between K and $\Delta t, \Delta z_i$. In other words, it is possible to have K large in the Weighted Particle Method, only under the condition that $K\Delta t$ and $K\Delta z_i$ remain bounded. This imposes the following bounds: $\Delta t, \Delta z_i \leq C/K$ for some constant $C > 0$.

The error estimate (III-4.26) is between the true and approximate characteristics. One may be interested in the error between the exact electric field evaluated at the value of the exact characteristics, and the approximate electric field at the value of the approximate characteristics. Using results proven in Section III-5, we get the following corollary:

Corollary III.1

Under the assumptions of Theorem III.2,

$$\left|E(t^n, X(t^n, 0, x_p, v_p)) - E^K(t^n, X_p^{K,n})\right| \leq C\left|X_p^{K,n} - X(t^n; 0, x_p, v_p)\right|.$$

Proof. It is rather straightforward using results proven later. We have

$$\begin{aligned}&\left|E(t^n, X(t^n; 0, x_p, v_p)) - E^K(t^n, X_p^{K,n})\right|\\ &\leq \left|E(t^n, X(t^n; 0, x_p, v_p)) - E(t^n, X_p^{K,n})\right| + \left|E(t^n, X_p^{K,n}) - E^K(t^n, X_p^{K,n})\right|.\end{aligned}$$

We first note that, by continuity, $X([0,T]; 0, x_p, v_p)$ is a compact set. Hence, using the error estimate (III-4.26), $X_p^{K,n}$ also belongs to a compact set for all n, and so does $X(t^n; 0, x_p, v_p) - X_p^{K,n}$. Moreover, E is differentiable with respect to the space variable x as proven in Proposition III.4, thus there exists a constant $C > 0$ so that

$$\left|E(t^n, X(t^n; 0, x_p, v_p)) - E(t^n, X_p^{K,n})\right| \leq C\left|X(t^n; 0, x_p, v_p) - X_p^{K,n}\right|.$$

This can be bounded using estimate (III-4.26). It remains to estimate the quantity

$$\left|E(t^n, X_p^{K,n}) - E^K(t^n, X_p^{K,n})\right|. \tag{III-4.28}$$

With the notations introduced in Proposition III.3, we have $E = E[f]$ and $E^K = E^K[f^K]$.

This Proposition states that if $\|f - f^K\|^2_{\mathcal{H}^0_\nu} \leq \frac{C}{(1+K)^\alpha}$ for some $C > 0$, then

$$\left|E[g](t,x) - E^K[h](t,x)\right| \leq \frac{C}{(1+K)^{\frac{\alpha+1}{2}-d}} \leq \left|X(t^n;0,x_p,v_p) - X_p^{K,n}\right|. \qquad \text{(III-4.29)}$$

However, Proposition III.2 yields the desired bound on $\|f - f^K\|^2_{\mathcal{H}^0_\nu}$, which finishes the proof. □

We recall that $(X_p^K(t), V_p^K(t))_p$ are the solutions to (III-4.25), that $(X^K(t;0,x_p,v_p)$ and $V^K(t;0,x_p,v_p))$ are the solutions to (III-4.20), and that $(X(t;0,x_p,v_p)$ and $V(t;0,x_p,v_p))$ are the solutions to (III-2.10).

Each line from the RHS of (III-4.27) corresponds to a different type of approximation: the first one is the time discretization error, the second one is the phase-space discretization error (i.e. the quadrature error), and the third one is the kernel truncature error.

Before proving our main result, which is achieved through several estimates and lengthy computations, we illustrate numerically the efficiency of our method.

III-4.4 Numerical Simulations

In this section we will give illustrations on how the Weighted Particle method performs on two standard one-dimensional benchmarks: Weak Landau damping and Two-Stream instability. By *one-dimensional*, we mean one dimension of space and one dimension of velocity. The time integration scheme for all simulations is a symplectic, explicit, 3-stage Runge-Kutta-Nyström method of order 4. Its Butcher tableau was taken from [66, p.327]. The Weighted Particle method is defined by some parameters:

— the truncation parameter K.

— the quadratures in x-space and v-space. We consider the rectangle rule in both dimensions, and let N_1, N_2 be the number of points for each quadrature. The total number of particles is given as $P = N_1 N_2$.

— the compact interval I_v for the v-quadrature. We consider an interval $I_v = [-v_{\max}, v_{\max}]$, where $v_{\max}$ is our parameter.

— the time step Δt of the time integration scheme.

Remark III.8

We recall that the trapezoidal rule on the torus converges exponentially fast for C^∞ functions (see [8, Sect. 5.4, Thm. 5.5]), and notice that when periodicity is considered in the trapezoidal rule, we recover the rectangle rule. This motivates the choice of the rectangle rule in x-space. This argument seems not to hold *a priori* for the rectangle rule in v-space because the initial conditions are not periodic *per se*. However, they converge exponentially fast to zero as $|v| \to \infty$ because of the gaussian enveloppe, hence $f_0(\pm v_{\max}) \approx 0$ for $v_{\max} > 0$ large enough. Numerically, this is the same as if f_0 vanished at $\pm v_{\max}$. Thus, we can conceptually extend f_0 by periodicity from $I_v = [-v_{\max}, v_{\max}]$ to $\mathbb{R}$. This new function is then periodic on $\mathbb{R}$ with period $[-v_{\max}, v_{\max}]$, and on that interval it cannot be distinguished numerically from f_0. This explains why the rectangle rule in v-space is also appropriate, and this holds for both of our initial conditions.

For each example, we display the time evolution of the electrical energy obtained with the WPM method. Moreover the total energy and momentum are conserved for the exact Vlasov-Poisson system, hence we can compare our WPM results with the exact quantities (computed exactly at time $t = 0$) and display the error. We also display the electrical energy as well as the errors obtained with a "reference solution": a Backward semi-Lagrangian scheme (abbrev. BSL) which uses B-splines of degree 5 for the interpolation of the remapping step, and N_1^{BSL}, N_2^{BSL} points. The approximate solution obtained with BSL is an approximation to the solution of (III-2.2a). However, the Poisson equation cannot be solved exactly numerically because all Fourier modes cannot be computed. Actually, we can only compute N_1^{BSL} modes for a Fourier transform along x. The usual Fast Fourier Transform is used to approximately solve the Poisson equation (III-2.2b) on the first N_1^{BSL} Fourier modes.

For this Backward semi-Lagrangian scheme, we have always used $N_1^{BSL} = 512$ points in the x-direction and $N_2^{BSL} = 512$ points in the v-direction. Moreover, it uses the usual Strang splitting procedure for the time integration.

We do not give the evolution of the $\mathbb{L}^p$ norms from the WPM method because they are all conserved with respect to time by construction of the Weighted Particle method: the $\mathbb{L}^p$ norm of the approximate solution is $\left(\sum_{j\in J} f_0(z^j)^p w^j\right)^{1/p}$, and this does not depend on time. Hence the error between the true $\mathbb{L}^p$ norms and the numerical ones are simply the quadrature error at time $t = 0$.

III-4.4.1 Computations of Observables

In our numerical examples to follow, we will monitor the behavior of some quantities for which we know that they are either conserved or know the expected behavior. We give in this section the expressions for some quantities, using the particle representation of the solution.

Electrical energy This quantity is not conserved, but we can compare its behavior with the literature. Recall Equation (III-4.23):

$$E^{K,h}(t,x) = \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{2\pi\left|\frac{k}{L}\right|^2} \frac{k}{L} \left[\sin\left(2\pi k\cdot\frac{x}{L}\right) C_k^{K,h}(t) - \cos\left(2\pi k\cdot\frac{x}{L}\right) S_k^{K,h}(t)\right]$$

In order to compute $\int_{\mathbb{T}_L^d} |E^{K,h}(t,x)|^2 dx$, we first need to compute the following integrals:

$$\begin{aligned}
&\int_{\mathbb{T}_L^d} \sin\left(2\pi k\cdot\frac{x}{L}\right)\sin\left(2\pi l\cdot\frac{x}{L}\right) dx \\
&\qquad = \int_{\mathbb{T}_L^d} \frac{e^{2i\pi k\cdot\frac{x}{L}} - e^{-2\pi k\cdot\frac{x}{L}}}{2i} \frac{e^{2\pi l\cdot\frac{x}{L}} - e^{-2\pi l\cdot\frac{x}{L}}}{2i} dx \\
&\qquad = -\frac{1}{2}\int_{\mathbb{T}_L^d} \left(\cos\left(2\pi(k+l)\cdot\frac{x}{L}\right) - \cos\left(2\pi(k-l)\cdot\frac{x}{L}\right)\right) dx \\
&\qquad = -\frac{\left|\mathbb{T}_L^d\right|}{2}\left(\delta_{k+l,0} - \delta_{k-l,0}\right),
\end{aligned}$$

$$\begin{aligned}
&\int_{\mathbb{T}_L^d} \cos\left(2\pi k\cdot\frac{x}{L}\right)\sin\left(2\pi l\cdot\frac{x}{L}\right) dx \\
&\qquad = \int_{\mathbb{T}_L^d} \frac{e^{2i\pi k\cdot\frac{x}{L}} + e^{-2\pi k\cdot\frac{x}{L}}}{2} \frac{e^{2\pi l\cdot\frac{x}{L}} - e^{-2\pi l\cdot\frac{x}{L}}}{2i} dx \\
&\qquad = \frac{1}{4i}\int_{\mathbb{T}_L^d} \left(2i\sin\left(2\pi(k+l)\cdot\frac{x}{L}\right) - 2i\sin\left(2\pi(k-l)\cdot\frac{x}{L}\right)\right) dx \\
&\qquad = 0,
\end{aligned}$$

$$\begin{aligned}
&\int_{\mathbb{T}_L^d} \cos\left(2\pi k\cdot\frac{x}{L}\right)\cos\left(2\pi l\cdot\frac{x}{L}\right) dx \\
&\qquad = \int_{\mathbb{T}_L^d} \frac{e^{2i\pi k\cdot\frac{x}{L}} + e^{-2\pi k\cdot\frac{x}{L}}}{2} \frac{e^{2\pi l\cdot\frac{x}{L}} + e^{-2\pi l\cdot\frac{x}{L}}}{2} dx \\
&\qquad = \frac{1}{4}\int_{\mathbb{T}_L^d} \left(2\cos\left(2\pi(k+l)\cdot\frac{x}{L}\right) + 2\cos\left(2\pi(k-l)\cdot\frac{x}{L}\right)\right) dx \\
&\qquad = \frac{\left|\mathbb{T}_L^d\right|}{2}\left(\delta_{k+l,0} + \delta_{k-l,0}\right).
\end{aligned}$$

We can now compute the electrical energy:

$$
\begin{aligned}
&\int_{\mathbb{T}_L^d} |E^{K,h}(t,x)|^2 dx \\
&= \frac{1}{\left|\mathbb{T}_L^d\right|^2} \sum_{\substack{k,l\in(\mathbb{Z}^d)^* \\ |k|,|l|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2 \left|\frac{l}{L}\right|^2} \frac{k\cdot l}{L^2} \frac{|\mathbb{T}_L^d|}{2} \Big[-C_k^{K,h} C_l^{K,h} \left(\delta_{k+l,0} - \delta_{k-l,0}\right) \\
&\qquad\qquad + S_k^{K,h} S_l^{K,h} \left(\delta_{k+l,0} + \delta_{k-l,0}\right)\Big] \\
&= \frac{1}{\left|\mathbb{T}_L^d\right|^2} \sum_{\substack{k,l\in(\mathbb{Z}^d)^* \\ |k|,|l|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2 \left|\frac{l}{L}\right|^2} \frac{k\cdot l}{L^2} \frac{|\mathbb{T}_L^d|}{2} \Big[-C_k^{K,h} C_l^{K,h} \delta_{k+l,0} + C_k^{K,h} C_l^{K,h} \delta_{k-l,0} \\
&\qquad\qquad + S_k^{K,h} S_l^{K,h} \delta_{k+l,0} + S_k^{K,h} S_l^{K,h} \delta_{k-l,0}\Big] \\
&= \frac{1}{\left|\mathbb{T}_L^d\right|^2} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2 \left|\frac{-k}{L}\right|^2} \frac{k\cdot(-k)}{L^2} \frac{|\mathbb{T}_L^d|}{2} \left[-C_k^{K,h} C_{-k}^{K,h} + S_k^{K,h} S_{-k}^{K,h} \right] \\
&\quad + \frac{1}{\left|\mathbb{T}_L^d\right|^2} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2 \left|\frac{k}{L}\right|^2} \frac{k\cdot k}{L^2} \frac{|\mathbb{T}_L^d|}{2} \left[C_k^{K,h} C_k^{K,h} + S_k^{K,h} S_k^{K,h} \right] \\
&= \frac{1}{\left|\mathbb{T}_L^d\right|^2} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2 \left|\frac{-k}{L}\right|^2} \frac{|k|^2}{L^2} \frac{|\mathbb{T}_L^d|}{2} \left[(C_k^{K,h})^2 + (S_k^{K,h})^2 \right] \\
&\quad + \frac{1}{\left|\mathbb{T}_L^d\right|^2} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2 \left|\frac{k}{L}\right|^2} \frac{|k|^2}{L^2} \frac{|\mathbb{T}_L^d|}{2} \left[(C_k^{K,h})^2 + (S_k^{K,h})^2 \right] \\
&= \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{4\pi^2 \left|\frac{k}{L}\right|^2} \left[(C_k^{K,h})^2 + (S_k^{K,h})^2 \right]
\end{aligned}
$$

III-4.4.2 Weak Landau damping

Description It is for example a test case in [131, p.54, Sect.4.4.2]. The initial condition writes:

$$f_0(x,v) = (1+\alpha\cos(k_x x))\exp(-v^2/2)\frac{1}{\sqrt{2\pi}}, \quad x\in[0,L], v\in[-v_{\max}, v_{\max}], \tag{III-4.30}$$

where $L := 2\pi/k_x$. This is one of the most famous examples. A numerical scheme has to recover accurately the damping rate and the period between oscillations in the electrical energy. There exists a theoretical formula giving the electrical energy for the dominating Fourier mode (see [131, p.56]). As other modes decay much faster, this formula is a good approximation of the exact electrical energy $E_{elec}^{th}(t)$ after a short time. For $k_x = 0.5$, the formula reads

$$E_{elec}^{th}(t) \approx 0.004 \times 0.3677 e^{-0.1533t} \left|\cos(1.4156t - 0.536245)\right| \sqrt{L/2}. \qquad \text{(III-4.31)}$$

WPM results The numerical parameters are $v_{\max} = 12, k_x = 0.5, \alpha = 0.001, \Delta t = 0.1$. We have used $K = 1$ and $K = 15$ to compare the effects of small and large K. The results are given in Figure III-4.1.

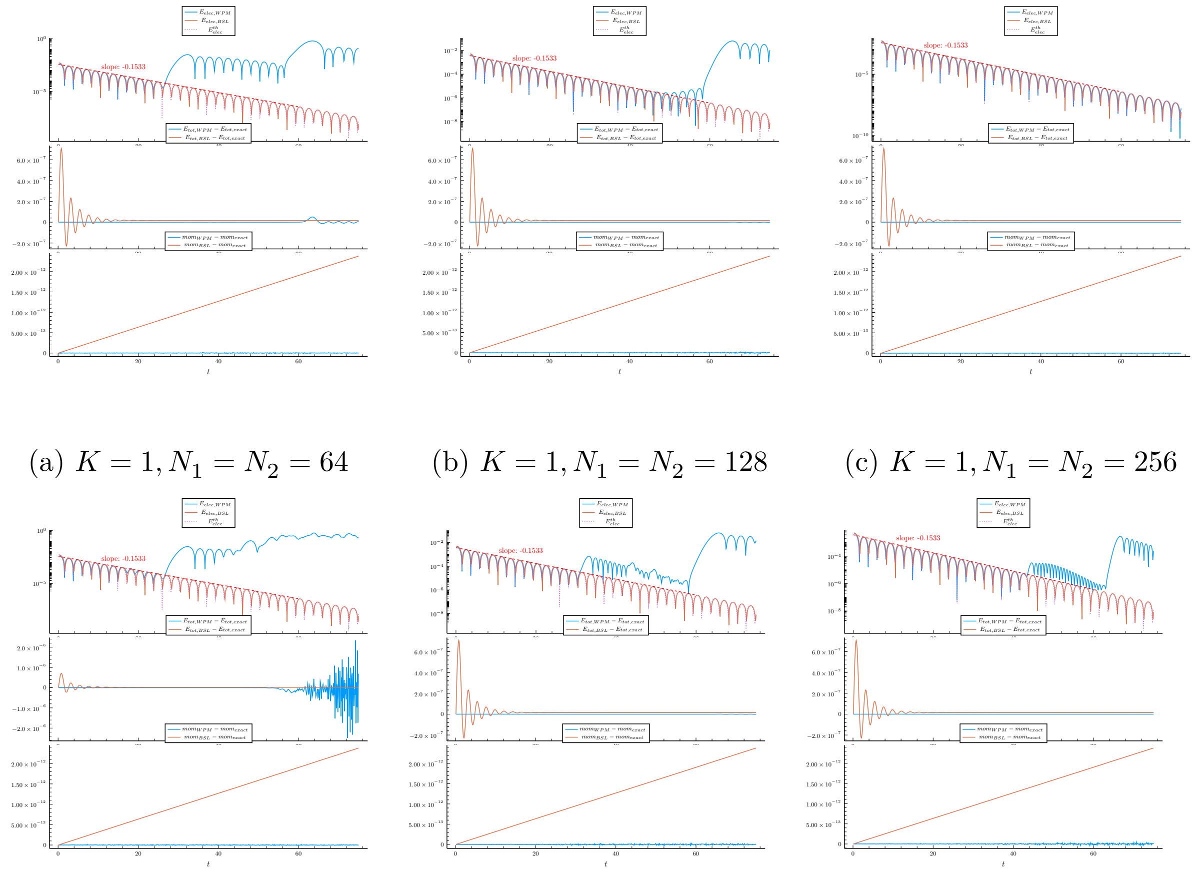
In the top row of each subfigure, we draw the results obtained with WPM (solid blue curve), the expected damping rate (red dashes), and the theoretical electrical energy of the dominating Fourier mode (purple dots). The results of BSL are also given (solid orange curve). For times up to $t \approx 45$ (or $t \approx 25$ in the case $N_1 = N_2 = 64$), the Weighted Particle method can recover the electrical energy with a very good accuracy. In the second row of each subfigure, we draw the difference between the theoretical total energy and the total energy computed from WPM or BSL. We observe that the Weighted Particle method can recover the total energy with a very good accuracy (the difference is of order 10^{-10}), even better than the semi-Lagrangian scheme. In the third row, we compare the exact momentum with the momentum obtained from WPM and BSL. Here as well, the momentum is very well recovered (the difference is of order 10^{-14} for the example with the smallest number of particles), which is again better than BSL.

For this example we also observe an expected jump called the “Poincare recurrence”, which is linked to the compact support in velocity (see [37, 131, 60]). However, we are not able to explain the amplitude increase after the jump. The recurrence with the Weighted Particle Method may be due to considering a quadrature rule with uniformly spaced points. It is possible that a quadrature with non-uniformly spaced points diminishes the effects of the Poincare recurrence, as suggested in [1, 110]. The relative $\mathbb{L}^2$ norm error is of order 10^{-14}.

We can observe on this example that the number of points N_1, N_2 needed to obtain satisfying results increases with K. This was expected from the error estimate of Theorem III.2.

III-4.4.3 Strong Landau damping

The initial condition is again given by (III-4.30). This testcase is given for example in [123, Sect. 5.1]. The electrical energy first decreases from $t = 0$ to $t \approx 15$, then increases until $t \approx 40$, and then stabilizes. Approximate slopes for the decrease and increase in energy can be found in the literature [37, 100].



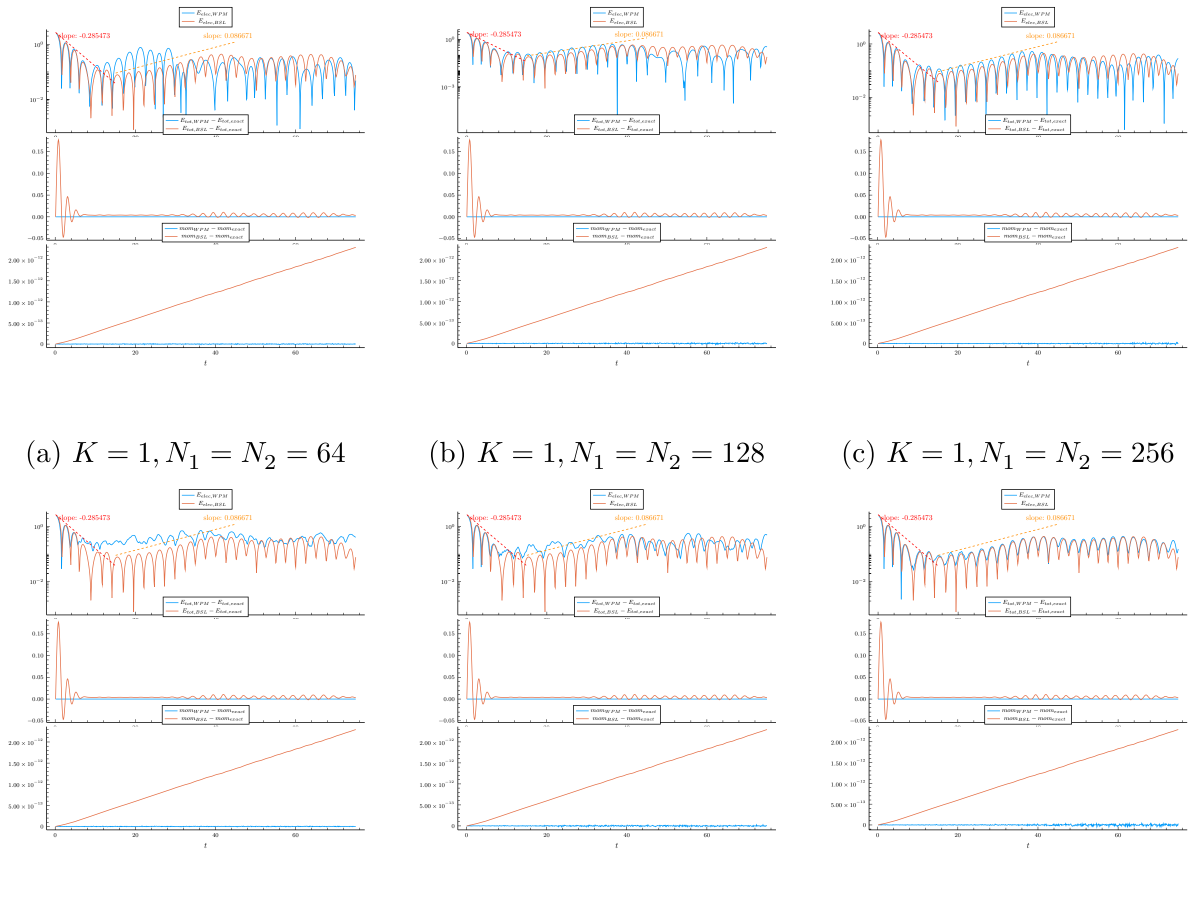
(a) $K = 1, N_1 = N_2 = 64$ (b) $K = 1, N_1 = N_2 = 128$ (c) $K = 1, N_1 = N_2 = 256$

(d) $K = 15, N_1 = N_2 = 64$ (e) $K = 15, N_1 = N_2 = 128$ (f) $K = 15, N_1 = N_2 = 256$

Figure III-4.1 – Results for the Weak Landau damping. Top row (log-scale) : Electrical energy from WPM (resp. BSL), in blue (resp. red). Below: error between WPM (resp. BSL) results and exact quantities, in blue (resp. red) – middle row: total energy, bottom row: momentum.

WPM results The numerical parameters are $v_{\text{max}} = 12, k_x = 0.5, \alpha = 0.5, \Delta t = 0.1$. We have used $K = 1$ and $K = 15$ to compare the effects of small and large K. The results are given in Figure III-4.2.

We can observe that the first results – with $N_1 = N_2 = 64$ and $K = 1$ – are very poor. This can *a priori* have two possible causes: (i) there are not enough particles, and (ii) the number of Fourier modes chosen (here, $K = 1$) is not sufficient. Looking at the results for larger N_1, N_2, we deduce that the first reason seems to be the main one: the results get better when more particles are introduced. However, with $K = 1$, the results are satisfying only until $t \approx 10$, even when a large number of particles are introduced. For $t > 10$, the WPM and BSL results have the same qualitative behavior but not quantitative. This issue is fixed by choosing a larger K. With $K = 15$, the error is first larger than for $K = 1$ – with $N_1, N_2 = 64$ – but gets much better as the number of particles increases: the WPM result is more satisfying than BSL up to time $t \approx 70$, when $K = 15$ and $N_1 = N_2 = 256$.



(a) $K = 1, N_1 = N_2 = 64$ (b) $K = 1, N_1 = N_2 = 128$ (c) $K = 1, N_1 = N_2 = 256$

(d) $K = 15, N_1 = N_2 = 64$ (e) $K = 15, N_1 = N_2 = 128$ (f) $K = 15, N_1 = N_2 = 256$

Figure III-4.2 – Results for the Strong Landau damping. Top row (log-scale) : Electrical energy from WPM (resp. BSL), in blue (resp. red). Below: error between WPM (resp. BSL) results and exact quantities, in blue (resp. red) – middle row: total energy, bottom row: momentum.

III-4.4.4 Two-Stream Instability

Description

This example can be found in [131, p.57] or [53, p.1738]. Depending on the reference, the initial condition may be different. The idea of this example in both cases is to have two streams with opposite velocities. We will consider the formulation from [131]. The initial condition then reads:

$$f_0(x, v) = (1 + \alpha \cos(k_x x)) \frac{1}{2\sqrt{2\pi}} (\exp(-(v - v_0)^2/2) + \exp(-(v + v_0)^2/2)), \quad \text{(III-4.32)}$$

for $x \in [0, 2\pi/k_x], v \in [-v_{\max}, v_{\max}]$.

WPM results The numerical parameters are $\alpha = 0.001, v_{\max} = 12, k_x = 0.2, v_0 = 3, \Delta t = 0.1$. We have used $K = 1$ and $K = 15$ to compare the effects of small and large K. The results are given in Figure III-4.3.

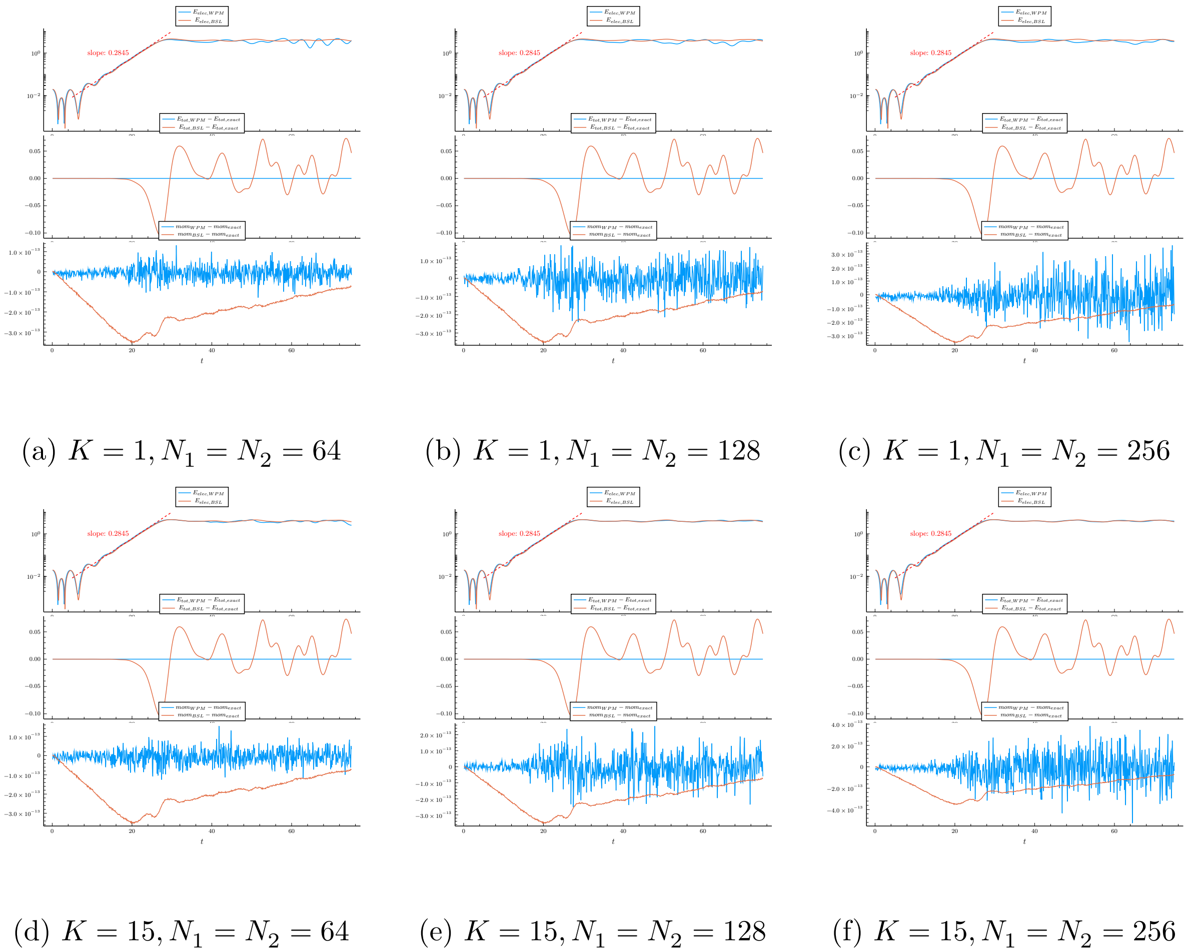
It is known that the Two-Stream instability first exhibits a short transition state, followed by an instability, and then some periodic behavior. The instability rate is 0.2845.

As in the previous example, the first row of each subfigure corresponds to the electrical energy obtained with the Weighted Particle method (solid blue curve), Backward Semi-Lagrangian (solid orange curve), and we display the expected instability rate (red dashes). We can observe that the instability rate is recovered accurately with both WPM and BSL. In the second row of each subfigure, we display the error between the theoretical total energy and the total energy obtained with WPM and BSL. The total energy is also recovered accurately with WPM (the difference is of order 10^{-6}), much more accurately than with BSL. In the third row, we compare the exact momentum with the momentum obtained from WPM and BSL. Here as well, the momentum is very well recovered (e.g. the difference is of order 10^{-13} for the example with the smallest number of particles). The relative $\mathbb{L}^2$ norm error is of order 10^{-14}.

The results for this test case are all satisfying in regard to the “interesting” part, corresponding to the instability and transition states, which happen for $t \leq 30$. The results differ after this time. Once again, we can observe that increasing K or (N_1, N_2) yields better results, closer to what is expected. For once, the results with $K = 15, N_1 = N_2 = 64$ are not worse than the results with $K = 1, N_1 = N_2 = 64$. This is not to be expected for all initial conditions as can be seen from the error estimate of Theorem III.2: for given $\Delta t, \Delta z_i$, if K increases the error bound increases.

For all those examples we were able to recover very accurately the exact momentum, electrical energy and total energy. Relatively few particles were needed, compared to the usual PIC methods. As a comparison, we can cite for instance the paper [116] which uses a Particle-In-Wavelets scheme, where 2^{19} particles were necessary in order to obtain satisfying results with a tolerable statistical noise on the Landau Damping and Two-Stream instability examples. The authors of [37] have done some Particle-In-Cell simulations and show that, on the Strong and Weak Landau damping examples after a short time, the statistical noise with 256×256 particles prevents from drawing conclusions from the results. The method presented in [37] does not have such a problem and can predict accurately the damping rates, but requires frequent remapping.

Moreover we have not displayed here the results of the comparison between a symplectic time integrator and a non-symplectic one, but experiments show that using a



(a) $K = 1, N_1 = N_2 = 64$ (b) $K = 1, N_1 = N_2 = 128$ (c) $K = 1, N_1 = N_2 = 256$

(d) $K = 15, N_1 = N_2 = 64$ (e) $K = 15, N_1 = N_2 = 128$ (f) $K = 15, N_1 = N_2 = 256$

Figure III-4.3 – Results for the Two-Stream Instability. Top row (log-scale) : Electrical energy from WPM (resp. BSL), in blue (resp. red). Below: error between WPM (resp. BSL) results and exact quantities, in blue (resp. red) – middle row: total energy, bottom row: momentum.

symplectic time integrator prevents from obtaining a drift in conservative quantities (e.g. total energy) which otherwise occurs. For this comparison, we have tested symplectic and non-symplectic versions of a 4th order Runge-Kutta-Nyström time integrator.

Part III

CHAPTER 5

Proof of the convergence theorem III.2

We prove in this section the convergence of Theorem III.2. The proof ends on Page 128.

The first thing to show is that the truncation of the kernel does not modify the existence result given by Theorem III.1.

We recall that the spaces $\mathcal{H}^r_\nu$ are defined by (III-4.1). For functions in $\mathcal{H}^r_\nu$, we consider the Fourier transform along the space variable $x \in \mathbb{T}^d_L$ and denote this transform $\mathcal{F}_x$. Let P_K be the projection on the Fourier modes with frequency $|k| \leq K$.

We have the following lemma:

Lemma III.2

Let $K \in \mathbb{N}^*$, define Φ, Φ^K as in (III-4.14) and (III-4.16). Then, for all $\nu, r \in \mathbb{N}$, we have

$$\forall g \in \mathcal{H}^r_\nu, \quad P_K\Phi[g] = \Phi^K[g] = \Phi[P_K g] \tag{III-5.1}$$

and

$$\forall g \in \mathcal{H}^r_\nu, \quad ||P_K g||_{\mathcal{H}^r_\nu} \leq ||g||_{\mathcal{H}^r_\nu}. \tag{III-5.2}$$

Proof. The first equality of (III-5.1) is just the definition of Φ^K. The second equality is straightforward by noting that $\Phi^K[g] = P_K\Phi[g]$, that the mapping $g \mapsto \Phi[g]$ is linear, and that the only dependance in the space variable x of $\Phi[g]$ is the dependance on x of g. It can also be shown by computing $P_K\Phi[g]$ and $\Phi[P_K g]$ explicitely and comparing the expressions.

For the estimate (III-5.2), we have by the Parseval equality

$$\begin{aligned} ||P_K g||^2_{\mathcal{H}^r_\nu} &= \sum_{\substack{(m,p,q)\in(\mathbb{N}^d)^3 \\ |p|+|q|\leq r \\ |m|\leq\nu}} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \int_{\mathbb{R}^d} \left| v^m \partial^q_v \mathcal{F}_x(g)(k,v) k^p \right|^2 dv \\ &\leq \sum_{\substack{(m,p,q)\in(\mathbb{N}^d)^3 \\ |p|+|q|\leq r \\ |m|\leq\nu}} \sum_{k\in(\mathbb{Z}^d)^*} \int_{\mathbb{R}^d} \left| v^m \partial^q_v \mathcal{F}_x(g)(k,v) k^p \right|^2 dv = ||g||^2_{\mathcal{H}^r_\nu}, \end{aligned}$$

and we recall that $k^p := k_1^{p_1}...k_d^{p_d}$. □

It is now possible to follow step by step the proofs of [38, Thm. 5.1, Lemma 5.3], with the estimates holding thanks to Lemma III.2, and we will obtain the following existence result:

Proposition III.1

Let $K \in \mathbb{N}^*$, $\nu, r \in \mathbb{N}$, with $\nu > d/2$ and $r \geq 3\nu$. There exist constants $C_{r,\nu}$ and $L_{r,\nu}$ such that for all given $B > 0$ and $f_0 \in \mathcal{H}_\nu^{r+2\nu+1}$ with $||f_0||_{\mathcal{H}_\nu^{r+2\nu+1}} \leq B$, for all $\alpha, \beta \in [0,1]$, there exists a solution $f^K(t,x,v)$ of the Vlasov-Poisson equation with truncated kernel (III-4.19a)

$$\begin{cases} \partial_t f^K + \alpha v \cdot \nabla_x f^K + \beta \nabla_x \Phi^K[f^K] \cdot \nabla_v f^K = 0 \\ f(0,x,v) = f_0(x,v) \end{cases}$$

on the interval $[0,T]$ with

$$T := \frac{C_{r,\nu}}{1+B},$$

and we have the estimate

$$\forall t \in [0,T], \quad \left|\left|f^K(t)\right|\right|_{\mathcal{H}_\nu^{r+2\nu+1}} \leq \min\left(2B, e^{L_{r,\nu}(1+B)t}\right) ||f_0||_{\mathcal{H}_\nu^{r+2\nu+1}}. \tag{III-5.3}$$

Moreover, for two initial conditions f_0 and g_0 satisfying the previous hypotheses, we have

$$\forall t \in [0,T], \quad \left|\left|f^K(t) - g^K(t)\right|\right|_{\mathcal{H}_\nu^r} \leq e^{L_{r,\nu}(1+B)t} ||f_0 - g_0||_{\mathcal{H}_\nu^r}. \tag{III-5.4}$$

We do not give the proof here as it would amount to copy *verbatim* the proof from [38], and we refer the reader to this paper for a self-contained proof. We have the following lemma:

Lemma III.3

Let $\nu, r_1, r_2 \in \mathbb{N}$ such that $r_2 \geq r_1$. For all $f \in \mathcal{H}_\nu^{r_2}$, there exists a constant $C > 0$ such that for all $k \in \mathbb{Z}^d$, and all $q \in \mathbb{N}^d$ such that $|q| \leq r_2$,

$$\forall m \in \mathbb{N}^d, \ |m| \leq \nu, \quad \int_{\mathbb{R}^d} \left|v^m \partial_v^q \mathcal{F}_x(f)(k,v)\right|^2 dv \leq \frac{C}{(1+|k|)^{2(r_2-|q|)}}$$

and, for all $K \in \mathbb{N}^*$,

$$||(I - P_K)f||^2_{\mathcal{H}^{r_1}_\nu} \leq \frac{C||f||^2_{\mathcal{H}^{r_2}_\nu}}{(1+K)^{2(r_2-r_1)}}.$$

Proof. Recall the definition of the $\mathcal{H}^{r_2}_\nu$ norm:

$$||f||^2_{\mathcal{H}^{r_2}_\nu} = \sum_{\substack{(m,p,q)\in(\mathbb{N}^d)^3\\|p|+|q|\leq r_2\\|m|\leq\nu}} \int_{\mathbb{R}^d}\int_{\mathbb{T}^d} \left|v^m \partial^p_x \partial^q_v f(x,v)\right|^2 dxdv$$

By the Parseval equality,

$$\begin{aligned}||f||^2_{\mathcal{H}^{r_2}_\nu} &= \left|\mathbb{T}^d_L\right| \sum_{\substack{(m,\tilde{p},q)\in(\mathbb{N}^d)^3\\|\tilde{p}|+|q|\leq r_2\\|m|\leq\nu}} \int_{\mathbb{R}^d} \sum_{k\in\mathbb{Z}^d} \left|\mathcal{F}_x\left(v^m\partial^{\tilde{p}}_x\partial^q_v f\right)(k,v)\right|^2 dv\\ &= \left|\mathbb{T}^d_L\right| \sum_{\substack{(m,q)\in(\mathbb{N}^d)^2\\|q|\leq r_2\\|m|\leq\nu}} \sum_{k\in\mathbb{Z}^d} \sum_{\substack{\tilde{p}\in\mathbb{N}^d\\|\tilde{p}|\leq r_2-|q|}} (2\pi)^{2|\tilde{p}|}\left(\frac{k}{L}\right)^{2\tilde{p}} \int_{\mathbb{R}^d} \left|v^m\partial^q_v\mathcal{F}_x(f)(k,v)\right|^2 dv.\end{aligned} \tag{III-5.5}$$

We recall that with our convention, as $\tilde{p}\in\mathbb{N}^d,\ k,L\in\mathbb{R}^d$ we let

$$\left(\frac{k}{L}\right)^{2\tilde{p}} = \left(\frac{k_1}{L_1}\right)^{2\tilde{p}_1} \cdots \left(\frac{k_d}{L_d}\right)^{2\tilde{p}_d}.$$

A by-product of (III-5.5) is that, since the right-hand side is finite, the sum over k is also finite for every m,q. In the sum over $\tilde{p}\in\mathbb{N}^d$ with $|\tilde{p}|\leq r_2-|q|$, we have in particular for each $i=1,\dots,d$, the term $\tilde{p}=(0,\cdots,0,r_2-|q|,0,\cdots,0)$ where only the $i-th$ coordinate is nonzero and its value is $r_2-|q|$. There is as well $\tilde{p}=(0,\dots,0)$. Thus, for some constant C that does not depend on k,

$$\sum_{\substack{\tilde{p}\in\mathbb{N}^d\\|\tilde{p}|\leq r_2-|q|}} (2\pi)^{2|\tilde{p}|}\left(\frac{k}{L}\right)^{2\tilde{p}} \geq C\left(1+\sum_{i=1}^d k_i^{2(r_2-|q|)}\right).$$

The right-hand side of (III-5.5) being finite for every m,q, we then have

$$\left(1+\sum_{i=1}^d k_i^{2(r_2-|q|)}\right)\int_{\mathbb{R}^d}\left|v^m\partial^q_v\mathcal{F}_x(f)(k,v)\right|^2 dv \leq C,$$

for some C large enough. Finally, for all $|q| \leq r_2,\ |m| \leq \nu$, we have

$$\int_{\mathbb{R}^d} |v^m \partial_v^q \mathcal{F}_x(f)(k,v)|^2 \, dv \leq \frac{C}{1 + \sum_{i=1}^d k_i^{2(r_2-|q|)}} \leq \frac{C}{(1+|k|)^{2(r_2-|q|)}},$$

where the last equality is a consequence of Jensen's inequality. This shows the first estimate we claim.

We now proceed to showing the second estimate. Coming back to (III-5.5), let $\tilde{p} = p+s$, where $p, s \in \mathbb{N}^d$ are such that $|s| \leq r_2 - r_1$ and $|p| \leq r_1 - |q|$. A given value of $\tilde{p}$ may be obtained by several combinations of s and p. However, the maximal number M of combinations yielding the same $\tilde{p}$ is finite and depends only on d, r_1, r_2. Therefore,

$$||f||^2_{\mathcal{H}^{r_2}_\nu} \geq \frac{|\mathbb{T}^d_L|}{M} \sum_{\substack{(m,q)\in(\mathbb{N}^d)^2 \\ |q|\leq r_2 \\ |m|\leq\nu}} \sum_{k\in\mathbb{Z}^d} \sum_{\substack{s\in\mathbb{N}^d \\ |s|\leq r_2-r_1}} (2\pi)^{2|s|} \left(\frac{k}{L}\right)^{2s} \sum_{\substack{p\in\mathbb{N}^d \\ |p|\leq r_1-|q|}} (2\pi)^{2|p|} \left(\frac{k}{L}\right)^{2p} \times \int_{\mathbb{R}^d} |v^m \partial_v^q \mathcal{F}_x(f)(k,v)|^2 \, dv$$

In the sum over $s \in \mathbb{N}^d$ with $|s| \leq r_2 - r_1$, we have in particular for each $i = 1, \dots, d$ the term $s = (0, \cdots, 0, r_2 - r_1, 0, \cdots, 0)$ where only the $i-th$ coordinate is nonzero and its value is $r_2 - r_1$. Thus,

$$||f||^2_{\mathcal{H}^{r_2}_\nu} \geq \frac{|\mathbb{T}^d_L|}{M} \sum_{\substack{(m,q)\in(\mathbb{N}^d)^2 \\ |q|\leq r_2 \\ |m|\leq\nu}} \sum_{k\in\mathbb{Z}^d} \left(\sum_{i=1}^d \left[2\pi \frac{k_i}{L_i}\right]^{2(r_2-r_1)} \right) \sum_{\substack{p\in\mathbb{N}^d \\ |p|\leq r_1-|q|}} (2\pi)^{2|p|} \left(\frac{k}{L}\right)^{2p} \times \int_{\mathbb{R}^d} |v^m \partial_v^q \mathcal{F}_x(f)(k,v)|^2 \, dv.$$

Again, by the Jensen inequality, there exists a constant $C_1 > 0$ such that

$$\sum_{i=1}^d \left[2\pi \frac{k_i}{L_i}\right]^{2(r_2-r_1)} \geq C_1 |k|^{2(r_2-r_1)}.$$

Hence

$$||f||^2_{\mathcal{H}^{r_2}_\nu} \geq C \sum_{\substack{(m,q)\in(\mathbb{N}^d)^2 \\ |q|\leq r_2 \\ |m|\leq\nu}} \sum_{k\in\mathbb{Z}^d} |k|^{2(r_2-r_1)} \sum_{\substack{p\in\mathbb{N}^d \\ |p|\leq r_1-|q|}} (2\pi)^{2|p|} \left(\frac{k}{L}\right)^{2p} \int_{\mathbb{R}^d} |v^m \partial_v^q \mathcal{F}_x(f)(k,v)|^2 \, dv,$$

where we let $C := C_1 \frac{|\mathbb{T}^d_L|}{M}$.

In the sum over q, we can drop the terms corresponding to $|q| > r_1$ because it yields an empty set $\{p \in \mathbb{N}^d : |p| \leq r_1 - |q|\}$. Thus,

$$||f||^2_{\mathcal{H}^{r_2}_\nu} \geq C \sum_{\substack{(m,q)\in(\mathbb{N}^d)^2 \\ |q|\leq r_1 \\ |m|\leq \nu}} \sum_{k\in\mathbb{Z}^d} |k|^{2(r_2-r_1)} \sum_{\substack{p\in\mathbb{N}^d \\ |p|\leq r_1-|q|}} (2\pi)^{2|p|} \left(\frac{k}{L}\right)^{2p} \int_{\mathbb{R}^d} |v^m \partial^q_v \mathcal{F}_x(f)(k,v)|^2 \, dv. \tag{III-5.6}$$

Now, if $|k| > K$, then $\mathbb{N} \ni |k|^2 > K^2 \geq 1 + K^2$. As we want an estimate that depends on $(1+K)^{2(r_2-r_1)}$ and not on $(1+K^2)^{r_2-r_1}$, we use the following inequality:

$$(K-1)^2 + (K+1)^2 = 2(1+K^2) \implies (1+K)^2 \leq 2(1+K^2).$$

We truncate the sum over $k \in \mathbb{Z}^d$ to $|k| > K$ in (III-5.6), and we get

$$||f||^2_{\mathcal{H}^{r_2}_\nu} \geq C \left(\frac{(1+K)^2}{2}\right)^{r_2-r_1} \sum_{\substack{(m,q)\in(\mathbb{N}^d)^2 \\ |q|\leq r_1 \\ |m|\leq \nu}} \sum_{\substack{k\in\mathbb{Z}^d \\ |k|>K}} \sum_{\substack{p\in\mathbb{N}^d \\ |p|\leq r_1-|q|}} (2\pi)^{2|p|} \left(\frac{k}{L}\right)^{2p} \int_{\mathbb{R}^d} |v^m \partial^q_v \mathcal{F}_x(f)(k,v)|^2 \, dv.$$

Finally we can compare this expression to the one we had in (III-5.5), and obtain

$$||f||^2_{\mathcal{H}^{r_2}_\nu} \geq C(1+K)^{2(r_2-r_1)} \, ||(I-P_K)f||^2_{\mathcal{H}^{r_1}_\nu}.$$

□

We now have bounds for f and f^K, uniform in K, so we are able to obtain an estimate on their difference.

Proposition III.2

Let $\nu, r \in \mathbb{N}$, with $\nu > d/2$, $r \geq 3\nu$, and $\alpha \geq 2\nu+1$. Let f be the solution to the Vlasov-Poisson equation (III-2.2a), and f^K be the solution to the Vlasov-Poisson equation with truncated kernel (III-4.19a), both with the same initial condition $f_0 \in \mathcal{H}^{r+\alpha}_\nu$, such that $||f_0||_{\mathcal{H}^{r+\alpha}_\nu} \leq B$ for some $B > 0$. Then, there exists a constant $C > 0$ such that for all $K \in \mathbb{N}^*$ and all $t \in [0,T]$,

$$\left|\left|(f-f^K)(t)\right|\right|^2_{\mathcal{H}^r_\nu} \leq \frac{C}{(1+K)^\alpha} \tag{III-5.7}$$

Proof. We follow the end of the proof of Theorem 5.1 from [38].

By taking the difference (III-2.2a) - (III-4.19a), we obtain

$$\partial_t(f-f^K)+v\cdot\nabla_x(f-f^K)-\nabla_x\Phi\cdot\nabla_v(f-f^K)=\nabla_x\Phi[P_Kf^K-f]\cdot\nabla_v f^K.$$

We have by previous estimates

$$\forall t\in[0,T],\qquad \begin{cases} ||f(t)||_{\mathcal{H}^{r+\alpha}_\nu}\le C(t,r,\nu,B)||f_0||_{\mathcal{H}^{r+\alpha}_\nu} \\ ||f^K(t)||_{\mathcal{H}^{r+\alpha}_\nu}\le C(t,r,\nu,B)||f_0||_{\mathcal{H}^{r+\alpha}_\nu} \end{cases}. \tag{III-5.8}$$

Since $\alpha\ge 2\nu+1$, Lemma 5.3 from [38], gives for all $t\in[0,T]$

$$\begin{aligned}||(f-f^K)(t)||^2_{\mathcal{H}^r_\nu}\le ||(f-f^K)(0)||^2_{\mathcal{H}^r_\nu}+C\int_0^t\left(1+||f(\sigma)||_{\mathcal{H}^r_\nu}\right)\left|\left|(f-f^K)(\sigma)\right|\right|^2_{\mathcal{H}^r_\nu}d\sigma \\ +2\int_0^t\left|\left|\nabla_x\Phi[P_Kf^K-f]\cdot\nabla_v f^K(\sigma)\right|\right|_{\mathcal{H}^r_\nu}\left|\left|(f-f^K)(\sigma)\right|\right|_{\mathcal{H}^r_\nu}d\sigma\end{aligned} \tag{III-5.9}$$

We have (we skip the details since they are given in [38])

$$\left|\left|v^m\partial_x^p\partial_v^q\left(\nabla_x\Phi[P_Kf^K-f]\cdot\nabla_v f^K\right)\right|\right|_{\mathbb{L}^2}\le C_{r,\nu}||f||_{\mathcal{H}^{r+2\nu+1}_\nu}||P_Kf^K-f||_{\mathcal{H}^r_\nu}. \tag{III-5.10}$$

Moreover, from the decomposition $P_Kf^K-f=P_K(f^K-f)+(P^K-I)f$ we have, using Lemma III.3,

$$\begin{aligned}||P_Kf^K-f||_{\mathcal{H}^r_\nu}&\le ||P_K(f^K-f)||_{\mathcal{H}^r_\nu}+||(I-P_K)f||_{\mathcal{H}^r_\nu}\\ &\le ||f^K-f||_{\mathcal{H}^r_\nu}+||(I-P_K)f||_{\mathcal{H}^r_\nu}\\ &\le ||f^K-f||_{\mathcal{H}^r_\nu}+\frac{C}{(1+K)^\alpha}.\end{aligned} \tag{III-5.11}$$

Then (III-5.9) becomes, with the help of (III-5.10),

$$\begin{aligned}\forall t\in[0,T],\quad ||(f-f^K)(t)||^2_{\mathcal{H}^r_\nu}\le ||(f-f^K)(0)||^2_{\mathcal{H}^r_\nu}+C(f_0)\int_0^t||(f-f^K)(\sigma)||^2_{\mathcal{H}^r_\nu}d\sigma\\ +2C_{r,\nu}(f_0)\int_0^t\left(||(f^K-f)(\sigma)||^2_{\mathcal{H}^r_\nu}+\frac{C}{(1+K)^\alpha}\right)d\sigma.\end{aligned} \tag{III-5.12}$$

Since (III-2.2b) and (III-4.19a) have the same initial condition, we obtain by the Grönwall lemma (see Lemma II.1) the existence of a time-dependent function C, independent of K, that depends on r,ν,f_0, such that

$$\forall t\in[0,T],\quad \left|\left|(f-f^K)(t)\right|\right|^2_{\mathcal{H}^r_\nu}\le\frac{C(t)}{(1+K)^\alpha}.$$

Since the function $C(t)$ depends continuously on $t \in [0,T]$, we get the result. □

Proposition III.3

Let $c \in \mathbb{N}^d$, $\nu \in \mathbb{N}, \alpha \in \mathbb{N}^*$, with $\nu > d/2$. Let $E[g] := \nabla_x \Phi[g]$ be the kernel to the Vlasov-Poisson equation (III-2.2a), computed with some function $g \in \mathcal{H}^\alpha_\nu$, and let $E^K[h] := \nabla_x \Phi^K[h]$ be the kernel to the Vlasov-Poisson equation with truncated kernel (III-4.19a), computed with $h \in \mathcal{H}^\alpha_\nu$. We do not require g and h to be respectively solutions of (III-2.2a) and (III-4.19a). Assume there exists a constant $C > 0$ such that for all $K \in \mathbb{N}^*$, $||(g-h)(t)||^2_{\mathcal{H}^0_\nu} \leq \frac{C}{(1+K)^\alpha}$. Then, for all $t \in [0,T]$ and all $x \in \mathbb{T}^d_L$,

$$\left|\partial^c_x \left(E[g](t,x) - E^K[h](t,x)\right)\right| \leq \frac{C}{(1+K)^{\frac{\alpha+1}{2}-d-\sum_i c_i}}. \tag{III-5.13}$$

Proof. For any $t \in [0,T]$ and $x \in \mathbb{T}^d_L$,

$$\begin{aligned}
&\partial^c_x \left(E^K[h](t,x) - E[g](t,x)\right) \\
&= \frac{1}{|\mathbb{T}^d_L|} \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} \frac{\frac{k}{L}}{2\pi \left|\frac{k}{L}\right|^2} \left(2\pi \frac{k}{L}\right)^c \sin\left(2\pi \frac{k}{L} \cdot y\right) \left(C^K_k(t) - C_k(t)\right) \\
&\quad - \frac{1}{|\mathbb{T}^d_L|} \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} \frac{\frac{k}{L}}{2\pi \left|\frac{k}{L}\right|^2} \left(2\pi \frac{k}{L}\right)^c \cos\left(2\pi \frac{k}{L} \cdot y\right) \left(S^K_k(t) - S_k(t)\right) \\
&\quad + \frac{1}{|\mathbb{T}^d_L|} \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| > K}} \frac{\frac{k}{L}}{2\pi \left|\frac{k}{L}\right|^2} \left(2\pi \frac{k}{L}\right)^c \left(\sin\left(2\pi \frac{k}{L} \cdot y\right) C_k(t) - \cos\left(2\pi \frac{k}{L} \cdot y\right) S_k(t)\right)
\end{aligned}$$

Let $\bar{c} := \sum_i c_i$. Note that $|k^c| \leq |k|^{\bar{c}}$, therefore

$$\begin{aligned}
\left|\partial_x^c \left(E^K[h](t,x) - E[g](t,x)\right)\right| &\leq \frac{(2\pi)^{\bar{c}}}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{\left|\frac{k}{L}\right|^{\bar{c}+1}}{2\pi\left|\frac{k}{L}\right|^2} \left(\left|C_k^K(t) - C_k(t)\right| + \left|S_k^K(t) - S_k(t)\right|\right) \\
&\quad + \frac{(2\pi)^{\bar{c}}}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|> K}} \frac{\left|\frac{k}{L}\right|^{\bar{c}+1}}{2\pi\left|\frac{k}{L}\right|^2} \left(|C_k(t)| + |S_k(t)|\right) \\
&\leq \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{2\pi\left|\frac{k}{L}\right|^{1-\bar{c}}} \left(\left|C_k^K(t) - C_k(t)\right| + \left|S_k^K(t) - S_k(t)\right|\right) \\
&\quad + \frac{1}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|> K}} \frac{1}{2\pi\left|\frac{k}{L}\right|^{1-\bar{c}}} \left(|C_k(t)| + |S_k(t)|\right).
\end{aligned} \tag{III-5.14}$$

We have, for $g, h \in \mathcal{H}_\nu^r$,

$$\begin{aligned}
\left|C_k^K(t) - C_k(t)\right| &\leq \int_{\mathbb{T}_L^d\times\mathbb{R}^d} |g(t,y,v) - h(t,y,v)|\, dydv \\
&\leq \int_{\mathbb{T}_L^d} \left(\int_{\mathbb{R}^d} \frac{dv}{(1+|v|^2)^\nu}\right)^{1/2} \left(\int_{\mathbb{R}^d} (1+|v|^2)^\nu \, |g(t,y,v) - h(t,y,v)|^2\, dv\right)^{1/2} dy \\
&\leq C \left(\int_{\mathbb{T}_L^d\times\mathbb{R}^d} (1+|v|^2)^\nu \, |g(t,y,v) - h(t,y,v)|^2\, dydv\right)^{1/2} \\
&\leq C \, ||(g-h)(t)||_{\mathcal{H}_\nu^0},
\end{aligned}$$

for some constant C that does not depend on K or t, thanks to the assumption $\nu > d/2$. The same estimate holds naturally for $\left|S_k^K(t) - S_k(t)\right|$. Therefore, using our hypothesis $||(g-h)(t)||^2_{\mathcal{H}_\nu^0} \leq \frac{C}{(1+K)^\alpha}$, we get

$$\left|C_k(t) - C_k^K(t)\right|^2 \leq \frac{C}{(1+K)^\alpha}.$$

The same estimate holds for $\left|S_k^K(t) - S_k(t)\right|^2$. Then, summing over k and applying a discrete Cauchy-Schwarz inequality, we obtain for any $\mu > d+2(\bar{c}-1)$, i.e. $2-2\bar{c}+\mu > d$,

$$\begin{aligned}
&\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\leq K}} \frac{1}{2\pi\left|\frac{k}{L}\right|^{1-\bar{c}+\mu/2}}|k|^{\mu/2}\left(\left|C_k^K(t)-C_k(t)\right|+\left|S_k^K(t)-S_k(t)\right|\right)\\
&\leq \left(\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\leq K}} \frac{1}{4\pi^2\left|\frac{k}{L}\right|^{2-2\bar{c}+\mu}}\right)^{1/2}\left(\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\leq K}} |k|^{\mu}\left[\left|C_k^K(t)-C_k(t)\right|+\left|S_k^K(t)-S_k(t)\right|\right]^2\right)^{1/2}\\
&\leq C\left(\frac{K^{d+\mu}}{(1+K)^{\alpha}}\right)^{1/2}\\
&\leq \frac{C}{(1+K)^{(\alpha-d-\mu)/2}} \qquad\qquad \text{(III-5.15)}
\end{aligned}$$

where the constant C does not depend on K or t.

The second sum in (III-5.14) can be estimated with Lemma III.3 by using the fact that $g\in\mathcal{H}^{\alpha}_{\nu}$. Indeed, we have

$$\begin{aligned}
|C_k(t)| &= \left|\frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{T}^d_L}\left(e^{2i\pi\frac{k}{L}\cdot y}+e^{-2i\pi\frac{k}{L}\cdot y}\right)g(t,y,v)dydv\right|\\
&= \left|\frac{|\mathbb{T}^d_L|}{2}\int_{\mathbb{R}^d}\left(\mathcal{F}_x(g)(k,v)+\mathcal{F}_x(g)(-k,v)\right)dv\right|\\
&\leq C\left(\left[\int_{\mathbb{R}^d}(1+|v|^2)^{\nu}\left|\mathcal{F}_x(g)(k,v)\right|^2dv\right]^{1/2}+\left[\int_{\mathbb{R}^d}(1+|v|^2)^{\nu}\left|\mathcal{F}_x(g)(-k,v)\right|^2dv\right]^{1/2}\right).
\end{aligned}$$

Now apply Lemma III.3 to obtain, for all $|k|>K$,

$$|C_k(t)|\leq\frac{C}{(1+K)^{\alpha}}, \qquad \text{(III-5.16)}$$

the same estimate holding for $|S_k(t)|$. Hence,

$$\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|>K}}\frac{1}{|k|^{1-\bar{c}}}\left(|C_k(t)|+|S_k(t)|\right)\leq\frac{C}{(1+K)^{\alpha+1-\bar{c}-(d+1)}}=\frac{C}{(1+K)^{\alpha-d-\bar{c}}}. \qquad \text{(III-5.17)}$$

It remains to compare the exponents in (III-5.15) and (III-5.17). Under the condition $\mu > d+2(\bar{c}-1)$, we have

$$\alpha-d-\bar{c}-\frac{\alpha-d-\mu}{2}=\frac{\alpha}{2}-\bar{c}-\frac{d}{2}+\frac{\mu}{2}>\frac{\alpha}{2}-\bar{c}+\bar{c}-1=\frac{\alpha}{2}-1.$$

Since $\alpha\geq 1$, we have $\frac{\alpha}{2}-1\geq-\frac{1}{2}$. However, the quantities α, d and $\bar{c}$ are all integers, hence the condition $\alpha - d - \bar{c} > -\frac{1}{2}$ implies $\alpha - d - \bar{c} \geq 0$. Hence, the error in (III-5.14) is dominated by the error of the first sum. Taking for instance $\mu = d + 2\bar{c} - 1$, we obtain

$$\left|\partial_x^c \left(E[g](t,x) - E^K[h](t,x)\right)\right| \leq \frac{C}{(1+K)^{\frac{\alpha+1}{2}-d-\bar{c}}}$$

□

We will need at some point regularity in time for f^K, E^K, and this can be obtained at the expense of additional space regularity. The following lemma shows how to "exchange" space regularity with time regularity:

Proposition III.4

Let $j \in \mathbb{N}^*$, $\nu, r, \alpha \in \mathbb{N}$ such that $\nu + j > d/2$, $r \geq \max\left(3(\nu+j), (j-1)(d+1)\right)$ and $\alpha \geq 2(r+d)$. Let $K \in \mathbb{N}^*$. If $f_0 \in \mathcal{H}^{r+\alpha}_{\nu+j}$, then the solution f^K to (III-4.19a), as well as the solution f to (III-2.2a), are smooth with respect to time in $\mathcal{H}^r_\nu$. That is, for all $l \in \mathbb{N}$ with $l \leq j$,

$$\partial_t^l f^K \in \mathcal{H}^{r-(l-1)(d+1)}_{\nu+j-l}, \qquad \partial_t^l f \in \mathcal{H}^{r-(l-1)(d+1)}_{\nu+j-l},$$

and we have

$$E^K \in C^j([0,T] \times \mathbb{R}^d), \qquad E \in C^j([0,T] \times \mathbb{R}^d).$$

Proof. Thanks to the way the kernel E^K is defined, the joint regularity in (t,x) can be studied by studying the regularity in t and the regularity in x. Note that $(x \mapsto E^K(t,x))$ is $C^\infty(\mathbb{R}^d)$ and periodic with period $\mathbb{T}^d_L$, so it only remains to study the regularity with respect to time of the kernel, which boils down to studying the regularity with respect to time of the coefficients $C_k^K(t), S_k^K(t)$. Our proof will be done by induction on the derivative.

Base case With our assumptions we get $r + \alpha - 2(\nu+j) - 1 \geq 3(\nu+j)$, so that by Proposition III.1 we have $f^K(t) \in \mathcal{H}^{r+\alpha}_{\nu+j}$ for short enough times. Thus,

$$\begin{aligned}
\partial_t C_k^K(t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi \frac{k}{L} \cdot y\right) \partial_t f^K(t,y,v)\, dydv \\
&= -\int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi \frac{k}{L} \cdot y\right) \left(v \cdot \nabla_x f^K(t,y,v) + E^K(t,y) \cdot \nabla_v f^K(t,y,v)\right) dydv \\
&= -\int_{\mathbb{T}^d \times \mathbb{R}^d} v \cdot \frac{2\pi k}{L} \sin\left(2\pi \frac{k}{L} \cdot y\right) f^K(t,y,v)\, dydv,
\end{aligned}$$

since

$$-\int_{\mathbb{T}^d\times\mathbb{R}^d}\cos\left(2\pi\frac{k}{L}\cdot y\right)E^K(t,y)\cdot\nabla_v f^K(t,y,v)dydv=0.$$

This can be rewritten

$$\begin{aligned}\partial_t C_k^K(t)&=-\frac{1}{2i}\int_{\mathbb{R}^d}v\cdot\frac{2\pi k}{L}\left(\int_{\mathbb{T}_L^d}e^{2i\pi\frac{k}{L}\cdot y}f^K(t,y,v)dy-\int_{\mathbb{T}_L^d}e^{-2i\pi\frac{k}{L}\cdot y}f^K(t,y,v)dy\right)dv\\&=-\frac{|\mathbb{T}_L^d|}{2i}\int_{\mathbb{R}^d}v\cdot\frac{2\pi k}{L}\left(\mathcal{F}_x(f^K)(-k,v)-\mathcal{F}_x(f^K)(k,v)\right)dv.\end{aligned}$$

Therefore,

$$\left|\partial_t C_k^K(t)\right|\leq\frac{|\mathbb{T}_L^d|}{2}\left|\frac{2\pi k}{L}\right|\int_{\mathbb{R}^d}|v|\left(\left|\mathcal{F}_x(f^K)(-k,v)\right|+\left|\mathcal{F}_x(f^K)(k,v)\right|\right)dv.$$

Apply the Cauchy-Schwarz inequality with

$$|v|\left|\mathcal{F}_x(f^K)(k,v)\right|=\frac{(1+|v|)^\nu}{(1+|v|)^\nu}|v|\left|\mathcal{F}_x(f^K)(k,v)\right|,$$

to obtain, for some C which does not depend on K:

$$\begin{aligned}\left|\partial_t C_k^K(t)\right|&\leq C|k|\left(\begin{aligned}&\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f^K)(k,v)\right|^2\right]^{1/2}\\&\qquad+\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f^K)(-k,v)\right|^2\right]^{1/2}\end{aligned}\right)\\&\leq C|k|\left(\begin{aligned}&\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f-f^K)(k,v)\right|\right]^{1/2}+\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f)(k,v)\right|\right]^{1/2}\\&+\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f-f^K)(-k,v)\right|^2\right]^{1/2}+\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f)(-k,v)\right|^2\right]^{1/2}\end{aligned}\right)\end{aligned}$$

The same estimate holds for $|\partial_t S_k^K(t)|$. The second and fourth terms are estimated by Lemma III.3, using that $f\in\mathcal{H}_{\nu+j}^{r+\alpha}$:

$$\left[\int_{\mathbb{R}^d}|v|^{2(1+\nu)}\left|\mathcal{F}_x(f)(-k,v)\right|^2\right]^{1/2}\leq\frac{1}{(1+|k|)^{r+\alpha}}.$$

Let $c \in \mathbb{N}^d$ and let $\bar{c} := \sum_i c_i$. We assume $\bar{c} \leq r + \alpha - 2(\nu + j) - 1$, so that

$$\begin{aligned}
\left|\partial_x^c \partial_t E^K(t,x)\right| &\leq \frac{(2\pi)^{\bar{c}}}{\left|\mathbb{T}_L^d\right|} \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} \frac{1}{2\pi \left|\frac{k}{L}\right|^{1-\bar{c}}} \left(\left|\partial_t C_k^K(t)\right| + \left|\partial_t S_k^K(t)\right|\right) \\
&\leq C \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} |k|^{\bar{c}} \left(\begin{array}{l} \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(f - f^K)(k,v)\right|^2 dv\right]^{1/2} + \frac{1}{(1+|k|)^{r+\alpha}} \\ + \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(f - f^K)(-k,v)\right|^2 dv\right]^{1/2} + \frac{1}{(1+|k|)^{r+\alpha}} \end{array} \right).
\end{aligned}$$

The sum

$$\sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} |k|^{\bar{c}} \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(f - f^K)(k,v)\right|^2 dv\right]^{1/2}$$

can be bounded by some quantity equivalent to $\left\|f - f^K\right\|_{\mathcal{H}^{\bar{c}}_{\nu+1}} \leq \left\|f - f^K\right\|_{\mathcal{H}^{r+\alpha-2(\nu+j)-1}_{\nu+j}}$. Since $f - f^K \in \mathcal{H}^{r+\alpha}_{\nu+j}$, by Proposition III.2 we get

$$\sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} |k|^{\bar{c}} \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(f - f^K)(k,v)\right|^2 dv\right]^{1/2} \leq \frac{C}{(1+K)^{\nu+j+1/2}}.$$

Hence,

$$\left|\partial_x^c \partial_t E^K(t,x)\right| \leq \frac{C}{(1+K)^{\nu+j+1/2}} + C \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} \frac{1}{(1+|k|)^{2(\nu+j)+1}}.$$

This can be bounded by a constant $C > 0$ which does not depend on K, by using the fact that $2(\nu + j) + 1 > d + 1$. Hence, for $\beta_1 \in \mathbb{N}$,

$$\left\|\partial_t f^K\right\|_{\mathcal{H}^{\beta_1}_{\nu+j-1}} \leq \left\|v \cdot f^K\right\|_{\mathcal{H}^{\beta_1}_{\nu+j-1}} + \left\|E^K \cdot \nabla_v f^K\right\|_{\mathcal{H}^{\beta_1}_{\nu+j-1}} \leq C \left\|f^K\right\|_{\mathcal{H}^{\beta_1+1}_{\nu+j}},$$

where the last inequality holds if

$$\beta_1 \leq r + \alpha - 2(\nu + j) - 1. \qquad \text{(III-5.18)}$$

For the right-hand side of the estimate to be finite, we need to have

$$\beta_1 + 1 \leq r + \alpha,$$

since we only have $f^K \in \mathcal{H}^{r+\alpha}_{\nu+j}$. However this is already satisfied by (III-5.18) since

$\nu + j \geq 0$. From now on let $\beta_1 = r$, so that

$$\partial_t f^K \in \mathcal{H}^{\beta_1}_{\nu+j-1}.$$

This holds true for any $K \in \mathbb{N}^*$, let's now show this estimate also holds with the solution f to the non-truncated Vlasov-Poisson equation. Let $p \leq \beta_1 - 1$,

$$\begin{aligned}
&\left|\left|\partial_t (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-1}} \\
&\leq \left|\left|v \cdot \nabla_x (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-1}} + \left|\left|E^K \cdot \nabla_v f^K - E \cdot \nabla_v f\right|\right|_{\mathcal{H}^p_{\nu+j-1}} \\
&\leq \left|\left|v \cdot \nabla_x (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-1}} + \left|\left|E^K \cdot \nabla_v (f^K - f)\right|\right|_{\mathcal{H}^p_{\nu+j-1}} + \left|\left|(E - E^K) \cdot \nabla_v f\right|\right|_{\mathcal{H}^p_{\nu+j-1}} \\
&\leq \left|\left|f - f^K\right|\right|_{\mathcal{H}^{p+1}_{\nu+j}} + C \left|\left|f^K - f\right|\right|_{\mathcal{H}^{p+1}_{\nu+j}} + \max_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \left(E^K - E\right)\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \left|\left|f\right|\right|_{\mathcal{H}^{p+1}_{\nu+j}}. \quad \text{(III-5.19)}
\end{aligned}$$

Because $p + 1 \leq \beta_1 \leq r$, we have

$$\left|\left|f - f^K\right|\right|_{\mathcal{H}^{p+1}_{\nu+j}} \leq \left|\left|f - f^K\right|\right|_{\mathcal{H}^r_{\nu+j}} \leq \frac{C}{(1+K)^{\alpha/2}},$$

where the first inequality is clear and the second one comes from Proposition III.2. For the third term of (III-5.19), we have

$$\left|\left|f - f^K\right|\right|_{\mathcal{H}^0_{\nu+j}} \leq \left|\left|f - f^K\right|\right|_{\mathcal{H}^r_{\nu+j}} \leq \frac{C}{(1+K)^{\alpha/2}},$$

so that, by Proposition III.3,

$$\max_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \left(E^K - E\right)\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \leq \max_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \frac{C}{(1+K)^{(\alpha+1)/2 - d - \bar{c}}} \leq \frac{C}{(1+K)^{(\alpha+1)/2 - d - p}}.$$

Hence, (III-5.19) yields

$$\begin{aligned}
\left|\left|\partial_t (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-1}} &\leq (C+1) \left|\left|f^K - f\right|\right|_{\mathcal{H}^{p+1}_{\nu+j}} + \max_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \left(E^K - E\right)\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \left|\left|f\right|\right|_{\mathcal{H}^{p+1}_{\nu+j}} \\
&\leq \frac{C}{(1+K)^{\alpha/2}} + \frac{C}{(1+K)^{(\alpha+1)/2 - d - p}}.
\end{aligned}$$

Thus,

$$\left|\left|\partial_t (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-1}} \leq \frac{C}{(1+K)^{(\alpha+1)/2 - d - p}} = \frac{C}{(1+K)^{\gamma_1 + \beta_1 - p}},$$

where γ_1 is defined by the relation

$$\begin{aligned} &\frac{\alpha+1}{2} - d - p = \gamma_1 + \beta_1 - p \\ \iff \gamma_1 = &\frac{\alpha+1}{2} - \beta_1 - d = \frac{\alpha+1}{2} - r - d. \end{aligned}$$

Requiring $\gamma_1 > 0$ yields the condition

$$\alpha > 2(\beta_1 + d) - 1 = 2(r+d) - 1.$$

With our assumption $\alpha \geq 2(r+d)$, the above inequality is satisfied.

Induction Let's now turn to the higher derivatives. Let $l \in \mathbb{N}$ with $l \leq j$, suppose that for any $m \leq l-1$, $\partial_t^m f^K, \partial_t^m f \in \mathcal{H}^{\beta_m}_{\nu+j-m}$ for some $r = \beta_1 \geq \dots \geq \beta_{l-1} > 0$, and assume there exists C and $0 < \gamma_1 \leq \dots \leq \gamma_{l-1}$ such that for all $m \leq l-1$, $p \leq \beta_m$,

$$||\partial_t^m (f - f^K)||_{\mathcal{H}^p_{\nu+j-m}} \leq \frac{C}{(1+K)^{\gamma_m + \beta_m - p}}. \tag{III-5.20}$$

Let $m \leq l$, we have

$$\begin{aligned} \partial_t^m C_k^K(t) &= \int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi \frac{k}{L} \cdot y\right) \partial_t^m f^K(t,y,v) dydv \\ &= - \int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi \frac{k}{L} \cdot y\right) \begin{pmatrix} v \cdot \nabla_x \partial_t^{m-1} f^K(t,y,v) \\ \qquad + \partial_t^{m-1} \left[E^K(t,y) \cdot \nabla_v f^K(t,y,v) \right] \end{pmatrix} dydv \\ &= - \int_{\mathbb{T}^d \times \mathbb{R}^d} \cos\left(2\pi \frac{k}{L} \cdot y\right) v \cdot \nabla_x \partial_t^{m-1} f^K(t,y,v) dydv, \end{aligned}$$

since

$$\int_{\mathbb{T}^d} \cos\left(2\pi \frac{k}{L} \cdot y\right) E^K(t,y) \cdot \left(\int_{\mathbb{R}^d} \nabla_v f^K(t,y,v) dv \right) dy = 0.$$

As in the case $l = 1$, we have

$$\left|\partial_t^m C_k^K(t)\right| \leq C|k| \left(\left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1} f^K)(k,v)\right|^2 dv\right]^{1/2} + \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1} f^K)(-k,v)\right|^2 dv\right]^{1/2} \right)$$

$$\leq C|k| \left(\begin{aligned} &\left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1}(f - f^K))(k,v)\right| dv\right]^{1/2} \\ &+ \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1} f)(k,v)\right| dv\right]^{1/2} \\ &+ \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1}(f - f^K))(-k,v)\right|^2 dv\right]^{1/2} \\ &+ \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1} f)(-k,v)\right|^2 dv\right]^{1/2} \end{aligned} \right).$$

The same estimate holds for $|\partial_t^m S_k^K(t)|$.

The second and fourth terms are estimated by Lemma III.3, using that $\partial_t^{m-1} f \in \mathcal{H}^{\beta_{m-1}}_{\nu+j-(m-1)}$:

$$\left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1} f)(-k,v) dv\right|^2\right]^{1/2} \leq \frac{1}{(1+|k|)^{\beta_{m-1}}}.$$

For $c \in \mathbb{N}^d$, $\bar{c} \leq \beta_{m-1} - d - 1$, we have

$$\left|\partial_x^c \partial_t^m E^K(t,x)\right| \leq \frac{(2\pi)^{\bar{c}}}{|\mathbb{T}_L^d|} \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} \frac{1}{2\pi \left|\frac{k}{L}\right|^{1-\bar{c}}} \left(\left|\partial_t^{m-1} C_k^K(t)\right| + \left|\partial_t^{m-1} S_k^K(t)\right|\right)$$

$$\leq C \sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} |k|^{\bar{c}} \left(\begin{aligned} &\left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1}(f - f^K))(k,v)\right|^2 dv\right]^{1/2} + \frac{1}{(1+|k|)^{\beta_{m-1}}} \\ &+ \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1}(f - f^K))(-k,v)\right|^2 dv\right]^{1/2} + \frac{1}{(1+|k|)^{\beta_{m-1}}} \end{aligned} \right).$$

The sum

$$\sum_{\substack{k \in (\mathbb{Z}^d)^* \\ |k| \leq K}} |k|^{\bar{c}} \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1}(f - f^K))(k,v)\right|^2 dv\right]^{1/2}$$

can be bounded by some quantity equivalent to

$$||\partial_t^{m-1}(f-f^K)||_{\mathcal{H}^{\bar{c}}_{\nu+1}} \leq ||\partial_t^{m-1}(f-f^K)||_{\mathcal{H}^{\beta_{m-1}-d-1}_{\nu+j-(m-1)}} \leq \frac{C}{(1+K)^{\gamma_{m-1}+d+1}},$$

where the last inequality is given by our induction hypothesis (III-5.20). That is,

$$\sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} |k|^{\bar{c}} \left[\int_{\mathbb{R}^d} |v|^{2(1+\nu)} \left|\mathcal{F}_x(\partial_t^{m-1}(f-f^K))(k,v)\right|^2 dv\right]^{1/2} \leq \frac{C}{(1+K)^{\gamma_{m-1}+d+1}}.$$

Hence, for all $m \leq l$ and $\bar{c} \leq \beta_{m-1} - d - 1$,

$$\left|\partial_x^c \partial_t^{m-1} E^K(t,x)\right| \leq \frac{C}{(1+K)^{\gamma_{m-1}+d+1}} + C \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\leq K}} \frac{1}{(1+|k|)^{d+1}}.$$

This can be bounded by a constant $C > 0$ which does not depend on K. Thus, there exists a constant $C > 0$ such that for all $m \leq l$, all $c \in \mathbb{N}^d$ with $\bar{c} \leq \beta_{m-1} - d - 1$ and all $K \in \mathbb{N}$,

$$\left|\partial_x^c \partial_t^{m-1} E^K(t,x)\right| \leq C$$

for some constant $C > 0$ which is independent of t, K, x.

Hence, for $\beta_l \in \mathbb{N}$,

$$\begin{aligned} ||\partial_t^l f^K||_{\mathcal{H}^{\beta_l}_{\nu+j-l}} &\leq ||v\cdot\partial_t^{l-1}\nabla_x f^K||_{\mathcal{H}^{\beta_l}_{\nu+j-l}} + \sum_{m=0}^{l-1}\binom{l-1}{m} ||\partial_t^m E^K \cdot \nabla_v \partial_t^{l-1-m} f^K||_{\mathcal{H}^{\beta_l}_{\nu+j-l}} \\ &\leq C \sum_{m=0}^{l-1}\binom{l-1}{m} ||\partial_t^{l-1-m} f^K||_{\mathcal{H}^{\beta_l+1}_{\nu+j-(l-1)}} \end{aligned}$$

where the last inequality holds when

$$\begin{gathered} \beta_l \leq \beta_{l-1} - d - 1 \\ \vdots \\ \beta_l \leq \beta_1 - d - 1, \end{gathered}$$

which reduces to

$$\beta_l \leq \beta_{l-1} - d - 1 \tag{III-5.21}$$

thanks to our assumption $\beta_1 \geq \cdots \geq \beta_{l-1}$. By induction on $l = j, j-1, \ldots, 1$, we get conditions on the β_l (recall we let $\beta_1 = r$):

$$\beta_l \leq \beta_1 - (l-1)(d+1) = r - (l-1)(d+1).$$

In order to have $\beta_j \geq 0$, we need to have $r \geq (j-1)(d+1)$ which is one of our assumptions on r. Let $p \leq \beta_l$,

$$\begin{aligned}
&\left|\left|\partial_t^l (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \leq \left|\left|v \cdot \nabla_x \partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-l}} + \left|\left|\partial_t^{l-1}\left(E \cdot \nabla_v f\right) - \partial_t^{l-1}\left(E^K \cdot \nabla_v f^K\right)\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \\
&\leq \left|\left|\partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \left|\left|\partial_t^m E \cdot \nabla_v \partial_t^{l-1-m} f - \partial_t^m E^K \cdot \nabla_v \partial_t^{l-1-m} f^K\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \\
&\leq \left|\left|\partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \left|\left|\partial_t^m E \cdot \nabla_v \partial_t^{l-1-m} f - \partial_t^m E^K \cdot \nabla_v \partial_t^{l-1-m} f^K\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \\
&\leq \left|\left|\partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \begin{pmatrix} \left|\left|(\partial_t^m E - \partial_t^m E^K) \cdot \nabla_v \partial_t^{l-1-m} f\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \\ + \left|\left|\partial_t^m E^K \cdot \nabla_v \partial_t^{l-1-m} (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \end{pmatrix} \\
&\leq \left|\left|\partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \begin{pmatrix} \left|\left|(E[\partial_t^m f] - E^K[\partial_t^m f^K]) \cdot \nabla_v \partial_t^{l-1-m} f\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \\ + \left|\left|\partial_t^m E^K \cdot \nabla_v \partial_t^{l-1-m} (f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-l}} \end{pmatrix} \\
&\leq \left|\left|\partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} \\
&\qquad + \sum_{m=0}^{l-1} \binom{l-1}{m} \begin{pmatrix} \max\limits_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \left(E[\partial_t^m f] - E^K[\partial_t^m f^K]\right)\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \left|\left|\partial_t^{l-1-m} f\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-l}} \\ + \max\limits_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \left|\left|\partial_t^{l-1-m}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-l}} \end{pmatrix}.
\end{aligned} \tag{III-5.22}$$

Recall the assumption (III-5.20):

$$\forall m \leq l-1,\, p \leq \beta_m,\, \left|\left|\partial_t^m (f-f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-m}} \leq \frac{C}{(1+K)^{\gamma_m + \beta_m - p}},$$

thus

$$\begin{aligned}
&\left|\left|\partial_t^{l-1}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \max_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \left|\left|\partial_t^{l-1-m}(f-f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-l}} \\
&\leq \frac{C}{(1+K)^{\gamma_{l-1} + \beta_{l-1} - p - 1}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \max_{\substack{c \in \mathbb{N}^d \\ \bar{c} \leq p}} \left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \frac{C}{(1+K)^{\gamma_{l-1-m} + \beta_{l-1-m} - p - 1}}.
\end{aligned}$$

By our previous estimates, $\left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}^d_L)} \leq C$ for any $c \in \mathbb{N}^d$ with $\bar{c} \leq \beta_m - d - 1$.

Thanks to the ordering $\beta_1 \geq \dots \geq \beta_{l-1}$, this implies $\left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}_L^d)} \leq C$ for any $c \in \mathbb{N}^d$ with $\bar{c} \leq \beta_{l-1} - d - 1$. Now, since we are considering $p \leq \beta_l \leq \beta_l - d - 1$, we obtain

$$\begin{aligned}&\left|\left|\partial_t^{l-1}(f - f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \max_{\substack{c\in\mathbb{N}^d \\ \bar{c}\leq p}} \left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}_L^d)} \left|\left|\partial_t^{l-1-m}(f - f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-l}} \\ &\leq \frac{C}{(1+K)^{\gamma_{l-1}+\beta_{l-1}-p-1}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \frac{C}{(1+K)^{\gamma_{l-1-m}+\beta_{l-1-m}-p-1}}.\end{aligned}$$

Moreover, for any $m \leq l-1$, suppose that

$$\gamma_{l-1} + \beta_{l-1} \leq \gamma_m + \beta_m, \tag{III-5.23}$$

so that by (III-5.20),

$$\begin{aligned}&\left|\left|\partial_t^{l-1}(f - f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-(l-1)}} + \sum_{m=0}^{l-1} \binom{l-1}{m} \max_{\substack{c\in\mathbb{N}^d \\ \bar{c}\leq p}} \left|\left|\partial_x^c \partial_t^m E^K\right|\right|_{\mathbb{L}^\infty(\mathbb{T}_L^d)} \left|\left|\partial_t^{l-1-m}(f - f^K)\right|\right|_{\mathcal{H}^{p+1}_{\nu+j-l}} \\ &\leq \frac{C}{(1+K)^{\gamma_{l-1}+\beta_{l-1}-p-1}}.\end{aligned}$$

It remains to estimate the first term in the sum of (III-5.22). Again, by (III-5.20) we have

$$\left|\left|\partial_t^m(f - f^K)\right|\right|_{\mathcal{H}^0_{\nu+j-m}} \leq \frac{C}{(1+K)^{\gamma_m+\beta_m}},$$

for $m \leq l-1$, therefore Proposition III.3 gives

$$\begin{aligned}\max_{\substack{c\in\mathbb{N}^d \\ \bar{c}\leq p}} \left|\left|\partial_x^c\left(E[\partial_t^m f] - E^K[\partial_t^m f^K]\right)\right|\right|_{\mathbb{L}^\infty(\mathbb{T}_L^d)} &\leq \frac{C}{(1+K)^{\gamma_m+\beta_m+\frac{1}{2}-d-p}} \\ &\leq \frac{C}{(1+K)^{\gamma_{l-1}+\beta_{l-1}+\frac{1}{2}-d-p}}.\end{aligned}$$

Finally, using the condition (III-5.21),

$$\begin{aligned}\left|\left|\partial_t^l(f - f^K)\right|\right|_{\mathcal{H}^p_{\nu+j-l}} &\leq \frac{C}{(1+K)^{\gamma_{l-1}+\beta_{l-1}+\min(1/2-d,-1)-p}} \\ &\leq \frac{C}{(1+K)^{\gamma_{l-1}+\beta_l+d+1+\min(1/2-d,-1)-p}} \\ &\leq \frac{C}{(1+K)^{\gamma_{l-1}+\beta_l+\min(3/2,d)-p}}\end{aligned}$$

Set $\gamma_l = \gamma_{l-1} + \min(3/2, d)$. We check that this choice of γ_l satisfies (III-5.23):

$$\begin{aligned}\gamma_{l-1} + \beta_{l-1} &= \gamma_m + (l-1-m)\min\left(\frac{3}{2}, d\right) + \beta_{l-1}\\ &\leq \gamma_m + (l-1-m)\min\left(\frac{3}{2}, d\right) + \beta_m - (l-1-m)(d+1)\\ &= \gamma_m + \beta_m + (l-1-m)\left[\min\left(\frac{3}{2}, d\right) - (d+1)\right].\end{aligned}$$

Since $d \in \mathbb{N}^*$, we have $\min(3/2, d) \leq d+1$ and therefore

$$\gamma_{l-1} + \beta_{l-1} \leq \gamma_m + \beta_m.$$

Moreover, $\gamma_1 = \frac{\alpha+1}{2} - \beta_1 - d$ and we require $\gamma_1 > 0$, i.e.

$$\frac{\alpha+1}{2} - \beta_1 - d > 0 \qquad \Longleftrightarrow \quad \alpha > 2(\beta_1 + d) - 1 = 2(r+d) - 1,$$

which is guaranteed to hold since we assume $\alpha \geq 2(r+d)$. □

We recall from Section III-4.2.1 that q_i is the order of the quadrature along dimension i and that Δz_i is the quadrature step along the i-th dimension, $1 \leq i \leq 2d$. We recall as well that the coefficients C_k^K, S_k^K are defined by (III-4.21), and the coefficients $C_k^{K,h}, S_k^{K,h}$ by (III-4.22). We have the following estimates on the quadrature error:

Proposition III.5

Let $j \in \mathbb{N}$ such that $j \geq 1 + \max_i q_i$, and $\nu, r, \alpha \in \mathbb{N}$ such that $\nu + j > d/2$, $r \geq \max\left(3(\nu+j), (j-1)(d+1)\right)$, and $\alpha \geq 2(r+d)$. Let $K \in \mathbb{N}$, and assume $f_0 \in \mathcal{H}^{r+\alpha}_{\nu+j}$. Then there exists a constant $C > 0$ such that the following holds: for $\delta \geq 0$, define finite intervals $I_{d+1} := [a_1, b_1], \dots, I_{2d} = [a_d, b_d]$ and $I_v := I_{d+1} \times \dots \times I_{2d}$ such that

$$||f_0||_{\mathcal{H}^0_\nu(\mathbb{T}^d_L \times (\mathbb{R}^d \setminus I_v))} \leq \delta.$$

Then for all $k \in (\mathbb{Z}^d)^*$ and $K \in \mathbb{N}^*$, we have

$$\left|C_k^K(t) - C_k^{K,h}(t)\right| \leq C\delta + C\sum_{i=1}^{2d}\left(1 + C\frac{2\pi(q_i+1)}{\ln(q_i+2)}\left|\frac{k}{L}\right|\right)^{q_i+1}\Delta z_i^{q_i} \qquad \text{(III-5.24)}$$

and

$$\left|S_k^K(t) - S_k^{K,h}(t)\right| \le C\delta + C\sum_{i=1}^{2d}\left(1 + C\frac{2\pi(q_i+1)}{\ln(q_i+2)}\left|\frac{k}{L}\right|\right)^{q_i+1}\Delta z_i^{q_i} \qquad \text{(III-5.25)}$$

where the estimates are uniform in time, and C does not depend on Δz_i.
As a consequence, for C large enough, we have the following estimates:

$$\left|C_k^K(t) - C_k^{K,h}(t)\right| \le C\delta + C\sum_{i=1}^{2d}(1 + C|k|)^{q_i+1}\Delta z_i^{q_i} \qquad \text{(III-5.26)}$$

and

$$\left|S_k^K(t) - S_k^{K,h}(t)\right| \le C\delta + C\sum_{i=1}^{2d}(1 + C|k|)^{q_i+1}\Delta z_i^{q_i}. \qquad \text{(III-5.27)}$$

Proof. We prove only the error estimate for $\left|C_k^K(t) - C_k^{K,h}(t)\right|$, since the treatment is exactly the same for $\left|S_k^K(t) - S_k^{K,h}(t)\right|$.

