

**HABILITATION À DIRIGER DES RECHERCHES**  
**en Mathématiques Appliquées**

présentée à

**L'UNIVERSITÉ PIERRE ET MARIE CURIE**

par

**Christophe Buet**

**Sujet :**

**Analyse mathématique et numérique de  
quelques modèles hydrodynamiques et cinétiques  
de la physique des plasmas**

**Rapporteurs :**

Laurent Desvillettes  
Giovanni Russo  
Shi Jin

**Soutenue le 23 novembre 2005 devant le jury composé de :**

François Bouchut  
Laurent Desvillettes  
Pierre Degond  
Thierry Goudon  
Benoît Perthame  
Laure Saint-Raymond  
Rémi Sentis



*À Christine Pétry.*

*À mes enfants, Simon et Valentin.*



# Remerciements

Je remercie tout d'abord Stéphane Cordier et Pierre Degond avec qui j'ai eu très plaisir à travailler, qui m'ont beaucoup apporté par leur compétence, leur dynamisme et leur intuition, et aussi pour leur patience et leur disponibilité. Pierre Degond m'a permis de découvrir les équations cinétiques et les problèmes liés à leur approximation numérique. C'est avec Stéphane Cordier que j'ai travaillé pendant de nombreuses années sur ce problème. Je leur exprime encore une fois toute ma gratitude.

Je remercie vivement

- Benoît Perthame pour avoir accepté de parrainer mon dossier et d'avoir accepté d'être membre du jury.
- Laurent Desvillettes, Shi Jin et Giovanni Russo qui ont accepté de rapporter sur ce travail et ce, en dépit de leur lourdes charges. Laurent Desvillettes qui a aussi accepté de faire partie du jury.
- François Bouchut, Thierry Goudon, Laure Saint-Raymond et Rémi Sentis d'avoir accepté de faire partie du jury.

Merci aussi à

- Stéphane Dellacherie, Bruno Despres, Brigitte Lucquin, Simona Mancini, et Pierre-Arnaud Raviart avec qui j'ai eu l'occasion de travailler.
- Hervé Jourdain et Bruno Scheurer pour leurs nombreux encouragements, d'abord pour m'inciter à écrire ce mémoire, puis pendant sa rédaction.

Mes derniers remerciements, et non les moindres, vont à Christine, pour m'avoir accompagné dans la région parisienne pour la fin de mes études il y a près de vingt ans, d'avoir accepté d'y rester par la suite et pour m'avoir supporté pendant toutes ces années.



# Table des matières

<b>Introduction</b>	<b>3</b>
<b>1 Méthodes numériques déterministes conservatives et entropiques pour les opérateurs de collisions.</b>	<b>5</b>
1.1 Introduction.	5
1.2 Une méthode entropique et conservative pour l'équation de Boltzmann dans le cas d'un gaz monoatomique [a1, c1]	9
1.3 Opérateur de Boltzmann discret pour un modèle de gaz polyatomique [a2]	11
1.4 Régularisation de l'équation de Boltzmann [a4]	12
1.5 Équation de Fokker-Planck-Landau.	14
1.5.1 Algorithmes rapides [a3]	15
1.5.2 Existence de solutions et propriétés des schémas [a6]	17
1.6 Équation de Fokker-Planck-Landau isotrope.	18
1.6.1 Existence de solutions pour FPL log [a5]	19
1.6.2 Forme sans log et schémas en temps [a9]	20
1.7 Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique [a7, cr1]	21
1.8 Méthode numérique pour l'opérateur de scattering de Compton [a12]	22
1.9 Analyse spectrale de l'opérateur de Lorentz [a8]	23
1.10 Méthode numérique pour une équation de type Fokker-Planck modélisant les milieux granulaires [a14]	24
1.11 Conclusions	27
<b>2 Méthodes de moments et régimes diffusifs pour le transfert radiatif: modélisation et aspects numériques</b>	<b>29</b>
2.1 Introduction	29
2.2 Limite de diffusion du modèle de Lorentz: schémas préservant l'asymptotique [a11]	32
2.3 Analyse asymptotique pour l'hydrodynamique radiative [a13]	33
2.4 Analyse asymptotique et méthodes numériques pour les méthodes de moments en hydrodynamique radiative [cr2, s1]	35
2.5 Un modèle à flux limité pour l'hydrodynamique radiative et un schéma de splitting associé préservant l'asymptotique de diffusion [s2]	37
2.6 Conclusions	39
<b>3 Autres travaux</b>	<b>41</b>
3.1 Ionisation multi-espèces [a10]	41
3.2 Schémas monotones et équations parabolique linéaires [cr3]	42
<b>Listes des Travaux Présentés</b>	<b>45</b>
<b>Bibliographie</b>	<b>47</b>





# Introduction

Mes recherches au Commissariat à l'Énergie Atomique concernent principalement la modélisation mathématique et la simulation numérique pour la physique des plasmas. Ce mémoire présente mes contributions dans ce domaine.

Dans un premier temps j'ai étudié la discrétisation d'opérateurs de collisions en théorie cinétique.

J'ai commencé à travailler avec Pierre Degond sur la résolution numérique déterministe de l'équation de Boltzmann, pour laquelle j'ai développé une méthode conservative, entropique et à coûts réduits pour le cas mono-atomique [a1]. J'ai ensuite étendu l'opérateur de collision discret au cas d'un modèle simple de gaz polyatomique [a2]. Et avec S. Cordier et P. Degond, j'ai travaillé sur une régularisation de l'opérateur de Boltzmann, [a4], pour pouvoir soit définir une méthode particulière déterministe, soit pour traiter des sections efficaces non isotropes ou des collisions entre espèces de masse différentes dans une méthode à répartition discrète de vitesse.

Par la suite je me suis intéressé aux équations de type Fokker-Planck. J'ai travaillé sur l'équation de Fokker-Planck-Landau (homogène) pour laquelle une discrétisation entropique venait d'être proposée par B. Lucquin et P. Degond [92, 57]. Nous avons d'abord réduit le coût quadratique de l'évaluation de cet opérateur par des méthodes de type sous-réseaux et multigrille qui ont fait l'objet d'une publication dans JCP en collaboration avec S.Cordier, P. Degond et M. Lemou [a3]. Puis, avec S. Cordier nous avons justifié l'existence de solutions pour les problèmes semi-discrétisés (i.e. uniquement dans l'espace des vitesses) et discrets (i.e. en temps et en espace) [a6]. Toujours avec S. Cordier, nous avons analysé en détail le cas des fonctions isotropes [a5, a9] pour lequel les résultats et les méthodes numériques peuvent être améliorés.

J'ai travaillé aussi avec S.Cordier, R.Sentis et S. Dellacherie sur diverses équations de type Fokker-Planck modélisant les collisions ions-électrons [a7] ou un modèle simplifié de milieu granulaire [a14], ou encore le scattering de type Compton des photons (équation de Kompaneets) [a12], équations pour lesquelles nous avons obtenus des schémas numériques entropiques et conservatifs.

Avec S. Cordier et B. Lucquin-Desreux, nous avons considéré le modèle de Lorentz [a8] et notamment la limite "collisions rasantes". Par une analyse spectrale des opérateurs nous avons montré que la convergence des solutions dans cette asymptotique est uniforme en temps et que l'on contrôle les vitesses de retour vers l'équilibre.

Depuis quelques temps mes recherches concernent la discrétisation de modèles simplifiés hyperboliques en transfert radiatif et notamment la capture correcte du régime de diffusion pour éviter les couplages hyperbolique-parabolique.

Avec B. Despres, je travaille sur l'hydrodynamique radiative. Nous avons obtenu un modèle simplifié de moments dans le cadre relativiste [a13]. Nous travaillons maintenant sur les aspects numériques de ce modèle : couplage avec l'hydrodynamique [s2] et extension multi-dimensionnelle.

Et avec S. Cordier, je travaille sur des schémas numériques préservant l'asymptotique de diffusion pour des équations de transport ou pour de modèles aux moments issus du cinétique (transfert radiatif, plasmas d'électrons). Nous avons ainsi obtenu, en dimension 1 d'espace, des schémas simples mais non monotones dans le cadre du modèle de Lorentz pour les électrons [a11], et un

schéma préservant la limitation de flux pour un modèle de moments en transfert radiatif, [cr2, s1], basé sur un schéma de relaxation et sur un schéma “well-balanced”. Nous travaillons maintenant sur l’extension au cas multidimensionnel de ces schémas.

## Plan de ce mémoire.

Il est organisé en trois parties.

Dans la première partie, je résume les travaux effectués pour la résolution numérique des opérateurs de collisions. Le fil conducteur de cette partie est la dérivation de schémas numériques préservant les propriétés des opérateurs à savoir conservation des invariants physiques (masse, impulsion, énergie), des états d’équilibre et décroissance de l’entropie. Ces propriétés, si elles sont facilement vérifiées formellement pour les opérateurs continus, posent parfois des difficultés au niveau discret. De plus, ces propriétés sont nécessaires pour assurer le retour des solutions approchées vers l’état d’équilibre thermodynamique local (ETL) et donc d’avoir, en temps grand, le comportement hydrodynamique souhaité.

Dans la seconde partie je résume les travaux concernant les méthodes de moments pour le transfert radiatif et les problèmes numériques liés à leur discrétisation. Le fil conducteur de cette partie est essentiellement la dérivation de schémas numériques préservant l’asymptotique de diffusion et les domaines invariants pour des modèles de lorentz ou pour des équations du type télégraphe non linéaires (transfert radiatif, plasma électronique).

Dans une plus courte troisième partie je présente deux travaux n’entrant pas directement dans mes thématiques principales de recherche mais qui sont quand même en connexion avec elles.

Je présente enfin ma liste de publications et la bibliographie.

*Convention de notations : Mes publications sont repérées par des lettres. [a1-a14] les articles parus ou acceptés, [cr1-cr3], les notes au comptes rendus, [s1,s2] les articles soumis, les actes de congrès [c1] et les travaux non publiés [np1].*

# Chapitre 1

## Méthodes numériques déterministes conservatives et entropiques pour les opérateurs de collisions.

### 1.1 Introduction.

Cette partie décrit mes travaux sur les équations cinétiques et plus particulièrement pour la résolution numérique des opérateurs de collisions de type Boltzmann ou Fokker-Planck.

#### Quelques généralités sur la théorie cinétique et sur les méthodes numériques dans ce domaine.

Dans la théorie cinétique, chaque espèce est caractérisée par sa fonction de distribution  $f$  qui est une fonction positive des variables d'espace  $x$ , de vitesse  $v$  et du temps  $t$ . La mesure  $fdxdv$  représente la probabilité de présence d'une particule au point  $x$  avec une vitesse  $v$  à l'instant  $t$ . La fonction de distribution  $f$  est solution d'une équation de transport dans l'espace des phases  $(x,v)$ , appelée équation de Vlasov, dans laquelle apparaît un terme de force qui est un champ "moyen" ne prenant en compte que les interactions collectives à longue portée. Pour des particules chargées par exemple, il s'agit de la force due aux champs électromagnétiques appliqués (par exemple, un champ électrique ou magnétique extérieur...) ou auto-consistants (i.e. générés par les particules elles-mêmes). Cette équation peut éventuellement comprendre des termes sources afin de modéliser les collisions entre particules ou la création de nouvelles particules par ionisation (voir section 3.1, partie III). Pour une seule espèce de particules de masse  $m$  sous l'action d'un champ de force  $F = F(t,x,v)$ , l'équation de Vlasov s'écrit

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + \frac{F}{m} \cdot \nabla_v f = Q(f,f), \quad (x,v) \in \mathbb{R}^3 \times \mathbb{R}^3, t > 0. \quad (1.1)$$

Dans le cas de la force électromagnétique ( $F = q(E + v \wedge B)$ ,  $q$  désignant la charge de la particule,  $E = E(t,x)$  le champ électrique et  $B = B(t,x)$  le champ magnétique), cette équation est fortement non linéaire. En effet, la partie auto-consistante du champ de force est reliée aux grandeurs macroscopiques associées à  $f$  par l'intermédiaire des équations de Maxwell. Dans le cas électrostatique ( $B = 0$ ), il faut coupler l'équation de Vlasov avec celle de Poisson.

Dans l'équation cinétique ci-dessus, le membre de droite  $Q(f,f)$  désigne le terme de collision (quadratique relativement à  $f$ , dans le cas de Boltzmann ou Landau). Il est donné par un modèle lié à la physique du problème. Ainsi pour les gaz raréfiés, il s'agit de l'opérateur de Boltzmann décrivant des collisions binaires entre particules non nécessairement chargées. Pour les plasmas, l'opérateur couramment utilisé est l'opérateur de Fokker-Planck-Landau (FPL). Il existe autant

d'équations de Boltzmann ou de FPL que de potentiels d'interaction entre les particules. Nous renvoyons à [41] pour une taxonomie des opérateurs de collisions. Le cas le plus intéressant physiquement est le cas Coulombien qui correspond à des potentiels dits mous qui décroissent en  $1/r^2$ . On parle alors d'équations de Vlasov-Boltzmann ou de Vlasov-FPL, selon les cas. Il existe bien entendu d'autres modèles d'opérateurs de collision, comme le modèle B.G.K., le modèle de Lorentz, etc.

La théorie mathématique pour cette équation, dans le cas Boltzmann, est par exemple due à L. Arkeryd [1, 2], R.E. Caflish [11], L. Desvillettes [67], T. Elmroth [18], T. Gustaffson [24] et B. Wennberg [42]. Dans [44], A.A. Arseniev et N.V. Peskov montrent l'existence, pour des intervalles de temps courts, de solutions faibles à l'équation de Fokker-Planck homogène dans le cas Coulombien. En dehors de ce cas, A.A. Arseniev et O.E. Buryak montrent dans [43] l'existence globale en temps de solutions, mais sous des hypothèses fortes sur le noyau et sur la donnée initiale. Signalons aussi des travaux très récents d'existence de solutions classiques, dans le cas de potentiels durs, établis par L. Desvillettes et C. Villani [70, 71] pour une classe assez large de données initiales. Pour le cas non homogène, des théorèmes d'existence globale ne nécessitant pas d'hypothèses fortes sur les données initiales ont pu être obtenus, grâce à la notion de solution renormalisée introduite par R.Di Perna et P.L. Lions [15] dans le cas Boltzmann. L'extension au cas Fokker-Planck résulte des travaux de P.L. Lions [89] et, plus récemment, de C. Villani [105]. Nous renvoyons à [41] pour une présentation de résultats récents.

La résolution numérique de ce système d'équations repose sur un algorithme de décomposition ou "splitting" : on résout pendant un pas de temps la partie transport i.e. le membre de gauche de (1.1) puis on calcule pendant un pas de temps l'effet des collisions ; si nécessaire, on résout les équations électromagnétiques (Maxwell, Poisson...) et on itère le procédé. L'évaluation numérique des opérateurs de collisions est généralement plus coûteuse que la phase de transport en raison du caractère quadratique de l'opérateur. Les travaux présentés dans cette partie sont consacrés à la résolution de la phase collision ou de façon équivalente du cas homogène (lorsque que  $f$  est indépendante de  $x$  et que  $F = 0$ ). La convergence d'un tel algorithme a d'ailleurs été obtenue par L. Desvillettes [66] pour l'opérateur de transfert radiatif, et par L. Desvillettes et S. Mischler pour le cas Boltzmann [69].

Les deux façons les plus utilisées (par les physiciens par exemple) pour représenter les fonctions de distribution sont les méthodes dites particulières et les méthodes à répartition uniforme des vitesses.

Dans le premier cas, la fonction de distribution est approchée par une somme de masses de Dirac de poids  $f_i(t)$  et situées dans l'espace des phases en  $(X_i(t), V_i(t))$ .

$$f(x, v, t) \approx \sum_i f_i(t) \delta(x - X_i(t)) \delta(v - V_i(t)). \quad (1.2)$$

où  $\delta$  désigne la masse de Dirac. Dans cette approche, la phase transport est très simple à résoudre car il suffit d'intégrer l'équation des trajectoires des particules sachant la position, la vitesse et la force qui s'applique sur la particule  $i$ . Le calcul de la force  $F_i$  nécessite de connaître les grandeurs macroscopiques et donc d'intégrer dans l'espace des vitesses pour les particules situées dans la même cellule d'un maillage en espace (méthodes PIC).

Le traitement de la phase de collision se fait généralement par une méthode de Monte-Carlo donc probabiliste. À la fin de la phase de collisions les particules auront changés de position  $V_i(t)$  dans l'espace des phases, les poids  $f_i(t)$  n'évoluant généralement pas. On donc une représentation dynamique de l'espace des vitesses. C'est ce qui fait la force de ce type de méthodes pour assurer par exemple la conservation de la quantité de mouvement et de l'énergie pour l'opérateur de Boltzmann pour des collisions binaires : le processus de collisions imite ce qui se passe pour des vraies particules. L'inconvénient majeur d'une méthode Monte Carlo est le bruit numérique. Pour réduire ce bruit, les deux méthodes utilisés sont soit de prendre un grand nombre de particules par maille d'espace (par exemple calcul de solutions instationnaires), soit de moyenner des résultats indépendants (calculs de solutions stationnaires). Typiquement en gaz raréfiés pour calculer l'écoulement stationnaire autour d'un objet rentrant dans l'atmosphère, on prend de l'ordre de

10 à 20 particules par maille et dès que l'on estime que la solution calculée est stationnaire, on moyenne des instantanés de la solution pris à intervalles réguliers, les intervalles de temps étant pris suffisamment grands pour supposer ces instantanés indépendants. Avec une moyenne de 20 particules par mailles le nombre d'instantanés est généralement de l'ordre de plusieurs centaines. Le grand avantage de ces méthodes est bien entendu le faible coût en temps CPU et mémoire. Au regard du développement des ordinateurs on peut estimer que dans le domaine des gaz raréfiés c'était la seule méthode praticable jusqu'au début des années 1990.

Les méthodes particulaires déterministes, voir par exemple [206] pour le principe général, où l'on fait évoluer les poids  $f_i$  et non les vitesses  $V_i$  pendant la phase de collision ne sont guère utilisées car elles ne permettent pas d'assurer la conservation de la quantité de mouvement ou de l'énergie dans le cas de Boltzmann ou Landau par exemple.

À la fin des années 1990 les ordinateurs étant alors de plus en plus performants il paraissait concevable pour l'opérateur de Boltzmann en gaz raréfiés de développer des méthodes dites déterministes. Le choix fait par tous ceux ayant travaillé sur le sujet a été celui de méthodes avec répartition uniforme et discrète des vitesses [39, 40, 37, a1, 28].

Pour ces méthodes, la fonction de distribution est connue par sa valeur aux points d'un maillage uniforme et indépendant du temps de l'espace des phases. On peut toujours écrire (1.2), mais cette fois ce sont les poids  $f_i(t)$  qui vont évoluer et non les positions  $X_i$  ou les vitesses  $V_i$ , et l'évolution sera de nature déterministe. On a donc une représentation statique de l'espace des vitesses. Une méthode de type différences finies ou volumes finis rentre dans ce cadre là.

Mais être déterministe ne suffit pas si l'on veut minimiser la taille du maillage dans l'espace des vitesses et donc le coût mémoire et CPU de l'algorithme.

Par exemple les équations de Boltzmann homogène ou de Fokker-Planck-Landau homogène possèdent des propriétés physiques et mathématiques communes, comme celle de conservation (de la masse, de l'impulsion et de l'énergie) et de (dé)croissance de l'entropie (mathématique). Ces propriétés sont souvent plus faciles à énoncer sur la formulation faible de l'opérateur de collisions : soit  $\psi$  une fonction test, on montre que les seuls invariants de collisions

$$\int Q(f, f)(v) \psi(v) dv = 0 \Leftrightarrow \exists a, b, c \psi(v) = a + bv + cv^2,$$

qui correspondent respectivement à la masse, l'impulsion et l'énergie. Le théorème H s'obtient en choisissant  $\psi = \log(f)$  dans la formulation faible (de Boltzmann ou de FPL) et on obtient

$$\int Q(f, f)(v) \log(f(v)) dv \leq 0.$$

La fonctionnelle d'entropie  $H = \int f \log(f)$  décroît et son minimum correspond aux états d'ETL ou Maxwellienne locale

$$f(t, x, v) = n(t, x) \frac{e^{-|v - u(t, x)|^2 / 2T(t, x)}}{(2\pi T(t, x))^{3/2}}.$$

Lorsqu'on suppose que le milieu est à l'ETL, on obtient en prenant les 5 premiers moments de l'équation de Vlasov (1.1) et compte tenu des invariants de collision, les équations de l'hydrodynamique. Il s'agit donc d'une fermeture du système d'équations de moments basée sur l'hypothèse que la fonction de distribution est une Maxwellienne. Pour d'autres échelles de temps, on obtient des modèles de diffusion. La justification de ces passages à la limite a fait l'objet de très nombreuses études [3, 4, 14].

Une discrétisation déterministe mais non conservative de ces opérateurs de collision engendrera de grandes erreurs de calculs à moins de prendre un maillage suffisamment fin de l'espace des vitesses, ce qui peut conduire à des coûts prohibitifs. C'est par exemple le cas des méthodes pour l'opérateur de Landau basées sur la forme convection diffusion avec potentiels de Rosenbluth, [79].

Vu aussi sous l'angle du couplage entre une équation cinétique et sa limite hydrodynamique, il sera toujours plus facile de coupler une méthode conservative et entropique avec un schéma pour le modèle fluide qu'avec une méthode bruitée ou non conservative et entropique.

C'est pourquoi nous nous sommes donc attachés à construire des solutions approchées de ces équations préservant ces propriétés de conservation et qui assurent le retour vers l'ETL, des solutions. C'est ce type de méthodes que l'on appellera dans ce qui suit DVM pour "Discrete Velocity Models". La totalité de mes travaux pour le traitement numérique des termes de collision ont été faits dans le cadre des DVM.

**Remarque 1** *Les méthodes spectrales, voir les travaux de Pareschi et Russo dans ce domaine [97, 35, 36, 95, 98, 34], ne sont guère utilisées pour le traitement des termes de collisions, car non conservatives, peu pratiques pour représenter des fonctions de distribution piqués.*

## Principe et mise en oeuvre des DVM dans le cas homogène en espace.

Le premier objectif est de construire une version discrète de l'opérateur considéré ayant les propriétés de conservation, retour à l'équilibre et solutions positives du modèle continu.

Les DVM que je présente sont en fait toutes construites sur le même principe. L'espace des vitesses est donc discrétisé de façon uniforme, presque toujours représenté par un maillage carré :  $v_i = i\Delta v$ ,  $\Delta v$  étant le pas de maillage, et  $i$  dans  $\mathbb{Z}^3$ .

Pour obtenir un modèle de collision discret préservant les invariants de collision et assurant la décroissance de l'entropie c'est la formulation faible symétrisée et "entropique" de l'opérateur qui doit, de préférence, être discrétisée. Comme en numérique il n'y a généralement pas équivalence entre les discrétisations des différentes formes d'une même expression, le moyen le plus simple d'avoir les propriétés souhaitées de conservation et de décroissance de l'entropie, c'est donc de discrétiser la forme de l'opérateur la plus "adéquante". On peut illustrer cela sur l'équation de Landau isotrope pour des potentiels Maxwelliens. L'équation en variable d'énergie  $\varepsilon$  s'écrit

$$\frac{\partial f(\varepsilon)}{\partial t} = \frac{1}{\sqrt{\varepsilon}} \frac{d}{d\varepsilon} \int_0^{\varepsilon_0} \left( f(\varepsilon') \frac{d}{d\varepsilon} f(\varepsilon) - f(\varepsilon) \frac{d}{d\varepsilon'} f(\varepsilon') \right) \varepsilon^{3/2} \varepsilon'^{3/2} d\varepsilon'. \quad (1.3)$$

Soit  $\rho = \int_{\varepsilon} f(\varepsilon) \sqrt{\varepsilon} d\varepsilon$  le nombre de particules et  $\rho E = \int_{\varepsilon} f(\varepsilon) \varepsilon \sqrt{\varepsilon} d\varepsilon$  l'énergie totale alors (1.3) s'écrit aussi tout simplement

$$\frac{\partial f(\varepsilon)}{\partial t} = \rho \frac{1}{\sqrt{\varepsilon}} \frac{d}{d\varepsilon} \varepsilon^{3/2} \left( \frac{d}{d\varepsilon} f(\varepsilon) + E f(\varepsilon') \right) \quad (1.4)$$

ou alors sous la formulation faible symétrisée et "entropique" suivante,

$$\begin{aligned} \int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon} d\varepsilon &= -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} f(\varepsilon) f(\varepsilon') \left( \frac{\partial \phi(\varepsilon)}{\partial \varepsilon} - \frac{\partial \phi(\varepsilon')}{\partial \varepsilon'} \right) \\ &\quad \left( \frac{\partial \ln f(\varepsilon)}{\partial \varepsilon} - \frac{\partial \ln f(\varepsilon')}{\partial \varepsilon'} \right) \varepsilon^{3/2} \varepsilon'^{3/2} d\varepsilon' d\varepsilon, \end{aligned} \quad (1.5)$$

pour toute fonction test  $\phi$ . Des trois formes de cette équation de Landau, c'est sur la dernière, (1.5), qu'on lit aisément la conservation du nombre de particules, de l'énergie et de la décroissance de l'entropie  $\rho = \int_{\varepsilon} \log(f(\varepsilon)) f(\varepsilon) \sqrt{\varepsilon} d\varepsilon$ . C'est donc cette forme que l'on discrétisera de préférence à la forme (1.4) bien plus simple. Sur cet exemple on peut penser que l'on augmente le coût de l'évaluation de l'opérateur, mais en fait il n'en est rien comme on le verra par la suite, paragraphe (1.6) pour l'équation de Landau isotrope pour des potentiels Coulombiens ou (1.10) pour une équation de Fokker-Planck pour les milieux granulaires..

Pour les opérateurs de Boltzmann ou Landau, on peut dire aussi que c'est grâce à un maillage uniforme que l'on arrive à construire de tels opérateurs discrets.

Cette façon d'obtenir le schéma est donc aux antipodes des schémas construits par les physiciens, par exemple pour les équations de type Fokker-Planck. Pour l'équation de Landau, les physiciens préféreront discrétiser la formulation avec potentiels de Rosenbluth, voir par exemple [79], c'est à dire la forme convection diffusion, les deux potentiels de Rosenbluth seront calculés par les méthodes standards de calculs des potentiels ou par résolution d'équations de Poisson, les

opérateurs d'ordre 1 et 2 seront discrétisés par les formules de différences finies classiques. Sur l'exemple ci dessus les physiciens utiliseront la forme (1.4). Qu'en est-il alors des propriétés de conservation et de retour à l'équilibre. Dans le cas d'une équation plus simple de type Fokker-Planck linéaire c'est aussi la méthodologie retenue, même si dans ce cas on peut facilement forcer la méthode à préserver les états d'équilibre, voir Chang et Cooper [49].

Comme on cherche à calculer une approximation d'une fonction de distribution, donc une fonction positive, il faut aussi s'assurer de la positivité des solutions du problème homogène discret associé. Si pour des opérateurs de type Boltzmann ou pour de opérateurs de type Fokker-Planck linéaire cela ne pose pas de problèmes, en revanche pour l'équation de Landau ce n'est pas le cas, comme on le verra, pour l'opérateur discret originel [57]. Le problème de la positivité des solutions est évidemment relié à l'existence d'un pas de temps minimal garantissant la positivité pour une discrétisation temporelle explicite. Dans le cas du schéma proposé pour l'équation de Landau, voir [57], cela se traduit donc par la difficulté de choisir un pas de temps assurant la positivité dans un code de calcul.

Ayant obtenu notre opérateur de collisions discret, il reste donc à le mettre en oeuvre. Et ce n'est pas aussi simple que ça.

Par exemple pour l'opérateur de Boltzmann ou Landau, ces opérateurs étant quadratiques le coût de leur évaluation en chaque point du maillage est prohibitif. De part la nature parabolique des équations de type Fokker-Planck, pour assurer la positivité de la solution discrète c'est la restriction sur le pas de temps qui devient rédhibitoire pour un schéma explicite. Il convient donc de trouver aussi des solutions soit pour évaluer à moindre coût ces opérateurs soit pour impliciter le schéma en temps. Notons que c'est pour l'équation de Landau que le problème représente le plus de difficultés, car l'opérateur est quadratique et le problème de nature parabolique.

C'est cette méthodologie que j'ai essayé d'appliquer à quelques types d'opérateurs de collisions.

## 1.2 Une méthode entropique et conservative pour l'équation de Boltzmann dans le cas d'un gaz monoatomique [a1, c1]

Dans cet article j'ai étudié une méthode conservative et entropique pour l'équation de Boltzmann pour un gaz monoatomique. L'opérateur de boltzmann peut s'écrire sous la forme faible et symétrisée, [13]

$$\begin{aligned} \int_{\mathbb{R}^3} Q(f,f)\psi dv &= \\ &= -\frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \left( \int_{S^2} q(v-v_*,\omega)(f'f'_* - ff_*)(\psi' + \psi'_* - \psi - \psi_*)d\omega \right) dv dv_* \end{aligned} \quad (1.6)$$

$$f = f(x,v,t), f_* = f(x,v_*,t), f' = f(x,v',t), f'_* = f(x,v'_*,t)$$

et

$$v' = \frac{v+v_*}{2} + R_\omega\left(\frac{v-v_*}{2}\right), \quad v'_* = \frac{v+v_*}{2} - R_\omega\left(\frac{v-v_*}{2}\right),$$

C'est sur cette forme de l'opérateur que l'on montre facilement la conservation en temps, du nombre de particules  $\int_v f dv$ , de la quantité de mouvement  $\int_v v f dv$ , de l'énergie  $\int_v \|v\|^2 f dv$

et la décroissance de l'entropie  $\int_v f \log(f) dv$ . C'est donc cette forme de l'opérateur que l'on discrétise.

Les difficultés que l'on rencontre pour définir un opérateur discret proviennent de l'intégrale sur la sphère unité.

Étant donné un maillage uniforme  $v_i = i\Delta v$ ,  $i = (i^1, i^2, i^3) \in \mathbb{Z}^3$ , on définit une approximation de  $f$  en résolvant le système

pour tout  $i \in \mathbb{Z}^3$ ,  $\frac{df_i}{dt} = B_i$ ,  $f_i(0) = f_0(v_i)$ ,

$$\begin{aligned} Q(f, f)(v_i) &\simeq \frac{1}{(\Delta v)^3} \bar{Q}(\bar{f}, \bar{f})_i \\ &= \frac{1}{(\Delta v)^3} \sum_{j \in \mathbb{Z}^3} \sum_{(k, l) \in S_{ij}} \frac{4\pi}{\text{Card}(S_{ij})} q(v_i - v_j, \omega_{ij}^{kl}) (f_k f_l - f_i f_j) \end{aligned} \quad (1.7)$$

$$S_{ij} = \{(k, l) \in \mathbb{Z}^3 \times \mathbb{Z}^3, \quad k + l = i + j, \quad |k|^2 + |l|^2 = |i|^2 + |j|^2\}.$$

ce qui correspond à l'approximation suivante de la formulation faible de l'opérateur

$$\begin{aligned} \int_{\mathbb{R}^3} Q(f, f) \psi dv &\simeq \frac{1}{4(\Delta v)^3} \sum_{i, j \in \mathbb{Z}^3} \sum_{(k, l) \in S_{ij}} \frac{4\pi}{\text{Card}(S_{ij})} (\psi_k + \psi_l \\ &\quad - \psi_i - \psi_j) q(v_i - v_j, \omega_{ij}^{kl}) (f_k f_l - f_i f_j) \end{aligned}$$

On peut vérifier facilement sur (1.7) que la méthode développée est bien conservative et entropique.

L'intégrale sur la sphère unité dans (1.6) a donc été discrétisée de la façon la plus simple qui soit. Cela revient effectivement à considérer un gaz à répartition discrète des vitesses, voir [20] : quand deux particules de vitesses  $i$  et  $j$  collisionnent, leur vitesses après collision sont les points d'intersection du maillage et de la sphère de collision et les probabilités de transition sont toutes égales. Cela fournit une méthode bien plus simple et au moins aussi efficace que celle proposée par Rogier et Schneider [37, 39].

