Méthodes numériques hybrides cinétique/fluide et préservant la structure pour des équations cinétiques collisionnelles

Tino Laidin

Laboratoire Paul Painlevé

Université de Lille

Université de Lille

École doctorale **ED MADIS**

Unité de recherche **Laboratoire Paul Painlevé**

Thèse présentée par **Tino Laidin**

Soutenue le **27 septembre 2024**

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématiques et leurs interactions**

Méthodes numériques hybrides cinétique/fluide et préservant la structure pour des équations cinétiques collisionnelles

Thèse dirigée par	Thomas Rey	directeur
	Marianne Bessemoulin-Chatard	co-directrice

Composition du jury

Rapporteurs	Bruno Després	professeur au Laboratoire Jacques-Louis Lions	
	Giacomo Dimarco	professeur à l'University of Ferrara	
Examinateurs	Vincent Calvez	directeur de recherche au Laboratoire de Mathématiques de Bretagne Atlantique	
	Claire Chainais-Hillairet	professeure à l'Université de Lille	
	Nicolas Crouseilles	directeur de recherche au Centre Inria de l'université de Rennes	président du jury
	Hélène Hivert	chargée de recherche à l'Inria - Géosciences Rennes	
Directeurs de thèse	Thomas Rey	professeur à l'Université Côte d'Azur	
	Marianne Bessemoulin-Chatard	directrice de recherche à Nantes Université	

Laboratoire Paul Painlevé

Université de Lille

Université de Lille

Doctoral School **ED MADIS**

University Department **Laboratoire Paul Painlevé**

Thesis defended by **Tino Laidin**

Defended on **September 27, 2024**

In order to become Doctor from Université de Lille

Academic Field **Mathematics and their interactions**

Hybrid kinetic/fluid and structure preserving numerical methods for collisional kinetic equations

Thesis supervised by	Thomas Rey	Supervisor
	Marianne Bessemoulin-Chatard	Co-Supervisor

Committee members

Referees	Bruno Després	Professor at Laboratoire Jacques-Louis Lions	
	Giacomo Dimarco	Professor at University of Ferrara	
Examiners	Vincent Calvez	Senior Researcher at Laboratoire de Mathématiques de Bretagne Atlantique	
	Claire Chainais-Hillairet	Professor at Université de Lille	
	Nicolas Crouseilles	Senior Researcher at Centre Inria de l'université de Rennes	Committee President
	Hélène Hivert	Junior Researcher at Inria - Géosciences Rennes	
Supervisors	Thomas Rey	Professor at Université Côte d'Azur	
	Marianne Bessemoulin-Chatard	Senior Researcher at Nantes Université	

Mots clés : équations cinétiques, hypocoercivité, méthodes hybrides, méthodes spectrales
Keywords: kinetic equations, hypocoercivity, hybrid methods, spectral methods

Cette thèse a été préparée dans les laboratoires suivants.

Laboratoire Paul Painlevé
Université de Lille
Cité Scientifique - Bâtiment M2
F-59655 VILLENEUVE d'ASCQ
France

Site `https://math.univ-lille.fr/`

Centre Inria de l'Université de Lille
Parc scientifique de la Haute-Borne
40, avenue Halley - Bât. A
59650 Villeneuve d'Ascq
France

Site `https://www.inria.fr/fr/centre-inria-de-luniversite-de-lille`

Cette thèse a été financée par le LabEx CEMPI et la région Hauts-de-France.

À Claudine et Maurice,

Méthodes numériques hybrides cinétique/fluide et préservant la structure pour des équations cinétiques collisionnelles

Résumé

Cette thèse porte sur le développement et l'analyse de méthodes numériques performantes pour l'approximation des solutions d'équations cinétiques collisionnelles éventuellement non linéaires. Ces équations apparaissent dans divers domaines tels que la physique, notamment dans l'étude des semi-conducteurs et de la dynamique des gaz. Elles apparaissent aussi en biologie dans la modélisation du mouvement de cellules dans un tissu. Ces modèles présentent un aspect multi-échelle. D'une part, on a une description mésoscopique (ou cinétique) qui donne l'évolution de la fonction de distribution des particules, molécules ou cellules. D'autre part, par un processus de moyennisation, on obtient l'échelle dite macroscopique (ou fluide) qui permet de suivre l'évolution de grandeurs physiques observables : les moments de la fonction de distribution. Ces derniers correspondent notamment à la densité, la vitesse moyenne et la température des particules considérées. Nous présentons dans ce manuscrit différentes façons de tirer parti de la dynamique fluide afin de construire et étudier des méthodes numériques efficaces pour l'échelle cinétique.
Dans la première partie, nous explorons des méthodes de discrétisation visant à préserver la structure des équations continues. Nous commençons par introduire un schéma volumes finis implicite en temps pour un modèle cinétique de réaction non linéaire. Nous étudions le comportement en temps long de la solution discrète par une méthode d'hypocoercivité. Ensuite, nous examinons une méthode spectrale, basée sur des polynômes orthogonaux généraux, capable de préserver les moments de la solution, tout en assurant de bonnes propriétés de convergence.
La seconde partie est consacrée à la conception de méthodes numériques visant à réduire le coût des simulations cinétiques. Pour ce faire, nous étudions deux approches exploitant l'évolution des moments de l'inconnue. La première, une méthode dite hybride cinétique/fluide, consiste à adopter dynamiquement et localement en espace une description fluide moins coûteuse du système au lieu de la description cinétique plus onéreuse. La seconde approche repose également sur l'utilisation d'un modèle fluide, mais cette fois-ci pour accélérer les itérations temporelles de la méthode. Nous proposons ici un prototype de méthode pararéelle multi-échelle, utilisant un modèle fluide comme solveur grossier et un modèle cinétique comme solveur fin.

Mots clés : équations cinétiques, hypocoercivité, méthodes hybrides, méthodes spectrales

---

Hybrid kinetic/fluid and structure preserving numerical methods for collisional kinetic equations

Abstract

This thesis focuses on the development and analysis of efficient numerical methods for approximating solutions of potentially nonlinear kinetic collisional equations. These equations arise in various fields such as physics, notably in the study of semiconductors and gas dynamics. They also appear in biology in modelling the movement of cells within tissue. These models exhibit a multiscale aspect where there is, on one hand, a mesoscopic (or kinetic) description that gives the evolution of the distribution function of particles, molecules, or cells. On the other hand, through a process of averaging, we obtain the so-called macroscopic (or fluid) scale which allows to track the evolution of observable physical quantities: the moments of the distribution function. These moments correspond to the density, average velocity, and temperature of the considered particles. Throughout this manuscript, we present various ways to take advantage of fluid dynamics to construct and study efficient numerical methods for the kinetic scale.
In the first part, we explore discretization methods aiming to preserve the structure of continuous equations. We begin by introducing an implicit finite volume scheme for a nonlinear reaction kinetic model. We study the long-time behaviour of the discrete solution using hypocoercivity methods. Then, we examine a spectral method, based on general orthogonal polynomials, capable of preserving the moments of the solution while ensuring good convergence properties.
The second part is dedicated to the design of numerical methods aiming to reduce the cost of kinetic simulations. To do this, we study two approaches exploiting the evolution of the unknown's moments. The first, a hybrid kinetic/fluid method, involves adopting dynamically and locally in position a less costly fluid description of the system instead of the more expensive kinetic one. The second approach also relies on the use of a fluid model, but this time to accelerate the temporal iterations of the method. Here, we propose a prototype of a multiscale parareal method, using a fluid model as a coarse solver and a kinetic model as a fine solver.

Keywords: kinetic equations, hypocoercivity, hybrid methods, spectral methods

Remerciements

La tradition l'oblige, ce manuscrit débute par le difficile exercice des remerciements. Les quelques mots qui vont suivre sont destinés à exprimer ma gratitude envers toutes les personnes qui ont contribué, de près ou de loin, à ces trois années de thèse.

Je souhaite naturellement commencer par remercier mes directeur.ice.s de thèse, Marianne Bessemoulin-Chatard et Thomas Rey. Je vous suis très reconnaissant de m'avoir fait confiance dès mon stage de master il y a maintenant plus de quatre ans, et je ne saurais comment exprimer toute ma gratitude pour ce que vous m'avez apporté ces dernières années. Que ce soit dans la recherche ou sur un plan plus personnel, chacun m'a offert son expertise, son soutien et ses encouragements avec une réelle bienveillance. Même si nombre de nos discussions ont eu lieu à distance, vous avez toujours été disponibles et attentifs et vous avez été, je le pense, une source d'inspiration à de nombreuses reprises quand j'en avais le plus besoin.

Je tiens ensuite à remercier Bruno Després et Giacomo Dimarco qui m'ont fait l'honneur de rapporter ma thèse. Leurs observations et questionnements ont permis d'approfondir et d'améliorer le manuscrits. Je remercie aussi l'ensemble membres de mon jury de thèse, Vincent Calvez, Claire Chainais-Hillairet, Nicolas Crouseilles et Hélène Hivert d'avoir accepté de consacrer de leur temps pour être présents lors de ma soutenance.

J'ai eu la chance pendant cette thèse côtoyer à la fois le Laboratoire Paul Painlevé et le centre Inria de l'université de Lille. Cela m'a permis de rencontrer et d'interagir avec de nombreux chercheurs et ingénieurs qui tous ont rendu mon expérience de jeune doctorant nouvellement arrivé dans le nord des plus agréable. J'en profite aussi pour dire merci aux personnels administratifs et techniques du LPP et du centre Inria pour leur aide précieuse et leur efficacité. Leur travail dans l'ombre a été essentiel pour toutes ces petites choses du quotidien qui au final représentent beaucoup.

Il est maintenant temps de donner quelques noms et je m'excuse d'avance pour ceux que j'aurai pu oublier. Merci à l'équipe ANEDP du LPP et aux équipes RAPSODI et PARADYSE d'inria. L'ambiance chaleureuse qui règne dans ces équipes fait qu'on s'y sentait chez soi. Merci à Clément, Simon, Maxime, Juliette, Federica, André, Andréa, Guillaume, Alexandre et à tous les autres pour votre sympathie au quotidien. Je tiens aussi à dire un merci tout particulier à Claire pour ta sagesse et bienveillance autour du thé du midi qui, bien souvent, était trop chaud pour pour moi.

Je tiens aussi à remercier chaleureusement Lorenzo Pareschi pour m'avoir accueilli à Ferrare et pour les très stimulantes discussions et projets à venir.

Je tiens également à remercier chaleureusement les doctorants du LPP avec qui j'ai partagé le bureau 320 et de nombreux moments après les séminaires aux 3B : Nicolas, Vincent, Ivan, Thomas, Iacopo, Gabriel, et bien d'autres. Vos échanges et votre camaraderie ont rendu cette expérience encore plus enrichissante. Je n'oublie évidemment pas Justine, dont la gentillesse et la disponibilité ont été précieuses pour le jeune doctorant que j'étais.

Je tiens aussi à dire un grand merci aux précaires du Village pour ce bout d'aventure passé ensembles. Merci Julien, Jules, Maxime et Quentin je n'oublierai jamais toutes nos discussions plus ou moins sérieuses et surtout nos moments de grande compétitivité à base d'IPA, de coca zéro ou

encore de loco lada. Et sans oublier la dernière arrivée, merci Amélie pour toutes ces soirées de rigolade et de découvertes fivoises. Cette fois Bon Jovi c'est pour moi ;)

Je remercie également mes amis de toujours, Alban, Constant, Loïc, Noémie et Jess'. Même si on est tous aux quatre coins de la France, vous m'avez toujours permis de m'évader et je suis certain qu'on a encore beaucoup à faire ensembles !

Il me parait évident de remercier ma famille : Papa, Maman, Flavie, Jade et bien-sûr Isko. Votre soutien inconditionnel depuis toujours m'a aidé à perséver sur ce chemin assez unique. Même si ce n'était pas toujours très clair quand j'expliquais ce que je faisais, votre enthousiasme et votre intérêt sincère m'ont souvent remotivé lors de mes retours à la maison. Chaque moment passé avec vous m'a permis de recharger les batteries et de repartir plus en forme que jamais.

Enfin, je ne saurais terminer sans dire merci à celle qui m'accompagne au quotidien. Merci, Jeanne, de m'avoir rejoins à l'autre bout de la France pour cette aventure. Ton amour, ton soutien et le bonheur d'être à tes côtés signifient tellement pour moi que des mots ne suffisent pas, à part peut être simplement ... toujours toujours ♡

Sommaire

Avant-propos

Ce manuscrit présente les résultats obtenus au cours de ma thèse réalisée sous la direction de Marianne Bessemoulin-Chatard et Thomas Rey. Ce travail est organisé en deux parties correspondant à deux aspects de l'étude numérique des équations cinétiques collisionnelles.

Dans un premier temps, nous présentons des travaux portant sur l'analyse de méthodes capables de préserver la structure des équations continues sous-jacentes. Plus particulièrement, les propriétés étudiées sont le comportement en temps long et la conservation des moments des solutions discrètes.

Dans un second temps, nous nous intéressons à la construction de méthodes numériques permettant de réduire le coût de calcul des simulations cinétiques. Deux approches utilisant la dynamique portée par les moments de l'inconnue sont présentées. La première repose sur une réduction du coût de calcul dans le plan de phase et la seconde vise à réduire le coût de l'intégration temporelle.

Travaux réalisés pendant la thèse :

Chapitre 1 : *Discrete hypocoercivity for a nonlinear kinetic reaction model* ; en collaboration avec Marianne Bessemoulin-Chatard et Thomas Rey ; accepté pour publication dans *IMA Journal of Numerical Analysis* [22].
Chapitre 2 : *Conservative polynomial approximations and applications to Fokker-Planck equations* ; en collaboration avec Lorenzo Pareschi ; soumis pour publication [149].
Chapitre 3 :

- *Hybrid Kinetic/Fluid numerical method for the Vlasov-BGK equation in the diffusive scaling* ; paru [148] dans *Kinetic & Related Models*.
- *Hybrid Kinetic/Fluid numerical method for the Vlasov-Poisson-BGK equation in the diffusive scaling* ; en collaboration avec Thomas Rey ; paru [150] dans *FVCA 10 - 2023 - International Conference on Finite Volumes for Complex Applications X*.

Chapitre 4 : *Multiscale parareal algorithm for kinetic equations* ; travail en cours.

Introduction générale

Sommaire du présent chapitre

De nombreux problèmes contemporains reposent sur la compréhension de systèmes faisant interagir de façon complexe un grand nombre d'objets élémentaires. Les comportements qui en émergent sont souvent difficiles à prévoir et, selon le contexte, ces derniers peuvent décrire des phénomènes bien différents. Ils représentent aussi des défis modernes pour nos sociétés, dans des domaines aussi variés que la transition écologique, la propagation d'épidémies ou encore la gestion des ressources.

Dans le cadre des sciences physiques, on peut par exemple considérer les molécules composant un gaz et on s'intéressera alors à des écoulements d'air dans une pièce ou autour d'un obstacle. Un autre exemple, qui a motivé le début de cette thèse, est celui des particules chargées comme des électrons ou des ions. Dans ce cas, on cherchera plutôt à comprendre le comportement de composants électriques comme les semi-conducteurs ou encore la dynamique d'un plasma dans un tokamak. Un autre domaine qui se développe activement est celui de l'étude de populations au sein de laquelle s'échangent des savoirs, des opinions ou encore des ressources financières. Ici, l'objet élémentaire considéré n'est autre que l'individu qui, au cours du temps, va interagir dans un milieu socio-économique et dont la situation va évoluer en conséquence de ses interactions. Un dernier point de vue concerne l'étude mathématique de phénomènes biologiques. Ici, nous mentionnons par exemple le chimiotactisme, c'est-à-dire la description du déplacement de cellules dans un tissu sous l'effet de la chimie de leur environnement.

Bien que ces domaines puissent sembler éloignés, on peut passer par une étape de modélisation, c'est-à-dire une mise en équations des phénomènes observés, et trouver des structures mathématiques communes. Ainsi, en étudiant et développant des outils d'analyse pour, par exemple, comprendre la dynamique des gaz, il est alors possible de réutiliser ces outils dans le cadre des sciences sociales. Grâce à un cadre mathématique bien défini, on peut ensuite réaliser des simulations numériques précises et prédire l'évolution de nombreux systèmes complexes pour des applications réelles. Par la suite, nous utiliserons le terme physique "particules" pour faire référence aux objets élémentaires composant notre système d'étude.

Dans cette thèse, nous nous intéressons à des modèles mathématiques dits *cinétiques* qui permettent de décrire de nombreux systèmes de particules. Ces derniers donnent naissance à une grande variété de phénomènes ayant lieu à plusieurs échelles d'observations et nous nous concentrons sur la construction et l'analyse de leur simulation numérique.

Théorie cinétique

La théorie cinétique a débuté au XIX$^{\text{ème}}$ avec les travaux de Carnot [44] et Clausius [52] sur la réversibilité des systèmes microscopiques et plus particulièrement avec l'étude de la description de la dynamique des gaz à un niveau atomique par J.C. Maxwell [166] et L. Boltzmann [29]. Leur objectif a été de comprendre des phénomènes observables à une échelle dite *macroscopique* grâce à la mise en équations des interactions *microscopiques* entre les particules. Cette notion d'échelles de description va jouer un rôle central dans notre étude et nous allons maintenant introduire ces différents points de vue à l'aide de modèles mathématiques.

L'échelle microscopique

Une première manière de décrire un système de particules, et peut-être la plus intuitive, est de s'intéresser à l'évolution dans le temps de la position et de la vitesse de chacune de ces dernières. De cette façon, on peut obtenir une description exacte du système, dans le sens où l'on dispose de toute l'information nécessaire pour prédire son évolution. Cependant, comme nous le verrons plus tard, cette précision entraîne des difficultés du point de vue de l'analyse et de la simulation numérique du modèle. Avant d'introduire quelques notations, il est important de donner une idée du nombre de particules à considérer pour une représentation réaliste d'un système. Par exemple, un litre d'air dans les conditions de pression à la surface de la Terre contient environ 10^{22} molécules et cet ordre de grandeur reste vrai pour l'étude d'un plasma par exemple. Pour la suite, notons maintenant N ce nombre de particules interagissant dans notre système. La description microscopique revient à écrire des équations pour $X_i(t) \in \Omega_x \subset \mathbb{R}^3$ et $V_i(t) \in \mathbb{R}^3$ qui sont respectivement la position et la vitesse de la $i^{\text{ème}}$ particule à un temps $t \in [0,T]$ se déplaçant dans un domaine Ω_x. On peut alors utiliser les lois de la mécanique classique, et en particulier la seconde loi de Newton pour décrire les variations temporelles des $X_i(t)$ et $V_i(t)$:

$$\forall\, i = 1,\dots,N, \quad \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\,t} X_i(t) = V_i(t), \\ \dfrac{\mathrm{d}}{\mathrm{d}\,t} V_i(t) = F_i(t, X_i(t), V_i(t)). \end{cases} \tag{1}$$

Dans la seconde équation, le terme F_i permet de prendre en compte les éventuelles forces extérieures s'appliquant au système ainsi que les interactions entre particules. Dans ce dernier cas, l'exemple typique est celui d'interactions gravitationnelles ou électrostatiques au travers d'un potentiel $\phi(x)$, $x \in \mathbb{R}^3$, et on a alors,

$$F_i(t, X_i(t)) = -\sum_{i \neq j} \nabla\phi(X_i(t) - X_j(t)).$$

Le système microscopique (1) est finalement composé de $2N$ Équations Différentielles Ordinaires (EDO) couplées. Les inconnues ne dépendent ici que du temps t mais, comme mentionné précédemment, N est typiquement extrêmement grand. Ainsi, du point de vue de l'analyse mathématique, il est très compliqué d'en faire émerger des comportements qualitatifs. En particulier, ce système est en fait chaotique dans le sens où de petites perturbations dans le système peuvent avoir un impact significatif sur la dynamique globale. L'exemple le plus célèbre est celui du problème à N corps qui pour $N = 3$ présente déjà ce type de comportement [185]. De plus, d'un point de vue numérique, cette description est aussi problématique. En effet, le stockage en mémoire et le calcul des interactions entre N particules représentent un coût numérique très élevé, d'une complexité de l'ordre de $\mathcal{O}(N^2)$ pour des méthodes classiques, qui est le plus souvent prohibitif pour des applications réelles. C'est pourquoi, bien que fournissant une description complète du système, on préfère d'autres échelles de description.

L'échelle macroscopique

À l'opposé de la description microscopique, où l'on considère le système via ses particules discrètes, on peut plutôt le décrire par un milieu continu. Ce dernier va évoluer dans le temps au travers d'Équations aux Dérivées Partielles (EDP) sur des quantités macroscopiques observables. Typiquement, on s'intéresse à la densité, à la vitesse moyenne et à la température du fluide. La mise en équation à cette échelle remonte aux travaux d'Euler à la fin de XVIII$^{\text{ème}}$ siècle sur les fluides incompressibles. Dans le contexte de la dynamique des gaz, en dimension $d = 3$ et pour un gaz parfait monoatomique, un exemple est celui des *équations d'Euler* :

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u) + \nabla_x(\rho\theta) = 0, \\ \partial_t (\rho(\frac{1}{2}|u|^2 + \frac{3}{2}\theta)) + \nabla_x \cdot (\rho u(\frac{1}{2}|u|^2 + \frac{5}{2}\theta)) = 0. \end{cases} \tag{2}$$

Dans ce système d'EDP couplées, les quantités ρ, u et θ sont respectivement les observables mentionnés précédemment : la densité, la vitesse moyenne et la température du gaz. En particulier, le système (2) traduit des lois de conservation pour ces observables. Nous reviendrons sur cette notion un peu plus tard.

Un autre exemple de modèle macroscopique, dont nous étudierons une version simplifiée, est le modèle de *dérive-diffusion*. Ce dernier apparaît dans la modélisation mathématique du

comportement des paires électrons/trous dans un semi-conducteur :

$$\begin{cases} \partial_t \rho_e - \nabla_x \cdot (\nabla_x \rho_e + \rho_e \nabla_x \phi) = -R(\rho_e, \rho_t), \\ \partial_t \rho_t - \nabla_x \cdot (\nabla_x \rho_t - \rho_t \nabla_x \phi) = -R(\rho_e, \rho_t), \\ -\Delta \phi = \rho_t - \rho_e + C. \end{cases} \tag{3}$$

Ce système d'EDP décrit l'évolution des densités ρ_e et ρ_t des électrons et trous avec l'effet d'un potentiel électrique ϕ auto-consistant solution d'une équation de Poisson. Le terme R décrit quant à lui la recombinaison des deux espèces.

À cette échelle, on constate que le nombre d'équations pour décrire le système est fortement réduit comparé à une description microscopique. Cependant, l'espace des solutions est maintenant de dimension infinie. Les inconnues ne dépendent plus seulement du temps t, mais aussi de la position x dans l'espace. Malgré cela, l'étude qualitative de ce type de modèle est plus réalisable. De la même manière, les simulations numériques, bien que comportant certaines contraintes, sont aussi plus intéressantes du point de vue coût de calcul. Ces avantages font que ce type de modèles est déjà largement utilisé dans des applications réelles depuis de nombreuses années. Cependant, il est important de noter que cette échelle de description ne convient pas pour décrire certains systèmes, par exemple la haute atmosphère [158] ou bien le plasma généré dans un tokamak [51] où les interactions entre particules ne sont pas assez nombreuses pour satisfaire une description fluide. Trop d'informations venant de l' "individualité" des particules ont été perdues dans l'obtention du modèle macroscopique. Il est alors nécessaire d'utiliser une description plus précise, mais étant tout de même adaptée pour l'analyse et la simulation.

L'échelle mésoscopique

Nous avons vu que les échelles microscopiques et macroscopiques présentent chacune des avantages et des inconvénients, que ce soit pour l'analyse ou bien pour la simulation numérique. Un point de vue intermédiaire, l'échelle dite *mésoscopique* (ou cinétique), utilise une approche statistique pour décrire le système. L'idée est ici de considérer une description continue du système à la fois en espace, comme un modèle fluide, mais aussi en vitesse. Ainsi il est possible, d'une certaine manière, de conserver la notion d' "individualité" des particules tout en obtenant un modèle mathématique que l'on peut analyser et simuler relativement efficacement. Avec ce point de vue, on s'intéresse alors à l'évolution de la fonction de distribution des particules $f = f(t,x,v)$ qui représente le nombre de particules ayant une position $x \in \Omega_x$ et une vitesse $v \in \mathbb{R}^3$ au temps $t > 0$. Plus précisément, si l'on prend $\omega \subset \Omega_x$ et $\mathcal{V} \subset \mathbb{R}^3$ des sous-ensembles mesurables en espace et vitesse, le nombre de particules se trouvant dans ω et ayant des vitesses dans $\mathcal{V}$ au temps t est donné par

$$\iint_{\omega \times \mathcal{V}} f(t,x,v)\, \mathrm{d}x\, \mathrm{d}v.$$

Dans un cadre général, l'échelle mésoscopique peut être modélisée par une équation cinétique de la forme :

$$\partial_t f(t,x,v) + v \cdot \nabla_x f(t,x,v) - \nabla_x \phi(t,x) \cdot \nabla_v f(t,x,v) = \mathcal{Q}(f). \tag{4}$$

Sous cette forme, (4) porte le nom d'équation de Vlasov collisionnelle. Ici, on modélise des particules se déplaçant librement à leur vitesse v et qui sont soumises à deux types d'interactions. D'une part, celles dites à longue portée affectent la vitesse des particules via un potentiel ϕ qui peut lui-même dépendre de l'environnement extérieur et/ou des autres particules. D'autre part, celles dites à courte portée, le plus souvent appelées collisions, sont décrites par un opérateur $\mathcal{Q}$, qui affecte seulement la vitesse des particules.

Moments. Une première propriété de la description cinétique est que l'on peut retrouver les quantités macroscopiques à partir de la fonction de distribution. En effet, les observables tels que la densité, la vitesse moyenne et l'énergie sont obtenus en calculant les *moments en vitesse* de $f(t,x,v)$:

— Densité :
$$\rho(t,x) = \int_{\mathbb{R}^3} f(t,x,v)\,\mathrm{d}\,v;$$

— Vitesse moyenne et quantité de mouvement :
$$u(t,x) = \frac{1}{\rho}\int_{\mathbb{R}^3} v f(t,x,v)\,\mathrm{d}\,v, \quad (\rho u)(t,x) = \int_{\mathbb{R}^3} v f(t,x,v)\,\mathrm{d}\,v;$$

— Température :
$$\theta(t,x) = \frac{1}{3\rho}\int_{\mathbb{R}^3} |v-u|^2 f(t,x,v)\,\mathrm{d}\,v;$$

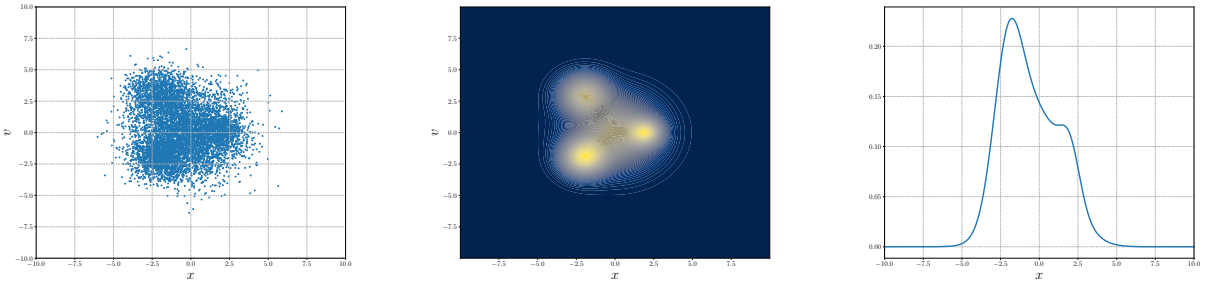
— Énergie cinétique locale :
$$\mathcal{E}(t,x) = \int_{\mathbb{R}^3} \frac{|v|^2}{2} f(t,x,v)\,\mathrm{d}\,v.$$

On peut ensuite intégrer ces moments en espace pour obtenir des quantités globales comme la masse $M(t)$ et l'énergie cinétique totale $E_c(t)$ du système :
$$M(t) = \int_{\Omega_x} \rho(t,x)\,\mathrm{d}\,x, \quad E_c(t) = \int_{\Omega_x} \mathcal{E}(t,x)\,\mathrm{d}\,x.$$

On introduit aussi l'entropie globale du système $H(t)$:
$$H(t) = \int_{\Omega_x\times\mathbb{R}^3} f(t,x,v)\log(f(t,x,v))\,\mathrm{d}\,x\,\mathrm{d}\,v.$$

Maxwelliennes. Un type de distributions particulièrement intéressant en théorie cinétique est celui des *Maxwelliennes*. Ces dernières sont des Gaussiennes déterminées par les quantités macroscopiques :
$$\mathcal{M}_{\rho,u,\theta} := \frac{\rho}{(2\pi\theta)^{d/2}}\exp\left(-\frac{|v-u|^2}{2\theta}\right).$$

Il découle de cette définition que les trois premiers moments d'une telle distribution sont exactement les quantités macroscopiques ρ, ρu et $\mathcal{E}$. En physique, lorsque la distribution des particules est donnée par une Maxwellienne, on dit que le système est à l'*équilibre thermodynamique* ou simplement à l'*équilibre local*.



Microscopique
$(X_i(t), V_i(t))_{i=1,\dots,N}$

Mésoscopique
$f(t,x,v)$

Macroscopique
$\rho(t,x)$, $u(t,x)$, $\theta(t,x)$

FIGURE 1 – Les différentes échelles de description.

Quelques propriétés des collisions. La nature des collisions dépend bien sûr du type d'interaction entre les particules du système. Dans le cas classique de la dynamique des gaz, ce qui revient à prendre $\nabla_x \phi = 0$ dans (4), l'opérateur $\mathcal{Q}$ vérifie plusieurs propriétés fondamentales. Premièrement, la masse, la quantité de mouvement ainsi que l'énergie du système sont conservées au cours du temps. On parle ici de *collisions élastiques*. Cela se traduit par la relation

$$\int_{\mathbb{R}^3} \mathcal{Q}(f) \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} \mathrm{d}\, v = 0_{\mathbb{R}^5}. \tag{5}$$

Une autre propriété fondamentale est la dissipation de l'entropie :

$$\int_{\mathbb{R}^3} \mathcal{Q}(f) \log(f)\, \mathrm{d}\, v \leq 0. \tag{6}$$

Cette relation est connue en physique sous le nom de théorème H, forme statistique de la seconde loi de la thermodynamique. Enfin, la dernière propriété impose la forme de la distribution annulant l'opérateur de collision. Dans notre contexte, ce sont les Maxwelliennes :

$$\mathcal{Q} = 0 \iff f = \mathcal{M}_{\rho,u,\theta}. \tag{7}$$

Plus précisément ces distributions appartiennent au noyau de l'opérateur de collision. L'exemple le plus connu satisfaisant ces hypothèses et reposant sur des interactions binaires entre particules est l'*opérateur de Boltzmann* :

$$\mathcal{Q}_{\mathrm{Boltz}}(f,f) = \iint_{\mathbb{R}^3 \times \mathbb{S}^2} \left[f(v') f(v'_*) - f(v) f(v_*) \right] B(v - v_*, \sigma)\, \mathrm{d}\, v_*\, \mathrm{d}\, \sigma. \tag{8}$$

Ici, v' et $v'_* \in \mathbb{R}^3$ sont les vitesses post-collisionnelles des particules et v, $v_* \in \mathbb{R}^3$ sont les vitesses pré-collisionnelles. Les collisions élastiques sont alors données par les relations

$$\begin{cases} v + v_* = v' + v'_*, \\ |v|^2 + |v_*|^2 = |v'|^2 + |v'_*|^2. \end{cases} \tag{9}$$

Ce système de 4 équations à 6 inconnues est sous-résolu, mais on peut écrire les solutions sous la forme dite *σ-représentation*, pour $\sigma \in \mathbb{S}^2$:

$$\begin{cases} v' = \dfrac{v+v_*}{2} + \sigma \dfrac{|v-v_*|}{2}, \\ v'_* = \dfrac{v+v_*}{2} - \sigma \dfrac{|v-v_*|}{2}. \end{cases} \tag{10}$$

L'opérateur de Boltzmann (8) est très précis dans sa description des interactions entre particules dans le sens où ces dernières sont décrites exactement. Cependant, d'un point de vue numérique, son approximation par des méthodes classiques à vitesses discrètes peut être extrêmement coûteuse [36, 167, 175]. Une large littérature de méthodes performantes dites *spectrales* s'est développée pour réduire ce coût [26, 169, 170, 176], mais, en pratique, une simplification de l'opérateur de Boltzmann est souvent utilisée : l'opérateur *BGK* [23, 184],

$$\mathcal{Q}_{\mathrm{BGK}}(f) = \frac{1}{\tau}\Big(\mathcal{M}_{\rho,u,\theta} - f\Big), \tag{11}$$

où τ est un temps caractéristique entre deux collisions. Cette description est bien plus simple à approcher, mais elle ne permet pas d'obtenir un modèle macroscopique avec le bon nombre de Prandtl, qui correspond au rapport entre la convection et la diffusion dans le système. Il est cependant possible de corriger ce défaut en utilisant l'opérateur ES-BGK [5, 128].

Il est intéressant de noter que dans d'autres contextes, en modifiant la façon dont les particules interagissent au niveau binaire, il est possible de modéliser d'autres phénomènes. Par exemple, on peut décrire les collisions inélastiques dans un gaz granulaire [35] ou encore les échanges d'opinions ou de richesses dans une population [93]. Les modèles que l'on obtient alors ne présentent pas forcément les propriétés mentionnées précédemment, en particulier celles portant sur les conservations. Il arrive que la modélisation mathématique amène seulement la masse à être conservée. C'est le cas pour l'opérateur de relaxation linéaire qui modélise des collisions avec un milieu ambiant statique :

$$\mathcal{Q}_{\mathrm{Relax}}(f) = \mathcal{M}_{\rho,0,1} - f, \tag{12}$$

ou encore des opérateurs de type Fokker-Planck apparaissant par exemple en modélisation socio-économique [93]

$$\mathcal{Q}_{\mathrm{Opinion}}(f) = \frac{\lambda}{2}\partial_{vv}\Big((1-v^2)f\Big) + \partial_v((v-m)f), \quad v \in \mathbb{R}.$$

De plus, la notion de Maxwellienne ne convient plus pour certaines applications, notamment en physique quantique avec l'étude des fermions où la distribution d'intérêt sera plutôt de type *Fermi-Dirac* :

$$F(t,k,x) = \Big[1 + \exp\Big(-\lambda_0(t,x) + |k|^2/2\Big)\Big]^{-1},$$

où $k \in \mathbb{R}^3$ correspond aux niveaux d'énergie des fermions. La fonction λ_0 est alors choisie de telle sorte que $\int_{\mathbb{R}^3} F \,\mathrm{d}\, k = \rho$. Nous renvoyons à [137] pour une plus complète description des modèles pour les semi-conducteurs.

Lien entre les différentes échelles

Étant donné que les échelles de description permettent chacune de décrire des systèmes de particules, il semble naturel de se poser la question de l'existence d'un lien entre ces dernières et de sa démonstration rigoureuse. Ce problème, fortement lié au sixième problème de Hilbert sur l'axiomatisation de la physique, n'est, encore à ce jour, pas totalement résolu.

De microscopique à mésoscopique

La dérivation rigoureuse des équations cinétiques à partir d'un système à N particules peut être obtenue par plusieurs techniques, mais l'ingrédient fondamental consiste à considérer la limite $N \to +\infty$.

Dans un premier temps, nous allons présenter l'obtention d'un modèle cinétique pour un système de particules qui n'interagissent pas entre elles, mais qui sont soumises à une force extérieure. Ces hypothèses reviennent à considérer le système (1) avec $F_i = F$ pour tout $i \in [\![1;N]\!]$ et on peut alors rigoureusement relier les échelles microscopique et mésoscopique. Dans ce but, on introduit la *distribution empirique* des particules

$$\mu_N(t) = \sum_{i=1}^{N} \alpha_i \delta_{(X_i(t),V_i(t))},$$

où pour $(x,v) \in \Omega_x \times \mathbb{R}^3$, $\delta_{x,v}$ est la distribution de Dirac en (x,v) et $\alpha_i \in \mathbb{R}$. Étant donné que μ_N est un élément de $\mathcal{D}'((0,T)\times\Omega_x\times\mathbb{R}^3)$, on a pour toute fonction test $\varphi \in \mathcal{D}((0,T)\times\Omega_x\times\mathbb{R}^3)$, à support compact dans $(0,T)$:

$$\langle \mu_N, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \sum_{i=1}^{N} \alpha_i \int_0^T \varphi(t, X_i(t), V_i(t))\,\mathrm{d}\,t,$$

où $\langle\cdot,\cdot\rangle_{\mathcal{D}',\mathcal{D}}$ désigne le crochet de dualité entre $\mathcal{D}'$ et $\mathcal{D}$. Supposons maintenant que $(X_i(t), V_i(t))$, $i \in [\![1;N]\!]$ et soit solution de (1) avec $F_i = F$. On a alors

$$\begin{aligned}
\langle \partial_t \mu_N, \varphi\rangle_{\mathcal{D}',\mathcal{D}} &= -\langle \mu_N, \partial_t \varphi\rangle_{\mathcal{D}',\mathcal{D}} \\
&= -\sum_{i=1}^{N} \alpha_i \int_0^T \partial_t \varphi(t, X_i(t), V_i(t))\,\mathrm{d}\,t \\
&= -\sum_{i=1}^{N} \alpha_i \int_0^T \Big(\frac{\mathrm{d}}{\mathrm{d}\,t}\varphi(t, X_i(t), V_i(t)) - X_i'(t)\cdot\nabla_x \varphi(t, X_i(t), V_i(t)) \\
&\qquad\qquad - V_i'(t)\cdot\nabla_v \varphi(t, X_i(t), V_i(t))\Big)\,\mathrm{d}\,t \\
&= \sum_{i=1}^{N} \alpha_i \int_0^T \Big(V_i(t)\cdot\nabla_x \varphi(t, X_i(t), V_i(t)) \\
&\qquad\qquad + F(t, X_i(t), V_i(t))\cdot\nabla_v \varphi(t, X_i(t), V_i(t))\Big)\,\mathrm{d}\,t \\
&= \langle \mu_N, v\cdot\nabla_x\varphi\rangle_{\mathcal{D}',\mathcal{D}} + \langle \mu_N, F\cdot\nabla_v\varphi\rangle_{\mathcal{D}',\mathcal{D}} \\
&= -\langle \nabla_x\cdot(v\mu_N), \varphi\rangle_{\mathcal{D}',\mathcal{D}} - \langle \nabla_v\cdot(F\mu_N), \varphi\rangle_{\mathcal{D}',\mathcal{D}}
\end{aligned}$$

soit encore,

$$\langle \partial_t \mu_N + \nabla_x \cdot (v\mu_N) + \nabla_v \cdot (F\mu_N), \varphi \rangle_{\mathcal{D}',\mathcal{D}} = 0.$$

Nous avons ainsi montré que μ_N est solution dans $\mathcal{D}'((0,T)\times\Omega_x\times\mathbb{R}^3)$ de l'équation de Vlasov linéaire

$$\partial_t \mu_N + v\cdot\nabla_x \mu_N + \nabla_v \cdot (F\mu_N) = 0.$$

De Newton vers des modèles non-linéaires

Dans le calcul précédent, les particules n'interagissaient pas entre elles et le modèle cinétique obtenu était alors linéaire. Cependant, les interactions sont dans de nombreuses applications au cœur des phénomènes étudiés. Dans ce cas, l'obtention rigoureuse d'un modèle cinétique est potentiellement très ardue et les équations obtenues sont non-linéaires.

Historiquement, deux approches ont été considérées pour obtenir des modèles mésoscopiques. La première repose sur le point de vue de la dynamique stochastique des particules [138]. La seconde, initiée dans [115] et basée sur la hiérarchie BBGKY [28, 30, 138, 139, 206], adopte quant à elle une description déterministe du système.

Afin de montrer ce type de résultat, on considère la distribution jointe des N particules, supposées indistinguables, du système $F_N(t,x_1,\dots,x_N,v_1,\dots,v_N)$ qui sera alors solution d'une équation de Liouville. Ensuite, en intégrant successivement pour les $N-1$ premières particules, on se ramène à étudier l'équation satisfaite par la marginale d'une unique particule :

$$F_N^{(1)}(t,x,v) = \int_{\Omega_x\times\mathbb{R}^3} F_N(t,x_1,\dots,x_N,v_1,\dots,v_N)\,\mathrm{d}\,x_2\dots\mathrm{d}\,x_N\,\mathrm{d}\,v_2\dots\mathrm{d}\,v_N.$$

Cette quantité est pertinente, car on a supposé toutes les particules indistinguables. Le passage à la limite $N\to+\infty$ nécessite un grand nombre d'hypothèses. Celle que nous souhaitons mentionner et qui sert en fait à fermer le système final est la *propagation du chaos*. L'idée est de supposer qu'à la limite, les particules ne sont pas corrélées :

$$F_N^{(2)} \underset{N\to+\infty}{\sim} F\otimes F, \tag{13}$$

où $F_N^{(2)}$ n'est autre que la marginale de deux particules.

On peut alors distinguer deux types de modèles :

— Le premier, les équations *cinétiques collisionnelles*, modélise les interactions à courte portée. Le point central est ici le caractère binaire des collisions [27, 45, 94, 151, 189].

— Le second type, à l'inverse, porte sur les effets à longue distance qu'ont les particules entre elles. On parle ici de modèles de *champ moyen* [104, 106, 133, 134].

Il est important de noter qu'en pratique, des modèles combinant collisions et champ moyen sont souvent utilisés. Par exemple, au c œur d'un tokamak, le plasma est essentiellement non collisionnel. Cependant, il en est tout autre aux bords du plasma où les collisions, notamment avec des impuretés, jouent un rôle critique.

De mésoscopique à macroscopique

Afin de mieux comprendre le lien entre les descriptions cinétiques et fluides, on introduit le nombre de Knudsen ε qui quantifie la fréquence des collisions entre les particules. Il correspond au ratio entre le libre parcours moyen, c'est-à-dire la distance moyenne parcourue entre deux collisions, et la longueur caractéristique du système. Dans le cas où ce paramètre est grand, $\varepsilon \sim 1$, le système est peu collisionnel et la description cinétique prévaut sur celle fluide. À l'inverse, un nombre de Knudsen petit correspond à un grand nombre de collisions et les modèles macroscopiques sont alors plus adaptés. On peut maintenant écrire l'équation de Vlasov collisionnelle mise à l'échelle :

$$\partial_t f^\varepsilon(t,x,v) + \frac{1}{\varepsilon^\alpha}\Big(v\cdot\nabla_x f^\varepsilon(t,x,v) - \nabla_x \phi(t,x)\cdot\nabla_v f^\varepsilon(t,x,v)\Big) = \frac{1}{\varepsilon^{\alpha+1}}\mathcal{Q}(f^\varepsilon). \tag{14}$$

Ici, le choix du paramètre α permet de considérer le système sous différentes échelles temporelles. Le cas $\alpha = 0$ correspond à la mise à l'échelle dite *hydrodynamique*. En prenant $\alpha = 1$, on parle cette fois d'échelle *diffusive* qui correspond à regarder notre système pour des temps très grands et grandes variations spatiales.

Nous avons vu précédemment que les distributions Maxwelliennes annulent l'opérateur de collision $\mathcal{Q}$. Il est alors assez naturel de considérer les perturbations autour de ces dernières. Il s'agit en fait de l'idée du développement de Chapman-Enskog [49] de la fonction de distribution :

$$f^\varepsilon = \mathcal{M}_{\rho^\varepsilon,u^\varepsilon,\theta^\varepsilon} + \varepsilon g^{(1)} + \varepsilon^2 g^{(2)} + \dots, \tag{15}$$

où les fonctions $g^{(k)}(t,x,v)$, $k = 1,2,\dots$ représentent les perturbations d'ordre ε^k par rapport à la Maxwellienne $\mathcal{M}_{\rho^\epsilon,u^\epsilon,\theta^\epsilon}$ et ne dépendent de f^ε qu'à travers ses trois premiers moments.

Avant de traiter le cas plus complexe de l'EDP (14), il est intéressant de considérer le système jouet d'EDO couplées suivant :

$$\begin{cases} x' = -\dfrac{1}{\varepsilon}x + y, \\ y' = \varepsilon y. \end{cases} \tag{16}$$

Par analogie avec (14), x jouerait le rôle de la direction orthogonale au noyau de l'opérateur de collision $\mathcal{Q}$ (7) et y serait la direction de la projection sur $\mathrm{Ker}(\mathcal{Q})$, aussi appelé variété d'équilibre, car contenant les distributions éponymes. Le paramètre ε serait alors l'équivalent du nombre de Knudsen. En notant $U(t) = (x(t), y(t))^T$, la solution de (16) est donnée par

$$U(t) = e^{A_\varepsilon t}U(0), \quad \text{avec} \quad A_\varepsilon = \begin{pmatrix} \frac{-1}{\varepsilon} & 1 \\ 0 & \varepsilon \end{pmatrix},$$

et on a

$$e^{A_\varepsilon t} = \begin{pmatrix} e^{-t/\varepsilon} & \frac{\varepsilon(e^{\varepsilon t}-1)}{\varepsilon^2+1} \\ 0 & e^{\varepsilon t} \end{pmatrix}.$$

Ainsi, en prenant la limite $\varepsilon \to 0$, on obtient

$$e^{A_\varepsilon t} \underset{\varepsilon\to 0}{\longrightarrow} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

On observe alors qu'à la limite, la solution du système ne dépend plus que de la variable y. De manière analogue, dans le cadre de l'équation cinétique (14), on peut effectivement montrer que la dynamique des solutions ne dépendra plus que de la variable d'espace à la limite $\varepsilon \to 0$. Nous allons maintenant décrire l'obtention des limites *hydrodynamique* et *diffusive* de (14).

Équations d'Euler

La dérivation rigoureuse des équations de la mécanique des fluides (Euler, Navier-Stokes) a historiquement été obtenue au travers d'une formulation continue des conservations de la masse, de la quantité de mouvement et de l'énergie. Mais alors, il est naturel d'espérer pouvoir retrouver ces lois à partir de l'équation de Boltzmann, elle-même obtenue à partir des lois de Newton. De nombreux travaux [11, 12, 95, 105, 107, 108, 109, 110, 160, 190] ont permis de grandes avancées en ce sens et ce domaine reste à ce jour très actif. Nous proposons maintenant une preuve formelle de la dérivation des équations d'Euler avec force à partir de (14) sous la mise à l'échelle hydrodynamique.

Une première observation est que formellement, lorsque $\varepsilon \to 0$, $f^\varepsilon \to f$ et

$$\mathcal{Q}(f) = 0,$$

ce qui, d'après l'hypothèse (7) implique que $f = \mathcal{M}_{\rho,u,\theta}$. Supposons maintenant que la distribution limite f soit donnée par une Maxwellienne (ce qui revient à tronquer le développement de Chapman-Enskog à l'ordre 0). On a alors par définition que

$$\int_{\mathbb{R}^3} f \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} \mathrm{d}\, v = \int_{\mathbb{R}^3} \mathcal{M}_{\rho,u,\theta} \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} \mathrm{d}\, v = \begin{pmatrix} \rho \\ \rho u \\ \mathcal{E} \end{pmatrix}.$$

De plus, en supposant que l'opérateur $\mathcal{Q}$ vérifie les conservations (5), et en intégrant (14) contre $\left(1, v, \frac{|v|^2}{2}\right)$, on obtient alors le système macroscopique :

$$\begin{cases} \partial_t \displaystyle\int_{R^3} f \,\mathrm{d}\, v + \nabla_x \cdot \int_{\mathbb{R}^3} v f \,\mathrm{d}\, v + \nabla_x \phi \cdot \int_{R^3} \nabla_v f \,\mathrm{d}\, v = 0, \\ \partial_t \displaystyle\int_{R^3} v f \,\mathrm{d}\, v + \nabla_x \cdot \int_{\mathbb{R}^3} v \otimes v f \,\mathrm{d}\, v + \nabla_x \phi \cdot \int_{R^3} \nabla_v(v) f \,\mathrm{d}\, v = 0, \\ \partial_t \displaystyle\int_{R^3} \frac{|v|^2}{2} f \,\mathrm{d}\, v + \nabla_x \cdot \int_{\mathbb{R}^3} \frac{|v|^2}{2} v f \,\mathrm{d}\, v + \nabla_x \phi \cdot \int_{R^3} \nabla_v \left(\frac{|v|^2}{2}\right) f \,\mathrm{d}\, v = 0, \end{cases}$$

où "⊗" désigne le produit tensoriel. Un rapide calcul montre alors que,

$$\int_{\mathbb{R}^3} v \otimes v f \,\mathrm{d}\, v = \mathbb{P} + \rho u \otimes u,$$

où $\mathbb{P} = \int_{\mathbb{R}^3} (v-u) \otimes (v-u) f \,\mathrm{d}\, v$ est appelé *tenseur de pression*. De plus, on a aussi

$$\int_{\mathbb{R}^3} \frac{|v|^2}{2} v f \,\mathrm{d}\, v = q + (\mathbb{P} + \mathcal{E}\mathrm{Id})u,$$

où $q = \int_{\mathbb{R}^3} \frac{1}{2}(v-u)|v-u|^2 f \,\mathrm{d}\, v$ est ce qu'on appelle le *flux de chaleur* et Id désigne la matrice identité.

Dans notre cas particulier où f est une Maxwellienne on a que $\mathbb{P} = \rho\theta$ et par imparité en la variable $(v-u)$ de l'intégrande de q, on obtient que ce dernier s'annule. Plus précisément, à l'équilibre thermodynamique, on obtient finalement un système macroscopique fermé dit d'*Euler avec force*

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0 \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u) + \nabla_x(\rho\theta) + \rho\nabla_x\phi = 0 \\ \partial_t \mathcal{E} + \nabla_x \cdot (\rho u(\mathcal{E}+\theta)) + \rho u \cdot \nabla_x\phi = 0. \end{cases}$$

Ici, le potentiel ϕ peut être donné ou bien être auto-consistant en étant solution de l'équation de Poisson $-\Delta\phi = \rho$ par exemple. Il est important de noter que ce système n'est pas le seul modèle macroscopique pouvant être obtenu à partir de (14) en régime hydrodynamique. En particulier, en considérant $f^\varepsilon = \mathcal{M}_{\rho^\varepsilon,u^\varepsilon,\theta^\varepsilon} + \varepsilon g^{(1)}$, c'est-à-dire en tronquant (15) à l'ordre 1, on peut obtenir les équations de Navier-Stokes qui peuvent être vues comme une correction des équations d'Euler.

Équation de dérive-diffusion

L'étude de (14) en régime diffusif a d'abord été considérée dans [17] et le développement de la fonction de distribution f^ε a ensuite été justifié dans [10] pour le transport de neutrons puis dans [187] pour l'équation de Boltzmann linéaire. Dans [62], le résultat a pu être étendu à une large classe d'opérateurs de collision. Enfin, dans [114], l'utilisation d'une approximation diffusive de l'équation cinétique a été justifiée via des méthodes d'homogénéisation.

Considérons maintenant le modèle simplifié de particules soumises à un champ extérieur donné $E(x)$ et dont les collisions préservent la masse. C'est le cas de l'opérateur de relaxation

$$\mathcal{Q}(f) = \mathcal{M}_{\rho,0,1} - f. \tag{17}$$

La fonction de distribution des particules est alors solution de

$$\begin{cases} \partial_t f^\varepsilon + \dfrac{v}{\varepsilon} \cdot \nabla_x f^\varepsilon + \dfrac{E}{\varepsilon} \cdot \nabla_v f^\varepsilon = \dfrac{1}{\varepsilon^2}\mathcal{Q}(f^\varepsilon), \\ f^\varepsilon(0,x,v) = f_0(x,v). \end{cases} \tag{18}$$

On considère maintenant le développement de Chapman-Enskog (15) de f^ε autour de la Maxwellienne de vitesse moyenne nulle et de température 1 : $\mathcal{M}_{\rho,0,1}$. Afin d'obtenir un modèle de type dérive-diffusion, on tronque à l'ordre $K = 1$ si bien que

$$f^\varepsilon = \mathcal{M}_{\rho,0,1} + \varepsilon g^{(1)}. \tag{19}$$

En injectant (19) dans (18), on peut identifier selon les puissances de ε et obtenir la relation $g^{(1)} = -(v \cdot \nabla_x + E \cdot \nabla_v)(\rho^\varepsilon \mathcal{M})$ soit encore

$$g^{(1)} = -v\mathcal{M} \cdot J^\varepsilon \quad \text{avec} \quad J^\varepsilon = \nabla_x \rho^\varepsilon - E\rho^\varepsilon.$$

En intégrant ensuite (14) en vitesse, on obtient :

$$\partial_t \rho^\varepsilon + \frac{1}{\varepsilon}\mathrm{div}_x\langle v f^\varepsilon \rangle = 0.$$

La distribution est alors remplacée par la troncature (19) et

$$\partial_t \rho^\varepsilon + \frac{1}{\varepsilon}\mathrm{div}_x\left(\rho^\varepsilon\langle v\mathcal{M}\rangle + \varepsilon\langle v g^{(1)}\rangle\right) = 0.$$

Enfin, comme ρ^ε ne dépend pas de la vitesse et que les moments impairs de la Maxwellienne sont nuls, nous obtenons en utilisant l'expression de $g^{(1)}$:

$$\partial_t \rho^\varepsilon - \langle v \otimes v\mathcal{M}\rangle : \nabla_x J^\varepsilon = 0,$$

où ":" est la contraction tensorielle d'ordre 2. Le tenseur de diffusion est alors donné par $\langle v \otimes v\mathcal{M}\rangle = m_2\mathrm{Id}$. Enfin, en supposant que $\rho^\varepsilon \to \rho$ lorsque $\varepsilon \to 0$, nous obtenons formellement le modèle de dérive-diffusion :

$$\partial_t \rho - m_2\mathrm{div}_x J = 0, \text{ où } J = \nabla_x\rho - E\rho. \tag{20}$$

Une extension de cette procédure dans le cadre des semi-conducteurs consiste à s'intéresser à la dynamique non seulement de la densité ρ, mais aussi de l'énergie $\mathcal{E}$. On parle alors de modèle de *transport d'énergie*. Cependant, la dérivation nécessite une prise en compte plus fine des interactions à courte portée des électrons avec leur environnement. Plus précisément elles se décomposent en trois types de collisions : élastiques, électron-électron et inélastiques. Nous renvoyons à [14, 15, 64, 137] pour une discussion plus approfondie sur le sujet et notamment sur le choix de la mise à l'échelle dans ce contexte.

Aperçu des travaux de la thèse

L'échelle mésoscopique représente un outil puissant pour décrire des systèmes de particules en interaction. Cependant, il est le plus souvent impossible d'obtenir des solutions exactes pour ce type d'équation. Ainsi la simulation numérique, c'est-à-dire la construction de méthodes d'approximation des solutions de ces équations, apparaît comme un outil indispensable pour des applications réelles. Cette thèse porte plus particulièrement sur le développement et l'analyse de méthodes numériques performantes pour l'approximation des solutions d'équations cinétiques collisionnelles éventuellement non-linéaires.

Nous présentons maintenant différentes méthodes permettant de tirer parti de la dynamique fluide afin de construire et étudier des méthodes numériques efficaces pour l'échelle cinétique.

Méthode d'hypocoercivité discrète

Un premier aspect de cette thèse porte sur la construction et l'analyse de méthodes numériques capables de préserver la structure de l'équation continue sous-jacente. Plus particulièrement, on s'intéresse ici à assurer le bon comportement en temps long des solutions discrètes. Pour les modèles cinétiques qui nous intéressent, on peut en effet observer que les solutions convergent à une vitesse bien définie vers un état d'équilibre. Pour montrer ce type de résultat on s'appuie

notamment sur des techniques d'*hypocoercivité* [204]. Afin de donner une idée de cette notion, commençons par considérer le problème jouet suivant

$$\begin{cases} x' = -y, \\ y' = x - y. \end{cases} \tag{21}$$

Un calcul direct sur l'évolution en temps de la quantité $x^2 + y^2$ donne la relation

$$\frac{\mathrm{d}}{\mathrm{d}t}(x^2 + y^2) = -2y^2, \tag{22}$$

qui n'est en fait pas suffisante pour conclure à la décroissance exponentielle de $x(t)$ et $y(t)$ vers 0 lorsque $t \to +\infty$. En effet, ce modèle présente à première vue de la *dissipation* uniquement dans la seconde variable. Il est cependant possible d'utiliser le couplage entre x et y pour obtenir de la dissipation dans la première. On introduit dans ce but une quantité que l'on nommera *entropie modifiée* :

$$\mathcal{H}(t) = x^2 + y^2 - \delta xy, \tag{23}$$

où δ est un petit paramètre à déterminer. On peut maintenant calculer

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(t) = (\delta - 2)y^2 - \delta x^2 + \delta xy.$$

En choisissant δ suffisamment petit, on peut alors déterminer une constante κ de telle sorte que

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(t) \leq -\kappa\,\mathcal{H}(t).$$

Enfin, en remarquant que $\mathcal{H}(t)$ est équivalent à $x^2 + y^2$, on peut conclure sur la convergence exponentielle de la solution (x, y) de (21) vers 0 :

$$x^2 + y^2 \lesssim e^{-\kappa t}.$$

De nombreux systèmes physiques présentent une dynamique de convergence vers un état stationnaire. Plus particulièrement, démontrer et quantifier la vitesse de convergence d'un système perturbé vers un état d'équilibre représente un domaine de recherche très actif. Considérons maintenant une équation cinétique abstraite de la forme

$$\begin{cases} \partial_t f + \mathsf{T} f = \mathsf{L} f, \\ f(0, x, v) = f_I(x, v), \end{cases} \tag{24}$$

où T est un opérateur de transport, typiquement $\mathsf{T} = v \cdot \nabla_x$ et L est un opérateur de collision linéaire. On peut penser au simple opérateur de relaxation (12). Dans ce cadre, on s'attend à observer un retour vers l'équilibre de la solution de (24). Cependant, à cause de l'action uniquement en vitesse des collisions L, ce résultat n'est pas immédiat. Comme nous l'avons montré au travers du problème jouet (21) une première observation est qu'une certaine quantité, ici la norme de la solution, ne décroît que selon l'une des variables du système, ici la vitesse v. On parle alors de *coercivité microscopique* et cela ne permet pas de conclure directement sur le type de convergence de

la solution vers un état stationnaire. Pour remédier à ce problème, on introduit une fonctionnelle bien adaptée qui non seulement sera décroissante le long des solutions de (24), mais qui pourra aussi être contrôlée par la norme des solutions dans un espace de Hilbert bien choisi $\mathcal{X}$. On peut alors trouver des constantes explicites $\kappa > 0$ et $C \geq 1$ telles que

$$\|f(t) - f_\infty\|_{\mathcal{X}} \leq C \|f_I - f_\infty\|_{\mathcal{X}} e^{-\kappa t}.$$

On appelle ce type de comportement *hypocoercif* [204] en référence à la propriété plus classique de *coercivité* qui correspondrait au cas $C = 1$. Une première version partielle de ce type de résultats a été obtenue dans [17] et par la suite, de nombreux opérateurs de collisions ont pu être traités notamment dans [120, 121]. Les techniques utilisées remontent aux travaux sur *l'hypoellipticité* des opérateurs linéaires [129] et plus récemment une technique très robuste, celle que nous avons esquissée dans notre exemple jouet, a été introduite dans [73]. L'idée est d'introduire une *entropie modifiée* équivalente à la norme L^2 sur $\mathcal{X}$ et de montrer la convergence exponentielle de la solution vers un équilibre global. Dans un cadre linéaire tel que (24) l'entropie modifiée n'est autre que la norme $\|\cdot\|_{\mathcal{X}}$ à laquelle on ajoute un terme qui va permettre de tirer parti du mélange induit par l'opérateur de transport T dans le plan de phase (x, v). Le point crucial est donc de bien choisir ce terme additionnel. Une manière empirique de l'obtenir est de regarder l'évolution des moments de l'inconnue f. On peut alors obtenir la propriété dite de *coercivité macroscopique*, c'est-à-dire de la dissipation aussi dans la variable x, grâce à l'évolution de la densité ρ dont l'intégrale en x correspond à la quantité conservée par ce système. Elle correspond, dans ce contexte, à la projection sur le noyau de L et la coercivité microscopique quant à elle n'est autre que de la dissipation dans la direction orthogonale à ce noyau.

Grâce à la robustesse de la méthode introduite dans [73], de nombreux modèles linéaires ont été étudiés dans un cadre continu. Dans le cadre de l'équation de *Fokker-Planck* une classification des types de convergences est notamment présentée dans [34]. L'équation de Fokker-Planck *fractionnaire* a été étudiée dans [32] et dans ce cas les équilibres ne sont plus des Maxwelliennes (4), mais plutôt des fonctions à *queues lourdes* qui décroissent polynomialement vers 0 quand $v \to +\infty$. On mentionne aussi [41] où un modèle de chimiotactisme de type *run and tumble* a été considéré. Un raffinement de la technique a récemment été proposé dans [66] motivé par la théorie du contrôle. Si l'on s'intéresse maintenant à des modèles présentant des non-linéarités, un premier résultat pour le modèle de Vlasov-Poisson-Fokker-Planck avec de petites non-linéarités a été obtenu dans [122] puis raffiné dans [123]. Enfin, la convergence exponentielle a été obtenue uniformément en la limite diffusive dans [2]. Cette liste n'est bien sûr pas exhaustive, mais permet tout de même de dresser un panorama des divers champs d'applications.

Lorsque l'on s'intéresse à démontrer des résultats de convergence exponentielle vers un équilibre pour un schéma numérique, les techniques maintenant classiques reposent sur des structures d'entropie discrète [19, 38, 46, 112]. La discrétisation de l'état d'équilibre peut aussi se révéler cruciale dans l'analyse [86, 179] et la question des conditions au bord a également été considérée dans [47, 84].

Sur la discrétisation des équations cinétiques plus précisément, la question d'*hypo-coercivité discrète* se pose alors naturellement. Dans ce sens, les premiers résultats portent sur un modèle cinétique simplifié : l'équation de Kolmogorov. À la fois des discrétisations de type différences

finies et éléments finis ont été utilisées [103, 186] et le type de schéma obtenu a récemment permis de stabiliser une méthode éléments finis [74]. Pour une discrétisation de type différences finies, une technique d'hypocoercivité H^1 [204] a été adaptée pour l'équation de Fokker-Planck [76]. À la différence de la technique dans L^2 [73], il faut dans ce cas en plus considérer l'estimation de gradients qui, au niveau discret, peut se révéler compliquée notamment par l'absence d'un équivalent discret de la propriété de dérivation d'un produit ou d'une fonction composée. La méthode L^2 a ensuite été étendue au cas discret dans [18] pour une discrétisation de type volumes finis des équations de Fokker-Planck et BGK linéaire. Un résultat similaire a aussi été obtenu pour l'équation de Fokker-Planck fractionnaire dans [7]. Enfin, l'équation de Vlasov-Fokker-Planck a été considérée pour un champ extérieur donné et pour le modèle linéarisé de Vlasov-Poisson-Fokker-Planck dans [24, 25] avec une discrétisation de type Hermite en vitesse et volumes finis en espace.

Contributions

Dans ce contexte, le Chapitre 1 de cette thèse porte sur un modèle de relaxation cinétique non-linéaire unidimensionnel décrivant une réaction de génération-recombinaison entre deux espèces introduit dans [171] :

$$\begin{cases} \partial_t \mathfrak{f} + v\,\partial_x \mathfrak{f} = \chi_1 - \rho_{\mathfrak{g}}\mathfrak{f}, \\ \partial_t \mathfrak{g} + v\,\partial_x \mathfrak{g} = \chi_2 - \rho_{\mathfrak{f}}\mathfrak{g}, \end{cases} \tag{25}$$

avec $\rho_{\mathfrak{f}} = \int_{\mathbb{R}} \mathfrak{f}\,\mathrm{d}v$ et $\rho_{\mathfrak{g}} = \int_{\mathbb{R}} \mathfrak{g}\,\mathrm{d}v$. Les fonctions $\mathfrak{f}(t,x,v)$ et $\mathfrak{g}(t,x,v)$, $(t,x,v) \in \mathbb{R}^+ \times \mathbb{T} \times \mathbb{R}$, sont les densités dans le plan de phase des composants d'une réaction chimique réversible $A + B \leftrightarrow C$ où la substance C est supposée en excès et qui est donc de quantité supposée constante. En particulier, cette hypothèse de modélisation entraîne que la quantité conservée par le système est la différence des masses de $\mathfrak{f}$ et $\mathfrak{g}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{T}\times\mathbb{R}} (\mathfrak{f} - \mathfrak{g})\,\mathrm{d}v\,\mathrm{d}x = 0.$$

Ce modèle peut aussi être vu comme une version simplifiée d'un système de génération et recombinaison électrons/trous dans un semi-conducteur [60]. L'étude du comportement en temps long des solutions de (25) a été conduite dans [171] et la convergence exponentielle vers un état d'équilibre

$$F_\infty = (\mathfrak{f}_\infty, \mathfrak{g}_\infty) = (\rho_\infty \chi_1, \rho_\infty^{-1} \chi_2)$$

a été obtenue pour des données initiales proches de ce dernier. Ici, la constante $\rho_\infty > 0$ dépend de la donnée initiale au travers de la relation :

$$\int_{\mathbb{T}\times\mathbb{R}} (\mathfrak{f}_I - \mathfrak{g}_I)\,\mathrm{d}v\,\mathrm{d}x = |\mathbb{T}|\left(\rho_\infty - \frac{1}{\rho_\infty}\right).$$

Nous proposons maintenant de donner un aperçu du résultat d'hypocoercivité non-linéaire discret qui fait l'objet du Chapitre 1. Ce résultat est, à notre connaissance, le premier résultat de ce type.

Dans le même esprit que [18], on choisit d'adopter une discrétisation de type volumes finis de (25). Le plan de phase (x,v) est divisé en volumes de contrôle K_{ij} et la donnée initiale du problème

$\mathfrak{F}_I = (\mathfrak{f}_I, \mathfrak{g}_I)$ est alors approchée par

$$\mathfrak{f}_{ij}^0 = \frac{1}{\Delta x \Delta v} \int_{K_{ij}} \mathfrak{f}_I(x,v)\,\mathrm{d}x\,\mathrm{d}v, \quad \mathfrak{g}_{ij}^0 = \frac{1}{\Delta x \Delta v} \int_{K_{ij}} \mathfrak{g}_I(x,v)\,\mathrm{d}x\,\mathrm{d}v \qquad \forall (i,j) \in \mathcal{I} \times \mathcal{J}.$$

En intégrant (25) sur chaque volume de contrôle K_{ij} et en utilisant une méthode d'Euler implicite en temps, on obtient alors le schéma :

$$\begin{cases} \dfrac{\mathfrak{f}_{ij}^{n+1} - \mathfrak{f}_{ij}^n}{\Delta t} + \dfrac{1}{\Delta x \Delta v}\left(\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1}\right) = \chi_{1,j} - \rho_{\mathfrak{g},i}^{n+1} \mathfrak{f}_{ij}^{n+1}, \\ \dfrac{\mathfrak{g}_{ij}^{n+1} - \mathfrak{g}_{ij}^n}{\Delta t} + \dfrac{1}{\Delta x \Delta v}\left(\mathcal{G}_{i+\frac{1}{2},j}^{n+1} - \mathcal{G}_{i-\frac{1}{2},j}^{n+1}\right) = \chi_{2,j} - \rho_{\mathfrak{f},i}^{n+1} \mathfrak{g}_{ij}^{n+1}, \end{cases} \tag{26}$$

où $\mathcal{F}$ et $\mathcal{G}$ sont des flux numériques à deux points monotones. La nouveauté de ce travail est triple. Dans un premier temps, il s'agit de montrer la convergence exponentielle vers 0 des perturbations discrètes $F = (f, g)$, solutions d'un schéma pour le modèle linéarisé de (25). Pour ce faire, on adopte une version discrète de la méthode L^2 [73] qui a notamment été utilisée dans [18]. Dans notre contexte, étant donné que la quantité conservée par le système est la masse de $h := f - g$, on définit une entropie modifiée basée sur l'évolution de ses moments :

$$H_\delta^\Delta[F^n] := \frac{1}{2}\|F^n\|_\Delta^2 + \delta \langle J_h^n, D_x^c \Phi^n \rangle_2 + \frac{\delta}{2\Delta t} \sum_{i \in \mathcal{I}} \Delta x \left((D_x^c \Phi^n)_i - (D_x^c \Phi^{n-1})_i \right)^2,$$

où J_h est le flux associé à h, $J_h = \int_{\mathbb{R}} v h \,\mathrm{d}v$ en continu, et Φ est solution d'une équation de Poisson discrète. Cette fonctionnelle discrète comporte trois termes. Le premier est celui qui, une fois dérivé en temps, procure la dissipation dans la direction v : la coercivité microscopique. À l'opposé, le second terme va permettre d'obtenir la dissipation dans la variable x grâce à l'évolution des moments de h : la coercivité macroscopique. Enfin, le dernier terme est purement numérique et a pour but de compenser des termes venant de la discrétisation temporelle. En établissant alors l'équivalence entre H_δ^Δ et la norme $\|\cdot\|_\Delta$ et en choisissant δ suffisamment petit, on peut alors obtenir un premier résultat sur la décroissance des perturbations :

Théorème 1. *Sous des hypothèses de sommabilité des profils en vitesse discrets χ_1 et χ_2 et en supposant que le nombre de cellules en x est impair, il existe des constantes $C \geq 1$ et $\kappa > 0$ de telle sorte que pour tout $\Delta t \leq \Delta t_{\max}$ et toute donnée initiale $F^0 = (f_{ij}^0, g_{ij}^0)_{(i,j)\in\mathcal{I}\times\mathcal{J}}$ telle que $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x \Delta v\,(f_{ij}^0 - g_{ij}^0) = 0$, la solution $F^n = (f_{ij}^n, g_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$ du schéma linéarisé vérifie pour tout $n \geq 0$*

$$\|F^{n+1}\|_\Delta \leq C \|F^0\|_\Delta e^{-\kappa t^n}.$$

Les constantes C et κ ne dépendent pas de la taille de la discrétisation et sont explicites. Le choix de $\Delta t_{\max}$ est arbitraire.

Ensuite, on montre que le schéma non-linéaire vérifie un principe du maximum. Plus particulièrement, on montre que des solutions initialement comprises entre deux profils d'équilibre restent entre ces derniers pour tous temps. Cette étape repose notamment sur la monotonie des flux $\mathcal{F}$ et $\mathcal{G}$ pour borner le transport par des arguments classiques (voir par exemple [203, Chapitre 13, Section 5]). La subtilité du résultat repose plutôt sur l'obtention de bornes pour la partie réaction.

Ces dernières ont été obtenues en montrant qu'une version tronquée du schéma propage les bornes de la solution. On montre alors le résultat :

Théorème 2. *Il existe des constantes positives $\gamma_1 < \rho_\infty^*$ et γ_2 de telle sorte que si la donnée initiale $\mathsf{F}^0 = (\mathfrak{f}_{ij}^0, \mathfrak{g}_{ij}^0)_{i\in\mathcal{I}, j\in\mathcal{J}}$ satisfait pour tous $i \in \mathcal{I}$, $j \in \mathcal{J}$*

$$(\rho_\infty^* - \gamma_1)\,\chi_{1,j} \le \mathfrak{f}_{ij}^0 \le (\rho_\infty^* + \gamma_2)\,\chi_{1,j},$$
$$(\rho_\infty^* + \gamma_2)^{-1}\,\chi_{2,j} \le \mathfrak{g}_{ij}^0 \le (\rho_\infty^* - \gamma_1)^{-1}\,\chi_{2,j},$$

alors le schéma non-linéaire (26) admet une solution $(\mathfrak{f}_{ij}^n, \mathfrak{g}_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}, n\ge 0}$ telle que pour tous $i \in \mathcal{I}$, $j \in \mathcal{J}$, $n \ge 0$,

$$(\rho_\infty^* - \gamma_1)\,\chi_{1,j} \le \mathfrak{f}_{ij}^n \le (\rho_\infty^* + \gamma_2)\,\chi_{1,j}, \tag{27}$$
$$(\rho_\infty^* + \gamma_2)^{-1}\,\chi_{2,j} \le \mathfrak{g}_{ij}^n \le (\rho_\infty^* - \gamma_1)^{-1}\,\chi_{2,j}. \tag{28}$$

Enfin, la dernière étape pour obtenir le résultat d'hypocoercivité discrète local, c'est-à-dire en considérant des données initiales proches de l'équilibre, consiste à montrer que pour de telles données initiales, la convergence obtenue pour le modèle linéarisé suffit à contrebalancer les effets non-linéaires. On a alors le théorème suivant :

Théorème 3. *Sous les hypothèses des théorèmes précédents, une solution $\mathsf{F}^n = (\mathfrak{f}_{ij}^n, \mathfrak{g}_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$ du schéma non-linéaire (26) satisfait pour tout $n \ge 0$*

$$\|\mathsf{F}^n - \mathsf{F}^\infty\|_\Delta \le C\,\|\mathsf{F}^0 - \mathsf{F}^\infty\|_\Delta\, e^{-\kappa t^n}. \tag{29}$$

Les constantes C et κ ne dépendent pas de la taille de la discrétisation et sont explicites.

Remarques sur l'implémentation. Le code correspondant à ce travail a été réalisé en Python et un soin particulier a été apporté à la résolution du système non linéaire associé à (26) par la méthode de Newton. Plus particulièrement, un pas de temps adaptatif a été mis en œuvre pour aider la convergence de l'algorithme dans les premières itérations temporelles. Le code est entièrement modulable et robuste face aux choix de conditions initiales et de profils en vitesse χ_1 et χ_2. La Figure 2 illustre la convergence en temps long vers l'équilibre pour différents types de flux numériques :

— Lax-Friedrichs : $\mathcal{F}_{i+\frac{1}{2},j}^{n+1} = \Delta v\,\frac{v_j}{2}(\mathfrak{f}_{i+1,j}^{n+1} + \mathfrak{f}_{ij}^{n+1}) - \Delta v\,\lambda(\mathfrak{f}_{i+1,j}^{n+1} - \mathfrak{f}_{ij}^{n+1})$, $\lambda = \Delta x/2\,\Delta t$;
— Upwind : $\mathcal{F}_{i+\frac{1}{2},j}^{n+1} = \Delta v\left(v_j^+\,\mathfrak{f}_{i,j}^{n+1} - v_j^-\,\mathfrak{f}_{i+1,j}^{n+1}\right)$;
— Centré : $\mathcal{F}_{i+\frac{1}{2},j}^{n+1} = \Delta v\,\frac{v_j}{2}(\mathfrak{f}_{i+1,j}^{n+1} + \mathfrak{f}_{ij}^{n+1})$.

Approximation spectrale conservative

Une autre propriété structurelle des équations cinétiques est la conservation d'un certain nombre de moments de l'inconnue. Comme nous avons commencé à le voir dans les sections précédentes, ces derniers représentent des quantités observables et leur intérêt est donc tout particulier [21, 67, 147, 183], notamment dans l'étude du comportement en temps long de la

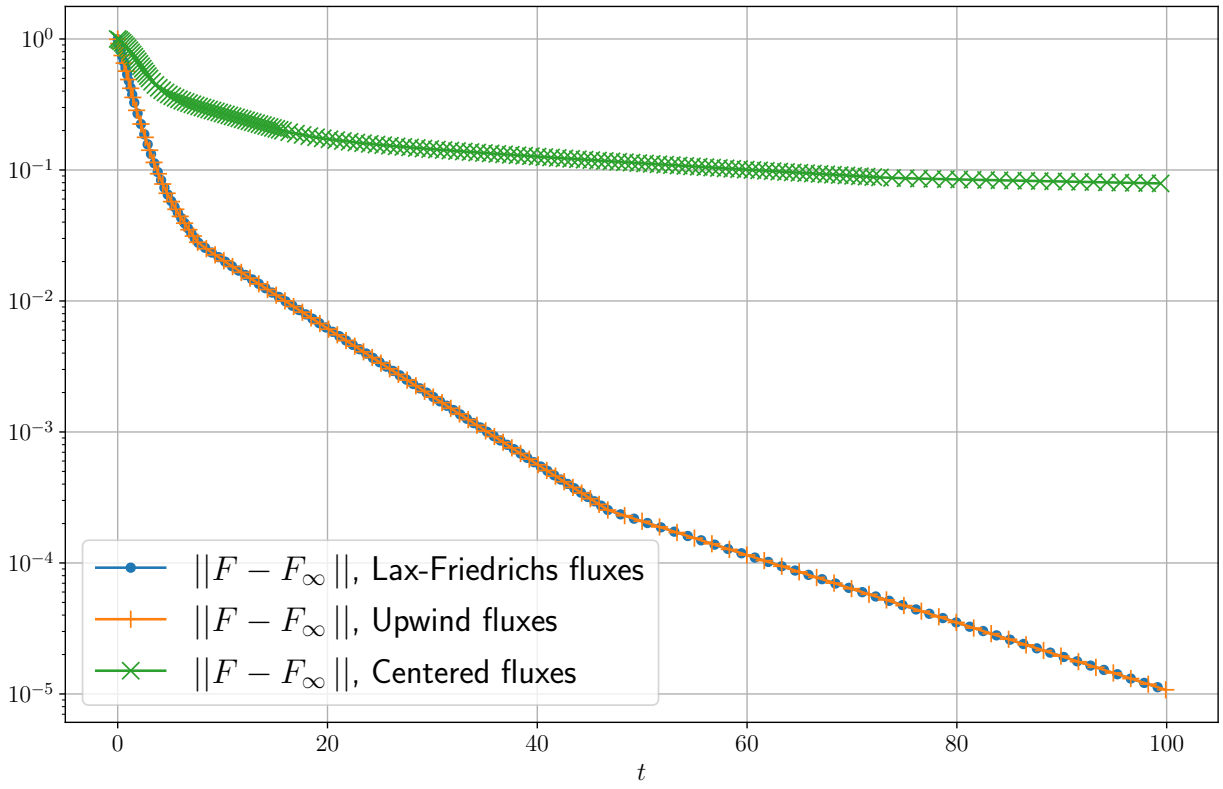


FIGURE 2 – Convergence de F vers F_∞ dans une norme L^2 à poids.

solution. Il est alors naturel de vouloir construire et analyser des méthodes numériques à la fois très précises, mais aussi capables de préserver ces moments. Par ailleurs, il est important de noter que les discrétisations capables de préserver la structure des équations sous-jacentes représentent un enjeu crucial. En effet, dans un grand nombre d'applications, les méthodes analytiques peuvent ne pas encore être accessibles pour étudier certains phénomènes complexes [111, 119, 135].

La littérature sur les méthodes préservant la structure est assez développée, et ce, pour des propriétés assez variées allant de la positivité au comportement en temps long des solutions. Si l'on s'intéresse plus particulièrement à la conservation des moments dans le cadre de l'équation de Fokker-Planck cinétique, plusieurs approches ont visé à préserver les moments dans le but de mieux décrire l'état stationnaire du système [8, 37, 48, 152, 168, 183].