First of all, by the regularity assumption on f_0 we know from Proposition III.4 that $E^K \in C^j([0,T]\times\mathbb{R}^d)$. Therefore, the characteristics $(X^K, V^K) \in C^j([0,T]\times\mathbb{T}^d\times\mathbb{R}^d)$, so that the j-th space derivative is continuous in time.

The quadratures in velocity will be performed on the intervals $I_{d+i} = [a_i, b_i], i = 1, \dots, d$, and the quadratures in space will be performed on $\mathbb{T}^d$. To make notations clearer and more general, define $I_i := [0, L_i]$ for $i = 1, \dots, d$.

For $n = 1, \cdots, 2d$, we define

$$\tilde{z}_n := (z_n, \dots, z_{2d}) \in I_n \times \cdots \times I_{2d},$$

$$g_t(\tilde{z}_n) := \int_{I_1\times\cdots\times I_{n-1}} \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right) f_0(z)dz_1\cdots dz_{n-1},$$

$$h_t(\tilde{z}_n) = \sum_{j_1,\dots,j_{n-1}} w_1^{j_1}\cdots w_{n-1}^{j_{n-1}}\cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z_1^{j_1},\cdots,z_{n-1}^{j_{n-1}},\tilde{z}_n)\right) f_0(z_1^{j_1},\cdots,z_{n-1}^{j_{n-1}},\tilde{z}_n).$$

We will prove the estimates (III-5.24) and (III-5.25) by induction on the number of dimensions.

Base case For a fixed $\tilde{z}_2 \in I_2 \times \dots \times I_{2d}$, the quadrature along the first dimension gives

$$\left| \int_{I_1} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z_1,\tilde{z}_2)\right) f(0,z_1,\tilde{z}_2) dz_1 - \sum_{j_1} w_1^{j_1} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z_1^{j_1},\tilde{z}_2)\right) f_0(z_1^{j_1},\tilde{z}_2) \right| \\ \leq C \Delta z_1{}^{q_1} \left\| \partial_{z_1}^{q_1+1} \left[\cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,\cdot,\tilde{z}_2)\right) f_0(\cdot,\tilde{z}_2)\right] \right\|_{\mathbb{L}^\infty(I_1)} . \tag{III-5.28}$$

We proceed to estimate the right-hand side, and consider a derivative along the n-th dimension instead of only along the first dimension:

$$\partial_{z_n}^{q_n+1} \left[\cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z)\right) f_0(z)\right] = \sum_{l=0}^{q_n+1} \binom{q_n+1}{l} \partial_{z_n}^{q_n+1-l} f_0(z) \partial_{z_n}^{l} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z)\right). \tag{III-5.29}$$

By the Faà di Bruno formula (see [2, Sect. 24.1.2]), we have

$$\partial_{z_n}^{l} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z)\right) \\ = \sum_{m=0}^{l} \cos^{(m)}\left(2\pi \frac{k}{L} \cdot X^K(t;0,z)\right) \sum (l;a_1,\dots,a_l)' \prod_{c=1}^{l} \left(2\pi \frac{k}{L} \cdot \partial_{z_n}^{c} X^K(t;0,z)\right)^{a_c},$$

where the unindexed sum is performed over all l-tuples $(a_1,\dots,a_l)$ such that

$$a_1 + 2a_2 + \dots + l a_l = l \quad \text{and} \quad a_1 + a_2 + \dots + a_l = m.$$

The sum $\sum (l;a_1,\dots,a_l)'$ is also called a Stirling number of the Second kind, of parameters (n,m). It counts the number of ways of partitioning a set of l elements into m non-empty subsets. We have

$$\begin{aligned} \left| \partial_{z_n}^{l} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z)\right)\right| &\leq \sum_{m=0}^{l} \sum (l;a_1,\dots,a_l)' \prod_{c=1}^{l} \left| 2\pi \frac{k}{L} \cdot \partial_{z_n}^{c} X^K(t;0,z)\right|^{a_c} \\ &\leq \sum_{m=0}^{l} \sum (l;a_1,\dots,a_l)' \prod_{c=1}^{l} \left| 2\pi \frac{k}{L}\right|_2^{a_c} \left| \partial_{z_n}^{c} X^K(t;0,z)\right|^{a_c} \\ &\leq (2\pi)^l \left|\frac{k}{L}\right|_2^l \sum_{m=0}^{l} \sum (l;a_1,\dots,a_l)' \prod_{c=1}^{l} \left| \partial_{z_n}^{c} X^K(t;0,y,z)\right|^{a_c}, \end{aligned}$$

where the second inequality has been obtained by the discrete Cauchy-Schwarz inequality.

Since the characteristics X^K is of class $C^j([0,T] \times \mathbb{T}_L^d \times \mathbb{R}^d)$, with $j \geq \max_i q_i + 1$, we know there exists a constant $C_{f_0}(K)$ that depends possibly on K such that for all $\mathbb{N} \ni c \leq q_n + 1$,

$$\left\| \partial_{z_n}^{c} X^K \right\|_{\mathbb{L}^\infty([0,T]\times\mathbb{T}_L^d \times I_v)} \leq C_{f_0}(K).$$

However, we want this constant C_{f_0} to be independent of K, and to be able to choose such a constant, we notice that as $K \to \infty$, we recover the non-truncated Vlasov-Poisson system's characteristics. For these characteristics, thanks to the regularity assumption, we know that there exists a constant denoted $C_{f_0}(\infty)$ such that

$$\left\|\partial_{z_n}^c X\right\|_{\mathbb{L}^\infty([0,T]\times\mathbb{T}_L^d\times I_v)} \leq C_{f_0}(\infty) < \infty.$$

So we can build a sequence of constants $\{C_{f_0}(K)\}_{K\geq 1}$ which is bounded. Then define $C_{f_0} := \max_{K\geq 1} C_{f_0}(K)$, and we have

$$\left\|\partial_{z_n}^c X^K\right\|_{\mathbb{L}^\infty([0,T]\times\mathbb{T}_L^d\times I_v)} \leq C_{f_0}.$$

Hence, for all $t \in [0,T]$,

$$\left|\partial_{z_n}^l \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right)\right| \leq C_{f_0}^l\,(2\pi)^l\left|\frac{k}{L}\right|^l \sum_{m=0}^{l}\sum(l;a_1,\dots,a_l)'.$$

The remaining sums correspond to the Bell number B_l, and it counts the number of ways to partition a set that has exactly l elements. We have the following bound (see [17]):

$$B_l \leq \left(\frac{0.792l}{\ln(l+1)}\right)^l \leq \frac{l^l}{(\ln(l+1))^l}.$$

Therefore,

$$\left|\partial_{z_n}^l \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right)\right| \leq C_{f_0}^l\left(\frac{2\pi l}{\ln(l+1)}\right)^l\left|\frac{k}{L}\right|^l \leq C_{f_0}^l\left(\frac{2\pi(q_n+1)}{\ln(q_n+2)}\right)^l\left|\frac{k}{L}\right|^l.$$

By regularity of the initial condition f_0, we can choose the constant C_{f_0} large enough so that for all $n = 1,\cdots,2d$,

$$\left\|\partial_{z_n}^l f_0\right\|_{\mathbb{L}^\infty(I_1\times\cdots\times I_{2d})} \leq C_{f_0}, \quad l = 0,\dots,q_n+1.$$

We then get from (III-5.29)

$$\begin{aligned}
&\left|\partial_{z_n}^{q_n+1}\left[\cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right)f_0(z)\right]\right| \\
&\leq C_{f_0}\sum_{l=0}^{q_n+1}\binom{q_n+1}{l}\left\|\partial_{z_n}^l \cos\left(2\pi\frac{k}{L}\cdot X(t;0,z)\right)\right\|_{\mathbb{L}^\infty(I_1\times\cdots\times I_{2d})} \\
&\leq C_{f_0}\left(1 + C_{f_0}\frac{2\pi(q_n+1)}{\ln(q_n+2)}\left|\frac{k}{L}\right|\right)^{q_n+1}
\end{aligned}$$

Note that the right-hand side does not depend on y or v, hence

$$\left\| \partial_{z_n}^{q_n+1} \left[\cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,\cdot) \right) f_0(\cdot) \right] \right\|_{\mathbb{L}^\infty(I_1\times\cdots\times I_{2d})} \leq C_{f_0} \left(1 + C_{f_0} \frac{2\pi(q_n+1)}{\ln(q_n+2)} \left| \frac{k}{L} \right| \right)^{q_n+1}. \tag{III-5.30}$$

Plugging this estimate with $n = 1$ back into (III-5.28), we obtain

$$\left| \int_{I_1} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z_1,\tilde{z}_2) \right) f_0(z_1,\tilde{z}_2) dz_1 - \sum_{j_1} w_1^{j_1} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z_1^{j_1},\tilde{z}_2) \right) f_0(z_1^{j_1},\tilde{z}_2) \right|$$
$$= |g_t(\tilde{z}_2) - h_t(\tilde{z}_2)| \leq C \left(1 + \frac{2\pi(q_1+1)}{\ln(q_1+2)} \left| \frac{k}{L} \right| \right)^{q_1+1} \Delta z_1^{q_1}, \tag{III-5.31}$$

where the constant C does not depend on $k, \Delta z_1, q_1, \tilde{z}_2$.

Induction step We have

$$|g_t(\tilde{z}_{n+1}) - h_t(\tilde{z}_{n+1})| = \left| \int_{I_n} g_t(z_n, \tilde{z}_{n+1}) dz_n - \sum_{j_n} w_n^{j_n} h_t(z_n^{j_n}, \tilde{z}_{n+1}) \right|$$
$$\leq \int_{I_n} |g_t(z_n,\tilde{z}_{n+1}) - h_t(z_n,\tilde{z}_{n+1})| \, dz_n + \left| \int_{I_n} h_t(z_n,\tilde{z}_{n+1}) dz_n - \sum_{j_n} w_n^{j_n} h_t(z_n^{j_n}, \tilde{z}_{n+1}) \right|. \tag{III-5.32}$$

The first term on the right-hand side can be bounded using the previous step in the induction, which is assumed to give the following estimate:

$$|g_t(\tilde{z}_n) - h_t(\tilde{z}_n)| \leq C \sum_{i=1}^{n-1} \left(1 + C \frac{2\pi(q_i+1)}{\ln(q_i+2)} \left| \frac{k}{L} \right| \right)^{q_i+1} \Delta z_i^{q_i}.$$

Since the right-hand side does not depend on $\tilde{z}_n$, we get

$$\int_{I_n} |g_t(z_n,\tilde{z}_{n+1}) - h_t(z_n,\tilde{z}_{n+1})| \, dz_n \leq |I_n| \, ||g_t(\tilde{z}_n) - h_t(\tilde{z}_n)||_{\mathbb{L}^\infty(I_n\times\cdots\times I_{2d})}$$
$$\leq C \sum_{i=1}^{n-1} \left(1 + C \frac{2\pi(q_i+1)}{\ln(q_i+2)} \left| \frac{k}{L} \right| \right)^{q_i+1} \Delta z_i^{q_i},$$

where the constant C does not depend on $k, \Delta z_i, q_i, \tilde{z}_{n+1}$.

It remains only to estimate the second term on the right-hand side of (III-5.32). We notice that it correspond to the quadrature error of the function $z_n \mapsto h_t(z_n, \tilde{z}_{n+1})$ over

I_n. Thus,

$$\left|\int_{I_n} h_t(z_n, \tilde{z}_{n+1}) dz_n - \sum_{j_n} w_n^{j_n} h_t(z_n^{j_n}, \tilde{z}_{n+1})\right| \leq C \left\|\partial_{z_n}^{q_n+1} h_t(\cdot, \tilde{z}_{n+1})\right\|_{\mathbb{L}^\infty(I_n)} \Delta z_n^{q_n}. \tag{III-5.33}$$

We have

$$\begin{aligned}
&\partial_{z_n}^{q_n+1} h_t(z_n, \tilde{z}_{n+1}) = \partial_{z_n}^{q_n+1} h_t(\tilde{z}_n) \\
&= \sum_{j_1,\cdots,j_{n-1}} w_1^{j_1} \cdots w_{n-1}^{j_{n-1}} \partial_{z_n}^{q_n+1} \left[\cos\left(2\pi \frac{k}{L} \cdot X^K(t; 0, z_1^{j_1}, \cdots, z_{n-1}^{j_{n-1}}, \tilde{z}_n)\right) f_0(z_1^{j_1}, \cdots, z_{n-1}^{j_{n-1}}, \tilde{z}_n)\right],
\end{aligned}$$

and hence

$$\begin{aligned}
&\left\|\partial_{z_n}^{q_n+1} h_t(\cdot, \tilde{z}_{n+1})\right\|_{\mathbb{L}^\infty(I_n)} \\
&\leq \sum_{j_1,\cdots,j_{n-1}} \left|w_1^{j_1} \cdots w_{n_1}^{j_{n-1}}\right| \\
&\qquad\qquad \left\|\partial_{z_n}^{q_n+1} \left[\cos\left(2\pi \frac{k}{L} \cdot X^K(t; 0, z_1^{j_1}, \cdots, z_{n-1}^{j_{n-1}}, \tilde{z}_n)\right) f_0(z_1^{j_1}, \cdots, z_{n-1}^{j_{n-1}}, \tilde{z}_n)\right]\right\|_{\mathbb{L}^\infty_{z_n}(I_n)} \\
&\leq \sum_{j_1,\cdots,j_{n-1}} \left|w_1^{j_1} \cdots w_{n-1}^{j_{n-1}}\right| \left\|\partial_{z_n}^{q_n+1} \left[\cos\left(2\pi \frac{k}{L} \cdot X^K(t; 0, \cdot)\right) f_0(\cdot)\right]\right\|_{\mathbb{L}^\infty(I_1 \times \cdots \times I_{2d})}.
\end{aligned}$$

By (III-5.30), we get

$$\left\|\partial_{z_n}^{q_n+1} h_t(\cdot, \tilde{z}_{n+1})\right\|_{\mathbb{L}^\infty(I_n)} \leq C \left(1 + C \frac{2\pi(q_n+1)}{\ln(q_n+2)} \left|\frac{k}{L}\right|\right)^{q_n+1} \sum_{j_1,\cdots,j_{n-1}} \left|w_1^{j_1} \cdots w_{n-1}^{j_{n-1}}\right|.$$

Moreover, since the weights are nonnegative,

$$\sum_{j_1,\cdots,j_{n-1}} \left|w_1^{j_1} \cdots w_{n-1}^{j_{n-1}}\right| = \sum_{j_1,\cdots,j_{n-1}} w_1^{j_1} \cdots w_{n-1}^{j_{n-1}}.$$

The right-hand side corresponds to an approximation of the constant function equal to one on the hyperrectangle $I_1 \times \cdots \times I_{n-1}$, hence the quadrature is exact and the value of the sum corresponds to the volume of the hyperrectangle. Therefore,

$$\left\|\partial_{z_n}^{q_n+1} h_t(\cdot, \tilde{z}_{n+1})\right\|_{\mathbb{L}^\infty(I_n)} \leq C \left(1 + C \frac{2\pi(q_n+1)}{\ln(q_n+2)} \left|\frac{k}{L}\right|\right)^{q_n+1}.$$

We can plug this into (III-5.33) to get

$$\left|\int_{I_n} h_t(z_n, \tilde{z}_{n+1}) dz_n - \sum_{j_n} w_n^{j_n} h_t(z^{j_n}, \tilde{z}_{n+1})\right| \leq C \left(1 + C \frac{2\pi(q_n+1)}{\ln(q_n+2)} \left|\frac{k}{L}\right|\right)^{q_n+1} \Delta z_n^{q_n}.$$

Finally, we obtain from (III-5.32)

$$\left|g_t(\tilde{z}_{n+1}) - h_t(\tilde{z}_{n+1})\right| \leq C \sum_{i=1}^{n} \left(1 + C\frac{2\pi(q_i+1)}{\ln(q_i+2)}\left|\frac{k}{L}\right|\right)^{q_i+1} \Delta z_i^{q_i}.$$

This achieves the induction step, so that this inequality holds for all $n = 1, \dots, 2d$.

When $n = 2d$,

$$\left|\int_{\mathbb{T}_L^d \times I_v} \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right) f_0(z)dz - C_k^{K,h}(t)\right| \leq C\sum_{i=1}^{2d}\left(1 + C\frac{2\pi(q_i+1)}{\ln(q_i+2)}\left|\frac{k}{L}\right|\right)^{q_i+1} \Delta z_i^{q_i},$$

where the constant C does not depend on $k, \Delta x, \Delta v, q_x, q_v$. Finally, by definition of the intervals I_i, we have

$$C_k(t) = \int_{\mathbb{T}^d \times I_v} \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right) f_0(z)dz + \int_{\mathbb{T}^d \times (\mathbb{R}^d \setminus I_v)} \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right) f_0(z)dz.$$

The second term on the left-hand side can be handled by using the fact that $f_0 \in \mathcal{H}_\nu^{r+2\nu+1}$, so that

$$\begin{aligned}
&\left|\int_{\mathbb{T}^d \times (\mathbb{R}^d \setminus I_v)} \cos\left(2\pi\frac{k}{L}\cdot X^K(t;0,z)\right) f_0(z)dz\right| \\
&\leq \int_{\mathbb{T}^d \times (\mathbb{R}^d \setminus I_v)} |f_0(z)|dz \\
&\leq \left(\int_{\mathbb{T}^d \times (\mathbb{R}^d \setminus I_v)} \frac{1}{(1+|v|^2)^\nu} dxdv\right)\left(\int_{\mathbb{T}^d \times (\mathbb{R}^d \setminus I_v)} (1+|v|^2)^\nu |f_0(x,v)|^2 dxdv\right) \\
&\leq C||f_0||_{\mathcal{H}_\nu^0(\mathbb{T}_L^d \times (\mathbb{R}^d \setminus I_v))} \leq C\delta
\end{aligned}$$

This achieves to show our claimed estimates. □

Finally, we are able to prove the convergence result.

Proof of Theorem III.2. We first show that any r-order time integration scheme for second order ODEs can be applied, then proceed to the claimed estimate. Throughout this proof we denote by C a quantity which is independent from $t, n, \Delta t, \Delta z_i, K$, its value may change from line to line.

Recall the the characteristics of the Vlasov equation with a truncated Fourier kernel:

$$\begin{cases} \dfrac{d}{dt}X^K(t;0,x,v) = V^K(t;0,x,v) \\ \dfrac{d}{dt}V^K(t;0,x,v) = E^K(t, X^K(t;0,x,v)) \end{cases}$$

where E^K is defined by (III-4.17). Therefore,

$$\frac{d^2}{dt^2} X^K(t;0,x,v) = E^K(t, X^K(t;0,x,v)).$$

However this function E^K is not a function we can compute in practice in the Weighted Particle method, since it requires a knowledge of the mapping $(x,v) \to (X^K, V^K)(t;t^0,x,v)$ for all $(x,v) \in \mathbb{T}^d_L \times \mathbb{R}^d$ in order to compute $C^K_k(t)$ and $S^K_k(t)$. We instead use the approximations $C^{K,h}_k, S^{K,h}_k$ of C^K_k, S^K_k, given in (III-4.22):

$$\begin{aligned} C^{K,h}_k(t) &= \sum_{j\in J} \cos\left(2\pi \frac{k}{L} \cdot X^K(t;0,z^j)\right) f(0,z^j) w^j, \\ S^{K,h}_k(t) &= \sum_{j\in J} \sin\left(2\pi \frac{k}{L} \cdot X^K(t;0,z^j)\right) f(0,z^j) w^j. \end{aligned} \tag{III-5.34}$$

We recall that from these approximate coefficients, we defined in (III-4.23) an approximate kernel $E^{K,h}$:

$$E^{K,h}(t,x) = \frac{1}{|\mathbb{T}^d_L|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\le K}} \frac{1}{2\pi \left|\frac{k}{L}\right|^2} \frac{k}{L} \left[\sin\left(2\pi k \cdot \frac{x}{L}\right) C^{K,h}_k(t) - \cos\left(2\pi k \cdot \frac{x}{L}\right) S^{K,h}_k(t)\right].$$

Let $p = 1, \dots, P$, the quantity $X^K_p(t^n)$, defined in (III-4.25), is the solution to the second-order ODE:

$$\frac{d^2}{dt^2} X^K_p(t) = E^{K,h}(t, X^K_p(t)), \qquad X^K_p(t^0) = x_p.$$

Moreover we have

$$E^K(t,x) = E^{K,h}(t,x) + (\delta E)^K(t,x) \tag{III-5.35}$$

where

$$\begin{aligned} &(\delta E)^K(t,y) := E^K(t,y) - E^{K,h}(t,y) \\ &= \frac{1}{|\mathbb{T}^d_L|} \sum_{\substack{k\in(\mathbb{Z}^d)^* \\ |k|\le K}} \frac{1}{2\pi \left|\frac{k}{L}\right|^2} \frac{k}{L} \left[\sin\left(2\pi k \cdot \frac{y}{L}\right) (C^{K,h}_k - C^K_k)(t) - \cos\left(2\pi k \cdot \frac{y}{L}\right) (S^{K,h}_k - S^K_k)(t)\right] \end{aligned}$$

Thus we deduce

$$
\begin{aligned}
|(\delta E)^K(t,y)| &\le \frac{1}{|\mathbb{T}_L^d|} \sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\le K}} \frac{1}{2\pi\left|\frac{k}{L}\right|^2}\left|\frac{k}{L}\right|\left(\left|C_k^{K,h}-C_k^K\right|(t)+\left|S_k^{K,h}-S_k^K\right|(t)\right)\\
&\le C\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\le K}}\frac{1}{|k|}\left(\left|(C_k^{K,h}-C_k^K)\right|(t)+\left|(S_k^{K,h}-S_k^K)\right|(t)\right)\\
&\le CK^d\delta + C\sum_{i=1}^{2d}\Delta z_i^{q_i}\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\le K}}\frac{(1+C|k|)^{q_i+1}}{|k|}\\
&\le CK^d\left(\delta+\sum_{i=1}^{2d}\Delta z_i^{q_i}K^{q_i}\right)=:\mathcal{E}(K,\Delta x,\Delta v)
\end{aligned}
\tag{III-5.36}
$$

where the third inequality is from (III-5.26) and (III-5.27). The exact characteristics $X^K(t;t^0,x_p,v_p)$, defined in (III-4.20), then satisfy

$$
\frac{d^2}{dt^2}X^K(t;t^0,x_p,v_p)=E^{K,h}(X^K(t;t^0,x_p,v_p))+(\delta E)^K(t,X^K(t;t^0,x_p,v_p)).
$$

We recall inequality (III-4.27), so that we can prove the claimed result in three steps, each one corresponding to a line of this inequality. Each line corresponds to a different type of approximation: the first one is the time discretization error, the second one the phase-space discretization error, and the third one the kernel truncature error.

Because $E^{K,h}$ is more appropriately dealt with by vector variables, we use the following notations: $\mathbb{V}^K(t):=\frac{d}{dt}\mathbb{X}^K(t)$, $\mathcal{X}^K(t):=(X^K(t;t^0,x_1,v_1),\dots,X^K(t;t^0,x_P,v_P))$, and $\mathcal{V}^K(t):=\frac{d}{dt}\mathcal{X}^K(t)$.

Step 1: time discretization error Notice that the time dependence of the function $E^{K,h}$ is only due to the time dependence of the finite-dimensional vector $\mathbb{X}^K(t)\in\mathbb{R}^{dP}$. Therefore we may write $C_k^{K,h}(t)\equiv C_k^{K,h}(\mathbb{X}^K(t))$ by abuse of notations, in which case $C_k^{K,h}(\mathbb{x})$ is a $C^\infty(\mathbb{R}^{dP},\mathbb{R})$ function of $\mathbb{x}$. It is possible to write $\frac{d^2}{dt^2}\mathbb{X}^K(t)=\mathbf{E}^{K,h}(\mathbb{X}^K(t))$ for some function

$$
\begin{aligned}
&\mathbb{R}^{dP}\to\mathbb{R}^{dP}\\
\mathbb{x}=(x_1,\dots,x_P)&\mapsto\mathbf{E}^{K,h}(\mathbb{x})=\left(\mathbf{E}_1^{K,h}(\mathbb{x}),\dots,\mathbf{E}_P^{K,h}(\mathbb{x})\right)
\end{aligned}
$$

where, for $i=1,\dots,P$ we let $x_i\in\mathbb{R}^d$ and

$$
\mathbb{R}^d\ni\mathbf{E}_i^{K,h}(\mathbb{x})=\frac{1}{|\mathbb{T}_L^d|}\sum_{\substack{k\in(\mathbb{Z}^d)^*\\|k|\le K}}\frac{1}{2\pi\left|\frac{k}{L}\right|^2}\frac{k}{L}\left[\sin\left(2\pi k\cdot\frac{x_i}{L}\right)C_k^{K,h}(\mathbb{x})-\cos\left(2\pi k\cdot\frac{x_i}{L}\right)S_k^{K,h}(\mathbb{x})\right].
$$

The coefficients $C_k^{K,h}(\mathbb{x})$ and $S_k^{K,h}(\mathbb{x})$ are defined by

$$C_k^{K,h}(\mathbb{x}) = \sum_{p=1}^{P} \cos\left(2\pi k \cdot \frac{\mathbb{x}_p}{L}\right)\beta_p,$$
$$S_k^{K,h}(\mathbb{x}) = \sum_{p=1}^{P} \sin\left(2\pi k \cdot \frac{\mathbb{x}_p}{L}\right)\beta_p.$$

Therefore, the mapping $\left(\mathbb{x} \mapsto \mathbf{E}^K(\mathbb{x})\right) \in C^\infty(\mathbb{R}^{dP}, \mathbb{R}^{dP})$. Moreover, from the definition of the characteristics $(\mathbb{X}^K, \mathbb{V}^K)$, we have

$$\begin{cases} \dfrac{d}{dt}\mathbb{X}^K(t) = \mathbb{V}^K(t) \\ \dfrac{d}{dt}\mathbb{V}^K(t) = \mathbf{E}^{K,h}(\mathbb{X}^K(t)) \end{cases} \tag{III-5.37}$$

The right-hand side is a $C^\infty(\mathbb{R}^{2dP}, \mathbb{R}^{2dP})$ function of $(\mathbb{X}^K(t), \mathbb{V}^K(t))$, therefore we know that the characteristics $t \mapsto (\mathbb{X}^K(t), \mathbb{V}^K(t))$ are $C^\infty([0,T])$.

In order to apply the error estimate for the time integration scheme to solve second-order ODE, we recall that the error depends on the $(\gamma+1)$-th derivative of the function $x \mapsto \mathbf{E}^{K,h}(x)$. If the time integration scheme solves first-order ODEs, the error would depend on the $(\gamma+1) - th$ derivative of the function $(x, v) \mapsto (v, \mathbf{E}^{K,h}(x))$.

It can be shown with the Faà di Bruno formula that for any $l \in \mathbb{N}^{dP}, |l| \leq \gamma$,

$$\left\|\partial_{\mathbb{x}}^l \left[\sin\left(2\pi k \cdot \frac{x_i}{L}\right) C_k^{K,h}(\mathbb{x}) - \cos\left(2\pi k \cdot \frac{x_i}{L}\right) S_k^{K,h}(\mathbb{x})\right]\right\|_{\mathbb{L}^\infty(\mathbb{R}^{dP})} \leq CK^{\gamma+1}, \tag{III-5.38}$$

where the constant C does not depend on K.

Therefore, no matter if the time integration scheme approximates first-order or second-order ODEs, we obtain for any $n = 1, \ldots, N_t$

$$\max_{p=1,\ldots,P}\left(\left|X_p^{K,n} - X_p^K(t^n)\right| + \left|V_p^{K,n} - V_p^K(t^n)\right|\right) \leq CK^{d+\gamma+1}\Delta t^\gamma \tag{III-5.39}$$

where the constant C does not depend on K or Δt.

Step 2: phase-space discretization The assumptions that characteristics and their approximations have the same initial conditions can be rewritten as $\mathbb{X}^K(t^0) = \mathcal{X}^K(t^0)$

and $\mathbb{V}^K(t^0) = \mathcal{V}^K(t^0)$. We have, for $s \in [t^0, t^0+T]$,

$$\begin{aligned}\begin{pmatrix}\mathbb{X}^K(s)\\ \mathbb{V}^K(s)\end{pmatrix} &= \begin{pmatrix}\mathbb{X}^K(t^0)\\ \mathbb{V}^K(t^0)\end{pmatrix} + \int_{t^0}^s \begin{pmatrix}\mathbb{V}^K(\tau)\\ \mathbf{E}^{K,h}(\mathbb{X}^K(\tau))\end{pmatrix} d\tau\\ &= \begin{pmatrix}\mathbb{X}^K(t^0)\\ \mathbb{V}^K(t^0)\end{pmatrix} + \int_{t^0}^s \begin{pmatrix}\mathbb{V}^K(\tau)\\ E^K(\tau,\mathbb{X}^K(\tau))\end{pmatrix} d\tau + \int_{t^0}^s \begin{pmatrix}0\\ (\delta E)^K(\tau,\mathbb{X}^K(\tau))\end{pmatrix} d\tau.\end{aligned}$$

Note that we also have

$$\begin{pmatrix}\mathcal{X}^K(s)\\ \mathcal{V}^K(s)\end{pmatrix} = \begin{pmatrix}\mathbb{X}^K(t^0)\\ \mathbb{V}^K(t^0)\end{pmatrix} + \int_{t^0}^s \begin{pmatrix}\mathcal{V}^K(\tau)\\ E^K(\tau,\mathcal{X}^K(\tau))\end{pmatrix} d\tau,$$

so that

$$\begin{pmatrix}\mathbb{X}^K(s)\\ \mathbb{V}^K(s)\end{pmatrix} = \begin{pmatrix}\mathcal{X}^K(s)\\ \mathcal{V}^K(s)\end{pmatrix} + \int_{t^0}^s \begin{pmatrix}\mathbb{V}^K(\tau) - \mathcal{V}^K(\tau)\\ E^K(\mathbb{X}^K(\tau)) - E^K(\mathcal{X}^K(\tau))\end{pmatrix} d\tau + \int_{t^0}^s \begin{pmatrix}0\\ (\delta E)^K(\tau,\mathbb{X}^K(\tau))\end{pmatrix} d\tau.$$

From the mean value theorem we get:

$$\left|E^K(\tau,\mathbb{X}^K(\tau)) - E^K(\tau,\mathcal{X}^K(\tau))\right| \leq C \left(\sum_{\substack{k\in(\mathbb{Z}^d)^*\\ |k|\leq K}} |C_k^K(t)| + |S_k^K(t)| \right) \left|\mathbb{X}^K(\tau) - \mathcal{X}^K(\tau)\right|.$$

Using the fact that the function $f_0 \in \mathcal{H}^{r+\alpha}_{\nu+j}$, we can apply the same ideas as those leading to (III-5.16), in order to obtain

$$|C_k(t)| \leq \frac{C}{(1+|k|)^{r+\alpha}}$$

for some $C > 0$ which does not depend on k. The same estimate holds for $\left|S_k^K(t)\right|$. Hence

$$\begin{aligned}\left|E^K(\tau,\mathbb{X}^K(\tau)) - E^K(\tau,\mathcal{X}^K(\tau))\right| &\leq C \sum_{\substack{k\in(\mathbb{Z}^d)^*\\ |k|\leq K}} \frac{1}{(1+|k|)^{r+\alpha}} \left|\mathbb{X}^K(\tau) - \mathcal{X}^K(\tau)\right|\\ &\leq C\left|\mathbb{X}^K(\tau) - \mathcal{X}^K(\tau)\right| \qquad \text{(III-5.40)}\end{aligned}$$

where the constant C can be taken independent of K because $r+\alpha > d+1$. Thus, using

(III-5.36),

$$\left|\begin{pmatrix}\mathbb{X}^K(s)-\mathcal{X}^K(s)\\ \mathbb{V}^K(s)-\mathcal{V}^K(s)\end{pmatrix}\right| \le \int_{t^0}^{s}\left|\begin{pmatrix}\mathbb{V}^K(\tau)-\mathcal{V}^K(\tau)\\ E^K(\mathbb{X}^K(\tau))-E^K(\mathcal{X}^K(\tau))\end{pmatrix}\right| d\tau + T|\mathcal{E}(K,\Delta x,\Delta v)|$$
$$\le C\int_{t^0}^{s}\left|\begin{pmatrix}\mathbb{V}^K(\tau)-\mathcal{V}^K(\tau)\\ \mathbb{X}^K(\tau)-\mathcal{X}^K(\tau)\end{pmatrix}\right| d\tau + T|\mathcal{E}(K,\Delta x,\Delta v)|\,,$$

and we conclude by using the Grönwall lemma II.1:

$$\left|\begin{pmatrix}\mathbb{X}^K(s)-\mathcal{X}^K(s)\\ \mathbb{V}^K(s)-\mathcal{V}^K(s)\end{pmatrix}\right| \le CTe^{CT}K^d\left(\delta+\sum_{i=1}^{2d}\Delta z_i^{q_i}K^{q_i}\right),$$

where C is independent of $K, \Delta z_i, s$.

Step 3: kernel truncature error We estimate the approximation in the characteristics that is due to the truncation error in the Fourier Kernel. For $p = 1, \dots, P$,

$$X^K(t;t^0,x_p,v_p) = x_p + \int_{t^0}^{t} V^K(\tau;t^0,x_p,v_p)d\tau,$$
$$X(t;t^0,x_p,v_p) = x_p + \int_{t^0}^{t} V(\tau;t^0,x_p,v_p)d\tau,$$
$$V^K(t;t^0,x_p,v_p) = v_p + \int_{t^0}^{t} E^K(\tau,X^K(t^0,x_p,v_p))d\tau,$$
$$V(t;t^0,x_p,v_p) = v_p + \int_{t^0}^{t} E(\tau;X(\tau;t^0,x_p,v_p))d\tau,$$

so that we have

$$\begin{pmatrix}\mathcal{X}^K(s)\\ \mathcal{V}^K(s)\end{pmatrix} = \begin{pmatrix}\mathcal{X}^K(t^0)\\ \mathcal{V}^K(t^0)\end{pmatrix} + \int_{t^0}^{s}\begin{pmatrix}\mathcal{V}^K(s)\\ E^K(\tau,\mathcal{X}^K(\tau))\end{pmatrix} d\tau$$
$$= \begin{pmatrix}\mathcal{X}(s)\\ \mathcal{V}(s)\end{pmatrix} + \int_{t^0}^{s}\begin{pmatrix}\mathcal{V}^K(s)-\mathcal{V}(s)\\ E^K(\tau,\mathcal{X}^K(\tau))-E(\tau,\mathcal{X}(\tau))\end{pmatrix} d\tau$$
$$= \begin{pmatrix}\mathcal{X}(s)\\ \mathcal{V}(s)\end{pmatrix} + \int_{t^0}^{s}\begin{pmatrix}\mathcal{V}^K(s)-\mathcal{V}(s)\\ E^K(\tau,\mathcal{X}^K(\tau))-E(\tau,\mathcal{X}^K(\tau))\end{pmatrix}$$
$$+\begin{pmatrix}0\\ E(\tau,\mathcal{X}^K(\tau))-E(\tau,\mathcal{X}(\tau))\end{pmatrix} d\tau.$$

Thus,

$$\left|\begin{pmatrix}\mathcal{X}^K(s)-\mathcal{X}(s)\\ \mathcal{V}^K(s)-\mathcal{V}(s)\end{pmatrix}\right| \leq \int_{t^0}^{s}\left|\begin{pmatrix}\mathcal{V}^K(s)-\mathcal{V}(s)\\ E^K(\tau,\mathcal{X}^K(\tau))-E^K(\tau,\mathcal{X}(\tau))\end{pmatrix}\right| + \left|\begin{pmatrix}0\\ E^K(\tau,\mathcal{X}(\tau))-E(\tau,\mathcal{X}(\tau))\end{pmatrix}\right| d\tau.$$

Since $f_0 \in \mathcal{H}^{r+\alpha}_{\nu+j}$, by Proposition III.2 we have $\left\|(f-f^K)(t)\right\|^2_{\mathcal{H}^r_\nu} \leq \frac{C}{(1+K)^\alpha}$, hence by Proposition III.3 we obtain

$$\left|\begin{pmatrix}0\\ E^K(\tau,\mathcal{X}(\tau))-E(\tau,\mathcal{X}(\tau))\end{pmatrix}\right| \leq \frac{C}{(1+K)^{\frac{\alpha+1}{2}-d}},$$

where C does not depend on K. We get

$$\left|\begin{pmatrix}\mathcal{X}^K(s)-\mathcal{X}(s)\\ \mathcal{V}^K(s)-\mathcal{V}(s)\end{pmatrix}\right| \leq \int_{t^0}^{s}\left|\begin{pmatrix}\mathcal{V}^K(s)-\mathcal{V}(s)\\ E^K(\tau,\mathcal{X}^K(\tau))-E^K(\tau,\mathcal{X}(\tau))\end{pmatrix}\right| d\tau + \frac{C}{(1+K)^{\frac{\alpha+1}{2}-d}}$$

For the same reasons as those leading to (III-5.40), we obtain

$$\left|\begin{pmatrix}\mathcal{X}^K(s)-\mathcal{X}(s)\\ \mathcal{V}^K(s)-\mathcal{V}(s)\end{pmatrix}\right| \leq C\int_{t^0}^{s}\left|\begin{pmatrix}\mathcal{V}^K(s)-\mathcal{V}(s)\\ \mathcal{X}^K(\tau)-\mathcal{X}(\tau)\end{pmatrix}\right| d\tau + \frac{CT}{(1+K)^{\frac{\alpha+1}{2}-d}}.$$

Finally, the Grönwall lemma yields

$$\left|\begin{pmatrix}\mathcal{X}^K(s)-\mathcal{X}(s)\\ \mathcal{V}^K(s)-\mathcal{V}(s)\end{pmatrix}\right| \leq \frac{CTe^{CT}}{(1+K)^{\frac{\alpha+1}{2}-d}}$$

which completes the proof. □

Part III

CHAPTER 6 Conclusion

In this Part of the manuscript we have studied a particle method that can be used to simulate the solution to the Vlasov-Poisson system. It was first described in [13], but unfortunately no detailed analysis of the method had been given. Moreover, they only considered the Vlasov-HMF[1] system. We presented an extension of their ideas to the Vlasov-Poisson system.

The numerical scheme can be understood as a semi-Lagrangian scheme where, instead of going back one timestep and perform the interpolation, we go back to the initial time of the simulation. Since we know the function at initial time, no interpolation is needed. It is, in some sense, related to the Vortex methods used to simulate numerically the solution to 2D-Euler equations.

The purpose of the paper [105], from which this Part of the manuscript is based on, was first to fully describe the method, and then to give a detailed analysis. By using the fact that the scheme can be decomposed into elementary, well-known components (namely numerical quadrature, discrete Fourier transform and time integration), we were able to obtain an error bound for the method that writes as a sum of the error bounds associated to each component of the method.

The method is applied to standard one-dimensional numerical examples (i.e. $1x-1v$), and compared to a semi-Lagrangian scheme. It is shown that the results are very satisfying for short times. For long times, the solution is less satisfying, and we explain this by the presence of vortices in the solution. Indeed, the method can intuitively be understood as "moving" the quadrature points along the flow of the equation, and if vortices are created then for long times all the quadrature points lie in the vortices. Thus, the vortices are well represented numerically by having many quadrature points, but the region of the computational domain outside the vortices is very poorly accounted for.

The numerical method presents other issues. The main one is probably the computational complexity, which makes it hardly applicable for higher dimensions. This is partly due to the use of a numerical grid, because the number of points grows exponentially with respect to the dimension. One other issue is the dependance of the error on the number K of Fourier modes used. Indeed, the error bound we use involves an increasing function of

[1] Hamiltonian Mean-Field

K as well as a decreasing function. Hence, it should be doable to find an optimal choice of K. This optimal choice would depend on the number of points used in the grid. However, if the solution has huge variations, we need K large, which means that the number of grid points must be large as well, and thus computationally expensive.

The issues mentioned are mainly related to the fact that a grid is used for the whole phase-space. Part of the solution to circumvent them would be to use Monte-Carlo integration (and thus a random quadrature grid), but the error bound would likely be harder to obtain, and involve probabilities as well as some different notions of convergence.

IV

The modulation of the Schrödinger equation

Part IV

Introduction

CHAPTER 1

Erwin Schrödinger (1887–1961). Credit to wikipedia.org.

The XVII$^{\text{th}}$ century saw a great deal of scientific advances and discoveries. But there were also quite a few misunderstandings of nature, and the wave-particle duality is one of them. At first, Isaac Newton thought that light was made of particles, but an opposite idea quickly surfaced: Christiaan Huygens imagined that light was made only of waves. A few years later in 1801, Thomas Young's interference experiments helped validate the wave model proposed by Huygens. Fast forward one century, Max Planck derived a model for which energy could only change by a minimal increment, just like if particles were emitted. This gave again some momentum to the "particle" understanding of light, but the two interpretations of light were still opposite to each other.

In 1925, Louis De Broglie wrote his doctoral thesis *Recherche sur la théorie des quanta*, and proposed that particles are bundles of waves which move with a group velocity, and which possess an effective mass [1].

In 1926, the physicist Erwin Schrödinger wrote the fundamental article [72]. In this work, he starts from ideas developed by Louis De Broglie one year earlier, who considered atoms and electrons as material points, and generalizes it. Schrödinger's work is the first account of what is now called the *time-independent Schrödinger equation*, which was written at that time

$$\Delta\psi + 8\pi^2 m(E-V)\psi/h^2 = 0, \tag{IV-1.1}$$

where Δ is the Laplacian operator, ψ is the wave function, $E-V$ is the kinetic energy, and $h \approx 6.6261 \cdot 10^{-34}$ Joule $\cdot$ Hz^{-1} is Planck's constant.

In order to see how his equation was able to solve physical problems, Schrödinger adapted equation (IV-1.1) to a simplified model of the hydrogen atom. The wave function for this model now has to satisfy the following equation:

$$\Delta\psi + 8\pi^2 m(E+e^2/r)\psi/h^2 = 0, \tag{IV-1.2}$$

[1] The effective mass of an object is the mass it *seems* to have when responding to forces or interacting with other identical objects.

where e is the electronic charge, and $r = (x^2 + y^2 + z^2)^{1/2}$. Owing to the fact that the Laplace operator Δ is diagonal in the Fourier space, one can study the values of the energy E which give existence of a solution, and for which the solution is finite and single-valued on the whole space $\mathbb{R}^3$. By doing so, Schrödinger got the following set of admissible E-values:

— $E > 0$
— $E = -2\pi^2 me^4/(h^2n^2)$, $n = 1, 2, \ldots$.

The first condition corresponds to “hyperbolic orbits” in ordinary mechanics, and Schrödinger argues that they are generally not interesting in quantum theory. Thus, the only interesting E-values for which (IV-1.2) has an unique, finite and singe-valued solution are discrete energy values. Moreover, these energy values correspond exactly to Bohr's stationary energy levels.

This explanation of the discrete energy levels convinced many physicists, and the interest for the Schrödinger equation grew over the following decades to be now of utter importance, for physicists, chemists, and mathematicians.

It is now generally accepted that the Schrödinger equation can explain electronic dynamics, and by understanding this equation better one can understand our world better. As simple as it is.

In 1926, Max Born gave a *statistical interpretation* of quantum mechanics: the square modulus of the wave function $|\psi(t,\cdot)|^2$ can be understood as a probability density for the position of a particle. See [16], or [67] for an English translation.

The rest of this chapter is a presentation of the modern-day Schrödinger equation, heavily based on [55, Chapter I]. The *time-dependent* Schrödinger equation writes

$$i\hbar\frac{\partial\psi}{\partial t} = H\psi, \tag{IV-1.3}$$

where $i^2 = -1$, and $H = T + V$ is the Hamiltonian operator depending on a kinetic operator T and a potential V. The quantity $\hbar = \frac{h}{2\pi}$ is the reduced Planck's constant, and its value in the standard international system of units is $\hbar \approx 1.0546 \cdot 10^{-34}$ Joule $\cdot$ Hz^{-1}. The kinetic operator is usually defined as

$$T\psi = -\frac{\hbar^2}{2m}\Delta\psi,$$

where m is the mass of the particle considered and where $\Delta = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}$ is the usual Laplace operator, while the potential simply is a multiplication operator:

$$(V\psi)(x) = V(x)\psi(x).$$

We supply this equation with the initial condition at time $t = 0$:

$$\psi(t = 0, \cdot) = \psi_0,$$

for some function ψ_0, called the *initial condition.*

The above description holds for one particle, but since H is linear, one can construct a multi-particle Hamiltonian operator as the sum of each individual Hamiltonian operator. Each individual Hamiltonian operator has a potential V which can be understood as an external force. When several particles are considered, the potential part of the multi-particle Hamiltonian operator can also take into account interactions between particles. The Schrödinger equation in the N-particle case is then

$$\begin{aligned} &i\hbar \partial_t \psi(t,x) = -\frac{\hbar^2}{2m}\Delta_x \psi(t,x) + V(t,x,\psi)\psi(t,x), \\ &\psi(t=0,\cdot) = \psi_0, \end{aligned} \tag{IV-1.4}$$

where $x = \left(x^1, \dots, x^N\right) \in \mathbb{R}^{Nd}$ and $t \geq 0$. The variable $x^j = \left(x^j_1, \dots, x^j_d\right) \in \mathbb{R}^d$ is the position variable of the j-th particle.

If the potential is real-valued and depends only on the space variable $x \in \mathbb{R}^{Nd}$, it is said to be *confining* if $V(x) \to +\infty$ when $|x| \to +\infty$. Not all potentials yield an unique solution to the Schrödinger equation (IV-1.4).

When considering the many-particle Hamiltonian, it is often to model electron and nuclei motions in atoms or molecules. If we denote by m the electrons' mass and by M the nuclei mass, we have

$$0 < \frac{m}{M} := \varepsilon^2 \ll 1.$$

Using physical considerations, it can be argued that motion of the nuclei over a distance ~ 1 can be expected on a time scale $\sim \varepsilon^{-1}$. The time is then rescaled to $t \to t/\varepsilon$, so that the Schrödinger equation then takes the form

$$i\varepsilon \frac{\partial \psi}{\partial t} = -\frac{\varepsilon^2}{2}\Delta\psi + V\psi. \tag{IV-1.5}$$

This is called the *semi-classical scaling of the Schrödinger equation.*

Other important applications of the Schrödinger equation include laser beam propagation and quantum optics [77, 62, 33]. In this case, the nonlinear Schrödinger equation writes

$$i\partial_t \psi + \Delta\psi + |\psi|^2\psi = 0.$$

The nonlinear Schrödinger equation also appears in plasma physics [64] or fiber optics [2].

We begin with Chapter IV-2 – Review of the Schrödinger equation, which is a quick overview of the literature treating the Schrödinger equation. After giving some properties and facts from the theoretical point of view, we review four main families of numerical schemes that are – or have been – used to approximate the solution to the Schrödinger equation. We will see that all but one have limitations that prevent them from being used in real life situations. The promising family of schemes consists in *Variational methods*. However, nowadays they are only used in the linear framework. Thus, there is room for works which focus on nonlinear equations but retain the advantages of variational methods.

Before presenting the main content of the current Part, we will see in Chapter IV-3 – Modulation of solutions as a theoretical tool the additional ideas that we use compared to those of the usual variational methods. We are motivated by some very recent theoretical works to look for solutions to the Schrödinger equation under some specific form. The idea is to have a function which depends on a small number of parameters, and the time evolution of these parameters is chosen appropriately so that the parametrized function is solution to the Schrödinger equation. This is what we call *modulation*. We will see how the exact dynamics of the linear Schrödinger equation with quadratic potential can be recovered exactly with a very small number of parameters. Then comes the main dish: Chapter IV-4 – Nonlinear Schrödinger equation, which deals with the cubic nonlinear Schrödinger equation. We can no longer recover the exact dynamics of the solution using a small number of parameters because of the cubic interactions in the equation, but it is possible to obtain numerically an approximate solution. We use a variational approach for this, and it seems to be the first account of a variational approach used in the nonlinear case. The variational method is the Dirac-Frenkel principle. The proposed method is then studied on some numerical examples against a grid-based spectral method, and it is observed that the Dirac-Frenkel principle suffers from the same issues than in the linear setting.

Finally, this Part ends with Chapter IV-5 which is a discussion on the limitations of the proposed method, as well as perspectives and ways to improve it.

The novel content of this Part of the manuscript is heavily based on some yet unpublished work [30], which is a joint work between Erwan Faou, Pierre Raphaël, and the author of the present manuscript.

Review of the Schrödinger equation

Part IV

CHAPTER 2

In this Chapter we will review some basic facts and results about the time-dependent Schrödinger equation (IV-1.4). Equation (IV-1.4) with total dimension Nd was a "physical" equation, we now consider a "mathematical" version: $x \in \mathbb{R}^d$, and set the physical constants $\hbar, m$ respectively to 1 and 1/2. The equation at hand is now:

$$
\begin{aligned}
&i\partial_t \psi(t,x) = \left(-\Delta_x + V(t,x,\psi)\right)\psi(t,x) = H(t,x,\psi)\psi(t,x), \qquad x \in \mathbb{R}^d, t \geq 0,\\
&\psi(t=0,\cdot) = \psi_0.
\end{aligned}
\tag{IV-2.1}
$$

We recall that $\Delta_x = \sum_{j=1}^d \frac{\partial^2}{\partial x_i^2}$ is the usual Laplacian with respect to variable x, and $H(t,x,\psi) := -\Delta_x + V(t,x,\psi)$ is called the *Hamiltonian operator*. We may use the notation $\psi(t)(x) = \psi(t,x)$, and $\psi(t) = \psi(t,\cdot)$.

IV-2.1 Some results about the Schrödinger equation

As usual when studying partial differential equations, the first things to ask are: is there a solution to (IV-2.1)? If the answer is "yes", is it unique? How does the solution depend on the initial condition ψ_0? Does the existence and/or uniqueness of the solution depend on the initial condition ψ_0? Does the solution exist for all times $t \geq 0$, or is there a bound $T > 0$ such that the solution exists only for times $0 \leq t < T$?

IV-2.1.1 Existence, uniqueness, ...

Potential $V = 0$

In the absence of potential, i.e. $V = 0$, it can be shown that a solution exists for all $\psi_0 \in \mathbb{L}^2(\mathbb{R}^d)$. This can be done for instance as explained in [20] or [55]: solve the equation for initial condition $\psi_0 \in \mathcal{S}$ using the Fourier transform, where $\mathcal{S}$ is the Schwartz space. The solution is unique and well-defined for all times $t \geq 0$. Then, use the fact that the Schwartz space is dense in $\mathbb{L}^2(\mathbb{R}^d)$ so that the solution can be defined by density for $\psi_0 \in \mathbb{L}^2(\mathbb{R}^d)$.

Potential $V = V(x)$

When a nonzero potential $V = V(x)$ is considered, the existence and uniqueness of a solution to (IV-2.1) generally relies on the self-adjointness of the Hamiltonian operator H.

Definition IV.1

Let $\mathscr{H}$ a complex Hilbert space with inner product $(\cdot, \cdot)$, taken antilinear in its first argument and linear in its second one. A linear operator $H : D(H) \to \mathscr{H}$, defined on a domain $D(H)$ dense in $\mathscr{H}$, is called *symmetric* if

$$(H\psi, \varphi) = (\psi, H\varphi)\,, \quad \forall \psi, \varphi \in D(H).$$

The operator H is *self-adjoint* if for any $\varphi, \eta \in \mathscr{H}$, the relation

$$(H\psi, \varphi) = (\psi, \eta)\,,\ \forall \psi \in D(H) \quad \text{implies} \quad \varphi \in D(H) \text{ and } \eta = H\varphi.$$

In order to show the existence of a solution to (IV-2.1), one can use the following theorem (see for instance [38, Theorem 2.16]):

Theorem IV.1

If H is a self-adjoint operator, then there is a unique family of bounded operators, $U(t) := e^{-itH}$, having the following properties for $t, s \in \mathbb{R}$:

$$i\frac{\partial}{\partial t}U(t) = HU(t) = U(t)H, \tag{IV-2.2}$$

$$U(0) = 1, \tag{IV-2.3}$$

$$U(t)U(s) = U(t+s), \tag{IV-2.4}$$

$$\|U(t)\psi\| = \|\psi\|. \tag{IV-2.5}$$

If the potential V is such that Theorem IV.1 can be applied (i.e. $-\Delta + V$ is a self-adjoint operator), then the unique solution to (IV-2.1) is given by

$$\psi(t) = U(t)\psi_0.$$

The question is now: when can Theorem IV.1 be applied? Among the different possible criteria, there is one rather simple, called the *Kato-Rellich theorem* (see [46, Section V.4.1, Theorem 4.3], or [38, Theorem 2.9]):

Theorem IV.2

Let T be a self-adjoint operator on a Hilbert space, and V a symmetric operator bounded by

$$\|V\psi\| \leq a\|\psi\| + b\|T\psi\|$$

for all $\psi \in D(T)$, with $0 < b < 1$. Then, $H = T + V$ is self-adjoint with domain $D(H) = D(T)$.

Potential $V = V(x, \psi)$

The results and properties given in this section are based on [20] and [18].

When the potential depends on the wave function ψ, the PDE (IV-2.1) is not linear anymore. It is then called a *nonlinear PDE*. Generally, properties like existence or uniqueness are harder to prove in the nonlinear case, and we can expect that interactions (or nonlinearities) will be the source of many issues (theoretical as well as numerical, as we will see in the next section).

The nonlinearity can be either local or nonlocal. A *global* or *nonlocal nonlinearity* means that the value of ψ over the whole domain $\mathbb{R}^d$ is required to compute the nonlinearity at each point $x \in \mathbb{R}^d$. Such examples are the Schrödinger-Poisson system:

$$i\partial_t\psi = -\Delta\psi + V(x)\psi + V_p\psi, \quad \Delta V_p = \lambda\left(|\psi|^2 - c\right)$$

for some appropriate constants λ, c, or the Hartree equation:

$$i\partial_t\psi = -\Delta\psi + V(x)\psi + \lambda\left(\frac{1}{|x|^\gamma} * |\psi|^2\right)\psi,$$

with some exponent γ. A *local nonlinearity* only involves the value $\psi(x)$ in order to compute the nonlinearity at $x \in \mathbb{R}^d$. It can for instance be polynomial:

$$i\partial_t\psi = -\Delta\psi + V(x)\psi + |\psi|^{2\sigma}\psi,$$

for some $\sigma > 0$, or logarithmic:

$$i\partial_t\psi = -\Delta\psi + V(x)\psi + \log(|\psi|^2)\psi.$$

See [18] for a recent and unified presentation of results about the Schrödinger equation with local nonlinearities.

We will focus later on the cubic nonlinear Schrödinger equation with quadratic potential, thus we give now some precise statements concerning the following problem:

$$\begin{cases} i\partial_t u + (\Delta - |x|^2)u = |u|^2 u, \\ u(0) = \varphi. \end{cases} \tag{IV-2.6}$$

But first, some notations and vocabulary. Consider a real-valued potential $U \in C^\infty(\mathbb{R}^d)$, such that $U \geq 0$ and

$$D^\alpha U \in \mathbb{L}^\infty(\mathbb{R}^d), \qquad \text{for all } \alpha \in \mathbb{N}^d \text{ such that } |\alpha| > 2,$$

where $D^\alpha := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}$. We define the operator A on $\mathbb{L}^2(\mathbb{R}^d)$ by

$$\begin{cases} D(A) = \left\{u \in H^1(\mathbb{R}^d) : U|u|^2 \in \mathbb{L}^1(\mathbb{R}^d) \text{ and } \Delta u - Uu \in \mathbb{L}^2(\mathbb{R}^d)\right\}, \\ \quad Au = \Delta u - Uu \quad \text{for } u \in D(A). \end{cases}$$

It can be shown that A is a self-adjoint and negative operator, see for instance [20, Lemma 9.2.1]. The total energy E is defined as

$$E(u) = \frac{1}{2}\int_{\mathbb{R}^d} \left(|\nabla u|^2 + U|u|^2 - \frac{1}{2}|u|^4\right) dx.$$

Also define

$$X_A := \left\{u \in H^1(\mathbb{R}^d) : U|u|^2 \in \mathbb{L}^1(\mathbb{R}^d)\right\},$$

and the associated norm:

$$\|u\|_{X_A}^2 = \|\nabla u\|_{\mathbb{L}^2}^2 + \|u\|_{\mathbb{L}^2}^2 + \int_{\mathbb{R}^d} U|u(x)|^2 dx.$$

We denote by X_A^* the dual of the space X_A.

For initial datum $\varphi \in X_A$, we have the following result (it can be found for instance in [20, Theorem 9.2.6]):

Theorem IV.3

The following properties hold:

— For every φ in X_A, there exist $T_{\min}(\varphi), T_{\max}(\varphi) > 0$ and a unique, maximal solution $u \in C((-T_{\min}, T_{\max}), X_A) \cap C^1((-T_{\min}, T_{\max}), X_A^*)$ of (IV-2.6). The solution u is maximal in the sense that if $T_{\max} < \infty$ (resp. $T_{\min} < \infty$), then

$\|u(t)\|_A \to \infty$ as $t \to T_{\max}$ (resp. as $t \to T_{\min}$).

— There is conservation of charge and energy, that is

$$\|u(t)\|_{\mathbb{L}^2} = \|\varphi\|_{\mathbb{L}^2} \quad \text{and} \quad E(u(t)) = E(\varphi) \quad \text{for all } t \in (-T_{\min}, T_{\max}).$$

— There is continuous dependence of the solution on the initial value in the sense that both function $T_{\min}(\varphi)$ and $T_{\max}(\varphi)$ are lower semicontinuous, and that if $\varphi_m \to \varphi$ in X_A and if $[-T_1, T_2] \subset (-T_{\min}(\varphi), T_{\max}(\varphi))$, then $u_m \to u$ in $C([-T_1, T_2], X_A)$, where u_m is the maximal solution of (IV-2.6) with initial value φ_m.

— If $\varphi \in D(A)$, then $u \in C((-T_{\min}, T_{\max}), D(A)) \cap C^1((-T_{\min}, T_{\max}), \mathbb{L}^2(\mathbb{R}^d))$.

In the following, the initial condition φ will be assumed *smooth enough*, so that the above existence result is enough for our purposes.

IV-2.1.2 Imaginary time method

The Schrödinger equation can be shown to be related to diffusive equations. This can be seen either by separating the real and imaginary parts (as in done for instance in [84]), or by using the so-called *Imaginary time method* (see for instance [4, 23]). This latter transformation is sometimes called *Wick rotational transformation from real time to imaginary time*, due to Wick who first used this idea in [85].

> "While the concept of an imaginary relative time variable does not help physical intuition, it has mathematically several advantages."
>
> G. C. Wick (1954)

Some properties of the Schrödinger equation can then be deduced from the corresponding diffusive equations. The Wick rotational tranform from real time to imaginary time is simply obtained by letting $\tau := it$, replacing t by $-i\tau$ in (IV-2.1), and then looking at τ just like if it was a real variable. We get

$$\frac{\partial}{\partial \tau}\psi(\tau, x) = \Delta\psi(t, x) - V(x)\psi(t, x). \tag{IV-2.7}$$

The variable τ is supposedly an imaginary number, but once we have (IV-2.7) we can study the equation by considering $\tau \in \mathbb{R}$. In particular, knowing the functions $\{\varphi_n\}_{n\geq 0}$, eigenfunctions of $H = -\Delta + V$, is enough to solve the Schrödinger equation. This opens some perspectives from the theoretical point of view, as well as from the numerical point of view. One of the main uses of this method is the Normalized Gradient method, but it has also been used in other various contexts, see for instance [82].

Normalized Gradient method

Suppose that the potential V is real-valued, depends only on space x, is such that $H = -\Delta + V$ is self-adjoint, and such that $V(x) \to \infty$ as $|x| \to \infty$. It can be shown that there exists an orthonormal basis $\{\varphi_n\}_{n\in\mathbb{N}}$ of eigenfunctions of the operator H, each φ_n being associated to an eigenvalue λ_n. Moreover, the eigenvalues are real and can be ordered so that $\lambda_0 \leq \lambda_1 \leq ....$ Therefore, we can decompose ψ solution to (IV-2.7) into this basis:

$$\psi(\tau, x) = \sum_{n=0}^{\infty} c_n(\tau)\varphi_n(x),$$

where the expansion coefficients at time $\tau = 0$ are given by $c_n(0) = (\psi_0, \varphi_n)_{\mathbb{L}^2(\mathbb{R}^d)}$. Plugging this ansatz into (IV-2.7), we obtain

$$c_n'(\tau) = -\lambda_n c_n(\tau) \implies c_n(\tau) = c_n(0)e^{-\lambda_n \tau},$$

and thus

$$\psi(\tau, x) = \sum_{n=0}^{\infty} c_n(0)\varphi_n(x)e^{-\lambda_n \tau}. \tag{IV-2.8}$$

As $\tau \to \infty$, all contributions in the sum become negligible except for the lowest integer n such that $c_n(0) \neq 0$. For simplicity, assume $c_0(0) \neq 0$. Then

$$\psi(\tau, x) \sim_{t\to\infty} c_0(0)\varphi(x)e^{-\lambda_0 \tau}. \tag{IV-2.9}$$

This means that, after a long-time, the space behavior of ψ is governed by the first eigenfunction of the operator H. In order to obtain the eigenfunction associated to the next eigenvalue λ_1, it suffices to start from an initial condition ψ_0 such that $(\psi_0, \varphi_0)_{\mathbb{L}^2(\mathbb{R}^d)} = 0$ and $(\psi_0, \varphi_1)_{\mathbb{L}^2(\mathbb{R}^d)} \neq 0$, and so on for the next ones.

Of course, as $\tau \to \infty$, we get $\|\psi(\tau)\|_2 \to 0$, so the main idea of the Normalized Gradient method is to renormalize ψ by its $\mathbb{L}^2$-norm at all times. In the Normalized Gradient method (see e.g. [7]), a time sequence $\{\tau_n\}_{n\in\mathbb{N}}$ is considered. For each time instant τ_n, the following steps are done:

1. Given an approximation of the function $\psi(\tau_{n-1})$, obtain an approximation $\tilde{\psi}(\tau_n)$ of $\psi(\tau_n)$;
2. Normalize $\tilde{\psi}(\tau_n)$ by its $\mathbb{L}^2$-norm in space;

3. Use this normalized approximation $\tilde{\psi}(\tau_n)$ for the starting point of the next step.

In essence, it consists in getting rid of the multiplication by $e^{-\lambda_0 \tau}$ in (IV-2.9) without the knowledge of λ_0. As $n \to \infty$, the approximation $\tilde{\psi}(\tau_n)$ becomes an approximation of the ground state φ (more precisely, it becomes an approximation of $\frac{c_0(0)}{|c_0(0)|}\varphi$ but this is φ up to a phase shift).

Remark IV.1

The main application of the Normalized Gradient method consists in looking for eigenfunctions of H, by looking at $\lim_{\tau\to\infty} \frac{\psi(\tau,x)}{\|\psi(\tau,x)\|_2}$. Therefore, one has to numerically discretize time and apply successively the numerical approximation of the operator H. Even though the methods involved are quite different, one can think of the discretized Normalized Gradient method as some kind of power method used to obtain the eigenvalues of a matrix. Some authors have not called it a “Normalized Gradient method” but applied the same idea, see for instance [78, 76].

The fact that the long-time space behavior of ψ is governed only by the first eigenmode of the operator H has lead some authors to derive the Normalized Gradient method by writing

$$\psi(t,x) = e^{-i\mu t}\varphi(x),$$

with μ and φ to be determined, and then to plug this into (IV-2.1).

Note that the above ideas also work in the nonlinear case, i.e. if the potential depends on the solution as well. The imaginary time method is for instance used in [7, 29] as the core idea of a generalized version of the Normalized Gradient method: the nonlinear cubic Schrödinger equation

$$i\partial_t\psi + \Delta\psi - |x|^2\psi = |\psi|^2\psi \tag{IV-2.10}$$

is used as means to obtain the *ground states* that minimize the energy functional defined by

$$E(\phi) := \int_{\mathbb{R}^d} \left(\frac{1}{2}|\nabla\phi|^2 - \frac{1}{2}|x|^2|\phi|^2 + \frac{1}{4}|\phi|^4\right).$$

These ground states are called *solitons*. In other words, the solitons solution to

$$\Delta Q - |x|^2 Q - |Q|^2 Q = \lambda Q,$$

are obtained by looking at the Wick transform of (IV-2.10). The pair (λ, Q) is an eigenpair of the nonlinear operator appearing on the left-hand side.

We have mentioned in this section that the operator $H = -\Delta + V$ needed to be discretized at some point. This holds for the Normalized Gradient method, but in general if one is not interested in the ground states of the Schrödinger equation but in the time evolution of the solution $\psi(t,x)$, then the time domain has to be discretized as well as the space domain. This is easier said than done. There are several challenges one will face when discretizing (IV-2.1):

- the high dimensionality of the space domain;
- the fast oscillations in the solutions;
- the different scales occuring in the equation: (IV-1.3) is used as a model for the motion of electrons and nuclei in atoms or molecules, but the mass of electrons is much smaller than that of nuclei particles. A satisfying numerical scheme has to take this discrepancy into account.

IV-2.2 Numerical schemes

When trying to obtain an approximate numerical solution to a PDE, the first and easiest scheme generally consists in using finite differences. We explained in Section IV-1 – Introduction that, in most physical situations, the dimension of the space variable x is $3N$ with N the number of particles considered. A finite-difference scheme needs to discretize the $3N$-dimensional phase space, which is already a computational challenge in itself. Furthermore, the grid has to be relatively fine so that the results are meaningful. Hence, there is no way one can simulate physical scenarii using finite differences. This is summed up in [55]:

> “Computations with direct finite-difference discretizations of Schrödinger's equation are out of reach for more than two or three particles.”
>
> C. Lubich (2008)

We note that, even for simple molecules like the one of carbon dioxyde CO_2, the dimension is much too large to even fit a grid onto any computer. For CO_2, there are three nuclei and twenty-two electrons, which yields twenty-five particles and thus a space variable $x \in \mathbb{R}^{75}$.

The **huge** space dimensionality then rules out any grid-based method for simulating the “physical” equation. However, the mathematical version (IV-2.1) of the Schrödinger equation may be simpler to study since it can be formulated in a d-dimensional setting with d possibly small. This (simpler) mathematical version can be justified as a way to gain insight into the physics happening in real situations.

Remark IV.2

From now on, we consider the mathematical and low-dimensional version of the Schrödinger equations, with dimension $d \geq 1$. If we need to refer to the "physical" Schrödinger equations, we will state it explicitly.

The question is now: are grid-based methods an interesting approach to simulating the Schrödinger equations? And the next question is: if grid-based methods are not appropriate, what better options are available?

IV-2.2.1 FDTD

In [78], a Finite-Difference Time-Domain (FDTD) method is applied to the Schrödinger equation. It consists in discretizing both the time and space domains using finite-differences, and was already largely used in electromagnetic simulations but not yet used in the context of the Schrödinger equation. It focuses on the Schrödinger operator $i\partial_t + \Delta - V$, which is discretized using finite-differences in the (t, x) domain. The method is a two-step process: the first step obtains the eigenvalues of the operator, while the second step obtains the eigenvectors by using the eigenvalues obtained during the first step.

It may be useful now to have in mind the computations we did in Section IV-2.1.2 – Imaginary time method, and specifically equation (IV-2.8). During the first step of the FDTD method, an observation point x_{obs} in the space domain is picked, and the initial condition is chosen to be a very localized function around x_{obs}. By having an initial condition very localized, we know that it involves many eigenfunctions of the operator H. Then equation (IV-2.1) is simulated over a long time using finite differences in time and space, and once enough iterations are done the simulation is stopped. Denote N_{final} the total number of iterations in time. In particular, we have a sequence $\left(\tilde{\psi}(t_n, x_{obs})\right)_{n=0,\ldots,N_{final}}$ which corresponds to approximate values of the solution ψ to the Schrödinger equation, evaluated at the observation point x_{obs} and for all instants t_n of the numerical simulation. By performing a (discrete) Fourier transform of this sequence, we can determine the first eigenvalues of H: indeed, equation (IV-2.8) indicates that $\psi(t, x_{obs})$ simply is a weighted sum of complex exponentials, each one having a phase corresponding to an eigenvalue of H. The second step of the FDTD procedure consists in obtaining the eigenfunctions from the eigenvalues, and [78] gives a way to do so. Their numerical examples are two-dimensional, and they use finite-differences of order one in time and two in space.

Another way of using the FDTD decomposition consists in first transforming the Schrödinger equation into a diffusion equation by using the imaginary time method, and then applying the FDTD method to the diffusion equation. This idea is for instance used in [76]. The resulting algorithm is based on the expression (IV-2.8), but since the amplitude of the function decreases with time, it is necessary to renormalize the solution at each timestep.

Some other examples of the use of finite differences applied to the Schrödinger equation can be found in [74, 22, 71, 70, 87, 63, 25].

IV-2.2.2 Grid-based spectral methods

One issue of finite difference methods is that they are usually not very accurate. Most of the works mentioned in the previous section used a discretization of order 2 in space, and order 1 or 2 in time. Thus, in order to obtain satisfying results, the grid needs to be fine. An alternative option consists in using spectral approximations, which can be much more accurate.

One of the first such schemes used in the context of the Schrödinger equation is probably due to Feit, Fleck and Steiger [32], who used a Strang splitting in time, combined with a Fourier transform in space in order to approximate the Laplacian operator. They already noted that the Fast Fourier Transform[1] could be used in order to have a fast and efficient algorithm. One year later, in [49], the FFT in space is used in combination with a finite difference discretization in time. Leforestier *et al.* compare in [54] several time-propagation schemes combined with a Fourier spectral method in space. We refer to [56] for a convergence analysis of the Strang splitting method applied to the nonlinear Schrödinger equation.

The grid-based spectral schemes have also been applied to the nonlinear Schrödinger equation, see for instance [8].

Recently, grid-based methods were investigated again thanks to the discovery of an exact splitting formula between the kinetic and potential parts. The exact splitting formula is due to Bernier [13, 3], and a Fourier-based spectral method using the exact splitting is presented in [14]. By separating the kinetic and potential parts, the kinetic part $(-\Delta)$ can be solved approximately in the Fourier domain and the potential part can be computed exactly in the space domain.

For additional numerical experiments using grid-based spectral methods, we refer the reader to [9, 8].

[1] See Section II-5.2 – Discrete Fourier transform (DFT) for more details.

IV-2.2.3 Gridless spectral methods

Even though the grid-based spectral schemes seemed promising, the only numerical experiments that could have been performed were at most three-dimensional. A part of this limitation is due to the fact that a grid is used, which makes the algorithm computationally expensive. In order to avoid having to rely on a grid, which is not efficient in high dimensions, some gridless spectral methods were devised. Contrary to every numerical method we mentioned until now, there is need for a fine grid. The basic idea is that, for some equations, an appropriate $\mathbb{L}^2$ basis exists. Once we know an appropriate basis, we can decompose the solution to the equation in this basis, and simply look at the evolution of the expansion coefficients in order to solve the equation.

For the sake of clarity, we explain how gridless spectral methods work with a simple example.

Example IV.1

We consider in this example the one-dimensional linear Schrödinger equation

$$i\partial_t \psi(t,x) = -\psi''(t,x) + |x|^2\psi(t,x), \quad x \in \mathbb{R}. \tag{IV-2.11}$$

It is a well known fact (see for instance [73, Section 7.2.1.2]) that the Hermite functions H_n satisfy the following ODE:

$$H_n''(x) + (2n+1-x^2)H_n(x) = 0, \quad n \in \mathbb{N}.$$

Moreover, $\{H_n : n \in \mathbb{N}\}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R})$. Therefore, we can decompose

$$\psi(t,x) = \sum_{n=0}^{\infty} c_n(t)H_n(x).$$

By plugging this expression of ψ into (IV-2.11), one gets

$$\sum_{n=0}^{\infty} ic_n'(t)H_n(x) = \sum_{n=0}^{\infty} c_n(t)(2n+1)H_n(x).$$

The orthogonality of the family $\{H_n : n \in \mathbb{N}\}$ yields the following ODEs:

$$ic_n'(t) = (2n+1)c_n(t) \quad \Longleftrightarrow \quad c_n(t) = e^{-i(2n+1)t}c_n(0).$$

Moreover, the coefficients $c_n(0)$ are known by decomposing the initial condition at time $t = 0$ into the Hermite basis. If this decomposition at time $t = 0$ is known, we can know exactly the solution $\psi(t)$ for any t, and without having to use a grid!

It is quite obvious that, in Example IV.1, the Hermite basis has been chosen because we are interested in the linear Schrödinger equation on $\mathbb{R}$ and the Hermite functions are eigenfunctions of the operator $-\Delta + |x|^2$. If the domain in which lives the space variable is different, the appropriate basis will also be different. For instance, in [11, 10], Bao *et al.* used a Generalized-Laguerre, Fourier and Hermite combination in order to solve the cubic nonlinear Schrödinger equation (also known as the Gross-Pitaevskii equation). In their study, the treat the two- and three-dimensional cases. In the two-dimensional case, they use a polar decomposition $(r, \theta) \in \mathbb{R}_+^* \times (0, 2\pi)$, which explains the use of the Generalized-Laguerre basis for r and the Fourier basis for θ. In the three-dimensional case, they use a cylindrical decomposition $(r, \theta, z) \in \mathbb{R}_+^* \times (0, 2\pi) \times \mathbb{R}$. This explains the addition of the Hermite basis to their two-dimensional results.

In [80], Thalhammer *et al.* use ideas similar to those of Bao *et al.* but consider only one tensorized basis: either the Hermite basis, tensorized so that it becomes a basis of $\mathbb{L}^2(\mathbb{R}^d)$, or the Fourier basis by assuming that the space $\mathbb{R}^d$ is restricted to a bounding box $[-a, a]^d$, with $a > 0$ sufficiently large.

One of the main issues with the gridless spectral methods is that they are designed for specific equations. For instance, if the potential V is not “nice”, finding an explicit $\mathbb{L}^2(\mathbb{R}^d)$ basis of which all elements are eigenfunctions of $-\Delta + V$ may be difficult.

Moreover, even in the cases where there exists an explicit appropriate basis, it may be relatively expensive to use the method, depending on the initial condition. For instance, let us go back to Example IV.1. Suppose the initial condition is a standard Gaussian function (i.e. centered with variance 1), then only one Hermite mode is required since $H_0(x) = e^{-x^2/2}$. In this case the gridless spectral method will be very efficient. On the other hand, it the initial condition is a sum of two Gaussian functions with unit variance and mean $\pm\mu$, then the number of Hermite modes in its Hermite decomposition will grow as $|\mu| \to \infty$. This can be seen in Figure IV-2.1, where we plot the number of Hermite modes needed to accurately decompose $\psi_0(x) = e^{-\frac{(x+\mu)^2}{2}} + e^{-\frac{(x-\mu)^2}{2}}$, for some values of μ. The key takeaway is that, for this example, the gridless spectral method will be efficient and cheap for initial conditions localized near the origin. The computational cost will increase when the initial condition is not close to the origin, because many more modes will be nonzero. Hence, the gridless spectral methods are adapted to certain special equations, and when they are easily applicable they can be very efficient.

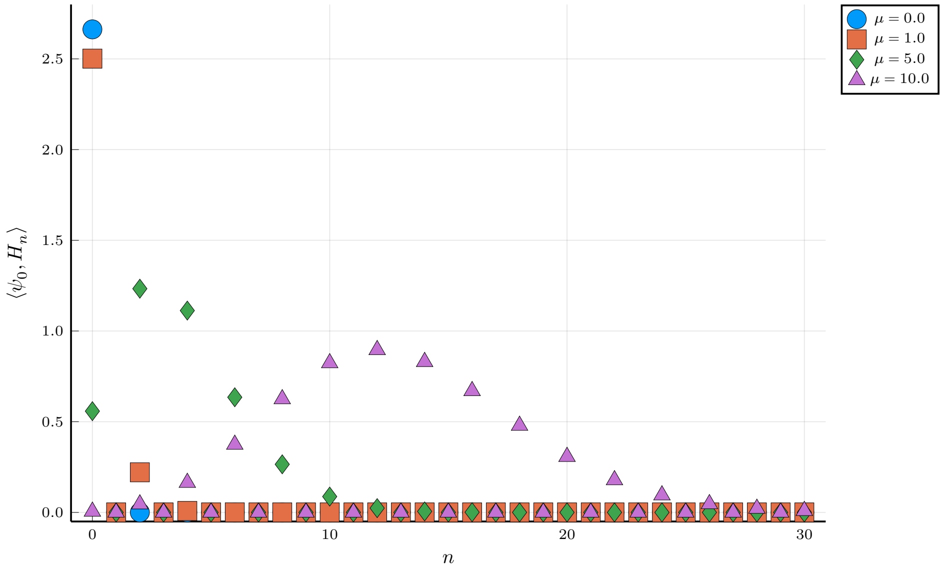


Figure IV-2.1 – Approximate Hermite coefficients of $\psi_0(x) = e^{-\frac{(x+\mu)^2}{2}} + e^{-\frac{(x-\mu)^2}{2}}$.

As we just explained, there may be situations were an appropriate basis really makes the algorithm gridless. If it happens, it generally is for linear equations. This was the case for instance of the linear Schrödinger equation. It gets a little trickier for nonlinear equations, and in this case we may need a grid the handle the nonlinear interactions. This is for example the case of [11, 80, 10], where the authors are interested in (IV-2.6) and resort to a collocation grid in order to simulate the time-evolution of the nonlinear part. A collocation grid may be smarter and cheaper than a uniform grid, but it still suffers for the well-known “curse of dimensionality”...

IV-2.2.4 Variational Gaussian wavepackets

An alternative method to the previously mentioned ones, introduced early and which is still widely used today, is the time-dependent variational approach proposed by Heller [43, 42]. For this method, Heller started by discretizing the solution as a sum of Gaussian functions, the so-called *Gaussian wavepackets*. The Gaussian functions possess many favorable properties, and one of them is the following: assume that $V(t, x, \psi) = V(x)$ is quadratic in (IV-2.1), if the initial condition ψ_0 is a Gaussian function, then the solution will remain Gaussian. The main component of this method is to let the Gaussian have time-dependent mean, moment, and width matrix.

In order to know the time-evolution of these Gaussian parameters, the Dirac-Frenkel variational principle is used. We note that recently, an alternative has been proposed, see [50]. We will use again later the Dirac-Frenkel principle, so it is essential to give some details about it. We postpone this detailed presentation to the end of the section.

When the potential is not quadratic, Heller proposed using a locally quadratic approximation of the potential around each wavepacket and he argues that, since the Gaussian functions are localized, the quadratic approximation of V is enough to recover accurately the dynamics of the wavepackets.

Remark IV.3

Even though the version by Heller is by far the most popular today, an early version of the variational wavepackets is due to Lebedeff [53], who only allowed a linear phase, and a single wavepacket. The work of Heller improves it by allowing a quadratic phase and several wavepackets.

Nowadays the variational Gaussian wavepackets are widely used, and this can be seen by the huge number of different methods based on the initial one by Heller, which is nowadays called the *Thawed Gaussian approximation*. As we explained previously, the Thawed Gaussian approximation works by letting the Gaussian's mean, momentum and width matrix be time-dependent. In contrast, the Frozen Gaussian approximation has been developed [41]. It consists in having Gaussian wavepackets for which the width matrix is fixed, and only the mean, momentum and complex phase are time-dependent.

> [...] the source of the term "Frozen Gaussian": the Gaussian packet moves along with its classical trajectory without changing shape. Like a rigid snowball in flight, the frozen Gaussian moves with the average position and momentum.
>
> Eric Heller (1981)

These two schemes have some remarkable properties: the first one is that both have the same $\mathbb{L}^2$ error bounds, even though one clearly has more freedom. A second remarkable property is that the collective oscillations appearing in both approximations allow some kind of averaging of the error, and thus the error of the superposition of Gaussians is lower than the sum of each individual error. However, these two methods are not mass or energy conserving for a non quadratic potential. This has to be put in perspective with the Dirac-Frenkel variational approach which is mass and energy conserving. See [51, Section 5.1] for a nice presentation of both methods and their properties, including error bounds.

Another important variant of the Gaussian wavepackets is due to Hagedorn and is now called *Hagedorn wavepackets* [39]. We know that the Hermite functions are the eigenfunctions of the harmonic oscillator $-\Delta + |x|^2$, what are the eigenfunctions for $-\Delta + V$ if V is allowed to be any quadratic function? The answer is Hagedorn's functions. Moreover, Hagedorn's functions allow more flexibility in terms of position and momentum than the usual Hermite functions, and most importantly they are multidimensional. Since Hagedorn wavepackets mimic Hermite functions, we can expect that they also satisfy some relations like sum rules and Rodriguez formula. Some of their properties have been proven in [52]. Some more details about the Hagedorn wavepackets can be found in [51, Section 4].

We refer to [51] for a recent, detailed and comprehensive review of the variational Gaussian wavepackets.

The variational methods mentioned above were studied and used with the linear Schrödinger equation (IV-2.1), or its semiclassical scaling (IV-1.5). To the author's knowledge, they were never used yet in the nonlinear setting. One of the aims of the work presented in Chapter IV-4 is to show that the variational methods can also be used in the nonlinear setting. Moreover, we will only deal with Gaussian functions, and this is mainly because they are well adapted for the linear Schrödinger equation. If one is interested in the cubic nonlinear Schrödinger equation, Gaussian functions are not adapted anymore but good candidates are the solitons. Hence, Chapter IV-4 can be understood as a first step towards the numerical simulation of the cubic nonlinear Schrödinger equation using solitons, which will be studied in future works.

The Dirac-Frenkel principle

The presentation given here follows the line of [51, Section 3.1]. In the context of this section, we are using the Dirac-Frenkel principle in order to numerically solve the linear Schrödinger equation (IV-2.1). The "Variational Gaussian wavepacket" method consists in seeking an approximation $u(t)$ to the solution $\psi(t)$, such that $u(t)$ lies in the following manifold:

$$\mathcal{M} = \left\{ v \in \mathbb{L}^2(\mathbb{R}^d) \middle| \begin{array}{c} v(x) = \exp\left[i\left((x-q)^T C(x-q) + p^T(x-q) + \zeta\right)\right] \\ p, q \in \mathbb{R}^d, \zeta \in \mathbb{C}, C = C^T \in \mathbb{C}^{d\times d}, \operatorname{Im} C \text{ positive definite.} \end{array} \right\}.$$

In order to have $u(t) \in \mathcal{M}$ at all times, we use the Dirac-Frenkel principle: it imposes that the residual of the Schrödinger equation is orthogonal to the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$ of the manifold $\mathcal{M}$ at the point $u(t)$. In other words, we are imposing the following condition:

$$\begin{aligned} &\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M} \quad \text{such that} \\ &\qquad (v, -i\partial_t u(t) - Hu(t))_{\mathbb{L}^2(\mathbb{R}^d)} = 0, \quad \forall v \in \mathcal{T}_{u(t)}\mathcal{M}. \end{aligned} \tag{IV-2.12}$$

The quantity $\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M}$ is an approximation to the true time derivative of $u(t)$, and (IV-2.12) translates the fact that we are orthogonally projecting the Schrödinger equation onto the tangent space of the manifold at point $u(t)$. Said differently, what is the best approximation of the time derivative of $u(t)$ such that $u(t)$ remains in the manifold? The answer is given by the orthogonal projection onto $\mathcal{T}_{u(t)}\mathcal{M}$ of the exact time derivative of $u(t)$, i.e. by the quantity $\partial_t u(t)$ that satisfies (IV-2.12).

Moreover, the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$ consists of derivatives of paths on $\mathcal{M}$ passing through $u(t)$, thus

$$\begin{aligned}\mathcal{T}_{u(t)}\mathcal{M} &= \left\{ \begin{array}{c} \frac{i}{2}\left(-2\dot{q}C(x-q) + (x-q)^T\dot{C}(x-q) + \dot{p}^T(x-q) - p^T\dot{q} + \dot{\zeta}\right) u(t) \\ \dot{p}, \dot{q} \in \mathbb{R}^d, \dot{\zeta} \in \mathbb{C}, \dot{C} = \dot{C}^T \in \mathbb{C}^{d\times d} \end{array} \right\} \\ &= \{\eta u \,|\, \eta \text{ is a complex } d\text{-variate polynomial of order at most } 2\}\,.\end{aligned}$$

We note that condition (IV-2.12) is equivalent in this case to the McLachlan approach [58], which consists in finding a minimizer of the following problem:

$$\min_{\varphi \in \mathcal{T}_{u(t)}\mathcal{M}} \|i\varphi - Hu(t)\|_{\mathbb{L}^2(\mathbb{R}^d)}.$$

When a minimizer φ is obtained, exactly or approximately, we then enforce the time derivative: $\partial_t u(t) = \varphi$. The equivalence between Dirac-Frenkel and McLachlan principles is due to [17], and the fact that the manifold can be reparametrized as follows:

$$\mathcal{M} = \left\{ v \in \mathbb{L}^2(\mathbb{R}^d) \,\middle|\, \begin{array}{c} v(x) = \exp\left[x^T(C_{\mathrm{Re}} + iC_{\mathrm{Im}})x + (\tilde{p}_{\mathrm{Re}} + i\tilde{p}_{\mathrm{Im}})x + \gamma_{\mathrm{Re}} + i\gamma_{\mathrm{Im}}\right] \\ \tilde{p}_{\mathrm{Re}}, \tilde{p}_{\mathrm{Im}}, \gamma_{\mathrm{Re}}, \gamma_{\mathrm{Im}} \in \mathbb{R}^d, C = (C_{\mathrm{Re}} + iC_{\mathrm{Im}}) = C^T \in \mathbb{C}^{d\times d}, \\ \operatorname{Im} C \text{ positive definite.} \end{array} \right\}.$$

Since the manifold $\mathcal{M}$ can be parametrized by pairs of complementary parameters (in the sense of [17]), then McLachlan and Dirac-Frenkel principles are equivalent on $\mathcal{M}$.