Bobylev et Schneider ont montré [6] que (1.7) est une approximation consistante de l'opérateur (1.6). On peut remarquer aussi que l'opérateur discret est une somme d'opérateurs de Broadwell à quatre vitesses.

La seconde difficulté que l'on rencontre dans la discrétisation de l'équation de Boltzmann est le coût exorbitant de l'évaluation du terme de collision. Pour réduire ce coût CPU, j'ai mis en oeuvre deux techniques. Elles constituent l'apport essentiel de ce travail. C'est ce qui a permis de faire des calculs en dimension 2 d'espace à un coût inférieur à celui d'une simulation Monte-Carlo, et ce pour de meilleurs résultats.

La première est une technique de sous-réseaux, qui sera d'ailleurs réutilisée dans le cas de l'opérateur de Landau. À chaque pas de temps on évalue en tout point le terme de collisions avec un sous-réseau du réseau initial, le choix du sous-réseau se faisant de façon aléatoire ou cyclique, je renvoie le lecteur au paragraphe (1.5.1) pour une explication plus détaillée.

La seconde est basée sur une formule de quadrature de type Monte Carlo pour l'opérateur de collision.

Dans tous le cas on garde une approximation conservative et entropique du terme de collision. L'opérateur reste toujours une somme de modèle de Broadwell à quatre vitesses.

La discrétisation en temps est faite soit par un schéma explicite soit par splitting et résolution exacte de modèle de Broadwell. Ce second choix permet d'avoir un schéma positif, implicite et entropique. le pas de temps est de l'ordre de quelques temps de vol libre moyen.

Pour des calculs non homogène en espace, j'ai utilisé un schéma de type volume fini classique pour la partie transport.

Les résultats numériques ont été obtenus en dimension un ou deux d'espace sur des cas réalistes modélisant des écoulements hypersoniques de gaz raréfiés. Ces résultats ont été comparés avec ceux obtenus par un code Monte-Carlo. On montre ainsi qu'avec la méthode à répartition discrète de vitesses on obtient des résultats en adéquation avec ceux d'un code de calcul Monte Carlo, et ce pour un coût CPU moindre et sans bruit aléatoire.



### 1.3 Opérateur de Boltzmann discret pour un modèle de gaz polyatomique [a2]

Cet article constitue la suite logique de l'article précédent sur la discrétisation de l'équation de Boltzmann. Dans le travail précédent on considérait un gaz monoatomique, il était donc naturel d'étendre la méthode numérique à une équation de Boltzmann modélisant un gaz polyatomique. Le choix s'est porté sur le modèle le plus simple qui soit, ne nécessitant que l'introduction d'une variable supplémentaire pour modéliser tous les phénomènes de vibration rotation, le modèle de Larsen-Borgnakke [17, 8, 7].

Soit  $\delta$  le nombre de degrés internes de liberté,  $f(x, v, I, t)$  la fonction de distribution. La densité, la quantité de mouvement et l'énergie totale sont définies respectivement par

$$\begin{pmatrix} n(x, t) \\ n(x, t)U(x, t) \\ n(x, t)E(x, t) \end{pmatrix} = \int_{\mathbb{R}^3 \times \mathbb{R}^+} \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} + I^2 \end{pmatrix} f(x, v, I, t) dv I^{\delta-1} dI.$$

L'opérateur de Larsen-Borgnakke-Boltzmann est défini par, voir [8] :

$$Q_\delta(f, f) = \int_{\Delta} B(f' f'_* - f f_*) dv_* I_*^{\delta-1} dI_* d\eta (r(1-r))^{\frac{\delta}{2}-1} dr R^2 (1-R^2)^{\delta-1} dR$$

avec

$$g = \frac{v - v_*}{2} = \text{vitesse relative},$$

$$E^2 = |g|^2 + I^2 + I_*^2 = \text{énergie totale},$$

$$(v_*, I_*, \eta, r, R) \in \Delta = \mathbb{R}^3 \times \mathbb{R}^+ \times S^{2,+} \times [0, 1]^2,$$

$$B = B(E, |Rg|, |Rg_*|, I^2 r(1-R^2), I_*^2 (1-r)(1-R^2)) > 0,$$

$$f = f(x, v, I, t), f_* = f(x, v_*, I_*, t), f' = f(x, v', I', t), f'_* = f(x, v'_*, I'_*, t),$$

et le processus de collision est défini de la façon suivante

$$\begin{cases} v + v_* = v' + v'_* \\ g' = \frac{RE}{|g|} \{g - 2(g \cdot \eta) \eta\} \\ I' = \sqrt{r(1-R^2)} E \\ I'_* = \sqrt{(1-r)(1-R^2)} E \end{cases}$$

$S^2$  est la sphère unité de  $\mathbb{R}^3$ .

L'équation de Boltzmann s'écrit donc :

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q_\delta(f, f)$$

Les états d'équilibre sont de la forme

$$f(v) = C_\delta \frac{\rho}{(RT)^{(3+\delta)/2}} \exp\left(-\frac{|v - u|^2 + 2I^2}{2RT}\right),$$

Par exemple pour le modèle "variable hard sphere" (VHS)  $B$  est donné par :

$$B = CR^{1-2\alpha} |g|^{-2\alpha} |g \cdot \eta|$$

avec  $\alpha \in [0, \frac{1}{2}]$ .

Dans cet article j'ai proposé deux approximations discrètes de cette équation vérifiant les propriétés de conservation et de décroissance de l'entropie. Le principe de construction bien que

plus compliqué à mettre en oeuvre est celui employé pour le cas monoatomique. Dans le cas où le nombre de degrés internes de liberté est nul on retrouve bien dans les deux cas l'opérateur de Boltzmann discret pour un gaz mono-atomique décrit dans le paragraphe précédent (voir [a1]).

Les techniques de réduction de coût pour l'évaluation du terme de collision et de discrétisation en temps utilisées pour le cas mono-atomique peuvent s'appliquer sans problèmes dans le cas polyatomique. Les modèles de Broadwell sous jacents peuvent encore se résoudre exactement.

Par suite d'une ré-orientation des objectifs de recherche du CEA il n'a pas été possible de tester cette méthode. Mes travaux se sont trouvés recentrés sur la physique des plasmas et plus particulièrement sur la discrétisation de l'équation de Fokker-Planck-Landau.

## 1.4 Régularisation de l'équation de Boltzmann [a4]

*En coll. avec S. Cordier et P. Degond.*

Dans les méthodes à répartition discrète de vitesse (DVM)(voir [21, a1, 37]), les vitesses sont sur un maillage fixe de  $\mathbb{R}^3$ . La consistance de ces méthodes est liée à la répartition des solutions entières de l'équation  $a^2 + b^2 + c^2 = n$ , qui vient de la conservation de l'énergie pour un quadruplet de vitesses sur la grille uniforme. Des résultats partiels ont été obtenus en utilisant des techniques issues de la théorie des nombres (par exemple [6]).

Du point de vue numérique et pratique, la principale difficulté avec les méthodes DVM est liée au petit nombre de paires de vitesses post-collisionnelles pour un couple de vitesses donné. En effet, le nombre de points d'intersection entre la sphère de collision et le maillage peut être très petit [22]. Dans de telles circonstances, la grille doit être raffinée et le coût devient exorbitant. Pour pallier cette difficulté, nous avons étudié une régularisation de la sphère de collision.

La seconde motivation de ce travail était liée à l'approximation de l'opérateur de Boltzmann par des méthodes particulières [16, 30, 33]. Les méthodes de type Monte Carlo permettent de traiter les phases de transport et de collision de façon naturelle et relativement facile à mettre en oeuvre [5, 31, 32]. Cependant, le calcul par une méthode Monte Carlo des intégrales de collisions génère un fort bruit numérique. Il serait donc intéressant de trouver une résolution déterministe des intégrales de collisions. Cet objectif a été atteint pour la théorie du transport linéaire [16, 30, 33] mais cela nécessite de trouver une régularisation du processus de collision microscopique afin que l'opérateur soit globalement conservatif.

Dans cet article nous avons étudié deux stratégies possibles. La première consiste à "épaissir" la sphère de collisions et à autoriser les vitesses post-collisionnelles à être situées sur une coquille sphérique. L'opérateur de Boltzmann peut s'écrire sous la forme :

$$Q[f](v) = \int_{(\mathbb{R}^3)^3} c(v, v_1, v', v'_1) (f' f'_1 - f f_1) dv' dv'_1 dv_1,$$

où l'intégration est prise sur l'ensemble des triplets de vitesses et  $c$  est défini par

$$c(v, v_1, v', v'_1) = \delta_0(v + v_1 - v' - v'_1) \delta_0(|v - v_1|^2 - |v' - v'_1|^2) C\left(|v - v_1|, \frac{(v - v_1, \Omega)}{|v - v_1|}\right),$$

où  $\Omega = \frac{(v' - v'_1) - (v - v_1)}{|(v' - v'_1) - (v - v_1)|}$  et  $\delta_0$  représente la mesure de Dirac en  $x = 0$ . Notons que les fonctions  $c$  et  $C$  ne sont définies que pour les vitesses  $v, v_1, v'$  et  $v'_1$  satisfaisant les propriétés de conservation de l'impulsion et de l'énergie

$$v + v_1 = v' + v'_1, \quad |v|^2 + |v_1|^2 = |v'|^2 + |v'_1|^2. \quad (1.8)$$

Dans cette formulation, que l'on trouve par exemple dans [13], les conservations (1.8) sont assurées par les mesures de Dirac et nous avons cherché à régulariser ces mesures de façon à augmenter le nombre de vitesses post-collisionnelles admissibles. L'opérateur régularisé  $\tilde{Q}$  est écrit sous forme

faible symétrisée pour une fonction test  $\Psi$

$$\begin{aligned} (\tilde{Q}[f], \Psi) &= \frac{-1}{4} \int_{(\mathbb{R}^3)^4} \tilde{C} \left( \frac{v-v_1}{2}, \frac{v'-v'_1}{2} \right) (\Psi' + \Psi'_1 - \Psi - \Psi_1) \\ &\quad \left( f' f'_1 \delta \left( \frac{v+v_1}{2}, \frac{v-v_1}{2}, \frac{v'+v'_1}{2}, \frac{v'-v'_1}{2} \right) \right. \\ &\quad \left. - f f_1 \delta \left( \frac{v'+v'_1}{2}, \frac{v'-v'_1}{2}, \frac{v+v_1}{2}, \frac{v-v_1}{2} \right) \right) dv dv' dv'_1 dv_1, \end{aligned}$$

avec  $\tilde{C}(z, z')$  défini par  $C(\overline{|z|}, \frac{|z-z'|}{2|z|})$ , où  $\overline{|z|}$  est une valeur moyenne de  $|z|$  et  $|z'|$ . On obtient d'abord des conditions nécessaires sur la fonction  $\delta$  pour que les propriétés (conservations, décroissance de l'entropie et états d'équilibre) soient satisfaites. Lorsque seule la condition sur l'énergie est régularisée, on construit un tel opérateur mais lorsque les deux contraintes (impulsion et énergie) sont relaxées, la construction nécessite des conditions sur la fonction de distribution elle-même afin d'assurer que lorsque le paramètre de régularisation tend vers 0, on retrouve l'opérateur de Boltzmann usuel. De plus, la section efficace régularisée dépend de la fonction de distribution ce qui complique la structure de l'opérateur et son implémentation. Rappelons que les propriétés requises sont indispensables pour garantir le retour vers la Maxwellienne d'équilibre en temps grand.

La seconde approche consiste à "modifier les masses" lors du processus de collisions tout en préservant les propriétés macroscopiques. Précisons la démarche. Pour un couple  $v, v_1$  de vitesses, on définit un centre de masse approché par

$$V(x) = xv + (1-x)v_1,$$

et les vitesses post-collisionnelles

$$v' = V(x) + (1-y)r\omega, \quad v'_1 = V(x) - yr\omega, \quad (1.9)$$

où  $x$  et  $y$  sont deux réels proches de  $1/2$ . La conservation de l'énergie permet de déterminer  $r$  en fonction de  $x, y, v$  et  $v_1$ . On définit une suite régularisante  $h_\varepsilon(x - 1/2) = \xi((x - 1/2)/\varepsilon)/\varepsilon$ , avec  $\xi$  une fonction paire, positive, régulière telle que  $\int_{z \in \mathcal{I}} \xi(z - 1/2) dz = 1$  et pour tout  $\varepsilon > 0$  et toute fonction  $f$ , on pose

$$\begin{aligned} \mathcal{C}_\varepsilon(f, f) &= \int_{\mathbb{R}^3} \int_{S^2} \int_{(x, y) \in \mathcal{I}^2} q \left( \frac{|v-v_1| + |v'-v'_1|}{2}, \omega \right) \chi \left( \frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y \right) \\ &\quad \left( \frac{M_1 M f' f'_1 - M'_1 M' f f_1}{\sqrt{M_1 M M'_1 M'}} \right) h_\varepsilon(x - \frac{1}{2}) h_\varepsilon(y - \frac{1}{2}) dv_1 d\omega p(x) dx dy, \end{aligned}$$

où  $(v', v'_1)$  est calculé à partir de  $(v, v_1, \omega, x, y)$  à partir des relations (1.9) et  $q(u, \omega) = u\sigma(u, \omega)$  et  $\sigma$  est la section efficace différentielle de collision,  $M = M^f$  est la Maxwellienne d'équilibre associée à  $f$ ,  $\chi$  est une fonction permettant d'assurer la décroissance de l'entropie et  $p(x) = 64x(x-x^2)^{3/2}$  la microréversibilité de l'opérateur. On montre formellement la convergence de cet opérateur régularisé vers l'opérateur de Boltzmann

$$\lim_{\varepsilon \rightarrow 0} \mathcal{C}_\varepsilon(f, f) = \int_{\mathbb{R}^3} \int_{S^2} (f' f'_1 - f f_1) q(|v-v_1|, \omega) dv_1 d\omega.$$

Nous vérifions qu'il satisfait les lois de conservations et la décroissance de l'entropie. Il est immédiat d'après la définition de  $\mathcal{C}_\varepsilon$  que les Maxwelliennes sont des états d'équilibres. L'implication inverse n'est pas claire mais nous montrons comment modifier cet opérateur pour éliminer les éventuels invariants non physiques. L'intérêt de cette méthode par rapport à la précédente est que la modification de la section efficace de collision ne dépend que de la Maxwellienne d'équilibre et non de la fonction de distribution.

## 1.5 Équation de Fokker-Planck-Landau.

Avant de présenter mes travaux sur l'analyse numérique de l'opérateur de Fokker-Planck-Landau, nous allons d'abord en rappeler quelques propriétés.

La dérivation statistique de l'**opérateur de Fokker-Planck** que l'on trouve par exemple dans le chapitre 6 de [113], repose sur le fait que les grandes déviations subies par une particule résultent essentiellement d'une succession de "collisions rasantes" i.e. de faibles déviations (voir section (1.9) pour la justification de cette limite pour un opérateur simplifié).

Citons les travaux de Lucquin, Degond et Desvillettes sur la justification de cette limite [65, 56]. Dans [56], on montre que, moyennant une adimensionnalisation convenable de l'opérateur de Boltzmann (dans le cas d'une loi d'attraction Coulombienne), il est possible de mettre en évidence un petit paramètre physique, appelé **paramètre plasma**. Nous constatons ensuite que le premier terme du développement asymptotique de cet opérateur de collision en fonction du petit paramètre est précisément l'opérateur de Fokker-Planck (voir la section (1.9) pour une étude similaire dans le cas de l'opérateur de Lorentz).

L'opérateur de Fokker-Planck, que nous noterons de manière spécifique  $Q^{FP}(f, f)$ , s'écrit dans le cas général

$$Q^{FP}(f, f)(v) = \nabla_v \cdot \left( \int_{\mathbb{R}^3} \Phi(v - v_1) (\nabla_v f f_1 - \nabla_{v_1} f_1 f) dv_1 \right),$$

où, pour simplifier les notations, la variable  $t$  est omise, et :  $f = f(v)$ ,  $f_1 = f(v_1)$ ,  $\nabla_v f = (\nabla f)(v)$ ,  $\nabla_{v_1} f = (\nabla f)(v_1)$ . Pour tout  $w \in \mathbb{R}^3$ , le potentiel  $\Phi(w) = \Phi(-w)$  est une matrice carrée d'ordre trois symétrique et semi-définie positive qui s'écrit

$$\Phi(w) = (\mathcal{B}S)(w), \quad S(w) = Id - \frac{w \otimes w}{|w|^2},$$

$Id$  désignant la matrice identité dans  $\mathbb{R}^3$ . Le noyau  $\mathcal{B}$  qui apparaît dans cet opérateur est le produit de  $|v|^3$  par la section efficace différentielle de collision, quantité statistique qui est elle-même étroitement liée au potentiel d'interaction entre les particules. Plus précisément, si la force d'interaction est de la forme  $r^{-s}$ , avec  $s \geq 2$  ( $r$  désigne la distance entre deux particules), le noyau s'écrit  $\mathcal{B}(v) = C|v|^{\gamma+2}$ , où  $\gamma = \frac{s-5}{s-1}$ , et  $C$  est une constante positive. On appelle potentiels durs les potentiels pour lesquels  $\gamma > 0$ , et potentiels mous ceux pour lesquels  $\gamma < 0$ . Enfin le cas particulier  $\gamma = 0$  correspond aux molécules Maxwelliennes. De cette distinction résultent des propriétés mathématiques, qui sont en général beaucoup plus difficiles à obtenir dans le cas de potentiels mous. Le cas Coulombien correspond à  $s = 2$ , soit  $\gamma = -3$ .

Nous avons écrit l'opérateur de Fokker-Planck sous une forme conservative, parfois appelée forme de Landau ; c'est une forme très agréable d'un point de vue mathématique. On rencontrera aussi dans la littérature physique une autre écriture de cet opérateur, appelée forme de Rosenbluth : en développant la forme de Landau précédente, l'opérateur de Fokker-Planck apparaît comme une combinaison (non linéaire) entre un opérateur de diffusion et un opérateur de friction ; les coefficients de cette combinaison s'appellent potentiels de Rosenbluth.

Si toutes les formes de l'opérateur sont équivalentes au niveau continu, une fois discrétisé, il n'en est pas de même. Des schémas de différences finies conservatifs ont été décrits antérieurement, soit dans le cas totalement isotrope (la fonction de distribution ne dépend de la vitesse que par l'intermédiaire de son module) dans [47, 48], soit dans le cas d'une géométrie à symétrie azimutale (i.e. en variables sphériques, la fonction de distribution est supposée indépendante de l'angle azimuthal) par différents auteurs [99, 100, 106]. La conservation des invariants collisionnels y est en général assurée, la décroissance de l'entropie parfois établie. Mais nulle part n'est vérifié le fait que les seuls invariants collisionnels soient les invariants "physiquement acceptables", à savoir la masse, la quantité de mouvement et l'énergie. Concernant les simulations numériques effectives de l'équation de Fokker-Planck, les premières remontent à 1957 pour le cas isotrope, [102] (voir aussi la section 1.6) et 1987 pour le cas axisymétrique [78]. Plus récemment, O. Larroche implémente dans [79] un schéma volumes finis qui ne préserve que la masse. Ce schéma est ensuite amélioré,

selon une méthode de correction inspirée de [45], de manière à ce qu'il conserve aussi l'impulsion et l'énergie [54]. Toutes les travaux cités ci dessus sont basées sur la forme avec potentiels de Rosenbluth.

Citons pour conclure les travaux de Lucquin, Degond [57, 59], Epperlein [72] (isotrope), Frenod-Lucquin (axisymétrique [76]) et Lemou [84, 85], G. Russo et L. Pareschi (méthodes spectrales [96]) et aussi ceux de Dellacherie [64] pour les plasmas électrons-ions.

Les travaux que je vais présenter maintenant portent sur la formulation "Landau" de l'opérateur : la première section est consacrée aux algorithmes rapides (multigrilles et sous-réseaux), la seconde à l'existence de solutions du problème semi-discret et discret.

### 1.5.1 Algorithmes rapides [a3]

*En coll. avec S. Cordier, M. Lemou et P. Degond.*

Nous avons mis au point des algorithmes rapides (de type multigrille et sous-réseaux) pour résoudre l'équation de Fokker Planck à partir d'une discrétisation de cet opérateur proposée par P. Degond et B. Lucquin [92, 57]. Ceci permet de réduire le coût a priori quadratique de telles simulations numériques à un coût d'ordre  $N \ln(N)$  et donc d'envisager de simuler des phénomènes beaucoup plus complexes.

Plus précisément, nous avons travaillé sur la formulation dite Landau-Log de l'opérateur. Un schéma d'approximation par différences finies de l'équation de Fokker-Planck homogène préservant au niveau discret toutes les propriétés physiques (conservations de la masse, l'impulsion et l'énergie, décroissance de l'entropie et états d'équilibre Maxwelliens et uniquement ceux là!) a été proposé par P. Degond et B. Lucquin dans [92, 57]. Rappelons en la structure. Étant donné un maillage uniforme  $v_i = i\Delta v$ ,  $i = (i^1, i^2, i^3) \in \mathbb{Z}^3$  de  $\mathbb{R}^3$ , on définit une approximation de  $f$  en résolvant le système

$$\text{pour tout } i \in \mathbb{Z}^3, \quad \frac{df_i}{dt} = FPL_i, \quad f_i(0) = f_0(v_i), \quad (1.10)$$

où  $FPL_i \simeq Q^{FP}(f, f)(v_i)$  est défini par :

$$FPL_i = -D^* \cdot \left[ \sum_{j \in \mathbb{Z}^3} \phi(v_i - v_j) f_i f_j (D \log f_i - D \log f_j) \Delta v^3 \right].$$

Dans cette expression,  $D$  est un opérateur de différences finies, défini de manière uniforme sur tout le maillage, et approchant l'opérateur gradient au moins à l'ordre 1 et  $D^*$  est l'opérateur discret adjoint. On peut l'écrire sous forme faible :

$$\begin{aligned} \sum_{i \in \mathbb{Z}^3} FPL_i \psi_i &= \frac{-\Delta v^3}{2} \sum_{(i,j) \in \mathbb{Z}^6} f_i f_j \\ &\quad ((D\psi)_i - (D\psi)_j)^T \Phi(v_i - v_j) ((D \ln f)_i - (D \ln f)_j). \end{aligned} \quad (1.11)$$

Il est clair sur la formule (1.11) que le coût de cet opérateur sera quadratique. Nous rappelons comment se ramener à un domaine borné de l'espace des vitesses en suivant les idées présentées dans [57]. Puis, nous expliquons comment cette discrétisation peut être étendue de façon immédiate au cas multi-espèces. Rappelons que cette extension, indispensable pour traiter les cas physiquement intéressants est un obstacle majeur lorsqu'on utilise une discrétisation basée sur la formulation de Rosenbluth [79].

#### Choix de l'opérateur de différence finies.

Il a été montré dans [56] que pour l'opérateur obtenu avec le schéma différence fini décentré (avec  $\varepsilon_i \in \{\pm 1\}$ )

$$(D_s \psi)_i = \varepsilon_i \frac{\psi_{i+\varepsilon_i e_s} - \psi_i}{\Delta v}, \quad s = 1, 2, 3,$$

les seuls invariants de collisions sont les combinaisons linéaires de  $1$ ,  $v$  et  $v^2$  (correspondant respectivement à la masse, l'impulsion et l'énergie). Avec le schéma centré,

$$(D_c^s \psi)_i = \frac{\psi_{i+e_s} - \psi_{i-e_s}}{2\Delta v}, \quad s = 1, 2, 3, \quad (1.12)$$

l'opérateur possède des invariants supplémentaires associés à des états d'équilibre parasites. En revanche, les schémas décentrés ne sont que d'ordre 1 et ils introduisent une dissymétrie artificielle dans la discrétisation. Nous avons montré qu'il était possible d'éliminer les invariants parasites tout en conservant la symétrie de l'opérateur en prenant la moyenne arithmétique des opérateurs obtenus pour chacun des 8 choix (2 possibilités pour 3 dimensions) possibles de direction pour les opérateurs décentrés. L'opérateur obtenu s'écrit comme celui obtenu avec le schéma centré avec une correction d'ordre 2 qui élimine les invariants parasites du schéma (1.12).

#### Réduction du coût calcul : sous-réseaux.

La première méthode que nous avons considérée est inspirée de mes travaux pour l'opérateur de Boltzmann. Cette méthode appelée méthode des sous-réseaux consiste à ne retenir dans la somme double que les indices  $i$  et  $j$  telles que le vecteur  $(i - j)$  soit un multiple de  $a$ . Nous montrons que l'opérateur  $(Q_i[a] + Q_i[b])/2$  avec  $a$  et  $b$  premiers entre eux possède les mêmes propriétés que  $Q$ , en particulier, les seuls invariants sont les invariants physiques. Le coût de l'évaluation est multiplié par  $\frac{1}{a^3} + \frac{1}{b^3}$ . Par exemple, avec  $a = 7$  et  $b = 8$ , il est divisé par environ 25.

#### Méthodes multigrilles.

On découpe le domaine cubique unité de calcul  $C_0 = [0, 1]^3$  (le "père") d'arête 1 en 8 cubes réguliers  $C_1^r$  (les "enfants") d'arête  $1/2$  et de centre

$$O_1^r = \left( \frac{1}{2^2} + \frac{r_1}{2}, \frac{1}{2^2} + \frac{r_2}{2}, \frac{1}{2^2} + \frac{r_3}{2} \right),$$

avec  $r = (r_1, r_2, r_3) \in I_1 \stackrel{\text{def}}{=} \{0, 1\}^3$ . On écrit ensuite la somme (ou l'intégrale) double

$$\int_{C_0} Q(f, f)(v) \psi(v) dv = \sum_{(r, r') \in I_1^2} \int_{C_1^r \times C_1^{r'}} H(v, w) dv dw.$$

On itère le procédé en divisant à nouveau chaque sous-cube en 8 cubes d'arêtes de longueur  $1/4$  etc... jusqu'à un niveau  $K$ . On définit ensuite la notion de cubes de niveau  $k \in \{2 \dots K\}$  - i.e. d'arêtes de longueur  $2^{-k}$  bien séparés. On dit que deux cubes de niveau  $k$  sont bien séparés s'ils ne sont pas voisins (pas de faces ou de sommets communs) mais que leurs "parents" le sont. On obtient ainsi une partition de  $C_0 \times C_0$  en prenant la réunion des sous-cubes bien séparés de niveau  $k$  variant de 2 à  $K$  et des sous-cubes voisins au niveau le plus fin  $K$ . On utilise ensuite une méthode aléatoire de type Monte-Carlo d'autant plus précise que le niveau est grand (exacte au niveau  $K$ ) pour évaluer la contribution entre deux sous-cubes. Cette méthode est bien adaptée au cas Coulombien pour lequel la section efficace décroît avec la vitesse relative de sorte que la contribution entre les sous-cubes des premiers niveaux est moins importante que celle des niveaux élevés. Le coût d'un tel algorithme est  $N \ln(N)$  avec  $N = 2^{3K}$  le nombre de points de discrétisation.

Une autre méthode a été étudiée par M. Lemou dans [84] : il s'agit d'une méthode de type multipolaire. On conserve la décomposition en niveaux et la hiérarchie multigrille mais on remplace le calcul Monte-Carlo par un développement multipolaire tronqué. Il s'agit d'une approximation (à cause de la troncature) qui est exacte dans le cas maxwellien. La complexité de l'algorithme est comparable. Ces algorithmes rapides (sous-réseaux, multigrilles et multipolaires) ont été adaptés au cas axisymétrique par M. Lemou [85].

Récemment, un autre type de méthode, dite spectrale, a été proposé par G. Russo et L. Pareschi [96]. Celle-ci est également en  $N \ln N$  grâce à l'utilisation de la transformée de Fourier rapide. De plus, il est possible d'en contrôler la précision. En revanche, le seul invariant est la masse et la fonction de distribution tend vers une valeur constante en temps grand.

### 1.5.2 Existence de solutions et propriétés des schémas [a6]

*En coll. avec S. Cordier.*

Nous avons montré l'existence de solutions pour le système d'équations différentielles non linéaires couplées correspondant à l'opérateur de Fokker-Planck discrétisé dans l'espace des vitesses. Nous montrons également que le problème peut être discrétisé explicitement en temps moyennant une condition sur le pas de temps de type CFL.

Rappelons que les algorithmes basés sur la formulation dite Landau-log de l'opérateur de Fokker-Planck permettent de vérifier les propriétés physiques (conservation, entropie et états d'équilibre). Ces propriétés sont indispensables pour éviter un chauffage ou un refroidissement artificiel de la fonction de distribution, comme cela a été noté pour l'équation de Boltzmann [25]. En d'autres termes, la décroissance de l'entropie est importante pour assurer la thermalisation du plasma et les conservations pour qu'elle se fasse vers l'ETL (équilibre thermodynamique local). De plus, comme nous l'avons montré dans [a3], il est possible de généraliser ces schémas au cas multi-espèces.

Numériquement, nous avons observé dans cet article que la décroissance de l'entropie et la positivité de la fonction de distribution généraient des solutions sans oscillation. Cette propriété de type principe du maximum est démontrée dans le cas de FPL linéaire. Dans le cas isotrope (section suivante et [a5]), nous avons montré sur des exemples numériques l'existence de telles oscillations dans le cas non linéaire.

Comme nous l'avons rappelé, ces propriétés de conservation et d'entropie pour les schémas FPL sont satisfaites en utilisant la formulation dite Landau-log qui a fait l'objet de nombreux travaux [57, 47, 100, 48, 92, 76, 99]. Dans ce papier, nous montrons que le schéma basé sur la formulation "sans"-log est bien conservatif mais qu'il existe des fonctions de distribution initiales positives qui conduisent à une solution négative après un temps arbitrairement petit.

**Remarque 2** *D'ailleurs un résultat récent que j'ai obtenu avec S. Cordier, voir [cr3] ou voir paragraphe (3.2) semble en fait indiquer qu'il n'existerait pas de schémas positifs pour la formulation "sans"-log. Une solution pour avoir la positivité est de prendre un schéma non-liéaire par exemple en partant de la formulation Landau-log.*

Les résultats principaux de ce travail sont l'existence de solutions positives pour le système d'équations différentielles ordinaires correspondant au problème semi-discrétisé (1.11) i.e. uniquement en vitesses et la vérification de la décroissance de l'entropie pour le schéma totalement discrétisé (i.e. en temps et en vitesse, avec un schéma explicite). En effet, les travaux cités précédemment vérifient que l'entropie de la solution du problème semi-discrétisé décroît mais pas celle du problème discrétisé en temps qui est effectivement implémenté dans les codes de calcul. Nous montrons que ces propriétés sont bien vérifiées pour une variante du schéma exposé auparavant.