Parmi les nombreuses méthodes de discrétisation, celles dites *spectrales*, qui reposent sur les *polynômes orthogonaux*, se sont révélées particulièrement efficaces pour résoudre des équations cinétiques collisionnelles comme l'équation de Boltzmann. En effet, elles peuvent se montrer extrêmement précises comparées à d'autres méthodes au coût numérique similaire. Cet avantage permet notamment de capturer avec précision des dynamiques loin de l'état d'équilibre [132, 169, 170, 176, 181, 182]. Dans le contexte de l'équation de Vlasov sans collisions, très utilisée pour modéliser le plasma dans un tokamak, les méthodes spectrales de type Hermite en vitesse exhibent aussi de très bonnes propriétés algébriques aidant à la préservation de la structure de l'équation [20, 21, 24, 50, 92, 127, 146, 153, 164, 193]. On renvoie ainsi à l'article de revue sur les méthodes numériques pour les plasmas [72] et aux références qu'il contient.

Malgré l'attrait que représentent les méthodes spectrales, l'un de leurs défauts est qu'elles

ne permettent pas de conserver les moments de l'inconnue. Ainsi, le comportement en temps long se retrouve fortement dégradé par l'accumulation d'erreurs machine qui contrebalancent la dynamique du système. Une solution à ce problème repose sur des techniques de minimisation L^2. Ces approches ont récemment fait leurs preuves dans plusieurs travaux [4, 96, 97, 177].

Contributions

L'idée d'une méthode spectrale repose sur la projection de l'inconnue dans un espace de polynômes orthogonaux pour un certain produit scalaire. Considérons ainsi l'approximation de la fonction $f(x) \in L^2_\omega(\Omega)$ où $\Omega \subseteq \mathbb{R}$ et $\omega(x) > 0$ est une fonction positive que l'on appellera *poids*. Pour une base polynomiale (ϕ_k), $k \in \mathbb{N}$ de $L^2_\omega(\Omega)$, on peut alors décomposer f de façon unique comme

$$f(x) = \sum_{k=0}^{\infty} \hat{f}_k \phi_k(x). \tag{30}$$

Si l'on note $\langle f, g \rangle_\omega = \int_{\mathbb{R}} f g \omega \, \mathrm{d}x$ le produit scalaire entre deux fonctions f et g de $L^2_\omega(\Omega)$, dire que les polynômes ϕ_k sont orthogonaux revient à la relation $\langle \phi_h, \phi_k \rangle_\omega = 0$ pour $h \neq k$. En particulier, les moments de f sont simplement donnés par

$$m_q = \langle f, x^q \rangle = \int_{\mathbb{R}} f x^q \, \mathrm{d}x.$$

Le choix du poids ω permet de considérer différentes familles de polynômes qui sont en particulier définis sur des domaines Ω bornés ou non, offrant ainsi de puissants outils dans un grand nombre de situations. On renvoie à la Table 2.1 du Chapitre 2 pour des exemples de polynômes orthogonaux.

Nous avons mentionné précédemment la précision des méthodes spectrales. Étant donné qu'il n'est pas possible numériquement de considérer le développement infini de f dans la base des ϕ_k, il est nécessaire de tronquer (30) à un ordre noté N. On note alors f_N cette troncature qui n'est en fait rien d'autre que la projection orthogonale de f sur l'espace S_N engendré par les N premiers polynômes ϕ_k. Dans le cas des polynômes de type Jacobi ($\omega(x) = (1-x)^\alpha (1+x)^\beta$, $\alpha, \beta > -1$), nous avons par exemple le résultat suivant [91] :

Théorème 4. *Si $f \in H^r_\omega(\Omega)$, où $r \geq 0$ est un entier, alors il existe une constante $C > 0$ qui dépend de α, β et r telle que*

$$\|f - f_N\|_{L^2_\omega} \leq \frac{C}{N^r} \left\| (1-x^2)^{r/2} \frac{d^r f}{dx^r} \right\|_{L^2_\omega}, \qquad N > r. \tag{31}$$

Une observation clé ici est que la précision de la méthode dépend directement de la régularité de la fonction à approcher. Plus elle est lisse (r grand), plus la convergence dans le nombre de modes N sera rapide. On parle alors de convergence *spectrale*.

L'objectif de ce travail, présenté dans le Chapitre 2, a été de développer dans un cadre général des méthodes spectrales préservant les moments. Dans ce but, on étend l'approche introduite dans [177] pour des polynômes trigonométriques à des familles générales, basées sur un poids ω. L'observation cruciale est, qu'en général, les moments ne sont pas préservés par la projection. Plus précisément, on a

$$m_q - m_{q,N} = \langle f - f_N, x^q \rangle = \left\langle f - f_N, \frac{x^q}{\omega(x)} \right\rangle_\omega \tag{32}$$

N	$\lvert m_0 - m_{0,N}\rvert$	$\lvert m_1 - m_{1,N}\rvert$	$\lvert m_2 - m_{2,N}\rvert$	$\lvert m_3 - m_{3,N}\rvert$
8	1.746e-03	1.404e-02	1.958e-02	1.538e-02
16	1.813e-07	6.126e-09	1.858e-07	6.264e-09
32	1.388e-17	5.551e-17	2.776e-17	5.551e-17

Tableau 1 – Erreur sur les 4 premiers moments pour une approximation de type Chebyshev de première espèce pour $N = 8, 16$ et 32 modes.

qui ne vaut zéro que si $x^q/\omega(x)$ appartient à l'espace S_N. Le Tableau 1 illustre ce phénomène pour l'approximation de la fonction $f(x) = \sin(2\pi x) + x^2\cos(2\pi x)$ à l'aide de polynômes de Chebyshev de première espèce ($\omega(x) = (1-x^2)^{-1/2}$).

En remarquant que la projection orthogonale sur S_N se reformule comme un problème de minimisation L^2, il semble alors naturel de contraindre cette minimisation afin de préserver les moments de la solution. On cherche alors à résoudre

$$f_N^c = \operatorname{argmin}\left\{\|g_N - f\|_{L^2_\omega} \;:\; g_N \in S_N, \langle g_N, x^q\rangle = \langle f, x^q\rangle,\ q = 0, 1, \dots, M\right\}, \tag{33}$$

où M est le nombre de moments que l'on cherche à préserver. La solution contrainte f_N^c préserve maintenant les moments par définition, mais la propriété de convergence spectrale de l'approximation n'est a priori plus garantie. Heureusement, le problème (33) peut être résolu explicitement par la méthode des multiplicateurs de Lagrange. La solution s'écrit alors sous la forme

$$f_N^c(x) = \sum_{k=0}^{N} \hat{f}_k^c \phi_k(x).$$

Les coefficients contraints $\hat{f}_k^c$ sont quant à eux donnés par les coefficients de l'approximation classique auxquels on ajoute un terme de correction non-local en le nombre de modes :

$$\hat{f}_k^c = \hat{f}_k + \hat{C}_k^T(\langle f, \Phi\rangle - \langle f_N, \Phi\rangle), \qquad \hat{C}_k^T = \frac{1}{\|\phi_k\|^2_{L^2_\omega}}\hat{\Phi}_k^T\left(\sum_{h=0}^{N}\frac{1}{\|\phi_h\|^2_{L^2_\omega}}\hat{\Phi}_h\hat{\Phi}_h^T\right)^{-1}, \tag{34}$$

où $\Phi = (1, x, x^2, \dots, x^M)^T$ et $\hat{\Phi}_k = (\mu_{0,k}, \mu_{1,k}, \dots, \mu_{M,k})^T$. En supposant que la fonction f admette des moments dans une norme appropriée, il est alors possible de montrer que l'approximation contrainte conserve bien la propriété de convergence spectrale :

Théorème 5. *Si $f \in H^r_\omega(\Omega)$, où $r \geq 0$ est un entier, on a*

$$\|f - f_N^c\|_{L^2_\omega} \leq \frac{C_\Phi}{N^r}\|f\|_{H^r_\omega} \tag{35}$$

où la constante C_Φ dépend de la norme du terme de correction dans (34).

L'approximation contrainte montre de très bons résultats quand elle est appliquée à des modèles de type Fokker-Planck issus de la physique, des sciences sociales et des sciences économiques [93]. Cette variété de cas tests a permis de montrer la robustesse de la méthode lorsqu'elle est combinée à des techniques d'imposition de conditions aux bords (non triviales pour les méthodes spectrales) et

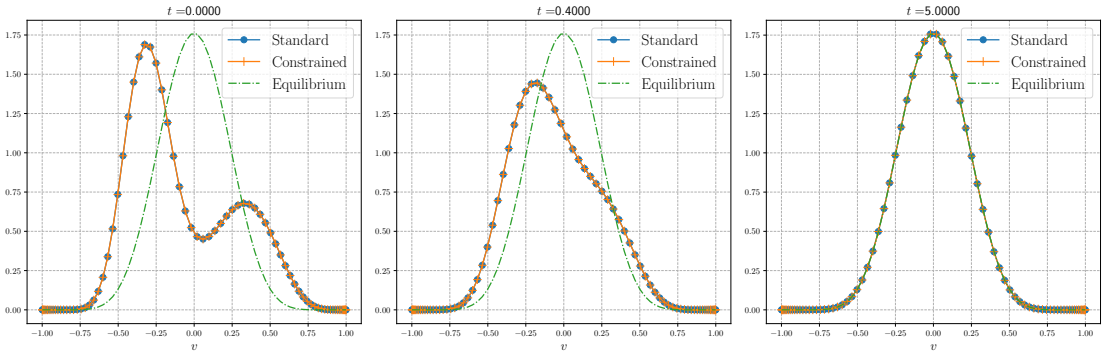


FIGURE 3 – Formation d'un consensus dans un modèle d'opinion.

à une approche micro-macro. La Figure 3 illustre la formation d'un consensus dans une population avec imposition de conditions de Dirichlet homogènes où l'on compare une approche classique et notre nouvelle méthode contrainte.

Méthodes numériques hybrides

On s'intéresse maintenant à la simulation d'une équation cinétique collisionnelle en régime diffusif de type (14) :

$$\begin{cases} \partial_t f^\varepsilon + \dfrac{v}{\varepsilon} \cdot \nabla_x f^\varepsilon + \dfrac{E}{\varepsilon} \cdot \nabla_v f^\varepsilon = \dfrac{1}{\varepsilon^2}\left(\mathcal{M}_{\rho,0,1} - f\right), \\ f^\varepsilon(0,x,v) = f_0(x,v). \end{cases} \tag{36}$$

Nous avons déjà rappelé qu'à la limite $\varepsilon \to 0$, les solutions de ce modèle étaient de la forme $f(t,x,v) = \rho(t,x)\mathcal{M}_{\rho,0,1}(v)$ où ρ est solution de

$$\partial_t \rho - \mathrm{div}_x J = 0, \text{ avec } J = \nabla_x \rho - E\rho. \tag{37}$$

Lorsque l'on cherche à approcher les solutions d'une équation cinétique dépendant d'un petit paramètre, le coût d'une simulation vient de deux facteurs. Le premier est issu de la condition de stabilité du schéma numérique utilisé. En effet, formellement, lorsque le nombre de Knudsen ε tend vers 0, la vitesse du transport tend vers l'infini. Numériquement, cela impose pour une méthode naïve de devoir choisir un pas de temps Δt extrêmement petit. La conséquence est alors l'augmentation significative du nombre d'itérations temporelles nécessaires pour atteindre un temps final donné. Une solution est d'utiliser des méthodes préservant l'asymptotique (dites *Asymptotic Preserving*, AP) [71, 135, 136, 142]. Ces schémas reposent sur trois caractéristiques :

1. La condition de stabilité du schéma doit être indépendante du paramètre ε ;
2. Le schéma doit converger vers une discrétisation du modèle limite lorsque $\varepsilon \to 0$;
3. Il est possible de prendre explicitement $\varepsilon = 0$ dans le schéma.

La seconde propriété est le plus souvent illustrée par le Diagramme 4. Les problèmes $\mathcal{P}^\varepsilon$ et $\mathcal{P}$ correspondent aux équations (36) et (37) et le nombre h désigne la taille de la discrétisation des problèmes discrets $\mathcal{P}_h^\varepsilon$ et $\mathcal{P}_h$, associés à (36) et (37).

Grâce à leur robustesse, les schémas AP peuvent être utilisés dans de nombreux contextes. On peut citer leur utilisation pour des équations cinétiques en régime diffusif [55, 59, 68, 124, 125], en

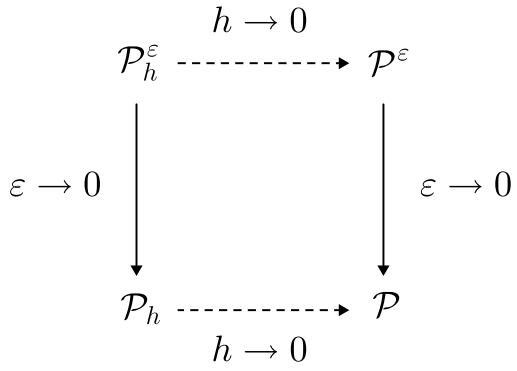


FIGURE 4 – Diagramme d'un schéma préservant l'asymptotique.

limite quasi-neutre [58] ou encore pour des équations de type Lotka-Voltera [40] par exemple. On renvoie aussi à l'article de review [135] ainsi qu'aux références qu'il contient.

Bien que les schémas AP garantissent un bon comportement quelle que soit la valeur du paramètre ε, ils peuvent néanmoins rester très coûteux numériquement pour les équations cinétiques. En effet, dans un cadre général, les variables d'espace x et de vitesse v appartiennent à des domaines en trois dimensions. Ainsi, la taille typique du système pour une application réaliste est en 7 dimensions en comptant la variable de temps. En pratique, de telles simulations sont extrêmement coûteuses. De plus, la précision d'une description cinétique n'est parfois pas nécessaire dans l'ensemble du domaine de calcul et une approche fluide peut se révéler suffisante.

De nombreuses techniques dites *hybrides* utilisant à la fois les descriptions cinétiques et fluides ont été développées pour réduire le coût des simulations. Une première approche, introduite pour la mise à l'échelle hydrodynamique dans [63], repose sur l'adaptation du domaine spatial en ajoutant une zone tampon entre des sous-domaines cinétiques et fluides en utilisant une fonction de transition dans le modèle continu. Elle a été étendue à un cadre dynamique avec des interfaces mobiles dans [61] et appliquée à la mise à l'échelle diffusive dans [69]. Toujours en considérant l'adaptation du domaine en position, il est aussi possible de diviser la fonction de distribution en une partie macroscopique résolue via la méthode des volumes finis et une partie microscopique résolue à l'aide d'une méthode de Monte-Carlo [55, 70]. Un cas stationnaire est aussi considéré dans [130, 131]. La *Heterogeneous Multiscale Method* (HMM) introduite dans [77, 78] repose sur l'aspect multi-échelles du problème avec en particulier la question du lien de reconstruction/projection entre des données microscopiques et macroscopiques. Nous faisons également référence à [145] où, en plus d'un couplage entre modèles cinétiques et fluides, un raffinement de maillage adaptatif est également utilisé. Une technique ne reposant pas sur une décomposition de domaine, mais similaire à [69] a aussi été introduite dans [65] où le modèle fluide est utilisé dans tout le domaine avec des évolutions localisées de la perturbation. Enfin, une approche alternative pour réduire le coût numérique repose sur les méthodes *Dynamical Low Rank* (DLR) [80, 143], où l'idée est de réduire la matrice du système linéaire associé au problème discret en projetant l'équation sur une variété de rang faible.

Contributions

Le cœur de ce travail, présenté dans le Chapitre 3, porte sur la conception d'une méthode hybride où le modèle utilisé est adapté dynamiquement et localement dans le domaine de calcul spatial. Afin de gagner en coût de calcul, on cherche à utiliser le modèle fluide asymptotique moins coûteux là où cette description est valide, dans un sens à déterminer, avec une approche cinétique dans le reste du domaine. Un point crucial pour mettre en œuvre une telle méthode consiste à déterminer de façon précise la validité d'une description fluide. Un certain nombre de critères ont déjà été étudiés dans ce but [61, 145, 199, 200, 202]. Ces derniers varient, certains étant basés uniquement sur des quantités macroscopiques, d'autres sur la distance entre la solution et l'équilibre thermodynamique.

Nous choisissons pour notre méthode hybride d'utiliser deux critères. Le premier est un critère mésoscopique qui n'est autre que la distance avec un équilibre local en vitesse. En effet, nous avons vu que la densité de la Maxwellienne de vitesse moyenne nulle et température 1 était solution du modèle fluide de dérive-diffusion (37). Le second critère est quant à lui motivé par les travaux [87, 145, 158, 201] basés sur la *matrice de réalisabilité* des moments. Dans le cas de l'équation (36), comme seule la masse est conservée, on s'intéresse uniquement à l'évolution de la densité dans le modèle macroscopique (37). Le critère de réalisabilité se réduit alors à une simple condition scalaire en chaque point du maillage spatial.

Nous avons vu précédemment dans la Section 4 que le modèle macroscopique de dérive-diffusion (37) pouvait s'obtenir en tronquant le développement de Chapman-Enskog à l'ordre 1 :

$$f = \mathcal{M}_{\rho,0,1} + \varepsilon g^{(1)}.$$

Afin d'estimer la validité d'une telle description, on choisit de considérer les perturbations d'ordre supérieur en tronquant maintenant à l'ordre 3 :

$$f = \mathcal{M}_{\rho,0,1} + \varepsilon g^{(1)} + \varepsilon^2 g^{(2)} + \varepsilon^3 g^{(3)}.$$

En introduisant cette expression dans l'équation cinétique (36) et en intégrant en vitesse, on obtient alors un modèle macroscopique d'ordre supérieur qui n'est autre qu'une correction à l'ordre 2 de (37) :

$$\partial_t \rho^\varepsilon - \partial_x J^\varepsilon + \varepsilon^2 \partial_x \left(2\,\partial_x (E J^\varepsilon) - E\,\partial_x J^\varepsilon - \partial_{xx} J^\varepsilon \right) = \mathcal{O}(\varepsilon^4), \tag{38}$$

avec le flux $J^\varepsilon = \partial_x \rho^\varepsilon - E\rho^\varepsilon$. En notant maintenant

$$\mathcal{R}^\varepsilon = \varepsilon^2 \partial_x \left(2\,\partial_x (E J^\varepsilon) - E\,\partial_x J^\varepsilon - \partial_{xx} J^\varepsilon \right),$$

on dispose alors d'une quantité macroscopique, indépendante de la variable vitesse et calculable sur tout le domaine, qui permet de quantifier la validité du régime fluide. Lorsque $\mathcal{R}^\varepsilon$ est petit, la solution est localement proche de la dynamique asymptotique. À l'inverse, si $\mathcal{R}^\varepsilon$ est grand, les perturbations ont beaucoup d'impact sur les quantités macroscopiques. Il est important de noter que l'utilisation de deux critères permet d'ajuster le modèle utilisé à la fois de cinétique vers fluide, mais aussi de fluide vers cinétique permettant un ajustement très fin.

Numériquement, le schéma utilisé repose sur l'approche micro-macro [59, 156] et les in-

tégrateurs exponentiels [126, 155]. L'idée est de considérer l'évolution de la décomposition $f^\varepsilon = \mathcal{M}_{\rho,0,1} + g^\varepsilon$. Après une réécriture de (36), on obtient le modèle micro-macro :

$$\begin{cases} \partial_t g^\varepsilon + \dfrac{1}{\varepsilon}\Big((v\cdot\nabla_x + E\cdot\nabla_v)g^\varepsilon - \mathrm{div}_x\langle v g^\varepsilon\rangle\mathcal{M} + v\mathcal{M}\cdot J^\varepsilon\Big) = \dfrac{-1}{\varepsilon^2}g^\varepsilon, \\ \partial_t \rho^\varepsilon + \dfrac{1}{\varepsilon}\mathrm{div}_x\langle v g^\varepsilon\rangle = 0. \end{cases}$$

Ce modèle tient son nom de la première équation, appelée *micro*, qui décrit l'évolution de la perturbation g^ε, et de la seconde sur la densité ρ^ε, appelée *macro* en référence à la partie se trouvant à l'équilibre thermodynamique. Avec cette reformulation, une discrétisation temporelle bien choisie [59, 156] permet d'obtenir une méthode AP pour la limite de diffusion. Grâce à un tel schéma, on dispose alors d'un solveur cinétique qui est robuste pour n'importe quelle valeur du paramètre ε, mais néanmoins coûteux même en régime fluide. Ce dernier est cependant naturellement approché par le schéma limite dont le coût est pratiquement invisible face au cinétique.

Du point de vue de l'implémentation de la méthode, un enjeu crucial est la gestion des interfaces entre domaines cinétiques et fluides. En effet, une technique classique de discrétisation spatiale du modèle micro-macro consiste à approcher la perturbation sur une grille décalée. Cela permet d'obtenir un stencil plus compact pour le schéma asymptotique. Plus précisément, le schéma obtenu pour la partie diffusion est l'approximation classique à trois points là où l'absence de grille décalée engendrerait un stencil à cinq points [59]. La nouveauté a donc été non seulement de faire communiquer des données à des échelles différentes, mais aussi sur des maillages décalés. La méthode a été implémentée en Fortran 90 dans un cadre $1D_x/3D_v$ et s'est montrée particulièrement robuste, notamment dans un cas où le nombre de Knudsen n'était pas constant en espace. On réfère à la Figure 6 pour un aperçu de l'évolution de la densité solution de (36). La Figure 5 illustre une accélération du temps de calcul. Plus précisément, le ratio

$$\text{Speedup} = \frac{\text{Temps de simulation hybride}}{\text{Temps de simulation cinétique}}$$

atteint jusqu'à un facteur 100 dans le cas de l'équation de Vlasov-Poisson-BGK.

Algorithme pararéel multi-échelle

La méthode pararéelle est une méthode introduite dans [159, 163] qui consiste à paralléliser en temps le calcul de la solution d'un système dynamique. Elle s'est notamment énormément développée ces dernières années avec l'évolution des architectures de calcul intensif. Il est donc intéressant d'appliquer une telle approche aux équations cinétiques qui, comme nous l'avons mentionné dans les sections précédentes, sont très coûteuses numériquement.

Commençons par considérer une simple EDO de la forme

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\,t}u(t) = f(u), \quad t\in[0,T], \\ u(0) = u^0. \end{cases}$$

Dans la méthode pararéelle, on s'intéresse à l'approximation de la solution à des temps T_n fixés qui sera raffinée par un processus itératif. La première étape est donc de décomposer l'intervalle

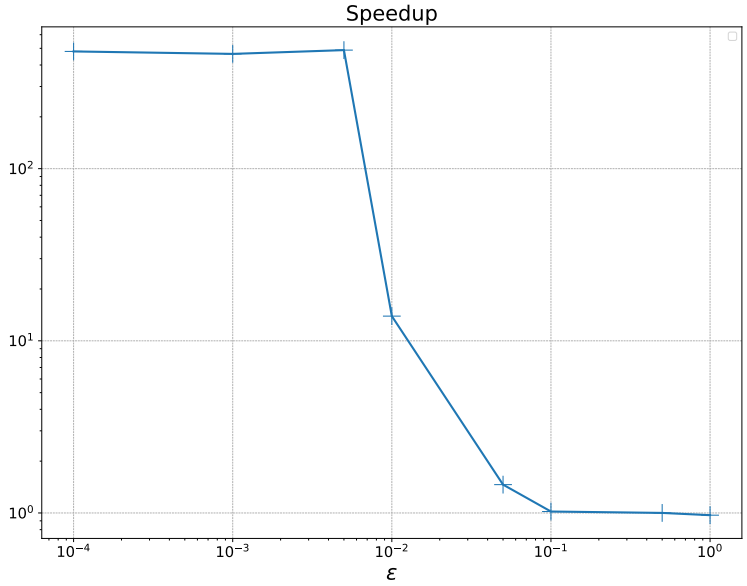


FIGURE 5 – Accélération du temps de calcul de la méthode hybride pour l'équation de Vlasov-Poisson-BGK en fonction du nombre de Knudsen.

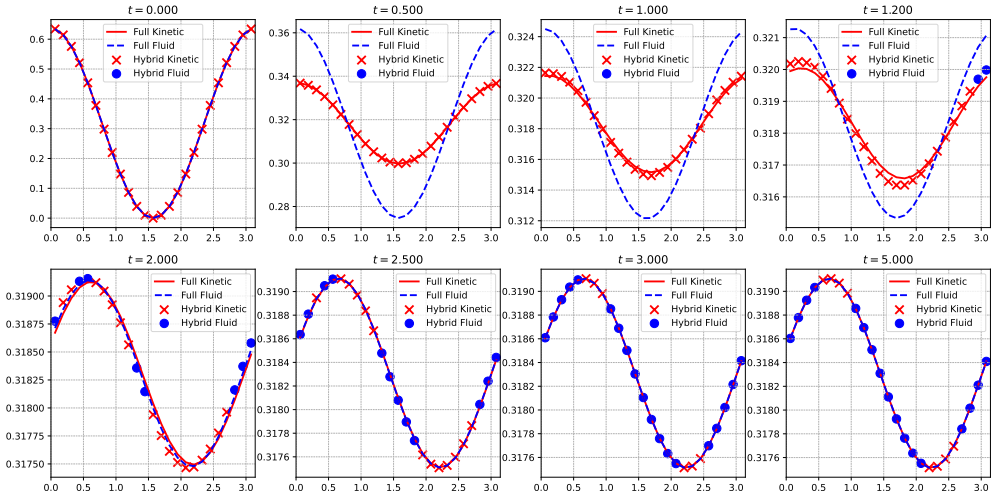


FIGURE 6 – Évolution de la densité approchée par des méthodes cinétique, hybride et fluide.

de temps $[0,T]$ en $N_g \in \mathbb{N}$ sous-intervalles uniformes $[T^n, T^{n+1}]$, $n \in \{0,\dots,N_g-1\}$. Le cœur de l'algorithme repose ensuite sur deux types de solveurs aussi appelés propagateurs. Le premier est grossier, mais rapide, noté $\mathcal{G}(T^n, T^{n+1}, U^n)$, et le second est un solveur fin, mais coûteux, noté $\mathcal{F}(T^n, T^{n+1}, U^n)$. Ces solveurs sont utilisés pour propager une donnée U^n du temps T^n au temps T^{n+1}. L'algorithme est initialisé par une première prédiction (peu coûteuse) avec le solveur grossier :

$$U^{n+1,0} = \mathcal{G}(T^n, T^{n+1}, U^{n,0}), \quad U^{0,0} = u^0.$$

Ensuite, pour $k = 1, 2, \dots$, l'approximation aux temps T^n est raffinée jusqu'à une tolérance souhaitée grâce à processus de prédiction-correction, appelé *itérations pararéelles*, donné par la relation :

$$U^{n+1,k+1} = \mathcal{G}(T^n, T^{n+1}, U^{n,k+1}) + \mathcal{F}(T^n, T^{n+1}, U^{n,k}) - \mathcal{G}(T^n, T^{n+1}, U^{n,k}),$$

avec $U^{0,k} = u^0$ pour tout k. L'intérêt de cet algorithme itératif est de pouvoir calculer en parallèle les propagations fines $\mathcal{F}(T^n, T^{n+1}, U^{n,k})$. Ainsi, tant que le nombre d'itérations pararéelles k reste suffisamment petit pour obtenir une précision donnée, on peut espérer un gain en temps de calcul sur l'intégration temporelle.

Les méthodes *Parallel IN Time* (PINT) ont été largement utilisées et analysées pour des systèmes variés. La convergence de l'algorithme a été étudiée dans [75, 100, 101] notamment pour des variations de l'algorithme original. Les applications sont nombreuses et on renvoie par exemple à quelques travaux pour le transport de neutrons [13], les équations de Navier-Stokes [89, 195], des systèmes Hamiltoniens [99] ou encore des turbulences dans des plasmas [191]. On mentionne aussi de récents travaux sur l'accélération de la méthode [162, 172].

Dans un autre contexte, celui des systèmes multi-échelles, l'algorithme pararéel a été utilisé dans [31, 75, 116, 117, 154, 192] notamment à l'aide de modèles réduits. L'idée est d'utiliser un modèle plus simple comme prédicteur grossier d'un modèle plus riche. Ce type d'approche reprend notamment des principes de la méthode HMM [1, 77, 78].

Contributions

Nous avons vu dans la Section 4 précédente une technique pour coupler un modèle cinétique et son modèle fluide asymptotique qui permet de réduire significativement le coût d'une simulation mésoscopique. Le couplage avait alors lieu dans la variable de position. Le dernier chapitre de cette thèse porte sur l'exploration d'une méthode pararéelle où, contrairement à ce qui a pu être fait dans d'autres travaux, le couplage n'aura pas lieu en espace ou en vitesse, mais en temps. La construction d'une telle méthode pararéelle pour un modèle cinétique n'a, à notre connaissance, par encore été présentée.

En s'inspirant des travaux sur des méthodes pararéelles multi-échelles, on s'intéresse à accélérer la simulation d'un modèle cinétique collisionnel $1D_x/3D_v$ avec mise à l'échelle hydrodynamique (14) en utilisant les équations d'Euler avec force comme modèle réduit ou avec une mise à l'échelle diffusive avec comme modèle fluide celui de dérive-diffusion. Le solveur cinétique consiste en une discrétisation volumes finis upwind en espace et vitesse avec une intégration de type Euler explicite en temps. Le solveur fluide est aussi discrétisé par une méthode volumes finis cette fois avec des flux de type Rusanov et toujours une intégration temporelle de type Euler explicite.

Il a été remarqué que l'algorithme pararéel appliqué à des systèmes hydrodynamiques, où la convection domine la diffusion, pouvait converger difficilement [9, 79, 83, 98, 173]. Il a aussi été noté [100] dans le cas de l'équation de Burgers que plus l'intervalle de simulation est grand, moins la convergence est rapide, et ce, pour une discrétisation identique. En vue de ces observations, on peut alors s'attendre à ce que la méthode pararéelle appliquée à une équation cinétique en régime hydrodynamique et avec un petit nombre de Knudsen soit complexe. Une seconde piste de réflexion serait de plutôt considérer une mise à l'échelle diffusive, régime dans lequel la méthode pararéelle est réputée mieux fonctionner.

La question cruciale des méthodes multi-échelles est celle du lien entre les modèles. Dans notre contexte, on choisit simplement de définir le passage de la distribution $f(t,x,v)$ aux moments $U = (\rho(t,x), u(t,x), \theta(t,x))^T$ par la projection

$$U = \mathcal{P}f = \begin{pmatrix} \int_{\mathbb{R}^3} f \,\mathrm{d}v \\ \frac{1}{\rho}\int_{\mathbb{R}^3} v f \,\mathrm{d}v \\ \frac{1}{3\rho}\int_{\mathbb{R}^3} |v-u|^2 f \,\mathrm{d}v \end{pmatrix}.$$

À l'inverse, pour reconstruire la fonction de distribution à partir des moments, une approche naturelle, décrite par la validité des équations d'Euler, est de relever les moments U vers la Maxwellienne associée :

$$f = \mathcal{L}U = \frac{\rho}{(\sqrt{2\pi\theta})^3}\exp\left(-\frac{|v-u|^2}{2\theta}\right),$$

où $\mathcal{L}$ est appelé opérateur de relèvement. Cette méthode renvoie notamment à un récent travail où des maillages fins et grossiers sont utilisés dans le processus de prédiction / correction [6]. On note maintenant $U^{n,k}$ les moments au temps T^n et à l'itération pararéelle k. Afin d'alléger la notation, on pose aussi $\Gamma = \mathcal{P}\mathcal{F}\mathcal{L}$ l'opérateur de Projection-Évolution-Relèvement. La méthode peut alors être résumée par l'Algorithme 1. Les premiers résultats de cette approche sont présentés dans le

Algorithm 1 Multi-scale kinetic parareal Algorithm

Require: $U^{0,0}$
  for $n = 1, \dots N_g$ **do** ▷ First coarse guess
    $U^{n,0} \leftarrow \mathcal{G}\left(U^{n-1,0}\right)$
  end for
  while $k \leq K$ **or** error $\geq$ tol **do** ▷ Parareal iterations
    for $n = 1, \dots N_g$ **do** ▷ Compute the jumps in parallel
      $\Delta^n = \Gamma(U^{n-1,k-1}) - \mathcal{G}\left(U^{n-1,k-1}\right)$
    end for
    for $n = 1, \dots N_g$ **do** ▷ Sequential correction
      $U^{n,k+1} = \mathcal{G}\left(U^{n-1,k}\right) + \Delta^n$
    end for
    Compute successive error on the moments and $k \leftarrow k+1$
  end while

Chapitre 4. L'implémentation de la méthode a été réalisée en Fortran 90 et la parallélisation de la méthode est en cours d'étude. En effet, l'intérêt principal de cette méthode est le gain en temps de calcul. Ainsi, il est prévu de déployer le code sur un grand nombre de cœurs.

Part I

Structure preserving numerical methods

Chapter 1

Discrete Hypocoercivity for a nonlinear kinetic reaction model

In this chapter, I present a finite volume discretization of a one dimensional nonlinear kinetic model which describes a 2-species recombination-generation process. Specifically, the long-time convergence of approximate solutions towards equilibrium is investigated. The study is based on an adaptation for a discretization of the linearized problem of an L^2 hypocoercivity method. From this, a local result for the discrete nonlinear problem can be deduced.
The content of this chapter covers the work published in *IMA Journal of Numerical Analysis*:

— M. Bessemoulin-Chatard, T. Laidin, and T. Rey. "Discrete hypocoercivity for a nonlinear kinetic reaction model". In: *IMA JNA* (Sept. 2024), drae058. ISSN: 0272-4979. DOI: 10.1093/imanum/drae058

Outline of the current chapter

1.1 Introduction

1.1.1 On a kinetic generation-recombination model

In this chapter, we analyze the decay to equilibrium of a finite volume scheme for a one-dimensional kinetic relaxation model describing a generation-recombination reaction of two species proposed in [171], which can be seen as a simplified version of models describing generation and recombination of electron-hole pairs in semiconductors (see *e.g.* [60]). More precisely, we consider the system

$$\partial_t \mathfrak{f} + v\,\partial_x \mathfrak{f} = \chi_1 - \rho_{\mathfrak{g}}\mathfrak{f}, \tag{1.1}$$

$$\partial_t \mathfrak{g} + v\,\partial_x \mathfrak{g} = \chi_2 - \rho_{\mathfrak{f}}\mathfrak{g}, \tag{1.2}$$

where $\mathfrak{f}$ and $\mathfrak{g}$ represent the phase space densities of two chemical reactants A and B. These reactants are produced by the decomposition of a substance C (whose density is assumed to be fixed) with nonnegative velocity profiles χ_1 and χ_2, and can also recombine to form C and thus be eliminated from the system. The densities $\mathfrak{f}$ and $\mathfrak{g}$ depend on time $t \geq 0$, position $x \in \mathbb{T}$ the one-dimensional torus, and velocity $v \in \mathbb{R}$. The probability of the reaction depends on the position density of the reaction partner, defined by

$$\rho_{\mathfrak{h}}(t,x) := \int_{\mathbb{R}} \mathfrak{h}(t,x,v)\,\mathrm{d}\,v, \qquad \mathfrak{h} = \mathfrak{f}, \mathfrak{g}. \tag{1.3}$$

The system (1.1)–(1.2) is completed by the initial condition

$$\mathfrak{f}(0,x,v) = \mathfrak{f}_I(x,v) \geq 0, \quad \mathfrak{g}(0,x,v) = \mathfrak{g}_I(x,v) \geq 0. \tag{1.4}$$

Since the chemical reaction under consideration is assumed to be reversible, it is required that $\int_{\mathbb{R}}(\chi_1 - \chi_2)\,\mathrm{d}\,v = 0$. We moreover make the following assumptions about the moments of the velocity profiles, which are given positive functions of v: for $k = 1, 2$,

$$\begin{aligned} \int_{\mathbb{R}} \chi_k\,\mathrm{d}\,v = 1, \qquad \int_{\mathbb{R}} v\,\chi_k\,\mathrm{d}\,v = 0, \\ D_k := \int_{\mathbb{R}} v^2\,\chi_k\,\mathrm{d}\,v < \infty, \quad Q_k := \int_{\mathbb{R}} v^4\,\chi_k\,\mathrm{d}\,v < \infty. \end{aligned} \tag{1.5}$$

The purpose of this chapter is to propose a numerical scheme for the system (1.1)–(1.2) for which we are able to rigorously study the long-time behavior. More precisely, remarking that the mass difference is conserved:

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\int_{\mathbb{T}\times\mathbb{R}} (\mathfrak{f} - \mathfrak{g})\,\mathrm{d}\,v\,\mathrm{d}\,x = 0,$$

let us introduce the unique constant $\rho_\infty > 0$ such that

$$\int_{\mathbb{T}\times\mathbb{R}} (\mathfrak{f}_I - \mathfrak{g}_I)\,\mathrm{d}\,v\,\mathrm{d}\,x = |\mathbb{T}|\left(\rho_\infty - \frac{1}{\rho_\infty}\right). \tag{1.6}$$

The equilibrium state $\mathsf{F}_\infty = (\mathfrak{f}_\infty, \mathfrak{g}_\infty)$, depending only on velocity, is then defined by

$$\mathfrak{f}_\infty(x,v) := \rho_\infty\,\chi_1(v), \qquad \mathfrak{g}_\infty(x,v) := \frac{1}{\rho_\infty}\chi_2(v). \tag{1.7}$$

In [171], the exponential decay to equilibrium (what we shall call in this chapter hypocoercivity) of the solution to the linearization of (1.1)–(1.2) is established by applying the method proposed in [73]. Then, this result is extended to a local result for the nonlinear case, thanks to the proof of maximum principle estimates for the nonlinear problem.

1.1.2 Hypocoercivity

Due to the relaxation structure of the right-hand side of system (1.1)–(1.2) and to the mixing properties of the free transport operator on the torus, one can naturally expect that the solutions to this type of problem will exhibit a fast relaxation towards F_∞. This is a classical problem, which can be traced back to the seminal work of Hörmander [129] on the hypoellipticity of linear operators.

The first partial proof of this large time behavior can be found in [17]. It was then proved for a very large class of collision operators (with or without confinement in velocity) in [120, 121], and in the series of papers [33, 73] (see also the references therein) that for a suitable norm, the rate of convergence towards the equilibrium is exponential: there exists some constants $\lambda > 0$ and $C \geq 1$ such that

$$\|\mathfrak{h}(t) - \mathfrak{h}_\infty\|_{\mathcal{X}} \leq C\,\|\mathfrak{h}_I - \mathfrak{h}_\infty\|_{\mathcal{X}}\,e^{-\lambda t},$$

in a well-chosen Hilbert space $\mathcal{X}$. We shall call this type of behavior *hypocoercivity* [204] (the case $C = 1$ corresponding to classical coercive behavior).

A very robust proof for establishing such a result can be found in the seminal work [73]. It introduces a so-called "modified entropy functional" that is equivalent to a weighted L^2 norm and decays exponentially fast toward the global Maxwellian equilibrium. This technique was then used extensively in many applications, showing its robustness and practicality. One can cite for example its use on the kinetic Keller-Segel-type chemotaxis models (the so-called Othmer, Dunbar and Alt model) in [41]; on the study of the fractional Fokker-Planck equation in [32]; on the fine properties of a large class of Vlasov-Fokker-Planck models in [34]; on the large time behavior of the Vlasov-Poisson-Fokker-Planck equation in [2]. Note that this technique has been very recently refined to singular problems, with motivation in control theory in [66].

The goal of the current work is to establish hypocoercive properties of finite volume types discretizations of the nonlinear system (1.1)–(1.2). Such an objective has been classical in the last decade for numerical approximations of macroscopic models exhibiting an entropic structure, see for example the classical works [19, 38, 46, 112]. A particular emphasis on the accurate discretization of steady states was done in [86, 179]. These results were then extended to problems with nonhomogeneous boundary conditions using the so-called ϕ-entropy method in [47, 84].

Extension to numerical discretizations of kinetic models which are able to reproduce accurately this type of behavior then becomes natural. This can be considered as a special case of Asymptotic Preserving (AP) schemes, where the asymptotic regime preserved is the large time behavior of the model. The first work on the subject considered the discretization of the so-called Kolmogorov master equation, which is a simplified kinetic model where the collision operator is only a Laplacian in velocity. Both finite differences and finite elements approaches were considered in the papers [103, 186]. Such AP schemes were then recently used to stabilize a finite element solver in [74]. The first numerical work "à la Villani" about the H^1-hypocoercivity of the linear kinetic Fokker-Planck equation using a finite difference approach was then published in [76]. It uses the framework introduced in [204] to establish the hypocoercive properties of the numerical scheme. The L^2 method from [73] was then used to establish both hypocoercivity properties and uniform stability of a finite volume scheme for both the linear BGK and kinetic Fokker-Planck equations in [18]. This is the approach we will generalize in the current work. Let us also mention the recent similar work about the hypocoercive properties of a finite volume method for the fractional Fokker-Planck equation in [7]. Finally, extensions of the L^2 method to the Vlasov equation (both with an external electric field or the linearized Vlasov-Poisson system), using an Hermite expansion in the velocity space was recently developed in [24, 25].

The outline of this chapter is as follows. In Section 1.2, we briefly present the results in the continuous setting, since we are going to adapt them in the discrete framework in the following. In Section 1.3, we introduce the discrete setting, in particular the definition of the numerical schemes for the nonlinear problem (1.1)–(1.2) and for its linearization around the equilibrium. Section 1.4 is devoted to the adaptation of the L^2-hypocoercivity method of [73] for the discretization of the linearized problem. Then in Section 1.5, we establish the discrete counterpart of the local result for the nonlinear case. As in the continuous framework, it requires maximum principle estimates, which necessitates the use of monotone numerical fluxes. This motivates our choice of the Lax-Friedrichs fluxes for the transport terms, unlike the centered fluxes used in [18]. Finally, in Section 1.6, we present some numerical experiments to illustrate the obtained theoretical results.

1.2 The continuous setting

In this section, we recall the main lines of the result in the continuous framework [171]. The decay towards equilibrium will be estimated quantitatively in the following weighted L^2 space:

$$\mathcal{H} := L^2(\mathbb{T}\times\mathbb{R}, \mathrm{d}x\,\mathrm{d}v/\chi_1)\times L^2(\mathbb{T}\times\mathbb{R}, \mathrm{d}x\,\mathrm{d}v/\chi_2), \tag{1.8}$$

endowed with the scalar product weighted with the steady state measure:

$$\langle F_1, F_2\rangle = \int_{\mathbb{T}\times\mathbb{R}} \left(\frac{f_1 f_2}{\rho_\infty \chi_1} + \frac{g_1 g_2 \rho_\infty}{\chi_2}\right) \mathrm{d}v\,\mathrm{d}x, \quad \text{for } F_k = \begin{pmatrix} f_k \\ g_k \end{pmatrix}.$$

The corresponding norm is denoted by $\|\cdot\|$.

1.2.1 The linearized case

Let us now introduce the linearization of the system (1.1)–(1.2) around the equilibrium f_∞. We shall denote by (f,g) the perturbations $\mathsf{f} = \mathsf{f}_\infty + \varepsilon f$, $\mathsf{g} = \mathsf{g}_\infty + \varepsilon g$. Formally, in the limit $\varepsilon \to 0$, one gets the following linearized problem:

$$\partial_t f + v\,\partial_x f = -\rho_\infty \chi_1 \rho_g - \rho_\infty^{-1} f, \tag{1.9}$$

$$\partial_t g + v\,\partial_x g = -\rho_\infty^{-1} \chi_2 \rho_f - \rho_\infty g. \tag{1.10}$$

Remark that the perturbations f and g now satisfy

$$\int_{\mathbb{T}\times\mathbb{R}} (f-g)\,\mathrm{d}x\,\mathrm{d}v = \int_{\mathbb{T}} (\rho_f - \rho_g)\,\mathrm{d}x = 0.$$

The orthogonal projection onto the null space of the linearized collision operator

$$\mathsf{L}F = \begin{pmatrix} -\rho_\infty \chi_1 \rho_g - \rho_\infty^{-1} f \\ -\rho_\infty^{-1} \chi_2 \rho_f - \rho_\infty g \end{pmatrix}, \quad \text{for } F = \begin{pmatrix} f \\ g \end{pmatrix},$$

is given by

$$\Pi F = \frac{\rho_f - \rho_g}{\rho_\infty^2 + 1} \begin{pmatrix} \rho_\infty^2 \chi_1 \\ -\chi_2 \end{pmatrix}.$$

First, the following microscopic coercivity estimate can be established.

Lemma 1 (Microscopic coercivity). *Let (1.5) hold and let $F = (f,g)$ be the solution to the linearized system (1.9)–(1.10) with initial data $F_I = (f_I, g_I) \in \mathcal{H}$ satisfying $\int_{\mathbb{T}\times\mathbb{R}} (f_I - g_I)\,\mathrm{d}x\,\mathrm{d}v = 0$.*

Then for every $t \geq 0$,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|F(t)\|^2 + C_{mc}\|(I-\Pi)F(t)\|^2 \leq 0, \quad C_{mc} = \min(\rho_\infty, \rho_\infty^{-1}).$$

Proof. Since F is the solution to (1.9)–(1.10),

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|F(t)\|^2 = \langle \mathsf{L}F, F\rangle.$$

Then straightforward computations using only the fact that $\int_{\mathbb{R}} \chi_k\,\mathrm{d}v = 1$ (see [171]) show that

$$-\langle \mathsf{L}F, F\rangle \geq \min(\rho_\infty, \rho_\infty^{-1})\,\|(I-\Pi)F\|^2, \tag{1.11}$$

which concludes the proof. □

Let us now introduce the moments of $h := f - g$:

$$u_h := \int_{\mathbb{R}} h\,\mathrm{d}v, \quad J_h := \int_{\mathbb{R}} v\,h\,\mathrm{d}v, \quad S_h := \int_{\mathbb{R}} (v^2 - D_0)\,h\,\mathrm{d}v, \tag{1.12}$$

where

$$D_0 = \frac{\rho_\infty^2 D_1 + D_2}{\rho_\infty^2 + 1}, \tag{1.13}$$

D_k being defined in (1.5).

By subtracting (1.10) from (1.9), multiplying by $(1, v)$ and integrating with respect to v, one obtains the following moments equations for h:

$$\partial_t u_h + \partial_x J_h = 0, \tag{1.14}$$
$$\partial_t J_h + \partial_x S_h + D_0 \partial_x u_h = -(\rho_\infty^{-1} J_f - \rho_\infty J_g), \tag{1.15}$$

where J_f and J_g are the first order moments of f and g respectively.

Lemma 2 (Moments estimates)**.** *Under assumptions* (1.5)*, the moments u_h, J_h and S_h satisfy the following estimates:*

$$\|u_h\|_{L^2(\mathbb{T})} = C_u \|\Pi F\|, \tag{1.16}$$
$$\|J_h\|_{L^2(\mathbb{T})} \le C_{J1} \|F\|, \tag{1.17}$$
$$\|J_h\|_{L^2(\mathbb{T})} \le C_{J1} \|(I-\Pi)F\|, \tag{1.18}$$
$$\|S_h\|_{L^2(\mathbb{T})} \le C_S \|(I-\Pi)F\|, \tag{1.19}$$
$$\|\rho_\infty^{-1} J_f - \rho_\infty J_g\|_{L^2(\mathbb{T})} \le C_{J2} \|(I-\Pi)F\|, \tag{1.20}$$

where the constants are given by:

$$C_u = \sqrt{\frac{\rho_\infty^2 + 1}{\rho_\infty}}, \tag{1.21}$$
$$C_{J1} = \sqrt{2\max(\rho_\infty D_1, \rho_\infty^{-1} D_2)}, \tag{1.22}$$
$$C_S = \sqrt{2\max(\rho_\infty(Q_1 - 2D_0D_1 + D_0^2), \rho_\infty^{-1}(Q_2 - 2D_0D_2 + D_0^2))}, \tag{1.23}$$
$$C_{J2} = \max(\rho_\infty^{-1}, \rho_\infty) C_{J1}. \tag{1.24}$$

Proof. Let us start with equality (1.16). By definition of $\|\Pi F\|$ and using the first property in (1.5) one has:

$$\begin{aligned}
\|\Pi F\|^2 &= \frac{1}{(\rho_\infty^2+1)^2} \int_{\mathbb{T}} \int_{\mathbb{R}} (\rho_f - \rho_g)^2 \left(\frac{\rho_\infty^4 \chi_1^2}{\chi_1 \rho_\infty} + \frac{\chi_2^2 \rho_\infty}{\chi_2} \right) \mathrm{d}v \, \mathrm{d}x \\
&= \frac{\rho_\infty}{(\rho_\infty^2+1)^2} \int_{\mathbb{T}} (\rho_f - \rho_g)^2 (\rho_\infty^2 + 1) \, \mathrm{d}x \\
&= \frac{\rho_\infty}{\rho_\infty^2+1} \|u_h\|^2_{L^2(\mathbb{T})}.
\end{aligned}$$

Now for (1.17), by definition of J_h and using the identity $(a-b)^2 \le 2(a^2+b^2)$ one has

$$\|J_h\|^2_{L^2(\mathbb{T})} \le 2 \int_{\mathbb{T}} \left(\int_{\mathbb{R}} v f \frac{\sqrt{\rho_\infty \chi_1}}{\sqrt{\rho_\infty \chi_1}} \mathrm{d}v \right)^2 \mathrm{d}x + 2 \int_{\mathbb{T}} \left(\int_{\mathbb{R}} v g \sqrt{\frac{\rho_\infty}{\chi_2}} \sqrt{\frac{\chi_2}{\rho_\infty}} \mathrm{d}v \right)^2 \mathrm{d}x. \tag{1.25}$$

This yields thanks to Cauchy-Schwarz inequality

$$\|J_h\|^2_{L^2(\mathbb{T})} \le 2 \int_{\mathbb{T}} \left(\rho_\infty D_1 \int_{\mathbb{R}} \frac{f^2}{\rho_\infty \chi_1} \mathrm{d}v + \frac{D_2}{\rho_\infty} \int_{\mathbb{R}} \frac{g^2 \rho_\infty}{\chi_2} \mathrm{d}v \right) \mathrm{d}x,$$

from which we deduce (1.17).

The second estimate (1.18) on $\|J_h\|_{L^2(\mathbb{T})}$ is obtained by observing that using the second property in (1.5), one has

$$\|J_h\|^2_{L^2(\mathbb{T})} \le 2\int_{\mathbb{T}}\left[\left(\int_{\mathbb{R}} v\left(f - \frac{u_h}{\rho_\infty^2+1}\rho_\infty^2\chi_1\right)\mathrm{d}\,v\right)^2 + \left(\int_{\mathbb{R}} v\left(g + \frac{u_h}{\rho_\infty^2+1}\chi_2\right)\mathrm{d}\,v\right)^2\right]\mathrm{d}\,x,$$

from which we deduce using Cauchy-Schwarz inequality

$$\begin{aligned}\|J_h\|^2_{L^2(\mathbb{T})} \le{}& 2\int_{\mathbb{T}}\rho_\infty D_1\int_{\mathbb{R}}\left(f - \frac{u_h}{\rho_\infty^2+1}\rho_\infty^2\chi_1\right)^2\frac{1}{\rho_\infty\chi_1}\mathrm{d}\,v\,\mathrm{d}\,x\\ &+ 2\int_{\mathbb{T}}\frac{D_2}{\rho_\infty}\int_{\mathbb{R}}\left(g + \frac{u_h}{\rho_\infty^2+1}\chi_2\right)^2\frac{\rho_\infty}{\chi_2}\mathrm{d}\,v\,\mathrm{d}\,x,\end{aligned}$$

from which we obtain (1.18).

Regarding the estimation (1.19) of $\|S_h\|_{L^2(\mathbb{T})}$, we notice that by definition (1.13) of D_0

$$\int_{\mathbb{R}}(v^2 - D_0)\frac{u_h}{\rho_\infty^2+1}(\rho_\infty^2\chi_1 + \chi_2)\mathrm{d}\,v = 0,$$

and as previously, we use that $(a-b)^2 \le 2(a^2+b^2)$, leading to the estimate

$$\begin{aligned}\|S_h\|^2_{L^2(\mathbb{T})} \le{}& 2\int_{\mathbb{T}}\left(\int_{\mathbb{R}}(v^2 - D_0)\left(f - \frac{u_h}{\rho_\infty^2+1}\rho_\infty^2\chi_1\right)\mathrm{d}\,v\right)^2\mathrm{d}\,x\\ &+2\int_{\mathbb{T}}\left(\int_{\mathbb{R}}(v^2 - D_0)\left(g + \frac{u_h}{\rho_\infty^2+1}\chi_2\right)\mathrm{d}\,v\right)^2\mathrm{d}\,x.\end{aligned}$$

The right-hand side can then be rewritten

$$\begin{aligned}\|S_h\|^2_{L^2(\mathbb{T})} \le{}& 2\int_{\mathbb{T}}\left(\int_{\mathbb{R}}(v^2 - D_0)\frac{\sqrt{\chi_1\rho_\infty}}{\sqrt{\chi_1\rho_\infty}}\left(f - \frac{u_h}{\rho_\infty^2+1}\rho_\infty^2\chi_1\right)\mathrm{d}\,v\right)^2\mathrm{d}\,x\\ &+ 2\int_{\mathbb{T}}\left(\int_{\mathbb{R}}(v^2 - D_0)\sqrt{\frac{\rho_\infty}{\chi_2}}\sqrt{\frac{\chi_2}{\rho_\infty}}\left(g + \frac{u_h}{\rho_\infty^2+1}\chi_2\right)\mathrm{d}\,v\right)^2\mathrm{d}\,x,\end{aligned}$$

and (1.19) is obtained after applying Cauchy-Schwarz inequality.

Let us finally turn to the last estimate (1.20). We have

$$\begin{aligned}\|\rho_\infty^{-1}J_f - \rho_\infty J_g\|^2_{L^2(\mathbb{T})} &= \int_{\mathbb{T}}\left(\rho_\infty^{-1}\int_{\mathbb{R}} v f\,\mathrm{d}\,v - \rho_\infty\int_{\mathbb{R}} v g\,\mathrm{d}\,v\right)^2\mathrm{d}\,x\\ &\le 2\max\{\rho_\infty^{-2},\rho_\infty^2\}\int_{\mathbb{T}}\left[\left(\int_{\mathbb{R}} v f\,\mathrm{d}\,v\right)^2 + \left(\int_{\mathbb{R}} v g\,\mathrm{d}\,v\right)^2\right]\mathrm{d}\,x.\end{aligned}$$

The right-hand side is then treated as previously. □

Let us now state the hypocoercivity result for the linearized problem.

Proposition 1. *There are constants $C \geq 1$ and $\kappa > 0$ such that for all initial data $F_I = (f_I, g_I) \in \mathcal{H}$ such that $\int_{\mathbb{T}\times\mathbb{R}}(f_I - g_I)\,\mathrm{d}x\,\mathrm{d}v = 0$, the solution $F = (f, g)$ of (1.9)–(1.10) satisfies*

$$\|F(t)\| \leq C\,\|F_I\|\,e^{-\kappa t}.$$

In order to prove this proposition, we introduce following [73, 171] a modified entropy functional, which is as in [18] a slight simplification of the original version using the fact that we work on a bounded space domain. It reads

$$H_\delta[F] := \frac{1}{2}\|F\|^2 + \delta\,\langle J_h, \partial_x \Phi\rangle_{L^2(\mathbb{T})}, \tag{1.26}$$

where $\Phi(t,x)$ is the solution to the Poisson equation

$$-\partial_{xx}\Phi = u_h, \qquad \int_{\mathbb{T}} \Phi\,\mathrm{d}x = 0, \tag{1.27}$$

and $\delta > 0$ is a small parameter to be chosen later.

Lemma 3. *The function Φ satisfies for all $t \geq 0$*

$$\|\partial_x \Phi(t)\|_{L^2(\mathbb{T})} \leq C_P\,C_u\|\Pi F(t)\|, \tag{1.28}$$

$$\|\partial_{tx} \Phi(t)\|_{L^2(\mathbb{T})} \leq C_{J1}\|(I-\Pi)F(t)\|, \tag{1.29}$$

where C_P is the Poincaré constant of $\mathbb{T}$.

Proof. The first estimate is obtained by multiplying the auxiliary equation (1.27) by Φ, integrating on $\mathbb{T}$, applying the Poincaré inequality and equality (1.16)

$$\|\partial_x \Phi\|^2_{L^2(\mathbb{T})} = \langle -\partial_{xx}\Phi, \Phi\rangle_{L^2(\mathbb{T})} \leq \|u_h\|_{L^2(\mathbb{T})}\|\Phi\|_{L^2(\mathbb{T})} \leq C_u C_P \|\Pi F\|\,\|\partial_x \Phi\|_{L^2(\mathbb{T})}.$$

For the second estimate, we differentiate the auxiliary equation with respect to time and use the continuity equation (1.14) to obtain $-\partial_t\partial_{xx}\Phi = -\partial_x J_h$. Then we multiply by $\partial_t\Phi$ and integrate to get, thanks to (1.18),

$$\begin{aligned}\|\partial_{tx}\Phi(t)\|^2_{L^2(\mathbb{T})} = \langle -\partial_x J_h, \partial_t \Phi\rangle_{L^2(\mathbb{T})} &= \langle J_h, \partial_{tx}\Phi\rangle_{L^2(\mathbb{T})} \\ &\leq C_{J1}\|(I-\Pi)F(t)\|\,\|\partial_{tx}\Phi(t)\|_{L^2(\mathbb{T})}.\end{aligned}$$

□

Thanks to the moments estimates, for small enough $\delta > 0$, the square root of the modified entropy defines an equivalent norm on $\mathcal{H}$, as stated in the following lemma.

Lemma 4 (Equivalent norm)**.** *There is $\delta_1 > 0$ such that for all $\delta \in (0, \delta_1)$, there are positive constants $0 < c_\delta < C_\delta$ such that for all $F \in \mathcal{H}$, one has*

$$c_\delta\|F\|^2 \leq H_\delta[F] \leq C_\delta\|F\|^2.$$

Proof. The result follows from the definition of the modified entropy (1.26) and the estimate

$$|\langle J_h, \partial_x \Phi\rangle_{L^2(\mathbb{T})}| \leq \|J_h\|_{L^2(\mathbb{T})} \|\partial_x \Phi\|_{L^2(\mathbb{T})} \leq C_{J1}\, C_P\, C_u \|F\| \|\Pi F\|.$$

Using the Young inequality and the fact that $\|\Pi F\| \leq \|F\|$, we obtain

$$|\langle J_h, \partial_x \Phi\rangle_{L^2(\mathbb{T})}| \leq C_{J1}\, C_u\, C_P \|F\|^2.$$

□

With the previous lemmas, it is then possible to establish the proof of Proposition 1.

Proof. By using the moment equation (1.15), the time derivative of the modified entropy becomes a sum of five terms,

$$\frac{\mathrm{d}}{\mathrm{d}\,t} H_\delta[F] = T_1 + \delta T_2 + \delta T_3 + \delta T_4 + \delta T_5.$$

Using Lemmas 1, 2 and 3, one has

$$\begin{aligned}
T_1 &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\,t}\|F\|^2 \leq -C_{mc}\|(I-\Pi)F\|^2,\\
T_2 &= -\langle \partial_x S_h, \partial_x \Phi\rangle_{L^2(\mathbb{T})} \leq \left|\langle S_h, \partial_{xx}\Phi\rangle_{L^2(\mathbb{T})}\right| \leq C_S\, C_u \|(I-\Pi)F\| \|\Pi F\|,\\
T_3 &= -D_0\langle \partial_x u_h, \partial_x \Phi\rangle_{L^2(\mathbb{T})} = D_0\langle u_h, \partial_{xx}\Phi\rangle_{L^2(\mathbb{T})} = -D_0\, C_u^2 \|\Pi F\|^2,\\
T_4 &= -\langle \rho_\infty^{-1} J_f - \rho_\infty J_g, \partial_x \Phi\rangle_{L^2(\mathbb{T})} \leq C_{J2}\, C_P\, C_u \|(I-\Pi)F\| \|\Pi F\|,\\
T_5 &= \langle J_h, \partial_{tx}\Phi\rangle_{L^2(\mathbb{T})} \leq C_{J1}^2 \|(I-\Pi)F\|^2.
\end{aligned}$$

Combining these estimates together, one has

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,t} H_\delta[F] + (C_{mc} - \delta C_{J1}^2)\|(I-\Pi)F\|^2 + \delta D_0\, C_u^2 \|\Pi F\|^2 \\
\leq \delta\Big(C_S\, C_u + C_{J2}\, C_P\, C_u\Big)\|(I-\Pi)F\| \|\Pi F\|.
\end{aligned}$$

Then, let us set $\delta \in (0, \min(\delta_1, \delta_2))$ where δ_1 appears in Lemma 4 and $\delta_2 < C_{mc}/C_{J1}^2$ ensures the positivity of $(C_{mc} - \delta C_{J1}^2)$. We also set $\delta < \delta_3$ where

$$\delta_3 = C_{mc} D_0\, C_u^2 \Big((C_S\, C_u + C_{J2}\, C_P\, C_u)^2 + C_{J1}^2 D_0\, C_u^2\Big)^{-1}$$

allows us to obtain:

$$\frac{\mathrm{d}}{\mathrm{d}\,t} H_\delta[F] + K_\delta\Big(\|(I-\Pi)F\|^2 + \|\Pi F\|^2\Big) \leq 0,$$

where $K_\delta = \min(C_{mc} - \delta C_{J1}^2, \delta D_0\, C_u^2)/2$. One can then use that Π is an orthogonal projector and Lemma 4 to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\,t} H_\delta[F] + \frac{K_\delta}{C_\delta} H_\delta[F] \leq 0.$$

This gives the exponential decay of $H_\delta[F]$, which allows to conclude the proof of Proposition 1 by using the lower bound in Lemma 4 and setting $\kappa = \frac{K_\delta}{2C_\delta}$, $C = \sqrt{C_\delta/c_\delta}$. □

1.2.2 Extension to the nonlinear setting

From this exponential decay result for the linearized system (1.9)–(1.10), a local result can be established for the nonlinear case by the same method. It is based on global existence result and maximum principle estimates as stated in [171, Theorem 1.1], and recalled in the following proposition.

Proposition 2 (Global existence result and maximum principle)**.** *Under assumptions* (1.5)*, let* ρ_∞ *be defined by* (1.6)*, and assume that there exist positive constants* $\gamma_1 < \rho_\infty$ *and* γ_2 *such that the initial datum* $\mathsf{F}_I = (\mathfrak{f}_I, \mathfrak{g}_I) \in L^\infty(\mathbb{T}\times\mathbb{R})$ *satisfies for all* $(x,v) \in \mathbb{T}\times\mathbb{R}$

$$(\rho_\infty - \gamma_1)\chi_1(v) \le \mathfrak{f}_I(x,v) \le (\rho_\infty + \gamma_2)\chi_1(v),$$
$$(\rho_\infty + \gamma_2)^{-1}\chi_2(v) \le \mathfrak{g}_I(x,v) \le (\rho_\infty - \gamma_1)^{-1}\chi_2(v).$$

Then the system (1.1)–(1.2) *with initial condition* F_I *admits a unique global mild solution* $\mathsf{F} \in \mathcal{C}([0,\infty), L^\infty(\mathbb{T}\times\mathbb{R}))^2$ *satisfying for all* $(t,x,v) \in [0,\infty)\times\mathbb{T}\times\mathbb{R}$*:*

$$(\rho_\infty - \gamma_1)\chi_1(v) \le \mathfrak{f}(t,x,v) \le (\rho_\infty + \gamma_2)\chi_1(v),$$
$$(\rho_\infty + \gamma_2)^{-1}\chi_2(v) \le \mathfrak{g}(t,x,v) \le (\rho_\infty - \gamma_1)^{-1}\chi_2(v).$$

From this stability result, exponential decay of small perturbations to equilibrium can be shown for the full nonlinear problem.

Theorem 1. *Let* (1.5) *hold and let* F_I *satisfies the assumptions of Proposition 2 with* γ_1 *and* γ_2 *small enough.*

Then the solution F *to* (1.1)–(1.2) *satisfies*

$$\|\mathsf{F}(t) - \mathsf{F}_\infty\| \le C\,\|\mathsf{F}_I - \mathsf{F}_\infty\|\, e^{-\kappa t},$$

with positive constants C *and* κ*.*

Proof. Let us first denote by $Q(\mathfrak{f},\mathfrak{g})$ the nonlinear collision operator given by

$$Q(\mathfrak{f},\mathfrak{g}) = \begin{pmatrix} \chi_1 - \rho_\mathfrak{g}\mathfrak{f} \\ \chi_2 - \rho_\mathfrak{f}\mathfrak{g} \end{pmatrix}. \tag{1.30}$$

Let us also introduce $\widetilde{\mathsf{F}} = \mathsf{F} - \mathsf{F}_\infty$. The nonlinear system (1.1)-(1.2) can now be rewritten in terms of $\widetilde{\mathsf{F}}$. Since the equilibrium F_∞ does not depend on the spatial nor time variables, $\widetilde{\mathsf{F}}$ satisfies

$$\partial_t\widetilde{\mathsf{F}} + v\,\partial_x\widetilde{\mathsf{F}} = \mathsf{L}\widetilde{\mathsf{F}} + \left(Q(\mathfrak{f},\mathfrak{g}) - \mathsf{L}\widetilde{\mathsf{F}}\right).$$

In addition, one can compute the following relation

$$Q(\mathfrak{f},\mathfrak{g}) - \mathsf{L}\widetilde{\mathsf{F}} = \begin{pmatrix} -(\rho_\mathfrak{g} - \rho_\infty^{-1})(\mathfrak{f} - \rho_\infty\chi_1) \\ -(\rho_\mathfrak{f} - \rho_\infty)(\mathfrak{g} - \rho_\infty^{-1}\chi_2) \end{pmatrix}. \tag{1.31}$$

Applying Proposition 2 one has the following bound

$$\left\| Q(\mathfrak{f},\mathfrak{g}) - \mathsf{L}\widetilde{\mathsf{F}} \right\| \leq \gamma \|\widetilde{\mathsf{F}}\|.$$

Consequently, the estimate on the entropy becomes

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\,t}\|\widetilde{\mathsf{F}}(t)\|^2 + C_{mc}\|(I-\Pi)\widetilde{\mathsf{F}}(t)\|^2 - \gamma\|\widetilde{\mathsf{F}}\|^2 \leq 0.$$

This ensures that for γ (related to γ_1 and γ_2) small enough, the entropy dissipation of the linearized system given in Proposition 1 is enough to overcome the nonlinear dynamics. □

1.3 The discrete setting

In this section, we present numerical schemes for systems (1.1)–(1.2) and (1.9)–(1.10). The schemes are implicit in time and of finite volume type in the (x,v)-phase space.

1.3.1 Notations

Mesh

Since it is in practice not possible to implement a numerical scheme on an unbounded domain, we first have to restrict the velocity domain to a bounded symmetric segment $[-v^*, v^*]$. We consider a mesh of this interval composed of $2L$ control volumes arranged symmetrically around $v = 0$. We denote $v_{j+\frac{1}{2}}$ the $2L+1$ interface points, with $j \in \mathcal{J} := \{-L, \cdots, L\}$. In this way,

$$v_{-L+\frac{1}{2}} = -v^*, \quad v_{\frac{1}{2}} = 0, \quad v_{j+\frac{1}{2}} = -v_{j-\frac{1}{2}} \qquad \forall j = 0, \cdots, L.$$

For the sake of simplicity, we assume that the velocity mesh is uniform, namely that every cell $\mathcal{V}_j = (v_{j-\frac{1}{2}}, v_{j+\frac{1}{2}})$ has constant length Δv. Denoting v_j the midpoint of the cell $\mathcal{V}_j$, we also have $v_j = -v_{-j+1}$ for all $j = 1, \cdots, L$.

In space, we consider a uniform discretization of the torus $\mathbb{T}$ into N cells

$$\mathcal{X}_i := (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}), \quad i \in \mathcal{I} := \mathbb{Z}/N\mathbb{Z}$$

of length Δx. Once again, we assume uniformity of the mesh for simplicity's sake, but our results generalize to a non-uniform setting. As in [18], we have to impose that N is odd in order to have, among others, a discrete Poincaré inequality on the torus.

The control volumes in phase space are defined by

$$K_{ij} := \mathcal{X}_i \times \mathcal{V}_j, \qquad \forall (i,j) \in \mathcal{I} \times \mathcal{J}.$$

The size of this phase space discretization is defined by $\Delta = (\Delta x, \Delta v)$. Finally, we set $\Delta t > 0$ the time step, and $t^n = n\Delta t$ for all $n \geq 0$.

Discrete velocity profiles

For $k = 1, 2$, we assume that we are given cell values $(\chi_{k,j})_{j\in\mathcal{J}} \in \mathbb{R}^{\mathcal{J}}$ such that the following assumptions are fulfilled

$$\begin{aligned} &\chi_{k,j} > 0, \quad \chi_{k,j} = \chi_{k,-j+1} \qquad \forall j = 1, \cdots, L, \\ &\sum_{j\in\mathcal{J}} \Delta v\, \chi_{k,j} = 1, \\ &0 < \underline{D}_k \le D_k^\Delta \le \overline{D}_k < \infty, \qquad Q_k^\Delta \le \overline{Q}_k < \infty, \end{aligned} \tag{1.32}$$

where

$$D_k^\Delta := \sum_{j\in\mathcal{J}} \Delta v\, |v_j|^2\, \chi_{k,j}, \qquad Q_k^\Delta := \sum_{j\in\mathcal{J}} \Delta v\, |v_j|^4\, \chi_{k,j},$$

and $\underline{D}_k$, $\overline{D}_k$, $\overline{Q}_k$ are universal constants. Typically, we define $\chi_{k,j} = c_{\Delta v}\chi_k(v_j)$ and compute $c_{\Delta v}$ in such a way that the mass of $(\chi_{k,j})_j$ is 1. Note also that the symmetry properties imply

$$\sum_{j\in\mathcal{J}} \Delta v\, v_j\, \chi_{k,j} = 0. \tag{1.33}$$

Discrete gradients and functional setting in space

Due to our choices of discretization, we need to define several discrete gradients in space. Given a macroscopic quantity $u = (u_i)_{i\in\mathcal{I}}$, we define

— the discrete centered gradient $D_x^c u \in \mathbb{R}^{\mathcal{I}}$ given by

$$(D_x^c u)_i = \frac{u_{i+1} - u_{i-1}}{2\Delta x} \qquad \forall i \in \mathcal{I},$$

— the discrete downstream gradient $D_x^- u \in \mathbb{R}^{\mathcal{I}}$ given by

$$(D_x^- u)_i = \frac{u_i - u_{i-1}}{\Delta x} \qquad \forall i \in \mathcal{I},$$

— the discrete upstream gradient $D_x^+ u \in \mathbb{R}^{\mathcal{I}}$ given by

$$(D_x^+ u)_i = \frac{u_{i+1} - u_i}{\Delta x} \qquad \forall i \in \mathcal{I}.$$

It is straightforward to see that these discrete gradients satisfy the following properties:

$$\frac{1}{2}(D_x^- + D_x^+) = D_x^c, \qquad D_x^+ D_x^- = D_x^- D_x^+. \tag{1.34}$$

For $u_k = (u_{k,i})_{i\in\mathcal{I}}$, $k = 1, 2$, we define the discrete L^2 scalar product by

$$\langle u_1, u_2\rangle_2 := \sum_{i\in\mathcal{I}} \Delta x\, u_{1,i} u_{2,i}, \tag{1.35}$$

and denote $\|\cdot\|_2$ the corresponding norm.

Using the definition of the discrete gradients and the periodic boundary conditions, we imme-

diately have the following algebraic properties.

Lemma 5. *For all $u = (u_i)_{i\in\mathcal{I}}$, $\bar{u} = (\bar{u}_i)_{i\in\mathcal{I}}$, it holds*

$$\langle D_x^c u, \bar{u}\rangle_2 = -\langle u, D_x^c \bar{u}\rangle_2, \tag{1.36}$$

$$\langle D_x^+ u, \bar{u}\rangle_2 = -\langle u, D_x^- \bar{u}\rangle_2, \tag{1.37}$$

$$\langle (D_x^+ D_x^- + D_x^- D_x^+)u, \bar{u}\rangle_2 = -4\langle D_x^c u, D_x^c \bar{u}\rangle_2, \tag{1.38}$$

$$\Delta x \|D_x^c u\|_2 \le \|u\|_2. \tag{1.39}$$

Let us finally recall the discrete Poincaré inequality on the torus (see for example [18, Lemma 6] for a proof of this result).

Lemma 6 (Discrete Poincaré inequality on the torus)**.** *Assume that the number of points N in the space discretization of the torus is odd. Then, there is a constant $C_P > 0$ independent on Δx such that for all $u = (u_i)_{i\in\mathcal{I}}$ satisfying $\sum_{i\in\mathcal{I}} \Delta x\, u_i = 0$,*

$$\|u\|_2 \le C_P \|D_x^c u\|_2.$$

Note that in the above lemma, the constant C_P converges towards $\frac{1}{\pi}$ while in the continuous case the Poincaré constant equals $\frac{1}{2\pi}$. For the sake of simplicity, we use the same notation. Actually, the difference originates from the choice of the discrete gradient, and we refer to the discussion on the matter in [18] for further details.

1.3.2 Definition of the numerical scheme for the nonlinear system

The numerical scheme for approximating (1.1)–(1.2) is based on a finite volume discretization in phase space and backward Euler discretization in time. The initial datum $\mathsf{F}_I = (\mathfrak{f}_I, \mathfrak{g}_I)$ is discretized by

$$\mathfrak{f}_{ij}^0 = \frac{1}{\Delta x\,\Delta v}\int_{K_{ij}} \mathfrak{f}_I(x,v)\,\mathrm{d}x\,\mathrm{d}v, \quad \mathfrak{g}_{ij}^0 = \frac{1}{\Delta x\,\Delta v}\int_{K_{ij}} \mathfrak{g}_I(x,v)\,\mathrm{d}x\,\mathrm{d}v \qquad \forall (i,j)\in\mathcal{I}\times\mathcal{J}.$$

Then, by integrating (1.1)–(1.2) on each cell K_{ij}, the following numerical scheme is obtained: for all $n \ge 0$, $i \in \mathcal{I}$, $j \in \mathcal{J}$,

$$\frac{\mathfrak{f}_{ij}^{n+1} - \mathfrak{f}_{ij}^n}{\Delta t} + \frac{1}{\Delta x\,\Delta v}\left(\mathcal{F}_{i+\frac12,j}^{n+1} - \mathcal{F}_{i-\frac12,j}^{n+1}\right) = \chi_{1,j} - \rho_{\mathfrak{g},i}^{n+1}\mathfrak{f}_{ij}^{n+1}, \tag{1.40}$$

$$\frac{\mathfrak{g}_{ij}^{n+1} - \mathfrak{g}_{ij}^n}{\Delta t} + \frac{1}{\Delta x\,\Delta v}\left(\mathcal{G}_{i+\frac12,j}^{n+1} - \mathcal{G}_{i-\frac12,j}^{n+1}\right) = \chi_{2,j} - \rho_{\mathfrak{f},i}^{n+1}\mathfrak{g}_{ij}^{n+1}, \tag{1.41}$$

with the so-called Lax-Friedrichs fluxes

$$\mathcal{F}_{i+\frac12,j}^{n+1} = \Delta v\,\frac{v_j}{2}(\mathfrak{f}_{i+1,j}^{n+1} + \mathfrak{f}_{ij}^{n+1}) - \Delta v\,\lambda(\mathfrak{f}_{i+1,j}^{n+1} - \mathfrak{f}_{ij}^{n+1}), \tag{1.42}$$

$$\mathcal{G}_{i+\frac12,j}^{n+1} = \Delta v\,\frac{v_j}{2}(\mathfrak{g}_{i+1,j}^{n+1} + \mathfrak{g}_{ij}^{n+1}) - \Delta v\,\lambda(\mathfrak{g}_{i+1,j}^{n+1} - \mathfrak{g}_{ij}^{n+1}), \tag{1.43}$$

where $\lambda = \Delta x/2\,\Delta t$ is assumed to be a fixed constant. For all $n \ge 0$ and $i \in \mathcal{I}$, the discrete macroscopic

densities are given by

$$\rho^n_{\mathfrak{f},i} = \sum_{j\in\mathcal{J}} \Delta v\, \mathfrak{f}^n_{ij}, \qquad \rho^n_{\mathfrak{g},i} = \sum_{j\in\mathcal{J}} \Delta v\, \mathfrak{g}^n_{ij}. \tag{1.44}$$

Remark that the scheme (1.40)–(1.41) clearly satisfies the discrete mass conservation of $\mathfrak{f}-\mathfrak{g}$:

$$\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x\,\Delta v\,(\mathfrak{f}^n_{ij}-\mathfrak{g}^n_{ij}) = \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x\,\Delta v\,(\mathfrak{f}^0_{ij}-\mathfrak{g}^0_{ij}) \quad \forall n\geq 0.$$

Then, we define $\rho^*_\infty > 0$ as the unique constant such that

$$M_0 := \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x\,\Delta v\,(\mathfrak{f}^0_{ij}-\mathfrak{g}^0_{ij}) = |\mathbb{T}|\left(\rho^*_\infty - \frac{1}{\rho^*_\infty}\right). \tag{1.45}$$

In particular, we take,

$$\rho^*_\infty = \frac{M_0+\sqrt{M_0^2+4|\mathbb{T}|}}{2|\mathbb{T}|}. \tag{1.46}$$

It is clear that $\mathfrak{F}^\infty = (\mathfrak{f}^\infty, \mathfrak{g}^\infty)$ defined by

$$\mathfrak{f}^\infty_{ij} = \rho^*_\infty \chi_{1,j}, \qquad \mathfrak{g}^\infty_{ij} = \frac{1}{\rho^*_\infty}\chi_{2,j} \qquad \forall (i,j)\in\mathcal{I}\times\mathcal{J} \tag{1.47}$$

is an equilibrium for the scheme (1.40)–(1.41). As in the continuous framework, the study of the exponential convergence of the approximate solutions to this discrete equilibrium is done by analyzing the discretization of the linearized problem that we now introduce.

Remark 1. *Note that one could choose to use many types of numerical fluxes in* (1.40)–(1.41)*, such as the classical central fluxes (as was done in [18]) given by*

$$\mathcal{F}^{n+1,C}_{i+\frac12,j} = \Delta v\,\frac{v_j}{2}(\mathfrak{f}^{n+1}_{i+1,j}+\mathfrak{f}^{n+1}_{ij}), \tag{1.48}$$

$$\mathcal{G}^{n+1,C}_{i+\frac12,j} = \Delta v\,\frac{v_j}{2}(\mathfrak{g}^{n+1}_{i+1,j}+\mathfrak{g}^{n+1}_{ij}), \tag{1.49}$$

or the upwind fluxes, defined for $a^+ = \max(a,0)$ and $a^- = -\min(a,0)$ by

$$\mathcal{F}^{n+1,U}_{i+\frac12,j} = \Delta v\left(v_j^+\,\mathfrak{f}^{n+1}_{i,j} - v_j^-\,\mathfrak{f}^{n+1}_{i+1,j}\right), \tag{1.50}$$

$$\mathcal{G}^{n+1,U}_{i+\frac12,j} = \Delta v\left(v_j^+\,\mathfrak{g}^{n+1}_{i,j} - v_j^-\,\mathfrak{g}^{n+1}_{i+1,j}\right). \tag{1.51}$$

The use of the seemingly more complicated Lax-Friedrichs fluxes stems from the fact that central fluxes are not monotone in the sense of Crandall and Majda [54]. This property ensures a maximum principle of the nonlinear scheme, which we will see later is needed for our main result. Upwind fluxes do enjoy such monotonicity properties, but complicates too much the analysis.