A remarkable thing is that, if the potential V is quadratic, then the approximation $u(t)$ of $\psi(t)$ is exact. Indeed, $-\Delta u(t) + Vu(t)$ is a complex d-variate polynomial or order 2, and therefore belongs to the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$. Thus, the true time derivative of $u(t)$ belongs to the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$, and thus the best approximation $\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M}$ of the true time derivative is the true time derivative itself. Therefore, $u(t)$ and $\psi(t)$ satisfy exactly the same equation, and no approximation is done here. This is probably the reason that lead Heller to consider local quadratic approximations of the potential if V is not quadratic.

Another great property of the Dirac-Frenkel principle is that it conserves mass and energy. We do not give the proof here, since it will be proven later in one setting of interest to us (see Section IV-4.2.2).

We also note that one general way to obtain $\partial_t u(t)$ in (IV-2.12) consists in obtaining a basis of the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$ and to compute exactly the $\mathbb{L}^2$ inner products involved. If we denote by $B_{u(t)}$ a basis of the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$, then (IV-2.12) is equivalent to

$$\begin{aligned} &\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M} \quad \text{such that} \\ &\qquad (b, -i\partial_t u(t))_{\mathbb{L}^2(\mathbb{R}^d)} = (b, Hu(t))_{\mathbb{L}^2(\mathbb{R}^d)}, \quad \forall b \in B_{u(t)}. \end{aligned}$$

The basis is finite-dimensional, so the above conditions can be expressed using a finite linear system of the form $\mathbf{AE} = \mathbf{S}$. Here, $\mathbf{A}$ is the *projection matrix*, $\mathbf{E}$ is a vector containing the time derivatives of the wavepackets parameters p, q, C, ζ, and $\mathbf{S}$ is a vector containing the $\mathbb{L}^2$ inner products appearing on the right-hand side. It is important to note that, if the family $B_{u(t)}$ has some redundancy, i.e. linear dependence between its functions, then the Dirac-Frenkel principle is known to yield unsatisfying results [47]. One way to overcome these issues is to drop the Thawed Gaussian approximation and to use the Frozen Gaussian approximation, or to use a matrix pseudoinverse instead of the inverse.

We refer to [51, Section 3] for more details about the Dirac-Frenkel method, including some error bounds. Some notable references are [26, 35, 68].

Part IV

CHAPTER 3 Modulation of solutions as a theoretical tool

This Chapter and the following are based on an unpublished joint work with Erwan Faou and Pierre Raphaël [30].

The basic idea is to consider functions that depends on a small number of time-dependent parameters, and to find the time derivatives of the parameters so that the function satisfies the linear Schrödinger equation. This function is called a *modulated solution*, and the general idea is *modulation*.

The initial idea comes from the works of Merle and Raphaël [60], Martel and Raphaël [57], and Faou and Raphaël [31], who study the infinite-time blow-up of the Schrödinger equation using modulated solutions. Borrowing the vocabulary from Faou and Raphaël, the modulated functions will be called *bubbles*. Inspired by these theoretical works, we use their ideas in order to devise an exact numerical algorithm. Basically, the algorithm consists in first decomposing the initial condition ψ_0 into a modulated Hermite basis, and then use the fact that the Hermite functions are eigenfunctions of the Quantum harmonic oscillator. This allows us to find conditions on the time-derivative of the modulation parameters, under the form of ODEs. It happens that these ODEs can be integrated exactly in this case, and this gives an exact numerical algorithm for the simulation of the time-dependent harmonic oscillator with arbitrary initial data in $\mathbb{L}^2(\mathbb{R}^d)$.

This Chapter only deals with the linear case, the next Chapter will treat the case of the cubic nonlinear Schrödinger equation:

$$i\partial_t\psi + \Delta_x\psi - |x|^2\psi = \psi|\psi|^2, \quad x \in \mathbb{R}^d. \tag{cNLS}$$

The end goal of the work done in the linear case is to make the nonlinear case a little simpler, as well as to introduce gently the idea behind the modulation.

Anticipating for the work to be done in the nonlinear case, we are motivated by recent works [57, 31] to discretize ψ, the solution to the Schrödinger equation (cNLS), as a sum of N modulated functions:

$$\psi(t,x) \approx u(t,x) := \sum_{j=1}^{N} u^j(t,x), \tag{IV-3.1}$$

where

$$u_j(t,x) := \frac{A^j}{L^j} e^{i\gamma^j + iL^j\beta^j\cdot y^j - i\frac{B^j}{4}|y^j|^2} v^j(s^j, y^j), \quad \text{with} \quad \left| \begin{aligned} \frac{\mathrm{d}s^j}{\mathrm{d}t} &:= \frac{1}{(L^j)^2}, \\ y^j &:= \frac{x - X^j}{L^j}, \end{aligned} \right. \tag{IV-3.2}$$

and $N \in \mathbb{N}^*$. In the cited works, the modulated functions u_j are called *bubbles*. Throughout this work, we may refer to the variables (s^j, y^j) as the *modulation frame* of the bubble labelled j.

The time dependence of the parameters $A^j, L^j, B^j, X^j, \beta^j, \gamma^j$ has not been written in (IV-3.2) for the sake of clarity, but it is one of the main ingredients of the approach. More precisely, the core idea is to plug the ansatz (IV-3.1)-(IV-3.2) into (cNLS) in order to obtain ODEs for the parameters.

Inspired by these successful theoretical works, we retain the idea of approximating solutions to (cNLS) by modulating the parameters A^j, L^j, B^j, X^j, β^j, γ^j in such a way that $v^j(s^j, y^j)$ satisfies a *smoother in time* equation. – typically a stationary soliton equation such as

$$-\Delta_y v + |y|^2 v + |v|^2 v = \lambda v. \tag{IV-3.3}$$

However, from the numerical point of view, choosing the v_j as stationary solitons would require first to solve explicitly the nonlinear equation (IV-3.3) and more problematically, to estimate numerically the nonlinear interactions between the modulated solitons by using the Dirac-Frenkel-MacLachlan principle. The latter consists essentially in a projection onto the manifold of modulated solitons, which is in practice very difficult to evaluate numerically. Moreover, one is naturally interested in using a splitting strategy between the linear and nonlinear parts, which would typically destroy the soliton structure in the equation. Following this idea, we split the Schrödinger equation (cNLS) into the linear part

$$i\partial_t\psi + \Delta_x\psi - |x|^2\psi = 0, \tag{HO}$$

and the nonlinear part

$$i\partial_t\psi = \psi|\psi|^2. \tag{NL}$$

The linear equation (HO) is also called the Quantum harmonic oscillator. For brevity, we will call it simply "harmonic oscillator". Traditional well-known numerical schemes are based on this abstract decomposition and it is easy to determine high-order splitting methods obtained by solving alternately the linear and nonlinear parts, like Lie, Strang Splitting or triple jump composition, see for instance [59, 40, 19]. However, the approximation of the solution to each of these two parts remains to be done using time and space discretizations. They are traditionally solved using grid-based numerical schemes (see for instance [8, 66, 28, 83, 14]). The computational complexity of grid-based methods is always an issue due to the bad scaling with respect to the dimension. Fortunately, using the modulation techniques given above, the solution to the linear part (HO) can be simulated exactly, in a straightforward manner, and very efficiently by considering Hermite decomposition of the functions v_j. The computational cost for the simulation of the linear part only is $\mathcal{O}(N \cdot d)$ – recall N is the number of bubbles and d the dimension – to be compared with grid-based complexities of order $\mathcal{O}(M^d)$ where M would be the number of discretization points in each dimension.

Notations

When dealing with the parameters of bubble labelled j, we write j as an exponent. For instance L^j is the parameter L corresponding to bubble j, and not some quantity L to the power j. If L^j has to be exponentiated, we will write $(L^j)^p$ to denote L^j to the power p.

We write X^j_k to denote the k-th component of the vector X^j (which is the X parameter of bubble j). Same goes for β^j and y^j.

When used as subscripts, t and s will always denote a time derivative (either with respect to time t or s). For instance, X^j_s denotes the derivative of the vector X^j with respect to time. We may also write $X^j_{k,s}$ to denote the derivative with respect to time s of the k-th component of the vector X^j (same goes for β^j and y^j). More explicitely,

$$X^j_{l,s}(s) = \frac{\mathrm{d}X^j_l(s)}{\mathrm{d}s}.$$

We use the shorthands $\partial_k = \frac{\partial}{\partial x_k}$ or $\partial_k = \frac{\partial}{\partial y_k}$, and the context will clarify which one of the two we use. We also use $\partial_t = \frac{\partial}{\partial t}$ and $\partial_s = \frac{\partial}{\partial s}$.

IV-3.1 Modulation

The idea of relying on time-dependent parameters to represent the solution, or an approximation, is not new and has been widely studied in the linear case, *i.e.* when the cubic nonlinearity is replaced by some multiplication with a potential. When the v_j are chosen as Gaussian functions, it has been called *Variational Gaussian wavepackets* and extensively analyzed by Lasser and Lubich [51], where they applied the Dirac-Frenkel-MacLachlan principe (DF principle) to the linear Schrödinger equation with potential.

More generally, this type of method using Gaussian functions is widely used in the field of Chemical Physics [42, 44, 24, 86, 1]. The different methods used are variations of the same idea, and possess many names: superposition of Gaussian Wavepackets, Gaussian beams, Thawed Gaussians, Frozen Gaussian...We refer to Section IV-2.2.4 for more details about variational Gaussian wavepackets.

At the end of this Chapter we obtain an algorithm which yields exact solutions to the harmonic oscillator (HO) by using modulated Gauss-Hermite functions. This algorithm can be easily implemented numerically, is grid-free, and is also able to capture high oscillations of the solution.

IV-3.1.1 Conservation Laws

We recall classical laws for the harmonic oscillator and cubic nonlinear Schrödinger equations (see for instance [79, 48]).

Before proceeding to the main conservation result, we will need an intermediate result, known as the Pohozaev identity.

Lemma IV.1: Pohozaev Identity

Let $x \in \mathbb{R}^d$, and $f \in H^1(\mathbb{R}^d)$ such that $xf \in \mathbb{L}^2(\mathbb{R}^d)$. Then

$$\int \Delta f \overline{\left(\frac{d}{2} f + x \cdot \nabla f\right)} dx = - \int |\nabla f|^2 dx. \tag{IV-3.4}$$

Proof. By density, we only need to prove equation (IV-3.4) for $f \in C_c^\infty(\mathbb{R}^d)$, where $C_c^\infty(\mathbb{R}^d)$ denotes the space of infinitely smooth functions with compact support in $\mathbb{R}^d$. Let

$$f_\lambda(x) := \lambda^{\frac{d}{2}} f(\lambda x),$$

then

$$\int |\nabla f_\lambda|^2 dx = \lambda^2 \int |\nabla f|^2 dx.$$

Differentiating this identity with respect to λ and evaluating the result at $\lambda = 1$ yields

$$\int \nabla f \cdot \overline{\nabla \left(\frac{d}{2} f + x \cdot \nabla f\right)} dx = \int |\nabla f|^2 dx.$$

We integrate by parts the LHS, and obtain (IV-3.4). □

We can now state the result about conserved quantities in (cNLS):

Lemma IV.2: Conserved quantities in dimension $d = 2$

We consider a two-parameter family of equations containing (HO) and (cNLS):

$$i\partial_t\psi + \mu(\Delta\psi - |x|^2\psi) = \lambda|\psi|^2\psi, \quad \mu, \lambda \in \mathbb{R}.$$

The (radial) conservation laws are mass $\|\psi\|_{\mathbb{L}^2}$, energy

$$E_{\mu,\lambda} = \frac{\mu}{2}\left\langle H\psi, \psi\right\rangle + \frac{\lambda}{4}\left\langle |\psi|^2\psi, \psi\right\rangle,$$

where $H = -\Delta + |x|^2$ and $\langle f, g\rangle := \int_{\mathbb{R}^d} f\bar{g}$, and momentum

$$M_{\mu,\lambda} = \left(E_{\mu,\lambda} - \mu\|x\psi\|_{\mathbb{L}^2}^2\right)^2 + \mu^2\left(\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi}\right)^2,$$

and the same applied to any power $(-H)^r\psi$, for $r \geq 1$. There also holds the non radial conservation law

$$\mathcal{P}^j = \frac{1}{4}\left(\mathrm{Im}\int \partial_j\psi\bar{\psi}\right)^2 + \mu^2\left(\int x^j|\psi|^2\right)^2, \quad j = 1, 2.$$

Proof. The proof is long, but since it is mostly composed of algebraic manipulations we can summarize now the main tools: the mass and energy conservations are obtained by simply differentiating the expression of these quantities with respect to time. The non radial conservation law is obtained by integrating by parts and differentiating $\int x^j|\psi|^2$. For the momentum conservation, we first differentiate with respect to time, then integrate by parts. After some algebraic manipulations, we use the Pohozaev identity (Lemma IV.1), and use the conservation of energy to conclude.

The proof ends on page 162.

Mass conservation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\|\psi\|_{\mathbb{L}^2}^2 &= \frac{\mathrm{d}}{\mathrm{dt}}\int|\psi|^2 = 2\mathrm{Re}\int\bar{\psi}\partial_t\psi = 2\mathrm{Re}\int -i\bar{\psi}\left(-\mu\Delta\psi + \mu|x|^2\psi + \lambda|\psi|^2\psi\right)\\ &= 2\mathrm{Re}\int -i\left(\mu\bar{\psi}\Delta\psi + \mu|x|^2|\psi|^2 + \lambda|\psi|^4\right) = 2\mu\mathrm{Re}\int -i\bar{\psi}\Delta\psi\\ &= 2\mu\mathrm{Re}\int i|\nabla\psi|^2 = 0.\end{aligned}$$

Energy conservation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E_{\mu,\lambda} &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle -\mu\Delta\psi + \mu|x|^2\psi + \frac{\lambda}{2}|\psi|^2\psi, \psi\right\rangle \\ &= \frac{1}{2}\left(-2\mu\mathrm{Re}\,\langle\Delta\psi, \partial_t\psi\rangle + 2\mu\mathrm{Re}\,\langle|x|^2\psi, \partial_t\psi\rangle + 4\mathrm{Re}\,\left\langle\frac{\lambda}{2}|\psi|^2\psi, \partial_t\psi\right\rangle\right) \\ &= \mathrm{Re}\,\left(i\langle\partial_t\psi, \partial_t\psi\rangle\right) = 0.\end{aligned}$$

For the momentum, we compute

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int|x|^2|\psi|^2 &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle|x|^2\psi, \psi\rangle = \mathrm{Re}\,\langle|x|^2\psi, \partial_t\psi\rangle = \mathrm{Re}\,\langle|x|^2\psi, i\mu\Delta\psi - i\mu|x|^2\psi - i\lambda|\psi|^2\psi\rangle \\ &= \mu\mathrm{Im}\,\langle|x|^2\psi, \Delta\psi\rangle = \mu\mathrm{Im}\,\int|x|^2\psi\Delta\bar{\psi} = -\mu\mathrm{Im}\,\int\nabla\bar{\psi}\cdot\nabla\left(|x|^2\psi\right) \\ &= -\mu\mathrm{Im}\,\int\nabla\bar{\psi}\cdot 2x\psi - \mu\mathrm{Im}\,\int\nabla\bar{\psi}\cdot\nabla\psi|x|^2 = -2\mu\mathrm{Im}\,\int x\cdot\nabla\bar{\psi}\psi \\ &= 2\mu\mathrm{Im}\,\int x\cdot\nabla\psi\bar{\psi}, \qquad \text{(IV-3.5)}\end{aligned}$$

and

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\,\int x\cdot\nabla\psi\bar{\psi} = \frac{1}{2}\mathrm{Im}\,\int\left(x\cdot\nabla\partial_t\psi\bar{\psi} + x\cdot\nabla\psi\partial_t\bar{\psi}\right).$$

An integration by parts gives

$$\int x\cdot\nabla\phi\psi = -\int\phi\nabla\cdot(x\psi) = -\int\phi\left(\psi d + x\cdot\nabla\psi\right),$$

hence

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\,\int x\cdot\nabla\psi\bar{\psi} &= \frac{1}{2}\mathrm{Im}\,\int\left(-\partial_t\psi\left(\bar{\psi}d + x\cdot\nabla\bar{\psi}\right) + x\cdot\nabla\psi\partial_t\bar{\psi}\right) \\ &= \frac{1}{2}\mathrm{Im}\,\int\left(-\partial_t\psi\bar{\psi}d - \partial_t\psi x\cdot\nabla\bar{\psi} + \partial_t\bar{\psi}x\cdot\nabla\psi\right) \\ &= \frac{1}{2}\mathrm{Im}\,\int\left(-\partial_t\psi\bar{\psi}d + 2i\mathrm{Im}\,\left[\partial_t\bar{\psi}x\cdot\nabla\psi\right]\right) \\ &= -\frac{d}{2}\mathrm{Im}\,\int\partial_t\psi\bar{\psi} + \mathrm{Im}\,\int\partial_t\bar{\psi}x\cdot\nabla\psi.\end{aligned}$$

Recall the equation satisfied by ψ:

$$\partial_t\psi = i\mu\Delta\psi - i\mu|x|^2\psi - i\lambda|\psi|^2\psi,$$

therefore

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi} = -\frac{d}{2}\mathrm{Im}\int i\left[\mu\Delta\psi-\mu|x|^2\psi-\lambda|\psi|^2\psi\right]\bar{\psi}\\ +\mathrm{Im}\int i\left[-\mu\Delta\bar{\psi}+\mu|x|^2\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi.\end{aligned}$$

We have

$$-\frac{d}{2}\mathrm{Im}\int i\left[\mu\Delta\psi-\mu|x|^2\psi-\lambda|\psi|^2\psi\right]\bar{\psi} = \frac{d}{2}\int\left[\mu|\nabla\psi|^2+\mu|x|^2|\psi|^2+\lambda|\psi|^4\right],$$

and

$$\mathrm{Im}\int i\left[-\mu\Delta\bar{\psi}+\mu|x|^2\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi = \mathrm{Re}\int\left[-\mu\Delta\bar{\psi}+\mu|x|^2\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi.$$

Moreover,

$$\begin{aligned}&\int|x|^2\bar{\psi}x\cdot\nabla\psi = -\int\psi\nabla\cdot\left(x|x|^2\bar{\psi}\right) = -\int\psi\left(d|x|^2\bar{\psi}+2|x|^2\bar{\psi}+x|x|^2\cdot\nabla\bar{\psi}\right)\\ \Longleftrightarrow\ &\int|x|^2\bar{\psi}x\cdot\nabla\psi+\overline{\int|x|^2\bar{\psi}x\cdot\nabla\psi} = -\int\psi\left(d|x|^2\bar{\psi}+2|x|^2\bar{\psi}\right)\\ &\Longleftrightarrow\ \mathrm{Re}\int|x|^2\bar{\psi}x\cdot\nabla\psi = -\int\psi\left(\frac{d}{2}|x|^2\bar{\psi}+|x|^2\bar{\psi}\right).\end{aligned}$$

Finally,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi} &= \frac{d}{2}\int\left[\mu|\nabla\psi|^2+\mu|x|^2|\psi|^2+\lambda|\psi|^4\right]\\ &\quad+\mathrm{Re}\int\left[-\mu\Delta\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi-\mu\int\psi\left(\frac{d}{2}|x|^2\bar{\psi}+|x|^2\bar{\psi}\right)\\ &= \frac{d}{2}\int\left[\mu|\nabla\psi|^2+\lambda|\psi|^4\right]+\mathrm{Re}\int\left[-\mu\Delta\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi-\mu\int|x|^2|\psi|^2\\ &= \frac{d}{2}\mu\int|\nabla\psi|^2-\mu\int|x|^2|\psi|^2+\frac{d}{2}\lambda\int|\psi|^4+\mathrm{Re}\int\left[-\mu\Delta\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi.\end{aligned}$$

We are in the two-dimensional case $d=2$, hence

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi} &= \int\mu|\nabla\psi|^2-\mu\int|x|^2|\psi|^2+\lambda\int|\psi|^4+\mathrm{Re}\int\left[-\mu\Delta\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi\\ &= 2E_\lambda+\frac{\lambda}{2}\int|\psi|^4-2\mu\int|x|^2|\psi|^2+\mathrm{Re}\int\left[-\mu\Delta\bar{\psi}+\lambda|\psi|^2\bar{\psi}\right]x\cdot\nabla\psi.\end{aligned}$$

Moreover,

$$\begin{aligned}\int |\psi|^2\bar{\psi}x\cdot\nabla\psi &= -\int \psi\nabla\cdot\left(|\psi|^2\bar{\psi}x\right)\\ &= -\int \psi\left(2\mathrm{Re}\,\left(\bar{\psi}\nabla\psi\right)\cdot\bar{\psi}x + |\psi|^2\nabla\bar{\psi}\cdot x + d|\psi|^2\bar{\psi}\right)\\ &= -\int \left(2\mathrm{Re}\,\left(\bar{\psi}\nabla\psi\right)\cdot|\psi|^2x + \psi|\psi|^2\nabla\bar{\psi}\cdot x + 2|\psi|^4\right)\\ \Longleftrightarrow\ 2\mathrm{Re}\int |\psi|^2\bar{\psi}x\cdot\nabla\psi &= -2\mathrm{Re}\int \bar{\psi}\nabla\psi\cdot|\psi|^2x - 2\int|\psi|^4\\ \Longleftrightarrow\ \mathrm{Re}\int |\psi|^2\bar{\psi}x\cdot\nabla\psi &= -\frac{1}{2}\int|\psi|^4,\end{aligned}$$

Finally,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi} &= 2E_\lambda + \frac{\lambda}{2}\int|\psi|^4 - 2\mu\int|x|^2|\psi|^2 - \mu\mathrm{Re}\int\Delta\bar{\psi}x\cdot\nabla\psi - \frac{\lambda}{2}\int|\psi|^4\\ &= 2E_\lambda - 2\mu\int|x|^2|\psi|^2 - \mu\mathrm{Re}\int\Delta\bar{\psi}x\cdot\nabla\psi.\end{aligned}$$

We then use the Pohozaev identity (IV-3.4) in dimension $d = 2$, which yields

$$\mathrm{Re}\left(\int x\cdot\nabla\psi\Delta\bar{\psi}\right) = 0.$$

Therefore,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi} = 2E_{\mu,\lambda} - 2\mu\int|x|^2|\psi|^2.$$

From the conservation of the energy E_λ and equation (IV-3.5),

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi} = -16\mu^2\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi}.$$

Hence, the conservation laws

$$\begin{aligned}&\frac{1}{16}\left(\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi}\right]\right)^2 + \mu^2\left(\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi}\right)^2\\ &= \left(E_{\mu,\lambda} - \mu\|x\psi\|_{\mathbb{L}^2}^2\right)^2 + \mu^2\left(\mathrm{Im}\int x\cdot\nabla\psi\bar{\psi}\right)^2.\end{aligned}$$

For the non radial conservation law:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\int\partial_j\psi\bar{\psi} &= -2\mathrm{Im}\int\partial_t\psi\overline{\partial_j\psi} = 2\mathrm{Re}\int i\partial_t\psi\overline{\partial_j\psi}\\ &= 2\mu\int|x|^2\mathrm{Re}\,\left(\psi\overline{\partial_j\psi}\right) = -2\mu\int x_j|\psi|^2,\end{aligned}$$

owing to the facts that integrations by parts yield

$$-\mathrm{Re}\int \Delta\psi\partial_j\bar{\psi} = \mathrm{Re}\int \partial_j\psi\Delta\bar{\psi},$$

and

$$2\mathrm{Re}\int |\psi|^2\psi\partial_j\bar{\psi} = \int |\psi|^2\partial_j|\psi|^2 = -\int |\psi|^2\partial_j|\psi|^2 \implies \mathrm{Re}\int |\psi|^2\psi\partial_j\bar{\psi} = 0.$$

We also have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int x_j|\psi|^2 = \mathrm{Re}\int x_j\partial_t\psi\bar{\psi} = \mathrm{Im}\int x_j i\partial_t\psi\bar{\psi} = \mu\mathrm{Im}\int -\Delta\psi x_j\bar{\psi} = \mu\mathrm{Im}\int \partial_j\psi\bar{\psi}.$$

Hence the relations

$$\left|\begin{array}{l}\dfrac{\mathrm{d}}{\mathrm{d}t}\displaystyle\int x_j|\psi|^2 = 2\mu\mathrm{Im}\int \partial_j\psi\bar{\psi} \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\mathrm{Im}\displaystyle\int \partial_j\psi\bar{\psi} = -2\mu\int x_j|\psi|^2,\end{array}\right.$$

which have the conservation law

$$\mathcal{P}_j = \frac{1}{4}\left(\mathrm{Im}\int \partial_j\psi\bar{\psi}\right)^2 + \mu^2\left(\int x_j|\psi|^2\right)^2.$$

□

IV-3.1.2 Obtaining modulation equations

By linearity of the harmonic oscillator, we can reduce the problem to calculating the evolution of the decomposition (IV-3.2) for only one bubble v^j. Throughout this Section, we fix an index $j \in [\![1, N]\!]$, and since this index is fixed we will not write it anymore. Hence, we can omit the superscript ^{j} for the time being. The notation u now denotes u^j since we are focusing on bubble labelled j and omitting the superscript. Our aim here is to see how the parameters and their time derivative are involved when plugging the decomposition (IV-3.2) into (HO).

Recall the expression of $u(t,x)$:

$$u(t,x) = \frac{A}{L}e^{i\gamma+iL\beta\cdot y-i\frac{B}{4}|y|^2}v(s,y), \quad y = \frac{x-X(t)}{L(t)}, \ \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{1}{L(t)^2}. \tag{IV-3.6}$$

We start by computing, in dimension $d \geq 1$,

$$\Delta_x u = \frac{Ae^{i\gamma}}{L^3}\Delta_y\left[e^{iL\beta\cdot y-i\frac{B}{4}|y|^2}v(s,y)\right].$$

By differentiating with respect to the k-th coordinate, we get

$$\partial_k \left[e^{iL\beta\cdot y - i\frac{B}{4}|y|^2} v\right] = e^{iL\beta\cdot y - i\frac{B}{4}|y|^2} \left[\partial_k v + i\left(L\beta_k - \frac{B}{2}y_k\right)v\right], \quad k = 1, \dots, d. \tag{IV-3.7}$$

Differentiating again with respect to the k-th coordinate:

$$\begin{aligned}
&\partial_k^2 \left[e^{iL\beta\cdot y - i\frac{B}{4}|y|^2} v\right] \\
&\quad = e^{iL\beta\cdot y - i\frac{B}{4}|y|^2} \begin{bmatrix} \partial_k^2 v + i\left(L\beta_k - \frac{B}{2}y_k\right)\partial_k v - i\frac{B}{2}v \\ \quad + i\left(L\beta_k - \frac{B}{2}y_k\right)\left[\partial_k v + i\left(L\beta_k - \frac{B}{2}y_k\right)v\right] \end{bmatrix} \\
&\quad = e^{iL\beta\cdot y - i\frac{B}{4}|y|^2} \begin{bmatrix} \partial_k^2 v + i\left(2L\beta_k - By_k\right)\partial_k v \\ \quad + \left(-i\frac{B}{2} - L^2\beta_k^2 + LB\beta_k y_k - \frac{B^2}{4}y_k^2\right)v \end{bmatrix}.
\end{aligned} \tag{IV-3.8}$$

By summing (IV-3.8) over $k = 1, \dots, d$, we get

$$\begin{aligned}
\Delta_x u = {} & \frac{A}{L^3} e^{i\gamma + iL\beta\cdot y - i\frac{B}{4}|y|^2} \\
& \times \left[\Delta_y v + i\left(2L\beta - By\right)\cdot\nabla v + \left(-i\frac{B}{2}d - L^2|\beta|^2 + LB\beta\cdot y - \frac{B^2}{4}|y|^2\right)v\right].
\end{aligned}$$

We have

$$\begin{aligned}
-|x|^2 u &= -\frac{A}{L} e^{i\gamma + iL\beta\cdot y - i\frac{B}{4}|y|^2} \left|Ly + X\right|^2 v \\
&= \frac{A}{L^3} e^{i\gamma + iL\beta\cdot y - i\frac{B}{4}|y|^2} \left(-L^4|y|^2 - 2L^3 X\cdot y - L^2|X|^2\right)v,
\end{aligned}$$

thus

$$\begin{aligned}
(\Delta_x - |x|^2)u = {} & \frac{A}{L^3} e^{i\gamma + iL\beta\cdot y - i\frac{B}{4}|y|^2} \left\{\Delta_y v - iB\left(\frac{d}{2}v + \Lambda v\right) - L^2\left(|\beta|^2 + |X|^2\right)v\right. \\
& \left. + 2iL\beta\cdot\nabla v + \left(LB\beta - 2L^3 X\right)\cdot yv + \left(-\frac{B^2}{4} - L^4\right)|y|^2 v\right\},
\end{aligned} \tag{IV-3.9}$$

where we denoted $\Lambda v := y \cdot \nabla v$. We now differentiate u with respect to time t:

$$\begin{aligned}
\partial_t u &= \partial_t \left(e^{i\gamma + i\beta\cdot(x-X) - i\frac{B}{4L^2}|x-X|^2} \frac{A}{L} v(s,y) \right) \\
&= e^{i\gamma + i\beta\cdot(x-X) - i\frac{B}{4L^2}|x-X|^2} \frac{A}{L} \left[\partial_t v + \frac{A_t}{A} v - \frac{L_t}{L}(v + \Lambda v) - \frac{X_t}{L} \cdot \nabla v \right] \\
&\quad + iu \begin{bmatrix} \gamma_t + \beta_t \cdot (x-X) - \beta \cdot X_t - \frac{B_t}{4L^2}|x-X|^2 \\ + \frac{2L_t B}{4L^3}|x - X(t)|^2 + \frac{2B}{4L^2}(x-X)\cdot X_t \end{bmatrix} \\
&= e^{i\gamma + i\beta\cdot(x-X) - i\frac{B}{4L^2}|x-X|^2} \frac{A}{L^3} \left[\partial_s v + \frac{A_s}{A} v - \frac{L_s}{L}(v + \Lambda v) - \frac{X_s}{L} \cdot \nabla v \right] \\
&\quad + \frac{1}{L^2} iu \left[\gamma_s + L\beta_s \cdot y - \beta \cdot X_s - \frac{B_s}{4}|y|^2 + \frac{2L_s B}{4L}|y|^2 + \frac{B}{2} y \cdot \frac{X_s}{L} \right],
\end{aligned}$$

where we recall that the subscripts $_t$ or $_s$ denote a time-differentiation with respect to time t or time s. Hence,

$$i\partial_t u = e^{i\gamma + i\beta\cdot(x-X) - i\frac{B}{4L^2}|x-X|^2} \frac{A}{L^3} \begin{bmatrix} i\partial_s v + (-\gamma_s + \beta \cdot X_s)v + \left(\frac{A_s}{A} - \frac{L_s}{L} \right) iv \\ - \frac{L_s}{L} i\Lambda v - i\frac{X_s}{L} \cdot \nabla v \\ + \left(-L\beta_s - \frac{BX_s}{2L} \right) \cdot yv + \left(\frac{B_s}{4} - \frac{B}{2}\frac{L_s}{L} \right) |y|^2 v \end{bmatrix}. \tag{IV-3.10}$$

This yields

$$\begin{aligned}
&(i\partial_t u + \Delta_x u - |x|^2 u)(t,x) \\
&= \frac{A}{L^3} e^{i\gamma + iL\beta\cdot y - i\frac{B}{4}|y|^2} \\
&\quad \times \begin{bmatrix} i\partial_s v + \left(-\gamma_s + \beta \cdot X_s - L^2 \left(|\beta|^2 + |X|^2 \right)\right) v \\ + \left(\frac{A_s}{A} - \frac{L_s}{L} - B\frac{d}{2} \right) iv + \left(-\frac{L_s}{L} - B \right) i\Lambda v \\ + i \left(2L\beta - \frac{X_s}{L} \right) \cdot \nabla v \\ + \left(-2L^3 X + LB\beta - L\beta_s - \frac{B}{2}\frac{X_s}{L} \right) \cdot yv \\ + \Delta_y v + \left[\frac{B_s}{4} - \left(\frac{B^2}{4} + L^4 \right) - \frac{B}{2}\frac{L_s}{L} \right] |y|^2 v \end{bmatrix} (s,y).
\end{aligned} \tag{IV-3.11}$$

Once we have Equation (IV-3.11), we are free to choose the parameters as we wish, Until now, the parameters have been completely free, with no condition whatsoever on them or their time derivatives. The main idea is to now choose conditions on them so that (IV-3.11) becomes an equation on v in variables (s,y) that is "easy" to solve. A natural choice is to conjugate the equation back to the original one in variables (s, y), *i.e.* to take

$$\begin{aligned} &\gamma_s - \beta\cdot X_s + L^2\left(|\beta|^2+|X|^2\right)=0, && 2L\beta - \frac{X_s}{L}=0,\\ &\frac{A_s}{A}-\frac{L_s}{L}-\frac{B}{2}d=0, && -2L^3X+LB\beta-L\beta_s-\frac{BX_s}{2L}=0,\\ &-\frac{L_s}{L}-B=0, && \frac{B_s}{4}-\left(\frac{B^2}{4}+L^4\right)-\frac{B}{2}\frac{L_s}{L}=-1, \end{aligned} \tag{IV-3.12}$$

so that v itself has to solve (HO) in variables (s, y). Then (IV-3.11) gives us the following equivalence:

$$(i\partial_t + \Delta_x - |x|^2)u(t,x)=0 \qquad \Longleftrightarrow \qquad (i\partial_s + \Delta_y - |y|^2)v(s,y)=0. \tag{IV-3.13}$$

The equivalence is straightforward if we assume (IV-3.12) to be satisfied. Now we see that if v is decomposed into the Hermite basis, we can solve explicitly (HO) for the function v in variables (s, y), and obtain the solution $u(t, x)$ after solving the differential system (IV-3.12).

Any function satisfying equation (IV-3.13) in variables (s, y) can be decomposed in the Hermite basis

$$\left\{\varphi_n(y) := H_{n_1}(y_1)\cdots H_{n_d}(y_d) : n\in\mathbb{N}^d\right\},$$

where the function $H_k(z)$ denotes the Hermite function of order $k \in \mathbb{N}$, which satisfies the following diffferential equation:

$$H_k''(z) + (2k+1-z^2)H_k(z)=0, \quad z\in\mathbb{R}.$$

A straightforward calculation shows that

$$(-\Delta_y + |y|^2)\varphi_n = (2|n|+d)\varphi_n,$$

where $|n| := \sum_{k=1}^d n_k$. Hence from a decomposition

$$v(0,y)=\sum_{n\in\mathbb{N}^d} v_n\varphi_n(y) \tag{IV-3.14}$$

with $v_n \in \mathbb{C}$, we calculate that

$$v(s,y)=\sum_{n\in\mathbb{N}^d} v_n e^{-(2|n|+d)is}\varphi_n(y) \tag{IV-3.15}$$

is the solution of (IV-3.13) in variables (s, y). It remains to obtain $u(t, x)$ solution of (HO). In order to do this, one simply needs to integrate (IV-3.12) (which is independent of n in the Hermite decomposition), and to plug (IV-3.15) into (IV-3.6). In particular, we need to calculate the time $s(t)$ as a function of the original time t.

IV-3.1.3 Integrability of the modulation equations

Our aim here is to show that the ordinary differential system (IV-3.12) can be integrated exactly with explicit formulas for the parameters. We rewrite (IV-3.12) as

$$\begin{aligned} A_s &= \frac{AB}{2}(d-2), & L_s &= -BL \\ B_s &= -4+4L^4-B^2, & X_s &= 2L^2\beta \\ \beta_s &= -2L^2X, & \gamma_s &= L^2\left(|\beta|^2-|X|^2\right). \end{aligned} \tag{IV-3.16}$$

In time t, as $\frac{\mathrm{d}}{\mathrm{d}s} = L^2\frac{\mathrm{d}}{\mathrm{d}t}$, this system is

$$\begin{aligned} A_t &= \frac{AB}{2L^2}(d-2), & L_t &= -\frac{B}{L} = -2L\partial_B\mathcal{E} \\ B_t &= -\frac{4}{L^2}+4L^2-\frac{B^2}{L^2} = 2L\partial_L\mathcal{E}, & X_t &= 2\beta = \nabla_\beta\mathcal{R} \\ \beta_t &= -2X = -\nabla_X\mathcal{R}, & \gamma_t &= |\beta|^2-|X|^2, \end{aligned} \tag{IV-3.17}$$

with

$$\mathcal{E}(B,L) = \frac{1}{L^2}\left(1+\frac{B^2}{4}\right)+L^2, \qquad \text{and} \qquad \mathcal{R}(X,\beta) = |X|^2+|\beta|^2.$$

Let us write explicitly the Darboux-Lie transformation associated with the previous Poisson system (see e.g. [40]), to obtain canonical Hamiltonian coordinates. We set

$$k = \frac{1}{2}\log L, \quad L = e^{2k},$$

and the system becomes

$$\left|\begin{aligned} k_t &= -\partial_B\mathcal{H} \\ B_t &= \partial_k\mathcal{H} \\ X_t &= \nabla_\beta\mathcal{H} \\ \beta_t &= -\nabla_X\mathcal{H} \\ A_t &= \frac{AB}{2}(d-2)e^{-4k} \\ \gamma_t &= |\beta|^2-|X|^2, \end{aligned}\right. \tag{IV-3.18}$$

with

$$\mathcal{H}(k,B,X,\beta) = \mathcal{E}(k,B)+\mathcal{R}(X,\beta) = e^{-4k}\left(\frac{B^2}{4}+1\right)+e^{4k}+|X|^2+|\beta|^2.$$

Lemma IV.3: Action-angle variables and exact update formulas

There exists a symplectic change of variable $(X, B, k, \beta) \mapsto (h, a, \xi, \theta) \in \mathbb{R} \times \mathbb{R}^d \times [0, 2\pi] \times [0, 2\pi]^d$, such that the Hamiltonian in these variables is given by

$$E(h, a, \xi, \theta) = 4h + 2|a|^2, \tag{IV-3.19}$$

so that the flow in variable (h, a, ξ, θ) is given by

$$\begin{aligned} a(t) &= a(0), & \theta(t) &= \theta(0) + 2t, \\ h(t) &= h(0), & \xi(t) &= \xi(0) - 4t. \end{aligned} \tag{IV-3.20}$$

We have the explicit formulas:

$$\begin{aligned} A(t) &= A(0) \left(\frac{L(t)}{L(0)} \right)^{\frac{2-d}{2}}, \\ e^{4k(t)} &= L(t)^2 = 2h(t) - \cos(\xi(t))\sqrt{4h(t)^2 - 1}, \\ B(t) &= 2\sin(\xi(t))\sqrt{4h(t)^2 - 1}, \\ X_i(t) &= \sin(\theta_i(t))\sqrt{2a_i(t)}, \quad i = 1, \dots, d, \\ \beta_i(t) &= \cos(\theta_i(t))\sqrt{2a_i(t)}, \quad i = 1, \dots, d, \\ \gamma(t) &= \gamma(0) + \sum_{l=1}^{d} \frac{a_l(0)}{2} \left[\sin(2\theta_l(t)) - \sin(2\theta_l(0))\right] \\ s(t) &= -\frac{1}{2}\arctan\left(\left(2h(0) + \sqrt{4h(0)^2 - 1}\right) \tan\left(\frac{\xi(0)}{2} - 2t \right) \right) \\ &\quad + \frac{1}{2}\arctan\left(\left(2h(0) + \sqrt{4h(0)^2 - 1}\right) \tan\left(\frac{\xi(0)}{2} \right) \right) + m_t \frac{\pi}{2}, \end{aligned} \tag{IV-3.21}$$

where, if $m_0 \in \mathbb{Z}$ is such that $\frac{\xi(0)}{2} \in m_0\pi + [-\frac{\pi}{2}, \frac{\pi}{2}]$, then $m_t \in \mathbb{Z}$ is defined by $\frac{\xi(t)}{2} \in (m_0 - m_t)\pi + [-\frac{\pi}{2}, \frac{\pi}{2}]$.

Proof. The proof of this Lemma consists in using action-angle variables, and then using explicit integrals in order to obtain the exact update formulas. The proof ends on page 174.

For the (X, β) part, it suffices to check that the change of variable $(X, \beta) \mapsto (a, \theta)$ defined by

$$X_i = \sqrt{2a_i}\sin(\theta_i) \quad \text{and} \quad \beta_i = \sqrt{2a_i}\cos(\theta_i) \qquad i = 1, \dots, d,$$

is symplectic and that

$$X_i = \sqrt{2a_i(0)}\sin(2t + \theta_i(0)) \quad \text{and} \quad \beta_i = \sqrt{2a_i(0)}\cos(2t + \theta_i(0)) \qquad i = 1, \dots, d,$$

are solutions. Thus,

$$a_i(t) = a_i(0) \qquad \text{and} \qquad \theta_i(t) = \theta_i(0) + 2t. \tag{IV-3.22}$$

For the (k, B) part we use the method of generating functions, described e.g. in [40, Sect. VI.5]. We can express B in terms of k and the Hamiltonian $\mathcal{E}$, so that on the set $\{B > 0\}$ we have:

$$B = 2\sqrt{e^{4k}\mathcal{E} - e^{8k} - 1}. \tag{IV-3.23}$$

This equality holds for $e^{4k} \in [e^{4k_0}, e^{4k_1}]$, where e^{4k_0}, e^{4k_1} are the reals roots of the polynomial $-z^2 + \mathcal{E}z - 1$,

$$e^{4k_0} = \frac{1}{2}\left(\mathcal{E} - \sqrt{\mathcal{E}^2 - 4}\right), \quad e^{4k_1} = \frac{1}{2}\left(\mathcal{E} + \sqrt{\mathcal{E}^2 - 4}\right). \tag{IV-3.24}$$

In order to obtain a symplectic change of variables, we look for a function $S(k, \mathcal{E})$ such that

$$B = \frac{\partial S}{\partial k}(k, \mathcal{E}).$$

We easily obtain $S(k, \mathcal{E})$, by integrating on $[k_0, k]$:

$$S(k, \mathcal{E}) = 2\int_{k_0}^{k} \sqrt{e^{4z}\mathcal{E} - e^{8z} - 1}dz.$$

The variable ϕ which makes the mapping $(B, k) \mapsto (\phi, \mathcal{E})$ symplectic is defined by

$$\phi = \frac{\partial S}{\partial \mathcal{E}}(k, \mathcal{E}) = \int_{k_0}^{k} \frac{e^{4z}}{\sqrt{e^{4z}\mathcal{E} - e^{8z} - 1}}dz.$$

We have

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{e^{4k}k_t}{\sqrt{e^{4k} - e^{8k} - 1}} = \frac{-e^{4k}\partial^B\mathcal{E}}{\frac{B}{2}} = \frac{-e^{-4k}\frac{B}{2}e^{4k}}{\frac{B}{2}} = -1.$$

We now proceed to obtaining an explicit expression for ψ:

$$\begin{aligned}\phi &= \int_{k_0}^{k} \frac{e^{4z}}{\sqrt{e^{4z}\mathcal{E} - e^{8z} - 1}} dz = \frac{1}{4}\int_{e^{4k_0}}^{e^{4k}} \frac{1}{\sqrt{\mathcal{E}u - u^2 - 1}} du \\ &= \frac{1}{4\sqrt{\frac{\mathcal{E}^2}{4} - 1}} \int_{e^{4k_0}}^{e^{4k}} \frac{1}{\sqrt{1 - \left(\frac{u - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}\right)^2}} du \\ &= \frac{1}{4}\int_{\frac{e^{4k_0} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}}^{\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}} \frac{1}{\sqrt{1 - u^2}} du.\end{aligned}$$

Recall the definition (IV-3.24) of k_0, which yields

$$e^{4k_0} - \frac{\mathcal{E}}{2} = -\sqrt{\frac{\mathcal{E}^2}{4} - 1}.$$

Therefore,

$$\begin{aligned}\phi &= \frac{1}{4}\int_{-1}^{\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}} \frac{1}{\sqrt{1 - u^2}} du = \frac{1}{4}\left(\arcsin\left(\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}\right) + \frac{\pi}{2}\right) \\ &= \frac{1}{4}\arcsin\left(\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}\right) + \frac{\pi}{8} \in \left[0, \frac{\pi}{4}\right].\end{aligned}$$

We want the angle variable to lie in $[0, 2\pi]$ so the above expression describes an eighth of a period. But we are only considering the set $\{B > 0\}$, thus the angle ξ we are looking for must lie only in $[0, \pi]$. Hence we set $(\xi, h) = (4\phi, \mathcal{E}/4)$ and let the Hamiltonian $\mathcal{E}(\xi, h) = 4h$ with a slight abuse of notation. It is then clear that $\frac{\mathrm{d}h}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}\xi}{\mathrm{d}t} = -4$. Moreover,

$$\xi = \arcsin\left(\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}\right) + \frac{\pi}{2} \in [0, \pi], \tag{IV-3.25}$$

and hence

$$\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}} = \sin\left(\xi - \frac{\pi}{2}\right) = -\cos(\xi).$$

We obtain

$$\begin{aligned} e^{4k} = L^2 = \frac{\mathcal{E}}{2} - \cos(\xi)\sqrt{\frac{\mathcal{E}^2}{4} - 1} = 2h - \cos(\xi)\sqrt{4h^2-1} \\ = 2h\left(1 - \cos(\xi)\sqrt{1 - \frac{1}{4h^2}}\right). \end{aligned}$$

With this formula, we have

$$0 < L^2 < 4h = \mathcal{E},$$

and (IV-3.23) becomes

$$\begin{aligned} B = 2\sqrt{\mathcal{E}e^{4k} - e^{8k} - 1} = 2\sqrt{4he^{4k} - (e^{4k})^2 - 1} = 2\sqrt{(4h^2-1)\sin^2(\xi)} \\ = 2\sin(\xi)\sqrt{4h^2-1}, \end{aligned}$$

where the last equality holds for $\xi \in [0, \pi]$.

We can now integrate the equations for A, and γ. The first one is

$$A_t = \frac{AB}{2}(d-2)e^{-4k}.$$

From the expressions we just obtained we get

$$A_t = A(d-2)\frac{\sin(\xi)\sqrt{4h^2-1}}{2h - \cos(\xi)\sqrt{4h^2-1}}.$$

The solution to this equation is of the form

$$A(t) = A(0)\exp\left\{(d-2)\int_0^t \frac{\sin(\xi(\sigma))\sqrt{4h(\sigma)^2-1}}{2h(\sigma) - \cos(\xi(\sigma))\sqrt{4h(\sigma)^2-1}}\mathrm{d}\sigma\right\}.$$

Moreover, we know that $\sigma \mapsto h(\sigma)$ is constant, and that $\xi(\sigma) = \xi(0) - 4\sigma$. Hence we have to solve

$$A(t) = A(0)\exp\left\{(d-2)\int_0^t \frac{\sin(\xi(0)-4\sigma)\sqrt{4h(0)^2-1}}{2h(0) - \cos(\xi(0)-4\sigma)\sqrt{4h(0)^2-1}}\mathrm{d}\sigma\right\}.$$

One can easily check that we have the following equality:

$$\begin{aligned}
&\int_0^t \frac{\sin(\xi(0)-4\sigma)\sqrt{4h(0)^2-1}}{2h(0)-\cos(\xi(0)-4\sigma)\sqrt{4h(0)^2-1}}\mathrm{d}\sigma \\
&= -\frac{1}{4}\left[\log\left(2h(t)-\cos(\xi(t))\sqrt{4h(t)^2-1}\right)-\log\left(2h(0)-\cos(\xi(0))\sqrt{4h(0)^2-1}\right)\right].
\end{aligned}$$

Note that, unless $h(0)=\frac{1}{2}$ or $h(t)=\frac{1}{2}$, these quantities are well-defined since $2h(\sigma) > \sqrt{4h(\sigma)^2-1}, \sigma \in \{0,t\}$. Thus, we obtain

$$\begin{aligned}
A(t) &= A(0)e^{\frac{2-d}{4}\left[\log\left(2h(t)-\cos(\xi(t))\sqrt{4h(t)^2-1}\right)-\log\left(2h(0)-\cos(\xi(0))\sqrt{4h(0)^2-1}\right)\right]} \\
&= C\left(2h(0)-\cos(\xi(0)-4t)\sqrt{4h(0)^2-1}\right)^{\frac{2-d}{4}},
\end{aligned}$$

where we defined $C := A(0)\left(2h(0)-\cos(\xi(0))\sqrt{4h(0)^2-1}\right)^{\frac{d-2}{4}}$. We recognize here the expressions for $L(0)^2$ and $L(t)^2$.

Let us finally turn to the expression for $\gamma(t)$. We proceed to the direct integration of γ_t. We have

$$\begin{aligned}
\gamma(t)-\gamma(0) &= \int_0^t \dot{\gamma}(\tau)d\tau = \int_0^t \left[|\beta(\tau)|^2-|X(\tau)|^2\right]d\tau \\
&= \int_0^t \left\{\sum_{l=1}^d 2a_l\cos(\theta_l(\tau))^2 - \sum_{l=1}^d 2a_l\sin(\theta_l(\tau))^2\right\}d\tau \\
&= \int_0^t 2\sum_{l=1}^d a_l\left(\cos(\theta_l(\tau))^2-\sin(\theta_l(\tau))^2\right)d\tau \\
&= \int_0^t \sum_{l=1}^d 2a_l\cos(2\theta_l(\tau))d\tau \\
&= \sum_{l=1}^d \frac{a_l}{2}\left[\sin(2\theta_l(t))-\sin(2\theta_l(0))\right],
\end{aligned}$$

where the last equality has been obtained using (IV-3.22).

Finally, we calculate the evolution of the time $s(t)$ in term of the original time t. Owing to the expression of $L(t)$ we obtained earlier,

$$\begin{aligned} s(t) := \int_0^t \frac{1}{L(\tau)^2} dt &= \int_0^t \frac{1}{\underbrace{2h(0)}_{=:c_1} - \underbrace{\sqrt{4h(0)^2-1}}_{=:c_2} \cos(\xi(0)-4\tau)} d\tau \\ &= \int_0^t \frac{1}{c_1 - c_2\cos(\xi(0)-4\tau)} d\tau \\ &= \frac{1}{4}\int_{\xi(0)-4t}^{\xi(0)} \frac{1}{c_1 - c_2\cos(\tau)} d\tau. \end{aligned}$$

Recall the following trigonometric identity:

$$\cos(2\tau) = \frac{1-\tan(\tau)^2}{1+\tan(\tau)^2}, \quad \tau \in \mathbb{R},$$

hence

$$\begin{aligned} &\int_0^t \frac{1}{c_1 - c_2\cos(\xi(0)-4\tau)} d\tau \\ &= \frac{1}{4}\int_{\xi(0)-4t}^{\xi(0)} \frac{1}{c_1 - c_2\frac{1-\tan(\tau/2)^2}{1+\tan(\tau/2)^2}} d\tau \\ &= \frac{1}{4}\int_{\xi(0)-4t}^{\xi(0)} \frac{1+\tan(\tau/2)^2}{c_1(1+\tan(\tau/2)^2) - c_2(1-\tan(\tau/2)^2)} d\tau \\ &= \frac{1}{4}\int_{\xi(0)-4t}^{\xi(0)} \frac{1+\tan(\tau/2)^2}{(c_1+c_2)\tan(\tau/2)^2 + c_1 - c_2} d\tau \\ &= \frac{1}{4(c_1-c_2)}\int_{\xi(0)-4t}^{\xi(0)} \frac{1+\tan(\tau/2)^2}{\frac{c_1+c_2}{c_1-c_2}\tan(\tau/2)^2 + 1} d\tau \\ &= \frac{1}{2(c_1-c_2)}\int_{\frac{\xi(0)}{2}-2t}^{\frac{\xi(0)}{2}} \frac{1+\tan(\tau)^2}{\frac{c_1+c_2}{c_1-c_2}\tan(\tau)^2 + 1} d\tau \\ &= \frac{1}{2(c_1-c_2)}\int_{\frac{\xi(0)}{2}-2t}^{\frac{\xi(0)}{2}} \frac{\frac{d}{d\tau}(\tan(\tau))}{\frac{c_1+c_2}{c_1-c_2}\tan(\tau)^2 + 1} d\tau \\ &= \frac{1}{2(c_1-c_2)} \frac{1}{\sqrt{\frac{c_1+c_2}{c_1-c_2}}} \int_{\frac{\xi(0)}{2}-2t}^{\frac{\xi(0)}{2}} \frac{\frac{d}{d\tau}\left(\sqrt{\frac{c_1+c_2}{c_1-c_2}}\tan(\tau)\right)}{\left[\sqrt{\frac{c_1+c_2}{c_1-c_2}}\tan(\tau)\right]^2 + 1} d\tau. \end{aligned}$$

Moreover, $(c_1-c_2)(c_1+c_2) = c_1^2 - c_2^2 = (2h)^2 - (4h^2-1) = 1$ and $c_1 - c_2 > 0$, thus

$\sqrt{\frac{c_1+c_2}{c_1-c_2}} = (c_1+c_2)$ and

$$\int_0^t \frac{1}{L(\tau)^2} d\tau = \frac{1}{2} \int_{\frac{\xi(0)}{2}-2t}^{\frac{\xi(0)}{2}} \frac{\frac{d}{d\tau}\left((c_1+c_2)\tan(\tau)\right)}{\left((c_1+c_2)\tan(\tau)\right)^2+1} d\tau.$$

Now let $m_0 \in \mathbb{Z}$ such that $\frac{\xi(0)}{2} \in m_0\pi + (-\frac{\pi}{2}, \frac{\pi}{2}]$, and $m_t \in \mathbb{Z}$ such that $\frac{\xi(t)}{2} \in (m_0 - m_t)\pi + (-\frac{\pi}{2}, \frac{\pi}{2}]$. We recall that $\xi(t) = \xi(0) - 4t$. Then

$$\begin{aligned} \int_0^t \frac{1}{L(\tau)^2} d\tau &= \frac{1}{2} \int_{\frac{\xi(0)}{2}-2t}^{\frac{\xi(0)}{2}} \underbrace{\frac{\frac{d}{d\tau}\left((c_1+c_2)\tan(\tau)\right)}{\left((c_1+c_2)\tan(\tau)\right)^2+1}}_{=:f(\tau)} d\tau \\ &= \frac{1}{2}\int_{m_0\pi-\frac{\pi}{2}}^{\frac{\xi(0)}{2}} f(\tau)d\tau + \frac{1}{2}\int_{(m_0-1)\pi-\frac{\pi}{2}}^{m_0\pi-\frac{\pi}{2}} f(\tau)d\tau + \cdots + \frac{1}{2}\int_{\frac{\xi(0)}{2}-2t}^{(m_0-m_t)\pi+\frac{\pi}{2}} f(\tau)d\tau. \end{aligned}$$

For $m \in \mathbb{Z}$, we have

$$\begin{aligned} \int_{m\pi-\frac{\pi}{2}}^{m\pi+\frac{\pi}{2}} f(\tau)d\tau &= \left[\arctan\left((c_1+c_2)\tan(\tau)\right)\right]_{m\pi-\frac{\pi}{2}}^{m\pi+\frac{\pi}{2}} \\ &= \left[\arctan\left((c_1+c_2)\tan(\tau)\right)\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \pi. \end{aligned}$$

Now write $\widetilde{\frac{\xi(0)}{2}} := \frac{\xi(0)}{2} - m_0\pi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$, and $\widetilde{\frac{\xi(\tau)}{2}} := \frac{\xi(\tau)}{2} - (m_0 - m_t)\pi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$. Then,

$$\begin{aligned}
&\int_0^t \frac{1}{L(\tau)^2} d\tau \\
&= \frac{1}{2}(m_t - 1)\pi + \frac{1}{2}\int_{m_0\pi - \frac{\pi}{2}}^{\frac{\xi(0)}{2}} f(\tau)d\tau + \frac{1}{2}\int_{\frac{\xi(0)}{2} - 2t}^{(m_0 - m_t)\pi + \frac{\pi}{2}} f(\tau)d\tau \\
&= (m_t - 1)\frac{\pi}{2} + \frac{1}{2}\int_{-\frac{\pi}{2}}^{\frac{\widetilde{\xi(0)}}{2}} f(\tau)d\tau + \frac{1}{2}\int_{\frac{\widetilde{\xi(t)}}{2}}^{\frac{\pi}{2}} f(\tau)d\tau \\
&= (m_t - 1)\frac{\pi}{2} + \frac{1}{2}\left[\arctan\left((c_1 + c_2)\tan(\tau)\right)\right]_{-\frac{\pi}{2}}^{\frac{\widetilde{\xi(0)}}{2}} + \frac{1}{2}\left[\arctan\left((c_1 + c_2)\tan(\tau)\right)\right]_{\frac{\widetilde{\xi(t)}}{2}}^{\frac{\pi}{2}} \\
&= (m_t - 1)\frac{\pi}{2} + \frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\widetilde{\xi(0)}}{2}\right)\right) + \frac{\pi}{2} \\
&\qquad + \frac{\pi}{2} - \arctan\left((c_1 + c_2)\tan\left(\frac{\widetilde{\xi(t)}}{2}\right)\right) \\
&= m_t\frac{\pi}{2} + \frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\widetilde{\xi(0)}}{2}\right)\right) - \frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\widetilde{\xi(t)}}{2}\right)\right) \\
&= m_t\frac{\pi}{2} + \frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\xi(0)}{2}\right)\right) - \frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\xi(0)}{2} - 2t\right)\right).
\end{aligned}$$

Hence

$$\begin{aligned}
s(t) = \int_0^t \frac{1}{L(\tau)^2} d\tau = &-\frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\xi(0)}{2} - 2t\right)\right) \\
&+ \frac{1}{2}\arctan\left((c_1 + c_2)\tan\left(\frac{\xi(0)}{2}\right)\right) + m_t\frac{\pi}{2}.
\end{aligned}$$

□

If one knows the parameters $(A, L, B, X, \beta, \gamma)$, it suffices to apply (IV-3.21) in order to update them. Then we combine these expressions with the decomposition (IV-3.15) and the expression of $s(t)$ to obtain the expression of $u(t, x)$.

In practice, we first perform a bubble decomposition of the initial condition in order to write it under the form (IV-3.1)-(IV-3.2). This gives the value of parameters at time $t = 0$. Then, in order to update the modulation parameters using (IV-3.21), we first have to compute the corresponding action-angle variables. We have the following result:

Lemma IV.4: Action-angle variables from the parameters

The change of variables $(L, B, X, \beta) \mapsto (h, a, \xi, \theta)$ is explicit, and at time $t = 0$ we have

$$\begin{aligned} a_i(0) &= \frac{1}{2}\left(X_i(0)^2 + \beta_i(0)^2\right), \quad i = 1, \dots, d, \\ \theta_i(0) &= \arctan\left(\frac{X_i(0)}{\beta_i(0)}\right), \quad i = 1, \dots, d, \\ h(0) &= \frac{L(0)^4 + 1 + \frac{B(0)^2}{4}}{4L(0)^2}, \\ \xi(0) &= \arctan\left(\frac{B(0)}{4h(0) - 2L(0)^2}\right), \end{aligned} \tag{IV-3.26}$$

whenever $\theta_i(0)$ and $\xi(0)$ are well-defined. When any one of them is ill-defined – which happens when $X_i(0) = \beta_i(0) = 0, i \in \{1, \dots, d\}$ or when $L(0) = 1$ and $B(0) = 0$ – any value can be taken and the time-evolution of $A(t)$, $L(t)$, $B(t)$, $X(t)$, $\beta(t)$ and $\gamma(t)$ will not depend on the value. Moreover, in the cases where $a_i(0) = 0$, $i \in \{1, \dots, d\}$ or $h(0) = \frac{1}{2}$, the formula (IV-3.26) for $\theta_i(0)$, $i \in \{1, \dots, d\}$ or $\xi(0)$ are ill-defined, but any value can be taken as a substitution and this will not affect the behavior of the mappings $t \mapsto \gamma(t)$ and $t \mapsto s(t)$.

Proof. The proof of this Lemma consists in inverting the expressions (IV-3.21) using algebraic manipulations. The proof ends on page 177.

We have $a_i(0) = \frac{1}{2}\left(X_i(0)^2 + \beta_i(0)^2\right), i = 1, \dots, d$. If $a_i(0) > 0$ we can define $\theta_i(0)$ as $\theta_i(0) = \arctan\left(\frac{X_i(0)}{\beta_i(0)}\right)$. Otherwise, if $a_i(0) = 0$, then we recall that $a(t) = a(0)$ and hence – whatever $\theta(0)$ – we have $X_i(t) = 0$ and $\beta_i(t) = 0$. Therefore, in the case $a_i(0) = 0$, the exact value of $\theta_i(0)$ does not change the behavior of $t \mapsto (X_i(t), \beta_i(t))$.

For the (L, B) part,

$$L(0)^2 - 2h(0) = -\cos(\xi(0))\sqrt{4h(0)^2 - 1},$$

hence

$$(L(0)^2 - 2h(0))^2 = L(0)^4 - 4L(0)^2 h(0) + 4h(0)^2 = \cos(\xi(0))^2\left(4h(0)^2 - 1\right).$$

We also have

$$\left(\frac{B(0)}{2}\right)^2 = \frac{B(0)^2}{4} = \sin(\xi(0))^2\left(4h(0)^2 - 1\right).$$

Then,

$$L(0)^4 - 4L(0)^2 h(0) + 4h(0)^2 + \frac{B(0)^2}{4} = 4h(0)^2 - 1,$$

that is

$$4L(0)^2h(0) = L(0)^4 + \frac{B(0)^2}{4} + 1.$$

We deduce that $h(0), L(0) \neq 0$, and therefore

$$h(0) = \frac{L(0)^4 + \frac{B(0)^2}{4} + 1}{4L(0)^2}.$$

Note that $h(0)$ is bounded from below by $\frac{1}{2}$. Indeed,

$$\begin{aligned} & L(0)^4 - 2L(0)^2 + 1 + \frac{B(0)^2}{4} = \left(L(0)^2 - 1\right)^2 + \frac{B(0)^2}{4} \geq 0 \\ \Longleftrightarrow\ & L(0)^4 + 1 + \frac{B(0)^2}{4} \geq 2L(0)^2 \\ \Longleftrightarrow\ & h(0) \geq \frac{1}{2}. \end{aligned}$$

From this we also get that $h(0) = \frac{1}{2} \iff L(0)^2 = 1$ and $B(0) = 0$.

If $h(0) > \frac{1}{2}$, we have

$$\begin{cases} 2h(0) - L(0)^2 = \cos(\xi(0))\sqrt{4h(0)^2 - 1} \\ \qquad\quad \dfrac{B(0)}{2} = \sin(\xi(0))\sqrt{4h(0)^2 - 1}, \end{cases} \implies \frac{B(0)/2}{2h(0) - L(0)^2} = \tan(\xi(0)),$$

hence

$$\xi(0) = \arctan\left(\frac{B(0)/2}{2h(0) - L(0)^2}\right).$$

Otherwise, in the case $h(0) = \frac{1}{2}$, the value of $\xi(0)$ is not rigourously defined. However, as previously, the exact value of $\xi(0)$ is not important because $h(t) = h(0) = \frac{1}{2}$, which means that $L(t)^2 = 1$ and $B(t) = 0$. Therefore, in the case $h(0) = \frac{1}{2}$, the mapping $t \mapsto (L(t), B(t))$ does not depend on the value of $\xi(0)$. Finally, since the mapping $t \mapsto L(t)$ does not depend on $\xi(0)$ in the case $h(0) = \frac{1}{2}$, we also have that $t \mapsto A(t)$ does not depend on the exact value of $\xi(0)$, thanks to the expression of $A(t) = A(0)\left(L(t)/L(0)\right)^{\frac{2-d}{2}}$.

Finally, it remains to show that if $a_i(0) = 0$, $i \in \{1, \dots, d\}$ or $h(0) = \frac{1}{2}$, then the behavior of the mappings $t \mapsto \gamma(t)$ and $t \mapsto s(t)$ do not depend on the exact value of

$\theta_i(0)$, $i \in \{1, \dots, d\}$ or $\xi(0)$. The exact formulae for $\gamma(t)$ and $s(t)$ are:

$$\begin{aligned}
\gamma(t) &= \gamma(0) + \sum_{l=1}^{d} \frac{a_l(0)}{2} \left[\sin(2\theta_l(t)) - \sin(2\theta_l(0))\right] \\
s(t) &= \frac{1}{2} \arctan\left(\left(2h(0) + \sqrt{4h(0)^2 - 1}\right) \tan\left(\frac{\xi(0)}{2} - 2t\right)\right) \\
&\qquad - \frac{1}{2} \arctan\left(\left(2h(0) + \sqrt{4h(0)^2 - 1}\right) \tan\left(\frac{\xi(0)}{2}\right)\right) - m_t \frac{\pi}{2}.
\end{aligned}$$

It is clear that if $a_i(0) = 0$ then $\gamma(t)$ does not depend on $\theta_i(0)$ nor $\theta_i(t)$, $i \in \{1, \dots, d\}$. If $h(0) = \frac{1}{2}$, then

$$2h(0) + \sqrt{4h(0)^2 - 1} = 1,$$

so that

$$\begin{aligned}
&\frac{1}{2} \arctan\left(\left(2h(0) + \sqrt{4h(0)^2 - 1}\right) \tan\left(\frac{\xi(0)}{2} - 2t\right)\right) \\
&\qquad - \frac{1}{2} \arctan\left(\left(2h(0) + \sqrt{4h(0)^2 - 1}\right) \tan\left(\frac{\xi(0)}{2}\right)\right) - m_t \frac{\pi}{2} \\
&= \frac{1}{2} \arctan\left(\tan\left(\frac{\xi(t)}{2}\right)\right) - \frac{1}{2} \arctan\left(\tan\left(\frac{\xi(0)}{2}\right)\right) - m_t \frac{\pi}{2}.
\end{aligned}$$

Since $\arctan : \mathbb{R} \mapsto (-\frac{\pi}{2}, \frac{\pi}{2}]$, we write $\widetilde{\frac{\xi(0)}{2}} := \frac{\xi(0)}{2} - m_0\pi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$, and $\widetilde{\frac{\xi(\tau)}{2}} := \frac{\xi(\tau)}{2} - (m_0 - m_t)\pi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$. Then we have

$$\begin{aligned}
&\frac{1}{2} \arctan\left(\tan\left(\frac{\xi(t)}{2}\right)\right) - \frac{1}{2} \arctan\left(\tan\left(\frac{\xi(0)}{2}\right)\right) - m_t \frac{\pi}{2} \\
&= \frac{1}{2} \widetilde{\frac{\xi(t)}{2}} - \frac{1}{2} \widetilde{\frac{\xi(0)}{2}} - m_t \frac{\pi}{2} \\
&= \frac{1}{2} \frac{\xi(t)}{2} - (m_0 - m_t)\frac{\pi}{2} - \frac{1}{2}\left(\frac{\xi(0)}{2} - m_0\pi\right) - m_t \frac{\pi}{2} \\
&= \frac{1}{2}\left(\frac{\xi(t)}{2} - \frac{\xi(0)}{2}\right) = -t.
\end{aligned}$$

This shows that, in the case $h(0) = \frac{1}{2}$, the mapping $t \mapsto \gamma(t)$ does not depend on the value chosen for the ill-defined quantity $\xi(0)$. □

Note that, in order to define $\xi(0)$, we could also use the expression (IV-3.25) obtained from the proof of Lemma IV.3:

$$\xi = \arcsin\left(\frac{e^{4k} - \frac{\mathcal{E}}{2}}{\sqrt{\frac{\mathcal{E}^2}{4} - 1}}\right) + \frac{\pi}{2} \in [0, \pi],$$

However this is not appropriate from a computational point of view, some more details about this are given in Remark IV.4.

Using Lemmata IV.3 and IV.4, we are now able to obtain a straightforward numerical algorithm which simulates exactly the evolution of bubbles according to the harmonic oscillator on a time interval $[0, T]$. It is described in Algorithm 2.

Algorithm 2 Solving the harmonic oscillator with Bubbles

```
Input:
— The bubble discretization of (IV-3.1), which gives the functions $u^j$.
— For each function $u^j$, its modulation parameters $(A^j, B^j, L^j, \beta^j, X^j, \gamma^j)$.
for $j = 1, \dots, N$ do                                      ▷ $j$ denotes a bubble's index
    Use (IV-3.26) to get the action-angle variables $(h, a, \xi, \theta)$ at time 0.
    Use (IV-3.20) to update the variables $(h, a, \xi, \theta)$ up to time $T$.
    Use (IV-3.21) to get the parameters of bubble $u^j$ at time $T$.
end for
Use (IV-3.15) to update the Hermite decomposition for each bubble.
Output:
— The bubbles $u^j(T, \cdot)$ given by (IV-3.2), solution to (HO).
```

Remark IV.4: Numerical considerations

Here are a few remarks about Algorithm 2:

- When applying equation (IV-3.26) to obtain the action-angle variables from the bubbles' parameters, it is advised to use the function arctan2 (y, x) instead of $\arctan(y/x)$ because it allows to obtain an angle lying in $(-\pi, \pi]$ instead of $(-\pi/2, \pi/2]$, by taking into account the signs of both x and y. The arctan2 (y, x) function is defined as follows:

$$\text{arctan2 }(y, x) := \begin{cases} \arctan(y/x) & \text{if } x > 0 \\ \arctan(y/x) + \pi & \text{if } x < 0 \text{ and } y \geq 0 \\ \arctan(y/x) - \pi & \text{if } x < 0 \text{ and } y < 0 \\ \dfrac{\pi}{2} & \text{if } x = 0 \text{ and } y > 0 \\ -\dfrac{\pi}{2} & \text{if } x = 0 \text{ and } y < 0 \\ \text{undefined} & \text{if } x = 0 \text{ and } y = 0. \end{cases}$$

  This is also the reason why we do not define $\xi(0)$ by (IV-3.25). Moreover most numerical implementations of arctan2 return a finite value for arctan2 $(0, 0)$, which avoids the manual tuning of a numerical threshold to know whether $a_i(0)$

or $h(0)$ vanish numerically or not. We recall that in this case the exact value returned does not impact the behavior of $t \mapsto (L(t), B(t), X(t), \beta(t), \gamma(t), s(t))$.

- The family $\left\{\varphi_n = H_{n_1} \cdots H_{n_d} : n = (n_1, \dots, n_d) \in \mathbb{N}^d\right\}$ is an orthonormal family of $\mathbb{L}^2(\mathbb{R}^d)$, hence the discretization of any initial condition is done by calculating (or approximating) the Hermite coefficients of the functions v^j in the decomposition (IV-3.14).
- The algorithm yields an *exact* solution as soon as the initial data is a sum of bubbles. If not, then the only error committed is the discretization error when approximating the initial condition $\psi(t = 0)$ by the ansatz (IV-3.1).
- This numerical algorithm does not need any discretization in time nor in space, since the integration in time is performed exactly and the function in space is known analytically.
- The solution obtained is the exact solution of (HO) (after the discretization of $\psi(0, \cdot)$ into a sum of bubbles), defined on the whole space $\mathbb{R}^d$, and no numerical boundary conditions are needed.
- If $M \geq 1$ Hermite modes are used in each dimension, then the computational complexity is $\mathcal{O}(N(M^d + d))$.

IV-3.2 Numerical examples

In this section we will assess the efficiency of the bubble approach described by Algorithm 2 against a spectral method, in the linear two-dimensional case.

IV-3.2.1 Grid-based spectral scheme of reference

We start by discussing the spectral method we shall use to compare with the results of Algorithm 2. We refer to [34] for a general introduction to spectral methods for the Schrödinger equation, and to [5] for grid-based schemes applied to the Gross-Pitaevskii equation. See also Section IV-2 – Review of the Schrödinger equation.

We now present a method which can be understood as the application of [14] to a simpler equation, namely the harmonic oscillator. We use a splitting method to simulate the linear part (HO), and thanks to [13, 3] we have:

$$\begin{aligned} e^{-it(-\Delta+|x|^2)} &= e^{-\frac{1}{2}\tanh(it)|x|^2} e^{\frac{1}{2}\sinh(2it)\Delta_x} e^{-\frac{1}{2}\tanh(it)|x|^2} \\ &= e^{-\frac{i}{2}\tan(t)|x|^2} e^{\frac{i}{2}\sin(2t)\Delta_x} e^{-\frac{i}{2}\tan(t)|x|^2}. \end{aligned} \tag{IV-3.27}$$

We can cite [45] which also presents a spectral method based on the Fourier transform with time splitting, however the above method is different in that (IV-3.27) is exact and hence we do not have any time-splitting error.

The first and third exponentials on the RHS are straightforward to compute on a grid. For the second one, we use a Fourier transform: $e^{\frac{i}{2}\sin(2t)\Delta_x}$ is the propagator of the following equation:

$$\partial_t \psi = i\cos(2t)\Delta_x \psi.$$

By using a Fourier transform, we get

$$\partial_t \mathscr{F}(\psi)(\xi) = i\cos(2t)\mathscr{F}\left(\Delta_x\psi\right)(\xi) = -i\cos(2t)|\xi|^2\mathscr{F}\left(\psi\right)(\xi).$$

Hence,

$$\mathscr{F}(\psi(t,\cdot))(\xi) = e^{-\frac{i}{2}\sin(2t)|\xi|^2}\mathscr{F}(\psi(0,\cdot))(\xi).$$

The RHS exponential is straightforward to compute in the Fourier space. Hence, an exact-time spectral approximation of the solution to (HO) is given by Algorithm 3.

Algorithm 3 Spectral solver for (HO), with an exact time resolution for each splitting step.

```
Input:
— GRID: a uniform discretization of a finite volume subset of $\mathbb{R}^d$
— the initial data $\psi_0$.
Discretize the initial data $\psi_0$ on GRID $\subset \mathbb{R}^d$, and let $\eta$ be this discretization.
for Each timestep of size $\Delta t$ do
    for $x \in$ GRID do                                      ▷ $x \in \mathbb{R}^d$.
        Multiply $\eta(x)$ by $e^{-\frac{i}{2}\tan(\Delta t)|x|^2}$.
    end for
    Apply a FFT to $\eta$.                     ▷ FFT: Fast Fourier Transform.
    for $\xi \in$ FOURIER GRID do                            ▷ $\xi \in \mathbb{R}^d$.
        Multiply $\hat{\eta}(\xi)$ by $e^{-\frac{i}{2}\sin(2\Delta t)|\xi|^2}$.
    end for
    Apply an inverse FFT to $\hat{\eta}$.
    for $x \in$ GRID do
        Multiply $\eta(x)$ by $e^{-\frac{i}{2}\tan(\Delta t)|x|^2}$.
    end for
end for
Output:
— $\eta$ is an approximation on GRID of $\psi(T,\cdot)$, where $\psi$ is the solution to (HO).
```

Of course, in pratical applications one is not able to define a grid over $\mathbb{R}^d$. Hence, Algorithm 3 has to be modified by defining GRID as a discretization of a finite-volume subset of $\mathbb{R}^d$, typically a product of intervals in each dimension. For all of our numerical examples, this will $[-15,15]\times[-15,15]$, discretized using $N_x \times N_y$ points. In order to have an easily computable FFT, one has to use a spatial uniform grid, which then defines the Fourier Grid. Special care has to be paid when choosing the number of points: if there are Fourier frequencies in the solution ψ larger than the *Nyquist frequency*, then we will observe a phenomenon known as *aliasing*. This may not be problematic for the harmonic oscillator (HO) depending on the initial condition, but will eventually become an issue when simulating (cNLS) in the next Chapter, because it involves interactions and hence an infinite number of frequencies. Moreover, by using an FFT-based algorithm we implicitly impose periodic boundary conditions.

IV-3.2.2 Discretization into a sum of Bubbles

We need to decompose any arbitrary function into a finite sum of N bubbles. A solution to this question has been proposed in [66], but it involves integrals over the whole phase space, which is something we want to avoid.

We could also use a nonlinear least squares approach, but our experimental results showed that it tends to yield spread out Gaussians, which may present huge overlaps between them. We will observe in the next Chapter that huge overlaps do not mix well with our nonlinear solution. The issue of discretizing an arbitrary function into a sum of bubbles without too much overlap is not the main concern of this work, hence we will use a visual trial-and-error discretization.

IV-3.2.3 Observables

In order to compare the bubble scheme against the spectral method, we compare them both in the absence of interactions, i.e. on the harmonic oscillator (HO). We showed in Lemma IV.2 the conservation of some quantities for (HO), we will focus on mass, energy and momentum. When computing the observables for the spectral solution, we noted that the approximation of the gradient using finite differences with periodic boundary conditions yielded very rough results while the gradient approximation using the Fast Fourier Transform gave more accurate results. We use the latter approximation in the Figures of Section IV-4.3.2. When reporting the results in the following log-plots, all values with an amplitude smaller than 10^{-16} were set to be equal to 10^{-16}.

For all of the results shown, the spectral scheme is supplied with the exact initial condition and not a projection on the grid of the bubbles discretization. We display the relative evolution of each observable quantity, that is the evolution of each quantity relative to the value of the quantity at time 0. In other words, for an observable quantity

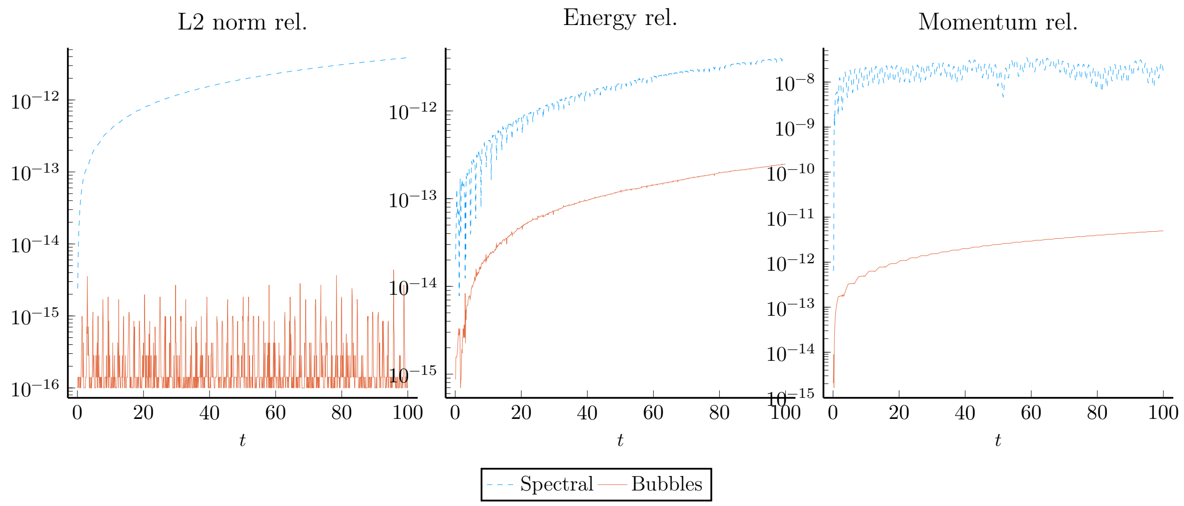


Figure IV-3.1 – Test case 1. Relative evolution of mass, energy and momentum with bubbles and spectral methods. $\Delta t = 10^{-2}$. Spectral scheme with $N_x = 256, N_y = 256$.

Q, computed at each time $\{t^n\}_{n\geq 0}$, its relative evolution is given by:

$$(rel_Q)(t^n) = \left|\frac{Q(t^n) - Q(t^0)}{Q(t^0)}\right|$$

Since $Q(t^0)$ is computed **after** the discretization of the initial condition, this means that the discretization error is not reported in the Figures of Section IV-4.3.2.

IV-3.2.4 Numerical experiments

Test case 1: Weak interactions

The initial condition reads

$$\psi(t=0,x) = e^{-|x-\mu_3|^2} e^{i\cosh|x-\mu_3|}, \quad x \in \mathbb{R}^2, \quad \mu_3 = (1,1). \tag{IV-3.28}$$

The approximation of this function as a sum of bubbles is pretty straightforward: we know that for x small, $\cosh x \approx 1 + \frac{x^2}{2}$, hence

$$\psi(t=0,x) \approx e^{-|x-\mu_3|^2} e^{i + i\frac{|x-\mu_3|^2}{2}}, \quad x \in \mathbb{R}^2.$$

The results are displayed in Figure IV-3.1. The solution obtained with the bubble scheme is at least one order of magnitude better than the spectral scheme.

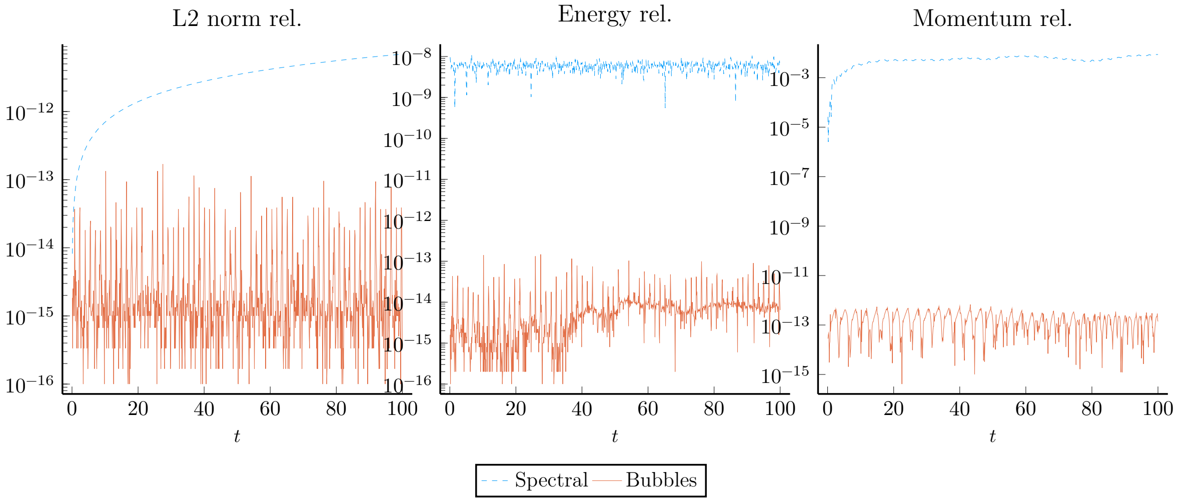


Figure IV-3.2 – Test case 2. Relative evolution of mass, energy and momentum with bubbles and spectral methods. $\Delta t = 10^{-2}$. Time-integrator for the nonlinear part of the splitting: Runge-Kutta of order 2. Spectral scheme with $N_x = 256, N_y = 256$.

Test case 2: Rotating Gaussians

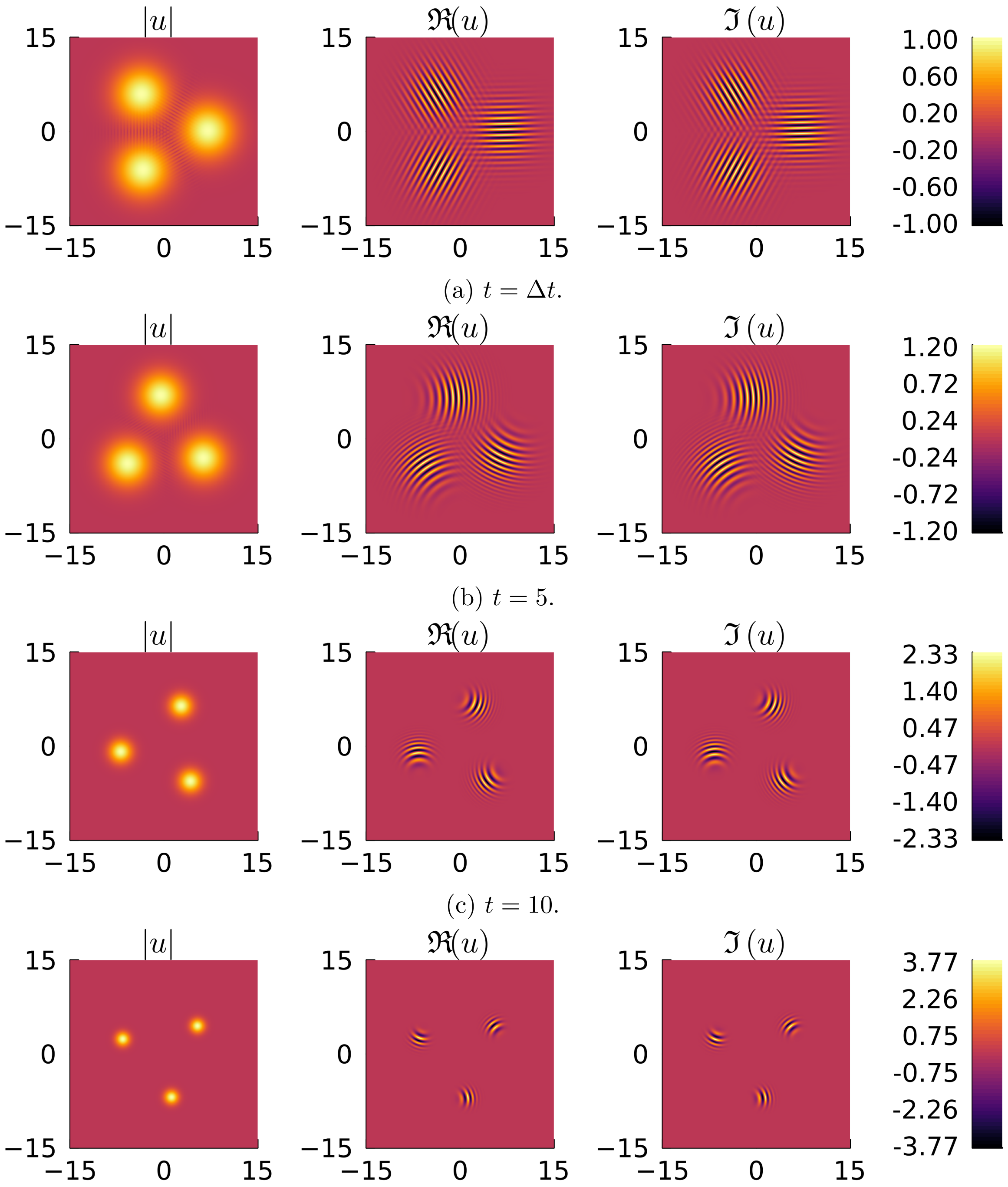
This test case is an illustration of the good conservation properties of the modulation algorithm, including the nonlinearities, as soon as the bubbles don't have too much overlap. The initial condition reads:

$$\psi(t=0,x) = \sum_{i=1}^{3} e^{\gamma^j + i\beta^j\cdot(x-X^j) - \frac{|x-X^j|^2}{2(L^j)^2}},$$

where

$$\begin{aligned}
&L^1 = 3 && \gamma^1 = 5 && X^1 = 7(1,0) && \beta^1 = (X^1)^\perp,\\
&L^2 = 3 && \gamma^2 = -5 && X^2 = 7\left(-\frac{\sqrt{3}}{2}, \frac{1}{2}\right) && \beta^2 = (X^2)^\perp,\\
&L^3 = 3 && \gamma^3 = 0 && X^3 = 7\left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) && \beta^3 = (X^3)^\perp.
\end{aligned}$$

The numerical results are given in Figure IV-3.2. The spectral scheme (blue dash line) is outperformed by the bubbles scheme (orange solid line) for all times t. The time evolution of the discretized solution $u = \sum_{j=1}^{3} u^j$ is given in Figure IV-3.3 at times $t \in \{\Delta t, 5, 10, 15\}$. Note that, in this figure, the bounding box $[-15, 15] \times [-15, 15]$ is only here for plotting purposes, and the solution u is known analytically on the whole $\mathbb{R}^2$ plane.



(a) $t = \Delta t$.

(b) $t = 5$.

(c) $t = 10$.

(d) $t = 15$.