On considère dans un premier temps l'équation de FPL linéaire (en dimension  $d$ ) qui décrit l'effet sur les électrons des collisions électrons-ions et que l'on peut écrire sous la forme

$$\frac{\partial f}{\partial t} = \nabla \cdot \left( T f \vec{\nabla}_v \log(f/M_f) \right),$$

On discrétise cette formulation par

$$\frac{df_i}{dt} = F P_i^L = (D^* \cdot p)_i, \quad p_i^s = g_i^s (D^s \log(f/M))_i, \quad (1.13)$$

où  $M$  est la maxwellienne discrète centrée de même masse et température que  $f$  et  $(D, D^*)$  sont des opérateurs de différences finis adjoints. Les coefficients  $g_i^s$  sont des approximations de  $f_i$  définis par

$$g_i^s = \frac{(\sharp N^s) \prod_{k \in N^s} f_{i+k}}{\sum_{k' \in N^s} \left( \prod_{k \in N^s - \{k'\}} f_{i+k} \right)}, \quad i \in \mathbb{Z}^d, \quad (1.14)$$

où  $N^s$  est l'ensemble des points pris en compte pour calculer l'opérateur de différences finis dans la direction  $s$  et  $(\sharp N)$  est son cardinal. Le schéma obtenu peut s'interpréter comme un schéma de type

volumes finis : en intégrant sur une maille  $C_i$  cubique centrée en  $v_i$ , avec formule de quadrature au point milieu, on obtient une formulation équivalente à (1.13)

$$\frac{df_i}{dt} = FP_i^L = \frac{1}{\Delta v^2} \sum_{\mu \in \{-1,1\}} \sum_{s=1 \dots d} g_{i,i+\mu e_s} \left( \log \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right),$$

en utilisant une approximation bien connue dans le cas des équations de diffusion [74] qui consiste à prendre la moyenne harmonique

$$g_{i,i+\mu e_s} = \frac{2f_i f_{i+\mu e_s}}{f_i + f_{i+\mu e_s}} \approx f \left( v = \frac{v_i + v_{i+\mu e_s}}{2} \right),$$

qui garantit la continuité des flux aux interfaces. On montre ensuite comment se ramener à un domaine borné  $I$  dans l'espace des vitesses et également que le problème de Cauchy est bien posé : soit  $(f_i^0) \in \mathbb{R}^{\#I}$  avec  $f_i^0 > 0$  pour tout  $i \in I$ ; le système (1.13) avec  $f_i(t=0) = f_i^0$  possède une solution positive, entropique, globale en temps telle que  $\forall i \in I, \lim_{t \rightarrow \infty} f_i(t) = M_i$ . On montre qu'il existe une constante  $C > 0$  - dépendant de la donnée initiale - telle que pour des pas de temps de la forme  $C/\Delta v^2$  la solution discrète en temps est positive, entropique et converge vers la Maxwellienne discrète en temps grand.

Dans le cas non linéaire, en utilisant l'approximation des  $f_i$  par les moyennes harmoniques généralisées  $g_i^s$  et en étudiant l'évolution de

$$K \stackrel{def}{=} \sup_{i \in I, k \in N} \left| \frac{f_i}{f_{i+k}} \right|,$$

on montre qu'il existe une solution entropique et positive pour des temps arbitrairement grands au problème de Cauchy analogue à (1.13) avec  $FPL$  défini cette fois par (1.10) où le produit  $f_i f_j$  est remplacé par  $g_i g_j$  défini par (1.14). Dans le cas linéaire, la fonction  $K$  est bornée ; dans le cas non linéaire, elle est bornée en temps fini. En particulier, la fonction de distribution pourrait s'annuler en quelques points (nécessairement multiples) lorsque  $t \rightarrow \infty$  et le problème de la convergence vers l'ETL pour le problème semi-discret reste ouvert.

De façon analogue, pour le problème discrétisé en temps, on n'a pas une borne supérieure du pas de temps qui assure la décroissance de l'entropie et la positivité mais une suite de pas de temps dont la série diverge. On sait donc qu'il existe une solution pour le problème semi-discret et discrétisé en temps (avec la modification de  $f$  en  $g$ ) qui est positive, entropique et globale en temps mais on ne peut montrer le retour vers l'ETL. En pratique, les pas de temps restent bornés inférieurement et la solution s'approche de la Maxwellienne discrète. Les pas de temps entropiques sont donnés par la condition

$$\Delta t_H \stackrel{def}{=} \frac{-\sum_{i \in I} FP_i \log(f_i)}{\sum_{i \in I} FP_i^2 / f_i}.$$

## 1.6 Équation de Fokker-Planck-Landau isotrope.

Nous avons également travaillé sur un modèle simplifié (symétrie sphérique des fonctions de distribution) pour lequel les résultats du cas tri-dimensionnel peuvent être améliorés [a5].

L'opérateur de Fokker-Planck-Landau peut s'écrire sous une forme simplifiée pour des fonctions de distribution possédant des propriétés de symétrie. En particulier, lorsque la fonction de distribution a un axe de révolution, ce qui est le cas en présence de champ magnétique par exemple; ce cas appelé cas axisymétrique a fait l'objet de travaux récents de Frenod-Lucquin [76] et Lemou [85]. Lorsque la fonction de distribution possède un centre de symétrie, c'est à dire quand la fonction est indépendante de la direction de la vitesse, on obtient le modèle isotrope considéré ici. L'opérateur satisfait alors les mêmes symétries et la solution reste donc symétrique. L'opérateur de FPL dans le cas isotrope est utilisé pour la modélisation des phénomènes de fusion par confinement inertiel (FCI). Plus précisément, il s'agit de décrire précisément le transport d'énergie dans



un plasma produit par un laser. Dans certaines conditions, il est admis que la théorie du transport fluide (dans lequel on ferme les équations hydrodynamiques par une loi pour le flux de chaleur, voir Spitzer-Harm [119]) n'est pas valable. L'opérateur de FPL isotrope peut également être considéré comme le premier terme du développement de FPL en harmonique sphérique (modèles SHE). Nous renvoyons à [72, 73] pour une présentation des modèles physiques et des méthodes numériques pour les résoudre. Outre l'intérêt intrinsèque de l'opérateur isotrope, que l'on rencontre également en astrophysique [50, 51], celui-ci sert à calculer des solutions de références dans le cas Coulombien [101, 102] pour lequel il n'existe pas de solutions exactes contrairement au cas Maxwellien [83].

Au niveau numérique, un schéma conservatif et entropique a été proposé dans le cas isotrope [47]. Les auteurs donnent des conditions sur les pas de temps pour assurer la décroissance de l'entropie sans justifier ces propriétés. Comme on l'a déjà dit, la décroissance de l'entropie est importante pour assurer le retour vers l'équilibre mais aussi pour empêcher la formation d'oscillations parasites. Celles-ci sont particulièrement visibles sur une fonctionnelle appelée information de Fischer. Les solutions discrètes doivent également rester positives et cela n'apparaît pas dans [47]. Rappelons qu'un schéma peut être conservatif et ne pas préserver la positivité, comme on l'a vu avec le schéma sans log dans le cas tridimensionnel. Il est annoncé dans [47] que dans le cas isotrope, les propriétés (conservation, entropie) sont satisfaites sur la forme sans log. Nous en donnons la preuve moyennant une modification utilisant les moyennes entropiques. Ceci nous permet également de montrer que la solution discrète tend en temps grand vers l'ETL.

Les deux sections suivantes décrivent les résultats obtenus : la première est une adaptation et une amélioration des résultats obtenus dans le cas tridimensionnel (section 1.5.2) pour le schéma basé sur la formulation log; la deuxième est basée sur autre formulation, sans log, pour laquelle on vérifie les propriétés de conservation, entropie et états d'équilibre et qui permet de proposer de nouveaux schémas en temps.

### 1.6.1 Existence de solutions pour FPL log [a5]

*En coll. avec S. Cordier.*

Nous nous intéressons à l'opérateur de FPL pour les fonctions de distribution isotropes i.e. qui ne dépendent que de la variable d'énergie  $\varepsilon$  et du temps  $f(\varepsilon, t)$ . On ne note pas la dépendance en temps pour simplifier. Cet opérateur s'écrit (voir [47]) une fois ramené à un domaine borné en énergie et adimensionné, dans le cas Coulombien

$$\frac{\partial f(\varepsilon)}{\partial t} = \frac{1}{\sqrt{\varepsilon}} \frac{d}{d\varepsilon} \int_0^{\varepsilon_0} f(\varepsilon) f(\varepsilon') \left( \frac{d}{d\varepsilon} \ln f(\varepsilon) - \frac{d}{d\varepsilon} \ln f(\varepsilon') \right) k(\varepsilon, \varepsilon') d\varepsilon',$$

avec  $k(\varepsilon, \varepsilon') = \inf(\varepsilon^{3/2}, (\varepsilon')^{3/2})$  et  $\varepsilon_0$  assez grand pour que la fonction  $f$  soit bien représentée. On considère une forme faible de cet opérateur

$$\begin{aligned} \int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon} d\varepsilon &= -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} f(\varepsilon) f(\varepsilon') \left( \frac{\partial \phi(\varepsilon)}{\partial \varepsilon} - \frac{\partial \phi(\varepsilon')}{\partial \varepsilon} \right) \\ &\quad \left( \frac{\partial \ln f(\varepsilon)}{\partial \varepsilon} - \frac{\partial \ln f(\varepsilon')}{\partial \varepsilon} \right) k(\varepsilon, \varepsilon') d\varepsilon' d\varepsilon, \end{aligned} \quad (1.15)$$

qui vérifie les conservations de la masse (resp. l'énergie) en prenant  $\phi = 1$  (resp.  $\phi = \varepsilon$ ) dans (1.15). L'entropie définie par

$$H = \int_0^{\varepsilon_0} f(\varepsilon) \ln(f(\varepsilon)) \sqrt{\varepsilon} d\varepsilon,$$

décroît en temps (prendre  $\phi = \ln(f)$  dans la formulation faible) et on a le théorème  $H$ . Les états d'équilibre sont de la forme

$$\partial_t H = 0 \Leftrightarrow f(\varepsilon) = \exp(-A\varepsilon + B).$$

Le problème est plus simple que dans le cas tridimensionnel. Nous montrons d'abord l'existence d'une unique solution, globale en temps pour le problème semi-discret qui est décrit dans la section

suivante. Ce résultat est obtenu en considérant à nouveau une moyenne harmonique comme dans le cas tridimensionnel (1.14). Ensuite, pour le problème discrétisé en temps, nous obtenons une borne sur les pas de temps pour assurer la positivité et la décroissance de l'entropie.

En outre, nous montrons que l'évaluation de cet opérateur peut-être réalisée avec un coût proportionnel au nombre de points de maillage malgré le caractère quadratique de l'opérateur. Nous expliquons également qu'il est possible dans ce cas (isotrope) de considérer un maillage arbitraire alors que les travaux précédents dans le cas 3D [a3, a6, 84] nécessitent un maillage uniforme. Ceci permet de raffiner le maillage pour les faibles énergies et donc d'obtenir des solutions très précises. On montre également sur quelques tests numériques que si la condition sur le pas de temps pour rester entropique est relaxée, des oscillations apparaissent sur la fonction. Ces oscillations sont particulièrement visibles sur la fonctionnelle de Linnick ou information de Fischer

$$L(t) = \int_{\varepsilon \geq 0} \left( \frac{\partial f}{\partial \varepsilon} \right)^2 \frac{\varepsilon^{3/2}}{f} d\varepsilon,$$

dont on ne sait pas montrer si elle est monotone sauf dans le cas linéaire [104]. Quelques questions restent ouvertes comme le comportement en temps grand de la solution à la fois pour le problème semi-discrétisé et discrétisé en temps, bien qu'on observe le retour vers l'ETL de la solution.

### 1.6.2 Forme sans log et schémas en temps [a9]

*En coll. avec S. Cordier.*

Dans cette partie, nous nous intéressons à nouveau aux opérateurs de FPL dans le cas isotrope. Nous considérons une autre moyenne que la moyenne harmonique, la moyenne dite "entropique" :

$$g_{i,j} \stackrel{def}{=} \frac{f_i Df_j - f_j Df_i}{D(\ln f)_j - D(\ln f)_i},$$

si  $D(\ln f)_j \neq D(\ln f)_i$  et  $f_i f_j$  sinon (mais les termes correspondants ont une contribution nulle à l'opérateur de collisions). Cette expression est une approximation d'ordre 1 du produit  $f_i f_j$  sauf pour une grille uniforme où elle est d'ordre 2. Cette moyenne a également été utilisée dans [a7]. Pour un tel choix, l'opérateur semi-discrétisé qui s'écrivait :

$$\sum_{i=1}^N c_i \frac{\partial f_i}{\partial t} \phi_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} \Delta \varepsilon_i \Delta \varepsilon_j (D\phi_i - D\phi_j) (D(\ln f)_i - D(\ln f)_j), \quad (1.16)$$

devient

$$\sum_{i=1}^N c_i \frac{\partial f_i}{\partial t} \phi_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} k_{i,j} \Delta \varepsilon_i \Delta \varepsilon_j (D\phi_i - D\phi_j) (f_j Df_i - f_i Df_j).$$

On a donc dans ce cas une façon de passer de la formulation discrète avec "log" (qui permet de montrer la décroissance de l'entropie) à la formulation sans log qui a une structure quadratique. Ceci n'est valable que pour un maillage uniforme en énergie et n'est pas généralisable facilement au cas tridimensionnel. Nous montrons l'existence d'une solution strictement positive en utilisant la décomposition de l'opérateur en terme de perte et de gain qui est classique pour l'opérateur de Boltzmann. Nous en déduisons une borne inférieure en utilisant l'inégalité de Csizar-Kullback

$$\|f - M\|_{L^1}^2 \leq 2H(f\|M),$$

où  $H(f\|M)$  est l'entropie relative discrète. Ceci nous permet de conclure que  $f(t)$  tend vers l'ETL pour une suite de temps  $t$  tendant vers  $+\infty$ . Il s'agit du premier exemple à notre connaissance de discrétisation de l'opérateur de FPL pour lequel on sache montrer cette propriété.

Nous nous intéressons ensuite à la discrétisation en temps de ces opérateurs. On considère dans un premier temps un schéma explicite basé sur la structure quadratique de l'opérateur. En effet, l'opérateur défini par (1.16) peut s'écrire comme une somme de systèmes à quatre vitesses

généralisés [a1, a2]. Nous obtenons une limitation sur le pas de temps pour que le schéma soit positif et entropique qui dépend de la norme sup de la solution (pour laquelle il n'existe pas d'estimations). Le schéma explicite dont le coût est un  $O(N^3)$  n'est en fait pas plus coûteux que le schéma implicite basé sur une linéarisation de l'opérateur de collision, proposé par Epperlein [73], et qui nécessite l'inversion d'une matrice pleine.

Nous étudions aussi des schémas d'ordre supérieur en temps pour lesquels le schéma explicite fournit une phase de prédiction qui est corrigée. Ces schémas vérifient également la positivité et la décroissance de l'entropie.

Nous présentons quelques résultats numériques en particulier pour une distribution initiale très singulière, de type Dirac en énergie, qu'on ne pouvait traiter avec les schémas précédents basés sur la formulation log.

**Remarque 3** *En fait il est possible de définir un schéma en temps implicite et positif suivant en cela ce que j'ai fait pour l'équation de Fokker-Planck pour les milieux granulaires, paragraphe (1.10) ou [a14]. Les schémas en temps implicite proposés par [73] ou par [86] ne sont pas positifs et pour le cas du schéma proposé dans [73] aussi coûteux que le schéma explicite en temps. Le coût du procédé itératif que nous avons proposé pour l'équation de Fokker-Planck des milieux granulaires est quant à lui linéaire par rapport au nombre de points.*

## 1.7 Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique [a7, cr1]

*En coll. avec S. Dellacherie et R. Sentis.*

Cet article décrit un schéma numérique pour le traitement d'un opérateur de collision ion-électron de type Fokker-Planck ; pour cela on introduit la notion de *moyenne entropique* de deux quantités positives. Ce schéma a la propriété d'être entropique au sens du théorème H de Boltzmann sous un critère de type *CFL*. On montre de plus que la solution converge en temps grand vers un unique état d'équilibre Maxwellien.

L'évolution d'une population  $f = f(t, x, v)$  d'ions (de masse  $m$  et de charge  $Z$ ) et de la température électronique  $T_e = T_e(t, x)$  (où  $x \in \mathbf{R}^3$  et  $\vec{v} \in \mathbf{R}^3$ ) est régie par le système

$$\frac{\partial}{\partial t} f = S(f), \quad (1.17)$$

$$\frac{\partial}{\partial t} \mathcal{E}_e(T_e) = -\frac{m}{2} \langle v^2 S(f) \rangle \quad (1.18)$$

où  $\mathcal{E}_e(T_e) = \frac{3}{2} Z N T_e$ ,  $P_e = Z N T_e$ ,  $N = \langle f \rangle$ ,  $N \vec{U} = \langle f \vec{v} \rangle$ , et

$$S(f)(\vec{v}) = \Omega \nabla_v \cdot \left[ (\vec{v} - \vec{U}) f(v) + \frac{T_e}{m} \nabla_v f \right].$$

En introduisant une température ionique  $T$  définie par  $3NT = m \langle (\vec{v} - \vec{U})^2 f \rangle$ , l'opérateur  $S$  vérifie

$$\langle S(f) \rangle = 0, \quad \langle S(f) \vec{v} \rangle = 0, \quad \frac{m}{2} \langle S(f) v^2 \rangle = 3\Omega N (T_e - T).$$

Définissons la Maxwellienne  $\mathbf{M}_{N, \vec{U}, T}(\vec{v}) = \frac{N}{(2\pi T/m)^{3/2}} \exp \left[ -\frac{m(\vec{v} - \vec{U})^2}{2T} \right]$ .

Le schéma entropique et conservatif pour (1.17-1.18) est construit à partir de la formulation faible logarithmique de l'opérateur de collision  $S$ :

$$S(f) = \Omega \nabla_v \cdot [f \nabla_v \log(f/\mathbf{M})]$$

Pour simplifier la présentation plaçons-nous dans un cadre monodimensionnel en  $\vec{v}$  et  $\mathcal{E}_e(T_e)$  devient alors  $\frac{1}{2} Z N T_e$ . L'extension au 3-D est triviale.

On définit une approximation discrète de  $S$  en tout point d'une grille uniforme  $i\Delta v$ ,  $i \in \mathbb{Z}$  par :

$$S(f) \frac{m\Delta v^2}{\Omega T_e} = \tilde{f}_{j+1/2} \left( \log\left(\frac{f}{\mathcal{M}_{\tilde{U}, T_e}}\right)_{j+1} - \log\left(\frac{f}{\mathcal{M}_{\tilde{U}, T_e}}\right)_j \right) - \tilde{f}_{j-1/2} \left( \log\left(\frac{f}{\mathcal{M}_{\tilde{U}, T_e}}\right)_j - \log\left(\frac{f}{\mathcal{M}_{\tilde{U}, T_e}}\right)_{j-1} \right) \quad (1.19)$$

Pour éliminer les logarithmes de  $f_{j+1/2}$  qui ne sont pas souhaitables pour traiter des distributions de type masse de Dirac et aussi pour avoir un schéma sous la forme plus standard diffusion + convection, nous avons pour cela introduit la moyenne entropique de deux quantités positives  $x$  et  $y$

$$\tilde{m}_{x,y} = \frac{x-y}{\log x - \log y} \text{ si } x \neq y, \quad \tilde{m}_{x,y} = x \text{ sinon.}$$

On prend donc

$$f_{j+1/2} = m \left( \log\left(\frac{f}{\mathcal{M}_{\tilde{U}, T_e}}\right)_{j+1}, \log\left(\frac{f}{\mathcal{M}_{\tilde{U}, T_e}}\right)_j \right)$$

ce qui fournit donc le schéma entropique et conservatif sans logarithme suivant

$$\left\{ \begin{array}{l} \frac{\partial}{\partial t} f_j = \frac{\Omega}{\Delta v} \left[ (v_{j+1/2} - \tilde{U}) \tilde{f}_{j+1/2} - (v_{j-1/2} - \tilde{U}) \tilde{f}_{j-1/2} \right] + \frac{\Omega T_e}{m \Delta v^2} (Af)_i \\ \frac{\partial}{\partial t} \mathcal{E}_e(T_e) + \frac{m}{2} \langle v_j^2 S(f) \rangle = 0 \end{array} \right. \quad (1.20)$$

et  $(Af)_i = a_j(f_{j+1} - f_j) - b_j(f_j - f_{j-1})$  avec  $a_j = 1$  (sauf  $a_{j_{\max}} = 0$ ),  $b_j = 1$  (sauf  $b_1 = 0$ ) n'est rien d'autre que le laplacien discret usuel.

Nous avons comparé les résultats obtenus avec le schéma utilisant la moyenne entropique, et la moyenne arithmétique, et aussi avec le schéma de Chang et Cooper [49] sur un maillage fin et sur un maillage grossier. Pour la simulation appliquée à la Fusion par Confinement Inertiel, comme la température est très forte en fin de calcul, il faut avoir un  $\Delta v$  adapté, ce qui conduit à une discrétisation sur quelques mailles seulement de la fonction de répartition initiale  $f^0$  où la température est beaucoup plus faible. Lorsque le maillage est grossier, on constate que la moyenne entropique donne un résultat plus précis que le schéma de Chang et Cooper et que la moyenne arithmétique ne preserve plus la positivité.

Le schéma proposé s'étend en géométrie bidimensionnelle axisymétrique sans difficulté. Par ailleurs, on peut impliciter facilement la partie diffusion dans (1.20) ce qui permet des pas de temps plus grands tout en gardant d'excellents résultats numériques; on montre que l'équilibre thermodynamique  $f^\infty$  est encore solution stationnaire avec ce schéma semi-implicite.

**Remarque 4** Avec  $S$ , Dellacherie nous avons montré plus tard que le schéma de Chang et Cooper est en fait un schéma entropique de la forme 1.19 (voir [np1]).

## 1.8 Méthode numérique pour l'opérateur de scattering de Compton [a12]

*En coll. avec S. Cordier.*

On s'intéresse à un gaz de photons isotrope et homogène, décrit par la densité  $f = f(t, k) \geq 0$  de photons qui à l'instant  $t \geq 0$  possèdent l'énergie  $k > 0$  qui vérifie l'équation de Boltzmann quantique

$$k^2 \frac{\partial f}{\partial t} = \int_0^\infty (f'(1+f)B(k', k) - f(1+f')B(k, k')) dk', \quad (1.21)$$

où l'on note  $f' = f(t, k')$ . La section efficace  $B(k, k')/k^2$  représente la probabilité de transition par scattering de l'état d'énergie  $k$  à l'état d'énergie  $k'$ . On vérifie alors, au moins formellement, que pour toute solution  $f$  de (1.21), on a conservation de la masse  $N(f)$  et décroissance d'une entropie  $S(f)$  définie dans [127]. Il est alors naturel de penser que les états d'équilibre de (1.21) sont les états qui maximisent l'entropie à masse donnée. Les distributions de Bose-Einstein (cas  $\mu > 0$ ) et de Planck (cas  $\mu = 0$ ) définies par

$$f_\mu(k) = \frac{1}{e^{k+\mu} - 1}.$$

On vérifie aisément que les  $f_\mu$  sont des solutions stationnaires de (1.21) i.e.  $Q(f_\mu, f_\mu) = 0$  et que  $f_\mu$  est solution du problème de maximisation  $S(f_\mu) = \max_{N(f)=N} S(f)$ , avec  $\mu$  défini par  $N_\mu = N(f_\mu) = N$  lorsque  $N \leq N_0$ . De plus, l'état de Planck  $f_0$  est le maximum global de l'entropie, i.e.  $S(f_0) = \max S(f)$  et satisfait  $N(f_0) < \infty$ . On peut alors se demander quelle est la solution du problème de maximisation de l'entropie pour  $N > N_0$ . Caflisch et Levermore montrent dans [123] que les fonctions de masse supérieure à celle de l'état de Planck  $f_0$  qui maximisent l'entropie sont égales à la somme de  $f_0$  et d'une masse de Dirac en 0. Mischler et Escobedo étudient le problème d'évolution et montrent la convergence faible des solutions vers ces états d'équilibre singuliers à l'origine et la convergence forte dans  $L^1([k_0, \infty))$  fort (pour tout  $k_0 > 0$ ) lorsque  $t \rightarrow \infty$  [127]. Soulignons deux conséquences de leur théorème. Si l'on part d'une donnée initiale régulière, la solution reste régulière pour tout temps. De plus, si  $N = N(f(t=0)) > N_0$  alors  $f(t, \cdot) \rightarrow f_0 + \alpha \delta_0$  et  $\alpha = N - N_0 > 0$ , i.e. un état initial régulier de masse supérieure à la masse de l'état de Planck  $N_0$  "condense à l'origine" en temps infini.

Nous nous sommes intéressées plus particulièrement à la discrétisation de la limite collision rasantes de (1.21) qui est l'équation de Kompaneets [128]:

$$\partial_t f(k, t) = k^{-2} \partial_k (k^4 (f + f^2 + \partial_k f)).$$

Cette équation présente un autre comportement singulier par rapport à l'équation de Boltzmann (1.21) comportement mis en évidence dans [126]: pour un certain type de conditions initiales vérifiant quand même  $N < N_0$  on a un phénomène d'explosion en temps fini à l'origine.

La méthode développée est évidemment entropique et conserve l'énergie des photons. La discrétisation en temps est implicite.

Les résultats numériques montrent bien les deux comportements singuliers attendus: lorsque la masse de la condition initiale dépasse celle de l'état de Planck, on voit apparaître une concentration à l'origine en temps grand et pour les conditions initiales permettant une explosion en temps fini on assiste bien à la formation d'une concentration de masse à l'origine en temps fini.

## 1.9 Analyse spectrale de l'opérateur de Lorentz [a8]

*En coll. avec S. Cordier et B. Lucquin.*

Dans ce travail, on s'intéresse à la limite collision rasante d'un opérateur de collision élastique de type Boltzmann:

$$Q^\varepsilon(f) = \int_{S^{d-1}} B^\varepsilon(\omega - \omega') [f(\omega') - f(\omega)] d\omega', \quad (1.22)$$

où  $S^{d-1}$  est la sphère unité  $\mathbb{R}^d$  de dimension  $d = 2, 3$  et  $B^\varepsilon$  est une suite de sections efficaces de collisions qui se "concentrent sur les faibles déviations". Plus précisément,  $B^\varepsilon$  est une fonction positive qui ne dépend que de l'angle de déviation  $\theta$  entre les vitesses  $\omega$  et  $\omega'$  de la forme (voir [65]) (cas 1):

$$B^\varepsilon(\theta) = \frac{1}{\varepsilon^3} \sigma\left(\frac{\theta}{\varepsilon}\right) \sin\left(\frac{\theta}{\varepsilon}\right) \chi_{[0, \varepsilon\pi]}(\theta),$$

où  $\chi_{[a,b]}$  est la fonction caractéristique de  $[a,b]$  et  $\sigma$  une fonction positive. Ce choix exprime que les collisions se concentrent bien vers les petits angles. Cependant, cela ne permet de traiter le cas

Coulombien qui correspond à des sections efficaces de la forme (cas 2) :

$$B^\varepsilon(\theta) = \sigma(\theta) \frac{1}{\log\left(\frac{1}{\sin(\frac{\theta}{2})}\right)} \frac{\sin(\theta)}{[\sin(\frac{\theta}{2})]^4} \chi_{[\varepsilon, \pi]}(\theta).$$

Dans ce cas, le petit paramètre  $\varepsilon$  que les physiciens appellent "paramètre plasma" a une signification physique, liée au nombre de particules dans la sphère de Debye (voir [56]).

Lorsque la fonction  $f$  est suffisamment régulière (au moins  $C^3$ ) et que les sections efficaces possèdent des moments bornés

$$\int_0^\pi \sigma(\theta) \sin(\theta) \theta^2 d\theta < +\infty, \quad \text{cas 1,}$$

on montre que

$$\lim_{\varepsilon \rightarrow 0} (Q^\varepsilon(f)(\omega) = C \Delta_\omega f(\omega)),$$

où  $C$  dépend des moments de  $\sigma$  et  $\Delta_\omega$  est l'opérateur de Laplace-Beltrami. Ceci permet de montrer que la solution du problème homogène sur  $S^d$  associée à l'opérateur de Boltzmann-Lorentz (1.22) converge vers celle associée à l'opérateur de Laplace-Beltrami pour tout  $t \in [0, T]$  où  $T$  est un temps arbitraire.

Ces opérateurs peuvent être dérivés à partir des opérateurs de Boltzmann ou Fokker-Planck en considérant des mélanges d'espèces de masses différentes [59, 61, 62]. On rappelle que la limite collisions rasantes de Boltzmann vers Fokker-Planck a été étudiée dans [65, 56]. Dans notre cas, l'opérateur est plus simple mais d'une part, nous pouvons améliorer le résultat en obtenant une convergence de  $f^\varepsilon$  vers  $f$  uniforme en temps et d'autre part, on contrôle la vitesse de retour vers les fonctions d'équilibre : on montre que la vitesse de relaxation des solutions à  $\varepsilon$  donné tend vers celle du système limite. Ce résultat est basé sur une analyse spectrale des opérateurs. Notons que les deux types d'opérateur ont la même base de fonctions propres. En dimension trois, il s'agit des harmoniques sphériques  $Y_{l,m}$  qui forment une base orthonormée de l'espace  $L^2(S^2)$ . Il est bien connu que ce sont des fonctions propres de l'opérateur de Laplace-Beltrami associées à des valeurs propres  $\nu_l = -l(l+1)$ . Elles sont également fonction propres de  $Q^\varepsilon$  (voir [122, 110]) avec des valeurs propres qui ne dépendent que de  $l$  données par

$$\nu_l^\varepsilon = 2\pi \int_0^\pi [1 - P_l(\cos(\theta))] B^\varepsilon(\theta) d\theta.$$

où  $P_l$  sont les polynômes de Legendre. On montre qu'il existe des constantes  $C^\pm$  telles que

$$C^- \nu_l^\varepsilon \leq \nu_l \leq C^+ \nu_l^\varepsilon, \quad \lim_{\varepsilon \rightarrow 0} \nu_l^\varepsilon = \nu_l.$$

Ces travaux ont fait l'objet d'applications numériques [52].

## 1.10 Méthode numérique pour une équation de type Fokker-Planck modélisant les milieux granulaires [a14]

*En coll. avec S. Cordier et V. Dos Santos.*

Dans un milieu granulaire, Chaque grain interagit par des collisions quasi instantanées, proche du modèle classique d'un gaz. Comme les grains peuvent être sans cohésion, ils doivent interagir comme des sphères dures, sans forces de grande portée.

La différence importante entre les collisions du granulaire et les particules des gaz parfaits, se trouve dans l'inélasticité des collisions entre les grains. L'énergie perdue lors d'une collision est exprimé par la diminution de la vitesse relative, dans le repère du centre de masse :

$$v' - u' = -h(v - u)$$

où  $v, u$  et  $v', u'$  sont les vitesses des grains, respectivement avant et après collision, et  $0 \leq h \leq 1$  est le coefficient de restitution. Les collisions parfaitement élastiques correspondent à  $h = 1$ , alors que les collisions parfaitement inélastiques sont obtenues pour  $h = 0$ . Dans le premier cas, les particules ont après collisions les mêmes vitesses de centre de masse.

Ce travail porte sur l'étude d'une méthode numérique pour une équation de type Fokker-Planck pour un milieu granulaire monodimensionnel ([134, 130, 131]). Cette équation représente une approximation cinétique d'un système de particules quasi-élastique dans un bain thermique.

Dans cette méthode, on utilise un "splitting" entre l'équation de transport et l'opérateur de collisions. La partie transport peut être traitée en utilisant des schémas "upwind". Par conséquent, nous nous intéressons uniquement à l'équation homogène.

Le modèle considéré, est une équation de type Fokker-Planck, représentant la limite quasi élastique d'un modèle granulaire de type Boltzmann

$$d_t f = \partial_v (\lambda F f + \beta(v - u)f + \sigma \partial_v f),$$

où  $F$  est le terme du granulaire pur, définie ci-dessus comme l'accélération, et  $F$  s'écrit

$$F(v) = \int_{\mathbb{R}} |v - v'| (v - v') f(v') dv'$$

et  $\lambda, \beta, u, \sigma$  sont des constantes arbitraires. D'autre part  $\sigma \partial_v^2 f$  représente le bain thermique, où  $\sigma$  est lié à cette température (apport d'énergie par agitation). Ce terme "remplace" le terme de diffusion  $D$ .  $\beta \partial_v (v - u)f$  est le terme de friction. Définissons  $\rho$  et  $u_f$  comme la masse et la vitesse moyenne respectivement

$$\rho = \int_{\mathbb{R}} f(v') dv', u_f = 1/\rho \int_{\mathbb{R}} f(v') v' dv'$$

Les propriétés de ce modèle sont la conservation de la masse (et du moment quand  $\beta = 0$ ) et la décroissance de l'énergie (pour  $\beta = \sigma = 0$ ) et de l'entropie définie par :

$$H = \int \sigma [v^2 + \ln(f(v))] f(v) dv + \frac{1}{6} \int_v \int_{v'} |v - v'|^3 f(v') f(v) dv' dv$$

Pour ce modèle d'équation de Fokker-Planck, on a bien l'existence et l'unicité d'un état d'équilibre (cf. [132]) mais on n'a pas d'expression analytique de l'état d'équilibre en fonction des quantités conservées.