We will see in Section 1.6 that the three choices yield similar numerical results in the linear setting, where monotonicity is not mandatory for the hypocoercive behavior of our method. It will nevertheless be crucial in the nonlinear case, where exponential decay do not occur with central fluxes.

1.3.3 Definition of the numerical scheme for the linearized problem

As in the continuous framework, the perturbations are again denoted by f and g. By integrating the linearized system (1.9)–(1.10) on each cell K_{ij}, the following numerical scheme is obtained: for all $n \geq 0$, $i \in \mathcal{I}$, $j \in \mathcal{J}$,

$$\frac{f_{ij}^{n+1} - f_{ij}^n}{\Delta t} + \frac{1}{\Delta x \Delta v}\left(\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1}\right) = -\rho_\infty^* \chi_{1,j} \rho_{g,i}^{n+1} - (\rho_\infty^*)^{-1} f_{ij}^{n+1}, \tag{1.52}$$

$$\frac{g_{ij}^{n+1} - g_{ij}^n}{\Delta t} + \frac{1}{\Delta x \Delta v}\left(\mathcal{G}_{i+\frac{1}{2},j}^{n+1} - \mathcal{G}_{i-\frac{1}{2},j}^{n+1}\right) = -(\rho_\infty^*)^{-1} \chi_{2,j} \rho_{f,i}^{n+1} - \rho_\infty^* g_{ij}^{n+1}, \tag{1.53}$$

where the numerical fluxes are still defined by (1.42)–(1.43) but replacing $\mathfrak{f}$ by f and $\mathfrak{g}$ by g, $\lambda = \Delta x / 2\Delta t$ is assumed to be a fixed constant, and the discrete macroscopic densities are given by (1.44).

For future use, we also need to define the discrete velocity moments of $h = (h_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$ given by $h_{ij}^n = f_{ij}^n - g_{ij}^n$: for all $n \geq 0$, $i \in \mathcal{I}$,

$$u_{h,i}^n := \sum_{j\in\mathcal{J}} \Delta v\, h_{ij}^n, \quad J_{h,i}^n := \sum_{j\in\mathcal{J}} \Delta v\, v_j\, h_{ij}^n, \quad S_{h,i}^n := \sum_{j\in\mathcal{J}} \Delta v\, (v_j^2 - D_0^\Delta)\, h_{ij}^n, \tag{1.54}$$

with

$$D_0^\Delta = \frac{(\rho_\infty^*)^2 D_1^\Delta + D_2^\Delta}{(\rho_\infty^*)^2 + 1}. \tag{1.55}$$

Note in particular that one obtains from (1.32) the existence of some universal constants $\underline{D_0}$ and $\overline{D_0}$ such that $\underline{D_0} \leq D_0^\Delta \leq \overline{D_0}$.

1.4 Numerical hypocoercivity for the linearized problem

We now adapt the study of the linearized system (1.9)–(1.10) to the discrete setting, following the hypocoercivity method proposed in [73] and briefly described in Section 1.2. In order to estimate the decay towards the equilibrium, we introduce the following weighted scalar product: for microscopic quantities $F_k = \begin{pmatrix} f_{k,ij} \\ g_{k,ij} \end{pmatrix}_{i\in\mathcal{I}, j\in\mathcal{J}}$, $k = 1, 2$,

$$\langle F_1, F_2 \rangle_\Delta := \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \Delta x \Delta v \left(\frac{f_{1,ij} f_{2,ij}}{\chi_{1,j} \rho_\infty^*} + \frac{g_{1,ij} g_{2,ij} \rho_\infty^*}{\chi_{2,j}} \right). \tag{1.56}$$

We denote by $\|\cdot\|_\Delta$ the corresponding norm.

For $(F_{ij})_{ij} = (f_{ij}, g_{ij})_{ij}$, we also define the discrete counterpart of the linear collision operator

$$(\mathsf{L}^\Delta F)_{ij} := \begin{pmatrix} -\rho_\infty^* \chi_{1,j} \rho_{g,i} - (\rho_\infty^*)^{-1} f_{ij} \\ -(\rho_\infty^*)^{-1} \chi_{2,j} \rho_{f,i} - \rho_\infty^* g_{ij} \end{pmatrix}, \tag{1.57}$$

and the orthogonal projection onto its null space:

$$(\Pi^{\Delta} F)_{ij} := \frac{\rho_{f,i} - \rho_{g,i}}{(\rho_\infty^*)^2 + 1} \begin{pmatrix} (\rho_\infty^*)^2 \chi_{1,j} \\ -\chi_{2,j} \end{pmatrix}. \tag{1.58}$$

In this section, we derive the discrete counterparts of estimates given in Lemmas 1 and 2.

Lemma 7 (Discrete microscopic coercivity). *Let (1.32) hold and let $(f_{ij}^n, g_{ij}^n)_{n\geq 0, i\in\mathcal{I}, j\in\mathcal{J}}$ solves the scheme (1.52)–(1.53). Then for all $n \geq 0$,*

$$\frac{1}{2}\left(\|F^{n+1}\|_\Delta^2 - \|F^n\|_\Delta^2\right) + \Delta t\, C_{mc}^* \|(I - \Pi^\Delta) F^{n+1}\|_\Delta^2 \leq 0, \tag{1.59}$$

where $C_{mc}^* = \min((\rho_\infty^*)^{-1}, \rho_\infty^*)$.

Proof. We multiply (1.52) by $f_{ij}^{n+1}/(\rho_\infty^* \chi_{1,j})$ and (1.53) by $g_{ij}^{n+1}\rho_\infty^*/\chi_{2,j}$, and then sum the two resulting expressions and sum over $(i,j) \in \mathcal{I}\times\mathcal{J}$. The obtained expression is of the form $A_1 + A_2 = A_3$ with

$$\begin{aligned}
A_1 &= \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x \Delta v \left((f_{ij}^{n+1} - f_{ij}^n) \frac{f_{ij}^{n+1}}{\rho_\infty^* \chi_{1,j}} + (g_{ij}^{n+1} - g_{ij}^n) \frac{g_{ij}^{n+1}\rho_\infty^*}{\chi_{2,j}} \right), \\
A_2 &= \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \left(\left(\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1} \right) \frac{f_{ij}^{n+1}}{\rho_\infty^* \chi_{1,j}} + \left(\mathcal{G}_{i+\frac{1}{2},j}^{n+1} - \mathcal{G}_{i-\frac{1}{2},j}^{n+1} \right) \frac{g_{ij}^{n+1}\rho_\infty^*}{\chi_{2,j}} \right), \\
A_3 &= \langle \mathsf{L}^\Delta F^{n+1}, F^{n+1} \rangle_\Delta.
\end{aligned}$$

Let us first deal with the transport part A_2. By definition of the numerical flux (1.42):

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \left(\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1} \right) \frac{f_{ij}^{n+1}}{\rho_\infty^* \chi_{1,j}} \\
&= \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \frac{v_j \Delta v}{2} \left(f_{i+1,j}^{n+1} f_{ij}^{n+1} - f_{ij}^{n+1} f_{i-1,j}^{n+1} \right) \frac{1}{\rho_\infty^* \chi_{1,j}} \\
&\quad - \lambda \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x^2 \left(D_x^+ D_x^- f_j^{n+1} \right)_i \frac{f_{ij}^{n+1}}{\rho_\infty^* \chi_{1,j}} \Delta v.
\end{aligned}$$

Due to the spatial periodic boundary conditions, the first term on the right-hand side vanishes. Then, using the properties of the discrete gradients (1.34) and a discrete integration by parts (1.38) on the second term, it yields

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \left(\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1} \right) \frac{f_{ij}^{n+1}}{\rho_\infty^* \chi_{1,j}} \\
&= -\lambda \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \frac{\Delta x^2}{2} (D_x^+ D_x^- + D_x^- D_x^+) \left(f_j^{n+1} \right)_i \frac{f_{ij}^{n+1}}{\rho_\infty^* \chi_{1,j}} \Delta v \\
&= 2\lambda \Delta x \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \left(D_x^c f_j^{n+1} \right)_i^2 \frac{\Delta x \Delta v}{\rho_\infty^* \chi_{1,j}}.
\end{aligned}$$

Therefore,

$$\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}\left(\mathcal{F}^{n+1}_{i+\frac{1}{2},j}-\mathcal{F}^{n+1}_{i-\frac{1}{2},j}\right)\frac{f^{n+1}_{ij}}{\rho^*_\infty\chi_{1,j}}\geq 0.$$

The same computations applied to (1.43) yield

$$\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}\left(\mathcal{G}^{n+1}_{i+\frac{1}{2},j}-\mathcal{G}^{n+1}_{i-\frac{1}{2},j}\right)\frac{g^{n+1}_{ij}\rho^*_\infty}{\chi_{2,j}}\geq 0.$$

Consequently, one has

$$A_2\geq 0 \tag{1.60}$$

Note that the choice of the Lax-Friedrichs numerical flux leads to an inequality where in previous works, the use of a centered scheme yielded exactly 0. This was expected since the Lax-Friedrichs scheme classically introduces numerical diffusion in system. Then, using the same computations as in the continuous case, one can obtain from A_3 a discrete counterpart of (1.11):

$$A_3=\langle \mathsf{L}^\Delta F^{n+1},F^{n+1}\rangle_\Delta\leq -C^*_{mc}\|(I-\Pi^\Delta)F^{n+1}\|^2_\Delta, \tag{1.61}$$

where we set $C^*_{mc}=\min(\rho^*_\infty,(\rho^*_\infty)^{-1})$. The term A_1 can then be bounded from below using the relation $(a^2-b^2)/2\leq a(a-b)$, yielding

$$\frac{1}{2}\left(\|F^{n+1}\|^2_\Delta-\|F^n\|^2_\Delta\right)\leq A_1. \tag{1.62}$$

Finally, combining (1.60), (1.61) and (1.62) we obtain the desired estimate:

$$\frac{1}{2}\left(\|F^{n+1}\|^2_\Delta-\|F^n\|^2_\Delta\right)+\Delta t\,C^*_{mc}\|(I-\Pi^\Delta)F^{n+1}\|^2_\Delta\leq 0.$$

□

From this lemma, we can deduce the uniqueness and then the existence of a solution to the scheme (1.52)–(1.53) since it is a finite dimensional linear system.

Corollary 1. *The scheme* (1.52)–(1.53) *admits a unique solution* $(F^n_{ij})_{n\geq 0,i\in\mathcal{I},j\in\mathcal{J}}$.

Let us now give the discrete counterpart of Lemma 2, that is uniform $L^2(\mathbb{T})$ bounds on the discrete moments.

Lemma 8 (Discrete moments estimates)**.** *Under the assumptions of Lemma 7, the discrete moments* $(u^n_{h,i})_{i\in\mathcal{I}}$, $(J^n_{h,i})_{i\in\mathcal{I}}$ *and* $(S^n_{h,i})_{i\in\mathcal{I}}$ *satisfy the following estimates for all* $n\geq 0$*:*

$$\|u^n_h\|_2=C^*_u\|\Pi^\Delta F^n\|_\Delta, \tag{1.63}$$

$$\|J^n_h\|_2\leq C^*_{J1}\|F^n\|_\Delta, \tag{1.64}$$

$$\|J^n_h\|_2\leq C^*_{J1}\|(I-\Pi^\Delta)F^n\|_\Delta, \tag{1.65}$$

$$\|S^n_h\|_2\leq C^*_S\|(I-\Pi^\Delta)F^n\|_\Delta, \tag{1.66}$$

$$\|(\rho^*_\infty)^{-1}J^n_f-\rho^*_\infty J^n_g\|_2\leq C^*_{J2}\|(I-\Pi^\Delta)F^n\|_\Delta, \tag{1.67}$$

where the constants are given by:

$$C_u^* = \sqrt{\frac{(\rho_\infty^*)^2+1}{\rho_\infty^*}}, \tag{1.68}$$

$$C_{J1}^* = \sqrt{2\max(\rho_\infty^* \overline{D_1}, (\rho_\infty^*)^{-1}\overline{D_2})}, \tag{1.69}$$

$$C_S^* = \sqrt{2\max(\rho_\infty^*(\overline{Q_1} - 2\underline{D_0}\,\underline{D_1} + \overline{D_0}^2), (\rho_\infty^*)^{-1}(\overline{Q_2} - 2\underline{D_0}\,\underline{D_2} + \overline{D_0}^2))}, \tag{1.70}$$

$$C_{J2}^* = \max((\rho_\infty^*)^{-1}, \rho_\infty^*) C_{J1}^*. \tag{1.71}$$

Proof. The strategy is very similar to the proof in continuous case. Estimate (1.63) is directly obtained from the definition of the discrete projector Π^Δ:

$$\begin{aligned}\|\Pi^\Delta F^n\|_\Delta^2 &= \frac{1}{((\rho_\infty^*)^2+1)^2} \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \left[(\rho_{f,i}^n - \rho_{g,i}^n)^2 \frac{(\rho_\infty^*)^4 \chi_{1,j}^2}{\chi_{1,j}\rho_\infty^*} + (\rho_{f,i}^n - \rho_{g,i}^n)^2 \frac{\chi_{2,j}^2 \rho_\infty^*}{\chi_{2,j}} \right] \Delta v \Delta x \\ &= \frac{\rho_\infty^*}{(\rho_\infty^*)^2+1} \|u_h^n\|_2^2.\end{aligned}$$

The remaining estimates rely on the discrete properties of unit mass (1.32) and null odd-moments (1.33) of the discrete velocity profiles. Estimates (1.64) and (1.65) are obtained noticing that

$$\|J_h^n\|_2^2 \le 2 \sum_{i\in\mathcal{I}} \left[\left(\sum_{j\in\mathcal{J}} v_j \frac{\sqrt{\chi_{1,j}\rho_\infty^*}}{\sqrt{\chi_{1,j}\rho_\infty^*}} f_{ij}^n \Delta v \right)^2 + \left(\sum_{j\in\mathcal{J}} v_j \frac{\sqrt{\chi_{2,j}}}{\sqrt{\rho_\infty^*}} \sqrt{\frac{\rho_\infty^*}{\chi_{2,j}}} g_{ij}^n \Delta v \right)^2 \right] \Delta x$$

and

$$\begin{aligned}\|J_h^n\|_2^2 \le 2 \sum_{i\in\mathcal{I}} \Bigg[&\left(\sum_{j\in\mathcal{J}} v_j \frac{\sqrt{\chi_{1,j}\rho_\infty^*}}{\sqrt{\chi_{1,j}\rho_\infty^*}} \left(f_{ij}^n - \frac{u_{h,i}^n}{(\rho_\infty^*)^2+1} (\rho_\infty^*)^2 \chi_{1,j} \right) \Delta v \right)^2 \\ &+ \left(\sum_{j\in\mathcal{J}} v_j \frac{\sqrt{\chi_{2,j}}}{\sqrt{\rho_\infty^*}} \frac{\sqrt{\rho_\infty^*}}{\sqrt{\chi_{2,j}}} \left(g_{ij}^n + \frac{u_{h,i}^n}{(\rho_\infty^*)^2+1} \chi_{2,j} \right) \Delta v \right)^2 \Bigg] \Delta x.\end{aligned}$$

Using definition (1.55) of D_0^Δ, one has for all $i \in \mathcal{I}$

$$\sum_{j\in\mathcal{J}} (v_j^2 - D_0^\Delta) \frac{u_{h,i}^n}{(\rho_\infty^*)^2+1} \left((\rho_\infty^*)^2 \chi_{1,j} + \chi_{2,j} \right) \Delta v = 0.$$

Estimate (1.66) is then obtained from

$$\begin{aligned}\|S_h^n\|_2^2 \le 2 \sum_{i\in\mathcal{I}} \Bigg[&\left(\sum_{j\in\mathcal{J}} (v_j^2 - D_0^\Delta) \frac{\sqrt{\chi_{1,j}\rho_\infty^*}}{\sqrt{\chi_{1,j}\rho_\infty^*}} \left(f_{ij}^n - \frac{u_{h,i}^n}{(\rho_\infty^*)^2+1} (\rho_\infty^*)^2 \chi_{1,j} \right) \Delta v \right)^2 \\ &+ \left(\sum_{j\in\mathcal{J}} (v_j^2 - D_0^\Delta) \sqrt{\frac{\chi_{2,j}}{\rho_\infty^*}} \sqrt{\frac{\rho_\infty^*}{\chi_{2,j}}} \left(g_{ij}^n + \frac{u_{h,i}^n}{(\rho_\infty^*)^2+1} \chi_{2,j} \right) \Delta v \right)^2 \Bigg] \Delta x.\end{aligned}$$

One can then apply Cauchy-Schwarz inequalities and the boundedness of the moments of the

velocity profiles to obtain the result. Finally, the last estimate (1.67) is obtained by following exactly the same reasoning as in the continuous case to obtain (1.20). □

Lemma 9 (Moments schemes)**.** *Under the assumptions of Lemma 7, the discrete moments satisfy the following equations: for all $i \in \mathcal{I}$, $n \geq 0$,*

$$\frac{u_{h,i}^{n+1} - u_{h,i}^{n}}{\Delta t} + (D_x^c J_h^{n+1})_i - \frac{\Delta x\, \lambda}{2}\Big((D_x^+ D_x^- + D_x^- D_x^+) u_h^{n+1}\Big)_i = 0, \tag{1.72}$$

$$\frac{J_{h,i}^{n+1} - J_{h,i}^{n}}{\Delta t} + (D_x^c S_h^{n+1})_i + D_0^{\Delta}(D_x^c u_h^{n+1})_i - \frac{\Delta x\, \lambda}{2}\Big((D_x^+ D_x^- + D_x^- D_x^+) J_h^{n+1}\Big)_i = - ((\rho_\infty^*)^{-1} J_{f,i}^{n+1} - \rho_\infty^* J_{g,i}^{n+1}). \tag{1.73}$$

Proof. Let us start by subtracting (1.53) from (1.52). By linearity of the numerical scheme, we obtain

$$\frac{h_{ij}^{n+1} - h_{ij}^{n}}{\Delta t} + \frac{1}{\Delta x \Delta v}\Big(\mathcal{F}_{i+\frac12,j}^{n+1} - \mathcal{F}_{i-\frac12,j}^{n+1} - (\mathcal{G}_{i+\frac12,j}^{n+1} - \mathcal{G}_{i-\frac12,j}^{n+1})\Big) = \rho_\infty^*(g_{ij}^{n+1} - \chi_{1,j}\rho_{g,i}^{n+1}) - (\rho_\infty^*)^{-1}(f_{ij}^{n+1} - \chi_{2,j}\rho_{f,i}^{n+1}).$$

Using the definition of the numerical fluxes (1.42)–(1.43) and the properties of the discrete gradients (1.34), one gets

$$\frac{h_{ij}^{n+1} - h_{ij}^{n}}{\Delta t} + v_j (D_x^c h_j^{n+1})_i - \frac{\lambda \Delta x}{2}\Big((D_x^+ D_x^- + D_x^- D_x^+) h_j^{n+1}\Big)_i = \rho_\infty^*(g_{ij}^{n+1} - \chi_{1,j}\rho_{g,i}^{n+1}) - (\rho_\infty^*)^{-1}(f_{ij}^{n+1} - \chi_{2,j}\rho_{f,i}^{n+1}). \tag{1.74}$$

Finally, the moment schemes are obtained by multiplying (1.74) by $(\Delta v, v_j \Delta v)$, summing over $j \in \mathcal{J}$ and applying definitions (1.54) of the discrete moments. □

Thanks to these three lemmas, we are now in position to establish the discrete counterpart of Proposition 1, namely the exponential decay to equilibrium for the discrete linearized problem.

Theorem 2. *Assuming that (1.32) is fulfilled and that the number of points N of the spatial discretization is odd, there exist constants $C \geq 1$ and $\kappa > 0$ such that for all $\Delta t \leq \Delta t_{\max}$ and all initial data $F^0 = (f_{ij}^0, g_{ij}^0)_{(i,j)\in\mathcal{I}\times\mathcal{J}}$ such that $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x \Delta v\, (f_{ij}^0 - g_{ij}^0) = 0$, the solution $F^n = (f_{ij}^n, g_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$ of (1.52)–(1.53) satisfies for all $n \geq 0$*

$$\|F^{n+1}\|_\Delta \leq C \|F^0\|_\Delta\, e^{-\kappa t^n}. \tag{1.75}$$

The constants C and κ do not depend on the size of the discretization Δ, and $\Delta t_{\max}$ can be arbitrarily chosen.

To prove this result, let us introduce the discrete modified entropy functional

$$H_\delta^\Delta[F^n] := \frac12 \|F^n\|_\Delta^2 + \delta \langle J_h^n, D_x^c \Phi^n\rangle_2 + \frac{\delta}{2\Delta t} \sum_{i\in\mathcal{I}} \Delta x \Big((D_x^c \Phi^n)_i - (D_x^c \Phi^{n-1})_i\Big)^2, \tag{1.76}$$

where $\delta > 0$ will be determined later and $(\Phi_i^n)_{i\in\mathcal{I}}$ is the solution to the following discrete Poisson

equation:

$$(D_x^c D_x^c \Phi^n)_i = -u_{h,i}^n \quad \forall i \in \mathcal{I}, \qquad \sum_{i\in\mathcal{I}} \Delta x \, \Phi_i^n = 0. \tag{1.77}$$

For an odd number of points N of the spatial discretization, existence and uniqueness of $(\Phi_i^n)_{i\in\mathcal{I}}$ satisfying (1.77) is obtained (see [18]).

Let us now derive some discrete estimates on $(\Phi_i^n)_{i\in\mathcal{I}}$.

Lemma 10. *Under the assumptions of Theorem 2, one has for all $n \geq 0$*

$$\|D_x^c \Phi^n\|_2 \leq C_P \, C_u^* \, \|\Pi^\Delta F^n\|_\Delta, \tag{1.78}$$

$$\|D_x^c \Phi^{n+1} - D_x^c \Phi^n\|_2 \leq \Delta t \, \|J_h^{n+1}\|_2 + 2\,\Delta t \, \lambda \, C_u^* \, \|\Pi^\Delta F^{n+1}\|_\Delta, \tag{1.79}$$

where C_P is the discrete Poincaré constant of Lemma 6.

Proof. To show (1.78), we multiply (1.77) by $\Phi_i^n \Delta x$ and sum over $i \in \mathcal{I}$. Then, discrete integration by parts together with Lemmas 6 and 8 are used to obtain:

$$\|D_x^c \Phi^n\|_2^2 = \langle -D_x^c D_x^c \Phi^n, \Phi^n \rangle_2 = \langle u_h^n, \Phi^n \rangle_2 \leq \|u_h^n\|_2 \|\Phi_n\|_2 \leq C_P C_u^* \|\Pi^\Delta F^n\|_\Delta \|D_x^c \Phi_n\|_2.$$

Let us now prove (1.79). The first step is to subtract (1.77) at time t^n from (1.77) at time t^{n+1}, multiply by $(\Phi_i^{n+1} - \Phi_i^n)\Delta x$ and sum over $i \in \mathcal{I}$:

$$\langle D_x^c D_x^c \Phi^{n+1} - D_x^c D_x^c \Phi^n, \Phi^{n+1} - \Phi^n \rangle_2 = -\langle u_h^{n+1} - u_h^n, \Phi^{n+1} - \Phi^n \rangle_2.$$

After an integration by parts on the left-hand side and plugging the continuity scheme (1.72) in the right-hand side, one has

$$\begin{aligned} \|D_x^c \Phi^{n+1} - D_x^c \Phi^n\|_2^2 &= -\Delta t \langle D_x^c J_h^{n+1}, \Phi^{n+1} - \Phi^n \rangle_2 \\ &\quad + \frac{\lambda \Delta t \Delta x}{2} \langle (D_x^+ D_x^- + D_x^- D_x^+) u_h^{n+1}, \Phi^{n+1} - \Phi^n \rangle_2. \end{aligned}$$

Applying integrations by parts (1.36) and (1.38) on the right-hand side together with Cauchy-Schwarz inequality, one obtains

$$\begin{aligned} \|D_x^c \Phi^{n+1} - D_x^c \Phi^n\|_2^2 &= \Delta t \langle J_h^{n+1}, D_x^c \Phi^{n+1} - D_x^c \Phi^n \rangle_2 \\ &\quad - 2\lambda \Delta t \Delta x \langle D_x^c u_h^{n+1}, D_x^c \Phi^{n+1} - D_x^c \Phi^n \rangle_2 \\ &\leq \Delta t (\|J_h^{n+1}\|_2 + 2\lambda \Delta x \|D_x^c u_h^{n+1}\|_2) \|D_x^c \Phi^{n+1} - D_x^c \Phi^n\|_2. \end{aligned}$$

One can conclude by dividing both sides by $\|D_x^c \Phi^{n+1} - D_x^c \Phi^n\|_2$ and using estimates (1.39) and (1.63). □

In the following lemma, we establish that for $\delta > 0$ small enough, the modified entropy functional is an equivalent $\|\cdot\|_\Delta$ norm.

Lemma 11. *Under the assumptions of Theorem 2 and assuming that $\Delta t \leq \Delta t_{\max}$, there exists $\alpha_1^* > 0$*

such that for all $n \geq 1$

$$\left(\frac{1}{2} - \delta\, C_{J1}^*\, C_u^*\, C_P\right)\|F\|_\Delta^2 \leq H_\delta^\Delta[F^n] \leq \left(\frac{1}{2} + \delta(\, C_{J1}^*\, C_u^*\, C_P + \alpha_1^*\, \Delta t_{\max})\right)\|F^n\|_\Delta^2,$$

where α_1^* *depends on* D_1^Δ, D_2^Δ *and* ρ_∞^*.

Proof. Let us start by estimating the second term of the discrete modified entropy (1.76). The Cauchy-Schwarz inequality followed by (1.64) and (1.78) yield

$$\begin{aligned}|\langle J_h^n, D_x^c\Phi^n\rangle_2| &\leq \|J_h^n\|_2\|D_x^c\Phi^n\|_2\\ &\leq C_{J1}^*\, C_P\, C_u^*\|F^n\|_\Delta\|\Pi^\Delta F^n\|_\Delta\\ &\leq C_{J1}^*\, C_P\, C_u^*\|F^n\|_\Delta^2.\end{aligned}$$

The last term of (1.76) shall be estimated using (1.79) and (1.64), together with Young inequality:

$$\begin{aligned}\sum_{i\in\mathcal{I}}\frac{\Delta x}{2\Delta t}\left((D_x^c\Phi^n)_i - (D_x^c\Phi^{n-1})_i\right)^2 &\leq \frac{\Delta t}{2}\left(\|J_h^n\|_2 + 2\,\lambda\, C_u^*\,\|\Pi^\Delta F^n\|_\Delta\right)^2\\ &\leq \Delta t\left((C_{J1}^*)^2 + 4(\lambda C_u^*)^2\right)\|F^n\|_\Delta^2.\end{aligned}$$

Setting

$$\alpha_1^* = (C_{J1}^*)^2 + 4(\lambda C_u^*)^2,$$

one obtains

$$0 \leq \sum_{i\in\mathcal{I}}\frac{\Delta x}{2\Delta t}\left((D_x^c\Phi^n)_i - (D_x^c\Phi^{n-1})_i\right)^2 \leq \alpha_1^*\Delta t\|F^n\|_\Delta^2.$$

Consequently, the modified entropy is bounded from below by a positive quantity as long as $\delta < \left(2C_{J1}^*\, C_u^*\, C_P\right)^{-1} =: \delta_3$. Finally, the upper bound is obtained using that $\Delta t \leq \Delta t_{max}$. □

Proposition 3. *Under the assumptions of Theorem 2, there is* $\delta_2 > 0$ *such that for all* $\Delta t \leq \Delta t_{\max}$ *and* $\delta \leq \delta_2$, *there exists* $K_\delta > 0$ *such that*

$$H_\delta^\Delta[F^{n+1}] - H_\delta^\Delta[F^n] \leq -\Delta t\, K_\delta\, \|F^{n+1}\|_\Delta^2. \tag{1.80}$$

Proof. Taking the difference between the modified entropy at time t^{n+1} and t^n we get

$$H_\delta[F^{n+1}] - H_\delta[F^n] = \frac{1}{2}\left(\|F^{n+1}\|_\Delta^2 - \|F^n\|_\Delta^2\right) + \delta T_1^n + \delta T_2^n, \tag{1.81}$$

where

$$\begin{aligned}T_1^n &= \sum_{i\in\mathcal{I}}\left(J_{h,i}^{n+1}(D_x^c\Phi^{n+1})_i - J_{h,i}^n(D_x^c\Phi^n)_i\right)\Delta x,\\ T_2^n &= \sum_{i\in\mathcal{I}}\frac{\Delta x}{2\Delta t}\left[\left((D_x^c\Phi^{n+1})_i - (D_x^c\Phi^n)_i\right)^2 - \left((D_x^c\Phi^n)_i - (D_x^c\Phi^{n-1})_i\right)^2\right].\end{aligned}$$

We already showed through Lemma 7 that

$$\frac{1}{2}\Big(\|F^{n+1}\|_\Delta^2 - \|F^n\|_\Delta^2\Big) \le -\Delta t\, C_{mc}^* \|(I-\Pi^\Delta)F^n\|_\Delta^2.$$

It remains to deal with the last two terms. First, let us remark that $T_1^n = T_{11}^n + T_{12}^n + T_{13}^n$, with

$$\begin{aligned}
T_{11}^n &= \langle J_h^{n+1} - J_h^n, D_x^c \Phi^{n+1}\rangle_2,\\
T_{12}^n &= \langle J_h^{n+1}, D_x^c \Phi^{n+1} - D_x^c \Phi^n\rangle_2,\\
T_{13}^n &= \sum_{i\in\mathcal{I}} \Delta x \Big[J_{h,i}^n (D_x^c\Phi^{n+1})_i - J_{h,i}^{n+1}(D_x^c\Phi^{n+1})_i + J_{h,i}^{n+1}(D_x^c\Phi^n)_i - J_{h,i}^n (D_x^c\Phi^n)_i\Big].
\end{aligned}$$

It is worth noticing that the first two terms are discrete equivalent of

$$\Big(\partial_t \langle J_h, \partial_x \Phi\rangle_{L^2(\mathbb{T})} + \langle J_h, \partial_{tx}\Phi\rangle_{L^2(\mathbb{T})}\Big)\Delta t$$

and will therefore be treated in the same manner as their continuous counterparts. Let us first deal with the residual term T_{13}^n. After reorganizing the terms and integrating by parts, it becomes

$$\begin{aligned}
T_{13}^n &= \sum_{i\in\mathcal{I}} \Delta x (J_{h,i}^n - J_{h,i}^{n+1})\Big((D_x^c\Phi^{n+1})_i - (D_x^c\Phi^n)_i\Big)\\
&= \sum_{i\in\mathcal{I}} \Delta x \Big((D_x^c J_h^{n+1})_i - (D_x^c J_h^n)_i\Big)(\Phi_i^{n+1} - \Phi_i^n).
\end{aligned}$$

The next step is to replace the discrete space derivatives of J_h using the continuity scheme (1.72), followed by several uses of the auxiliary scheme (1.77) and integrations by parts:

$$\begin{aligned}
T_{13}^n &= -\frac{1}{\Delta t}\sum_{i\in\mathcal{I}} \Delta x (u_{h,i}^{n+1} - 2u_{h,i}^n + u_{h,i}^{n-1})(\Phi_i^{n+1} - \Phi_i^n)\\
&\quad + \frac{\lambda\Delta x}{2}\sum_{i\in\mathcal{I}} \Delta x \Big[\Big((D_x^+D_x^- + D_x^-D_x^+)u_h^{n+1}\Big)_i - \Big((D_x^+D_x^- + D_x^-D_x^+)u_h^n\Big)_i\Big](\Phi_i^{n+1} - \Phi_i^n)\\
&= \frac{1}{\Delta t}\sum_{i\in\mathcal{I}} \Delta x \Big((D_x^cD_x^c\Phi^{n+1})_i - 2(D_x^cD_x^c\Phi^n)_i + (D_x^cD_x^c\Phi^{n-1})_i\Big)(\Phi_i^{n+1} - \Phi_i^n)\\
&\quad - 2\lambda\Delta x \sum_{i\in\mathcal{I}} \Delta x \Big((D_x^c u_h^{n+1})_i - (D_x^c u_h^n)_i\Big)\Big((D_x^c\Phi^{n+1})_i - (D_x^c\Phi^n)_i\Big)\\
&= -\frac{1}{\Delta t}\sum_{i\in\mathcal{I}} \Delta x \Big((D_x^c\Phi^{n+1})_i - 2(D_x^c\Phi^n)_i + (D_x^c\Phi^{n-1})_i\Big)\Big((D_x^c\Phi^{n+1})_i - (D_x^c\Phi^n)_i\Big)\\
&\quad - 2\lambda\Delta x \sum_{i\in\mathcal{I}} \Delta x (u_{h,i}^{n+1} - u_{h,i}^n)^2.
\end{aligned}$$

The second term is clearly nonpositive. The first term is combined with T_2^n using the identity $-a(a-b)+(a^2-b^2)/2 = -(a-b)^2/2$ with $a = (D_x^c\Phi^{n+1})_i - (D_x^c\Phi^n)_i$ and $b = (D_x^c\Phi^n)_i - (D_x^c\Phi^{n-1})_i$, therefore yielding $T_2^n + T_{13}^n \le 0$, and finally

$$H_\delta[F^{n+1}] - H_\delta[F^n] \le \frac{1}{2}\Big(\|F^{n+1}\|_\Delta^2 - \|F^n\|_\Delta^2\Big) + \delta T_{11}^n + \delta T_{12}^n. \tag{1.82}$$

The rest of the proof follows the same steps as in the continuous setting. The term T_{11}^n can be expanded using scheme (1.73) so that

$$T_{11}^n = T_{111}^n + T_{112}^n + T_{113}^n + T_{114}^n,$$

where

$$\begin{aligned}
T_{111}^n &= -\Delta t \langle D_x^c S_h^{n+1}, D_x^c \Phi^{n+1} \rangle_2, \\
T_{112}^n &= -\Delta t D_0 \langle D_x^c u_h^{n+1}, D_x^c \Phi^{n+1} \rangle_2, \\
T_{113}^n &= -\Delta t \langle (\rho_\infty^*)^{-1} J_f^{n+1} - \rho_\infty^* J_g^{n+1}, D_x^c \Phi^{n+1} \rangle_2, \\
T_{114}^n &= \frac{\lambda \Delta t \Delta x}{2} \langle (D_x^+ D_x^- + D_x^- D_x^+) J_h^{n+1}, D_x^c \Phi^{n+1} \rangle_2.
\end{aligned}$$

Thanks to Lemmas 8 and 10 we can estimate T_{11k}, $k = 1, \dots, 4$. After an integration by parts, using the auxiliary scheme (1.77) and applying the Cauchy-Schwarz inequality, one has

$$\begin{aligned}
\left| T_{111}^n \right| = \Delta t \left| \langle S_h^{n+1}, u_h^{n+1} \rangle_2 \right| &\leq \Delta t \| S_h^{n+1} \|_2 \| u_h^{n+1} \|_2 \\
&\leq \Delta t C_S^* C_u^* \| (I - \Pi^\Delta) F^{n+1} \|_\Delta \| \Pi^\Delta F^{n+1} \|_\Delta.
\end{aligned} \tag{1.83}$$

The second term is treated similarly and actually provides the macroscopic coercivity:

$$T_{112}^n = -\Delta t D_0^\Delta \langle u_h^{n+1}, u_h^{n+1} \rangle_2 = -\Delta t D_0^\Delta \| u_h^{n+1} \|_2^2 \leq -\Delta t \underline{D_0} (C_u^*)^2 \| \Pi^\Delta F^{n+1} \|_\Delta^2. \tag{1.84}$$

The next estimate follows directly from the Cauchy-Schwarz inequality:

$$\begin{aligned}
\left| T_{113}^n \right| &= \Delta t \left| \langle (\rho_\infty^*)^{-1} J_f^{n+1} - \rho_\infty^* J_g^{n+1}, D_x^c \Phi^{n+1} \rangle_2 \right| \\
&\leq \Delta t \| (\rho_\infty^*)^{-1} J_f^{n+1} - \rho_\infty^* J_g^{n+1} \|_2 \| D_x^c \Phi^{n+1} \|_2 \\
&\leq \Delta t C_{J2}^* C_P C_u^* \| (I - \Pi^\Delta) F^{n+1} \|_\Delta \| \Pi^\Delta F^{n+1} \|_\Delta.
\end{aligned} \tag{1.85}$$

The estimation of T_{114}^n is obtained through several integrations by parts, the Cauchy-Schwarz inequality and (1.39):

$$\begin{aligned}
\left| T_{114}^n \right| &= 2\lambda \Delta t \Delta x \left| \langle D_x^c J_h^{n+1}, D_x^c D_x^c \Phi^{n+1} \rangle_2 \right| \\
&= 2\lambda \Delta t \Delta x \left| \langle D_x^c J_h^{n+1}, u_h^{n+1} \rangle_2 \right| \\
&\leq 2\lambda \Delta t \Delta x \| J_h^{n+1} \|_2 \| D_x^c u_h^{n+1} \|_2 \\
&\leq 2\lambda \Delta t C_{J1}^* C_u^* \| (I - \Pi^\Delta) F^{n+1} \|_\Delta \| \Pi^\Delta F^{n+1} \|_\Delta.
\end{aligned} \tag{1.86}$$

It remains now to estimate T_{12}^n. After applying a Cauchy-Schwarz inequality followed by (1.79), one obtains:

$$\begin{aligned}
\left| T_{12}^n \right| &\leq \| J_h^{n+1} \|_2 \| D_x^c \Phi^{n+1} - D_x^c \Phi^n \|_2 \\
&\leq \Delta t (C_{J1}^*)^2 \| (I - \Pi^\Delta) F^{n+1} \|_\Delta^2 + 2\Delta t \lambda C_{J1}^* C_u^* \| (I - \Pi^\Delta) F^{n+1} \|_\Delta \| \Pi^\Delta F^{n+1} \|_\Delta.
\end{aligned} \tag{1.87}$$

Summarizing, combining estimates (1.83), (1.84), (1.85), (1.86), (1.87) and (1.59) in (1.82), it

yields

$$\begin{aligned} H_\delta[F^{n+1}] - H_\delta[F^n] \leq & -\Delta t (C^*_{mc} - \delta (C^*_{J1})^2) \|(I-\Pi^\Delta) F^{n+1}\|^2_\Delta \\ & - \Delta t \delta \underline{D_0} (C^*_u)^2 \|\Pi^\Delta F^{n+1}\|^2_\Delta \\ & + \Delta t \delta (C^*_S C^*_u + C^*_{J2} C_P C^*_u + 4\lambda C^*_{J1} C^*_u) \|(I-\Pi^\Delta) F^{n+1}\|_\Delta \|\Pi^\Delta F^{n+1}\|_\Delta. \end{aligned} \tag{1.88}$$

Let us first set $\delta \in (0, \delta_1)$ with $\delta_1 = C^*_{mc}(C^*_{J1})^{-2}$ to ensure the positivity of $C^*_{mc} - \delta (C^*_{J1})^2$. We can then rewrite the right-hand side of (1.88):

$$\begin{aligned} H_\delta[F^{n+1}] - H_\delta[F^n] \leq & -\frac{\Delta t}{2} (C^*_{mc} - \delta (C^*_{J1})^2) \|(I-\Pi^\Delta) F^{n+1}\|^2_\Delta \\ & - \frac{\Delta t}{2} \delta \underline{D_0} (C^*_u)^2 \|\Pi^\Delta F^{n+1}\|^2_\Delta \\ & - \Delta t \|\Pi^\Delta F^{n+1}\|^2_\Delta P(\|(I-\Pi^\Delta) F^{n+1}\|_\Delta \|\Pi^\Delta F^{n+1}\|^{-2}_\Delta), \end{aligned} \tag{1.89}$$

where P is the polynomial given by

$$P(X) = \frac{1}{2}(C^*_{mc} - \delta (C^*_{J1})^2) X^2 - \delta \widetilde{C}^\Delta X + \frac{1}{2} \delta \underline{D_0} (C^*_u)^2,$$

with $\widetilde{C}^\Delta = C^*_S C^*_u + C^*_{J2} C_P C^*_u + 4\lambda C^*_{J1} C^*_u$.

The sum of the first two terms in (1.89) is nonpositive thanks to our choice of δ. It remains to impose a second condition on δ to ensure that the polynomial P is positive. Since the leading order coefficient of P is positive, it suffices to choose δ in such a way that the discriminant is negative. It yields the following condition on δ:

$$\delta_2 := \frac{C^*_{mc} \underline{D_0} (C^*_u)^2}{(\widetilde{C}^\Delta)^2 + (C^*_{J1})^2 \underline{D_0} (C^*_u)^2} > \delta. \tag{1.90}$$

Assuming that $\delta \in (0, \min(\delta_1, \delta_2))$, one then has

$$H_\delta[F^{n+1}] - H_\delta[F^n] \leq -\frac{\Delta t}{2} \Big[(C^*_{mc} - \delta (C^*_{J1})^2) \|(I-\Pi^\Delta) F^{n+1}\|^2_\Delta + \delta \underline{D_0} (C^*_u)^2 \|\Pi^\Delta F^{n+1}\|^2_\Delta \Big].$$

Finally, setting $K_\delta = \frac{1}{2} \min(C^*_{mc} - \delta (C^*_{J1})^2, \delta \underline{D_0} (C^*_u)^2)$ and using orthogonality properties, one concludes the proof. □

Proof of Theorem 2. Starting from the result of Proposition 3, we use Lemma 11 to bound $\|F^{n+1}\|^2_\Delta$ from below by $H_\delta[F^{n+1}]$ and obtain

$$(1 + \Delta t\, \kappa) H^\Delta_\delta[F^{n+1}] \leq H^\Delta_\delta[F^n],$$

where we set $\kappa = \frac{K_\delta}{C_\delta}$ and $C_\delta = \frac{1}{2} + \delta\, C^*_{J1}\, C^*_u\, C_P + \delta\, \alpha^*_1\, \Delta t_{\max}$. It implies that for all $n \geq 1$,

$$H^\Delta_\delta[F^{n+1}] \leq H^\Delta_\delta[F^1](1 + \Delta t \kappa)^{-n} = H^\Delta_\delta[F^1] \exp\left(-t^n \log(1 + \Delta t \kappa)/\Delta t\right).$$

Since $s \mapsto \log(1 + s\kappa)/s$ is nonincreasing on $]0, +\infty[$ and $\Delta t \leq \Delta t_{max}$, we obtain

$$H^\Delta_\delta[F^{n+1}] \leq H^\Delta_\delta[F^1]\, e^{-\beta\, t^n},$$

where $\beta = \log(1+\Delta t_{max}\kappa)/\Delta t_{max}$. Then, using lemmas 11 and 7, we finally get

$$H_\delta^\Delta[F^{n+1}] \le C_\delta \|F^1\|_\Delta^2 \, e^{-\beta t^n} \le C_\delta \|F^0\|_\Delta^2 \, e^{-\beta t^n}.$$

Finally, choosing

$$\delta \in (0, \min(\delta_1, \delta_2, \delta_3)),$$

the constant on the left-hand side in Lemma 11 is positive and one can conclude the proof. □

1.5 The nonlinear problem

In this section, we prove that approximate solutions to the nonlinear system with initial data sufficiently close to equilibrium converge exponentially towards this equilibrium as time goes to infinity. The idea of the proof is as follows: we first establish in Section 1.5.1 that solutions to the numerical scheme (1.40)–(1.41) satisfy the maximum principle (and also that they exist), and then in Section 1.5.2, using the fact that the entropy dissipation for the nonlinear problem is a small perturbation of the entropy dissipation of the linearized problem for initial data sufficiently close to equilibrium, the nonlinear discrete hypocoercivity is obtained.

1.5.1 Existence and maximum principle

This section is devoted to the proof of the following result.

Theorem 3. *Under assumptions (1.32), let ρ_∞^* be defined by (1.46) and assume that there exists positive constants $\gamma_1 < \rho_\infty^*$ and γ_2 such that the initial datum $\mathsf{F}^0 = (\mathfrak{f}_{ij}^0, \mathfrak{g}_{ij}^0)_{i\in\mathcal{I}, j\in\mathcal{J}}$ satisfies for all $i \in \mathcal{I}$, $j \in \mathcal{J}$*

$$(\rho_\infty^* - \gamma_1)\chi_{1,j} \le \mathfrak{f}_{ij}^0 \le (\rho_\infty^* + \gamma_2)\chi_{1,j},$$
$$(\rho_\infty^* + \gamma_2)^{-1}\chi_{2,j} \le \mathfrak{g}_{ij}^0 \le (\rho_\infty^* - \gamma_1)^{-1}\chi_{2,j}.$$

Assuming moreover that $\lambda = \Delta x/2\Delta t \ge v^*/2$, the scheme (1.40)–(1.41) admits a solution $(\mathfrak{f}_{ij}^n, \mathfrak{g}_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}, n\ge 0}$ such that for all $i \in \mathcal{I}$, $j \in \mathcal{J}$, $n \ge 0$,

$$(\rho_\infty^* - \gamma_1)\chi_{1,j} \le \mathfrak{f}_{ij}^n \le (\rho_\infty^* + \gamma_2)\chi_{1,j}, \tag{1.91}$$
$$(\rho_\infty^* + \gamma_2)^{-1}\chi_{2,j} \le \mathfrak{g}_{ij}^n \le (\rho_\infty^* - \gamma_1)^{-1}\chi_{2,j}. \tag{1.92}$$

Remark 2. *The condition $\lambda \ge v^*/2$ is there to ensure the monotonicity of the Lax-Friedrichs fluxes, which is necessary in our proof. In practice, we observe that this restriction is necessary in the numerical experiments presented in Section 1.6.*

To prove this result, we introduce the following truncated version of the scheme (1.40)–(1.41):

$$\frac{\mathfrak{f}_{ij}^{n+1} - \mathfrak{f}_{ij}^n}{\Delta t} + \frac{1}{\Delta x \Delta v}\left(\mathcal{F}_{i+\frac{1}{2},}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1}\right) = \chi_{1,j} - \tilde{\rho}_{\mathfrak{g},i}^{n+1}\tilde{\mathfrak{f}}_{ij}^{n+1}, \tag{1.93}$$
$$\frac{\mathfrak{g}_{ij}^{n+1} - \mathfrak{g}_{ij}^n}{\Delta t} + \frac{1}{\Delta x \Delta v}\left(\mathcal{G}_{i+\frac{1}{2},j}^{n+1} - \mathcal{G}_{i-\frac{1}{2},j}^{n+1}\right) = \chi_{2,j} - \tilde{\rho}_{\mathfrak{f},i}^{n+1}\tilde{\mathfrak{g}}_{ij}^{n+1}, \tag{1.94}$$

with the Lax-Friedrichs fluxes (1.42)–(1.43), and where the truncated quantities are defined by

$$\tilde{\mathfrak{f}}_{ij} := \begin{cases} (\rho_\infty^* - \gamma_1)\chi_{1,j} & \text{if } \mathfrak{f}_{ij} \le (\rho_\infty^* - \gamma_1)\chi_{1,j}, \\ \mathfrak{f}_{ij} & \text{if } (\rho_\infty^* - \gamma_1)\chi_{1,j} \le \mathfrak{f}_{ij} \le (\rho_\infty^* + \gamma_2)\chi_{1,j}, \\ (\rho_\infty^* + \gamma_2)\chi_{1,j} & \text{if } \mathfrak{f}_{ij} \ge (\rho_\infty^* + \gamma_2)\chi_{1,j}, \end{cases} \tag{1.95}$$

and

$$\tilde{\mathfrak{g}}_{ij} := \begin{cases} (\rho_\infty^* + \gamma_2)^{-1}\chi_{2,j} & \text{if } \mathfrak{g}_{ij} \le (\rho_\infty^* + \gamma_2)^{-1}\chi_{2,j}, \\ \mathfrak{g}_{ij} & \text{if } (\rho_\infty^* + \gamma_2)^{-1}\chi_{2,j} \le \mathfrak{g}_{ij} \le (\rho_\infty^* - \gamma_1)^{-1}\chi_{2,j}, \\ (\rho_\infty^* - \gamma_1)^{-1}\chi_{2,j} & \text{if } \mathfrak{g}_{ij} \ge (\rho_\infty^* - \gamma_1)^{-1}\chi_{2,j}. \end{cases} \tag{1.96}$$

The corresponding truncated densities are then given by

$$\tilde{\rho}_{\mathfrak{f},i} = \sum_{j\in\mathcal{J}} \Delta v\, \tilde{\mathfrak{f}}_{ij}, \qquad \tilde{\rho}_{\mathfrak{g},i} = \sum_{j\in\mathcal{J}} \Delta v\, \tilde{\mathfrak{g}}_{ij} \qquad \forall i \in \mathcal{I}.$$

To prove Theorem 3, we start by establishing the existence of a solution to the truncated scheme (Lemma 12) and then the fact that a solution to the truncated scheme verifies the estimates (1.91)–(1.92) (Lemma 14).

Lemma 12. *Given any* $(\mathfrak{f}_{ij}^n, \mathfrak{g}_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$, *the truncated scheme* (1.93)–(1.94) *admits at least one solution* $(\mathfrak{f}_{ij}^{n+1}, \mathfrak{g}_{ij}^{n+1})_{i\in\mathcal{I}, j\in\mathcal{J}}$.

Proof. The proof relies on the following result which is proven for example in [197, Lemma 1.4, Chapter II].

Lemma 13. *Let* X *be a finite dimensional Hilbert space with scalar product* $\langle\cdot,\cdot\rangle_X$ *and associated norm* $\|\cdot\|_X$. *Let* $P : X \to X$ *be a continuous mapping such that* $\langle P(\xi), \xi\rangle_X > 0$ *for all* $\xi \in X$ *such that* $\|\xi\|_X = k$ *for some fixed* $k > 0$. *Then there exists* $\xi_0 \in X$ *such that* $\|\xi_0\|_X \le k$ *and* $P(\xi_0) = 0$.

We apply this result with $X = \left(\mathbb{R}^{(2L+1)N}\right)^2$ where for $u^1 = (f^1, g^1) \in X$ and $u^2 = (f^2, g^2) \in X$,

$$\langle u_\Delta^1, u_\Delta^2\rangle_X = \langle f_\Delta^1, f_\Delta^2\rangle + \langle g_\Delta^1, g_\Delta^2\rangle,$$

$\langle\cdot,\cdot\rangle$ being the classical Euclidean dot product on $\mathbb{R}^{(2L+1)N}$. Then, let us define

$$P : \begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix} \mapsto \begin{pmatrix} \Delta x\Delta v(\mathfrak{f}_{ij} - \mathfrak{f}_{ij}^n) + \Delta t(\mathcal{F}_{i+\frac{1}{2},j} - \mathcal{F}_{i-\frac{1}{2},j}) - \Delta x\Delta v\Delta t(\chi_{1,j} - \tilde{\rho}_{\mathfrak{g},i}\tilde{\mathfrak{f}}_{ij}) \\ \Delta x\Delta v(\mathfrak{g}_{ij} - \mathfrak{g}_{ij}^n) + \Delta t(\mathcal{G}_{i+\frac{1}{2},j} - \mathcal{G}_{i-\frac{1}{2},j}) - \Delta x\Delta v\Delta t(\chi_{1,j} - \tilde{\rho}_{\mathfrak{f},i}\tilde{\mathfrak{g}}_{ij}) \end{pmatrix},$$

where $\mathcal{F}$ and $\mathcal{G}$ are the numerical fluxes (1.42)–(1.43) computed with $\mathfrak{f}_{ij}$ and $\mathfrak{g}_{ij}$. Let us now show that

$$\left\langle P\left(\begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix}\right), \begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix}\right\rangle_X > 0, \quad \text{for all } \begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix} \text{ such that } \left\|\begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix}\right\|_X = k,$$

where k is taken large enough. This scalar product splits into 3 terms,

$$\left\langle P\left(\begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix}\right), \begin{pmatrix} \mathfrak{f} \\ \mathfrak{g} \end{pmatrix}\right\rangle_X = A_1 + A_2 + A_3,$$

where

$$A_1 = \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x\Delta v\Big((\mathfrak{f}_{ij}-\mathfrak{f}_{ij}^n)\mathfrak{f}_{ij}+(\mathfrak{g}_{ij}-\mathfrak{g}_{ij}^n)\mathfrak{g}_{ij}\Big),$$
$$A_2 = \Delta t \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Big((\mathcal{F}_{i+\frac{1}{2},j}-\mathcal{F}_{i-\frac{1}{2},j})\mathfrak{f}_{ij}+(\mathcal{G}_{i+\frac{1}{2},j}-\mathcal{G}_{i-\frac{1}{2},j})\mathfrak{g}_{ij}\Big),$$
$$A_3 = -\Delta t \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Delta x\Delta v\Big((\chi_{1,j}-\tilde{\rho}_{\mathfrak{g},i}\tilde{\mathfrak{f}}_{ij})\mathfrak{f}_{ij}+(\chi_{2,j}-\tilde{\rho}_{\mathfrak{f},i}\tilde{\mathfrak{g}}_{ij})\mathfrak{g}_{ij}\Big).$$

Using the relation $a(a-b)\geq(a^2-b^2)/2$ one gets

$$A_1 \geq \frac{1}{2}\left(\left\|\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\|_X^2 - \left\|\begin{pmatrix}\mathfrak{f}^n\\ \mathfrak{g}^n\end{pmatrix}\right\|_X^2\right). \tag{1.97}$$

Then, by definition of the numerical fluxes, one can use the same computations as in the proof of Lemma 7 to obtain

$$A_2 = 2\lambda\Delta t\Delta x \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} \Big((D_x^c\mathfrak{f}_j)_i^2+(D_x^c\mathfrak{g}_j)_i^2\Big)\Delta x\Delta v \geq 0. \tag{1.98}$$

Now, by assumptions (1.32), the velocity profiles are bounded:

$$\exists C_\chi>0 \quad \text{such that} \quad \forall j\in\mathcal{J},\quad 0\leq\chi_{1,j},\chi_{2,j}\leq C_\chi.$$

Then, by definition of the truncated quantities, there exists a constant $C>0$ such that

$$|\chi_{1,j}-\tilde{\rho}_{\mathfrak{g},i}\tilde{\mathfrak{f}}_{ij}|\leq C \quad \text{and} \quad |\chi_{2,j}-\tilde{\rho}_{\mathfrak{f},i}\tilde{\mathfrak{g}}_{ij}|\leq C.$$

Applying the Cauchy-Schwarz inequality on A_3, one gets

$$A_3 \geq -\Delta t C\left\|\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\|_X. \tag{1.99}$$

Combining (1.97), (1.98) and (1.99) yields the estimate

$$\left\langle P(\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}),\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\rangle_X \geq \frac{1}{2}\left(\left\|\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\|_X^2 - \left\|\begin{pmatrix}\mathfrak{f}^n\\ \mathfrak{g}^n\end{pmatrix}\right\|_X^2\right) - \Delta t C\left\|\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\|_X.$$

The right-hand side is a second order polynomial in $\left\|\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\|_X$ with a positive leading coefficient. Since $\left\|\begin{pmatrix}\mathfrak{f}^n\\ \mathfrak{g}^n\end{pmatrix}\right\|_X$ is a constant in this context, there exists $k>0$ such that if $\left\|\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\|_X \geq k$ then

$$\left\langle P(\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}),\begin{pmatrix}\mathfrak{f}\\ \mathfrak{g}\end{pmatrix}\right\rangle_X > 0.$$

Finally, we can apply Lemma 13 to obtain existence of $\begin{pmatrix} \mathfrak{f}^{n+1} \\ \mathfrak{g}^{n+1} \end{pmatrix}$ such that $P(\begin{pmatrix} \mathfrak{f}^{n+1} \\ \mathfrak{g}^{n+1} \end{pmatrix}) = 0$, therefore ensuring existence of a solution to the truncated scheme (1.93)-(1.94). □

Lemma 14. *If $(\mathfrak{f}_{ij}^n, \mathfrak{g}_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$ satisfies the estimates (1.91)–(1.92), then any solution $(\mathfrak{f}_{ij}^{n+1}, \mathfrak{g}_{ij}^{n+1})_{i\in\mathcal{I}, j\in\mathcal{J}}$ to the truncated scheme (1.93)–(1.94) also satisfies these estimates.*

Proof. Let us focus on showing that a solution to the nonlinear truncated scheme (1.93)-(1.94) satisfies

$$\mathfrak{f}_{ij}^{n+1} \geq (\rho_\infty^* - \gamma_1)\chi_{1,j}.$$

We start by setting $(i,j) \in \mathcal{I}\times\mathcal{J}$ such that

$$\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j} = \min_{(k,l)\in\mathcal{I}\times\mathcal{J}} (\mathfrak{f}_{kl}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,l}). \tag{1.100}$$

Our aim is now to show that this quantity is nonnegative. We begin by multiplying the line of (1.93) corresponding to the fixed pair (i,j) by

$$(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})^- = \min(0, \mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j}) \leq 0. \tag{1.101}$$

It yields an expression of the form $B_1 = B_2 + B_3$ where

$$\begin{aligned}
B_1 &= \Delta x\Delta v(\mathfrak{f}_{ij}^{n+1} - \mathfrak{f}_{ij}^n)(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})^-, \\
B_2 &= -\Delta t(\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1})(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})^-, \\
B_3 &= \Delta t\Delta x\Delta v(\chi_{1,j} - \tilde{\rho}_{\mathfrak{g},i}^{n+1}\tilde{\mathfrak{f}}_{ij}^{n+1})(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})^-.
\end{aligned}$$

Starting with B_1, we add and subtract $(\rho_\infty^* - \gamma_1)\chi_{1,j}$ to obtain

$$B_1 = \Delta x\Delta v((\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j}) - (\mathfrak{f}_{ij}^n - (\rho_\infty^* - \gamma_1)\chi_{1,j}))(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})^-.$$

Then, under the hypothesis that $\mathfrak{f}_{ij}^n$ satisfies the bounds (1.91) and by definition (1.101) of the negative part,

$$B_1 \geq \Delta x\Delta v\,(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})(\mathfrak{f}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j})^- \geq 0. \tag{1.102}$$

Next, we turn our attention to B_3. We need to consider two cases:

— If $\mathfrak{f}_{ij}^{n+1} \geq (\rho_\infty^* - \gamma_1)\chi_{1,j}$ then $B_3 = 0$.

— Else, by definition of the truncated quantities (1.95) and (1.96), one has $\tilde{\mathfrak{f}}_{ij}^{n+1} = (\rho_\infty^* - \gamma_1)\chi_{1,j}$ and since the discrete velocity profiles have unit mass, $\tilde{\rho}_{\mathfrak{g},i}^{n+1} \leq (\rho_\infty^* - \gamma_1)^{-1}$. Therefore,

$$\chi_{1,j} - \tilde{\rho}_{\mathfrak{g},i}^{n+1}\tilde{\mathfrak{f}}_{ij}^{n+1} \geq \chi_{1,j} - \frac{\rho_\infty^* - \gamma_1}{\rho_\infty^* - \gamma_1}\chi_{1,j} \geq 0.$$

As a result, one gets

$$B_3 \leq 0. \tag{1.103}$$

Finally, the sign of B_2 relies on the monotonicity of the numerical flux. More precisely, let us

consider a monotone numerical flux in a general two-point approximation form:

$$\mathcal{F}_{i+\frac{1}{2},j}^{n+1} = \varphi_j(\mathfrak{k}_{ij}^{n+1}, \mathfrak{k}_{i+1,j}^{n+1}),$$

where $a \mapsto \varphi_j(a,\cdot)$ is assumed to be nondecreasing and $b \mapsto \varphi_j(\cdot,b)$ is assumed to be nonincreasing. Then, the balance of fluxes at cell $\mathcal{X}_i$ rewrites as

$$\begin{aligned}\mathcal{F}_{i+\frac{1}{2},j}^{n+1} - \mathcal{F}_{i-\frac{1}{2},j}^{n+1} = \varphi_j(\mathfrak{k}_{ij}^{n+1}, \mathfrak{k}_{i+1,j}^{n+1}) - \varphi_j(\mathfrak{k}_{ij}^{n+1}, \mathfrak{k}_{i,j}^{n+1}) \\ + \varphi_j(\mathfrak{k}_{ij}^{n+1}, \mathfrak{k}_{ij}^{n+1}) - \varphi_j(\mathfrak{k}_{i-1,j}^{n+1}, \mathfrak{k}_{ij}^{n+1}).\end{aligned} \tag{1.104}$$

From our choice (1.100) of the pair (i,j), we deduce that

$$\mathfrak{k}_{ij}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j} \le \mathfrak{k}_{kl}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,l}, \quad \forall (k,l) \in \mathcal{I} \times \mathcal{J},$$

therefore in particular, for every $k \in \mathcal{I}$, $\mathfrak{k}_{ij}^{n+1} \le \mathfrak{k}_{kj}^{n+1}$. Consequently, due to the monotonicity of the function φ_j, it yields

$$B_2 \le 0. \tag{1.105}$$

At this point, one could select any flux satisfying the monotonicity property. In our framework, we opted for the Lax-Friedrichs flux

$$\varphi_j(a,b) = \frac{v_j \Delta v}{2}(a+b) - \lambda \Delta v (b-a)$$

which is monotone under condition. More precisely, we want to ensure the nonnegativity of the first partial derivative of φ_j and the nonpositivity of the second. Therefore, one imposes $\lambda = \Delta x / 2\Delta t$ such that

$$\begin{aligned}\partial_a \varphi_j(a,b) \ge 0 &\iff \frac{v_j}{2} + \lambda \ge 0 \iff -\frac{v_j}{2} \le \lambda, \\ \partial_b \varphi_j(a,b) \le 0 &\iff \frac{v_j}{2} - \lambda \ge 0 \iff \frac{v_j}{2} \le \lambda.\end{aligned}$$

In the end, the choice $\lambda \ge v_\star/2$ suffices.

Gathering (1.102), (1.103) and (1.105) into $B_1 = B_2 + B_3$, we get that $B_1 = 0$, and consequently that

$$\mathfrak{k}_{i,j}^{n+1} - (\rho_\infty^* - \gamma_1)\chi_{1,j} \ge 0.$$

Lastly, since we chose (i,j) satisfying (1.100),

$$\mathfrak{k}_{i,j}^{n+1} \ge (\rho_\infty^* - \gamma_1)\chi_{1,j}, \quad \forall (i,j) \in \mathcal{I} \times \mathcal{J}.$$

The remaining bounds can then be obtained following the same steps. □

Proof of Theorem 3. We proceed by induction on n. The case $n = 0$ is satisfied by assumption. Suppose now that there exists $(\mathfrak{k}^n, \mathfrak{g}^n)$ satisfying bounds (1.91) and (1.92). Then, we obtain from Lemma 12 the existence of $(\mathfrak{k}^{n+1}, \mathfrak{g}^{n+1})$ solution to the truncated scheme (1.93)-(1.94), which satisfies

(1.91) and (1.92) according to Lemma 14. But then, we have

$$\tilde{\mathfrak{f}}^{n+1} = \mathfrak{f}^{n+1}, \quad \tilde{\mathfrak{g}}^{n+1} = \mathfrak{g}^{n+1},$$
$$\tilde{\rho_{\mathfrak{f}}}^{n+1} = \rho_{\mathfrak{f}}^{n+1}, \quad \tilde{\rho_{\mathfrak{g}}}^{n+1} = \rho_{\mathfrak{g}}^{n+1},$$

meaning that $(\mathfrak{f}^{n+1}, \mathfrak{g}^{n+1})$ is indeed a solution to the original nonlinear scheme (1.40)-(1.41), which concludes the proof. □

1.5.2 Local hypocoercivity result

Thanks to the maximum principle estimates obtained in Theorem 3, we can now extend the decay result for the discrete linearized problem to a local result in the nonlinear case with the same method.

Theorem 4. *Under the assumptions of Theorem 2 and Theorem 3, with γ_1 and γ_2 small enough, a solution $\mathsf{F}^n = (\mathfrak{f}_{ij}^n, \mathfrak{g}_{ij}^n)_{i\in\mathcal{I}, j\in\mathcal{J}}$ of the nonlinear scheme (1.40)–(1.41) satisfies for all $n \geq 0$*

$$\|\mathsf{F}^n - \mathsf{F}^\infty\|_\Delta \leq C \|\mathsf{F}^0 - \mathsf{F}^\infty\|_\Delta \, e^{-\kappa t^n}, \tag{1.106}$$

where the equilibrium F^∞ is defined by (1.47).

The constants C and κ do not depend on the size of the discretization Δ.

Proof. Let us first denote by $Q^\Delta(\mathfrak{f}, \mathfrak{g})$ the discrete nonlinear collision operator given by

$$Q^\Delta(\mathfrak{f}, \mathfrak{g}) = \begin{pmatrix} \chi_{1,j} - \rho_{\mathfrak{g},i} \mathfrak{f}_{ij} \\ \chi_{2,j} - \rho_{\mathfrak{f},i} \mathfrak{g}_{ij} \end{pmatrix}. \tag{1.107}$$

Let us now rewrite the nonlinear scheme (1.40)-(1.41) in terms of $\widetilde{\mathsf{F}}_{ij}^n = \mathsf{F}_{ij}^n - \mathsf{F}_j^\infty$. Since the equilibrium F_j^∞ does not depend on position nor time, the discrete time derivative of $\widetilde{\mathsf{F}}$ satisfies

$$\frac{\widetilde{\mathsf{F}}_{ij}^{n+1} - \widetilde{\mathsf{F}}_{ij}^n}{\Delta t} + \frac{1}{\Delta x \Delta v} \begin{pmatrix} \widetilde{\mathcal{F}}_{i+\frac{1}{2},j}^{n+1} - \widetilde{\mathcal{F}}_{i-\frac{1}{2},j}^{n+1} \\ \widetilde{\mathcal{G}}_{i+\frac{1}{2},j}^{n+1} - \widetilde{\mathcal{G}}_{i-\frac{1}{2},j}^{n+1} \end{pmatrix} = Q^\Delta(\mathfrak{f}^{n+1}, \mathfrak{g}^{n+1}), \tag{1.108}$$

where $\widetilde{\mathcal{F}}$ and $\widetilde{\mathcal{G}}$ are the numerical fluxes applied to $\widetilde{\mathfrak{f}}$ and $\widetilde{\mathfrak{g}}$ respectively. In addition, one can compute the following relation

$$Q^\Delta(\mathfrak{f}, \mathfrak{g}) - \mathsf{L}^\Delta \widetilde{\mathsf{F}} = \begin{pmatrix} -(\rho_{\mathfrak{g},i} - (\rho_\infty^*)^{-1})(\mathfrak{f} - \rho_\infty^* \chi_{1,j}) \\ -(\rho_{\mathfrak{f},i} - \rho_\infty^*)(\mathfrak{g} - (\rho_\infty^*)^{-1} \chi_{2,j}) \end{pmatrix} = \begin{pmatrix} -(\rho_{\mathfrak{g},i} - (\rho_\infty^*)^{-1})\tilde{\mathfrak{f}} \\ -(\rho_{\mathfrak{f},i} - \rho_\infty^*)\tilde{\mathfrak{g}} \end{pmatrix}. \tag{1.109}$$

Then, multiplying by Δv and summing over $j \in \mathcal{J}$ the bounds (1.91) and (1.92) obtained in Theorem 3, one gets

$$\begin{cases} \rho_\infty^* - \gamma_1 \leq \rho_{\mathfrak{f},i}^{n+1} \leq \rho_\infty^* + \gamma_2, \\ (\rho_\infty^* + \gamma_2)^{-1} \leq \rho_{\mathfrak{g},i}^{n+1} \leq (\rho_\infty^* - \gamma_1)^{-1}. \end{cases}$$

Therefore, by setting

$$\gamma_3 = \max(\gamma_1, \gamma_2) \quad \text{and} \quad \gamma_4 = \max\left(\frac{\gamma_2}{\rho_\infty^*(\rho_\infty^* + \gamma_2)}, \frac{\gamma_1}{\rho_\infty^*(\rho_\infty^* - \gamma_1)}\right),$$

one has for all $i \in \mathcal{I}$

$$|\rho_{\mathfrak{f},i}^{n+1} - \rho_\infty^*| \leq \gamma_3 \quad \text{and} \quad |\rho_{\mathfrak{g},i}^{n+1} - (\rho_\infty^*)^{-1}| \leq \gamma_4. \tag{1.110}$$

We are now able to estimate the norm of (1.109). Using the previous estimation (1.110) and setting $\gamma = \max(\gamma_3, \gamma_4)$ we obtain

$$\left\| Q^\Delta(\mathfrak{f}^{n+1}, \mathfrak{g}^{n+1}) - \mathsf{L}^\Delta \widetilde{\mathsf{F}}^{n+1} \right\|_\Delta \leq \gamma \|\widetilde{\mathsf{F}}^{n+1}\|_\Delta. \tag{1.111}$$

Then, by combining this estimate with Lemma 7, we obtain the following dissipation estimate

$$\frac{1}{2}\left(\|\widetilde{\mathsf{F}}^{n+1}\|_\Delta^2 - \|\widetilde{\mathsf{F}}^{n}\|_\Delta^2\right) \leq -\Delta t C_{mc}^* \|(I - \Pi^\Delta)\widetilde{\mathsf{F}}^{n+1}\|_\Delta^2 + \Delta t \gamma \|\widetilde{\mathsf{F}}^{n+1}\|_\Delta^2. \tag{1.112}$$

Defining, $\tilde{\mathfrak{h}} := \tilde{\mathfrak{f}} - \tilde{\mathfrak{g}}$, the rest of the proof unfolds as in the linear setting, except a slight modification on the right-hand side of the moment scheme on $J_{\tilde{\mathfrak{h}}}$:

$$\frac{u_{\tilde{\mathfrak{h}},i}^{n+1} - u_{\tilde{\mathfrak{h}},i}^{n}}{\Delta t} + (D_x^c J_{\tilde{\mathfrak{h}}}^{n+1})_i - \frac{\Delta x\, \lambda}{2}\left((D_x^+ D_x^- + D_x^- D_x^+) u_{\tilde{\mathfrak{h}}}^{n+1}\right)_i = 0, \tag{1.113}$$

$$\begin{aligned}\frac{J_{\tilde{\mathfrak{h}},i}^{n+1} - J_{\tilde{\mathfrak{h}},i}^{n}}{\Delta t} &+ (D_x^c S_{\tilde{\mathfrak{h}}}^{n+1})_i + D_0^\Delta (D_x^c u_{\tilde{\mathfrak{h}}}^{n+1})_i - \frac{\Delta x\, \lambda}{2}\left((D_x^+ D_x^- + D_x^- D_x^+) J_{\tilde{\mathfrak{h}}}^{n+1}\right)_i \\ &= -((\rho_\infty^*)^{-1} J_{\tilde{\mathfrak{f}},i}^{n+1} - \rho_\infty^* J_{\tilde{\mathfrak{g}},i}^{n+1}) - \left((\rho_{\mathfrak{g},i}^{n+1} - (\rho_\infty^*)^{-1}) J_{\tilde{\mathfrak{f}},i}^{n+1} - (\rho_{\mathfrak{f},i}^{n+1} - \rho_\infty^*) J_{\tilde{\mathfrak{g}},i}^{n+1}\right).\end{aligned} \tag{1.114}$$

Using the bounds (1.110) and the same computations as in the proof of (1.67), one can estimate the norm of the additional term so that

$$\|(\rho_{\mathfrak{g}}^{n+1} - (\rho_\infty^*)^{-1}) J_{\tilde{\mathfrak{f}}}^{n+1} - (\rho_{\mathfrak{f}}^{n+1} - \rho_\infty^*) J_{\tilde{\mathfrak{g}}}^{n+1}\|_2 \leq \gamma C_{J1}^* \|(I - \Pi^\Delta)\widetilde{\mathsf{F}}^{n+1}\|_\Delta. \tag{1.115}$$

Next, looking at the discrete time derivative of H_δ^Δ, the same computations as in the linearized setting lead to the relation

$$H_\delta[\widetilde{\mathsf{F}}^{n+1}] - H_\delta[\widetilde{\mathsf{F}}^{n}] \leq \frac{1}{2}\left(\|\widetilde{\mathsf{F}}^{n+1}\|_\Delta^2 - \|\widetilde{\mathsf{F}}^{n}\|_\Delta^2\right) + \delta \sum_{k=1}^{4} \tilde{T}_{11k}^n + \delta \tilde{T}_{115}^n + \delta \tilde{T}_{12}^n,$$

where $\tilde{T}_{11k}^n$, $k = 1, \dots, 4$ and $\tilde{T}_{12}^n$ are defined respectively as T_{11k}^n, $k = 1, \dots, 4$ and T_{12}^n in the proof of Proposition 3, replacing F by $\widetilde{\mathsf{F}}$. The additional term $\tilde{T}_{115}^n$ comes from the right-hand side of (1.114) and is given by

$$\tilde{T}_{115}^n = \langle (\rho_{\mathfrak{g}}^{n+1} - (\rho_\infty^*)^{-1}) J_{\tilde{\mathfrak{f}}}^{n+1} - (\rho_{\mathfrak{f}}^{n+1} - \rho_\infty^*) J_{\tilde{\mathfrak{g}}}^{n+1}, D_x^c \Phi^{n+1} \rangle_2.$$

This term can then be estimated using the Cauchy-Schwarz inequality, (1.115) and (1.78), producing a cross-term

$$\left|\tilde{T}_{115}^n\right| \leq \gamma C_{J1}^* C_P C_u^* \|(I - \Pi^\Delta)\widetilde{\mathsf{F}}^{n+1}\|_\Delta \|\Pi^\Delta \widetilde{\mathsf{F}}^{n+1}\|_\Delta. \tag{1.116}$$

Proceeding as in the proof of Proposition 3, choosing an admissible $\delta > 0$, we can then establish a

similar entropy dissipation as before, but with an additional positive term

$$H_\delta^\Delta\left[\widetilde{\mathsf{F}}^{n+1}\right] - H_\delta^\Delta\left[\widetilde{\mathsf{F}}^{n}\right] \leq -\Delta t\, K_\delta \left\|\widetilde{\mathsf{F}}^{n+1}\right\|_\Delta^2 + \Delta t \gamma \left\|\widetilde{\mathsf{F}}^{n+1}\right\|_\Delta^2. \tag{1.117}$$

Finally, choosing γ small enough, that is such that $\gamma < K_\delta$, we can conclude the proof as before, using Lemma 11. □

1.6 Numerical results

In this section we shall present numerical results for both the linearized and nonlinear schemes we presented and analyzed in the previous sections. We shall do it for a variety of initial data and equilibrium profiles, in order to showcase both the accuracy and the robustness of our approach. Unless stated otherwise, we will always take $v^* = 12$ and $L = 16$ half points in velocity (and have then 32 control volumes in the velocity space) and $N = 101$ spatial cells for the torus $[0, \pi]$. Due to the unconditional stability of our implicit approach, we shall take the rather large time step $\Delta t = 0.1$ for the linear case. The nonlinear case being more intricate, we shall use an adaptive time stepping to optimize the number of iterations needed in the Newton-Raphson solver for the scheme (1.40)–(1.41). Note that due to the relative simplicity of this method, the Jacobian involved in this solver is exact.

Equilibrium profiles. We shall consider three different equilibrium densities in our upcoming numerical experiments: a classical centered reduced Gaussian

$$\chi_{\mathcal{M}}(v) := \frac{1}{\sqrt{2\pi}} e^{-|v|^2/2}, \quad \forall v \in [-v^*, v^*]; \tag{1.118}$$

a polynomially decaying function

$$\chi_{\mathcal{P}}(v) := \frac{1}{1+|v|^4}, \quad \forall v \in [-v^*, v^*]; \tag{1.119}$$

and a polynomially oscillating function

$$\chi_{\mathcal{O}}(v) := \frac{\cos(\pi v) + 1.1}{1+|v|^6}, \quad \forall v \in [-v^*, v^*]. \tag{1.120}$$

Remark 3. *Note that the distribution $\chi_{\mathcal{P}}$ has only bounded moments up to order 3, instead of the 4 needed for Theorem 2 to hold. Nevertheless, we observe in the upcoming numerical experiments that this is not an issue for recovering an exponential decay of our discrete solutions. We infer that this hypothesis is only technical. This is in total agreement with the continuous theorem from [171] where this hypothesis is not required.*

1.6.1 Discrete hypocoercivity of the linearized scheme

Let us start by investigating the hypocoercive behavior of the linear scheme (1.52)-(1.53). We shall compare the three fluxes introduced in Section 1.3, namely the monotone Lax-Friedrichs (1.42)-(1.43) and upwind (1.50)-(1.51), and the nonmonotone central fluxes (1.48)-(1.49).

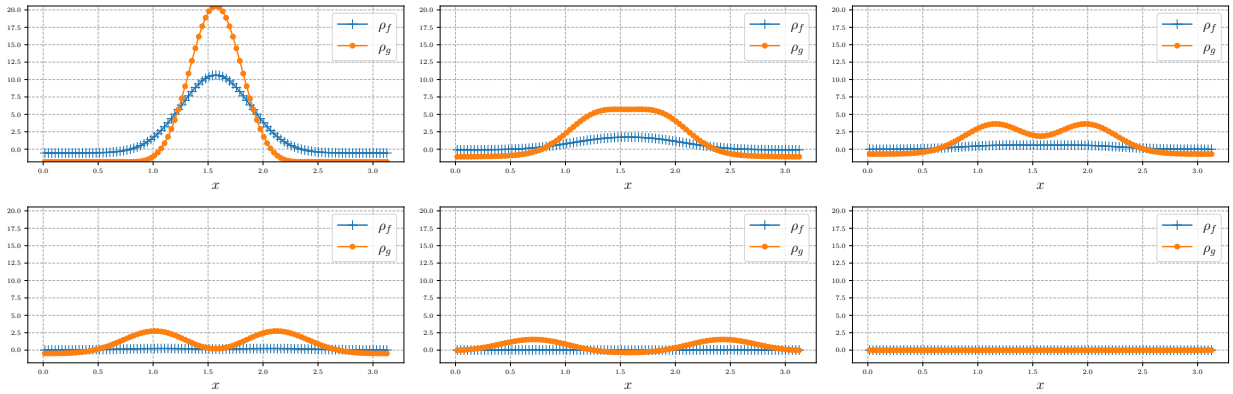


Figure 1.1 – **Test 1. Large time behavior of the linearized scheme, same equilibria $\chi_{\mathcal{P}}$.** Snapshots of the densities of the two species at time $t = 0, 0.8, 1.2, 1.6, 2.5, 50$, using the Lax-Friedrichs fluxes (1.42)-(1.43).