Figure IV-3.3 – Solution $u(t, x)$ for $x \in \mathbb{R}^2$, at different times t. For each panel corresponding to a time t, we display the module of u (left), the real part of u (middle), and the imaginary part of u (right).

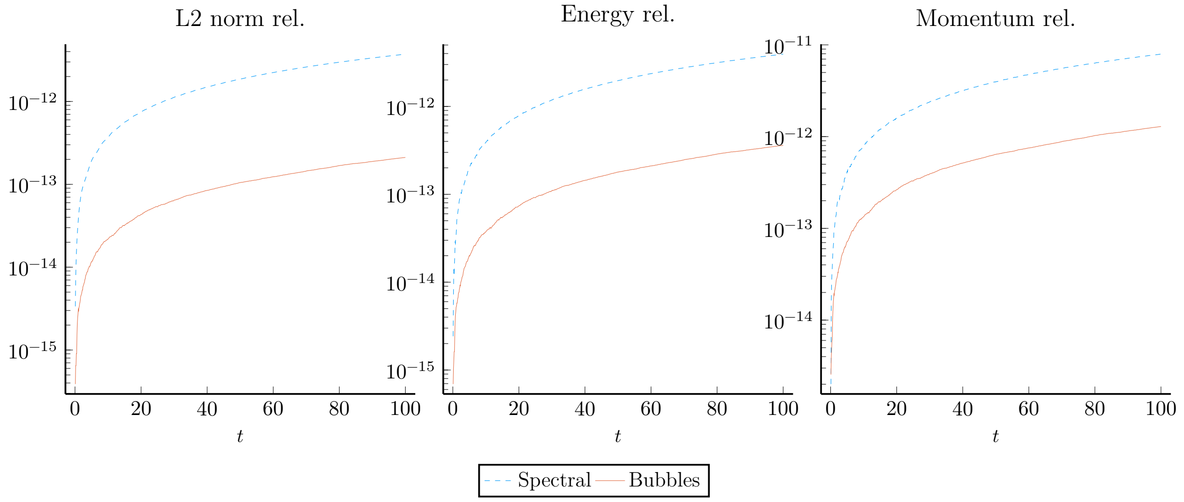


Figure IV-3.4 – Test case 3. Relative evolution of mass, energy and momentum with bubbles and spectral methods. $\Delta t = 10^{-2}$. Time-integrator for the nonlinear part of the splitting: Runge-Kutta of order 2. Spectral scheme with $N_x = 256, N_y = 256$.

Test case 3: Zero phase initial data

The initial condition reads

$$\psi(t=0,x) = \pi e^{-\frac{|x-\mu_1|^2}{2}} + 2e^{-\frac{|x-\mu_2|^2}{2}}, \quad x \in \mathbb{R}^2, \quad \mu_1 = (0,2),\ \mu_2 = (1,0). \qquad \text{(IV-3.29)}$$

The results are displayed in Figure IV-3.4. The bubble solution (orange solid line) outperforms the spectral method (blue dash line) on the harmonic oscillator, and the solution obtained with the Bubbles scheme is about one order of magnitude better than the spectral scheme.

Part IV

CHAPTER 4 Nonlinear Schrödinger equation

The case of the linear Schrödinger equation has just been treated in the previous Chapter, and we now want to spice things up by considering the nonlinear Schrödinger equation. It is straightforward to see that the computation $|u(t)|^2 u(t)$ involves $\mathcal{O}(N^3)$ bubbles. While it is technically possible to use N^3 bubbles in the linear setting, it means that the nonlinearity will then involve $\mathcal{O}(N^9)$ bubbles...In just a few iterations of the method, the number of bubbles becomes unmanageable!

Therefore, some approximate solution is seeked, and it *cannot* be better than an approximation. We will see in this Chapter how the Dirac-Frenkel principle – which has always been applied in other works to the linear setting – can be used in the nonlinear setting. To the author's knowledge, it is the first account of the Dirac-Frenkel method being applied with a nonlinearity.

We start this Chapter by presenting quickly the Dirac-Frenkel principle, and then how it can be reformulated into a matrix problem. We note that, using a simplifying assumption that we think is not totally restrictive, the coefficients of the linear system can be computed analytically. Therefore, there is no need for a grid to evaluate the Dirac-Frenkel $\mathbb{L}^2$ inner products, and the presented method is completely gridless. We end this Chapter with some numerical experiments that showcase the advantages and limitations of the method. Among the limitations, we can find the computational complexity which depends polynomially on the number of bubbles, and the invertibility issues of the matrix reformulation of the Dirac-Frenkel principle. While the first issue really appears when polynomial nonlinearities are considered, the second issue is inherent to the Dirac-Frenkel principle and we note that it has already been reported in previous works for the linear setting.

All the content of this Chapter is also part of the joint, unpublished, work with Erwan Faou and Pierre Raphaël [30].

IV-4.1 Motivation

We are interested in approximating numerically the solution $\psi(t,x)$ to the cubic nonlinear Schrödinger equation with harmonic potential,

$$i\partial_t\psi + \Delta_x\psi - |x|^2\psi = \psi|\psi|^2, \quad x \in \mathbb{R}^d, \tag{cNLS}$$

where $d \geq 1$, $|\cdot|$ denotes the usual Euclidian norm over $\mathbb{R}^d$, and Δ_x denotes the Laplace operator over $\mathbb{R}^d$: $\Delta_x = \sum_{i=1}^d \partial_{x_i}^2$. This equation is also sometimes called *time-dependent Gross-Pitaevskii equation* [21, 8, 83, 81]. We focus on a cubic nonlinearity for the sake of clarity, but we emphasize the fact that everything we present is also applicable to other types of polynomial nonlinearities, *mutatis mudandis*. Similarly, the extension to the equation (IV-4.1) – which is (cNLS) without the harmonic potential – is also straightforward.

Let us now explain the main ideas underlying the full modulation (IV-3.2) – developed in various works, see for instance [57, 31] and the references therein – and why it is particularly adapted to the nonlinear case.

Consider for instance the case of one bubble, *i.e.* $N = 1$. When plugging the ansatz (IV-3.2) into (cNLS), we obtain an equation of the form

$$i\partial_s v + \Delta_y v - |y|^2 v - |v|^2 v + P(s; y, \partial_y)v = 0,$$

where $P(s; y, \partial_y)$ is a quadratic operator in y and ∂_y, which depends on time s through the parameters $(A, L, B, X, \beta, \gamma)$ and their time derivatives with respect to s. See (IV-3.11) for more precise detail. It is then possible to choose the parameters in such a way that for instance $P(s; y, \partial_y)v = -\lambda v$ for some $\lambda \in \mathbb{R}$, and to take v as a soliton solution of the stationary equation

$$-\Delta_y v + |y|^2 v + |v|^2 v = \lambda v.$$

This yields a differential system to be solved by the parameters $(A, L, B, X, \beta, \gamma)$ which is given below by (IV-3.16). It turns out that these equations form a *completely integrable Poisson system* that can be solved, and the solution for a single bubble can be thus taken as a *modulated soliton*.

In other words, taking $v_j = v$ when $N = 1$, a solution of the nonlinear equation (IV-3.3) yields an exact solution $u_j = u$ under the form (IV-3.2) of (cNLS).

This kind of approach has been used successfully in various situations from a theoretical point of view, see [60, 57, 31, 61] and the references therein. Typically, when $N \geq 2$, several modulated solitons interact and this can produce finite time blow-up of growth of Sobolev norm phenomena. A large part of the analysis relies on the ability of calculating nonlinear interactions between two modulated solitons. This can be done for instance in an integrable situation, e.g. the Szegö equation [36].

Another byproduct of these modulation techniques in 2D is to make a link between (cNLS) on a finite time interval and the Schrödinger equation without harmonic potential

$$i\partial_s\psi + \Delta_x\psi = \psi|\psi|^2, \quad x \in \mathbb{R}^d \tag{IV-4.1}$$

on an unbounded time interval. In this case, the modulation equations generate the so-called *lens* transform, see for instance [18]. Note that our algorithms could be also be applied to the latter equation but we will restrict our analysis to the Harmonic case. Let us note as well that such modulation techniques can also be related with the families of exact splitting introduced in [13], where the time coefficients can be seen as specific time changes s in the modulation equations.

To approximate the solution to the nonlinear part (NL), we use the Dirac-Frenkel-MacLachlan principle. In theory, when the v_j are finite sums of Hermite polynomials, the calculation of the interactions boils down to the computation of integrals of products of Hermite functions in different modulation frames, which *a priori* can be done in a systematic way. In practice, these computations can get heavy and to simplify them we will give the explicit result of the Gaussian case.

IV-4.2 The Dirac-Frenkel principle

In this section we consider the Schrödinger equation (cNLS). As it has been explained before, the equation consists in two parts: the linear part (HO), and the nonlinear part (NL). Chapter IV-3 was dedicated to solving the harmonic oscillator. We are interested now in solving the nonlinear part, and as it is usually done for numerical simulations, we will use a splitting method (see for instance [59, 40, 19]). This will allow us to solve (cNLS) by solving separately (HO) and (NL), one after the other. By doing so, a splitting error is made, which depends on the timestep Δt, and the order of the error depends on the specific splitting method. It is also possible to use high-order splitting methods.

We focus on approximating numerically the solution to (NL):

$$i\partial_t\psi = \psi|\psi|^2.$$

We are free to use any method we want, but one has to keep in mind that Algorithm 2 solves (HO) exactly when ψ is expressed under the form (IV-3.1), i.e. as a sum of bubbles. Therefore we would like the approximate solution to (NL) to keep this particular form.

This naturally calls for the use of the Dirac-Frenkel principle (abbreviated DF principle). For more details, see [51, Sect. 3].

In theory, the computations can be performed in a very general situation, when all the v^j involved in (IV-3.1)-(IV-3.2) are given in terms of Hermite polynomials. In essence, the only difficulty lies in the evaluation of general integrals of products of Hermite functions in different modulation frames, which can be done using generating functions techniques for instance. In order to simplify the presentation, we will only consider the first Hermite mode, a.k.a the Gaussian function. From now on, we consider

$$v(s,y) = e^{-|y|^2/2}. \tag{IV-4.2}$$

Remark IV.5

Note that the functions $e^{-\frac{|y|^2}{2}}$ are simply

$$\varphi_{(0,\ldots,0)}(y) = H_0(y_1)\cdots H_0(y_d),$$

hence we can use Section IV-3.1.3 for the linear part.

Another alternative would be to use nonlinear solitons and rely on numerical evaluations of the corresponding integrals, but we will not use this approach in this work.

In the remainder of this Chapter, we consider $\mathcal{M}$ a manifold of the sum of N complex-valued Gaussian functions:

$$\mathcal{M} := \left\{ u \in \mathbb{L}^2(\mathbb{R}^d) \,\middle|\, \begin{array}{c} u(x) = \displaystyle\sum_{j=1}^{N} \frac{A^j}{L^j} e^{i\gamma^j + i\beta^j\cdot(x-X^j) - \frac{2+iB^j}{4(L^j)^2}|x-X^j|^2}, \\ A^j, B^j, \gamma^j \in \mathbb{R},\ L^j \in \mathbb{R}_+^*,\ X^j, \beta^j \in \mathbb{R}^d \end{array} \right\}. \tag{IV-4.3}$$

We look for a function $u \in \mathcal{M}$ that solves (NL) on $\mathcal{M}$. More precisely, u is defined such that its time derivative lies in the tangent space of $\mathcal{M}$ at u, denoted $\mathcal{T}_{u(t)}\mathcal{M}$, and such that the residual of equation (NL) is orthogonal to the tangent space. That is,

$$\begin{aligned} &\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M}, \quad \text{such that} \\ &\qquad \langle f, i\partial_t u(t) - u(t)|u(t)|^2\rangle = 0,\ \forall f \in \mathcal{T}_{u(t)}\mathcal{M}. \end{aligned} \tag{IV-4.4}$$

Let $B_{u(t)}$ be a basis of $\mathcal{T}_{u(t)}\mathcal{M}$, then (IV-4.4) is equivalent to

$$\begin{aligned} &\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M}, \quad \text{such that} \\ &\qquad \langle f, i\partial_t u(t)\rangle = \langle f, u(t)|u(t)|^2\rangle,\ \forall f \in B_{u(t)}. \end{aligned} \tag{IV-4.5}$$

A family (which may happen to be linearly dependent!) spanning the tangent space $\mathcal{T}_{u(t)}\mathcal{M}$ is given by

$$\begin{aligned} B_{u(t)} = \Big\{ & e^{i\Gamma^j(y^j)-\frac{|y|^2}{2}}, (y_1^j)e^{i\Gamma^j(y^j)-\frac{|y^j|^2}{2}}, \dots, (y_d^j)e^{i\Gamma^j(y^j)-\frac{|y^j|^2}{2}}, \\ & |y^j|^2 e^{i\Gamma^j(y^j)-\frac{|y^j|^2}{2}} : j = 1, \dots, N \Big\}, \\ =: & \left\{ b_{j,1}, b_{j,2}, \dots, b_{j,d+1}, b_{j,d+2} : j = 1, \dots, N \right\}, \end{aligned} \tag{IV-4.6}$$

where we defined

$$\Gamma^j(y^j) := \gamma^j + L^j\beta^j \cdot y^j - \frac{B^j}{4}|y^j|^2.$$

We recall that the functions of $B_{u(t)}$ are obtained by differentiating $u(t) \in \mathcal{M}$ with respect to each parameter. Thus, (IV-4.5) is equivalent to

$$\begin{aligned} & \partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M}, \quad \text{such that} \\ & \quad \langle i\partial_t u(t), b_{j,l}\rangle = \langle u|u|^2, b_{j,l}\rangle, \quad j = 1, \dots, N, \quad l = 1, \dots, d+2. \end{aligned} \tag{IV-4.7}$$

The next step consists in expressing (IV-4.7) as a linear system involving the parameters of the bubbles and their time derivative. We then solve the linear system, which yields ODEs on the parameters that we can integrate numerically. The main advantage of this approach is that it guarantees to keep the approximate solution of (NL) as a sum of N bubbles. There are however some issues in practice with the application of the Dirac-Frenkel principle, we will discuss them in more details in Section IV-4.3.

In order to obtain the linear system, we first have to get the expression of $i\partial_t u(t)$ when $v(y) = e^{-\frac{|y|^2}{2}}$: by summing (IV-3.10) over $j = 1, \dots, N$, and using that $\partial_s v^j = 0$ (since v^j is defined by (IV-4.2)), one has

$$\begin{aligned} i\partial_t u = \sum_{j=1}^{N} \frac{u^j}{(L^j)^2} \Bigg\{ & |y^j|^2 \left(i\frac{(L^j)_s}{L^j} - \frac{B^j(L^j)_s}{2L^j} + \frac{(B^j)_s}{4} \right) \\ & + y^j \cdot \left(-L^j(\beta^j)_s + i\frac{(X^j)_s}{L^j} - \frac{B^j}{2L^j}(X^j)_s \right) \\ & + i\frac{(A^j)_s}{A^j} - i\frac{(L^j)_s}{L^j} + \beta \cdot (X^j)_s - (\gamma^j)_s \Bigg\}. \end{aligned} \tag{IV-4.8}$$

We recall that the subscript $_s$ denotes the time-differentiation with respect to time s.

More concisely, we have

$$\begin{aligned} i\partial_t u = & \sum_{j=1}^{N} \frac{A^j}{(L_j)^3} e^{i\Gamma^j - \frac{|y^j|^2}{2}} \left\{ |y^j|^2 \left(E^{j,(5)} + iE^{j,(6)} \right) \right. \\ & \qquad + y^j \cdot \left(E^{j,(3)} + iE^{j,(4)} \right) \\ & \qquad \left. + \left(E^{j,(1)} + iE^{j,(2)} \right) \right\} \\ = & \sum_{j=1}^{N} \frac{A^j}{(L_j)^3} \left\{ b_{j,1} \left(E^{j,(1)} + iE^{j,(2)} \right) + b_{j,2} \left(E_1^{j,(3)} + iE_1^{j,(4)} \right) \right. \\ & \qquad \left. \cdots + b_{j,d+1} \left(E_d^{j,(3)} + iE_d^{j,(4)} \right) + b_{j,d+2} \left(E^{j,(5)} + iE^{j,(6)} \right) \right\}, \end{aligned} \tag{IV-4.9}$$

where

$$\begin{aligned} E^{j,(1)} &:= \beta^j \cdot X_s^j - \gamma_s^j, & E^{j,(2)} &:= \frac{A_s^j}{A^j} - \frac{L_s^j}{L^j}, \\ E_l^{j,(3)} &:= -L^j \beta_{l,s}^j - \frac{B^j}{2L^j} X_{l,s}^j, & E_l^{j,(4)} &:= \frac{X_{l,s}^j}{L^j}, \qquad l = 1, \dots, d, \\ E^{j,(5)} &:= \frac{B_s^j}{4} - \frac{B^j L_s^j}{2L^j}, & E^{j,(6)} &:= \frac{L_s^j}{L^j}, \end{aligned} \tag{IV-4.10}$$

Following our notation convention (given in the paragraph before Section IV-3.1), we recall that a subscript $_t$ or $_s$ *always* denotes a time derivative (either with respect to time t or s), the exponent j denotes the bubble's label, and the subscript l denotes the l-th component of a vector.

According to (IV-4.7), we then want to project $i\partial_t u(t)$ against every element of $B_{u(t)}$. We obtain the following linear system:

$$\mathbf{AE} = \mathbf{S}, \tag{IV-4.11}$$

where

$$\mathbf{A} := \begin{pmatrix} \langle b_{1,1}, b_{1,1} \rangle & \langle b_{1,2}, b_{1,1} \rangle & \dots & \langle b_{N,d+1}, b_{1,1} \rangle & \langle b_{N,d+2}, b_{1,1} \rangle \\ \vdots & & & & \vdots \\ \langle b_{1,1}, b_{N,d+2} \rangle & \langle b_{1,2}, b_{N,d+2} \rangle & \dots & \langle b_{N,d+1}, b_{N,d+2} \rangle & \langle b_{N,d+2}, b_{N,d+2} \rangle \end{pmatrix},$$

$$\mathbf{E} := \begin{pmatrix} \frac{A^1}{(L^1)^3}\left(E^{1,(1)} + iE^{1,(2)}\right) \\ \frac{A^1}{(L^1)^3}\left(E_1^{1,(3)} + iE_1^{1,(4)}\right) \\ \vdots \\ \frac{A^1}{(L^1)^3}\left(E_d^{1,(3)} + iE_d^{1,(4)}\right) \\ \frac{A^1}{(L^1)^3}\left(E^{1,(5)} + iE^{1,(6)}\right) \\ \vdots \\ \frac{A^j}{(L^j)^3}\left(E^{j,(1)} + iE^{j,(2)}\right) \\ \frac{A^j}{(L^j)^3}\left(E_1^{j,(3)} + iE_1^{j,(4)}\right) \\ \vdots \\ \frac{A^j}{(L^j)^3}\left(E_d^{j,(3)} + iE_d^{j,(4)}\right) \\ \frac{A^j}{(L^j)^3}\left(E^{j,(5)} + iE^{j,(6)}\right) \\ \vdots \\ \frac{A^N}{(L^N)^3}\left(E_d^{N,(3)} + iE_d^{N,(4)}\right) \\ \frac{A^N}{(L^N)^3}\left(E^{N,(5)} + iE^{N,(6)}\right) \end{pmatrix} \qquad \text{and} \qquad \mathbf{S} := \begin{pmatrix} \langle u|u|^2, b_{1,1}\rangle \\ \vdots \\ \langle u|u|^2, b_{N,d+2}\rangle \end{pmatrix}.$$

The matrix $\mathbf{A} \in \mathbb{C}^{(d+2)N,(d+2)N}$ is the Gram matrix of the family $B_{u(t)}$, which obviously depends on time. We have $\mathbf{E} \in \mathbb{C}^{(d+2)N}$ and $\mathbf{S} \in \mathbb{C}^{(d+2)N}$. In order to solve the linear system (IV-4.11) we shall use the Moore-Penrose pseudoinverse which always exists, and which corresponds to the Least Squares solution if the matrix $\mathbf{A}^*\mathbf{A}$ is invertible. The matrix $\mathbf{A}$ is invertible if and only if $B_{u(t)}$ is a linearly independent family of $\mathbb{L}^2(\mathbb{R}^d)$. We can already notice that if two bubbles have the same parameters then the family will be linearly dependent: this is why the Moore-Penrose pseudoinverse is used, instead of $\mathbf{A}^{-1}$ which is not always well-defined. A variety of numerical techniques exist for solving an ill-conditioned or singular linear system, see for instance [15, 37, 27, 75]. We have chosen the Moore-Penrose pseudoinverse, because as G. Strang wrote in [75]: “When A^{-1} fails to exist, the best substitute is the pseudoinverse”. Other numerical techniques aim to approximate efficiently the pseudoinverse solution, but the computation itself has not been a computational burden during our numerical experiments so we haven't looked into more advanced techniques. As we already mentioned, the Dirac-Frenkel suffers from some inherent issues, and the main problem is due to the non-invertibility of the matrix $\mathbf{A}$. Even when using the pseudoinverse these issues arise, as we will see in Section IV-4.3.

Once the linear system (IV-4.11) is solved, we obtain $\mathbf{E}$, from which we can update the modulation parameters. In order to solve numerically the linear system, we shall rewrite it under a more convenient form. Let $\mathbf{A}_{\text{Re}} := \text{Re}\,(\mathbf{A})$, $\mathbf{A}_{\text{Im}} := \text{Im}\,(\mathbf{A})$, $\mathbf{E}_{\text{Re}} := \text{Re}\,(\mathbf{E})$,

$\mathbf{E}_{\mathrm{Im}} := \mathrm{Im}(\mathbf{E})$, $\mathbf{S}_{\mathrm{Re}} := \mathrm{Re}(\mathbf{S})$, and $\mathbf{S}_{\mathrm{Im}} := \mathrm{Im}(\mathbf{S})$. Then, (IV-4.11) writes:

$$\begin{aligned}\mathbf{AE} = \mathbf{S} \iff & (\mathbf{A}_{\mathrm{Re}} + i\mathbf{A}_{\mathrm{Im}})(\mathbf{E}_{\mathrm{Re}} + i\mathbf{E}_{\mathrm{Im}}) = \mathbf{S}_{\mathrm{Re}} + i\mathbf{S}_{\mathrm{Im}} \\ \iff & \begin{cases}\mathbf{A}_{\mathrm{Re}}\mathbf{E}_{\mathrm{Re}} - \mathbf{A}_{\mathrm{Im}}\mathbf{E}_{\mathrm{Im}} = \mathbf{S}_{\mathrm{Re}} \\ \mathbf{A}_{\mathrm{Im}}\mathbf{E}_{\mathrm{Re}} + \mathbf{A}_{\mathrm{Re}}\mathbf{E}_{\mathrm{Im}} = \mathbf{S}_{\mathrm{Im}}\end{cases} \\ \iff & \begin{pmatrix}\mathbf{A}_{\mathrm{Re}} & -\mathbf{A}_{\mathrm{Im}} \\ \mathbf{A}_{\mathrm{Im}} & \mathbf{A}_{\mathrm{Re}}\end{pmatrix}\begin{pmatrix}\mathbf{E}_{\mathrm{Re}} \\ \mathbf{E}_{\mathrm{Im}}\end{pmatrix} = \begin{pmatrix}\mathbf{S}_{\mathrm{Re}} \\ \mathbf{S}_{\mathrm{Im}}\end{pmatrix}.\end{aligned} \tag{IV-4.12}$$

It is more convenient to solve (IV-4.12) than (IV-4.11), because we only have to deal with real matrices and vectors.

Remark IV.6

We first tried to solve (IV-4.11) using the Moore-Penrose pseudoinverse, however it yielded very poor and seemingly wrong results. After some investigation, we found out that the issue seemed to be the complex numbers involved, and that they do not mix well with the pseudoinverse. Note that it is purely a numerical issue, and most probably due to the specific language or library used (the language used is `Julia`). After separating the real and imaginary parts (i.e. solving the linear system (IV-4.12)), we observed much better results.

Once (IV-4.12) is solved, we get the vector $\mathbf{E} = \mathbf{E}_{\mathrm{Re}} + i\mathbf{E}_{\mathrm{Im}}$. Multiplying the components corresponding to bubble j by $(L^j)^3/A^j$, we get the quantities $(E^{j,(k)})_{k=1,\dots,6}$.

By rearranging (IV-4.10), we are able to obtain the approximate update of the modulation parameters with respect to time t:

$$\begin{cases} A_t^j = \dfrac{A^j}{(L^j)^2}\left(E^{j,(2)} + E^{j,(6)}\right), \\ L_t^j = \dfrac{1}{L^j}E^{j,(6)}, \\ B_t^j = \dfrac{4}{(L^j)^2}E^{j,(5)} + \dfrac{2}{(L^j)^2}B^jE^{j,(6)}, \\ X_t^j = \dfrac{1}{L^j}E^{j,(4)}, \\ \beta_t^j = -\dfrac{1}{(L^j)^3}E^{j,(3)} - \dfrac{B^j}{2(L^j)^3}E^{j,(4)}, \\ \gamma_t^j = \dfrac{1}{L^j}\beta^j \cdot E^{j,(4)} - \dfrac{1}{(L^j)^2}E^{j,(1)}. \end{cases} \tag{IV-4.13}$$

Let us clarify now the procedure for obtaining the approximate update of the parameters: the DF principle tells us that we need to solve (IV-4.11). In Remark IV.6, we explained that it was better to solve (IV-4.12). Once (IV-4.12) is solved, we can get the quantities $(E^{j,(k)})_{j=1,\dots,N,k=1,\dots,6}$. Then, we can update the modulation parameters of all bubbles via the numerical integration of (IV-4.13).

IV-4.2.1 Computing coefficients of the linear system (IV-4.11)

In order to be able to compute $\mathbf{A}$ and $\mathbf{S}$, we give the exact expression of the inner products involved. We recall that these exact expressions have been obtained thanks to the assumption $v(s,y) = e^{-|y|^2/2}$. They involve the Fourier transform of some functions, let us first give the Fourier transform used. For $f \in \mathbb{L}^2(\mathbb{R}^d)$, the Fourier transform of f is denoted $\hat{f}$, with the following convention:

$$\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-i\xi\cdot x} dx.$$

For $j, l = 1, \dots, N$, let

$$\left|\begin{aligned} z_{j,l} &:= \frac{2+iB^l}{4(L^l)^2} + \frac{2-iB^j}{4(L^j)^2}, \\ a_{j,l} &:= \frac{X^l}{(L^l)^2} + \frac{X^j}{(L^j)^2}, \\ \xi_{j,l} &:= \frac{B^j}{2(L^j)^2}X^j + \beta^j - \frac{B^l}{2(L^l)^2}X^l - \beta_l, \\ C_{j,l} &= \exp\left\{ i(\gamma^l - \gamma^j) - \frac{2+iB^l}{4(L^l)^2}|X^l|^2 - \frac{2-iB^j}{4(L^j)^2}|X^j|^2 - i\beta^l\cdot X^l + i\beta^j \cdot X^j \right\}. \end{aligned}\right. \tag{IV-4.14}$$

Define

$$f_{j,l} : x \in \mathbb{R}^d \mapsto \exp(-z_{j,l}|x|^2 + a_{j,l}\cdot x) \in \mathbb{C},$$

then, for $n, m = 1, \dots, d$,

$$\begin{aligned} \langle b_{l,1}, b_{j,1}\rangle &= C_{j,l}\widehat{f_{j,l}}(\xi_{j,l}) \\ \langle b_{l,n+1}, b_{j,1}\rangle &= \frac{C_{j,l}}{L^l}\left(\widehat{xf_{j,l}}_n - X^l_n\widehat{f_{j,l}}\right)(\xi_{j,l}) \\ \langle b_{l,d+2}, b_{j,1}\rangle &= \frac{C_{j,l}}{(L^l)^2}\left(\widehat{|x|^2f_{j,l}} - 2X^l\cdot\widehat{xf_{j,l}} + |X^l|^2\widehat{f_{j,l}}\right)(\xi_{j,l}) \\ \langle b_{l,n+1}, b_{j,m+1}\rangle &= \frac{C_{j,l}}{L^jL^l}\left[x_n\widehat{x_mf_{j,l}} - X^l_n\widehat{x_mf_{j,l}} - X^j_m\widehat{x_nf_{j,l}} + X^l_nX^j_m\widehat{f_{j,l}}\right](\xi_{j,l}) \\ \langle b_{l,d+2}, b_{j,m+1}\rangle &= \frac{C_{j,l}}{(L^l)^2L^j}\begin{bmatrix} x_m\widehat{|x|^2f_{j,l}} - 2X^l\cdot x_m\widehat{xf_{j,l}} + |X^l|^2\widehat{x_mf_{j,l}} \\ -X^j_m\widehat{|x|^2f_{j,l}} + 2X^j_mX^l\cdot\widehat{xf_{j,l}} - |X^l|^2X^j_m\widehat{f_{j,l}}\end{bmatrix}(\xi_{j,l}), \end{aligned}$$

and

$$\langle b_{l,d+2}, b_{j,d+2}\rangle = \frac{C_{j,l}}{(L^l)^2(L^j)^2}\begin{bmatrix} \widehat{|x|^4 f_{j,l}} - 2X^l\cdot\widehat{|x|^2 x f_{j,l}} + |X^l|^2\widehat{|x|^2 f_{j,l}} \\ -2X^j\cdot x\widehat{|x|^2 f_{j,l}} + 4\displaystyle\sum_{n,m=1}^{d} X^l_n X^j_m x_n\widehat{x_m f_{j,l}} \\ -2|X^l|^2 X^j\cdot\widehat{x f_{j,l}} + \widehat{|x|^2 f_{j,l}}|X^j|^2 \\ -2|X^j|^2 X^l\cdot\widehat{x f_{j,l}} + |X^l|^2|X^j|^2\widehat{f_{j,l}} \end{bmatrix}(\xi_{j,l}). \tag{IV-4.15}$$

Moreover, we recall that $\mathbf{A}$ is Hermitian, so the above relations allow us to obtain all components of the matrix $\mathbf{A}$.

We now compute the components of the vector $\mathbf{S}$. For $j,k,l,m = 1,\dots,N$, let

$$\left|\begin{aligned} (C_{\mathrm{Im}})_{j,k,l,m} &:= \exp\left\{i\left(\gamma^k+\gamma^l-\gamma^m-\gamma^j\right)\right\} \\ &\quad\times\exp\left\{i\left(\beta^j\cdot X^j+\beta^m\cdot X^m-\beta^l\cdot X^l-\beta^k\cdot X^k\right)\right\} \\ &\quad\times\exp\left\{-i\left(\frac{B^k}{4(L^k)^2}|X^k|^2+\frac{B^l}{4(L^l)^2}|X^l|^2-\frac{B^m}{4(L^m)^2}|X^m|^2-\frac{B^j}{4(L^j)^2}|X^j|^2\right)\right\}, \\ (C_{\mathrm{Re}})_{j,k,l,m} &:= \exp\left\{-\frac{1}{2}\left(\frac{|X^k|^2}{(L^k)^2}+\frac{|X^l|^2}{(L^l)^2}+\frac{|X^m|^2}{(L^m)^2}+\frac{|X^j|^2}{(L^j)^2}\right)\right\}, \\ C_{j,k,l,m} &:= \frac{A^kA^lA^m}{L^kL^lL^m}(C_{\mathrm{Im}})_{j,k,l,m}(C_{\mathrm{Re}})_{j,k,l,m}, \\ \xi_{j,k,l,m} &:= -\left[\beta^k+\beta^l-\beta^m-\beta^j+\frac{B^k}{2(L^k)^2}X^k+\frac{B^l}{2(L^l)^2}X^l-\frac{B^m}{2(L^m)^2}X^m-\frac{B^j}{2(L^j)^2}X^j\right], \\ z_{j,k,l,m} &:= \frac{1}{2}\left(\frac{1}{(L^k)^2}+\frac{1}{(L^l)^2}+\frac{1}{(L^m)^2}+\frac{1}{(L^j)^2}\right) \\ &\quad+\frac{i}{4}\left(\frac{B^k}{(L^k)^2}+\frac{B^l}{(L^l)^2}-\frac{B^m}{(L^m)^2}-\frac{B^j}{(L^j)^2}\right), \\ a_{j,k,l,m} &:= \frac{1}{(L^k)^2}X^k+\frac{1}{(L^l)^2}X^l+\frac{1}{(L^m)^2}X^m+\frac{1}{(L^j)^2}X^j. \end{aligned}\right. \tag{IV-4.16}$$

Define

$$f_{j,k,l,m} : x\in\mathbb{R}^d \mapsto \exp(-z_{j,k,l,m}|x|^2 + a_{j,k,l,m}\cdot x)\in\mathbb{C},$$

then, for $1 \le r \le d$,

$$\begin{aligned} \langle u|u|^2, b_{j,1}\rangle &= \sum_{k,l,m} C_{j,k,l,m} \widehat{f_{j,k,l,m}}(\xi_{j,k,l,m}) \\ \langle u|u|^2, b_{j,r+1}\rangle &= \sum_{k,l,m} \frac{C_{j,k,l,m}}{L^j} \left(\widehat{x_r f_{j,k,l,m}}(\xi_{j,k,l,m}) - X_r^j \widehat{f_{j,k,l,m}}(\xi_{j,k,l,m}) \right) \qquad \text{(IV-4.17)} \\ \langle u|u|^2, b_{j,d+2}\rangle &= \sum_{k,l,m} \frac{C_{j,k,l,m}}{(L^j)^2} \begin{pmatrix} |x|^2 \widehat{f_{j,k,l,m}}(\xi_{j,k,l,m}) - 2X^j \cdot \widehat{x f_{j,k,l,m}}(\xi_{j,k,l,m}) \\ + |X^j|^2 \widehat{f_{j,k,l,m}}(\xi_{j,k,l,m}) \end{pmatrix}. \end{aligned}$$

$$\text{(IV-4.18)}$$

We refer to Section IV-6.2 for more details. Moreover, Lemma IV.5 gives the needed Fourier transform.

Lemma IV.5: Fourier transform of complex Gaussians

Let $z \in \mathbb{C}$, $\operatorname{Re}(z) \ge 0$. Then,

$$\widehat{e^{-z|\cdot|^2}}(\xi) = \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{-\frac{|\xi|^2}{4z}}, \quad \xi \in \mathbb{R}^d. \tag{IV-4.19}$$

More generally, let $z = z_1 + iz_2 \in \mathbb{C}$, $z_1, z_2 \in \mathbb{R}$, $z_1 > 0$, $a \in \mathbb{R}^d$ and

$$f: \ x \in \mathbb{R}^d \mapsto \exp\left(-z|x|^2 + a \cdot x\right) \in \mathbb{C}, \tag{IV-4.20}$$

then we have the Fourier transforms given by Table IV-4.1.

Proof. The proof relies on straightforward but lengthy computations. Details are given in Section IV-6.1. □

Remark IV.7: Computational complexity

Throughout this section, we have chosen

$$v^j(s^j, y^j) = e^{-\frac{1}{2}|y^j|^2}.$$

This choice was made so that the inner products involved in the application of the DF principle are easily computable in an exact way. Therefore we do not rely on numerical integration to compute the coefficients of the linear system (IV-4.11). In particular, this shows that the computational effort required to obtain the linear system is

$h(x)$	$\widehat{h}(\xi)/e^{-\frac{(\xi+ia)\cdot(\xi+ia)}{4z}}$
f	$\left(\frac{\pi}{z}\right)^{\frac{d}{2}}$
xf	$-i\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\frac{\xi+ia}{2z}$
$x_m x_n f$	$-\frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}(\xi_n+ia_n)(\xi_m+ia_m)$
$x_m^2 f$	$\frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[1-\frac{(\xi_m+ia_m)^2}{2z}\right]$
$\lvert x\rvert^2 f$	$\frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[d-\frac{\lvert\xi\rvert^2+2ia\cdot\xi-\lvert a\rvert^2}{2z}\right]$
$x_m\lvert x\rvert^2 f$	$-\frac{i}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}(\xi_m+ia_m)\left[d+2-\frac{\lvert\xi\rvert^2+2ia\cdot\xi-\lvert a\rvert^2}{2z}\right]$
$x_m^2 x_n^2 f$	$\frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left(1-\frac{(\xi_n+ia_n)^2}{2z}\right)\left(1-\frac{(\xi_m+ia_m)^2}{2z}\right)$
$x_m^4 f$	$\frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[3-6\frac{(\xi_m+ia_m)^2}{2z}+\frac{(\xi_m+ia_m)^4}{4z^2}\right]$

Table IV-4.1 – Fourier Transform of some polynomials in $x=(x_1,\dots,x_d)\in\mathbb{R}^d$ multiplied by $f(x)=e^{-z|x|^2+a\cdot x}$, $z\in\mathbb{C},\operatorname{Re}(z)>0, a\in\mathbb{R}^d$.

$\mathcal{O}(N^4d+N^2d^2)$. This complexity can be obtained by simply counting the elementary operations (or equivalently, *flops*) needed to fill the linear system (IV-4.11): in order to compute all the coefficients of the matrix $\mathbf{A}$, the most costful operation is (IV-4.15), which is $\mathcal{O}(d^2)$ for each inner product $\langle b_{l,d+2}, b_{j,d+2}\rangle$. Since the indices l, j range from 1 to N, there are N^2 such inner products to compute. This gives the $\mathcal{O}(N^2d^2)$ cost. We also have to compute the coefficients of the vector of interactions $\mathbf{S}$, and the most costful computations are given by (IV-4.17): it is a sum over 3 indices ranging from 1 to N, hence an $\mathcal{O}(N^3)$ cost per inner product $\langle u|u|^2, b_{j,r+1}\rangle$. The index j ranges from 1 to N, and r from 1 to d. Thus, there are Nd such inner products to compute, which gives a cost for $\mathbf{S}$ of $\mathcal{O}(N^4d)$. To obtain the total complexity, we have to add the cost of computing the pseudoinverse of the Hermitian matrix $\mathbf{A}\in\mathbb{C}^{(d+2)N,(d+2)N}$, which is $\mathcal{O}(N^3d^3)$ (see for instance [37, Fig. 8.6.1] for the computational complexity of the pseudoinverse). We also mention that one possible way to improve (from a computational point of view) the numerical solution of (IV-4.11) is to consider low-rank approximations, see for example [65]. This yields the overall computational complexity: $\mathcal{O}(N^4d+N^3d^3)$. In a more general setting, one could use the Hermite basis decomposition (IV-3.14) and perform all computations exactly. This would yield more involved computations and we chose the easy way out by experimenting only with Gaussian functions, but this is completely doable. By using the full Hermite basis, the complexity would also grow with the number of Hermite modes used.

Remark IV.8

If the computations can be performed by hand, as it is the case with Gaussian functions v (and *a priori* with Hermite functions), the method proposed truly is *grid-free*. If arbitrary functions v are used, then we cannot expect to be able to compute the inner products analytically by hand. We then have to resort to numerical integration, and the curse of dimensionality occurs. This is *only* due to being able to do most of the computations manually beforehand. If they cannot be done, they have to be approximated within the algorithm and this will yield a much more expensive algorithm.

By gathering the ideas mentioned previously, we obtain Algorithm 4 which can be used to obtain an approximate solution to (cNLS) as a sum of bubbles, using the Strang splitting between the linear and nonlinear parts, and using an arbitrary explicit time-integrator for the nonlinear part. The splitting error can be analyzed separately from the other types of errors, and we refer to [59, 40, 28, 19] for its analysis.

Algorithm 4 Approximating a solution to (cNLS) as a sum of bubbles.

```
Input:
— The bubble discretization of (IV-3.1), which gives the functions $u^j$.
— For each function $u^j$, its modulation parameters $(A^j, B^j, L^j, \beta^j, X^j, \gamma^j)$.
for  Each timestep of size $dt$  do
    for  $j = 1, \dots, N$  do                              ▷ $j$ denotes a bubble's index.
        Use Algorithm 2 to update the bubbles over a timestep of size $dt/2$.
        for each stage of a time-integrator do
            Compute the coefficients of the linear system (IV-4.11).
            Solve the linear system (IV-4.11) to obtain $\mathbf{E}$.
            Use (IV-4.13) to update the parameters over a timestep whose length depends
on the stage of the time-integrator.
        end for
        Use Algorithm 2 to update the bubbles over a timestep of size $dt/2$.
    end for
    Output:
end for
— The bubbles $u^j(T, \cdot)$ given by (IV-3.2), solution to (cNLS).
```

IV-4.2.2 Hamiltonian and norm conservation for the interactions

When solving (NL) via the DF principle, i.e. when solving the linear system (IV-4.11), a Hamiltonian is conserved.

Lemma IV.6

Let $u(t)$ be the approximation to (NL) obtained by applying the Dirac-Frenkel principle, and define

$$H_{\text{interactions}}(t) := \frac{1}{4}\langle u(t), u(t)|u(t)|^2\rangle = \frac{1}{4}\langle u(t)^2, u(t)^2\rangle.$$

Then $H_{\text{interactions}}$ is conserved, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t}H_{\text{interactions}}(t) = 0,$$

and the $\mathbb{L}^2$ norm of u is also conserved.

Proof. We have

$$H_{\text{interactions}}(t) := \frac{1}{4}\langle u(t), u(t)|u(t)|^2\rangle = \frac{1}{4}\langle u(t)^2, u(t)^2\rangle,$$

by using the Hermitian property of the inner product $\langle\cdot,\cdot\rangle$. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H_{\text{interactions}}(t) &= \frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t)^2, u(t)^2\rangle \\ &= \frac{1}{4}\left\langle 2u(t)\partial_t u(t), u(t)^2\right\rangle + \frac{1}{4}\left\langle u(t)^2, 2u(t)\partial_t u(t)\right\rangle \\ &= \operatorname{Re}\left\langle u(t)\partial_t u(t), u(t)^2\right\rangle \\ &= \operatorname{Re}\left\langle \partial_t u(t), u(t)|u(t)|^2\right\rangle.\end{aligned}$$

By definition of $\partial_t u(t)$, we have $\partial_t u(t) \in \mathcal{T}_{u(t)}\mathcal{M}$, hence we can take $f = \partial_t u(t)$ in (IV-4.4). We obtain the following equality:

$$\langle\partial_t u(t), u(t)|u(t)|^2\rangle = \langle\partial_t u(t), i\partial_t u(t)\rangle = -i\|\partial_t u(t)\|^2.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}H_{\text{interactions}}(t) = \operatorname{Re}\left(-i\|\partial_t u(t)\|^2\right) = 0.$$

Using similar ideas, we can easily show the conservation of the $\mathbb{L}^2$ norm: we obviously have $u(t) \in \mathcal{T}_{u(t)}\mathcal{M}$, hence

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2 &= 2\mathrm{Re}\,\langle u(t), \partial_t u(t)\rangle = 2\mathrm{Re}\,\langle u(t), -iu(t)|u(t)|^2\rangle \\ &= 2\mathrm{Re}\,\left(i\langle |u(t)|^2, |u(t)|^2\rangle\right) = 0.\end{aligned}$$

□

We have established the conservation of some quantities for both the linear and nonlinear part, but we are ultimately interested in the conservation properties when the two parts are combined. The composition of the linear and nonlinear parts conserve exactly the $\mathbb{L}^2$ norm, thus it is conserved through splitting. For the Hamiltonian of (cNLS), the Hamiltonian of the linear part is conserved exactly in the linear part, and the Hamiltonian of the nonlinear part is conserved exactly in the nonlinear part. When combining these two parts *via* splitting, we get that the Hamiltonian of (cNLS) is conserved up to the splitting error.

IV-4.2.3 Recovering the harmonic oscillator equations

Suppose the family $B_{u(t)} \subset \mathbb{L}^2(\mathbb{R}^d)$ defined by (IV-4.6) is linearly independent, and consider the equation (HO). By summing equation (IV-3.11) over $j = 1, \dots, N$ with $v^j(s^j, y^j) = e^{-\frac{|y^j|^2}{2}}$, and letting this sum be equal to zero, we obtain an equation of the form

$$\sum_{j=1}^{N} \left(c_{j,1}b_{j,1} + c_{j,2}b_{j,2} + \dots + c_{j,d+1}b_{j,d+1} + c_{j,d+2}b_{j,d+2}\right) = 0. \tag{IV-4.21}$$

The coefficients $c_{k,l}$ are, for instance, obtained by identifying powers of y and ∇_y in (IV-3.11). Thanks to the assumption that $B_{u(t)}$ is a linearly independent family, we know that we must have

$$c_{k,1} = c_{k,2} = \dots = c_{k,d+1} = c_{k,d+2} = 0, \qquad k = 1, \dots, (d+2)N. \tag{IV-4.22}$$

This yields exactly the system of equations (IV-3.16). Indeed, in (IV-3.16), we set the coefficient for each power of y and ∇_y to 0, except for the “-1”. The “-1” can be understood as needed to “compensate” the Laplacian operator applied to v, in the case of v a Hermite function, because then

$$\Delta_y H_m(y) = |y|^2 H_m(y).$$

In other words, the DF principle approach gives the same equations as those given in Section IV-3.1.3 when $B_{u(t)}$ is a linearly independent family. However, our approach as described in Section IV-3.1.3 allows to solve them exactly and not only numerically with some numerical time-integrator.

Finally, if the family $B_{u(t)}$ is linearly dependent, then we cannot write equation (IV-4.22) anymore, hence the DF principle approach in the linear case fails. Our approach avoids this issue by naturally imposing conditions (IV-4.22) (which are the same as (IV-3.12)).

IV-4.3 Numerical examples

In this Section we will assess the numerical efficiency of the bubble scheme in its nonlinear setting. The reference scheme will be a slight variation of the spectral scheme given in Section IV-3.2. More precisely, we use a Strang splitting on (cNLS), so that we can reuse the spectral scheme from the linear Chapter. For the grid approximation of the solution to (NL), we use the fact that the modulus $|\psi(\cdot, x)|$ is conserved when solving (NL). Indeed, if we multiply (NL) by $\overline{\psi}$, we get

$$i\overline{\psi}\partial_t\psi = |\psi|^4.$$

By taking the complex conjugate of this equation,

$$-i\psi\partial_t\overline{\psi} = |\psi|^4.$$

Finally,

$$\begin{aligned} i\partial_t|\psi|^2 &= i\partial_t\left(\psi\overline{\psi}\right) = i\overline{\psi}\partial_t\psi + i\psi\partial_t\overline{\psi} \\ &= |\psi|^4 - |\psi|^4 = 0, \end{aligned}$$

which implies that the modulus of ψ is constant with respect to time t. The spectral scheme is the nonlinear setting is then fully described by Algorithm 5.

We use the results of Section IV-4.2.1 in order to compute the linear system (IV-4.11). We recall that those computations were performed analytically since we used the assumption $v(s, y) = e^{-\frac{|y|^2}{2}}$.

IV-4.3.1 Discretization into a sum of Bubbles

We need to decompose any arbitrary function into a finite sum of N bubbles. A solution to this question has been proposed in [66], but it involves integrals over the whole phase space, which is something we want to avoid.

Algorithm 5 Spectral solver for (cNLS), with a Strang Splitting method.

```
Input:
— GRID: a uniform discretization of a finite volume subset of $\mathbb{R}^d$
— the initial data $\psi_0$.
Discretize the initial data $\psi_0$ on GRID $\subset \mathbb{R}^d$, and let $\eta$ be this discretization.
for Each timestep of size $\Delta t$ do
    Use Algorithm 3 with a stepsize $\Delta t/2$.
    for $x \in$ GRID do                                  ▷ Add interactions.
        Multiply $\eta(x)$ by $e^{-i\,\Delta t\,|\eta(x)|^2}$.
    end for
    Use Algorithm 3 with a stepsize $\Delta t/2$.
end for
Output:
— $\eta$ is approximation on GRID of $\psi(T,\cdot)$, where $\psi$ is the solution to (cNLS).
```

We could also use a nonlinear least squares approach, but our experimental results showed that it tends to yield spread out Gaussians, which may present huge overlaps between them. The overlaps cause issues with the DF principle, for instance a blow-up of the conservative quantities. This has been observed during our experiments but the results are not reported in the numerical results. The issue of discretizing an arbitrary function into a sum of bubbles is not the main concern of this work, hence we will consider initial conditions whose bubble discretization is natural and straightforward. Another possible way of discretizing the initial data is outlined in [1]. Generally, the discretization of an arbitrary function as a sum of Gaussian functions is an active research area, and we can cite for instance [12] who studies several discretization methods. This work also has some perspectives about the high-dimensional case, where the discretization using quadrature rules faces the curse of dimensionality. Finally, if we do not restrict ourselves to Gaussian functions and allow general Hermite functions, then the discretization simply consists in projecting the initial condition onto this basis, and truncating the highest modes if necessary.

IV-4.3.2 Results

Some of the given examples are adapted from [8]. Note that for each example, the solution of the harmonic oscillator is computed for each time step $\{t^n\}_{n\geq 0}$. In practice, if one needs the solution of (HO) at time t^n, they can simply use Algorithm 2 with $T = t^n$, and there is no need for a time discretization.

In this regard, the numerical simulation of (IV-3.2) outperforms the spectral scheme of reference. When $M \geq 1$ Hermite modes are considered in each dimension for each bubble, the computational complexity simply is $\mathcal{O}(N(M^d + d))$. This is very favorable for high-dimensional use if the number of Hermite modes is small. In contrast, when the nonlinear part is introduced, the use of the Dirac-Frenkel principle makes the computational complexity grow polynomially with N and d. It gets even worse if we allow Hermite modes other than the Gaussian function. Furthermore, we have to take into account the numerical issues caused by the DF principle (see for instance [47, 69]).

In our numerical results, we have used $T = 100$. We note that it is on the same order of magnitude as the final times from [14], and much larger than the simulation times in [5].

When solving the nonlinear part (NL) of (cNLS) using the Dirac-Frenkel approach, we have to integrate (IV-4.13). In practice, we used a second-order Runge-Kutta time-integrator (see for instance [6, Eqn. (6.10.10)]). Since the Strang splitting error is of order 2 in time, a second-order time-integrator is enough.

Our numerical implementation has been done in `Julia`.

Test case 1: Weak interactions

The initial condition is given in Section IV-3.2.4, and the results for the non linear case are displayed in Figure IV-4.1. This example shows the performance of the DF principle approach in its most efficient setting: it only has one bubble. This explains the very good conservation results obtained: the bubbles scheme (orange solid line) outperforms the spectral scheme (blue dash line) on (cNLS), except for the energy. However, in this case, the error of the DF principle method remains globally less than one order of magnitude larger than the error from the spectral method.

Test case 2: Rotating Gaussians

This test case is an illustration of the good conservation properties of the modulation algorithm, including the nonlinearities, as soon as the bubbles don't have too much overlap. The initial condition is given in Section IV-3.2.4, and the results are given in Figure IV-4.2. We can observe that, for this test case, the Dirac-Frenkel principle works well, since the bubbles do not present too much overlap. The spectral scheme (blue dash line) is outperformed by the bubbles scheme (orange solid line) for all times t, and the spectral scheme performs poorly for large times in the presence of cubic interactions. Note that, in this figure, the bounding box $[-15, 15] \times [-15, 15]$ is only here for plotting purposes, and the solution u is known analytically in the whole $\mathbb{R}^2$ plane.

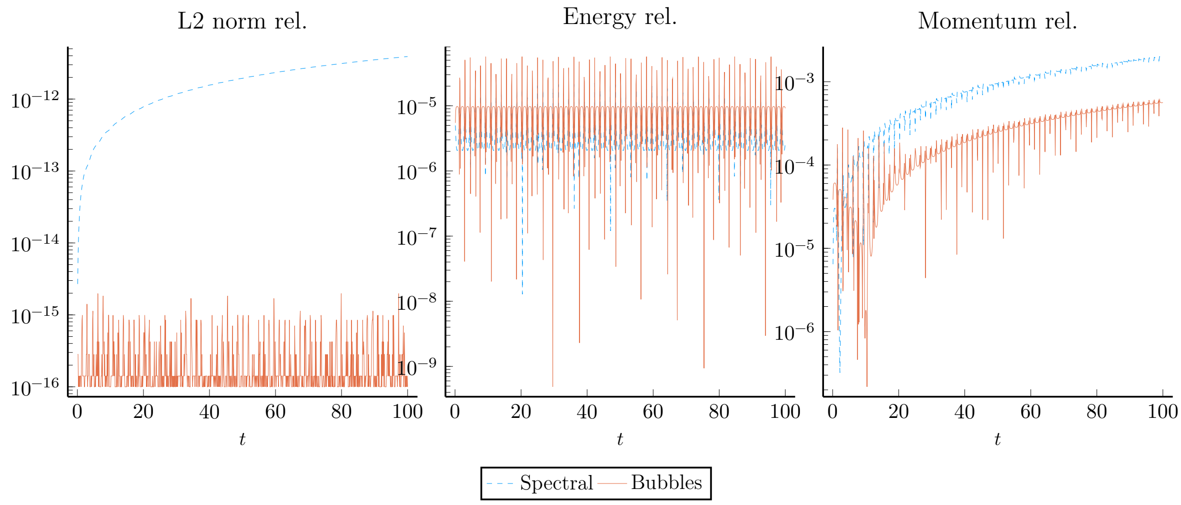


Figure IV-4.1 – Test case 1. Relative evolution of mass, energy and momentum with bubbles and spectral methods. $\Delta t = 10^{-2}$. Time-integrator for the nonlinear part of the splitting: Runge-Kutta of order 2. Spectral scheme with $N_x = 256, N_y = 256$.

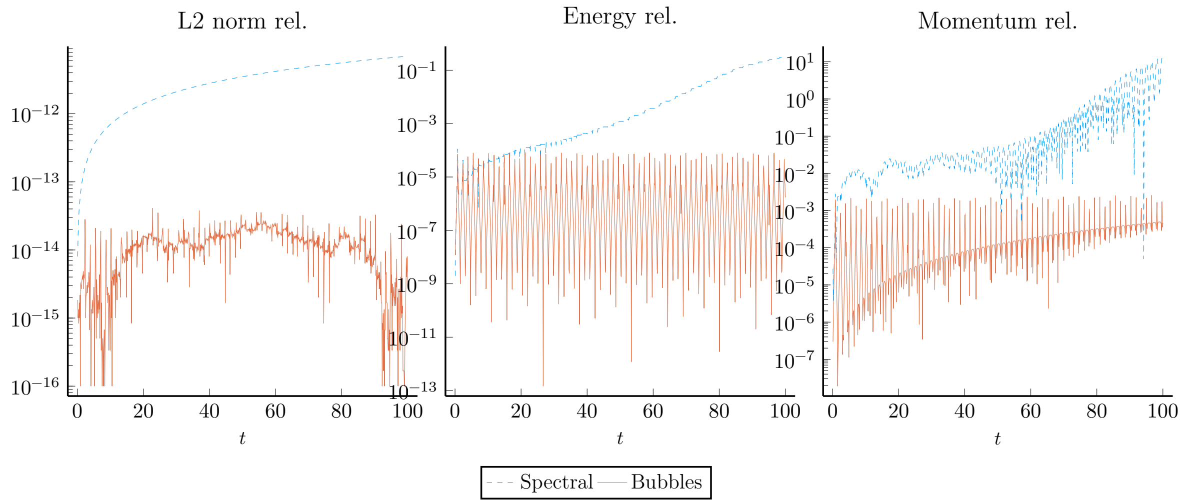


Figure IV-4.2 – Test case 2. Relative evolution of mass, energy and momentum with bubbles and spectral methods. $\Delta t = 10^{-2}$. Time-integrator for the nonlinear part of the splitting: Runge-Kutta of order 2. Spectral scheme with $N_x = 256, N_y = 256$.

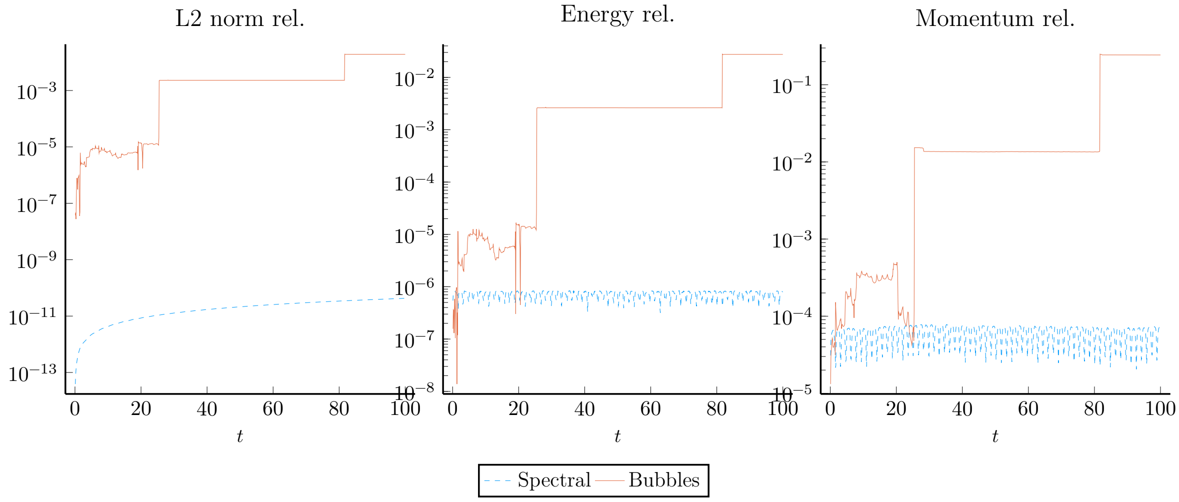


Figure IV-4.3 – Test case 3. Relative evolution of mass, energy and momentum with bubbles and spectral methods. $\Delta t = 10^{-3}$. Time-integrator for the nonlinear part of the splitting: Runge-Kutta of order 2. Spectral scheme with $N_x = 256, N_y = 256$.

Test case 3: Zero phase initial data

The initial condition is given in Section IV-3.2.4 and the results are given in Figure IV-4.3. We can note that the bubble solution (orange solid line) is outperformed by the spectral method (blue dash line). When we compare them on (cNLS), the DF principle suffers from inherent issues (already mentioned in [47, 69, 50] for instance) and gives very poor results. This can be explained by the overlapping of bubbles, which then gives a badly-conditioned matrix $\mathbf{A}$. Note that, in order to be able to use DF principle for this test case, the time step Δt had to be lowered from $\Delta t = 10^{-2}$ (the timestep used for all other numerical results) to $\Delta t = 10^{-3}$. Otherwise, the pseudoinverse of the matrix $\mathbf{A}$ was too badly conditioned. Even then, the results are very poor and illustrate the issues inherent to the Dirac-Frenkel principle.

Part IV

CHAPTER 5 Conclusions and perspectives

We presented in this work an approach based on recent results from [60, 57, 31]. It allows to solve exactly the harmonic oscillator (HO) for initial functions that can be represented as a sum of modulated functions (the *bubbles*), for a certain kind of modulation.

In this context we focused on a particular subclass of such functions, modulated Hermite functions, which have the advantage of allowing explicit computations. This is particularly interesting since we do not have to rely on any sort of discretization of the phase space, which is usually the main computational burden in numerical simulations. We obtain an algorithm which yields an exact solution as soon as the initial data is a sum of modulated Hermite functions. If we consider an arbitrary initial function, it suffices to project it into onto the Hermite basis and to perform analytical time-evolution. Moreover, the algorithm only relies on a small number of parameters whose time-evolution is explicit, making it very fast and computationally efficient. However, the algorithm possesses some limitations. The most obvious one is that we solved the harmonic oscillator, which allowed us to use Hermite functions and to integrate exactly the modulation parameters with respect to time. This is very restrictive, and one could wonder if such results hold for more general potential functions V. It seems easy and straightforward to obtain modulation equations for a quadratic potential V using the results given here, but it is much less clear how the Hermite decomposition will work with arbitrary potentials. One option could be to use Hagedorn functions, see Section IV-2.2.4. If one considers a non quadratic potential, the simplification mentioned by Heller [43, 42] could be interesting: it consists in using local quadratic approximations of the potential. In this case, an extended numerical analysis would be required to see how this local approximation affects the results.

On our numerical tests, the bubble algorithm outperforms a spectral method when compared on the harmonic oscillator. Moreover, any grid-based method is inherently bound to a finite subset of $\mathbb{R}^d$ to which we have to add boundary conditions, while the bubble approach does not have such restrictions. We emphasize the fact that the algorithm presented here in the case of Gaussian functions extends in a natural manner when dealing with complex modulated Hermite functions. The presented algorithm is also more general in some sense than other grid-free spectral methods: indeed, we allow multiple Hermite basis (basically one basis per bubble), so that the number of Hermite modes for a given initial condition can remain low. This constrasts with, for example, the schemes presented in [11, 10, 80], where only one Hermite basis is considered. Depending on the initial condition, we may need many more Hermite modes with one basis than with several bases (for example, if the initial condition is made of two bumps far away from each other).

We also extended the results from [31] by allowing cubic interactions, at the cost of approximating the solution to the cubic nonlinear equation (NL) via the Dirac-Frenkel principle. We only considered modulated Gaussian functions, because they allowed us to easily perform explicit computations and to obtain a numerical algorithm whose computational complexity is $\mathcal{O}(N^4 d + N^3 d^3)$. Here d is the dimension and N is the number of bubbles. The most critical parameter is N, which corresponds roughly to the precision of the discretization when considering arbitrary initial data. For any given function, the higher N, the better we can approximate it as a sum of modulated Gaussian functions. We then have a clear trade-off between the speed of the algorithm and the precision of the discretization.

The algorithm for the nonlinear part makes use of the Dirac-Frenkel variational principle, and it appears in practice that some inherent issue may arise. This issue makes the results be very unsatisfying if the bubbles overlap at some point during the simulation. If the bubbles do not present overlap, the Dirac-Frenkel principle works well. We have to underline some limitations of the algorithm as presented here. First of all, the discretization error has not been analyzed in detail, and this is crucial when studying the error for arbitrary initial condition. A second limitation, probably the biggest issue, is due to the Dirac-Frenkel principle, which displays the same issues as those observed when it is applied in the linear setting: in order to obtain an approximate time-derivative, the interactions $|u|^2 u$ need to be orthogonally projected onto some tangent space, and we use an explicit basis of the tangent space in order to perform that projection. However, it appears in practice that the family we use as a basis for the tangent space may not be a basis and may present some linear dependences. In this case, the projection matrix in the Dirac-Frenkel principle is very ill-conditioned or noninvertible, and this is the cause of the main numerical issues observed. This situation occurs for example when two bubbles overlap too much, which happens in practice. It would be interesting to have some procedure that allow overlaps, and maybe study the alternative used by [50] and see if it applies to polynomial interactions. It would be interesting as well to perform an exhaustive numerical analysis, and in particular to study the error terms coming from the Dirac-Frenkel principle. Another issue is that we do not allow here the number of bubbles to grow and diminish with time.

As a final note, we can say that the ideas presented here seem promising, but a lot of work has to be made in order to improve them to the point that they are applicable in most situations, with performance comparable to other state-of-the-art schemes.

Part IV

6 CHAPTER

Miscellanenous computations

IV-6.1 Fourier transforms of Gaussians

For the sake of clarity, for $\xi, a \in \mathbb{R}^d$ and $z \in \mathbb{C}$, let

$$E(\xi, a, z) := \exp\left\{-\frac{|\xi|^2 + 2ia\cdot\xi - |a|^2}{4z}\right\} = \exp\left\{-\frac{(\xi + ia)\cdot(\xi + ia)}{4z}\right\}.$$

Computation of $\widehat{f}$

We have

$$-z|x|^2 + a\cdot x = -z\left|x - \frac{a}{2z_1}\right|^2 - i\frac{z_2 a}{z_1}\cdot x + \frac{z|a|^2}{4z_1^2}.$$

Recall the following usual properties on Fourier transform:

$$f\widehat{(x-a)} = \widehat{f}(\xi)e^{-ia\cdot\xi}, \quad f\widehat{e^{-ia\cdot x}} = \widehat{f}(\xi + a).$$

Let

$$g(x) = e^{-z\left|x - \frac{a}{2z_1}\right|^2},$$

then

$$\widehat{g}(\xi) = \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{-\frac{|\xi|^2}{4z} - \frac{ia\cdot\xi}{2z_1}}$$

and

$$f(x) = g(x)e^{-\frac{iz_2}{z_1}a\cdot x + \frac{z|a|^2}{4z_1^2}}.$$

Hence,

$$\begin{aligned}
\hat{f}(\xi) &= e^{\frac{z|a|^2}{4z_1^2}}\,\hat{g}\left(\xi + \frac{z_2}{z_1}a\right) = \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{\frac{z|a|^2}{4z_1^2}}\, e^{-\frac{1}{4z}\left|\xi+\frac{z_2}{z_1}a\right|^2-\frac{ia}{2z_1}\cdot\left(\xi+\frac{z_2}{z_1}a\right)}\\
&= \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{\frac{z|a|^2}{4z_1^2}-\frac{1}{4z}\left(|\xi|^2+2\frac{z_2}{z_1}a\cdot\xi+\frac{z_2^2}{z_1^2}|a|^2\right)-\frac{ia\cdot\xi}{2z_1}-\frac{i|a|^2z_2}{2z_1^2}}\\
&= \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{-\frac{|\xi|^2}{4z}+(a\cdot\xi)\left(-\frac{z_2}{2zz_1}-\frac{i}{2z_1}\right)+|a|^2\left(\frac{z}{4z_1^2}-\frac{z_2^2}{4zz_1^2}-\frac{iz_2}{2z_1^2}\right)}\\
&= \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{-\frac{|\xi|^2}{4z}-\frac{a\cdot\xi}{2zz_1}[z_2+i(z_1+iz_2)]+\frac{|a|^2}{4zz_1^2}[(z_1+iz_2)^2-z_2^2-2iz_2(z_1+iz_2)]}\\
&= \left(\frac{\pi}{z}\right)^{\frac{d}{2}} e^{-\frac{|\xi|^2}{4z}-i\frac{a\cdot\xi}{2z}+\frac{|a|^2}{4z}}\\
&= \left(\frac{\pi}{z}\right)^{\frac{d}{2}} E(\xi, a, z).
\end{aligned}$$

Computation of $\widehat{xf}$

$$\begin{aligned}
\widehat{xf}(\xi) &= i\nabla_\xi \hat{f} = i\nabla_\xi\left[\left(\frac{\pi}{z}\right)^{\frac{d}{2}} E(\xi,a,z)\right] = i\left(\frac{\pi}{z}\right)^{\frac{d}{2}} E(\xi,a,z)\left[-\frac{\xi}{2z}-\frac{ia}{2z}\right]\\
&= -i\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\frac{\xi+ia}{2z}E(\xi,a,z).
\end{aligned}$$

Computation of $\widehat{x_m^2 f}$, $m = 1, \dots, d$

$$\begin{aligned}
\widehat{x_m^2 f}(\xi) &= i\partial^{\xi_m}\left(\widehat{xf}\right)_m = i\partial^{\xi_m}\left[-i\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\frac{(\xi+ia)_m}{2z}E(\xi,a,z)\right]\\
&= \frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[E(\xi,a,z)+(\xi_m+ia_m)\,E(\xi,a,z)\left(-\frac{\xi_m}{2z}-\frac{ia_m}{2z}\right)\right]\\
&= \frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[1-\frac{(\xi_m+ia_m)^2}{2z}\right]E(\xi,a,z).
\end{aligned}$$

Computation of $\widehat{x_m x_n f}$, $m, n = 1, \dots, d$, $n \neq m$

$$\begin{aligned}\widehat{x_m x_n f}(\xi) &= i\partial^{\xi_m}\left(\widehat{xf}\right)_n = i\partial^{\xi_m}\left[-i\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\frac{\xi_n + ia_n}{2z}E(\xi, a, z)\right]\\ &= \frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}(\xi_n + ia_n)\left[-\frac{\xi_m + ia_m}{2z}\right]E(\xi, a, z)\\ &= -\frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}(\xi_n + ia_n)(\xi_m + ia_m)E(\xi, a, z).\end{aligned}$$

Computation of $\widehat{|x|^2 f}$

$$\begin{aligned}\widehat{|x|^2 f}(\xi) &= \widehat{x_1^2 f}(\xi) + \dots + \widehat{x_d^2 f}(\xi)\\ &= \frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[d - \frac{(\xi_1 + ia_1)^2 + \dots + (\xi_d + ia_d)^2}{2z}\right]E(\xi, a, z)\\ &= \frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[d - \frac{|\xi|^2 + 2ia\cdot\xi - |a|^2}{2z}\right]E(\xi, a, z).\end{aligned}$$

Computation of $\widehat{x_m |x|^2 f}$, $m = 1, \dots, d$

$$\begin{aligned}&\widehat{x_m |x|^2 f}(\xi)\\ &= i\partial^{\xi_m}\left[\widehat{|x|^2 f}(\xi)\right] = i\partial^{\xi_m}\left[\frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left(d - \frac{|\xi|^2 + 2ia\cdot\xi - |a|^2}{2z}\right)E(\xi, a, z)\right]\\ &= \frac{i}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[-2\frac{\xi_m + ia_m}{2z} + \left(d - \frac{|\xi|^2 + 2ia\cdot\xi - |a|^2}{2z}\right)\left(-\frac{\xi_m + ia_m}{2z}\right)\right]E(\xi, a, z)\\ &= -\frac{i}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}(\xi_m + ia_m)\left[d + 2 - \frac{|\xi|^2 + 2ia\cdot\xi - |a|^2}{2z}\right]E(\xi, a, z).\end{aligned}$$

Computation of $\widehat{x_m^3 f}$, $m = 1, \dots, d$

$$\begin{aligned}
\widehat{x_m^3 f}(\xi) &= i\partial^{\xi_m}\left[\widehat{x_m^2 f}(\xi)\right] = i\partial^{\xi_m}\left[\frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left(1-\frac{(\xi_m+ia_m)^2}{2z}\right)E(\xi,a,z)\right]\\
&= \frac{i}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[-2\frac{\xi_m+ia_m}{2z}+\left(-\frac{\xi_m+ia_m}{2z}\right)\left(1-\frac{(\xi_m+ia_m)^2}{2z}\right)\right]E(\xi,a,z)\\
&= -\frac{i}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}(\xi_m+ia_m)\left[3-\frac{(\xi_m+ia_m)^2}{2z}\right]E(\xi,a,z)\\
&= -\frac{i}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[3(\xi_m+ia_m)-\frac{(\xi_m+ia_m)^3}{2z}\right]E(\xi,a,z).
\end{aligned}$$

Computation of $\widehat{x_m x_n^2 f}$, $m, n = 1, \dots, d$, $n \neq m$

$$\begin{aligned}
\widehat{x_m x_n^2 f}(\xi) &= i\partial^{\xi_m}\left(\widehat{x_n^2 f}\right) = i\partial^{\xi_m}\left[\frac{1}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left(1-\frac{(\xi_n+ia_n)^2}{2z}\right)E(\xi,a,z)\right]\\
&= -\frac{i}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left(1-\frac{(\xi_n+ia_n)^2}{2z}\right)\frac{\xi_m+ia_m}{2z}E(\xi,a,z).
\end{aligned}$$

Computation of $\widehat{x_m^4 f}$, $m = 1, \dots, d$

$$\begin{aligned}
&\widehat{x_m^4 f}(\xi)\\
&= i\partial^{\xi_m}\left[\widehat{x_m^3 f}(\xi)\right] = i\partial^{\xi_m}\left[-\frac{i}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left(3(\xi_m+ia_m)-\frac{(\xi_m+ia_m)^3}{2z}\right)E(\xi,a,z)\right]\\
&= \frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[3-3\frac{(\xi+ia_m)^2}{2z}+\left(3(\xi_m+ia_m)-\frac{(\xi_m+ia_m)^3}{2z}\right)\left(-\frac{\xi_m+ia_m}{2z}\right)\right]E(\xi,a,z)\\
&= \frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}}\left[3-6\frac{(\xi_m+ia_m)^2}{2z}+\frac{(\xi_m+ia_m)^4}{4z^2}\right]E(\xi,a,z).
\end{aligned}$$

Computation of $\widehat{x_m^2 x_n^2 f}$, $m = 1, \dots, d$, $n \neq m$

$$\begin{aligned}
\widehat{x_m^2 x_n^2 f}(\xi) &= i\partial^{\xi_m} \left(\widehat{x_m x_n^2 f}\right)_n = i\partial^{\xi_m} \left[-\frac{i}{2z}\left(\frac{\pi}{z}\right)^{\frac{d}{2}} \left(1 - \frac{(\xi_n + ia_n)^2}{2z}\right) \frac{\xi_m + ia_m}{2z} E(\xi, a, z)\right] \\
&= \frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}} \left(1 - \frac{(\xi_n + ia_n)^2}{2z}\right) \partial^{\xi_m}\left[(\xi_m + ia_m)E(\xi, a, z)\right] \\
&= \frac{1}{4z^2}\left(\frac{\pi}{z}\right)^{\frac{d}{2}} \left(1 - \frac{(\xi_n + ia_n)^2}{2z}\right)\left(1 - \frac{(\xi_m + ia_m)^2}{2z}\right) E(\xi, a, z).
\end{aligned}$$

IV-6.2 Computing the coefficients of the Dirac-Frenkel linear system

Coefficients of the matrix A

Computation of $\langle b_{l,1}, b_{j,1}\rangle$

$$\begin{aligned}
\langle b_{l,1}, b_{j,1}\rangle &= e^{i\gamma^l - i\gamma^j} \int_{\mathbb{R}^d} e^{iL^l\beta^l\cdot\frac{x-X^l}{L^l} - i\frac{B^l}{4}\left|\frac{x-X^l}{L^l}\right|^2} e^{-\frac{1}{2}\left|\frac{x-X^l}{L^l}\right|^2} \\
&\qquad\qquad \times e^{-iL^j\beta^j\cdot\frac{x-X^j}{L^j} + i\frac{B^j}{4}\left|\frac{x-X^j}{L^j}\right|^2} e^{-\frac{1}{2}\left|\frac{x-X^j}{L^j}\right|^2} dx \\
&= e^{i(\gamma^l - \gamma^j)} \int_{\mathbb{R}^d} e^{i\beta^l\cdot(x-X^l) - i\beta^j\cdot(x-X^j)} e^{-\frac{2+iB^l}{4}\left|\frac{x-X^l}{L^l}\right|^2} e^{-\frac{2-iB^j}{4}\left|\frac{x-X^j}{L^j}\right|^2} dx \\
&= e^{i(\gamma^l-\gamma^j) - \frac{2+iB^l}{4(L^l)^2}|X^l|^2 - \frac{2-iB^j}{4(L^j)^2}|X^j|^2 - i\beta^l\cdot X^l + i\beta^j\cdot X^j} \\
&\quad \times \int_{\mathbb{R}^d} e^{i(\beta^l - \beta^j)\cdot x} e^{-\frac{2+iB^l}{4(L^l)^2}(|x|^2 - 2x\cdot X^l)} e^{-\frac{2-iB^j}{4(L^j)^2}(|x|^2 - 2x\cdot X^j)} dx \\
&= e^{i(\gamma^l-\gamma^j) - \frac{2+iB^l}{4(L^l)^2}|X^l|^2 - \frac{2-iB^j}{4(L^j)^2}|X^j|^2 - i\beta^l\cdot X^l + i\beta^j\cdot X^j} \\
&\quad \times \int_{\mathbb{R}^d} e^{i(\beta^l - \beta^j + \frac{B^l}{2(L^l)^2}X^l - \frac{B^j}{2(L^j)^2}X^j)\cdot x} e^{x\cdot\left(\frac{1}{(L^l)^2}X^l + \frac{1}{(L^j)^2}X^j\right)} e^{-\left(\frac{2+iB^l}{4(L^l)^2} + \frac{2-iB^j}{4(L^j)^2}\right)|x|^2} dx
\end{aligned}$$

Let

$$\left|\begin{aligned}
&z := \frac{2+iB^l}{4(L^l)^2} + \frac{2-iB^j}{4(L^j)^2},\\
&a := \frac{X^l}{(L^l)^2} + \frac{X^j}{(L^j)^2},\\
&\xi := \frac{B^j}{2(L^j)^2}X^j + \beta^j - \frac{B^l}{2(L^l)^2}X^l - \beta^l,\\
&C = \exp\left\{i(\gamma^l - \gamma^j) - \frac{2+iB^l}{4(L^l)^2}|X^l|^2 - \frac{2-iB^j}{4(L^j)^2}|X^j|^2 - i\beta^l\cdot X^l + i\beta^j \cdot X^j\right\},
\end{aligned}\right. \tag{IV-6.1}$$

and $f(x) := e^{-z|x|^2 + a\cdot x}$. Then

$$\langle b_{l,1}, b_{j,1}\rangle = C\int_{\mathbb{R}^d} e^{-i\xi\cdot x} f(x)dx = C\hat{f}(\xi)$$

Computation of $\langle b_{l,n+1}, b_{j,1}\rangle$, $1 \le n \le d$

$$\begin{aligned}
\langle b_{l,n+1}, b_{j,1}\rangle &= C\int_{\mathbb{R}^d} \frac{(x-X^l)_n}{L^l} e^{-i\xi\cdot x} f(x)dx\\
&= \frac{C}{L^l}\left(\widehat{xf}_n - X^l_n \hat{f}\right)(\xi)
\end{aligned}$$

Computation of $\langle b_{l,d+2}, b_{j,1}\rangle$

$$\begin{aligned}
\langle b_{l,d+2}, b_{j,1}\rangle &= C\int_{\mathbb{R}^d} e^{-i\xi\cdot x} f(x)\frac{|x-X^l|^2}{(L^l)^2}dx\\
&= \frac{C}{(L^l)^2}\int_{\mathbb{R}^d} e^{-i\xi\cdot x} f(x)\left(|x|^2 - 2x\cdot X^l + |X^l|^2\right)dx\\
&= \frac{C}{(L^l)^2}\left(\widehat{|x|^2 f} - 2X^l\cdot\widehat{xf} + |X^l|^2\hat{f}\right)(\xi)
\end{aligned}$$

Computation of $\langle b_{l,n+1}, b_{j,m+1}\rangle$, $1 \leq n, m \leq d$

$$\begin{aligned}
\langle b_{l,n+1}, b_{j,m+1}\rangle &= C\int_{\mathbb{R}^d} \frac{x_n - X^l_n}{L^l}\frac{x_m - X^j_m}{L^j} e^{-i\xi\cdot x} f(x)dx \\
&= \frac{C}{L^j L^l}\int_{\mathbb{R}^d} (x_n - X^l_n)(x_m - X^j_m) e^{-i\xi\cdot x} f(x)dx \\
&= \frac{C}{L^j L^l}\int_{\mathbb{R}^d} \left[x_n x_m - x_n X^j_m - x_m X^l_n + X^l_n X^j_m\right] \\
&\qquad\qquad \times e^{-i\xi\cdot x} f(x)dx \\
&= \frac{C}{L^j L^l}\left[\widehat{x_n x_m f} - X^l_n \widehat{x_m f} - X^j_m \widehat{x_n f} + X^l_n X^j_m \widehat{f}\right](\xi).
\end{aligned}$$

Computation of $\langle b_{l,d+2}, b_{j,m+1}\rangle$, $1 \leq m \leq d$

$$\begin{aligned}
&\langle b_{l,d+2}, b_{j,m+1}\rangle \\
&= C\int_{\mathbb{R}^d} e^{-i\xi\cdot x} e^{-z|x|^2 + a\cdot x}\frac{|x - X^l|^2}{(L^l)^2}\frac{x_m - X^j_m}{L^j} dx \\
&= \frac{C}{(L^l)^2 L^j}\int_{\mathbb{R}^d} e^{-i\xi\cdot x} e^{-z|x|^2 + a\cdot x}\left(|x|^2 - 2x\cdot X^l + |X^l|^2\right)\left(x_m - X^j_m\right) dx \\
&= \frac{C}{(L^l)^2 L^j}\left[\widehat{x_m |x|^2 f} - 2X^l\cdot\widehat{x_m x f} + |X^l|^2\widehat{x_m f}\right. \\
&\qquad \left. - X^j_m\widehat{|x|^2 f} + 2X^j_m X^l\cdot\widehat{xf} - |X^l|^2 X^j_m\widehat{f}\right](\xi).
\end{aligned}$$

Computation of $\langle b_{l,d+2}, b_{j,d+2}\rangle$

$$\begin{aligned}
\langle b_{l,d+2}, b_{j,d+2}\rangle &= C\int_{\mathbb{R}^d} e^{-i\xi\cdot x} e^{-z|x|^2+a\cdot x}\frac{|x-X^l|^2}{(L^l)^2}\frac{|x-X^j|^2}{(L^j)^2}dx \\
&= \frac{C}{(L^l)^2(L^j)^2}\int_{\mathbb{R}^d} e^{-i\xi\cdot x} e^{-z|x|^2+a\cdot x}\left(|x|^2-2x\cdot X^l+|X^l|^2\right) \\
&\qquad \times\left(|x|^2-2x\cdot X^j+|X^j|^2\right)dx \\
&= \frac{C}{(L^l)^2(L^j)^2}\int_{\mathbb{R}^d} e^{-i\xi\cdot x} e^{-z|x|^2+a\cdot x}\left(|x|^4-2|x|^2x\cdot X^l+|X^l|^2|x|^2\right. \\
&\qquad -2(x\cdot X^j)|x|^2+4(x\cdot X^l)(x\cdot X^j)-2(x\cdot X^j)|X^l|^2 \\
&\qquad \left.+|x|^2|X^j|^2-2(x\cdot X^l)|X^j|^2+|X^l|^2|X^j|^2\right)dx \\
&= \frac{C}{(L^l)^2(L^j)^2}\left[\widehat{|x|^4f}-2(X^l+X^j)\cdot|\widehat{x|^2x}f+\left(|X^l|^2+|X^j|^2\right)\widehat{|x|^2f}\right. \\
&\qquad +4(x\cdot\widehat{X^l)(x}\cdot X^j)f-2\left(|X^l|^2X^j+|X^j|^2X^l\right)\cdot\widehat{xf} \\
&\qquad \left.+|X^l|^2|X^j|^2\widehat{f}\right](\xi)
\end{aligned}$$

Moreover,

$$\begin{aligned}
(x\cdot X^l)(x\cdot X^j) &= \left(\sum_{n=1}^d x_nX_n^l\right)\left(\sum_{m=1}^d x_mX_m^j\right) \\
&= \sum_{n,m=1}^d X_n^lX_m^jx_nx_m,
\end{aligned}$$

Hence

$$(x\cdot\widehat{X^l)(x}\cdot X^j)f=\sum_{n,m=1}^d X_n^lX_m^j\widehat{x_nx_m}f.$$

Coefficients of the vector of interactions S

Computation of $\langle u|u|^2, b_{j,1}\rangle$

$$\langle u|u|^2, b_{j,1}\rangle = \sum_{k,l,m}\frac{A^kA^lA^m}{L^kL^lL^m}\left\langle e^{i\Gamma^k+i\Gamma^l-i\Gamma^m}e^{-\frac{|y_k|^2+|y_l|^2+|y_m|^2}{2}}, e^{i\Gamma^j}e^{-\frac{1}{2}|y_j|^2}\right\rangle$$

We recall the previously defined notations:

$$\left| \begin{aligned} y_k &= \frac{x - X^k}{L^k}, \\ \Gamma^k(x) &= \gamma^k + \beta^k \cdot (x - X^k) - \frac{B^k}{4(L^k)^2}|x - X^k|^2. \end{aligned} \right.$$

Then,

$$\begin{aligned} -\frac{1}{2}\left(|y_k|^2 + |y_l|^2 + |y_m|^2 + |y_j|^2\right) &= -\frac{1}{2(L^k)^2}|x - X^k|^2 - \frac{1}{2(L^l)^2}|x - X^l|^2 - \frac{1}{2(L^m)^2}|x - X^m|^2 \\ &\quad - \frac{1}{2(L^j)^2}|x - X^j|^2 \\ = -\frac{1}{2}\left(\frac{1}{(L^k)^2} + \frac{1}{(L^l)^2} + \frac{1}{(L^m)^2} + \frac{1}{(L^j)^2}\right)|x|^2 &+ \left(\frac{1}{(L^k)^2}X^k + \frac{1}{(L^l)^2}X^l + \frac{1}{(L^m)^2}X^m + \frac{1}{(L^j)^2}X^j\right) \\ - \frac{1}{2}\left(\frac{|X^k|^2}{(L^k)^2} + \frac{|X^l|^2}{(L^l)^2} + \frac{|X^m|^2}{(L^m)^2} + \frac{|X^j|^2}{(L^j)^2}\right), \end{aligned}$$

and

$$\begin{aligned} &(\Gamma^k + \Gamma^l - \Gamma^m - \Gamma^j) \\ &= \gamma^k + \beta^k \cdot (x - X^k) - \frac{B^k}{4(L^k)^2}|x - X^k|^2 + \gamma^l + \beta^l \cdot (x - X^l) - \frac{B^l}{4(L^l)^2}|x - X^l|^2 \\ &\quad - \gamma^m - \beta^m \cdot (x - X^m) + \frac{B^m}{4(L^m)^2}|x - X^m|^2 - \gamma^j - \beta^j \cdot (x - X^j) + \frac{B^j}{4(L^j)^2}|x - X^j|^2 \\ &= (\gamma^k + \gamma^l - \gamma^m - \gamma^j) + (\beta^j \cdot X^j + \beta^m \cdot X^m - \beta^l \cdot X^l - \beta^k \cdot X^k) \\ &\quad - \left(\frac{B^k}{4(L^k)^2}|X^k|^2 + \frac{B^l}{4(L^l)^2}|X^l|^2 - \frac{B^m}{4(L^m)^2}|X^m|^2 - \frac{B^j}{4(L^j)^2}|X^j|^2\right) \\ &\quad + x \cdot \left(\beta^k + \beta^l - \beta^m - \beta^j + \frac{B^k}{2(L^k)^2}X^k + \frac{B^l}{2(L^l)^2}X^l - \frac{B^m}{2(L^m)^2}X^m - \frac{B^j}{2(L^j)^2}X^j\right) \\ &\quad - \left(\frac{B^k}{4(L^k)^2} + \frac{B^l}{4(L^l)^2} - \frac{B^m}{4(L^m)^2} - \frac{B^j}{4(L^j)^2}\right)|x|^2 \end{aligned}$$

Define

$$\left|\begin{aligned}
C_{\mathrm{Im}} &:= \exp\left\{i\left(\gamma^k+\gamma^l-\gamma^m-\gamma^j\right)\right\}\\
&\quad\times\exp\left\{i\left(\beta^j\cdot X^j+\beta^m\cdot X^m-\beta^l\cdot X^l-\beta^k\cdot X^k\right)\right\}\\
&\quad\times\exp\left\{-i\left(\frac{B^k}{4(L^k)^2}|X^k|^2+\frac{B^l}{4(L^l)^2}|X^l|^2-\frac{B^m}{4(L^m)^2}|X^m|^2-\frac{B^j}{4(L^j)^2}|X^j|^2\right)\right\}\\
C_{\mathrm{Re}} &:= \exp\left\{-\frac{1}{2}\left(\frac{|X^k|^2}{(L^k)^2}+\frac{|X^l|^2}{(L^l)^2}+\frac{|X^m|^2}{(L^m)^2}+\frac{|X^j|^2}{(L^j)^2}\right)\right\}\\
C &:= \frac{A^kA^lA^m}{L^kL^lL^m}C_{\mathrm{Im}}C_{\mathrm{Re}}\\
\xi &:= -\left[\beta^k+\beta^l-\beta^m-\beta^j+\frac{B^k}{2(L^k)^2}X^k+\frac{B^l}{2(L^l)^2}X^l-\frac{B^m}{2(L^m)^2}X^m-\frac{B^j}{2(L^j)^2}X^j\right]\\
z &:= \frac{1}{2}\left(\frac{1}{(L^k)^2}+\frac{1}{(L^l)^2}+\frac{1}{(L^m)^2}+\frac{1}{(L^j)^2}\right)+i\left(\frac{B^k}{4(L^k)^2}+\frac{B^l}{4(L^l)^2}-\frac{B^m}{4(L^m)^2}-\frac{B^j}{4(L^j)^2}\right)\\
a &:= \frac{1}{(L^k)^2}X^k+\frac{1}{(L^l)^2}X^l+\frac{1}{(L^m)^2}X^m+\frac{1}{(L^j)^2}X^j
\end{aligned}\right.$$

and $f(x) := e^{-z|x|^2+a\cdot x}$. Then

$$\langle u|u|^2, b_{j,1}\rangle = \sum_{k,l,m} C\hat{f}(\xi). \tag{IV-6.2}$$

Computation of $\langle u|u|^2, b_{j,r+1}\rangle$, $r=1,\dots,d$

$$\begin{aligned}
&\langle u|u|^2, b_{j,r+1}\rangle\\
&= \sum_{k,l,m}\frac{A^kA^lA^m}{L^kL^lL^m}\left\langle e^{i\Gamma^k+i\Gamma^l-i\Gamma^m}e^{-\frac{|y_k|^2+|y_l|^2+|y_m|^2}{2}}, e^{i\Gamma^j}e^{-\frac{1}{2}|y_j|^2}\frac{x_r-X^j_r}{L^j}\right\rangle\\
&= \sum_{k,l,m}\frac{C}{L^j}\left(\widehat{x_rf}-X^j_r\hat{f}\right).
\end{aligned}$$

Computation of $\langle u|u|^2, b_{j,d+2}\rangle$

$$
\begin{aligned}
&\langle u|u|^2, b_{j,d+2}\rangle \\
&= \sum_{k,l,m} \frac{A^k A^l A^m}{L^k L^l L^m} \left\langle e^{i\Gamma^k + i\Gamma^l - i\Gamma^m} e^{-\frac{|y_k|^2+|y_l|^2+|y_m|^2}{2}}, e^{i\Gamma^j} e^{-\frac{1}{2}|y_j|^2} \left|\frac{x - X^j}{L^j}\right|^2 \right\rangle \\
&= \sum_{k,l,m} \frac{C}{(L^j)^2} \left(\widehat{|x|^2 f} - 2X^j \cdot \widehat{xf} + |X^j|^2 \hat{f} \right).
\end{aligned}
$$

IV-6.3 Conservative quantities in dimension $d = 2$

We provide in this section some miscellaneous computations, which hold in dimension $d = 2$ as long as $v_j(s_j, y_j) = e^{-\frac{|y_j|^2}{2}}$, $j = 1, \dots, N$. We give the explicit expressions for the conserved quantities involved in Lemma IV.2, in the two-dimensional case.