Les difficultés liées à la discrétisation de cette équation sont la prise compte de la conservation de la quantité de mouvement et à l'absence d'expression analytique pour l'état d'équilibre. D'autre part il nous paraissait nécessaire d'obtenir une méthode qui pour un maillage fixé nous permettent de traiter des valeurs de la température du bain thermique quelconque, donc éventuellement nulle. Une méthode numérique analogue à celle utilisée pour l'équation de Landau ne permet pas d'une façon générale d'assurer ce second point.

Le schéma proposé est construit par une méthode inspirée de celle de Chang et Cooper, méthode utilisée couramment pour les équations de Fokker-Planck linéaire.

Les atouts bien connus de méthode de Chang et Cooper sont sa simplicité (pas de logarithme), c'est un schéma monotone et qui préserve les états d'équilibre. Mais en fait on peut montrer que ce schéma est entropique, [np1] :

Soit l'équation de type Fokker Planck linéaire (FPL), similaire à celle du granulaire, mais avec  $F = v$ ,  $\beta = 0$

$$\partial_t f = \partial_v (v f + \sigma \partial_v f), \tag{1.23}$$

qu'on peut réécrire sous la forme

$$\partial_t f = \sigma \partial_v (M \partial_v (f/M))$$

où  $M$  est une Maxwellienne, donnée par  $M(v) = \exp(-\|v\|^2/2\sigma)$ .

On recherche alors une discrétisation de la seconde forme (formulation faible) i.e.

$$\partial_t f_i \phi_i = \sigma(D\phi)_{i+1/2}(M_{i+1/2} D(f/M)_{i+1/2})$$

où  $Dg$  représente la différence centrée finie i.e.  $D(g)_{i+1/2} = (g_{i+1} - g_i)/\Delta v$  et les coefficients  $M_{i+1/2}$  constituent une moyenne de la valeur entre  $M_i$  et  $M_{i+1}$  à définir. En prenant

$$M_{i+1/2} = \frac{M_i M_{i+1}}{M_{i+1} - M_i} \log(M_{i+1}/M_i) \quad (1.24)$$

On trouve une discrétisation de (1.23) de la forme

$$\partial_t f_i = \frac{F_{i+1/2} - F_{i-1/2}}{\Delta v}$$

avec  $F_{i+1/2} = \frac{\sigma}{\Delta v} \tilde{M}_{i+1/2} (\frac{f_{i+1}}{\tilde{M}_{i+1}} - \frac{f_i}{\tilde{M}_i})$ . On pose  $\tilde{M}_{i+1/2} = \frac{M_i M_{i+1}}{M_{i+1} - M_i} (\ln M_{i+1} - \ln M_i)$ , d'où

$$F_{i+1/2} = \frac{\sigma}{\Delta v} (f_{i+1} - f_i) + \frac{\sigma}{\Delta v} (-1 + \frac{\tilde{M}_{i+1/2}}{M_{i+1}}) f_{i+1} + \frac{\sigma}{\Delta v} (1 - \frac{\tilde{M}_{i+1/2}}{M_i}) f_i.$$

Analysons chaque terme :

$$\begin{aligned} \frac{\sigma}{\Delta v} (1 - \frac{\tilde{M}_{i+1/2}}{M_i}) &= \frac{\sigma}{\Delta v} (1 - \frac{M_{i+1}}{M_{i+1} - M_i} (\ln M_{i+1} - \ln M_i)) \\ &= \frac{\sigma}{\Delta v} \ln \frac{M_{i+1}}{M_i} (\frac{1}{\ln M_{i+1} - \ln M_i} - \frac{M_{i+1}}{M_{i+1} - M_i}) \\ &= -\frac{\sigma}{\Delta v} \ln \frac{M_{i+1}}{M_i} (\frac{1}{\ln \frac{M_i}{M_{i+1}}} - \frac{1}{\frac{M_i}{M_{i+1}} - 1}) \end{aligned}$$

Si on pose  $w = \ln(M_i/M_{i+1})$ , on obtient :

$$\begin{aligned} \frac{\sigma}{\Delta v} (1 - \frac{\tilde{M}_{i+1/2}}{M_i}) &= \frac{\sigma}{\Delta v} w (\frac{1}{w} - \frac{1}{\exp w - 1}) \\ &= \frac{\sigma}{\Delta v} w h(w) = \frac{\sigma}{\Delta v} w \theta \end{aligned}$$

où  $\theta = h(w)$  et  $h$  est la fonction

$$h(x) = \frac{1}{x} - \frac{1}{e^x - 1}.$$

Or,  $h$  est positive sur  $\mathbb{R}$ , décroissante et comprise entre 0 et 1.

On fait de même pour le second terme :

$$\frac{\sigma}{\Delta v} (-1 + \frac{\tilde{M}_{i+1/2}}{M_{i+1}}) = \frac{\sigma}{\Delta v} w h(-w),$$

et  $h(-w) = -h(w) + 1$ , d'où

$$\frac{\sigma}{\Delta v} (-1 + \frac{\tilde{M}_{i+1/2}}{M_{i+1}}) = -\frac{\sigma}{\Delta v} w (\theta - 1).$$

Alors

$$\begin{aligned} F_{i+1/2} &= \frac{\sigma}{\Delta v} (f_{i+1} - f_i) + \frac{\sigma}{\Delta v} (-w(\theta - 1)) f_{i+1} + \frac{\sigma}{\Delta v} (w\theta) f_i \\ &= \frac{\sigma}{\Delta v} (f_{i+1} - f_i) + \frac{\sigma w}{\Delta v} (\theta f_i + (1 - \theta) f_{i+1}). \end{aligned}$$



C'est la définition du schéma de Chang et Cooper, construit à l'origine pour préserver les états d'équilibre, voir [49]. Cette construction du schéma de Chang et Cooper, que j'ai obtenu avec S. Dellacherie voir [np1], montre en fait que ce schéma est entropique. On vérifie aussi que ce schéma permet de traiter une température de bain quelconque, c'est à dire que dans la limite  $\sigma \rightarrow 0$  on obtient un schéma upwind pour

$$\partial_t f = \partial_v(vf),$$

Ce qui n'est pas le cas pour d'autres choix des poids  $M_{i+1/2}$ .

Dans le cas du granulaire, pour assurer la conservation de mouvement l'opérateur a été symétrisé,

$$\partial_t \int f \phi dv = \frac{-\sigma}{2\rho} \iint (\partial_v \phi - \partial_{v'} \phi') f f' (\partial_v \log(f/M) - \partial_{v'} \log(f/M)') dv dv'.$$

puis sur cette forme, on définit un schéma de type Chang et Cooper en partant de la formulation faible et en utilisant les moyennes de type (1.24).

L'inconvénient du schéma de Chang et Cooper est qu'il requiert la connaissance explicite de l'état d'équilibre. Dans le cas du granulaire ce n'est en fait pas gênant, car on utilise en réalité un pseudo état d'équilibre.

On a montré que le schéma ainsi construit a toutes les qualités requises, conservatif, entropique et qu'il redonne une discrétisation correcte pour une température de bain nulle.

Comme pour FPL isotrope le coût de l'évaluation de l'opérateur est linéaire bien qu'à priori quadratique(en fait les deux termes de collisions ont la même structure).

Mais plus intéressant, on a utilisé une discrétisation temporelle implicite qui peut aussi être utilisée dans le cas de FPL isotrope. L'implication est basée sur une méthode itérative non conservative en quantité de mouvement où les coefficients de drift et de diffusion sont pris à l'itérée précédente. À chaque itération la solution obtenue est positive. La conservation de la quantité de mouvement est obtenue à la convergence du processus itératif. Numériquement on montre que si l'on accepte une erreur  $\epsilon$  à chaque pas de temps sur la quantité de mouvement alors l'erreur reste en  $O(\epsilon)$  pendant toute la phase transitoire et l'on converge bien vers un état d'équilibre. Ce bon comportement nous a semblé lié à la nature entropique et conservative de l'opérateur discret.

**Remarque 5** *Nous avons testé par la suite cette façon de faire l'implicite sur l'équation de Fokker-Planck-Landau isotrope. Cela marche tout aussi bien que dans le cas du granulaire. Et contrairement aux schémas implicites proposés dans [86, 73], on est assuré de la positivité de la solution. De plus c'est moins coûteux car la matrice à inverser est tridiagonale et non pleine.*

## 1.11 Conclusions

Quelles conclusions tirer aujourd'hui de ces études sur la discrétisation de quelques opérateurs de collision par des méthodes à répartition discrète de vitesses?

Ces schémas constituent bien une alternative sérieuse aux méthodes de Monte-Carlo ou aux schémas de type différence finies classiques, même pour l'équation de Boltzmann, mais il y a une exception : l'opérateur de Landau dans le cas non isotrope.

Pour l'opérateur de Landau 3-d ou 2-d axisymétrique il faudrait revoir la discrétisation de cet opérateur pour pouvoir traiter par exemple des fonctions de distribution très piquées, par exemple des masses de Dirac, ou localement nulles. Donc obtenir un schéma sans logarithmes serait une solution. Mais comme signalé au paragraphe (1.5.2) un résultat, [cr3], concernant les schémas pour l'équation de la chaleur. anisotrope avec une direction de diffusion quelconque, semble en fait indiquer qu'il n'y aurait pas de schémas quadratiques positifs sur un maillage uniforme pour l'équation de Landau. Le choix d'un schéma avec logarithmes ne semble pas non plus, garantir la positivité de la solution, voir (1.5.2), et la modification proposée pour garantir la positivité a pour effet néfaste d'empêcher l'évolution d'une fonction de distribution qui serait par exemple la somme de deux masses de Dirac. Il serait aussi souhaitable, toujours dans le cas de l'équation de Landau 3-d ou 2-d axisymétrique, de pouvoir utiliser un schéma implicite en temps. Autrement,

sans ces deux améliorations, ce type de discrétisation de l'opérateur de Landau risque fort de rester un travail académique.

Autre petite difficulté pour l'équation de Landau, le cas isotrope, pour lequel il semble impossible, bien que ce soit un problème 1-D, d'étendre la méthode conservative et entropique sans logarithme au cas d'un maillage non uniforme en énergie (pour la formulation avec "log" c'est trivial). On peut définir une méthode positive, sans logarithme, conservative et préservant les états d'équilibre, mais le théorème H n'est pas vérifié au sens strict. On obtient un résultat plus faible du type

$$\sum_i Q_i(f) \overline{\log}(f_i) \leq 0,$$

où  $\overline{\log}(x)$  est une approximation de la fonction  $\log(x)$ , ce qui est doit être suffisant pour assurer le retour à l'équilibre.

## Chapitre 2

# Méthodes de moments et régimes diffusifs pour le transfert radiatif : modélisation et aspects numériques

Cette partie concerne mon nouvel axe de recherche et mes travaux les plus récents.

### 2.1 Introduction

Le fil conducteur de ce chapitre est la simulation numérique en hydrodynamique radiative en utilisant des modèles aux moments pour la partie radiative, voir par exemple [147]. Pour la simulation d'un système de moments en radiatif, il semble qu'il n'existe pas pour l'instant de codes permettant de traiter de façon satisfaisante, à la fois les zones transparentes et les zones de diffusion, [147, 143, 142], et ce, même en dimension 1 d'espace. On peut déjà dire que soit les schémas utilisés font de façon cachée l'approximation de diffusion, soit la méthode numérique ne permet pas de capturer correctement le régime de diffusion, soit la méthode capture correctement le régime de diffusion mais dans les régimes intermédiaires le schéma n'est pas positif.

Examinons donc un peu plus en détail les problèmes rencontrés et les solutions existantes.

Dans certaines situations physiques, quand les collisions prépondérantes pour les particules légères sont les collisions avec des particules lourdes, par exemple collision électron-ion ou photon-électron, il est possible de modéliser le terme de collision lourd-léger pour les espèces légères par un opérateur d'isotropisation, voire même de modéliser le milieu léger par un système d'équations de moments basées sur une approximation de type quasi-isotrope. Par exemple en transfert radiatif, en 1-d et pour une vitesse matière nulle l'équation de transport pour le transfert radiatif s'écrit, voir par exemple [148, 153, 147] :

$$\frac{1}{c} \frac{\partial}{\partial t} I + \vec{n} \cdot \frac{\partial}{\partial x} I = S_t(\nu, \vec{n}), \quad (2.1)$$

$S_t = S_a + S_s$  est la somme de deux contributions. Le premier prend en compte l'émission-absorption de photons par la matière

$$S_a(\nu, \vec{n}) = \sigma_a(\nu) [B(\nu, T) - I].$$

où  $B(\nu, T)$  est la Planckienne

$$B(\nu, T) = \frac{2h\nu^3}{c^2} \left( e^{h\nu/kT} - 1 \right)^{-1}$$

et  $\sigma_a(\nu) \geq 0$  est le coefficient d'émission-absorption.

Le second terme prend en compte le scattering des photons par la matière [148]

$$S_s(\nu, \vec{n}) = \sigma_s(\nu) \frac{1}{4\pi} \int I(\nu, \vec{n}) d\vec{n} - I(\nu, \vec{n})$$

Une classe bien connue de modèles approchés pour l'équation du transfert radiatif sont les modèles à deux moments. En prenant les moments de l'équation de transfert radiatif (2.1) contre 1 and  $\vec{n}$  on obtient la forme générale de modèles à deux moments

$$\frac{1}{c} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} F_r = \sigma_a(aT^4 - E_r), \frac{1}{c} \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} P_r = -(\sigma_a + \sigma_s) F_r \quad (2.2)$$

la fermeture pour le terme de pression  $P_r$  étant généralement obtenu soit par minimisation d'entropie soit empiriquement et est de la forme

$$P_r = E_r \chi(|\frac{F_r}{E_r}|)$$

$\chi$  étant le facteur d'Eddington et vérifie  $\chi(0) = \frac{1}{3}$ . En intégrant en fréquence ou non on obtient soit un modèle dit "gris", soit un modèle "multi-groupes". L'avantage de tels modèles étant évidemment le faible coût par rapport à celui de la résolution directe de l'équation du transfert radiatif. Concernant les propriétés de ce type de modèle on peut montrer que si le facteur d'Eddington vérifie  $x^2 < \chi(x) < 1$  alors le système (2.2) est hyperbolique et a pour domaine invariant  $E_r > 0$  et  $|F_r/E_r| \leq 1$ , la deuxième inégalité montrant que le flux d'énergie radiative est limité ce qui est naturellement le cas pour une solution positive de l'équation du transfert radiatif.

On peut aussi obtenir le même type de modèle simplifié pour les électrons en partant de l'équation de Landau et en supposant toujours que la fonction de distribution des électrons est quasi isotrope. Le terme de collision electron-electron se réduit alors à Fokker-Planck-Landau isotrope pour lequel nous avons étudié la discrétisation. Dans la terminologie du transfert radiatif c'est en fait un modèle multigroupe, on a donc un système infini de systèmes hyperbolique  $2 \times 2$  de type équations du télégraphe couplées par FPL isotrope qui assure la thermalisation de la fonction de distribution.

Une simplification supplémentaire peut être obtenue en faisant l'approximation de type diffusion  $\frac{\partial}{\partial t} F_r = 0$ , ce qui permet de remplacer un problème hyperbolique par un problème de nature parabolique (modèle de diffusion hors équilibre) plus facile à traiter au niveau numérique :

$$\frac{\partial}{\partial t} E_r - \frac{1}{3} \frac{\partial}{\partial x} (\frac{1}{\sigma_a + \sigma_s} E_r) = \sigma_a(aT^4 - E_r)$$

Dans certains régimes cette approximation de diffusion peut se justifier par une analyse asymptotique. Signalons que par cette approximation une propriété fondamentale de l'équation est perdue c'est à dire que le flux n'est plus limité, propriété qui résulte de la positivité de la solution et du choix des moments. On a aussi propagation à vitesse infinie de l'information. Il existe bien une technique, limitation de flux, pour remédier à ces problèmes. Mais le meilleur moyen de les éviter est de ne pas faire l'approximation de diffusion, Mihalas [148].

Du point de vue numérique les difficultés principales de la discrétisation d'un système de type (2.2) sont dues à l'implication d'un schéma pour un système hyperbolique non linéaire et à une diffusion numérique grande devant la diffusion physique du régime asymptotique pour toutes méthodes numériques classiques ce qui ne permet pas de capturer le régime asymptotique de diffusion. Considérons un modèle de transfert radiatif simplifié

$$\frac{1}{\varepsilon} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} F_r = 0, \quad \frac{1}{\varepsilon} \frac{\partial}{\partial t} F_r + \frac{1}{3} \frac{\partial}{\partial x} E_r = -\frac{\sigma}{\varepsilon} F_r \quad (2.3)$$

En effet dans les applications physiques intéressantes  $\varepsilon \equiv .001$ . Pour laisser l'hydrodynamique piloter le pas de temps il est nécessaire de traiter la partie radiative en implicite. D'autre part quand  $\varepsilon \equiv .001$  et  $\sigma_s = O(1)$  le modèle dégénère vers le modèle de diffusion

$$\frac{\partial}{\partial t} E_r - \frac{1}{3} \frac{\partial^2}{\partial x^2} E_r = 0. \quad (2.4)$$

Or pour une méthode de Godunov classique pour (2.3), la diffusion numérique est de l'ordre de  $\Delta x/\varepsilon$ ,  $\Delta x$  étant le pas d'espace. Si l'on veut traiter toutes les zones d'espace par la même méthode et ce quelque soit la valeur du paramètre de relaxation  $\varepsilon$  il est impératif que le modèle discret ainsi obtenu ait la même asymptotique que le modèle continu, c'est à dire que le régime limite soit une discrétisation de 2.4.

Cette problématique se retrouve d'ailleurs dans la discrétisation d'équations cinétiques du type modèle de Lorentz :

$$\frac{1}{\varepsilon} \frac{\partial}{\partial t} f + \omega \cdot \nabla f = Q(f), \quad (2.5)$$

le terme de collision  $Q(f)$  étant un opérateur d'isotropisation sur la sphère, soit Laplace Beltrami,  $\Delta f$  sur la sphère unité, ou de type BGK  $\langle af \rangle - f$  avec  $\langle a \rangle = 1$ . Notons qu'en prenant  $Pr = 1$  dans (2.2) on rentre dans ce cas, c'est en fait l'équation du télégraphe. Pour ce type de modèles la limite de diffusion est bien établie du point de vue mathématique.  $\rho = \langle f \rangle$  vérifie

$$\frac{\partial}{\partial t} \rho - \frac{1}{D} \Delta \rho = 0.$$

$D$  étant la dimension de l'espace physique dans lequel on se place.

L'équation de transfert radiatif (2.1) est évidemment du type (2.5).

Cela a conduit ces dernières années, pour ces problèmes hyperboliques, à la construction de schémas permettant de capturer correctement le régime de diffusion, "asymptotic preserving schemes", [174, 175, 172]. On peut aussi citer les travaux de Bouchut, Perthame et autres [158, 161] sur quelques cas particuliers, équations de Saint-Venant ou Euler isentropique dans lesquels il est proposé un schéma avec recentrage du terme source et permettant aussi de capturer le régime asymptotique de diffusion.

Mais ces méthodes n'apportent pas de réponses satisfaisantes dans le cadre radiatif, le flux n'étant toujours pas limité, ces schémas n'étant pas positifs. On peut illustrer ce problème par l'application de la méthode proposée par Jin, Pareschi et Toscani, [174, 175], pour l'équation du télégraphe

$$\varepsilon \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} F_r = 0, \varepsilon \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} E_r = -\frac{1}{\varepsilon} F_r \quad (2.6)$$

On notera si les données initiales  $\rho^0, j^0$  sont positives, alors les solutions de (2.6)  $\rho, j$  sont positives, ce qui dans le contexte radiatif correspond à la limitation de flux.

Après rescaling  $\bar{F}_r = \frac{F_r}{\varepsilon}$  le problème s'écrit

$$\frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \bar{F}_r = 0, \frac{\partial}{\partial t} \bar{F}_r + \frac{\partial}{\partial x} E_r = -\frac{1}{\varepsilon^2} (F_r - (1 - \varepsilon^2) E_r) \quad (2.7)$$

La dérivée en espace dans le terme source de (2.7) est alors discrétisée par une différence centrée. Sans splitting et avec un schéma upwind classique pour le terme de transport de la partie gauche de (2.7), dans les variables d'origine cela revient à prendre tout simplement des flux de la forme  $(1 - \theta) F_c + \theta F_u$ ,  $F_c$  et  $F_u$  sont respectivement les flux centrés et décentrés amont, pour (2.6) et  $\theta = \varepsilon/(\Delta x)$ . Dans les régimes intermédiaires pour  $\varepsilon$  le schéma ne peut donc pas être positif.

Cette conclusion vaut aussi pour l'application de ce schéma au modèle de Lorentz (2.5). Par contre la généralisation au cas de flux non-linéaire ou au cas multidimensionnel est triviale.

Un autre approche plus séduisante car respectant la positivité des solutions, donc limitant le flux, est le schéma de type well-balanced développé par Greenberg et Leroux, [203], et ensuite par Gosse, [164, 166, 167], et basée sur la construction d'un schéma de type Godunov après localisations des termes sources aux interfaces (en 1-d). L'inconvénient est que ce type de schéma ne se généralise pas facilement au cas de fermeture non linéaire et l'extension multi-dimensionnelle n'est pas simple.

Mon travail de recherche actuel est donc axé sur les modèles approchés en hydrodynamique radiative, modèles Euleriens et relativistes, et sur les problèmes numériques liés à leur discrétisation. En particulier je cherche à construire des schémas numériques multidimensionnels préservant l'asymptotique de diffusion et les domaines invariants pour des équations de transport ou, dans le contexte du radiatif, pour les modèles aux moments. Je présente les quelques résultats que nous avons déjà obtenus dans cette direction en collaboration avec S. Cordier ou B. Depres.

## 2.2 Limite de diffusion du modèle de Lorentz : schémas préservant l'asymptotique [a11]

*En coll. avec S. Cordier, B. Lucquin et S. Mancini.*

Dans cet article nous avons étudié des schémas numériques pour le modèle de Lorentz dans les régimes diffusifs

$$\varepsilon \partial_t f + \cos \theta \partial_x f = \frac{1}{\varepsilon} \mathcal{L}(f). \quad (2.8)$$

Ce problème est mono-dimensionnel en espace et bi-dimensionnel dans la variable vitesse  $v = (\cos \theta, \sin \theta)$ .

La fonction de distribution  $f = f(x, \theta, t)$  est une de  $(x \in \mathbb{R})$  et de l'angle de la vitesse  $\theta \in [-\pi, \pi]$  et du temps  $t > 0$ .

L'opérateur  $\mathcal{L}(f)$  est un opérateur de collision du type Lorentz. Les opérateurs de Lorentz apparaissent quand par exemple on considère des collisions élastiques entre des particules lourdes et des particules légères : c'est le terme prédominant de collisions inter-espèces pour les particules légères du à l'effet des collisions avec les lourds et que l'on voit apparaître quand on fait un développement asymptotique en termes du rapport de masse (voir [58, 91]). Il est défini dans le cas de Boltzmann par :

$$\mathcal{L}(f)(\theta) = \int_{S^1} K(\theta' - \theta) [f(\theta') - f(\theta)] d\theta' \quad (2.9)$$

et dans le cas de Fokker-Planck par :

$$\mathcal{L}(f)(\theta) = \partial_{\theta\theta}^2 f. \quad (2.10)$$

Il est bien connu dans la littérature, que lorsque  $\varepsilon \ll 1$  les solutions de (2.8) convergent vers la solution d'un problème de diffusion en espace :

$$\partial_t f^0 - \frac{1}{2} \partial_{xx}^2 f^0 = 0.$$

Notre objectif était d'obtenir un schéma utilisable pour toutes les valeurs de  $\varepsilon$ , donc compatible avec la limite de diffusion ("Asymptotic Preserving Scheme").

Cette étude est faite sur l'opérateur de Fokker-Planck-Lorentz (2.10) mais les résultats s'étendent au cas de l'opérateur de Boltzmann-Lorentz (2.9).

Dans une première partie, suivant en cela les idées de Jin et Levermore pour l'opérateur de Boltzmann-Lorentz isotropique ([170, 171]), nous avons discrétisé en vitesse sur un maillage non uniforme mais symétrique pour avoir exactement le bon coefficient de diffusion avec un faible nombre de points de quadrature. En particulier nous traitons le cas de 4 et 8 points de discrétisation.

Dans une seconde partie nous considérons la discrétisation spatiale, en domaine infini.

L'approche naturelle consiste à utiliser le schéma upwind, mais dans le régime de diffusion le coefficient de diffusion est de l'ordre de  $\Delta x / \varepsilon$ .

A contrario le schéma centré converge une discrétisation du laplacien avec le bon coefficient de diffusion mais agit sur un maillage double, ce qui se passe sur les mailles paires étant totalement découplés du maillage formé des mailles impaires ce qui conduit à des modes parasites.

On considère alors un  $\theta$ -schema ( $\varepsilon$  pour les flux upwind et  $(1 - \varepsilon)$  pour les flux centrés) : la première partie recouple tous les points de discrétisation mais introduit une erreur de l'ordre de  $\Delta x$  sur le coefficient de diffusion.

On a proposé aussi un autre schéma basé sur les idées de Jin et Levermore ([177, 172]), c'est à dire on exprime le flux aux interfaces dans une méthode de volume finis en utilisant l'équation stationnaire. Nous avons donné une interprétation de ce schéma comme un schéma d'éléments finis discontinus  $P^1$  sur un maillage de maille  $\Delta x / 2$  (voir [186]).

La discrétisation en temps est totalement implicite et sans splitting entre le transport et la phase de collision. Cela requiert l'inversion d'une très grosse matrice mais cette matrice est creuse. Toutes méthodes capturant le bon régime de diffusion (voir par exemple [198, 173, 174, 172]), requièrent l'inversion d'une matrice de taille  $N_x \times N_\theta$  où  $N_x$  est le nombre de points de discrétisation dans l'espace et  $N_\theta$  est le nombre de points de discrétisation dans l'espace des vitesses, car toutes ces méthodes, voir par exemple ([173, 174, 172]), pour l'opérateur de Lorentz, nécessitent la résolution d'un problème stationnaire du type

$$v \cdot \nabla_x f^0 = L(f)$$

et un tel problème est vraiment multi-dimensionnel, i.e. on ne peut splitter le transport en espace et la phase de collision sinon on n'a pas les bons états d'équilibre.

Signalons aussi que le schéma "well balanced" proposé récemment par Gosse et Toscani [167] pour le transfert radiatif et avec approximation de Rosseland semble impossible à étendre pour un opérateur du type Boltzmann-Lorentz (2.9) avec une section efficace générale et donc en particulier pour l'opérateur (2.10). Donc les schémas présentés dans ce travail gardent tout leur intérêt.

Des résultats numériques ont validé nos deux schémas.

L'intérêt des schémas que nous proposons est que leur extension au cas multidimensionnel en espace est triviale pour un maillage cartésien. Ce qui peut s'avérer un inconvénient c'est que ces schémas ne sont pas positifs. Mais c'est aussi le cas pour le schéma décrit dans [173, 174, 172]). Seule la discrétisation proposée par Gosse et Toscani [198, 167] est positive mais quid de son extension au cas multi-dimensionnel.

## 2.3 Analyse asymptotique pour l'hydrodynamique radiative [a13]

*En coll. avec B. Despres.*

Dans ce travail nous sommes intéressés à certains régimes asymptotiques pour le couplage fluide-radiatif. Puis nous proposons un modèle de moments dans le référentiel eulérien permettant de retrouver dans ces régimes asymptotiques les modèles standards de diffusion à l'équilibre ou hors équilibre. La motivation est le développement de méthodes numériques euleriennes modernes, précises et robustes pour un modèle simplifié de transfert radiatif.

Le système considéré couplant le transfert radiatif et l'hydrodynamique s'écrit, dans le référentiel du laboratoire :

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) = 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I}) = -\vec{S}_F, \\ \frac{\partial}{\partial t}(\rho E) + \nabla \cdot (\rho E \vec{v} + p \vec{v}) = -S_E, \\ \frac{1}{c} \frac{\partial}{\partial t} I(\nu, \vec{n}) + \vec{n} \cdot \nabla I(\nu, \vec{n}) = S_t(\nu, \vec{n}), \quad \forall \nu, \vec{n} \end{cases} \quad (2.11)$$

avec

$$S_E = \int \int S_t d\nu d\vec{n} \text{ et } \vec{S}_F = \frac{1}{c} \int \int \vec{n} S_t d\nu d\vec{n}. \quad (2.12)$$

et  $S_t = S_a + S_s$  est la somme de deux contributions.

La première prend en compte l'émission-absorption des photons par la matière qui est considérée sous la forme suivante

$$S_a(\nu, \vec{n}) = \frac{\nu_0}{\nu} \sigma_a(\nu_0) \left[ \left( \frac{\nu}{\nu_0} \right)^3 B(\nu_0, T) - I \right]. \quad (2.13)$$

$B(\nu_0, T)$  est la Planckienne

$$B(\nu_0, T) = \frac{2h\nu_0^3}{c^2} \left( e^{h\nu_0/kT} - 1 \right)^{-1} \quad (2.14)$$

Les quantités  $q_0$  correspondent aux quantités mesurées dans le référentiel comobile. Les définitions (2.13-2.14), utilisent la fréquence  $\nu_0$  et la direction du photon  $\vec{n}_0$  dans le repère comobile et qui s'expriment par

$$\nu_0 = \gamma\nu(1 - \frac{\vec{n} \cdot \vec{v}}{c}) \text{ et } \vec{n}_0 = \left(\frac{\nu}{\nu_0}\right) \left[\vec{n} - \frac{\gamma}{c}\vec{v}\left(1 - \frac{\vec{n} \cdot \vec{v}}{c}\left(\frac{\gamma}{\gamma+1}\right)\right)\right]. \quad (2.15)$$

Le second terme prend en compte le scattering des photons par la matière, que l'on considère sous la forme simplifiée suivante [148]

$$S_s(\nu, \vec{n}) = \frac{\nu^2}{\nu_0^2} (S_s)_0(\nu_0, \vec{n}_0), \quad (2.16)$$

et le scattering mesuré dans le référentiel comobile est donné par

$$(S_s)_0(\nu_0, \vec{n}_0) = \sigma_s(\nu_0) \left[ \frac{1}{4\pi} \int I(\nu_0, \vec{n}'_0) d\vec{n}'_0 - I_0(\nu_0, \vec{n}_0) \right]. \quad (2.17)$$

Dans une première partie de ce travail nous utilisons une analyse asymptotique rigoureuse et retrouvons le modèle de diffusion à l'équilibre ( $T = T_r$ ) quand l'émission absorption est prédominante

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) = 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + (p + p_r)\mathbf{I}) = 0, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \nabla \cdot ((\rho E + E_r)\vec{v} + (p + p_r)\vec{v}) = \nabla \cdot (\frac{1}{3\sigma_a} \nabla T^4), \end{cases}$$

where

$$E_r = T^4, \quad p_r = \frac{1}{3}T^4.$$

et le modèle de diffusion hors équilibre ( $T \neq T_r$ ) pour l'énergie radiative quand le scattering est prédominant :

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) = 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + (p + p_r)\mathbf{I}) = 0, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \nabla \cdot ((\rho E + E_r)\vec{v} + (p + p_r)\vec{v}) = \nabla \cdot (\frac{1}{3\sigma_s} \nabla T^4), \\ \frac{\partial}{\partial t}E_r + \nabla \cdot (\vec{v}E_r) + p_r \nabla \cdot \vec{v} = \nabla \cdot (\frac{1}{3\sigma_s} \nabla T^4) + \sigma_a(T^4 - T_r^4). \end{cases} \quad (2.18)$$

where

$$E_r = T_r^4, \quad p_r = \frac{1}{3}T_r^4.$$

A notre connaissance l'analyse asymptotique que nous proposons en variables d'Euler pour la diffusion hors équilibre est nouvelle.