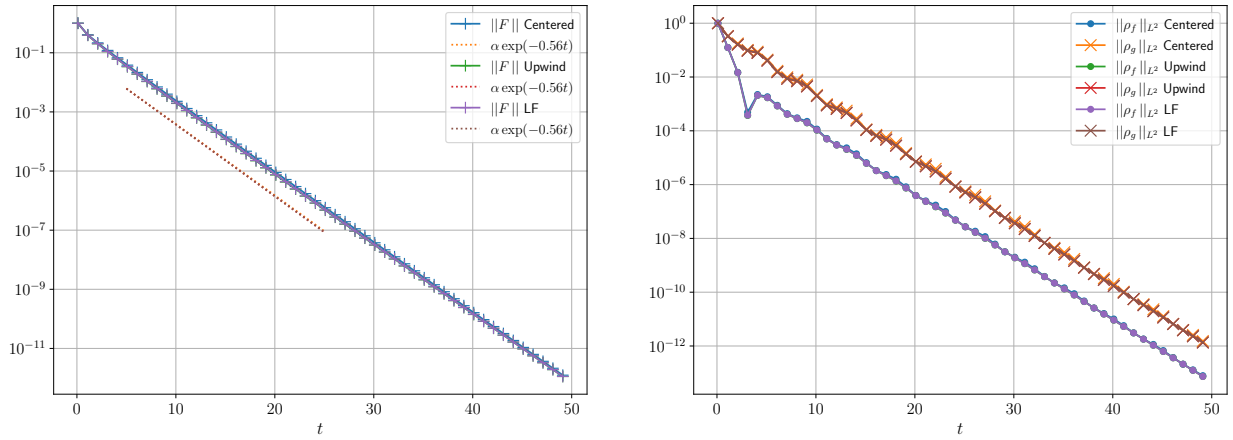


Figure 1.2 – **Test 1. Large time behavior of the linearized scheme, same equilibria $\chi_{\mathcal{P}}$.** Time evolution of the weighted L^2 norm of the solution F to the linearized problem (1.52)-(1.53) (left) and L^2 norm of the densities, for three different flux functions (right).

Test 1. Large time behavior with the same equilibrium. We first present an experiment where the distributions χ_1 and χ_2 are both fixed as the heavy-tailed profile $\chi_{\mathcal{P}}$. We choose as initial data some far from equilibrium distributions given for $x \in [0, \pi]$, $v \in [-v^*, v^*]$ by

$$\begin{cases} f_I(x,v) = \exp\Big(-((x-\pi/2)^2 + v^2/2)/0.2\Big)/0.1, \\ g_I(x,v) = (1+\cos(4x))\exp\Big(-((x-\pi/2)^2 + v^2/2)/0.2\Big). \end{cases} \tag{1.121}$$

We observe in Figure 1.1 that the densities associated with F both converge nicely toward the global equilibrium, without spurious oscillations or loss of the global mass difference.

Figure 1.2 presents the hypocoercive behavior of our numerical method. We observe on the left part of the figure that the expected exponential decay of the weighted L^2-norm of the solution F to the linear scheme (1.52)-(1.53) holds. We notice that the rate of decay is independent on the choice of the numerical flux. This is not unexpected, since monotonicity properties were not crucial in our proof of Theorem 2. We shall see in the next subsection that this conclusion is not valid in the nonlinear case, where monotonicity was paramount for obtaining the result.

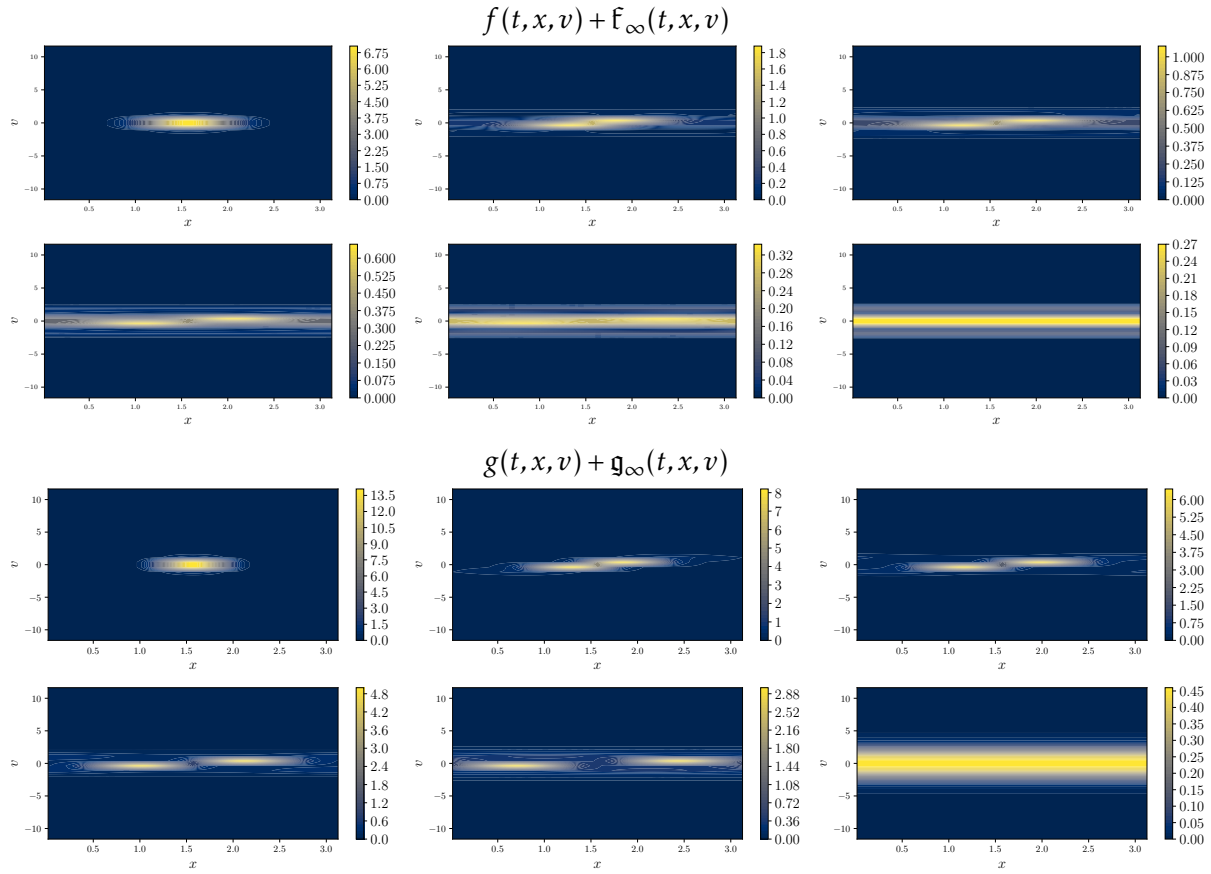


Figure 1.3 – **Test 2. Large time behavior of the linearized scheme, different equilibria.** Snapshots of the distribution function of each species at time $t = 0, 0.8, 1.2, 1.6, 2.5, 50$, using the Lax-Friedrichs fluxes (1.42)-(1.43).

The right part of the figure illustrates the time evolution of the L^2-norms of the densities ρ_f and ρ_g. These objects are not expected to decay monotonically whatsoever. We nevertheless observe an exponential decay of the upper envelope of these quantities, which expresses the rapid equilibration of the densities toward the same global mass, a consequence of the global preservation of the mass difference.

Test 2. Large time behavior with different equilibria. Using the same initial datum (1.121) than in the previous test case, we present now simulations of our scheme with $\chi_1 = \chi_{\mathcal{P}}$ and $\chi_2 = \chi_{\mathcal{O}}$. The use of the two different equilibrium profiles will enrich the dynamics of this linear problem, showcasing the ability of our approach to deal with multiple species of particles.

Figure 1.3 presents the isosurfaces in the (x,v)-plane of the distribution functions $f + \mathfrak{f}_\infty$ and $g + \mathfrak{g}_\infty$ at different times, using the Lax-Friedrichs flux (1.42)-(1.43). We observe the expected relaxation of the initial data of each species towards the correct global equilibrium in large time, without any oscillations or loss of the global mass difference.

The left part of Figure 1.4 presents the large time behavior of the weighted L^2-norm of F, for the three possible choices of flux functions. Once more, we observe the result predicted by Theorem 2, namely exponential decay of this quantity towards 0, even with the heavy-tailed distribution $\chi_{\mathcal{P}}$. The choice of the flux (monotone or not) has no impact on this rate of decay. The right part of this

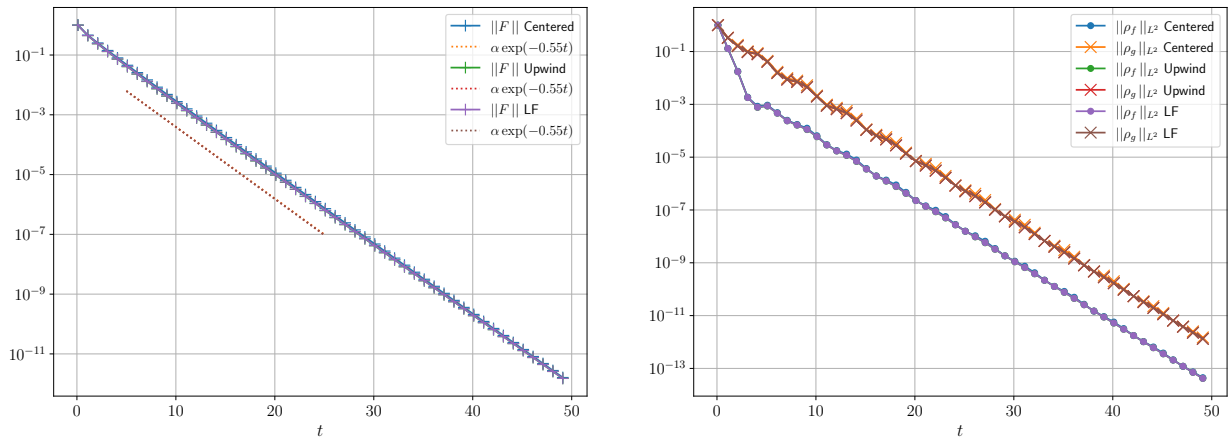


Figure 1.4 – **Test 2. Large time behavior of the linearized scheme, different equilibria.** Time evolution of the weighted L^2 norm of the solution F to the linearized problem (1.52)-(1.53) (left) and L^2 norm of the densities, for three different fluxes.

figure presents the time evolution of the L^2-norms of ρ_f and ρ_g. We observe once more some small oscillations on these quantities, that are not expected to decay monotonically. Nevertheless, a rapid exponential-like convergence towards 0 happens also in this more complicated test case.

1.6.2 Discrete hypocoercivity of the nonlinear scheme

We now present simulations of the full nonlinear scheme (1.40)-(1.41). The numerical parameters are the same as before, except for the time step, that we choose to be adaptive for the sake of stability of the Newton-Raphson solver. The initial time step is set up at $\Delta t = 10^{-3}$, and it is iteratively doubled according to the behavior of the solver. Indeed, the Newton-Raphson solver used to update the solution to the scheme may break the global conservation of mass difference between the species for a large time step when the solution is far from equilibrium if its tolerance is too low. Nevertheless, we observe a rapid growth of the time step during the length of the simulation, up to values of order 0.3. Note that during this refinement process, this value always remains smaller than the critical one needed for monotonicity of the Lax-Friedrichs flux in Theorem 4, hence enforcing the idea that monotonicity is necessary for discrete hypocoercivity.

Test 3. Large time behavior with the same equilibrium. As a first numerical experiment for this nonlinear model we consider the same Gaussian equilibrium $\chi_{\mathcal{M}}$ for both species, and a smooth initial data:

$$\begin{cases} f_I(x,v) = \chi_{\mathcal{M}}(v)\, v^4 (1+\cos(2x)), \\ g_I(x,v) = (1+\cos(4x)) \exp\left(-((x-\pi/2)^2 + v^2/2)/0.2\right). \end{cases}$$

We present in Figure 1.5 some snapshots of the time evolution of the particle distribution function in the (x,v)-phase space, for each species $\mathfrak{f}$ and $\mathfrak{g}$. As in the linear case, we observe a rapid convergence of the far from equilibrium initial data towards the equilibrium distributions $\mathfrak{f}_\infty$ and $\mathfrak{g}_\infty$, without any spurious oscillations or loss of global mass difference.

Figure 1.6 illustrates the nonlinear hypocoercivity result established in Theorem 4. It presents the large time behavior of the weighted L^2-norm of $\mathfrak{F}$, for the three possible choices of flux functions.

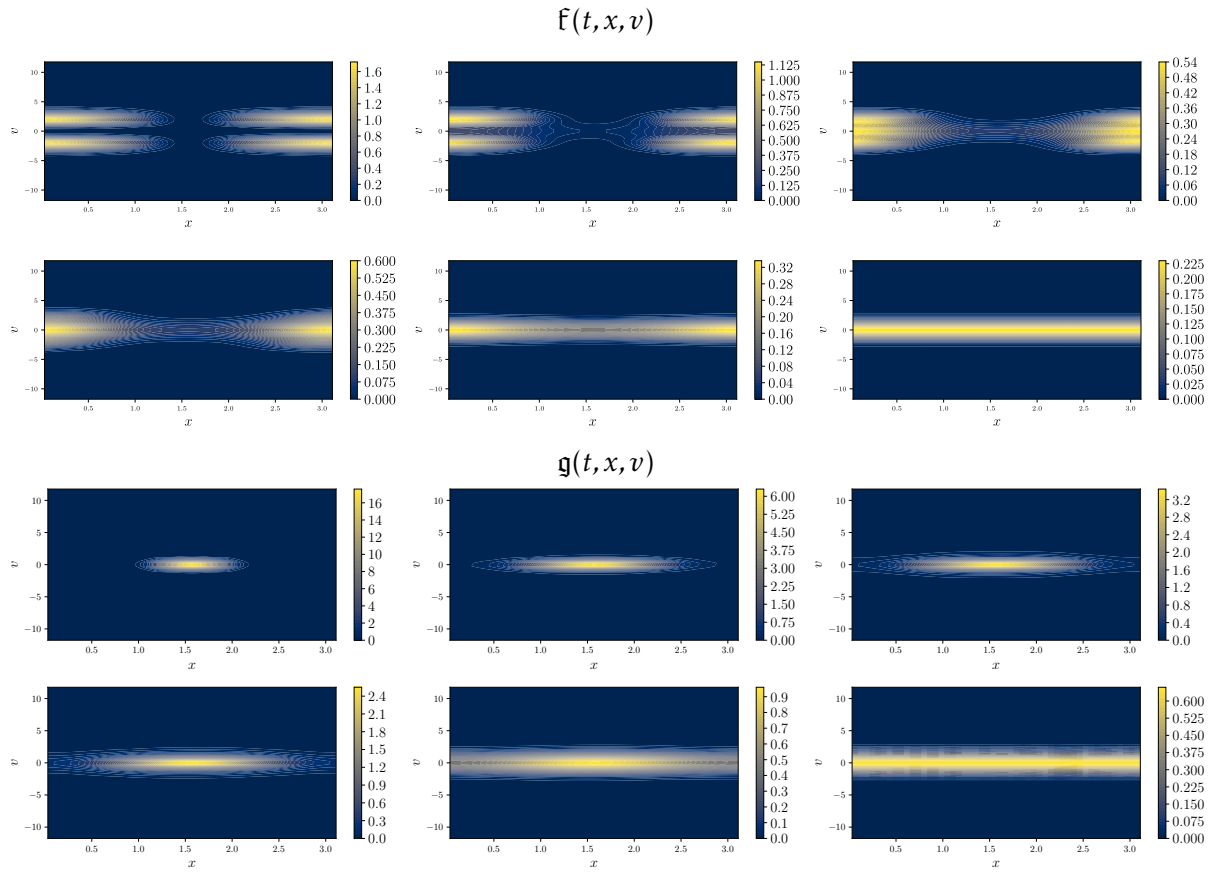


Figure 1.5 – **Test 3. Large time behavior of the nonlinear scheme.** Snapshots of the distribution function of each species at time $t = 0, 0.83, 2.25, 3.35, 9.67, 100$, using the Lax-Friedrichs fluxes (1.42)-(1.43).

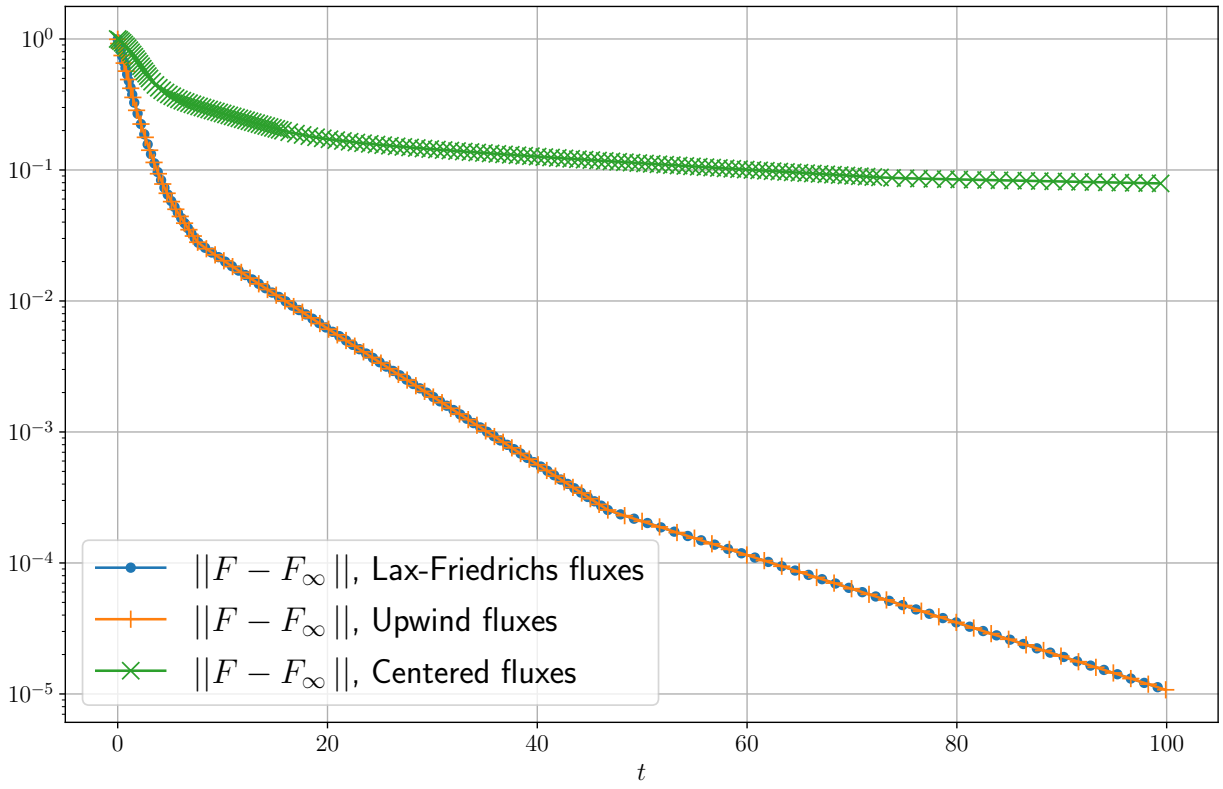


Figure 1.6 – **Test 3. Large time behavior of the nonlinear scheme.** Trends to equilibrium of F towards F_∞ in the weighted L^2-norm, for the monotone fluxes.

We observe in that particular numerical experiment a huge departure from the simpler linear setting, as well as a confirmation of the fact that monotone fluxes would be an optimal assumption in that result. Indeed, one can see on that figure that the exponential decay (1.106) of $\|\mathsf{F}^n - \mathsf{F}^\infty\|_\Delta$ occurs only for the monotone Lax-Friedrichs (1.42)-(1.43) and upwind (1.50)-(1.51) fluxes, and not for the nonmonotone centered fluxes. While the former exhibits the expected exponential decay, the latter will saturate at a value which is dictated by the loss of global mass difference induced by the choice of flux.

Test 4. Large time behavior with different equilibria. In order to showcase the robustness of our scheme, we finally choose as an initial condition a fully random uniform initial data for both species. Note here that such initial data does not fulfil the bounds needed in Theorem 4 for the result to hold, making them seemingly purely technical. The profiles chosen in this test are the heavy-tailed $\chi_1 = \chi_{\mathcal{P}}$ and $\chi_2 = \chi_{\mathcal{O}}$.

We present in Figure 1.7 some snapshots of the time evolution of the particle distribution function in the (x, v)-phase space, for each species f and g. We observe again a rapid convergence of the far from equilibrium initial data towards the equilibrium distributions f_∞ and g_∞. No spurious oscillations appear with the monotone Lax-Friedrichs fluxes and the global mass difference is preserved up to machine precision in this test.

These observations are confirmed in Figure 1.8, where we present the time evolution of $\|\mathsf{F}^n - \mathsf{F}^\infty\|_\Delta$. Note that this far from equilibrium case is not covered by Theorem 4. Nevertheless, this quantity exhibits a fast exponential decay towards 0, but only for monotone fluxes. We did not

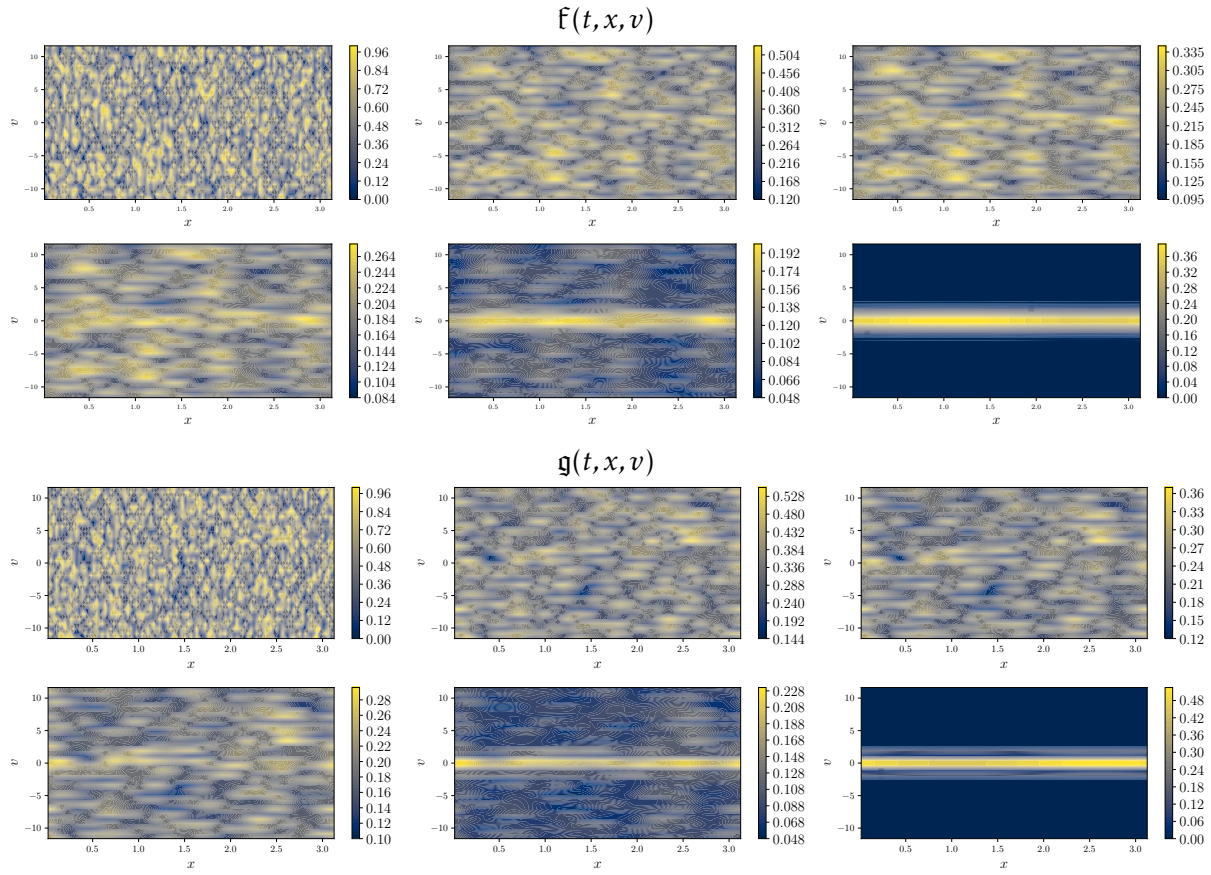


Figure 1.7 – **Test 4. Large time behavior of the nonlinear scheme.** Snapshots of the distribution function of each species at time $t = 0, 0.16, 0.38, 0.77, 1.66, 49.9$, using the Lax-Friedrichs fluxes (1.42)-(1.43).

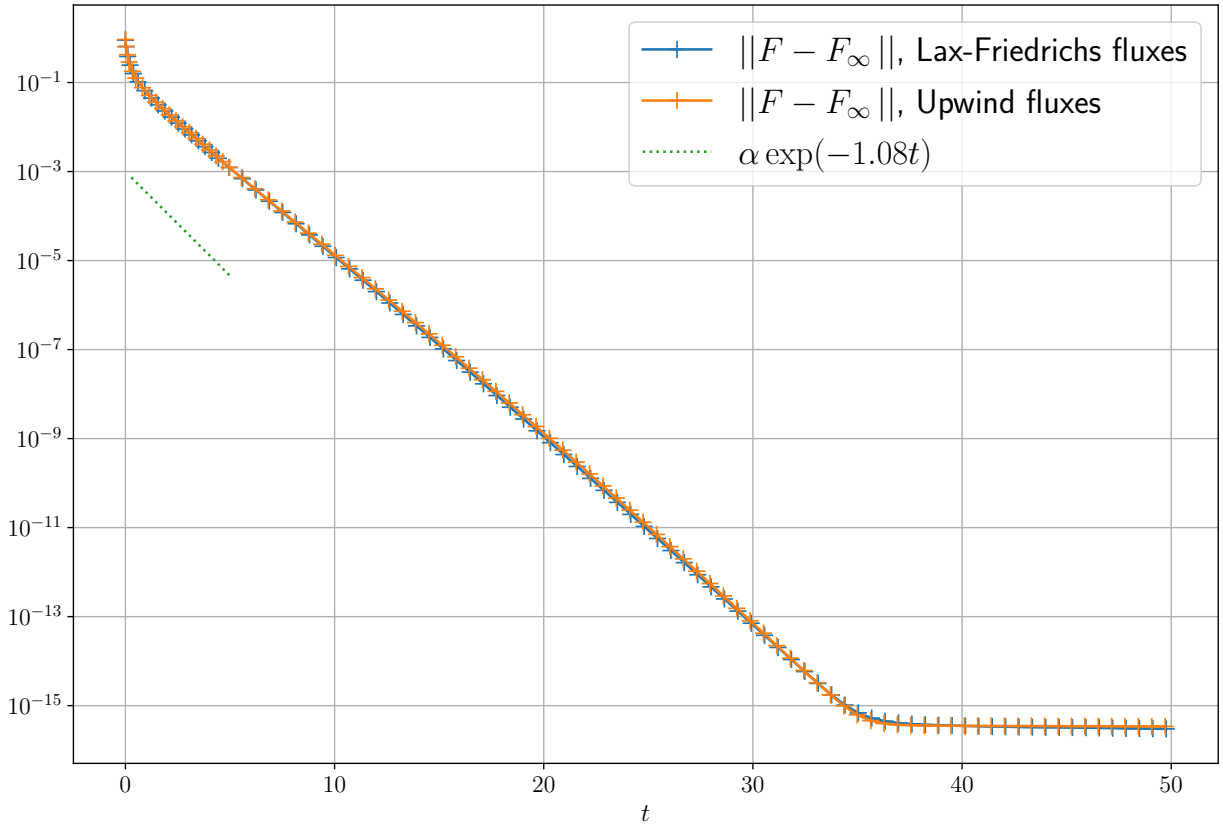


Figure 1.8 – **Test 4. Large time behavior of the nonlinear scheme.** Trends to equilibrium of F towards F_∞ in the weighted L^2-norm, for the monotone fluxes.

present the case of centered fluxes, because we were not able to have a convergent Newton solver in that particular case. This is certainly due to the loss of mass induced by nonmonotone fluxes with such a singular initial data. This test case hence emphasizes our claim about the nonoptimality of the bounds needed in Theorem 3.

Chapter 2

Conservative polynomial approximations and applications to Fokker-Planck equations

In this chapter, I consider the problem of constructing approximations based on orthogonal polynomials that preserve an arbitrary set of moments of a given function without losing the spectral convergence property. To this aim, the constrained polynomial of best approximation for a generic basis of orthogonal polynomials can be computed. The construction is entirely general and allows to derive structure preserving numerical methods for partial differential equations.
The content of this chapter covers the submitted work :

— T. Laidin and L. Pareschi. "Conservative polynomial approximations and applications to Fokker-Planck equations". Preprint arXiv:2402.06473. 2024

Outline of the current chapter

2.1 Introduction

Computational techniques that maintain certain fundamental properties or structures of the underlying mathematical problem in their discrete approximations play a crucial role in the study

and analysis of ODEs and PDEs, offering insights into complex physical phenomena that are often inaccessible through analytical means alone [111, 119, 135]. These structure-preserving properties could include symmetries, conservation laws, or other structural characteristics that are crucial for accurately representing the behavior of the system being modelled. In recent years, such methods, referred to as structure-preserving, have emerged as new paradigm in numerical solution of PDEs being able of accurately capturing the behavior of the solution while maintaining the structural properties of the equations.

Capturing the long-time behavior of solutions to PDEs is also intricately linked to the structure-preserving properties of numerical methods. In fact, preserving key structural properties at a discrete level, such as conservation laws and invariant quantities, ensures that the numerical approximation retains the essential characteristics of the original problem over extended time intervals.

In this regard, the preservation of moments of the solution stands out as a fundamental aspect, especially when considering kinetic and mean-field equations, where the moments represent the macroscopic observable physical quantities [21, 67, 147, 183]. Regarding kinetic equations, for example, the long time behavior of the systems is fully determined by the knowledge of some invariant moments. Achieving such properties, however, is particularly challenging due to the inherent complexity of many physical systems and the necessity to accurately capture these dynamics.

A substantial body of literature has been devoted to the development of structure-preserving methods for various types of PDEs, with a focus on preserving specific physical properties. For instance, in the context of Fokker-Planck equations, numerous studies have explored different numerical strategies aimed at maintaining conservation of mass, momentum and energy with the aim to describe accurately the steady state solution of the problem [8, 37, 48, 152, 168, 183].

Spectral methods, which rely on expansions in terms of orthogonal polynomials, have garnered significant attention for their effectiveness in solving collisional kinetic equations of Boltzmann-type. The spectral accuracy and efficiency of these methods make them particularly well-suited for capturing the non equilibrium dynamics of such systems [132, 169, 170, 176, 181, 182]. However, due to the lack of conservations, the long time behavior of such methods may lead to accumulation of errors and their extension to a conservative setting has represented a challenging task in numerical simulations. Recent advancements in conservative spectral methods based on L^2-minimization frameworks have shown promise in addressing these challenges [4, 96, 97, 177].

Building upon these developments, our main goal in this manuscript is to extend the L^2-minimization setting to derive families of orthogonal polynomials capable of preserving the moments of the solution, thereby enhancing the accuracy and efficiency of structure-preserving numerical methods. To this end, we draw upon the well established theory of orthogonal polynomials, which provides a powerful framework for constructing accurate approximations of PDEs [43, 91, 102, 113, 144]. In particular, we have developed a general framework for constructing constrained orthogonal polynomial approximations which maintain the accuracy properties of the unconstrained polynomials for smooth solutions. This framework enabled us to construct spectrally accurate moment-preserving Galerkin-type approximations for several Fokker-Planck equations originating from various fields, ranging from classical physics to social sciences and

operations research. Consequently, these equations are defined in different domains, both bounded and unbounded, and possess different steady states, requiring the adoption of suitable orthogonal polynomial bases.

Let us also remark, that the connection between orthogonal polynomials and probability theory to design uncertainty quantification methods further underscores the importance of our approach, as classical families of continuous and discrete orthogonal polynomials are intimately linked to probability distributions and our considerations naturally generalize to such contexts [205].

In the following sections, we provide a detailed exposition of our methodology and present numerical results demonstrating the efficacy of our approach. The rest of the chapter is organized as follows. In Section 2, we introduce some notations and present our moment-preserving approach based on a constrained L_2-minimization setting. We also study the convergence properties and prove a general result on spectral accuracy. Section 3 is then devoted to testing the novel polynomial approximation, first by approximating probability densities with given moments and subsequently by considering different Fokker-Planck equations on bounded and unbounded domains. Some concluding remarks are provided in the last section.

2.2 Conservative approximations by orthogonal polynomials

In order to get conservative approximations, we construct a conservative projection on the space of orthogonal polynomials by generalizing the constrained formulation approach introduced in [177] for trigonometric polynomials. In particular, we will give an explicit formulation of the orthogonal polynomial of best approximation in the weighted least square sense, constrained by preservation of moments, and show that it preserves spectral accuracy for smooth solutions like the classical orthogonal polynomial approximation.

2.2.1 Orthogonal polynomials

Let us first set up the mathematical framework of our analysis. To simplify our treatment we will consider the one-dimensional case. Extension to the multidimensional setting can be handled via tensorization. The context of multivariate polynomials is not investigated here but could represent an interesting development. Given a real function $f(x) \in L^2_\omega(\Omega)$, with $\Omega \subseteq \mathbb{R}$ and $\omega(x) > 0$ a positive function, we define

$$\|f\|_{L^2_\omega} = \left(\int_\Omega |f(x)|^2 \omega(x)\, dx \right)^{1/2}, \tag{2.1}$$

which has the associated inner product

$$\langle f, g \rangle_\omega = \int_\Omega f(x) g(x) \omega(x)\, dx. \tag{2.2}$$

We assume that $\omega(x) > 0$ is such that

$$\int_\Omega |x|^q \omega(x)\, dx < \infty, \quad q = 0, 1, 2, \dots \tag{2.3}$$

We consider the problem of approximating functions in the weighted $L^2_\omega(\Omega)$ space defined by the norm (2.1). In the case $\omega(x) = 1$ we will use standard notations $\|\cdot\|_{L^2}$ and $\langle\cdot,\cdot\rangle$ to denote the L^2 norm and the associated inner product.

A polynomial basis $\{\phi_k\}$, $k = 0, 1, \ldots$ for $L^2_\omega(\Omega)$ is a set of functions such that any f in the space can be expressed uniquely as

$$f(x) = \sum_{k=0}^{\infty} \hat{f}_k \phi_k(x). \tag{2.4}$$

A set of polynomials $\{\phi_k\}$ is called *orthogonal* if $\langle\phi_h, \phi_k\rangle_\omega = 0$ for $h \neq k$. We refer to Table 2.1 for some examples of standard polynomial families.

Orthogonal bases are particularly nice, both for theory and numerical approximation. In fact, from

$$\langle f, \phi_k\rangle_\omega = \sum_{h=0}^{\infty} \hat{f}_h \langle\phi_h, \phi_k\rangle_\omega = \hat{f}_k \langle\phi_k, \phi_k\rangle_\omega,$$

the coefficients in (2.4) are easily represented as

$$\hat{f}_k = \frac{\langle f, \phi_k\rangle_\omega}{\langle\phi_k, \phi_k\rangle_\omega} = \frac{\langle f, \phi_k\rangle_\omega}{\|\phi_k\|^2_{L^2_\omega}}. \tag{2.5}$$

A first result that we recall is that the truncated approximation

$$f_N(x) = \sum_{k=0}^{N} \hat{f}_k \phi_k(x) \tag{2.6}$$

is the best approximation of f in the subspace $S_N = \text{span}\{\phi_k \,|\, k = 0, 1, \ldots, N\}$. More precisely, let $\mathcal{P}_{\omega,N} : L^2_\omega(\Omega) \to S_N$ be the orthogonal projection upon S_N in the inner product of $L^2_\omega(\Omega)$

$$\langle f - \mathcal{P}_{\omega,N} f, \phi_k\rangle_\omega = 0, \qquad \forall\, \phi_k \in S_N. \tag{2.7}$$

With these definitions, $\mathcal{P}_{\omega,N} f = f_N$ is the solution of the following minimization problem

$$f_N = \text{argmin}\left\{\|g_N - f\|_{L^2_\omega} \,:\, g_N \in S_N\right\}.$$

By orthogonality, one also obtains the Parseval's identities

$$\|f\|^2_{L^2_\omega} = \sum_{k=0}^{\infty} |\hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega}, \qquad \|f_N\|^2_{L^2_\omega} = \sum_{k=0}^{N} |\hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega},$$

which gives the error estimate

$$\|f - f_N\|^2_{L^2_\omega} = \sum_{k=N+1}^{\infty} |\hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega}.$$

An important feature of the orthogonal polynomial approximations on S_N is related to their spectral convergence properties for smooth solutions. We report here a result for Jacobi type approximations [91].

Theorem 5. *If $f \in H^r_\omega([-1,1])$, where $r \geq 0$ is an integer, then there exists a constant $C > 0$ only dependent on α, β and r such that*

$$\|f - f_N\|_{L^2_\omega} \leq \frac{C}{N^r} \left\| (1-x^2)^{r/2} \frac{d^r f}{dx^r} \right\|_{L^2_\omega}, \qquad N > r. \tag{2.8}$$

This type of convergence corresponds to the fact that the more the function is regular, that is $f \in H^r_\omega([-1,1])$ with r large, the faster the approximation error converges towards 0.

Table 2.1 – Families of orthogonal polynomials

Name	$\omega(x)$	Ω
Jacobi	$(1-x)^\alpha (1+x)^\beta$, $\alpha, \beta > -1$	$[-1,1]$
- Legendre	1, $\alpha = \beta = 0$	$[-1,1]$
- Chebyshev 1st kind	$\frac{1}{\sqrt{1-x^2}}$, $\alpha = \beta = -\frac{1}{2}$	$[-1,1]$
- Chebyshev 2nd kind	$\sqrt{1-x^2}$, $\alpha = \beta = \frac{1}{2}$	$[-1,1]$
Laguerre	$x^\alpha e^{-x}$, $\alpha > -1$	$\mathbb{R}^+$
Hermite	$\lvert x\rvert^{2\alpha} e^{-x^2}$, $\alpha > -1$	$\mathbb{R}$

2.2.2 Approximation of moments

Next, we discuss some properties of the projection operator $\mathcal{P}_{\omega,N}$, in particular those concerning approximation of moments of nonnegative functions. We will denote for $q = 0, 1, 2, \ldots$

$$m_q = \int_\Omega f(x) x^q \, dx = \langle f, x^q \rangle, \qquad m_{q,N} = \int_\Omega f_N(x) x^q \, dx = \langle f_N, x^q \rangle. \tag{2.9}$$

Note that

$$m_q = \sum_{k=0}^{\infty} \hat{f}_k \mu_{q,k}, \qquad m_{q,N} = \sum_{k=0}^{N} \hat{f}_k \mu_{q,k},$$

where for $k, q \geq 0$

$$\mu_{q,k} = \langle \phi_k, x^q \rangle, \tag{2.10}$$

is the q-th moment of the k-th order polynomial.

It is well-known that moments of the weight function $\omega(x)$ permits to generate explicitly the orthogonal polynomials [144]. In our case, however, we are interested in the behavior of the moments of the orthogonal polynomials themselves defined in (2.10). An important characterization of monic orthogonal polynomials is the classical three-term recurrence relation [91, 102, 113, 144]

$$\phi_k(x) = (a_k x + b_k)\phi_{k-1}(x) + c_k \phi_{k-2}(x), \quad k \geq 1, \tag{2.11}$$

where we assumed $\phi_0(x) = 1$, $\phi_{-1}(x) = 0$ and the sequences $\{a_k\}_{k\geq 1}$, $\{b_k\}_{k\geq 1}$ and $\{c_k\}_{k\geq 1}$ depend on the particular polynomial family under consideration.

Equation (2.11) permits to recursively compute the value of $\phi_k(x)$ starting from the values of $\phi_0(x)$ and $\phi_{-1}(x)$.

From (2.11) one gets the following relations for the moments coefficients

$$\mu_{q,k} = a_k \mu_{q+1,k-1} + b_k \mu_{q,k-1} + c_k \mu_{q,k-2}, \quad k \geq 1, \quad q \geq 0, \tag{2.12}$$

with

$$\mu_{q,0} = \frac{x^{q+1}}{q+1}\Big|_\Omega, \quad \mu_{q,-1} = 0, \quad q \geq 0,$$

which, therefore, can also be computed recursively from (2.12) up to the moment q provided we know $q+h$ moments of $\phi_{k-h}(x)$ for $h = 1, \ldots, k$.

Unbounded domains

Note, however, that for unbounded domains Ω, like the case of Laguerre or Hermite polynomials (see Table 2.1), the moments $\mu_{q,0}$, $q \geq 0$, are not finite. In such cases, one resorts on suitable orthogonal function families with respect to the inner product $\langle \cdot, \cdot \rangle$ defined in the symmetrically weighted case [21, 127, 164] as

$$P_k(x) = \phi_k(x) w^{1/2}(x),$$

so that $\langle P_h, P_k \rangle = \langle \phi_h, \phi_k \rangle_\omega$, $\forall\, h, k \geq 0$.

For functions that we write in the form $f(x) = h(x) w^{1/2}(x)$ we have

$$f(x) = \sum_{k=0}^{\infty} \hat{h}_k P_k(x), \tag{2.13}$$

with

$$\hat{h}_k = \frac{\langle f, P_k \rangle}{\langle P_k, P_k \rangle} = \frac{\langle h, \phi_k \rangle_\omega}{\|\phi_k\|^2_{L^2_\omega}}. \tag{2.14}$$

Clearly, the orthogonal functions P_k satisfy an analogous three term relation as (2.11) with $P_0(x) = w^{1/2}(x)$, and similarly their moments satisfy (2.12) with

$$\langle P_0, x^q \rangle = \frac{x^{q+1}}{q+1} w^{1/2}(x)\Big|_\Omega, \quad \mu_{q,-1} = 0, \quad q \geq 0.$$

The above quantity is bounded for Laguerre and Hermite polynomials thanks to (2.3).

We can then define $f_N(x)$ as the truncated approximation to the first $N+1$ terms of (2.13). We have

$$\langle f - f_N, P_k \rangle = 0, \qquad \forall\, P_k = \phi_k w^{1/2},\ \phi_k \in S_N. \tag{2.15}$$

In the weighted function family representation, Parseval identity reads

$$\|f\|_2 = \sum_{k=0}^{\infty} |\hat{h}_k|^2 \|P_k\|_2^2.$$

Furthermore, we report here a result for Hermite approximations [90, 91].

Theorem 6. *If $f \in H^r_\omega(\Omega)$, where $r \geq 0$ is an integer, then there exist a constant $C > 0$ such that*

$$\|f - f_N\|_{L^2_\omega} \leq \frac{C}{N^{r/2}} \|f\|_{H^r_\omega}, \qquad N > r. \tag{2.16}$$

Let us remark that, in general, moments are not preserved by the projection. Neither in standard nor in the weighted case. In fact, we have

$$m_q - m_{q,N} = \langle f - f_N, x^q \rangle = \left\langle f - f_N, \frac{x^q}{\omega(x)} \right\rangle_\omega \tag{2.17}$$

which, by (2.7), is equal to zero only if $x^q/\omega(x)$ belongs to S_N. A special case is represented by the Legendre polynomials which by construction preserve all moments of the approximated function $f(x)$ up to the order $q \leq N$.

Finally, we have for each $\varphi \in L^2_\omega(\Omega)$ by Cauchy-Schwarz inequality

$$|\langle f, \varphi \rangle_\omega - \langle f_N, \varphi \rangle_\omega| \leq \|\varphi\|_{L^2_\omega} \|f - f_N\|_{L^2_\omega}. \tag{2.18}$$

Therefore, the projection error on the moments decays faster than algebraically, for the weighted norms, when the solution is infinitely smooth. Although this is in general a guarantee of the accuracy of the methods, in practice the loss of conservations can lead to accumulations of error in the approximation of PDEs over long time scales, resulting in the determination of inaccurate steady states [111, 183].

In the following, we focus on the context and formalism of bounded domains. The same computations can be done in the unbounded case by considering the family of functions P_k instead of the polynomials ϕ_k.

2.2.3 Moment preserving orthogonal polynomials

Although the error on moments is spectrally small for smooth functions, in many applications moments represent structural properties of the PDE and are related to conservation of physical properties such as mass, momentum and energy. In such cases, the design of a numerical method that is structure preserving requires the construction of a projection operator that does not affect the relevant moments. To this end, we will construct a different projection operator on the space of polynomials, $\mathcal{P}^c_{\omega,N} : L^2_\omega(\Omega) \to S_N$ such that it preserves a finite number of moments. More precisely, we require the projection to satisfy for $q = 0, 1, \ldots, Q$

$$\left\langle \mathcal{P}^c_{\omega,N} f, x^q \right\rangle = \langle f, x^q \rangle,$$

but maintaining the convergence properties of the original orthogonal projection.

To this aim, it is natural to consider the following constrained best approximation problem

$$f^c_N = \operatorname{argmin} \left\{ \|g_N - f\|^2_{L^2_\omega} : g_N \in S_N,\ \langle g_N, x^q \rangle = \langle f, x^q \rangle,\ q = 0, 1, \ldots, Q \right\}. \tag{2.19}$$

Now, since $g_N \in S_N$ we can represent it in the form

$$g_N(x) = \sum_{k=0}^{N} \hat{g}_k \phi_k(x)$$

and then by Parseval's identity

$$\|g_N - f\|^2_{L^2_\omega} = \sum_{k=0}^{\infty} |\hat{g}_k - \hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega},$$

where we assumed $\hat{g}_k = 0$, $k > N$.

We require that

$$\langle g_N, x^q \rangle = \sum_{k=0}^{N} \hat{g}_k \mu_{q,k} = \langle f, x^q \rangle, \quad q = 0, 1, \ldots, Q \tag{2.20}$$

or, after introducing the vector of moments $\Phi = (1, x, x^2, \ldots, x^Q)^T$, in vector form

$$\langle g_N, \Phi \rangle = \sum_{k=0}^{N} \hat{g}_k \hat{\Phi}_k = \langle f, \Phi \rangle,$$

where $\hat{\Phi}_k = (\mu_{0,k}, \mu_{1,k}, \ldots, \mu_{Q,k})^T$.

Let us now solve the minimization problem (2.19) using the Lagrange multiplier method. Let $\lambda \in \mathbb{R}^{Q+1}$ be the vector of Lagrange multipliers, we consider the objective function

$$\mathsf{L}(\hat{g}, \lambda) = \sum_{k=0}^{\infty} |\hat{g}_k - \hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega} + \lambda^T \left(\sum_{k=0}^{N} \hat{g}_k \hat{\Phi}_k - \langle f, \Phi \rangle \right) = 0,$$

with $\hat{g} \in \mathbb{R}^{N+1}$ the vector of coefficients $\hat{g}_k$, $k = 0, \ldots, N$.

Stationary points are found by imposing

$$\frac{\partial \mathsf{L}(\hat{g}, \lambda)}{\partial \hat{g}_k} = 0, \quad k = 0, 1, \ldots, N; \qquad \frac{\partial \mathsf{L}(\hat{g}, \lambda)}{\partial \lambda_q} = 0, \quad q = 0, 1, \ldots, Q.$$

From the first condition, one gets

$$2(\hat{g}_k - \hat{f}_k)\|\phi_k\|^2_{L^2_\omega} + \lambda^T \hat{\Phi}_k = 0, \tag{2.21}$$

whereas the second condition corresponds to (2.20).

Multiplying the above equation by $\hat{\Phi}_k / \|\phi_k\|^2_{L^2_\omega}$ and summing up over k we can write

$$2 \sum_{k=0}^{N} (\hat{g}_k - \hat{f}_k) \hat{\Phi}_k + \sum_{k=0}^{N} \frac{1}{\|\phi_k\|^2_{L^2_\omega}} \hat{\Phi}_k \hat{\Phi}_k^T \lambda = 0$$

and, using (2.20) as well as the fact that $\hat{\Phi}_k \hat{\Phi}_k^T$ are symmetric and positive definite matrices of size

$Q+1$ one obtains

$$\lambda = -2\left(\sum_{k=0}^{N} \frac{1}{\|\phi_k\|^2_{L^2_\omega}} \hat{\Phi}_k \hat{\Phi}_k^T\right)^{-1} (\langle f, \Phi\rangle - \langle f_N, \Phi\rangle). \tag{2.22}$$

Now, rewriting the first condition (2.21) as

$$\hat{g}_k = \hat{f}_k - \frac{1}{2\|\phi_k\|^2_{L^2_\omega}} \hat{\Phi}_k^T \lambda,$$

and plugging the expression (2.22) of λ in it, one obtains that the minimum is achieved for $\hat{g}_k = \hat{f}_k^c$, given by the following definition.

Definition 1. *For $f \in L^2_\omega(\Omega)$, we define the* conservative orthogonal projection $\mathcal{P}^c_{N,\omega} f = f_N^c$ *in S_N, where f_N^c is given by*

$$f_N^c(x) = \sum_{k=0}^{N} \hat{f}_k^c \phi_k(x), \tag{2.23}$$

and we define the moment constrained coefficients *as*

$$\hat{f}_k^c = \hat{f}_k + \hat{C}_k^T (\langle f, \Phi\rangle - \langle f_N, \Phi\rangle), \qquad \hat{C}_k^T = \frac{1}{\|\phi_k\|^2_{L^2_\omega}} \hat{\Phi}_k^T \left(\sum_{h=0}^{N} \frac{1}{\|\phi_h\|^2_{L^2_\omega}} \hat{\Phi}_h \hat{\Phi}_h^T\right)^{-1}. \tag{2.24}$$

The following result states the spectral accuracy of the conservative best approximation in the least square sense (2.23), and generalizes Theorem 5. In order to prove this result, one needs the following hypothesis:

Hypothesis 1. *The moments of the approximated solution are spectrally accurate in the classical L^2-norm. Namely, there exists $C > 0$ such that*

$$\|U - U_N\| \leq \frac{C}{N^r} \|f\|_{H^r_\omega} \|\Phi\|_{2,L^2_\omega}, \tag{2.25}$$

where U and U_N are the moment vectors of f and f_N:

$$U = (m_1, \dots, m_Q), \quad U_N = (m_{1,N}, \dots, m_{Q,N}),$$

and the norm $\|\cdot\|_{2,L^2_\omega}$ is defined by:

$$\|\Phi\|^2_{2,L^2_\omega} = \sum_{q=0}^{Q} \|\Phi_q\|^2_{L^2_\omega}$$

This assumption is different from the spectral accuracy on the moment, in weighted norm, one could obtain from (2.18). This property depends heavily on the function under consideration, and we will see in numerical simulations that it may only be a technical detail.

Theorem 7. *If $f \in H^r_\omega(\Omega)$, where $r \geq 0$ is an integer, we have*

$$\|f - f_N^c\|_{L^2_\omega} \leq \frac{C_\Phi}{N^r} \|f\|_{H^r_\omega} \tag{2.26}$$

where the constant C_Φ depends on the spectral radius of the matrix

$$M = \sum_{h=0}^{N} \frac{1}{\|\phi_h\|^2_{L^2_\omega}} \hat{\Phi}_h \hat{\Phi}_h^T,$$

and on $\|\Phi\|^2_{2,L^2_\omega}$ where Φ_q, $q = 0, 1, \ldots, Q$ are the components of the vector Φ.

Proof. We can split

$$\|f - f_N^c\|^2_{L^2_\omega} \leq \|f - f_N\|^2_{L^2_\omega} + \|f_N - f_N^c\|^2_{L^2_\omega}.$$

The first term is bounded by the spectral estimate of truncated approximation of the form (2.8), whereas for the second term, by Parseval's identity, we have

$$\|f_N^c - f_N\|^2_{L^2_\omega} = \sum_{k=0}^{N} |\hat{f}_k^c - \hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega}.$$

Now, using the definition (2.24) we get

$$\sum_{k=0}^{N} |\hat{f}_k^c - \hat{f}_k|^2 \|\phi_k\|^2_{L^2_\omega} = \sum_{k=0}^{N} |\hat{C}_k^T (U - U_N)|^2 \|\phi_k\|^2_{L^2_\omega} \leq \|U - U_N\|_2^2 \sum_{k=0}^{N} \|\hat{C}_k^T\|_2^2 \|\phi_k\|^2_{L^2_\omega},$$

where $\|\cdot\|_2$ denotes the euclidean norm of the vector. The Hypothesis 1 of spectral accuracy, in classical L^2 norm, on the moments implies

$$\|U - U_N\|_2 \leq \frac{C}{N^r} \|f\|_{H^r_\omega} \left(\sum_{q=0}^{Q} \|\Phi_q\|^2_{L^2_\omega} \right)^{1/2} = \frac{C}{N^r} \|f\|_{H^r_\omega} \|\Phi\|_{2,L^2_\omega}$$

where for $\Phi = (\Phi_0, \ldots, \Phi_Q)^T$ we defined $\|\Phi\|^2_{2,L^2_\omega} = \sum_{q=0}^{Q} \|\Phi_q\|^2_{L^2_\omega}$. Finally, by definition of $\hat{C}_k^T = \frac{1}{\|\phi_k\|^2_{L^2_\omega}} \hat{\Phi}_k^T \left(\sum_{h=0}^{N} \frac{1}{\|\phi_h\|^2_{L^2_\omega}} \hat{\Phi}_h \hat{\Phi}_h^T \right)^{-1}$, one has

$$\sum_{k=0}^{N} \|\hat{C}_k^T\|_2^2 \|\phi_k\|^2_{L^2_\omega} \leq \sum_{k=0}^{N} \|\hat{\Phi}_k\|_2^2 \left\| \left(\sum_{h=0}^{N} \frac{1}{\|\phi_h\|^2_{L^2_\omega}} \hat{\Phi}_h \hat{\Phi}_h^T \right)^{-1} \right\|_2^2.$$

Now let $M = \sum_{h=0}^{N} \hat{\Phi}_h \hat{\Phi}_h^T / \|\phi_h\|^2_{L^2_\omega}$. It follows that

$$\sum_{k=0}^{N} \|\hat{C}_k^T\|_2^2 \|\phi_k\|^2_{L^2_\omega} \leq \|\Phi\|^2_{2,L^2_\omega} \left\| M^{-1} \right\|_2^2. \tag{2.27}$$

As a sum of positive definite symmetric matrices, M enjoys the same properties. Consequently,

$$\sum_{k=0}^{N} \|\hat{C}_k^T\|_2^2 \|\phi_k\|^2_{L^2_\omega} \leq \|\Phi\|^2_{2,L^2_\omega} \rho^2(M^{-1}), \tag{2.28}$$

where $\rho(M^{-1})$ denotes the spectral radius of the matrix M^{-1}. Combining everything, one finally

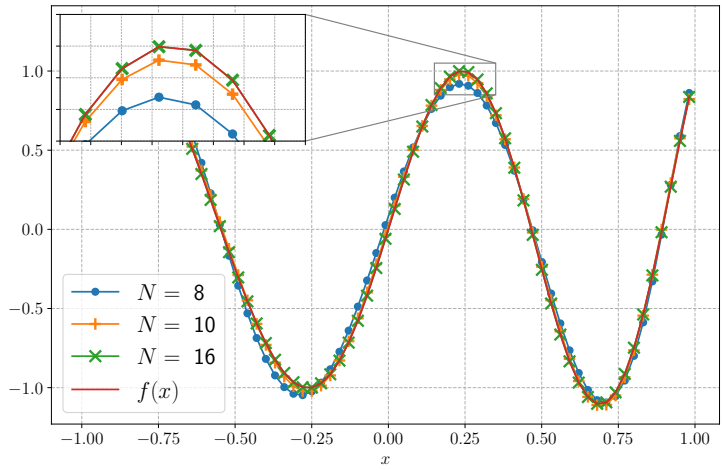


Figure 2.1 – Constrained Legendre approximation for (2.30).

obtains

$$\|f_N^c - f_N\|_{L^2_\omega} \leq \frac{C}{N^r}\|f\|_{H^r_\omega}\|\Phi\|_{2,L^2_\omega}\rho(M^{-1})$$

which proves (2.26). □

Remark 4. *The conservative best approximation in least square* (2.24) *can be represented in terms of the standard projection as*

$$\mathcal{P}_N^c f = \mathcal{P}_N f + \sum_{k=0}^{N} \hat{C}_k^T \langle f - \mathcal{P}_N f, \Phi\rangle. \tag{2.29}$$

2.3 Numerical examples and applications

In this section, we conduct numerical experiments to illustrate the properties of the proposed method. The initial test focuses on evaluating the accuracy of the moment-preserving approach. Subsequently, in Tests 2 to 4, we employ the conservative method within a PDE framework, considering scenarios involving both bounded and unbounded domains. The ensuing analysis includes the assessment of the numerical scheme's accuracy, its conservation properties, and an exploration of the long-time behavior of solutions.

2.3.1 Test 1: Approximation of functions

Bounded domain $[-1,1]$. Let us consider the approximation of the oscillating and non-symmetric function

$$f(x) = \sin(2\pi x) + x^2\cos(2\pi x), \quad x \in [-1,1]. \tag{2.30}$$

We first consider a standard approximation using Legendre polynomials. Figure 2.1 illustrates the approximate solutions and Table 2.2 shows that the standard method is indeed spectrally accurate. In addition, we observe that moments are indeed preserved without constraining the projection as predicted by (2.17). Note that errors are sometimes extremely close to 0 but not exactly 0 because the Gaussian quadrature used may introduce small machine epsilons.

N	$\|f - f_N\|_2$	$\|m_0 - m_{0,N}\|$	$\|m_1 - m_{1,N}\|$	$\|m_2 - m_{2,N}\|$	$\|m_3 - m_{3,N}\|$
8	9.004e-03	2.429e-15	2.387e-15	2.387e-15	2.331e-15
16	2.197e-05	1.388e-17	0	0	0
32	1.050e-13	8.327e-17	1.110e-16	0	0

Table 2.2 – Error Analysis: Legendre approximation for (2.30).

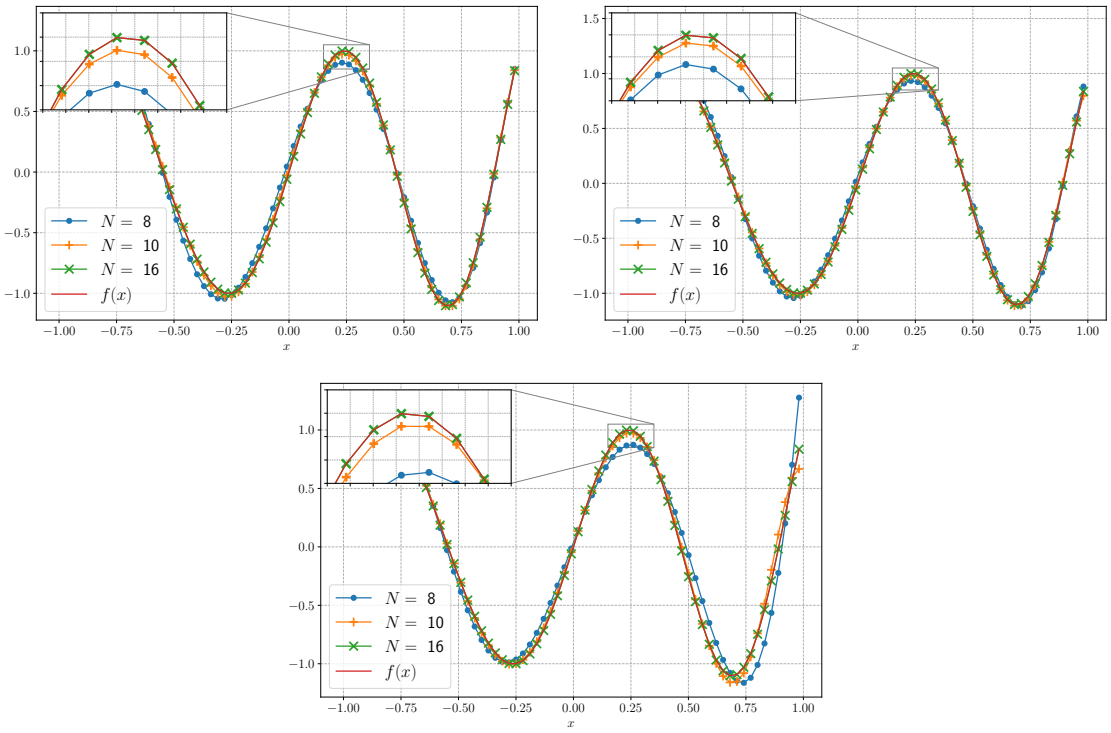


Figure 2.2 – Constrained approximations for Chebyshev 1st kind (Top Left), Chebyshev 2nd kind (Top Right) and Jacobi $\alpha = 1$, $\beta = -1/2$ (Bottom) for (2.30).

Let us now compare the standard and constrained approximations for three polynomial families: Chebyshev 1st kind, Chebyshev 2nd kind, and general Jacobi with parameters $\alpha = 1$ and $\beta = -1/2$. The associated outcomes are illustrated in Figure 2.2 and errors are detailed in Tables 2.3, 2.4, and 2.5. Within the Jacobi family, the deliberate choice of α and β serves the purpose of disrupting the basis symmetry to evaluate the method's robustness in this context. Across all three cases, it is clear that standard approximations exhibit spectral accuracy, with moments converging in a spectral manner. Consequently, our analysis falls within the range of Theorem 7, affirming spectral accuracy for constrained approximations and exact conservation of the first $Q = 4$ moments. This property shows clearly in the tables mentioned earlier. It is noteworthy that the constraining matrix $\hat{C}_k$ may exhibit pronounced ill-conditioning for large values of Q, potentially introducing numerical artifacts. Preconditioners may therefore be required to improve numerical stability.

Unbounded domain $\mathbb{R}$. Continuing, we extend our approximation to $\mathbb{R}$ utilizing Hermite functions. Let H_k denote the k^{th} Hermite polynomial. We define the k^{th} symmetrically weighted Hermite functions as

$$\psi_k = H_k e^{-x^2/2}. \tag{2.31}$$

Standard

N	$\|f - f_N\|_2$	$\|m_0 - m_{0,N}\|$	$\|m_1 - m_{1,N}\|$	$\|m_2 - m_{2,N}\|$	$\|m_3 - m_{3,N}\|$
8	1.034e-01	1.746e-03	1.404e-02	1.958e-02	1.538e-02
16	2.477e-05	1.813e-07	6.126e-09	1.858e-07	6.264e-09
32	2.094e-14	1.388e-17	5.551e-17	2.776e-17	5.551e-17

Constrained

N	$\|f - f_N^c\|_2$	$\|m_0 - m_{0,N}^c\|$	$\|m_1 - m_{1,N}^c\|$	$\|m_2 - m_{2,N}^c\|$	$\|m_3 - m_{3,N}^c\|$
8	1.039e-01	5.551e-17	5.551e-17	0.	5.551e-17
16	2.477e-05	0.	0.	2.776e-17	0.
32	2.093e-14	2.776e-17	5.551e-17	2.776e-17	5.551e-17

Table 2.3 – Error Analysis: Standard vs. Conservative with Chebyshev 1st kind approximation of (2.30).

Standard

N	$\|f - f_N\|_2$	$\|m_0 - m_{0,N}\|$	$\|m_1 - m_{1,N}\|$	$\|m_2 - m_{2,N}\|$	$\|m_3 - m_{3,N}\|$
8	1.081e-01	8.262e-03	5.914e-03	8.499e-03	6.049e-03
16	2.701e-05	1.388e-06	4.889e-08	1.398e-06	4.921e-08
32	1.132e-14	8.327e-17	0	2.776e-17	5.551e-17

Constrained

N	$\|f - f_N^c\|_2$	$\|m_0 - m_{0,N}^c\|$	$\|m_1 - m_{1,N}^c\|$	$\|m_2 - m_{2,N}^c\|$	$\|m_3 - m_{3,N}^c\|$
8	1.029e-01	1.388e-17	0	0	0
16	2.629e-05	8.327e-17	0	2.776e-17	5.551e-17
32	1.135e-14	1.388e-17	0	0	5.551e-17

Table 2.4 – Error Analysis: Standard vs. Conservative with Chebyshev 2nd kind approximation of (2.30).

Standard

N	$\|f - f_N\|_2$	$\|m_0 - m_{0,N}\|$	$\|m_1 - m_{1,N}\|$	$\|m_2 - m_{2,N}\|$	$\|m_3 - m_{3,N}\|$
8	2.738e-01	4.374e-02	4.507e-02	0.437e-02	4.514e-02
16	8.873e-05	7.374e-06	7.433e-06	7.373e-06	7.434e-06
32	2.237e-12	8.263e-14	9.082e-14	8.271e-14	9.093e-14

Constrained

N	$\|f - f_N^c\|_2$	$\|m_0 - m_{0,N}^c\|$	$\|m_1 - m_{1,N}^c\|$	$\|m_2 - m_{2,N}^c\|$	$\|m_3 - m_{3,N}^c\|$
8	1.787e-01	6.939e-17	0	2.776e-17	5.551e-17
16	6.695e-05	1.110e-16	0	1.110e-16	5.551e-17
32	1.868e-12	1.388e-16	5.551e-17	5.551e-17	0

Table 2.5 – Error Analysis: Standard vs. Conservative with Jacobi ($\alpha = 1$, $\beta = -\frac{1}{2}$) approximation of (2.30).

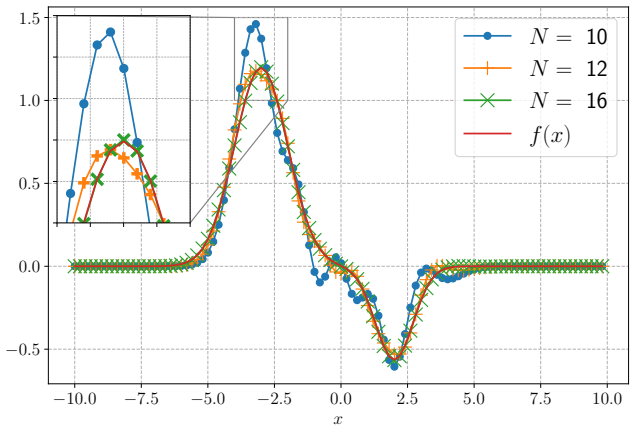


Figure 2.3 – Constrained Hermite approximation for (2.32).

Standard					
N	$\|f - f_N\|_2$	$\|m_0 - m_{0,N}\|$	$\|m_1 - m_{1,N}\|$	$\|m_2 - m_{2,N}\|$	$\|m_3 - m_{3,N}\|$
8	1.264e-02	0.583e-01	1.718e00	1.130e+01	3.686e+01
16	1.716e-03	6.692e-03	2.060e-02	2.293e-01	7.457e-01
32	4.995e-09	1.698e-09	5.175e-09	9.470e-08	3.552e-07
Constrained					
N	$\|f - f_N^c\|_2$	$\|m_0 - m_{0,N}^c\|$	$\|m_1 - m_{1,N}^c\|$	$\|m_2 - m_{2,N}^c\|$	$\|m_3 - m_{3,N}^c\|$
8	4.324e-01	0	1.776e-15	3.553e-15	2.842e-14
16	2.879e-03	2.220e-16	1.776e-15	0	0
32	5.055e-09	2.220e-16	1.776e-15	0	0

Table 2.6 – Error Analysis: Standard vs. Conservative Hermite function approximation for (2.32).

Specifically, we want to approximate the difference between two Gaussian functions deliberately crafting a non-symmetric outcome:

$$f(x) = \frac{3}{\sqrt{2\pi}} e^{(x+3)/2} - \frac{1}{\sqrt{\pi}} e^{(x-2)}, \quad x \in \mathbb{R}. \tag{2.32}$$

We illustrate the results in Figure 2.3 and, again, the results (See Table 2.6) ensure that we are within the scope of Theorem 7. As predicted, the conservative method exhibits spectral accuracy, along with conservation of moments.

Unbounded domain $\mathbb{R}^+$. We now look at the approximation on $\mathbb{R}^+$ using symmetric Laguerre functions. We define for a Laguerre polynomial $\mathcal{L}_k$ the k^{th} Laguerre function as

$$\xi_k = \mathcal{L}_k e^{-x/2} \tag{2.33}$$

Let us consider the function

$$f(x) = (x^3 - 2x + \sin(x))e^{-x}, \quad x \in \mathbb{R}^+. \tag{2.34}$$

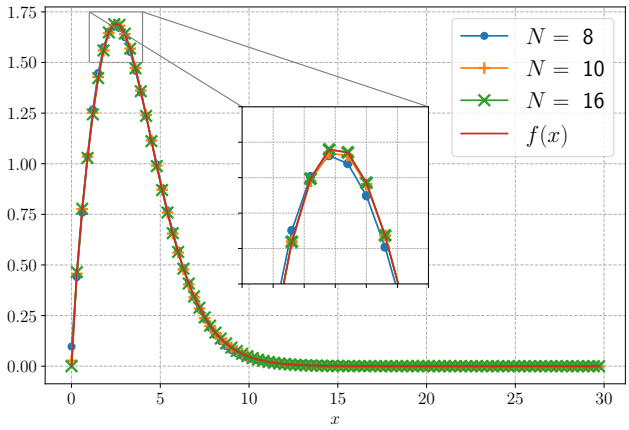


Figure 2.4 – Constrained Laguerre approximation for (2.34).

Standard					
N	$\|f - f_N\|_2$	$\|m_0 - m_{0,N}\|$	$\|m_1 - m_{1,N}\|$	$\|m_2 - m_{2,N}\|$	$\|m_3 - m_{3,N}\|$
8	3.761e-02	3.780e-02	1.225e+00	3.968e+01	1.295e+03
16	9.552e-05	1.227e-04	7.954e-03	5.157e-01	3.350e+01
32	2.337e-11	3.174e-11	4.092e-09	5.276e-07	6.807e-05
Constrained					
N	$\|f - f_N^c\|_2$	$\|m_0 - m_{0,N}^c\|$	$\|m_1 - m_{1,N}^c\|$	$\|m_2 - m_{2,N}^c\|$	$\|m_3 - m_{3,N}^c\|$
8	4.834e-02	8.882e-16	3.553e-15	3.126e-13	3.638e-12
16	1.174e-04	8.882e-16	0	5.684e-14	9.095e-13
32	2.622e-11	0	3.553e-15	2.842e-14	2.274e-13

Table 2.7 – Error Analysis: Standard vs. Conservative using Laguerre Function approximation for (2.34).

The approximation is illustrated in Figure 2.4 and, consistent with prior cases, the observations in Table 2.7 show the spectral accuracy of the conservative method. Although errors on the moments exhibit a slight increase compared to previous cases, it is noteworthy that this discrepancy can be attributed to numerical artifacts arising from the ill-conditioning of the constraining matrix $\hat{C}_k$ and numerical integration intricacies.

To conclude this first test let us again consider Laguerre functions but for the approximation of

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}v} \exp\left(-\frac{(\ln(v)-\mu)^2}{2\sigma}\right), \quad x \in \mathbb{R}^{+,*}, \tag{2.35}$$

where $\sigma = 0.2$ and $\mu = \ln(40) - 0.2$. This function is badly approximated by negative exponentials at infinity, yet it holds significance for Test 4. As depicted in Figure 2.5, spectral accuracy is not attained in the approximation of (2.35). Consequently, this experiment falls outside the assumptions of Theorem 7. Nevertheless, it is noteworthy that the resultant conservative approximation successfully preserves the 0^{th} order moment and achieves fifth-order accuracy.

In the next tests, we apply the conservative method in a PDE setting to various Fokker-Planck equations.

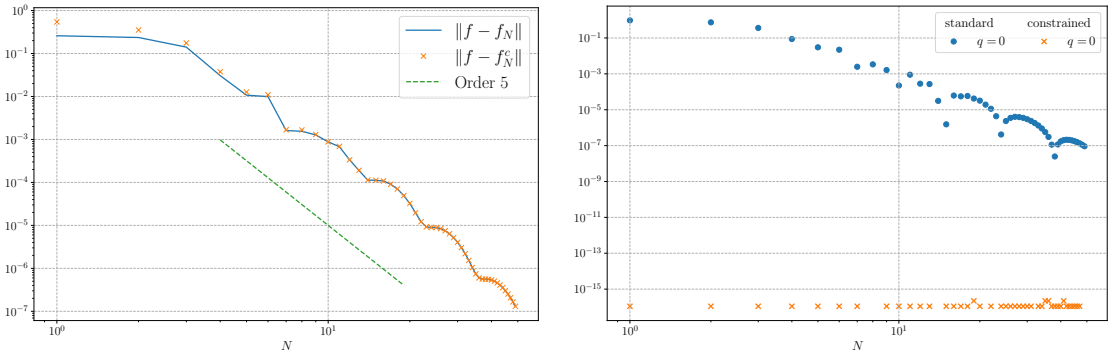


Figure 2.5 – Error Analysis: Solution and first Moment using Laguerre functions for (2.35).

2.3.2 Test 2: Kinetic Fokker-Planck equation

Let us now consider the following kinetic Fokker-Planck equation on $\mathbb{R}$:

$$\begin{cases} \partial_t f = \partial_v \left((v - \mu) f + T \partial_v f \right) := L_{FP}(f), \\ f(0, v) = f^0(v). \end{cases} \tag{2.36}$$

In this case, the unknown $f(t, v)$ describes the amount of particles moving with velocity v at time t. In this setting, the particle will be forced towards a mean velocity μ and an average temperature T. It is well known that this equation admits a steady state f_∞ given by a Maxwellian distribution:

$$f_\infty(v) = \frac{\rho}{\sqrt{2\pi T}} \exp\left(\frac{-(v - \mu)^2}{2T} \right), \tag{2.37}$$

where ρ is the total mass of particles. In the following, let us consider the initial data

$$f^0(v) = \frac{1}{\sqrt{\pi}} \left(e^{-(v+2)^2} + e^{-(v-2)^2} \right). \tag{2.38}$$

This distribution admits a mass ρ^0, mean velocity μ^0 and temperature T^0 given by

$$\rho^0 = \int_{\mathbb{R}} f^0(v) \, \mathrm{d}v, \quad \mu^0 = \frac{1}{\rho^0} \int_{\mathbb{R}} v f^0(v) \, \mathrm{d}v, \quad T^0 = \int_{\mathbb{R}} (v - \mu^0)^2 f^0(v) \, \mathrm{d}v. \tag{2.39}$$

In addition, we set the parameters of L_{FP} in (2.36) as

$$\mu = \mu^0 \text{ and } \quad T = T^0. \tag{2.40}$$

With this choice of parameters, one can in particular expect preservation of mass, mean velocity and temperature. We now consider the symmetrically weighted Hermite functions: $\psi_k = H_k e^{-x^2/2}$. By definition of the Hermite polynomials, we have the relations:

$$\begin{aligned} H_k &= 2v H_{k-1} - 2(k-1) H_{k-2}, \\ H_k' &= 2k H_{k-1}. \end{aligned} \tag{2.41}$$

Therefore, the Hermite functions ψ_k satisfy:

$$\begin{aligned}
\psi_k &= 2v\psi_{k-1} - 2(k-1)\psi_{k-2}, \\
\psi_k' &= k\psi_{k-1} - \frac{1}{2}\psi_{k+1}, \\
\psi_k'' &= k(k-1)\psi_{k-2} - \left(k + \frac{1}{2}\right)\psi_k + \frac{1}{4}\psi_{k+2}.
\end{aligned} \tag{2.42}$$

Using these relations, one can explicitly compute

$$\begin{aligned}
L_{FP}(\psi_k) &= \psi_k + (v-\mu)\psi_k' + T\psi_k'' \\
&= k(k-1)(1+T)\psi_{k-2} - \frac{\mu}{2}\psi_{k-1} + \left(-kT - \frac{(T-1)}{2}\right)\psi_k \\
&\quad + \frac{\mu}{2}\psi_{k+1} + \frac{1}{4}(T-1)\psi_{k+2}.
\end{aligned} \tag{2.43}$$

We can now proceed to use a Galerkin method to solve (2.36). Let us denote by $\mathcal{P}_N$ and $\mathcal{P}_N^c$ the standard and constrained Hermite function approximation. We want to find $f_N \in S_N$ solution to

$$\begin{cases} \partial_t f_N = \mathcal{P}_N\left(L_{FP}(f_N)\right), \\ f_N(0,v) = \mathcal{P}_N(f^0)(v). \end{cases} \tag{2.44}$$

The moment constrained problem writes

$$\begin{cases} \partial_t f_N^c = L_{FP,N}^c(f_N^c), \\ f_N^c(0,v) = \mathcal{P}_N^c(f^0)(v), \end{cases} \tag{2.45}$$

where $L_{FP,N}^c$ is solution to

$$L_{FP,N}^c(f) = \operatorname{argmin}\left\{\|g_N - L_{FP}(f)\|_{L^2_\omega}^2 \,:\, g_N \in S_N,\ \langle g_N, x^q\rangle = 0,\ q = 0,1,2,3\right\}. \tag{2.46}$$

Let us emphasize that (2.46) corresponds to (2.19) where we ensure the first three moments of L_{FP} to be zero. For the time stepping method, we utilize a classical forth order Runge-Kutta (RK4) method with fixed time step $\Delta t = 10^{-4}$. In Figure 2.6 one can observe a very good agreement between the two solutions as well as a trend towards the steady state f_∞. This observation suggests that the constrained approximation may be able to finely capture long-time properties of the solution.

Spectral accuracy and conservations. Let us now investigate the accuracy of the method. In Figure 2.7 we set a final time $T_f = 0.1$ and compare the solution to a reference one computed with $N = 32$ modes. We do the same comparison for the first three moments. A first observation is that the standard method does exhibit spectral accuracy both on the solution and on the first three moments. This therefore falls within the scope of Theorem 7 and the constrained method is indeed spectrally accurate. Note that in this particular case, the mean velocity is 0, hence the machine accuracy in the bottom left panel of Figure 2.7.

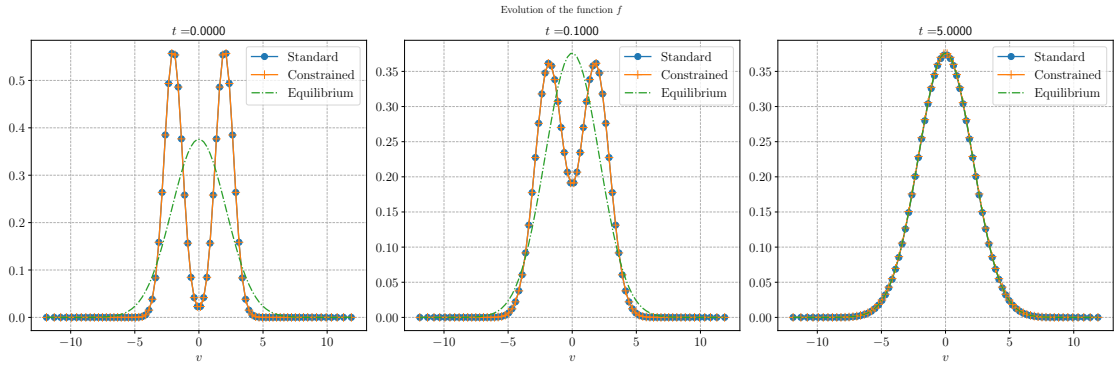


Figure 2.6 – Fokker-Planck model: Snapshots of the distribution at $t = 0, 0.1$ and 5, $N = 32$.