The $\mathbb{L}^2$ norm of a sum of N bubbles is given by

$$
\|u\|_{\mathbb{L}^2}^2 = \sum_{k,l=1}^{N} \frac{A^k A^l}{L^k L^l} \langle b_{k,1}, b_{l,1}\rangle.
$$

The energy of a sum of bubbles is given by

$$
E_{\mu,\lambda} = \frac{\mu}{2} \left\langle -\Delta u + |x|^2 u, u \right\rangle + \frac{\lambda}{4} \left\langle |u|^2 u, u \right\rangle = E_{\mu,0} + E_{0,\lambda} = \mu E_{1,0} + \lambda E_{0,1}.
$$

We have

$$
2E_{1,0} = \langle Hu, u\rangle = \langle -\Delta u, u\rangle + \langle |x|^2 u, u\rangle = \sum_{j,k=1}^{N} \langle \nabla_x u_j, \nabla_x u_k\rangle + \sum_{j,k=1}^{N} \langle |x|^2 u_j, u_k\rangle.
$$

Furthermore,

$$
\begin{aligned}
\langle \nabla_x u_j, \nabla_x u_k\rangle &= \frac{A^j A^k}{L^j L^k} \left\langle \left(i\beta^j - \frac{2 + iB^j}{2L^j} y_j \right) b_{j,1}, \left(i\beta^k - \frac{2 + iB^k}{2L^k} y_k \right) b_{k,1} \right\rangle \\
&= \frac{A^j A^k}{L^j L^k} \left\{ \beta^j \cdot \beta^k \left\langle b_{j,1}, b_{k,1}\right\rangle + i \frac{2 + iB^j}{2L^j} \beta^k \cdot \begin{pmatrix} \left\langle b_{j,2}, b_{k,1}\right\rangle \\ \left\langle b_{j,3}, b_{k,1}\right\rangle \end{pmatrix} \right. \\
&\qquad - i \frac{2 - iB^k}{2L^k} \beta^j \cdot \begin{pmatrix} \left\langle b_{j,1}, b_{k,2}\right\rangle \\ \left\langle b_{j,1}, b_{k,3}\right\rangle \end{pmatrix} \\
&\qquad \left. + \frac{2 + iB^j}{2L^j} \frac{2 - iB^k}{2L^k} \left(\left\langle b_{j,2}, b_{k,2}\right\rangle + \left\langle b_{j,3}, b_{k,3}\right\rangle \right) \right\},
\end{aligned}
$$

and

$$\begin{aligned}\langle |x|^2 u_j, u_k\rangle &= \frac{A^j A^k}{L^j L^k} \left\langle \left((L^j)^2 |y_j|^2 + 2L^j y_j \cdot X^j + |X^j|^2\right) b_{j,1}, b_{k,1}\right\rangle \\ &= \frac{A^j A^k}{L^j L^k} \left\{ (L^j)^2 \langle b_{j,4}, b_{k,1}\rangle + 2L^j X^j \cdot \begin{pmatrix} \langle b_{j,2}, b_{k,1}\rangle \\ \langle b_{j,3}, b_{k,1}\rangle \end{pmatrix} + |X^j|^2 \langle b_{j,1}, b_{k,1}\rangle \right\}.\end{aligned}$$

We also have

$$E_{0,1} = \langle u|u|^2, u\rangle = \sum_{j=1}^{N} \frac{A^j}{L^j} \langle u|u|^2, b_{j,1}\rangle.$$

We now proceed to computing the momentum, given by

$$M_{\mu,\lambda} = \left(E_{\mu,\lambda} - \mu \|xu\|_{\mathbb{L}^2}^2\right)^2 + \mu^2 \left(\operatorname{Im} \int x \cdot \nabla u \bar{u}\right)^2.$$

We know how to compute $E_{\mu,\lambda}$ from previously, as well as $\|xu\|_{\mathbb{L}^2}^2 = \langle |x|^2 u, u\rangle$. It only remains to compute

$$\begin{aligned}\int x \cdot \nabla u \bar{u} &= \sum_{j,k=1}^{N} \frac{A^j A^k}{L^j L^k} \left\langle (L^j y_j + X^j) \cdot \left(i\beta^j - \frac{2 + iB^j}{2L^j} y_j \right) b_{j,1}, b_{k,1} \right\rangle \\ &= \sum_{j,k=1}^{N} \frac{A^j A^k}{L^j L^k} \left\{ iL^j \beta^j \cdot \begin{pmatrix} \langle b_{j,2}, b_{k,1}\rangle \\ \langle b_{j,3}, b_{k,1}\rangle \end{pmatrix} - \frac{2 + iB^j}{2} \langle b_{j,4}, b_{k,1}\rangle \right. \\ &\qquad \left. + i\beta^j \cdot X^j \langle b_{j,1}, b_{k,1}\rangle - \frac{2 + iB^j}{2L^j} X^j \cdot \begin{pmatrix} \langle b_{j,2}, b_{k,1}\rangle \\ \langle b_{j,3}, b_{k,1}\rangle \end{pmatrix} \right\}.\end{aligned}$$

Note that all the inner products involved have already been computed when creating the matrix associated to the Dirac-Frenkel principle, see Section IV-4.2.1.

V

The spectral concentration problem

Part V

CHAPTER 1 Motivation

For once, mathematical work can be motivated by a real-life example 🎉

Imagine Alice wants to record a voice memo for her friend Bob. She will take out her smartphone or any voice recorder she has available, start the recording, and then talk. Once she is finished talking, she stops the recording.

The recorder's microphone has recorded a human voice for a few seconds, or minutes. The signal is virtually zero before she started talking, and zero after she finished. Any mathematician or physicist would approximate this signal using a compactly supported function, very originally named f. Moreover, the frequencies of a human voice typically range from a few tens to a few hundreds of Hertz. These frequencies correspond to the Fourier transform [1] of the function f.

The function f would then be a compactly supported function whose Fourier transform is also compactly supported. This means [2] that the function f has to be zero, meaning Alice actually produced no sound, but she did!

The Fourier transform is not suited to deal with this particular type of functions – compactly supported in both space and frequencies – even though it is a very powerful tool used every day to handle signals and telecommunications.

In the 1960s, D. Slepian, H. J. Landau, and H. O. Pollak asked something along the lines of: *A function cannot be compactly supported in both domains. What if we look for functions that are compactly supported in one domain, and among all of them take the one that is the «most concentrated» in the other domain?* They then answered in a very convincing and efficient manner to this problem in a series of five papers [46, 29, 30, 47, 45]. However, their framework is quite constraining and one may want to generalize those results.

It will be our aim in this Part of the manuscript to give some elements towards the

[1] Any person with a maths background will eventually ask "In what space lies f? In the Fourier transform well-defined on this space?" These are legitimate questions, but we are also only in the introductory example...

[2] [22, Theorem 7.1.14]: The Fourier transform $\hat{f}$ of a distribution $f \in \mathcal{D}'(\mathbb{R}^d)$ with compact support is an entire analytic function over $\mathbb{C}^d$. Moreover, [41, Proposition 10.23]: Every bounded entire function is constant. If $\hat{f}$, an entire function, was compactly supported it would be bounded and hence constant. Moreover this constant would be zero: $\hat{f}$ being compactly supported means $\hat{f}(z) = 0$ for $|z| > R$, $R > 0$ large enough, and $\hat{f}$ being constant means in particular $\hat{f} \equiv \hat{f}(2R) = 0$.

generalization of the problem and its answer.

David Slepian (1923–2007). Credit to itsoc.org.

In Chapter V-2, we start by giving some more details and a precise formulation of the *Spectral Concentration Problem.* We will then explain the "very convincing and efficient" solution proposed by Slepian *et al.* After this, we give an overview of the literature treating the spectral concentration problem. We will see that most works remain very close to the framework Slepian *et al.* studied, and even by taking this into account, the problem is not well understood nor is Slepian's answer. It looks like the elegant answer he proposed is just an accident, which could not have been reproduced in other situations. We will also discuss why the problem is difficult to solve from the numerical point of view, and more specifically why a straightforward, direct, brute-force solution is not satisfying.

Chapter V-3 is dedicated to presenting very briefly the Prolate Spheroidal Wave Functions, which are essential to the elegant solution proposed by Slepian.

We then proceed in Chapter V-4 to propose a generalized formulation of the spectral concentration problem. In particular, we will see that the problem Slepian *et al.* considered can be recovered from this generalized problem. We give some basic properties about this situation, and then derive the discrete formulation of the problem. We will see that the discrete problem as considered by Slepian in [45] is recovered by our discrete framework.

Now that we have set up our theoretical foundations, we can try to solve the spectral concentration problem in Chapter V-5. We were not able to do much from the theoretical point of view, but we were able to find an algorithm that bypasses the limitations one usually encounters in the discrete setting. Even though we do not recover exactly the same eigenvectors as those of Slepian et al., they are qualitatively satisfying, obtained in a deterministic way, and most importantly they are guaranteed to be a good approximation of the true solutions we were looking for. We then proceed to solving the spectral concentration problems in previously unstudied situations. To the author's knowledge, the numerical procedure proposed is the only existing way to solve the generalized spectral concentration problem without suffering from the issues we will have discussed.

Remark V.1

The algorithm is devised here in the context of the spectral concentration problem, but one could see it in a more general way: given a matrix with simple eigenvalues very close together, how to recover the associated eigenvectors without confusing them numerically? The confusion is due to the eigenvalues being too close to each other, we'll explain that in more details later. One possible answer to this question is the procedure we will describe in Chapter V-5, which allows one to recover approximately the eigenvectors with no confusion.

Review of the spectral concentration problem

Part V

2 CHAPTER

V-2.1 The Slepian "toy model"

The (continuous) Slepian Concentration Problem, as it was first formulated in [46], consists in looking for functions $f \in \mathbb{L}^2(\mathbb{R})$, $\|f\|_{\mathbb{L}^2(\mathbb{R})} = 1$, having a Fourier transform with compact support in $[-\Omega, \Omega]$, which have the largest $\mathbb{L}^2([-T,T])$ norm, for some $T, \Omega > 0$.

This simultaneous space-frequency localization problem is obviously linked to Heisenberg's Uncertainty principle, and the link is discussed in [29].

A function $f \in \mathbb{L}^2(\mathbb{R})$ having the support of its Fourier transform $\hat{f}$ in $[-\Omega, \Omega]$ can be written

$$f(x) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \hat{f}(\xi) e^{i\xi \cdot x} d\xi.$$

This simply comes from the usual inverse Fourier transform, and using the compact support of $\hat{f}$: $\hat{f}_{|(-\infty,-\Omega)\cup(\Omega,+\infty)} = 0$ almost everywhere. The $\mathbb{L}^2([-T,T])$ norm of such function is then given as

$$\begin{aligned}
\|f\|^2_{\mathbb{L}^2([-T,T])} &= \frac{1}{4\pi^2} \int_{-T}^{T} \left(\int_{-\Omega}^{\Omega} \hat{f}(\xi) e^{i\xi \cdot x} d\xi \right) \overline{\left(\int_{-\Omega}^{\Omega} \hat{f}(\eta) e^{i\eta \cdot x} d\eta \right)} dx \\
&= \frac{1}{4\pi^2} \int_{-\Omega}^{\Omega} \int_{-\Omega}^{\Omega} \hat{f}(\xi) \overline{\hat{f}(\eta)} \left(\int_{-T}^{T} e^{i(\xi-\eta)x} dx \right) d\eta d\xi \\
&= \frac{1}{4\pi^2} \int_{-\Omega}^{\Omega} \int_{-\Omega}^{\Omega} \hat{f}(\xi) \overline{\hat{f}(\eta)} \frac{e^{iT(\xi-\eta)} - e^{-iT(\xi-\eta)}}{i(\xi-\eta)} d\eta d\xi \\
&= \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \int_{-\Omega}^{\Omega} \hat{f}(\xi) \overline{\hat{f}(\eta)} \frac{\sin(T(\xi-\eta))}{\pi(\xi-\eta)} d\eta d\xi \\
&= \frac{\Omega^2}{2\pi} \int_{-1}^{1} \int_{-1}^{1} \hat{f}(\xi) \overline{\hat{f}(\eta)} \frac{\sin(\Omega T(\xi-\eta))}{\pi(\xi-\eta)} d\eta d\xi. \qquad \text{(V-2.1)}
\end{aligned}$$

Finally, the functions $f \in \mathbb{L}^2(\mathbb{R})$ we are looking for must maximize the quantity given by (V-2.1). It can be shown [1] that the $\mathbb{L}^2(\mathbb{R})$ function of norm 1 maximizing (V-2.1) is an

[1] The derivation for the general case will be done later in Section V-4 – Generalized framework.

eigenfunction ψ_1 of the operator

$$g \mapsto (\mathcal{K}g)(x) := \int_{-1}^{1} g(y)\frac{\sin\left(\Omega T(y-x)\right)}{\pi(y-x)}dy, \tag{V-2.2}$$

associated to the eigenvalue λ_1 of largest magnitude. Moreover, there exists [2] a countable infinity of eigenpairs (λ_i, ψ_i) with real eigenvalues for the operator $\mathcal{K}$, and one can order them according to the magnitude of eigenvalues:

$$1 > \lambda_1 > \lambda_2 > \cdots > \lambda_k > \cdots > 0$$

Remark V.2: Vocabulary

The operator $\mathcal{K}$ from (V-2.2), and later its generalization (V-4.5), will be called the *concentration operator*. The eigenvalues λ_i are called the *concentration ratios*, since they measure how well the eigenvector ψ_i is concentrated in both Fourier and space domains. The name “ratio” is inspired from [46, p. 55], where the highest eigenvalue λ_1 was defined as the ratio of the $\mathbb{L}^2$ norm of the concentrated function and the $\mathbb{L}^2$ norm of the unrestricted function:

$$\lambda_1 := \frac{\|\mathcal{K}g\|_{\mathbb{L}^2([-T,T])}}{\|g\|_{\mathbb{L}^2(\mathbb{R})}}.$$

It is actually a consequence of λ_1 being an eigenvalue of the concentration operator $\mathcal{K}$. Finally, we may refer to the eigenfunctions ψ_i as *Slepian modes*.

It happens that approximately the first $\lfloor \frac{2}{\pi}\Omega T \rfloor + 1$ eigenvalues are all close to one, followed by a few eigenvalues which are far from one and zero, and then there are infinitely many eigenvalues close to zero. This phenomenon was first studied in [30]. The number of eigenvalues between zero and one in the discrete setting (i.e. $[-T, T]$ is discretized using N uniformly spaced points) has also been studied, and [24] is the first work to have derived a bound that depends on the product $N\Omega$.

The behavior of the eigenvalues is *very* interesting from the applied mathematical point of view: indeed, if one decides to use the so-called Slepian basis $\{\psi_i\}_{i\geq 1}$ made of the eigenfunctions of $\mathcal{K}$, then in practice only the eigenfunctions corresponding to eigenvalues far from zero are useful. The eigenvalues close to zero carry almost no information. It gives a natural threshold on where to truncate the Slepian basis, and this is very different from other usual basis (e.g. Hermite or Fourier). This is one of the reasons why the Slepian

[2] Also shown in the generalized setting.

basis is interesting, and probably why it deserves some attention.

We will refer to the above case as the "toy model". This is not to undermine of the work of D. Slepian, H. J. Landau and H. O. Pollak, but only to convey the fact that this problem is known relatively well, and that this is the problem upon which we shall build the generalized theory in Section V-4 – Generalized framework.

V-2.2 Known results

The toy model was studied in detail by D. Slepian and H. O. Pollak in [46], who, in particular, showed that the eigenfunctions of $\mathscr{K}$ are orthogonal and complete in both spaces $\mathbb{L}^2(\mathbb{R})$ and $\mathbb{L}^2([-T,T])$. This is also the first account of the commuting property between a certain differential operator $\mathscr{P}$ and the kernel operator $\mathscr{K}$. More specifically, it was shown that the differential operator

$$\mathscr{P}(x,\partial_x) := \frac{d}{dx}\left(1-x^2\right)\frac{d}{dx} - c^2x^2 \tag{V-2.3}$$

commutes with the integral operator $\mathscr{K}$ defined in (V-2.2), where c is a multiple of the product ΩT. The eigenfunctions of $\mathscr{P}$ were known before D. Slepian realized this connection, and are called *Prolate Spheroidal Wave Functions*[3]. They are studied for instance in [11, 26, 35, 32, 50], and as early as 1880 [33]. We present them briefly in Chapter V-3 – Prolate spheroidal wave functions.

The interesting fact about this commutation is that it allows one to look for eigenfunctions of $\mathscr{K}$ by looking for eigenfunctions of $\mathscr{P}$:

Lemma V.1

Let $A, B : \mathbb{L}^2(\mathbb{R}^d) \to \mathbb{L}^2(\mathbb{R}^d)$ two commuting operators acting on $\mathbb{L}^2(\mathbb{R}^d)$, that is $AB = BA$. Suppose that each eigenfunction φ_i of A is associated to an eigenvalue κ_i of multiplicity one, and that Span $\{\varphi_i : i \in \mathbb{N}\} = \mathbb{L}^2(\mathbb{R}^d)$. Then A and B have the same eigenfunctions.

Proof. Using the commutation relation between A and B, one obtains

$$A\varphi_i = \kappa_i\varphi_i \implies BA\varphi_i = \kappa_i B\varphi_i \implies AB\varphi_i = \kappa_i B\varphi_i.$$

This means that $B\varphi_i$ also is an eigenfunction of A, and since κ_i is an eigenvalue of

[3] Not to be confused with "prolapse", which is a totally different subject... 🤡

multiplicity one we must have $B\varphi_i = c\varphi_i$ for some constant $c \in \mathbb{C}$. In other words, all eigenfunctions of A are eigenfunctions of B. Moreover, since the eigenfunctions of A form a complete family of $\mathbb{L}^2(\mathbb{R}^d)$, we deduce that the eigenfunctions of B are exactly the eigenfunctions of A. □

Then, [30] quickly followed, which showed how well an arbitrary $\mathbb{L}^2(\mathbb{R})$ function can be approximated using the first $\lfloor \frac{2}{\pi}\Omega T \rfloor$ eigenfunctions of $\mathcal{K}$, and in what sense this basis is better than "sampling functions" of the form $t \mapsto \operatorname{sinc}(2Wt - r)$.

Thus far, Slepian, Landau and Pollak studied the continuous, one-dimensional case [46, 29, 30]. A main component for the study of so-called Slepian modes – the eigenfunctions of the integral operator $\mathcal{K}$ – is the commutation property with a differential operator, which is *a priori* unexpected.

> There was a lot of serendipity here, clearly. And then our solution, too, seemed to hinge on a lucky accident [...]
>
> D. Slepian (1983)

This "lucky accident" is precisely the commutation property.

The continuous 2-dimensional case was then studied in [47], and it was shown that there also exists a differential operator commuting with the multi-dimensionl counterpart of the integral operator. For this 2-dimensional case, the Fourier domain is restricted to $B(0,1)$ and the space domain is restricted to $B(0,c)$ for $c > 0$. We denote by $B(x_0, r)$ the 2-dimensional ball centered at $x_0 \in \mathbb{R}^d$ of radius $r > 0$. The multidimensional counterpart of the operator $\mathcal{P}$ writes

$$\mathcal{P}_{\dim 2}(r, \partial_r) = \frac{d}{dr}\left(1 - r^2\right)\frac{d}{dr} - \left(c^2 r^2 + \frac{\frac{1}{4} - N^2}{r^2}\right), \tag{V-2.4}$$

and applies to radial functions. The quantity N denotes an integer, and for each N there exists an orthonormal family of eigenfunctions of the concentration operator. In polar coordinates, they write

$$\left\{R_{N,n}(r)\cos(N\theta)\right\}_{n\in\mathbb{N}} \cup \left\{R_{N,n}(r)\sin(N\theta)\right\}_{n\in\mathbb{N}},$$

where $R_{N,n}$ is the n-th eigenfunction of $\mathcal{P}_{\dim 2}$. The work [47] is achieved by treating the general d-dimensional case in a very similar fashion (just replace N by $N + \frac{d-2}{2}$ in

(V-2.4)).

Finally, the Slepian-Landau-Pollak series of five papers was achieved a few years later with [45], which treated the discrete-space continuous-Fourier case. In this situation, there exists as well a second-order differential operator which commutes with the discrete-space continuous-Fourier operator.

However, all the studies so far used the integral kernel $\mathscr{K}$ given in (V-2.2), or its higher-dimensional counterpart. There is a lot of symmetry involved in considering unit balls in both space and Fourier domains, and one can wonder if such commutation property holds with other domains of restrictions[4]. In Section V-2.2.4 – Alternative way to recover the differential operator, we give a new way to obtain the commuting differential operator in the cases for which it is already known. Unfortunately, neither old nor new approaches seem to be applicable to other geometries, and the profound reason why it does not work remains mysterious.

Of course, other authors have worked on this subject since the 1960s. We give below a few references on generalizations to the Spectral Concentration Problem, but few of them treats our main concern, which are the eigenfunctions in the generalized case, with the generalized kernel given by (V-4.5). The work closest to our concerns is by Grünbaum [14], but unfortunately he was unable to exhibit an elegant[5] solution like in the toy model. Among all works we are aware of, perhaps the result furthest away from Slepian's framework but with an equally interesting solution is due to Brander and DeFacio [4], who studied the case where the space and Fourier restrictions are Gaussian functions instead of indicator functions.

In addition to its very interesting mathematical study, the spectral concentration problem has been used in several other fields of science. We refer to [44] and the references therein for a large variety of examples of applications.

V-2.2.1 One-dimensional works

One of the first attemps at extending the results from Slepian to other examples in 1D is due to Grünbaum [15]. It is shown that, if $k = \mathscr{F}^{-1}[\mu]$, with μ an even Lebesgue integrable nonnegative function with compact support, then second- or fourth-order differential operators can be found to commute with the integral operator

$$(\mathscr{K}f)(x) = \int_{-1}^{1} k(x-y)f(y)dy,$$

[4] When $d = 2$, the case of a disc in space and Fourier was studied, but what if we wanted to restrict the function in space to a cat-head shape and in Fourier to a duck-head shape?

[5] *Elegant* means here the existence of a commuting differential operator.

only if $k(x) = \text{sinc}\ (\Omega x)$, up to some multiplicative constant. In other words, the Fourier restriction can only be (up to a multiplicative constant) the indicator function of an interval if one wants to find a commuting operator. The assumptions on μ are stronger than those from a previous, unpublished result by Morrison, cited in [52] and [13, p. 119]:

Theorem V.1: Morrison (1962)

For any constants b and c, the eigenfunctions of the integral operator on $\mathbb{L}^2(-1,1)$ with kernel

$$\frac{b \sin c(x-y)}{c \sinh b(x-y)}$$

are the eigenfunctions of the differential operator

$$\frac{d}{dx}\left(1 - \frac{\sinh^2 bx}{\sinh^2 b}\right)\frac{d}{dx} - (b^2+c^2)\frac{\sinh^2 bx}{\sinh^2 b},$$

where the eigenfunctions are required to be continuous at $x = \pm 1$.

The characterization of the "admissible pairs" of integral and differential operators has recently been extended to the complex case in [12]. They obtain the following result:

Theorem V.2: Grabovsky, Hovspeyan (2021)

Let K and L be given by

$$(Ku)(x) = \int_{-1}^{1} k(x-y)u(y)dy, \quad Lu = -au''+bu'+cu, \quad a(\pm 1) = 0, b(\pm 1) = a'(\pm 1),$$

with a, b, c smooth in $[-2,2]$. Assume k is smooth in $[-2,2] \setminus \{0\}$ and either:

1. k is analytic at 0, not identically zero near 0, and cannot be written as a finite linear combination of exponentials $e^{\alpha z}$ or be written under the form $e^{\alpha z}p(z)$, with p polynomial;
2. k has a simple pole at 0.

If $KL = LK$, then

$$k(z) = \frac{\lambda}{\sinh\left(\frac{\lambda}{2}z\right)}\left(\alpha_1 \frac{\sinh(\mu z)}{\mu} + \alpha_2 \cosh(\mu z)\right).$$

and

$$\begin{cases} a(y) = \dfrac{1}{\lambda^2}\left[\cosh(\lambda y) - \cosh(\lambda)\right], \\ b(y) = a'(y), \\ c(y) = \left(\dfrac{\lambda^2}{4} - \mu^2\right) a(y). \end{cases}$$

These results were obtained by Taylor expansion of the kernel k and brute-force exploitation of the commutation relation.

Note that all the above cited works are concerned with an indicator function in space, and let the kernel k free. If one decides forget about the space indicator function, no result is known.

Furthermore, these results tell us what kernel k can be used if one wants to find a commuting differential operator, but we are not able to choose k. This goes against the idea that one restricts in both space and Fourier domains as they want, and only then looks for eigenvectors.

We also refer to the work of Papoulis [36] for band-filtering methods and applications of the Prolate Spheroidal Wave Functions. Reconstruction of a function given its values on a compact domain is also discussed by Grünbaum in [14].

V-2.2.2 Multi-dimensional works

One of the first attempts at finding a commuting differential operator for a domain of restriction other than the unit ball is due to Grünbaum *et al.* [19]. They are able to recover the d-dimensional toy model commuting differential operator, and expect the existence of a commuting differential operator to fail for other geometries, like the torus for example.

In [4], the case of Gaussian filters in both space and Fourier domains is treated, and the following result is obtained:

Theorem V.3: Brander, DeFacio (1986)

Let Q a real, piecewise twice continuously differentiable function that either has finite support or goes to zero at infinity sufficiently fast, and such that $Q(x) = Q(-x)$ for $x \in \mathbb{R}^d$. Let K the integral operator defined by

$$(Kf)(x) := \int_{\mathbb{R}^d} Q(x) e^{-icx\cdot y} Q(y) f(y) dy,$$

and D the differential operator defined by

$$D := -\nabla \cdot (\alpha(x)\nabla) + U(x).$$

Assume also that α and Q are angle independent and twice differentiable, that is $\alpha, Q \in C^2(\mathbb{R}_+)$.
Then, the most general function α for which K and D commute is given by

$$\alpha(x) = a + b|x|^2,$$

where a and b are arbitrary constants, $a \neq 0$. The corresponding function Q is given by

$$Q(x) = \frac{1}{\sqrt{\alpha(x)}} \exp\left(-\gamma \int_0^x \frac{u}{\alpha(u)} du\right),$$

with another arbitrary constant γ, $\gamma \geq |b|$.

Again, this states that there exists some kind of compatibility required between the space and Fourier restrictions.

Of utter importance is the work of Grünbaum [17], which shows that, in dimension $d = 2$, the search for a commuting differential operator may be vain. He starts by restricting the Fourier domain in a ball centered at origin of finite radius, and restricts the space variable to a two-dimensional torus (i.e. a square with periodic boundary conditions). It is then shown that the only second-order differential operator which can commute with the integral operator is actually a scalar matrix, which commutes trivially. This is achieved, again, by writing out explicitely the commutation relation, and finding explicit conditions that are or are not satisfied.

In a recent review, Wang [51] gives a large number of references related to the spectral concentration problem or Prolate Spheroidal Wave Functions (PSWF):

[10] an analoguous to the PSWF is derived when looking for them under polynomial form. The study is done for the unit ball in the one- and three-dimensional cases.

[43] presents an analoguous to the Shannon number, which is roughly the number of "useful" Slepian modes. It also gives a detailed study of the computation of Slepian function on the sphere, with an arbitrary region of interest, or with an axisymmetric polar cap. It relies heavily on the spherical geometry, by using spherical harmonics.

[40, 39] the first work presents a wavelet version of the Slepian functions on the sphere, for incomplete data reconstruction. It computes the Slepian functions on the sphere using the method described in [43]. The second work does essentially the

same study when restricting to a manifold instead of the sphere. In this case, the Slepian eigenvectors are obtained by solving a matrix eigenproblem. Even though the authors do not mention it, their results seem to suffer from the same "eigenvalue clustering" issue detailed later in Section V-2.2.3: the manifold on which they compute the Slepian function is symmetric but their Slepian functions are not symmetric, which indicates that they might be linear combinations of all the relevant eigenvectors.

[44] studies Slepian function on a disc. The eigenvalues are distinct enough so that the eigenvectors can be well recovered. The same remark can be done with their arbitrary two-dimensional shape, where all eigenvalues are distinct. It has to be noted though that their arbitrary two-dimensional shape is perhaps the framework closest to ours, since it does not rely at all on Slepian's ideas and onyl uses numerical integration.

[31] studies the case of the unit ball with a polar cap. The author says that "attempts to compute the eigenfunctions of the integral operator $\mathbf{K}F_n(u) = \mu_n F_n(u)$ directly have not been fruitful", then compares the integral discretization with the commuting differential operator from [19].

In [7], another approach to time-frequency localization is presented, and it is based on so-called *coherent states*. This is a direction which we will not explore here.

The only work we are aware of that treats the case of completely arbitrary space and Fourier restrictions is due to Simons and Wang [44], but their brief study of the generalized situation is only numerical. Moreover, the parameters in their numerical experiments are such that there is no confusion possible between the eigenvalues, which is an ideal situation not always occuring.

Numerical works

After realizing the commutation property, D. Slepian *et al.* put it to good use by describing a method for computing numerically the eigenvectors of the discretized concentration operator $\mathscr{K}$ from (V-2.2). Once again, the method used in the discretized case relies heavily on the explicit expression for a commuting differential operator.

Ever since, all efficient computations of the eigenvectors of (the discretized version of) $\mathscr{K}$ have used the differential operator $\mathscr{P}$ from (V-2.3). Among many other works, we can cite [45, 16], and [37, Section 8.3].

These efficient algorithms are efficient in two senses: first, the matrix representing the differential operator is tridiagonal, thus it can be manipulated and diagonalized quickly. Second, its eigenvalues are distinct and the eigenvectors can be obtained easily[6]. The

[6] We will show later that this second fact fails when diagonalizing the full concentration matrix, and

reason why an eigendecomposition is never applied to the discretized version of the concentration operator $\mathscr{K}$ is explained in Section V-2.2.3 – Numerical difficulties.

In [25], Karnik *et al.* propose a Fast Slepian Transform, i.e. an efficient way of projecting onto the Slepian basis. Their Slepian modes are, however, only in 1D and treat the Slepian toy model.

The fast computation of Slepian modes may become one day important, for the reason underlined by Boyd in [34]:

> Because the prolate functions are orthogonal with the weight function of unity, just like Legendre polynomials, the prolate functions are the basis that is "plug-and-play" compatible with finite elements or spectral element or other programs that employ Legendre polynomials. The claimed advantage of prolate functions is that they can resolve wavy, bandlimited signals with only two points per wavelength, whereas Legendre polynomials and Chebyshev polynomials require a minimum of π degrees of freedom per wavelength.
>
> J. P. Boyd (2013)

One of the most recent works concerning the study of eigenvalues is [24], where they study the eigenvalues of the discretized operator, i.e. of the band and time limiting matrix $\mathbf{K}$, for the toy model. In order to obtain a bound on the number of eigenvalues λ_k such that $\varepsilon < 1-\lambda_k < 1-\varepsilon$ for $\varepsilon \in (0,1)$, they show that the matrix $\mathbf{K}-\mathbf{K}^2$ has low numerical rank. This means that few eigenvalues of $\mathbf{K}-\mathbf{K}^2$ are far from zero, which in turn means that $\mathbf{K}$ has few eigenvalues far away from zero or one.

Remark V.3

There exist some other generalizations of the spectral concentration problem (e.g. [8]), but we will not focus on them because they do not fit into the generalized framework we will describe in Chapter V-4.

special attention has to be paid when designing numerical algorithms for the full concentration matrix.

V-2.2.3 Numerical difficulties

To us, the quantities of interest are the eigenfunctions of the concentration operator $\mathcal{K}$, defined either by (V-2.2) for the toy model, or by (V-4.5) for the more general case. Thus we are interested in cheap, fast, and precise algorithms to obtain them.

For the toy model (V-2.2), this "graal" has been obtained thanks to the "lucky accident", i.e. the commutation property with the differential operator $\mathcal{P}$ from (V-2.3).

As we mentioned previously, Grünbaum [14] studied a generalization of the Slepian concentration problem that is of interest to us. Unable to find a relation similar to the commutation relation, the author had to resort to "full-matrix" algorithms, and so do we. This work was however interested mostly in the behavior of eigenvalues rather than eigenvectors, and the main difficulty we face when diagonalizing the full matrix is not studied. One issue that [14] did not explain, is that the computation of eigenvectors needs much more precision than the computation of eigenvalues. This is purely a "numerical linear algebra" issue, let us explain why.

It is linked to the particular behavior of eigenvalues that we mentioned earlier: some of them are close to 1, the others are close to 0. By "close", we mean that the difference between two successive eigenvalues can be only a few orders of magnitude greater than the machine precision[7]. In such situations, numerical algorithms used for the eigendecomposition of a matrix may consider that the eigenvalues close to 1 are not several eigenvalues close to others, but only a multiple eigenvalue. This is actually the behavior observed on the one-dimensional toy model. Our main problem is that any linear combination of eigenvectors associated to the same eigenvalue is also an eigenvector, thus two eigenalgorithms may yield very different results while they theoretically should be the same (because theoretically, the eigenvalues are all distinct).

Numerically, we are in the following situation:

Lemma V.2

Let $\mathbf{A}$ a $n \times n$ matrix, with an eigenvalue λ of multiplicity $m \leq n$. Let $u_1, \dots, u_m$, m independant eigenvectors of $\mathbf{A}$ associated to the eigenvalue λ. Then any linear combination of $u_1, \dots, u_m$ is also an eigenvector of $\mathbf{A}$ associated to λ.

Proof. Let $c_1, \dots, c_m \in \mathbb{C}$,

$$\mathbf{A}\left(\sum_{i=1}^{m} c_i u_i\right) = \sum_{i=1}^{m} c_i \mathbf{A} u_i = \sum_{i=1}^{m} c_i \lambda u_i = \lambda \left(\sum_{i=1}^{m} c_i u_i\right).$$

[7] The meaning of "machine precision" has been explained in Section II-1 – Representation of continuous problems on a computer.

□

We display in Figure V-2.1a the eigenvector of $\mathbf{K}$ associated to the largest eigenvalue λ_1, obtained using three different methods. We recall that $\mathbf{K}$ is the discretized version of the operator $\mathscr{K}$ defined in (V-2.2), using N discretization points spread uniformly. Figure V-2.1b shows how close the first eigenvalues are to each other. The first method [8] used to compute the eigenvector (solid blue curve) is called `eig`, and corresponds to the `_geev` LAPACK routine. It can be used for general square matrices. The second method [9] used to compute the eigenvector (dash orange curve) is called `eigh`, and corresponds to the `_heevd` LAPACK routine. It specializes to symmetric or Hermitian matrices. The last method [10] (dot-dash green curve) is the tri-diagonal formulation obtained from the work of Slepian, see the above Section V-2.2.2 – Numerical works. The two LAPACK routines used are described in the following paragraph.

The question of the precision of the computation of eigenvalues has been discussed in [37, Section 8.1], and they also show that a lack of precision yields the wrong eigenvectors.

Used LAPACK routines They are detailed in [1]. The first routine we describe is `_geev`. It performs an eigendecomposition on a general square matrix A, using the following steps (see [1, Section 2.4.5]):

1. Reduce A to upper Hessenberg form:

$$A = QHQ^H,$$

   where Q is unitary and H is zero below its first subdiagonal.

2. Reduce the upper Hessenberg matrix H to Schur form:

$$H = STS^H,$$

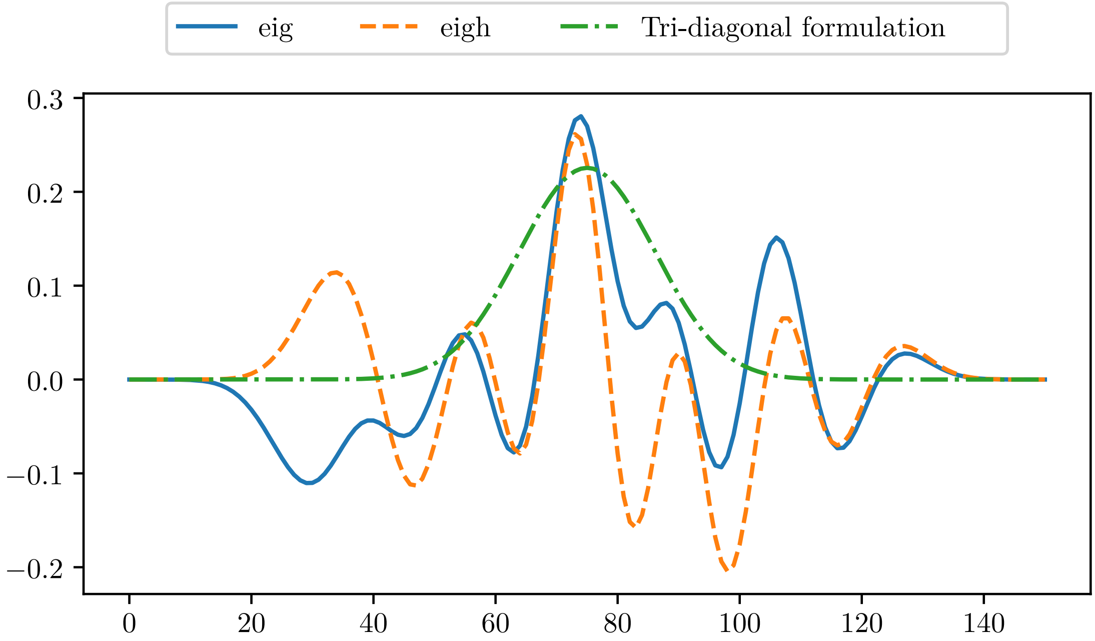
   where S is an orthonormal matrix and T is upper triangular. The eigenvalues of A are given in the diagonal of T, and once they are known one can obtain the eigenvectors. For instance, one can use inverse interation to obtain the eigenvectors from H and then multiply them with Q.

The second routine used is `_heevd`, and it specializes to Hermitian (or symmetric) matrices. The steps are as follows (see [1, Section 2.4.4]):

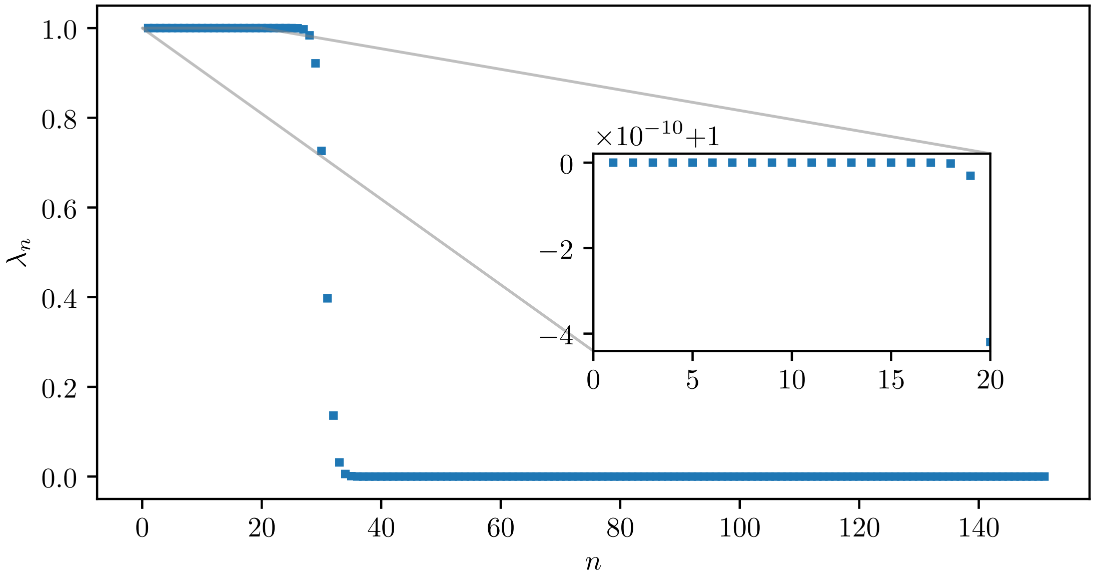
[8] https://numpy.org/doc/stable/reference/generated/numpy.linalg.eig.html

[9] https://numpy.org/doc/stable/reference/generated/numpy.linalg.eigh.html

[10] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.windows.dpss.html



(a) Eigenvector associated to the largest eigenvalue λ_1, using three different algorithms.



(b) All eigenvalues.

Figure V-2.1 – $N = 151$, $\Omega = 0.1 \cdot 2\pi$.

1. Reduce A to real tridiagonal form:

$$A = QTQ^H,$$

   where Q is unitary and T real symmetric tridiagonal.

2. Compute eigenvalues and eigenvectors of the real symmetric tridiagonal matrix T:

$$T = S\Lambda S^T,$$

   where S is orthogonal and Λ diagonal. The desired eigenvectors are then QS.

V-2.2.4 Alternative way to recover the differential operator

The commuting property of the integral operator $\mathcal{K}$ (given in (V-2.2)) and the differential operator $\mathcal{P}$ (given in (V-2.3)) was stated by Slepian. Actually, he just checked that both operators commute, and did not explain where the differential operator came from. It was found again later with brute-force computations, we refer to Section V-2 – Review of the spectral concentration problem for relevant works.

In this section, we give an new way of obtaining the commuting differential operator from $\mathcal{K}$. The ideas used were designed in order to get some insight on how to generalize the commuting property for more general integral kernels, and more specifically the generalized concentration kernels $\mathcal{K}$ defined by (V-4.5).

It consists in assuming that the commuting differential operator is of order 2, and to use basic Fourier relations: multiplication by x in x-space is differentiation in Fourier, and vice-versa.

Pseudodifferential approach In this subsection, we assume that the space domain is restricted to $D_1 \subset \mathbb{R}^d$ and the Fourier domain is restricted to $D_2 \subset \mathbb{R}^d$. It is the analoguous situation to the one-dimensional toy model, for which $D_1 = [-T, T]$ and $D_2 = [-\Omega, \Omega]$. We are looking for conditions on D_1 and D_2 such that there exists a (self-adjoint) differential operator $\mathcal{P}$ commuting with the integral operator $\mathcal{K}$ defined by

$$(\mathcal{K}f)(x) := \int_{D_1} f(y) \int_{D_2} e^{i\xi\cdot(x-y)} d\xi dy.$$

We will restrict ourselves to self-adjoint second-order differential operators of the form:

$$\mathcal{P}(x, \partial_x) = \operatorname{div}(\mathbf{A}(x)\nabla) + C(x),$$

where $\mathbf{A}$ is a matrix such that

$$\mathbf{A}(x) = 0 \quad \text{on} \quad \partial D_1.$$

Because of this boundary condition, the commutation relation is equivalent to

$$\mathcal{P}(x, \partial_x)K(x,y) = \mathcal{P}(y, \partial_y)K(x,y).$$

We write $\varphi(x,\xi) = e^{ix\cdot\xi}$. Using the fact that

$$\partial_x \varphi(x,\xi) = i\xi\varphi(x,\xi) \quad \text{and} \quad x\varphi(x,\xi) = -i\partial_\xi \varphi(x,\xi), \tag{V-2.5}$$

the previous relation writes

$$\begin{aligned}\int_{D_2} \mathcal{P}(-i\partial_\xi, i\xi)\varphi(x,\xi)\overline{\varphi(y,\xi)}\mathrm{d}\xi &= \int_{D_2} \varphi(x,\xi)\mathcal{P}(i\partial_\xi, -i\xi)\overline{\varphi(y,\xi)}\mathrm{d}\xi \\ &= \int_{D_2} \varphi(x,\xi)\overline{\mathcal{P}(-i\partial_\xi, i\xi)\varphi(y,\xi)}\mathrm{d}\xi.\end{aligned}$$

Thus, the commutation relation is equivalent to

$$\langle \mathcal{P}(-i\partial_\xi, i\xi)f, g\rangle_{L^2(D_2)} = \langle f, \mathcal{P}(-i\partial_\xi, i\xi)g\rangle_{L^2(D_2)},$$

for the Hermitian scalar product.

Let us now show using the pseudodifferential approach that, in the two-dimensional case, the commutation results obtained by Slepian when D_1 and D_2 are both the unit ball, can be generalized to the case of ellipses. Suppose $D_1 = \mathit{Ellipse}(0,a,b)$ and $D_2 = \mathit{Ellipse}(0,k_1,k_2)$, where

$$\mathit{Ellipse}(0,a,b) := \left\{x \in \mathbb{R}^2 : ax_1^2 + bx_2^2 \leq 1\right\}, \quad a,b \geq 0.$$

Since we require $\mathbf{A}$ to vanish on ∂D_1, we choose

$$\mathbf{A}(x) = \begin{pmatrix} k_1(ax_1^2 + bx_2^2 - 1) & 0 \\ 0 & k_2(ax_1^2 + bx_2^2 - 1)\end{pmatrix}.$$

Thus, using relations (V-2.5),

$$\mathcal{P}(-i\partial_\xi, i\xi) = \xi^T \mathbf{A}(-i\partial_\xi)\xi + C(-i\partial_\xi) = \sum_{j=1}^{2} k_j \xi_j (a\partial_{\xi_1}^2 + b\partial_{\xi_2}^2 + 1)\xi_j + C(-i\partial_\xi).$$

Remark V.4

We emphasize the fact that $\partial_{\xi_j}\xi_j$ is actually the operator given for $h \in C^\infty(\mathbb{R}^d)$ by:

$$\partial_{\xi_j}\xi_j h = \partial_{\xi_j}(\xi_j h).$$

In particular, we have

$$\partial_{\xi_j}\xi_j - \xi_j\partial_{\xi_j} = 1, \tag{V-2.6}$$

since

$$\partial_{\xi_j}(\xi_j h) = h + \xi_j\partial_{\xi_j} h \quad \Longleftrightarrow \quad \left(\partial_{\xi_j}\xi_j - \xi_j\partial_{\xi_j}\right) h = h.$$

Now we use the following relation:

$$\nabla_\xi(\xi_j f(\xi)) = \xi_j \nabla_\xi f + e_j f(\xi),$$

where $e_1 = \begin{pmatrix}1\\0\end{pmatrix}$ and $e_2 = \begin{pmatrix}0\\1\end{pmatrix}$, from which we deduce

$$\xi_j\partial^2_{\xi_j}\xi_j = \left(\partial_{\xi_j}\xi_j - 1\right)\left(\xi_j\partial_{\xi_j} + 1\right) = \partial_{\xi_j}\xi_j^2\partial_{\xi_j} + \partial_{\xi_j}\xi_j - \xi_j\partial_{\xi_j} - 1 = \partial_{\xi_j}\xi_j^2\partial_{\xi_j}.$$

The last equality follows from (V-2.6). Hence

$$\begin{aligned}\sum_{j=1}^2 k_j\xi_j(a\partial^2_{\xi_1} + b\partial^2_{\xi_2} + 1)\xi_j &= k_1\left(a\xi_1\partial^2_{\xi_1}\xi_1 + b\xi_1\partial^2_{\xi_2}\xi_1 + \xi_1^2\right) + k_2\left(a\xi_2\partial^2_{\xi_1}\xi_2 + b\xi_2\partial^2_{\xi_2}\xi_2 + \xi_2^2\right)\\ &= k_1\left(a\partial_{\xi_1}\xi_1^2\partial_{\xi_1} + b\partial_{\xi_2}\xi_1^2\partial_{\xi_2} + \xi_1^2\right) + k_2\left(a\partial_{\xi_1}\xi_2^2\partial_{\xi_1} + b\partial_{\xi_2}\xi_2^2\partial_{\xi_2} + \xi_2^2\right)\\ &= a\partial_{\xi_1}\left(k_1\xi_1^2 + k_2\xi_2^2\right)\partial_{\xi_1} + b\partial_{\xi_2}\left(k_1\xi_1^2 + k_2\xi_2^2\right)\partial_{\xi_2} + k_1\xi_1^2 + k_2\xi_2^2.\end{aligned}$$

Thus we have

$$\mathcal{P}(-i\partial_\xi, i\xi) = \operatorname{div}\left(\begin{pmatrix}a(k_1\xi_1^2 + k_2\xi_2^2) & 0\\ 0 & b(k_1\xi_1^2 + k_2\xi_2^2)\end{pmatrix}\nabla\right) + k_1\xi_1^2 + k_2\xi_2^2 + C(-i\partial_\xi)$$

Let $C(x) = ax_1^2 + bx_2^2$, we have

$$\begin{aligned}\mathcal{P}(-i\partial_\xi, i\xi) &= \operatorname{div}\left(\begin{pmatrix}a(k_1\xi_1^2 + k_2\xi_2^2) & 0\\ 0 & b(k_1\xi_1^2 + k_2\xi_2^2)\end{pmatrix}\nabla\right) + k_1\xi_1^2 + k_2\xi_2^2 - a\partial^2_{\xi_1} - b\partial^2_{\xi_2}\\ &= \operatorname{div}\left(\begin{pmatrix}a(k_1\xi_1^2 + k_2\xi_2^2 - 1) & 0\\ 0 & b(k_1\xi_1^2 + k_2\xi_2^2 - 1)\end{pmatrix}\nabla\right) + k_1\xi_1^2 + k_2\xi_2^2.\end{aligned}$$

This operator is self-adjoint if $\begin{pmatrix} a(k_1\xi_1^2 + k_2\xi_2^2 - 1) & 0 \\ 0 & b(k_1\xi_1^2 + k_2\xi_2^2 - 1) \end{pmatrix}$ vanishes on ∂D_2. This happens when $D_2 = \mathit{Ellipse}(0, k_1, k_2)$.

When $a = b$ and $k_1 = k_2$, we recover the radial differential operator $\mathcal{P}_{\text{dim } 2}$ from (V-2.4). For $a \neq b$ or $k_1 \neq k_2$, the result is new and cannot be found in the literature, to the author's knowledge.

However, this approach fails when considering more general geometrical shapes, and also seems to fail if one is looking for higher-order differential operators.

Part V

CHAPTER 3 Prolate spheroidal wave functions

The Prolate Spheroidal Wave Functions are encountered when one solves the Helmholtz equation by separation of variables in prolate spheroidal coordinates (see [11, 35]). Their first mention is in [33].

The scalar Helmholtz equation is

$$(\Delta + k^2)\psi = 0, \tag{V-3.1}$$

and the change of variables from the Cartesian coordinates (x, y, z) to the prolate spheroidal coordinates (η, ξ, φ) is given by:

$$\begin{aligned} x &= \frac{d}{2}\left[(1-\eta^2)(\xi^2-1)\right]^{1/2}\cos\varphi, \\ y &= \frac{d}{2}\left[(1-\eta^2)(\xi^2-1)\right]^{1/2}\sin\varphi, \\ z &= \frac{d}{2}\eta\xi. \end{aligned}$$

This change of variables is illustrated in Figure V-3.1.

After a change of coordinates, the functions ψ solutions to (V-3.1) in Cartesian coordinates become solutions, in the Prolate Spheroidal coordinates, of the following equation:

$$\left[\frac{\partial}{\partial\eta}(1-\eta^2)\frac{\partial}{\partial\eta} + \frac{\partial}{\partial\xi}(\xi^2-1)\frac{\partial}{\partial\xi} + \frac{\xi^2-\eta^2}{(\xi^2-1)(1-\eta^2)}\frac{\partial^2}{\partial\varphi^2} + c^2(\xi^2-\eta^2)\right]\psi = 0, \tag{V-3.2}$$

where $c > 0$.

By separation of variables, we can write the solutions to (V-3.2) as

$$\psi_{m,n}(c,\eta,\xi,\varphi) = S_{m,n}(c,\eta)R_{m,n}(c,\xi)\cos(m\varphi)$$

and

$$\psi_{m,n}(c,\eta,\xi,\varphi) = S_{m,n}(c,\eta)R_{m,n}(c,\xi)\sin(m\varphi),$$

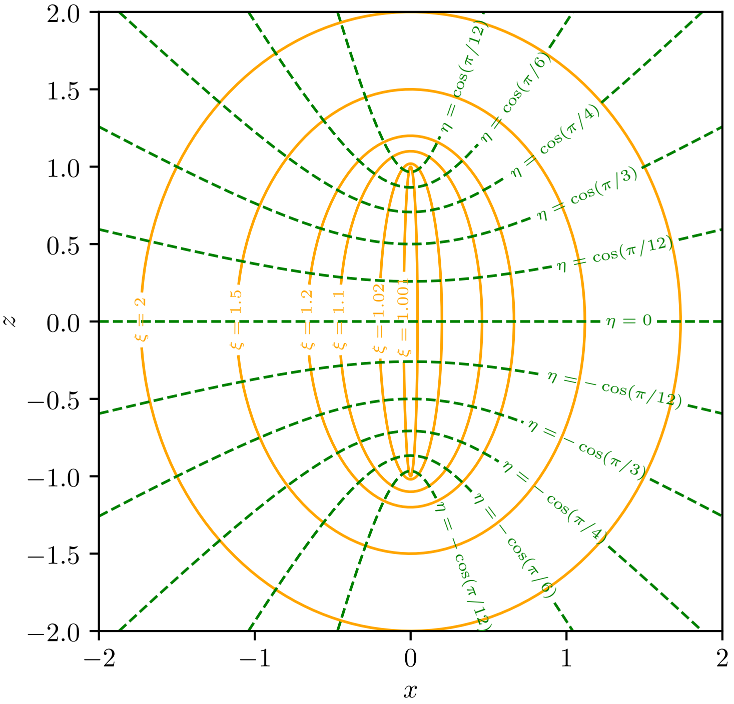


Figure V-3.1 – Prolate Spheroidal coordinates. The surfaces of constant η and ξ are obtained by rotation around the x-axis, so this diagram is valid for any plane containing the x-axis, hence it is sufficient to show only the x, z-plane.

where the functions S and R solve respectively

$$\frac{d}{d\eta}(1-\eta^2)\frac{d}{d\eta}S_{m,n}(c,\eta) + \left[\lambda_{m,n} - c^2\eta^2 - \frac{m^2}{1-\eta^2}\right] S_{m,n}(c,\eta) = 0,$$

and

$$\frac{d}{d\xi}(\xi^2-1)\frac{d}{d\xi}R_{m,n}(c,\xi) - \left[\lambda_{m,n} - c^2\xi^2 + \frac{m^2}{\xi^2-1}\right] R_{m,n}(c,\xi) = 0.$$

In order to obtain the functions $S_{m,n}, R_{m,n}$, it is sufficient to study the following one-dimensional equation:

$$\frac{d}{dz}(1-z^2)\frac{d}{dz}u(c,z) + \left[\lambda - c^2z^2 - \frac{\mu^2}{1-z^2}\right] u(c,z) = 0, \quad z \in (-1,1).$$

When $\mu = 0$, we are looking for u an eigenfunction of the differential operator $\mathscr{P}$ from (V-2.3). The function u is then said to be a PSWF of order zero.

We refer to [11] for a detailed study of Prolate Spheroidal Wave Function, and to [35] for a more recent work.

The Prolate Spheroidal Wave Functions are a one-parameter (c) family of functions, studied on the real interval $[-1,1]$. In [20], they are studied over $[b,1]$, and it is shown that a kernel operator exists and commutes with the modified differential operator. This allows to obtain Prolate Spheroidal Wave Function on a spherical cap of the three-dimensional unit ball. In [49], the Prolate Spheroidal Wave Functions are studied on triangles. In [32], the reproducing-kernel approach is generalized and known cases are recovered.

For other various applications and generalizations, we refer to [21, 26, 27].

The importance of PSWF in numerical schemes has been assessed in [5] and references therein, and it was shown that using the PSWF as a basis for spectral methods require less grid points and is numerically more appropriate for band-limited functions. The use of the Prolate Spheroidal Wave functions from a numerical perspective is also studied in [28, 3].

Also quite unexpectedly, the Prolate Spheroidal Wave functions seem to be linked to the zeros of the Riemann *zeta* function [6, 18]. This is merely anedoctal in this work, and we will not focus in this work on this aspect of the Prolate Spheroidal Wave Functions.

Part V

CHAPTER 4

Generalized framework

In the toy model from Section V-2.1 – The Slepian "toy model", it was chosen to have a Fourier transform with compact support in $[-\Omega, \Omega]$, and to maximize the $\mathbb{L}^2([-T, T])$ norm. The same "importance" is given to every point in the Fourier domain and to every point in the space domain. A natural question is the following: can we draw similar conclusions if one can choose the importance of every point in the Fourier and space domains? In other words, the toy model corresponds to filters being applied to the function f in space and Fourier domains, these filters being respectively $\mathbf{1}_{[-T,T]}$ and $\mathbf{1}_{[-\Omega,\Omega]}$. What can we say if these space and Fourier filters are now arbitrary?

Let us write m_S and $\widehat{m_F}$ the space and Fourier filters. They are the two inputs of the generalized spectral concentration problem. We assume that $m_S, \widehat{m_F} \in \mathbb{L}^2(\mathbb{R}^d; \mathbb{C})$ are not identically equal to zero. We also assume they are such that the kernel K defined later in (V-4.5) is not identically zero.

To the author's knowledge, this very general question has not been studied before. The existing situation closest to this one is that of Brander and DeFacio [4], who consider Gaussian filters. But even then, it is very restrictive. We allow much more freedom and derive a theoretical framework that allows our filters to be completely arbitrary, except for some mild integrability condition.

V-4.1 Derivation of the kernel K

We associate to each one of the filters an operator $\mathbb{L}^2(\mathbb{R}^d; \mathbb{C}) \to \mathbb{L}^2(\mathbb{R}^d; \mathbb{C})$, defined as follows: for $g \in \mathbb{L}^2(\mathbb{R}^d; \mathbb{C})$,

$$(\mathcal{M}_S g)(x) := m_S(x) g(x), \quad (\mathcal{M}_F g)(x) := \mathcal{F}^{-1}\left[\widehat{m_F}\mathcal{F}[g]\right](x),$$

where $\mathcal{F}[h]$ denotes the Fourier transform [1] of the function $h \in \mathbb{L}^2(\mathbb{R}^d; \mathbb{C})$:

$$\mathcal{F}[h](\xi) := \int_{\mathbb{R}^d} h(x) e^{-i\xi\cdot x} dx. \tag{V-4.1}$$

[1] See Section II-5 – The Fourier transforms for more details.

The notation $\hat{h}$ may also be used interchangeably. The Fourier transform is invertible, and we use the following convention:

$$\mathcal{F}^{-1}[h](x) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} h(\xi) e^{ix\cdot\xi} d\xi. \tag{V-4.2}$$

The Slepian Concentration Problem with arbitrary filters is now the following:

$$\arg \max_{f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} \frac{\|\mathcal{M}_F \mathcal{M}_S f\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}}{\|f\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}}. \tag{V-4.3}$$

We have

$$(\mathcal{M}_F \mathcal{M}_S f)(x) = \mathcal{F}^{-1}\left[\widehat{m_F} \mathcal{F}[m_S f]\right](x) = (m_F \star (m_S f))(x),$$

where $\star$ denotes the convolution operator, and thus

$$\begin{aligned}
&\|\mathcal{M}_F \mathcal{M}_S f\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} \\
&\quad = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} m_F(x-y)(m_S f)(y) dy \right) \overline{\left(\int_{\mathbb{R}^d} m_F(x-z)(m_S f)(z)(z) dz \right)} dx \\
&\quad = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} m_S(y) f(y) \overline{m_S(z) f(z)} \left(\int_{\mathbb{R}^d} m_F(x-y) \overline{m_F(x-z)} dx \right) dy dz.
\end{aligned}$$

The Parseval identity (see e.g. [48], or Theorem II.1) yields

$$\int_{\mathbb{R}^d} m_F(x-y) \overline{m_F(x-z)} dx = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\xi\cdot(z-y)} \left|\widehat{m_F}(\xi)\right|^2 d\xi,$$

hence

$$\begin{aligned}
&\|\mathcal{M}_F \mathcal{M}_S f\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} \\
&\quad = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} m_S(y) f(y) \overline{m_S(z) f(z)} \left(\int_{\mathbb{R}^d} e^{i\xi\cdot(z-y)} \left|\widehat{m_F}(\xi)\right|^2 d\xi \right) dy dz \\
&\quad = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x) K(x,y) \overline{f(y)} dx dy, \qquad \text{(V-4.4)}
\end{aligned}$$

where

$$\begin{aligned}
K(x,y) &:= \frac{1}{(2\pi)^d} m_S(x) \overline{m_S(y)} \int_{\mathbb{R}^d} e^{i\xi\cdot(y-x)} \left|\widehat{m_F}(\xi)\right|^2 d\xi \\
&= m_S(x) \overline{m_S(y)} \mathcal{F}^{-1}\left[\left|\widehat{m_F}\right|^2\right](y-x).
\end{aligned} \tag{V-4.5}$$

This kernel K is the generalized spectral concentration kernel, with filters m_S in space and $\widehat{m_F}$ in Fourier domain.

Lemma V.3

If $m_S, \widehat{m_F} \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$, then the kernel K is square integrable:

$$\|K\|_{\mathbb{L}^2(\mathbb{R}^d\times\mathbb{R}^d;\mathbb{C})} \leq \|m_S\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}\|\widehat{m_F}\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}.$$

Proof. We have:

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} |K(x,y)|^2\,dxdy = \int_{\mathbb{R}^d\times\mathbb{R}^d} |m_S(x)|^2|m_S(y)|^2 \left|\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](y-x)\right|^2 dxdy,$$

and

$$\left|\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](y-x)\right|^2 = \left|\int_{\mathbb{R}^d} |\widehat{m_F}(\xi)|^2 e^{i\xi\cdot(y-x)}d\xi\right|^2 \leq \left(\int_{\mathbb{R}^d} |\widehat{m_F}(\xi)|^2 d\xi\right)^2 = \|\widehat{m_F}\|^4_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}$$

Hence,

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d} |K(x,y)|^2\,dxdy &= \int_{\mathbb{R}^d\times\mathbb{R}^d} |m_S(x)|^2|m_S(y)|^2\|\widehat{m_F}\|^4_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}dxdy \\ &= \|m_S\|^4_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}\|\widehat{m_F}\|^4_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})},\end{aligned}$$

and finally

$$\|K\|_{\mathbb{L}^2(\mathbb{R}^d\times\mathbb{R}^d;\mathbb{C})} = \|m_S\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}\|\widehat{m_F}\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}. \tag{V-4.6}$$

□

Finally, the optimization problem (V-4.3) becomes:

$$\arg \max_{\substack{f\in\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})\\ \|f\|_2=1}} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x)K(x,y)\overline{f(y)}dxdy. \tag{V-4.7}$$

V-4.2 Eigenvectors of the kernel K

In this section, we show that the solutions to the maximization problem (V-4.7) form a countable subset of $\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$, and that these solutions are exactly the eigenvectors of the concentration operator. Moreover, for each one of those eigenvectors there exists an associated real eigenvalue lying in the open interval $(0,1)$.

We start by showing that the concentration operator is Hilbert-Schmidt, but first recall a few definitions and results from [38, Chapter VI]. In this section, we denote by H

a separable Hilbert space, and by $\mathcal{L}(H)$ the set of linear operators $H \to H$. We denote its inner product $(\cdot,\cdot)_H$.

Definition V.1: Trace of an operator, trace class

Let $\{\varphi_n\}_{n\in\mathbb{N}}$ an orthonormal basis. Then, for any positive operator $A \in \mathcal{L}(H)$, *trace* is defined by

$$\text{trace } A := \sum_{n\in\mathbb{N}} (\varphi_n, A\varphi_n)_H\,.$$

An operator $A \in \mathcal{L}(H)$ is called *trace class* if and only if

$$\text{trace } |A| < \infty.$$

Definition V.2: Hilbert-Schmidt operator

An operator $T \in \mathcal{L}(H)$ is called *Hilbert-Schmidt* if and only if

$$\text{trace } T^*T < \infty.$$

Theorem V.4

Let (M,μ) be a measure space and $H = \mathbb{L}^2(M, d\mu)$. Then $\mathcal{K} \in \mathcal{L}(H)$ is Hilbert-Schmidt if and only if there is a function

$$K \in \mathbb{L}^2(M \times M, d\mu \otimes d\mu),$$

with

$$(\mathcal{K}f)(x) = \int_M K(x,y)f(y)d\mu(y).$$

Moreover,

$$\|\mathcal{K}\|_2^2 = \int_{M\times M} |K(x,y)|^2 d\mu(x)d\mu(y).$$

Theorem V.5: Hilbert-Schmidt

Let A be a self-adjoint compact operator on H. Then there is a complete orthonormal basis $\{\psi_n\}_{n\geq 1}$ for H, so that $A\psi_n = \lambda_n\psi_n$ and $\lambda_n \to 0$ as $n \to \infty$.

We then have our main result about the concentration operator:

Proposition V.1

The concentration operator $\mathcal{K}$ defined for $f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$ by

$$(\mathcal{K}f)(x) := \int_{\mathbb{R}^d} K(x,y)f(y)dy, \tag{V-4.8}$$

where the kernel K is defined by (V-4.5), is a *Hilbert-Schmidt* operator and enjoys the following properties:

1. The kernel K is Hermitian, and the operator $\mathcal{K}$ is self-adjoint, compact, and positive semi-definite.
2. The countable family $\{\psi_i\}_{i=1}^{\infty}$ of eigenfunctions of $\mathcal{K}$ is orthonormal for the usual $\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$ inner product, the associated eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are real, nonnegative, and we can order them so that $1 > \lambda_i \geq \lambda_{i+1} \geq 0$, $i \geq 1$.
3. The orthonormal basis of eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ solve the maximization problem (V-4.7), and the maximal values attained are the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$.
4. For large n, $\lambda_n = o(n^{-1/2})$.
5. Suppose $|\widehat{m_F}|^2$ is even, and m_S is real, then $\mathcal{K}$ is real-valued for real inputs.

Proof. The fact that $\mathcal{K}$ is a Hilbert-Schmidt operator directly follows from Theorem V.4, since $K \in \mathbb{L}^2(\mathbb{R}^d \times \mathbb{R}^d;\mathbb{C})$ by Lemma V.3.

The Hermitian character is obtained easily:

$$\begin{aligned}\overline{K(x,y)} &= m_S(x)\overline{m_S(y)}\,\overline{\left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\widehat{m_F}(\xi)\right|^2 e^{i\xi\cdot(y-x)}\right)} \\ &= \overline{m_S(x)}m_S(y)\left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\widehat{m_F}(\xi)\right|^2 e^{i\xi\cdot(x-y)}\right) \\ &= K(y,x).\end{aligned}$$

For the self-adjointness, let $f, g \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$:

$$\begin{aligned}(\mathcal{K}f,g)_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} &= \int_{\mathbb{R}^d}(\mathcal{K}f)(x)\overline{g(x)}dx = \int_{\mathbb{R}^d}\overline{g(x)}\int_{\mathbb{R}^d}K(x,y)f(y)dydx \\ &= \int_{\mathbb{R}^d}\overline{g(x)}\int_{\mathbb{R}^d}\overline{K(y,x)}f(y)dydx = (f,\mathcal{K}g)_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}.\end{aligned}$$

The compactness follows from [38, Theorem VI.22(e)] which states that all Hilbert-

Schmidt operators are compact. The positive semi-definiteness is straightforward: let $f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$,

$$\begin{aligned}(f, \mathcal{K} f)_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x) m_S(x) \overline{m_S(y) f(y)} \left(\int_{\mathbb{R}^d} \left|\widehat{m_F}(\xi)\right|^2 e^{i\xi\cdot(y-x)} d\xi \right) dx dy \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\widehat{m_F}(\xi)|^2 \left(\int_{\mathbb{R}^d} f(x) m_S(x) e^{-i\xi\cdot x} dx \right) \overline{\left(\int_{\mathbb{R}^d} m_S(y) f(y) e^{-i\xi\cdot y} dy \right)} d\xi \\ &\geq 0.\end{aligned}$$

In particular, this implies that the eigenvalues are nonnegative. In order to obtain the orthonormal basis made of eigenfunctions of $\mathcal{K}$, we use Theorem V.5, which can be applied since we have just shown that $\mathcal{K}$ is a self-adjoint and compact operator.

We now show that the eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ are exactly solutions to the maximization problem (V-4.7).

Let $f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$, using the completeness of $\{\psi_i\}_{i=1}^{\infty}$ we can decompose

$$f = \sum_{i=1}^{\infty} c_i \psi_i. \tag{V-4.9}$$

We denote κ_1 the maximal value attainable in (V-4.7):

$$\kappa_1 = \max_{f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} \frac{(f, \mathcal{K} f)_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}}{\|f\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}}.$$

Using (V-4.9), we get

$$\kappa_1 = \max_{f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} \frac{1}{\|f\|^2_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})}} \sum_{i,j=1}^{\infty} c_i \lambda_j \overline{c_j} \left(\psi_i, \psi_j\right)_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} = \max_{f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})} \frac{1}{\sum_{i=1}^{\infty} |c_i|^2} \sum_{i=1}^{\infty} |c_i|^2 \lambda_i.$$

Now suppose that $\lambda_1 > \lambda_2$, then the maximal value attainable is $\kappa_1 = \lambda_1$ and it is attained for $f_1 = \psi_1$. The next function we are looking for in the maximization problem is one that maximizes (V-4.7), and which is orthogonal to ψ_1. If $\lambda_2 > \lambda_3$, the maximal value is $\kappa_2 = \lambda_2$ and it is attained for $f_2 = \psi_2$.

Now, if at any point there exists $i, m \geq 1$ such that $\lambda_i = \cdots = \lambda_{i+m}$, the maximal value attainable by the maximization problem is λ_i, and it is attained for $f \in \text{Span}\ \{\psi_i, \dots, \psi_{i+m}\}$. The m next maximization problems will also have as maximal value $\kappa_{i+l} = \lambda_i$, $l = 0, \dots, m$, and the functions for which this value is attained will be $f_{i+1}, \dots, f_{i+m} \in \text{Span}\ \{\psi_i, \dots, \psi_{i+m}\}$. Moreover, when looking for f_l, we impose that it is orthogonal to f_k, $k < l$. Thus, the functions $f_i, \dots, f_{i+m}$ will be pairwise orthogonal. We can obviously choose $f_i = \psi_i$. Finally, the orthonormal family of eigenvectors

$\{\psi_i\}_{i=1}^{\infty}$ of the concentration operator $\mathcal{K}$ solve the maximization problem (V-4.7), and the eigenvalues λ_i is the maximal value attained by (V-4.7) when imposing f orthogonal to $\{\psi_1, \dots, \psi_{i-1}\}$.

The fourth point follows from [38, Theorem VI.22(e)], which states that

$$\sum_{n=1}^{\infty} \lambda_n^2 < \infty.$$

This series is convergent only if $\lambda_n = o(n^{-1/2})$ for large n.

About the fifth point, it suffices to show that the inverse Fourier transform of $|\widehat{m_F}|^2$ is real. We have

$$\int_{\mathbb{R}^d} |\widehat{m_F}(\xi)|^2 e^{i\xi\cdot(y-x)} d\xi = \int_{\mathbb{R}^d} |\widehat{m_F}(\xi)|^2 \left(\cos(\xi\cdot(y-x)) + i\sin(\xi\cdot(y-x))\right) d\xi$$

When $|\widehat{m_F}|^2$ is even, the complex quantity vanishes as the integral of an odd function. Thus, only the real part remains. □

We are now led to study the following eigenproblem:

$$\begin{aligned} \lambda_i\psi_i(x) &= \int_{\mathbb{R}^d} K(x,y)\psi_i(y)dy \\ &= \int_{\mathbb{R}^d} m_S(x)\overline{m_S(y)}\psi_i(y)\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](y-x)dy. \end{aligned} \tag{V-4.10}$$

Remark V.5

To make ideas clearer, suppose that $m_S = \mathbf{1}_{D_1}$ and $\widehat{m_F} = \mathbf{1}_{D_2}$ for some finite-volume open sets $D_1, D_2 \in \mathbb{R}^d$. This means that we are looking for functions that are the most concentrated in D_1 and whose Fourier transform is the most concentrated in D_2. The eigenproblem (V-4.10) becomes

$$\lambda_i\psi_i(x) = \mathbf{1}_{D_1}(x)\int_{D_1} \psi_i(y)\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](y-x)dy.$$

Two situations can occur:

1. The eigenvector ψ_i is nonzero in some domain outside of D_1, and the RHS must be zero because of the indicator function. This implies that $\lambda_i = 0$. We note that, since the family of eigenvectors $\{\psi_i\}_{i=1}^{\infty}$ is a basis of $\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$, there is necessarily an infinity of eigenvalues which are zero.
2. The eigenvector ψ_i is compactly supported with supp $\psi_i \subset D_1$, in which case

λ_i can be "large" (i.e. close to 1).

The space mask $m_S = \mathbf{1}_{D_1}$ acts as a cut-off function for ψ_i: if we look at a nonzero eigenvalue λ_i, the corresponding eigenfunction ψ_i is contained in D_1. On the other hand, since ψ_i is compactly supported, the Fourier transform $\mathcal{F}[\psi_i]$ cannot be compactly supported in D_2. In other words, ψ_i is compactly supported in space and its Fourier transform is concentrated in D_2.

In the following lemma, for a function f and a matrix $\mathbf{A}$, we write $(f \circ \mathbf{A})(x) := f(\mathbf{A}x)$.

Lemma V.4: Symmetries

Suppose there exists a unitary matrix $\mathbf{S}$ and $\alpha \in \mathbb{C}$, $|\alpha| = 1$, such that $m_S \circ \mathbf{S} = \alpha m_S$ and $|\widehat{m_F} \circ \mathbf{S}| = |\widehat{m_F}|$. If ψ is an eigenvector of $\mathcal{K}$ associated to a nonzero eigenvalue of multiplicity one, then there exists $\beta \in \mathbb{C}$, $|\beta| = 1$, such that

$$\psi \circ \mathbf{S} = \beta\psi.$$

Proof. It follows from straightforward computations. First of all, since $\mathbf{S}$ is a unitary matrix, $|\det \mathbf{S}| = 1$. We now compute

$$\lambda\psi(\mathbf{S}x) = \int_{\mathbb{R}^d} m_S(\mathbf{S}x)\overline{m_S(y)}\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](y - \mathbf{S}x)\psi(y)dy.$$

The change of variables $y = \mathbf{S}\tilde{y}$ yields

$$\lambda\psi(\mathbf{S}x) = \int_{\mathbb{R}^d} m_S(\mathbf{S}x)\overline{m_S(\mathbf{S}\tilde{y})}\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\mathbf{S}\tilde{y} - \mathbf{S}x)\psi(\mathbf{S}\tilde{y})d\tilde{y},$$

where we have used $|\det \mathbf{S}| = 1$. Owing to the assumption $m_S \circ \mathbf{S} = \alpha m_S$, $|\alpha| = 1$,

$$\lambda\psi(\mathbf{S}x) = \int_{\mathbb{R}^d} m_S(x)\overline{m_S(\tilde{y})}\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\mathbf{S}\tilde{y} - \mathbf{S}x)\psi(\mathbf{S}\tilde{y})d\tilde{y}.$$

It only remains to show that $\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\mathbf{S}\tilde{y} - \mathbf{S}x) = \mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\tilde{y} - x)$. Letting $\xi = \mathbf{S}\tilde{\xi} \iff \tilde{\xi} = \mathbf{S}^*\xi$,

$$\begin{aligned}\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\mathbf{S}\tilde{y} - \mathbf{S}x) &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\widehat{m_F}(\xi)\right|^2 e^{i\xi\cdot\mathbf{S}(\tilde{y}-x)}d\xi \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\widehat{m_F}(\mathbf{S}\tilde{\xi})\right|^2 e^{i\mathbf{S}\tilde{\xi}\cdot\mathbf{S}(\tilde{y}-x)}d\tilde{\xi}.\end{aligned}$$

We have again used the fact that $|\det \mathbf{S}| = 1$. By the unitary property of the matrix $\mathbf{S}$, $\mathbf{S}\tilde{\xi} \cdot \mathbf{S}(\tilde{y} - x) = \tilde{\xi} \cdot (\tilde{y} - x)$, hence

$$\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\mathbf{S}\tilde{y}-\mathbf{S}x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\widehat{m_F}(\mathbf{S}\tilde{\xi})\right|^2 e^{i\tilde{\xi}\cdot(\tilde{y}-x)}d\tilde{\xi} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\widehat{m_F}(\tilde{\xi})\right|^2 e^{i\tilde{\xi}\cdot(\tilde{y}-x)}d\tilde{\xi},$$

where the last equality is due to the assumption $|\widehat{m_F} \circ \mathbf{S}| = |\widehat{m_F}|$. We finally obtain

$$\lambda\psi(\mathbf{S}x) = \int_{\mathbb{R}^d} m_S(x)\overline{m_S(\tilde{y})}\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](\tilde{y} - x)\psi(\mathbf{S}\tilde{y})d\tilde{y},$$

and since both $\psi \circ \mathbf{S}$ and ψ are eigenfunctions of $\mathcal{K}$ associated to a nonzero eigenvalue of multiplicity one, they must agree up to some multiplicative constant. Since they both have the same $\mathbb{L}^2(\mathbb{R}^d)$ norm, that constant must have modulus one. □

When λ is a multiple eigenvalue, we cannot show that all symmetries of the masks are recovered in the eigenfunctions. Indeed, suppose λ is an eigenvalue of multiplicity $p \in \mathbb{N}^*$, then there exist $\phi_1, \dots, \phi_p$, eigenfunctions of $\mathcal{K}$ associated to λ. For any $i = 1, \dots, p$, the same computations as above yield that $\phi_i \circ \mathbf{S}$ is an eigenfunction associated to λ. Therefore, we can only decompose

$$\phi_i \circ \mathbf{S} = \sum_{j=1}^{p} \beta_{i,j}\phi_j, \quad \beta_{i,j} \in \mathbb{C}, |\beta_{i,j}| = 1.$$

The case $p = 1$ is precisely the statement of Lemma V.4. For $p > 1$, we can expect that not all symmetries (and maybe none) will be recovered in the associated eigenfunctions.

V-4.3 Discretized problem

In this section we will discuss the discretized version of the continuous eigenproblem (V-4.10). As is generally done when dealing with continuous problems on a computer [2], we will discretize the problem.

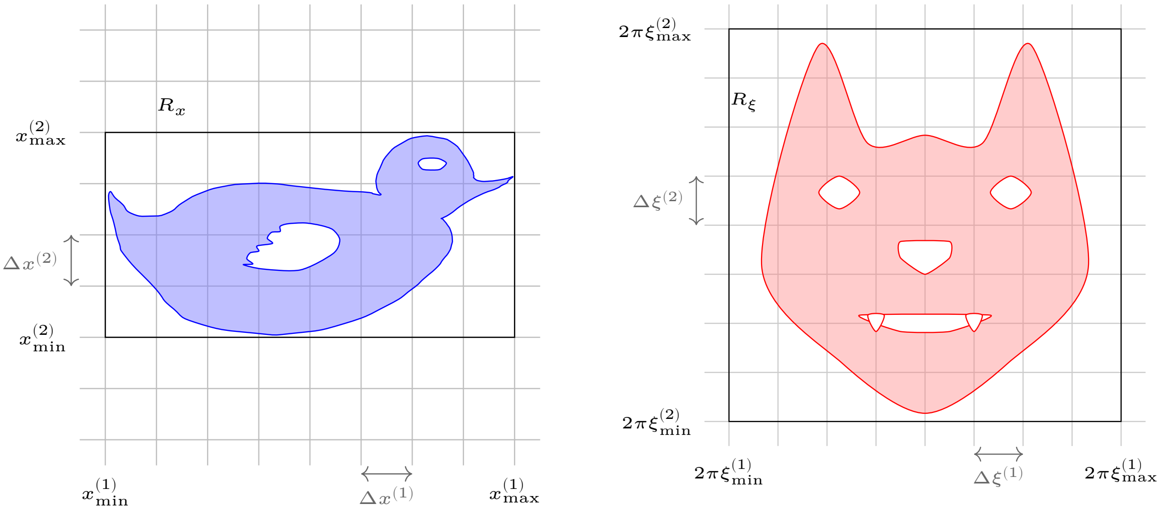
We start with a simplifying assumption: suppose there exists $x_{\text{min}}^{(k)}, x_{\text{max}}^{(k)} \in \mathbb{R}$, $k = 1, \dots, d$, such that

$$\text{supp } m_S \subset R_x := \left[x_{\text{min}}^{(1)}, x_{\text{max}}^{(1)}\right] \times \cdots \times \left[x_{\text{min}}^{(d)}, x_{\text{max}}^{(d)}\right],$$

and $\xi_{\text{min}}^{(k)}, \xi_{\text{max}}^{(k)} \in \mathbb{R}$, $k = 1, \dots, d$, such that

$$\text{supp } \widehat{m_F} \subset R_\xi := \left[\xi_{\text{min}}^{(1)}, \xi_{\text{max}}^{(1)}\right] \times \cdots \times \left[\xi_{\text{min}}^{(d)}, \xi_{\text{max}}^{(d)}\right].$$

[2] See Section II-1 – Representation of continuous problems on a computer for more details.



(a) Illustration of R_x.

(b) Illustration of R_ξ.

Figure V-4.1 – Illustration of R_x and R_ξ, when $m_S = \mathbf{1}_{\text{duck shape}}$ (left) and $\widehat{m_F} = \mathbf{1}_{\text{cat head}}$ (right).

For each dimension n of space, we use N_n+1 uniform discretization points, the stepsize is then $\Delta x^{(n)} = \frac{x_{\max}^{(n)} - x_{\min}^{(n)}}{N_n}$. We define

$$x_k^{(n)} := x_{\min}^{(n)} + k\Delta x^{(n)}, \quad k = 0, \dots, N_n,$$

and for $j = (j_1, \dots, j_d) \in \mathbb{N}^d$, define

$$\mathbf{x}_j := (x_{j_1}^{(1)}, \dots, x_{j_d}^{(d)}).$$

We write $J_x := [\![0, N_1]\!] \times \dots [\![0, N_d]\!] \subset \mathbb{N}^d$ the subset of all indices j such that $\mathbf{x}_j \in R_x$. The quantities R_x, R_ξ, $\Delta x^{(d)}$ and $\Delta\xi^{(d)}$ are illustrated in Figure V-4.1.

Similarly, for the Fourier variable, we use $2N_n + 1$ uniform discretization points in each Fourier dimension n, and the stepsize is $\Delta\xi^{(n)} = \frac{\xi_{\max}^{(n)} - \xi_{\min}^{(n)}}{2N_n}$. We define

$$\xi_k^{(n)} := \xi_{\min}^{(n)} + k\Delta\xi^{(n)}, \quad k = 0, \dots, 2N_n,$$

and for $j = (j_1, \dots, j_d) \in \mathbb{N}^d$, define

$$\Xi_j := (\xi_{j_1}^{(1)}, \dots, \xi_{j_d}^{(d)}).$$

We write $J_\xi := [\![0, 2N_1]\!] \times \dots [\![0, 2N_d]\!] \subset \mathbb{N}^d$ the subset of all indices j such that $\Xi_j \in R_\xi$. Unfortunately, since our continuous and discrete Fourier transform agree only up to a scaling of 2π , we need as well a variable taking into account this scaling of 2π. Thus, we let

$$\theta_{\min}^{(n)} := \frac{\xi_{\min}^{(n)}}{2\pi}, \quad \theta_{\max}^{(n)} := \frac{\xi_{\max}^{(n)}}{2\pi}, \quad \Delta\theta^{(n)} := \frac{1}{2\pi}\Delta\xi^{(n)},$$

for $n = 1, \dots, d$. We also write $\widehat{m}_{F,2\pi}$ the function such that

$$\widehat{m_{F,2\pi}}(\cdot) = \widehat{m_F}(2\pi\cdot).$$

Similarly to what we did for the ξ variable, we write

$$\theta_k^{(n)} := \theta_{\min}^{(n)} + k\Delta\theta^{(n)}, \quad k = 0, \dots, 2N_n,$$

and for $j = (j_1, \dots, j_d) \in \mathbb{N}^d$, we let

$$\Theta_j := \left(\theta_{j_1}^{(1)}, \dots, \theta_{j_d}^{(d)}\right).$$

The discrete version of the eigenproblem (V-4.10) is obtained by performing the numerical integration of (V-4.10) using the rectangle quadrature rule. In particular, the inverse Fourier transform has to be approximated. We have

$$\begin{aligned}
\mathcal{F}^{-1}\left[|\widehat{m_F}|^2\right](x) &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} |\widehat{m_F}(\xi)|^2 e^{i\xi\cdot x} d\xi \\
&= \int_{\mathbb{R}^d} |\widehat{m_F}(2\pi\theta)|^2 e^{2i\pi\theta\cdot x} d\theta, \quad \text{(by letting } \xi = 2\pi\theta) \\
&= \int_{\mathbb{R}^d} \left|\widehat{m_{F,2\pi}}(\theta)\right|^2 e^{2i\pi\theta\cdot x} d\theta \\
&\approx \left(\prod_{n=1}^{d} \Delta\theta^{(n)}\right) \sum_{l\in J_\xi} \left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\Theta_l\cdot x}
\end{aligned}$$

This yields, for every $j \in J_x$,

$$\lambda_i v_i(\mathbf{x}_j) = \left(\prod_{n=1}^{d} \Delta x^{(n)}\Delta\theta^{(n)}\right) m_S(\mathbf{x}_j) \sum_{k\in J_x} \overline{m_S}(\mathbf{x}_k) v_i(\mathbf{x}_k) \left[\sum_{l\in J_\xi} \left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\Theta_l\cdot(\mathbf{x}_k - \mathbf{x}_j)}\right]. \tag{V-4.11}$$

For the sake of simplicity, we will assume from now on that $\Delta x^{(n)} = 1$ for $n = 1, \dots, d$.

Remark V.6

This assumption is made without loss of generality. Indeed, if $\Delta x^{(n)} \neq 1$, we can consider the function $\widetilde{m}_S(y) := m_S(y_1\Delta x^{(1)}, \dots, y_d\Delta x^{(d)})$, and thus for an appropriate scaled version R_y of the grid R_x, we get $\Delta y^{(n)} = 1$.