Dans une deuxième partie nous dérivons un modèle aux moments "gris" relativiste par minimisation d'entropie dans le référentiel eulerien et redonnant la diffusion hors équilibre 2.18 :

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) = 0, \\ \frac{\partial}{\partial t}(\rho \vec{v} + \frac{\mathcal{P}}{c} \vec{F}_r) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p\mathbf{I} + \mathcal{P}\mathbf{P}_r) = 0, \\ \frac{\partial}{\partial t}(\rho E + \mathcal{P}E_r) + \nabla \cdot (\rho E \vec{v} + p\vec{v} + \mathcal{P}\mathcal{C} \vec{F}_r) = 0, \\ \frac{1}{c} \frac{\partial}{\partial t}E_r + \nabla \cdot \vec{F}_r = S_E, \\ \frac{1}{c} \frac{\partial}{\partial t}\vec{F}_r + \nabla \cdot \mathbf{P}_r = \vec{S}_F, \end{cases} \quad (2.19)$$

avec l'énergie radiative et le flux d'énergie radiative définis par  $E_r = \int \int I \, d\nu d\vec{n}$ ,  $F_r = \int \int \vec{n} I \, d\nu d\vec{n}$  et le tenseur de pression donné par

$$\mathbf{P}_r = E_r D_r = E_r \left( \frac{1-\chi}{2} \mathbf{I} + \frac{3\chi-1}{2} \frac{f}{|f|} \otimes \frac{f}{|f|} \right)$$



où  $\vec{f} = \vec{F}_r/E_r$  et  $\chi = \chi(\|f\|)$  est le facteur d'Eddington bien connu, voir [150, 144, 140],

$$\chi(x) = \frac{3 + 4x^2}{5 + 2\sqrt{4 - 3x^2}}.$$

L'expression des termes sources  $S_E$  et  $\vec{S}_F$  n'est pas donnée dans [a13], mais tous les éléments permettant de les calculer y sont. Je renvoie le lecteur au paragraphe (2.5) pour leur expression en 1-d.

Puis nous étudions la compatibilité des modèles à deux moments avec la diffusion hors équilibre. Le résultat principal est que le modèle à deux moments  $M^1$  (2.19) permet de retrouver les solutions continues de la diffusion hors équilibre mais ne permet pas de retrouver les solutions chocs. En revanche si nous utilisons l'entropie radiative  $S_r = S_r = -\frac{2k}{c^3} \int \int \nu^2 [n \log n - (n+1) \log(n+1)] d\nu d\vec{n}$  à la place de l'énergie radiative  $E_r$ , nous obtenons le modèle

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) = 0, \\ \frac{\partial}{\partial t}(\rho \vec{v} + \frac{\mathcal{P}}{c} \vec{F}_r) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I} + \mathcal{P} \mathbf{P}_r) = 0, \\ \frac{\partial}{\partial t}(\rho E + \mathcal{P} E_r) + \nabla \cdot (\rho E \vec{v} + p \vec{v} + \mathcal{P} \mathcal{C} \vec{F}_r) = 0, \\ \frac{1}{c} \frac{\partial}{\partial t} S_r + \nabla \cdot \vec{Q}_r = \frac{1}{\Theta_r} (S_E + b \cdot \vec{S}_F), \\ \frac{1}{c} \frac{\partial}{\partial t} \vec{F}_r + \nabla \cdot \mathbf{P}_r = \vec{S}_F. \end{cases}$$

Ce modèle présente l'avantage de retrouver les solutions chocs et les solutions continues de la diffusion hors équilibre.

## 2.4 Analyse asymptotique et méthodes numériques pour les méthodes de moments en hydrodynamique radiative [cr2, s1]

*En coll. avec S. Cordier.*

Ce travail présente une discrétisation d'un système hyperboliques de lois de conservation qui soit compatible avec le régime asymptotique de diffusion. On s'intéresse à un système de la forme (2.20)

$$\varepsilon \partial_t U + \partial_x F(U) = \frac{1}{\varepsilon} R(U), \quad (2.20)$$

with  $U = (\rho, j)$ ,  $F(U) = (j, \rho h(j/\rho))$ ,  $R(U) = (0, -\sigma j)$  où  $\rho$  et  $j$  sont les deux premiers moments de la solution d'une équation de transport,  $\sigma(x) > 0$  est la section efficace,  $\varepsilon$  est un petit paramètre et  $h$  est une fonction paire positive et convexe.

Dans les problèmes de transfert radiatif,  $h$  est le facteur d'Eddington

$h$  satisfait de plus les propriétés suivantes

$$h(0) = \frac{1}{3}, \quad u^2 \leq h(u) \leq 1.$$

Sous ces hypothèses on peut montrer que l'on a le domaine invariant suivant

$$\rho \geq 0, \quad \|j\| \leq \rho. \quad (2.21)$$

C'est à dire que le flux d'énergie est limité.

Lorsque  $\varepsilon \rightarrow 0$ , le système (2.20) se comporte comme une équation de diffusion dont le coefficient de diffusion est  $h(0)\sigma$ ,  $\sigma$  étant la section efficace.

$$\partial_t \rho - h(0) \partial_x \left( \frac{1}{\sigma} \partial_x \rho \right) = 0. \quad (2.22)$$

$$j = -\frac{\varepsilon}{\sigma} \partial_x (h(0) \rho). \quad (2.23)$$

On peut remarquer que pour  $\rho$  solution de (2.22) et  $j$  défini par (2.23) peuvent ne pas satisfaire (2.21), dans le cas de forts gradients de  $\rho$ .

La méthode que nous présentons ici est constituée de deux étapes: la première consiste à transformer le système de deux équations non linéaires en un double système (2.24) de deux équations, linéaires et découplées, connues sous le nom d'équations du télégraphe. Il s'agit d'une application des méthodes de relaxation [177] et cela conduit à doubler le nombre d'inconnues.

Le schéma relaxé (transport-projection) est le suivant:

on résout la partie transport:

$$\begin{cases} \partial_t \rho + \frac{1}{\varepsilon} \partial_x z = 0 \\ \partial_t z + \frac{a}{\varepsilon} \partial_x \rho + \frac{\sigma}{\varepsilon^2} z = 0 \\ \partial_t w + \frac{a}{\varepsilon} \partial_x j = 0 \\ \partial_t j + \frac{1}{\varepsilon} \partial_x w + \frac{\sigma}{\varepsilon^2} j = 0. \end{cases} \quad (2.24)$$

puis on projette

$$z = j, \quad w = \rho h\left(\frac{j}{\rho}\right).$$

On peut remarquer que la phase transport (2.24) consiste à résoudre deux systèmes linéaires indépendants le premier pour  $(\rho \text{ and } z)$ , le second pour

Pour chacun des systèmes obtenus, on utilise ensuite le schéma "well-balanced" proposée dans [164, 165, 166, 167] et que l'on généralise pour une section efficace non constante et un maillage non uniforme. On peut rappeler brièvement que ce schéma "well-balanced" est un schéma de Godunov pour un système hyperbolique linéaire avec des termes sources localisés aux interfaces. Dans le cas linéaire la résolution du problème de Riemann est triviale [s1]. Le schéma s'écrit

$$\begin{cases} \frac{du_i}{dt} + M_{i-\frac{1}{2}} \frac{\sqrt{a}}{\varepsilon \Delta x_i} (u_i - u_{i-1}) = M_{i-\frac{1}{2}} \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} \frac{\sigma_{i-\frac{1}{2}}}{2\varepsilon^2} (v_i - u_i) \\ \frac{dv_i}{dt} - M_{i+\frac{1}{2}} \frac{\sqrt{a}}{\varepsilon \Delta x_i} (v_{i+1} - v_i) = M_{i+\frac{1}{2}} \frac{\Delta x_{i+\frac{1}{2}}}{\Delta x_i} \frac{\sigma_{i+\frac{1}{2}}}{2\varepsilon^2} (u_i - v_i) \end{cases} \quad (2.25)$$

et les coefficients  $M_{i+\frac{1}{2}}$  sont définis par

$$M_{i+\frac{1}{2}} = \frac{2\sqrt{a}\varepsilon}{\sigma_{i+\frac{1}{2}} \Delta x_{i+\frac{1}{2}} + 2\sqrt{a}\varepsilon},$$

et  $\sigma_{i+\frac{1}{2}}$  est une moyenne de  $\sigma$  aux interfaces.

Dans les variables d'origine  $(\rho, z)$  et aussi pour  $(w, j)$

$$\begin{cases} \frac{d\rho_i}{dt} + \frac{1}{\varepsilon \Delta x_i} (M_{i+\frac{1}{2}} z_{i+\frac{1}{2}} - M_{i-\frac{1}{2}} z_{i-\frac{1}{2}}) = 0, \\ \frac{dz_i}{dt} + \frac{a}{\varepsilon \Delta x_i} (M_{i+\frac{1}{2}} \rho_{i+\frac{1}{2}} - M_{i-\frac{1}{2}} \rho_{i-\frac{1}{2}}) = \frac{-\lambda_i}{2\varepsilon^2} z_i + \frac{M_{i+\frac{1}{2}} - M_{i-\frac{1}{2}}}{\varepsilon \Delta x_i} (a\rho_i) \end{cases}$$

avec

$$z_{i+\frac{1}{2}} = (z_i + z_{i+1} + \rho_{i+1} - \rho_i)/2, \quad \rho_{i+\frac{1}{2}} = (\rho_i + \rho_{i+1} + z_{i+1} - z_i)/2,$$

et

$$\lambda_i = \frac{\Delta x_{i+\frac{1}{2}}}{\Delta x_i} M_{i+\frac{1}{2}} \sigma_{i+\frac{1}{2}} + \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} M_{i-\frac{1}{2}} \sigma_{i-\frac{1}{2}}.$$

On considère une discrétisation en temps totalement implicite Le schéma ainsi obtenu a toutes les propriétés requises: consistance lorsque  $\Delta x \rightarrow 0$ , comportement asymptotique  $\varepsilon \rightarrow 0$ , préservation du domaine invariant (ce qui revient dans les nouvelles variables à garantir la positivité des solutions).

Le schéma peut paraître bien compliqué, mais c'est le résultat de trois contraintes: on veut un schéma totalement implicite, qui a pour domaine invariant (2.21) et qui a la bonne limite de diffusion. Le schéma "well-balanced" proposé par Gosse et Toscani, [164, 165, 166, 167], est certes monotone et donne la bonne limite de diffusion mais il ne s'applique pas à un système hyperbolique non-linéaire, notamment avec un point résonnant, ce qui est le cas du système que l'on considère.

La phase de relaxation (2.24) rend le système linéaire ce qui permet, d'une part, une implication en temps facile et d'autre part permet l'utilisation de ce schéma "well-balanced".

L'avantage de la méthode présentée ici est qu'elle peut être utilisée avec une section efficace  $\sigma$  variable, ce qui est très important en pratique car de tels problèmes de transfert radiatif sont couplés avec un modèle hydrodynamique qui va déterminer la valeur de  $\sigma$ . Celle-ci sera d'ordre 1 dans les zones denses ou opaques et pourra être très faible dans les zones dites transparentes. Pour pouvoir utiliser un schéma sans restriction sur les valeurs de  $\sigma$  il est donc indispensable que la discrétisation ait un bon comportement y compris dans les zones transparentes.

Nous nous sommes également attachés à présenter une méthode utilisable avec un maillage non uniforme car les codes de calcul utilisent des techniques de raffinement de maillage automatique et il est donc important de pouvoir traiter de tels maillages.

## 2.5 Un modèle à flux limité pour l'hydrodynamique radiative et un schéma de splitting associé préservant l'asymptotique de diffusion [s2]

*En coll. avec B.Despres.*

Dans ce travail nous étudions la discrétisation d'un modèle à deux moments pour l'hydrodynamique radiative par un schéma avec flux limité préservant l'asymptotique de diffusion hors équilibre. Le modèle utilisé est le modèle  $M^1$  (2.19) décrit dans le paragraphe (2.3).

Pour simplifier la présentation écrivons le modèle  $M^1$  en question en dimension 1 d'espace

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho v + \frac{p}{c} F_r) + \frac{\partial}{\partial x}(\rho v^2 + p + \mathcal{P} P_r) = 0, \\ \frac{\partial}{\partial t}(\rho E + \mathcal{P} E_r) + \frac{\partial}{\partial x}(\rho E v + p v + \mathcal{P} C F_r) = 0, \\ \frac{1}{c} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} F_r = S_E, \\ \frac{1}{c} \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} P_r = S_F, \end{cases} \quad (2.26)$$

Le modèle est écrit en variables adimensionnelles.  $\rho \geq 0$  représente la densité du fluide,  $\rho v$  est la quantité de mouvement du fluide,  $\rho E$  est l'énergie du fluide et  $p$  est la pression du fluide.  $E_r$  est la densité d'énergie radiative,  $F_r$  est le flux d'énergie radiative et  $P_r$  est la pression radiative. Ces trois quantités sont les moments de  $I$  qui vérifie l'équation de transfert radiatif (2.11). Les termes sources  $S_E$  et  $S_F$  prennent en compte l'émission-absorption et le scattering, et sont donnés par (2.12-2.17), paragraphe (2.3).

L'application de la méthode des moments proposée par Levermore [145, 146, a13] conduit une intensité de la lumière  $I$  donnée par la Planckienne

$$\frac{4\pi^5}{15} \frac{I}{\nu^3} = \frac{1}{e^{\frac{\nu}{T_r} + \frac{\nu b n}{T_r}} - 1}. \quad (2.27)$$

C'est une fonction de la fréquence  $\nu$  et de la direction  $n$  des photons et dépendant de deux paramètres  $T_r$  et  $b \in ]-1, 1[$ .  $T_r$  représente la "température" du flux radiatif,  $b$  représente l'anisotropie de ce flux radiatif.

En utilisant (2.27) on obtient en 1-D

$$\begin{cases} E_r = \int \int I(\nu, n) d\nu dn = T_r^4 \frac{3+b^2}{3(1-b^2)^3}, \\ F_r = \int \int n I(\nu, n) d\nu dn = -T_r^4 \frac{4b}{3(1-b^2)^3} \end{cases}$$

En 3-D la pression radiative  $P_r = \int \int n \otimes n I(\nu, n) d\nu dn$  est donnée par le tenseur, voir Levermore [144],

$$\mathbf{P}_r = E_r \left( \frac{1-\chi}{2} \mathbf{I} + \frac{3\chi-1}{2} \frac{f}{|f|} \otimes \frac{f}{|f|} \right)$$

avec  $f = \frac{F_r}{E_r}$  et le facteur d'Eddington  $\chi$  est donné par

$$\chi(f) = \frac{3 + 4f^2}{5 + 2\sqrt{4 - 3f^2}}$$

En dimension 1 cela se réduit à  $\frac{f}{|f|} = \pm 1$  et  $P_r = \chi(f)E_r$ .

En utilisant (2.27) et les expressions (2.12-2.17) pour l'émission-absorption et le scattering, les termes sources sont alors donnés par  $S_E = S_E^a + S_E^s$  et  $S_F = S_F^a + S_F^s$  avec pour l'émission-absorption

$$\begin{cases} S_E^a = \frac{\gamma \sigma_a}{\varepsilon^2} (T^4 - E_r + \varepsilon v F_r), \\ S_F^a = -\frac{\gamma \sigma_a}{\varepsilon^2} (F_r - \varepsilon v (T^4 + P_r)), \end{cases}$$

et pour le scattering, en 1-D

$$\begin{cases} S_E^s = -\frac{\gamma \sigma_s}{\varepsilon} v F_r^0, \\ S_F^s = -\frac{\gamma \sigma_s}{\varepsilon^2} F_r^0. \end{cases}$$

avec  $\gamma = 1/\sqrt{1 - |v|^2/\mathcal{C}^2}$ . L'expression du flux d'énergie radiative dans le repère comobile  $F_r^0$  est quant à elle donnée par

$$F_r^0 = \gamma^2 ((1 + \varepsilon^2 v^2) F_r - \varepsilon v (E_r + P_r)).$$

Deux paramètres sans dimension apparaissent dans le système.  $\mathcal{C} \geq \infty$  est le ratio de la vitesse de la lumière sur la vitesse du son du fluide. La plupart du temps  $\varepsilon = \frac{1}{\mathcal{C}}$  est un petit paramètre. L'autre paramètre  $\mathcal{P}$  est le rapport de la pression radiative sur la pression du fluide. Suivant en cela le travail [147] nous ne nous intéressons qu'au régime  $\mathcal{P} = 1$ .

Nous proposons une approximation en  $O(\varepsilon^2)$  du modèle (2.26) ce qui est satisfaisant pour le domaine d'applications qui nous intéresse où  $\varepsilon$  est toujours un petit paramètre, c'est-à-dire que la vitesse du fluide est toujours petite devant la vitesse de la lumière :

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho v^2 + p + \frac{E_r}{3}) = 0, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \frac{\partial}{\partial x}(\rho E v + p v + \frac{E_r}{3} u + \frac{1}{\varepsilon} F_r) = 0, \\ \frac{1}{\varepsilon} \frac{\partial}{\partial t} E_r + \partial_x(v E_r) + \frac{E_r}{3} \partial_x v + \frac{\partial}{\partial x} F_r = \frac{\sigma_a}{\varepsilon^2} (T^4 - E_r), \\ \frac{1}{\varepsilon} \frac{\partial}{\partial t} F_r + \partial_x(v F_r) + \frac{F_r}{3} \partial_x v + \frac{\partial}{\partial x} P_r = -\frac{\sigma_a + \sigma_s}{\varepsilon^2} F_r, \end{cases} \quad (2.28)$$

Ce modèle n'est pas celui qui est couramment utilisé en transfert radiatif, voir par exemple [148, 137]. Il en est différent par  $\frac{E_r}{3} \partial_x v$  au lieu de  $F_r \partial_x v$  et  $\frac{E_r}{3} \partial_x v$  au lieu de  $P_r \partial_x v$ .

Nous proposons alors le splitting suivant du modèle (2.28) :

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho v^2 + p + \frac{E_r}{3}) = 0, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \frac{\partial}{\partial x}(\rho E v + p v + \frac{E_r}{3} v) = 0, \\ \frac{\partial}{\partial t} S_r + \frac{\partial}{\partial x}(u S_r) = 0, & S_r = E_r^{\frac{3}{4}}, \\ \frac{\partial}{\partial t} Q_r + \frac{\partial}{\partial x}(u Q_r) = 0, & Q_r = -b S_r, \end{cases} \quad (2.29)$$

suivi de

$$\begin{cases} \partial_t E_r + \frac{1}{\varepsilon} \partial_x F_r = \sigma_a \frac{T^4 - E_r}{\varepsilon^2}, \\ \partial_t F_r + \frac{1}{\varepsilon} \partial_x P_r = -\sigma_t \frac{F_r}{\varepsilon^2}, & \sigma_t = \sigma_a + \sigma_s \\ \rho C_v \partial_t T = \sigma_a \frac{E_r - T^4}{\varepsilon^2}, \end{cases} \quad (2.30)$$

voir [a13].

Ce splitting permet de retrouver les "bonnes" relations de saut de Rankine-Hugoniot, voir [a13], et consiste donc à transporter les photons par la matière via l'entropie et le flux d'entropie, système (2.29), et la phase radiative se fait dans le référentiel de la matière, système (2.30), donc c'est le système  $M^1$  à vitesse matière nulle. Il permet aussi une implémentation plus simple que la résolution directe de (2.26).

La phase radiative est résolu avec une amélioration du schéma proposé dans mon travail précédent [cr2, s1]. Rappelons que ce schéma repose sur un méthode de de relaxation, [177], et sur le schéma "Well Balanced", [165, 166]. L'amélioration apportée ici concerne la phase de relaxation et permet, contrairement à ce qui été proposé dans [cr2, s1], une dépendance en espace et en temps du coefficient  $a$  dans (2.24). En effet la phase 2.24) est remplacée par

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} aX = 0, \\ \frac{\partial}{\partial t} X + \frac{1}{\varepsilon} \frac{\partial}{\partial x} aE_r = -\frac{\sigma}{\varepsilon^2} X, \end{cases}$$

et

$$\begin{cases} \frac{\partial}{\partial t} Y + \frac{1}{\varepsilon} \frac{\partial}{\partial x} aF_r = 0, \\ \frac{\partial}{\partial t} F_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} aY = -\frac{\sigma}{\varepsilon^2} F_r. \end{cases}$$

avec  $a = \sqrt{\chi}$ . Chacun des deux systèmes est alors résolu par le schéma "Well Balanced" (2.25). La discrétisation en temps pour la partie radiative est totalement implicite en prenant la valeur de  $a$  au début du pas de temps.

Les principales propriétés du schéma que nous montrons dans ce travail sont les suivantes

a) Au niveau discret le flux radiatif reste limité

$$\frac{|F_r|}{E_r} \leq 1$$

b) La limite de diffusion est correcte même sur des maillages grossiers.

c) Le traitement numérique du terme non conservatif  $P_r \partial_x u$  de la diffusion hors-équilibre par le transport de l'entropie ce qui permet d'obtenir les bonnes relations de Rankine-Hugoniot.

Le schéma numérique est donc robuste.

De nombreux résultats numériques illustrent la précision et la robustesse de la méthode numérique.

La suite de ce travail concernera évidemment l'extension multi-dimensionnelle du schéma.

## 2.6 Conclusions

En conclusion de ce chapitre, on peut dire que dans le cadre de la simulation du transfert radiatif via l'utilisation d'un modèle aux moments de type (2.19), l'approche "asymptotic preserving schemes" basée sur le schéma "well-balanced", [165, 166], donne de bons résultats en dimension 1 d'espace. Le prochain challenge sera d'étendre, pour le modèle de type 2.19), cette méthode en dimension 2 ou 3 d'espace, sur un maillage cartésien conforme ou non (méthode de raffinement de maillage adaptatif). La principale difficulté, en dimension 2 ou 3, est d'assurer la limitation de flux par un schéma totalement implicite. Ce sera la majeure partie de mon travail de recherche au CEA dans un futur proche.

Pour les modèles de Lorentz (2.5) en théorie cinétique, si l'on désire un schéma "asymptotic preserving" positif, une voie envisageable semble être le schéma "well balanced" [166, 167] mais au vue des résultats présentés dans [167] il y a encore fort à faire pour, ne serait-ce qu'en dimension 1 d'espace, étendre ce type de schéma à des sections efficaces non isotropes. Quant à l'extension multi dimensionnelle de ce type de schémas, elle ne semble toujours pas être à l'ordre du jour. C'est pourtant une voie à défricher.



## Chapitre 3

### Autres travaux

Dans cette partie je présente deux travaux n'entrant pas dans mes thématiques principales de recherche. Le premier a quand même un rapport, certes lointain avec la théorie cinétique et les termes de collisions. Le second aurait plutôt des conséquences sur la discrétisation de certains opérateurs de collision de type parabolique, voire sur les schémas monotones préservant la limite de diffusion.

#### 3.1 Ionisation multi-espèces [a10]

*En coll. avec S. Cordier et P.A. Raviart.*

Ce travail porte sur un modèle fluide multi-espèce d'ionisation qui est détaillé dans [a10]. Ce modèle conduit à un système d'équations différentielles ordinaires, mais singulières à l'origine. L'existence de solutions maximales pour ce système a pu être démontrée et des simulations numériques permettent de calculer les courants émis par chaque espèce d'un faisceau d'ions, ce qui intéresse les physiciens. De nouvelles perspectives sont envisagées pour mieux comprendre les propriétés mathématiquement surprenantes mais physiquement raisonnables de ce type de modèle.

L'étude d'un modèle cinétique d'ionisation montre que les températures ioniques restent faibles dans la zone d'ionisation. Cela conduit à considérer un modèle fluide où les ions sont froids, plus facile à étudier numériquement que le modèle cinétique. On cherche des solutions stationnaires du système Euler-Poisson qui s'écrit cette fois

$$\frac{d}{dx}(n_\alpha u_\alpha) = g_\alpha(n_e), \quad \alpha = 1, \dots, N-1, \quad (3.1)$$

$$\frac{d}{dx}(n_\alpha u_\alpha^2) - n_\alpha \frac{d\phi}{dx} = 0 \quad (3.2)$$

$$\lambda \frac{d^2\phi}{dx^2} = \sum_\alpha n_\alpha - n_e, \quad n_e = \exp(-\phi), \quad (3.3)$$

avec les conditions aux limites et la condition de quasineutralité en  $x = 0$

$$u_\alpha(0) = 0, \phi(0) = \frac{d\phi}{dx}(0) = 0, \sum_\alpha n_\alpha(0) = n_e(0) = 1. \quad (3.4)$$

Nous allons étudier plus particulièrement l'approximation plasma obtenue en supposant la quasineutralité du plasma ( $\lambda = 0$ ). L'équation d'impulsion ionique s'écrit alors

$$\frac{d}{dx}(n_\alpha u_\alpha^2) + \frac{n_\alpha}{n_e} \frac{dn_e}{dx} = 0, \quad (3.5)$$

avec  $n_e = \exp(-\phi) = \sum_\alpha n_\alpha$ . Il s'agit d'abord d'étudier l'existence et l'unicité de la solution du problème ainsi que ses propriétés qualitatives. Dans le cas  $N = 2$ , on sait calculer explicitement

la solution maximale (et les courants émis qui sont indépendants du taux d'ionisation). On veut généraliser ce résultat. Posons  $p = N - 1$  le nombre d'espèces d'ions. On appelle solution (physiquement admissible) du système différentiel (3.1-3.5), une fonction  $U = (n_1, \dots, n_p, j_1, \dots, j_p)$ , où  $U : x \in \mathbb{R}_+ \rightarrow U(x) \in \mathbb{R}^{2p}$  de classe  $C^1$  solution de (3.1-3.5) et vérifiant  $n_\alpha(x) > 0$  pour tout  $\alpha = 1, \dots, p$ .

Pour fixer les idées, on suppose que chaque taux d'ionisation (adimensionné)  $g_\alpha : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  est une fonction de classe  $C^1$  croissante qui vérifie  $g_\alpha(n_e) > 0 \ \forall n_e > 0$ . En pratique, les taux d'ionisation  $g$  dépendent des processus d'ionisation et sont des polynômes de  $n_e$ .

On montre alors que le modèle fluide dans l'approximation plasma admet une solution (physiquement admissible) unique définie dans un intervalle maximal  $[0, x_0[$  où  $x_0 < +\infty$  est tel que  $U(x_0) = \lim_{x \rightarrow x_0} U(x)$  existe et vérifie (en  $x_0$ )

$$\sum_{\alpha} \frac{n_{\alpha}}{u_{\alpha}^2} = n_e, \quad \sum_{\alpha} n_{\alpha} u_{\alpha}^2 = 1 - n_e;$$

de plus, les dérivées de  $U$  explosent en  $x_0$

$$\lim_{x \rightarrow x_0} \frac{dn_{\alpha}}{dx}(x) = -\infty, \quad \lim_{x \rightarrow x_0} \frac{du_{\alpha}}{dx}(x) = +\infty;$$

et la fonction  $n_e$  est strictement décroissante dans  $[0, x_0]$  et  $n_e(x_0) \leq 1/2$ .

La première étape de la démonstration consiste à prouver que  $n_e$  est solution de l'équation intégrale non linéaire

$$n_e(x) = 1 - \sum_{\alpha} \sqrt{-2 \int_0^x \left( \frac{j_{\alpha}^2}{n_e} \frac{dn_e}{dx} \right)(y) dy}, \quad (3.6)$$

où  $j_{\alpha}$  est donné en fonction de  $n_e$  par

$$j_{\alpha}(x) = \int_0^x g_{\alpha}(n_e(y)) dy.$$

Cela va nous servir d'une part à obtenir un résultat d'existence et d'unicité locale de la solution de au voisinage de  $x = 0$  et d'autre part à calculer numériquement la solution.

L'existence d'une solution maximale pour l'équation plasma (3.6) n'est pas immédiate. Celle-ci repose sur la construction de solutions approchées qui permettent de supprimer la singularité du problème à l'origine. Une fois que les solutions ont "décollé", il faut vérifier qu'elles ont le bon comportement et qu'elles convergent vers une solution du problème. Cette technique est également utile du point de vue numérique. Nous montrons que les solutions numériques sont telles que les vitesses des différentes espèces d'ions sont très proches (mais distinctes) et le point  $x_0$  est caractérisé par

$$n_e(x_0) \approx 1/2, \quad u_{\alpha}(x_0) \approx 1,$$

ce qui correspond au cas mono-espèce et qui est physiquement raisonnable. Nous comparons les résultats de ce modèle avec d'autres (cinétique, mono-cinétique) et nous montrons comment traiter les termes de friction, le système quasi-neutre ( $\lambda \ll 1$ ) et enfin, nous montrons, en reprenant des résultats de [190], que l'hyperbolicité du système d'évolution associé à (3.1-3.5) impose que les vitesses des différentes espèces soient égales.

### 3.2 Schémas monotones et équations parabolique linéaires [cr3]

*En coll. avec S. Cordier.*

Ce travail présente un résultat de non existence de schémas linéaires monotones pour certaines équations paraboliques linéaires avec diffusion anisotrope.



On s'intéresse à l'approximation numérique de l'équation parabolique de la forme (3.7).

$$\partial_t f - \frac{1}{2} \nabla \cdot K \nabla f = 0, \quad (3.7)$$

Dans le cas où  $c = ab$ , cette équation représente l'équation de la chaleur mono dimensionnelle immergée en dimension 2. La diffusion agit donc uniquement dans la direction  $(a, b)$ . Il est bien connu que cette équation vérifie un principe du maximum : soit deux données initiales telles que  $f_0 \geq g_0$  alors  $f \geq g$ .

On considère une grille de calcul cartésienne à mailles carrées de pas d'espace  $h$ .

Nous montrons alors que, pour un stencil fixe de la forme

$$S = \{j \text{ such that } -N \leq j_k \leq N, k = 1, 2\},$$

il existe toujours des directions  $(a, b)$  de diffusion pour lesquelles on ne peut construire de schéma linéaire monotone. La preuve est basée sur l'analyse de l'erreur de consistance.

Ce résultat est encore valide pour une matrice de diffusion dépendant de la variable d'espace. donc en particulier pour l'équation de Fokker-Planck-Lorentz

$$K(x, y) = \Psi\left(\frac{1}{x^2 + y^2}\right) (\mathbf{Id} - (x, y)^t \otimes (x, y))$$

issue de la physique des plasmas et représentant un modèle simplifié des collisions avec les ions pour les électrons. Cela montre aussi que pour certains maillages quadrangulaires, en se ramenant à un maillage carré par homéomorphisme, on ne pourra pas toujours avoir de schémas monotones pour (3.7).

Nous montrons de plus que, pour l'équation du télégraphe

$$\begin{aligned} \varepsilon \partial_t u + a \partial_x u + b \partial_y u &= \frac{1}{\varepsilon} (v - u) \\ \varepsilon \partial_t v - a \partial_x v - b \partial_y v &= \frac{1}{\varepsilon} (u - v) \end{aligned}$$

, il ne peut y avoir de schémas linéaires monotones consistant avec la limite de diffusion (3.7) avec  $c = ab$ , pour une direction arbitraire de propagation.

Nous indiquons enfin comment ce résultat pourrait signifier que pour l'équation de Fokker-Planck-Landau

$$\frac{d}{dt} f = \nabla_v \cdot \int_{v'} \phi(v - v') (f' \nabla_v f - f \nabla_{v'} f') dv',$$

il ne peut y avoir de schémas positifs pour une discrétisation de la forme naturelle c'est à dire de schémas quadratiques positifs, ce qui justifierait l'emploi d'algorithmes vraiment non linéaires par exemple basés sur la forme dite logarithmique

$$\frac{d}{dt} f = \nabla_v \cdot \int_{v'} f f' \phi(v - v') (\nabla_v \log(f) - f \nabla_{v'} \log(f')) dv'$$

voir par exemple [57].

Une solution pour contourner ce problème est soit d'utiliser un schéma non linéaire, soit de faire en sorte que la taille du stencil puisse croître quand le pas de maillage tends vers 0.

**Remarque 6** On peut citer encore deux exemples autour de l'équation de Fokker-Planck-Landau, pour lesquels ce papier montre que l'on ne peut construire de schémas positifs. Le premier concerne les modèles du type "Spherical Harmonic Expansion" pour les électrons, voir [72, 82], pour la partie Vlasov de l'équation et ce à cause du champ électrique.