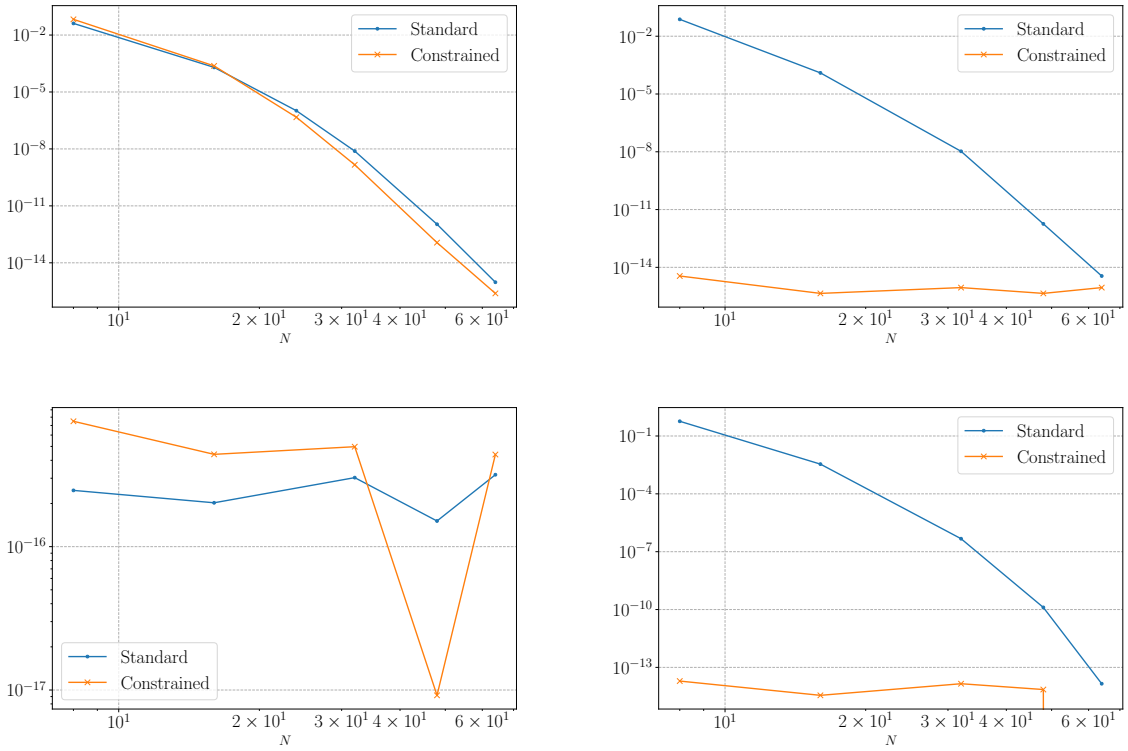


Figure 2.7 – Fokker-Planck model: Approximation error on the solution (Top Left) and absolute error on the mass (Top Right), mean velocity (Bottom Left) and temperature (Bottom Right) at $T_f = 0.1$.

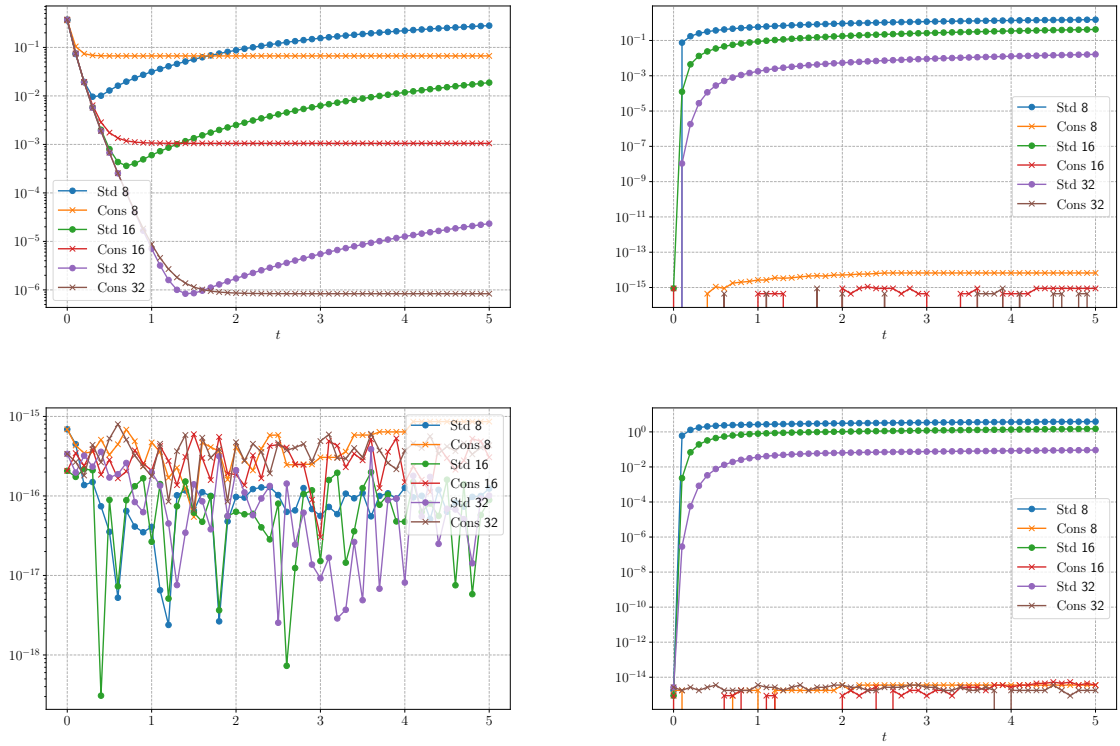


Figure 2.8 – Fokker Planck model: Convergence towards the steady state (Top Left) and absolute error on the mass (Top Right), mean velocity (Bottom Left) and temperature (Bottom Right).

Long-time behavior. Let us conclude this test by investigating the long time behavior of the solution observed in Figure 2.6. In Figure 2.8, the top left panel shows the evolution of the norms $\|f_N(t^n) - f_\infty\|_2$ and $\|f_N^c(t^n) - f_\infty\|_2$ while the remaining panels show the variation of the moments $|\rho^n - \rho^0|$, $|\mu^n - \mu^0|$ and $|T^n - T^0|$. We observe that the constrained method indeed preserves the first three moments up to machine accuracy. As a consequence, the long-time behavior should be better approximated. The standard approximation does seem to converge towards the steady state but ultimately, accumulation of errors on the moments overcomes the dynamic of the solution, pushing it away from the equilibrium. On the contrary, our proposed method does not exhibit this re-bounce and converges to a plateau that decreases as one increases the number of modes. One can mention that equilibrium preserving techniques [86, 178] have been developed to overcome this saturation, but these are beyond the scope of this chapter.

2.3.3 Test 3: A model of Opinion formation

For this second test, we turn our attention to the following opinion model [93] where $v \in [-1, 1]$:

$$\begin{cases} \partial_t g = \dfrac{\lambda}{2} \partial_{vv} \left((1 - v^2) g \right) + \partial_v ((v - m) g) := L_{Op} g, \\ g(0, v) = g^0(v). \end{cases} \tag{2.47}$$

In this case, the unknown $g(t, v)$ describes the distribution of individuals at time t with an opinion trait v. The modelling parameters are chosen so that $|m| < 1$ and $\lambda < 1 + |m|$ to prevent blow-ups as

$v \to \pm 1$. The value of m corresponds to a consensual opinion acting like the mean velocity in the kinetic setting. The parameter λ will control the spread of the equilibrium opinion distribution like how the temperature affects the variance of the particle velocities. In addition, we supplement this equation with a no flux boundary condition to ensure the conservation of mass:

$$\frac{\lambda}{2}\partial_v\Big((1-v^2)g\Big) + (v-m)g\Big|_{v=\pm 1} = 0. \tag{2.48}$$

In particular, one can show that this condition in fact boils down to the homogeneous Dirichlet boundary condition $g(t,\pm 1) = 0$. This equation also admits a steady state given by

$$g_\infty(v) = c_{m,\lambda}(1+v)^{\frac{1+m}{\lambda}-1}(1-v)^{\frac{1-m}{\lambda}-1} \tag{2.49}$$

where the constant $c_{m,\lambda}$ is for normalization. In order to approximate solutions to (2.47) we want to consider general Jacobi polynomials $\left(p_k^{\alpha,\beta}\right)_{k=0,\dots,N}$ associated to the weight $\omega(v) = (1-v)^\alpha(1+v)^\beta$. However, this polynomial basis does not satisfy the homogeneous Dirichlet boundary conditions. To solve this issue, we construct a new polynomial basis obtained from the original Jacobi polynomials (see *e.g.* [194]). For $\alpha, \beta > -1$, let

$$\zeta_k = p_k^{\alpha,\beta} + a_k p_{k+1}^{\alpha,\beta} + b_k p_{k+2}^{\alpha,\beta}, \quad k = 0,\dots,N-2 \tag{2.50}$$

where the constants a_k and b_k are solutions to the linear system:

$$\begin{cases} \zeta_k(-1) = 0, \\ \zeta_k(1) = 0. \end{cases} \tag{2.51}$$

This new polynomial family now satisfies the boundary conditions, but it is, by construction, no longer orthogonal. Since orthogonality is a key ingredient of the conservative spectral method, we need to apply a Gram-Schmidt algorithm to obtain an orthonormal basis. Note that in practice the process of orthonormalization may induce an accumulation of machine errors that can lead to a significant loss in orthogonality for large number of modes. We can now expand g in the basis of polynomials $(\zeta_k)_{k=0,\dots,N-2}$:

$$g(t,v) = \sum_{k=0}^{N-2} \hat{g}_k(t)\zeta_k(v), \tag{2.52}$$

and then proceed as before to solve (2.47) using a Galerkin method. We define the standard and constrained problems as

$$\begin{cases} \partial_t g_N = \mathcal{P}_N\Big(L_{Op}(g_N)\Big), \\ g_N(0,v) = \mathcal{P}_N(g^0)(v), \end{cases} \tag{2.53}$$

and

$$\begin{cases} \partial_t g_N^c = L_{Op,N}^c(g_N^c), \\ g_N^c(0,v) = \mathcal{P}_N^c(g^0)(v). \end{cases} \tag{2.54}$$

Since (2.47) preserves only the first moment, we define $L_{Op,N}^c$ as the solution to

$$L_{Op,N}^c(g) = \operatorname{argmin}\Big\{\|g_N - L_{Op}(g)\|_{L^2_\omega}^2 \,:\, g_N \in S_N,\ \langle g_N, 1\rangle = 0\Big\}. \tag{2.55}$$

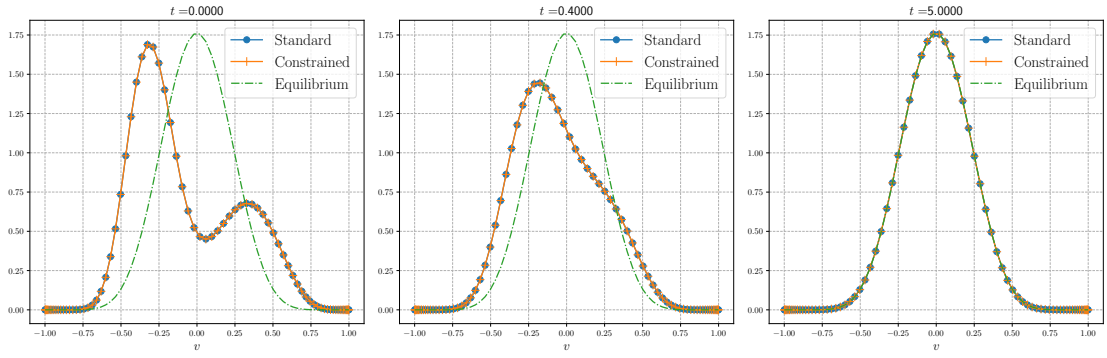


Figure 2.9 – Opinion model: Snapshots of the distribution at times $t = 0, 0.4$ and 5, $N = 24$.

The time stepping is again achieved using an RK4 time integrator with time step $\Delta t = 10^{-4}$. Before presenting numerical results, let us mention that the space spanned by the newly constructed basis does not contain the monomials v^q, $q = 0, 1, 2, \dots$ anymore. As a consequence, regarding (2.17), even if one considers Legendre polynomials ($\omega = 1$) to construct the new basis, one does not expect exact conservation of moments. However, as will be shown through numerical experiments, one still recovers spectral accuracy on the moments for the standard method which ensures spectral accuracy on the constrained solution. In regard to this discussion, we now present only the choice of Legendre polynomials for the construction of the new basis $(\zeta_k)_{k=0,\dots,N-2}$. Note however that classical families such as Chebyshev first and second kind provide similar results. As an initial data, we take

$$g^0(v) = c_0\Big((1+v)^{12}(1-v)^6 + (1+v)^{13}(1-v)^{25}\Big), \tag{2.56}$$

where c_0 is for normalizing the first moment. Then we set $m = 0$ and $\lambda = 0.1$ in the definition of L_{Op} in (2.47), therefore expecting an equilibrium of the form

$$g_\infty(v) = c_\infty(1-v^2)^9.$$

Figure 2.9 shows the evolution of the approximate distributions g_N and g_N^c. As in the previous case, we observe a very good agreement between the standard and constrained method. Moreover, the solutions also appear to converge towards the steady state.

Spectral accuracy and conservations. Let us now investigate the accuracy of the method and its mass conservation properties. Indeed, at the continuous level, equation (2.47) along with no flux boundary conditions preserves only the $0th$ order moment. In the next study, we set a final time $T_f = 0.1$ and compare the approximation error between the standard and constrained method. We observe in Figure 2.10 that we again fall into the framework of the convergence theorem 7 and the constrained method behaves as predicted.

Long-time behavior. Let us conclude this test by looking at the long time behavior of the approximated solution to (2.47). In order to quantify the observation made on Figure 2.9, we show in Figure 2.11 the evolution of the norms $\|g_N - g_\infty\|_2$ and $\|g_N^c - g_\infty\|_2$ as well as the evolution of the mass variation $|m_0^0 - m_0(t^n)|$. Similarly, as before, we observe that the conservation of moments

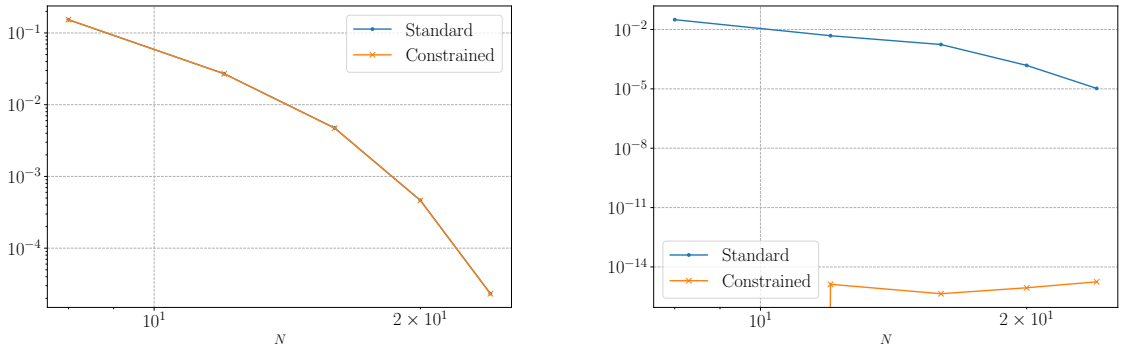


Figure 2.10 – Opinion model: Approximation error on the solution (Left) and on the 0^{th} order moment (Right) at $T_f = 0.1$.

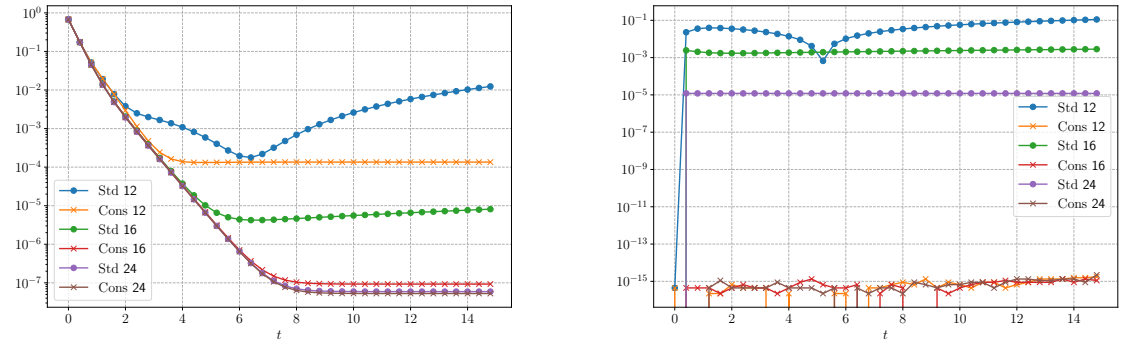


Figure 2.11 – Opinion model: Convergence towards the steady state (Left) and mass variation (Right).

allows to stabilize the long time dynamic of the solution.

2.3.4 Test 4: A Call center service time model

This final test deals with a call center service time model introduced in [118]:

$$\begin{cases} \partial_t h = \dfrac{\lambda}{2}\partial_{vv}\left(v^2 h\right) + \dfrac{\gamma}{2}\partial_v\left(v\ln\left(\dfrac{v}{v_L}\right)h\right) := L_{CC}h, \quad v \in \mathbb{R}^+ \\ h(0,v) = h^0(v). \end{cases} \tag{2.57}$$

In this context, the distribution density $h(t,v)$ describes the service time of agents in a call center. The constants $0 < \gamma < 1$, $\lambda > 0$ and $v_L > 0$ are modelling parameters. In particular, the latter corresponds to an ideal time for agents to complete their task. The value γ will denote the maximal amount of change in service time that agents will be able to perform in a single operation and λ is the variance of the distribution. To supplement this equation, we consider a no-flux boundary condition

$$\frac{\lambda}{2}\partial_v\left(v^2 h\right) + \frac{\gamma}{2}\ln\left(\frac{v}{v_L}\right)h\Big|_{v=0} = 0. \tag{2.58}$$

Studying the limit $v \to 0$ in this condition, it is clear that it is automatically satisfied by (2.57). It can also be shown that (2.57) admits an equilibrium distribution of the form:

$$h_\infty(v) = \frac{1}{\sqrt{2\pi\sigma}v} \exp\left(-\frac{(\ln(v)-\mu)^2}{2\sigma}\right) \tag{2.59}$$

where $\sigma = \frac{\lambda}{\gamma}$ and $\mu = \ln(v_L) - \sigma$. As in Test 1, in order to deal with integrability at infinity, we want to consider the symmetrically weighted Laguerre functions : $\xi_k = \mathcal{L}_k \omega^{\frac{1}{2}}$ defined by (2.33). In addition, to deal with the poor approximation of the steady state observed in Figure 2.5 we consider a micro-macro approach. Since the equilibrium is known, we can decompose the unknown as

$$h = h_\infty + \tilde{h}, \tag{2.60}$$

where $\tilde{h}$ is the perturbation that can take negative values. Since (2.57) is linear in h, and by definition of the steady state, the perturbation also solves (2.57). The only difference is the initial data:

$$\begin{cases} \partial_t \tilde{h} = \dfrac{\lambda}{2}\partial_{vv}\left(v^2\tilde{h}\right) + \dfrac{\gamma}{2}\partial_v\left(v\ln\left(\dfrac{v}{v_L}\right)\tilde{h}\right), \\ \tilde{h}(0,v) = h^0(v) - h_\infty(v). \end{cases}$$

From this procedure, we expect to observe that $\tilde{h}$ converges towards 0 for long time. Consequently, the long time behavior of the solution should be better approximated since 0 now belongs to the approximation space. However, one cannot say anything about the short time behavior as the initial data could still be poorly approximated.

As previously, we can then define the standard and constrained problems on the perturbation as

$$\begin{cases} \partial_t \tilde{h}_N = \mathcal{P}_N\left(L_{CC}(\tilde{h}_N)\right), \\ \tilde{h}_N(0,v) = \mathcal{P}_N(\tilde{h}^0 - h_\infty)(v), \end{cases} \tag{2.61}$$

and

$$\begin{cases} \partial_t \tilde{h}^c_N = L^c_{CC,N}(\tilde{h}^c_N), \\ \tilde{h}^c_N(0,v) = \mathcal{P}^c_N(\tilde{h}^0 - h_\infty)(v). \end{cases} \tag{2.62}$$

The original distribution is then reconstructed using (2.60) for visualizations. Since (2.57) preserves only the first moment, we define $L^c_{CC,N}$ as the solution to

$$L^c_{CC,N}(h) = \operatorname{argmin}\left\{\|g_N - L_{CC}(h)\|^2_{L^2_\omega} \,:\, g_N \in S_N,\ \langle g_N, 1\rangle = 0\right\}. \tag{2.63}$$

The time stepping is yet again achieved through a RK4 method with time step $\Delta t = 10^{-4}$. Let us now fix the parameters $\lambda = 0.5$, $\gamma = 0.9$ and $v_L = 40$ in (2.57). As an initial data, we consider

$$h^0(v) = (x^3 - 2x + \sin(x))e^{-x}. \tag{2.64}$$

This function is actually well approximated by Laguerre functions. We show on Figure 2.12 the evolution of the standard and constrained distributions. A first observation is that there is again a good agreement between the two methods. In addition, as expected from the micro-macro

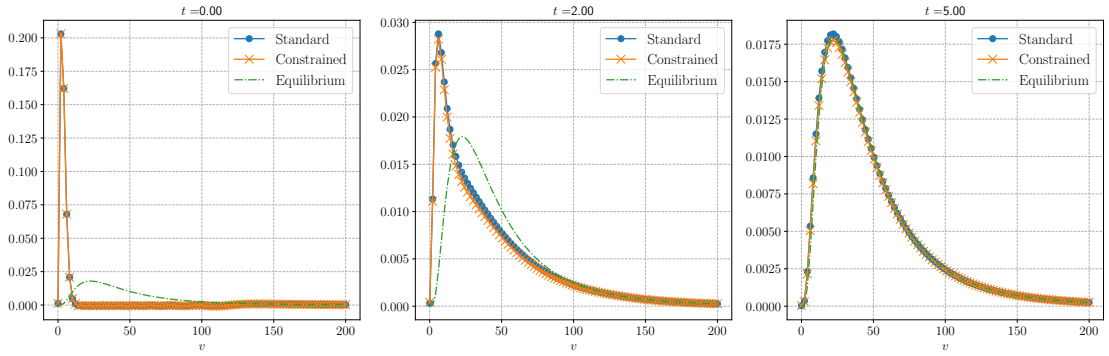


Figure 2.12 – Call Center model: Snapshots of the distributions at times $t = 0, 2$ and 5, $N = 32$.

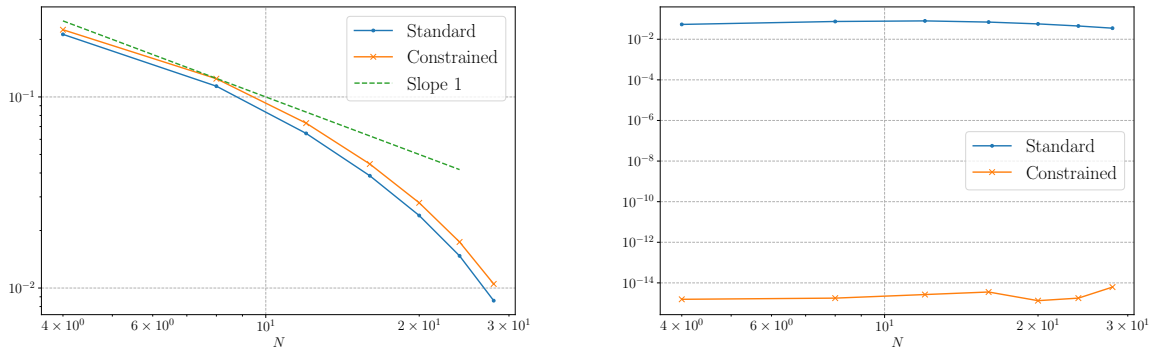


Figure 2.13 – Call Center model: Approximation error on the solution (Left) and on the 0^{th} order moment (Right) at $T_f = 0.1$.

approach, the steady state is well approximated.

Accuracy and conservation. To assess the spectral accuracy of the method, we consider a final time $T_f = 0.1$ along with a time step $\Delta t = 10^{-4}$ for the time integration. We present on Figure 2.13 the study of the convergence of the method. As expected, in short times, the distribution is not very well approximated in our chosen basis. As a consequence, we are only able to recover a convergence that is quite slow. However, the up-side is that the mass is still exactly preserved by the constrained approximation. Now, contrary to previous tests, it is important to note that the standard method is not spectrally accurate on the mass.

Long-time behavior. To conclude this section, and as in previous tests, we investigate the long-time behavior of the solution through the norms $\|h_N - h_\infty\|_2$ and $\|h_N^c - h_\infty\|_2$. We observe in Figure 2.14 that the micro-macro approach allows to approximate really well the steady state with both methods. In addition, the mass variation induced by the standard approach decays exponentially fast towards 0 therefore ensuring that it also converges towards h_∞ albeit more slowly than the conservative approach.

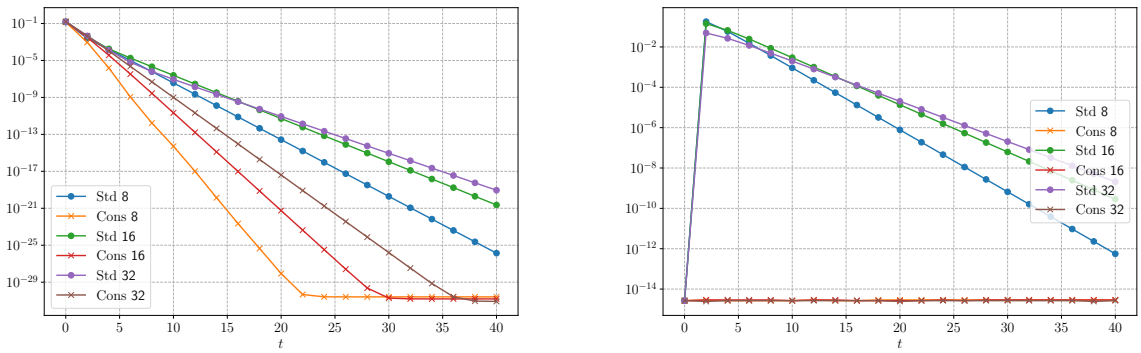


Figure 2.14 – Call Center model: Convergence towards the steady state (Left), and mass variation (Right).

Part II

Moment-driven efficient numerical methods

Chapter 3

Hybrid Kinetic/Fluid numerical method for the Vlasov-BGK equation in the diffusive scaling

In this chapter, I present a hybrid model-adaptation method for collisional kinetic equations in a diffusive scaling. Specifically, the method is applied to the linear Vlasov-BGK model and an extension to the Vlasov-Poisson-BGK system is considered at the end. The aim of the approach is to reduce the computational cost of kinetic simulations by taking advantage of the lower dimensionality of an asymptotic model while reducing the overall error. It relies on two criteria motivated by a perturbative approach to obtain a dynamic domain adaptation.
The content of this chapter covers the published works :

— T. Laidin. "Hybrid kinetic/fluid numerical method for the Vlasov-BGK equation in the diffusive scaling". In: *Kinet. Relat. Mod.* (2023). ISSN: 1937-5093. DOI: 10.3934/krm.2023013

— T. Laidin and T. Rey. "Hybrid Kinetic/Fluid numerical method for the Vlasov-Poisson-BGK equation in the diffusive scaling". In: *FVCA 10 - 2023 - International Conference on Finite Volumes for Complex Applications X*. Springer Proceedings in Mathematics & Statistics. 8 pages, 4 figures. Strasbourg, France, Oct. 2023, pp. 229–237. DOI: 10.1007/978-3-031-40860-1_24

Outline of the current chapter

3.1 Introduction

Systems of particles in interactions arise in many fields of science such as gas theory, plasmas physics or even semiconductors. The mathematical models for such systems can be classified in three major scales: particles (microscopic), kinetic (mesoscopic), and fluid (macroscopic). The first type of model is about describing the system as point particles interacting with each other via collisions and/or electromagnetic forces. Although the resulting system is the most realistic, it is in practice extremely large. Its study both theoretically and numerically becomes unattainable. In this work we consider a kinetic description of the system modelled by the Vlasov-BGK equation. The unknown is the probability distribution $f(t,x,v)$ solution to:

$$\begin{cases} \partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = \mathcal{Q}(f), \\ f(0,x,v) = f_0(x,v). \end{cases}$$

The short-range interactions between particles are taken into account through the collision operator $\mathcal{Q}(f)$, and the function $E(x)$ is a given exterior electrical field. While attainable, the simulation of this equation remains expensive in computational resources. Moreover, using the kinetic description of the system may not be necessary on the whole computing domain. The fluid one, that is less precise but much less costly, can be used where it is accurate enough. The aim of the chapter is therefore to design a hybrid kinetic/fluid scheme with an automatic domain adaptation method. It relies on a robust numerical scheme for the kinetic equation, on relevant criteria to carefully determine fluid and kinetic regions and on a smart implementation.

Diffusive scaling. In some applications, it is relevant to consider a scaled version of the Vlasov-BGK equation. Let us introduce the scaling parameter $\varepsilon > 0$. It is related to the Knudsen number: the ratio between the mean free path of the particles and the length scale of observation. This work will focus on the diffusive scaling. Let $d_x \geq 1$ and $d_v \in \mathbb{N}^*$ be integers. We denote by Ω_x a subset of $\mathbb{R}^{d_x}$. Let $t \geq 0$, $x \in \Omega_x$ and $v \in \mathbb{R}^{d_v}$. We look for a particle distribution function f^ε solution to the

following scaled equation:

$$\begin{cases} \partial_t f^\varepsilon + \dfrac{v}{\varepsilon} \cdot \nabla_x f^\varepsilon + \dfrac{E}{\varepsilon} \cdot \nabla_v f^\varepsilon = \dfrac{1}{\varepsilon^2} \mathcal{Q}(f^\varepsilon), \\ f^\varepsilon(0,x,v) = f_0(x,v). \end{cases} \tag{P^ε}$$

We assume that the initial condition f_0 is nonnegative and does not depend on ε. In practice, the Knudsen number can be of order 1 down to 0 depending on the physics being modelled. On the one hand, when $\varepsilon \sim 1$, the system is said to be in the kinetic regime. It models a system with few collisions between particles. On the other hand, when $\varepsilon \ll 1$, the system reaches the fluid regime. This scaling and its asymptotic limit have first been studied in [17]. The asymptotic expansion of the distribution function f^ε in ε is justified in [10] for the neutron transport and in [187] for the linear Boltzmann equation. In [62] a large class of linear collision operators is dealt with and in [114], the authors justified an approximation of the kinetic equation by diffusion using homogenization. In our setting, the limit case $\varepsilon = 0$ is described by a drift-diffusion equation on the density $\rho(t,x)$:

$$\begin{cases} \partial_t \rho - \operatorname{div}_x (\nabla_x \rho - E\rho) = 0, \\ \rho(0,x) = \rho_0(x). \end{cases} \tag{P}$$

Asymptotic preserving scheme. In order to design a numerical method that is efficient, one first needs a numerical scheme that performs well for any value of the parameter ε. Indeed, as ε tends to 0, the transport velocity in (P^ε) formally goes to infinity. Numerically, it translates to smaller and smaller time steps to guarantee the stability of a naive scheme and, consequently, a reasonable computation time cannot be ensured. A solution to this issue is to use schemes that remain stable in the diffusive limit $\varepsilon \to 0$. These schemes fall into the framework of Asymptotic Preserving (AP) schemes, a notion introduced in [141] and [136]. We also refer to the recent review articles [72, 135]. This AP property can be summarized by the diagram in Figure 3.1. In the diagram, ρ corresponds to a solution to the problem (P) and ρ_h is an approximation of ρ. On the other hand, f^ε is a solution to the problem (P^ε) and f_h^ε is an approximation of f^ε. The idea behind AP scheme is threefold. Firstly, the scheme for (P^ε) has to be a consistent discretization of the limit model as $\varepsilon \to 0$. Secondly, a scheme is considered truly AP only if the stability criterion on the time step is independent on the parameter ε. Thirdly, one can explicitly take $\varepsilon = 0$ in the scheme. The need of an AP scheme in the kinetic domain of the hybrid scheme is crucial. On the one hand, for computation time considerations. On the other hand, the limit scheme is used in the fluid regions of the domain adaptation to ensure good transitions between kinetic and fluid states. While AP schemes are designed to resolve both the mesoscopic and the macroscopic scales automatically, it often implies more expensive computations even in a fluid regime because of the resolution of the kinetic scale. By using a hybrid method, one can effectively take advantage of the properties of an AP scheme while limiting its use and therefore reduce the computation time.

Hybrid method. A strategy to take advantage of both the kinetic and fluid scales of description and reduce the computational time is to use a hybrid method. The notion of multiscale coupling has already been studied and a wide range of techniques has been developed. We will rely on the technique that consists in adapting the domain in the position variable. Kinetic and fluid regions

are created and move throughout the simulations and the appropriate solver can then be used in each subdomain. The key idea is the definition of the interfaces between subdomains. Various criteria have been investigated, and we refer to [61, 145, 199, 200, 202] and the references therein. These criteria vary from being based purely on macroscopic quantities (macroscopic criteria) to considering the deviation of the distribution function from a local equilibrium in velocity (kinetic criteria). In particular, the idea to use the asymptotic limit of the kinetic model to achieve a domain adaptation can be found in [142]. In this work, we shall adapt to the diffusive scaling the criterion introduced in [158] and later used in [201] and [87]. Unlike our work, the system studied in these articles considers several moments of the distribution function in a hydrodynamic scaling. It therefore relies on the study of the moment realizability matrix to assess the validity of fluid models. In this work, we only consider the first moment equation. To achieve the domain adaptation, we will use a combination of macroscopic and kinetic criteria. We also refer to [145] where in addition to the coupling between the kinetic and fluid scales, mesh refinement is also considered. Another technique, first introduced for the hydrodynamic scaling in [63], also relies on domain adaptation but consists in adding a buffer zone between kinetic and fluid subdomains using a transition function in the continuous model. It was extended to a dynamic setting with moving interfaces in [61] and applied to the diffusive scaling in [69]. Another approach that do not rely on the adaptation of the domain in position [55, 70] consists in splitting the distribution function into a macroscopic part solved via the finite volume method and a microscopic one solved using a Monte-Carlo method. In [130, 131] the splitting of the distribution function is also considered in a stationary setting. While the microscopic part is again solved via Monte-Carlo methods, the macroscopic one is solved using a moment method. Another technique introduced in [77, 78] relies on mesh free methods where the microscopic part is solved using particles and a Lagrangian scheme is used for the macroscopic part. Finally, an alternative approach to reduce the computational cost relies on dynamical low-rank methods [80] where the idea is to reduce the matrix of the linear system associated to the discrete problem. A question that arises when designing hybrid method is the one of conservation properties. While the conservation of mass was not investigated in [87], we shall mention that such a property for hybrid methods has been studied in the past. In a hybrid PIC/Finite volume setting in [56], the conservation of mass was ensured via a projection step so that the moments of the perturbation are preserved. In [39], a Monte Carlo method is used, and the property is obtained via a moment matching sampling method. Another approach is to use boundary conditions such that one obtains a natural balance of fluxes as it was done in [165]. Finally, let us mention [57] where a hybrid model is first constructed at the continuous level to satisfy conservation properties and then discretized.

Main contribution. In this work, we develop a hybrid kinetic/fluid numerical method with an automatic domain adaptation. In particular, no buffer zones are introduced, and we consider a full finite volume approach. The criteria for the dynamic interfaces are adapted from [87]. The domain adaptation is accomplished cell-wise and the mesh is fixed throughout the simulation. Domain adaptation methods are heavily dependent on the treatment of interfaces. One of the main novelty of this work is the way to deal with interface conditions between kinetic and fluid. This is achieved by using a micro-macro reformulation of the kinetic equation. Moreover, the conservation of mass of the hybrid method is thoroughly investigated for a relevant toy model. The accuracy

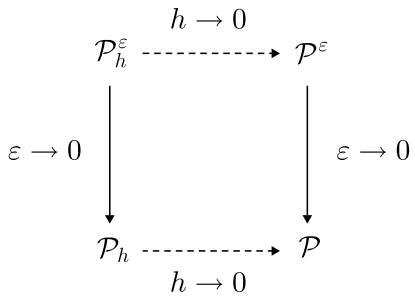


Figure 3.1 – The AP diagram (h denotes the size of the discretization)

and long-time behavior of the method are observed numerically. Finally, the speedup provided by the method is significant and illustrated through several test cases.

Plan of the paper. The outline is as follows. Section 3.2 is dedicated to the derivation of a hierarchy of macroscopic models based on the Chapman-Enskog expansion of the distribution function. In Section 3.3 we recall the micro-macro reformulation of the Vlasov-BGK equation. This reformulation is then used to develop an Asymptotic Preserving scheme with a finite volume approach. Section 3.4 is dedicated to the hybrid method. The coupling indicators based on the hierarchy introduced in Section 3.2 are presented and the implementation of the hybrid scheme is discussed. Finally, numerical experiments are performed in Section 3.5.

3.2 Chapman-Enskog expansion

The aim of this section is to derive a hierarchy of macroscopic models from which we will deduce a macroscopic criterion.

3.2.1 Notations and functional setting

From now on, we consider periodic boundary conditions in position and let $v \in \mathbb{R}^{d_v}$, $d_v \in \mathbb{N}^*$. We also assume that the electrical field E is periodic on $\Omega_x = [0, x_\star]^{d_x}$. From now on, the collision operator will be the linear BGK operator [23]:

$$\mathcal{Q}(f) = \rho\mathcal{M} - f, \quad \text{where } \rho = \langle f \rangle = \langle f \rangle = \int_{\mathbb{R}^{d_v}} f \, \mathrm{d}v. \tag{3.1}$$

The notation $\langle \cdot \rangle$ will be used either for scalars or component-wise for higher order tensors. Here $\mathcal{M}(v)$ denotes the so-called Maxwellian. A standard function that we consider in this work is the multidimensional centered Gaussian:

$$\mathcal{M}(v) = \frac{e^{-|v|^2/2}}{(2\pi)^{d/2}}.$$

Let us recall some properties of $\mathcal{M}$:

$$\begin{cases} \mathcal{M}(v) > 0, \quad \forall v \in \mathbb{R}^{d_v}, \\ \langle \mathcal{M} \rangle = 1. \end{cases}$$

Moreover, since $\mathcal{M}$ can be expressed as the product of d_v 1-dimensional Gaussians, it is isotropic and even in each direction. As a consequence of the symmetric domain in velocity and the symmetry of Gaussian, its odd moments vanish. The 1-dimensional Gaussian also admits finite zeroth, second, and fourth moments in velocity. We denote by m_k its k-th moment:

$$m_k = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v_x^k \exp\left(\frac{-v_x^2}{2}\right) \mathrm{d}v_x.$$

Let us recall the integro-differential problem we are interested in:

$$\begin{cases} \partial_t f^\varepsilon + \dfrac{1}{\varepsilon} \mathsf{T}(f^\varepsilon) = \dfrac{1}{\varepsilon^2}(\rho^\varepsilon \mathcal{M} - f^\varepsilon), \\ f^\varepsilon(0, x, v) = f_0(x, v), \end{cases} \tag{VBGK}$$

where

$$\mathsf{T}(f) = v \cdot \nabla_x f + E \cdot \nabla_v f.$$

Let us set $\gamma(v) = \frac{1}{\mathcal{M}(v)}$ and introduce the measure

$$\mathrm{d}\gamma = \gamma(v)\,\mathrm{d}v = \frac{\mathrm{d}v}{\mathcal{M}(v)}.$$

Let $L^2(\mathrm{d}x\,\mathrm{d}\gamma)$ be the space of square integrable functions against the measure $\mathrm{d}x\mathrm{d}\gamma$ equipped with the scalar product

$$(f_1, f_2)_{L^2(\mathrm{d}x\,\mathrm{d}\gamma)} = \int_{\Omega_x \times \mathbb{R}^{d_v}} f_1 f_2 \,\mathrm{d}x\,\mathrm{d}\gamma.$$

With an initial data in $L^2(\mathrm{d}x\,\mathrm{d}\gamma)$, there is a unique solution to (VBGK) (see, e.g., [3]) which conserves mass and nonnegativity.

One can define the null space of the linear BGK operator (3.1) as

$$\mathcal{N} = \Big\{ f = \rho\mathcal{M} \text{ where } f \in L^2(\mathrm{d}x\,\mathrm{d}\gamma),\, \rho = \langle f \rangle \Big\}.$$

The space $\mathcal{N}$ is sometimes referred to as the equilibrium manifold. In particular, one has that

$$\mathcal{N}^\perp = \Big\{ f \in L^2(\mathrm{d}x\,\mathrm{d}\gamma) \text{ such that } \langle f \rangle = 0 \Big\}.$$

With these notations, one can decompose f as its equilibrium part in $\mathcal{N}$ plus a perturbative part in $\mathcal{N}^\perp$. Note that the perturbative part is not necessarily small.

Let us now introduce the so-called Chapman-Enskog expansion of the distribution function f^ε:

$$f^\varepsilon(t, x, v) = \rho^\varepsilon(t, x)\mathcal{M}(v) + \sum_{k=1}^{\infty} \varepsilon^k h^{(k)}(t, x, v).$$

This expansion comes with the following assumptions. First, the functions $h^{(k)}$ do not depend on the parameter ε. Secondly, we assume that $h^{(k)} \in \mathcal{N}^\perp$ for all k and is therefore such that:

$$\langle h^{(k)} \rangle = 0, \quad \forall k \geq 1.$$

In particular, we will show that these functions can be expressed using the density ρ^ε, the electric field E, the velocity variable v and the Maxwellian $\mathcal{M}$.

3.2.2 Hierarchical truncations

To derive a hierarchy of models, let us consider truncations of order $K \in \mathbb{N}$ of the Chapman-Enskog expansion:

$$f^\varepsilon(t,x,v) \approx \rho^\varepsilon(t,x)\mathcal{M}(v) + \sum_{k=1}^{K} \varepsilon^k h^{(k)}(t,x,v).$$

Plugging this expansion in (VBGK) leads to

$$\partial_t(\rho^\varepsilon \mathcal{M}) + \partial_t \sum_{k=1}^{K} \varepsilon^k h^{(k)} = -\frac{1}{\varepsilon}\mathsf{T}(\rho^\varepsilon M) - \sum_{k=1}^{K} \varepsilon^{k-1}\mathsf{T}(h^{(k)}) - \frac{1}{\varepsilon}\sum_{k=1}^{K} \varepsilon^{k-1} h^{(k)}.$$

Multiplying by ε and rearranging the terms, one obtains

$$\sum_{k=0}^{K-1} \varepsilon^k h^{(k+1)} = -\mathsf{T}(\rho^\varepsilon \mathcal{M}) - \sum_{k=1}^{K} \varepsilon^k \mathsf{T}(h^{(k)}) - \partial_t \sum_{k=2}^{K+1} \varepsilon^k h^{(k-1)} - \varepsilon\,\partial_t(\rho^\varepsilon \mathcal{M}).$$

We now identify powers of ε:

$$k = 0: \qquad h^{(1)} = -\mathsf{T}(\rho^\varepsilon \mathcal{M}), \tag{3.2a}$$

$$k = 1: \qquad h^{(2)} = -\partial_t(\rho^\varepsilon \mathcal{M}) - \mathsf{T}(h^{(1)}), \tag{3.2b}$$

$$2 \leq k \leq K-1: \quad h^{(k+1)} = -\partial_t h^{(k-1)} - \mathsf{T}(h^{(k)}). \tag{3.2c}$$

Macroscopic model

To derive the fluid model, let us truncate the Chapman-Enskog expansion at first order $K = 1$:

$$f^\varepsilon \approx \rho^\varepsilon \mathcal{M} + \varepsilon h^{(1)}. \tag{3.3}$$

We start by integrating (VBGK) in velocity to obtain:

$$\partial_t \rho^\varepsilon + \frac{1}{\varepsilon}\mathrm{div}_x \langle v f^\varepsilon \rangle = 0.$$

Then f^ε is replaced by its expression (3.3):

$$\partial_t \rho^\varepsilon + \frac{1}{\varepsilon}\mathrm{div}_x\left(\rho^\varepsilon \langle v\mathcal{M}\rangle + \varepsilon \langle v h^{(1)} \rangle\right) = 0.$$

The function $h^{(1)}$ is given by the identification (3.2a) and can be simplified using the identity $\nabla_v \mathcal{M} = -v\mathcal{M}$:

$$h^{(1)} = -\mathsf{T}(\rho^\varepsilon \mathcal{M}) = -v\mathcal{M} \cdot J^\varepsilon \quad \text{where} \quad J^\varepsilon = \nabla_x \rho^\varepsilon - E\rho^\varepsilon.$$

Using the fact that ρ^ε does not depend on the velocity and that odd moments of the Maxwellian are zero, we obtain by plugging in the expression of $h^{(1)}$:

$$\partial_t \rho^\varepsilon - \langle v \otimes v\mathcal{M} \rangle : \nabla_x J^\varepsilon = 0, \tag{3.4}$$

where "$\otimes$" denotes the tensor product and ":" is the tensor contraction of order 2. The moment tensor is then computed:

$$\langle v \otimes v\mathcal{M} \rangle = m_2 I$$

with I the identity matrix. Finally, assuming that $\rho^\varepsilon \to \rho$ as $\varepsilon \to 0$ we formally obtain the drift-diffusion model:

$$\partial_t \rho - m_2 \mathrm{div}_x J = 0, \text{ where } J = \nabla_x \rho - E\rho. \tag{DD}$$

Higher order macroscopic model

To derive the third order fluid model, let us place ourselves in the $1D_x$-$1D_v$ setting to lighten the computations and focus on the various steps involved. The full $3D_x$-$3D_v$ framework is presented in Appendix A.

Let us start by truncating the Chapman-Enskog expansion at third order $K = 3$:

$$f^\varepsilon \approx \rho^\varepsilon \mathcal{M} + \varepsilon h^{(1)} + \varepsilon^2 h^{(2)} + \varepsilon^3 h^{(3)}. \tag{3.5}$$

We will see in the derivation process that the second order yields no additional information compared with the first order.

Again, we start by integrating (VBGK) in velocity, and we replace f^ε by its expansion (3.5):

$$\partial_t \rho^\varepsilon + \frac{1}{\varepsilon} \partial_x \langle v\rho^\varepsilon \mathcal{M} \rangle + \partial_x \left\langle v h^{(1)} + \varepsilon v h^{(2)} + \varepsilon^2 v h^{(3)} \right\rangle = 0. \tag{3.6}$$

At this point, we use the identification (3.2) to compute the perturbations $h^{(1)}$, $h^{(2)}$ and $h^{(3)}$. One obtains

$$h^{(1)} = -v\mathcal{M} J^\varepsilon,$$

$$h^{(2)} = -\mathcal{M}\partial_t \rho^\varepsilon + v^2 \mathcal{M} \partial_x J^\varepsilon + (1 - v^2)\mathcal{M} E J^\varepsilon.$$

We replace $h^{(1)}$ and $h^{(2)}$ by their expressions in $h^{(3)} = -\partial_t h^{(1)} - \mathsf{T}(h^{(2)})$ to obtain:

$$\begin{aligned} h^{(3)} = {} & 2v\mathcal{M}\partial_t J^\varepsilon - v^3 \mathcal{M} \partial_{xx} J^\varepsilon - (v\mathcal{M} - v^3\mathcal{M})\partial_x (E J^\varepsilon) \\ & - (2v\mathcal{M} - v^3\mathcal{M}) E \partial_x J^\varepsilon + (3v\mathcal{M} - v^3\mathcal{M}) E^2 J^\varepsilon. \end{aligned} \tag{3.7}$$

We simplify (3.6) by using the fact that $v\mathcal{M}$ and $vh^{(2)}$ are odd in v, obtaining

$$\partial_t \rho^\varepsilon + \partial_x \left\langle v h^{(1)} + \varepsilon^2 v h^{(3)} \right\rangle = 0. \tag{3.8}$$

Equation (3.8) shows that the truncation $K = 2$ gives no further information compared with $K = 1$. We have already shown in (3.4) that

$$\partial_x \left\langle v h^{(1)} \right\rangle = -m_2 \, \partial_x J^\varepsilon.$$

Therefore, (3.8) gives $\partial_t \rho^\varepsilon = m_2 \, \partial_x J^\varepsilon + \mathcal{O}(\varepsilon^2)$. It follows that

$$\partial_t J^\varepsilon = m_2 (\partial_{xx} J^\varepsilon - E \, \partial_x J^\varepsilon) + \mathcal{O}(\varepsilon^2).$$

We use this relation to replace the time derivative of J^ε in (3.7) which gives:

$$\begin{aligned} h^{(3)} = {} & 2m_2 v\mathcal{M} \, \partial_{xx} J^\varepsilon - 2m_2 v\mathcal{M} E \, \partial_x J^\varepsilon - v^3 \mathcal{M} \, \partial_{xx} J^\varepsilon - (v\mathcal{M} - v^3 \mathcal{M}) \, \partial_x (E J^\varepsilon) \\ & - (2v\mathcal{M} - v^3 \mathcal{M}) E \, \partial_x J^\varepsilon + (3v\mathcal{M} - v^3 \mathcal{M}) E^2 J^\varepsilon + \mathcal{O}(\varepsilon^2). \end{aligned}$$

The choice to replace the time derivative is motivated by the discrete setting. Indeed, we want to avoid the discretization of mixed derivatives to lighten the cost of the macroscopic criterion. Next, the remaining integral is computed:

$$\begin{aligned} \partial_x \left\langle v h^{(3)} \right\rangle = \partial_x \Big[& 2m_2^2 \, \partial_{xx} J^\varepsilon - 2m_2^2 E \, \partial_x J^\varepsilon - m_4 \, \partial_{xx} J^\varepsilon - (m_2 - m_4) \, \partial_x (E J^\varepsilon) \\ & - (2m_2 - m_4) E \, \partial_x J^\varepsilon + (3m_2 - m_4) E^2 J^\varepsilon + \mathcal{O}(\varepsilon^2) \Big]. \end{aligned}$$

With our choice of $\mathcal{M}(v)$, we can explicitly compute $m_2 = 1$ and $m_4 = 3$. Therefore, one has:

$$\partial_x \left\langle v h^{(3)} \right\rangle = \partial_x \Big[2 \, \partial_x (E J^\varepsilon) - E \, \partial_x J^\varepsilon - \partial_{xx} J^\varepsilon + \mathcal{O}(\varepsilon^2) \Big].$$

Finally, we obtain a higher order model in the drift-diffusion limit.

Proposition 4. *(formal) Let f^ε be a solution of (VBGK). Assuming that f^ε admits a Chapman-Enskog expansion of order $K = 3$, the truncated model up to order 2 in ε is given by a higher order drift-diffusion equation. The macroscopic density $\rho^\varepsilon = \langle f^\varepsilon \rangle$ is a solution to:*

$$\partial_t \rho^\varepsilon - \partial_x J^\varepsilon + \varepsilon^2 \, \partial_x (2 \, \partial_x (E J^\varepsilon) - E \, \partial_x J^\varepsilon - \partial_{xx} J^\varepsilon) = \mathcal{O}(\varepsilon^4), \qquad (\widetilde{DD})$$

where $J^\varepsilon = \partial_x \rho^\varepsilon - E \rho^\varepsilon$.

Remark 5. *($\widetilde{DD}$) is, as expected, a second order correction of (P).*

3.3 Micro-macro model

In this part, we recall the derivation of the micro-macro model for (VBGK). Then, we introduce a micro-macro finite volume scheme that enjoys the property of being Asymptotic Preserving which is a crucial point of the hybrid method we want to construct. The micro-macro approach was first used to derive AP schemes for the radiative heat transfer in [140]. It was then applied to the Boltzmann equation in [16, 156] and to the Vlasov-Poisson-BGK equation in [59].

3.3.1 Continuous setting

Let us decompose the distribution f^ε as follows:

$$f^\varepsilon = \rho^\varepsilon \mathcal{M} + g^\varepsilon. \tag{3.9}$$

We introduce the orthogonal projector Π in $L^2(\mathrm{d}x\mathrm{d}\gamma)$ on $\mathcal{N}$ defined for all $f \in L^2(\mathrm{d}x\mathrm{d}\gamma)$ by:

$$\Pi f = \langle f \rangle \mathcal{M}.$$

To help us in the derivation of the micro-macro model, let us first recall the following lemma.

Lemma 15. *Let $f^\varepsilon = \rho^\varepsilon \mathcal{M} + g^\varepsilon$ be a solution of (VBGK). One has:*

$$\Pi(g^\varepsilon) = \Pi(\partial_t g^\varepsilon) = \Pi(\mathsf{T}(\rho^\varepsilon \mathcal{M})) = (I - \Pi)(\partial_t (\rho^\varepsilon \mathcal{M})) = 0.$$

Moreover, one has $\Pi(\mathsf{T} g^\varepsilon) = \mathrm{div}_x \langle v g^\varepsilon \rangle \mathcal{M}$.

The proof of this Lemma relies on the orthogonality properties of the decomposition (3.9). More precisely, the perturbation lies in $\mathcal{N}^\perp$ and the function $\rho^\varepsilon \mathcal{M}$ belongs to $\mathcal{N}$. Recalling that Π denotes the projection on the equilibrium manifolds $\mathcal{N}$, the identities easily obtained.

To derive the micro-macro model, we start by injecting (3.9) in (VBGK):

$$\partial_t (\rho^\varepsilon \mathcal{M}) + \partial_t g^\varepsilon + \frac{1}{\varepsilon}\Big(\mathsf{T}(\rho^\varepsilon \mathcal{M}) + \mathsf{T} g^\varepsilon\Big) = \frac{-1}{\varepsilon^2} g^\varepsilon. \tag{3.10}$$

We then apply $(I - \Pi)$ to (3.10), and simplify using Lemma 15 to obtain the micro part of the model:

$$\partial_t g^\varepsilon + \frac{1}{\varepsilon}\Big(\mathsf{T} g^\varepsilon - \Pi(\mathsf{T} g^\varepsilon)\Big) + \frac{1}{\varepsilon} v\mathcal{M} \cdot J^\varepsilon = \frac{-1}{\varepsilon^2} g^\varepsilon.$$

The macro part is obtained by applying Π to (3.10) and using Lemma 15. It leads to:

$$\partial_t \rho^\varepsilon \mathcal{M} + \frac{1}{\varepsilon} \Pi(\mathsf{T} g^\varepsilon) = 0.$$

Finally, the micro-macro model is given by:

$$\partial_t g^\varepsilon + \frac{1}{\varepsilon}(\mathsf{T} g^\varepsilon - \mathrm{div}_x \langle v g^\varepsilon \rangle \mathcal{M} + v\mathcal{M} \cdot J^\varepsilon) = \frac{-1}{\varepsilon^2} g^\varepsilon, \tag{Micro}$$

$$\partial_t \rho^\varepsilon + \frac{1}{\varepsilon} \mathrm{div}_x \langle v g^\varepsilon \rangle = 0. \tag{Macro}$$

The following proposition states the equivalence between the (Micro)-(Macro) model and the original equation (VBGK) [59].

Proposition 5. *(formal)*

1. *If f^ε is a solution to (VBGK) with an initial data f_0 in $L^2(\mathrm{d}x\mathrm{d}\gamma)$, then $(\rho^\varepsilon, g^\varepsilon) = (\langle f^\varepsilon \rangle, f^\varepsilon - \langle f^\varepsilon \rangle \mathcal{M})$ is a solution to (Micro)-(Macro) with the associated initial data*

$$\rho_0 = \langle f_0 \rangle, \quad g_0 = f_0 - \rho_0 \mathcal{M}.$$

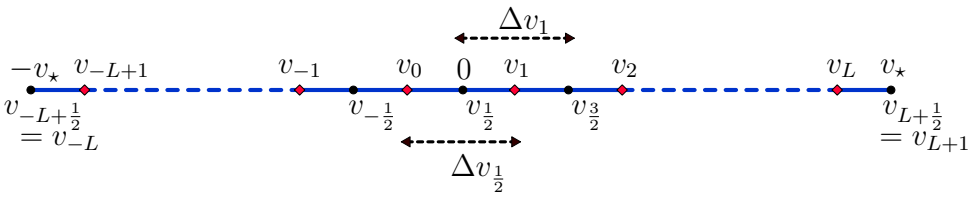


Figure 3.2 – Discretization of the velocity domain.

2. *Conversely, if* $(\rho^\varepsilon, g^\varepsilon)$ *is a solution to* (Micro)-(Macro) *with initial data* $\rho^\varepsilon(t=0) = \rho_0$ *and* $g^\varepsilon(t=0) = g_0$ *with* $\langle g_0 \rangle = 0$, *then* $\langle g^\varepsilon(t) \rangle = 0$, *for all* $t > 0$ *and* $f^\varepsilon = \rho^\varepsilon \mathcal{M} + g^\varepsilon$ *is a solution to* (VBGK) *with initial data* $f_0 = \rho_0 \mathcal{M} + g_0$.

3.3.2 Discrete setting

Let us now tackle the discretization of the (Micro)-(Macro) model. We shall adopt a finite volume approach to discretize the phase space. From now on, we restrict ourselves to the $1D_x$-$3D_v$ problem, namely one dimension in position and three in velocity. In particular, let $\Omega_x = [0, x_\star]$ with periodic boundary conditions. In this setting, the (Micro)-(Macro) equations and the transport operator T reduce to

$$\begin{cases} \partial_t g^\varepsilon + \dfrac{1}{\varepsilon^2} g^\varepsilon + \dfrac{1}{\varepsilon}\Big(\mathsf{T} g^\varepsilon - \partial_x \langle v_x g^\varepsilon \rangle \mathcal{M} + v_x \mathcal{M} J_x^\varepsilon\Big) = 0, \\ \partial_t \rho^\varepsilon + \dfrac{1}{\varepsilon} \partial_x \langle v_x g^\varepsilon \rangle = 0, \end{cases}$$

and

$$\mathsf{T}(g) = v_x \partial_x g + E_x \partial_{v_x} g,$$

where $v = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix}$, $E = \begin{pmatrix} E_x \\ 0 \\ 0 \end{pmatrix}$ and $J_x^\varepsilon = \partial_x \rho^\varepsilon - E_x \rho^\varepsilon$. In the following, we shall omit the subscript x for E_x and J_x^ε.

The mesh. The velocity domain is restricted to a bounded symmetric cube $[-v_\star, v_\star]^3$ as it is impractical to implement a numerical scheme on an unbounded domain. We consider a Cartesian mesh of the cube composed of $N_v = 2L$ velocity cells in each direction arranged symmetrically around $v = 0$. Let $\mathcal{J} = \{-L+1, \dots, L\}$ and let us denote by $j = (j_x, j_y, j_z) \in \mathcal{J}^3$ a multi-index. The cells of the velocity mesh are given by

$$\mathcal{V}_j = \mathcal{V}_{j_x} \times \mathcal{V}_{j_y} \times \mathcal{V}_{j_z} = (v_{j_x - \frac{1}{2}}, v_{j_x + \frac{1}{2}}) \times (v_{j_y - \frac{1}{2}}, v_{j_y + \frac{1}{2}}) \times (v_{j_z - \frac{1}{2}}, v_{j_z + \frac{1}{2}}), \quad j \in \mathcal{J}^3.$$

Each cell $\mathcal{V}_j$ has a constant volume Δv^3 and midpoint v_j. The velocity mesh is illustrated in one dimension in Figure 3.2.

In position, because of the periodic boundary conditions, we consider a discretization of the 1-D torus $\mathbb{T}$ into N_x primal cells

$$\mathcal{X}_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}), \quad i \in \mathcal{I} = \mathbb{Z}/N_x\mathbb{Z},$$

of constant length Δx and centers x_i. We also define the dual cells

$$\mathcal{X}_{i+\frac{1}{2}} = (x_i, x_{i+1}), \quad i \in \mathcal{I},$$

of constant length Δx and centers $x_{i+\frac{1}{2}}$. The primal control volumes in the phase space are defined by

$$K_{ij} = \mathcal{X}_i \times \mathcal{V}_j, \quad \forall (i,j) \in \mathcal{I} \times \mathcal{J}^3,$$

while the dual control volumes are given by

$$K_{i+\frac{1}{2},j} = \mathcal{X}_{i+\frac{1}{2}} \times \mathcal{V}_j, \quad \forall (i,j) \in \mathcal{I} \times \mathcal{J}^3.$$

Finally, we set a time step $\Delta t > 0$, and we define $t^n = n\Delta t, \quad n \in \mathbb{N}$.

The discrete Maxwellian. We assume that we have given cell values of the unidimensional Maxwellian $(M_l)_{l\in\mathcal{J}}$ satisfying:

$$\begin{cases} M_l > 0, \quad M_l = M_{-l+1} \quad \forall l \in \mathcal{J}, \\ M_L = M_{L+1}, \quad M_{-L} = M_{-L+1}, \\ \displaystyle\sum_{l\in\mathcal{J}} M_l \Delta v = 1. \end{cases} \tag{3.11}$$

These properties are a discrete version of the continuous ones, namely, the positivity, the parity, and the unit mass. The second line is a zero boundary flux on the Maxwellian. For a sufficiently large domain in velocity, it is relevant due to the fast decay of the Gaussian. We now define the discrete multidimensional Maxwellian as:

$$\mathcal{M}_j = M_{j_x} M_{j_y} M_{j_z}, \tag{3.12}$$

where $(M_{j_x})_{j_x\in\mathcal{J}}$, $(M_{j_y})_{j_y\in\mathcal{J}}$ and $(M_{j_z})_{j_z\in\mathcal{J}}$ satisfy (3.11) and j is the multi-index (j_x, j_y, j_z).

Let us introduce the discrete integration operator in velocity: for $f = \left(f_j\right)_{j\in\mathcal{J}^3}$

$$\langle f \rangle_\Delta = \sum_{j\in\mathcal{J}^3} f_j \Delta v^3. \tag{3.13}$$

We also introduce the discrete moments of the discrete 1D Maxwellian (3.11):

$$m_k^{\Delta v} = \sum_{l\in\mathcal{J}} v_l^k M_l \Delta v. \tag{3.14}$$

Semi-discretization in the phase-space. We start by considering a semi-discretization in the phase-space of the (Micro)-(Macro) model. Let $(i,j) \in \mathcal{I} \times \mathcal{J}^3$. We choose to approximate the perturbation g^ε on the dual cells while the density ρ^ε is approximated on the primal mesh:

$$g^\varepsilon_{i+\frac{1}{2},j}(t) \approx \frac{1}{\Delta x \Delta v^3} \int_{K_{i+\frac{1}{2},j}} g^\varepsilon(t,x,v)\mathrm{d}x\mathrm{d}v \text{ and } \rho^\varepsilon_i(t) \approx \frac{1}{\Delta x} \int_{\mathcal{X}_i} \rho^\varepsilon(t,x)\mathrm{d}x. \tag{3.15}$$

This choice of staggered meshes will result in a more compact stencil for the asymptotic scheme and is quite standard in the literature [59, 156, 161]. We start by integrating (Macro) on $\mathcal{X}_i$:

$$\int_{\mathcal{X}_i} \left[\partial_t \rho^\varepsilon + \frac{1}{\varepsilon} \partial_x \langle v_x g^\varepsilon \rangle \right] \mathrm{d}x = 0.$$

After integrating the space derivative, one then obtains a continuous in time finite volume scheme for (Macro):

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_i^\varepsilon = -\frac{1}{\Delta x}\left(\frac{1}{\varepsilon}\langle \xi g^\varepsilon_{i+\frac{1}{2}} \rangle_\Delta - \frac{1}{\varepsilon}\langle \xi g^\varepsilon_{i-\frac{1}{2}} \rangle_\Delta \right), \text{ where } \xi_j = \xi_{(j_x,j_y,j_z)} = v_{j_x} \quad \forall j \in \mathcal{J}^3. \tag{3.16}$$

Next, we deal with the (Micro) equation which is integrated on $K_{i+\frac{1}{2},j}$:

$$\begin{aligned} &\int_{K_{i+\frac{1}{2},j}} \left(\partial_t g^\varepsilon + \frac{1}{\varepsilon^2} g^\varepsilon \right) \mathrm{d}x\mathrm{d}v \\ &+ \frac{1}{\varepsilon} \int_{K_{i+\frac{1}{2},j}} [\mathsf{T} g^\varepsilon - \partial_x \langle v_x g^\varepsilon \rangle \mathcal{M} + v_x \mathcal{M} J^\varepsilon] \, \mathrm{d}x\mathrm{d}v = 0. \end{aligned} \tag{3.17}$$

One then obtains

$$\begin{aligned} &-\varepsilon \Delta x \Delta v^3 \left(\frac{\mathrm{d}}{\mathrm{d}t} g^\varepsilon_{i+\frac{1}{2},j} + \frac{1}{\varepsilon^2} g^\varepsilon_{i+\frac{1}{2},j} \right) \\ &= \underbrace{\int_{K_{i+\frac{1}{2},j}} \mathsf{T} g^\varepsilon \mathrm{d}x\mathrm{d}v}_{A} - \underbrace{\int_{K_{i+\frac{1}{2},j}} \partial_x \langle v_x g^\varepsilon \rangle \mathcal{M} \mathrm{d}x\mathrm{d}v}_{B} + \underbrace{\int_{K_{i+\frac{1}{2},j}} v_x \mathcal{M} J^\varepsilon \mathrm{d}x\mathrm{d}v}_{C}. \end{aligned} \tag{3.18}$$

Using the definition of the transport operator T, one has:

$$\begin{aligned} A &= \int_{\mathcal{V}_j} v_x \Big(g^\varepsilon(t, x_{i+1}, v) - g^\varepsilon(t, x_i, v) \Big) \mathrm{d}v \\ &\quad + \int_{\mathcal{X}_{i+\frac{1}{2}} \times \mathcal{V}_{j_y} \times \mathcal{V}_{j_z}} E \Big(g^\varepsilon(t, x, v_{j_x+\frac{1}{2}}, v_y, v_z) - g^\varepsilon(t, x, v_{j_x-\frac{1}{2}}, v_y, v_z) \Big) \mathrm{d}x\mathrm{d}v_y\mathrm{d}v_z, \\ B &= \int_{\mathcal{V}_j} \mathcal{M} \Big(\langle v_x g^\varepsilon(t, x_{i+1}, v) \rangle - \langle v_x g^\varepsilon(t, x_i, v) \rangle \Big) \mathrm{d}v, \\ C &= \int_{\mathcal{V}_j} v_x \mathcal{M} \mathrm{d}v \int_{\mathcal{X}_{i+\frac{1}{2}}} J^\varepsilon \mathrm{d}x. \end{aligned}$$

We now denote by $\left(\mathcal{F}^\varepsilon_{i,j}\right)_{ij}$ an approximation of the microscopic flux in position at interfaces $(x_i)_i$, namely

$$\mathcal{F}^\varepsilon_{i,j} \approx \int_{\mathcal{V}_j} [v_x g^\varepsilon(t, x_i, v) - \mathcal{M} \langle v_x g^\varepsilon(t, x_i, v) \rangle] \, \mathrm{d}v. \tag{3.19}$$

We also denote by $\left(\mathcal{G}^{\varepsilon}_{i+\frac{1}{2},j_x+\frac{1}{2},j_y,j_z}\right)_{ij}$ an approximation of the microscopic flux in velocity, namely

$$\mathcal{G}^{\varepsilon}_{i+\frac{1}{2},j_x+\frac{1}{2},j_y,j_z} \approx \int_{\mathcal{X}_{i+\frac{1}{2}}\times\mathcal{V}_{j_y}\times\mathcal{V}_{j_z}} E g^{\varepsilon}(t,x,v_{x,j+\frac{1}{2}},v_y,v_z)\mathrm{d}x\mathrm{d}v_x\mathrm{d}v_y. \tag{3.20}$$

In the following, to lighten the notation, we will denote by $\mathcal{G}^{\varepsilon}_{i+\frac{1}{2},j+\frac{1}{2}}$ this flux. Let us now present our choice of numerical fluxes. In position, we choose a first-order upwind approximation for both terms of (3.19):

$$\mathcal{F}^{\varepsilon}_{i,j} = \left(\xi^+_j g^{\varepsilon,n}_{i-\frac{1}{2},j} + \xi^-_j g^{\varepsilon}_{i+\frac{1}{2},j}\right)\Delta v^3 - \mathcal{M}_j\left\langle \xi^+ g^{\varepsilon}_{i-\frac{1}{2}} + \xi^- g^{\varepsilon}_{i+\frac{1}{2}}\right\rangle_{\Delta}\Delta v^3,$$

where the notation $r^{\pm} = \frac{r\pm|r|}{2}$ is used. At the boundaries in position, the periodic setting implies

$$\mathcal{F}^{\varepsilon}_{0,j} = \mathcal{F}^{\varepsilon}_{N_x,j}.$$

In velocity, a first-order upwind approximation is used and since E is given, it is explicitly discretized on the dual mesh, $E_{i+\frac{1}{2}} = E(x_{i+\frac{1}{2}})$. The numerical flux in velocity then reads:

$$\mathcal{G}^{\varepsilon}_{i+\frac{1}{2},j+\frac{1}{2}} = \left(E^+_{i+\frac{1}{2}} g^{\varepsilon,n}_{i+\frac{1}{2},j_x,j_y,j_z} + E^-_{i+\frac{1}{2}} g^{\varepsilon}_{i+\frac{1}{2},j_x+1,j_y,j_z}\right)\Delta x\Delta v^2.$$

Zero flux boundary conditions are applied in velocity, and therefore we set

$$\mathcal{G}^{\varepsilon}_{i+\frac{1}{2},-L+\frac{1}{2}} = \mathcal{G}^{\varepsilon}_{i+\frac{1}{2},L+\frac{1}{2}} = 0.$$

Finally, C is treated as a source term and approximated using first-order centered finite differences:

$$C \approx \xi_j \mathcal{M}_j J^{\varepsilon}_{i+\frac{1}{2}}\Delta x\Delta v^3,$$

with $J^{\varepsilon}_{i+\frac{1}{2}} = \frac{\rho^{\varepsilon}_{i+1}-\rho^{\varepsilon}_i}{\Delta x} - E_{i+\frac{1}{2}}\rho^{\varepsilon}_{i+\frac{1}{2}}$, and $\rho^{\varepsilon}_{i+\frac{1}{2}} = \frac{1}{2}(\rho^{\varepsilon}_i + \rho^{\varepsilon}_{i+1})$. A continuous in time finite volume scheme for equation (Micro) finally reads:

$$\frac{\mathrm{d}}{\mathrm{d}t} g^{\varepsilon}_{i+\frac{1}{2},j} + \frac{1}{\varepsilon^2} g^{\varepsilon}_{i+\frac{1}{2},j} = -\frac{1}{\varepsilon}\left(\frac{T^{\varepsilon}_{i+\frac{1}{2},j}}{\Delta x\Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon}_{i+\frac{1}{2}}\right), \tag{3.21}$$

where $T^{\varepsilon}_{i+\frac{1}{2},j} = \mathcal{F}^{\varepsilon}_{i+1,j} - \mathcal{F}^{\varepsilon}_{i,j} + \mathcal{G}^{\varepsilon}_{i+\frac{1}{2},j+\frac{1}{2}} - \mathcal{G}^{\varepsilon}_{i+\frac{1}{2},j-\frac{1}{2}}$.

Full discretization. In order to obtain an AP scheme, one must carefully choose the time discretization. Following [155], we adapt the so-called relaxed micro-macro scheme to our finite volume setting. This method falls into the framework of exponential time integrators [126]. Let $n \in \mathbb{N}$ and $(i,j) \in \mathcal{I}\times\mathcal{J}^3$. Let $\left(g^{\varepsilon,n}_{i+\frac{1}{2},j}\right)_{ij}$ be an approximation of $\left(g^{\varepsilon}_{i+\frac{1}{2},j}(t^n)\right)_{ij}$ and $\left(\rho^{\varepsilon,n}_i\right)_i$ an approximation of $\left(\rho^{\varepsilon}_i(t^n)\right)_i$. The first step is to multiply (3.21) by e^{t/ε^2} which gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(g^{\varepsilon}_{i+\frac{1}{2},j}(t)e^{t/\varepsilon^2}\right) = -\frac{e^{t/\varepsilon^2}}{\varepsilon}\left(\frac{T^{\varepsilon}_{i+\frac{1}{2},j}}{\Delta x\Delta v^3} + \xi_j\mathcal{M}_j J^{\varepsilon}_{i+\frac{1}{2}}\right).$$

Let us then integrate between t^n and t^{n+1} and divide by $e^{t^{n+1}/\varepsilon^2}$:

$$g^{\varepsilon}_{i+\frac{1}{2},j}(t^{n+1}) = g^{\varepsilon}_{i+\frac{1}{2},j}(t^n)e^{-\Delta t/\varepsilon^2} + \int_{t^n}^{t^{n+1}} -\frac{e^{(t-t^{n+1})/\varepsilon^2}}{\varepsilon}\left(\frac{T^{\varepsilon}_{i+\frac{1}{2},j}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon}_{i+\frac{1}{2}}\right)\mathrm{d}t.$$

Then, the transport and source terms are approximated at time t^n and the integral can be computed explicitly:

$$\begin{aligned}&\int_{t^n}^{t^{n+1}} -\frac{e^{(t-t^{n+1})/\varepsilon^2}}{\varepsilon}\left(\frac{T^{\varepsilon}_{i+\frac{1}{2},j}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon}_{i+\frac{1}{2}}\right)\mathrm{d}t \\ &= -\varepsilon(1-e^{-\Delta t/\varepsilon^2})\left(\frac{T^{\varepsilon,n}_{i+\frac{1}{2},j}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon,n}_{i+\frac{1}{2}}\right).\end{aligned}$$

Finally, the fully discretized microscopic equation reads:

$$g^{\varepsilon,n+1}_{i+\frac{1}{2},j} = g^{\varepsilon,n}_{i+\frac{1}{2},j}e^{-\Delta t/\varepsilon^2} - \varepsilon(1-e^{-\Delta t/\varepsilon^2})\left(\frac{T^{\varepsilon,n}_{i+\frac{1}{2},j}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon,n}_{i+\frac{1}{2}}\right), \tag{3.22}$$

where

$$T^{\varepsilon,n}_{i+\frac{1}{2},j} = \mathcal{F}^{\varepsilon,n}_{i+1,j} - \mathcal{F}^{\varepsilon,n}_{i,j} + \mathcal{G}^{\varepsilon,n}_{i+\frac{1}{2},j+\frac{1}{2}} - \mathcal{G}^{\varepsilon,n}_{i+\frac{1}{2},j-\frac{1}{2}}. \tag{3.23}$$

The discretization in time of (3.16) is quite standard, with an implicit discretization of the stiff term:

$$\rho^{\varepsilon,n+1}_i = \rho^{\varepsilon,n}_i - \frac{\Delta t}{\varepsilon \Delta x}\left(\langle \xi g^{\varepsilon,n+1}_{i+\frac{1}{2}}\rangle_\Delta - \langle \xi g^{\varepsilon,n+1}_{i-\frac{1}{2}}\rangle_\Delta\right). \tag{3.24}$$

Note that (3.22) defines an explicit scheme. Moreover, (3.24) does not require the inversion of a system. Indeed, (3.22) is explicitly computed at time t^{n+1} and is then used to update the density in (3.24). In practice, the method is therefore fully explicit.

Before stating the next proposition, let us introduce the following assumption:

Assumption 1. *Let $\left(\rho^{\varepsilon,n}_i\right)_{i\in\mathcal{I}}$ be given by (3.24). Then,*

$$\rho^{\varepsilon,n}_i \longrightarrow \rho^n_i, \quad \forall i \in \mathcal{I}. \tag{3.25}$$

This assumption corresponds to the convergence of ρ^ε to ρ as $\varepsilon \to 0$. Such property is not trivial to obtain in the discrete setting. A rigorous proof of this result requires, among other things, uniform estimates in ε of the discrete L^2-norm of ρ^ε, g^ε, and moments of g^ε. It is outside the scope of this chapter and may be thoroughly investigated in upcoming work.

The following proposition states the AP property of our discretization of the (Micro)-(Macro) model.

Proposition 6. *Let $n \in \mathbb{N}$. Let $\left(g^{\varepsilon,n}_{i+\frac{1}{2},j}\right)_{ij}$ and $\left(\rho^{\varepsilon,n}_i\right)_i$ be given by the following micro-macro finite volume scheme:*

$$g^{\varepsilon,n+1}_{i+\frac{1}{2},j} = g^{\varepsilon,n}_{i+\frac{1}{2},j}e^{-\Delta t/\varepsilon^2} - \varepsilon(1-e^{-\Delta t/\varepsilon^2})\left(\frac{T^{\varepsilon,n}_{i+\frac{1}{2},j}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon,n}_{i+\frac{1}{2}}\right), \tag{3.26}$$

$$\rho_i^{\varepsilon,n+1} = \rho_i^{\varepsilon,n} - \frac{\Delta t}{\varepsilon \Delta x}\left(\langle \xi g_{i+\frac{1}{2}}^{\varepsilon,n+1}\rangle_\Delta - \langle \xi g_{i-\frac{1}{2}}^{\varepsilon,n+1}\rangle_\Delta\right), \tag{3.27}$$

where $T_{i+\frac{1}{2},j}^{\varepsilon,n}$ *is given by* (3.23).

Assuming that Assumption 1 holds and for a fixed mesh size $\Delta x, \Delta v > 0$*, the scheme enjoys the AP property in the diffusion limit. This property does not depend on the initial data, and the associated limit scheme reads*

$$\rho_i^{n+1} = \rho_i^n + m_2^{\Delta v}\frac{\Delta t}{\Delta x}\left(J_{i+\frac{1}{2}}^n - J_{i-\frac{1}{2}}^n\right), \tag{S_{Lim}}$$

with the limit flux

$$J_{i+\frac{1}{2}}^n = \frac{\rho_{i+1}^n - \rho_i^n}{\Delta x} - E_{i+\frac{1}{2}}\rho_{i+\frac{1}{2}}^n, \tag{3.28}$$

where $m_2^{\Delta v}$ *is given by* (3.14).