Remark V.7

It is interesting to note that, in [45], D. Slepian defined the one-dimensional discrete problem on $\{1, \dots, N\}$ from scratch. The current section shows that the continuous and discrete problems are the same (up to space discretizations). Slepian's framework is indeed recovered if we assume $\prod_{n=1}^{d} \Delta x^{(n)} = 1$, since the discretization points are now all integers (up to a translation in the real space $\mathbb{R}^d$).

Moreover, we have

$$\Theta_l = \left(\theta_{\min}^{(1)} + l_1 \frac{\theta_{\max}^{(1)} - \theta_{\min}^{(1)}}{2N_1 + 1}, \dots, \theta_{\min}^{(d)} + l_d \frac{\theta_{\max}^{(d)} - \theta_{\min}^{(d)}}{2N_d + 1} \right).$$

We let

$$\Theta_{\min} := \left(\theta_{\min}^{(1)}, \dots, \theta_{\min}^{(d)}\right), \quad \Theta_{\max} := \left(\theta_{\max}^{(1)}, \dots, \theta_{\max}^{(d)}\right),$$

and

$$\Theta_l^{\max-\min} := \left(l_1 \frac{\theta_{\max}^{(1)} - \theta_{\min}^{(1)}}{2N_1 + 1}, \dots, l_d \frac{\theta_{\max}^{(d)} - \theta_{\min}^{(d)}}{2N_d + 1} \right).$$

Then,

$$\Theta_l = \Theta_{\min} + \Theta_l^{\max-\min},$$

and the eigenproblem (V-4.11) can be rewritten under the form:

$$\begin{aligned}
& \lambda_i v_i(\mathbf{x}_j) = m_S(\mathbf{x}_j) \sum_{k \in J_x} \overline{m_S}(\mathbf{x}_k) v_i(\mathbf{x}_k) e^{2i\pi\Theta_{\min}\cdot(\mathbf{x}_k - \mathbf{x}_j)} \\
& \qquad \times \left[\left(\prod_{n=1}^{d} \Delta\theta^{(n)} \right) \sum_{l \in J_\xi} \left| \widehat{m_{F,2\pi}}(\Theta_l) \right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(\mathbf{x}_k - \mathbf{x}_j)} \right] \\
\iff & \lambda_i v_i(\mathbf{x}_j) e^{2i\pi\Theta_{\min}\cdot\mathbf{x}_j} = m_S(\mathbf{x}_j) \sum_{k \in J_x} \overline{m_S}(\mathbf{x}_k) v_i(\mathbf{x}_k) e^{2i\pi\Theta_{\min}\cdot\mathbf{x}_k} \\
& \qquad \times \left[\left(\prod_{n=1}^{d} \Delta\theta^{(n)} \right) \sum_{l \in J_\xi} \left| \widehat{m_{F,2\pi}}(\Theta_l) \right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(\mathbf{x}_k - \mathbf{x}_j)} \right] \\
\iff & \lambda_i \tilde{v}_i(\mathbf{x}_j) = m_S(\mathbf{x}_j) \sum_{k \in J_x} \overline{m_S}(\mathbf{x}_k) \tilde{v}_i(\mathbf{x}_k) \left[\left(\prod_{n=1}^{d} \Delta\theta^{(n)} \right) \sum_{l \in J_\xi} \left| \widehat{m_{F,2\pi}}(\Theta_l) \right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(\mathbf{x}_k - \mathbf{x}_j)} \right],
\end{aligned} \tag{V-4.12}$$

where $\tilde{v}_i(x) := v_i(x)e^{2i\pi\Theta_{\min}\cdot x}$.

Thus, we are now looking for the vectors $\left\{\tilde{v}_i(\mathbf{x}_j)\right\}_{j\in J_x}$, which are eigenvectors of a matrix defined component-wise by:

$$m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\Delta\theta^{(n)}\right)\sum_{l\in J_\xi}\left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(\mathbf{x}_k-\mathbf{x}_j)}\right].$$

We can rewrite the matrix component (j,k) under a numerically more convenient form:

$$\begin{aligned}&m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\Delta\theta^{(n)}\right)\sum_{l\in J_\xi}\left|\widehat{m_F}(2\pi\Theta_l)\right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(\mathbf{x}_k-\mathbf{x}_j)}\right]\\&=m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\Delta\theta^{(n)}\right)\sum_{l\in J_\xi}\left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(\mathbf{x}_k-\mathbf{x}_j)}\right]\\&=m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\Delta\theta^{(n)}\right)\sum_{l\in J_\xi}\left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\Theta_l^{\max-\min}\cdot(k-j)}\right],\end{aligned}$$

where the last equality is due to

$$\mathbf{x}_k-\mathbf{x}_j=\left((k_1-j_1)\Delta x^{(1)},\dots,(k_d-j_d)\Delta x^{(d)}\right),$$

and we use the assumption that $\Delta x^{(n)}=1$ for $n=1,\dots,d$. Moreover,

$$\prod_{n=1}^{d}\Delta\theta^{(n)}=\prod_{n=1}^{d}\frac{\theta_{\max}^{(n)}-\theta_{\min}^{(n)}}{2N_n+1}$$

Again, for simplicity reasons, we assume that $\prod_{n=1}^{d}\left(\theta_{\max}^{(n)}-\theta_{\min}^{(n)}\right)=1$, since it is only a scaling of the eigenvalues and does not change the eigenvectors (which truly are the quantities of interest here). Thus,

$$\Theta_l^{\max-\min}=\left(\frac{l_1}{2N_1+1},\dots,\frac{l_d}{2N_d+1}\right),$$

and

$$\Delta\theta^{(n)}=\frac{1}{2N_n+1},\quad n=1,\dots,d.$$

Then,

$$\begin{aligned}
&m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\frac{1}{2N_n+1}\right)\sum_{l\in J_\xi}\left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\Theta_l\cdot(k-j)}\right]\\
&= m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\frac{1}{2N_n+1}\right)\sum_{l\in J_\xi}\left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\left(\frac{l_1}{2N_1+1},\dots,\frac{l_d}{2N_d+1}\right)\cdot(k-j)}\right]
\end{aligned}$$

We have finally obtained the matrix $\mathbf{K}$, of which we are looking for the eigenvectors $\tilde{v}$. It is defined component-wise by:

$$[\mathbf{K}]_{j,k} = m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\left[\left(\prod_{n=1}^{d}\frac{1}{2N_n+1}\right)\sum_{l\in J_\xi}\left|\widehat{m_{F,2\pi}}(\Theta_l)\right|^2 e^{2i\pi\left(\frac{l_1}{2N_1+1},\dots,\frac{l_d}{2N_d+1}\right)\cdot(k-j)}\right].$$

The attentive reader will have noticed that the expression inside square brackets corresponds exactly to the inverse Discrete Fourier Transform (IDFT, see Section II-5 – The Fourier transforms for more details).

We denote

$$\mathrm{IDFT}\left[|\widehat{m_{F,2\pi}}|^2\right]$$

the result obtained after the application of the inverse Discrete Fourier Transform to the function $|\widehat{m_{F,2\pi}}|^2$ evaluated at discretization points $\{\Theta_l\}_{l\in J_\xi}$. Then we can rewrite the matrix $\mathbf{K}$:

$$[\mathbf{K}]_{j,k} = m_S(\mathbf{x}_j)\overline{m_S(\mathbf{x}_k)}\,\mathrm{IDFT}\left[|\widehat{m_{F,2\pi}}|^2\right](k-j)\,. \qquad \text{(V-4.13)}$$

We emphasize that the matrix defined by (V-4.13) allows to compute the matrix $\mathbf{K}$ efficiently by using the Fast Fourier Transform (again, see Section II-5 for more details).

Remark V.8

We have not explained why we chose N_n+1 discretization points for x in dimension n, and $2N_n+1$ for ξ. It can be understood now. It suffices to consider the case $d=1$ in order to understand it. We have $j,k\in R_x = [\![0,N_1]\!]$, thus $k-j\in[\![-N_1,N_1]\!]$. We want to compute the DFT of $\left|\widehat{m_{F,2\pi}}\right|^2$ at every point $u\in[\![-N_1,N_1]\!]$, thus the definition of the DFT tells us that we need to discretize $\left|\widehat{m_{F,2\pi}}\right|^2$ using as many points as there are in $[\![-N_1,N_1]\!]$. In other words, the input and outputs of the DFT must have the same size for all dimensions. Since there are $2N_1+1$ integers in $[\![-N_1,N_1]\!]$, the ξ variable needs to be decomposed using $2N_1+1$ points of discretization. This holds for all dimensions. This explains why, if the space variable x is discretized using N_n discretization points along dimension n, then the variable ξ along dimension n has to be discretized using $2N_n + 1$ discretization points.

Moreover, in dimension greater than one, one cannot index the matrix $\mathbf{K}$ with j and k since they are not indices but multi-indices. Some convention thus has to be set in order to transform the d-dimensional index $j = (j_1, \dots, j_d)$ into a one-dimensional index $\tilde{j}$. The convention we choose for $\tilde{j}$ is to actually count the number of tuples $(j_1, \dots, j_d)$. But counting these tuples is not enough, we also need to order them so that a given tuple corresponds to a unique value of $\tilde{j}$, and vice-verse. To define this order relation, we compare the first index, then the second, and so on. Hence,

$\tilde{j}$	$j = (j_1, \dots, j_{d-1}, j_d)$
1	$(0, \dots, 0, 0)$
2	$(0, \dots, 0, 1)$
$\vdots$	$\vdots$
N_d	$(0, \dots, 0, N_d - 1)$
$N_d + 1$	$(0, \dots, 1, 0)$
$\vdots$	$\vdots$
$N_1 \cdots N_d$	$(N_d - 1, \dots, N_{d-1} - 1, N_d - 1)$

The explicit expression for $\tilde{j}$ is:

$$\tilde{j} = 1 + j_d + N_d j_{d-1} + N_d N_{d-1} j_{d-2} + \cdots + N_d \cdots N_2 j_1.$$

In order to simplify what follows, suppose $d = 2$. Then, $\mathbf{K}$ can be written as a block matrix:

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}^{(1,1)} & \mathbf{K}^{(1,2)} & \cdots & \mathbf{K}^{(1,N_2-2)} & \mathbf{K}^{(1,N_2-1)} \\ \mathbf{K}^{(2,1)} & \mathbf{K}^{(2,2)} & \cdots & \mathbf{K}^{(2,N_2-2)} & \mathbf{K}^{(2,N_2-1)} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{K}^{(N_1-2,1)} & \mathbf{K}^{(N_1-2,2)} & \cdots & \mathbf{K}^{(N_1-2,N_2-2)} & \mathbf{K}^{(N_1-2,N_2-1)} \\ \mathbf{K}^{(N_1-1,1)} & \mathbf{K}^{(N_1-1,2)} & \cdots & \mathbf{K}^{(N_1-1,N_2-2)} & \mathbf{K}^{(N_1-1,N_2-1)} \end{pmatrix}, \tag{V-4.14}$$

where each block $\mathbf{K}^{(r,c)}$ is defined component-wise by

$$[\mathbf{K}^{(r,c)}]_{m,n} := \mathbf{K}_{j=(r,m),k=(c,n)}.$$

Let us now explain why this indexing convention is particularly efficient. In (V-4.13), we can write the difference $k-j$ as:

$$\begin{pmatrix} k_1 - j_1 \\ k_2 - j_2 \end{pmatrix}.$$

With the chosen indexing convention, for each submatrix $\mathbf{K}^{(r,c)}$, we have

$$\left[\mathbf{K}^{(r,c)}\right]_{m,n} = m_S(\mathbf{x}_{j=(r,m)})\overline{m_S(\mathbf{x}_{k=(c,n)})}\,\mathrm{IDFT}\left[\left|\widehat{m_{F,2\pi}}\right|^2\right]\begin{pmatrix} c-r \\ n-m \end{pmatrix}.$$

This means that, along each diagonal of $\mathbf{K}^{(r,c)}$, the value of IDFT is constant: indeed, $c-r$ is constant in $\mathbf{K}^{(r,c)}$, and $n-m$ is constant along each diagonal of $\mathbf{K}^{(r,c)}$. In other words, each submatrix $\mathbf{K}^{(r,c)}$ is a Toeplitz matrix, multiplied row- and column-wise by the function m_S. The Toeplitz nature of each block $\mathbf{K}^{(r,c)}$ allows for efficient computational storage and complexity. Some more details are given in Section V-6.1 – Leveraging the Toeplitz nature of the kernel.

This two-dimensional discussion easily generalizes to the multidimensional case with the same indexing convention: only the last index varies within each block $\mathbf{K}^{(r_1,\dots,r_{d-1})}$ of $\mathbf{K}$, so each block can be expressed into a Toeplitz matrix and component-wise multiplications.

We have the following easy result:

Proposition V.2

The matrix $\mathbf{K}$ enjoys the following properties:

1. Hermitian: $\mathbf{K}^* = \mathbf{K}$
2. Structure: $\mathbf{K} = DBD^*$, where B is a block matrix where each block is Toeplitz, and D is a diagonal matrix.
3. Its eigenvalues are real and its eigenvectors form an unitary basis of $\mathbb{C}^{N_1 \cdots N_d}$.

Proof. For the first point, we use (V-4.13):

$$\begin{aligned}
[\mathbf{K}^*]_{k,j} = \overline{\mathbf{K}_{j,k}} &= \overline{m_S(\mathbf{x}_j)\overline{m_S}(\mathbf{x}_k)\,\mathrm{IDFT}\left[|\widehat{m_{F,2\pi}}|^2\right](k-j)} \\
&= \overline{m_S(\mathbf{x}_j)}m_S(\mathbf{x}_k)\overline{\mathrm{IDFT}\left[|\widehat{m_{F,2\pi}}|^2\right](k-j)}.
\end{aligned}$$

Moreover, since $|\widehat{m_{F,2\pi}}|^2$ is real, we obtain

$$\overline{\mathrm{IDFT}\left[|\widehat{m_{F,2\pi}}|^2\right](k-j)} = \mathrm{IDFT}\left[|\widehat{m_{F,2\pi}}|^2\right](j-k)\,.$$

Hence,

$$[\mathbf{K}^*]_{k,j} = \overline{m_S(\mathbf{x}_j)} m_S(\mathbf{x}_k) \, \mathrm{IDFT} \left[|\widehat{m_{F,2\pi}}|^2 \right] (j-k) = [\mathbf{K}]_{k,j}.$$

That is, $\mathbf{K}^* = \mathbf{K}$, i.e. $\mathbf{K}$ is an Hermitian matrix.

The second point has already been mentioned earlier: the diagonal matrix D follows from (V-4.13), and it corresponds to the component-wise multiplication by the function $m_S(\mathbf{x}_j)$ for each row j and by $m_S(\mathbf{x}_k)$ for each column k of $\mathbf{K}$. Hence $\mathbf{K}$ is of the form DBD^*, where B is some matrix. The block nature of B, where each block is a Toeplitz matrix, follows from (V-4.14).

The third point is a classical result in linear algebra: for any normal matrix, there exists an orthonormal basis of eigenvectors. See for instance [2, Theorem 7.31]. □

Remark V.9

In all of the numerical examples of Section V-5, we will use the simplifying assumptions that allowed us to write $\mathbf{K}$ as in (V-4.13). The assumption $\prod_{n=1}^{d} \left(\theta_{\max}^{(n)} - \theta_{\min}^{(n)} \right) = 1$ can simply be stated in the following way:

$$\mathrm{supp} \; \widehat{m_F} \subset [-\pi, \pi]^d \qquad \Longleftrightarrow \qquad \mathrm{supp} \; \widehat{m_{F,2\pi}} \subset \left[-\frac{1}{2}, \frac{1}{2} \right]^d .$$

The assumption $\Delta x^{(n)} = 1$ for $n = 1, \ldots, d$, simply is a change of coordinates: if $\Delta x^{(n)} \neq 1$, we consider a scaling of the space grid R_x, so that the grid size after scaling is 1.

We end this Chapter with a remark concerning notation: we are interested in eigenpairs of the matrix $\mathbf{K}$ of finite-dimension. With a slight abuse of notation, we'll denote ψ_i the eigenvectors, which is the same notation as used for the eigenfunctions of $\mathcal{K}$. Whether we are talking about an eigenvector or an eigenfunction will always be clear from the context: if we are talking about the continuous concentration operator $\mathcal{K}$, ψ_i will denote an eigenfunction, and if we are talking about the discretized version of $\mathcal{K}$ (i.e. the matrix $\mathbf{K}$), then ψ_i will denote an eigenvector.

Part V

CHAPTER 5 I like to move it move it

It has already been mentioned in Section V-2.2.3 – Numerical difficulties that the main difficulty in obtaining the eigenvectors of the matrix **K** defined by (V-4.13) is the fact that the eigenvalues are very tightly packed close to one or zero. Figure V-2.1b is an illustration of this phenomenon.

In this Chapter, we will see that the eigenvalues grouped together are often several simple eigenvalues, and not a single one with a high multiplicity. This is a very important fact, because it means we can generally associate one eigenvector to each eigenvalue. This would not be possible if there was a multiple eigenvalue, in this case we could only associate an eigenspace to the eigenvalue. The fact that several eigenvalues are the same in the Slepian toy-model can then be understood as a "degeneracy" of the generalized kernel (V-4.5), where the "degeneracy" is due to having indicator functions as filters m_S and $\widehat{m_F}$.

Once we know that the eigenvalues actually are of multiplicity one, we can hope for numerical schemes to be able to recover (at least approximately!) the desired eigenvectors.

We start by showing that indeed, each eigenvalue is of multiplicity one, but they get packed very tightly when indicator filters are chosen. Moreover, by approximating the indicator functions by some other function, we are able to "separate" the eigenvalues and get good approximations of the desired eigenvectors of the matrix **K**. This is the first numerical procedure we present. Then, we give a second numerical procedure to obtain approximate eigenvectors of **K**. It is more expensive than the first one, but the approximation can be quantified. We end this Chapter by obtaining eigenvectors for previously unstudied filters, with qualitatively good results. The "qualitative" property is unfortunately only visible on numerical results, since the eigenvectors obtained must answer the following question: *Do the eigenvectors look like linear combinations of some of them, or do they look like "single" ones?* This can be understood for instance on Figure V-2.1a, where the solid and dash curves seem more "mixed" than the dot-dash curve.

For the two procedures mentioned above, each family of approximate eigenvectors that we obtain is an acceptable answer to the Spectral Concentration Problem.

V-5.1 Perturbed operator

In order to show that the eigenvalues are each of multiplicity one, we consider a perturbation $K^{[\varepsilon]}$ of the concentration kernel K defined by (V-4.5), depending on the *perturbation parameter* $\varepsilon > 0$:

$$K^{[\varepsilon]}(x,y) := m_S^{[\varepsilon]}(x)\overline{m_S^{[\varepsilon]}(y)} \int_{\mathbb{R}^d} e^{i\xi\cdot(y-x)} \left|\widehat{m_F}(\xi)\right|^2 d\xi. \tag{V-5.1}$$

Only the space filter m_S is perturbed, its perturbation being written $m_S^{[\varepsilon]}$, and the Fourier filter $\widehat{m_F}$ is left unmodified[1]. The perturbed space filter $m_S^{[\varepsilon]}$ is assumed to be such that $m_S^{[\varepsilon]} \to m_S$ as $\varepsilon \to 0$, in the $\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$ norm.

We can assume that $m_S^{[\varepsilon]} \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$ since it holds for $\varepsilon > 0$ small enough by convergence. Thus, the kernel $K^{[\varepsilon]}$ satisfies the same hypotheses as K and the generalized framework from Section V-4 applies. In particular, Proposition V.1 applies to the perturbed operator $\mathcal{K}^{[\varepsilon]}$ defined for $f \in \mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$ by:

$$(\mathcal{K}^{[\varepsilon]}f)(x) := \int_{\mathbb{R}^d} K^{[\varepsilon]}(x,y)f(y)dy. \tag{V-5.2}$$

We obtain that the countable family of eigenfunctions of $\mathcal{K}^{[\varepsilon]}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}^d;\mathbb{C})$ of which all elements are concentrated in the Fourier domain. Moreover, depending on the application at hand, a small modification of the space mask may yield an acceptable answer. Hence, one could use the eigenfunctions obtained for some $\varepsilon > 0$ as an approximation of the true eigenfunctions (corresponding to $\varepsilon = 0$).

Now is an appropriate time to discuss why the perturbed operator is of particular interest. For the time being, we consider the one-dimensional Slepian toy model[2], and recall the behavior of eigenvalues which was noticed in numerical experiments (i.e. for the discretized toy model) by Slepian as early as 1961: when indicator filters are considered, most of the eigenvalues are very tightly packed close to zero or close to one, and only a very small number of eigenvalues is in-between. More rigorous studies have been done since, see for instance [24] for the best current bound on the number of eigenvalues between zero and one. Using his commuting differential operator, Slepian was able to find the correct eigenvectors. A qualitative aspect of the n-th eigenvector ψ_n is that, as n increases, the associated eigenvalue λ_n decreases and the support of ψ_n gets larger. To illustrate this

[1] This is mainly to have an easier numerical construction of the discretized operator, since the filter $\widehat{m_F}$ needs to be Fourier transformed. By keeping the same $\widehat{m_F}$, the Fourier transform only has to be done once for all perturbations.

[2] See Section V-2.1 – The Slepian "toy model".

qualitative aspect, the first 8 eigenvectors are given for the discretized Slepian toy model in Figure V-5.1.

The first five eigenvectors have very close eigenvalues because they can go to zero sufficiently fast at the boundaries of the interval. Thus, they are all seen as numerically very concentrated. On the other hand, the last three eigenvectors cannot go sufficiently fast to zero at the boundaries of the interval, and their $\mathbb{L}^2(\mathbb{R} \setminus [-1,1])$ norm must be nonzero. This is why their associated eigenvalues get smaller and smaller.

The heart of the problem lies in the first eigenvectors, because there is numerically no way to distinguish them using their concentration ratio alone. This is partly due to the indicator mask in space, because the same "importance" is given to the first and fifth eigenvectors, even though the former is qualitatively more concentrated than the latter. By modifying slightly the space mask, it is possible to change this and to impose the concentration around the origin (more generally, around any point in the interval).

Remark V.10

We now focus on the case of binary filters, that is

$$m_S = \mathbf{1}_{D_1} \quad \text{and} \quad \widehat{m_F} = \mathbf{1}_{D_2},$$

for some subsets $D_1, D_2 \subset \mathbb{R}^d$.

Take $m_S^{[\varepsilon]} = (1 + \varepsilon \cos(\omega \cdot))\, \mathbf{1}_{[-1,1]}$. The parameter ω will be called the *perturbation frequency*. The parameter ε can be understood as the space perturbation, while ω is the Fourier perturbation. Conceptually, we want ε, ω small: it should be clear for ε, and for ω it is due to the fact that the application of the operators $\mathcal{M}_S$ and $\mathcal{M}_F$ actually constrains the Fourier transform of $m_S f$, and not only the Fourier transform of f. Thus, to obtain results similar to those we know in the 1d case, we want the modification due to the perturbation to be as small as possible. This is why both ε and ω have to be small.

Using this perturbed filter $m_S^{[\varepsilon]}$, we can easily obtain numerically the eigenvalues. They are given in Figure V-5.2. To the contrary of the unperturbed eigenvalues given in Figure V-5.2a, those of the perturbed problem given in Figure V-5.2b are clearly separated.

Still with the same perturbed filter $m_S^{[\varepsilon]}$, the eigenvectors of the perturbed problem can be obtained more easily than those of the unperturbed problem. In Figure V-5.3 we plot the first sixteen eigenvectors (associated to the sixteen highest eigenvalues) in the space domain for the perturbed problem (blue solid curve). The yellow dash curve corresponds to the exact eigenvectors obtained using Slepian's commuting differential operator solution.

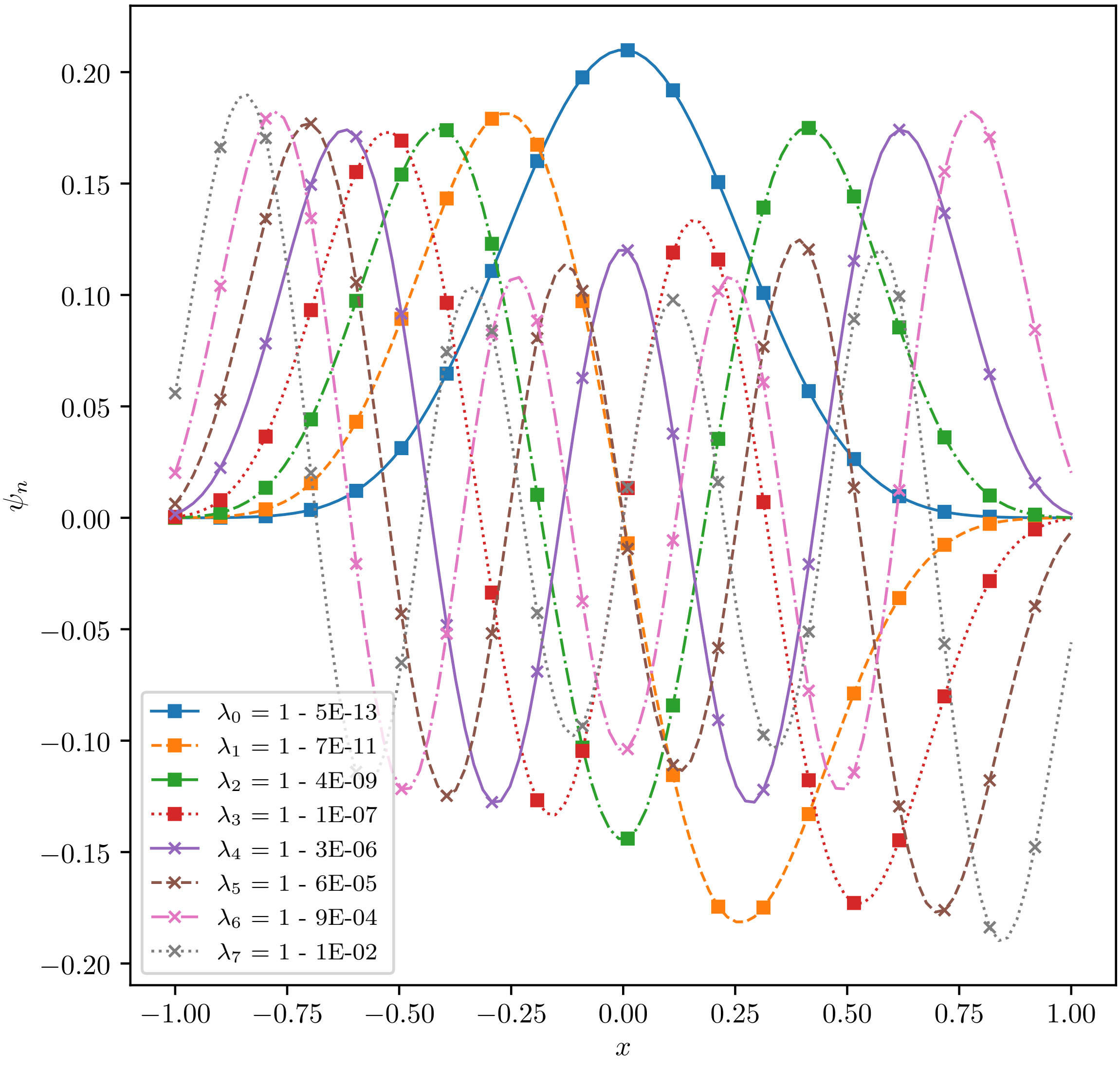
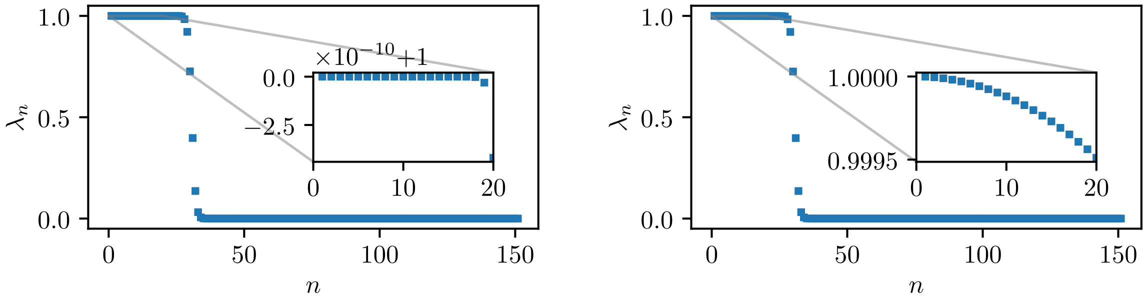


Figure V-5.1 – First eight eigenvectors for the Slepian toy model, obtained via the commuting differential operator. Numerical parameters are $N = 100$, $\Omega = 0.05 \cdot 2\pi$.



(a) $\varepsilon = 0$.

(b) $\varepsilon = 0.1$.

Figure V-5.2 – Eigenvalues for the perturbed filter $m_S^{[\varepsilon]} = (1 + \varepsilon \cos(\omega\cdot))\, \mathbf{1}_{[-1,1]}$, with $\omega = 0.1$ and $\varepsilon \in \{0, 0.1\}$.

In Figure V-5.4 we plot their Fourier transform, as well as vertical dash lines indicating the boundaries of the Fourier mask interval $[-\Omega, \Omega]$.

In Figures V-5.3 and V-5.4, one can see that the results are relatively satisfying: they are qualitatively good in the space domain (Figure V-5.3) and the Fourier transform is indeed restricted to the given interval $[-\Omega, \Omega]$ (Figure V-5.4). More precisely, in space, the eigenvectors are localized correctly, they exhibit the expected number of main "bumps", and the "spread" of the eigenvector labelled n is indeed growing with n: $\lambda_i > \lambda_j$ for $i < j$. Qualitatively, the eigenvector i is more localized than the eigenvector j. In the Fourier domain, the expected behavior is that the eigenvectors are restricted to the interval $[-\Omega, \Omega]$, which is indeed the case.

The fact that the eigenvectors of the perturbed problem correspond qualitatively to those we are looking for is due to the fact that the eigenvalues are "far" from one another in the perturbed case, given that ε is sufficiently far from zero.

There are however two main issues with this approach and the perturbed space mask $m_S^{[\varepsilon]}$ of the form given previously:

1. the eigenvectors to the perturbed problem exhibit oscillations which are absent from the solution to the nonperturbed problem,
2. the eigenvectors are more concentrated around the origin than expected.

The first point does not seem to be linked to the particular form of the perturbed filter $m_S^{[\varepsilon]}$, our experiments showed that oscillations also occur if one takes $m_S^{[\varepsilon]}(x) = \left(1 + \varepsilon e^{-\varepsilon \frac{x^2}{2}}\right) \mathbf{1}_{[-1,1]}$: the results are displayed in Figure V-5.5.

The second point is due to our specific perturbation function, and the fact that the amplitude of the filter is larger around the origin. Eigenvectors maximizing the concentration ratio with respect to this perturbed filter will then be concentrated around the

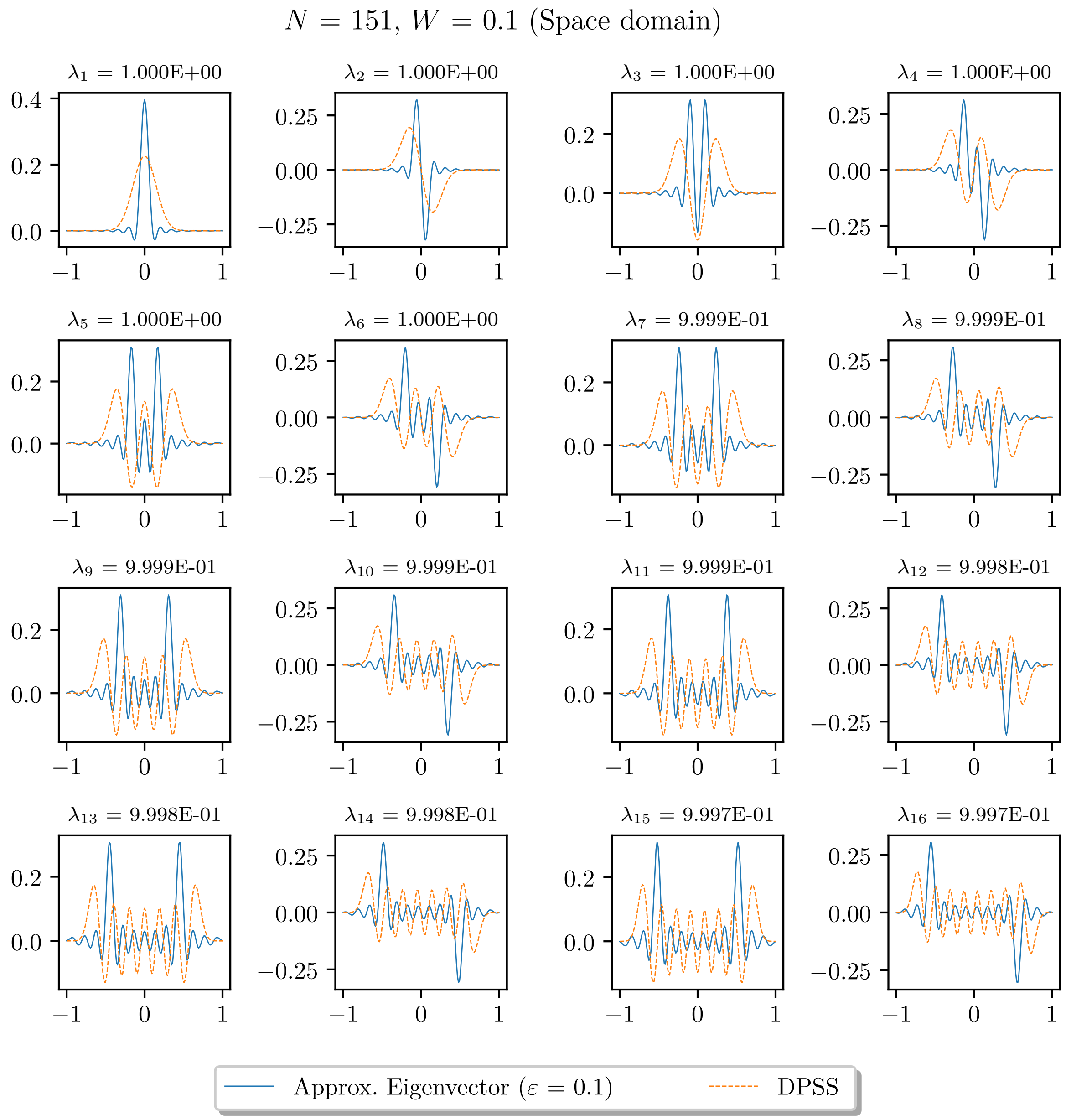


Figure V-5.3 – Results obtained with $N = 151$, $\Omega = 0.1 \cdot 2\pi$, with the following perturbation parameters: $\varepsilon = 0.1$ and $\omega = 0.1$. Eigenvectors are plotted in the space domain. The space filter is $m_S^{[\varepsilon]}(x) = (1 + \varepsilon \cos(\omega x))\mathbf{1}_{[-1,1]}$.

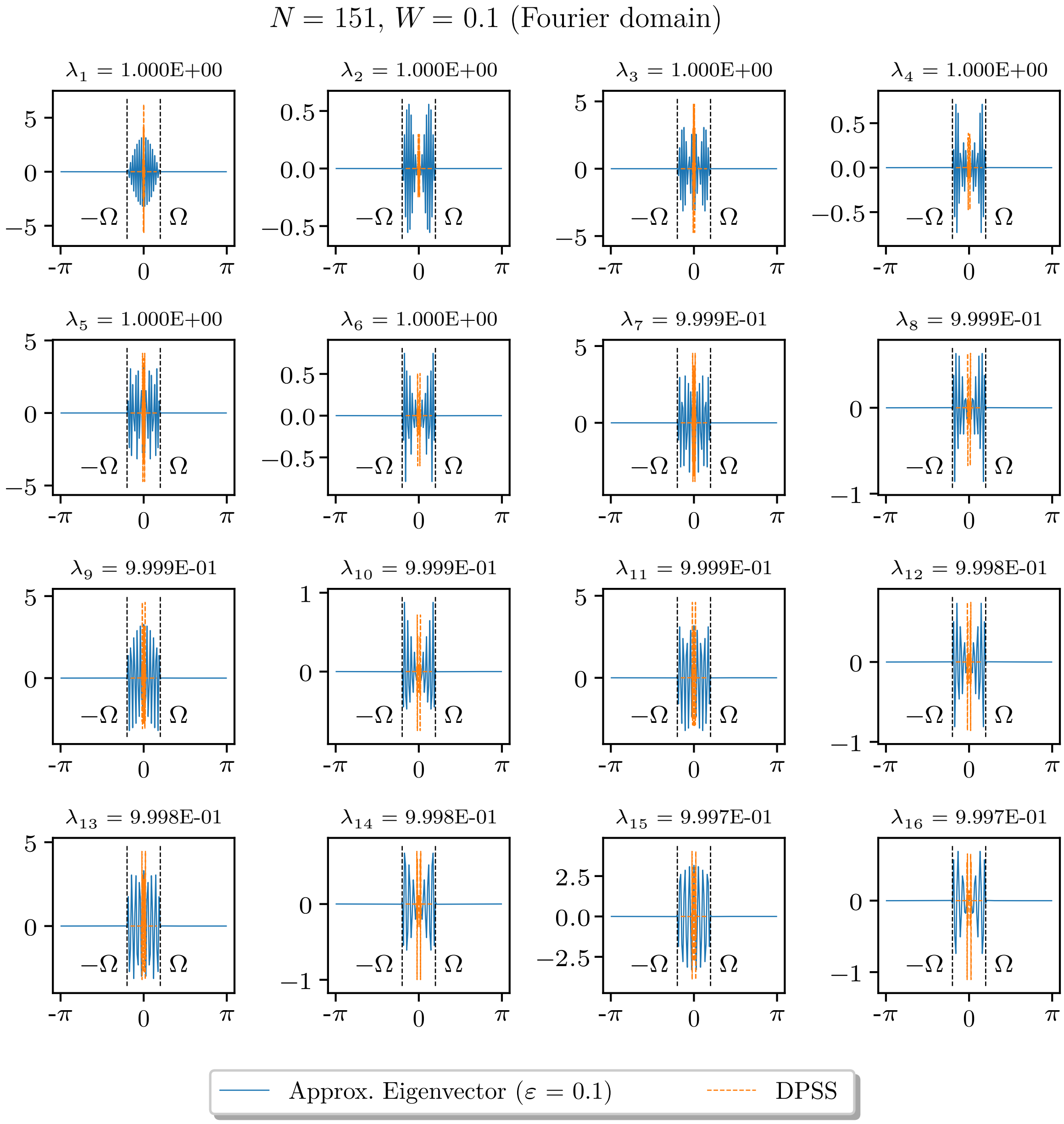


Figure V-5.4 – Results obtained with $N = 151$, $\Omega = 0.1 \cdot 2\pi$, with the following perturbation parameters: $\varepsilon = 0.1$ and $\omega = 0.1$. Eigenvectors are plotted in the Fourier domain. The space filter is $m_S^{[\varepsilon]}(x) = (1 + \varepsilon \cos(\omega x))\mathbf{1}_{[-1,1]}$.

origin. For instance, if one decides to take $m_S^{[\varepsilon]}(x) = (1 + \varepsilon \sin(\omega x))\, \mathbf{1}_{[-1,1]}$, the results are totally different, see Figure V-5.6. The difference is that the sine-based filter has higher amplitude close to 1, so the most concentrated eigenvectors will be in a region close to 1.

The difference in the results can be easily explained: the function $x \mapsto 1+\varepsilon\cos(\omega x)$ is highest around the origin, thus the corresponding eigenvectors will concentrate here. On the contrary, the function $x \mapsto 1 + \varepsilon \sin(\omega x)$ is highest at the right of the interval, thus the corresponding eigenvectors will concentrate on the right of the interval. See Figure V-5.7.

The key takeaway from these numerical experiments is that the seemingly multiple eigenvalue $\lambda \approx 1$ in Figure V-2.1b is simply several simple eigenvalues which coincidentally have the same value. This is very important in the sense that it justifies looking for a particular eigenvector associated to each eigenvalue. Having the perturbation in mind: if one perturbs the kernel, eigenvalues are distinct and eigenvectors have quantitatively the correct behavior, and as the perturbation parameter goes to zero we expect to be able to keep track of these particular eigenvectors. This naturally calls for the use of eigenvector continuation…

V-5.2 Eigenvector continuation

The *eigenvector continuation* is a technique which consists in relating the eigenvectors and eigenvalues derivatives, giving a system of ordinary differential equations (ODE), and then integrating these ODE. Basically, these equations are obtained by differentiating the (right) eigendecomposition and using left and right eigenvectors. We refer to [23] for the details.

The motivation in this Section is the following: we saw previously that we are able to obtain the eigenvectors and eigenvalues of the concentration operator $\mathcal{K}$ associated to a perturbed space filter $m_S^{[\varepsilon]}$ depending on parameters ε, ω. If $\varepsilon = 0$, we recover the unperturbed space filter. We then want to find the eigenvectors of the concentration operator $\mathcal{K}$ associated to $m_S^{[\varepsilon]}$, and let $\varepsilon \to 0$. Since the eigenvectors and eigenvalues will also depend on ε, we want to apply eigenvector continuation in the limit $\varepsilon \to 0$.

The main equations for an Hermitian matrix $M(\varepsilon) \in \mathbb{R}^{n\times n}$ which depends on a parameter $\varepsilon \in \mathbb{R}$, with eigenvectors $u_i(\varepsilon)$ associated to n **distinct** eigenvalues $\lambda_i(\varepsilon)$ are

$$\begin{aligned} \dot{\lambda}_i(\varepsilon) &= u_i^*(\varepsilon)\dot{M}(\varepsilon)u_i(\varepsilon), \\ \dot{u}_i(\varepsilon) &= \sum_{j\neq i} \frac{u_j^*(\varepsilon)\dot{M}(\varepsilon)u_i(\varepsilon)}{\lambda_i(\varepsilon)-\lambda_j(\varepsilon)} u_j(\varepsilon), \end{aligned} \tag{V-5.3}$$

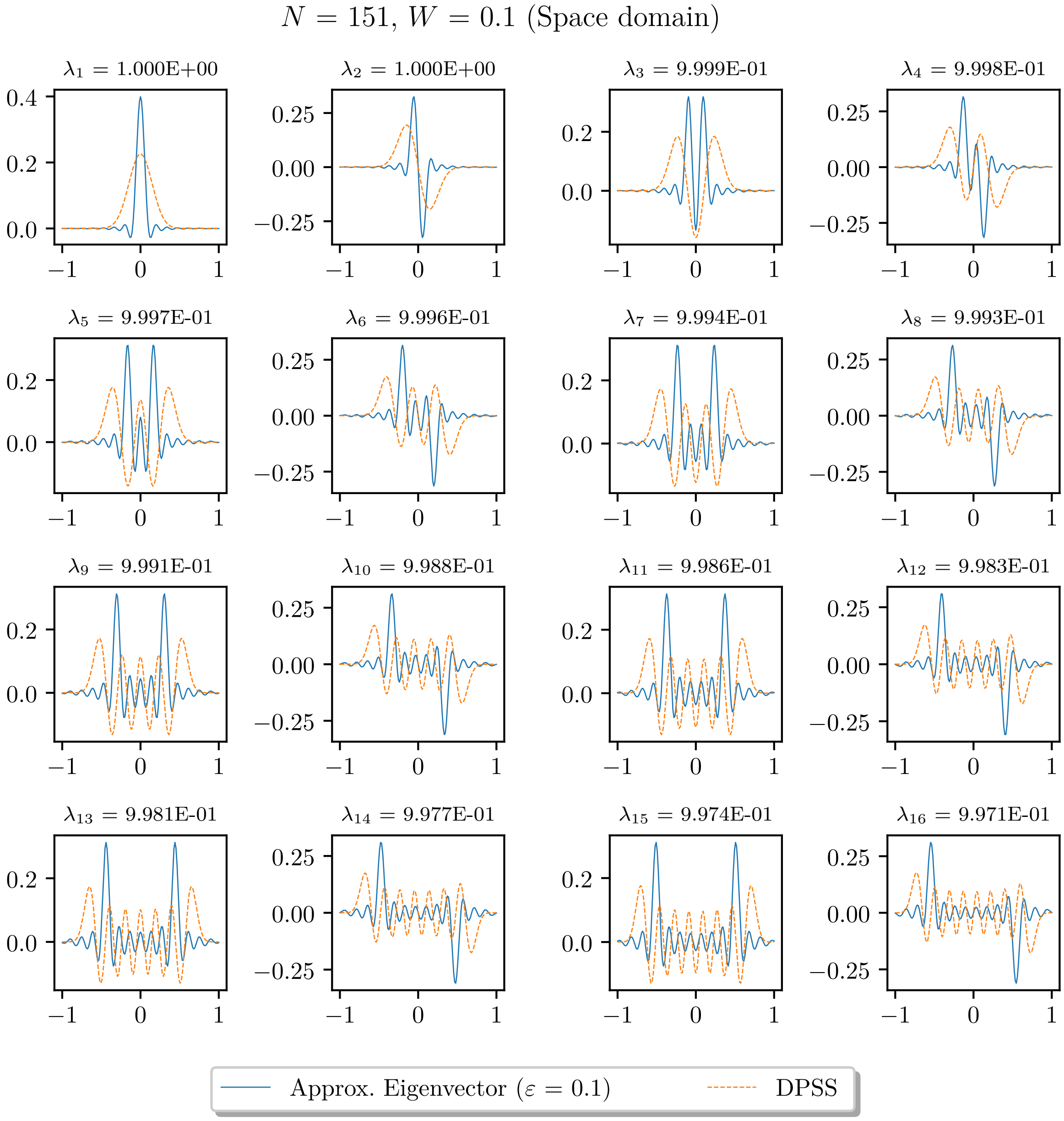


Figure V-5.5 – Results obtained with $N = 151$, $\Omega = 0.1 \cdot 2\pi$, with the following perturbation parameters: $\varepsilon = 10^{-1}$. Eigenvectors are plotted in the space domain. The perturbed space mask is $m_S^{[\varepsilon]}(x) = \left(1 + \varepsilon e^{-\varepsilon \frac{x^2}{2}}\right) \mathbf{1}_{[-1,1]}$.

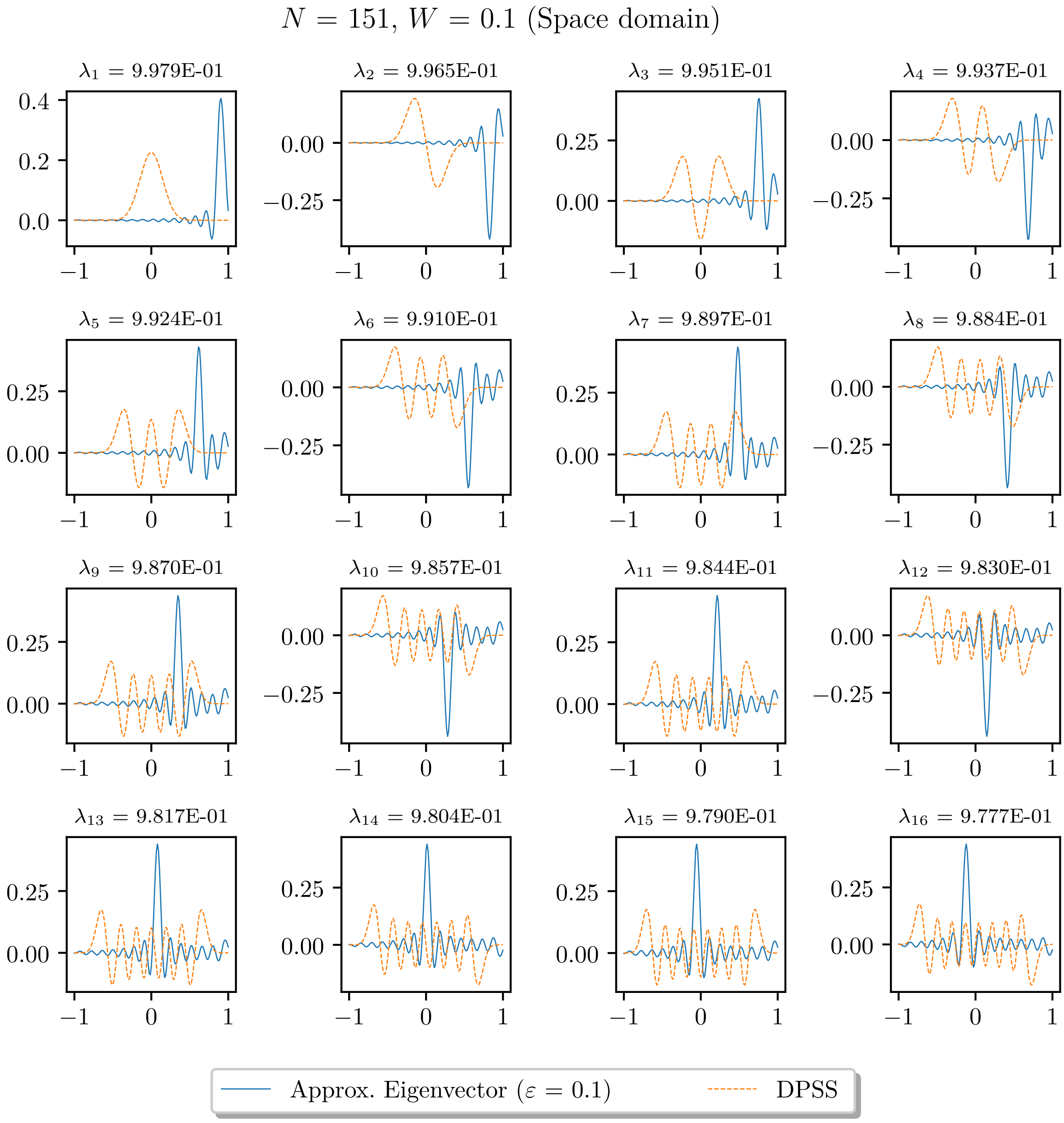
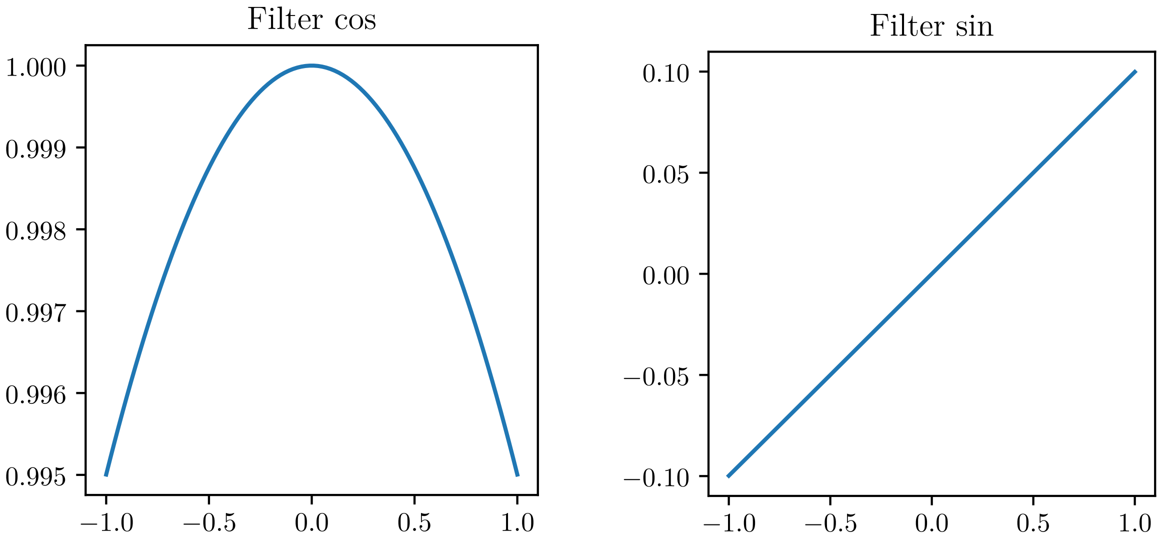


Figure V-5.6 – Results obtained with $N = 151$, $\Omega = 0.1 \cdot 2\pi$, with the following perturbation parameters: $\varepsilon = 0.1$ and $\omega = 0.1$. Eigenvectors are plotted in the space domain. The perturbation filter $m_S^{[\varepsilon]}$ uses a sine function.



(a) $x \mapsto \cos(\omega x)$.

(b) $x \mapsto \sin(\omega x)$.

Figure V-5.7 – Cosine and sine-based perturbations, with $\omega = 10^{-1}$.

for $i = 1, \dots, n$. The dot notation $\dot{A}(\varepsilon)$ denotes the derivative of the quantity A with respect to the variable ε.

Numerically, one would use these equations as follows:

1. Start from an eigendecomposition $M(\varepsilon)U(\varepsilon) = U(\varepsilon)\Lambda(\varepsilon)$ of the perturbed matrix $M(\varepsilon)$, with $\varepsilon > 0$. We denote $U(\varepsilon) = (u_1, \dots, u_n)(\varepsilon)$, and $\Lambda(\varepsilon) = \operatorname{diag}(\lambda_1, \dots, \lambda_n)(\varepsilon)$. This raises no issue since the eigenvalues are distinct and the eigenvectors cannot be linear combinations of the others, thus it is easy to obtain eigenvectors.

2. Compute the derivatives of the eigenvalues and eigenvectors using (V-5.3).

3. Use a numerical integrator over the interval $[\varepsilon - h, \varepsilon]$, in order to approximate $U(\varepsilon - h)$ and $\Lambda(\varepsilon - h)$.

4. Repeat this process until U and Λ are approximated for ε sufficiently close to zero.

When using the procedure described above, one may lose the unitary property that we initially had for $U(\varepsilon)$, namely $U(\varepsilon)^*U(\varepsilon) = I$, depending on the time-integrator used. In order to impose this property, we express $t \mapsto U(t)$ as a matrix exponential. This idea comes from [9, Sect. 2.2] and has proven useful in a theoretical context. Lemma V.5 shows the reasoning underlying the numerical algorithm we have in mind.

Lemma V.5

Let $U \in \mathbb{C}^{n\times n}$ a unitary matrix, and $A \in \mathbb{C}^{n\times n}$ an anti-Hermitian matrix. Then the matrix $e^A U$ is unitary.

Proof.

$$\left(e^A U\right)^* e^A U = U^* e^{A^*} e^A U = U^* e^{-A} e^A U = U^* U = I,$$

and

$$e^A U \left(e^A U\right)^* = e^A U U^* e^{A^*} = e^A e^{-A} = I.$$

□

We can interpret the matrix A as a kind of logarithm of U. For the numerical algorithm, we write

$$\dot{U}(\varepsilon) = U(\varepsilon) A(\varepsilon) \tag{V-5.4}$$

and we suppose A is an anti-hermitian matrix, i.e. $A^* = -A$. Equation (V-5.4) may seem like a condition imposed on U, but we are actually just using the fact that for any two vectors U, V, we can find an (invertible) matrix A such that $V = AU$. Thus, the only assumption is that A is anti-Hermitian, but we will check with its expression (V-5.7) that it is indeed the case.

Now differentiate the eigendecomposition of $M(\varepsilon)$:

$$\begin{aligned}
\frac{d}{d\varepsilon}\left(M(\varepsilon)U(\varepsilon)\right) &= \frac{d}{d\varepsilon}(U(\varepsilon)\Lambda(\varepsilon)) \\
\iff \dot{M}(\varepsilon)U(\varepsilon) + M(\varepsilon)\dot{U}(\varepsilon) &= \dot{U}(\varepsilon)\Lambda(\varepsilon) + U(\varepsilon)\dot{\Lambda}(\varepsilon) \\
\iff \dot{M}(\varepsilon)U(\varepsilon) + M(\varepsilon)U(\varepsilon)A(\varepsilon) &= U(\varepsilon)A(\varepsilon)\Lambda(\varepsilon) + U(\varepsilon)\dot{\Lambda}(\varepsilon).
\end{aligned}$$

Multiply by $U^*(\varepsilon)$ on the left, and use the relations $M(\varepsilon)U(\varepsilon) = U(\varepsilon)\Lambda(\varepsilon)$ and $U^*(\varepsilon)U(\varepsilon) = I$:

$$\begin{aligned}
U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon) + U^*(\varepsilon)M(\varepsilon)U(\varepsilon)A(\varepsilon) &= U^*(\varepsilon)U(\varepsilon)A(\varepsilon)\Lambda(\varepsilon) + U^*(\varepsilon)U(\varepsilon)\dot{\Lambda}(\varepsilon) \\
\iff U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon) + \Lambda(\varepsilon)A(\varepsilon) &= A(\varepsilon)\Lambda(\varepsilon) + \dot{\Lambda}(\varepsilon) \\
\iff U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon) &= [A(\varepsilon), \Lambda(\varepsilon)] + \dot{\Lambda}(\varepsilon), \qquad \text{(V-5.5)}
\end{aligned}$$

where $[A, B] := AB - BA$ is the matrix commutator. We have the following easy result:

Lemma V.6

Let $A \in \mathbb{C}^{n\times n}$ a matrix, and $D \in \mathbb{C}^{n\times n}$ a diagonal matrix. Then, for all $i, j = 1, \dots, n$,

$$[A, D]_{i,j} = A_{i,j}\left(d_j - d_i\right)$$

Proof. It is straightforward by writing explicitely the matrix products. □

Remark V.11: Notation

For a square matrix $C \in \mathbb{C}^{n\times n}$, we denote the vector composed of the diagonal elements of C by $\operatorname{diag}(C)$.

As a consequence of Lemma V.6, $\operatorname{diag}([A(\varepsilon), \Lambda(\varepsilon)]) = 0$, since $\Lambda(\varepsilon)$ is the diagonal matrix with eigenvalues of $M(\varepsilon)$ on its diagonal. Hence (V-5.5) yields

$$\dot{\Lambda}(\varepsilon) = U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon) \iff \operatorname{diag}\left(\dot{\Lambda}(\varepsilon)\right) = \operatorname{diag}\left(U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon)\right), \tag{V-5.6}$$

where the equivalence is due to $\Lambda(\varepsilon)$ being diagonal. On the other hand, for $i \neq j$,

$$\begin{aligned}[A(\varepsilon), \Lambda(\varepsilon)]_{i,j} &= \left(U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon) - \dot{\Lambda}(\varepsilon)\right)_{i,j} \\ &= \left(U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon)\right)_{i,j}.\end{aligned}$$

We then obtain:

$$A_{i,j}(\varepsilon) = \frac{1}{\lambda_j(\varepsilon) - \lambda_i(\varepsilon)}\left(U^*(\varepsilon)\dot{M}(\varepsilon)U(\varepsilon)\right)_{i,j}, \quad i \neq j. \tag{V-5.7}$$

Since the matrix $M(\varepsilon)$ is Hermitian, the matrix $\dot{M}(\varepsilon)$ is also Hermitian. Thus, using (V-5.7), we get that A is anti-Hermitian.

Over a small interval $[\varepsilon - h, \varepsilon]$, $0 < h < \varepsilon$, starting from $\Lambda(\varepsilon)$ we can get an approximation to $\Lambda(\varepsilon - h)$ using (V-5.6):

$$\Lambda(\varepsilon - h) \approx \Lambda(\varepsilon) - h\dot{\Lambda}(\varepsilon).$$

The diagonal matrix $\Lambda(\varepsilon - h)$ contains on its diagonal the approximate eigenvalues $\{\lambda_j(\varepsilon - h)\}_{j=1}^n$. In order to approximate the matrix $U(\varepsilon - h)$ containing the eigenvectors $\{u_j(\varepsilon - h)\}_{j=1}^n$, we shall use (V-5.4) and approximate $A(\sigma)$ by $A(\varepsilon)$ for $\sigma \in [\varepsilon - h, \varepsilon]$. We then obtain:

$$\dot{U}(\sigma) \approx U(\sigma)A(\varepsilon) \implies U(\varepsilon - h) \approx U(\varepsilon)e^{-hA(\varepsilon)} := \tilde{U}(\varepsilon - h). \tag{V-5.8}$$

By using the anti-Hermitian character of $A(\varepsilon)$ combined with Lemma V.5, we deduce that $\tilde{U}(\varepsilon - h)$ is unitary.

The interval $[\varepsilon - h, \varepsilon]$ above was not chosen at random: we want to obtain a numerical procedure, and we will need a discretization of the interval $[0, \varepsilon]$. Naturally, we divide this interval into L subintervals $[lh, (l+1)h]$, where $l = \frac{\varepsilon}{L+1}$, and use formulas (V-5.6), (V-5.7) and (V-5.8) over each subinterval.

The main advantange with this algorithm is that the unitary property is conserved at all times, but the cost to pay is to compute a matrix exponential for every subinterval. If the matrix size n becomes large, this computation may become very expensive.

In practice, in our case, one quickly faces an important issue: these equations hold for distinct eigenvalues, and if the eigenvalues are too close to each other, the division by $\lambda_i - \lambda_j$ fails. Even though this difference is nonzero, it may become too small to obtain reliable numerical results.

V-5.3 Varying the space mask

The content of this section has been designed specifically for space masks of the form $m_S = \mathbf{1}_{D_1}$ and Fourier masks of the form $\widehat{m_F} = \mathbf{1}_{D_2}$, for some some finite-volume subsets D_1 and D_2 of $\mathbb{R}^d$. We start by studying (again) the one-dimensional toy model, for which $D_1 = [-T, T]$ and $D_2 = [-\Omega, \Omega]$. We observe, through numerical experiments, that studying the spectral concentration problem on a scaled version of the interval $[-T, T]$ allows to separate eigenvalues enough so that eigenvectors can easily be obtained approximately. This fact is the core idea of the numerical procedure to be presented, which gives approximate eigenvectors.

This idea can be generalized to higher dimensional settings, the only difference is that it can be a little trickier to find a scaled version of an arbitrary finite-size subset $D_1 \subset \mathbb{R}^d$. Moreover, depending on the scaling of D_1, some parts of D_1 can be emphasized in the search of eigenvectors. An estimate of the error between the exact and approximate eigenvectors is also given.

We end this section with several one- and two-dimensional examples. They show that approximate eigenvectors can indeed be recovered in the cases where an application of classical eigensolvers fail to yield satisfying results.

V-5.3.1 Dimension $d = 1$

We consider in this section a particular type of perturbation, which consists in a scaling of the nonperturbed space mask. For the toy model in dimension one, the nonperturbed space mask is

$$m_S(x) = \mathbf{1}_{[-1,1]}.$$

We choose a perturbed space mask of the form

$$m_S^{[\varepsilon]}(x) = \mathbf{1}_{[-\mu(\varepsilon),\mu(\varepsilon)]}. \tag{V-5.9}$$

The function μ is such that $\mu \to 1$ as $\varepsilon \to 0$, and $\mu \to 0$ as $\varepsilon \to +\infty$. In the following, we have chosen

$$\mu(\varepsilon) = \frac{1}{(1+\varepsilon^4)^{1/4}},$$

and this function is plotted in Figure V-5.8. The choice of μ is rather arbitrary, but in practice it is desirable to have a flat curve around zero and a slope that is not too steep far from zero.

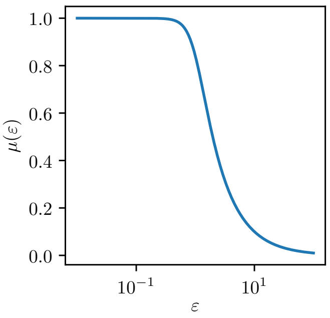


Figure V-5.8 – Scaling $\mu(\varepsilon) = (1+\varepsilon^4)^{-1/4}$.

A phenomenon that we have already illustrated is the following: when $\varepsilon \gg 1$ (i.e. $\mu(\varepsilon) \ll 1$), there is no problem in obtaining the eigenvectors since the eigenvalues are distinct and well separated. When $\varepsilon \to 0$ (i.e. $\mu(\varepsilon) \to 1$), most of the eigenvalues get packed together: some of them around one, and most of them around zero.

Let us try to understand (once again!) the reason behind this phenomenon. We consider the one-dimensional framework of Section V-2.1 – The Slepian "toy model", and a fixed given Fourier filter $\widehat{m_F} = \mathbf{1}_{D_2}$. Suppose that the perturbed space filter is of the form given by (V-5.9). If ε is small, the intervall $[-\mu(\varepsilon), \mu(\varepsilon)]$ will be "large", thus many orthogonal functions with Fourier support in D_2 will fit in $[-\mu(\varepsilon), \mu(\varepsilon)]$, and their *tails*[3] will be very small. In this case, we expect many eigenvalues (or *concentration ratios*, they are the same quantity) to be close to 1. On the other hand, if ε is large, the tails of the orthogonal functions with Fourier support in D_2 will be very big, and all the eigenvalues will be close to 0.

Another way of rephrasing this idea is that the Fourier filter $\widehat{m_F}$ gives a bound on how fast the function can vary: basically, if $D_2 = [-\Omega, \Omega]$, the variations of the function cannot be faster than those of $x \mapsto \cos(\Omega x)$.

[3] The part of the function which is outside $[-\mu(\varepsilon), \mu(\varepsilon)]$.

Remark V.12

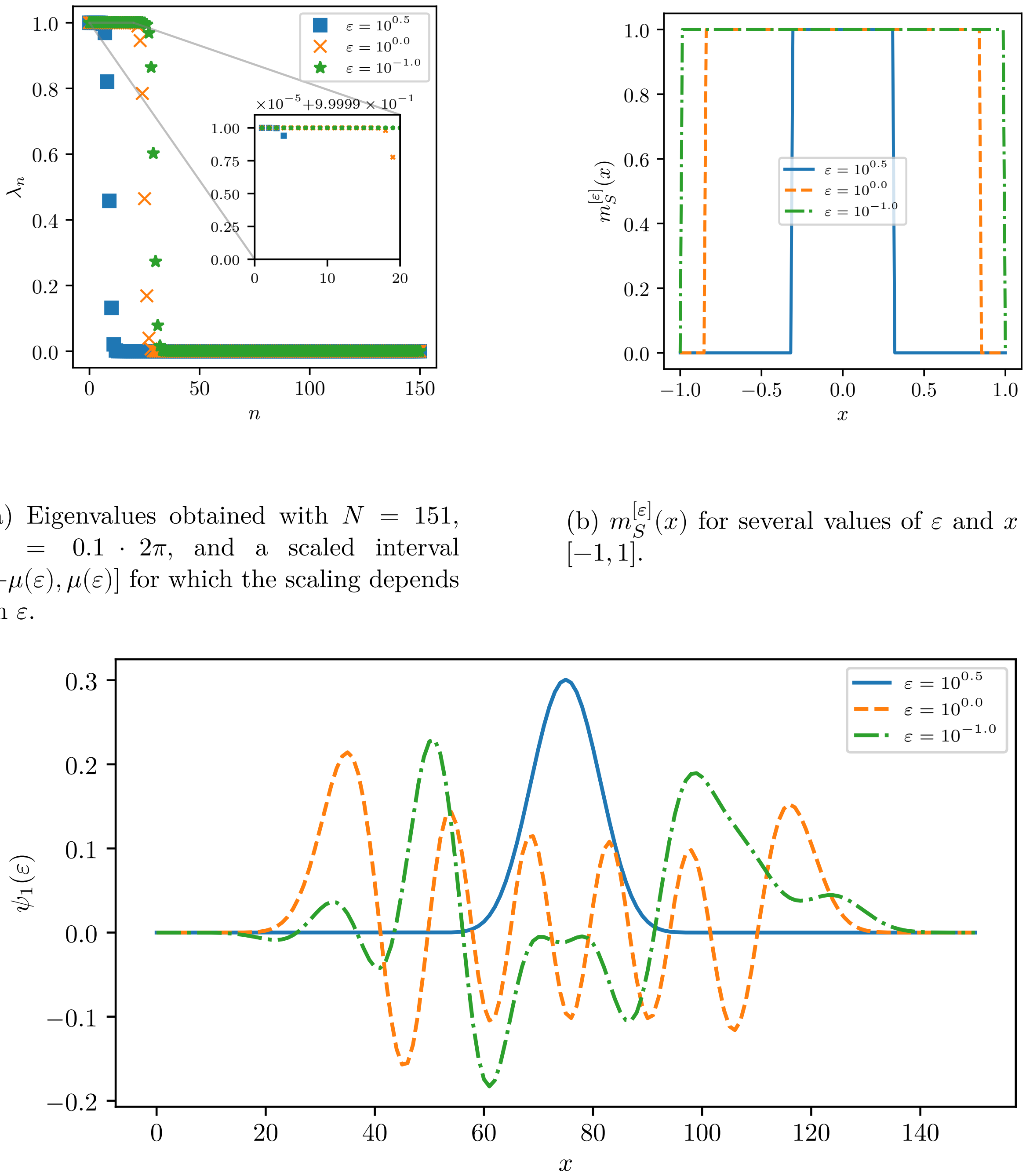
The above sentence is only true in the context of the one-dimensional toy model. Indeed when discretizing this particular case with space mask $m_S = \mathbf{1}_{[-T,T]}$, Slepian studied eigenvectors within the interval $[-T, T]$. Because only the interval of interest is studied, it is just as if there was no space mask, thus the Fourier mask applies only to the Fourier transform of the eigenvector ψ. In the general case, it should be applied to the Fourier transform of $m_S\psi$, and it is why the above sentence is specific to the one-dimensional toy model. However, it is a pretty good visual explanation of the phenomenon observed.

So, if $[-\mu(\varepsilon), \mu(\varepsilon)]$ is small, the function will not have "enough room" to go to zero on the boundary of the interval. We already illustrated this phenomenon in Figure V-5.1: the first five eigenvectors go sufficiently fast to zero so that their eigenvalues are $\lambda \approx 1$, while the three last eigenvectors cannot be zero at ±1, so their eigenvalues are $\lambda < 1$.

This "coordination" between the space and Fourier masks was already made clear in the first paper by Slepian and Pollak [46], where they noted that the eigenvalues only depend on the product WT.

What is particularly interesting to us is the behavior of those eigenvalues as functions of the product ΩT: they all converge to zero as this product goes to zero, but the smallest eigenvalues converge faster to zero. This is illustrated in Figure V-5.9a, where all the eigenvalues are plotted for several values of ε: there are more eigenvalues far from zero when $\Omega\mu(\varepsilon)$ is large than when $\Omega\mu(\varepsilon)$ is small. The space filters corresponding to these values of ε are given Figure V-5.9b.

When $\varepsilon \gg 1$, the support $[-\mu(\varepsilon), \mu(\varepsilon)]$ of $m_S^{[\varepsilon]}$ is very small, and only the eigenvectors very concentrated around the origin may have a large associated concentration ratio. For example, let's look at $\varepsilon = 10^{0.5}$ in Figure V-5.9: the space mask corresponds to the solid blue line in Figure V-5.9b, and its support is roughly $[-0.3, 0.3]$. This is the smallest support in this Figure. If we now look at Figure V-5.9a, the associated eigenvalues are represented by blue squares. There are only three eigenvalues close to one, and they can be distinguished very easily. In this case, it will be very easy to obtain the eigenvectors. On the contrary, if we look at the green dot-dash line ($\varepsilon = 10^{-1}$) in Figure V-5.9b, many



(a) Eigenvalues obtained with $N = 151$, $\Omega = 0.1 \cdot 2\pi$, and a scaled interval $[-\mu(\varepsilon), \mu(\varepsilon)]$ for which the scaling depends on ε.

(b) $m_S^{[\varepsilon]}(x)$ for several values of ε and $x \in [-1, 1]$.

(c) First eigenvector obtained with $N = 151$, $\Omega = 0.1 \cdot 2\pi$, and a scaled interval $[-\mu(\varepsilon), \mu(\varepsilon)]$ for which the scaling depends on ε.

Figure V-5.9 – Eigenvalues (left) and spacemask (right) corresponding to several values of the parameter ε.

eigenvalues (green asterisks[4] 🤡) are gathered close to one. In this case, it will be difficult to distinguish eigenvectors. Between these two extreme cases, we note that, as ε decreases, the support $[-\mu(\varepsilon), \mu(\varepsilon)]$ of $m_S^{[\varepsilon]}$ gets larger and larger, and the number of eigenvalues close to one increases.

This is at the heart of the procedure we will describe now: we start from a very narrow space mask, for which only a very small number of eigenvalues are close to one. We can easily obtain the most significant eigenvector v of the "scaled matrix" $\mathbf{K}^{[\varepsilon]}$, and then check how this eigenvector is concentrated by computing the concentration ratio

$$\alpha = \frac{v^* \mathbf{K}^{[0]} v}{v^* v}.$$

with respect to the true concentration matrix $\mathbf{K}^{[0]}$.

If this concentration ratio α is close enough to the exact eigenvalue, say within a tolerance $\eta > 0$, we save this eigenvector and proceed to the next one.

Remark V.13

We call *exact eigenvalues* the eigenvalues of the unperturbed operator $\mathbf{K}^{[0]}$. It is important to note and understand that the whole problem lies in the computation of eigenvectors, and that obtaining the eigenvalues precisely is not a problem at all. This is why we can safely compare the concentration ratio α with the largest eigenvalue of $\mathbf{K}^{[0]}$.

The next one is obtained by looking at an eigenvector associated to the eigenvalue of highest magnitude, and such that this eigenvector is orthogonal to the previously saved one. If the concentration ratio is not close enough to the exact eigenvalue, we let ε be a little smaller and repeat this. The whole procedure is summed up in Algorithm 6.

It is clear from the algorithm that the set of vectors obtained at the end is an orthonormal basis. Moreover, the concentration ratios we obtain are as close to the true eigenvalues of $\mathbf{K}^{[0]}$ as desired, thanks to the numerical threshold η. In Algorithm 6, the number of recorded vectors can also be chosen, so that one is able to only compute the

[4] And not Astérix

most significant eigenvectors. This can even be done “on-the-fly”, by stopping the algorithm as soon as the concentration ratio becomes too small. Moreover, since we are only interested in the eigenvector associated to the largest eigenvalue of $\mathbf{K}^{[\varepsilon]}$ (with orthogonality conditions), one can take advantage of the Toeplitz nature of the matrix and, for instance, use the Power Method efficiently.

Remark V.14

We can summarize Algorithm 6 as follows: it performs an approximate eigendecomposition of the unperturbed operator $\mathbf{K}^{[0]}$, and the approximation is made so that it bypasses the numerical issues caused by the very close eigenvalues of $\mathbf{K}^{[0]}$. It is more stable than usual eigendecomposition algorithms in the following sense: two different algorithms (and most possibly two applications of the same algorithm) will yield different sets of eigenvectors, while Algorithm 6 will always give the same eigenvectors. Moreover, Algorithm 6 is built upon usual eigenalgorithms which makes it easy to understand and use. The method used to obtain Algorithm 6 was based on the physical interpretation of the spectral concentration problem.
However, now that was have devised this algorithm, we could try to apply it to other situations where we lack physical meaning: given an arbitrary matrix with very close eigenvalues, how to obtain associated eigenvectors in a stable way? Algorithm 6 is a possible answer, and it will yield approximate eigenvectors. However, some numerical tests have to be performed in more general situations to check that Algorithm 6 indeed yields satisfying results.

Another way of looking at Algorithm 6 consists in saying that, among all linear combinations of eigenvectors associated to an eigenvalue ≈ 1, we are looking for the ones with increasing variance.

Let us discuss now the reasoning behind Algorithm 6. Before that, we note that for any vector $v \in \mathbb{C}^n$, we define $|v|_2 := \sqrt{v^* v}$.

Lemma V.7

Let A be an Hermitian matrix, and $w \in \mathbb{C}^N$ such that $|w|_2 = 1$. Let $\lambda_1 \in \mathbb{R}$ the largest eigenvalue (real) of A, then

$$w^* A w \leq \lambda_1 .$$

Algorithm 6 Varying space mask method (Dimension $d = 1$)

```
Input
— $\mu(\varepsilon)$: a scaling function, such that $\mu$ is decreasing, $\mu(0) = 1$, $\mu(+\infty) = 0$.
— $\varepsilon \mapsto \mathbf{K}^{[\varepsilon]}$: the perturbed concentration matrix, of size $N \times N$.
— $M$: number of eigenvectors we are looking for, $M \leq N$.
— $\varepsilon_{\max}$: maximum value of the concentration parameter.
— $T$: number of subdivisions of the interval $[0, \varepsilon_{\max}]$.
— $\lambda(0) = (\lambda_1(0), \dots, \lambda_N(0))$: vector of eigenvalues for the nonperturbed matrix $K$.
— $\eta$: numerical tolerance to compare two eigenvalues.
$h := \varepsilon_{\max}/T$
$q := 0$: this is the number of recorded eigenvectors.
$\alpha^{[saved]} = \left(\alpha^{[saved]}, \dots, \alpha_M^{[saved]}\right)$: the vector to hold all the concentration ratios.
$v^{[saved]} = \left(v_1^{[saved]}, \dots, v_M^{[saved]}\right)$: the matrix to hold all the recorded eigenvectors
for $\varepsilon = \varepsilon_{\max}, \varepsilon_{\max} - h, \dots, 0$ do
    Find $(\kappa, u)$ the most significant eigenpair of $\mathbf{K}^{[\varepsilon]}$ (i.e. the one associated to the eigenvalue of highest magnitude), where $u \perp \text{Span}\left\{v_1^{[saved]}, \dots, v_q^{[saved]}\right\}$.
    Compute the concentration ratio with respect to the unperturbed problem:
        $$\beta := \frac{u^* \mathbf{K}^{[0]} u}{u^* u}.$$
    Add complex scaling/orthonormalization.
    if $\left|\beta - \lambda_q(0)\right| \leq \eta$ then
        $v_q^{[saved]} \leftarrow u/\|u\|$.
        $\alpha_q^{[saved]} \leftarrow \beta$.
        $q \leftarrow q + 1$.
    end if
    Stop if $q > M$.
end for
```

Proof. Decompose w into a unitary basis of A:

$$w = \sum_{i=1}^{N} c_i v_i.$$

Then

$$\begin{aligned} w^* A w &= \left(\sum_{i=1}^{N} \overline{c_i} v_i^*\right) A \left(\sum_{j=1}^{N} c_j v_j\right) = \left(\sum_{i=1}^{N} \overline{c_i} v_i^*\right)\left(\sum_{j=1}^{N} c_j \lambda_j v_j\right) \\ &= \sum_{i,j=1}^{N} \overline{c_i} c_j \lambda_j v_i^* v_j = \sum_{i=1}^{N} |c_i|^2 \lambda_i \leq \lambda_1 \sum_{i=1}^{N} |c_i|^2. \end{aligned} \tag{V-5.10}$$

To obtain the claim, it only suffices to recall that $\sum_{i=1}^{N} |c_i|^2 = w^* w = 1$, since $\left\{v_j\right\}_j$ is an unitary basis of $\mathbb{C}^N$. □

Lemma V.8

Let $A \in \mathbb{R}^{N\times N}$ an Hermitian matrix, and denote $\{\lambda_i\}_{i=1}^N$ its eigenvalues (real), ordered so that $\lambda_i \geq \lambda_{i+1}$. Let $\{v_1, \dots, v_N\}$ an orthonormal basis of $\mathbb{C}^N$, where $v_i \in \mathbb{C}^N$ is an eigenvector of A associated to λ_i. Let $\eta > 0$, $w \in \mathbb{C}^N$ such that $|w|_2 = 1$, and

$$\left|w^T A w - \lambda_1\right|_2 \leq \eta.$$

Then

$$|w - v_1|_2^2 = \mathcal{O}\left(\frac{\eta}{\lambda_1 - \lambda_2}\right)$$

Proof. There exist coefficients $c_1, \dots, c_N \in \mathbb{C}^N$ such that

$$w = \sum_{i=1}^N c_i v_i.$$

Using (V-5.10), we get

$$\lambda_1 - w^T A w = \lambda_1 - \sum_{i=1}^N |c_i|^2 \lambda_i = \lambda_1 \left(1 - |c_1|^2\right) - \sum_{i=2}^N |c_i|^2 \lambda_i \geq \lambda_1 \left(1 - |c_1|^2\right) - \lambda_2 \sum_{i=2}^N |c_i|^2.$$

We use the fact that $|w|_2 = 1$, which implies

$$\sum_{i=1}^N |c_i|^2 = 1,$$

hence

$$\begin{aligned}\lambda_1 - w^T A w \geq \lambda_1 \left(1 - |c_1|^2\right) - \lambda_2 \sum_{i=2}^N |c_i|^2 &= \lambda_1 \left(1 - |c_1|^2\right) - \lambda_2 \left(1 - |c_1|^2\right) \\ &= \left(\lambda_1 - \lambda_2\right)\left(1 - |c_1|^2\right).\end{aligned}$$

Using Lemma V.7, our assumption gives

$$\eta \geq |w^T A w - \lambda_1|_2 = \lambda_1 - w^T A w \geq \left(\lambda_1 - \lambda_2\right)\left(1 - |c_1|^2\right),$$

hence

$$1 \geq |c_1|^2 \geq 1 - \frac{\eta}{\lambda_1 - \lambda_2}.$$

Furthermore,

$$\begin{aligned}|w - v_1|_2^2 &= \left|\sum_{i=1}^{N} c_i v_i - v_1\right|_2^2 = \left|(c_1 - 1)v_1 + \sum_{i=2}^{N} c_i v_i\right|_2^2 \\ &= |c_1 - 1|^2 + \sum_{i=2}^{N} |c_i|^2 = |c_1 - 1|^2 + 1 - |c_1|^2 \\ &= 2(1 - c_1).\end{aligned}$$

Up to a change of sign, we can assume $c_1 \geq 0$. Then,

$$|w - v_1|_2^2 \leq 2\left(1 - \sqrt{1 - \frac{\eta}{\lambda_1 - \lambda_2}}\right),$$

and thus, for $\frac{\eta}{\lambda_1 - \lambda_2}$ small,

$$|w - v_1|_2^2 = \mathcal{O}\left(\frac{\eta}{\lambda_1 - \lambda_2}\right)$$

□

One issue with Lemma V.8 is that it only cares about the first eigenpair. Fortunately, a generalization is straightforward when we suppose the first eigenvalues to be packed together, see Lemma V.9.

Lemma V.9

Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ an Hermitian matrix, and denote $\{\lambda_i\}_{i=1}^N$ its eigenvalues (real), ordered so that $\lambda_i \geq \lambda_{i+1}$. Let $\{v_1, \dots, v_N\}$ an orthonormal basis of $\mathbb{C}^N$, where $v_i \in \mathbb{C}^N$ is an eigenvector of $\mathbf{A}$ associated to λ_i. Let $\eta > 0$, $w \in \mathbb{C}^n$ such that $|w|_2 = 1$, and

$$|w^*\mathbf{A}w - \lambda_1|_2 \leq \eta.$$

We assume $\lambda_1 = \lambda_2 + \mathcal{O}(\tau) = \dots = \lambda_m + \mathcal{O}(\tau)$ for some $m \leq N$ and some $\tau > 0$. Then

$$\left|w - \mathrm{Proj}_{\mathrm{Span}\ \{v_1,\dots,v_m\}} w\right|_2^2 = \mathcal{O}\left(\frac{\eta + \mathcal{O}(\tau)}{\lambda_1 - \lambda_{m+1}}\right).$$

Proof. Decompose w into the $\{v_i\}_{i=1}^N$ basis:

$$w = \sum_{i=1}^{N} c_i v_i,$$

for some coefficients $c_i \in \mathbb{C}$. We have

$$\begin{aligned} w^* \mathbf{A} w &= \left(\sum_{i=1}^{N} c_i v_i^T \right) \left(\sum_{j=1}^{N} c_j \lambda_j v_j \right) = \sum_{i=1}^{N} |c_i|^2 \lambda_i \\ &= \lambda_1 \left(\sum_{i=1}^{m} |c_i|^2 \right) + \mathcal{O}(\tau) \sum_{i=2}^{m} |c_i|^2 + \sum_{i=m+1}^{N} |c_i|^2 \lambda_i \\ &= \lambda_1 \left(\sum_{i=1}^{m} |c_i|^2 \right) + \mathcal{O}(\tau) + \sum_{i=m+1}^{N} |c_i|^2 \lambda_i, \end{aligned}$$

where the last equality is due to having $|w|_2 = 1$. Then

$$\begin{aligned} \lambda_1 - w^* \mathbf{A} w &= \lambda_1 \left(1 - \sum_{i=1}^{m} |c_i|^2 \right) + \mathcal{O}(\tau) - \sum_{i=m+1}^{N} |c_i|^2 \lambda_i \\ &\geq \lambda_1 \left(1 - \sum_{i=1}^{m} |c_i|^2 \right) + \mathcal{O}(\tau) - \lambda_{m+1} \sum_{i=m+1}^{N} |c_i|^2 \\ &\geq \lambda_1 \left(1 - \sum_{i=1}^{m} |c_i|^2 \right) + \mathcal{O}(\tau) - \lambda_{m+1} \left(1 - \sum_{i=1}^{m} |c_i|^2 \right) \\ &\geq (\lambda_1 - \lambda_{m+1}) \left(1 - \sum_{i=1}^{m} |c_i|^2 \right) + \mathcal{O}(\tau), \end{aligned}$$

and thus

$$1 - \sum_{i=1}^{m} |c_i|^2 \leq \frac{\lambda_1 - w^* \mathbf{A} w}{\lambda_1 - \lambda_{m+1}} \leq \frac{\eta + \mathcal{O}(\tau)}{\lambda_1 - \lambda_{m+1}}.$$

We now compute

$$\left| w - \mathrm{Proj}_{\mathrm{Span}\ \{v_1, \ldots, v_m\}} w \right|_2^2 = \left| \sum_{i=m+1}^{N} c_i v_i \right|_2^2 = \sum_{i=m+1}^{N} |c_i|^2 = 1 - \sum_{i=1}^{m} |c_i|^2 \leq \frac{\eta + \mathcal{O}(\tau)}{\lambda_1 - \lambda_{m+1}}.$$

□

In Lemma V.9, the first eigenvalues are assumed to be bunched together while $\lambda_1 - \lambda_{m+1}$ is supposedly large (i.e. λ_{m+1} is supposedly "far" from λ_1). This situation is close to the framework exhibited in the spectral concentration problem, where eigenvalues are close to zero or one and only a very small number of them being in between.

All the discussion until now was concerned with the one-dimensional problem of an arbitrary Fourier mask $\widehat{m_F}$ and a space mask m_S, assumed of the form

$$m_S = \mathbf{1}_{D_1}, \quad \widehat{m_F} = \mathbf{1}_{D_2}$$

with $D_1 = [-1,1]$ and $D_2 = [-\Omega,\Omega]$ The ideas explained are not specific to the one-dimensional case, and we can apply them as well in the multidimensional case.

V-5.3.2 Dimension $d \geq 1$

We now suppose D_1 to be a finite-volume subset of $\mathbb{R}^d$, $d \geq 1$. We are interested in applying the same ideas as in the one-dimensional case, more specifically the fact that studying the eigenproblem on a scaled version of D_1 may help in having distinguishable eigenvalues (and thus, obtaining eigenvectors easily).

We write $D_1(\varepsilon)$ a set-valued scaling such that

$$D_1(0) = D_1, \qquad D_1(\varepsilon_1) \subset D_1(\varepsilon_2) \text{ for } \varepsilon_1 > \varepsilon_2, \qquad \text{and} \qquad D_1(+\infty) = \{0\}\,.$$

When ε is small, $|D_1(\varepsilon) - D_1|$ is assumed to be small as well. Note the abuse of notation here, where D_1 denotes a subset of $\mathbb{R}^d$ and $D_1(\cdot)$ a set-valued function.