Le deuxième provient d'une simplification du terme de collision de Landau pour traiter le ralentissement des particules en fusion par confinement inertiel, [80, 81].



# Listes des Travaux Présentés

## ARTICLES POSTÉRIEURS À LA THÈSE

- [a1] C. Buet, "A discrete-velocity scheme for the Boltzmann operator of rarefied gas dynamics", *Transp. Theory Stat. Phys.* 25, No.1, 33-60 (1996).
- [a2] C. Buet, "Conservative and entropy schemes for Boltzmann collision operator of polyatomic gases", *Math. Models Methods Appl. Sci.* 7, No.2, 165-192 (1997).
- [a3] C. Buet, S. Cordier, P. Degond, M. Lemou, "Fast algorithms for numerical, conservative, and entropy approximations of the Fokker-Planck-Landau equation", *J. Comput. Phys.* 133, No.2, 310-322 (1997).
- [a4] C. Buet, S. Cordier, P. Degond, "Regularized Boltzmann operators", *Comput. Math. Appl.* 35, No.1-2, 55-74 (1998).
- [a5] C. Buet, S. Cordier, "Conservative and entropy decaying numerical scheme for the isotropic Fokker-Planck-Landau equation", *J. Comput. Phys.* 145, No.1, 228-245 (1998).
- [a6] C. Buet, S. Cordier, "Numerical analysis of conservative and entropy schemes for the Fokker-Planck-Landau equation", *SIAM J. Numer. Anal.* 36, No.3, 953-973 (1999).
- [a7] C. Buet, S. Dellacherie, R. Sentis, "Numerical solution of an ionic Fokker-Planck equation with electronic temperature", *SIAM J. Numer. Anal.* 39, No.4, 1219-1253 (2001).
- [a8] C. Buet, S. Cordier, B. Lucquin, "The grazing collision limit for the Boltzmann-Lorentz model", *Asymptotic Anal.* 25, No.2, 93-107 (2001).
- [a9] C. Buet, S. Cordier, "Numerical analysis of the isotropic Fokker-Planck-Landau equation", *J. Comput. Phys.* 179, No.1, 43-67 (2002).
- [a10] C. Buet, S. Cordier, P.-A. Raviart, "Multifluid ionization models", *Math. Models Methods Appl. Sci.* 12, No.1, 17-48 (2002).
- [a11] C. Buet, S. Cordier, B. Lucquin-Desreux, S. Mancini, "Diffusion limit of the Lorentz model: asymptotic preserving schemes", *M2AN, Math. Model. Numer. Anal.* 36, No.4, 631-655 (2002).
- [a12] C. Buet, S. Cordier, "Numerical Method for the Compton Scattering Operator", *Series on Advances in Mathematics for Applied Sciences*, Vol. 63, World Scientific Publishing Co. (2003).
- [a13] C. Buet, B. Despres, "Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics", *Journal of Quantitative Spectroscopy and Radiative Transfer*, Volume 85, Issues 3-4, 385-418 (2004).
- [a14] C. Buet, S. Cordier, V. Dos Santos, "A Conservative and Entropy Scheme for a Simplified Model of Granular Media" *Transp. Theory Stat. Phys.* 33, No.2, 125-155 (2004).

## NOTES AUX COMPTES RENDUS DE L'ACADÉMIE DES SCIENCES

- [cr1] C. Buet, S. Dellacherie, R. Sentis, "Numerical solution of a ionic Fokker-Planck equation with electronic temperature. (Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique)" C. R. Acad. Sci., Paris, Sér. I, Math. 327, No.1, 93-98 (1998).
- [cr2] C. Buet, S. Cordier, "Asymptotic preserving scheme and numerical methods for radiative hydrodynamic models", Comptes Rendus Mathématique, Volume 338, 951-956 (2004).
- [cr3] C. Buet, S. Cordier, "On the non existence of monotone linear scheme for some linear parabolic equations (Sur la non existence de schémas linéaires monotones pour certaines équations paraboliques linéaires)", Comptes Rendus Mathématique, Volume 340, 399-404 (2005).

#### CONFÉRENCES AVEC COMITÉ DE LECTURE

- [c1] C. Buet, "A discrete-velocity scheme for the Boltzmann operator of rarefied gas dynamics", Proceedings of the 19th International Symposium of Rarefied Gas Dynamics, Oxford (1994).

#### ARTICLES SOUMIS.

- [s1] C. Buet, S. Cordier, "An asymptotic preserving scheme for Hydrodynamics Radiative Transfer Models", soumis à Num. Math.
- [s2] C. Buet, B. Despres, "Asymptotic Preserving and Positive Schemes for Radiation Hydrodynamics", soumis au JCP.

#### NON PUBLIÉ.

- [np1] C. Buet, S. Dellacherie, "About the Chang and Cooper method for linear Fokker-Planck equations".

# Bibliographie

## EQUATION DE BOLTZMANN

- [1] L. ARKERYD, *On the Boltzmann equation, I and II*, Arch. Rat. Mech. Anal., **45**, (1972), pp. 1-34.
- [2] L. ARKERYD, *Intermolecular forces of infinite range and the Boltzmann equation*, Arch. Rat. Mech. Anal., **77**, (1981), pp. 11-21.
- [3] C. BARDOS, F. GOLSE, and D. LEVERMORE. Fluid dynamical limits of kinetic equations, I: Formal derivation. *J. Stat. Phys.*, **63**: 323–344, 1991.
- [4] C. BARDOS, F. GOLSE and D. LEVERMORE, *Fluid dynamics limits of kinetic equations II; convergence proofs for the Boltzmann equation*, Comm. on Pure and Appl. Math. **46** (1993), pp. 667-753.
- [5] G. A. BIRD, *Molecular Gas Dynamics and the direct simulation of gas flows*, Clarendon Press, Oxford, 1994.
- [6] A. BOBYLEV, A. PALCZEWSKI, J. SCHNEIDER, *A consistency result for a discrete-velocity model of the Boltzmann equation*, Institute of Applied Mathematics, Warsaw University, 1995.
- [7] C. BORGNACKE, PS. LARSEN *Statistical model for Monte-Carlo simulation of polyatomic gas mixtures*, J. Comp. Phys., **18** (1975) 405-420.
- [8] J. F. BOURGAT, L.DESVILLETES, P. LE TALLEC, B.PERTHAME, *Microreversible collisions for polyatomic gases and Boltzmann's theorem*, Eur. J. Mech. B fluids, (1994).
- [9] A.V. BOBYLEV, TOSCANI, G., *On the generalization of the Boltzmann H-theorem for a spatially homogeneous Maxwell gas*, Jal of Math Phys, Vol 33, 7, 1992.
- [10] R. E. CAFLISH, *The fluid dynamic limit of the nonlinear Boltzmann equation*, Comm. Pure Appl. Math. **33** (1980), pp. 651-666.
- [11] R.E. CAFLISH, *The Boltzmann equation with a soft potential, I: Linear, spatially homogeneous*, Comm. Math. Phys., **74**, (1980), pp. 71-96.
- [12] R.E.CAFLISH, L. PARESCHI, *An implicit Monte Carlo method for rarefied gas dynamics I: The space homogeneous case*, J. Computational Physics, **154**, pp. 90-116, (1999).
- [13] C. CERCIGNANI, *The Boltzmann Equation and its Applications*, Springer, New York, (1988).
- [14] C. CERCIGNANI, R. ILLNER, M. PULVIRENTI, *The mathematical theory of dilute gases*, Appl. Math. Sc. 106, Springer (1994).
- [15] R. DI PERNA and P. L. LIONS, *On the Cauchy problem for the Boltzmann equations: global existence and weak stability*, Annals of Math. **130** (1990), pp. 321-366.
- [16] P. DEGOND, S. MAS GALLIC, *The weighted particle method for convection diffusion equations*, Math. Comp, Vol. 53, pp 485-507 (part 1) and pp 509-525 (part 2), (1989).
- [17] L. DESVILLETES, *Sur un modele de type Borgnakke-Larsen conduisant a des lois d'energie non-linéaires en température pour les gaz parfaits polyatomiques*, Actes du workshop du GDR SPARCH, (1995).

- [18] T. ELMROTH, *On the H-function and convergence towards equilibrium for a space homogeneous molecular density*, S.I.A.M. J. of Appl. Math., **44**, (1984), pp. 150-159.
- [19] E.GABETTA, L. PARESCHI, M.RONCONI, Central schemes for hydrodynamical limits of discrete-velocity kinetic equations, Transp. Theo. Stat. Phys. (to appear).
- [20] R. GATIGNOL, *Théorie cinétique des gaz à répartitions discrètes de vitesses*, Springer, New York, (1975).
- [21] D. GOLDSTEIN, B. STURTEVANT, J. E. BROADWELL, *Investigations of the Motion of Discrete-Velocity Gases*, in "Rarefied Gas Dynamics: Theoretical and Computational Techniques", E. P. Muntz, D. P. Weaver and D. H. Campbell (eds), Progress in Astronautics and Aeronautics, Vol. 118, AIAA, Washington DC, (1989).
- [22] D. B. GOLDSTEIN, *Discrete-Velocity collision dynamics for polyatomic molecules*, Phys. Fluids A4, pp 1831-1839, (1992).
- [23] F. GROPENGIESSER, H. NEUNZERT, J. STRUCKMEIER, *Computational methods for the Boltzmann equation*. Venice 1989: The state of Art in Appl. and Industrial math., eds. R. Spigler, Kluwer acad. publ., (1990).
- [24] T. GUSTAFFSON, *Global  $L^p$  properties for the spatially homogeneous Boltzmann equation*, Arch. Rat. Mech. Anal., **103**, (1988), pp. 1-37.
- [25] R. ILLNER, S. RJASANOW, Random discrete velocity method: possibles bridges between the Boltzmann equation, discrete velocity models and particle simulation, *Non linear kinetic theories and mathematical aspects of hyperbolic systems*, Proc. Rapallo, April (1992).
- [26] ILLNER, R., WAGNER, W., A random discrete velocity model and approximation of the Boltzmann equation, Jal of Stat Phys., vol 70, No 3/4, 1993.
- [27] ILLNER, R., WAGNER, W., A random discrete velocity model and approximation of the Boltzmann equation: conservation of momentum and energy, to appear in TTSP.
- [28] T. INAMURO, B. STURTEVANT, *Numerical Study of Discrete-Velocity Gases*, Phys. Fluids, Vol. A2 pp 2196-2203, (1990).
- [29] S.JIN, L. PARESCHI, G. TOSCANI, Uniformly accurate diffusive relaxation schemes for multiscale transport equations, Preprint CDNSNS98-314, GeorgiaTech, (1998). Submitted to SIAM J. Numerical Analysis.
- [30] S. MAS GALLIC, F. POUPAUD, *A deterministic particle method for the linearized Boltzmann operator*, Trans. Theory Stat. Phys. Vol. 17, 311-345., 4 (1987).
- [31] K. NANBU, *Direct simulation schemes derived from the Boltzmann equation*, J. Phys, Japan Vol. 49 p. 2042, (1980).
- [32] K. NANBU, *Model kinetic equation for the distribution of discretized internal energy*, Math Methods and Models in the Applied Sci, (1992).
- [33] B. NICLOT, P. DEGOND, F. POUPAUD, *Deterministic particles simulations of the Boltzmann transport equation of semiconductors*, J. Comp. Phys., Vol. 78, pp 313-345, (1988).
- [34] L. PARESCHI, G.RUSSO, Asymptotic preserving Monte Carlo methods for the Boltzmann equation, Transp. Theo. Stat. Phys. (to appear).
- [35] L. PARESCHI, G.RUSSO, Numerical solution of the Boltzmann equation I: Spectrally accurate approximation of the collision operator, SIAM J. Numerical Analysis (to appear).
- [36] L. PARESCHI, G.RUSSO, On the stability of spectral methods for the homogeneous Boltzmann equation, Transp. Theo. Stat. Phys. (to appear).
- [37] F. ROGIER, J. SCHNEIDER, *A direct Method for solving the Boltzmann Equation*, Transp. Theory. Stat. Phys, Vol. 23, pp 313-338 (1994).
- [38] Y. SHIZUTA: On the classical solutions of the Boltzmann equation, *Comm. Pure Appl. Math.*, 36 (1983), 705 – 754.
- [39] J. SCHNEIDER, *Une méthode déterministe pour la résolution de l'équation de Boltzmann*, Ph. D thesis, University Paris 6, (1993).
- [40] Z. TAN, P. L. VARGHESE, *The  $\Delta - \epsilon$  method for the Boltzmann equation*, J. Comput. Phys., Vol 110, (1994).

- [41] C. VILLANI, Contribution à l'étude mathématique des collisions en théorie cinétique, habilitation à diriger des recherches, 1999.
- [42] B. WENNERBERG *Stability and exponential convergence in  $L^p$  for the spatially homogeneous Boltzmann equation*, Nonlinear Analysis, Theory, Methods and Applications, 20 (1993),  $n^0$  8, pp. 935-964.

#### EQUATION DE FOKKER-PLANCK

- [43] A.A. ARSENEV and O.E. BURYAC. On the connection between a solution of the Boltzmann equation and a solution of the Landau-Fokker-Planck equation. Math. USSR Sbornik, Vol 69, No. 2 pp. 465-478, 1991.
- [44] A.A. ARSENEV and N.V. PESKOV. On the existence of a generalized solution of Landau's equation. USSR Comput. Math. Phys. Vol. 17, pp. 241-246, 1977.
- [45] V. V. ARISTOV, F. G. CHEREMISIN, *On the connection between a solution of the Boltzmann equation and a solution of the Landau-Fokker-Planck equation*, Math. USSR Sbornik, Vol. 69,  $N^0$  2, pp. 465-478 (1991).
- [46] R. BALESCU, Phys. Fluids, 3, 52, 1960.
- [47] Yu.A. BEREZIN, V.N. KHUDICK, M.S. PEKKER, Conservative finite difference schemes for the Fokker-Planck equation not violating the law of an increasing entropy, J. of Comp. Phys, Vol 69, pp. 163-174, 1987.
- [48] A.V. BOBILEV. I.F. POTAPENKO and V.A. CHUYANOV, Kinetic Equations of the Landau type as a model of the Boltzmann Equation and Completely Conservative Difference Schemes. U.S.S.R. Comput. Maths. Math. Phys. Vol. 20, Vol 4 pp. 190-201, 1981.
- [49] J. S. Chang et G. Cooper - A practical Difference Scheme for Fokker-Planck Equations - J. Comp. Physics, 6, p.1-16, (1970).
- [50] H. COHN. Numerical integration of the Fokker-Planck equation and the evolution of stars clusters. The Astrophysical Journal, 234, 1036-1053, 1979.
- [51] H. COHN. Late core collapse in star clusters and the gravothermal instability. The Astrophysical Journal, 242, 765-771, 1980.
- [52] S. CORDIER, B. LUCQUIN-DESREUX and A.SABRY, Numerical approximation of the Vlasov-Fokker-Planck-Lorentz model. *ESAIM: Proceed. CEMRACS 1999* (2001).
- [53] J.F., CLOUET, K., DOMELEVO, Solution of a kinetic stochastic equation modelling a spray in a turbulence gas flow, R.I. 330, C.M.A.P., Ecole Polytechnique, 1996.
- [54] D. DECK, G. SAMBA, *Le code Procions*, Note C.E.A.  $N^0$  2780, C.E.A./C.E.L.-V, F- 94195 Villeneuve St. Georges, (1994).
- [55] P. DEGOND and B. LUCQUIN-DESREUX, The Fokker-Planck asymptotics of the Boltzmann collision operator in the Coulomb case. *Math.Mod.Meth.Appl.Sci.* **2** 2 (1992) 167-182.
- [56] P. DEGOND and B. LUCQUIN-DESREUX. The Fokker-Planck asymptotics of the Boltzmann collision operator in the Coulomb case, Math. Models and Methods in Appl.Sci, vol. 2, No 2, p 167-182, 1992.
- [57] P. DEGOND and B. LUCQUIN-DESREUX. An entropy scheme for the Fokker-Planck collision of plasma kinetic theory. Numer. Math. vol. 68, pp 239-262, 1994.
- [58] P. DEGOND and B. LUCQUIN-DESREUX, The asymptotics of collision operators for two species of particles of disparate masses. *Math.Mod.Meth.Appl.Sci.* **6** 3 (1996) 405-436.
- [59] P. DEGOND and B. LUCQUIN-DESREUX *The asymptotics of collision operators for two species of particles of disparate masses*, Mathematical Models and Methods in Applied Sciences, vol.6,  $n^0$  3, (1996), pp 405-436.
- [60] P. DEGOND et B. LUCQUIN-DESREUX - Transport coefficients of plasmas and disparate mass binary gases - Transport Theory and Statistical Physics, 25, p.595 (1996).

- [61] P. DEGOND and B. LUCQUIN-DESREUX *Comportement hydrodynamique d'un mélange gazeux formé de deux espèces de particules de masses très différentes*, C. R. Acad. Sc. Paris, t.322, Série I, p 405-410, (1996).
- [62] P. DEGOND and B. LUCQUIN-DESREUX *Transport coefficients of plasmas and disparate mass binary gases*, Transp. Theory in Stat. Phys., Vol 25, n<sup>o</sup> 6, pp. 595-633 (1996).
- [63] P. DEGOND and S. MAS-GALLIC, *The weighted particle method for convection-diffusion equations, part I: the case of an isotropic viscosity, part II: the anisotropic case*, Math. Comput. **53** (1989), 485-526.
- [64] S. DELLACHERIE, Contribution à l'analyse et la simulation numérique des équations cinétiques décrivant les plasmas chauds, these de doctorat université paris VII, Nov. 1998.
- [65] L. DESVILLETES, On asymptotics of the Boltzmann equation when the collisions become grazing. *Transp.Theor.Stat.Phys.* **21** 3 (1992) 259-276.
- [66] L. DESVILLETES, On the convergence of the splitting algorithms for some kinetic equations, *Asympt. Anal.*, **6**, n<sup>o</sup> 4, (1993), pp. 315-333.
- [67] L. DESVILLETES, Some applications of the method of moments for the homogeneous Boltzmann and Kac equations, *Arch. Rat. Mech. Anal.*, **123**, n<sup>o</sup> 4, (1993), pp. 387-404.
- [68] L. DESVILLETES, Convergence to equilibrium in various situations for the solution of the Boltzmann equation, *Nonlinear Kin. Th. and Math. Aspects of Hyp. Syst., Series on Adv. in Math. for Appl. Sci.*, 9, (1992), 99. 101-114.
- [69] L. DESVILLETES, S. MISCHLER, About the splitting algorithm for Boltzmann and B.G.K. equations. [*J*] *Math. Models Methods Appl. Sci.* 6, No.8, 1079-1101 (1996).
- [70] L. DESVILLETES, C. VILLANI, On the spatially homogeneous Landau Equation for Hard potentials. Part I: Existence, Uniqueness and Smoothness, Preprint du DMI, ENS de Paris, 1998.
- [71] L. DESVILLETES, C. VILLANI, On the spatially homogeneous Landau Equation for Hard potentials. Part II: H-Theorem and Applications, Preprint du DMI, ENS de Paris, 1998.
- [72] EPPERLEIN, E.M., Fokker-Planck modelling of electrons transport in laser-produced plasmas, in *Laser and particles Beams*, Vol 2, No 2, p. 257-272, 1994.
- [73] EPPERLEIN, E.M., Implicit and conservative difference schemes for the Fokker-Planck equation, *J. Comp. Phys*, Vol. 112, p. 291, 1994.
- [74] R. EYMARD, T. GALLOUET and R. HERBIN. Schémas de type volume finis. Ecole ceaedf-inria, problemes non-linéaires appliqués, Paris, Octobre 1992.
- [75] F. FILBET, semi-lagrangian method for the Vlasov equations, 1999.
- [76] E. FRENOD and B. LUCQUIN-DESREUX. On conservative and entropy discrete axisymmetric Fokker-Planck operators. *M2AN*, Vol 33, p. 307, 1998.
- [77] L. GREENGARD and V. ROKHLIN. A fast algorithm for a particle simulation. *J. Comput.Phys* Vol. 73, 1987.
- [78] S. JORNA, L. WOOD, *Fokker-Planck calculations on relaxation of anisotropic velocity distributions in plasmas*, *Phys. Rev. A*, Vol. 36, N<sup>o</sup> 1, (1987).
- [79] O. LARROCHE. Kinetic Simulations of a plasma collision experiment. *Phys. Fluids B*, Vol. 5, No 8, 1993.
- [80] O. LARROCHE. Kinetic simulations of fuel ion transport in ICF target implosions, *Eur. Phys. J. D* 27, 131-146, 2003.
- [81] O. LARROCHE. An efficient explicit numerical scheme for hot fusion particle slowing-down in the Fokker-Planck formalism, 34th Anomalous Absorption Conference, 2004.
- [82] V. LATOCHA. Deux problemes en transport des particules chargees intervenant dans la modelisation d'un propulseur ionique. Thèse de doctorat 2001.
- [83] M. LEMOU. Exact solutions of the Fokker-Planck equation. *C.R. Acad. Sci.* t.319, Serie 1, pp. 579-583, 1994.
- [84] M. LEMOU. Multipole expansions for the Fokker-Planck equation, preprint of MIP, *Numer. Math.* 78, No.4, 597-618, 1998.



- [85] M. LEMOU. Numerical Algorithms for Axisymmetric Fokker Planck Landau Operators, Journal of Computational Physics 157, 762-786, 2000.
- [86] M. LEMOU, L. MIEUSSENS, Fast implicit schemes for the Fokker-Planck-Landau equation, C. R. Acad. Sci. Paris, Ser. I 338, 2004.
- [87] R. LENARD, Ann. Phys., Vol 3, p. 390, 1960.
- [88] O. LIE-SVENDSEN, M.H., REES, An improved kinetic model for the polar outflow of a minor ion, J. Geophys. Res., 1995.
- [89] P.L. LIONS, *On Boltzmann and Landau equations*, Phil. Trans. R. Soc. Lond. A, (1994), 346, pp. 191-204.
- [90] S. LIVI, E. MARSCH, Generation of solar wind proton tails and double beams by coulomb collisions, J. Geophys. Res., Vol. 92, No A7, pp. 7255-7261, 1987.
- [91] B. LUCQUIN-DESREUX, Diffusion of electrons by multicharged ions. *Math.Mod.Meth.Appl.Sci.* **10** 3 (2000) 409-440.
- [92] B. LUCQUIN-DESREUX. Discrétisation de l'opérateur de Fokker-Planck dans le cas homogène, C.R. Acad. Sci. Paris, A 314, série 1, pp. 407-411, 1992.
- [93] B. LUCQUIN-DESREUX and S.MANCINI, A finite element approximation of grazing collisions. (submitted).
- [94] L. MIEUSSENS, Thèse de l'université de Bordeaux, 1999.
- [95] L. PARESCHI, G. RUSSO, G.TOSCANI, Methode spectrale rapide pour l'equation de Fokker Planck Landau, C. R. Acad. Sci. Paris, t.330, Serie I, pp.517-522, (2000).
- [96] L. PARESCHI, G. RUSSO, Fast spectral methods for Boltzmann and Landau integral operators of gas and plasma kinetic theory, con G.Russo. Proceedings Analisi Numerica metodi e software matematico, Annali Universit  di Ferrara, Sez.VII - Sc. Mat., Vol.XLV, (2000), 329-341.
- [97] L. PARESCHI, G. RUSSO, G. TOSCANI, Fast spectral methods for the Fokker-Planck-Landau collision operator, J. Comp. Phys. 165, pp. 216-236, (2000).
- [98] L.Pareschi, G.Toscani, C.Villani, Spectral methods for the non cut-off Boltzmann equation and numerical grazing collision limit, Numerische Mathematik, 93, pp.527-548, (2003).
- [99] M.S. PEKKER and V.N. KHUEDIK. Conservative Difference Schemes for the Fokker-Planck equation, U.S.S.R. Comput. Maths. Math. Phys. Vol. 24,3, pp. 206-210, 1984.
- [100] I.F. POTAPENKO and V.A. CHUYANOV. A completely conservative difference scheme for the two-dimensional Landau equation. U.S.S.R. Comput. Maths. Math. Phys., Vol 20, 2, pp. 249-253, 1980.
- [101] M.N. ROSENBLUTH, W. MACDONALD and D.L. JUDD. Fokker-Planck equation for an Inverse-Square Force. The Physical Review Vol 107, 1, pp. 1-6, 1957.
- [102] W. M. MACDONALD, M. N. ROSENBLUTH et W. CHUCK, *Relaxation of a system of particles with Coulomb interactions*, Phys. Rev., Vol. 107, N  2, (1957).
- [103] J. SHAEFFER, Convergence of a difference scheme for the Vlasov-Poisson-Fokker-Planck system in one dimension, SIAM J of Num. Anal., Vol. 35, p. 1149, 1998.
- [104] G. TOSCANI, Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation, preprint of Univ. Pavia, Dept of Math., 1997.
- [105] C. VILLANI *On the Landau equation: weak stability and global existence*, Adv. Diff. Eq., **1**, (5): pp. 793-816, (1996).
- [106] J.C. WHITNEY. Finite Difference Methods for the Fokker-Planck Equation. J. Comput. Phys. Vol 6, pp. 483-509, 1970.

- [107] S. I. BRAGINSKII, *Transport processes in a plasma*, Reviews of Plasma Physics, Vol. 1, M. A. Leontovitch (ed.), (1965).
- [108] CHEN, F.F., *Introduction to plasma physics*, Plenum Press, 1977.
- [109] A. DECOSTER - Fluid Equations and Transport Coefficients of Plasmas - in Modeling of collisions, P.A. Raviart editor, Masson (1998).
- [110] DELCROIX, J.L., BEERS, A., Physique des plasmas, CNRS Editions, 1994.
- [111] ISHIMARU. Principles of Plasma physic
- [112] N.A. KRALL, D. A. TIDMANN, Shock waves in collisionless plasmas, Wiley-Interscience, 1971.
- [113] N. A. KRALL, A. W. TRIEVELPIECE, *Principles of plasma Physics*, Mc Graw- Hill, New-York, 1973.
- [114] L. LANDAU, E. LIFSCHITZ : Mécanique des fluides, *Mir*.
- [115] PALMADESSO, P.J., MITCHELL, H.G., GANGULI, S.B., multimoment fluid simulation of transport processes in the auroral zones, American geophysical union, (1988).
- [116] PANTELLINI, F., Etude de la structure des chocs non collisionnels dans les plasmas spatiaux, *Thèse de l'Université Paris VII*, 1992.
- [117] PARKS, G., *space plasmas an introduction* , Addison wesley, the advanced book program, (1991).
- [118] SCHUNCK, R.W.: Mathematical structure of transport equations for mutispecies flows, *Rev. geophys. space phys.*, 15, 429-445, 1977.
- [119] SPITZER, L, HARM, R, Phys. rev, vol 89, p.977, 1953.
- [120] G.R., WILSON, Semikinetik modeling of the outflow of ionospheric plasma through the topside collisional to collisionless transition region, *Jal. Geophys. Res.*, Vol. 97, pp. 10551, 1992.
- [121] ZAGDEEV, R., KENNEL, C., Des ondes de chocs sans collisions, *Pour la science*, p88, no 164, Juin 1991.

#### LORENTZ-COMPTON-KOMPANEETS

- [122] M. BAYET, J. L. DELCROIX, J. F. DENISSE *Théorie cinétique des plasmas homogènes faiblement ionisés, II*, *J. Phys. Rad.*, **16**, pp. 274-280, (1955).
- [123] R.E. CAFLISH, C.D. LEVERMORE, *Equilibrium for radiation in a homogeneous plasma*, *Phys. Fluids* **29** (1986) 748-752.
- [124] G. COOPER, *Compton Fokker-Planck equation for hot plasmas*, *Phys. Rev. D* **3** (1974) 2312-2316.
- [125] H. DREICER, *Kinetic Theory of an Electron-Photon Gas*, *Phys. Fluids* **7** (1964) 735-753.
- [126] M. ESCOBEDO, M.A. HERERO, J.J.L. VELASQUEZ, *A nonlinear Fokker-Planck equation modelling the approach to thermal equilibrium in a homogeneous plasma*, *Trans. Amer. Math. Soc.* **350** (1998) 3837-3901.
- [127] M. ESCOBEDO, S. MISCHLER, article en préparation et note aux CRAS.
- [128] A.S. KOMPANEETS, *The establishment of thermal equilibrium between quanta and electrons*, *Soviet Physics JETP* **4** (1957) 730-737.

#### MILIEUX GRANULAIRES

- [129] A. BALDASSARRI, A. PUGLISI, and U. MARCONI, Kinetics models of inelastic gases, *Math. Models Meth. Appl. Sci.*, 12, p.965-984, 2002.

- [130] D. BENEDETTO, E. CAGLIOTI, The collapse phenomenon in one -dimensional inelastic point particle systems, *Physica D*, Vol. 132, p 457, 1999
- [131] D. BENEDETTO, E. CAGLIOTI, M. PULVIRENTI, A kinetic equation for granular media, *Math. Mod. and Num. Anal, M2AN*, Vol 31, N 5, p 615, 1997.
- [132] D. BENEDETTO, E. CAGLIOTI, J.A. CARILLO, M. Pulvirenti, A non-Maxwellian equilibrium distribution for the one-dimensional granular media, *J. Stat. Phys.*, 91(5/6) p. 979, 1998.
- [133] D. BENEDETTO, E. CAGLIOTI, F. GOLSE, M. PULVIRENTI, A hydrodynamical model arising in the context of granular media, *Computers & Mathematics with Applications*, V. 38, Issues 7-8, p. 121-131, October 1999.
- [134] S. Mac Namara, W.R. Young, Inelastic collapse and clumping in a one-dimensional granular medium, *Phys of Fluids*, A4, Vol 3, p. 496, 1992.
- [135] S. Mac Namara, W.R. Young, Kinetics of a one-dimensional granular medium in the quasi-elastic limit, *Phys of Fluids*, A5, Vol 1, p. 1619, 1993.
- [136] S. Mac Namara, W.R. Young, Inelastic collapse in two dimension, *Phys. Rev. E*, 50, R28, 1994.

#### MÉTHODES DE MOMENTS POUR LE TRANSFERT RADIATIF, HYDRODYNAMIQUE RADIATIVE

- [137] E. Audit, P. Charrier, J.-P. Chieze and B. Dubroca, A radiation-hydrodynamics scheme valid from the transport to the diffusion limit, JCP. [lanl.arxiv.org/abs/astro-ph/0206281](https://arxiv.org/abs/astro-ph/0206281), (2002)
- [138] A.M. ANILE, S. PENNISI, M. SAMMARTINO. A thermodynamical approach to Eddington factors, *J. Math Phys.*, 31,(1992).
- [139] J.I. CASTOR, Radiation Hydrodynamics, Cambridge University Press, (2004).
- [140] J.L. FEUGEAS and B. DUBROCA, Entropy moment closure hierarchy for the radiative transfer equation, JCP.
- [141] B. DESPRES, Hyperbolic systems of conservation laws with equality or convex constraints and entropy, Rapport du laboratoire d'analyse numérique, R 00026 (2000).
- [142] N.Y. GNEDIN, T. ABEL, Multi-dimensionnal cosmological radiative transfer with a variable Eddington tensor formalism, *New astronomy*, 6, pp 437-455, 2001.
- [143] J.C. HAYES, M.L. NORMAN, Beyond flux limited diffusion: parallel algorithms for multi-dimensionnal radiation hydrodynamics, *The astrophysical journal*, 147, pp 197-220, 2003.
- [144] C. D. LEVERMORE, Relating Eddington factors to flux limiters, *J. Quant. Spectros. Radiat. Transfer*, Vol 31, No 2, pp. 149-160, 1984.
- [145] C. D. LEVERMORE, Moment closure hierarchies for kinetic theories. *J. Stat. Phys.* 83, No.5-6, 1021-1065 1996
- [146] C. D. LEVERMORE, Entropy-based moment closures for kinetic equations. *Transp. Theory Stat. Phys.* 26, No.4-5, 591-606 (1997)
- [147] R.B. LOWRIE, J.E. MOREL and J.A. HITTINGER, The coupling of radiation ad hydrodynamics, *the astrophysical journal*, 521, 432-450 (1999).
- [148] D. MIHALAS and B.W. MIHALAS, Foundations of radiation hydrodynamics, Oxford press university (1984).
- [149] G. N. MINERBO, Maximum entropy Eddington factors, *J. Quant. Spectros. Radiat. Transfer*, Vol 20 pp541-545, 1978
- [150] I. MÜLLER and T. RUGGERI, Rational extended thermodynamics. 2nd ed. Springer Tracts in Natural Philosophy. 37. New York, NY: Springer.
- [151] A. MUNIER and R. WEAVER, *Computer Physics Reports* 3, 125-164 (1986).
- [152] G. L. OLSON, L. H. AUER and M. L. HALL, Diffusion, P1, and other approximate forms of radiation transport. Los Alamos report LA-UR-99-471

- [153] G. C. POMRANING, The equations of radiation hydrodynamics, Pergamon Press, 1973.
- [154] R. SENTIS, Sur les équations de diffusion multigroupe en transfert radiatif, Note CEA 2603 (1989).
- [155] J.M. SMIT, J. CERNOHORSKY, C.P. DULLEMOND, hyperbolicity and critical points in two-moment approximate radiative transfer, *Astron. Astrophys.* **325**, 203-211, (1997).
- [156] J. TASSART, Transfert de Rayonnement, in *La Fusion Thermonucléaire inertielle par laser*, R. Dautray and J. P. Watteau editors, Eyrolles, 1993.