Proof. The mesh size $\Delta x, \Delta v > 0$ being set, let us emphasize that we consider only the pointwise convergence of the scheme as ε tends to 0. The first step is to study the asymptotic behavior of the perturbation $\left(g_{i+\frac{1}{2},j}^{\varepsilon,n+1}\right)_{i,j}$. By induction on n, let us show that $g_{i+\frac{1}{2},j}^{\varepsilon,n+1} \underset{\varepsilon\to 0}{\longrightarrow} 0$ for any initial data (ρ^0, g^0) and for all $(i,j) \in \mathcal{I}\times\mathcal{J}^3$. At $n=0$, one has

$$g_{i+\frac{1}{2},j}^{\varepsilon,1} = g_{i+\frac{1}{2},j}^0 e^{-\Delta t/\varepsilon^2} - \varepsilon(1-e^{-\Delta t/\varepsilon^2})\left(\frac{T_{i+\frac{1}{2},j}^{\varepsilon,0}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J_{i+\frac{1}{2}}^{\varepsilon,0}\right). \tag{3.29}$$

As $e^{-\Delta t/\varepsilon^2} \underset{\varepsilon\to 0}{\longrightarrow} 0$ and since $\left(\frac{T_{i+\frac{1}{2},j}^{\varepsilon,0}}{\Delta x\Delta v^3} + \xi_j \mathcal{M}_j J_{i+\frac{1}{2}}^{\varepsilon,0}\right)_{ij}$ depends only on the initial data which itself is independent of ε,

$$g_{i+\frac{1}{2},j}^{\varepsilon,1} \underset{\varepsilon\to 0}{\longrightarrow} 0 \quad \forall (i,j)\in\mathcal{I}\times\mathcal{J}^3.$$

Let us now assume that

$$g_{i+\frac{1}{2},j}^{\varepsilon,n} \underset{\varepsilon\to 0}{\longrightarrow} 0 \quad \forall (i,j)\in\mathcal{I}\times\mathcal{J}^3. \tag{3.30}$$

Under the hypothesis (3.30), one obtains that the transport term, that depends only on the perturbation, vanishes in the limit:

$$T_{i+\frac{1}{2},j}^{\varepsilon,n} \underset{\varepsilon\to 0}{\longrightarrow} 0 \quad \forall (i,j)\in\mathcal{I}\times\mathcal{J}^3.$$

From Assumption 1 one obtains the convergence of the source term:

$$\xi_j \mathcal{M}_j J_{i+\frac{1}{2}}^{\varepsilon,n} \underset{\varepsilon\to 0}{\longrightarrow} \xi_j \mathcal{M}_j J_{i+\frac{1}{2}}^n. \tag{3.31}$$

Now, the asymptotic limit of (3.26) can be computed. Since $\varepsilon(1-e^{-\Delta t/\varepsilon^2}) \underset{\varepsilon\to 0}{\longrightarrow} 0$ and $e^{-\Delta t/\varepsilon^2} \underset{\varepsilon\to 0}{\longrightarrow} 0$ we can use (3.31) and our induction hypothesis (3.30) to obtain:

$$g_{i+\frac{1}{2},j}^{\varepsilon,n+1} \underset{\varepsilon\to 0}{\longrightarrow} 0, \quad \forall n, \quad \forall (i,j)\in\mathcal{I}\times\mathcal{J}. \tag{3.32}$$

As a consequence, one also has that for all n,

$$\frac{T_{i+\frac{1}{2},j}^{\varepsilon,n}}{\Delta x\Delta v^3} + \xi_j \mathcal{M}_j J_{i+\frac{1}{2}}^{\varepsilon,n} \underset{\varepsilon\to 0}{\longrightarrow} \xi_j \mathcal{M}_j J_{i+\frac{1}{2}}^n, \quad \forall (i,j)\in\mathcal{I}\times\mathcal{J}. \tag{3.33}$$

The next step is to plug (3.26) into (3.27):

$$\begin{aligned}\rho_i^{\varepsilon,n+1} =& \rho_i^{\varepsilon,n} - \frac{\Delta t}{\varepsilon \Delta x}\left\langle \xi \left[e^{-\Delta t/\varepsilon^2}(g_{i+\frac{1}{2}}^{\varepsilon,n} - g_{i-\frac{1}{2}}^{\varepsilon,n}) \right.\right. \\ & \left.\left. - \varepsilon(1-e^{-\Delta t/\varepsilon^2})\left(\frac{T_{i+\frac{1}{2}}^{\varepsilon,n} - T_{i-\frac{1}{2}}^{\varepsilon,n}}{\Delta x \Delta v^3} + (\xi \mathcal{M} J_{i+\frac{1}{2}}^{\varepsilon,n} - \xi \mathcal{M} J_{i-\frac{1}{2}}^{\varepsilon,n}) \right) \right] \right\rangle_\Delta .\end{aligned} \tag{3.34}$$

We can then take the limit $\varepsilon \longrightarrow 0$ in (3.34) using the previous asymptotic limits (3.32) and (3.33):

$$\rho_i^{n+1} = \rho_i^n + \langle \xi^2 \mathcal{M} \rangle_\Delta \frac{\Delta t}{\Delta x}\left(J_{i+\frac{1}{2}}^n - J_{i-\frac{1}{2}}^n \right). \tag{3.35}$$

Then, using the assumptions on the discrete Maxwellian (3.11), (3.12) one can obtain:

$$\begin{aligned}\langle \xi^2 \mathcal{M} \rangle_\Delta &= \sum_{j \in \mathcal{J}^3} \xi_j^2 \mathcal{M}_j \Delta v^3 \\ &= \sum_{j_x \in \mathcal{J}} \sum_{j_y \in \mathcal{J}} \sum_{j_z \in \mathcal{J}} v_{j_x}^2 M_{j_x} M_{j_y} M_{j_z} \Delta v^3 \\ &= \sum_{j_x \in \mathcal{J}} v_{j_x}^2 M_{j_x} \Delta v \\ &= m_2^{\Delta v}.\end{aligned} \tag{3.36}$$

Finally, we obtain the asymptotic scheme (S_{Lim}):

$$\rho_i^{n+1} = \rho_i^n + m_2^{\Delta v} \frac{\Delta t}{\Delta x}\left(J_{i+\frac{1}{2}}^n - J_{i-\frac{1}{2}}^n \right),$$

which concludes the proof. □

In order to show that the scheme (3.26)-(3.27) is truly AP, one also needs the stability condition to be independent (or at least not degenerate) on ε. While we do not prove the stability of the scheme, in practice, we can indeed use the same time-step for both large and small values of ε.

3.4 Hybrid method

The aim of this section is to introduce a hybrid method between kinetic and fluid schemes. The goal is to obtain a coupled solver that is faster than a full kinetic one to solve (P^ε) while still being accurate. These methods come naturally when designing accurate numerical codes while ensuring reasonable computation times.

Following [87] we first construct a hybrid kinetic/fluid solver with a dynamic domain adaptation method and present its implementation. In the second part, we are interested in understanding the conservative aspect of the method. More precisely, we give a result on the mass variation induced by the coupling.

Derivative \ Index	-3	-2	-1	0	1	2	3
1	-1/60	3/20	-3/4	0	3/4	-3/20	1/60
2	1/90	-3/20	3/2	-49/18	3/2	-3/20	1/90
3	1/8	-1	13/8	0	-13/8	1	-1/8
4	-1/6	2	-13/2	28/3	-13/2	2	-1/6

Table 3.1 – Central finite differences coefficients.

3.4.1 Coupling criteria

The idea of the dynamic domain adaptation method is twofold. First, the subdomains must accurately describe the state of the solution. In particular, the fluid model is only valid where the solution is near the local equilibrium in velocity. Secondly, we want the method to be dynamic in the sense that the subdomains are adapted at each time step. For this purpose, let us introduce $\Omega_{\mathcal{K}}^n$ the kinetic domain and $\Omega_{\mathcal{F}}^n$ the fluid one at time t^n. To determine in which domain each cell lies, we introduce criteria based on the higher order fluid model introduced in Section 3.2.2 and the norm of the perturbation $g^\varepsilon = f^\varepsilon - \rho^\varepsilon \mathcal{M}$. Indeed, when g^ε is close to 0, it means that the solution is close to the local equilibrium in velocity.

Let us consider a fluid subdomain. In this subdomain, one only has access to the macroscopic quantity ρ and the given electrical field E. Therefore, one cannot consider the perturbation g^ε. The solution we propose is to use the higher order model ($\widetilde{DD}$) and derive a macroscopic criterion. We have formally shown that ($\widetilde{DD}$) can be written in the form:

$$\partial_t \rho^\varepsilon - m_2 \mathrm{div}_x J^\varepsilon = \mathcal{R}^\varepsilon,$$

where $\mathcal{R}^\varepsilon$ is a remainder that depends only on the density ρ^ε and the electrical field E. During the coupling procedure it will be computed using both the kinetic density ρ^ε in kinetic cells and using the fluid density ρ in fluid cells. In the $1D_x$-$3D_v$ setting, it is given by

$$\mathcal{R}^\varepsilon = -\varepsilon^2 \partial_x (2\,\partial_x (E J^\varepsilon) - E\,\partial_x J^\varepsilon - \partial_{xx} J^\varepsilon), \quad \text{where } J^\varepsilon = \partial_x \rho^\varepsilon - E\rho^\varepsilon.$$

Expanding $\mathcal{R}^\varepsilon$ shows that one needs derivatives of the density up to fourth order and of E up to third order:

$$\begin{aligned}
\mathcal{R}^\varepsilon = -\varepsilon^2 \Big(& - \partial_{xxxx} \rho^\varepsilon \\
& + E(2\,\partial_{xxx} \rho^\varepsilon - E\,\partial_{xx} \rho^\varepsilon) \\
& + \partial_x E(-3\rho^\varepsilon\,\partial_x E - 5E\,\partial_x \rho^\varepsilon + 6\,\partial_{xx} \rho^\varepsilon) \\
& + \partial_{xx} E(-3\rho^\varepsilon E + 5\,\partial_x \rho^\varepsilon) \\
& + \rho^\varepsilon\,\partial_{xxx} E \Big).
\end{aligned}$$

Let us denote by $\mathcal{R}_i^{\varepsilon,n}$ a discretization of the remainder $\mathcal{R}^\varepsilon$ at time t^n in cell $\mathcal{X}_i$. High order finite difference schemes are used (See Table 3.1) to compute $\mathcal{R}_i^{\varepsilon,n}$.

Let $\eta_0, \delta_0 > 0$ be the coupling thresholds. In a fluid domain, when $\mathcal{R}_i^{\varepsilon,n}$ is large, the model ($\widetilde{DD}$) is far from the limit model (P) and one must use the kinetic one instead. More specifically, consider

a fluid cell $\mathcal{X}_i \subset \Omega_{\mathcal{F}}^n$.

— If $\left|\mathcal{R}_i^{\varepsilon,n}\right| \leq \eta_0$, then the cell stays fluid at t^{n+1}.

— If $\left|\mathcal{R}_i^{\varepsilon,n}\right| > \eta_0$, then the cell becomes kinetic at t^{n+1}:

$$\mathcal{X}_i \not\subset \Omega_{\mathcal{F}}^{n+1} \quad \text{and} \quad \mathcal{X}_i \subset \Omega_{\mathcal{K}}^{n+1}.$$

In a kinetic subdomain, unlike the previous case, one has access to the perturbation g^ε. When this perturbation is small, it means that the solution is near a local equilibrium in velocity. As a consequence, the behavior of the system is close to the fluid one and one can use the limit model instead. Moreover, we also use the criterion that the remainder $\mathcal{R}^\varepsilon$ must be small. Consider now a kinetic cell $\mathcal{X}_i \subset \Omega_{\mathcal{K}}^n$:

— If $\|g_{i-\frac{1}{2}}^{\varepsilon,n}\|_2 > \delta_0$ and $\|g_{i+\frac{1}{2}}^{\varepsilon,n}\|_2 > \delta_0$ then the cell stays kinetic at t^{n+1}.

— If $\|g_{i-\frac{1}{2}}^{\varepsilon,n}\|_2 \leq \delta_0$, $\|g_{i+\frac{1}{2}}^{\varepsilon,n}\|_2 \leq \delta_0$ and $\left|\mathcal{R}_i^{\varepsilon,n}\right| > \eta_0$, then the cell stays kinetic at t^{n+1}.

— If $\|g_{i-\frac{1}{2}}^{\varepsilon,n}\|_2 \leq \delta_0$, $\|g_{i+\frac{1}{2}}^{\varepsilon,n}\|_2 \leq \delta_0$ and $\left|\mathcal{R}_i^{\varepsilon,n}\right| \leq \eta_0$, then the cell becomes fluid at t^{n+1}:

$$\mathcal{X}_i \not\subset \Omega_{\mathcal{K}}^{n+1} \quad \text{and} \quad \mathcal{X}_i \subset \Omega_{\mathcal{F}}^{n+1}.$$

The discrete norm $\|g_{i+\frac{1}{2}}^{\varepsilon,n}\|_2$ is the classic l^2-norm. It is defined for $\left(g_{i+\frac{1}{2},j}^{\varepsilon,n}\right)_{j\in\mathcal{J}^3}$ by:

$$\|g_{i+\frac{1}{2}}^{\varepsilon,n}\|_2 = \sum_{j\in\mathcal{J}^3} \left(g_{i+\frac{1}{2},j}^{\varepsilon,n}\right)^2 \Delta v^3.$$

Remark 6. *Note that in a kinetic cell, the criterion on the norm of g^ε is mandatory. Indeed, the remainder $\mathcal{R}_i^{\varepsilon,n}$ could be small because of small gradients, but the perturbation large. In this situation, one does not want to change from kinetic to fluid. As an example, one could take a distribution function at equilibrium in position and far from the Maxwellian in velocity.*

3.4.2 Implementation

We now present in more details the implementation of the hybrid method. An important part of this approach is the management of boundary conditions. When solving on the whole space domain, periodic boundary conditions are applied. However, when solving in the subdomains $\Omega_{\mathcal{K}}^n$ and $\Omega_{\mathcal{F}}^n$, we need to adapt our solvers. Our strategy is to use ghost cell values that are chosen appropriately. The difficulty lies in the fact that the limit scheme only computes the density ρ and not the pair $(\rho^\varepsilon, g^\varepsilon)$. Since the hybrid method is dynamic, one does not know in advance the state of the cells. As a consequence, one must be able to access all unknowns on the whole domain at any time. Our solution is to take advantage of the structure of the micro-macro scheme. Indeed, aside from visualization and diagnostics, an explicit discretization of the distribution function isn't necessary. We are working only with $\rho^{\varepsilon,n}$ and $g^{\varepsilon,n}$. Therefore, we have access to the macro unknown on the whole domain and there is no information missing in the arrays. The distribution f^ε is reconstructed using $f_{i,j}^{\varepsilon,n} = \rho_i^{\varepsilon,n}\mathcal{M}_j + \frac{1}{2}\left(g_{i-\frac{1}{2},j}^{\varepsilon,n} + g_{i+\frac{1}{2},j}^{\varepsilon,n}\right)$ only for posttreatment. However, the kinetic solver may still need values of g^ε on the whole space domain. In theory, the array storing g^ε must be filled with zeros in the fluid domain. However, to improve the performance, g^ε is in

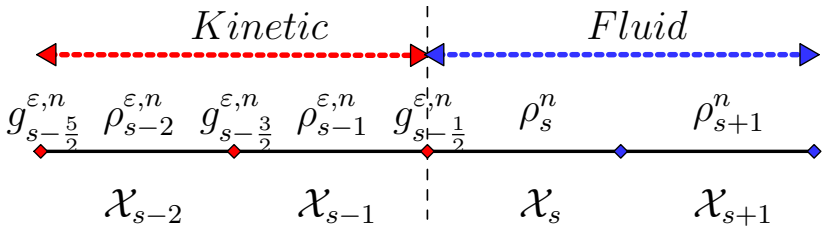


Figure 3.3 – Transition between kinetic and fluid cell for the micro-macro scheme.

practice not updated in fluid regions, and it is set to 0 only when needed. In particular, it only occurs when a fluid cell becomes kinetic and when saving data.

Another important remark is that since g^ε is approximated on the dual mesh, one must be careful at the interfaces between kinetic and fluid subdomains. To avoid any ambiguity on the state of an interface when updating the perturbation g^ε, we impose that a fluid subdomain is at least two cells wide. Under this condition the state of the ghost interface is well determined. See Figure 3.3 for an illustration of such a situation.

The algorithm can be summarized as follows:

Algorithm 2 Hybrid scheme

1. Set ε, δ_0, η_0 and a final time T.
2. Initialize micro-macro unknowns using the relations $\rho^0 = \langle f^0 \rangle$ and $g^0 = f^0 - \rho^0 M$.
3. Initialize $\Omega^0_{\mathcal{K}}$ as the whole space ($\Omega^0_{\mathcal{F}} = \emptyset$).
4. Compute $g^{\varepsilon,n+1}$ and $\rho^{\varepsilon,n+1}$ in $\Omega^n_{\mathcal{K}}$ using the kinetic scheme (3.26)-(3.27).
5. Compute ρ^{n+1} in $\Omega^n_{\mathcal{F}}$ using the limit scheme (S_{Lim}).
6. Set $g^{n+1} = 0$ in $\Omega^n_{\mathcal{F}}$.
7. Update $\Omega^n_{\mathcal{K}}$ and $\Omega^n_{\mathcal{F}}$ to $\Omega^{n+1}_{\mathcal{K}}$ and $\Omega^{n+1}_{\mathcal{F}}$ using the criteria presented above.
8. Increment time and repeat until $t^{n+1} = T$.

In particular, Algorithm 2 explicitly defines a numerical scheme on the hybrid density $\tilde{\rho}$:

$$\tilde{\rho}_i^{n+1} = \tilde{\rho}_i^n + \frac{\Delta t}{\Delta x} J_i^{H,n}, \tag{3.37}$$

where

$$J_i^{H,n} = \begin{cases} -\dfrac{1}{\varepsilon}\left(\langle \xi g^{\varepsilon,n+1}_{i+\frac{1}{2}} \rangle_\Delta - \langle \xi g^{\varepsilon,n+1}_{i-\frac{1}{2}} \rangle_\Delta\right) & \text{if} \quad \mathcal{X}_i \in \Omega^n_{\mathcal{K}}, \\ m_2^{\Delta v}\left(J^n_{i+\frac{1}{2}} - J^n_{i-\frac{1}{2}}\right) & \text{if} \quad \mathcal{X}_i \in \Omega^n_{\mathcal{F}}. \end{cases} \tag{3.38}$$

Note that we want to start the resolution with the approach containing the full information on the system. Hence, it makes sense to initialize our domain as fully kinetic. Moreover, let us emphasize again that the kinetic fluxes are in practice explicitly computed. Therefore, the hybrid setting remains an explicit method.

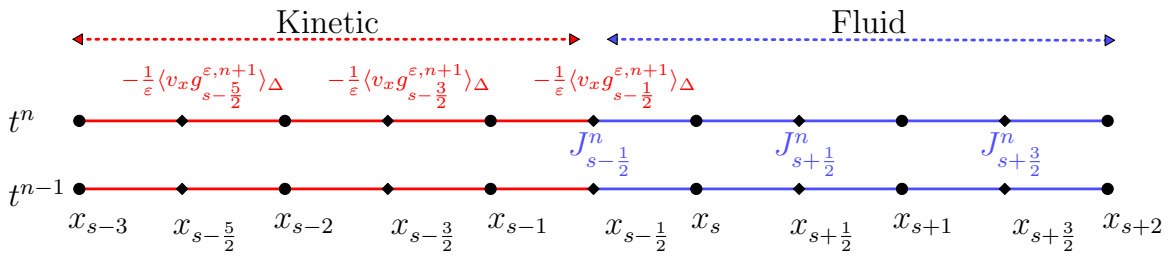


Figure 3.4 – Zoom on the interface of a steady domain adaptation.

3.4.3 Mass conservation

This section is dedicated to the study of the mass conservation of the hybrid method. This property being satisfied in the continuous case, one expects conservation in the discrete setting. Let us recall that techniques to ensure this property with hybrid method have been developed in the past [39, 56, 57, 165]. In our setting, each of the standard schemes is conservative on its own by construction. However, the question arises when considering the hybrid scheme.

In order to understand the variation of mass, we consider a toy model that isn't relevant in practice but will highlight the key elements to constrain the mass variation. Let us set the state of cells for every time step into two domains:

$$\Omega_{\mathcal{K}} = \bigcup_{i=1}^{s-1} \mathcal{X}_i \quad \text{and} \quad \Omega_{\mathcal{F}} = \bigcup_{i=s}^{Nx} \mathcal{X}_i. \tag{3.39}$$

Note that in the next result, we neglect what happens at the boundary. Our primary focus is to understand what happens at the interface $x_{s-\frac{1}{2}}$ between the two domains. Moreover, in that context and with periodic boundary conditions in position, the same analysis can be done at the interface $x_{\frac{1}{2}}$. Figure 3.4 illustrates this framework. We define the mass of the hybrid system as:

$$m^n = \sum_{i\in\mathcal{I}} \tilde{\rho}_i^n \Delta x.$$

Note that this definition suggests that the quantity $\langle g^n_{i+\frac{1}{2}}\rangle_\Delta$ is zero. This property holds in the continuous setting and will be numerically investigated in the next section. The following lemma quantifies the mass variation between two time steps.

Lemma 16. *Let* $\left(\tilde{\rho}_i^n\right)_i$ *and* $\left(g^{\varepsilon,n}_{i+\frac{1}{2},j}\right)_{ij}$ *be computed using the hybrid scheme* (3.26)-(3.37)*. Let the mass variation between* t^n *and* t^{n+1} *be defined as:*

$$\Delta m^{n+\frac{1}{2}} = \sum_{i\in\mathcal{I}} \Delta x \frac{(\tilde{\rho}_i^{n+1} - \tilde{\rho}_i^n)}{\Delta t}. \tag{3.40}$$

In the context of the steady domain adaptation (3.39) *and neglecting the boundaries, one has:*

$$\begin{aligned}\Delta m^{n+\frac{1}{2}} = &-\left\langle \xi g^{\varepsilon,n}_{s-\frac{1}{2}} \right\rangle_\Delta \frac{e^{-\Delta t/\varepsilon^2}}{\varepsilon} + \frac{1-e^{-\Delta t/\varepsilon^2}}{\Delta x \Delta v^3} \left\langle \xi T^{\varepsilon,n}_{s-\frac{1}{2}} \right\rangle_\Delta \\ &- m_2^{\Delta v} e^{-\Delta t/\varepsilon^2} J^{\varepsilon,n}_{s-\frac{1}{2}} + m_2^{\Delta v}\left(J^{\varepsilon,n}_{s-\frac{1}{2}} - J^{n}_{s-\frac{1}{2}} \right).\end{aligned} \tag{3.41}$$

An important consequence of this lemma is that thanks to (3.32) and Assumption 1, the mass variation converges to 0 as ε tends to 0.

Proof. Let us consider the hybrid scheme (3.37)-(3.38) on the density. Using the fixed domain adaptation (3.39) and neglecting the boundary, the mass variation writes:

$$\begin{aligned}\Delta m^{n+\frac{1}{2}} &= \sum_{i\in\mathcal{I}} \frac{\Delta x}{\Delta t}(\tilde\rho_i^{n+1} - \tilde\rho_i^n) \\ &= \sum_{i\in\mathcal{I}} J_i^{H,n} \\ &= -\frac{1}{\varepsilon}\sum_{i=1}^{s-1}\left(\langle \xi g^{\varepsilon,n+1}_{i+\frac{1}{2}} \rangle_\Delta - \langle \xi g^{\varepsilon,n+1}_{i-\frac{1}{2}} \rangle_\Delta \right) + m_2^{\Delta v} \sum_{i=s}^{N_x} \left(J^n_{i+\frac{1}{2}} - J^n_{i-\frac{1}{2}} \right) \\ &= -\frac{1}{\varepsilon}\left\langle \xi g^{\varepsilon,n+1}_{s-\frac{1}{2}} \right\rangle_\Delta - m_2^{\Delta v} J^n_{s-\frac{1}{2}}.\end{aligned} \tag{3.42}$$

Similarly, as in the proof of Proposition 6, $g^{\varepsilon,n+1}_{s-\frac{1}{2},j}$ is replaced by its expression (3.26). The quantity $\frac{1}{\varepsilon}\left\langle \xi g^{\varepsilon,n+1}_{s-\frac{1}{2}} \right\rangle_\Delta$ then reads:

$$\begin{aligned}\frac{1}{\varepsilon}\left\langle \xi g^{\varepsilon,n+1}_{s-\frac{1}{2}} \right\rangle_\Delta = &\left\langle \xi g^{\varepsilon,n}_{s-\frac{1}{2}} \right\rangle_\Delta \frac{e^{-\Delta t/\varepsilon^2}}{\varepsilon} - \left\langle \xi T^{\varepsilon,n}_{s-\frac{1}{2}} \right\rangle_\Delta \left(1-e^{-\Delta t/\varepsilon^2}\right)\frac{1}{\Delta x \Delta v^3} \\ &- \left\langle \xi^2 \mathcal{M}_j \right\rangle_\Delta J^{\varepsilon,n}_{s-\frac{1}{2}} \left(1-e^{-\Delta t/\varepsilon^2}\right).\end{aligned} \tag{3.43}$$

Finally, plugging (3.43) into (3.42) and using the computation (3.36) for the term $\langle \xi^2 \mathcal{M}_j \rangle_\Delta$, one obtains:

$$\begin{aligned}\Delta m^{n+\frac{1}{2}} = &-\left\langle \xi g^{\varepsilon,n}_{s-\frac{1}{2}} \right\rangle_\Delta \frac{e^{-\Delta t/\varepsilon^2}}{\varepsilon} + \frac{1-e^{-\Delta t/\varepsilon^2}}{\Delta x \Delta v^3} \left\langle \xi T^{\varepsilon,n}_{s-\frac{1}{2}} \right\rangle_\Delta \\ &- m_2^{\Delta v} e^{-\Delta t/\varepsilon^2} J^{\varepsilon,n}_{s-\frac{1}{2}} + m_2^{\Delta v}\left(J^{\varepsilon,n}_{s-\frac{1}{2}} - J^{n}_{s-\frac{1}{2}} \right).\end{aligned}$$

□

Remark 7. *The proof only holds in the context of the toy problem* (3.39)*. However, it can be extended to a more general setting seeing that the mass variation occurs at all interfaces between kinetic and fluid subdomains. Namely,*

$$\begin{aligned}\Delta t \Delta m^{n+\frac{1}{2}} = \sum_{\alpha\in\mathcal{S}} \beta\Big(&-\langle \xi g^{\varepsilon,n}_\alpha \rangle_\Delta \frac{e^{-\Delta t/\varepsilon^2}}{\varepsilon} + \frac{1-e^{-\Delta t/\varepsilon^2}}{\Delta x \Delta v^3} \langle \xi T^{\varepsilon,n}_\alpha \rangle_\Delta \\ &- m_2^{\Delta v} e^{-\Delta t/\varepsilon^2} J^{\varepsilon,n}_\alpha + m_2^{\Delta v}\left(J^{\varepsilon,n}_\alpha - J^n_\alpha \right)\Big),\end{aligned}$$

where $\mathcal{S}$ is the set of interfaces between kinetic and fluid subdomains and $\beta = \pm 1$ depends on the orientation of the subdomains.

3.5 Numerical simulations

Let us now present some numerical simulations. The properties of the fully-kinetic scheme are presented in Section 3.5.1. Section 3.5.2 is dedicated to the hybrid method. In the following, unless specified otherwise, the phase-space is discretized as follows:

$$N_v = 16, \quad N_x = 50, \quad v_\star = 10, \quad x_\star = \pi, \quad \Delta t = 10^{-4}.$$

The same time step is used for all schemes. Note that since the limit scheme is explicit, its stability is therefore guaranteed under a parabolic condition: $\Delta t \leq C\Delta x^2$. The hybrid scheme is therefore also constrained by a similar CFL condition. This could be overcome by plugging the micro scheme into the macro one and impliciting the appearing diffusive term. Such a procedure is detailed in [59] and its implementation in the hybrid setting is outside the scope of the chapter. Let us assume that the electrical field is the gradient of a potential V: $E = -\partial_x V$. To satisfy the periodicity on the domain, we choose $V(x) = -\frac{\sin(2x)}{4} \times 10^{-2}$ so $E(x) = \frac{1}{2}\cos(2x) \times 10^{-2}$. We also set

$$f_0^1 = \frac{1}{(2\pi)^{3/2}} e^{-|v|^2/2}(1 + \cos(2x)), \tag{3.44}$$

an initial data at local equilibrium in velocity and

$$f_0^2 = \frac{1}{(2\pi)^{3/2}} |v|^4 e^{-|v|^2/2}(1 + \cos(2x)), \tag{3.45}$$

an initial data far from the local equilibrium in velocity. Finally, we consider four configurations:

— Case 1: $E = 0$, with initial data (3.44);
— Case 2: $E \neq 0$, with initial data (3.44);
— Case 3: $E = 0$, with initial data (3.45);
— Case 4: $E \neq 0$, with initial data (3.45).

The properties of the fully kinetic implementation are shown in the $1D_x$-$1D_v$ setting. The performance and properties of the hybrid method are presented in the full $1D_x$-$3D_v$ one.

3.5.1 The full kinetic scheme

In this Section, we present the key properties of the micro macro scheme (3.26)-(3.27).

Convergence of the numerical scheme. We study the convergence in the phase space for $\varepsilon = 1.0$ and a final time T so that the solution is still far from equilibrium. For each discretization parameter, the error is computed against a reference solution obtained on a fine mesh and the other parameters are fixed. Let us set $\Delta t = 10^{-4}$ and $T = 0.01$. In position, $N_x = 1024$ points are used as reference and we set $N_v = 16$. In velocity, $N_v = 128$ points in each velocity direction are used as reference and we set $N_x = 32$. As we can see on Figure 3.5, the numerical scheme converges with

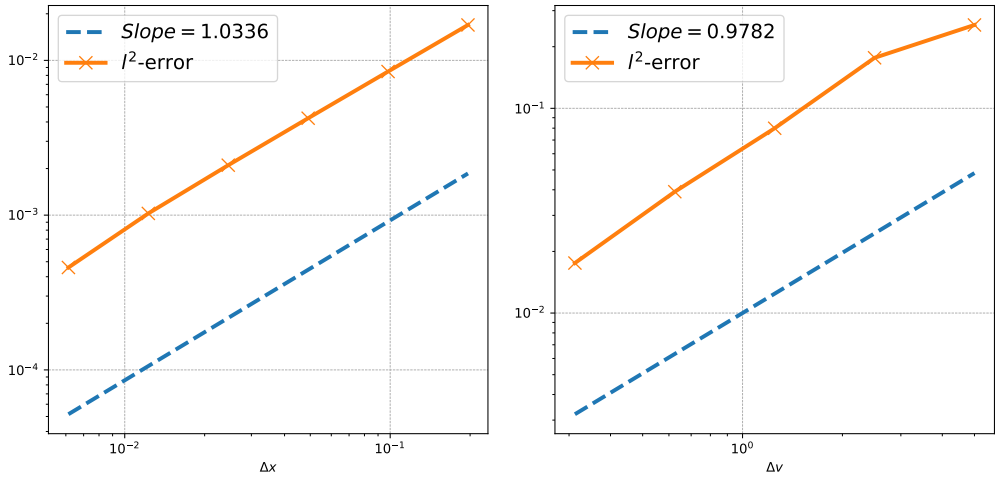


Figure 3.5 – Case 1. Convergence of the fully kinetic scheme (3.26)-(3.27) with respect to Δx (Left) and Δv (Right).

respect to the discretization parameters. In particular, an expected order 1 is obtained in position and velocity due to the first order upwind scheme for the transport.

Convergence towards the drift-diffusion equation. Let us now numerically investigate the AP property of the (Micro)-(Macro) scheme. We consider this analysis for the cases 1 and 2. The results can be found in Figure 3.6 for Case 1 and in Figure 3.7 for Case 2. We can observe a convergence of the kinetic scheme to the limit one as $\varepsilon \to 0$. In particular, the curves for $\varepsilon = 0.05$ and 10^{-4} overlap and are close to the limit case. This validates the asymptotic consistency of the (Micro)-(Macro) scheme. The stability is numerically verified as the same Δt is used for every ε.

Long time behavior. The long time behavior of solutions to (P^ε) has been extensively studied in the past decades. In a more general setting, the electric field E is the gradient of a potential $V \in \mathcal{C}^2(\Omega_x)$, $E = -\nabla V$, and (P^ε) admits a global equilibrium given by

$$F(x,v) = \frac{M_0}{\mu_0} e^{-(V(x)+\frac{|v|^2}{2})}, \quad (x,v) \in \Omega_x \times \mathbb{R}^{d_v}$$

where $\mu_0 = (2\pi)^{-d/2} \int_{\Omega_x \times \mathbb{R}^{d_v}} e^{-(V(x)+\frac{|v|^2}{2})} \mathrm{d}x\,\mathrm{d}v$ and $M_0 = \int_{\Omega_x \times \mathbb{R}^{d_v}} f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v$ is the mass of the initial condition. In particular, F can be written under a separate variable form:

$$F(x,v) = M_0 \phi(x) \mathcal{M}(v), \quad \text{where } \phi = \frac{e^{-V(x)}}{\int_{\Omega_x} e^{-V(x)}\,\mathrm{d}x}.$$

The functions $\mathcal{M}$ and ϕ are called local equilibria in velocity and position respectively. When one considers models such as (P^ε), there are various ways to show that there exists $\kappa(\varepsilon) > 0$ and $C(\varepsilon) > 0$, such that if f^ε is solution to (P^ε),

$$\|f^\varepsilon(t) - F\|_{\mathcal{V}} \leq C(\varepsilon)\|f_0 - F\|_{\mathcal{V}} e^{-\kappa(\varepsilon)t},$$

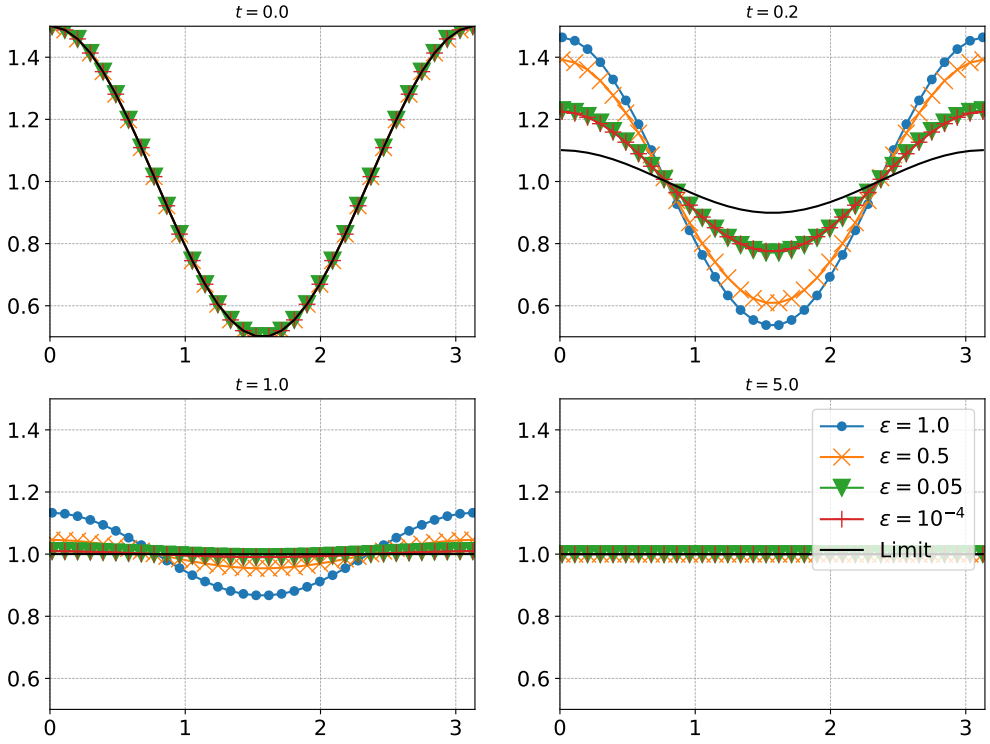


Figure 3.6 – Case 1. Comparison of the solution of the limit scheme (S_{Lim}) with the solution obtained with the (Micro)-(Macro) scheme with different ε, $t = 0.0$, 0.2, 1.0 and 5.0.

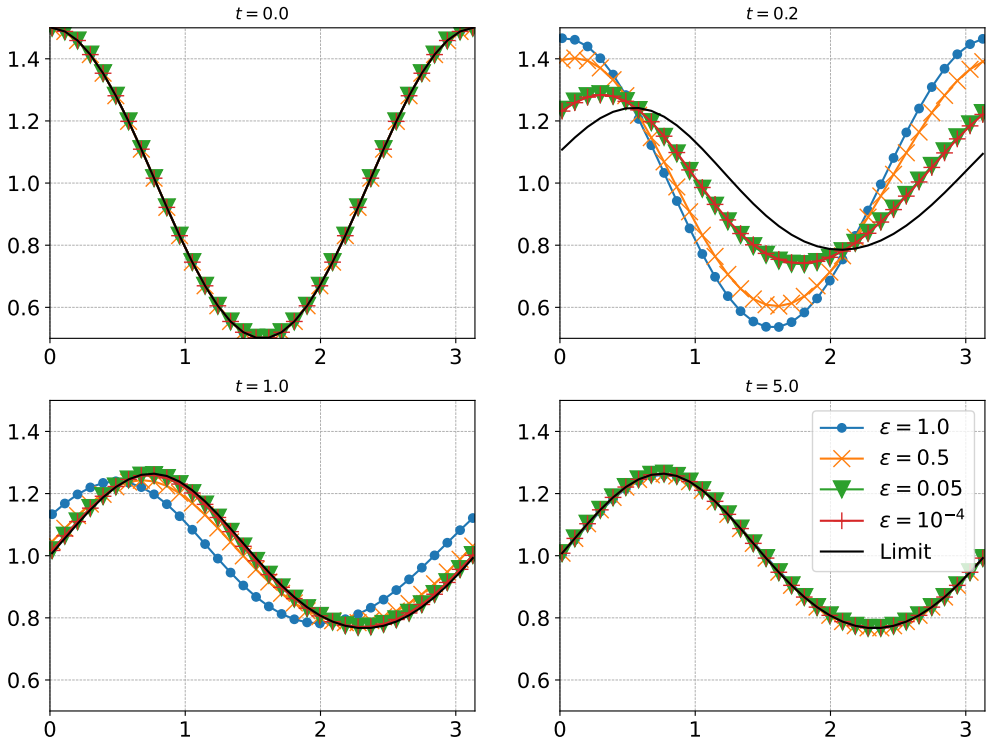


Figure 3.7 – Case 2. Comparison of the solution of the limit scheme (S_{Lim}) with the solution obtained with the (Micro)-(Macro) scheme with different ε, $t = 0.0$, 0.2, 1.0 and 5.0.

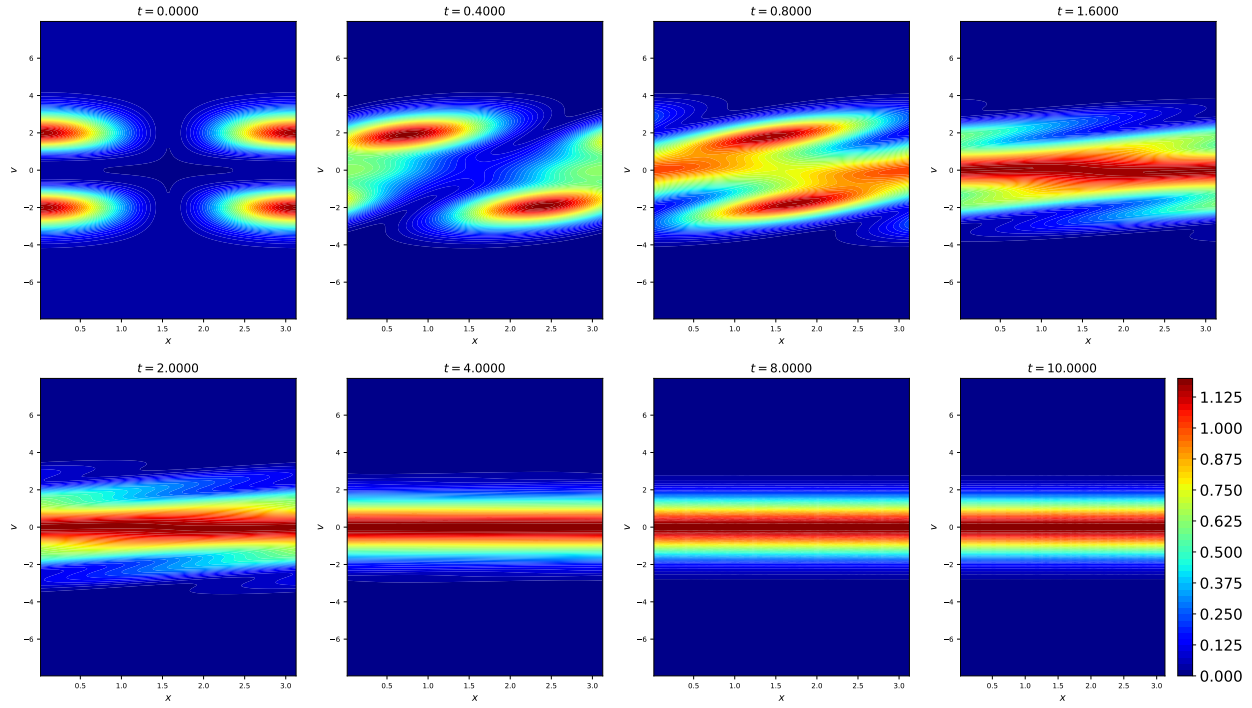


Figure 3.8 – Case 3. Snapshots of the distribution function f computed with the scheme (3.26)-(3.27), $\varepsilon = 1.0$, $N_x = 100$, $N_v = 128$.

where $\mathcal{V}$ is an appropriate functional space. A proof of this result was done in [121] in a setting without electric field. In recent years, the literature on the subject expanded a lot. Robust and systematic methods were developed to show the convergence to an equilibrium. Those are called hypocoercivity methods. We refer to the introduction on the subject and references in Section 4 and Chapter 1 for more details.

Following these ideas, we want to observe the convergence of the scheme (3.26)-(3.27) to equilibrium in a large timescale. Figure 3.8 shows the evolution of the marginal distribution $\bar{f}^\varepsilon = \int_{\mathbb{R}^2} f^\varepsilon \mathrm{d}v_y \mathrm{d}v_z$ as time increases (Case 3, $\varepsilon = 1.0$). In particular, the numerical solution indeed seems to converge to equilibrium. Let us introduce the following discrete norm for $f = \left(f_{ij}\right)_{ij}$:

$$\|f\|_\Delta^2 = \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}^3} f_{ij}^2 F_{ij}^{-1} \Delta x \Delta v^3,$$

where $\left(F_{ij}\right)_{ij}$ is a discretization of the global equilibrium F, $F_{ij} = F(x_i, v_j)$. For $(\rho_i)_{i\in\mathcal{I}}$, we also denote by $\|\rho\|_2 = \sum_{i\in\mathcal{I}} \rho_i^2 \Delta x$ the discrete L^2-norm in position. We now investigate the rate of convergence of the following discrete norms:

$$\|f - F\|_\Delta, \quad \|g\|_2, \quad \|\rho^\varepsilon - \langle F\rangle_\Delta\|_2 \quad \text{and} \quad \|\rho - \langle F\rangle_\Delta\|_2, \tag{3.46}$$

where ρ^ε is the density obtained with the kinetic scheme and ρ is obtained with the limit scheme. We consider Case 1. On Figure 3.9 we choose $\varepsilon = 0.5$ and 0.1, and show the norms (3.46) as functions of time in semilog scale. The exponential convergence of the various norms is clear. Moreover, the rates $\kappa(\varepsilon)$ observed are $\kappa(0.5) = 2.43$ and $\kappa(0.1) = 3.72$. The rate $\kappa(\varepsilon)$ increases as the

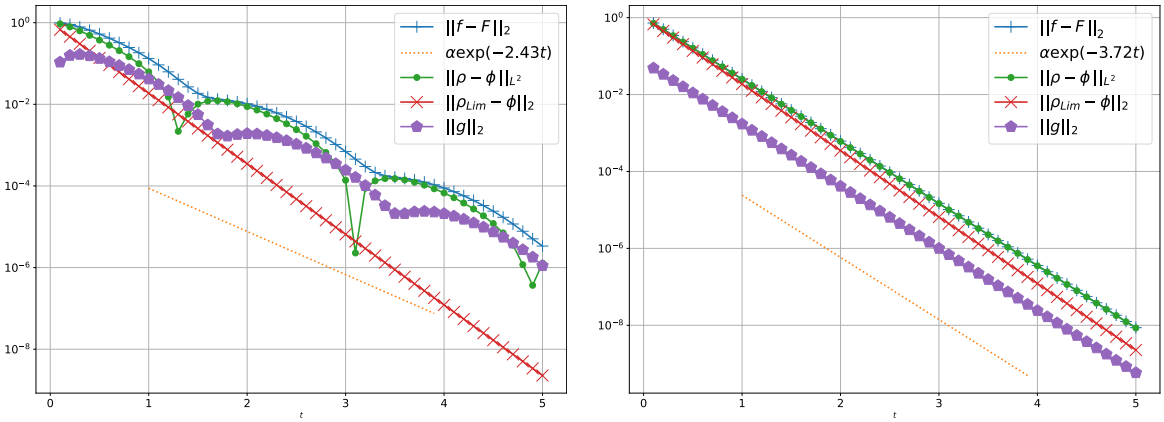


Figure 3.9 – Case 1. Time evolution of the norms (3.46) computed with the fully kinetic scheme and limit scheme, $\varepsilon = 0.5$ (Left), $\varepsilon = 0.1$ (Right).

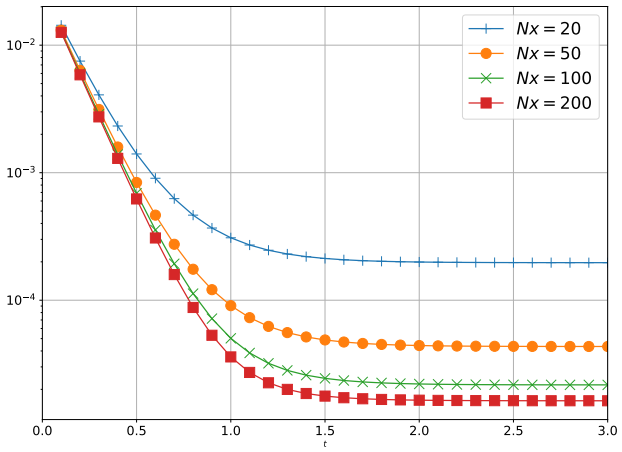


Figure 3.10 – Case 4. Time evolution of the norm $\|f^\varepsilon - F\|_\Delta$ computed with the fully kinetic scheme for $N_x = 20$, 50, 100 and 200, $\varepsilon = 0.1$.

Knudsen number gets smaller. In particular, we observe the same rate of convergence between the fully kinetic scheme and the limit one for small values of ε.

Let us point out that in the case of a non-zero electric field, we do not recover the same convergence to equilibrium. Indeed, our numerical scheme is not well-balanced, i.e. designed to preserve steady states. As a consequence, the numerical solution only converges to an equilibrium that is an approximation of the steady state. Figure 3.10 shows the convergence to the equilibrium as the number of cells in position increases for Case 4.

3.5.2 Properties of the hybrid scheme

Choice of the coupling parameters. Before investigating the properties of the hybrid scheme, a natural question is the choice of the coupling parameters η_0 and δ_0. Indeed, as we have seen earlier, that choice has an impact on the conservation of mass. The smaller the parameters, the later the coupling occurs, and the more one can control this variation. However, the bigger the parameters are, the faster is the resulting hybrid scheme as one allows more fluid cells to appear. Therefore, one must find a good balance between accuracy and computation time. To illustrate how the macroscopic indicator behaves, we compute it without updating the state of the cells.

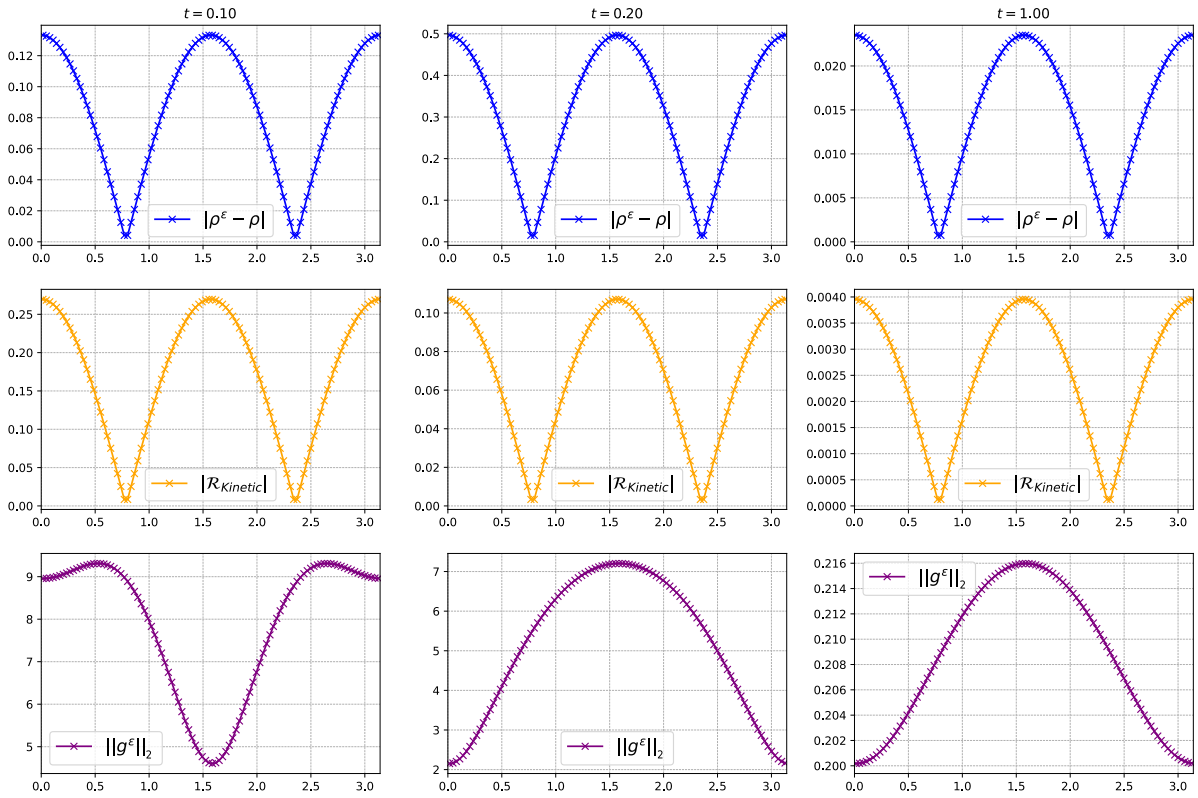


Figure 3.11 – Case 3. Snapshots of the difference between the densities computed with the kinetic and the limit schemes (Top), macroscopic indicator (Middle), $L^2(\mathrm{d}\gamma)$ norm of the perturbation g^ε (Bottom), $\varepsilon = 0.5$.

Figure 3.11 shows the indicator compared to the difference between the kinetic and fluid densities. One can observe that this indicator behaves as expected. When the kinetic and limit densities are close, the indicator is also small. Regarding the norm of g^ε, its behavior is also expected. Indeed, we chose an initial data far from the local equilibrium in velocity and therefore, the norm can be high even if the densities are close (See first column, third row in Figure 3.11). Lastly, both the macroscopic indicator and the norm of g^ε tend to 0 as time increases. As a consequence, the closer to the equilibrium the solution is, the more fluid cells will appear.

Qualitative comparison. Let us now compare the kinetic and the hybrid schemes. Figures 3.12 and 3.13 show the densities computed by the kinetic, hybrid and limit schemes for Cases 3 and 4 with $\varepsilon = 0.1$. We can see a good agreement between the three schemes. The domain adaptation works for $E \neq 0$ which was not investigated in previous works on the method. As time increases, the solution relaxes to an equilibrium and the domain becomes fully fluid. One can also observe that when both type of cells co-exists the hybrid density slightly deviates from the full kinetic model.

Conservation of mass. Let us now numerically investigate the conservation of mass. Indeed, we have shown in Lemma 16 that the hybrid method is not exactly conservative. However, it becomes conservative asymptotically. In addition, the hybrid method was constructed so that the cells become fluid when the solution is close to a local equilibrium in velocity. As a consequence, the perturbation is small when the coupling occurs and so is the mass variation. In practice, we can

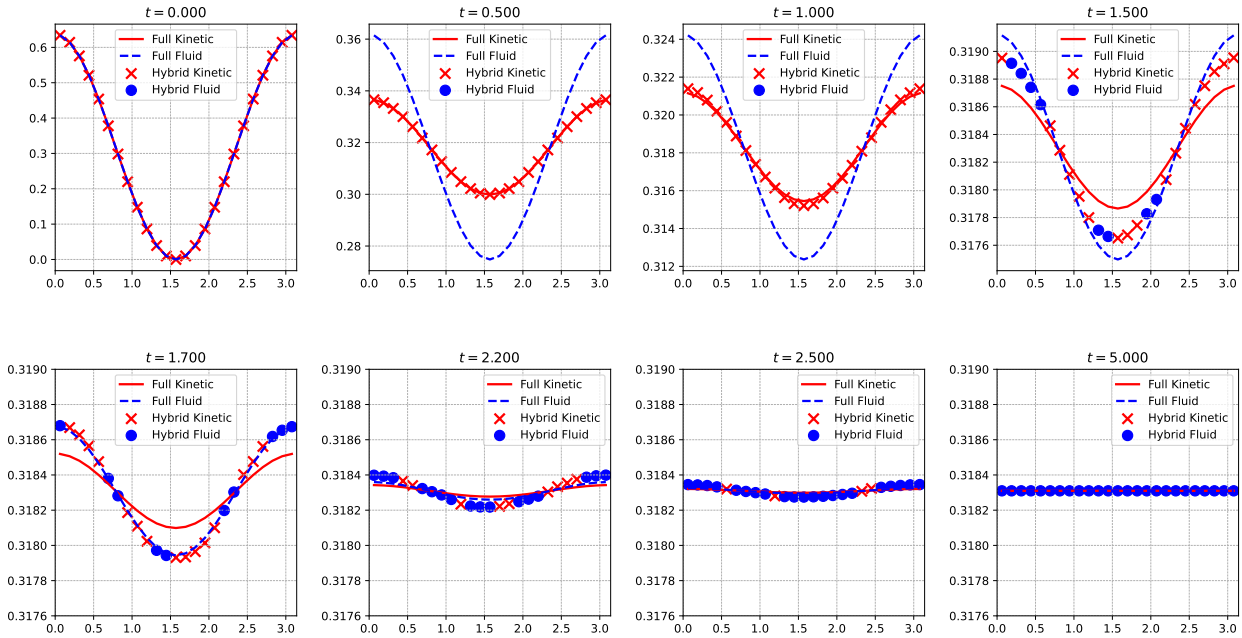


Figure 3.12 – Case 3. Snapshots of the densities computed using the full kinetic, hybrid and limit schemes, $\varepsilon = 0.1$, $\eta_0 = \delta_0 = 10^{-4}$.

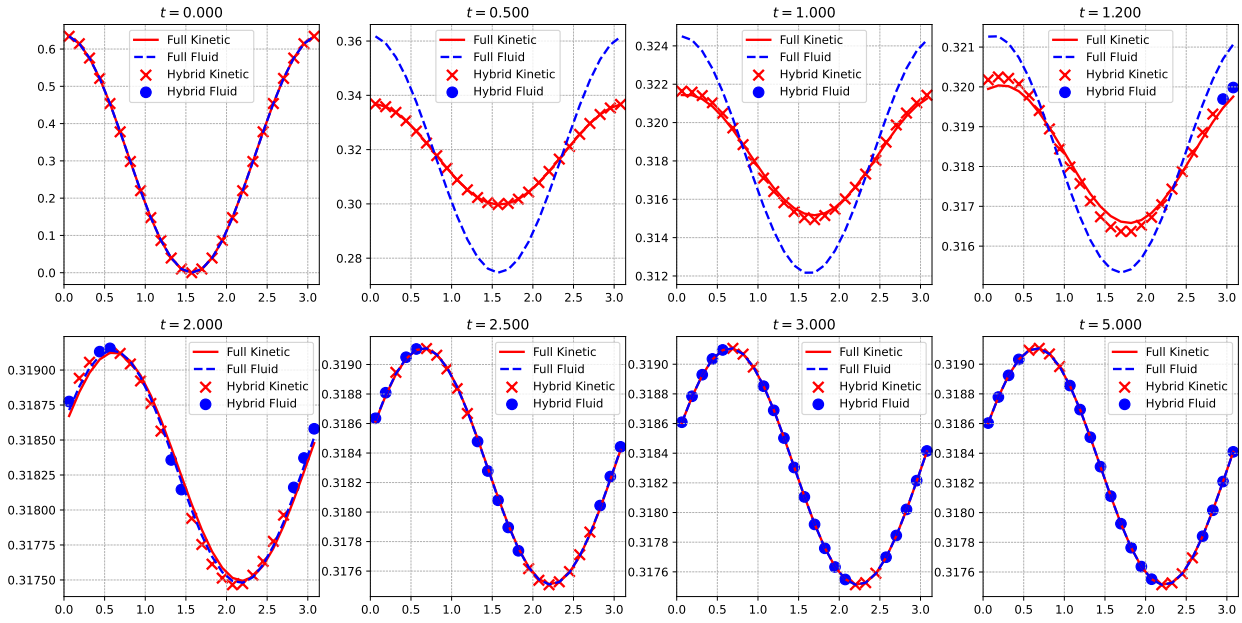


Figure 3.13 – Case 4. Snapshots of the densities computed using the full kinetic, hybrid and limit schemes, $\varepsilon = 0.1$, $\eta_0 = \delta_0 = 10^{-4}$.

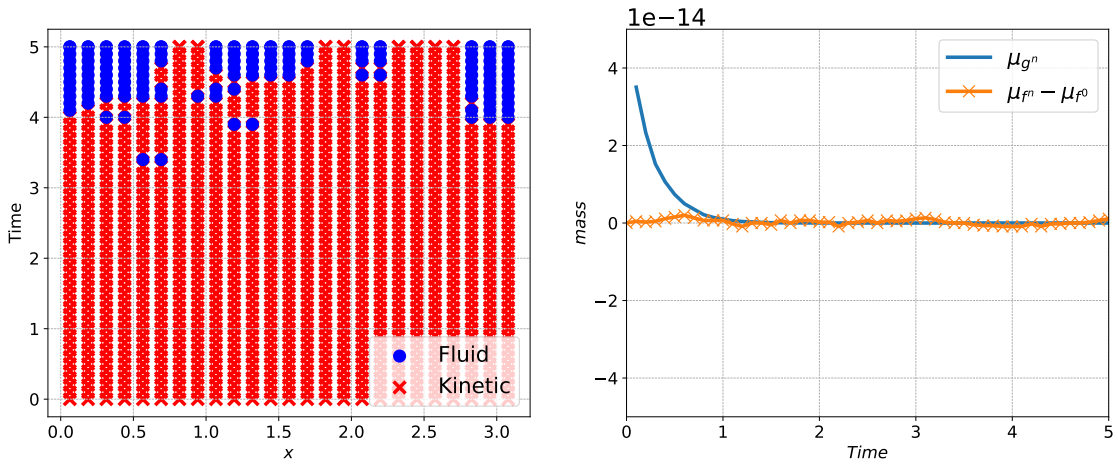


Figure 3.14 – Case 3. Time evolution of the state of the cells (Left), mass variation and mass of g^ε (Right), $\varepsilon = 0.5$, $\eta_0 = \delta_0 = 10^{-4}$.

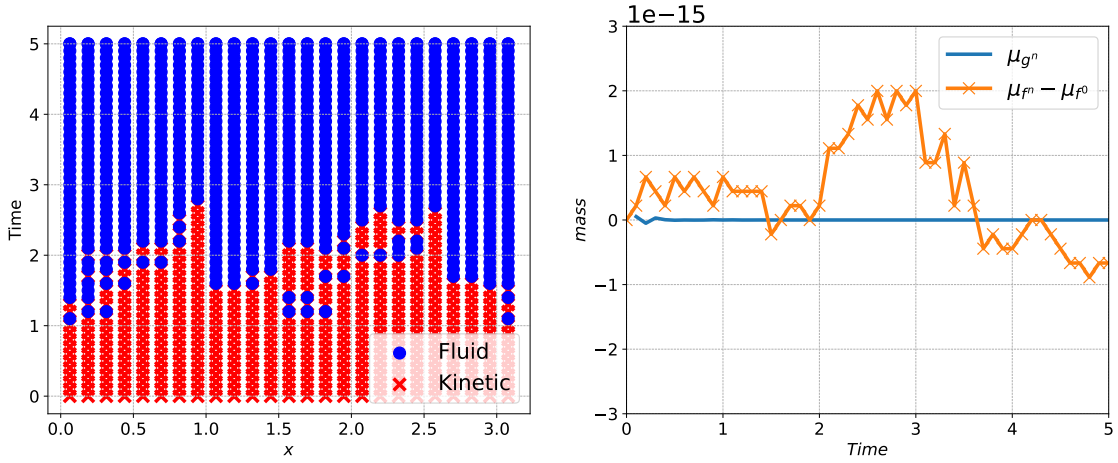


Figure 3.15 – Case 3. Time evolution of the state of the cells (Left), mass variation and mass of g^ε (Right), $\varepsilon = 0.1$, $\eta_0 = \delta_0 = 10^{-4}$.

observe a mass variation of the order of the machine accuracy. In addition, the zero-mass property of the perturbation

$$\mu_g = \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}^3} g_{i+\frac{1}{2},j}\Delta x\Delta v^3$$

is preserved. We illustrate the state of the cells, the corresponding mass variation as well as μ_g on Figures 3.14 and 3.15. Case 3 is considered for $\varepsilon = 0.5$ and 0.1.

Error analysis. We are now interested in the error induced by the hybrid method. In particular, we investigate the error between the full kinetic scheme and the hybrid method. The goal is to be more efficient than the full kinetic solver. However, the gain in computation time comes with a slight loss in accuracy. Let $f_{Kinetic}$ be the distribution computed using the full kinetic scheme and f_{Hybrid} be the distribution obtained from the hybrid scheme. On Figure 3.16 we compute the error between the two distributions in l^2-norm: $\|f_{Kinetic} - f_{Hybrid}\|_2$ at several times with $\varepsilon = 0.1$. The corresponding state of the cells can be found in Figure 3.15 for $\eta_0 = \delta_0 = 10^{-4}$. Quite expectedly, there indeed is a slight loss in accuracy as soon as the coupling occurs. However, it quickly

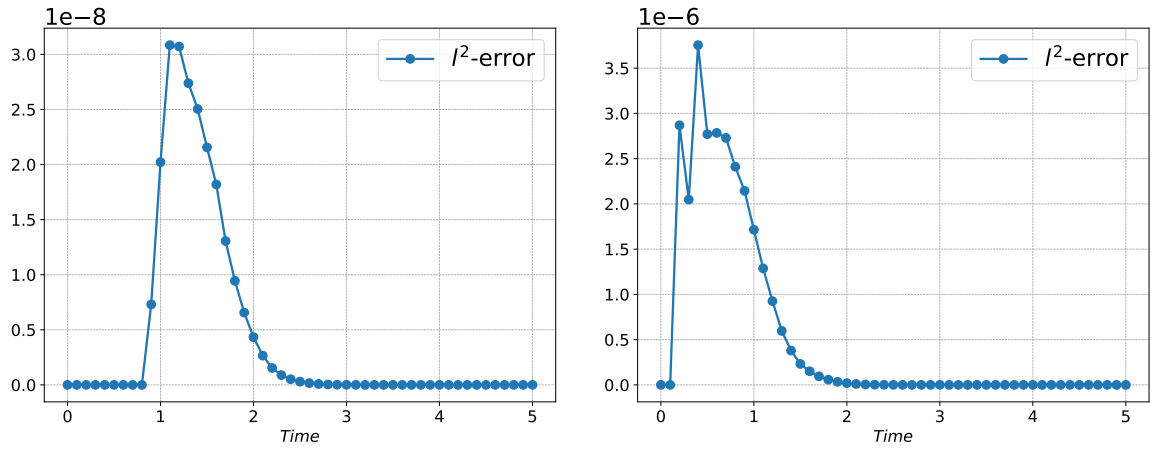


Figure 3.16 – Case 3. Time evolution of the l^2-error between the kinetic distribution $f_{Kinetic}$ and the hybrid one f_{Hybrid}, $\varepsilon = 0.1$, $\eta_0 = \delta_0 = 10^{-4}$ (Left), $\eta_0 = \delta_0 = 10^{-3}$ (Right).

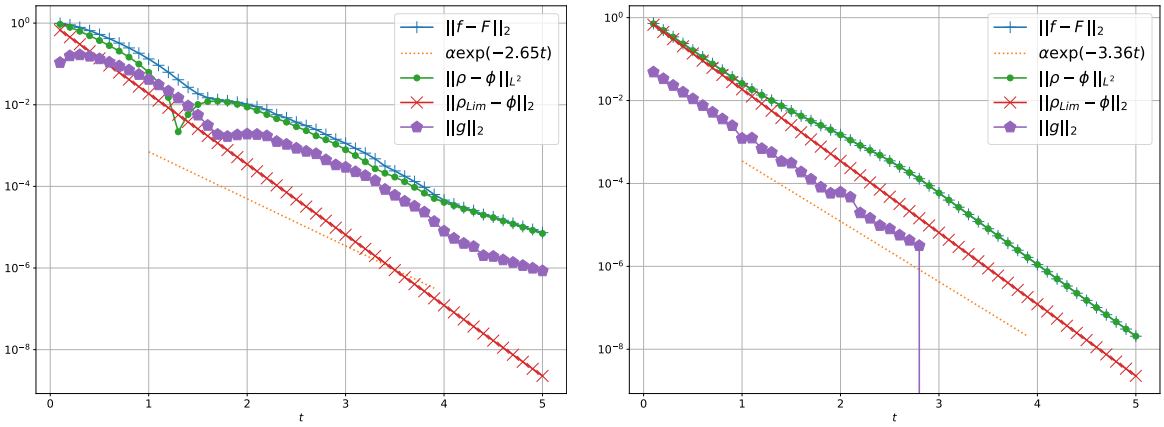


Figure 3.17 – Case 1. Time evolution of the norms (3.46) computed with the hybrid and limit schemes, $\varepsilon = 0.5$ (Left), $\varepsilon = 0.1$ (Right), $\eta_0 = \delta_0 = 10^{-4}$.

diminishes as the coupled solution relaxes to equilibrium. Moreover, one can lessen this error by tuning down the coupling parameters but at the expense of the computational gain.

Long time behavior. Similarly, as for the full kinetic scheme, we are interested in the long time behavior of the hybrid scheme. We shall focus on the case $E = 0$. Figure 3.17 shows the convergence of the norms (3.46) in the hybrid setting. As the perturbation is set to 0 in fluid subdomains, the norms of g is directly projected to 0 when all cells are fluid. In addition, one can observe the convergence of the density towards the global mass. The rates of convergence are not exactly recovered compared to the full kinetic scheme but remains close: $\kappa_{Hybrid}(0.5) = 2.65$ and $\kappa_{Hybrid}(0.1) = 3.36$.

Computation time. Let us now consider the efficiency of the hybrid method. We set the final time $T = 5.0$ to compare the computation time. Tables 3.2-3.3-3.4-3.5 show the computation time of the full kinetic, hybrid and limit scheme for different test cases with two sets of coupling parameters: $\eta_0 = \delta_0 = 10^{-4}$ and $\eta_0 = \delta_0 = 10^{-3}$. We recall that the same time step, $\Delta t = 10^{-4}$, is used for the three schemes.

	$\eta_0 = \delta_0 = 10^{-4}$				$\eta_0 = \delta_0 = 10^{-3}$			
Case	1	2	3	4	1	2	3	4
MM	298.2	285.8	287.7	281.0	298.2	285.8	287.7	281.0
Hybrid	291.5	282.2	283.5	303.1	278.0	281.2	276.0	280.3
Limit	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 3.2 – Test 1. Comparison of the computation time (sec), $\varepsilon = 1.0$, $T = 5.0$, $N_x = 50$, $N_v = 16$.

	$\eta_0 = \delta_0 = 10^{-4}$				$\eta_0 = \delta_0 = 10^{-3}$			
Case	1	2	3	4	1	2	3	4
MM	290.1	295.7	282.0	295.8	290.1	295.7	282.0	295.8
Hybrid	181.4	210.7	152.3	210.1	116.8	131.0	110.4	131.0
Limit	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 3.3 – Test 2. Comparison of the computation time (sec), $\varepsilon = 0.1$, $T = 5.0$, $N_x = 50$, $N_v = 16$.

	$\eta_0 = \delta_0 = 10^{-4}$				$\eta_0 = \delta_0 = 10^{-3}$			
Case	1	2	3	4	1	2	3	4
MM	289.0	281.7	299.4	293.6	289.0	281.7	299.4	293.6
Hybrid	110.6	116.6	111.9	130.1	0.09	0.09	79.4	73.0
Limit	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 3.4 – Test 2. Comparison of the computation time (sec), $\varepsilon = 0.05$, $T = 5.0$, $N_x = 50$, $N_v = 16$.

	$\eta_0 = \delta_0 = 10^{-4}$				$\eta_0 = \delta_0 = 10^{-3}$			
Case	1	2	3	4	1	2	3	4
MM	278.6	296.5	286.7	290.1	278.6	296.5	286.7	290.1
Hybrid	0.11	0.10	0.10	0.10	0.09	0.10	0.11	0.10
Limit	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 3.5 – Test 4. Comparison of the computation time (sec), $\varepsilon = 10^{-4}$, $T = 5.0$, $N_x = 50$, $N_v = 16$.

	$\eta_0 = \delta_0 = 10^{-4}$				$\eta_0 = \delta_0 = 10^{-3}$			
ε \ Case	1	2	3	4	1	2	3	4
1.0	1.02	1.01	1.01	0.92	1.07	1.01	1.04	1.00
0.1	1.59	1.40	1.94	1.41	2.48	2.26	2.68	2.26
0.05	2.6	2.42	2.68	2.26	3211	3130	3.77	4.02
10^{-4}	2533	2965	2867	2901	3096	2965	2606	2901

Table 3.6 – Speedup of the hybrid method compared to the full kinetic scheme, $N_x = 50$, $N_v = 16$.

We can make several observations. First, the fluid solver is as expected, much faster than the full kinetic one. Moreover, it is also always faster than the hybrid method. This can easily be explained by the additional cost of computing the indicators and the added cost of dealing with interfaces between kinetic and fluid. Table 3.6 shows the computational gain for the previous tests. In particular, the hybrid method offers no significant gain in very low collision regimes. Because of the slow convergence rate towards the equilibrium, the coupling occurs very late or not at all depending on the chosen final time. However, we can point out that the coupling algorithm doesn't add much cost relatively to the full kinetic scheme. This shows that the implementation is quasi-optimal. In addition, since the cost of the fluid model is negligible compared to the kinetic one, if fluid cells are used half the time, the computation time is essentially also reduced by half.

Another observation is that the speedup is test case dependent as it can be seen in Table 3.6, $\varepsilon = 0.05$ and $\delta_0 = \eta_0 = 10^{-3}$. When the parameter is small the gain becomes extremely significant and the hybrid method becomes competitive with the fluid solver. A final observation is that the choice of larger coupling parameters indeed speeds up the method. If we consider Test 3, with $\varepsilon = 0.1$ and $\delta_0 = \eta_0 = 10^{-4}$ the speedup is 1.94 while a choice of $\delta_0 = \eta_0 = 10^{-3}$ offers a speedup of 2.68. Figure 3.16 compares the error made as time increases for these two sets of parameters. One can observe that the error reaches a maximum at two different orders of magnitude: 3.10^{-8} with $\delta_0 = \eta_0 = 10^{-4}$ and 3.10^{-6} with $\delta_0 = \eta_0 = 10^{-3}$. This last point raises the question of optimal parameters relatively to the error between hybrid and full kinetic. This shall be addressed in the future but seems, again, problem-dependent.

Non-homogeneous Knudsen number. In this last numerical experiment, we place ourselves in the $1D_x$-$1D_v$ setting and consider a non-homogeneous Knudsen number in the physical domain. Let us define the function

$$e(x) = \frac{1}{2}(\arctan(5 + 10(x - \frac{\pi}{2})) + \arctan(5 - 10(x - \frac{\pi}{2}))).$$

In particular, we choose $\varepsilon = \varepsilon(x)$ as

$$\varepsilon(x) = \frac{e(x)}{\max(e(x))}.$$

Such a function admits a maximum of 1 in the center of the domain and decays to 0 near the boundaries. Physically, it corresponds to few collisions in the center of the domain and to a fluid behavior elsewhere. In the following simulations, $\Delta t = 5.10^{-5}$ and the coupling parameters are $\delta_0 = \eta_0 = 10^{-4}$. Note that depending on the choice of $\varepsilon(x)$ one may need to decrease the time step to ensure stability. From an implementation point of view, the constant ε is simply replaced by

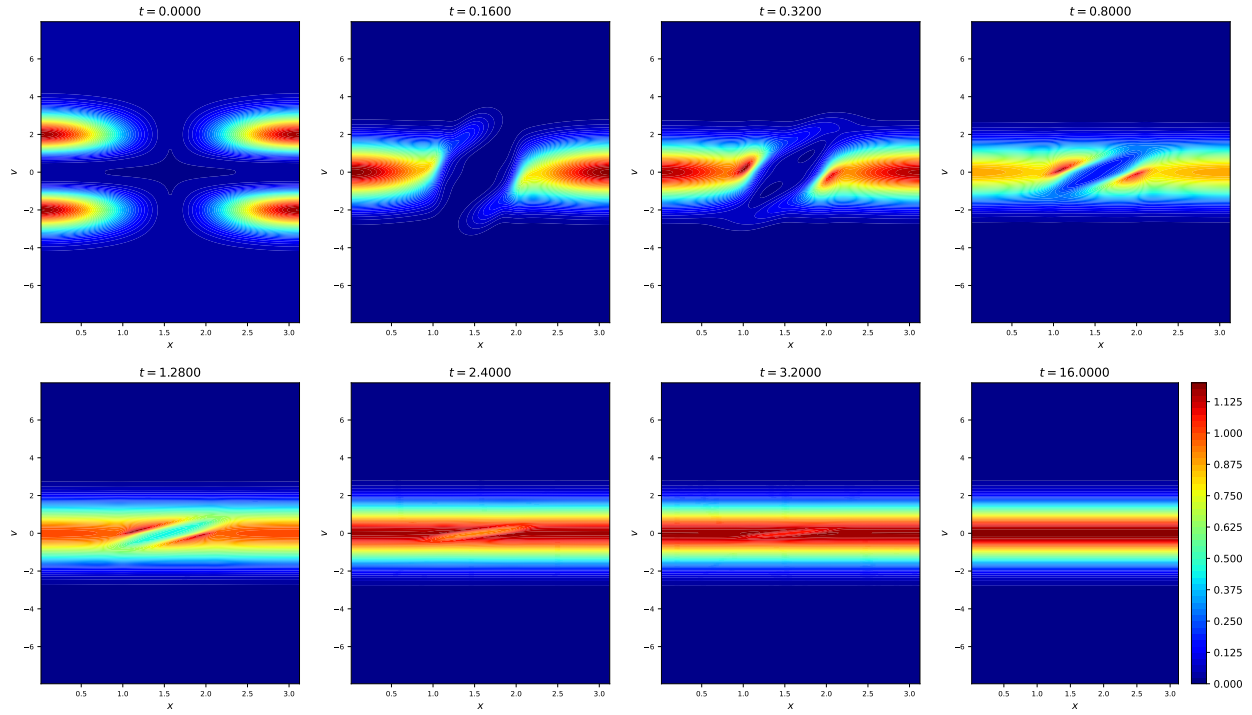


Figure 3.18 – Case 3. Snapshots of distribution f obtained with the hybrid scheme for a non-homogeneous Knudsen number.

$\varepsilon_{i+\frac{1}{2}} = \varepsilon(x_{i+\frac{1}{2}})$ without any change to the indicators.

Figure 3.18 shows that the hybrid scheme captures well the behavior of the distribution. Indeed, we observe a fast relaxation where $\varepsilon(x)$ is small and a much slower one in the center of the domain where $\varepsilon(x)$ is around 1. Regarding the state of the cells, one can see on Figures 3.19 and 3.20 that the fluid solver is quickly used where $\varepsilon(x)$ is small. Moreover, the last cells to become fluid are the ones where the gradient of $\varepsilon(x)$ is large. It is explained by the nature of the macroscopic indicator which uses derivatives up to order 4. One can also observe that the hybrid density starts deviate from the full kinetic one as more fluid cells appear. However, this deviation occurs at a small scale: between 10^{-8} and 10^{-10}. In this setting, the variation of mass was of order 10^{-11}. Finally, we looked at the convergence in time to the global equilibrium. On Figure 3.21, one can again observe the exponential convergence to equilibrium and the rate is slightly higher than the one obtained with an homogeneous value of $\varepsilon = 1$. Up to stability considerations, this experiment shows the robustness of the hybrid method. Performance-wise, considering Case 3 with $N_x = 200$ and $N_v = 256$, the full kinetic scheme takes 183.7 seconds to run while the hybrid one takes 147.6 seconds offering a speedup of 1.24.

3.6 Extension to the Vlasov-Poisson-BGK system

3.6.1 The model

In this section we extend the hybrid method to the kinetic description of a system of particles interacting via both mean-field electromagnetic interaction and collisions. Such system can be modelled using the Vlasov-Poisson-BGK equation. The unknown is the probability distribution

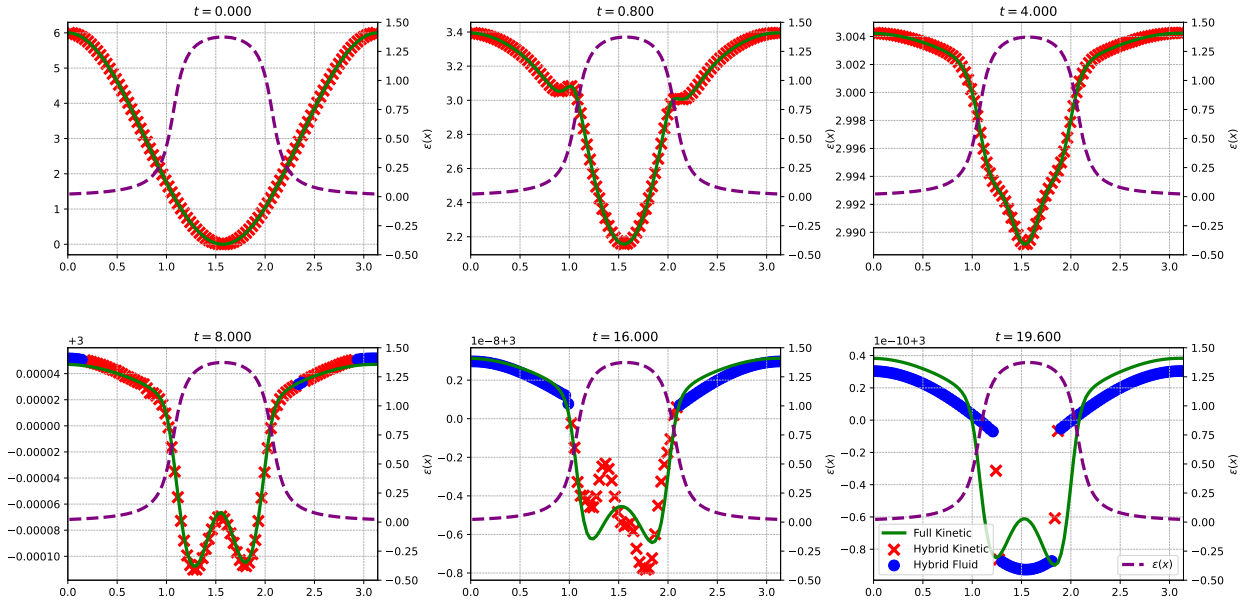


Figure 3.19 – Case 3. Snapshots of the densities computed using the full kinetic and hybrid schemes for a non-homogeneous Knudsen number.