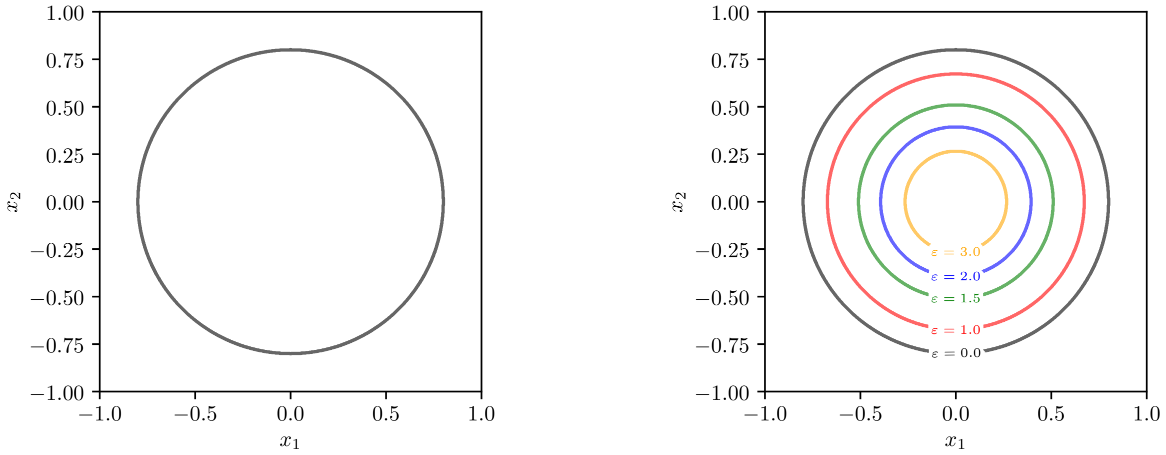
Remark V.15

In the one-dimensional case, using the previously defined notations, we would have

$$D_1 = [-1,1] \qquad \text{and} \qquad D_1(\varepsilon) = [-\mu(\varepsilon),\mu(\varepsilon)].$$

In the two-dimensional case, we give some examples of set-valued functions $D_1(\varepsilon)$ in Figures V-5.10, V-5.11, and V-5.12.

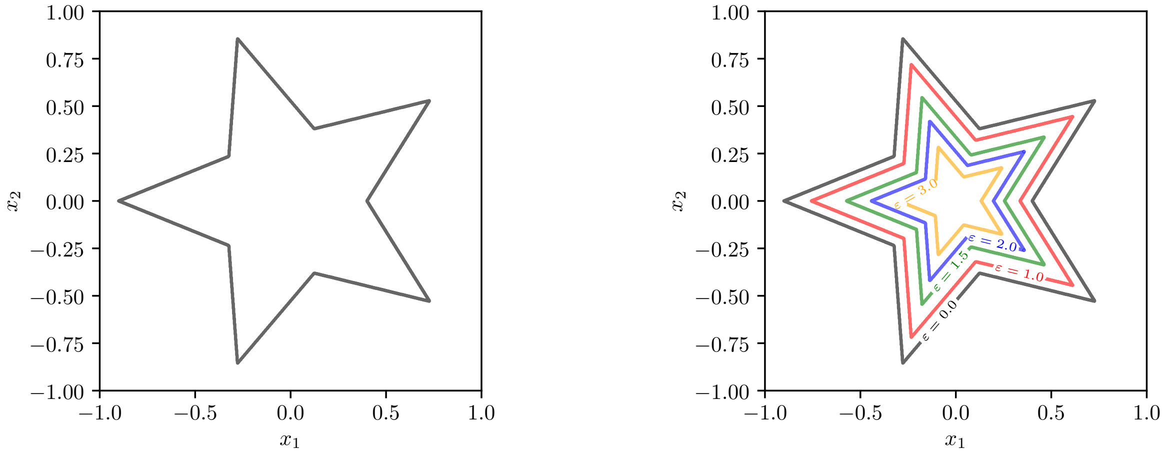
We can check numerically in Figures V-5.13, V-5.14 and V-5.15 the fact that, when $\varepsilon \to \infty$, the eigenvalues all converge to zero. This is expected from the eigenproblem (V-4.10), because the support of $m_S^{[\varepsilon]}$ becomes closer and closer to a null set. See Remark V.5 for more details. Since the behavior in the multidimensional case is qualitatively the same as in the one-dimensional case, we can apply the same ideas in order to obtain approximate eigenvectors: start from a large $\varepsilon > 0$, and look for the eigenvector associated to the largest eigenvalue of $\mathbf{K}^{[\varepsilon]}$. If the concentration ratio of this vector with respect to the unscaled matrix $\mathbf{K}^{[0]}$ is close enough to the first eigenvalue of $\mathbf{K}^{[0]}$, keep this vector v_1. Otherwise, decrease ε by a little amount and do the same steps. Once a vector has been saved, we look for the second eigenvector associated to the second largest eigenvalue of $\mathbf{K}^{[\varepsilon]}$, with an orthogonality condition with respect to v_1. The numerical procedure described in Algorithm 7 simply consists in applying these steps as many times as necessary.



(a) Disc$(0, 0.8)$

(b) Set-valued function $D_1(\varepsilon)$, with $D_1 = D_1(0) = \text{Disc}(0, 0.8)$.

Figure V-5.10 – The disc centered at origin and of radius 0.8, written Disc$(0, 0.8)$, as well as the set-valued function $D_1(\varepsilon)$, decreasing for the inclusion relation, and such that $D_1(0) = \text{Disc}(0, 0.8)$.



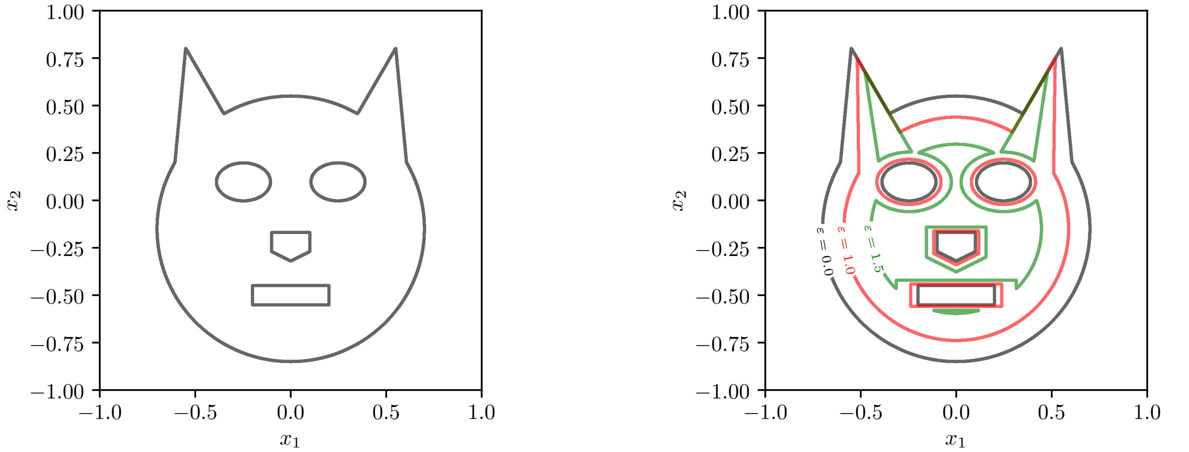
(a) Star.

(b) Set-valued function $D_1(\varepsilon)$, with $D_1 = D_1(0) = \text{Star}$.

Figure V-5.11 – A 5-branch star, as well as the set-valued function $D_1(\varepsilon)$, decreasing for the inclusion relation, and such that $D_1(0) = \text{Star}$.

Algorithm 7 Varying space mask method (Dimension $d \geq 1$)

```
Require:
  — $D_1(\varepsilon)$: a set-valued function, such that $D_1 : \varepsilon \mapsto D_1(\varepsilon)$ is decreasing (for the relation
    of set inclusion), $D_1(0) = D_1$, $D_1(+\infty) = 0$.
  — $\varepsilon \mapsto \mathbf{K}^{[\varepsilon]}$: the "scaled" concentration matrix, of size $N \times N$, with the space mask
    being $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$.
  — $M$: number of eigenvectors we are looking for, $M \leq N$.
  — $\varepsilon_{\max}$: maximum value of the parameter $\varepsilon$.
  — $\{\varepsilon_T, \dots, \varepsilon_0\} \subset [0, \varepsilon_{\max}]$: $T+1$ discretization points of the interval $[0, \varepsilon_{\max}]$ (can be a
    uniform discretization, log-uniform, ...). They are assumed to be such that $\varepsilon_t > \varepsilon_{t-1}$,
    $t \in [\![1, T]\!]$.
  — $\lambda_{approx}(0) = (\lambda_{approx,1}(0), \dots, \lambda_{approx,N}(0))$: vector of approximate eigenvalues for
    the nonperturbed matrix $\mathbf{K}^{[0]}$.
  — $\eta$: numerical tolerance to compare two eigenvalues.
  $q := 0$: this is the number of recorded eigenvectors.
  $\alpha_{saved} = (\alpha_{saved,1}, \dots, \alpha_{saved,M})$: the vector to hold all the concentration ratios.
  $v_{saved} = (v_{saved,1}, \dots, v_{saved,M})$: the matrix to hold all the recorded eigenvectors
  for $\varepsilon = \varepsilon_T, \dots, \varepsilon_0$ do    ▷ (i.e. first $\varepsilon = \varepsilon_T$, then $\varepsilon = \varepsilon_{T-1}$, ..., until $\varepsilon = \varepsilon_0$).
      Find $(\kappa, u)$ the most significant eigenpair of $\mathbf{K}^{[\varepsilon]}$ (i.e. the one associated to the
  eigenvalue of highest magnitude), where $u \perp \text{Span } \{v_{saved,1}, \dots, v_{saved,q}\}$.
      Compute the concentration ratio with respect to the unperturbed problem:
          $$\beta := \frac{u^* \mathbf{K}^{[0]} u}{u^* u}.$$
      Add complex orthonormalization.
      if $|\beta - \lambda_{approx,q}(0)| \leq \eta$ then
          $v_{saved,q} \leftarrow u / \|u\|$.
          $\alpha_{saved,q} \leftarrow \beta$.
          $q \leftarrow q + 1$.
      end if
      Stop if $q > M$.
  end for
```



(a) Cat-head shape.

(b) Set-valued function $D_1(\varepsilon)$, with $D_1 = D_1(0) = $ Cat-head.

Figure V-5.12 – A (poorly drawn) cat-head shape, as well as the set-valued function $D_1(\varepsilon)$, decreasing for the inclusion relation, and such that $D_1(0) = $ Cat-head.

Remark V.16

In practice, obtaining the "exact" eigenvalues $\lambda_i(0)$ (i.e. up to machine precision) can be difficult if the matrix is large in Algorithm 7. This is purely a problem of computational power available, which we did not mention in dimension one but would occur if N is really large. In dimension two and higher, this problem quickly occurs and this is why we mention it here, and also why we used approximate eigenvalues $\lambda_{approx,i}$ in Algorithm 7. In Algorithm 6, one can understand the "true" eigenvalues as approximate eigenvalues for which the approximation error is of the order of the machine epsilon. In order to obtain the approximate eigenvalues, one can for instance perform an iterative search of eigenvalues (e.g. the power method) and stop whenever the convergence of eigenvalue is considered good enough. A simple criterion could be when the difference between one approximation and the next one is smaller than some threshold. In our numerical experiments, we looked for approximate eigenvalues such that $|\lambda_{approx,i} - \lambda_i(0)| \leq \eta$, hence

$$\left|\alpha_{saved,i} - \lambda_i(0)\right| \leq 2\eta.$$

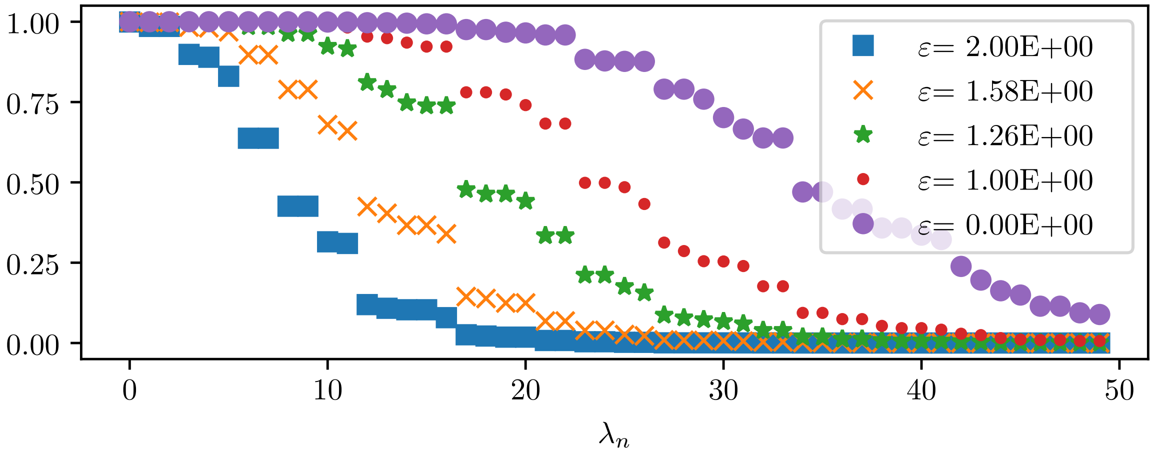


Figure V-5.13 – The 50 first largest eigenvalues of the concentration matrix $\mathbf{K}^{[\varepsilon]}$, for several values of ε. We recall that the space mask used for $\mathbf{K}^{[\varepsilon]}$ is $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$. The set-valued function $D_1(\varepsilon)$ is illustrated in Figure V-5.10. The Fourier space mask used here is $\widehat{m_F} = \mathbf{1}_{\text{Disc}(0,0.1\cdot 2\pi)}$. $N_1 = N_2 = 50$ discretization points were used in each dimension.

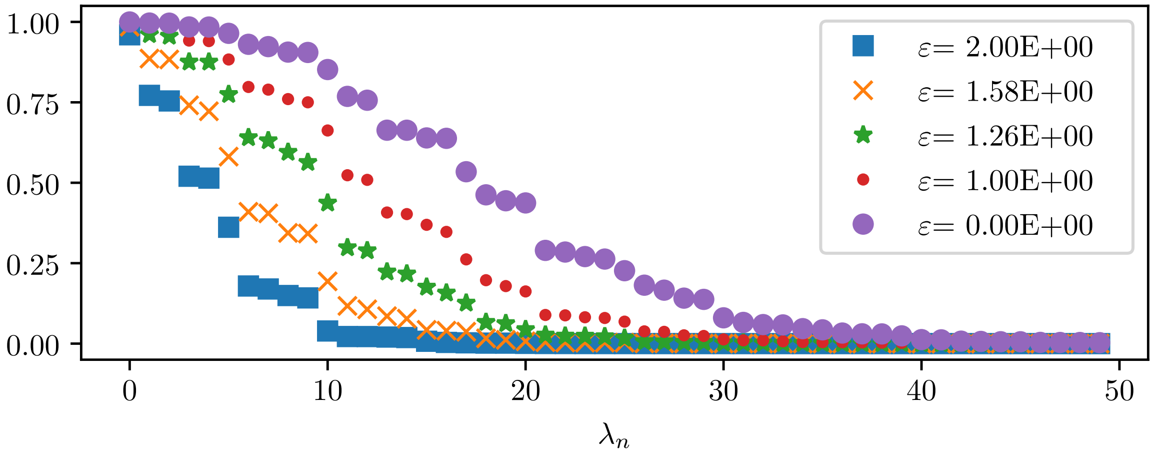


Figure V-5.14 – The 50 first largest eigenvalues of the concentration matrix $\mathbf{K}^{[\varepsilon]}$, for several values of ε. We recall that the space mask used for $\mathbf{K}^{[\varepsilon]}$ is $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$. The set-valued function $D_1(\varepsilon)$ is illustrated in Figure V-5.11. The Fourier space mask used here is $\widehat{m_F} = \mathbf{1}_{\text{Disc}(0,0.1\cdot 2\pi)}$. $N_1 = N_2 = 50$ discretization points were used in each dimension.

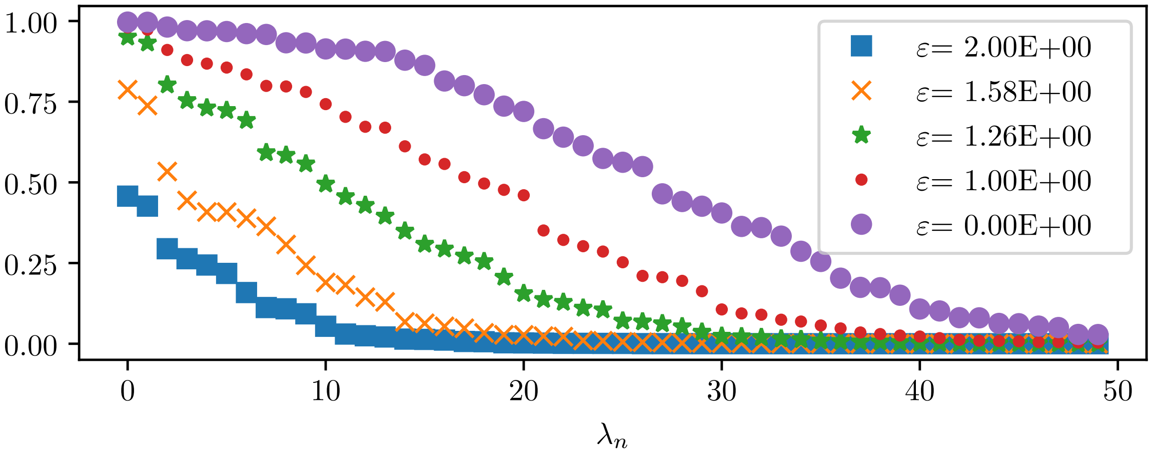


Figure V-5.15 – The 50 first largest eigenvalues of the concentration matrix $\mathbf{K}^{[\varepsilon]}$, for several values of ε. We recall that the space mask used for $\mathbf{K}^{[\varepsilon]}$ is $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$. The set-valued function $D_1(\varepsilon)$ is illustrated in Figure V-5.12. The Fourier space mask used here is $\widehat{m_F} = \mathbf{1}_{\mathrm{Disc}(0,0.1\cdot 2\pi)}$. $N_1 = N_2 = 50$ discretization points were used in each dimension.

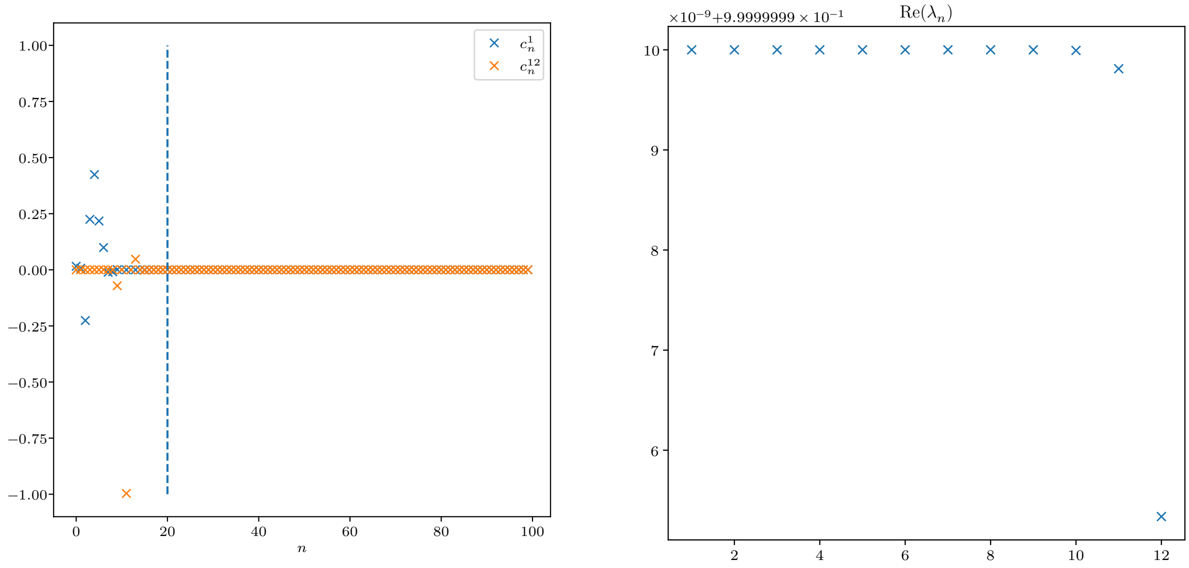
V-5.3.3 Numerical results

We present in this section the numerical results obtained with the procedure consisting in modifying the boundaries of the space mask. In dimension one, this procedure is described by Algorithm 6, and for dimension two by Algorithm 7 (which is simply a multidimensional generalization of Algorithm 6).

One-dimensional intervals

We give in Figure V-5.17 the first twelve eigenpairs of $\mathbf{K}^{[0]}$, with a Fourier restriction to $[-\Omega, \Omega]$, $\Omega = 0.1 \cdot 2\pi$, and a space restriction to $[-1, 1]$ with $N = 100$ discretization points in space. The solid blue curve gives the eigenvectors obtained using a classical eigenalgorithm, while the dash orange lines give the reference eigenvectors obtained from Slepian's commuting operator approach. We can see that the first seven solid blue eigenvectors seem like combinations of the dash orange ones. For $n \geq 8$, the eigenvalues λ_n are sufficiently far from one so that they are numerically distinct.

Slepian *et al.* showed in [46] that all eigenvalues of the toy model are unique, and one can then deduce from the toy-model kernel (V-2.2) that eigenfunctions are either even or odd. This can be understood as a specialized version of the symmetry Lemma V.4 in dimension one. This is also a proof that in this case, a direct eigendecomposition yields wrong eigenvectors: indeed, the eigenvectors recovered with a direct eigendecomposition are neither even nor odd.



(a) Coefficients of the expansion of v_1 and v_{12} into the Slepian basis $\{\psi_n\}_{n=1}^N$.

(b) First twelve exact eigenvalues (up to machine precision).

Figure V-5.16 – Eigenvalues and coefficient expansions for $N = 100$ and $\Omega = 0.1 \cdot 2\pi$, obtained with a classical eigendecomposition.

Let us check that the solid blue curves are indeed linear combinations of the dash orange curves, for the eigenvectors with indices 1 and 12. Let us denote by $\{\psi_n\}_{n=1}^N$ the eigenvector basis that one obtains from the solution given by Slepian, and $\{\lambda_n\}_{n=1}^N$ the corresponding eigenvalues (or concentration ratios). Let $\{v_n\}_{n=1}^N$ the eigenvectors obtained with a direct eigendecomposition. Since $\{\psi_n\}_{n=1}^N$ is a basis of $\mathbb{R}^N$, we can decompose

$$v_j = \sum_{i=1}^N c_i^j \psi_i.$$

Due to the numerical confusion among eigenvalues very close to 1, we expect that most expansion coefficients c_i^1 are nonzero for $i = 1, \dots, I$, and then all zero. This index I is the largest index such that a computer mistakes λ_I for λ_1. We know that $I \leq \lfloor \frac{N\Omega}{\pi} \rfloor$, and this upper bound is drawn using a vertical line. On the other hand, we expect $c_{12}^{12} \approx 1$. We give in Figure V-5.16a the expansion coefficients $\{c_i^1\}_i$ and $\{c_i^{12}\}_i$ of v_1 and v_{12} into the $\mathbb{R}^N$ basis $\{\psi_n\}_{n=1}^N$, and we can check that our intuition was indeed correct. In Figure V-5.16b, we display the eigenvalues corresponding to the eigenvectors $\{v_i\}_{i=1}^{12}$, obtained with a classical eigendecomposition.

On the other hand, Figure V-5.18 gives the approximate eigenvectors obtained using the varying space mask procedure. They are qualitatively much closer to the reference eigenvectors, and in particular exhibit the correct number of “bumps”. They are also more concentrated around the origin than expected, but this is due to the fact that the space mask gets larger and larger symetrically from the origin, so the blue eigenvectors displayed do not correspond exactly to the space mask $\mathbf{1}_{[-1,1]}$ but to $\mathbf{1}_{[-\mu(\varepsilon_n),\mu(\varepsilon_n)]}$, for some $\varepsilon_n > 0$ (and thus for some $\mu(\varepsilon_n) < 1$).

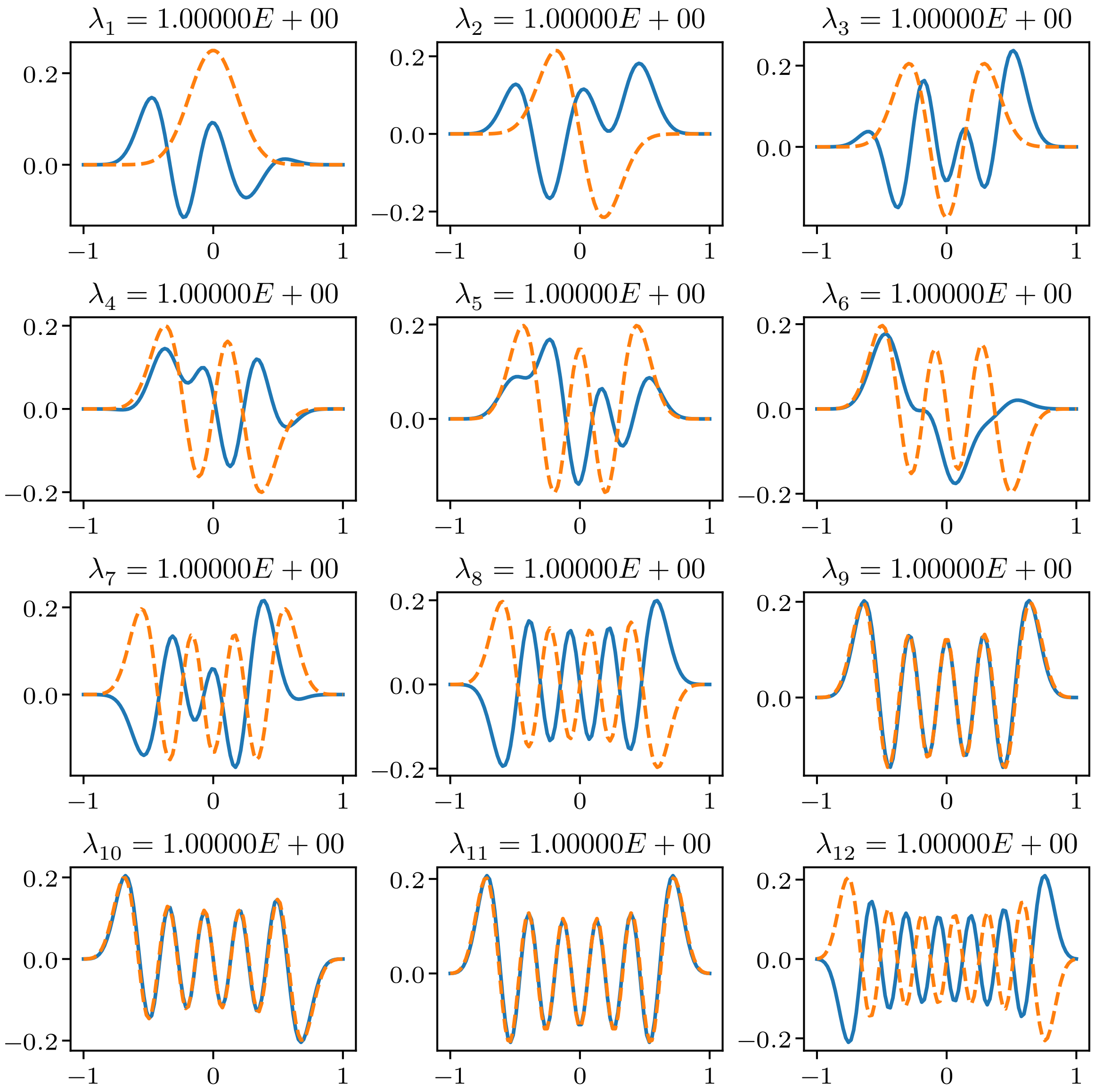


Figure V-5.17 – First 12 eigenpairs with an eigendecomposition of $\mathbf{K}^{[0]}$ (solid blue curve), in 1D, with $N = 100$, $\Omega = 0.1 \cdot 2\pi$. The exact eigenvectors are given by orange dash curves.

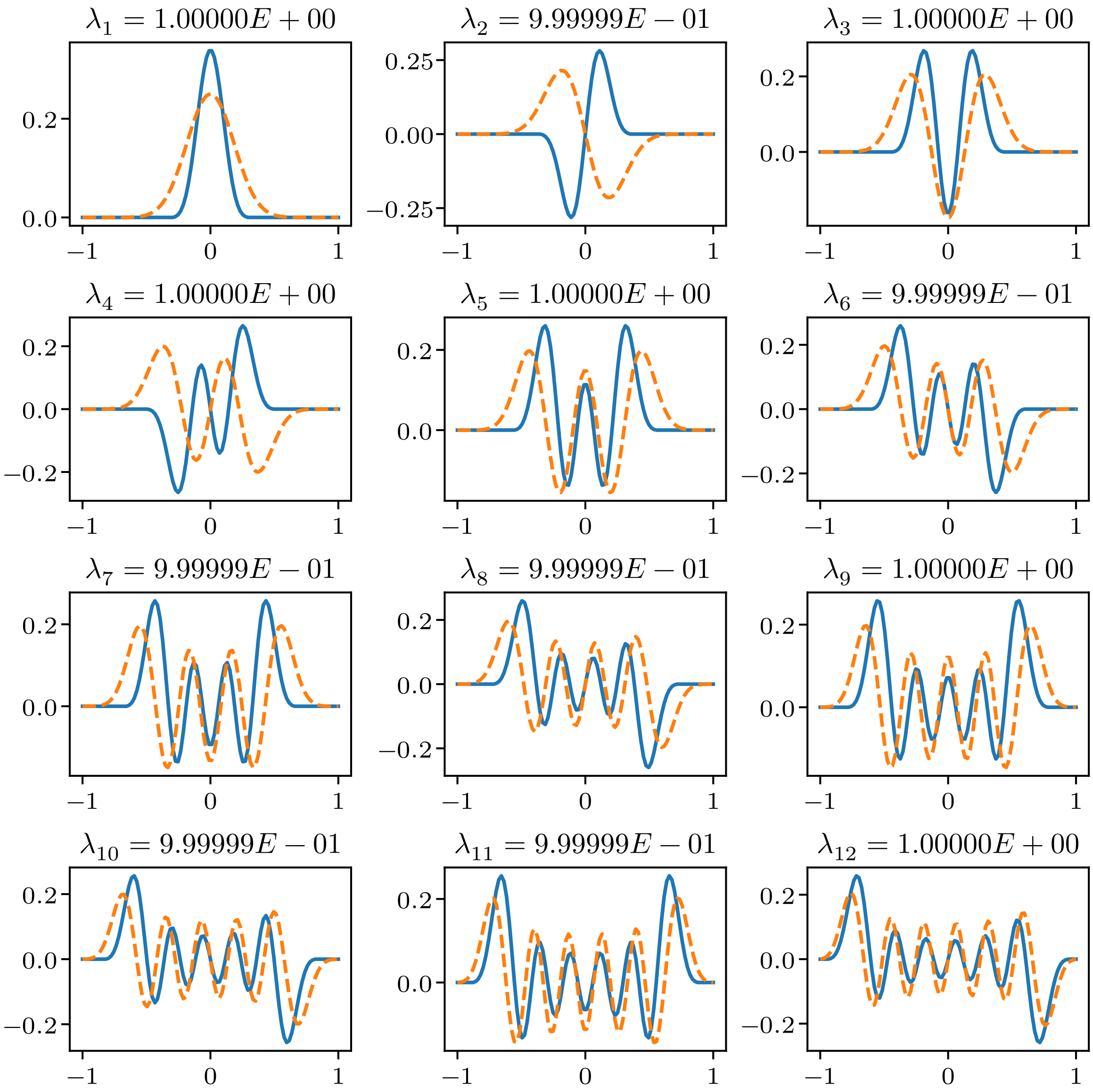


Figure V-5.18 – First 12 eigenpairs with the varying spacemask procedure (solid blue curve), in 1D, with $N = 100$, $\Omega = 0.1 \cdot 2\pi$. The exact eigenvectors are given by orange dash curves.

Two-dimensional examples

All the two-dimensional examples we present here are Fourier restricted to the same two-dimensional ball, written as $\mathrm{Disc}(c, \Omega)$ to denote a disc centered at $c \in \mathbb{R}^d$ with radius $\Omega > 0$. For the results shown, we have chosen $c = (0, 0)$ and $\Omega = 0.3$, i.e. $\widehat{m_{F,2\pi}} = \mathbf{1}_{\mathrm{Disc}(0,0.3)}$. For the ε-discretization, we have used a log-uniform discretization of the interval $[10^{-3}, 10^{1}]$ containing 500 points.

In the following, we give the results obtained with three differents space masks (disc, cat-head, and star), and for each example we show four figures:

— the first figure gives twelve eigenvectors corresponding to the sixteen largest eigenvalues of $\mathbf{K}^{[0]}$, obtained using a classical eigenalgorithm;

— the second figure shows the Fourier transform of the approximate eigenvectors obtained with a classical eigenalgorithm;

— the third figure shows twelve approximate eigenvectors corresponding to the twelve largest eigenvalues of $\mathbf{K}^{[0]}$, obtained using the varying space mask procedure described by Algorithm 7;

— the fourth figure shows the Fourier transform of the approximate eigenvectors obtained with the varying mask procedure.

For each space mask, the results are:

— $D_1 = \text{Disc}(0, 0.8)$: Figures V-5.19, V-5.20, V-5.21, V-5.22. The set-valued function $D_1(\varepsilon)$ that we used, such that $D_1(0) = \text{Disc}(0, 0.8)$, is illustrated in Figure V-5.10.

— $D_1 = \text{Star}$: Figures V-5.25, V-5.26, V-5.27, V-5.28. The set-valued function $D_1(\varepsilon)$ that we used, such that $D_1(0) = \text{Star}$, is illustrated in Figure V-5.11.

— $D_1 = \text{Cat-head}$: Figures V-5.31, V-5.32, V-5.33, V-5.34. The set-valued function $D_1(\varepsilon)$ that we used, such that $D_1(0) = \text{Cat-head}$, is illustrated in Figure V-5.12.

The first thing to see for each example is that the eigenvectors are indeed zero on $D_1^C = \mathbb{R}^d \setminus D_1$, for both the exact and approximate eigenvectors. The second thing to see is that the Fourier transform of the approximate eigenvectors are indeed concentrated with respect to the Fourier mask $\text{Disc}(0, 0.3)$. The third thing to notice is that a straightforward eigendecomposition of the matrices $\mathbf{K}^{[0]}$ yields eigenvectors that look like linear combinations of the "simpler" ones obtained via Algorithm 7.

We can also see the simple and multiple eigenvalues: by Lemma V.4, the eigenfunctions associated to simple eigenvalues must recover all symmetries present in the masks. For multiple eigenvalues, it is expected that all symmetries cannot be recovered.

In the case $D_1 = \text{Disc}(0, 0.8)$, the types of eigenvectors predicted by Slepian are polar ones of the form $(r, \theta) \mapsto R(r)\cos(m\theta)$ or $R(r)\sin(m\theta)$ with $m \geq 0$. They are not recovered via a direct eigendecomposition, but they are with the varying space mask procedure. In the cases $D_1 = \text{Star}$ and $D_1 = \text{Cat-head}$, we know nothing about the multiplicity of eigenvalues or expected form of eigenvectors. Therefore, we can't expect eigenvectors to recover all symmetries of the space mask. However, the numerical results show that, for most eigenvectors, we get some of the symmetries that are present in the space mask but not all of them, therefore they must correspond to eigenvalues of multiplicity $p > 1$. There are also some eigenvectors exhibiting all symmetries, and we conjecture that they correspond to eigenvalues of multiplicity $p = 1$.

The varying space mask procedure only yields approximate eigenvectors for the matrix $\mathbf{K}^{[0]}$, but their concentration ratio is 2η-close to the true eigenvalues (i.e. the concentration ratio of the exact eigenvectors), they have the advantage of recovering various symmetries that are present in D_1, and they do not depend on the eigenalgorithm used (as opposed to the straightforward eigendecomposition, see Section V-2.2.3 – Numerical difficulties). They are also visually more appealing!

In all of our two-dimensional numerical examples, we have used 50 points of discretization along each dimension. This choice was made in order to be able to compute approximate eigenvectors relatively quickly on a regular laptop. Most of the computational work is spent finding the first eigenpair of a given matrix, and this could surely be accelerated using GPU and parallel computations. This is not a direction we have explored.

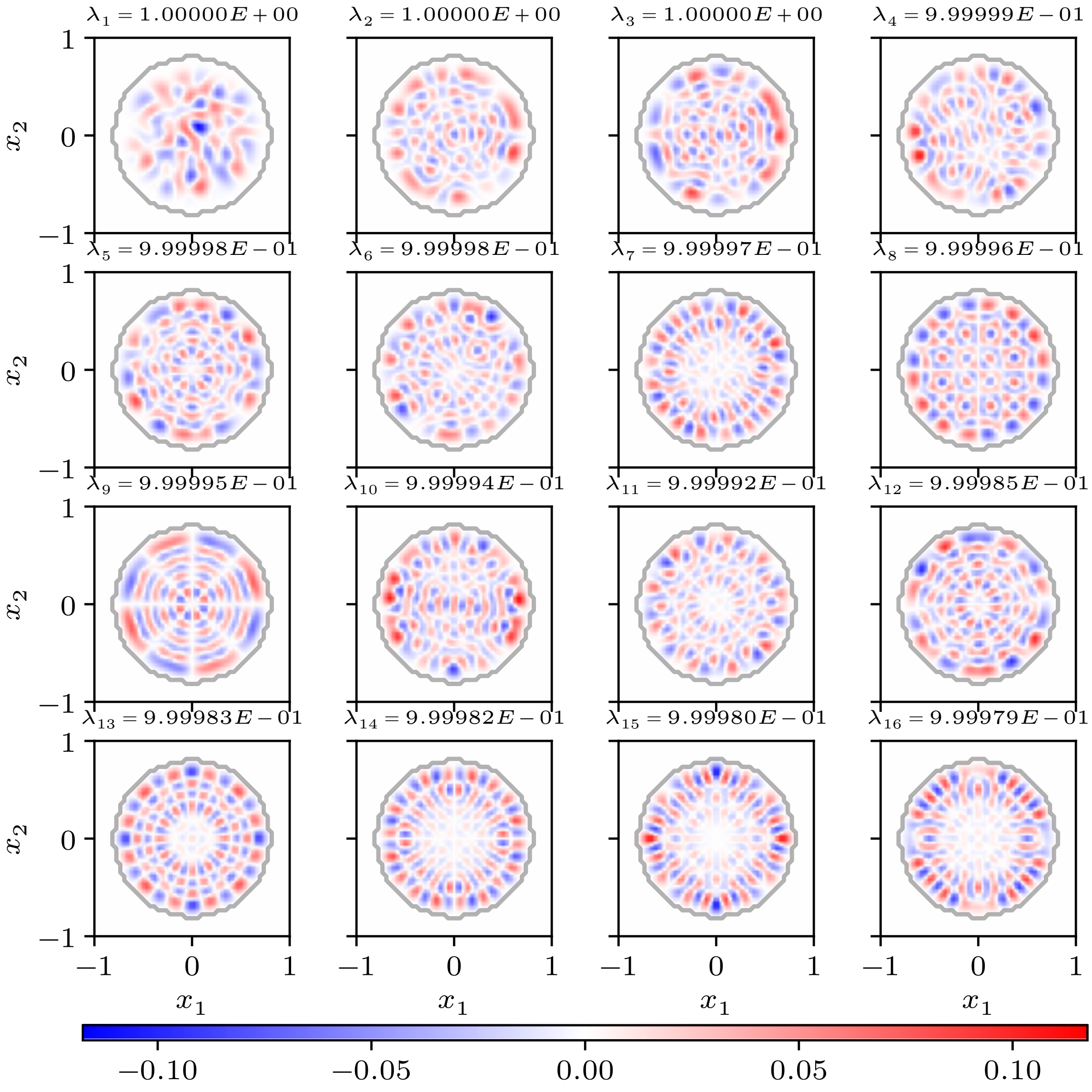


Figure V-5.19 – First 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with a classical eigendecomposition in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Disc}(0, 0.8)$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.10, and the boundary of $D_1(0) = \text{Disc}(0, 0.8)$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

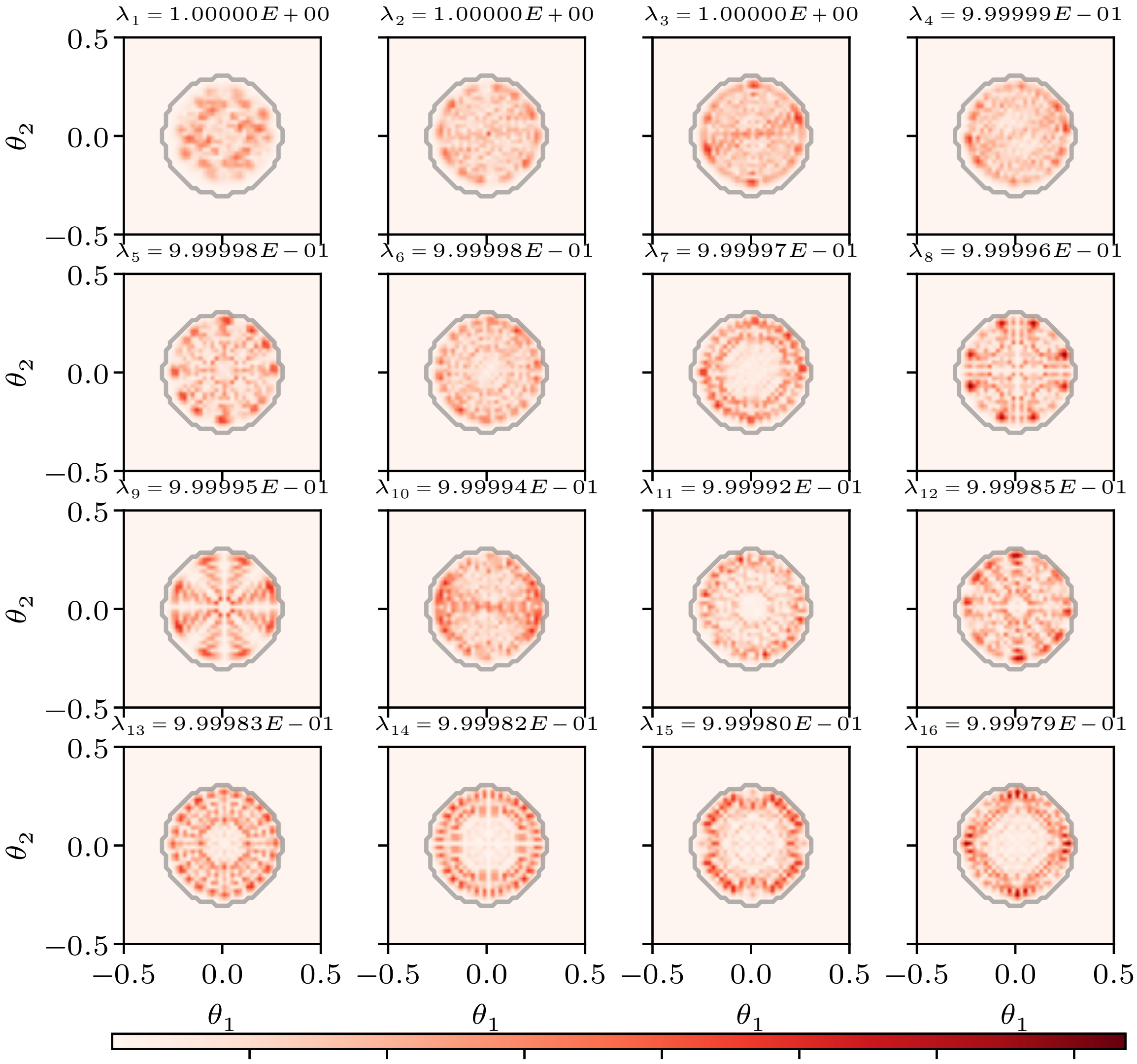


Figure V-5.20 – Absolute value of the Fourier transform of the first 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with a classical eigendecomposition in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\mathrm{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \mathrm{Disc}(0, 0.8)$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.10 and the boundary of $\mathrm{Disc}(0, 0.3)$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

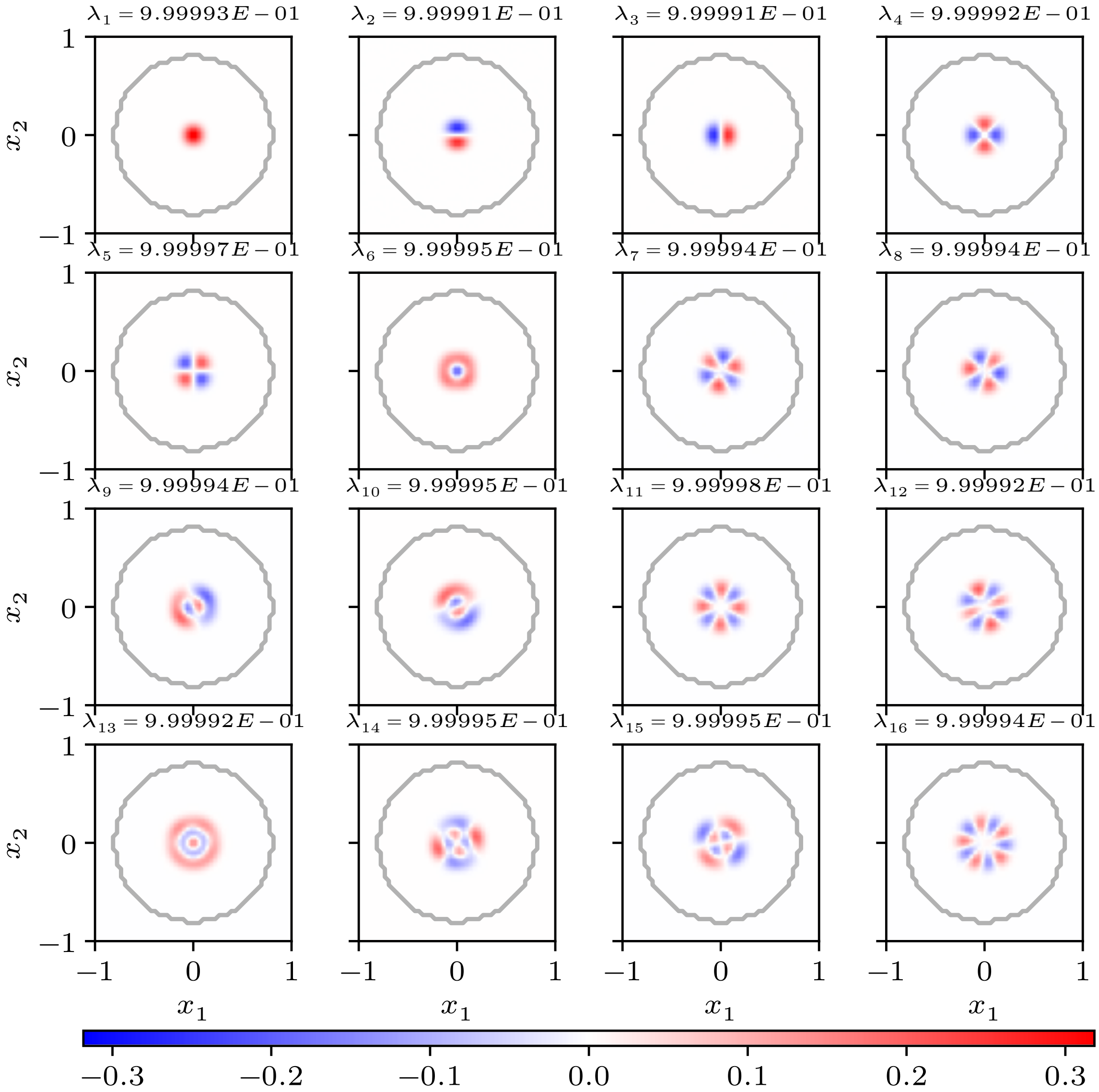


Figure V-5.21 – First 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with the varying mask procedure in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Disc}(0, 0.8)$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.10, and the boundary of $D_1(0) = \text{Disc}(0, 0.8)$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

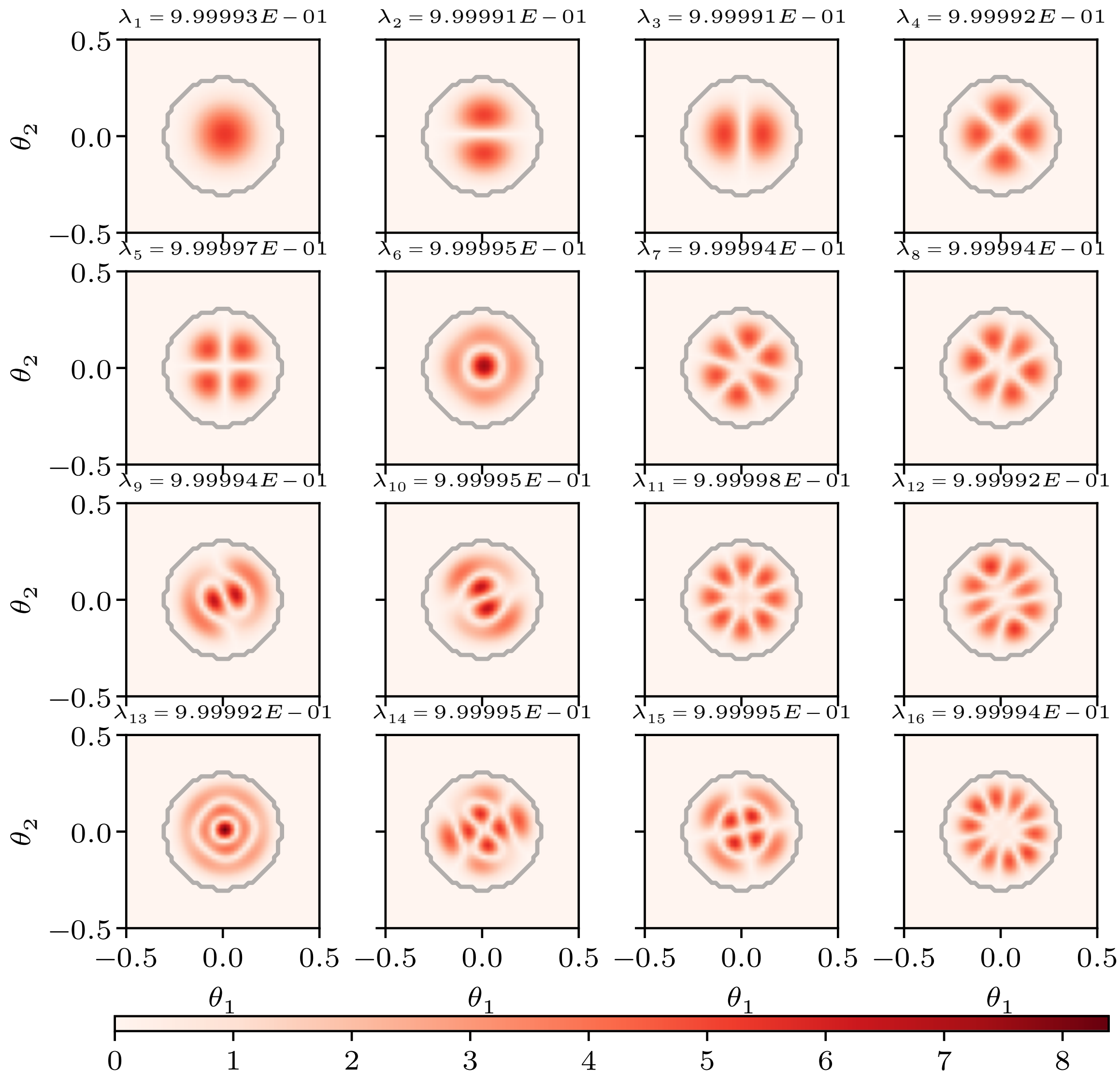


Figure V-5.22 – Absolute value of the Fourier transform of the first 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with the varying mask procedure in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Disc}(0, 0.8)$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.10 and the boundary of $\text{Disc}(0, 0.3)$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

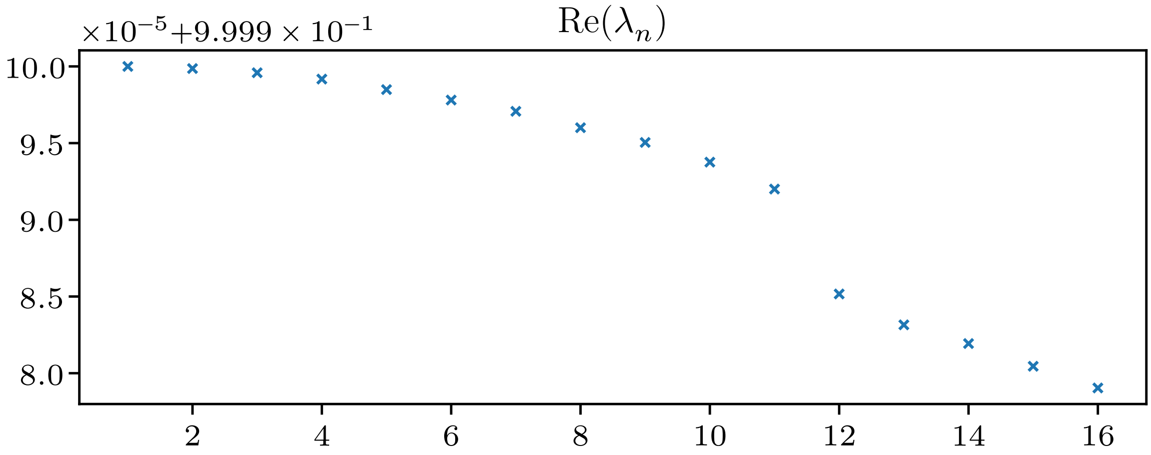


Figure V-5.23 – Eigenvalues with a direct decomposition, in the case $D_1(0) = \mathrm{Disc}(0, 0.8)$. They are the exact eigenvalues up to some tolerance $\eta = 10^{-5}$.

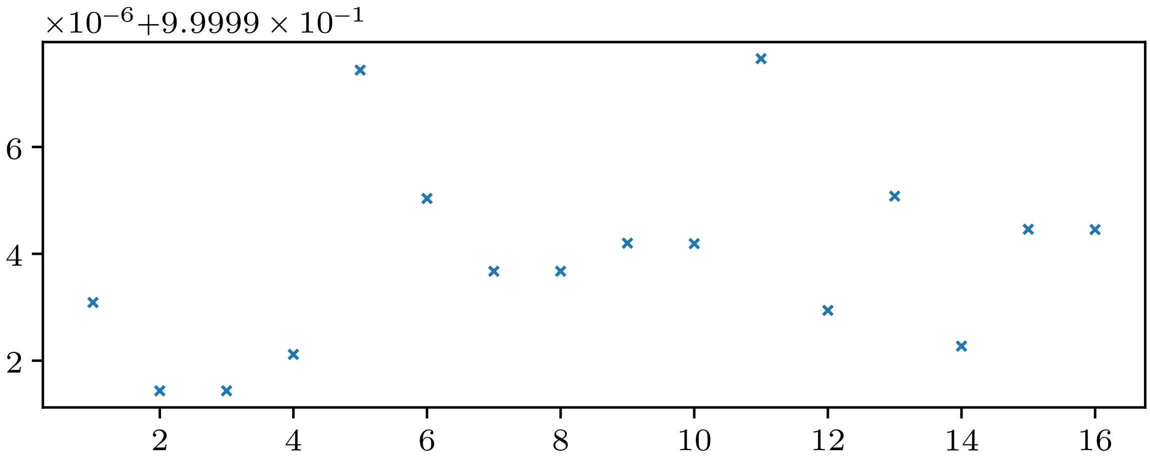


Figure V-5.24 – Eigenvalues obtained with the varying spacemask procedure, in the case $D_1(0) = \mathrm{Disc}(0, 0.8)$. They are the exact eigenvalues up to some tolerance $\eta = 10^{-5}$.

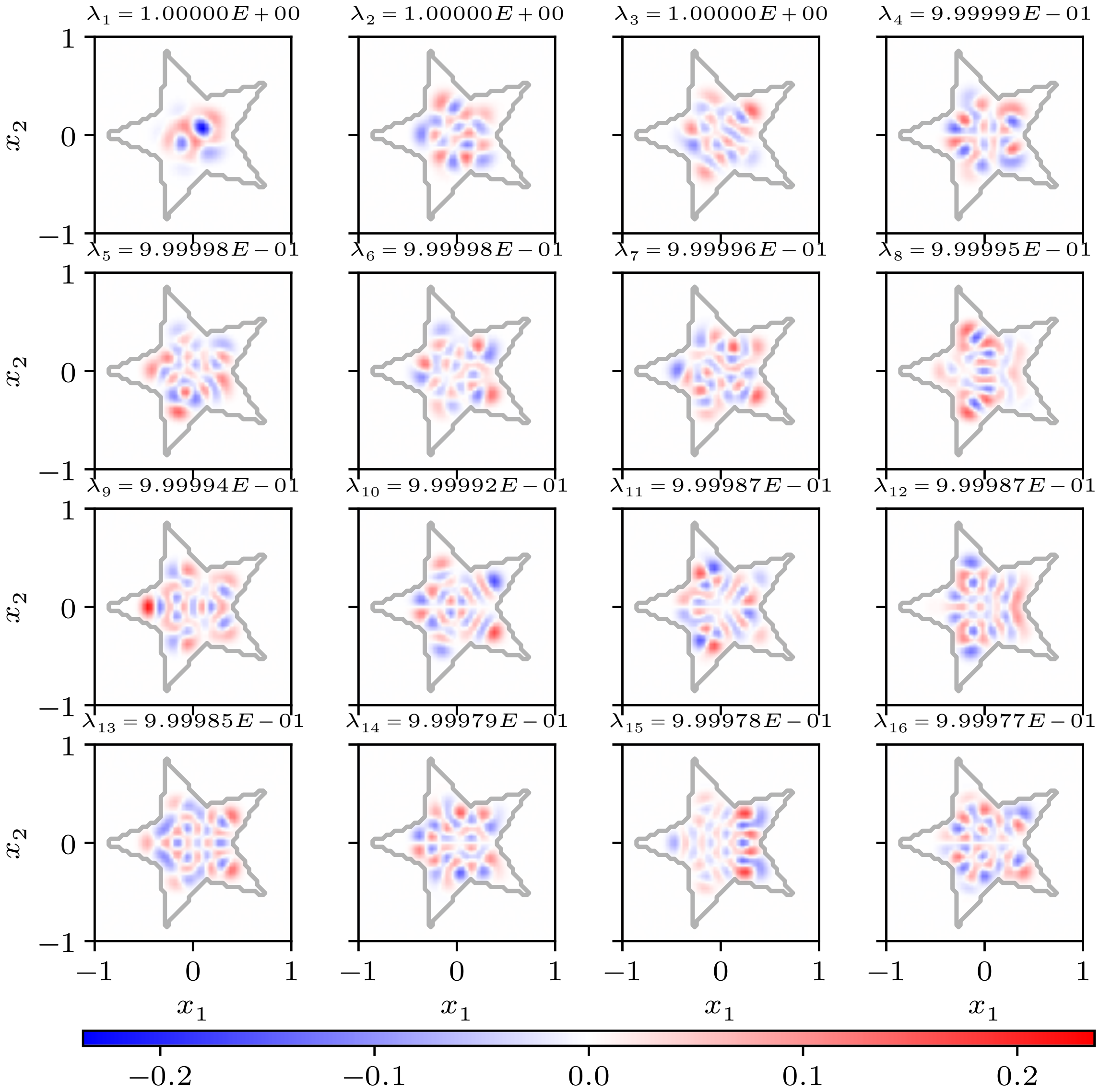


Figure V-5.25 – First 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with a classical eigendecomposition, in 2D with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Star}$ and $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.11. $N_1 = 50$, $N_2 = 50$. The boundary of $D_1(0) = \text{Star}$ is outlined in gray.

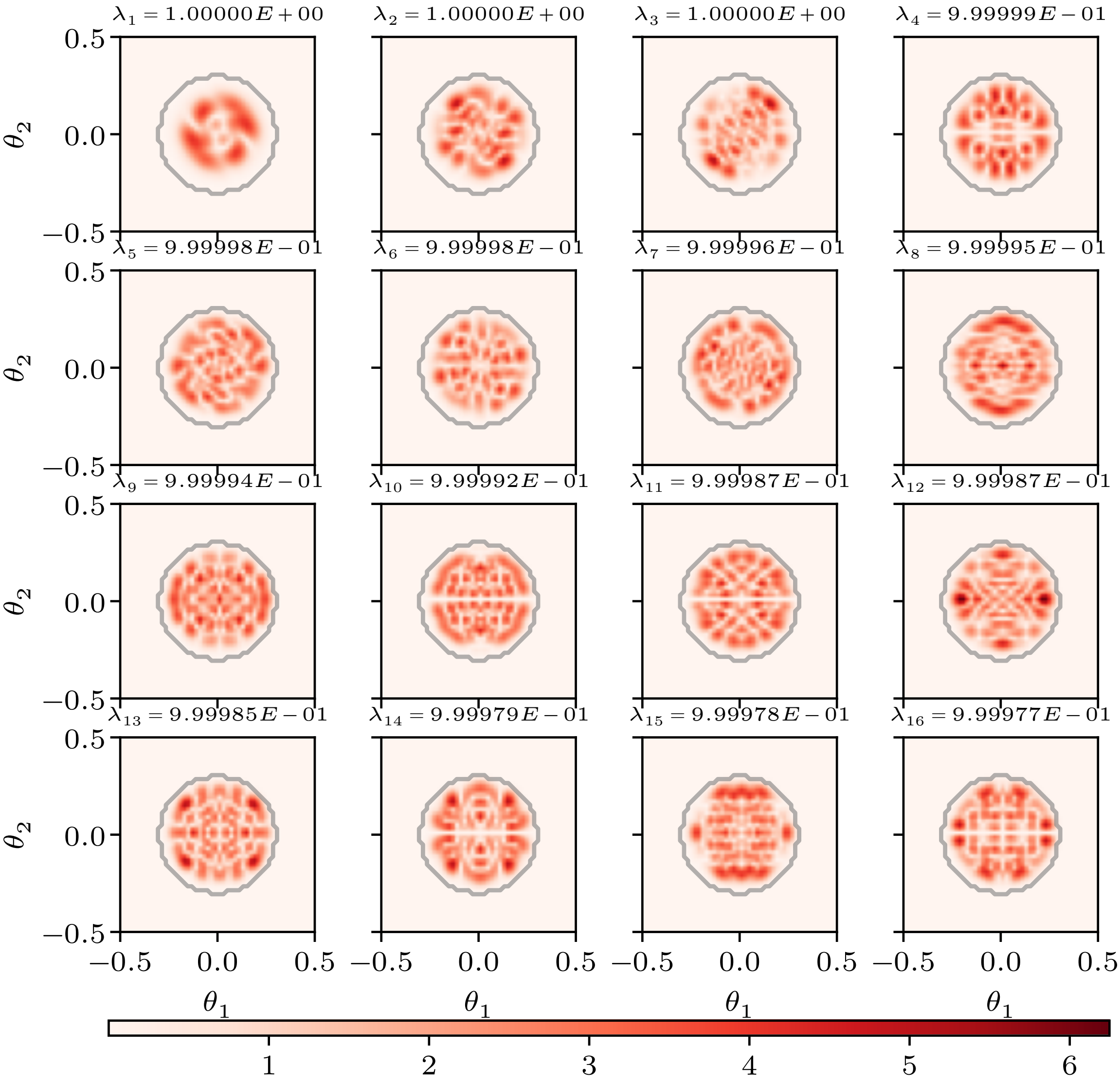


Figure V-5.26 – Absolute value of the Fourier transform of the first 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with a classical eigendecomposition, in 2D with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Star}$ and $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.11. $N_1 = 50$, $N_2 = 50$. The boundary of $\text{Disc}(0, 0.3)$ is outlined in gray.

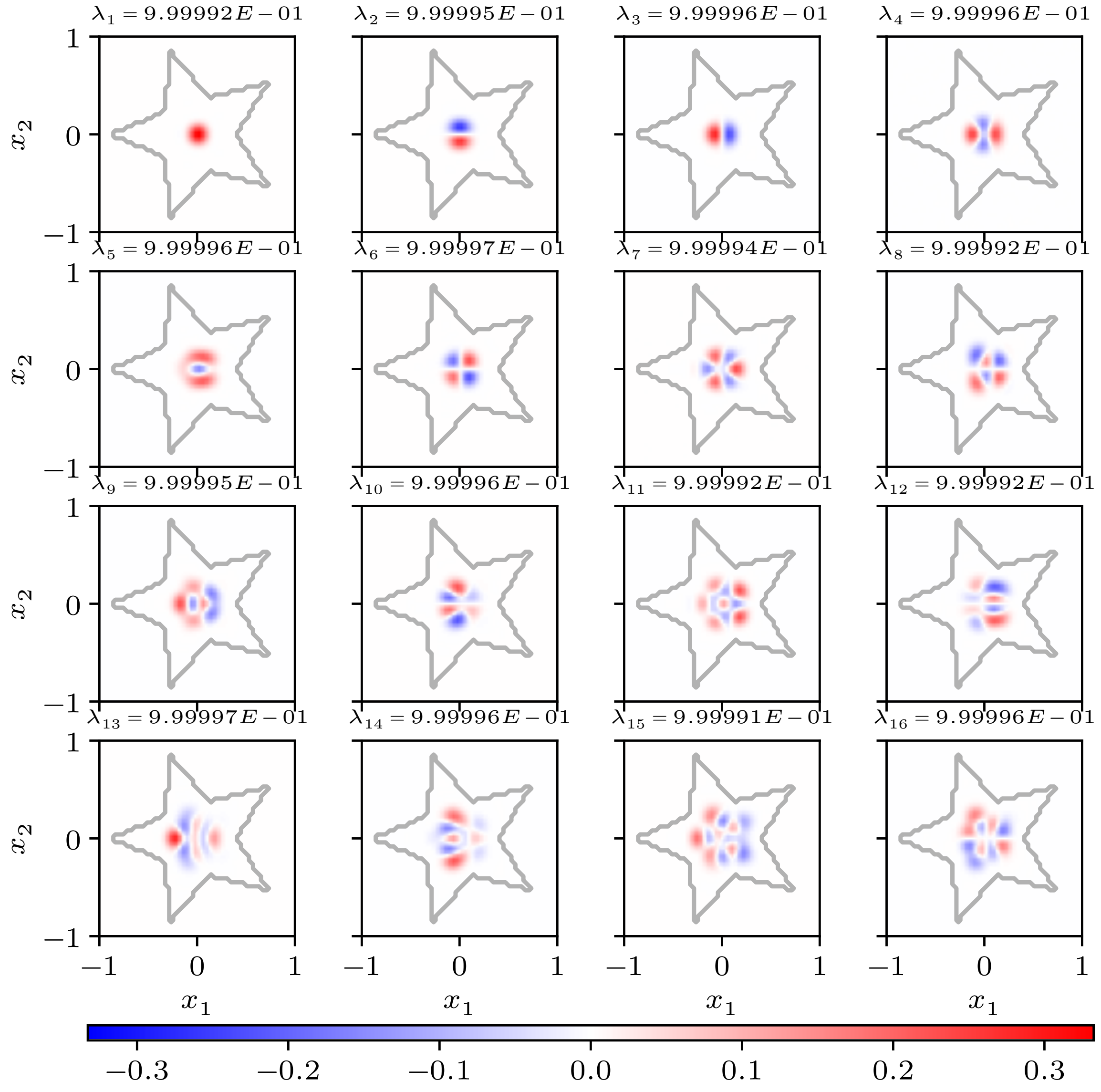


Figure V-5.27 – First 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with the varying mask procedure in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\mathrm{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Star}$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.11, and the boundary of $D_1(0) = \text{Star}$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

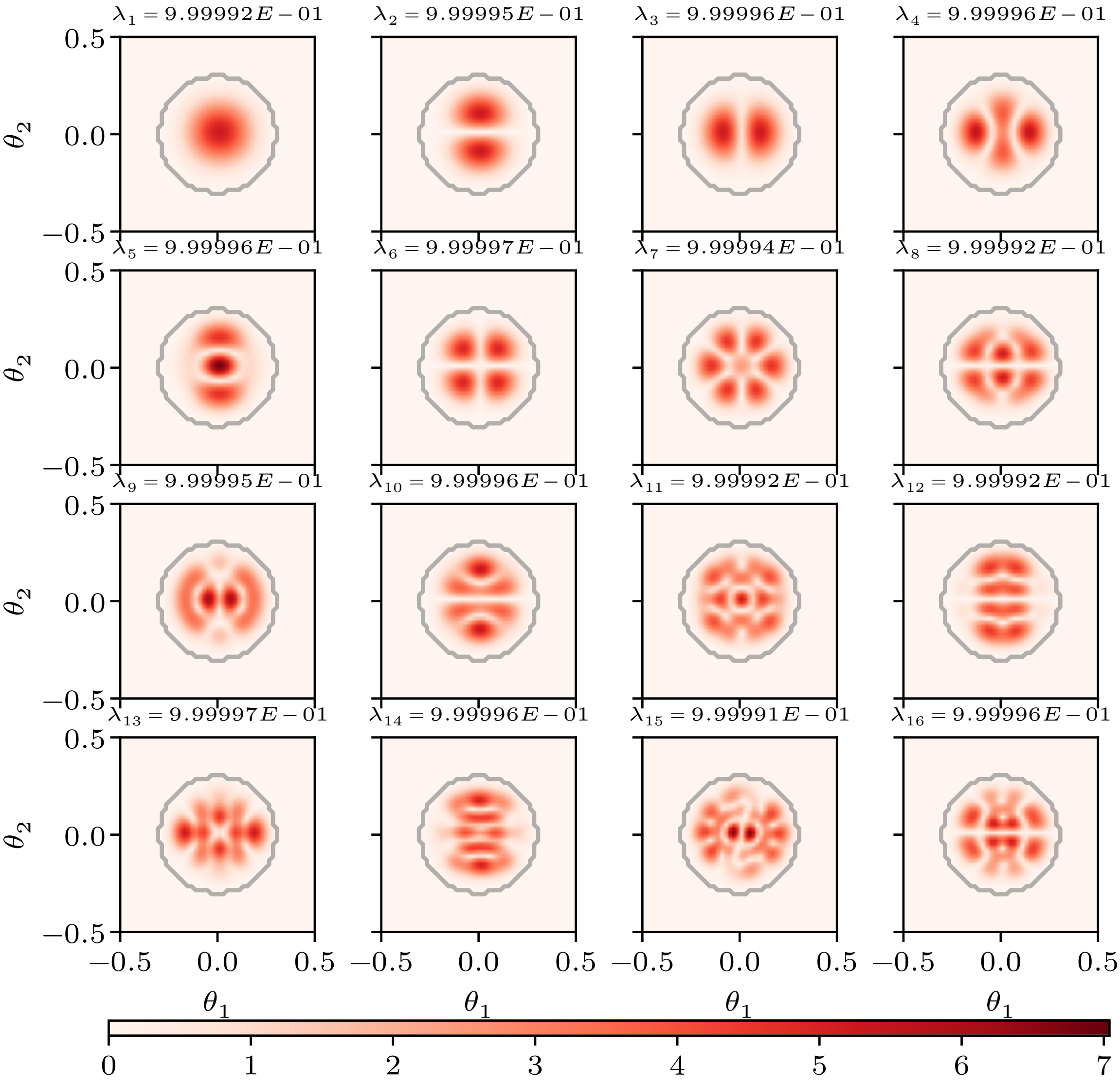


Figure V-5.28 – Absolute value of the Fourier transform of the first 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with the varying mask procedure in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Star}$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.11 and the boundary of $\text{Disc}(0, 0.3)$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

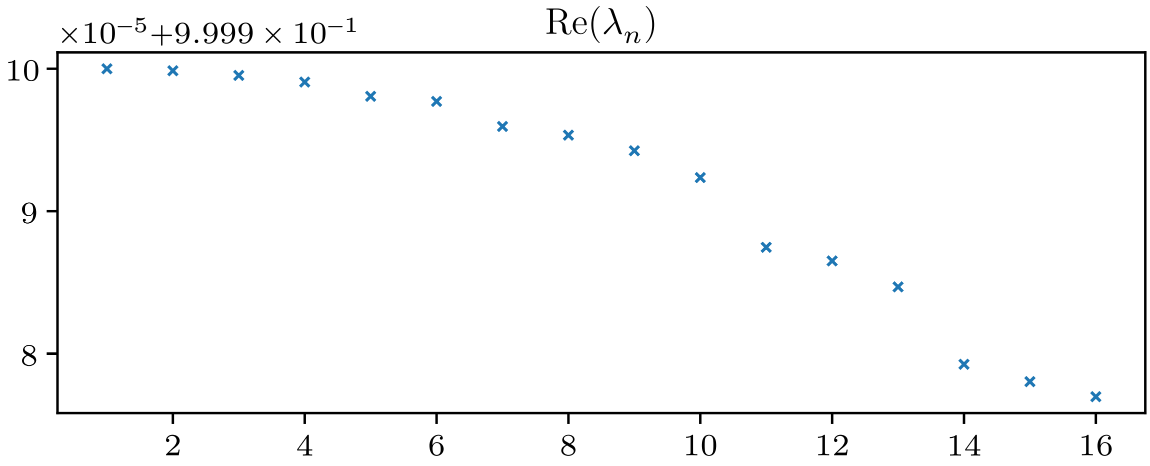


Figure V-5.29 – Eigenvalues with a direct decomposition, in the case $D_1(0) =$ Star. They are the exact eigenvalues up to some tolerance $\eta = 10^{-5}$.

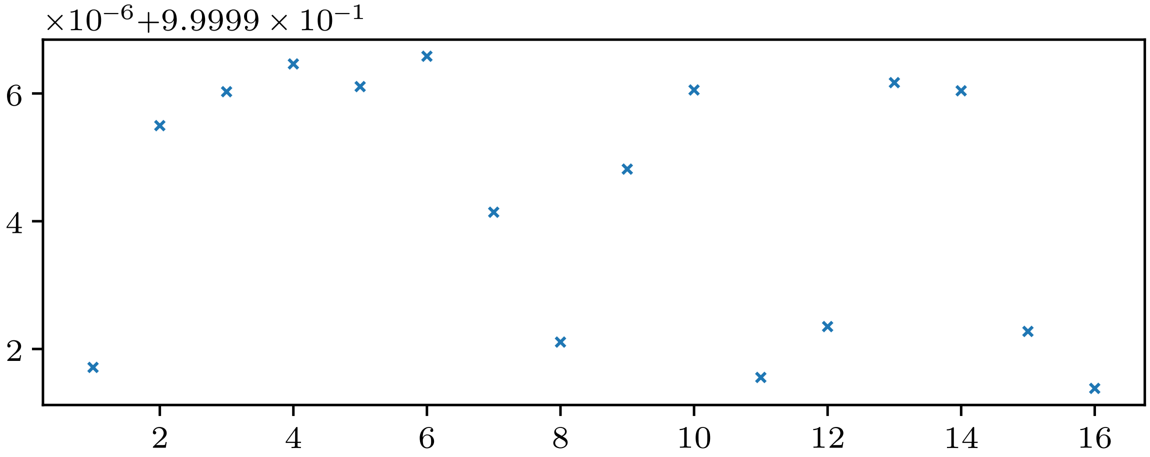


Figure V-5.30 – Eigenvalues obtained with the varying spacemask procedure, in the case $D_1(0) =$ Star. They are the exact eigenvalues up to some tolerance $\eta = 10^{-5}$.

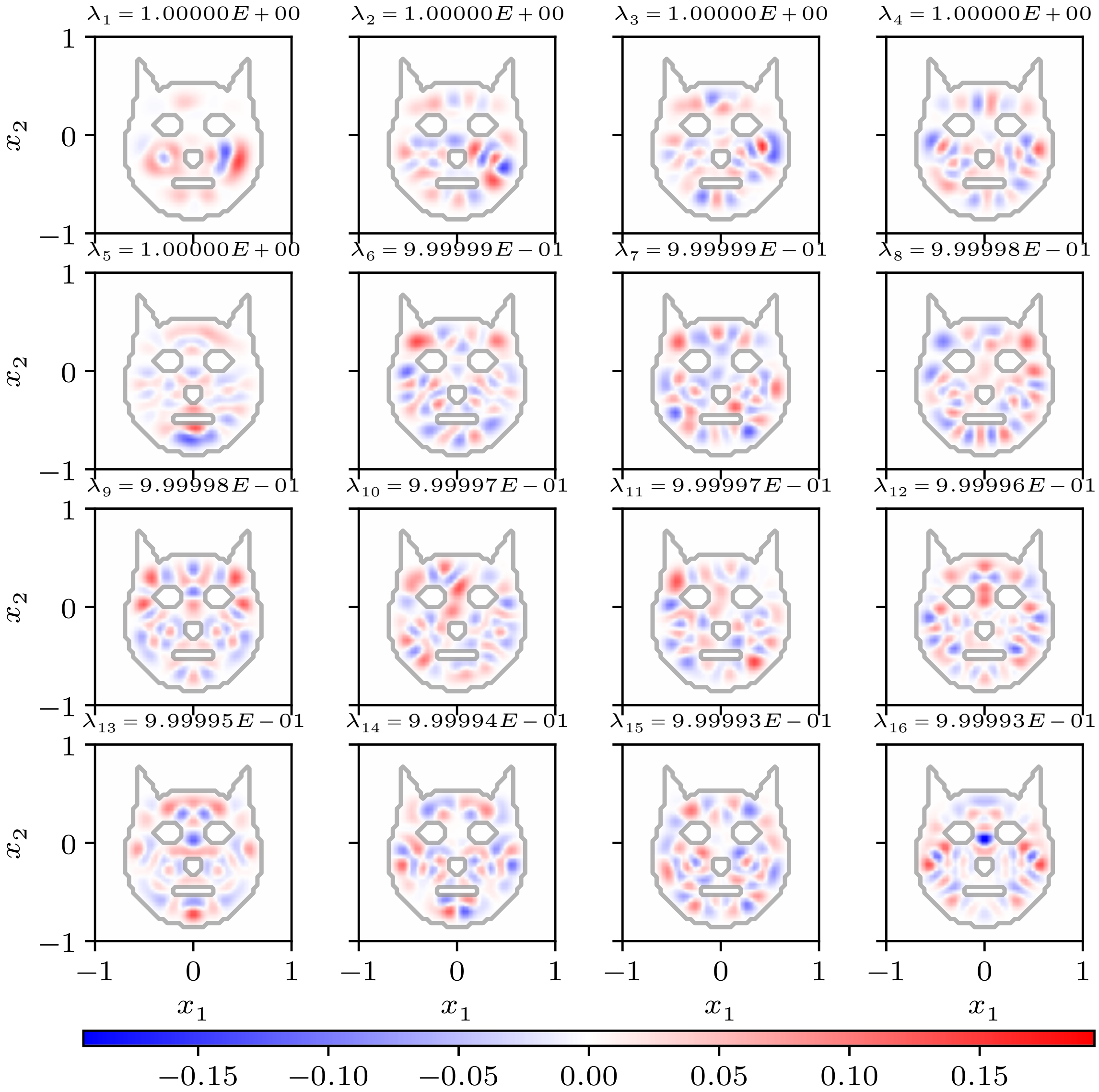


Figure V-5.31 – First 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with a classical eigendecomposition, in 2D with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) =$ Cat-head and $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.12. $N_1 = 50$, $N_2 = 50$. The boundary of $D_1(0) =$ Cat-head is outlined in gray.

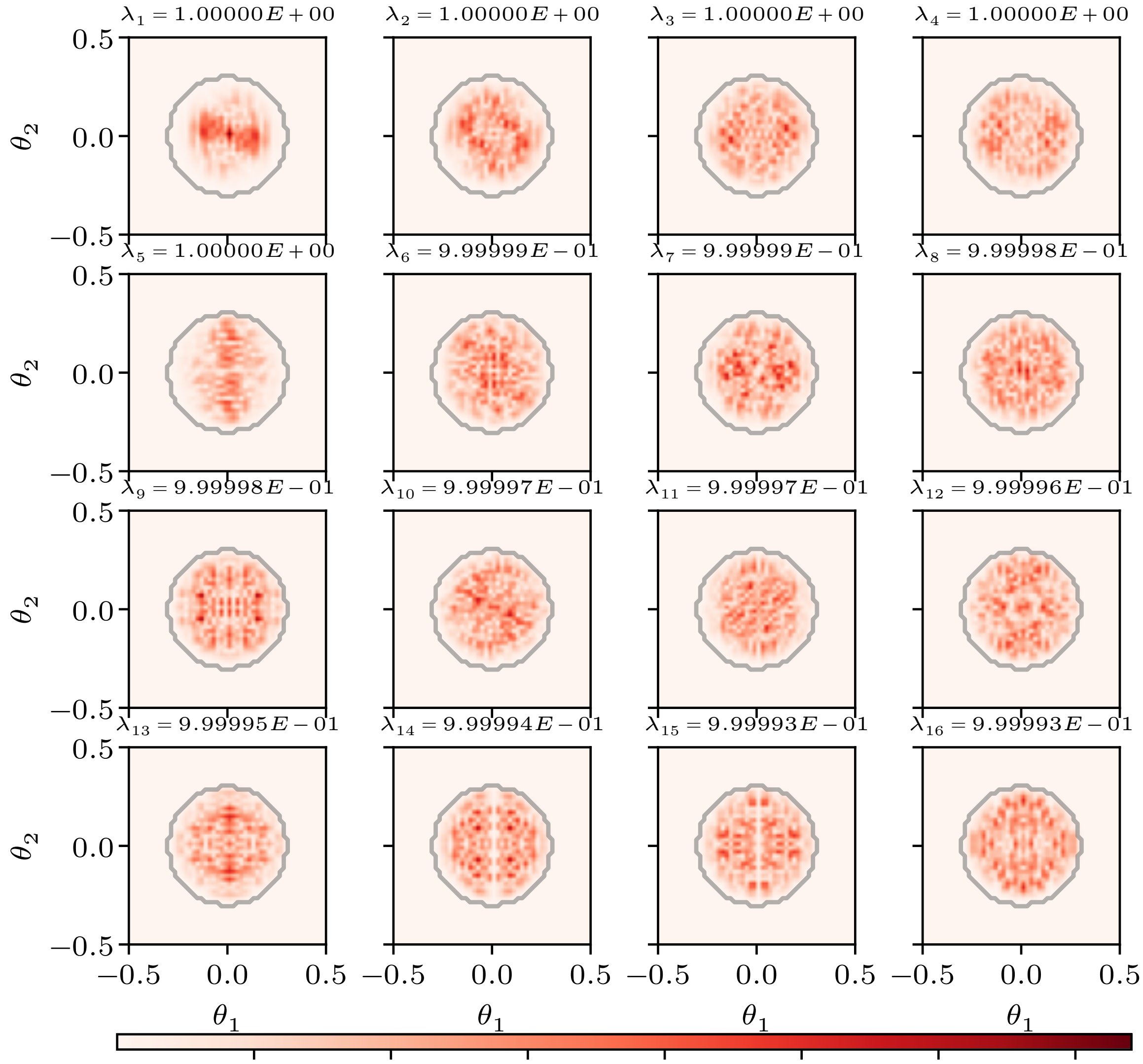


Figure V-5.32 – Absolute value of the Fourier transform of the first 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with a classical eigendecomposition, in 2D with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Cat-head}$ and $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.12. $N_1 = 50$, $N_2 = 50$. The boundary of Disc$(0, 0.3)$ is outlined in gray.

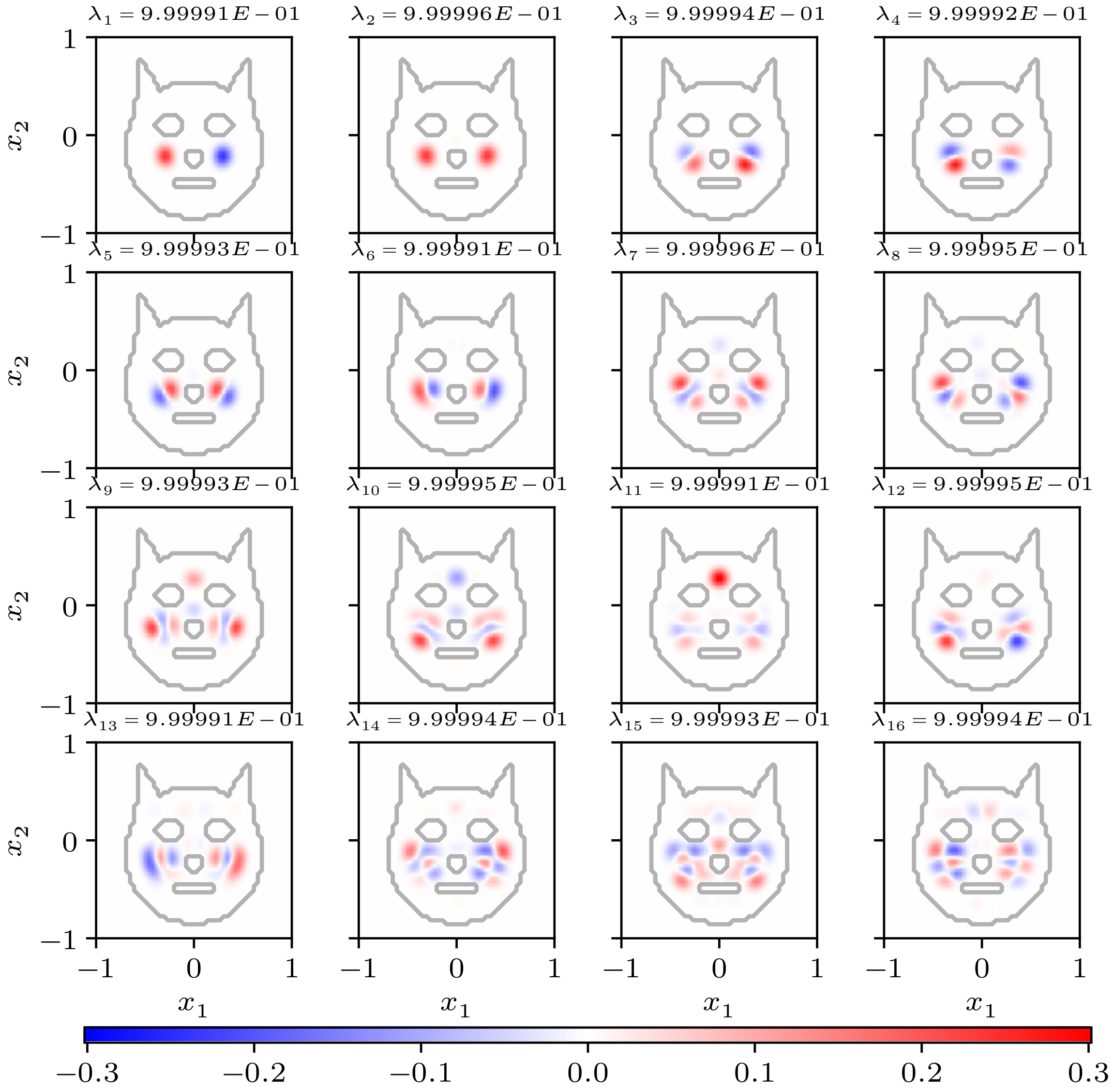


Figure V-5.33 – First 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with the varying mask procedure in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\text{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) = \text{Cat-head}$. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.12, and the boundary of $D_1(0) = \text{Cat-head}$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

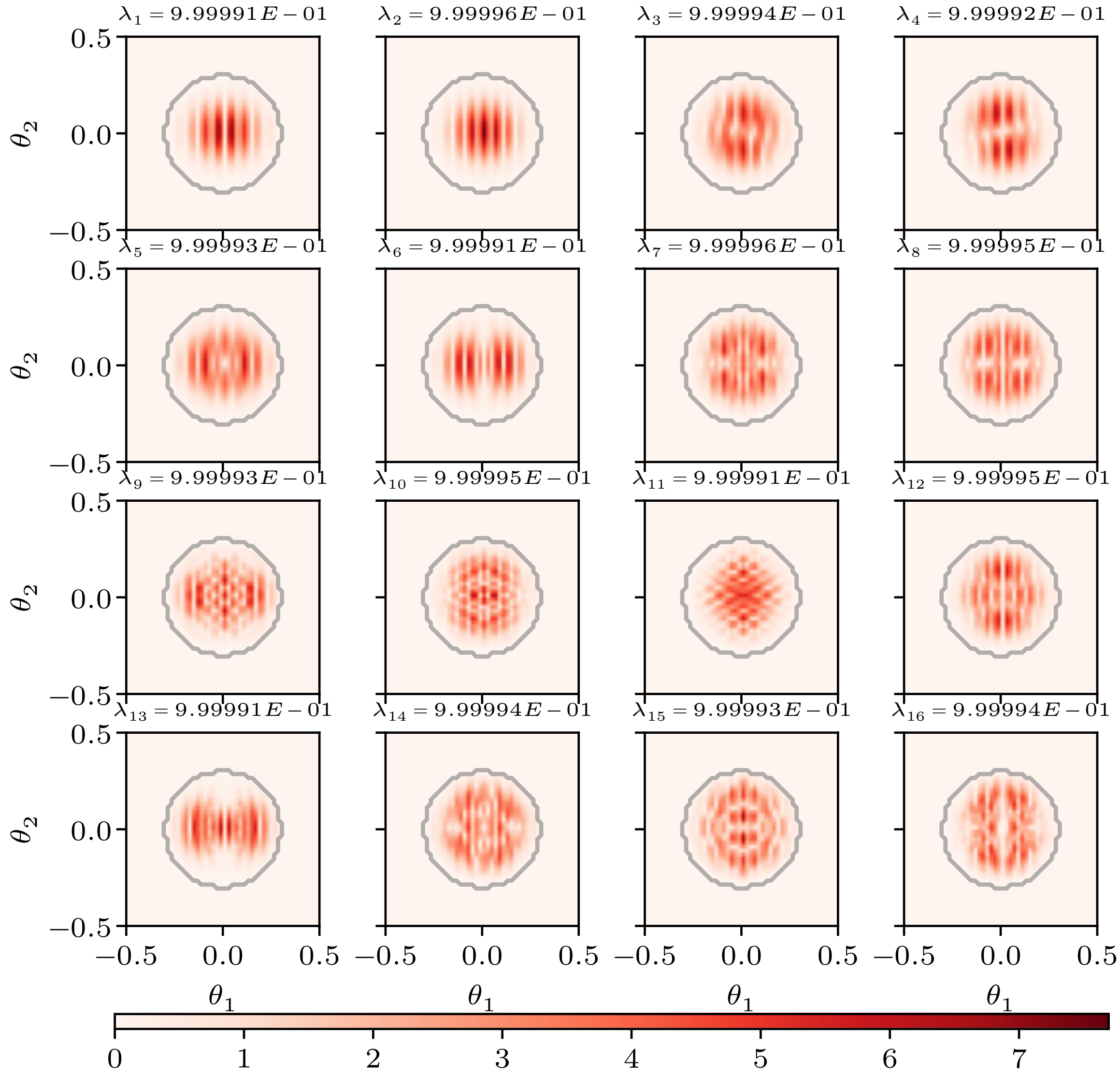


Figure V-5.34 – Absolute value of the Fourier transform of the first 16 eigenpairs of $\mathbf{K}^{[0]}$, obtained with the varying mask procedure in 2D, with $\eta = 10^{-5}$, $\widehat{m_{F,2\pi}} = \mathbf{1}_{\mathrm{Disc}(0,0.3)}$, and $m_S^{[\varepsilon]} = \mathbf{1}_{D_1(\varepsilon)}$ with $D_1(0) =$ Cat-head. The mapping $\varepsilon \mapsto D_1(\varepsilon)$ is illustrated in Figure V-5.12 and the boundary of Disc$(0, 0.3)$ is outlined in gray. $N_1 = 50$, $N_2 = 50$.