LIMITE DE DIFFUSION POUR LES SYSTÈMES HYPERBOLIQUES: SCHÉMAS PRÉSERVANT L'ASYMPTOTIQUE

- [157] M.L. ADAMS, Subcell balance methods for radiative transfer on arbitrary grids. *Transp.Theor.Stat.Phys.* **27** 4&5 (1997) 385-431.
- [158] E. AUDUSSE, F. BOUCHUT, M.-O. BRISTEAU, R. KLEIN, B. PERTHAME. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comp.*, **25**(6):2050–2065, (2004).
- [159] R. BOTCHORISHVILI, B. PERTHAME et A.VASSEUR, Equilibrium schemes for scalar conservation laws with stiff sources. *Inria report RR-3891* (2000).
- [160] F. BOUCHUT. Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources, *Frontiers in Mathematics series*, Birkhäuser, (2004).
- [161] F. BOUCHUT, H. OUNAÏSSA, B. PERTHAME. Upwinding of source term at interfaces for Euler equations with high friction. preprint, (2005).
- [162] R.E. CAFLISCH, S. JIN et G. RUSSO, Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM J.Numer.Anal.* **34** 1 (1997) 246-281.
- [163] F. GOLSE, S. JIN et C.D. LEVERMORE, The convergence of numerical transfer schemes in diffusive regimes I: discrete-ordinate method. *SIAM J.Numer.Anal.* **36** 5 (1999) 1333-1369.
- [164] L. GOSSE, Localization effects and measure source terms in numerical schemes for balance laws *Math. Comput.* **71**, No.238, 553-582 (2002).
- [165] L. GOSSE, G. TOSCANI, An asymptotic preserving well-balanced scheme for the hyperbolic heat equation, *CRAS Série I*, **334**, p. 1-6, 2002.
- [166] L. GOSSE, G. TOSCANI, Space localization and well-balanced schemes for discrete kinetic models in diffusive regime. *SIAM J.Numer. Anal.* Vol **41**, No 2641-658.
- [167] L. GOSSE, G. TOSCANI, Asymptotic-preserving and well-balanced schemes for radiative transfer and the Rosseland approximation. *Numer. Math.* **98**, No.2, 223-250 (2004)
- [168] S. JIN, Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J.Sci.Comput.* **21** 2 (1999) 441-454.
- [169] S. JIN, Numerical integrations of systems of conservation laws of mixed type. *SIAM J.Appl.Math.* **55** 6 (1995) 1536-1551.
- [170] S. JIN and C.D. LEVERMORE, The discrete-ordinate method in diffusive regimes. *Transp.Theor.Stat.Phys.* **20** 5&6 (1991) 413-439.
- [171] S. JIN and C.D. LEVERMORE, Fully-discrete numerical transfer in diffusive regimes. *Transp.Theor.Stat.Phys.* **22** 6 (1993) 739-791.
- [172] S. JIN and C.D. LEVERMORE, Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. *J.Comput.Phys.* **126** (1996) 449-467.
- [173] S. JIN and L. PARESCHI, Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes. *J.Comput.Phys.* **161** (2000) 312-330.
- [174] S. JIN, L. PARESCHI and G. TOSCANI, Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. *SIAM J.Numer.Anal.* **35** 6 (1998) 2405-2439.

- [175] S. JIN, L. PARESCHI and G.TOSCANI, Uniformly accurate diffusive relaxation schemes for multiscale transport equations. *SIAM J.Numer.Anal.* (2000).
- [176] JIN S., LEVERMORE C.D., Numerical Schemes for Hyperbolic Systems of Conservation Laws with Stiff Diffusive Relaxation, J.C.P., vol126, 1996.
- [177] S. JIN and Z. XIN, The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm.Pure Appl.Math.* **XLVIII** (1995) 235-276.
- [178] A.KLAR, An asymptotic-induced scheme for non stationary transport equations in the diffusive limit. *SIAM J.Numer.Anal* **35** 3 (1998) 1073-1094.
- [179] E.W. LARSEN, The asymptotic diffusion limit of discretized transport problems. *NSE* **112** (1992) 336-346.
- [180] E.W. LARSEN and J.E. MOREL, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. II. *J.Comput.Phys.* **83** 1 (1989) 212-236.
- [181] E.W. LARSEN, J.E. MOREL and W.F.MILLER Jr., Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *J.Comput.Phys.* **69** 2 (1987) 283-324.
- [182] W.F. MILLER Jr. and T. NOH, Finite differences versus finite elements in slab geometry, even-parity transport theory. *Transp.Theor.Stat.Phys.* **22** 2 & 3 (1993) 247-270.
- [183] J.E. MOREL, T.A. WAREING and K.SMITH, A linear-discontinuous spatial differencing scheme for  $S_n$  radiative transfer calculations. *J.Comput.Phys.* **128** (1996) 445-462.
- [184] NALDI G., PARESCHI L., Numerical Schemes for Hyperbolic Systems of Conservation Laws with Stiff Diffusive Relaxation, *SIAM Journal on Num. Anal.*, 37, p. 1246-1270, 2000
- [185] L.PARESCHI, Central differencing based numerical schemes for hyperbolic conservation laws with relaxation terms. Preprint.
- [186] G. SAMBA, Limite asymptotique d'un schéma d'éléments finis linéaires discontinus lumpés en régime diffusion. *Rapport CEA*.

#### SYSTÈMES HYPERBOLIQUES: THÉORIE ET MÉTHODES NUMÉRIQUES

- [187] AUDUSSE,E., BOUCHUT,F., BRISTEAU, M.-O., KLEIN, R., PERTHAME, B. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comp.*, 25(6):2050–2065, 2004.
- [188] F. BOUCHUT, Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources, *Frontiers in Mathematics series*, Birkhäuser, 2004.
- [189] CHEN, C.D. LEVERMORE and LIU, Hyperbolic conservation laws with stiff relaxation terms and entropy. *Comm.Pure Appl.* **47** (1994) 187-830.
- [190] S. CORDIER, "Hyperbolicity of the hydrodynamical model of plasmas under the quasineutrality hypothesis", *Mathematical Models in Applied Science (M2AS)*, Vol 18, p. 627-647, 1995.
- [191] DAL MASO, G., LE FLOCH, P., MURAT, F, Definition and weak stability of nonconservative product, Preprint Ecole Polytechnique, *J. Math. Pures Appliquées*, to appear, (1994).
- [192] GODLEWSKI, E., RAVIART, P.-A., *Hyperbolic systems of conservation laws*, Ellipse, (1991).
- [193] E. GODLEWSKI and P.A. RAVIART, *Numerical approximations of hyperbolic systems of conservation laws*. Applied Mathematical Sciences 118, Springer-Verlag New York (1996).
- [194] J. GLIMM, G. MARSHALL and B.J. PLOHR, A generalized Riemann problem for quasi one dimensional gas flows. *Adv.Appl.Math.* **5** (1984) 1-30.
- [195] GLIMM, J., Solutions in the large for nonlinear hyperbolic systems of equations, *Comm. Pure Appl. Math.*, 18: 698-715, 1965.

- [196] L. GOSSE and A.Y. LEROUX, A well-balanced scheme designed for inhomogeneous scalar conservation laws. *C.R.Acad.Sc.Paris* **I 323** (1996) 543-546.
- [197] L. GOSSE, Localization effects and measure source terms in numerical schemes for Balance laws, *Math of Comp*, Vol 71, No 238, pp 553-582.
- [198] L. GOSSE, A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. *Math.Mod.Meth.Appl.Sci.* **11** (2001) 339-365.
- [199] L. D. LIFSHITZ and E. LANDAU, *Mécanique des fluides*, Éditions MIR, Paris.
- [200] LE FLOCH, P., LIU, T.P., Existence theory for nonlinear hyperbolic systems in nonconservative form, in *Forum Mathematicum*, 5, 261-280, 1993.
- [201] R.J. LEVEQUE, Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *J.Comput.Phys.* **146** 1 (1998) 346-365.
- [202] J.M. GREENBERG and A.Y. LEROUX, A well balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J.Num.Anal.* **33** (1996) 1-16.
- [203] GREENBERG J-M., LE ROUX A-Y, A well-balanced scheme for the numerical processing of source terms in hyperbolic equations, *SIAM J. Numer. Anal.* 33, No.1, 1-16 (1996).
- [204] P.L. LIONS, B. PERTHAME and P.E. SOUGANIDIS, Existence of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates. *Comm.Pure Appl.Math.* **49** 6 (1996) 599-638.
- [205] A. MAJDA : Compressible fluid flow and systems of conservation laws in several space variables, *Appl. Math. Sci.* 53, *Springer* 184.
- [206] P.A. RAVIART, An analysis of particle methods, *Numerical methods in fluid dynamics*, Lect. 3rd 1983 Sess. C.I.M.E., Como/Italy 1983, *Lect. Notes Math.* 1127, 243-324 (1985).
- [207] D. SERRE : Systèmes de lois de conservation II, *Diderot editeur, Arts et Sciences*, 1996.
- [208] E.F.Toro : *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer-Verlag, Second Edition, Chapter 10, 1999.
- [209] B. PERTHAME, *An introduction to kinetic schemes for gas dynamics. An introduction to recent developments in theory and numerics for conservation laws*. L.N. in Computational Sc. and Eng., 5, D. Kroner, M. Oehlberger and C. Rohde eds, Springer (1998).
- [210] K.H. PRENDERGAST and K. XU, Numerical hydrodynamics for gas-kinetic theory. *J.Comput.Phys.* **109** (1993) 53-66.
- [211] K.H. PRENDERGAST and K. XU, Numerical Navier-Stokes solutions from gas kinetic theory. *J.Comput.Phys.* **114** (1994) 9-17.
- [212] B. VANLEER, On the relation between the upwind differencing schemes of Engquist-Osher, Godunov and Roe. *SIAM J.Sci.Stat.Comp.* **5** (1984) 1-20.

# TRAVAUX PRÉSENTÉS

- [a1] C. Buet, "A discrete-velocity scheme for the Boltzmann operator of rarefied gas dynamics", *Transp. Theory Stat. Phys.* 25, No.1, 33-60 (1996).
- [a2] C. Buet, "Conservative and entropy schemes for Boltzmann collision operator of polyatomic gases", *Math. Models Methods Appl. Sci.* 7, No.2, 165-192 (1997).
- [a3] C. Buet, S. Cordier, P. Degond, M. Lemou, "Fast algorithms for numerical, conservative, and entropy approximations of the Fokker-Planck-Landau equation", *J. Comput. Phys.* 133, No.2, 310-322 (1997).
- [a4] C. Buet, S. Cordier, P. Degond, "Regularized Boltzmann operators", *Comput. Math. Appl.* 35, No.1-2, 55-74 (1998).
- [a5] C. Buet, S. Cordier, "Conservative and entropy decaying numerical scheme for the isotropic Fokker-Planck-Landau equation", *J. Comput. Phys.* 145, No.1, 228-245 (1998).
- [a6] C. Buet, S. Cordier, "Numerical analysis of conservative and entropy schemes for the Fokker-Planck-Landau equation", *SIAM J. Numer. Anal.* 36, No.3, 953-973 (1999).
- [a7] C. Buet, S. Dellacherie, R. Sentis, "Numerical solution of an ionic Fokker-Planck equation with electronic temperature", *SIAM J. Numer. Anal.* 39, No.4, 1219-1253 (2001).
- [a8] C. Buet, S. Cordier, B. Lucquin, "The grazing collision limit for the Boltzmann-Lorentz model", *Asymptotic Anal.* 25, No.2, 93-107 (2001).
- [a9] C. Buet, S. Cordier, "Numerical analysis of the isotropic Fokker-Planck-Landau equation", *J. Comput. Phys.* 179, No.1, 43-67 (2002).
- [a10] C. Buet, S. Cordier, P.-A. Raviart, "Multifluid ionization models", *Math. Models Methods Appl. Sci.* 12, No.1, 17-48 (2002).
- [a11] C. Buet, S. Cordier, B. Lucquin-Desreux, S. Mancini, "Diffusion limit of the Lorentz model: asymptotic preserving schemes", *M2AN, Math. Model. Numer. Anal.* 36, No.4, 631-655 (2002).
- [a12] C. Buet, S. Cordier, "Numerical Method for the Compton Scattering Operator", *Series on Advances in Mathematics for Applied Sciences*, Vol. 63, World Scientific Publishing Co. (2003).
- [a13] C. Buet, B. Despres, "Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics", *Journal of Quantitative Spectroscopy and Radiative Transfer*, Volume 85, Issues 3-4, 385-418 (2004).
- [a14] C. Buet, S. Cordier, V. Dos Santos, "A Conservative and Entropy Scheme for a Simplified Model of Granular Media" *Transp. Theory Stat. Phys.* 33, No.2, 125-155 (2004).
- [cr1] C. Buet, S. Dellacherie, R. Sentis, "Numerical solution of a ionic Fokker-Planck equation with electronic temperature. (Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique)" *C. R. Acad. Sci., Paris, Sér. I, Math.* 327, No.1, 93-98 (1998).
- [cr2] C. Buet, S. Cordier, "Asymptotic preserving scheme and numerical methods for radiative hydrodynamic models", *Comptes Rendus Mathématique*, Volume 338, 951-956 (2004).
- [cr3] C. Buet, S. Cordier, "On the non existence of monotone linear scheme for some linear parabolic equations (Sur la non existence de schémas linéaires monotones pour certaines équations paraboliques linéaires)", *Comptes Rendus Mathématique*, Volume 340, 399-404 (2005).

- [np1] C. Buet, S. Dellacherie, "About the Chang and Cooper method for linear Fokker-Planck equations".
- [s1] C. Buet, S. Cordier, "An asymptotic preserving scheme for Hydrodynamics Radiative Transfer Models", soumis à Num. Math.
- [s2] C. Buet, B. Despres, "Asymptotic Preserving and Positive Schemes for Radiation Hydrodynamics", soumis au JCP.



# A Discrete-Velocity Scheme for the Boltzmann Operator of Rarefied Gas Dynamics

C. Buet

CEA-CEL-V

94195 Villeneuve Saint Georges Cedex, France

## Abstract

We propose a conservative and entropic discrete-velocity method to compute the solutions of the Boltzmann equation in the case of monoatomic species. We begin by defining a discrete collision kernel on a velocity lattice which verifies all the properties of the continuous kernel. The continuous Boltzmann equation will be replaced by a Boltzmann equation for a discrete velocity gas, which is a hyperbolic system. This equation will be discretized by a finite volume scheme. For the evaluation of the collision term we use acceleration procedures of Monte Carlo type. The possibilities of our scheme will be illustrated by numerical tests in 1 and 2 space dimensions.

## 1 Introduction

In rarefied aerodynamics the kinetic model which is currently used is the Boltzmann equation, because we consider only binary collisions. Most of the numerical codes for rarefied aerodynamics are based on Monte-Carlo procedures, following the earlier work of gas dynamicists [1]. A description of these methods is given, for example, in [7, 13, 1]. However, in many situations, the numerical fluctuations which originate in the use of random sequences, lead to extremely noisy results and motivate the search for alternate, more accurate methods [11, 14, 15, 9, 10]. The aim of this paper is to present some attempts in this direction. The discrete velocity model that we use is close to the model presented in [11]. The most important point of our work is the technique to reduce the cost of the collision phase.

## 2 Conservation properties of the Boltzmann equation

We consider the Boltzmann equation for a monoatomic gas

$$(1) \quad \frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q(f, f), \quad f|_{t=0} = f_0(x, v)$$

$$(2) \quad Q(f, f) = \int_{\mathbb{R}^3} \left( \int_{S^2} q(v - v_*, \omega) (f' f'_* - f f_*) d\omega \right) dv_*$$

with the following notations

$$S^2 = \{\omega \in \mathbb{R}^3, |\omega|^2 = 1\}$$

$$f = f(x, v, t), \quad f_* = f(x, v_*, t), \quad f' = f(x, v', t), \quad f'_* = f(x, v'_*, t)$$

$$(3) \quad v' = \frac{v + v_*}{2} + R_\omega\left(\frac{v - v_*}{2}\right), \quad v'_* = \frac{v + v_*}{2} - R_\omega\left(\frac{v - v_*}{2}\right),$$

where  $R_\omega(\vec{u})$  is defined by

$$R_\omega(\vec{u}) = \cos \theta \vec{u} + |\vec{u}| \sin \theta (\cos \varphi \vec{i}_{\vec{u}} + \sin \varphi \vec{j}_{\vec{u}}),$$

with  $\omega = (\cos \theta, \sin \theta \cos \varphi, \sin \theta \sin \varphi)$ ,  $|\vec{j}_{\vec{u}}| = |\vec{i}_{\vec{u}}| = 1$  and  $(\vec{u}, \vec{i}_{\vec{u}}, \vec{j}_{\vec{u}})$  form an orthogonal base of  $\mathbb{R}^3$ . The velocities  $v$  and  $v_*$  are pre-collisional velocities, while  $v'$  and  $v'_*$  are post-collisional velocities. They satisfy

$$(4) \quad v + v_* = v' + v'_* \quad (\text{conservation of momentum})$$

$$(5) \quad |v|^2 + |v_*|^2 = |v'|^2 + |v'_*|^2 \quad (\text{conservation of energy})$$

Finally,  $q(v, \omega)$  is defined by

$$(6) \quad q(v, \omega) = |v| \sigma(v, \omega),$$

where  $\sigma(v, \omega) = \sigma(|v|, \cos \theta)$  is the differential scattering cross section. Well-known properties of the Boltzmann collision operator (2) are conservation of mass, momentum and energy, and decay of entropy

$$(7) \quad \int_{\mathbb{R}^3} Q(f, f) \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv = 0$$

$$(8) \quad \int_{\mathbb{R}^3} Q(f, f) \log(f) dv \leq 0.$$

More precisely, let  $\psi(v)$  be any smooth test function. We have

$$(9) \quad \begin{aligned} \int_{\mathbb{R}^3} Q(f, f) \psi dv &= \\ &= -\frac{1}{4} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \left( \int_{S^2} q(v - v_*, \omega) (f' f'_* - f f_*) (\psi' + \psi'_* - \psi - \psi_*) d\omega \right) dv dv_* \end{aligned}$$

It is well known that the only functions  $\psi$  such that  $\int Q(f, f) \psi dv = 0$  are linear combinations of 1,  $v$ , and  $|v|^2$ .

Similarly, (8) follows from (9) (with  $\psi = \log f$ ). Finally, from (9), it follows that any equilibrium distribution function, i.e., any  $f$  satisfying  $Q(f, f) = 0$  is a Maxwellian

$$(10) \quad f(v) = \frac{\rho}{(2\pi RT)^{3/2}} \exp\left(-\frac{|v - u|^2}{2RT}\right),$$

where  $\rho, T \in \mathbb{R}$ ,  $\rho > 0, T > 0$ , and  $u \in \mathbb{R}^3$ . ( $\rho, u, T$ ) are the density, mean velocity and temperature of gas. In the homogeneous case, from (8) follows the  $H$ -theorem

$$(11) \quad \frac{d}{dt} \int \int_{\mathbb{R}^3} f(v, t) \log f(v, t) dx dv = \int_{\mathbb{R}^3} Q(f, f) (1 + \log f) dv \leq 0,$$

showing that the entropy  $H(f) = \int f \log f dv$  can only decrease during a time evolution. Furthermore,  $H(f)$  can only be minimal if  $f$  is an equilibrium distribution, i. e. if  $f$  is a Maxwellian. The same result holds in the inhomogeneous case in the absence of boundaries

$$(12) \quad \frac{d}{dt} \int \int_{\mathbb{R}^3 \times \mathbb{R}^3} f(x, v, t) \log f(x, v, t) dx dv = \int_{\mathbb{R}^3 \times \mathbb{R}^3} Q(f, f) (1 + \log f) dx dv \leq 0.$$

It is extremely important to preserve this latter property in any numerical discretization, because it expresses conditions that any dynamics must satisfy. Thus, a numerical scheme should satisfy discrete analogues of (7), (8), (9), should exhibit discrete invariants of collision, connected with discrete equilibrium distribution functions. The failure of deterministic particle methods to satisfy these requirements naturally leads to the following discrete-velocity model.

### 3 A discrete-velocity model

We first deal with the space homogeneous equation.

$$(13) \quad \frac{df}{dt} = Q(f, f), \quad f|_{t=0} = f_0(v),$$

where  $Q(f, f)$  is given by (2). We introduce a regular discretization of  $\mathbb{R}^3$ : Let  $\Delta v > 0$ ,  $v_i = i\Delta v$ ,  $i = (i_1, i_2, i_3) \in \mathbb{Z}^3$ , and  $f_i \simeq (\Delta v)^3 f(v_i)$ . We derive an approximation of (2) by using a quadrature formula for the integral with respect to  $v_*$ , the quadrature points of which are the lattice points of  $\Delta v \mathbb{Z}^3$ . We let

$$(14) \quad \begin{aligned} & \left[ \int_{\mathbb{R}^3} \left( \int_{S^2} q(v - v_*, \omega) (f' f'_* - f f_*) d\omega \right) dv_* \right]_{v=v_i} \\ & \simeq \sum_{j \in \mathbb{Z}^3} \left( \int_{S^2} q(v_i - v_j, \omega) (f(v'_i) f(v'_j) - f(v_i) f(v_j)) d\omega \right) \Delta v^3, \end{aligned}$$

where  $v'_i, v'_j$  are defined by (3) and where  $v$  and  $v_*$  are replaced by  $v_i$  and  $v_j$ . At this point, the loss term already depends on the values  $f(v_i), f(v_j)$  of the distribution function  $f$ , while the gain term still depends on  $f$ , through an integral over  $\omega \in S^2$ . We have to find a quadrature formula for the integral at the right-hand side of (14), which involves values of  $f$  at the lattice points  $\Delta v \mathbb{Z}^3$ . Formula (3) shows that, when  $\omega$  varies in  $S^2$ ,  $v'$  varies on the sphere of largest diameter  $(v, v_*)$ , and  $v'_*$  is diametrically opposed to  $v'$  (i. e.  $(v', v'_*)$  is another largest diameter of the same sphere). Furthermore, quite generically, the sphere of largest diameter  $(v_i, v_j)$  contains other lattice points  $v_k$  and such points appear in pairs of diametrically opposed points  $(v_k, v_l)$ . The set  $S_{ij}$  of such pairs can be defined by

$$S_{ij} = \{(k, l) \in \mathbb{Z}^3 \times \mathbb{Z}^3, \quad k + l = i + j, \quad |k|^2 + |l|^2 = |i|^2 + |j|^2\}.$$

Furthermore, for  $(k, l) \in S_{ij}$ , we can define a unique  $\omega_{ij}^{kl} \in S^2$ , such that formula (3) holds for  $v = v_i, v_* = v_j, v' = v_k, v'_* = v_l$ .

Now, for the integral with respect to  $\omega$  which appears at the right hand-side of (14), we can use a quadrature formula where the  $\omega_{ij}^{kl}$  are the quadrature points and the weights are all equal to  $4\pi/\text{Card}(S_{ij})$ . This gives

$$(15) \quad \int_{S^2} q(v_i - v_j, \omega) f(v'_i) f(v'_j) d\omega \simeq \sum_{(k, l) \in S_{ij}} \frac{4\pi}{\text{Card}(S_{ij})} q(v_i - v_j, \omega_{ij}^{kl}) f(v_k) f(v_l)$$

and in (15), the distribution function  $f$  has been replaced by its values  $f(v_k), f(v_l)$  at the lattice points. The overall approximate collision operator is now written by letting  $\bar{f} = \{f_i, i \in$

$\mathbb{Z}^3\}$ ,

$$(16) \quad \begin{aligned} Q(f, f)(v_i) &\simeq \frac{1}{(\Delta v)^3} \bar{Q}(\bar{f}, \bar{f})_i \\ &= \frac{1}{(\Delta v)^3} \sum_{j \in \mathbb{Z}^3} \sum_{(k, l) \in S_{ij}} \frac{4\pi}{\text{Card}(S_{ij})} q(v_i - v_j, \omega_{ij}^{kl}) (f_k f_l - f_i f_j) \end{aligned}$$

or

$$(17) \quad \bar{Q}(\bar{f}, \bar{f})_i = \frac{1}{2} \sum_{(j, k, l) \in (\mathbb{Z}^3)^3} (A_{k, l}^{i, j} f_k f_l - A_{ij}^{kl} f_i f_j),$$

with

$$(18) \quad A_{ij}^{kl} = \begin{cases} \frac{4\pi q(v_i - v_j, \omega_{ij}^{kl})}{\text{Card}(S_{ij})} & \text{if } (k, l) \in S_{ij}. \\ 0 & \text{otherwise} \end{cases}$$

Formula (17) shows that this approximation enters the class of discrete-velocity models. No error estimates for the quadrature formula (15) is available yet. One problem is that the number of quadrature points  $\omega_{ij}^{kl}$  on  $S^2$  is a function of  $v_j - v_i$ , which has an (almost) random behaviour. Another one is that, even if this number was known accurately, the location of the points  $\omega_{ij}^{kl}$  on  $S^2$  also varies (apparently) randomly. We have just an estimate of the number of quadrature points  $\omega_{ij}^{kl}$  by adapting a very classical theorem of number theory, on how to split an integer into a sum of three squares of integers (see [8]). Indeed the center of the collisions spheres  $S_{i, j}$  are in  $\frac{1}{2} \cdot \mathbb{Z}^3$ , given a center of collisions sphere  $\frac{\epsilon}{2}$  with  $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3) \in \mathbb{Z}^3$ , by translation, we can suppose that  $\epsilon \in \mathbb{N}^3$  and  $\max_{i=1}^3(\epsilon_i) \leq 2$ . All the spheres having  $\frac{\epsilon}{2}$  for center and which intersect  $\mathbb{Z}^3$  have a radius of the form  $|i - \frac{\epsilon}{2}|$  with  $i \in \mathbb{Z}^3$  such that  $|i - \frac{\epsilon}{2}|^2 \in \mathbb{N} + \left|\frac{\epsilon}{2}\right|^2$ . If we let

$$r_\epsilon(n) = \text{Card}\left(\left\{i \in \mathbb{Z}^3 \mid \left|i - \frac{\epsilon}{2}\right|^2 = n + \left|\frac{\epsilon}{2}\right|^2\right\}\right), \text{ with } n \in \mathbb{N}$$

be the number of points of  $\mathbb{Z}^3$  on the sphere having the center  $\frac{\epsilon}{2}$  and the radius  $(n + \left|\frac{\epsilon}{2}\right|^2)^{\frac{1}{2}}$ , we have the result:

**Lemma 1**

$$\sum_{k=0}^n r_\epsilon(k) = \frac{4}{3} \pi n^{\frac{3}{2}} + O(n)$$

and  $r_\epsilon(n) = O(n^{\frac{1}{2}+\delta})$  for all  $\delta > 0$  or equivalently,  $r_\epsilon(n) = o(n^{\frac{1}{2}+\delta})$  for all  $\delta > 0$ .

For the proof see [2].

Since

$$\text{Card}(S_{i, j}) = \frac{1}{2} r_\epsilon\left(\left|\frac{i - j}{2}\right|^2\right) \text{ with } \epsilon \equiv i - j \pmod{2}, \quad \max_{i=1}^3(\epsilon_i) \leq 2.$$

in the sense of the Cesaro mean value,  $\text{Card}(S_{i, j})$  behaves like  $\frac{4}{3} \pi \left|\frac{i - j}{2}\right|$ . In some sense, this tends to show that the approximation is “reasonably accurate”. But a rigorous proof of

the accuracy of the approximation is missing up to now. We must now bound the velocity domain. The issue is to replace the Boltzmann equation in the whole velocity space domain, by a bounded space one, for which the algebraic properties displayed in section 3.1 still hold. We do that as in [14, 15]. Let  $V$  be a bounded velocity domain, and let  $I(v, v_*, v', v'_*)$  be the following truncation function:

$$(19) \quad I(v, v_*, v', v'_*) = \begin{cases} 1 & \text{if } (v, v_*, v', v'_*) \in V^4 \\ 0 & \text{otherwise.} \end{cases}$$

Now, let us consider the Boltzmann operator:

$$(20) \quad Q(f, f)(v) = \int_V \int_{S^2} q(v - v_*, \omega) (f' f'_* - f f_*) I(v, v_*, v', v'_*) dv_* d\omega, \quad v \in V.$$

It is easily shown that properties (7) to (11) still hold, with the only difference that the coefficient of  $|v|^2$  in (10) is no more positive. Indeed, its positivity for the full space case follows from integrability requirements on the Maxwellian, which can no longer be used because of the boundedness of the domain. The discrete velocity model is now restricted to approximations  $f_i$  of  $(\Delta v)^3 f(v_i)$  for  $v_i \in \Delta v \mathbb{Z}^3 \cap V$  and the discrete Boltzmann operator is of the form (17)

$$(21) \quad \tilde{S}_{ij} = \{(k, l) \in S_{ij} \text{ such that } v_k, v_l \in V\},$$

with  $i, j$  such that  $v_i, v_j \in \Delta v \mathbb{Z}^3 \cap V$ .

$$(22) \quad \tilde{A}_{ij}^{kl} = \begin{cases} \frac{4\pi q(v_i - v_j, \omega_{ij}^{kl})}{\text{Card}(\tilde{S}_{ij})} & \text{if } v_i, v_j \in V \text{ and } (k, l) \in \tilde{S}_{ij} \\ 0 & \text{otherwise.} \end{cases}$$

In practice, all the discrete pre- and post-collisional velocities must be within the computational domain  $V$ , and the number of “allowed” post-collisional pairs must be considered in the quadrature formula for integrals on  $S^2$ .