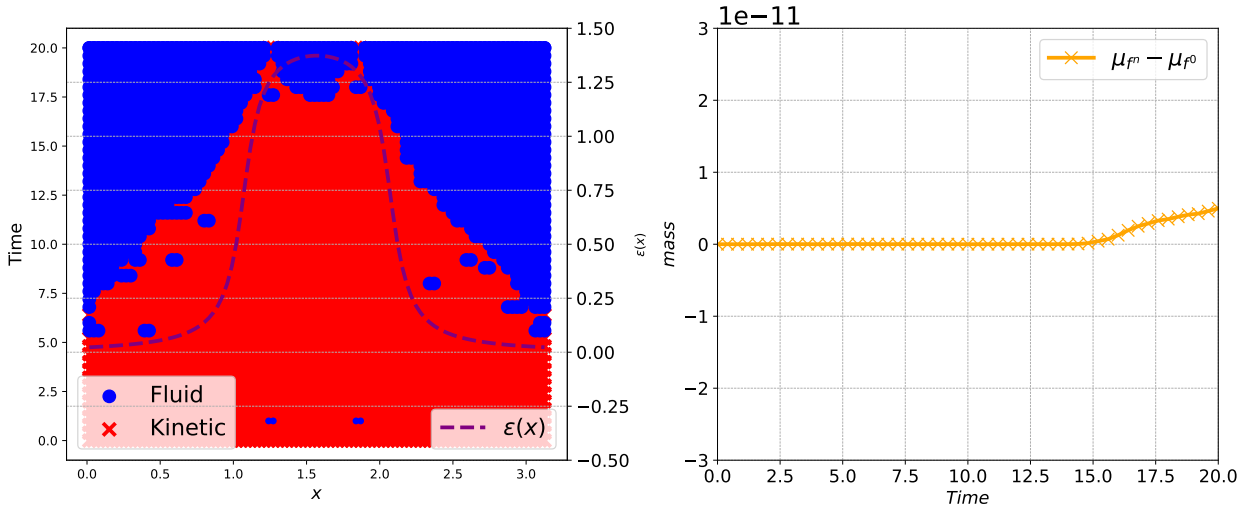


Figure 3.20 – Case 3. Time evolution of the state of the cells (Top) and mass variation (Bottom) for a non-homogeneous Knudsen number.

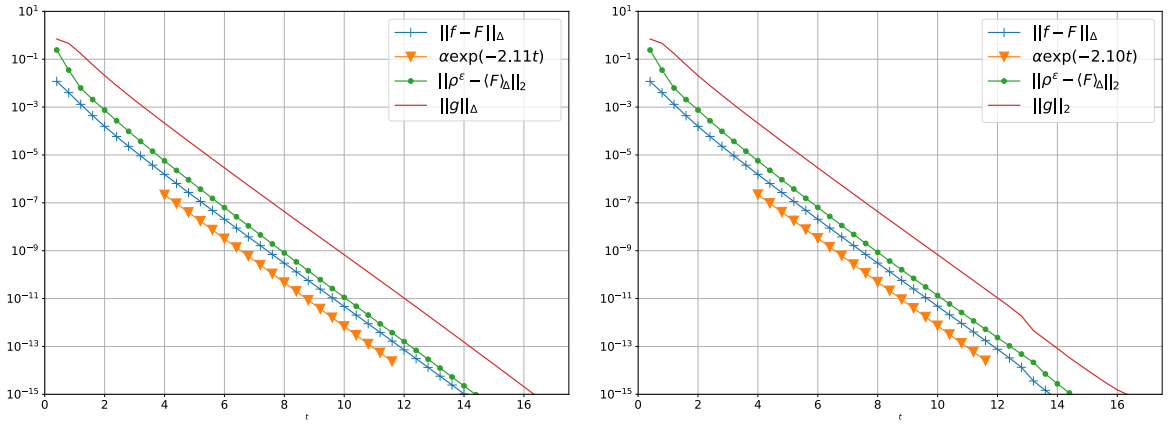


Figure 3.21 – Case 3. Convergence to a global equilibrium, full kinetic (Left), hybrid (Right) for a non-homogeneous Knudsen number.

function $f = f(t,x,v) \in \mathbb{R}^+$ solution to:

$$\begin{cases} \partial_t f^\varepsilon + \dfrac{v_x}{\varepsilon}\,\partial_x f^\varepsilon + \dfrac{E^\varepsilon}{\varepsilon}\,\partial_{v_x} f^\varepsilon = \dfrac{1}{\varepsilon^2}\mathcal{Q}(f^\varepsilon), \\ f(0,x,v) = f_0(x,v), \end{cases} \tag{3.47}$$

where $(t,x,v) \in \mathbb{R}^+ \times [0,x_*] \times \mathbb{R}^{d_v}$ with periodic boundary conditions in the physical space and $v = (v_x, v_y, v_z)$. The long range interactions are modelled via the self-consistent electrical field $E = E(t,x)$ solution to the Poisson equation

$$\partial_x E^\varepsilon = \rho^\varepsilon - \bar{\rho} \quad \text{with} \quad \bar{\rho} = \iint f_0 \,\mathrm{d}x\,\mathrm{d}v. \tag{3.48}$$

The short-range interactions between particles are taken into account through the linear BGK-like relaxation collision operator

$$\mathcal{Q}(f^\varepsilon)(t,x,v) = \rho^\varepsilon(t,x)\mathcal{M}(v) - f^\varepsilon(t,x,v), \quad \forall (t,x,v) \in \mathbb{R}^+ \times [0,x_*] \times \mathbb{R}^{d_v},$$

where the global Maxwellian and local density are respectively defined as

$$\mathcal{M}(v) = \frac{e^{-|v|^2/2}}{(2\pi)^{d_v/2}}, \quad \rho^\varepsilon(t,x) = \int f^\varepsilon(t,x,v)\,\mathrm{d}v =: \langle f^\varepsilon(t,x,\cdot)\rangle.$$

It is now well known [114] that the limit case $\varepsilon = 0$ is described by a drift-diffusion equation on the density ρ. More precisely, when $\varepsilon \to 0$, the distribution function f^ε converges towards a Maxwellian distribution $\rho\mathcal{M}$ whose density ρ is solution to

$$\begin{cases} \partial_t \rho - \partial_x J = 0, \quad J = \partial_x \rho - E\rho, \\ \partial_x E = \rho - \bar{\rho}, \\ \rho(0,x) = \rho_0(x), \quad \forall x \in [0,x_*]. \end{cases} \tag{P}$$

3.6.2 Macroscopic models

The aim of this section is to derive a higher-order macroscopic model from which we deduce a macroscopic coupling criterion. It generalizes the approach presented in the previous sections and [148]. We again consider the truncated Chapman-Enskog expansion of the distribution function f^ε at order $K \in \mathbb{N}^*$:

$$f^\varepsilon(t,x,v) = \rho^\varepsilon(t,x)\mathcal{M}(v) + \sum_{k=1}^{K} \varepsilon^k h^{(k)}(t,x,v). \tag{3.49}$$

By inserting (3.49) into the original equation (3.47), one can identify powers of epsilon to obtain

$$k = 0: \qquad h^{(1)} = -\mathsf{T}(\rho^\varepsilon \mathcal{M}), \tag{3.50a}$$

$$k = 1: \qquad h^{(2)} = -\partial_t(\rho^\varepsilon \mathcal{M}) - \mathsf{T}(h^{(1)}), \tag{3.50b}$$

$$2 \le k \le K-1: \quad h^{(k+1)} = -\partial_t h^{(k-1)} - \mathsf{T}(h^{(k)}), \tag{3.50c}$$

where $\mathsf{T}f = v_x \partial_x f + E\,\partial_{v_x} f$ is the transport operator and E is now time dependent. To obtain a hierarchy of macroscopic models, one can again consider different truncation orders K, then plug (3.49) into (3.47) and integrates in velocity. The order $K = 1$ allows us to (formally) recover the asymptotic model (P).

Let us now recall the idea behind the computations in a $1D_x$-$1D_v$ setting for the case $K = 3$. Note that the same method can be applied up to the full $3D/3D$ setting. The set of equations (3.50) allows us to compute the functions $h^{(k)}$, $k = 1,2,3$. Using the identity $\partial_v \mathcal{M} = -v\mathcal{M}$, one has

$$\begin{aligned}
h^{(1)} &= -v\mathcal{M}J^\varepsilon, \quad J^\varepsilon = \partial_x \rho^\varepsilon - E^\varepsilon \rho^\varepsilon,\\
h^{(2)} &= -\mathcal{M}\partial_t \rho^\varepsilon + v^2 \mathcal{M}\partial_x J^\varepsilon + (1-v^2)\mathcal{M}E^\varepsilon J^\varepsilon,\\
h^{(3)} &= v\mathcal{M}J^\varepsilon + v\mathcal{M}\partial_x(\partial_t \rho^\varepsilon) - v^3\mathcal{M}\partial_{xx}J^\varepsilon - (v-v^3)\mathcal{M}\partial_x(E^\varepsilon J)\\
&\quad - v\mathcal{M}E^\varepsilon \partial_t \rho^\varepsilon - (2v-v^3)\mathcal{M}E^\varepsilon \partial_x J^\varepsilon - (v^3-3v)(E^\varepsilon)^2 J^\varepsilon.
\end{aligned}$$

Integrating in velocity yields

$$\partial_t \rho^\varepsilon + \partial_x \langle v h^{(1)} \rangle + \varepsilon^2 \partial_x \langle v h^{(3)} \rangle = \mathcal{O}(\varepsilon^4). \tag{3.51}$$

It remains to compute the quantities $\partial_x \langle v h^{(1)} \rangle$ and $\partial_x \langle v h^{(3)} \rangle$. By the definition of $\mathcal{M}$, one has $m_2 = 1$, $m_4 = 3$ and $\partial_x \langle v h^{(1)} \rangle = \partial_x J^\varepsilon$. Moreover, to avoid some approximation of mixed derivatives, we observe from (3.51) that

$$\partial_t \rho^\varepsilon = \partial_x J^\varepsilon + \mathcal{O}(\varepsilon^2), \tag{3.52}$$

$$\partial_t J^\varepsilon = \partial_{xx} J^\varepsilon - E^\varepsilon \partial_x J^\varepsilon - \rho^\varepsilon \partial_t E^\varepsilon + \mathcal{O}(\varepsilon^2). \tag{3.53}$$

Note that compared to the case where E is not self-consistent, and in particular not time dependent, one now has to be careful in the computation of $\partial_t J^\varepsilon$. Replacing the time derivatives by their

approximations (3.52) and (3.53) yields

$$\partial_x \langle v h^{(3)} \rangle = -\partial_{xxx} J^\varepsilon + \partial_x \rho^\varepsilon \partial_t E^\varepsilon - \rho^\varepsilon \partial_t (\partial_x E^\varepsilon) + 3 \partial_x E^\varepsilon \partial_x J^\varepsilon + E^\varepsilon \partial_{xx} J^\varepsilon + 2 J^\varepsilon \partial_x (\partial_x E^\varepsilon) + \mathcal{O}(\varepsilon^2).$$

Finally, using the Poisson equation (3.48) and rearranging the terms, we obtain a higher order macroscopic model:

$$\partial_t \rho^\varepsilon - \partial_x J^\varepsilon = -\varepsilon^2 \mathcal{R} + \mathcal{O}(\varepsilon^4), \tag{3.54}$$

where the remainder $\mathcal{R}^\varepsilon$ is given by

$$\mathcal{R}^\varepsilon = -\partial_{xxx} J^\varepsilon + E^\varepsilon \partial_{xx} J^\varepsilon + (2\rho^\varepsilon - 3\bar{\rho}) \partial_x J^\varepsilon + 2 J^\varepsilon \partial_x \rho^\varepsilon - \partial_x \rho^\varepsilon \partial_t E^\varepsilon, \tag{3.55}$$

with $\partial_x E^\varepsilon = \rho^\varepsilon - \bar{\rho}$. Let us emphasize that this term does not depend on the velocity variable but on ρ and E. It quantifies the deviation from the thermodynamical equilibrium.

3.6.3 Numerical results

Let us now present some numerical simulations with our approach for the $1D_x$-$3D_v$ case. The numerical scheme used follows the same micro-macro framework as in the case without coupling with the Poisson equation and the electrical field is updated explicitly. For the sake of completeness, we state the following proposition :

Proposition 7. *Let $n \in \mathbb{N}$. Let $\left(g^{\varepsilon,n}_{i+\frac{1}{2},j}\right)_{ij}$ and $\left(\rho^{\varepsilon,n}_i\right)_i$ be given by the following micro-macro finite volume scheme:*

$$g^{\varepsilon,n+1}_{i+\frac{1}{2},j} = g^{\varepsilon,n}_{i+\frac{1}{2},j} e^{-\Delta t/\varepsilon^2} - \varepsilon(1 - e^{-\Delta t/\varepsilon^2}) \left(\frac{T^{\varepsilon,n}_{i+\frac{1}{2},j}}{\Delta x \Delta v^3} + \xi_j \mathcal{M}_j J^{\varepsilon,n}_{i+\frac{1}{2}} \right),$$

$$\rho^{\varepsilon,n+1}_i = \rho^{\varepsilon,n}_i - \frac{\Delta t}{\varepsilon \Delta x} \left(\langle \xi g^{\varepsilon,n+1}_{i+\frac{1}{2}} \rangle_\Delta - \langle \xi g^{\varepsilon,n+1}_{i-\frac{1}{2}} \rangle_\Delta \right),$$

$$E^{\varepsilon,n}_{i+\frac{1}{2}} - E^{\varepsilon,n}_{i-\frac{1}{2}} = (\rho^{\varepsilon,n}_i - \bar{\rho}) \Delta x,$$

where $\xi_j = \xi_{(j_x,j_y,j_z)} = v_{j_x}\ \forall j \in \mathcal{J}^3$ and $T^{\varepsilon,n}_{i+\frac{1}{2},j}$ is the discretization of the transport terms. Assuming some uniform bounds in ε on ρ^ε and for a fixed mesh size Δx, $\Delta v > 0$, the scheme enjoys the AP property in the diffusion limit. This property does not depend on the initial data, and the associated limit scheme reads for $m_2^{\Delta v} = \sum_{l \in \mathcal{J}} v_l^2 M_l \Delta v_l$ as

$$\rho^{n+1}_i = \rho^n_i + m_2^{\Delta v} \frac{\Delta t}{\Delta x} \left(J^n_{i+\frac{1}{2}} - J^n_{i-\frac{1}{2}} \right), \quad J^n_{i+\frac{1}{2}} = \frac{\rho^n_{i+1} - \rho^n_i}{\Delta x} - E_{i+\frac{1}{2}} \rho^n_{i+\frac{1}{2}},$$

$$E^n_{i+\frac{1}{2}} - E^n_{i-\frac{1}{2}} = (\rho^n_i - \bar{\rho}) \Delta x,$$

with the limit flux

$$J^n_{i+\frac{1}{2}} = \frac{\rho^n_{i+1} - \rho^n_i}{\Delta x} - E_{i+\frac{1}{2}} \rho^n_{i+\frac{1}{2}}.$$

In addition, the hybrid procedure is the same as before with a slight modification of the

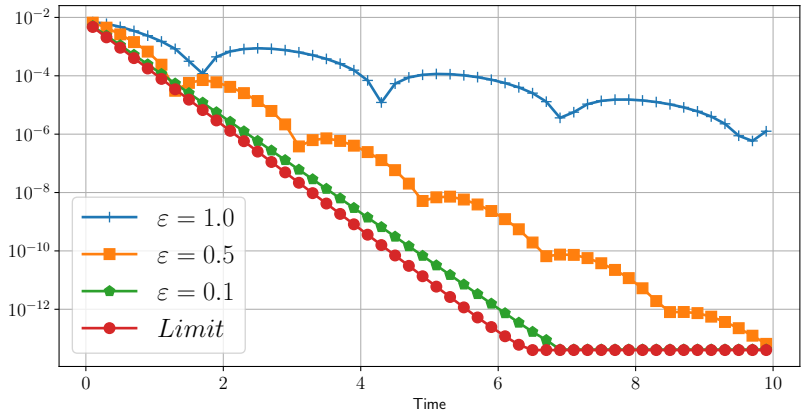


Figure 3.22 – **Fully kinetic scheme: Landau damping.** Time evolution of $\|E(t)\|_2$ for different values of ε.

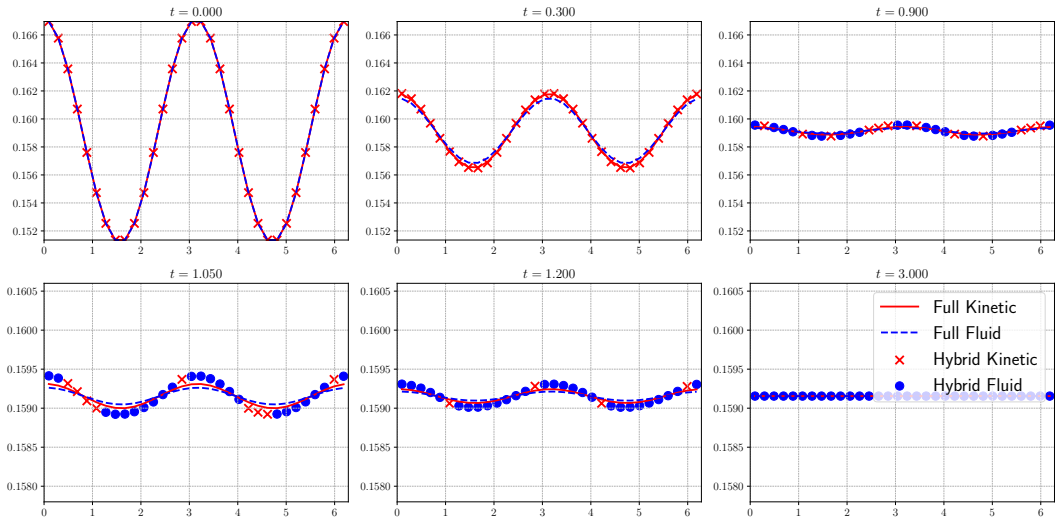


Figure 3.23 – **Comparison of the solvers.** Time evolution of the space densities for $\varepsilon = 0.1$ with a smooth initial data.

computation of the remainder $\mathcal{R}^\varepsilon$ that takes into account the time dependence of the electrical field.

We shall start with the validation of the AP property of the micro-macro Vlasov-Poisson-BGK solver with exponential integrator presented in Section 3.3.2. We recall that it is a combination of the methods presented in [59, 155] that have never been implemented in any work, to the best of our knowledge.

Figure 3.22 presents the time evolution of the L^2 norm of the electric field of solutions to equations (3.47–3.48) with different values of the relaxation parameter ε, for the seminal weak Landau damping initial data from [59]. We observe the convergence with respect to ε of the electric energy. The oscillations due to the Vlasov-Poisson transport term T occur only in the kinetic regime, when ε is large. They are then damped by the linear BGK term for smaller values of ε, where exponential decay of the electric field occurs.

We now turn our attention to the hybrid method. Figure 3.23 assesses the validity of this new method by computing the time evolution of the density of a smooth solution. The initial

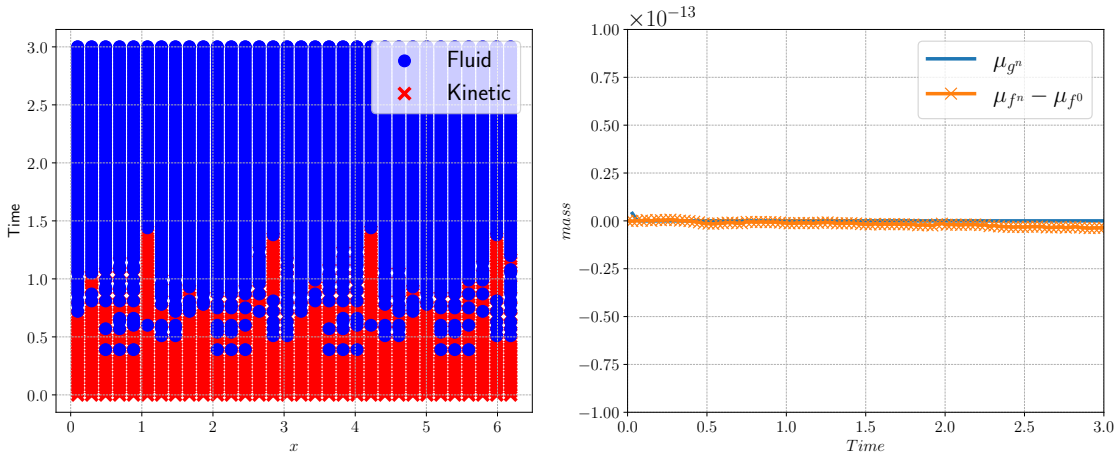


Figure 3.24 – **Hybrid scheme.** *Left.* Time evolution of the state of the cells. *Right.* Evolution of the mass variation (orange crosses) and mass of g^ε (solid blue line) for $\varepsilon = 0.1$.

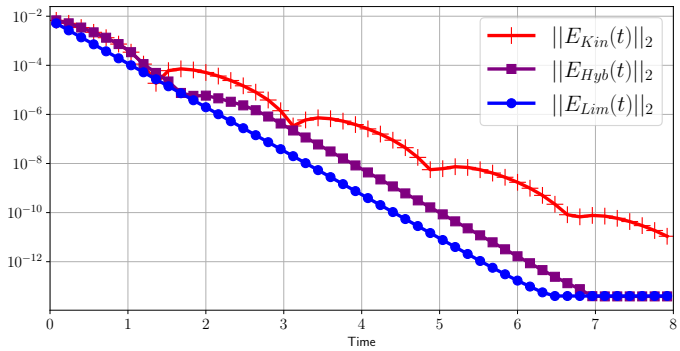


Figure 3.25 – **Landau damping revisited.** Time Evolution of $\|E(t)\|_2$ for $\varepsilon = 0.5$.

condition is a Maxwellian distribution $\rho\mathcal{M}$, whose initial space dependent density is given by $\rho(x) = 1 + 0.05\cos(2x)$ for $x_* = 2\pi$ with a fixed $\varepsilon = 0.1$. Although the solution is far from the fluid description because of this mild value of ε, we observe an almost perfect agreement between the fully kinetic and the hybrid kinetic-fluid solvers. We also observe the back-and-forth phenomenon between kinetic and fluid cells, resulting in large time in a full fluid (albeit accurate with the kinetic equation) solver for the global equilibrium. The speed-up tables are similar to those presented in [148], where factors up to 400 have been observed between the hybrid and the fully kinetic solver.

Then this domain adaptation is investigated in Figure 3.24. One can observe the precise domain adaptation during time. We notice in particular the quick vanishing of the kinetic cells in favor of the fluid ones (and hence a computational speedup). This adaptation phenomenon can bring mass variation as noticed in [148], but we observe that it remains very close to the machine precision.

Finally, we investigate again the case of the weak Landau damping using the new hybrid solver. We observe in Figure 3.25 that this new method is able to accurately capture the oscillations induced by the transport operator T in short time. Nevertheless, these oscillations are destroyed by the switch to a full fluid solver, which relaxes exponentially.

Chapter 4

Multiscale parareal algorithm for collisional kinetic equations

In this chapter, I present an ongoing work on the design of a multiscale parareal method for kinetic equations. The goal is to reduce the cost of a fully kinetic simulation using a parallel in time procedure. Using the multiscale property of kinetic models, the cheap, coarse propagator consists in a fluid solver and the fine (expensive) propagation is achieved through a kinetic solver for a collisional Vlasov equation.

Outline of the current chapter

4.1 Introduction

The simulation of many complex phenomena involving interacting particles can be modelled through either fluid or kinetic descriptions. However, the validity of the models varies highly depending on the application. In particular, to accurately describe rarefied gases or charged particles in a device, a fluid description given by the Euler, Navier-Stokes or drift-diffusion equations

may break down. Typically, it occurs around shocks or because of boundary layers and a kinetic description is therefore necessary. Nevertheless, the main drawback of the latter approach lies in its numerical cost. Indeed, deterministic kinetic simulations suffer even more strongly than usual solvers from the curse of dimensionality because of the large size of the phase space. As a consequence, one wants to resort to kinetic simulation sparingly.

In this work, we are interested in the approximation of solutions to the following scaled collisional Vlasov equation:

$$\begin{cases} \partial_t f^\varepsilon(t,x,v) + \dfrac{1}{\varepsilon^\alpha}\Big(v\cdot\nabla_x f^\varepsilon(t,x,v) - E\cdot\nabla_v f^\varepsilon(t,x,v)\Big) = \dfrac{1}{\varepsilon^{\alpha+1}}\mathcal{Q}(f^\varepsilon), \\ f(0,x,v) = f_0(x,v), \end{cases} \tag{4.1}$$

where $f(t,x,v)$ is the distribution function of particles, $t \geq 0$, $x \in \Omega_x \subset \mathbb{R}^{d_x}$, $v \in \mathbb{R}^{d_v}$. In position, the domain Ω_x is bounded, and boundary conditions will be presented later. This equation is used to model several phenomena including rarefied gases ($E = 0$) or charged particles in a tokamak, or electronic device where E typically is gradient of a potential that solves a Poisson equation. The parameter ε is the Knudsen number, which is the ratio between the mean free path of particles and the typical length scale of observation. Depending on the value of α, the asymptotic model as ε tends to 0 will either be given by a *hydrodynamic* system ($\alpha = 0$) [12, 108], such as Euler's equations, or a *diffusive* one ($\alpha = 1$) [187, 188], typically a drift-diffusion equation.

A main component of (4.1) is the Boltzmann-like collision operator $\mathcal{Q}$. In the following, we will assume that it satisfies the following classical properties:

— It preserves mass, momentum and energy:

$$\int_{\mathbb{R}^3} \mathcal{Q}(f) \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} \mathrm{d}\, v = 0_{\mathbb{R}^5}. \tag{4.2}$$

— It dissipates the Boltzmann entropy:

$$\int_{\mathbb{R}^3} \mathcal{Q}(f)\log(f)\,\mathrm{d}\, v \leq 0. \tag{4.3}$$

— Its equilibria are given by Maxwellian distributions:

$$\mathcal{Q}(f) = 0 \iff f = \mathcal{M}_{\rho,u,\theta} := \frac{\rho}{(2\pi\theta)^{d/2}} \exp\left(-\frac{|v-u|^2}{2\theta}\right), \tag{4.4}$$

where ρ, u and θ denote the density, mean velocity and temperature respectively. In regard to property (4.4) one obtains, at least formally [137], that in the limit $\varepsilon \to 0$, the distribution f^ε tends towards a Maxwellian distribution whose moments are solution to a fluid model. In the hydrodynamic scaling ($\alpha = 0$), the moments solve Euler's equations for a monoatomic gas:

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u) + \nabla_x(\mathbb{P}) - \rho E = 0, \\ \partial_t \mathcal{E} + \nabla_x \cdot ((\mathcal{E} + \mathbb{P})u) - \rho u E = 0, \end{cases} \tag{4.5}$$

where $\mathbb{P} = \rho\theta I$ is the pressure and the energy $\mathcal{E}$ is related to the temperature through the relation,

$$\mathcal{E} = \frac{1}{2}\Big(\rho|u|^2 + d_v\rho\theta\Big).$$

In the diffusive scaling ($\alpha = 1$), they solve a drift-diffusion equation ($u = 0$, $\theta = 1$):

$$\partial_t \rho - m_2 \mathrm{div}_x J = 0, \text{ where } J = \nabla_x \rho - E\rho. \tag{4.6}$$

A widely used collision operator that we consider in this work, and which is simpler than the quadratic Boltzmann operator, is the BGK operator [23]. It satisfies the above properties and is a relaxation operator with Maxwellian equilibrium:

$$\mathcal{Q}(f) = \frac{\tau}{\varepsilon}(\mathcal{M}_{\rho,u,\theta} - f), \tag{4.7}$$

where τ depends on the moments ρ, u and θ. While this simple model gives the right asymptotic models as $\varepsilon \to 0$, a careful study using a Chapman-Enskog expansion shows that higher order systems exhibit wrong viscosity and thermal conductivity. For example, in the hydrodynamic scaling, one can obtain the Navier-Stokes equations:

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u) + \nabla_x(\mathbb{P}) = \varepsilon \nabla_x \cdot (\mu\sigma(u)), \\ \partial_t \mathcal{E} + \nabla_x \cdot ((\mathcal{E} + \mathbb{P})u) = \varepsilon \nabla_x \cdot (\mu\sigma(u)u + \kappa\nabla_x\theta), \end{cases} \tag{4.8}$$

with the strain rate tensor $\sigma(u)$ given by

$$\sigma(u) = \Big(\nabla_x u + (\nabla_x u)^T\Big) - \frac{2}{3}(\nabla_x \cdot u)I.$$

However, the viscosity and thermal conductivity, denoted by μ and κ respectively, are in fact not adequate. In particular, the Prandtl number that is related to the ratio between μ and ν is equal to 1 while it should be less than 1 for monoatomic gases. One way to solve this issue is to rely on the ES-BGK operator introduced in [128]. The idea is to recover the correct asymptotic higher order model by modifying the classical Maxwellian $\mathcal{M}$ in (4.7) by an anisotropic one denoted by $\mathcal{G}$ that still satisfied the properties (4.2)-(4.3)-(4.4):

$$\mathcal{G}[f] = \frac{\rho}{\sqrt{\det(2\pi\mathcal{T})}} \exp\left(-\frac{(v-u)\mathcal{T}^{-1}(v-u)}{2}\right),$$

where for $\beta \in [-1/2, 1]$, free parameter of the model:

$$\mathcal{T} = (1-\beta)\theta I + \beta\Theta, \quad \rho\Theta = \int_{\mathbb{R}^3} (v-u)\otimes(v-u) f \,\mathrm{d}v.$$

The corresponding collision operator is then given by

$$\mathcal{Q}(f) = \frac{\nu(\rho,\theta)}{\varepsilon}(\mathcal{G}[f] - f),$$

where the collision frequency ν arises from the Boltzmann operator [196].

Several approaches have been designed to significantly reduce the cost of kinetic solvers. An inspiration for this work are the so-called *hybrid methods*. The idea behind these approaches is to achieve a coupling between a cheap, low dimensional, fluid model and the expensive, high dimensional, kinetic one. This type of techniques [57, 87, 200] mainly relies on some kind of domain decomposition in position or velocity. As in Chapter 3, to accurately describe the solution, domain indicators that assess the validity of the cheap fluid description over the expensive kinetic one are often needed [61, 145, 158, 199]. In addition to the computation of subdomains one must also deal with boundary conditions between different kinds of models. Another type of methods that aims at reducing the cost of kinetic simulations are the dynamical Low Rank Methods that directly deal with the matrix of the discretized system [80, 143].

The parareal method. Our goal in this chapter is to focus on the time integration of multiscale kinetic equations. More precisely, we design a *Parallel IN Time* (PINT) method that takes advantage of parallel computing architectures. This approach gained a lot of popularity in the past years with the introduction of the parareal method in [159, 163]. It consists in parallelizing the time integration of a dynamical system using an iterative process based on a predictor-corrector framework.

Let us now present it for a simple ODE:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} u(t) = f(u), \quad t \in [0,T], \\ u(0) = u^0. \end{cases}$$

The goal of the parareal algorithm is to approximate the solution at some fixed discrete times T^n,

$$U^n \approx u(T^n), \quad n \in \{0,\dots,N_g\}, N_g \in \mathbb{N}.$$

To do so, instead of using a single time integrator (or propagator), two solvers are coupled through an iterative process : the *parareal iterations*. The first one, denoted by $\mathcal{G}(T^n, T^{n+1}, U^n)$, must be cheap to compute while the second one, $\mathcal{F}(T^n, T^{n+1}, U^n)$, must be very accurate. While the propagator $\mathcal{G}$ is less costly, its drawback is naturally its precision. However, it will be corrected by the accurate, but expensive, propagator $\mathcal{F}$. The algorithm then unfolds as follows :

1. Divide the time domain in $N_g \in \mathbb{N}$ subintervals $[T^n, T^{n+1}]$, $n \in \{0,\dots,N_g - 1\}$;
2. Perform a first coarse guess:

$$U^{n+1,0} = \mathcal{G}(T^n, T^{n+1}, U^{n,0}) = u^0;$$

3. Refine the guess through the parareal iterations:

$$U^{n+1,k+1} = \mathcal{G}(T^n, T^{n+1}, U^{n,k+1}) + \mathcal{F}(T^n, T^{n+1}, U^{n,k}) - \mathcal{G}(T^n, T^{n+1}, U^{n,k}),$$

where $k = 1, 2, \dots$ denotes the k^{th} parareal iteration.

The advantage of this iterative algorithm is its ability to compute the expensive fine propagations

in parallel. Therefore, as long as the number of parareal iterations remains small enough to obtain a given accuracy, one can expect a reduction of the computational cost of the time integration.

Parallel in time methods are widely utilized and studied in various contexts. The convergence of the algorithm is investigated in [75, 100, 101] including variations of the original method. The uses are plenty, and we refer to [13] for neutron transport, [89, 195] for Navier-Stokes equations, [99] for Hamiltonian systems and [191] for turbulence in plasmas. We also refer to a recent work on the acceleration of the method [162, 172]. It was furthermore observed that the parareal method suffers from poor convergence for hydrodynamic systems where convection dominates diffusion [9, 79, 83, 98, 173]. In particular, in the case of Burgers' equation it was noticed that for a fixed discretization, the longer the integration domain, the slower was the convergence [100].

In the context of multiscale systems, the parareal algorithm has been used in [31, 75, 116, 117, 154, 192] by leveraging reduced order models as coarse solvers. The idea is to use a simpler model, instead of a simpler time integrator, as a coarse propagator and a richer model for the fine propagations. These ideas are similar to the HMM method introduced in [1, 77, 78].

Expanding on these advancements, our main goal in this work is to build a multiscale parareal method for kinetic equations to perform accurate simulations of the observable macroscopic moments. A fluid model, either given by a hydrodynamic or diffusive limit of the underlying kinetic equation, is used as a coarse propagator. The fine propagation is in turn achieved by the resolution of the kinetic model. To this aim, we resort to *Asymptotic Preserving* (AP) schemes [136, 142] that remain stable for any value of the Knudsen number in (4.1). While they provide an accurate approximation in the limit $\varepsilon \to 0$, their main drawback is that their cost remains the one of a kinetic scheme even in a fluid regime. We expect that our new method can solve this issue.

Plan of the chapter. This chapter is organized as follows. In Section 4.2 we present the multiscale parareal algorithm for kinetic equations and discuss its properties. Sections 4.3 is dedicated to the implementation of the method and in particular to the solvers used. We then discuss the parallelization of the method in Section 4.4. Finally, we thoroughly assess the accuracy and properties of the method through numerical experiments in Section 4.5.

4.2 Multiscale parareal algorithm

We introduce in this section the multiscale parareal algorithm. The formalism we employ follows [154] and the references therein.

4.2.1 Link between the scales

The crucial point of a multiscale method is the link between the different scales of the model. In the context of kinetic equations a natural way to go from kinetic to fluid is to consider the projection of a distribution f towards its moments $U = (\rho, u, \theta)^T$ defined in Definition 2.

Definition 2. *For a distribution function $f(t,x,v) \in L^1((1+v)^2 \,\mathrm{d}\, v)$, we define the projection $\mathcal{P}f(t,x) =$*

$U(t,x)$ *as*

$$\mathcal{P}f = U = \begin{pmatrix} \int_{\mathbb{R}^3} f \,\mathrm{d}v \\ \frac{1}{\rho}\int_{\mathbb{R}^3} v f \,\mathrm{d}v \\ \frac{1}{3\rho}\int_{\mathbb{R}^3} |v-u|^2 f \,\mathrm{d}v \end{pmatrix} = \begin{pmatrix} \rho \\ u \\ \theta \end{pmatrix}.$$

Note that in Definition 2, we start from a distribution that contains all the information and project it onto a manifold where the moments solve a system valid only for Maxwellians. On the opposite, to lift a macroscopic data $U(t,x) \in \mathbb{R}^5$ to a distribution $f(t,x,v)$ a first idea is to consider the validity of the fluid model used. In the case of Euler's equations, we therefore chose to reconstruct a Maxwellian whose moments are given by U.

Definition 3. *For a macroscopic data $U(t,x) \in \mathbb{R}^5$, we define the lifting $\mathcal{L}U(t,x,v) = f(t,x,v)$ as*

$$\mathcal{L}U = f = \frac{\rho}{\sqrt{2\pi\theta}} \exp\left(-\frac{|v-u|^2}{2\theta}\right).$$

Another finer way to handle the lifting and fluid propagations is to consider the Chapman-Enskog expansion of the distribution:

$$f^\varepsilon = \mathcal{M}_{\rho^\varepsilon, u^\varepsilon \theta^\varepsilon} + \sum_{l=1}^{\infty} \varepsilon^l g^{(l)}, \tag{4.9}$$

where the perturbations $g^{(l)}$ can be explicitly computed and depend on f only through its moments (see Chapter 3 for more details).

We have already mentioned that in the hydrodynamic scaling, the moments of a Maxwellian distribution are solution to Euler's equations. By considering the first order perturbation in (4.9), these moments can be shown to solve the Compressible Navier-Stokes equations (CNS) that can be understood as a first order correction of Euler's equations. Consequently, a finer method would be to use a CNS solver as a coarse propagator. Note that even with this choice, the cost of the coarse integration remains much lower than a kinetic propagation. In addition, instead of only lifting the moments to a Maxwellian, one can consider adding the perturbations $g^{(l)}$. It allows for a finer reconstruction of the distribution and the consideration of far from equilibrium phenomena.

Definition 4. *With the same hypothesis of Definition 3, we define the lifting $\mathcal{L}^{(L)}U(t,x,v) = f(t,x,v)$ of order L by*

$$f = \mathcal{L}^{(L)}U = \frac{\rho}{\sqrt{2\pi\theta}} \exp\left(-\frac{|v-u|^2}{2\theta}\right) + \sum_{l=1}^{L} \varepsilon^l g^{(l)}.$$

4.2.2 Semi-discretization in time

Let us now consider the time interval $[0,T]$ and divide it in $N_g \in \mathbb{N}$ uniform subintervals $[T^n, T^{n+1}]$. We are interested in the approximation of the moments at the fixed discrete times T^n. We denote by $U^{n,k}$ this approximation at a parareal iteration k. Furthermore, we also need to introduce a fine discretization that will be used by the fine propagator. Let $N_f \in \mathbb{N}$ denote the

number of uniform fine subintervals $[t^n, t^{n+1}]$. Typically, N_f shall be greater than N_g. We can then define the coarse (resp. fine) time steps and times:

$$T^n = n\Delta t_g, \quad \Delta t_g = \frac{T}{N_g}, \quad t^n = n\Delta t_f, \quad \Delta t_f = \frac{T}{N_f}.$$

Note that the output of the algorithm will therefore contain N_g fixed snapshots at times T^n of the moments. We briefly omit the definitions of the kinetic and fluid propagators $\mathcal{F}$ and $\mathcal{G}$ as they will be discussed bellow. However, it is important to mention that both the coarse and fine time steps are in practice the maximum time steps allowed within a propagator. The local time steps inside the solvers are naturally subject to a stability condition, and we define

$$\Delta t_{\text{loc}}^{n,\mathcal{G}} = \min\left\{\Delta t_{\text{stab}}^{n,\mathcal{G}}, \Delta t_g\right\}, \qquad \Delta t_{\text{loc}}^{n,\mathcal{F}} = \min\left\{\Delta t_{\text{stab}}^{n,\mathcal{F}}, \Delta t_f\right\},$$

where $\Delta t_{\text{stab}}^{n,\mathcal{G}}$ and $\Delta t_{\text{stab}}^{n,\mathcal{F}}$ denote the time steps prescribed by the stability conditions of the numerical schemes used. The method can then be summarized in Algorithm 3. As a stopping criterion of the

Algorithm 3 Multiscale kinetic parareal Algorithm

Require: $U^{0,0}$

1: **for** $n = 1, \dots, N_g$ **do** ▷ First coarse guess
2: $\quad U^{n,0} \leftarrow \mathcal{G}\left(U^{n-1,0}\right)$
3: **end for**
4: **while** $k \leq K$ **or** error $\geq$ tol **do** ▷ Parareal iterations
5: $\quad$ **for** $n = 1, \dots, N_g$ **do** ▷ Compute the jumps in parallel
6: $\quad\quad \Delta^n = \mathcal{P}\mathcal{F}\mathcal{L}(U^{n-1,k-1}) - \mathcal{G}\left(U^{n-1,k-1}\right)$
7: $\quad$ **end for**
8: $\quad$ **for** $n = 1, \dots, N_g$ **do** ▷ Sequential correction
9: $\quad\quad U^{n,k+1} = \mathcal{G}\left(U^{n-1,k}\right) + \Delta^n$
10: $\quad$ **end for**
11: $\quad$ Compute successive error on the moments and $k \leftarrow k+1$
12: **end while**

algorithm, we simply consider the error between two successive parareal iterations:

$$\text{error} = \max_n |U^{n,k+1} - U^{n,k}|,$$

or a fixed number of iterations $K \in \mathbb{N}$.

An interesting property of Algorithm 3 is that, at parareal iteration k, one can in fact start the loops (lines 5 and 8) at k instead of 1. Indeed, since one propagates a data from time 0 that is fixed, it is unnecessary to recompute the first 1 to k iterations as they would not be modified. This can be seen through a simple ODE example by focusing on the first time intervals and parareal iterations. For $n = 1$, to update $U^{1,k+1}$, one needs to compute

$$U^{1,k} = \mathcal{G}(T^0, T^{n+1}, U^{0,1}) + \mathcal{F}(T^0, T^1, U^{0,0}) - \mathcal{G}(T^0, T^1, U^{0,0}).$$

Since $U^{0,k} = U^0$ for all k, the two coarsely propagated data cancel each other, and we obtain that the first parareal iteration automatically corrects the initial guess towards the fine propagation:

$$U^{1,k} = \mathcal{F}(T^0, T^1, U^{0,0}), \quad \forall k \geq 1.$$

Consequently, when computing the next parareal iteration $k = 2$, $U^{1,k}$ is now fixed and one can make the same observation as before but now for $n = 2$:

$$U^{2,k} = \mathcal{F}(T^1, T^2, U^{1,k}) = \mathcal{F}(T^1, T^2, \mathcal{F}(T^0, T^1, U^{0,0})), \quad \forall k \geq 2.$$

By induction, at parareal iteration k, the 1 to k first times are therefore fixed, and don't need to be recomputed. A consequence is that after N_g parareal iterations, the outcome of Algorithm 3 is the same as a fully fine simulation.

Building upon these observations, an optimized version of this algorithm is given by Algorithm 4.

Algorithm 4 Multiscale kinetic parareal Algorithm

Require: $U^{0,0}$

1: **for** $n = 1, \dots, N_g$ **do** ▷ First coarse guess
2: $\quad U^{n,0} \leftarrow \mathcal{G}\left(U^{n-1,0}\right)$
3: **end for**
4: **while** $k \leq K$ **or** error $\geq$ tol **do** ▷ Parareal iterations
5: $\quad$ **for** $n = k, \dots, N_g$ **do** ▷ Compute the jumps in parallel
6: $\quad\quad \Delta^n = \mathcal{P}\mathcal{F}\mathcal{L}(U^{n-1,k-1}) - \mathcal{G}\left(U^{n-1,k-1}\right)$
7: $\quad$ **end for**
8: $\quad$ **for** $n = k, \dots, N_g$ **do** ▷ Sequential correction
9: $\quad\quad U^{n,k+1} = \mathcal{G}\left(U^{n-1,k}\right) + \Delta^n$
10: $\quad$ **end for**
11: $\quad$ Compute successive error on the moments and $k \leftarrow k + 1$
12: **end while**

From an optimization point of view, note also that steps 5-7 of both Algorithms 3 and 4 contain all the expensive computations. More particularly we chose to do the lifting and the projection alongside the kinetic propagation to better distribute the workload. Indeed, since the iterations of this loop are independant from one another, which is the core of the parareal method, it can be parallelized to reduce the computation time.

4.3 Numerical schemes

In this section, we detail our choice of fine and coarse propagators as well as the discrete version of the lifting and projection operators.

From now on, let us place ourselves in the $d_x = 1$ and $d_v = 3$ setting. The phase space (x, v) is discretized in a finite volume fashion using control volumes $K_{ij} = \mathcal{X}_i \otimes \mathcal{V}_j$, $i \in \{1, \dots, N_x\}t$ and

$j = (j_x, j_y, j_z) \in \{1, \dots, N_v\}^3$. The discrete unknowns are defined as

$$f_{ij}^n \approx \frac{1}{\Delta x_i \Delta v_j} \int_{K_{ij}} f(T^n, x, v) \mathrm{d}x \mathrm{d}v, \quad U_i^n \approx \frac{1}{\Delta x_i} \int_{\mathcal{X}_i} U(T^n, x) \mathrm{d}x,$$

with Δx_i the volume of $\mathcal{X}_i$ and Δv_j the volume of the cube $\mathcal{V}_J$.

4.3.1 Kinetic schemes

The main difficulty to approximate scaled kinetic equations is the stability of the time integration with respect to the small parameters. In both hydrodynamic and diffusive scaling, AP schemes must therefore be considered to ensure tractable simulations for any value of the Knudsen number. For the phase space discretizations, finite volume schemes either of upwind or Rusanov type are used [157, 203].

Hydrodynamic scaling. In the hydrodynamic scaling, several AP schemes have been developed, and they rely on an IMEX type time discretization [180]. For the classical BGK collision operator, we consider the scheme introduced in [53] and for the ESBGK case we refer to the method in [85]. These two approaches rely on an implicit treatment of the collision operator, but that can be solved explicitly, resulting in an efficient numerical scheme.

Diffusive scaling. In the diffusive scaling, we adopt the micro-macro approach introduced in [16, 59] together with an exponential time integrator for the microscopic equation [148, 155].

4.3.2 Fluid schemes

Coarse propagators aim at giving a decent guess of the behavior of the moments at a low numerical cost. We distinguish two types of schemes, depending on the scaling considered at the kinetic level.

Systems of conservation laws. Systems of conservation laws [157, 203] typically arise as hydrodynamic limits of kinetic equations. As a coarse propagator, we rely on a simple finite volume discretization with Rusanov-type fluxes and a first order forward Euler time integration.

Drift-diffusion equation. Explicit discretizations of diffusive systems usually suffer from a parabolic stability condition. However, since the cost of such a scheme is essentially negligible compared to a full kinetic simulation, we do not consider an implicit-in-time procedure. The fluid system is therefore approximated using a central finite difference scheme that is nothing but the asymptotic limit of the micro-macro scheme mentioned above [148, 155].

4.3.3 Discrete lifting and projection operators

A natural way to define the discrete lifting operator is to compute the pointwise discrete Maxwellian associated to the moments $U_i = (\rho_i, u_i, \theta_i)^T$, $i \in \{1, \dots, N_x\}$:

$$(\mathcal{L}U)_{ij} = \frac{\rho_i}{(2\pi\theta_i)^{3/2}} \exp\left(-\frac{|v_j - u_i|^2}{2\theta_i}\right), \quad j = (j_x, j_y, j_z) \in \{1, \dots, N_v\}^3. \tag{4.10}$$

We could also define higher order lifting as defined in the continuous case (4) and proceed in the same way using pointwise computations and finite difference approximations of the derivatives. To project a discrete distribution f_{ij} towards its discrete moments $(\rho_i, u_i, \theta_i)^T$, we consider a simple first order quadrature:

$$\rho_i = \sum_j f_{ij} \Delta v_j, \quad (\rho u)_i = \sum_j v_j f_{ij} \Delta v_j, \quad 3(\rho\theta)_i = \sum_j |v_j - u_i|^2 f_{ij} \Delta v_j.$$

Note that more accurate quadratures could be considered depending on the velocity discretization.

4.4 Parallelization of the method

In order to achieve a significant gain in computational time, one needs to rely on an efficient parallelization. The method is currently implemented using an OpenMP paradigm to validate its concept. Its real performance should stand out when properly deployed on a supercomputer along with an MPI parallelization over several processors. The parallelization of this type of method on distributed memory architecture is known to be challenging and such an implementation is discussed in Appendix B.

Theoretical speed up. In an ideal setting, *i.e.* by omitting the cost of communications/synchronizations between cores/threads, one can derive an estimate on the number of parareal iterations to obtain a speedup with the method. Let us denote by T_{Kin} (resp. T_{Fluid}) the maximum time for the kinetic (resp. fluid) solver to evolve an initial data from time T^n to time T^{n+1}. Note that we take the maximum over all propagations because, on each sub time-interval, the stability condition may vary, impacting the optimal local time step. In addition, one needs to take into account the cost of the projection and lifting operators since they are fully non-local and can therefore be significantly expensive. We shall denote by T_{Lift} and T_{Proj} these costs. The ideal numerical cost of Algorithm 3 on N_p processors with N_g sub-time intervals is then given by the formula:

$$T_{\text{Parareal}} = T_{\text{Fluid}} + N_g k \left(\frac{T_{\text{Lift}} + T_{\text{Proj}} + T_{\text{Kin}} + T_{\text{Fluid}}}{N_p} + T_{\text{Fluid}} \right).$$

Consequently, the ideal number of parareal iterations to be less costly than a fully kinetic simulation should satisfy:

$$T_{\text{Parareal}} \leq N_g T_{\text{Kin}}$$

# CPUs	2 x Intel Xeon Silver 4116
# cores	24
# threads	48
RAM	188Go

Table 4.1 – Computing architecture.

or stated differently,

$$k_{\text{opt}} \leq \left\lceil \frac{N_g T_{\text{Kin}} - T_{\text{Fluid}}}{N_g \left(\frac{T_{\text{Lift}} + T_{\text{Proj}} + T_{\text{Kin}} + T_{\text{Fluid}}}{N_p} + T_{\text{Fluid}} \right)} \right\rceil.$$

The performance of the method and scalability of the current implementation in Fortran90 (gfortran 11.4.0) with an OpenMP parallelization will be discussed in the next Section 4.5. These tests were ran on the architecture detailed in Table 4.1.

4.5 Numerical results

In this section, we present some numerical results for simulations of the 1D/3D Vlasov-BGK equation (4.1) in the hydrodynamic scaling. We consider Algorithm 4 and assess its accuracy as well as its performance through several experiments. In the following, unless specified otherwise, the phase space is uniformly discretized using $200 \times 32 \times 32 \times 32$ cells, namely 200 points in position and 32 in each velocity direction. Note that, the method was only implemented in the hydrodynamic scaling with an Euler solver at the fluid level and the lifting doesn't consider any perturbation. Extensions to the cases mentioned in Section 4.3 will be implemented in the future.

4.5.1 Test 1: Sod shock tube

We first consider a Riemann problem where the initial data is given by:

$$f_0(x,v) = \mathcal{M}_{\rho(x),u(x),\theta(x)}(v), \quad x \in [0,2], \quad v \in [-8,8],$$

where

$$(\rho(x),u(x),\theta(x)) = \begin{cases} (1,0,0,0,1) & \text{if } x < 1, \\ (0.125,0,0,0,0.8) & \text{if } x \geq 1. \end{cases}$$

The exterior force is set to 0, and the regime is rarefied: $\varepsilon = 10^{-2}$ at the kinetic level. Boundary conditions are absorbing. We set a simulation interval $[0,T]$ with $T = 0.5$ and discretize it uniformly using $N_g = 200$ points for the coarse grid and $N_f = 800$ points for the fine grid. As a stopping criterion, we set a maximum of $K = 80$ iterations or when then consecutive error (4.2.2) is smaller than 10^{-8}. Note that this threshold is much smaller than both the fine and coarse time increments. We present in Figure 4.1 snapshots of the moments obtained with Algorithm 4 as well as the ones obtained through fully kinetic and fully fluid simulations. Furthermore, we can observe a very good agreement between the kinetic and parareal densities. We can also notice that higher order moments are less well captured, and this observation also holds as we look for larger times. Nevertheless, considering we used the simplest fluid model and reconstruction, the results are

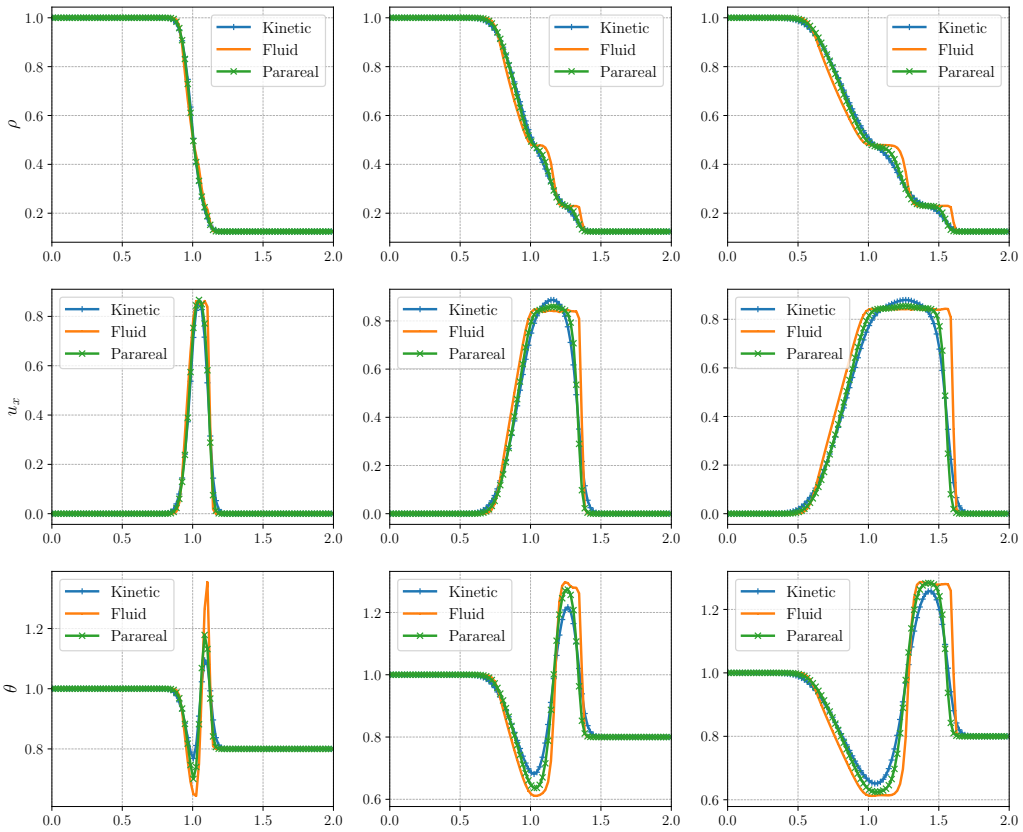


Figure 4.1 – **Test 1 - Sod shock tube**, $\varepsilon = 10^{-2}$: Snapshots of the density (Top), x mean velocity (Middle) and Temperature (Bottom) at times $T^n = 0.05$ (Left), 0.15 (Middle) and 0.25 (Right).

promising.

Convergence. Figure 4.2 illustrates the convergence of the algorithm, by plotting in loglog scale the successive errors (4.2.2) at each parareal iteration. A first observation is that the algorithm indeed converges well. In particular, the error decreases exponentially fast which is encouraging in the sense that one may need only a few iterations to reach the desired accuracy.

Performance. We conclude this test by first investigating the parallel efficiency of our implementation. The scalability is assessed by computing $K = 10$ parareal iterations. We report in Figures 4.3 the runtimes as well as the speedup per number of threads considered:

$$\text{speedup} = \frac{\text{Sequential runtime}}{\text{Runtime on } p \text{ threads}}.$$

The ideal runtime mentioned in Figure 4.3 is the sequential runtime divided by the number of threads. We observe that our OpenMP implementation indeed allows to reduce the computational

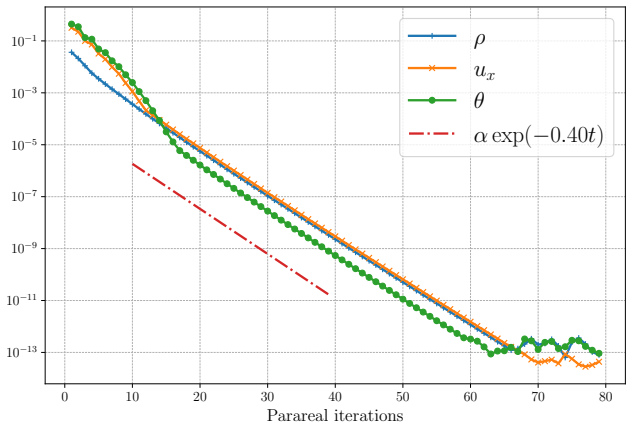


Figure 4.2 – **Test 1 - Sod shock tube**, $\varepsilon = 10^{-2}$: Convergence of the successive errors.

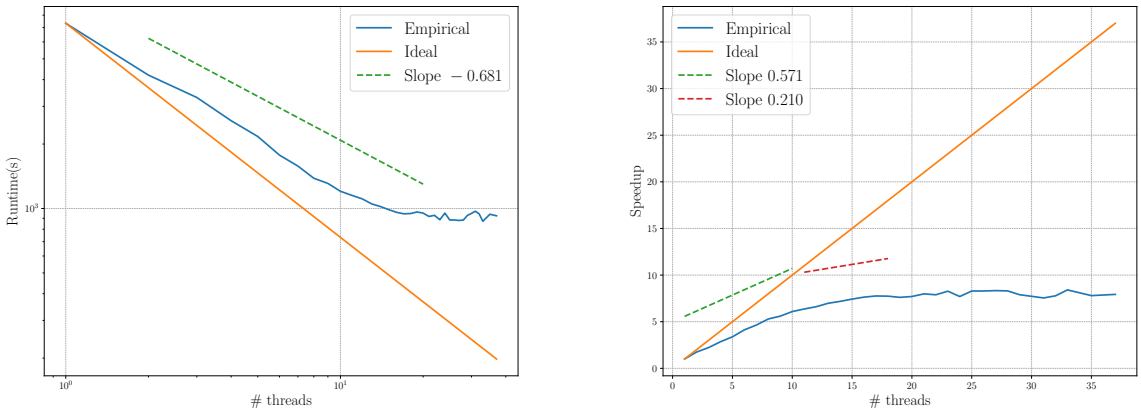


Figure 4.3 – **Test 1 - Sod shock tube**, $\varepsilon = 10^{-2}$: Runtime (Left) and parallel speedup (Right) as a function of the number of threads.

time. The empirical runtime scales as 0.68 with the number of threads. While one is far from the ideal speedup it is important to note that only a basic parallelization was considered, and one may therefore try to optimize more finely the workload distribution using an MPI paradigm. We discuss such an implementation in Appendix B.

Let us now assess the performance of the parareal algorithm compared to a kinetic simulation. We choose to use 24 threads. It is important to recall that such a method is designed to be heavily parallelized and an efficient, scalable parallelization is far from being straightforward. Therefore, the results below also reflect the implementation as well as the specific choice of test case. Indeed, as it was for example observed in [162], a speedup can be more significant for large time simulations. Moreover, as we illustrate in Table 4.2, the acceleration also naturally depends on the required accuracy. In this simple test case, we observe in particular no real speedup compared to a fully kinetic simulation.

While the first results in terms of acceleration are not as good as one could expect, we still showed that the algorithm is able to converge. These encouraging findings could now be refined

	Fluid	Kinetic	Parareal 1	Parareal 10	Parareal 20
Runtime (s)	2.73E-03	6.03E+02	1.11E+02	8.91+02	1.52E+03

Table 4.2 – **Test 1 - Sod shock tube**, Comparison of the computational time between fully fluid, fully kinetic, and the parareal Algorithm 4 on 24 threads.

in terms of parallel implementation and long-time simulation. Another idea would be to consider more sophisticated versions of the parareal algorithm constructed with improvement of performance in mind [162, 172].

4.5.2 Test 2: Blast waves

We now consider a second Riemann problem consisting of a blast wave. The initial data is at a local equilibrium in velocity.

$$f_0(x,v) = \mathcal{M}_{\rho(x),u(x),\theta(x)}(v), \quad x \in [0,2], \quad v \in [-8,8],$$

with the moments defined by:

$$(\rho(x),u(x),\theta(x)) = \begin{cases} (1,1,0,0,2) & \text{if } x < 0.4, \\ (1,0,0,0,0.25) & \text{if } 0.4 \le x < 1.6, \\ (1,-1,0,0,2) & \text{if } x \ge 1.6. \end{cases}$$

We consider the same setting as in Test 1, and we fix the number of parareal iterations to $K = 10$. We present in Figure 4.4 snapshots of the moments. Note that with only 10 iterations, the consecutive error is only of the order of 10^{-3}. We observe that the solution has not yet converged towards the kinetic one, and this is even more striking for large time. Nevertheless, this behavior is expected as the parareal algorithm first corrects for early times and then propagates this correction. In Figure 4.5, we investigate the main source of error by plotting the pointwise difference between the kinetic and parareal moments. According to Figure 4.4 we observe that the deviation between kinetic and parareal is essentially localized near shocks, where the regularity of the solution is low. Again, such behavior is not surprising as it was already observed in previous works [9, 79, 83, 98, 173]. We also report in Figure 4.6 the convergence of the algorithm where we observe the same behavior as in Test 1.

4.5.3 Test 3: Exterior forces

We conclude this investigation of the method by considering the case of a smooth initial data along with a non-zero exterior force. Moreover, we consider this time an initial data far from the Maxwellian equilibrium given by two opposite beams with constant densities:

$$f_0(x,v) = \mathcal{M}_{\rho_1(x),u_1(x),\theta_1(x)}(v) + \mathcal{M}_{\rho_2(x),u_2(x),\theta_2(x)}(v), \quad x \in [0,2], \quad v \in [-8,8],$$

where the two sets of moments are given by:

$$(\rho_1(x),u_1(x),\theta_1(x)) = (1,1,0,0,1) \quad .$$

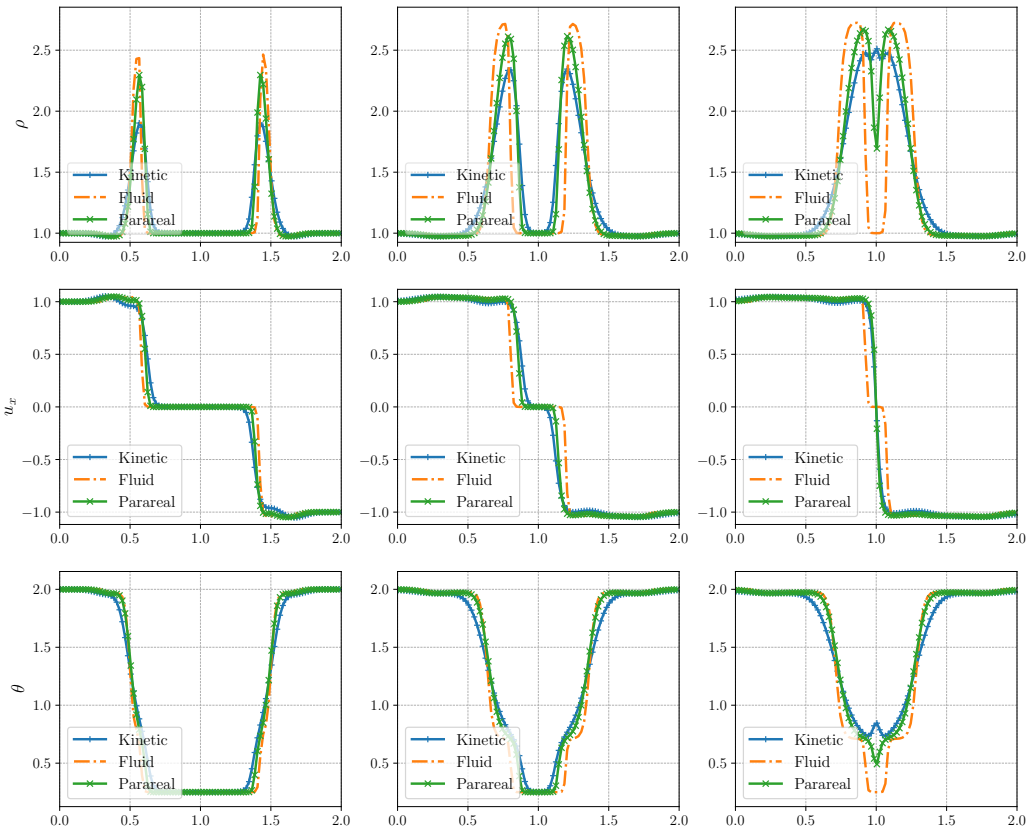


Figure 4.4 – **Test 2 - Blast waves**, $\varepsilon = 10^{-2}$: Snapshots of the density (Top), x mean velocity (Middle) and Temperature (Bottom) at times $T^n = 0.1$ (Left), 0.23 (Middle) and 0.3 (Right).

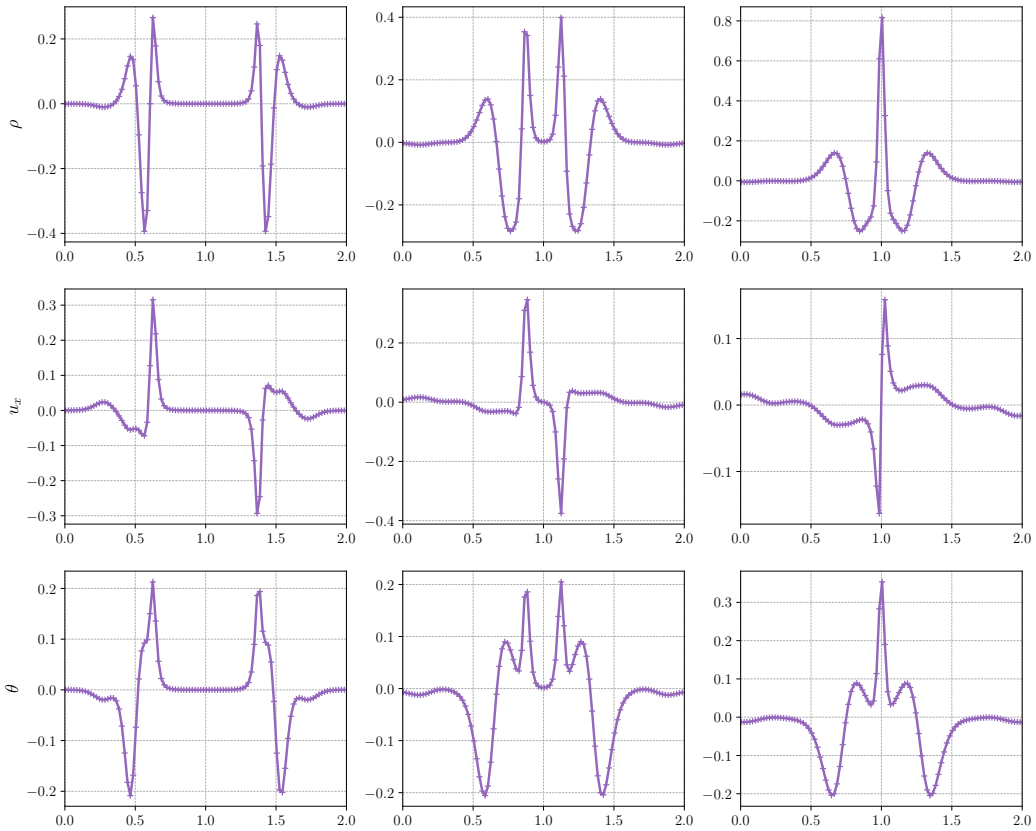


Figure 4.5 – **Test 2 - Blast waves**, $\varepsilon = 10^{-2}$: Snapshots of the pointwise difference on the density (Top), x mean velocity (Middle) and Temperature (Bottom) at times $T^n = 0.1$ (Left), 0.23 (Middle) and 0.3 (Right).

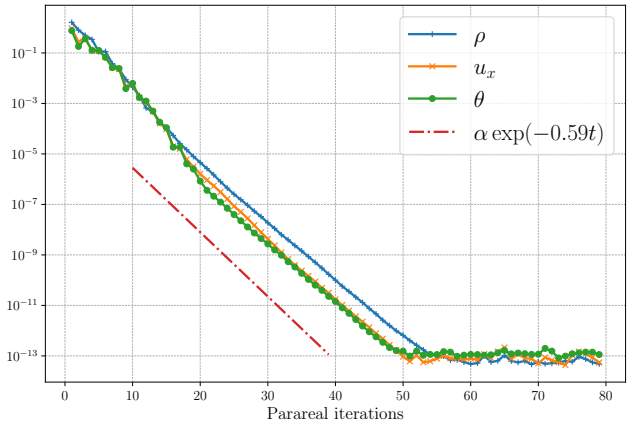


Figure 4.6 – **Test 2 - Blast waves**, $\varepsilon = 10^{-2}$: Convergence of the successive errors.

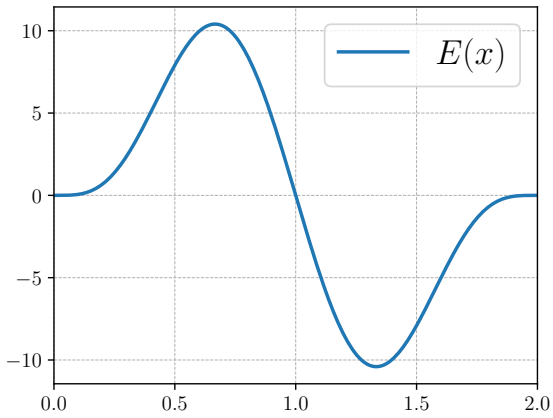


Figure 4.7 – **Test 3 - Exterior forces:** Plot of the external force $E(x)$.

and

$$(\rho_2(x), u_2(x), \theta_2(x)) = (1, -1, 0, 0, 1). \quad .$$

We set the exterior forces as

$$E(x) = -50x^4(x-2)^4(x-1).$$

This exterior force is illustrated in Figure 4.7 and will concentrate the mass towards the center of the domain $x = 1$.

The same discretization setting as the previous tests is considered with the difference that we assume periodic boundary conditions and a Knudsen number of 10^{-6}. The resulting moments are presented in Figure 4.8. We can observe that the mass indeed concentrates towards the center of the domain until the pressure is too high and discontinuities appear. In particular, the solution for short times being smooth, we obtain a very good agreement between kinetic and parareal moments. It holds even for times at the end of the simulation interval. In addition, we also illustrate the convergence of the algorithm in Figure 4.9. We observe a convergence rate that is similar to the less regular previous cases. A natural follow up to this test will be to consider a coupling with Poisson's equation.

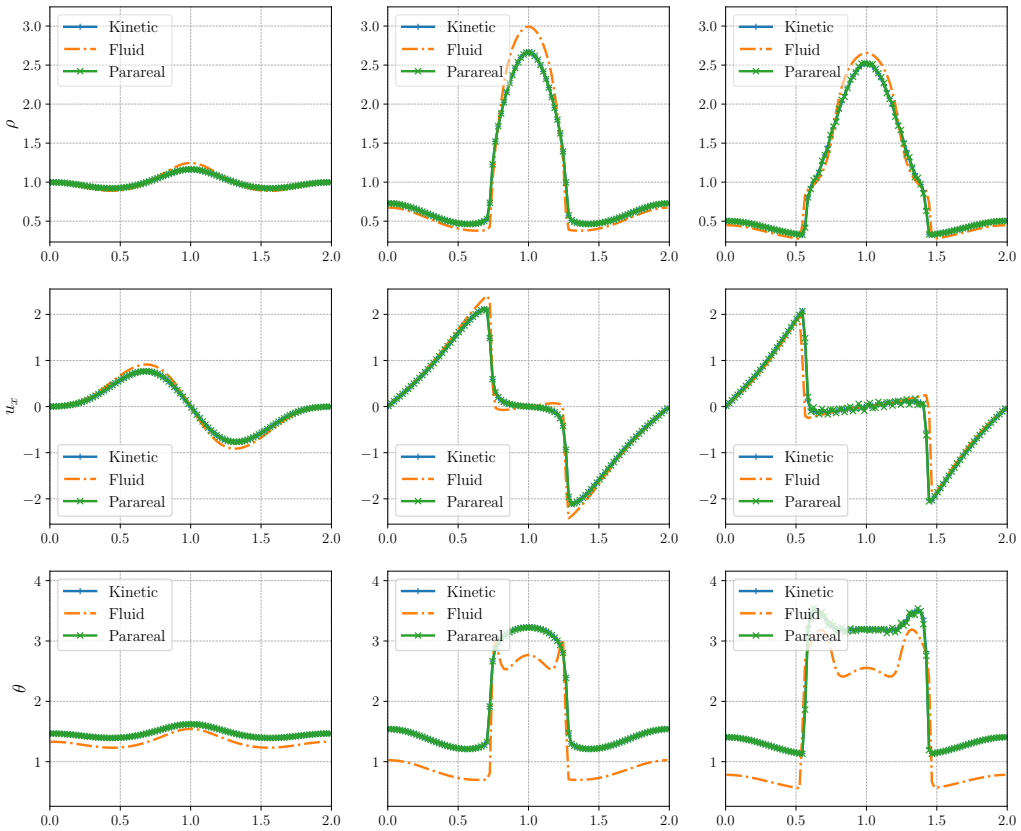


Figure 4.8 – **Test 3 - Exterior forces**, $\varepsilon = 10^{-6}$: Snapshots of the density (Top), x mean velocity (Middle) and Temperature (Bottom) at times $T^n = 0.075$ (Left), 0.3 (Middle) and 0.4 (Right).

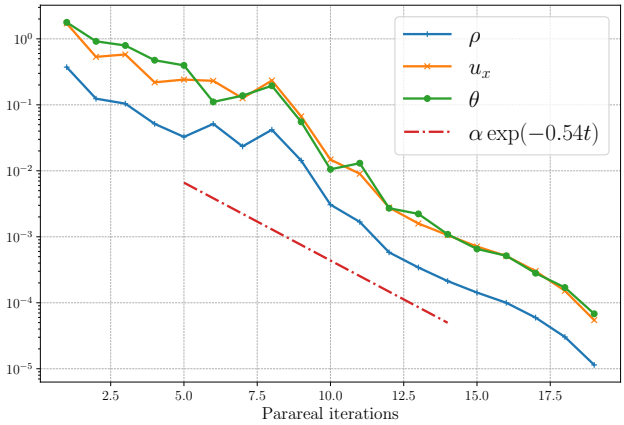


Figure 4.9 – **Test 3 - Exterior forces**, $\varepsilon = 10^{-6}$: Convergence of the successive errors.

Conclusions and perspectives

Outline of the current chapter

I presented in this manuscript several new numerical methods that leverage the multiscale nature of kinetic equations. From a theoretical point of view they are designed to preserve fundamental properties such as the long time behaviour and conservation of physical quantities. I also considered the computational aspect of kinetic simulations and developed hybrid kinetic/fluid techniques that significantly reduce their cost.

Structure preserving numerical methods

Discrete hypocoercivity

I presented a finite volume discretization of a one dimensional nonlinear kinetic model which describes a 2-species recombination-generation process. The long-time convergence of approximate solutions towards equilibrium is shown to occur at an explicit and quantified exponential rate. This study is based on an adaptation for a discretization of the linearized problem of an L^2 hypocoercivity method. The nonlinear local result is, to the best of our knowledge, the first result of this type for a discrete nonlinear kinetic problem.

In addition to this result, we observed numerically that convergence is exponential even for initial data far from equilibrium. This shows in particular that the implementation of the method is very robust, and that the result obtained is in fact more general. In line with this idea, we also observed that the nonlinear entropy decreases along discrete solutions

$$\mathcal{H}(f_1, f_2) = \sum_{i=1,2} \iint_{\mathbb{T}\times\mathbb{R}} \left[f_i \left(\ln\left(\frac{f_i}{f_{i,\infty}} - 1 \right) \right) + f_{i,\infty} \right] \mathrm{d}\, v \, \mathrm{d}\, x. \tag{4.11}$$

This suggests that a careful study of this quantity may allow to generalize our results. Still in

the context of physically motivated model, an interesting perspective is the study the of long time behaviour of a fully finite volume scheme for a collisional Vlasov system. Such a discrete hypocoercivity result has already been obtained in [24, 25] using a Hermite decomposition in velocity, but I believe that a full finite volume study would need the development of interesting techniques.

The long-time behaviour of models arising in biology is also an interesting topic, in particular the model introduced by Othmer, Dunbar and Alt [174]. Convergence towards a stationary state for a *run-and-tumble* model was studied in [41] using hypocoercivity techniques, the same technique that we adapted to the discrete setting. Note that this approach relies on a very good knowledge of the steady state. More recently, in [42, 81, 82], it has been shown that this type of hypothesis can be lifted by using the Harris Theorem, derived from the ergodic theory of Markov processes. It would therefore be interesting to explore this type of proof at the discrete level.

Moment preserving spectral methods

I presented an extension and analysis of a class of Galerkin spectral methods for PDEs, based on general families of orthogonal functions, that can preserve the moments of the solution. Leading examples of such problems arise in kinetic and mean-field theory, where preservation of the moments of the distribution function is of paramount importance to describe correctly the long-time behaviour. The methods were designed using an L_2 best constrained approximation formalism in the space of orthogonal polynomials. Thanks to this setting, the resulting methods allow the spectral accuracy property of the solution to be maintained for the conservative polynomial series. The new approximation was then used to derive moment preserving methods for several Fokker-Planck equations in distinct physical domains, and by adopting different standard Galerkin projections, including the use of micro-macro decompositions.

Given its generality, the method admits numerous extensions. Among the most interesting are certainly the construction of spectrally accurate and conservative methods for the Vlasov equation. Recent works considered this question [20, 21, 50, 127], and we believe that our approach could provide a nice framework to prove essential properties such as convergence and stability of the discrete system. Another interesting direction is the application to kinetic models in the socio-economic domain where equilibrium states are often unknown and, thanks to the present approach, can be computed with spectral accuracy.

Moment-driven efficient numerical methods

Hybrid numerical methods

In these works, a new hybrid numerical method for linear kinetic equations in the diffusive scaling was presented. The method relies on two criteria motivated by a perturbative approach. The first one quantifies the distance from a local equilibrium. The second criterion depends on the macroscopic quantities that are available on the whole computing domain. We have managed to quantify the mass variation induced by the method, and we have shown that it is in practice very small. The method has proven to be efficient through various numerical experiments: the computational gain compared to a full kinetic scheme is significant. Moreover, the method performs

well with a non-homogeneous Knudsen number and when the coupling with Poisson's equation is considered which is encouraging to tackle more physically motivated problems.

In future works, more general contexts shall be considered. In particular, the multi-dimensional setting should require a smart implementation to ensure well-defined subdomains with respect to the stencil of the schemes used. Regarding the hydrodynamic scaling such an extension was considered in [88]. In the diffusive scaling, since high order derivatives appear in the Chapman-Enskog expansion (See Appendix A), the extension to general meshes may be challenging and might require a deeper analysis of the criterion to ensure an efficient numerical strategy.

Another aspect that is of great interest is a well planned parallelization of the method. More precisely, the load distribution over several cores of computation of the hybrid method is not straightforward. Indeed, since the domain decomposition is dynamic in time, one cannot evenly distribute space regions in advance to specific cores.

Finally, from a modeling point of view, a more general collision operator shall be investigated. We are confident that the computational gain will be even more worthwhile in a full $3D_x - 3D_v$ setting in the case of the Boltzmann operator which is known to be costly numerically.