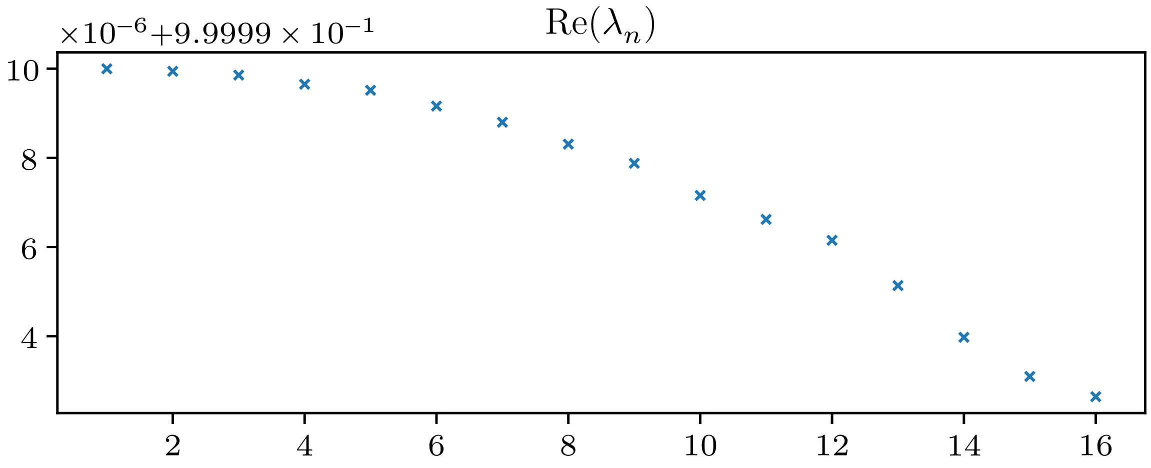


Figure V-5.35 – Eigenvalues with a direct decomposition, in the case $D_1(0) =$ Cat-head. They are the exact eigenvalues up to some tolerance $\eta = 10^{-5}$.

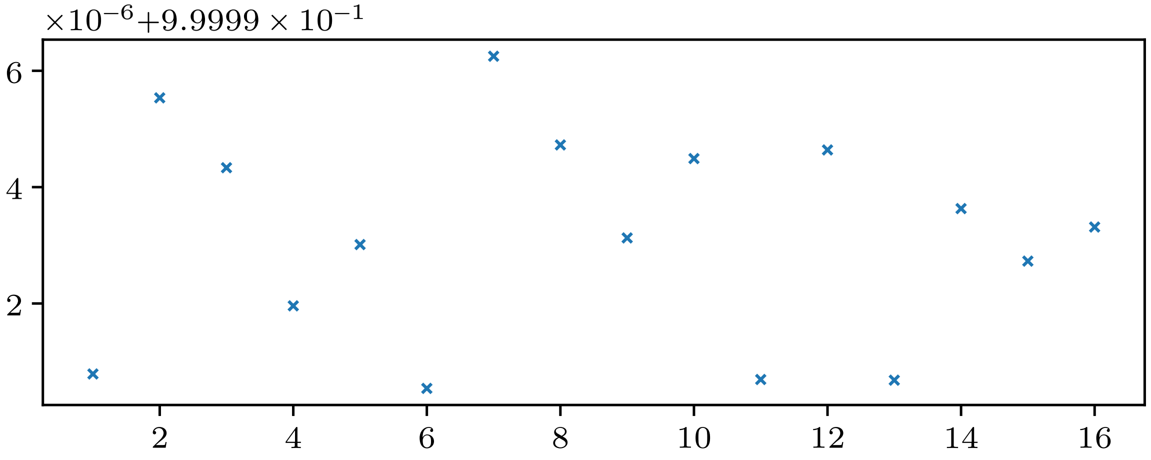


Figure V-5.36 – Eigenvalues obtained with the varying spacemask procedure, in the case $D_1(0) =$ Cat-head. They are the exact eigenvalues up to some tolerance $\eta = 10^{-5}$.

Part V

Numerical considerations

6 CHAPTER

In this Chapter we will discuss some numerical aspects of the concentration problem.

V-6.1 Leveraging the Toeplitz nature of the kernel

The fact that the concentration $\mathbf{K}$ as defined (V-4.13) is a block matrix where each block is a Toeplitz matrix, is interesting from the numerical point of view. The most interesting property for us is that the matrix-vector products can be done efficiently, both in memory and complexity. This is done by first transforming the Toeplitz matrix into a circulant matrix, and then computing efficient matrix-vector products for the circulant matrix.

First, let us recall what is a Toeplitz matrix: $T \in \mathcal{M}_{n_1,n_2}(\mathbb{C})$ is said to be Toeplitz if it is of the form

$$T = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & \cdots & t_{-n_2+1} \\ t_1 & t_0 & \ddots & \ddots & & \vdots \\ t_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & t_{-2} \\ \vdots & & \ddots & \ddots & \ddots & t_{-1} \\ t_{n_1-1} & \cdots & \cdots & t_2 & t_1 & t_0 \end{pmatrix}.$$

In other words, $T_{i,j}$ is constant for $i-j$ constant. The matrix T can be described using only its first row $\{t_k\}_{k=0,\ldots,-n_2+1}$ and first column $\{t_k\}_{k=0,\ldots,n_1-1}$. A circulant matrix is a particular case of a square Toeplitz matrix: $C \in \mathcal{M}_{n,n}(\mathbb{C})$ is said to be a circulant matrix if it can be written as

$$C = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & \cdots & c_1 \\ c_1 & c_0 & \ddots & \ddots & & \vdots \\ c_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & c_{n-2} \\ \vdots & & \ddots & \ddots & c_0 & c_{n-1} \\ c_{n-1} & \cdots & \cdots & c_2 & c_1 & c_0 \end{pmatrix}.$$

It is a square Toeplitz matrix of size n with the additinal property that $t_i = t_{i+n}$. For

$b \in \mathbb{C}^n$,

$$(Cb)_i = \sum_{j=0}^{n-1} c_{i-j} b_j.$$

That is, the left multiplication by the circulant matrix C corresponds to a circular convolution between the vector $(c_i)_{i=1}^n$ and the vector b. By means of the Fourier transform, we get

$$(Cb)_i = \mathcal{F}^{-1}[\mathcal{F}[c]\mathcal{F}[b]]_i,$$

where the product $\mathcal{F}[c]\mathcal{F}[b]$ is performed component by component.

Hence, a matrix-vector product between a circulant matrix and a given vector can be done using the Fourier transform. If this product was performed with no prior knowledge on C, it would require $\mathcal{O}(n^2)$ operations and $\mathcal{O}(n^2)$ memory space. By knowing that the matrix C is circulant, we only need $\mathcal{O}(n \log n)$ operations[1] and $\mathcal{O}(n)$ memory space.

Given a Toeplitz matrix T, we can create a circulant matrix M_C of twice the size, with the same information. The first row of this matrix M_C (which is enough to completely define it) is:

$$r := (t_0, t_1, \cdots, t_{n-1}, 0, t_{-n+1}, \cdots, t_{-1}) \in \mathbb{C}^{2n}.$$

The matrix M_C then has the following shape:

$$M_C = \begin{pmatrix} T & A \\ A & T \end{pmatrix},$$

where

$$A := \begin{pmatrix} 0 & t_{n-1} & \cdots & t_2 & t_1 \\ t_{-n+1} & \ddots & \ddots & & t_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ t_{-2} & & \ddots & \ddots & t_{n-1} \\ t_{-1} & t_{-2} & \cdots & t_{-n+1} & 0 \end{pmatrix}.$$

Now, for a vector $v \in \mathbb{C}^n$,

$$\begin{pmatrix} I_n & 0 \end{pmatrix} M_C \begin{pmatrix} v \\ 0_n \end{pmatrix} = \begin{pmatrix} I_n & 0 \end{pmatrix} \begin{pmatrix} T & A \\ A & T \end{pmatrix} \begin{pmatrix} v \\ 0_n \end{pmatrix} = \begin{pmatrix} I_n & 0 \end{pmatrix} \begin{pmatrix} Tv \\ Av \end{pmatrix} = Tv.$$

The product

$$M_C \begin{pmatrix} v \\ 0_n \end{pmatrix}$$

is a matrix-vector product between the circulant matrix M_C and a vector, so it can be

[1] We implicitely use the Fast Fourier Transform to obtain this numerical complexity.

done in $\mathcal{O}(n \log n)$ operations with $\mathcal{O}(n)$ memory storage. The product

$$\begin{pmatrix} I_n & 0 \end{pmatrix} \begin{pmatrix} Tv \\ Av \end{pmatrix}$$

simply consists in selecting the first n lines of $\begin{pmatrix} Tv \\ Av \end{pmatrix}$, so it can be done in only $\mathcal{O}(n)$ operations and memory storage.

This procedure thus shows that a matrix-vector product can be done in $\mathcal{O}(n \log n)$ operations and $\mathcal{O}(n)$ memory storage when the matrix is Toeplitz, by making use of the Fast Fourier Transform.

When we apply this to the concentration matrix $\mathbf{K}$ as defined in (V-4.14), there are N_1^2 Toeplitz blocks of size $N_2 \times N_2$. This results in $\mathcal{O}(N_1^2 N_2 \log N_2)$ operations for a product $\mathbf{K}v$, to be compared with $\mathcal{O}(N_1^2 N_2^2)$ operations without using the Toeplitz nature of each block. There is also a gain for the memory storage: instead of storing the whole matrix $\mathbf{K}$, with a memory usage of order $\mathcal{O}(N_1^2 N_2^2)$, it is sufficient to store two vectors of size N_2 for each one of the N_1^2 blocks. Thus, an efficient memory usage of order $\mathcal{O}(N_1^2 N_2)$.

Part V

CHAPTER 7 Conclusion

In this Part we started by reviewing the so-called Spectral Concentration problem. It is a problem popularized by Slepian and co-authors in the 1960s and 1970s. In a series of five papers, they obtained a very elegant solution for the one- and many-dimensional cases, as well as for the discrete case. Most of the subsequent work by other authors have used their results, and stayed within the framework set up during the 1960s and 1970s. The authors who tried to generalize the spectral concentration problem in the past have failed to find an equally elegant solution. This Part is not the end of never-ending series of failed attemps, we failed as well in this regard.

Since an elegant, theoretical solution could not be found, people have tried to find a numerical solution. In order to do this, they started by discretizing the concentration operator $\mathcal{K}$, in order to look for its eigenvectors and eigenvalues. In most situations, they used settings where they did not observe the main issue appearing with a discretization of the operator $\mathcal{K}$. More precisely, when the number of discretization points is too large, the highest eigenvalues gather very close to 1, and they cannot be distinguished anymore. Therefore, 1 looks like a multiple eigenvalue. This may not seem like a problem if one is only interested in eigenvalues, but it quickly becomes one if we are also interested in associated eigenvectors. Indeed, we are numerically looking for a basis of eigenvectors associated to a multiple eigenvalue. Therefore, one algorithm usually used to perform an eigendecomposition will yield a result, but another algorithm will yield another result. Both results are valid from the computer's point of view because the eigenvalues are too close to be distinguished, but they are not valid to us, since we get two different results for the same inputs.

This instability led us to study a generalized spectral concentration problem. After setting the theoretical foundations and obtaining a generalized concentration operator $\mathcal{K}$ (note the abuse of notation here), we discretized it. We observed again the same behavior of eigenvalues for previously unstudied situations. After this, we showed that 1 is not a multiple eigenvalue but only several distinct eigenvalues very close to 1, and that adding a perturbation to $\mathcal{K}$ allowed to perturb the eigenvalues as well.

Armed with physical intuition and interpretation of the spectral concentration problem, we managed to find an algorithm that bypasses the seemingly multiple eigenvalue, and which yields satisfying, deterministic eigenvectors. This algorithm works by shrinking and unshrinking the space mask little by little, and capturing the eigenvectors as soon as they are “satisfying enough”. We observed that the results this algorithm yields are not exactly the same as the results obtained by Slepian, but they are equally satisfying and do not rely on an “accidental property” of the concentration operator. Therefore, we could study approximate solutions to the generalized spectral concentration problem on previously unstudied situations.

Note that we only considered a Cartesian discretization of space and Fourier domains. If one only considers a binary mask in Fourier restricted to a two-dimensional ball, the Fourier kernel is known analytically. In this case, one can use the Gauss-Legendre integration method, as presented for instance in [42, 44]. We can hope that it would reduce the computational complexity and required grid points, while attaining a satisfying accuracy. It may however be troublesome to compute the kernel K is one considers arbitrary binary restrictions in Fourier and space, and a uniform discretization of the space and Fourier domain would then be the way to go for general problems.

VI

Conclusion

This is now in the last Part of the manuscript. Each Part is decorrelated from the others, so it wouldn't make much sense to recall here our main results and perspectives. We refer to Chapters III-6, IV-5 and V-7 for detailed discussions. However, we can give a brief summary of each Part.

In Part III, we have studied the Vlasov-Poisson system for plasma dynamics. More specifically, we have proved a convergence estimate for a gridfree particle method that was first introduced by Barré, Olivetti and Yamaguchi in 2011. It is called Weighted Particle method. It can be understood as being halfway between between Semi-Lagrangian and particle schemes. Unfortunately, no proof of convergence had been given when the Weighted Particle method was introduced, and the goal here was to obtain an error bound. It was relatively easy to do since the method is composed of fundational well-studied components, and we just had to gather the error bounds for each part. We also assessed the efficiency of the method on one-dimensional standard examples. This work has been published in a peer-reviewed journal.

In Part IV, we studied a numerical scheme for the Schrödinger equation. We started from several theoretical works which focused on *modulation* techniques, and we devised a numerical algorithm by using these techniques. The theoretical works are due to Faou, Martel, Merle and Raphaël. We saw that the modulation is very efficient when applied to the linear Schrödinger equation with quadratic potential, and yields the exact solution. Then, we tried to add a nonlinearity to the Schrödinger equation, and studied the cubic nonlinear Schrödinger equation. Unfortunately, the modulation couldn't be applied fully on this equation, and we had to resort to *splitting* techniques. We split the equation into its linear and nonlinear parts. The linear part was solved exactly using modulation, while the nonlinear part was approximately solved using the Dirac-Frenkel principle. To the best of our knowledge, this work is the first account of the Dirac-Frenkel principle being used in a nonlinear setting. We finally studied the results on some numerical examples, and we observed that the results were sometimes unsatisfying based on the initial condition. More precisely, we exhibited an initial condition where the Dirac-Frenkel principle worked well and yielded satisfying results, and one other initial condition for which it didn't work. The issues observed with the Dirac-Frenkel principle are not newly observed issues, and were already observed in some previous works.

Finally, in Part V, we studied the Spectral Concentration problem. It has been popularized by Slepian, Landau and Pollak in the 1960s and 1970s, and then found numerous applications in diverse fields of science. The problem they studied was relatively restrictive, but their solution was very elegant. Very few works have been concerned with a similar problem in different settings. After explaining why the problem was difficult from the theoretical and numerical points of view, we studied the numerical aspect. Using the physical intuition and interpretation of the problem, we have obtained an algorithm that bypasses the issues of a direct and straightforward approach. We compared the results of this algorithm in the well-studied framework, and observed that we obtain relatively satisfying approximations of the expected solutions. We then applied the algorithm to previously unstudied examples, and also observed that it yielded results much more satisfying than those of a direct and straightforward approach.

As a side-note, the ideas used in this Algorithm could probably be adapted to a more general context and are not necessarily restricted to the spectral concentration problem. Deep down, the numerical issues we overcame are numerical linear algebra problems.

Bibliography

References for Some preliminaries of Numerical Analysis

[1] P. Alphonse and J. Bernier, « Polar Decomposition of Semigroups Generated by Non-Selfadjoint Quadratic Differential Operators and Regularizing Effects », *in*: *Annales scientifiques de l'École Normale Supérieure* 56.*2* (2023) (referenced on page 31).

[2] H. F. Baker, « Alternants and Continuous Groups », *in*: *Proceedings of the London Mathematical Society* s2-3.*1* (1905) (referenced on page 30).

[3] A. H. Barnett, « Aliasing Error of the $\exp(\beta(1-Z^2))$ Kernel in the Nonuniform Fast Fourier Transform », *in*: *Applied and Computational Harmonic Analysis* 51 (Mar. 2021) (referenced on page 41).

[4] A. H. Barnett, J. Magland, and L. Af Klinteberg, « A Parallel Nonuniform Fast Fourier Transform Library Based on an "Exponential of Semicircle" Kernel », *in*: *SIAM Journal on Scientific Computing* 41.*5* (Jan. 2019) (referenced on page 41).

[5] J. Bernier, « Exact Splitting Methods for Semigroups Generated by Inhomogeneous Quadratic Differential Operators », *in*: *Foundations of Computational Mathematics* 21.*5* (Oct. 2021) (referenced on page 31).

[6] J. Bernier, N. Crouseilles, and Y. Li, « Exact Splitting Methods for Kinetic and Schrödinger Equations », *in*: *Journal of Scientific Computing* 86.*1* (Jan. 2021) (referenced on page 31).

[7] W. L. Briggs and V. E. Henson, *The DFT: An Owner's Manual for the Discrete Fourier Transform*, Society for Industrial and Applied Mathematics, Jan. 1995 (referenced on page 39).

[8] F. Casas, N. Crouseilles, E. Faou, and M. Mehrenberger, « High-Order Hamiltonian Splitting for the Vlasov–Poisson Equations », *in*: *Numerische Mathematik* 135.*3* (Mar. 2017) (referenced on page 31).

[9] D. Chiron, *Espace de Schwartz, distributions tempérées et transformation de Fourier*, Chemins d'analyse tome 1, Paris: Calvage & Mounet, 2021 (referenced on pages 37, 38).

[10] B. A. Cipra, « The Best of the 20th Century: Editors Name Top 10 Algorithms », *in*: *SIAM News* 33.*4* (May 2000) (referenced on page 41).

[11] J. W. Cooley and J. W. Tukey, « An Algorithm for the Machine Calculation of Complex Fourier Series », *in*: *Mathematics of Computation* 19.*90* (1965) (referenced on page 41).

[12] A. Dutt and V. Rokhlin, « Fast Fourier Transforms for Nonequispaced Data », *in*: *SIAM Journal on Scientific Computing* 14.*6* (Nov. 1993) (referenced on page 41).

[13] T. H. Gronwall, « Note on the Derivatives with Respect to a Parameter of the Solutions of a System of Differential Equations », *in*: *The Annals of Mathematics* 20.*4* (July 1919) (referenced on page 43).

[14] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed, Springer Series in Computational Mathematics 31, Berlin ; New York: Springer, 2006 (referenced on page 30).

[15] M. T. Heideman, D. H. Johnson, and C. S. Burrus, « Gauss and the History of the Fast Fourier Transform », *in*: *Archive for History of Exact Sciences* 34.*3* (1985) (referenced on page 41).

[16] M. Hochbruck and A. Ostermann, « Exponential Integrators », *in*: *Acta Numerica* 19 (May 2010) (referenced on page 30).

[17] G. Joyce, G. Knorr, and H. K. Meier, « Numerical Integration Methods of the Vlasov Equation », *in*: *Journal of Computational Physics* 8.*1* (Aug. 1971) (referenced on page 33).

[18] R. I. McLachlan and G. R. W. Quispel, « Splitting Methods », *in*: *Acta Numerica 2002*, 1st ed., Cambridge University Press, July 2002 (referenced on page 30).

[19] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Rev. and enl. ed, New York: Academic Press, 1980 (referenced on pages 37, 38).

[20] F. Rouvière, *Petit guide de calcul différentiel à l'usage de la licence et de l'agrégation*, 2e éd. rev. et augm, Enseignement des mathématiques 4, Paris: Cassini, 2003 (referenced on page 44).

[21] J. Shen, T. Tang, and L.-L. Wang, *Spectral Methods: Algorithms, Analysis and Applications*, vol. 41, Springer Series in Computational Mathematics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011 (referenced on page 33).

[22] H. F. Trotter, « On the Product of Semi-Groups of Operators », *in*: *Proceedings of the American Mathematical Society* 10.*4* (1959) (referenced on page 30).

[23] N. Wiener and R. Paley, *Fourier Transforms in the Complex Domain*, vol. 19, Colloquium Publications, Providence, Rhode Island: American Mathematical Society, Dec. 1934 (referenced on pages 37, 38).

[24] N. N. Yanenko, *The Method of Fractional Steps*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1971 (referenced on page 30).

References for The Vlasov-Poisson system

[1] H. Abbasi, M. H. Jenab, and H. H. Pajouh, « Preventing the Recurrence Effect in the Vlasov Simulation by Randomizing Phase-Point Velocities in Phase Space », *in*: *Physical Review E* 84.*3* (Sept. 2011) (referenced on page 93).

[2] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, vol. 55, US Government printing office, 1964 (referenced on page 118).

[3] C. Anderson and C. Greengard, « On Vortex Methods », *in*: *SIAM Journal on Numerical Analysis* 22.*3* (June 1985) (referenced on page 78).

[4] M. Antoni and S. Ruffo, « Clustering and Relaxation in Hamiltonian Long-Range Dynamics », *in*: *Physical Review E* 52.*3* (Sept. 1995) (referenced on page 86).

[5] T. P. Armstrong, « Numerical Studies of the Nonlinear Vlasov Equation », *in*: *The Physics of Fluids* 10.*6* (June 1967) (referenced on pages 64, 65).

[6] A. Arsen'ev, « Global Existence of a Weak Solution of Vlasov's System of Equations », *in*: *USSR Computational Mathematics and Mathematical Physics* 15.*1* (Jan. 1975) (referenced on page 63).

[7] D. Arsénio, E. Dormy, and C. Lacave, « The Vortex Method for Two-Dimensional Ideal Flows in Exterior Domains », *in*: *SIAM Journal on Mathematical Analysis* 52.*4* (Jan. 2020) (referenced on page 78).

[8] K. E. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed, New York: Wiley, 1989 (referenced on page 89).

[9] A. Y. Aydemir, « A Unified Monte-Carlo Interpretation of Particle Simulations and Applications to Non-neutral Plasmas », *in*: *Physics of Plasmas* 1.*4* (1994) (referenced on page 73).

[10] G. Backus, « Linearized Plasma Oscillations in Arbitrary Electron Velocity Distributions », *in*: *Journal of Mathematical Physics* 1.*3* (May 1960) (referenced on page 64).

[11] A. H. Barnett, « Aliasing Error of the $\exp(\beta(1-Z^2))$ Kernel in the Nonuniform Fast Fourier Transform », *in*: *Applied and Computational Harmonic Analysis* 51 (Mar. 2021) (referenced on pages 85, 86).

[12] A. H. Barnett, J. Magland, and L. Af Klinteberg, « A Parallel Nonuniform Fast Fourier Transform Library Based on an "Exponential of Semicircle" Kernel », *in*: *SIAM Journal on Scientific Computing* 41.*5* (Jan. 2019) (referenced on page 86).

[13] J. Barré, A. Olivetti, and Y. Y. Yamaguchi, « Algebraic Damping in the One-Dimensional Vlasov Equation », *in*: *Journal of Physics A: Mathematical and Theoretical* 44.*40* (Oct. 2011) (referenced on pages 54, 75, 85, 86, 129).

[14] J. Batt, « Global Symmetric Solutions of the Initial Value Problem of Stellar Dynamics », *in*: *Journal of Differential Equations* 25.*3* (Sept. 1977) (referenced on page 63).

[15] J. T. Beale, « A Convergent 3-D Vortex Method with Grid-Free Stretching », *in*: *Mathematics of Computation* 46.*174* (Apr. 1986), JSTOR: 2007984 (referenced on page 78).

[16] J. T. Beale and A. Majda, « Vortex Methods. i: Convergence in Three Dimensions », *in*: *Mathematics of Computation* 39.*159* (July 1982), JSTOR: 2007617 (referenced on page 78).

[17] D. Berend and T. Tassa, « Improved Bounds on Bell Numbers and on Moments of Sums of Random Variables », *in*: *Probability and Mathematical Statistics* 30.*2* (2010) (referenced on page 119).

[18] N. Besse, « Convergence of a Semi-Lagrangian Scheme for the One-Dimensional Vlasov-Poisson System », *in*: *SIAM Journal on Numerical Analysis* 42.*1* (Jan. 2004) (referenced on pages 68, 78).

[19] N. Besse and M. Mehrenberger, « Convergence of Classes of High-Order Semi-Lagrangian Schemes for the Vlasov-Poisson System », *in*: *Mathematics of Computation* 77.*261* (Jan. 2008) (referenced on page 68).

[20] M. Bessemoulin-Chatard and F. Filbet, « On the Convergence of Discontinuous Galerkin/Hermite Spectral Methods for the Vlasov–Poisson System », *in*: *SIAM Journal on Numerical Analysis* 61.*4* (Aug. 2023) (referenced on page 66).

[21] M. Bessemoulin-Chatard and F. Filbet, « On the Stability of Conservative Discontinuous Galerkin/Hermite Spectral Methods for the Vlasov-Poisson System », *in*: *Journal of Computational Physics* 451 (Feb. 2022) (referenced on pages 66, 67).

[22] C. Birdsall and A. Langdon, *Plasma Physics via Computer Simulation*, 1st ed., CRC Press, 1991 (referenced on pages 62, 73).

[23] C. K. Birdsall and D. Fuss, « Clouds-in-Clouds, Clouds-in-Cells Physics for Many-Body Plasma Simulation », *in*: *Journal of Computational Physics* 3.*4* (Apr. 1969) (referenced on page 71).

[24] A. Blaustein, *Structure Preserving Solver for Multi-dimensional Vlasov-Poisson Type Equations*, 2024 (referenced on page 66).

[25] A. Blaustein and F. Filbet, « A Structure and Asymptotic Preserving Scheme for the Vlasov-Poisson-Fokker-Planck Model », *in*: *Journal of Computational Physics* 498 (Feb. 2024) (referenced on page 66).

[26] J. U. Brackbill, « Particle Methods », *in*: *International Journal for Numerical Methods in Fluids* 47.*8-9* (Mar. 2005) (referenced on page 73).

[27] J. Brackbill, « The Ringing Instability in Particle-in-Cell Calculations of Low-Speed Flow », *in*: *Journal of Computational Physics* 75.*2* (Apr. 1988) (referenced on page 73).

[28] J. Brackbill, « On Energy and Momentum Conservation in Particle-in-Cell Plasma Simulation », *in*: *Journal of Computational Physics* 317 (July 2016) (referenced on pages 61, 73).

[29] J. Brackbill, D. Kothe, and H. Ruppel, « FLIP: A Low-Dissipation, Particle-in-Cell Method for Fluid Flow », *in*: *Computer Physics Communications* 48.*1* (Jan. 1988) (referenced on page 73).

[30] J. Brackbill and H. Ruppel, « FLIP: A Method for Adaptively Zoned, Particle-in-Cell Calculations of Fluid Flows in Two Dimensions », *in*: *Journal of Computational Physics* 65.*2* (Aug. 1986) (referenced on page 73).

[31] W. Braun and K. Hepp, « The Vlasov Dynamics and Its Fluctuations in the 1/N Limit of Interacting Classical Particles », *in*: *Communications in Mathematical Physics* 56.*2* (June 1977) (referenced on page 53).

[32] O. Buneman, « Dissipation of Currents in Ionized Media », *in*: *Physical Review* 115.*3* (Aug. 1959) (referenced on page 71).

[33] P. de Buyl, « Numerical Resolution of the Vlasov Equation for the Hamiltonian Mean-Field Model », *in*: *Communications in Nonlinear Science and Numerical Simulation* 15.*8* (Aug. 2010) (referenced on page 86).

[34] A. Caljub-Simon, « Existence globale d’une solution du problème de Cauchy pour le système d’équations aux dérivées partielles de Liouville-Newton », *in*: *Comptes Rendus de l’Académie des Sciences : série A* 276 (1973) (referenced on page 63).

[35] M. Campos Pinto, « Towards Smooth Particle Methods without Smoothing », *in*: *Journal of Scientific Computing* 65.*1* (Oct. 2015) (referenced on page 72).

[36] M. Campos Pinto and M. Mehrenberger, « Convergence of an Adaptive Semi-Lagrangian Scheme for the Vlasov-Poisson System », *in*: *Numerische Mathematik* 108.*3* (Jan. 2008) (referenced on page 68).

[37] M. Campos Pinto, E. Sonnendrücker, A. Friedman, D. P. Grote, and S. M. Lund, « Noiseless Vlasov–Poisson Simulations with Linearly Transformed Particles », *in*: *Journal of Computational Physics* 275 (Oct. 2014) (referenced on pages 72, 93, 96).

[38] F. Casas, N. Crouseilles, E. Faou, and M. Mehrenberger, « High-Order Hamiltonian Splitting for the Vlasov–Poisson Equations », *in*: *Numerische Mathematik* 135.*3* (Mar. 2017) (referenced on pages 75, 76, 84, 99, 102, 103).

[39] F. Charles, B. Després, R. Dai, and S. A. Hirstoaga, « Discrete Moments Models for Vlasov Equations with Non Constant Strong Magnetic Limit », *in*: *Comptes Rendus. Mécanique* 351.*S1* (Nov. 2023) (referenced on page 66).

[40] F. Charles, B. Després, and M. Mehrenberger, « Enhanced Convergence Estimates for Semi-Lagrangian Schemes Application to the Vlasov-Poisson Equation », *in*: *SIAM Journal on Numerical Analysis* 51.*2* (Jan. 2013) (referenced on page 68).

[41] P. Chartier, E. Darrigrand, and E. Faou, « A Regular Fast Multipole Method for Geometric Numerical Integrations of Hamiltonian Systems », *in*: *BIT Numerical Mathematics* 50.*1* (Mar. 2010) (referenced on page 73).

[42] F. F. Chen, *Introduction to Plasma Physics*, 1st ed., Boston, MA: Springer US, 1974 (referenced on page 48).

[43] C. Cheng and G. Knorr, « The Integration of the Vlasov Equation in Configuration Space », *in*: *Journal of Computational Physics* 22.*3* (Nov. 1976) (referenced on pages 64, 67).

[44] Y. Cheng, I. M. Gamba, F. Li, and P. J. Morrison, « Discontinuous Galerkin Methods for the Vlasov–Maxwell Equations », *in*: *SIAM Journal on Numerical Analysis* 52.*2* (Jan. 2014) (referenced on page 67).

[45] Y. Cheng, I. M. Gamba, and P. J. Morrison, « Study of Conservation and Recurrence of Runge–Kutta Discontinuous Galerkin Schemes for Vlasov–Poisson Systems », *in*: *Journal of Scientific Computing* 56.*2* (Aug. 2013) (referenced on page 67).

[46] A. Christlieb, R. Krasny, J. Verboncoeur, J. Emhoff, and I. Boyd, « Grid-Free Plasma Simulation Techniques », *in*: *IEEE Transactions on Plasma Science* 34.*2* (Apr. 2006) (referenced on page 73).

[47] B. A. Cipra, « The Best of the 20th Century: Editors Name Top 10 Algorithms », *in*: *SIAM News* 33.*4* (May 2000) (referenced on page 73).

[48] A. Cohen and B. Perthame, « Optimal Approximations of Transport Equations by Particle and Pseudoparticle Methods », *in*: *SIAM Journal on Mathematical Analysis* 32.*3* (Jan. 2000) (referenced on page 74).

[49] G. H. Cottet, « Convergence of a Vortex in Cell Method for the Two-Dimensional Euler Equations », *in*: (1987) (referenced on page 78).

[50] G.-H. Cottet, J. Goodman, and T. Y. Hou, « Convergence of the Grid-Free Point Vortex Method for the Three-Dimensional Euler Equations », *in*: *SIAM Journal on Numerical Analysis* 28.*2* (1991), JSTOR: 2157814 (referenced on page 78).

[51] W. Crookes, « On Radiant Matter », *in*: *Journal of the Franklin Institute* 108.*5* (Nov. 1879) (referenced on page 48).

[52] N. Crouseilles, L. Einkemmer, and E. Faou, « Hamiltonian Splitting for the Vlasov–Maxwell Equations », *in*: *Journal of Computational Physics* 283 (Feb. 2015) (referenced on page 62).

[53] N. Crouseilles, T. Respaud, and E. Sonnendrücker, « A Forward Semi-Lagrangian Method for the Numerical Solution of the Vlasov Equation », *in*: *Computer Physics Communications* 180.*10* (Oct. 2009) (referenced on pages 68, 95).

[54] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, « Scientific Machine Learning Through Physics–Informed Neural Networks: Where We Are and What's Next », *in*: *Journal of Scientific Computing* 92.*3* (Sept. 2022) (referenced on page 64).

[55] J. Dawson, « One-Dimensional Plasma Model », *in*: *The Physics of Fluids* 5.*4* (Apr. 1962) (referenced on page 71).

[56] J. Dawson, « Thermal Relaxation in a One-Species, One-Dimensional Plasma », *in*: *The Physics of Fluids* 7.*3* (Mar. 1964) (referenced on page 71).

[57] P. Degond and S. Mas-Gallic, « The Weighted Particle Method for Convection-Diffusion Equations Part 1: The Case of an Isotropie Viscosity », *in*: (1989) (referenced on pages 76, 86).

[58] P. Degond and C. Bardos, « Global Existence for the Vlasov-Poisson Equation in 3 Space Variables with Small Initial Data », *in*: *Annales de l'Institut Henri Poincaré C, Analyse non linéaire* 2.*2* (Apr. 1985) (referenced on page 63).

[59] G. Delzanno, « Multi-Dimensional, Fully-Implicit, Spectral Method for the Vlasov–Maxwell Equations with Exact Conservation Laws in Discrete Form », *in*: *Journal of Computational Physics* 301 (Nov. 2015) (referenced on page 66).

[60] L. Einkemmer and A. Ostermann, « A Strategy to Suppress Recurrence in Grid-Based Vlasov Solvers », *in*: *The European Physical Journal D* 68.*7* (July 2014), arXiv: 1401.4809 (referenced on page 93).

[61] C. Eldred, F. Gay-Balmaz, S. Huraka, and V. Putkaradze, « Lie–Poisson Neural Networks (LPNets): Data-based Computing of Hamiltonian Systems with Symmetries », *in*: *Neural Networks* 173 (May 2024) (referenced on page 64).

[62] B. Eliasson, « Numerical Modelling of the Two-Dimensional Fourier Transformed Vlasov–Maxwell System », *in*: *Journal of Computational Physics* 190.*2* (Sept. 2003) (referenced on page 66).

[63] B. Eliasson, « Outflow Boundary Conditions for the Fourier Transformed One-Dimensional Vlasov–Poisson System », *in*: *Journal of Scientific Computing* 16.*1* (2001) (referenced on page 66).

[64] B. Eliasson, « Outflow Boundary Conditions for the Fourier Transformed Two-Dimensional Vlasov Equation », *in*: *Journal of Computational Physics* 181.*1* (Sept. 2002) (referenced on page 66).

[65] E. G. Evstatiev and B. A. Shadwick, « Variational Formulation of Particle Algorithms for Kinetic Plasma Simulations », *in*: *Journal of Computational Physics* 245 (July 2013) (referenced on page 73).

[66] K. Feng and M. Qin, *Symplectic Geometric Algorithms for Hamiltonian Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010 (referenced on pages 84, 89).

[67] F. Filbet and E. Sonnendrücker, « Numerical Methods for the Vlasov Equation », *in*: *Numerical Mathematics and Advanced Applications*, Milano: Springer Milan, 2003 (referenced on page 67).

[68] F. Filbet, « Convergence of a Finite Volume Scheme for the Vlasov-Poisson System », *in*: *SIAM Journal on Numerical Analysis* 39.*4* (Jan. 2001) (referenced on page 67).

[69] A. M. Fridman and V. L. Polyachenko, *Physics of Gravitating Systems I*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1984 (referenced on page 53).

[70] C. Fu, Q. Guo, T. Gast, C. Jiang, and J. Teran, « A Polynomial Particle-in-Cell Method », *in*: *ACM Transactions on Graphics* 36.*6* (Nov. 2017) (referenced on page 72).

[71] R. R. Gagné and M. M. Shoucri, « A Splitting Scheme for the Numerical Solution of a One-Dimensional Vlasov Equation », *in*: *Journal of Computational Physics* 24.*4* (Aug. 1977) (referenced on page 67).

[72] R. A. Gingold and J. J. Monaghan, « Smoothed Particle Hydrodynamics: Theory and Application to Non-Spherical Stars », *in*: *Monthly Notices of the Royal Astronomical Society* 181.*3* (Dec. 1977) (referenced on page 73).

[73] R. T. Glassey, *The Cauchy Problem in Kinetic Theory*, Society for Industrial and Applied Mathematics, Jan. 1996 (referenced on pages 50–52, 61).

[74] J. Goodman, T. Y. Hou, and J. Lowengrub, « Convergence of the Point Vortex Method for the 2-D Euler Equations », *in*: *Communications on Pure and Applied Mathematics* 43.*3* (Apr. 1990) (referenced on page 78).

[75] H. Grad, « On the Kinetic Theory of Rarefied Gases », *in*: *Communications on Pure and Applied Mathematics* 2.*4* (Dec. 1949) (referenced on page 64).

[76] L. Greengard and V. Rokhlin, « A Fast Algorithm for Particle Simulations », *in*: *Journal of Computational Physics* 73.*2* (Dec. 1987) (referenced on page 73).

[77] E. Hairer and G. Wanner, « A Theory for Nyström Methods », *in*: *Numerische Mathematik* 25.*4* (Dec. 1975) (referenced on page 84).

[78] O. H. Hald, « Convergence of Vortex Methods for Euler's Equations. II », *in*: *SIAM Journal on Numerical Analysis* 16.*5* (Oct. 1979) (referenced on page 78).

[79] F. H. Harlow, *The Particle-in-Cell Method for Numerical Solution of Problems in Fluid Dynamics*, tech. rep. LADC-5288, 4769185, Mar. 1962 (referenced on page 71).

[80] R. Heath, I. Gamba, P. Morrison, and C. Michler, « A Discontinuous Galerkin Method for the Vlasov–Poisson System », *in*: *Journal of Computational Physics* 231.*4* (Feb. 2012) (referenced on page 67).

[81] M. M. Hejlesen, J. T. Rasmussen, P. Chatelain, and J. H. Walther, « A High Order Solver for the Unbounded Poisson Equation », *in*: *Journal of Computational Physics* 252 (Nov. 2013) (referenced on page 78).

[82] D. W. Hewett, « Fragmentation, Merging, and Internal Dynamics for PIC Simulation with Finite Size Particles », *in*: *Journal of Computational Physics* 189.*2* (Aug. 2003) (referenced on page 72).

[83] H. Higuchi, J. W. Pedersen, and A. Yoshikawa, *Quantum Calculation of Classical Kinetic Equations: A Novel Approach for Numerical Analysis of 6D Boltzmann-Maxwell Equations in Collisionless Plasmas Using Quantum Computing*, June 2023, arXiv: 2306.05967 (referenced on page 64).

[84] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, IOP Publishing Ltd, 1988 (referenced on pages 58, 61, 73).

[85] R. W. Hockney, « Computer Experiment of Anomalous Diffusion », *in*: *The Physics of Fluids* 9.*9* (Sept. 1966) (referenced on page 71).

[86] F. Holderied, S. Possanner, A. Ratnani, and X. Wang, « Structure-Preserving vs. Standard Particle-in-Cell Methods: The Case of an Electron Hybrid Model », *in*: *Journal of Computational Physics* 402 (Feb. 2020) (referenced on page 62).

[87] J. P. Holloway, « Spectral Velocity Discretizations for the Vlasov-Maxwell Equations », *in*: *Transport Theory and Statistical Physics* 25.*1* (Jan. 1996) (referenced on page 66).

[88] E. Horst, « On the Asymptotic Growth of the Solutions of the Vlasov-Poisson System », *in*: *Mathematical Methods in the Applied Sciences* 16.*2* (Feb. 1993) (referenced on page 63).

[89] E. Horst, « On the Classical Solutions of the Initial Value Problem for the Unmodified Non-Linear Vlasov Equation I General Theory », *in*: *Mathematical Methods in the Applied Sciences* 3.*1* (1981) (referenced on page 63).

[90] E. Horst, « On the Classical Solutions of the Initial Value Problem for the Unmodified Non-Linear Vlasov Equation II Special Cases », *in*: *Mathematical Methods in the Applied Sciences* 4.*1* (1982) (referenced on page 63).

[91] S. V. Iordanskii, « The Cauchy Problem for the Kinetic Equation of Plasma », *in*: *Twelve Papers on Analysis and Applied Mathematics*, vol. 35, American Mathematical Society Translations: Series 2, Providence, Rhode Island: American Mathematical Society, 1964 (referenced on page 63).

[92] B. Izrar, A. Ghizzo, P. Bertrand, E. Fijalkow, and M. Feix, « Integration of Vlasov Equation by a Fast Fourier Eulerian Code », *in*: *Computer Physics Communications* 52.*3* (Mar. 1989) (referenced on page 67).

[93] R. James, « The Solution of Poisson's Equation for Isolated Source Distributions », *in*: *Journal of Computational Physics* 25.*2* (Oct. 1977) (referenced on page 67).

[94] J. H. Jeans, « On the Theory of Star-Streaming and the Structure of the Universe », *in*: *Monthly Notices of the Royal Astronomical Society* 76.*2* (Dec. 1915) (referenced on page 53).

[95] G. Joyce, G. Knorr, and H. K. Meier, « Numerical Integration Methods of the Vlasov Equation », *in*: *Journal of Computational Physics* 8.*1* (Aug. 1971) (referenced on page 65).

[96] P. J. Kellogg, « Some Properties of the Two-Stream Instability at Large Amplitudes », *in*: *The Physics of Fluids* 8.*1* (Jan. 1965) (referenced on page 67).

[97] A. J. Klimas, « A Method for Overcoming the Velocity Space Filamentation Problem in Collisionless Plasma Model Solutions », *in*: *Journal of Computational Physics* 68.*1* (Jan. 1987) (referenced on pages 64, 67).

[98] A. J. Klimas, « A Numerical Method Based on the Fourier-Fourier Transform Approach for Modeling 1-D Electron Plasma Evolution », *in*: *Journal of Computational Physics* 50.*2* (May 1983) (referenced on page 66).

[99] G. Knorr, « Zur lösung der nicht-linearen Vlasov-gleichung », *in*: *Zeitschrift für Naturforschung A* 18.*12* (Dec. 1963) (referenced on pages 64, 66).

[100] M. Kraus, K. Kormann, P. J. Morrison, and E. Sonnendrücker, « GEMPIC: Geometric Electromagnetic Particle-in-Cell Methods », *in*: *Journal of Plasma Physics* 83.*4* (Aug. 2017) (referenced on pages 52, 62, 93).

[101] L. D. Landau, « On the Vibrations of the Electronic Plasma », *in*: *Collected Papers of L.D. Landau*, Elsevier, 1946 (referenced on page 50).

[102] A. B. Langdon, « Effects of the Spatial Grid in Simulation Plasmas », *in*: *Journal of Computational Physics* 6.*2* (Oct. 1970) (referenced on page 71).

[103] A. B. Langdon, « Theory of Plasma Simulation Using Finite-Size Particles », *in*: *Physics of Fluids* 13.*8* (1970) (referenced on page 71).

[104] I. Langmuir, « Oscillations in Ionized Gases », *in*: *Proceedings of the National Academy of Sciences* 14.*8* (Aug. 1928) (referenced on page 48).

[105] Y. Le Henaff, « Grid-Free Weighted Particle Method Applied to the Vlasov–Poisson Equation », *in*: *Numerische Mathematik* 155.*3-4* (Dec. 2023) (referenced on pages 54, 75, 129).

[106] P. L. Lions and B. Perthame, « Propagation of Moments and Regularity for the 3-Dimensional Vlasov-Poisson System », *in*: *Inventiones Mathematicae* 105.*1* (Dec. 1991) (referenced on page 63).

[107] B. M. Marder, « GAP–a PIC-type Fluid Code », *in*: *Mathematics of Computation* 29.*130* (Apr. 1975), JSTOR: 2005562 (referenced on page 72).

[108] J. C. Maxwell, « VIII. A Dynamical Theory of the Electromagnetic Field », *in*: *Philosophical Transactions of the Royal Society of London* 155 (Dec. 1865) (referenced on page 52).

[109] R. L. McCrory, R. L. Morse, and K. A. Taggart, « Growth and Saturation of Instability of Spherical Implosions Driven by Laser or Charged Particle Beams », *in*: *Nuclear Science and Engineering* 64.*1* (Sept. 1977) (referenced on page 72).

[110] M. Mehrenberger, « Recurrence Phenomenon for Vlasov-Poisson Simulations on Regular Finite Element Mesh », *in*: *Communications in Computational Physics* 28.*3* (June 2020) (referenced on page 93).

[111] M. S. Mitchell, M. T. Miecnikowski, G. Beylkin, and S. E. Parker, « Efficient Fourier Basis Particle Simulation », *in*: *Journal of Computational Physics* 396 (Nov. 2019) (referenced on pages 73, 75, 86).

[112] J. J. Monaghan, « Smoothed Particle Hydrodynamics », *in*: *Annual Review of Astronomy and Astrophysics* 30.*1* (Sept. 1992) (referenced on page 73).

[113] J. Monaghan, « Particle Methods for Hydrodynamics », *in*: *Computer Physics Reports* 3.*2* (Oct. 1985) (referenced on page 73).

[114] R. L. Morse and C. W. Nielson, « Numerical Simulation of Warm Two-Beam Plasma », *in*: *The Physics of Fluids* 12.*11* (Nov. 1969) (referenced on page 71).

[115] T. Nakamura and T. Yabe, « Cubic Interpolated Propagation Scheme for Solving the Hyper-Dimensional Vlasov-Poisson Equation in Phase Space », *in*: *Computer Physics Communications* 120.*2-3* (Aug. 1999) (referenced on page 67).

[116] R. Nguyen Van Yen, É. Sonnendrücker, K. Schneider, and M. Farge, « Particle-in-Wavelets Scheme for the 1D Vlasov-Poisson Equations », *in*: *ESAIM: Proceedings* 32 (Oct. 2011) (referenced on page 96).

[117] A. Nishiguchi and T. Yabe, « Second-Order Fluid Particle Scheme », *in*: *Journal of Computational Physics* 52.*2* (Nov. 1983) (referenced on page 73).

[118] J. Nührenberg, « A Difference Scheme for Vlasov's Equation », *in*: *Zeitschrift für angewandte Mathematik und Physik ZAMP* 22.*6* (Nov. 1971) (referenced on pages 64, 65, 67).

[119] H. Okuda, « Nonphysical Noises and Instabilities in Plasma Simulation Due to a Spatial Grid », *in*: *Journal of Computational Physics* 10.*3* (Dec. 1972) (referenced on page 71).

[120] L. Pareschi and T. Rey, *Moment Preserving Fourier-Galerkin Spectral Methods and Application to the Boltzmann Equation*, May 2021, arXiv: `2105.13158` (referenced on page 86).

[121] M. Perlman, « On the Accuracy of Vortex Methods », *in*: *Journal of Computational Physics* 59.*2* (June 1985) (referenced on page 78).

[122] K. Pfaffelmoser, « Global Classical Solutions of the Vlasov-Poisson System in Three Dimensions for General Initial Data », *in*: *Journal of Differential Equations* 95.*2* (Feb. 1992) (referenced on page 63).

[123] M. C. Pinto, J. Ameres, K. Kormann, and E. Sonnendrücker, *On Geometric Fourier Particle in Cell Methods*, Feb. 2021, arXiv: 2102.02106 (referenced on pages 62, 93).

[124] M. Qin, « Symplectic Schemes for Nonautonomous Hamiltonian System », *in*: *Acta Mathematicae Applicatae Sinica* 12.*3* (July 1996) (referenced on page 84).

[125] M. Raissi, P. Perdikaris, and G. Karniadakis, « Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations », *in*: *Journal of Computational Physics* 378 (Feb. 2019) (referenced on page 64).

[126] P. A. Raviart, « An Analysis of Particle Methods », *in*: *Numerical Methods in Fluid Dynamics*, vol. 1127, Springer Berlin Heidelberg, 1985 (referenced on page 74).

[127] E. Rein, « Collisionless Kinetic Equations from Astrophysics – the Vlasov–Poisson System », *in*: *Handbook of Differential Equations: Evolutionary Equations*, 1st ed, Amsterdam Boston: Elsevier/North Holland, 2004 (referenced on page 53).

[128] T. Respaud and E. Sonnendrücker, « Analysis of a New Class of Forward Semi-Lagrangian Schemes for the 1D Vlasov-Poisson Equations », *in*: *Numerische Mathematik* 118.*2* (June 2011) (referenced on page 78).

[129] J. Schaeffer, « Global Existence of Smooth Solutions to the Vlasov-Poisson System in Three Dimensions », *in*: *Communications in Partial Differential Equations* 16.*8-9* (Jan. 1991) (referenced on page 63).

[130] R. Schneider and R. Kleiber, « 16 Computational Plasma Physics », *in*: (2005) (referenced on page 74).

[131] E. Sonnendrucker, « Numerical Methods for the Vlasov-Maxwell Equations », *in*: (Jan. 2016) (referenced on pages 92, 93, 95).

[132] E. Sonnendrücker, J. Roche, P. Bertrand, and A. Ghizzo, « The Semi-Lagrangian Method for the Numerical Resolution of the Vlasov Equation », *in*: *Journal of Computational Physics* 149.*2* (Mar. 1999) (referenced on page 67).

[133] H. Spohn, *Large Scale Dynamics of Interacting Particles*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1991 (referenced on page 53).

[134] T. Tajima, J. Leboeuf, and J. Dawson, « A Magnetohydrodynamic Particle Code with Force Free Electrons for Fluid Simulations », *in*: *Journal of Computational Physics* 38.*2* (Nov. 1980) (referenced on page 73).

[135] V. I. Telegin, « A Difference Scheme for Vlasov's Equation », *in*: *Zh. Vychisl. Mat. mat. Fiz.* 16.*5* (1976) (referenced on page 67).

[136] L. Tonks, « The Birth of “Plasma” », *in*: *American Journal of Physics* 35.*9* (Sept. 1967) (referenced on page 48).

[137] S. Ukai and T. Okabe, « On Classical Solutions in the Large in Time of Two-Dimensional Vlasov's Equation », *in*: *Osaka Journal of Mathematics* 15.*2* (1978) (referenced on page 63).

[138] J. P. Verboncoeur, « Particle Simulation of Plasmas: Review and Advances », *in*: *Plasma Physics and Controlled Fusion* 47.*5A* (May 2005) (referenced on page 71).

[139] A. A. Vlasov, « The Vibrational Properties of an Electron Gas », *in*: *Soviet Physics Uspekhi* 10.*6* (June 1968) (referenced on page 52).

[140] S. Wollman, « On the Approximation of the Vlasov-Poisson System by Particle Methods », *in*: *SIAM Journal on Numerical Analysis* 37.*4* (Jan. 2000) (referenced on page 78).

[141] S. Wollman and E. Ozizmir, « Numerical Approximation of the One-Dimensional Vlasov–Poisson System with Periodic Boundary Conditions », *in*: *SIAM Journal on Numerical Analysis* 33.*4* (Aug. 1996) (referenced on pages 78, 86).

[142] E. Ye and N. F. G. Loureiro, « Quantum-Inspired Method for Solving the Vlasov-Poisson Equations », *in*: *Physical Review E* 106.*3* (Sept. 2022) (referenced on page 64).

[143] H. Yoshida, « Construction of Higher Order Symplectic Integrators », *in*: *Physics Letters A* 150.*5-7* (Nov. 1990) (referenced on page 84).

[144] B. Zhang, G. Cai, H. Weng, W. Wang, L. Liu, and B. He, « Physics-Informed Neural Networks for Solving Forward and Inverse Vlasov–Poisson Equation via Fully Kinetic Simulation », *in*: *Machine Learning: Science and Technology* 4.*4* (Dec. 2023) (referenced on page 64).

References for The Schrödinger equation

[1] L. Adamowicz, S. Kvaal, C. Lasser, and T. B. Pedersen, « Laser-Induced Dynamic Alignment of the HD Molecule without the Born–Oppenheimer Approximation », *in*: *The Journal of Chemical Physics* 157.*14* (Oct. 2022) (referenced on pages 157, 202).

[2] G. Agrawal, *Nonlinear Fiber Optics*, Elsevier, 2013 (referenced on page 135).

[3] P. Alphonse and J. Bernier, « Polar Decomposition of Semigroups Generated by Non-Selfadjoint Quadratic Differential Operators and Regularizing Effects », *in*: *Annales scientifiques de l'École Normale Supérieure* 56.*2* (2023) (referenced on pages 146, 179).

[4] J. B. Anderson, « A Random-Walk Simulation of the Schrödinger Equation: H+3 », *in*: *The Journal of Chemical Physics* 63.*4* (Aug. 1975) (referenced on page 141).

[5] X. Antoine, W. Bao, and C. Besse, « Computational Methods for the Dynamics of the Nonlinear Schrödinger/Gross-Pitaevskii Equations », *in*: *Computer Physics Communications* 184.*12* (Dec. 2013) (referenced on pages 179, 203).

[6] K. E. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed, New York: Wiley, 1989 (referenced on page 203).

[7] W. Bao and Q. Du, « Computing the Ground State Solution of Bose–Einstein Condensates by a Normalized Gradient Flow », *in*: *SIAM Journal on Scientific Computing* 25.*5* (Jan. 2004) (referenced on pages 142, 143).

[8] W. Bao, D. Jaksch, and P. A. Markowich, « Numerical Solution of the Gross–Pitaevskii Equation for Bose–Einstein Condensation », *in*: *Journal of Computational Physics* 187.*1* (May 2003) (referenced on pages 146, 156, 187, 202).

[9] W. Bao, S. Jin, and P. A. Markowich, « On Time-Splitting Spectral Approximations for the Schrödinger Equation in the Semiclassical Regime », *in*: *Journal of Computational Physics* 175.*2* (Jan. 2002) (referenced on page 146).

[10] W. Bao, H. Li, and J. Shen, « A Generalized-Laguerre–Fourier–Hermite Pseudospectral Method for Computing the Dynamics of Rotating Bose-Einstein Condensates », *in*: *SIAM Journal on Scientific Computing* 31.*5* (Jan. 2009) (referenced on pages 148, 149, 207).

[11] W. Bao and J. Shen, « A Fourth-Order Time-Splitting Laguerre–Hermite Pseudospectral Method for Bose–Einstein Condensates », *in*: *SIAM Journal on Scientific Computing* 26.*6* (Jan. 2005) (referenced on pages 148, 149, 207).

[12] P. Bergold and C. Lasser, *The Gaussian Wave Packet Transform via Quadrature Rules*, June 2023, arXiv: 2010.03478 (referenced on page 202).

[13] J. Bernier, « Exact Splitting Methods for Semigroups Generated by Inhomogeneous Quadratic Differential Operators », *in*: *Foundations of Computational Mathematics* 21.*5* (Oct. 2021) (referenced on pages 146, 179, 188).

[14] J. Bernier, N. Crouseilles, and Y. Li, « Exact Splitting Methods for Kinetic and Schrödinger Equations », *in*: *Journal of Scientific Computing* 86.*1* (Jan. 2021) (referenced on pages 146, 156, 179, 203).

[15] Å. Björck, « Least Squares Methods », *in*: *Handbook of Numerical Analysis*, vol. 1, Elsevier, 1990 (referenced on page 192).

[16] M. Born, « Quantenmechanik der stoßvorgänge », *in*: *Zeitschrift für Physik* 38.*11-12* (Nov. 1926) (referenced on page 134).

[17] J. Broeckhove, L. Lathouwers, E. Kesteloot, and P. Van Leuven, « On the Equivalence of Time-Dependent Variational Principles », *in*: *Chemical Physics Letters* 149.*5-6* (Sept. 1988) (referenced on page 152).

[18] R. Carles, *Semi-Classical Analysis for Nonlinear Schrödinger Equations: WKB Analysis, Focal Points, Coherent States*, 2nd edition, New Jersey: World Scientific, 2021 (referenced on pages 139, 188).

[19] F. Casas, N. Crouseilles, E. Faou, and M. Mehrenberger, « High-Order Hamiltonian Splitting for the Vlasov–Poisson Equations », *in*: *Numerische Mathematik* 135.*3* (Mar. 2017) (referenced on pages 155, 188, 198).

[20] T. Cazenave, *Semilinear Schrödinger Equations*, vol. 10, Courant Lecture Notes, Providence, Rhode Island: American Mathematical Society, Sept. 2003 (referenced on pages 137, 139, 140).

[21] M. M. Cerimele, M. L. Chiofalo, F. Pistella, S. Succi, and M. P. Tosi, « Numerical Solution of the Gross-Pitaevskii Equation Using an Explicit Finite-Difference Scheme: An Application to Trapped Bose-Einstein Condensates », *in*: *Physical Review E* 62.*1* (July 2000) (referenced on page 187).

[22] Q. Chang, E. Jia, and W. Sun, « Difference Schemes for Solving the Generalized Nonlinear Schrödinger Equation », *in*: *Journal of Computational Physics* 148.*2* (Jan. 1999) (referenced on page 146).

[23] M. L. Chiofalo, S. Succi, and M. P. Tosi, « Ground State of Trapped Interacting Bose-Einstein Condensates by an Explicit Imaginary-Time Algorithm », *in*: *Physical Review E* 62.*5* (Nov. 2000) (referenced on page 141).

[24] R. D. Coalson and M. Karplus, « Multidimensional Variational Gaussian Wave Packet Dynamics with Application to Photodissociation Spectroscopy », *in*: *The Journal of Chemical Physics* 93.*6* (Sept. 1990) (referenced on page 157).

[25] P. Decleer, A. Van Londersele, H. Rogier, and D. Vande Ginste, « An Alternating-Direction Hybrid Implicit-Explicit Finite-Difference Time-Domain Method for the Schrödinger Equation », *in*: *Journal of Computational and Applied Mathematics* 403 (Mar. 2022) (referenced on page 146).

[26] P. A. M. Dirac, « Note on Exchange Phenomena in the Thomas Atom », *in*: *Mathematical Proceedings of the Cambridge Philosophical Society* 26.*3* (July 1930) (referenced on page 153).

[27] C. C. Douglas, L. Lee, and M.-C. Yeung, « On Solving Ill Conditioned Linear Systems », *in*: *Procedia Computer Science* 80 (2016) (referenced on page 192).

[28] E. Faou, *Geometric Numerical Integration and Schrödinger Equations*, Zurich Lectures in Advanced Mathematics, Zürich, Switzerland: European Mathematical Society, 2012 (referenced on pages 156, 198).

[29] E. Faou and T. Jézéquel, « Convergence of a Normalized Gradient Algorithm for Computing Ground States », *in*: *IMA Journal of Numerical Analysis* 38.*1* (Jan. 2018) (referenced on page 143).

[30] E. Faou, Y. Le Henaff, and P. Raphaël, *Modulation Algorithm for the Nonlinear Schrödinger Equation*, Oct. 2023, arXiv: 2303.13969 (referenced on pages 136, 154, 186).

[31] E. Faou and P. Raphaël, « On Weakly Turbulent Solutions to the Perturbed Linear Harmonic Oscillator », *in*: *American Journal of Mathematics* 145.*5* (Oct. 2023) (referenced on pages 154, 187, 206, 207).

[32] M. Feit, J. Fleck, and A. Steiger, « Solution of the Schrödinger Equation by a Spectral Method », *in*: *Journal of Computational Physics* 47.*3* (Sept. 1982) (referenced on page 146).

[33] G. Fibich, *The Nonlinear Schrödinger Equation: Singular Solutions and Optical Collapse*, vol. 192, Applied Mathematical Sciences, Cham: Springer International Publishing, 2015 (referenced on page 135).

[34] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge Monographs on Applied and Computational Mathematics 1, Cambridge ; New York: Cambridge University Press, 1996 (referenced on page 179).

[35] J. Frenkel, *Wave Mechanics - Advanced General Theory*, 1st, Oxford University Press, 1934 (referenced on page 153).

[36] P. Gérard, E. Lenzmann, O. Pocovnicu, and P. Raphaël, « A Two-Soliton with Transient Turbulent Regime for the Cubic Half-Wave Equation on the Real Line », *in*: *Annals of PDE* 4.*1* (June 2018) (referenced on page 188).

[37] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Fourth edition, Johns Hopkins Studies in the Mathematical Sciences, Baltimore: The Johns Hopkins University Press, 2013 (referenced on pages 192, 197).

[38] S. J. Gustafson and I. M. Sigal, *Mathematical Concepts of Quantum Mechanics*, Universitext, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011 (referenced on page 138).

[39] G. A. Hagedorn, « Raising and Lowering Operators for Semiclassical Wave Packets », *in*: *Annals of Physics* 269.*1* (Oct. 1998) (referenced on page 151).

[40] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed, Springer Series in Computational Mathematics 31, Berlin ; New York: Springer, 2006 (referenced on pages 155, 166, 168, 188, 198).

[41] E. J. Heller, « Frozen Gaussians: A Very Simple Semiclassical Approximation », *in*: *The Journal of Chemical Physics* 75.*6* (Sept. 1981) (referenced on page 150).

[42] E. J. Heller, « Time Dependent Variational Approach to Semiclassical Dynamics », *in*: *The Journal of Chemical Physics* 64.*1* (Jan. 1976) (referenced on pages 149, 157, 206).

[43] E. J. Heller, « Time-Dependent Approach to Semiclassical Dynamics », *in*: *The Journal of Chemical Physics* 62.*4* (Feb. 1975) (referenced on pages 149, 206).

[44] D. Huber and E. J. Heller, « Generalized Gaussian Wave Packet Dynamics », *in*: *The Journal of Chemical Physics* 87.*9* (Nov. 1987) (referenced on page 157).

[45] S. Jin, P. Markowich, and C. Sparber, « Mathematical and Computational Methods for Semiclassical Schrödinger Equations », *in*: *Acta Numerica* 20 (May 2011) (referenced on page 179).

[46] T. Kato, *Perturbation Theory for Linear Operators*, vol. 132, Classics in Mathematics, Berlin, Heidelberg: Springer Berlin Heidelberg, 1995 (referenced on page 138).

[47] K. G. Kay, « The Matrix Singularity Problem in the Time-Dependent Variational Method », *in*: *Chemical Physics* 137.*1-3* (Oct. 1989) (referenced on pages 153, 203, 205).

[48] R. Killip and M. Visan, « Nonlinear Schrödinger Equations at Critical Regularity », 2013 (referenced on page 157).

[49] D. Kosloff and R. Kosloff, « A Fourier Method Solution for the Time Dependent Schrödinger Equation as a Tool in Molecular Dynamics », *in*: *Journal of Computational Physics* 52.*1* (Oct. 1983) (referenced on page 146).

[50] S. Kvaal, C. Lasser, T. B. Pedersen, and L. Adamowicz, *No Need for a Grid: Adaptive Fully-Flexible Gaussians for the Time-Dependent Schrödinger Equation*, Mar. 2023, arXiv: 2207.00271 (referenced on pages 149, 205, 207).

[51] C. Lasser and C. Lubich, « Computing Quantum Dynamics in the Semiclassical Regime », *in*: *Acta Numerica* 29 (May 2020) (referenced on pages 150, 151, 153, 156, 189).

[52] C. Lasser and S. Troppmann, « Hagedorn Wavepackets in Time-Frequency and Phase Space », *in*: *Journal of Fourier Analysis and Applications* 20.*4* (Aug. 2014) (referenced on page 151).

[53] S. A. Lebedeff, « Time-Dependent Formulation of the Semiclassical Approximation in Collision Theory », *in*: *Physical Review* 165.*5* (Jan. 1968) (referenced on page 150).

[54] C. Leforestier, R. Bisseling, C. Cerjan, M. Feit, R. Friesner, A. Guldberg, A. Hammerich, G. Jolicard, W. Karrlein, H.-D. Meyer, N. Lipkin, O. Roncero, and R. Kosloff, « A Comparison of Different Propagation Schemes for the Time Dependent Schrödinger Equation », *in*: *Journal of Computational Physics* 94.*1* (May 1991) (referenced on page 146).

[55] C. Lubich, *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*, 1st ed., EMS Press, Sept. 2008 (referenced on pages 134, 137, 144).

[56] C. Lubich, « On Splitting Methods for Schrödinger-Poisson and Cubic Nonlinear Schrödinger Equations », *in*: *Mathematics of Computation* 77.*264* (Feb. 2008) (referenced on page 146).

[57] Y. Martel and P. Raphaël, « Strongly Interacting Blow up Bubbles for the Mass Critical Nonlinear Schrödinger Equation », *in*: *Annales scientifiques de l'École normale supérieure* 51.*3* (2018) (referenced on pages 154, 187, 206).

[58] A. McLachlan, « A Variational Solution of the Time-Dependent Schrödinger Equation », *in*: *Molecular Physics* 8.*1* (Jan. 1964) (referenced on page 152).

[59] R. I. McLachlan and G. R. W. Quispel, « Splitting Methods », *in*: *Acta Numerica 2002*, 1st ed., Cambridge University Press, July 2002 (referenced on pages 155, 188, 198).

[60] F. Merle and P. Raphaël, « The Blow-up Dynamic and Upper Bound on the Blow-up Rate for Critical Nonlinear Schrödinger Equation », *in*: *Annals of Mathematics* 161.*1* (Jan. 2005) (referenced on pages 154, 187, 206).

[61] F. Merle, P. Raphaël, I. Rodnianski, and J. Szeftel, *On the Implosion of a Three Dimensional Compressible Fluid*, 2020 (referenced on page 187).

[62] L. F. Mollenhauer and J. Gordon P., *Solitons in Optical Fibers*, Elsevier, 2006 (referenced on page 135).

[63] F. I. Moxley, T. Byrnes, B. Ma, Y. Yan, and W. Dai, « A G-FDTD Scheme for Solving Multi-Dimensional Open Dissipative Gross–Pitaevskii Equations », *in*: *Journal of Computational Physics* 282 (Feb. 2015) (referenced on page 146).

[64] D. Oelz and S. Trabelsi, « Analysis of a Relaxation Scheme for a Nonlinear Schrödinger Equation Occurring in Plasma Physics », *in*: *Mathematical Modelling and Analysis* 19.*2* (Apr. 2014) (referenced on page 135).

[65] I. V. Oseledets, « Tensor-Train Decomposition », *in*: *SIAM Journal on Scientific Computing* 33.*5* (Jan. 2011) (referenced on page 197).

[66] J. Qian and L. Ying, « Fast Gaussian Wavepacket Transforms and Gaussian Beams for the Schrödinger Equation », *in*: *Journal of Computational Physics* 229.*20* (Oct. 2010) (referenced on pages 156, 181, 201).

[67] *Quantum Theory and Measurement:* Princeton University Press, Dec. 1983 (referenced on page 134).

[68] A. Raab, « On the Dirac–Frenkel/McLachlan Variational Principle », *in*: (2000) (referenced on page 153).

[69] K. Rowan, L. Schatzki, T. Zaklama, Y. Suzuki, K. Watanabe, and K. Varga, « Simulation of a Hydrogen Atom in a Laser Field Using the Time-Dependent Variational Principle », *in*: *Physical Review E* 101.*2* (Feb. 2020) (referenced on pages 203, 205).

[70] A. K. Roy, A. J. Thakkar, and B. M. Deb, « Low-Lying States of Two-Dimensional Double-Well Potentials », *in*: *Journal of Physics A: Mathematical and General* 38.*10* (Mar. 2005) (referenced on page 146).

[71] A. K. Roy, N. Gupta, and B. M. Deb, « Time-Dependent Quantum-Mechanical Calculation of Ground and Excited States of Anharmonic and Double-Well Oscillators », *in*: *Physical Review A* 65.*1* (Dec. 2001) (referenced on page 146).

[72] E. Schrödinger, « An Undulatory Theory of the Mechanics of Atoms and Molecules », *in*: *Physical Review* 28.*6* (Dec. 1926) (referenced on page 133).

[73] J. Shen, T. Tang, and L.-L. Wang, *Spectral Methods: Algorithms, Analysis and Applications*, vol. 41, Springer Series in Computational Mathematics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011 (referenced on page 147).

[74] T. Simos and P. Williams, « A Finite-Difference Method for the Numerical Solution of the Schrödinger Equation », *in*: *Journal of Computational and Applied Mathematics* 79.*2* (Mar. 1997) (referenced on page 146).

[75] G. Strang, *Linear Algebra and Its Applications*, 4th ed, Belmont, CA: Thomson, Brooks/Cole, 2006 (referenced on page 192).

[76] I. W. Sudiarta and D. J. W. Geldart, « Solving the Schrödinger Equation Using the Finite Difference Time Domain Method », *in*: *Journal of Physics A: Mathematical and Theoretical* 40.*8* (Feb. 2007) (referenced on pages 143, 146).

[77] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation: Self-Focusing and Wave Collapse*, Applied Mathematical Sciences 139, New York Berlin Heidelberg: Springer, 1999 (referenced on page 135).

[78] D. M. Sullivan and D. S. Citrin, « Determination of the Eigenfunctions of Arbitrary Nanostructures Using Time Domain Simulation », *in*: *Journal of Applied Physics* 91.*5* (Mar. 2002) (referenced on pages 143, 145).

[79] T. Tao, *Nonlinear Dispersive Equations*, vol. 106, CBMS Regional Conference Series in Mathematics, Providence, Rhode Island: American Mathematical Society, June 2006 (referenced on page 157).

[80] M. Thalhammer, M. Caliari, and C. Neuhauser, « High-Order Time-Splitting Hermite and Fourier Spectral Methods », *in*: *Journal of Computational Physics* 228.*3* (Feb. 2009) (referenced on pages 148, 149, 207).

[81] T. Tsednee, B. Tsednee, and T. Khinayat, *Numerical Solution to the Time-Dependent Gross-Pitaevskii Equation*, Nov. 2022, arXiv: 2211.03964 (referenced on page 187).

[82] K. Varga, « Optimization of the Nonlinear Parameters of the Correlated Gaussian Basis Functions with Imaginary-Time Propagation », *in*: *Physical Review A* 99.*1* (Jan. 2019) (referenced on page 142).

[83] H. Wang, « Numerical Simulation for the Gross-Pitaevskii Equation Based on the Lattice Boltzmann Method », *in*: *Advances in Space Research* 60.*6* (Sept. 2017) (referenced on pages 156, 187).

[84] M. I. Weinstein, « Modulational Stability of Ground States of Nonlinear Schrödinger Equations », *in*: *SIAM Journal on Mathematical Analysis* 16.*3* (May 1985) (referenced on page 141).

[85] G. C. Wick, « Properties of Bethe-Salpeter Wave Functions », *in*: *Physical Review* 96.*4* (Nov. 1954) (referenced on page 141).

[86] G. A. Worth, M. A. Robb, and I. Burghardt, « A Novel Algorithm for Non-Adiabatic Direct Dynamics Using Variational Gaussian Wavepackets », *in*: *Faraday Discussions* 127 (2004) (referenced on page 157).

[87] Y. Xu and L. Zhang, « Alternating Direction Implicit Method for Solving Two-Dimensional Cubic Nonlinear Schrödinger Equation », *in*: *Computer Physics Communications* 183.*5* (May 2012) (referenced on page 146).

References for The spectral concentration problem

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed, Software, Environments, Tools, Philadelphia: Society for Industrial and Applied Mathematics, 1999 (referenced on page 236).

[2] S. Axler, *Linear Algebra Done Right*, Undergraduate Texts in Mathematics, Cham: Springer International Publishing, 2024 (referenced on page 261).

[3] J. P. Boyd, « Algorithm 840: Computation of Grid Points, Quadrature Weights and Derivatives for Spectral Element Methods Using Prolate Spheroidal Wave Functions—Prolate Elements », *in*: *ACM Transactions on Mathematical Software* 31.*1* (Mar. 2005) (referenced on page 244).

[4] O. Brander and B. DeFacio, « A Generalisation of Slepian's Solution for the Singular Value Decomposition of Filtered Fourier Transforms », *in*: *Inverse Problems* 2.*2* (May 1986) (referenced on pages 229, 231, 245).

[5] Q. Y. Chen, D. Gottlieb, and J. S. Hesthaven, « Spectral Methods Based on Prolate Spheroidal Wave Functions for Hyperbolic PDEs », *in*: *SIAM Journal on Numerical Analysis* 43.*5* (Jan. 2005) (referenced on page 244).

[6] A. Connes and H. Moscovici, « The UV Prolate Spectrum Matches the Zeros of Zeta », *in*: *Proceedings of the National Academy of Sciences* 119.*22* (May 2022) (referenced on page 244).

[7] I. Daubechies, « Time-Frequency Localization Operators: A Geometric Phase Space Approach », *in*: *IEEE Transactions on Information Theory* 34.*4* (July 1988) (referenced on page 233).

[8] R. Demesmaeker, M. G. Preti, and D. Van De Ville, « Augmented Slepians: Bandlimited Functions That Counterbalance Energy in Selected Intervals », *in*: *IEEE Transactions on Signal Processing* 66.*15* (Aug. 2018), arXiv: 1710.11251 (referenced on page 234).

[9] G. Dujardin and E. Faou, « Normal Form and Long Time Analysis of Splitting Schemes for the Linear Schrödinger Equation with Small Potential », *in*: *Numerische Mathematik* 108.*2* (Nov. 2007) (referenced on page 272).

[10] W. Erb, *An Orthogonal Polynomial Analogue of the Landau-Pollak-Slepian Time-Frequency Analysis*, Mar. 2012, arXiv: 1104.0615 (referenced on page 232).

[11] C. Flammer, *Spheroidal Wave Functions*, Dover Publications, 1957 (referenced on pages 227, 242, 243).

[12] Y. Grabovsky and N. Hovsepyan, « On the Commutation Properties of Finite Convolution and Differential Operators I: Commutation. », *in*: *Results in Mathematics* 76.*3* (Aug. 2021) (referenced on page 230).

[13] J. W. Green and G. L. Walker, « The American Mathematical Society », *in*: () (referenced on page 230).

[14] F. A. Grünbaum, « A Study of Fourier Space Methods for "Limited Angle" Image Reconstruction », *in*: *Numerical Functional Analysis and Optimization* 2.*1* (Jan. 1980) (referenced on pages 229, 231, 235).

[15] F. A. Grünbaum, « Differential Operators Commuting with Convolution Integral Operators », *in*: *Journal of Mathematical Analysis and Applications* 91.*1* (Jan. 1983) (referenced on page 229).

[16] F. A. Grünbaum, « Eigenvectors of a Toeplitz Matrix: Discrete Version of the Prolate Spheroidal Wave Functions », *in*: *SIAM Journal on Algebraic Discrete Methods* 2.*2* (June 1981) (referenced on page 233).

[17] F. A. Grünbaum, « Finite Convolution Integral Operators Commuting with Differential Operators: Some Counterexamples », *in*: *Numerical Functional Analysis and Optimization* 3.*2* (Jan. 1981) (referenced on page 232).

[18] F. A. Grünbaum, « Serendipity Strikes Again », *in*: *Proceedings of the National Academy of Sciences* 119.*26* (June 2022) (referenced on page 244).

[19] F. A. Grünbaum, L. Longhi, and M. Perlstadt, « Differential Operators Commuting with Finite Convolution Integral Operators: Some Non-Abelian Examples », *in*: (1982) (referenced on pages 231, 233).

[20] F. A. Grünbaum and L. Miranian, « The Magic of the Prolate Spheroidal Functions in Various Setups », *in*: 4478 (2001) (referenced on page 244).

[21] J. A. Hogan and J. D. Lakey, *Duration and Bandwidth Limiting: Prolate Functions, Sampling, and Applications*, Applied and Numerical Harmonic Analysis, Boston: Birkhäuser Boston, 2012 (referenced on page 244).

[22] L. Hörmander, *The Analysis of Linear Partial Differential Operators I*, Classics in Mathematics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003 (referenced on page 222).

[23] R. Kalaba, K. Spingarn, and L. Tesfatsion, « Variational Equations for the Eigenvalues and Eigenvectors of Nonsymmetric Matrices », *in*: *Journal of Optimization Theory and Applications* 33.*1* (Jan. 1981) (referenced on page 269).

[24] S. Karnik, J. Romberg, and M. A. Davenport, « Improved Bounds for the Eigenvalues of Prolate Spheroidal Wave Functions and Discrete Prolate Spheroidal Sequences », *in*: *Applied and Computational Harmonic Analysis* 55 (Nov. 2021) (referenced on pages 226, 234, 263).

[25] S. Karnik, Z. Zhu, M. B. Wakin, J. Romberg, and M. A. Davenport, « The Fast Slepian Transform », *in*: *Applied and Computational Harmonic Analysis* 46.*3* (May 2019) (referenced on page 234).

[26] A. Karoui, « Uncertainty Principles, Prolate Spheroidal Wave Functions, and Applications », *in*: *Recent Developments in Fractals and Related Fields*, Boston: Birkhäuser Boston, 2010 (referenced on pages 227, 244).

[27] A. Karoui and I. Mehrzi, « Asymptotic Behaviors and Numerical Computations of the Eigenfunctions and Eigenvalues Associated with the Classical and Circular Prolate Spheroidal Wave Functions », *in*: *Applied Mathematics and Computation* 218.*22* (July 2012) (referenced on page 244).

[28] A. Karoui and T. Moumni, « New Efficient Methods of Computing the Prolate Spheroidal Wave Functions and Their Corresponding Eigenvalues », *in*: *Applied and Computational Harmonic Analysis* 24.*3* (May 2008) (referenced on page 244).

[29] H. J. Landau and H. O. Pollak, « Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - II », *in*: *Bell System Technical Journal* 40.*1* (Jan. 1961) (referenced on pages 222, 225, 228).

[30] H. J. Landau and H. O. Pollak, « Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-III: The Dimension of the Space of Essentially Time- and Band-Limited Signals », *in*: *Bell System Technical Journal* 41.*4* (July 1962) (referenced on pages 222, 226, 228).

[31] L. Miranian, « Slepian Functions on the Sphere, Generalized Gaussian Quadrature Rule », *in*: *Inverse Problems* 20.*3* (June 2004) (referenced on page 233).

[32] T. Moumni and A. I. Zayed, « A Generalization of the Prolate Spheroidal Wave Functions with Applications to Sampling », *in*: *Integral Transforms and Special Functions* 25.*6* (June 2014) (referenced on pages 227, 244).

[33] C. Niven, « V. on the Conduction of Heat in Ellipsoids of Revolution », *in*: *Philosophical Transactions of the Royal Society of London* 171 (Dec. 1880) (referenced on pages 227, 242).

[34] R. E. O'Malley, « Book Reviews », *in*: *SIAM Review* 55.*3* (Jan. 2013) (referenced on page 234).

[35] A. Osipov, V. Rokhlin, and H. Xiao, *Prolate Spheroidal Wave Functions of Order Zero: Mathematical Tools for Bandlimited Approximation*, vol. 187, Applied Mathematical Sciences, Boston, MA: Springer US, 2013 (referenced on pages 227, 242, 243).

[36] A. Papoulis, « A New Algorithm in Spectral Analysis and Band-Limited Extrapolation », *in*: *IEEE Transactions on Circuits and Systems* 22.*9* (Sept. 1975) (referenced on page 231).

[37] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*, 1st ed., Cambridge University Press, June 1993 (referenced on pages 233, 236).

[38] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Rev. and enl. ed, New York: Academic Press, 1980 (referenced on pages 247, 249, 251).

[39] P. J. Roddy and J. D. McEwen, *Slepian Scale-Discretised Wavelets on Manifolds*, Feb. 2023, arXiv: 2302.06006 (referenced on page 232).

[40] P. J. Roddy and J. D. McEwen, « Slepian Scale-Discretised Wavelets on the Sphere », *in*: *IEEE Transactions on Signal Processing* 70 (2022), arXiv: 2106.02023 (referenced on page 232).

[41] W. Rudin, *Real and Complex Analysis*, 3rd ed, New York: McGraw-Hill, 1987 (referenced on page 222).

[42] Y. Shkolnisky, « Prolate Spheroidal Wave Functions on a Disc—Integration and Approximation of Two-Dimensional Bandlimited Functions », *in*: *Applied and Computational Harmonic Analysis* 22.*2* (Mar. 2007) (referenced on page 315).

[43] F. J. Simons, F. A. Dahlen, and M. A. Wieczorek, « Spatiospectral Concentration on a Sphere », *in*: *SIAM Review* 48.*3* (Jan. 2006) (referenced on page 232).

[44] F. J. Simons and D. V. Wang, « Spatiospectral Concentration in the Cartesian Plane », *in*: *GEM - International Journal on Geomathematics* 2.*1* (June 2011) (referenced on pages 229, 233, 315).

[45] D. Slepian, « Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty-V: The Discrete Case », *in*: *Bell System Technical Journal* 57.*5* (May 1978) (referenced on pages 222, 223, 229, 233, 256).

[46] D. Slepian and H. O. Pollak, « Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - I », *in*: *Bell System Technical Journal* 40.*1* (Jan. 1961) (referenced on pages 222, 225–228, 277, 290).

[47] D. Slepian, « Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - IV: Extensions to Many Dimensions; Generalized Prolate Spheroidal Functions », *in*: *Bell System Technical Journal* 43.*6* (Nov. 1964) (referenced on pages 222, 228).

[48] E. M. Stein, R. Shakarchi, and E. M. Stein, *Fourier Analysis: An Introduction*, 15. Druck, Princeton Lectures in Analysis 1, Princeton Oxford: Princeton University Press, 2003 (referenced on page 246).

[49] M. A. Taylor and B. A. Wingate, « A Generalization of Prolate Spheroidal Functions with More Uniform Resolution to the Triangle », *in*: *Journal of Engineering Mathematics* 56.*3* (Jan. 2007) (referenced on page 244).

[50] H. Wang, « Numerical Simulation for the Gross-Pitaevskii Equation Based on the Lattice Boltzmann Method », *in*: *Advances in Space Research* 60.*6* (Sept. 2017) (referenced on page 227).

[51] L. Wang, « A Review of Prolate Spheroidal Wave Functions from the Perspective of Spectral Methods », *in*: *Journal of Mathematical Study* 50.*2* (June 2017) (referenced on page 232).

[52] H. Widom, « Asymptotic Behavior of the Eigenvalues of Certain Integral Equations. II », *in*: *Archive for Rational Mechanics and Analysis* 17.*3* (Jan. 1964) (referenced on page 230).

Index

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS INFORMATIQUE, SIGNAL SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Méthodes particulaires modulées et ordres élevés.

Mots clés : Vlasov-Poisson, Schrödinger, Slepian, méthodes numériques, équations différentielles partielles

Résumé : Dans cette thèse trois grands axes ont été étudiés. Le premier concerne le système de Vlasov-Poisson, pour lequel la convergence d'une méthode particulaire a été démontrée. Cette méthode particulaire fait en quelque sorte le lien entre les méthodes semi-lagrangiennes et celle du type PIC. La simplicité de cette méthode réside dans le fait qu'elle se base sur des briques existantes bien connues.

Le second axe étudié traite de l'équation de Schrödinger. En se basant sur des travaux récents de Faou, Merle et Raphaël, un algorithme de modulation est proposé pour la simulation numérique de l'oscillateur harmonique. En utilisant le principe de Dirac-Frenkel, cet algorithme a pu être étendu au cas de l'équation de Schrödinger cubique non linéaire.

Enfin, le troisième et dernier axe de cette thèse parle du problème de concentration spectrale, aussi appelé problème de Slepian. Des outils ont été mis en place pour étendre les travaux de Landau, Pollak et Slepian, et des soucis importants d'ordre numérique ont été illustrés. Un algorithme est proposé afin de résoudre approximativement le problème de façon plus robuste qu'une discrétisation directe.

Title: Modulated particle methods and high orders.

Keywords: Vlasov-Poisson, Schrödinger, Slepian, numerical methods, partial differential equations

Abstract: This thesis is made of three distinct parts. The first one concerns the Vlasov-Poisson system, for which the convergence of a particle method has been shown. It makes a sort of link between semi-Lagrangian methods and those of PIC-type. The simplicity of this numerical method lies in the fact that it is composed of well-studied building blocks.

The second part is about the Schrödinger equation. Based on recent works by Faou, Merle and Raphaël, a modulation algorithm is proposed for the numerical simulation of the harmonic oscillator. By using the Dirac-Frenkel principle, this algorithm has been extended to the cubic nonlinear Schrödinger equation.

Last but not least, the third part of this thesis treats the spectral concentration problem, also known as the Slepian problem. We extended the framework studied by Landau, Pollak and Slepian, and illustrated important numerical issues. An algorithm is proposed in order to solve approximately the problem in a more robust way than a straightforward discretization.