Multiscale parareal numerical method

I presented an ongoing work on the design of a multiscale parareal method for kinetic equations. The goal of this approach is to reduce the cost of the time integration of a fully kinetic simulation using a parallel in time procedure. Leveraging the multiscale property of kinetic models, the cheap coarse propagator consists in a fluid solver and the fine but expensive propagation is achieved through a kinetic solver for a collisional Vlasov equation. We showed that the algorithm performs well in a $1D_x$-$3D_v$ hydrodynamic setting that is a computationally very demanding problem. Initial testing of the method showed promising results as the parareal solutions are able to capture the kinetic behaviour. While the computational speedup are not as good as one could expect, we still showed that the algorithm is able to converge. These encouraging findings shall now be refined in terms of parallel implementation and long-time simulations. Another idea would be to consider more sophisticated versions of the parareal algorithm constructed with further acceleration in mind. In addition, while computationally expensive, the kinetic BGK model is still a lot less costly than the full Boltzmann operator that we would like to consider in a later work.

In the future, the method will also be implemented in the diffusive scaling. Since the fluid model will then be diffusion-dominated, we expect the parareal algorithm will be more robust. Indeed, parareal methods can struggle to capture shocks. Moreover, the diffusive scaling may be more suitable for long time simulations. In addition, since diffusion helps the convergence of the algorithm, extension to a Navier-Stokes-like fluid solver will be investigated in the hydrodynamic setting, along with higher order moment lifting.

It is important to mention that this new multiscale parareal method only acts on the time integration. Its coupling with any type of kinetic or fluid solver, both in the phase space and for the time iterations inside the propagators, is therefore straightforward, as long as the projection and lifting operators are well-defined. Consequently, one could aim at developing higher order versions of this approach.

Appendix A

Higher order drift diffusion model in the 3D-3D setting

This appendix is dedicated to the extension to the full 3D/3D setting of the derivation of a higher order diffusive model, presented in a 1D/1D setting in Chapter 3 Section 3.2.

Let us recall the Vlasov-Poisson-BGK model where the unknown is the distribution $f(t,x,v)$:

$$\begin{cases} \partial_t f^\varepsilon + \frac{1}{\varepsilon}\mathsf{T} f = \frac{1}{\varepsilon^2}(\mathcal{M}_{\rho,0,1}(v) - f^\varepsilon), \quad (t,x,v) \in \mathbb{R}^+ \times \mathbb{R}^3 \times \mathbb{R}^3, \\ f^\varepsilon(0,x,v) = f_0(x,v), \end{cases} \tag{A.1}$$

where the transport operator is given by $\mathsf{T} f = v \cdot \nabla_x f + E \cdot \nabla_v f$, $E \in \mathbb{R}^3$. We also denote the macroscopic flux $J = \nabla_x \rho - E\rho$ and recall the property $\nabla_v \mathcal{M}_{\rho,0,1}(v) = -v\mathcal{M}_{\rho,0,1}(v)$.

Notations. In order to alleviate the computations, let us introduce some compact notations:

— Tensor product of two vectors: $(v \otimes v)_{ij} = v_i v_j$;

— Contraction of two matrices: $A \cdot^2 B = \sum_{ij} a_{ij} b_{ij}$;

— Contraction of two rank N tensors indexed by $i_1, \dots, i_N$:

$$A \cdot^{(N)} B = \sum_{i_1,\dots,i_N} a_{i_1,\dots,i_N} b_{i_1,\dots,i_N}.$$

To further fix the ideas, we recall that the gradient of rank $N-1$ tensor yields a tensor of rank N. Note also that any such defined contraction is symmetric : $A \cdot^{(N)} B = B \cdot^{(N)} A$.

Lemma 17. *Consider a rank $N-1$ tensor A whose components depend only on velocity and another rank $N-1$ tensor B whose components depend only on position. The transport operator T applied to the contraction $A \cdot^{(N-1)} B$ rewrites as*

$$\mathsf{T}(A \cdot^{(N-1)} B) = v \otimes A \cdot^{(N)} \nabla_x B + \nabla_v A \cdot^{(N)} E \otimes B.$$

Proof. For simplicity, let us consider the case $N = 3$ and 2 dimensions in x and v. Extension to any rank and dimensions follows the same ideas. We also focus only on the term $v \cdot \nabla_x (A \cdot^{(2)} B)$ since the transport in velocity can be treated in the same way. We denote by $a_{ij}(v) = a_{ij}$ and $b_{ij}(x) = b_{ij}$

the entries of the matrices A and B. On the one hand, we can now compute

$$\begin{aligned} v \cdot \nabla_x (A \cdot^{(2)} B) &= \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \cdot \begin{pmatrix} \partial_{x_1} \sum_{ij} a_{ij} b_{ij} \\ \partial_{x_2} \sum_{ij} a_{ij} b_{ij} \end{pmatrix} \\ &= \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \cdot \begin{pmatrix} \sum_{ij} b_{ij} \partial_{x_1} a_{ij} \\ \sum_{ij} b_{ij} \partial_{x_2} a_{ij} \end{pmatrix} \\ &= v_1 \left(a_{11} \partial_{x_1} b_{11} + a_{12} \partial_{x_1} b_{12} + a_{21} \partial_{x_1} b_{21} + a_{22} \partial_{x_1} b_{22} \right) \\ &+ v_2 \left(a_{11} \partial_{x_2} b_{11} + a_{12} \partial_{x_2} b_{12} + a_{21} \partial_{x_2} b_{21} + a_{22} \partial_{x_2} b_{22} \right). \end{aligned}$$

On the other hand, the rank 3 tensors $v \otimes A$ and $\nabla_x B$ are given by

$$(v \otimes A)_{i,j,k} = v_i a_{j,k}, \quad \text{and} \quad (\nabla_x B)_{i,j,k} = \partial_{x_i} b_{j,k}.$$

Consequently, the contraction $v \otimes A \cdot^{(3)} \nabla_x B$ then yields

$$\begin{aligned} v \otimes A \cdot^{(3)} \nabla_x B &= \sum_{i,j,k} v_i a_{j,k} \partial_{x_i} b_{j,k} \\ &= v_1 \partial_{x_1} (a_{11} b_{11} + a_{12} b_{12} + a_{21} b_{21} + a_{22} b_{22}) \\ &+ v_2 \partial_{x_2} (a_{11} b_{11} + a_{12} b_{12} + a_{21} b_{21} + a_{22} b_{22}), \end{aligned}$$

which up to factorization is exactly the result of the previous computation. □

For the sake of readability, the velocity dependent tensors shall be written as the first term of the contractions in the following computations.

Step 1: Chapman Enskog expansion. We recall the Chapman Enskog expansion of f around a Maxwellian equilibrium $\rho \mathcal{M}_{1,0,1}$:

$$f^\varepsilon(t,x,v) \approx \rho \mathcal{M}_{1,0,1}(v) + \sum_{k=1}^{K} \varepsilon^k h^{(k)}(t,x,v).$$

In the following we omit the moment subscripts and simply use $\mathcal{M}$ to denote the Maxwellian $\mathcal{M}_{1,0,1}$. We can now inject the expansion into the kinetic system (A.1) and identify the perturbations through the powers of ε:

$$k = 0: \qquad h^{(1)} = -\mathsf{T}(\rho^\varepsilon \mathcal{M}), \tag{A.2a}$$

$$k = 1: \qquad h^{(2)} = -\partial_t (\rho^\varepsilon \mathcal{M}) - \mathsf{T}(h^{(1)}), \tag{A.2b}$$

$$2 \le k \le K-1: \quad h^{(k+1)} = -\partial_t h^{(k-1)} - \mathsf{T}(h^{(k)}). \tag{A.2c}$$

Consequently, we obtain

$$h^{(1)} = -\mathsf{T}(\rho \mathcal{M}) = -v \mathcal{M} \cdot J,$$

and

$$\begin{aligned} h^{(2)} &= -\partial_t (\rho \mathcal{M}) - \mathsf{T}(h^{(1)}) \\ &= \mathcal{M} \partial_t \rho + v \otimes v \mathcal{M} \cdot^{(2)} \nabla_x J + (I - v \otimes v) \mathcal{M} \cdot^{(2)} E \otimes J, \end{aligned}$$

where I denotes the identity matrix and

$$\begin{aligned} h^{(3)} &= -\partial_t(h^{(1)}) - \mathsf{T}(h^{(2)}) \\ &= v\mathcal{M}\cdot\,\partial_t J + v\mathcal{M}\cdot\nabla_x(\partial_t\rho) - v\otimes v\otimes v\mathcal{M}\cdot^{(3)}\nabla_x^2 J - v\otimes(I-v\otimes v)\mathcal{M}\cdot^{(3)}\nabla_x(E\otimes J) \\ &\quad - v\mathcal{M}\cdot E\,\partial_t\rho - \nabla_v((v\otimes v)\mathcal{M})\cdot^{(3)} E\otimes\nabla_x J - \nabla_v((I-v\otimes v)\mathcal{M})\cdot^{(3)} E\otimes E\otimes J. \end{aligned}$$

Since the Maxwellian $\mathcal{M}_{\rho,0,1}$ has the same moments as the distribution f, a straightforward consequence is the zero mean of the perturbations $h^{(k)}$:

$$\int_{\mathbb{R}^3} h^{(k)}\,\mathrm{d}v = 0, \quad \forall k = 1,2,3.$$

Moreover, we point out that the perturbation depends on the velocity only through the Maxwellian $\mathcal{M}$ and its derivatives.

Step 2: Apply transport to the perturbations. The next step is to inject the definitions of $h^{(1)}$, $h^{(2)}$ and $h^{(3)}$ into (A.1) and then integrate the resulting expression in velocity:

$$\partial_t\rho + \varepsilon^{-1}\int_{\mathbb{R}^3}\mathsf{T}\mathcal{M}\,\mathrm{d}v + \int_{\mathbb{R}^3}\mathsf{T}h^{(1)}\,\mathrm{d}v + \varepsilon\int_{\mathbb{R}^3}\mathsf{T}h^{(2)}\,\mathrm{d}v + \varepsilon^2\int_{\mathbb{R}^3}\mathsf{T}h^{(3)}\,\mathrm{d}v = 0.$$

In particular, one needs to apply the transport operator T to the perturbations. However, one only needs to apply the transport in position $v\cdot\nabla_x$ since the velocity gradients vanish thanks to the integrability of $\mathcal{M}$ on $\mathbb{R}^3$:

$$\partial_t\rho + \varepsilon^{-1}\int_{\mathbb{R}^3} v\cdot\nabla_x\mathcal{M}\,\mathrm{d}v + \int_{\mathbb{R}^3} v\cdot\nabla_x h^{(1)}\,\mathrm{d}v + \varepsilon\int_{\mathbb{R}^3} v\cdot\nabla_x h^{(2)}\,\mathrm{d}v + \varepsilon^2\int_{\mathbb{R}^3} v\cdot\nabla_x h^{(3)}\,\mathrm{d}v = 0. \quad \text{(A.3)}$$

Therefore, one can compute

$$\begin{aligned} v\cdot\nabla_x h^{(1)} &= -v\otimes v\mathcal{M}\cdot^{(2)}\nabla_x J, \\ v\cdot\nabla_x h^{(2)} &= v\mathcal{M}\cdot\nabla_x(\partial_t\rho) + v\otimes v\otimes v\mathcal{M}\cdot^{(2)}\nabla_x^2 J + (v\otimes(I-v\otimes v))\mathcal{M}\cdot^{(3)}\nabla_x(E\otimes J), \\ v\cdot\nabla_x h^{(3)} &= v\otimes v\mathcal{M}\cdot^{(2)}\nabla_x(\partial_t J) + v\otimes v\mathcal{M}\cdot^{(2)}\nabla_x^2(\partial_t\rho) - v\otimes v\otimes v\otimes v\mathcal{M}\cdot^{(4)}\nabla_x^3 J \\ &\quad - v\otimes v\otimes(I-v\otimes v)\mathcal{M}\cdot^{(4)}\nabla_x^2(E\otimes J) \\ &\quad - v\otimes v\mathcal{M}\cdot^{(2)}\nabla_x(E\,\partial_t\rho) - v\otimes\nabla_v((v\otimes v)\mathcal{M})\cdot^{(4)}\nabla_x(E\otimes\nabla_x J) \\ &\quad - v\otimes\nabla_v((I-v\otimes v)\mathcal{M})\cdot^{(4)}\nabla_x(E\otimes E\otimes J). \end{aligned}$$

Step 3: Integration in velocity. The next step is to compute the integrals in (A.3). A first observation is that the quantities $v\cdot\nabla_x\mathcal{M}$ and $v\cdot\nabla_x h^{(2)}$ contain only odd in v terms that therefore vanish through integration:

$$\int_{\mathbb{R}^3} v\cdot\nabla_x\mathcal{M}\,\mathrm{d}v = \int_{\mathbb{R}^3} v\cdot\nabla_x h^{(2)}\,\mathrm{d}v = 0.$$

The remaining terms $v\cdot\nabla_x h^{(1)}$ and $v\cdot\nabla_x h^{(0)}$ require the computation of moments of the Maxwellian $\mathcal{M}$ and its derivative. A first computation yields

$$\int_{\mathbb{R}^3}\mathcal{M}\,\mathrm{d}v = I.$$

Then, using the parity of $\mathcal{M}$ in each direction one obtains

$$\int_{\mathbb{R}^3} v \otimes v \mathcal{M} \, \mathrm{d}\, v = m_2 I,$$

where m_k is the k^{th} moment of the 1-dimensional Gaussian. Similarly, one can compute the integral of the following rank 4 tensor:

$$(A_1)_{i_1,i_2,i_3,i_4} = \left(\int_{\mathbb{R}^3} v \otimes v \otimes v \otimes v \mathcal{M} \, \mathrm{d}\, v \right)_{i_1,i_2,i_3,i_4} = \begin{cases} m_4 & \text{if } i_1 = i_2 = i_3 = i_4, \\ m_2 & \text{if } i_\alpha = i_\beta, \, 1 \leq \alpha, \beta \leq 4, \\ 0 & \text{elsewhere.} \end{cases}$$

Using the property of the gradient of $\mathcal{M}$, $\nabla_v \mathcal{M} = -v\mathcal{M}$ and working component wise we can define the following integral:

$$(A_2)_{i_1,i_2,i_3,i_4} = \left(\int_{\mathbb{R}^3} v \otimes \nabla_v ((v \otimes v)\mathcal{M}) \, \mathrm{d}\, v \right)_{i_1,i_2,i_3,i_4} = \begin{cases} 2m_2 - m_4 & \text{if } i_1 = i_2 = i_3 = i_4, \\ m_2 - m_{2,2} & \text{if } i_\alpha = i_\beta, \, 1 \leq \alpha, \beta \leq 4, \\ 0 & \text{elsewhere,} \end{cases}$$

with $m_{2,2} = \int_{\mathbb{R}^3} v_\alpha^2 v_\beta^2 \, \mathrm{d}\, v$, $\alpha \neq \beta$. Similarly, we also compute

$$(A_3)_{i_1,i_2,i_3,i_4} = \left(\int_{\mathbb{R}^3} (v \otimes v \otimes I\mathcal{M}) \, \mathrm{d}\, v \right)_{i_1,i_2,i_3,i_4} = \begin{cases} m_2 & \text{if } (i_1 = i_2 \text{ or } i_3 = i_4) \text{ and } i_1 = i_3, \\ 0 & \text{elsewhere.} \end{cases} \tag{A.4}$$

Final step. Finally, one obtains the following higher order diffusive limit of the Vlasov-BGK equation in the 3D/3D setting:

$$\begin{aligned} \partial_t \rho - m_2 I \cdot^{(2)} \nabla_x J + \varepsilon^2 \Big[& m_2 I \cdot^{(2)} \nabla_x^2 \, \partial_t \rho - A_1 \cdot^{(4)} \nabla_x^3 J \\ & + m_2 I \cdot^{(2)} \nabla_x \, \partial_t J - (A_3 - A_1) \cdot^{(4)} \nabla_x^2 (E \otimes J) + m_2 I \cdot^{(2)} \nabla_x (E \, \partial_t \rho) \\ & - A_2 \cdot^{(4)} \nabla_x (E \otimes \nabla_x J) - (A_4 - A_2) \cdot^{(4)} \nabla_x (E \otimes E \otimes J) \Big] = 0. \end{aligned}$$

Remark 8.

— *As it was done in Chapter 3, one can also obtain the following approximation of the time derivative of ρ:*

$$\partial_t \rho = m_2 I \cdot^{(2)} \nabla_x J + \mathcal{O}(\varepsilon^2),$$

that allows to avoid a potential mixed derivative to be approximated numerically.

— *One can also check that equation (A.4) is indeed consistent with ($\widetilde{DD}$) from Chapter 3, obtained in the 1D/1D setting.*

Appendix B

MPI implementation of the multiscale parareal method

This appendix is dedicated to proposing an MPI implementation of Algorithm 4. We present the procedure to be implemented and discuss the challenges it poses.

Workload distribution. Let us consider the parallelization of Algorithm 4 over $N_p \in \mathbb{N}$ processors with distributed memory. We recall that we aim at computing the expensive kinetic propagations in parallel to reduce the simulation time. A first observation is that after each parareal iteration, the actual work to be done diminishes. Therefore, one must adjust the load distribution at each iteration. Let us now consider a discretization of the position variable using N_x points and N_g points for the coarse time discretization. From a practical point of view, the memory allocation of the jumps arrays (Δ^n in Algorithm 4) should be done locally on each processor, and only once, to avoid multiple allocation/deallocation overheads. Indeed, since they depend on the size of the spatial discretization, their memory footprint is potentially large. Therefore, the master processor shall be allocated with the full time discretization which is of size $5N_xN_g$, where 5 corresponds to the number of moments. The remaining processors are allocated their maximum workload, namely their workload at the first parareal iteration. This quantity, denoted by chunk in the following, can be defined using integer arithmetic to evenly distribute the time iterations. The workload at parareal iteration k, the starting and ending time indices for each processor are then computed as

$$
\begin{aligned}
&\texttt{work} = N_g - k;\\
&\texttt{chunk} = \left\lfloor \texttt{work}/N_p \right\rfloor;\\
&\texttt{remainder} = \mathrm{mod}(\texttt{work}, N_p);\\
&\texttt{start} = \texttt{rank} * \texttt{chunk} + \min(\texttt{remainder}, \texttt{rank}) + 1;\\
&\texttt{end} = (\texttt{rank}+1) * \texttt{chunk} + \min(\texttt{remainder}, \texttt{rank}+1);\\
&\texttt{myChunk} = \texttt{end} - \texttt{start}.
\end{aligned}
$$

At later iterations, when each process has less fine propagations to deal with, one can adjust using array indexing instead of dealing with multiple allocation/deallocation. Our strategy is presented in Algorithm 5 and we shall now discuss the communication aspect of our approach.

Communications. The main challenge to implement this strategy will be to efficiently send time-chunks of data that are of size $5N_x \times \texttt{myChunk}$ which are potentially large. It is however important to note that since the workload (in time) diminishes with parareal iterations, an empirical threshold for the number of processors to use should be considered to keep the ammount of communications

Algorithm 5 MPI implementation: multiscale kinetic parareal Algorithm 4.

Require: $U^{0,0}$

1: **for** $n = 1,\dots,N_g$ **do** ▷ First coarse guess
2: $U^{n,0} = \mathcal{G}\left(U^{n-1,0}\right)$
3: **end for**
4: **while** $k \leq K$ **or** error $\geq$ tol **do** ▷ Parareal iterations
5: work $= N_g - k$ ▷ Work distribution
6: chunk $= \left\lfloor \text{work}/N_p \right\rfloor$
7: remainder $= \text{mod}(\text{work}, N_p)$
8: start = rank * chunk + min(remainder, rank) + 1
9: end = (rank+1) * chunk + min (remainder, rank+1)
10: myChunk = end - start
11: **for** $n = start,\dots,end$ **do** ▷ Compute the jumps in parallel
12: $\Delta^n = \mathcal{PFL}(U^{n-1,k-1}) - \mathcal{G}\left(U^{n-1,k-1}\right)$
13: **end for**
14: Gather chunks of jumps Δ^n to master process
15: **if** rank **is** master **then**
16: **for** $n = k,\dots,N_g$ **do** ▷ Sequential correction
17: $U^{n,k+1} = \mathcal{G}\left(U^{n-1,k}\right) + \Delta^n$
18: **end for**
19: **end if**
20: Broadcast updated solution $U^{n,k+1}$;
21: Compute successive error on the moments and
22: $k = k + 1$
23: **end while**

in check. Our approach is to gather all the local-in-time jumps on the master processor (step 14) and update the solution. Then, the full updated solution of size $5N_xN_g$ is broadcast to each processor (step 20) to perform its work for the next parareal iteration. Note that this strategy may not be the most efficient depending on the size of the problem and computing architecture. For example, we refer to [198] where message passing cost and optimization is discussed.

Finally, another computational difficulty will be the actual storage of such data that needs to be MPI friendly in order to avoid the use of many buffer arrays and have an optimized browsing of the memory.

Bibliography

[1] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden. "The heterogeneous multiscale method". In: *Acta Numer.* 21 (2012), pp. 1–87. DOI: 10.1017/S0962492912000025.

[2] L. Addala, J. Dolbeault, X. Li, and M. L. Tayeb. "L^2-hypocoercivity and large time asymptotics of the linearized Vlasov-Poisson-Fokker-Planck system". In: *J. Stat. Phys.* 184.1 (2021), Paper No. 4, 34. ISSN: 0022-4715,1572-9613. DOI: 10.1007/s10955-021-02784-4.

[3] G. Allaire, X. Blanc, B. Despres, and F. Golse. "Transport et diffusion". Lecture notes, downloaded from http://paestel.fr/content/transport-et-diffusion-g-allaire-x-blanc-b-despres-f-golse in August 2021.

[4] R. J. Alonso, I. M. Gamba, and S. H. Tharkabhushanam. "Convergence and error estimates for the Lagrangian-based conservative spectral method for Boltzmann equations". English. In: *SIAM J. Numer. Anal.* 56.6 (2018), pp. 3534–3579. ISSN: 0036-1429. DOI: 10.1137/18M1173332.

[5] P. Andries, P. Le Tallec, J.-P. Perlat, and B. Perthame. "The Gaussian-BGK model of Boltzmann equation with small Prandtl number". English. In: *Eur. J. Mech., B, Fluids* 19.6 (2000), pp. 813–830. ISSN: 0997-7546. DOI: 10.1016/S0997-7546(00)01103-1.

[6] J. Angel, S. Götschel, and D. Ruprecht. "Impact of spatial coarsening on Parareal convergence". Preprint arXiv:2111.10228. 2021. arXiv: 2111.10228.

[7] N. Ayi, M. Herda, H. Hivert, and I. Tristani. "On a structure-preserving numerical method for fractional Fokker-Planck equations". English. In: *Math. Comput.* 92.340 (2023), pp. 635–693. ISSN: 0025-5718. DOI: 10.1090/mcom/3789.

[8] R. Bailo, J. A. Carrillo, and J. Hu. "Fully discrete positivity-preserving and energy-dissipating schemes for aggregation-diffusion equations with a gradient-flow structure". English. In: *Commun. Math. Sci.* 18.5 (2020), pp. 1259–1303. ISSN: 1539-6746. DOI: 10.4310/CMS.2020.v18.n5.a5.

[9] G. Bal. "On the convergence and the stability of the parareal algorithm to solve partial differential equations". English. In: *Domain decomposition methods in science and engineering. XV. Selected papers of the 15th international conference on domain decomposition, Berlin, Germany, July 21–25, 2003.* Berlin: Springer, 2005, pp. 425–432. ISBN: 3-540-22523-4. DOI: 10.1007/3-540-26825-1_43.

[10] C. Bardos, R. Santos, and R. Sentis. "Diffusion Approximation and Computation of the Critical Size". In: *Trans. Amer. Math. Soc.* (1984).

[11] C. Bardos and S. Ukai. "The classical incompressible Navier-Stokes limit of the Boltzmann equation". English. In: *Math. Models Methods Appl. Sci.* 1.2 (1991), pp. 235–257. ISSN: 0218-2025. DOI: 10.1142/S0218202591000137.

[12] C. W. Bardos, F. Golse, and D. M. Levermore. "Fluid dynamic limits of kinetic equations. I. Formal derivations". In: *J. Statist. Phys.* 63 (1991), pp. 323–344.

[13] A.-M. Baudron, J.-J. Lautard, Y. Maday, and O. Mula. "The Parareal in Time Algorithm Applied to the Kinetic Neutron Diffusion Equation". In: *Domain Decomposition Methods in Science and Engineering XXI*. pringer International Publishing, 2014, pp. 437–445. ISBN: 978-3-319-05789-7.

[14] N. Ben Abdallah and P. Degond. "On a hierarchy of macroscopic models for semiconductors". English. In: *J. Math. Phys.* 37.7 (1996), pp. 3306–3333. ISSN: 0022-2488. DOI: 10.1063/1.531567.

[15] N. Ben Abdallah, P. Degond, and S. Genieys. "An energy-transport model for semiconductors derived from the Boltzmann equation." English. In: *J. Stat. Phys.* 84.1-2 (1996), pp. 205–231. ISSN: 0022-4715. DOI: 10.1007/BF02179583.

[16] M. Bennoune, M. Lemou, and L. Mieussens. "Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier-Stokes asymptotics". In: *J. Comput. Phys.* 227 (2008).

[17] A. Bensoussan, J.-L. Lions, and G. C. Papanicolaou. "Boundary Layers and Homogenization of Transport Processes". In: *Publ. Res. I. Math. Sci.* (1979).

[18] M. Bessemoulin-Chatard, M. Herda, and T. Rey. "Hypocoercivity and diffusion limit of a finite volume scheme for linear kinetic equations". In: *Math. Comput.* 89.323 (2020), pp. 1093–1133. ISSN: 0025-5718,1088-6842. DOI: 10.1090/mcom/3490.

[19] M. Bessemoulin-Chatard and F. Filbet. "A Finite Volume Scheme for Nonlinear Degenerate Parabolic Equations". In: *SIAM J. Sci. Comput.* 34.5 (2012), B559–B583. DOI: 10.1137/110853807.

[20] M. Bessemoulin-Chatard and F. Filbet. "On the convergence of discontinuous Galerkin/Hermite spectral methods for the Vlasov-Poisson system". In: *SIAM J. Numer. Anal.* 61.4 (Aug. 2023), pp. 1664–1688. DOI: 10.1137/22M1518232.

[21] M. Bessemoulin-Chatard and F. Filbet. "On the stability of conservative discontinuous Galerkin/Hermite spectral methods for the Vlasov-Poisson system". In: *J. Comput. Phys.* 451 (2022), p. 110881. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2021.110881.

[22] M. Bessemoulin-Chatard, T. Laidin, and T. Rey. "Discrete hypocoercivity for a nonlinear kinetic reaction model". In: *IMA JNA* (Sept. 2024), drae058. ISSN: 0272-4979. DOI: 10.1093/imanum/drae058.

[23] P. L. Bhatnagar, E. P. Gross, and M. Krook. "A model for collision processes in gases. I: Small amplitude processes in charged and neutral one-component systems". English. In: *Phys. Rev., II. Ser.* 94 (1954), pp. 511–525. ISSN: 0031-899X. DOI: 10.1103/PhysRev.94.511.

[24] A. Blaustein and F. Filbet. "A structure and asymptotic preserving scheme for the Vlasov-Poisson-Fokker-Planck model". In: *J. Comput. Phys.* 498 (2024), p. 112693. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2023.112693.

[25] A. Blaustein and F. Filbet. "On a discrete framework of hypocoercivity for kinetic equations". In: *Math. Comput.* 93.345 (2024), pp. 163–202. ISSN: 0025-5718. DOI: 10.1090/mcom/3862.

[26] A. Bobylev and S. Rjasanow. "Difference scheme for the Boltzmann equation based on the fast Fourier transform". English. In: *Eur. J. Mech., B* 16.2 (1997), pp. 293–306. ISSN: 0997-7546.

[27] T. Bodineau, I. Gallagher, L. Saint-Raymond, and S. Simonella. "One-sided convergence in the Boltzmann-Grad limit". English. In: *Ann. Fac. Sci. Toulouse, Math. (6)* 27.5 (2018), pp. 985–1022. ISSN: 0240-2963. DOI: 10.5802/afst.1589.

[28] N. N. Bogolyubov. *Problems of a dynamical theory in statistical physics*. English. Stud. Stat. Mech. 1, 1-118 (1962). 1962.

[29] L. Boltzmann. "Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen". In: 66 (1872), pp. 275–370.

[30] M. Born and H. S. Green. "A general kinetic theory of liquids. I. The molecular distribution functions". English. In: *Proc. R. Soc. Lond., Ser. A* 188 (1946), pp. 10–18. ISSN: 0080-4630. DOI: 10.1098/rspa.1946.0093.

[31] I. Bossuyt, S. Vandewalle, and G. Samaey. "Monte-Carlo/Moments micro-macro Parareal method for unimodal and bimodal scalar McKean-Vlasov SDEs". Preprint arXiv:2310.11365. 2023. arXiv: 2310.11365.

[32] E. Bouin, J. Dolbeault, and L. Lafleche. "Fractional hypocoercivity". English. In: *Commun. Math. Phys.* 390.3 (2022), pp. 1369–1411. ISSN: 0010-3616. DOI: 10.1007/s00220-021-04296-4.

[33] E. Bouin, J. Dolbeault, S. Mischler, C. Mouhot, and C. Schmeiser. "Hypocoercivity without confinement". English. In: *Pure Appl. Anal.* 2.2 (2020), pp. 203–232. ISSN: 2578-5893. DOI: 10.2140/paa.2020.2.203.

[34] E. Bouin, J. Dolbeault, and L. Ziviani. "L^2 Hypocoercivity methods for kinetic Fokker-Planck equations with factorised Gibbs states". Preprint arXiv:2304.12040. 2023.

[35] N. V. Brilliantov and T. Pöschel. *Kinetic theory of granular gases.* English. Reprint of the 2004 hardback ed. Oxf. Grad. Texts. Oxford: Oxford University Press, 2010. ISBN: 978-0-19-958813-8.

[36] C. Buet. "A discrete-velocity scheme for the Boltzmann operator of rarefied gas dynamics". English. In: *Transp. Theory Stat. Phys.* 25.1 (1996), pp. 33–60. ISSN: 0041-1450. DOI: 10.1080/00411459608204829.

[37] C. Buet and S. Dellacherie. "On the Chang and Cooper scheme applied to a linear Fokker-Planck equation". English. In: *Commun. Math. Sci.* 8.4 (2010), pp. 1079–1090. ISSN: 1539-6746. DOI: 10.4310/CMS.2010.v8.n4.a15.

[38] M. Burger, J. A. Carrillo, and M.-T. Wolfram. "A mixed finite element method for nonlinear diffusion equations". In: *Kinet. Relat. Mod.* 3.1 (2010), pp. 59–83. ISSN: 1937-5093. DOI: 10.3934/krm.2010.3.59.

[39] R. E. Caflisch, G. Dimarco, and L. Pareschi. "An hybrid method for the Boltzmann equation". In: *AIP Conf. Proc.* 1786.1 (2016), p. 180001. DOI: 10.1063/1.4967670.

[40] V. Calvez, H. Hivert, and H. Yoldaş. "Concentration in Lotka-Volterra parabolic equations: an asymptotic-preserving scheme". English. In: *Numer. Math.* 154.1-2 (2023), pp. 103–153. ISSN: 0029-599X. DOI: 10.1007/s00211-023-01362-y.

[41] V. Calvez, G. Raoul, and C. Schmeiser. "Confinement by biased velocity jumps: Aggregation of escherichia coli". In: *Kinet. Relat. Mod.* 8.4 (2015), pp. 651–666. ISSN: 1937-5093. DOI: 10.3934/krm.2015.8.651.

[42] J. A. Cañizo, C. Cao, J. Evans, and H. Yoldaş. "Hypocoercivity of linear kinetic equations via Harris's Theorem". In: *Kinetic and Related Models* 13.1 (2020), pp. 97–128. ISSN: 1937-5093. DOI: 10.3934/krm.2020004.

[43] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods in fluid dynamics.* English. New York etc.: Springer-Verlag, 1988. ISBN: 3-540-17371-4.

[44] S. Carnot. "Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance". In: (1824), pp. 1–38.

[45] C. Cercignani. "On the Boltzmann equation for rigid spheres". English. In: *Transp. Theory Stat. Phys.* 2 (1972), pp. 211–225. ISSN: 0041-1450. DOI: 10.1080/00411457208232538.

[46] C. Chainais-Hillairet and F. Filbet. "Asymptotic behaviour of a finite-volume scheme for the transient drift-diffusion model". In: *IMA J. Numer. Anal.* 27.4 (Feb. 2007), pp. 689–716. ISSN: 0272-4979. DOI: 10.1093/imanum/drl045.

[47] C. Chainais-Hillairet and M. Herda. "Large-time behaviour of a family of finite volume schemes for boundary-driven convection-diffusion equations". In: *IMA J. Numer. Anal.* 40.4 (Oct. 2020), pp. 2473–2505. DOI: 10.1093/imanum/drz037.

[48] J. S. Chang and G. Cooper. "A practical difference scheme for Fokker-Planck equations". English. In: *J. Comput. Phys.* 6 (1970), pp. 1–16. ISSN: 0021-9991. DOI: 10.1016/0021-9991(70)90001-X.

[49] S. Chapman and T. G. Cowling. *The mathematical theory of non-uniform gases. An account of the kinetic theory of viscosity, thermal conduction, and diffusion in gases.* English. Cambridge: Cambridge University Press; New York: The Macmillan Company. xxiv, 404 p. (1939). 1939.

[50] 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), pp. 1–23.

[51] F. F. Chen. *Introduction to Plasma Physics*. English. Springer New York, NY, 2012. ISBN: 978-1-4757-0461-7. DOI: https://doi.org/10.1007/978-1-4757-0459-4.

[52] R. Clausius. "Ueber einen auf die Wärme anwendbaren mechanischen Satz". In: *Annalen der Physik* 217.9 (1870), pp. 124–130. DOI: https://doi.org/10.1002/andp.18702170911.

[53] F. Coron and B. Perthame. "Numerical passage from kinetic to fluid equations". English. In: *SIAM J. Numer. Anal.* 28.1 (1991), pp. 26–42. ISSN: 0036-1429. DOI: 10.1137/0728002.

[54] M. G. Crandall and A. Majda. "Monotone difference approximations for scalar conservation laws". English. In: *Math. Comput.* 34 (1980), pp. 1–21. ISSN: 0025-5718. DOI: 10.2307/2006218.

[55] A. Crestetto, N. Crouseilles, G. Dimarco, and M. Lemou. "Asymptotically complexity diminishing schemes (ACDS) for kinetic equations in the diffusive scaling". English. In: *J. Comput. Phys.* 394 (2019), pp. 243–262. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2019.05.032.

[56] A. Crestetto, N. Crouseilles, and M. Lemou. "Kinetic/fluid micro-macro numerical schemes for Vlasov-Poisson-BGK equation using particles". English. In: *Kinet. Relat. Models* 5.4 (2012), pp. 787–816. ISSN: 1937-5093. DOI: 10.3934/krm.2012.5.787.

[57] N. Crouseilles, P. Degond, and M. Lemou. "A hybrid kinetic/fluid model for solving the gas dynamics Boltzmann-BGK equation". English. In: *J. Comput. Phys.* 199.2 (2004), pp. 776–808. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2004.03.007.

[58] N. Crouseilles, G. Dimarco, and M.-H. Vignal. "Multiscale schemes for the BGK-Vlasov-Poisson system in the quasi-neutral and fluid limits. stability analysis and first order schemes". English. In: *Multiscale Model. Simul.* 14.1 (2016), pp. 65–95. ISSN: 1540-3459. DOI: 10.1137/140991558.

[59] N. Crouseilles and M. Lemou. "An asymptotic preserving scheme based on a micro-macro decomposition for collisional Vlasov equations: diffusion and high-field scaling limits." In: *Kinet. Relat. Mod.* 4.2 (2011), pp. 441–477. DOI: 10.3934/krm.2011.4.441.

[60] P. Degond, A. Nouri, and C. Schmeiser. "Macroscopic models for ionization in the presence of strong electric fields". In: *Transport Theory Statist. Phys.* 29.3-5 (2000), pp. 551–561. ISSN: 0041-1450,1532-2424. DOI: 10.1080/00411450008205891.

[61] P. Degond, G. Dimarco, and L. Mieussens. "A moving interface method for dynamic kinetic-fluid coupling". English. In: *J. Comput. Phys.* 227.2 (2007), pp. 1176–1208. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2007.08.027.

[62] P. Degond, T. Goudon, and F. Poupaud. "Diffusion limit for non homogeneous and non-micro-reversible processes". In: *Indiana Univ. Math. J.* (2000).

[63] P. Degond, S. Jin, and L. Mieussens. "A smooth transition model between kinetic and hydrodynamic equations". In: *J. Comput. Phys.* 209 (2005), pp. 665–694.

[64] P. Degond, C. D. Levermore, and C. Schmeiser. "A Note on the Energy-Transport Limit of the Semiconductor Boltzmann Equation". In: *Transport in Transition Regimes*. Ed. by N. B. Abdallah, I. M. Gamba, C. Ringhofer, A. Arnold, R. T. Glassey, P. Degond, and C. D. Levermore. New York, NY: Springer New York, 2004, pp. 137–153. ISBN: 978-1-4613-0017-5.

[65] P. Degond, J.-G. Liu, and L. Mieussens. "Macroscopic fluid models with localized kinetic upscaling effects". English. In: *Multiscale Model. Simul.* 5.3 (2006), pp. 940–979. ISSN: 1540-3459. DOI: 10.1137/060651574.

[66] H. Dietert, F. Hérau, H. Hutridurga, and C. Mouhot. "Trajectorial hypocoercivity and application to control theory". English. In: *Sémin. Laurent Schwartz, EDP Appl.* 2021-2022 (2022), ex. ISSN: 2266-0607. DOI: 10.5802/slsedp.156.

[67] G. Dimarco, Q. Li, L. Pareschi, and B. Yan. "Numerical methods for plasma physics in collisional regimes". In: *J. Plasma Phys.* 81.1 (2015), p. 305810106. DOI: 10.1017/S0022377814000762.

[68] G. Dimarco, L. Pareschi, and G. Samaey. "Asymptotic-preserving Monte Carlo methods for transport equations in the diffusive limit". English. In: *SIAM J. Sci. Comput.* 40.1 (2018), a504–a528. ISSN: 1064-8275. DOI: 10.1137/17M1140741.

[69] G. Dimarco, L. Mieussens, and V. Rispoli. "An asymptotic preserving automatic domain decomposition method for the Vlasov-Poisson-BGK system with applications to plasmas". English. In: *J. Comput. Phys.* 274 (2014), pp. 122–139. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2014.06.002.

[70] G. Dimarco and L. Pareschi. "Hybrid multiscale methods. II: Kinetic equations". English. In: *Multiscale Model. Simul.* 6.4 (2008), pp. 1169–1197. ISSN: 1540-3459. DOI: 10.1137/070680916.

[71] G. Dimarco and L. Pareschi. "Implicit-explicit linear multistep methods for stiff kinetic equations". English. In: *SIAM J. Numer. Anal.* 55.2 (2017), pp. 664–690. ISSN: 0036-1429. DOI: 10.1137/16M1063824.

[72] G. Dimarco and L. Pareschi. "Numerical methods for kinetic equations". In: *Acta Numer.* 23 (2014), pp. 369–520.

[73] J. Dolbeault, C. Mouhot, and C. Schmeiser. "Hypocoercivity for linear kinetic equations conserving mass". English. In: *Trans. Am. Math. Soc.* 367.6 (2015), pp. 3807–3828. ISSN: 0002-9947. DOI: 10.1090/S0002-9947-2015-06012-7.

[74] Z. Dong, E. H. Georgoulis, and P. J. Herbert. "A hypocoercivity-exploiting stabilised finite element method for Kolmogorov equation". Preprint arXiv:2401.12921. 2024. eprint: 2401.12921.

[75] M. Duarte, M. Massot, and S. Descombes. "Parareal operator splitting techniques for multi-scale reaction waves: numerical analysis and strategies". English. In: *ESAIM, Math. Model. Numer. Anal.* 45.5 (2011), pp. 825–852. ISSN: 0764-583X. DOI: 10.1051/m2an/2010104.

[76] G. Dujardin, F. Hérau, and P. Lafitte-Godillon. "Coercivity, hypocoercivity, exponential time decay and simulations for discrete Fokker- Planck equations". In: *Numer. Math.* 144 (2018). DOI: 10.1007/s00211-019-01094-y.

[77] W. E. *Principles of multiscale modeling*. English. Cambridge: Cambridge University Press, 2011. ISBN: 978-1-107-09654-7.

[78] W. E and B. Engquist. "The heterogeneous multiscale methods". English. In: *Commun. Math. Sci.* 1.1 (2003), pp. 87–132. ISSN: 1539-6746. DOI: 10.4310/CMS.2003.v1.n1.a8.

[79] A. Eghbal, A. G. Gerber, and E. Aubanel. "Acceleration of unsteady hydrodynamic simulations using the parareal algorithm". In: *J. Comput. Sci.* 19 (2017), pp. 57–76. ISSN: 1877-7503. DOI: https://doi.org/10.1016/j.jocs.2016.12.006.

[80] L. Einkemmer, J. Hu, and Y. Wang. "An asymptotic-preserving dynamical low-rank method for the multi-scale multi-dimensional linear transport equation". English. In: *J. Comput. Phys.* 439 (2021). Id/No 110353, p. 21. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2021.110353.

[81] J. Evans and H. Yoldaş. "Trend to Equilibrium for Run and Tumble Equations with Non-uniform Tumbling Kernels". In: *Acta Applicandae Mathematicae* 191.1 (2024), p. 6. ISSN: 1572-9036. DOI: 10.1007/s10440-024-00657-y.

[82] J. A. Evans and H. Yoldaş. "On the Asymptotic Behavior of a Run and Tumble Equation for Bacterial Chemotaxis". In: *SIAM J. Math. Anal.* 55.6 (2023), pp. 7635–7664. DOI: 10.1137/22M1539332.

[83] C. Farhat and M. Chandesris. "Time-decomposed parallel time-integrators: Theory and feasibility studies for fluid, structure, and fluid-structure applications". English. In: *Int. J. Numer. Methods Eng.* 58.9 (2003), pp. 1397–1434. ISSN: 0029-5981. DOI: 10.1002/nme.860.

[84] F. Filbet and M. Herda. "A finite volume scheme for boundary-driven convection-diffusion equations with relative entropy structure". English. In: *Numer. Math.* 137.3 (2017), pp. 535–577. ISSN: 0029-599X. DOI: 10.1007/s00211-017-0885-7.

[85] F. Filbet and S. Jin. "An asymptotic preserving scheme for the ES-BGK model of the Boltzmann equation". English. In: *J. Sci. Comput.* 46.2 (2011), pp. 204–224. ISSN: 0885-7474. DOI: 10.1007/s10915-010-9394-x.

[86] F. Filbet, L. Pareschi, and T. Rey. "On steady-state preserving spectral methods for homogeneous Boltzmann equations". In: *C. R. Acad. Sci. Paris, Ser. I* 353.4 (2015), pp. 309–314. DOI: 10.1016/j.crma.2015.01.015.

[87] F. Filbet and T. Rey. "A hierarchy of hybrid numerical methods for multiscale kinetic equations". English. In: *SIAM J. Sci. Comput.* 37.3 (2015), a1218–a1247. ISSN: 1064-8275. DOI: 10.1137/140958773.

[88] F. Filbet and T. Xiong. "A hybrid discontinuous Galerkin scheme for multi-scale kinetic equations". English. In: *J. Comput. Phys.* 372 (2018), pp. 841–863. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2018.06.064.

[89] P. F. Fischer, F. Hecht, and Y. Maday. "A parareal in time semi-implicit approximation of the Navier-Stokes equations". English. In: *Domain decomposition methods in science and engineering. XV. Selected papers of the 15th international conference on domain decomposition, Berlin, Germany, July 21–25, 2003.* Berlin: Springer, 2005, pp. 433–440. ISBN: 3-540-22523-4. DOI: 10.1007/3-540-26825-1_44.

[90] J. C. M. Fok, B. Guo, and T. Tang. "Combined Hermite spectral-finite difference method for the Fokker-Planck equation". English. In: *Math. Comput.* 71.240 (2002), pp. 1497–1528. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-01-01365-5.

[91] D. Funaro. *Polynomial approximation of differential equations*. English. Vol. 8. Lect. Notes Phys., New Ser. m, Monogr. Berlin: Springer-Verlag, 1992. ISBN: 3-540-55230-8.

[92] D. Funaro and G. Manzini. "Stability and conservation properties of Hermite-based approximations of the Vlasov-Poisson system". English. In: *J. Sci. Comput.* 88.1 (2021). Id/No 29, p. 36. ISSN: 0885-7474. DOI: 10.1007/s10915-021-01537-5.

[93] G. Furioli, A. Pulvirenti, E. Terraneo, and G. Toscani. "Fokker-Planck equations in the modeling of socio-economic phenomena". English. In: *Math. Models Methods Appl. Sci.* 27.1 (2017), pp. 115–158. ISSN: 0218-2025. DOI: 10.1142/S0218202517400048.

[94] I. Gallagher, L. Saint-Raymond, and B. Texier. *From Newton to Boltzmann: hard spheres and short-range potentials*. English. Zur. Lect. Adv. Math. Zürich: European Mathematical Society (EMS), 2013. ISBN: 978-3-03719-129-3. DOI: 10.4171/129.

[95] I. Gallagher and I. Tristani. "On the convergence of smooth solutions from Boltzmann to Navier-Stokes". English. In: *Ann. Henri Lebesgue* 3 (2020), pp. 561–614. ISSN: 2644-9463. DOI: 10.5802/ahl.40.

[96] I. M. Gamba, J. R. Haack, C. D. Hauck, and J. Hu. "A fast spectral method for the Boltzmann collision operator with general collision kernels". English. In: *SIAM J. Sci. Comput.* 39.4 (2017), b658–b674. ISSN: 1064-8275. DOI: 10.1137/16M1096001.

[97] I. M. Gamba and S. H. Tharkabhushanam. "Shock and boundary structure formation by spectral-Lagrangian methods for the inhomogeneous Boltzmann transport equation". English. In: *J. Comput. Math.* 28.4 (2010), pp. 430–460. ISSN: 0254-9409. DOI: 10.4208/jcm.1003-m0011.

[98] M. J. Gander. "Analysis of the parareal algorithm applied to hyperbolic problems using characteristics". English. In: *Bol. Soc. Esp. Mat. Apl., SẽMA* 42 (2008), pp. 21–35. ISSN: 1575-9822.

[99] M. J. Gander and E. Hairer. "Analysis for parareal algorithms applied to Hamiltonian differential equations". English. In: *J. Comput. Appl. Math.* 259 (2014), pp. 2–13. ISSN: 0377-0427. DOI: 10.1016/j.cam.2013.01.011.

[100] M. J. Gander and E. Hairer. "Nonlinear convergence analysis for the parareal algorithm". English. In: *Domain decomposition methods in science and engineering XVII. Selected papers based on the presentations at the 17th international conference on domain decomposition methods, St. Wolfgang/Strobl, Austria, July 3–7, 2006*. Berlin: Springer, 2008, pp. 45–56. ISBN: 978-3-540-75198-4. DOI: 10.1007/978-3-540-75199-1_4.

[101] M. J. Gander and S. Vandewalle. "Analysis of the parareal time-parallel time-integration method". English. In: *SIAM J. Sci. Comput.* 29.2 (2007), pp. 556–578. ISSN: 1064-8275. DOI: 10.1137/05064607X.

[102] W. Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Oxford University Press, Apr. 2004. ISBN: 9780198506720. DOI: 10.1093/oso/9780198506720.001.0001.

[103] E. H. Georgoulis. "Hypocoercivity-compatible Finite Element Methods for the Long-time Computation of Kolmogorov's Equation". In: *SIAM J. Numer. Anal.* 59 (2018), pp. 173–194.

[104] F. Golse. "The mean-field limit for the dynamics of large particle systems". English. In: *Journées "Équations aux dérivées partielles", Forges-les-Eaux, France, 2 au 6 juin 2003. Exposés Nos. I-XV*. Nantes: Université de Nantes, 2003, ex. ISBN: 2-86939-207-9.

[105] F. Golse and C. D. Levermore. "The Stokes-Fourier limit for the Boltzmann equation". English. In: *C. R. Acad. Sci., Paris, Sér. I, Math.* 333.2 (2001), pp. 145–150. ISSN: 0764-4442. DOI: 10.1016/S0764-4442(01)01969-3.

[106] F. Golse, C. Mouhot, and V. Ricci. "Empirical measures and Vlasov hierarchies". English. In: *Kinet. Relat. Models* 6.4 (2013), pp. 919–943. ISSN: 1937-5093. DOI: 10.3934/krm.2013.6.919.

[107] F. Golse and L. Saint-Raymond. "The incompressible Navier-Stokes limit of the Boltzmann equation for hard cutoff potentials". English. In: *J. Math. Pures Appl. (9)* 91.5 (2009), pp. 508–552. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2009.01.013.

[108] F. Golse and L. Saint-Raymond. "The Navier-Stokes limit for the Boltzmann equation". English. In: *C. R. Acad. Sci., Paris, Sér. I, Math.* 333.9 (2001), pp. 897–902. ISSN: 0764-4442. DOI: 10.1016/S0764-4442(01)02136-X.

[109] F. Golse and L. Saint-Raymond. "The Navier-Stokes limit of the Boltzmann equation for bounded collision kernels". English. In: *Invent. Math.* 155.1 (2004), pp. 81–161. ISSN: 0020-9910. DOI: 10.1007/s00222-003-0316-5.

[110] F. Golse and L. Saint-Raymond. "Velocity averaging in L^1 for the transport equation". English. In: *C. R., Math., Acad. Sci. Paris* 334.7 (2002), pp. 557–562. ISSN: 1631-073X. DOI: 10.1016/S1631-073X(02)02302-6.

[111] L. Gosse. *Computing qualitatively correct approximations of balance laws. Exponential-fit, well-balanced and asymptotic-preserving*. English. Vol. 2. SIMAI Springer Ser. Milano: Springer, 2013. ISBN: 978-88-470-2891-3; 978-88-470-2892-0. DOI: 10.1007/978-88-470-2892-0.

[112] L. Gosse and G. Toscani. "Identification of Asymptotic Decay to Self-Similarity for One-Dimensional Filtration Equations". In: *SIAM J. Numer. Anal.* 43.6 (2006), pp. 2590–2606. DOI: 10.1137/040608672.

[113] D. Gottlieb and S. A. Orszag. *Numerical analysis of spectral methods: Theory and applications*. English. Vol. 26. CBMS-NSF Reg. Conf. Ser. Appl. Math. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1977.

[114] T. Goudon and F. Poupaud. "Approximation by Homogenization and Diffusion of Kinetic Equations". In: *Comm. Partial Differential Equations* (2001).

[115] H. Grad. "On the kinetic theory of rarefied gases". English. In: *Commun. Pure Appl. Math.* 2 (1949), pp. 331–407. ISSN: 0010-3640. DOI: 10.1002/cpa.3160020403.

[116] L. Grigori, S. A. Hirstoaga, V.-T. Nguyen, and J. Salomon. "Reduced model-based parareal simulations of oscillatory singularly perturbed ordinary differential equations". English. In: *J. Comput. Phys.* 436 (2021). Id/No 110282, p. 18. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2021.110282.

[117] L. Grigori, S. A. Hirstoaga, and J. Salomon. "A parareal algorithm for a highly oscillating Vlasov-Poisson system with reduced models for the coarse solving". English. In: *Comput. Math. Appl.* 130 (2023), pp. 137–148. ISSN: 0898-1221. DOI: 10.1016/j.camwa.2022.12.004.

[118] S. Gualandi and G. Toscani. "Call center service times are lognormal: a Fokker-Planck description". English. In: *Math. Models Methods Appl. Sci.* 28.8 (2018), pp. 1513–1527. ISSN: 0218-2025. DOI: 10.1142/S0218202518500410.

[119] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*. English. Vol. 31. Springer Ser. Comput. Math. Berlin: Springer, 2002. ISBN: 3-540-43003-2.

[120] F. Hérau. "Introduction to hypocoercive methods and applications for simple linear inhomogeneous kinetic models". In: *Lectures on the analysis of nonlinear partial differential equations. Part 5*. Somerville, MA: International Press; Beijing: Higher Education Press, 2018, pp. 119–147. ISBN: 978-1-57146-357-9.

[121] F. Hérau and F. Nier. "Isotropic hypoellipticity and trend to the equilibrium for the Fokker-Planck equation with high degree potential". In: *Arch. Rational Mech. Anal.* 171.2 (2004), pp. 151–218. DOI: 10.1007/s00205-003-0276-3.

[122] F. Hérau and L. Thomann. "On global existence and trend to the equilibrium for the Vlasov-Poisson-Fokker-Planck system with exterior confining potential". English. In: *J. Funct. Anal.* 271.5 (2016), pp. 1301–1340. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2016.04.030.

[123] M. Herda and L. M. Rodrigues. "Large-time behavior of solutions to Vlasov-Poisson-Fokker-Planck equations: from evanescent collisions to diffusive limit". English. In: *J. Stat. Phys.* 170.5 (2018), pp. 895–931. ISSN: 0022-4715. DOI: 10.1007/s10955-018-1963-7.

[124] H. Hivert. "A first-order asymptotic preserving scheme for front propagation in a one-dimensional kinetic reaction–transport equation". In: *J. Comput. Phys.* 367 (2018), pp. 253–278. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2018.04.036.

[125] H. Hivert. "Numerical schemes for kinetic equation with diffusion limit and anomalous time scale". English. In: *Kinet. Relat. Models* 11.2 (2018), pp. 409–439. ISSN: 1937-5093. DOI: 10.3934/krm.2018019.

[126] M. Hochbruck and A. Ostermann. "Exponential integrators". English. In: *Acta Numer.* 19 (2010), pp. 209–286. ISSN: 0962-4929. DOI: 10.1017/S0962492910000048.

[127] J. P. Holloway. "Spectral velocity discretizations for the Vlasov-Maxwell equations". English. In: *Transp. Theory Stat. Phys.* 25.1 (1996), pp. 1–32. ISSN: 0041-1450. DOI: 10.1080/00411459608204828.

[128] J. Holway Lowell H. "New Statistical Models for Kinetic Theory: Methods of Construction". In: *The Physics of Fluids* 9.9 (Sept. 1966), pp. 1658–1673. ISSN: 0031-9171. DOI: 10.1063/1.1761920.

[129] L. Hörmander. "Hypoelliptic second order differential equations". In: *Acta Math.* 119 (1967), pp. 147–171. DOI: 10.1007/BF02392081.

[130] N. Horsten, G. Samaey, and M. Baelmans. "Hybrid fluid-kinetic model for neutral particles in the plasma edge". In: *Nuclear Materials and Energy* 18 (2019), pp. 201–207. ISSN: 2352-1791. DOI: https://doi.org/10.1016/j.nme.2018.12.018. URL: https://www.sciencedirect.com/science/article/pii/S2352179118301194.

[131] N. Horsten, G. Samaey, and M. Baelmans. "A hybrid fluid-kinetic model for hydrogenic atoms in the plasma edge of tokamaks based on a micro-macro decomposition of the kinetic equation". English. In: *J. Comput. Phys.* 409 (2020). Id/No 109308, p. 23. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2020.109308. URL: lirias.kuleuven.be/handle/20.500.12942/694332.

[132] J. Hu and Z. Ma. "A fast spectral method for the inelastic Boltzmann collision operator and application to heated granular gases". English. In: *J. Comput. Phys.* 385 (2019), pp. 119–134. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2019.01.049.

[133] P.-E. Jabin. "A review of the mean field limits for Vlasov equations". English. In: *Kinet. Relat. Models* 7.4 (2014), pp. 661–711. ISSN: 1937-5093. DOI: 10.3934/krm.2014.7.661.

[134] P.-E. Jabin and Z. Wang. "Quantitative estimates of propagation of chaos for stochastic systems with $W^{-1,\infty}$ kernels". English. In: *Invent. Math.* 214.1 (2018), pp. 523–591. ISSN: 0020-9910. DOI: 10.1007/s00222-018-0808-y.

[135] S. Jin. "Asymptotic-preserving schemes for multiscale physical problems". English. In: *Acta Numer.* 31 (2022), pp. 415–489. ISSN: 0962-4929. DOI: 10.1017/S0962492922000010.

[136] S. Jin. "Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations". English. In: *SIAM J. Sci. Comput.* 21.2 (1999), pp. 441–454. ISSN: 1064-8275. DOI: 10.1137/S1064827598334599.

[137] A. Jüngel. *Transport Equations for Semiconductors*. en. Vol. 773. Lecture Notes in Physics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-89525-1 978-3-540-89526-8. DOI: 10.1007/978-3-540-89526-8.

[138] M. Kac. "Foundations of kinetic theory". English. In: *Proc. 3rd Berkeley Sympos. Math. Statist. Probability 3, 171-197 (1956).* 1956.

[139] J. G. Kirkwood. "The Statistical Mechanical Theory of Transport Processes I. General Theory". In: *J. Chem. Phys.* 14.3 (Mar. 1946), pp. 180–201. ISSN: 0021-9606. DOI: 10.1063/1.1724117.

[140] A. Klar and C. Schmeiser. "Numerical passage from radiative heat transfer to nonlinear diffusion models". English. In: *Math. Models Methods Appl. Sci.* 11.5 (2001), pp. 749–767. ISSN: 0218-2025. DOI: 10.1142/S0218202501001082.

[141] A. Klar. "A numerical method for kinetic semiconductor equations in the drift-diffusion limit". English. In: *SIAM J. Sci. Comput.* 20.5 (1999), pp. 1696–1712. ISSN: 1064-8275. DOI: 10.1137/S1064827597319258.

[142] A. Klar. "Asymptotic-induced domain decomposition methods for kinetic and drift diffusion semiconductor equations". English. In: *SIAM J. Sci. Comput.* 19.6 (1998), pp. 2032–2050. ISSN: 1064-8275. DOI: 10.1137/S1064827595286177.

[143] O. Koch and C. Lubich. "Dynamical Low-Rank Approximation". In: *SIAM Journal on Matrix Analysis and Applications* 29.2 (2007), pp. 434–454. DOI: 10.1137/050639703.

[144] R. Koekoek, P. A. Lesky, and R. F. Swarttouw. *Hypergeometric orthogonal polynomials and their q-analogues. With a foreword by Tom H. Koornwinder*. English. Springer Monogr. Math. Berlin: Springer, 2010. ISBN: 978-3-642-05013-8; 978-3-642-26351-4; 978-3-642-05014-5. DOI: 10.1007/978-3-642-05014-5.

[145] V. I. Kolobov, R. R. Arslanbekov, V. V. Aristov, A. A. Frolova, and S. A. Zabelok. "Unified solver for rarefied and continuum flows with adaptive mesh and algorithm refinement". English. In: *J. Comput. Phys.* 223.2 (2007), pp. 589–608. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2006.09.021.

[146] K. Kormann and A. Yurova. "A generalized Fourier-Hermite method for the Vlasov-Poisson system". English. In: *BIT* 61.3 (2021), pp. 881–909. ISSN: 0006-3835. DOI: 10.1007/s10543-021-00853-4.

[147] M. Kraus and E. Hirvijoki. "Metriplectic integrators for the Landau collision operator". In: *Phys. Plasmas* 24.10 (Oct. 2017), p. 102311. ISSN: 1070-664X. DOI: 10.1063/1.4998610.

[148] T. Laidin. "Hybrid kinetic/fluid numerical method for the Vlasov-BGK equation in the diffusive scaling". In: *Kinet. Relat. Mod.* (2023). ISSN: 1937-5093. DOI: 10.3934/krm.2023013.

[149] T. Laidin and L. Pareschi. "Conservative polynomial approximations and applications to Fokker-Planck equations". Preprint arXiv:2402.06473. 2024.

[150] T. Laidin and T. Rey. "Hybrid Kinetic/Fluid numerical method for the Vlasov-Poisson-BGK equation in the diffusive scaling". In: *FVCA 10 - 2023 - International Conference on Finite Volumes for Complex Applications X*. Springer Proceedings in Mathematics & Statistics. 8 pages, 4 figures. Strasbourg, France, Oct. 2023, pp. 229–237. DOI: 10.1007/978-3-031-40860-1_24.

[151] O. E. I. Lanford. *Time evolution of large classical systems*. English. Dyn. Syst., Theor. Appl., Battelle Seattle 1974 Renc., Lect. Notes Phys. 38, 1-111 (1975). 1975. DOI: 10.1007/3-540-07171-7_1.

[152] E. W. Larsen, C. D. Levermore, G. C. Pomraning, and J. G. Sanderson. "Discretization methods for one-dimensional Fokker-Planck operators". English. In: *J. Comput. Phys.* 61 (1985), pp. 359–390. ISSN: 0021-9991. DOI: 10.1016/0021-9991(85)90070-1.

[153] S. Le Bourdiec, F. de Vuyst, and L. Jacquet. "Numerical solution of the Vlasov–Poisson system using generalized Hermite functions". In: *Comput. Phys. Commun.* 175.8 (2006), pp. 528–544. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2006.07.004.

[154] F. Legoll, T. Lelièvre, and G. Samaey. "A micro-macro parareal algorithm: application to singularly perturbed ordinary differential equations". English. In: *SIAM J. Sci. Comput.* 35.4 (2013), a1951–a1986. ISSN: 1064-8275. DOI: 10.1137/120872681.

[155] M. Lemou. "Relaxed micro-macro schemes for kinetic equations". English. In: *C. R., Math., Acad. Sci. Paris* 348.7-8 (2010), pp. 455–460. ISSN: 1631-073X. DOI: 10.1016/j.crma.2010.02.017.

[156] M. Lemou and L. Mieussens. "A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit". In: *SIAM J. Sci. Comput.* 31.1 (2008), pp. 334–368. DOI: 10.1137/07069479X.

[157] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. English. Camb. Texts Appl. Math. Cambridge: Cambridge University Press, 2002. ISBN: 0-521-00924-3; 0-521-81087-6. DOI: 10.1017/CBO9780511791253.

[158] C. D. Levermore, W. J. Morokoff, and B. T. Nadiga. "Moment realizability and the validity of the Navier–Stokes equations for rarefied gas dynamics". In: *Physics of Fluids* 10 (1998), pp. 3214–3226.

[159] J.-L. Lions, Y. Maday, and G. Turinici. "A "parareal" in time discretization of PDE's". French. In: *C. R. Acad. Sci., Paris, Sér. I, Math.* 332.7 (2001), pp. 661–668. ISSN: 0764-4442. DOI: 10.1016/S0764-4442(00)01793-6.

[160] P.-L. Lions and N. Masmoudi. "A local approach to the incompressible limit". French. In: *C. R. Acad. Sci., Paris, Sér. I, Math.* 329.5 (1999), pp. 387–392. ISSN: 0764-4442. DOI: 10.1016/S0764-4442(00)88611-5.

[161] J.-G. Liu and L. Mieussens. "Analysis of an asymptotic preserving scheme for linear kinetic equations in the diffusion limit". In: *SIAM J. Numer. Anal.* 48.4 (2010), pp. 1474–1491. DOI: 10.1137/090772770.

[162] Y. Maday and O. Mula. "An adaptive parareal algorithm". English. In: *J. Comput. Appl. Math.* 377 (2020). Id/No 112915, p. 17. ISSN: 0377-0427. DOI: 10.1016/j.cam.2020.112915.

[163] Y. Maday and G. Turinici. "A parareal in time procedure for the control of partial differential equations". English. In: *C. R., Math., Acad. Sci. Paris* 335.4 (2002), pp. 387–392. ISSN: 1631-073X. DOI: 10.1016/S1631-073X(02)02467-6.

[164] G. Manzini, D. Funaro, and G. L. Delzanno. "Convergence of spectral discretizations of the Vlasov-Poisson system". English. In: *SIAM J. Numer. Anal.* 55.5 (2017), pp. 2312–2335. ISSN: 0036-1429. DOI: 10.1137/16M1076848.

[165] H. Mathis, C. Cancès, E. Godlewski, and N. Seguin. "Dynamic model adaptation for multiscale simulation of hyperbolic systems with relaxation". English. In: *J. Sci. Comput.* 63.3 (2015), pp. 820–861. ISSN: 0885-7474. DOI: 10.1007/s10915-014-9915-0.

[166] J. C. Maxwell. "IV. On the dynamical theory of gases". In: *Philosophical Transactions of the Royal Society of London* 157 (1867), pp. 49–88. DOI: 10.1098/rstl.1867.0004.

[167] P. Michel and J. Schneider. "Simultaneous approximation of real numbers by rational numbers and its application to the Boltzmann equation". French. In: *C. R. Acad. Sci., Paris, Sér. I, Math.* 330.9 (2000), pp. 857–862. ISSN: 0764-4442. DOI: 10.1016/S0764-4442(00)00258-5.

[168] M. Mohammadi and A. Borzì. "Analysis of the Chang-Cooper discretization scheme for a class of Fokker-Planck equations". English. In: *J. Numer. Math.* 23.3 (2015), pp. 271–288. ISSN: 1570-2820. DOI: 10.1515/jnma-2015-0018.

[169] C. Mouhot and L. Pareschi. "Fast algorithms for computing the Boltzmann collision operator". English. In: *Math. Comput.* 75.256 (2006), pp. 1833–1852. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-06-01874-6.

[170] C. Mouhot, L. Pareschi, and T. Rey. "Convolutive decomposition and fast summation methods for discrete-velocity approximations of the Boltzmann equation". English. In: *ESAIM, Math. Model. Numer. Anal.* 47.5 (2013), pp. 1515–1531. ISSN: 0764-583X. DOI: 10.1051/m2an/2013078.

[171] L. Neumann and C. Schmeiser. "A kinetic reaction model: Decay to equilibrium and macroscopic limit". In: *Kinet. Relat. Mod.* 9.3 (2016), pp. 571–585. ISSN: 1937-5093. DOI: 10.3934/krm.2016007.

[172] V.-T. Nguyen. "Acceleration techniques of the Parareal algorithm for solving some differential equations". Theses. Sorbonne Université, May 2022.

[173] A. S. Nielsen, G. Brunner, and J. S. Hesthaven. "Communication-aware adaptive parareal with application to a nonlinear hyperbolic system of partial differential equations". English. In: *J. Comput. Phys.* 371 (2018), pp. 483–505. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2018.04.056.

[174] H. G. Othmer, S. R. Dunbar, and W. Alt. "Models of dispersal in biological systems". English. In: *J. Math. Biol.* 26.3 (1988), pp. 263–298. ISSN: 0303-6812. DOI: 10.1007/BF00277392.

[175] V. A. Panferov and A. G. Heintz. "A new consistent discrete-velocity model for the Boltzmann equation". English. In: *Math. Methods Appl. Sci.* 25.7 (2002), pp. 571–593. ISSN: 0170-4214. DOI: 10.1002/mma.303.

[176] L. Pareschi, G. Russo, and G. Toscani. "Fast Spectral Methods for the Fokker-Planck-Landau Collision Operator". In: *J. Comput. Phys.* 165.1 (2000), pp. 216–236. ISSN: 0021-9991. DOI: https://doi.org/10.1006/jcph.2000.6612.

[177] L. Pareschi and T. Rey. "Moment Preserving Fourier–Galerkin Spectral Methods and Application to the Boltzmann Equation". In: *SIAM J. Numer. Anal.* 60.6 (2022), pp. 3216–3240. DOI: 10.1137/21M1423452.

[178] L. Pareschi and T. Rey. "On the stability of equilibrium preserving spectral methods for the homogeneous Boltzmann equation". English. In: *Appl. Math. Lett.* 120 (2021). Id/No 107187, p. 6. ISSN: 0893-9659. DOI: 10.1016/j.aml.2021.107187.

[179] L. Pareschi and T. Rey. "Residual equilibrium schemes for time dependent partial differential equations". In: *Comput. Fluids* 156 (2017), pp. 329–342. DOI: 10.1016/j.compfluid.2017.07.013.

[180] L. Pareschi and G. Russo. "Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation". English. In: *J. Sci. Comput.* 25.1-2 (2005), pp. 129–155. ISSN: 0885-7474. DOI: 10.1007/BF02728986.

[181] L. Pareschi and G. Russo. "Numerical Solution of the Boltzmann Equation I: Spectrally Accurate Approximation of the Collision Operator". In: *SIAM J. Numer. Anal.* 37.4 (2000), pp. 1217–1245. DOI: 10.1137/S0036142998343300.

[182] L. Pareschi and G. Russo. "On the stability of spectral methods for the homogeneous Boltzmann equation". In: *Transport Theory Statist. Phys.* 29.3-5 (2000), pp. 431–447. ISSN: 0041-1450. DOI: 10.1080/00411450008205883.

[183] L. Pareschi and M. Zanella. "Structure preserving schemes for nonlinear Fokker-Planck equations and applications". English. In: *J. Sci. Comput.* 74.3 (2018), pp. 1575–1600. ISSN: 0885-7474. DOI: 10.1007/s10915-017-0510-z.

[184] B. Perthame. "Global existence to the BGK model of Boltzmann equation". English. In: *J. Differ. Equations* 82.1 (1989), pp. 191–205. ISSN: 0022-0396. DOI: 10.1016/0022-0396(89)90173-3.

[185] H. Poincaré. *Les méthodes nouvelles de la mécanique céleste. Tome I. Solutions périodiques. Non-existence des intégrales uniformes. Solutions asymptotiques.* French. Paris: Gauthier-Villars et Fils. 385 p. gr. 8° (1892). 1892.

[186] A. Porretta and E. Zuazua. "Numerical hypocoercivity for the Kolmogorov equation". English. In: *Math. Comput.* 86.303 (2017), pp. 97–119. ISSN: 0025-5718. DOI: 10.1090/mcom/3157. URL: semanticscholar.org/paper/051a414ef1f34d1168a7a03e161dc3637a700416.

[187] F. Poupaud. "Diffusion approximation of the linear semiconductor Boltzmann equation: Analysis of boundary layers". English. In: *Asymptotic Anal.* 4.4 (1991), pp. 293–317. ISSN: 0921-7134.

[188] F. Poupaud and J. Soler. "Parabolic limit and stability of the Vlasov-Fokker-Planck system". English. In: *Math. Models Methods Appl. Sci.* 10.7 (2000), pp. 1027–1045. ISSN: 0218-2025. DOI: 10.1142/S0218202500000525.

[189] M. Pulvirenti, C. Saffirio, and S. Simonella. "On the validity of the Boltzmann equation for short range potentials". English. In: *Rev. Math. Phys.* 26.2 (2014). Id/No 1450001, p. 64. ISSN: 0129-055X. DOI: 10.1142/S0129055X14500019.

[190] L. Saint-Raymond. *Hydrodynamic limits of the Boltzmann equation*. English. Vol. 1971. Lect. Notes Math. Berlin: Springer, 2009. ISBN: 978-3-540-92846-1; 978-3-540-92847-8. DOI: 10.1007/978-3-540-92847-8. URL: hdl.handle.net/1903/20951.

[191] D. Samaddar, D. E. Newman, and R. Sánchez. "Parallelization in time of numerical simulations of fully-developed plasma turbulence using the parareal algorithm". English. In: *J. Comput. Phys.* 229.18 (2010), pp. 6558–6573. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2010.05.012.

[192] G. Samaey and T. Slawig. "A micro/macro parallel-in-time (parareal) algorithm applied to a climate model with discontinuous non-monotone coefficients and oscillatory forcing". In: (2023). Preprint arXiv:1806.04442. arXiv: 1806.04442.

[193] J. W. Schumer and J. P. Holloway. "Vlasov simulations using velocity-scaled Hermite representations". English. In: *J. Comput. Phys.* 144.2 (1998), pp. 626–661. ISSN: 0021-9991. DOI: 10.1006/jcph.1998.5925.

[194] J. Shen, T. Tang, and L.-L. Wang. *Spectral methods. Algorithms, analysis and applications.* English. Vol. 41. Springer Ser. Comput. Math. Berlin: Springer, 2011. ISBN: 978-3-540-71040-0; 978-3-540-71041-7. DOI: 10.1007/978-3-540-71041-7.

[195] J. Steiner, D. Ruprecht, R. Speck, and R. Krause. "Convergence of parareal for the Navier-Stokes equations depending on the Reynolds number". English. In: *Numerical mathematics and advanced applications – ENUMATH 2013. Proceedings of ENUMATH 2013, the 10th European conference on numerical mathematics and advanced applications, Lausanne, Switzerland, August 26–30, 2013*. Cham: Springer, 2015, pp. 195–202. ISBN: 978-3-319-10704-2; 978-3-319-10705-9. DOI: 10.1007/978-3-319-10705-9_19.

[196] H. Struchtrup. *Macroscopic transport equations for rarefied gas flows. Approximation methods in kinetic theory.* English. Interact. Mech. Math. Berlin: Springer, 2005. ISBN: 978-3-540-24542-1. DOI: 10.1007/3-540-32386-4.

[197] R. Temam. *Navier-Stokes equations. Theory and numerical analysis. 3rd (rev.) ed*. English. Vol. 2. Stud. Math. Appl. Elsevier, Amsterdam, 1984.

[198] R. Thakur, R. Rabenseifner, and W. Gropp. "Optimization of Collective Communication Operations in MPICH". In: *Int. J. High Perform. Comput. Appl.* 19.1 (2005), pp. 49–66. ISSN: 1094-3420. DOI: 10.1177/1094342005051521. URL: https://doi.org/10.1177/1094342005051521.

[199] S. Tiwari. "Coupling of the Boltzmann and Euler equations with automatic domain decomposition". English. In: *J. Comput. Phys.* 144.2 (1998), pp. 710–726. ISSN: 0021-9991. DOI: 10.1006/jcph.1998.6011.

[200] S. Tiwari and A. Klar. "An adaptive domain decomposition procedure for Boltzmann and Euler equations". English. In: *J. Comput. Appl. Math.* 90.2 (1998), pp. 223–237. ISSN: 0377-0427. DOI: 10.1016/S0377-0427(98)00027-2.

[201] S. Tiwari. "Application of moment realizability criteria for the coupling of the Boltzmann and Euler equations." English. In: *Transp. Theory Stat. Phys.* 29.7 (2000), pp. 759–783. ISSN: 0041-1450. DOI: 10.1080/00411450008200001. URL: nbn-resolving.de/urn/resolver.pl?urn:nbn:de:hbz:386-kluedo-5995.

[202] S. Tiwari, A. Klar, and S. Hardt. "A particle-particle hybrid method for kinetic and continuum equations". English. In: *J. Comput. Phys.* 228.18 (2009), pp. 7109–7124. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2009.06.019.

[203] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics. A practical introduction*. English. 3rd ed. Berlin: Springer, 2009. ISBN: 978-3-540-25202-3; 978-3-540-49834-6. DOI: 10.1007/b79761.

[204] C. Villani. *Hypocoercivity*. English. Vol. 950. Mem. Am. Math. Soc. Providence, RI: American Mathematical Society (AMS), 2009. ISBN: 978-0-8218-4498-4; 978-1-4704-0564-9. DOI: 10.1090/S0065-9266-09-00567-5.

[205] D. Xiu. *Numerical methods for stochastic computations. A spectral method approach*. English. Princeton, NJ: Princeton University Press, 2010. ISBN: 978-0-691-14212-8.

[206] J. Yvon. *La théorie statistique des fluides et l'équation d'état*. Actualités scientifiques et industrielles : hydrodynamique, acoustique: Théories mécaniques. Hermann & cie, 1935.

Table des matières

Méthodes numériques hybrides cinétique/fluide et préservant la structure pour des équations cinétiques collisionnelles

Résumé

Cette thèse porte sur le développement et l'analyse de méthodes numériques performantes pour l'approximation des solutions d'équations cinétiques collisionnelles éventuellement non linéaires. Ces équations apparaissent dans divers domaines tels que la physique, notamment dans l'étude des semi-conducteurs et de la dynamique des gaz. Elles apparaissent aussi en biologie dans la modélisation du mouvement de cellules dans un tissu. Ces modèles présentent un aspect multi-échelle. D'une part, on a une description mésoscopique (ou cinétique) qui donne l'évolution de la fonction de distribution des particules, molécules ou cellules. D'autre part, par un processus de moyennisation, on obtient l'échelle dite macroscopique (ou fluide) qui permet de suivre l'évolution de grandeurs physiques observables : les moments de la fonction de distribution. Ces derniers correspondent notamment à la densité, la vitesse moyenne et la température des particules considérées. Nous présentons dans ce manuscrit différentes façons de tirer parti de la dynamique fluide afin de construire et étudier des méthodes numériques efficaces pour l'échelle cinétique.
Dans la première partie, nous explorons des méthodes de discrétisation visant à préserver la structure des équations continues. Nous commençons par introduire un schéma volumes finis implicite en temps pour un modèle cinétique de réaction non linéaire. Nous étudions le comportement en temps long de la solution discrète par une méthode d'hypocoercivité. Ensuite, nous examinons une méthode spectrale, basée sur des polynômes orthogonaux généraux, capable de préserver les moments de la solution, tout en assurant de bonnes propriétés de convergence.
La seconde partie est consacrée à la conception de méthodes numériques visant à réduire le coût des simulations cinétiques. Pour ce faire, nous étudions deux approches exploitant l'évolution des moments de l'inconnue. La première, une méthode dite hybride cinétique/fluide, consiste à adopter dynamiquement et localement en espace une description fluide moins coûteuse du système au lieu de la description cinétique plus onéreuse. La seconde approche repose également sur l'utilisation d'un modèle fluide, mais cette fois-ci pour accélérer les itérations temporelles de la méthode. Nous proposons ici un prototype de méthode pararéelle multi-échelle, utilisant un modèle fluide comme solveur grossier et un modèle cinétique comme solveur fin.

Mots clés : équations cinétiques, hypocoercivité, méthodes hybrides, méthodes spectrales

---

Hybrid kinetic/fluid and structure preserving numerical methods for collisional kinetic equations

Abstract

This thesis focuses on the development and analysis of efficient numerical methods for approximating solutions of potentially nonlinear kinetic collisional equations. These equations arise in various fields such as physics, notably in the study of semiconductors and gas dynamics. They also appear in biology in modelling the movement of cells within tissue. These models exhibit a multiscale aspect where there is, on one hand, a mesoscopic (or kinetic) description that gives the evolution of the distribution function of particles, molecules, or cells. On the other hand, through a process of averaging, we obtain the so-called macroscopic (or fluid) scale which allows to track the evolution of observable physical quantities: the moments of the distribution function. These moments correspond to the density, average velocity, and temperature of the considered particles. Throughout this manuscript, we present various ways to take advantage of fluid dynamics to construct and study efficient numerical methods for the kinetic scale.
In the first part, we explore discretization methods aiming to preserve the structure of continuous equations. We begin by introducing an implicit finite volume scheme for a nonlinear reaction kinetic model. We study the long-time behaviour of the discrete solution using hypocoercivity methods. Then, we examine a spectral method, based on general orthogonal polynomials, capable of preserving the moments of the solution while ensuring good convergence properties.
The second part is dedicated to the design of numerical methods aiming to reduce the cost of kinetic simulations. To do this, we study two approaches exploiting the evolution of the unknown's moments. The first, a hybrid kinetic/fluid method, involves adopting dynamically and locally in position a less costly fluid description of the system instead of the more expensive kinetic one. The second approach also relies on the use of a fluid model, but this time to accelerate the temporal iterations of the method. Here, we propose a prototype of a multiscale parareal method, using a fluid model as a coarse solver and a kinetic model as a fine solver.

Keywords: kinetic equations, hypocoercivity, hybrid methods, spectral methods