## 4 Properties of the discrete-velocity model

First, since the cross section is supposed to depend only of  $|v|$  and  $\cos(\theta)$ , it is easy to see that the tensor  $A_{ij}^{kl}$  is positive and satisfies the following symmetry properties:

$$(23) \quad A_{ij}^{kl} = A_{ji}^{kl} = A_{ij}^{lk}$$

and

$$(24) \quad A_{ij}^{kl} = A_{kl}^{ij}.$$

Property (23) expresses that the two pre-collisional particles are undistinguishable. The same is also true for the two post-collisional particles. Property (24) is the microreversibility. We have the identity (9): let  $\bar{\psi} = (\psi_i)_{i \in \mathbb{Z}^3}$  be a test sequence. Then

$$(25) \quad \sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i \psi_i = -\frac{1}{8} \sum_{(i, j, k, l) \in (\mathbb{Z}^3)^4} (A_{kl}^{i, j} f_k f_l - A_{ij}^{kl} f_i f_j) (\psi_k + \psi_l - \psi_i - \psi_j).$$

Using the definition of the tensor  $A_{ij}^{kl}$  it is easy to see that we have the discrete analogue of conservation of mass, momentum and energy

$$(26) \quad \sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i \begin{pmatrix} 1 \\ v_i \\ |v_i|^2 \end{pmatrix} = 0,$$

and by the microreversibility property of the tensor, the decay of entropy hold (see [4]):

$$(27) \quad \sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i \log(f_i) \leq 0.$$

The equilibrium state  $\bar{f}^\infty$  is characterized by one of the following properties (see [4]):

1.  $\sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}^\infty, \bar{f}^\infty)_i \log(f_i^\infty) = 0$
2.  $\bar{Q}(\bar{f}^\infty, \bar{f}^\infty)_i = 0$  for all  $i$
3.  $\overline{\log f^\infty}$  is an invariant of collision that is  $\overline{\log f^\infty} \in \{\bar{\varphi} \text{ such that } \varphi_i + \varphi_j - \varphi_k - \varphi_l = 0 \text{ if } A_{ij}^{kl} \neq 0\}$
4.  $f_i^\infty f_j^\infty - f_k^\infty f_l^\infty = 0$  if  $A_{ij}^{kl} \neq 0$

Now, for the specific model given by (18), it is noticeable that the reciprocal of (26) holds, like in the continuous case and so the equilibrium states are discrete Maxwellians:

**Lemma 2** *With  $A_{ij}^{kl}$  defined by (18) the invariants of collisions are given by*

$$\psi_i = Av_i^2 + \langle B, v_i \rangle + C$$

with  $A$  et  $C \in \mathbb{R}$  and  $B \in \mathbb{R}^3$ .

**Proof** We say that  $(i, j) \rightarrow (k, l)$  is a possible collision if  $A_{i,j}^{k,l} \neq 0$ , that implies by the symmetry of  $\mathbb{Z}^3$  that the collision  $(-i, -j) \rightarrow (-k, -l)$  is also possible. Let  $e_1 = (1, 0, 0)$ ,  $e_2 = (0, 1, 0)$ ,  $e_3 = (0, 0, 1)$  be the canonical base of  $\mathbb{Z}^3$ . We search for  $\psi$  such that

$$\psi_i + \psi_j - \psi_k - \psi_l = 0 \text{ for all } A_{i,j}^{k,l} \neq 0.$$

We set

$$a_i = \psi_i + \psi_{-i} \text{ and } b_i = \psi_i - \psi_{-i}.$$

By construction we have  $a_{-i} = a_i$ ,  $b_{-i} = -b_i$  and then  $b_0 = 0$ . We show recursively on  $m = |i|_\infty = \max_{n=1}^3 |i_n|$  that

$$a_i - a_0 = |i|^2 \cdot (a_{e_1} - a_0) \quad b_i = i_1 b_{e_1} + i_2 b_{e_2} + i_3 b_{e_3}.$$

This is evidently true for  $a_i$  when  $i = e_1, e_3$  or  $e_3$  because  $(e_k, -e_k) \rightarrow (e_l, -e_l)$  is a possible collision. For  $b_i$  in this case this is trivial. We suppose now that it is true until rank  $m$ . Let  $i \in \mathbb{Z}^3$  such that  $|i|_\infty = m + 1$ . If  $(i, j) \rightarrow (k, l)$  is a possible collision then, by construction,  $a_i + a_j = a_k + a_l$  and  $b_i + b_j = b_k + b_l$ . Since the following collisions are possible

$$(i, 0) \rightarrow (i_1 e_1 + i_2 e_2, i_3 e_3), \quad (i_1 e_1 + i_2 e_2, 0) \rightarrow (i_1 e_1, i_2 e_2),$$

we have

$$a_i + a_0 = a_{i_1 e_1 + i_2 e_2} + a_{i_3 e_3} = a_{i_1 e_1} + a_{i_2 e_2} + a_{i_3 e_3} - a_0$$

and then

$$a_i - a_0 = (a_{i_1 e_1} - a_0) + (a_{i_2 e_2} - a_0) + (a_{i_3 e_3} - a_0)$$

and for  $b_i$

$$b_i = b_{i_1 e_1} + b_{i_2 e_2} + b_{i_3 e_3}.$$

It suffices then to verify that:

$$(28) \quad a_{(m+1)e_k} - a_0 = (m+1)^2(a_{e_1} - a_0) \quad , \quad b_{(m+1)e_k} = (m+1)b_{e_k}.$$

We define  $u = (m-1)e_k$ ,  $v = me_k + e_l$  and  $w = me_k - e_l$ . Since  $|u|_\infty = m-1$  and  $|v|_\infty = |w|_\infty = m$  the assertion holds for  $u, v, w$ . By the inductive hypothesis and since the following collision is possible

$$\left( (m+1)e_k, u \right) \rightarrow \left( v, w \right)$$

(28) is true: we have for  $a_i$

$$a_{(m+1)e_k} + a_u = a_v + a_w,$$

and then

$$\begin{aligned} a_{(m+1)e_k} - a_0 &= a_w - a_0 + a_v - a_0 - (a_u - a_0) \\ &= (|w|^2 + |v|^2 - |u|^2)(a_{e_1} - a_0) \\ &= (m^2 + 1 + m^2 + 1 - (m-1)^2)(a_{e_1} - a_0) \\ &= (m+1)^2(a_{e_1} - a_0), \end{aligned}$$

and for  $b_i$

$$b_{(m+1)e_k} = b_v + b_w - b_u = mb_{e_k} + b_{e_l} + mb_{e_k} - b_{e_l} - (m-1)b_{e_k} = (m+1)b_{e_k}.$$

Since  $\psi_i = \frac{a_i + b_i}{2}$  we have the result for  $\psi_i$  with  $A = \frac{a_{e_1} - a_0}{2\Delta v}$ ,  $C = \frac{a_0}{2\Delta v}$ ,  $B = (\frac{b_{e_1}}{2\Delta v}, \frac{b_{e_2}}{2\Delta v}, \frac{b_{e_3}}{2\Delta v})$ .  
□

**Remark 1** *This proof shows also that, in the case of a bounded domain for  $v$  of the form  $V = B(\vec{U}, R)$  or  $V = \vec{U} + [-R, R]^3$  (which we use in practice) which define, after a translation of vector  $\vec{U}$ , a bounded domain for  $i$  of the form  $\{i \in \mathbb{Z}^3 / i_1^2 + i_2^2 + i_3^2 \leq M\}$  or  $\{i \in \mathbb{Z}^3 / \sup_{k=1}^3 i_k \leq M\}$ , the result for the form of the invariants of collisions remains valid.*

Since the only invariants of collision are  $1, v, |v|^2$  the constants  $A, B, C$  only depend on the density, mean velocity, and temperature of  $\bar{f}$  (see [4]).

These properties show that the discrete models (17), (18), (17) or (22) satisfy the requirements that we have stated at the end of section (3.1).

## 5 Discretization in space and time

Now we define  $N$  as  $\text{Card}((\Delta v \mathbb{Z}^3) \cap V)$  and we let  $\mathcal{V}_N = \{v_i \in (\Delta v \mathbb{Z}^3) \cap V, i = 1, \dots, N\}$  be the finite set of velocities. We set now

$$S_{ij} = \{\{v_k, v_l\} \in (\Delta v \mathbb{Z}^3)^2, \quad v_k + v_l = v_i + v_j, \quad |v_k|^2 + |v_l|^2 = |v_i|^2 + |v_j|^2\}.$$

$$\tilde{S}_{i,j} = \{\{v_k, v_l\} \in S_{ij} \text{ and } v_k, v_l \in V\}$$

The problem is now to solve

$$(29) \quad \frac{\partial f_i}{\partial t} + v_i \cdot \nabla f_i = Q(\bar{f}, \bar{f})_i = \sum_{j=1}^N \sum_{\{v_k, v_l\} \in S_{ij}} (A_{kl}^{ij} f_k f_l - A_{ij}^{kl} f_i f_j)$$

where  $f_i = f_i(x, t)$  is an approximation of  $(\Delta v)^3 f(x, v, t)$  at the point  $v_i$ ,  $A_{ij}^{kl}$  is defined by (22) and  $\bar{f} = (f_1, \dots, f_N)$ . As usual, we use a splitting method between the transport and the collision phase.

### 5.1 Numerical transport

The numerical treatment of the convective term can be done in a variety of ways, e.g., finite differences, finite volumes, the method of characteristics, or particle methods. We have developed the second one. The first one must be associated with directional splitting which leads to unpleasant directional effects. The third one implies a large dependency upon the data at the previous time step, and would involve heavy storage. The fourth one needs the storage of the positions of the particles on top of the storage of the value of  $f$ . This would also imply too heavy storage. We restrict the presentation to 2-D computations on quadrangular meshes and to a single equation of convection

$$\frac{\partial f(x, t)}{\partial t} + v \cdot \nabla f(x, t) = 0.$$

We introduce a partitioning of the computation domain by a set  $\mathcal{M}$  of cells  $M$ , and take a time increment  $\Delta t > 0$ . Given a cell  $M$  we suppose that we know an approximation  $f_M^n$  of

$$\frac{1}{|M|} \cdot \int_M f(x, n\Delta t) dx.$$

Let  $A, B, C, D$  be the vertices of the cell  $M$ ,  $(A, B)$ ,  $(B, C)$ ,... the sides of the cell,  $n_{AB}$ ,  $n_{BC}$ ,... the unit outward normals of each side,  $l_{AB}$ ,  $l_{BC}$ ,... the length of the sides,  $M_{AB}$ ,  $M_{BC}$ ,... the nearest neighbors of  $M$  and  $AB$ ,  $BC$ ,... the midpoint of each side. On each cell  $M$  we define a function  $g_M$  which verifies

$$f_M^n = \frac{1}{|M|} \cdot \int_M g_M dx,$$

and we define  $g$  by  $g = g_M$  on each cell  $M$ .

We let  $g_{AB}^{in} = \lim_{x \rightarrow AB, x \in M} g(x)$  and  $g_{AB}^{out} = \lim_{x \rightarrow AB, x \notin M} g(x)$  and similarly for the other sides.

The finite volume scheme is defined by

$$f_M^{n+1} = f_M^n - \frac{\Delta t}{|M|} \left( flux(g_{AB}^{in}, g_{AB}^{out}) \cdot l_{AB} + flux(g_{BC}^{in}, g_{BC}^{out}) \cdot l_{BC} \right. \\ \left. + flux(g_{CD}^{in}, g_{CD}^{out}) \cdot l_{CD} + flux(g_{DA}^{in}, g_{DA}^{out}) \cdot l_{DA} \right),$$



where the function  $flux(q^{in}, q^{out})$  is defined by

$$flux(q^{in}, q^{out}) = [\max(v.n, 0)]q^{in} + [\min(v.n, 0)]q^{out},$$

and  $n$  is the outward unit normal for the considered side.

The usual first order finite volume scheme, obtained by taking  $g$  constant on each cell  $M$  i.e.,  $g_M(x) = f_M^n$ , is very dissipative and yields smoothed shock profiles, for which the Rayleigh line is no longer a straight line. We recall that the Rayleigh line is the set of  $\mathbb{R}^2$  which contains the pairs  $(\frac{1}{n}, p_{xx})$ , where  $n$  is the density and  $p_{xx}$  is the longitudinal component of the pressure tensor,  $p_{xx} = \int_{\mathbb{R}^3} (v_x - u_x)^2 f(x, v) dv$ . When  $x$  ranges over the shock region, this set is a straight line. This is a good test for numerical methods as it is noted in [5], because how close the numerical Rayleigh line is to a straight line tells us how good the method is. The scheme also dissipates too much on a nonuniform grid.

We thus turn to Van Leer's method [16] to achieve second order accuracy in space and this leads to much better results. For this we take  $g$  linear on each cell:

$$g_M(x) = f_M^n + (\nabla f)_M^n \cdot (x - x_M)$$

where  $x_M$  is the centroid of the cell and  $(\nabla f)_M^n$  is an approximation of the gradient of  $f$  on the cell  $M$  limited such that:

$$g_{AB}^{in} \in [\min(f_M^n, f_{M_{AB}}^n), \max(f_M^n, f_{M_{AB}}^n)]$$

and similarly for the other sides.

The use of the Van-Leer method gives much better results, in particular in the case of shock profile, for which the Rayleigh line is now really close to a straight line.

A sufficient condition for stability for these two schemes is the conservation of positivity of  $f_M^{n+1}$ . For the first order scheme, one can verify that under the CFL condition

$$\begin{aligned} \max_{M \in \mathcal{M}} \frac{\Delta t}{|M|} (\max(v.n_{AB}, 0).l_{AB} + \max(v.n_{BC}, 0).l_{BC} + \\ \max(v.n_{CD}, 0).l_{CD} + \max(v.n_{DA}, 0).l_{DA}) \leq 1, \end{aligned}$$

the scheme is positive. In 1-D this reduces to

$$\max_{M \in \mathcal{M}} \frac{v \Delta t}{(\Delta x)_M} \leq 1.$$

Under this condition for the time step  $f_M^{n+1}$  is then a convex linear combination of the values  $f_M^n, f_{M_{AB}}^n, f_{M_{BC}}^n, f_{M_{CD}}^n, f_{M_{DA}}^n$ . Using the convexity of the function  $x \rightarrow x \log x$  and in the absence of boundary we have then

$$\sum_{M \in \mathcal{M}} |M| f_M^{n+1} \log(f_M^{n+1}) \leq \sum_{M \in \mathcal{M}} |M| f_M^n \log(f_M^n),$$

which gives a discrete analogue of property (12) for the transport phase.

For the second order accuracy in space we have a much stronger CFL condition. By noting that by construction we have

$$g_{AB}^{in} + g_{BC}^{in} + g_{CD}^{in} + g_{DA}^{in} = 4f_M^n$$

and all of the values  $g_{AB}^{in}, g_{AB}^{in} \dots$  are positives, then under

$$\max_{M \in \mathcal{M}} \frac{\Delta t}{|M|} \sup (\max(v.n_{AB}, 0).l_{AB}, \max(v.n_{BC}, 0).l_{BC}, \\ \max(v.n_{CD}, 0).l_{CD}, \max(v.n_{DA}, 0).l_{DA}) \leq \frac{1}{4}$$

the scheme is positive. In 1-D we recall that  $\frac{1}{4}$  can be replaced by  $\frac{1}{2}$ .

## 5.2 Full collision phase

In each space cell, the problem is to compute an approximation  $\bar{f}^{n+1}$  to the solution of the Cauchy problem

$$(30) \quad \begin{cases} \frac{d\bar{f}(t)}{dt} = Q(\bar{f}, \bar{f}), \\ \bar{f}(0) = \bar{f}^n \end{cases}$$

where  $Q(\bar{f}, \bar{f}) = (Q(\bar{f}, \bar{f})_1, \dots, Q(\bar{f}, \bar{f})_N)$ . The requirements of the time discretization scheme are that the solution at time  $\Delta t$  has the same first five moments that  $\bar{f}^n$  and that the Maxwellians are steady state of the problem. An immediate solution is to use an Euler explicit scheme:

$$\bar{f}^{n+1} = \bar{f}^n + \Delta t Q(\bar{f}^n, \bar{f}^n).$$

The main questionable point about the method is its computational cost. The cost of the evaluation of the discrete collision operator in the right-hand side of (29) is of order  $N^{2+(1/3)+\delta}$  for general differential cross section and if the differential cross section  $\sigma$  does not depend on  $\omega$ , this cost can be reduced to order  $N^2$  (see [2]). A computational complexity of order  $N^2$  is much too large for a practical use of the algorithm. We propose different ways to reduce this cost. All these methods involve random choices (i.e., are of Monte Carlo type). The amount of “randomness” varies from one to the other.

## 5.3 Collision phase: first acceleration procedure using randomized sublattice.

The first acceleration technique (method A) which can be used is the following modification of the full method.

Let  $b \in \mathbb{N}$ ,  $b \geq 1$  and let  $a \in \mathbb{N}^3$ , such that

$$\max_{s=1,3} a_s \leq b - 1$$

(where  $a_s$  denotes the  $s$ -th coordinate of  $a$ ,  $s = 1, 2, 3$ ). Then,  $L_{a,b} = a\Delta v + b.\Delta v.\mathbb{Z}^3$  be a sublattice of  $\Delta v.\mathbb{Z}^3$ . We use it for the quadrature formula (14) instead of the original lattice  $\Delta v.\mathbb{Z}^3 = L_{0,1}$ . More precisely, we write

$$\begin{aligned} & \left[ \int_{\mathbb{R}^3} \int_{S^2} q(v - v_*, \omega) (f' f'_* - f f_*) d\omega dv_* \right]_{v=v_i} \\ & \simeq \sum_{j/v_i + v_j \in L_{a,b}} (b\Delta v)^3 \int_{S^2} q(v_i - v_j, \omega) (f(v'_i) f(v'_j) - f(v_i) f(v_j)) d\omega. \end{aligned}$$

We notice that the volume of the elementary cell of the coarser lattice  $L_{a,b}$  is now  $(b\Delta v)^3$ . Then, the evaluation of the  $\omega$ -integral is done exactly in the same way as previously, by taking the pairs  $\{v_k, v_l\}$  of points of the finer lattice  $\Delta v \mathbb{Z}^3$ , which belong to  $\tilde{S}_{i,j}$ . The resulting simplified operator of collisions can be written

$$(31) \quad Q(f, f)(v_i) \simeq \frac{1}{(\Delta v)^3} Q_{i,a,b}(\bar{f}, \bar{f}) = \sum_{j/v_i + v_j \in L_{a,b}} \sum_{\{v_k, v_l\} \in \tilde{S}_{i,j}} b^3 \{A_{kl}^{ij} f_k f_l - A_{ij}^{kl} f_i f_j\}.$$

We can write

$$Q_{i,a,b}(\bar{f}, \bar{f}) = G_{i,a,b}(\bar{f}, \bar{f}) - p_{i,a,b}(\bar{f}) \cdot f_i$$

with

$$G_{i,a,b}(\bar{f}, \bar{f}) = \sum_{j/v_i + v_j \in L_{a,b}} \sum_{\{v_k, v_l\} \in \tilde{S}_{i,j}} b^3 \{A_{kl}^{ij} f_k f_l\} \geq 0$$

which is the gain term for  $v_i$  and

$$p_{i,a,b}(\bar{f}) = \sum_{j/v_i + v_j \in L_{a,b}} \left( \sum_{\{v_k, v_l\} \in \tilde{S}_{i,j}} A_{ij}^{kl} \right) f_j \geq 0$$

is the collision frequency for  $v_i$ .

By using the facts that  $\bigcup_{a \in A_b} L_{a,b} = \Delta v \mathbb{Z}^3$  and if  $a_1 \neq a_2$  then  $L_{a_1,b} \cap L_{a_2,b} = \emptyset$ , where  $A_b = \{0, 1, \dots, b-1\}^3$ , we remark that we have the relations

$$(32) \quad \frac{1}{b^3} \sum_{a \in A_b} p_{i,a,b}(\bar{f}) = p_i(\bar{f}) = p_{i,0,1}(\bar{f})$$

and

$$(33) \quad \frac{1}{b^3} \sum_{a \in A_b} G_{i,a,b}(\bar{f}, \bar{f}) = G_i(\bar{f}, \bar{f}) = G_{i,0,1}(\bar{f}, \bar{f}),$$

that implies the following decomposition of  $Q(\bar{f}, \bar{f})$ :

$$Q(\bar{f}, \bar{f}) = Q_{0,1}(\bar{f}, \bar{f}) = \frac{1}{b^3} \sum_{a / \max_{s=1}^3 a_s < b-1}^{b^3} Q_{a,b}(\bar{f}, \bar{f}).$$

We use the same time discretization as for the full method:

$$\bar{f}^{n+1} = \bar{f}^n + \Delta t Q_{a,b}(\bar{f}^n, \bar{f}^n).$$

Now, let us assume that  $b$  has been chosen  $\geq 2$ . In order to preserve the accuracy of the finer lattice, (even though we use at some step of the discretization a coarser lattice), *the triple  $a$  is chosen randomly at each time step, in the set  $A_b$ .*

We can hope that the accuracy of the finer grid is reached for steady problems, if we look at a mean value result after some time steps, since the expectation of this random choice of  $Q_{a,b}(\bar{f}, \bar{f})$  is indeed  $Q_{0,1}(\bar{f}, \bar{f})$  as the relations (32) and (33) show it. In this sense, we can say that this scheme preserves the accuracy of the finer mesh.

It is important that the symmetry properties (23) and (24) are still satisfied. It is clear for (23). For (24), it suffices to notice that if  $(v_k, v_l) \in \tilde{S}_{ij}$ , then  $v_k + v_l = v_i + v_j \in L_{a,b}$ . Therefore,

**Table 1 computational costs**

	b=1	b=3	b=6
$8^3$	1	0.0 4	0.0 05
$16^3$	64	2. 5	0.3
$32^3$	4096	150	19

properties (25) to (27) still hold for the discrete collision operator (31). However, it is no longer obvious that the only invariants of collision, (which are sequences  $\bar{\psi} = (\psi_1, \dots, \psi_N)$ , such that  $\psi_k + \psi_l - \psi_i - \psi_j = 0$  for  $(v_i, v_j) \in (\Delta v \mathbb{Z}^3)^2$  s. t.  $v_i + v_j \in L_{a,b}$  and  $(v_k, v_l) \in \tilde{S}_{i,j}$ ) are linear combinations of 1,  $v_i$ ,  $|v_i|^2$  or that the system of equations is again coupled. With the choice of  $a$  at each time step, it is also clear, on homogeneous problems, that the only steady states are the same Maxwellians as those obtained with the full collision operator.

The same estimates of the computational efficiency can be made for this new method.

The evaluation of the collision operator over one time step costs of the order of  $\frac{(N^{2+(1/3)+\delta})}{b^3}$  operations in the general case and  $N^2/b^3$  in the case of an  $\omega$ -independent scattering cross section. For comparison, let us consider  $8^3$ ,  $16^3$  and  $32^3$  points discretizations of the velocity space, for an  $\omega$ -independent scattering cross section, without any sublattice ( $b = 1$ ) and with  $b = 3$  and  $b = 6$ . Assume that computational cost is 1 for  $8^3$  and  $b = 1$ . Then, the costs which are obtained for these various situations are given in table 1. Any doubling of the number of discretization points in one direction multiplies the computational cost by a factor 4096. However, if the doubling is associated with sublattice (with a doubling of the sublattice while doubling the number of points), the computational cost increases milder. It has been verified numerically that important sublattice ( $b$  of the order of 6) does not affect the results in any noticeable way, at least as far as moments of the distribution function are concerned, which are the most physically interesting quantities.

It is important for the problem that  $\bar{f}^{n+1}$  remains positive. Since the scheme can be written

$$f_i^{n+1} = f_i^n (1 - \Delta t p_{i,a,b}(\bar{f}^n)) + \Delta t G_{i,a,b}(\bar{f}^n, \bar{f}^n),$$

then we can see that under the condition

$$(34) \quad \sup_{i=1}^N p_{i,a,b}(\bar{f}^n) \Delta t \leq 1,$$

the scheme preserve the positivity of the solution.

#### 5.4 Collision phase: second acceleration procedure using a 4-velocity model and splitting of operator.

When the support of the distribution function is too small, that is, when almost all of the mass is concentrated on a small number of points, one can verify that the sublattice method leads to collision frequencies  $p_{i,a,b}(\bar{f}^n)$  which fluctuate too much. If we let  $\Delta t_{a,b}$  be the maximum time step allowed by (34) for the sublattice method with parameters  $b$  and  $a$ , and if we suppose that  $\bar{f}$  is like a  $\delta$  function we have

$$\Delta t_{a,b} \sim \frac{\Delta t_{0,1}}{b^3},$$

and in this case we gain nothing compared to the full method. On the other hand, when all the  $f_i$  are equal to a constant we have

$$\Delta t_{a,b} \sim \Delta t_{0,1},$$

then in this case we have the maximum efficiency for the sublattice method. Moreover, it is not clear that if  $\Delta t$  satisfies (34) we have the decay of entropy, that is

$$(35) \quad \sum_{i=1}^N f_i^{n+1} \log(f_i^{n+1}) \leq \sum_{i=1}^N f_i^n \log(f_i^n).$$

It would be nice to have an acceleration method which is unconditionally stable in time, produces less fluctuating collision frequencies and such that the decay of entropy holds. We propose two other acceleration techniques satisfying the first and third points. They are based on a splitting method of the operator in order to reduce the scheme to a 4-velocity model, indeed in this case we have an exact solution. We hope that the second one satisfies the second wish.

For the sake of simplicity, we explain these methods in the case of an isotropic cross section i.e.,  $\sigma$  does not depend on  $\omega$ , as it is the case for the VHS model used in aerodynamics.

To begin with, we must do some algebraic manipulations on the collision operator. We call  $S_V$  the collection of the spheres  $\tilde{S}_{i,j}$ . We have the partition of the set of pairs  $\{v_i, v_j\}$  where  $v_i, v_j \in \mathcal{V}_N$ :

$$\left\{ \{v_i, v_j\}, v_i \in \mathcal{V}_N, v_j \in \mathcal{V}_N \right\} = \bigcup_{S \in S_V} S.$$

We can write  $Q_{a,b}(\bar{f}, \bar{f}) = (Q_{1,a,b}(\bar{f}, \bar{f}), \dots, Q_{N,a,b}(\bar{f}, \bar{f}))$  defined by (31) as

$$Q_{a,b}(\bar{f}, \bar{f}) = \sum_{i=1}^N (Q_{i,a,b}) e_i = \sum_{S \in S_V} a_S Q_S(\bar{f}, \bar{f})$$

with  $a_S = b^3$  if the center of  $S$  is in  $L_{a,b}$  or  $a_S = 0$  if it is not the case, the operator  $Q_S(\bar{f}, \bar{f})$  is defined by

$$Q_S(\bar{f}, \bar{f}) = \sum_{\{v_i, v_j\} \in S} \frac{C_S}{\text{Card}(S)} \left( \sum_{\{v_k, v_l\} \in S} (f_k f_l - f_i f_j) \right) (e_i + e_j),$$

$(e_1, \dots, e_N)$  is the canonical base of  $\mathbb{R}^N$  and the constant  $C_S$  is defined by (see formula (22))

$$C_S = 4\pi q(\text{diam}(S)).$$

If for two pairs  $\{v_i, v_j\}$  and  $\{v_l, v_k\}$  in  $S$  we define the operator  $E_{i,j,k,l}(\bar{f}, \bar{f})$  by

$$E_{i,j,k,l}(\bar{f}, \bar{f}) = (f_k f_l - f_i f_j)(e_i + e_j - e_k - e_l),$$

using the symmetries properties

$$E_{i,j,k,l}(\bar{f}, \bar{f}) = E_{j,i,k,l}(\bar{f}, \bar{f}) = E_{i,j,l,k}(\bar{f}, \bar{f}) = E_{k,l,i,j}(\bar{f}, \bar{f}),$$

we can write

$$(36) \quad Q_S(\bar{f}, \bar{f}) = \frac{C_S}{\text{Card}(S)} \sum_{\{v_i, v_j\} \in S} \sum_{\{v_k, v_l\} \in S} \frac{1}{2} E_{i,j,k,l}(\bar{f}, \bar{f}),$$

or, if we let  $\mu = \{\mu_1, \dots, \mu_N\}$  so that  $\mu_i > 0$  for all  $i$ ,

$$Q_S(\bar{f}, \bar{f}) = \frac{C_S}{\text{Card}(S)} \sum_{\{v_i, v_j\} \in S} (\mu_i + \mu_j) \sum_{\{v_k, v_l\} \in S} \frac{1}{\mu_i + \mu_j + \mu_k + \mu_l} E_{i,j,k,l}(\bar{f}, \bar{f}).$$

With this last decomposition of  $Q_S(\bar{f}, \bar{f})$  for each sphere  $S$  of  $S_V$  we can also write  $Q(\bar{f}, \bar{f}) = Q_{0,1}(\bar{f}, \bar{f})$ , since we shall use this form later, as

$$(37) \quad Q(\bar{f}, \bar{f}) = \sum_{i=1}^N \sum_{j=1}^N \mu_j \sum_{\{v_k, v_l\} \in \tilde{S}_{ij}} \frac{C_{\tilde{S}_{ij}}}{\text{Card}(\tilde{S}_{ij})(\mu_i + \mu_j + \mu_k + \mu_l)} E_{i,j,k,l}(\bar{f}, \bar{f}).$$

Now we turn to the splitting method that we will use for time discretization. We suppose that we have defined an approximation  $\tilde{Q}(\bar{f}^n, \bar{f}^n)$  of  $Q(\bar{f}^n, \bar{f}^n)$  as a linear combination of terms  $E_{i,j,k,l}(\bar{f}^n, \bar{f}^n)$ , that is

$$\tilde{Q}(\bar{f}^n, \bar{f}^n) = \sum_{p=1}^P c_p E_{i_p, j_p, k_p, l_p}(\bar{f}^n, \bar{f}^n)$$

where  $P$  is some integer and the coefficients  $c_p$  are positive. With the definition of the operator  $E_{i,j,k,l}$ , it is clear that  $\tilde{Q}(\bar{f}, \bar{f})$  preserve the five first moments and the Maxwellians. In the Euler explicit scheme

$$\bar{f}^{n+1} = \bar{f}^n + \Delta t \tilde{Q}(\bar{f}^n, \bar{f}^n)$$

for solving problem (30),  $\bar{f}^{n+1}$  can also be viewed as an approximation for the solution at time  $\Delta t$  to the differential equation

$$\frac{d\bar{f}(t)}{dt} = \tilde{Q}(\bar{f}, \bar{f}),$$

with the initial condition

$$\bar{f}(0) = \bar{f}^n.$$

Another approximation of this solution can be obtained by the usual splitting technique for operator. Let  $\pi(p)$  be a permutation of the indices  $p$ , and let  $\bar{g}^p = (g_1^p, \dots, g_N^p)$  be the solution of the problem

$$(38) \quad \frac{d\bar{g}^p(t)}{dt} = c_{\pi(p)} E_{i_{\pi(p)}, j_{\pi(p)}, k_{\pi(p)}, l_{\pi(p)}}(\bar{g}^p, \bar{g}^p), \quad \bar{g}^p(0) = \bar{g}^{p-1}(\Delta t)$$

for  $p = 1, \dots, P$  and by defining  $\bar{g}^0$  as  $\bar{f}^n$ . Then  $\bar{g}^P(\Delta t)$  is an approximation of  $\bar{f}^{n+1}$ .

The interest of this is that we can exhibit the solution for each step of this splitting technique. Solving  $\frac{d\bar{g}}{dt} = c_{\pi(p)} E_{i_{\pi(p)}, j_{\pi(p)}, k_{\pi(p)}, l_{\pi(p)}}(\bar{g}, \bar{g})$  is equivalent to solving the homogeneous Boltzmann equation for a discrete velocity gas with only four velocities (four velocities Broadwell model). If we consider two pairs of velocities  $\{v_1, v_2\}$  and  $\{v_3, v_4\}$  which are two

diameters of a same sphere, the homogeneous Boltzmann equation for this discrete velocity gas is:

$$\begin{aligned}\frac{df_1(t)}{dt} &= C(f_3(t)f_4(t) - f_1(t)f_2(t)) \\ \frac{df_2(t)}{dt} &= C(f_3(t)f_4(t) - f_1(t)f_2(t)) \\ \frac{df_3(t)}{dt} &= -C(f_3(t)f_4(t) - f_1(t)f_2(t)) \\ \frac{df_4(t)}{dt} &= -C(f_3(t)f_4(t) - f_1(t)f_2(t)).\end{aligned}$$

The Cauchy problem for this Boltzmann equation with the initial data

$$f_1(0) = f_1^0, f_2(0) = f_2^0, f_3(0) = f_3^0, f_4(0) = f_4^0.$$

has for solution

$$f_1(t) = f_1^0 + A(t), f_2(t) = f_2^0 + A(t), f_3(t) = f_3^0 - A(t), f_4(t) = f_4^0 - A(t)$$

where

$$A(t) = \frac{(f_3^0 f_4^0 - f_1^0 f_2^0)}{\rho} (1 - e^{-\rho C t}) \quad \text{and} \quad \rho = \sum_{i=1}^4 f_i^0.$$

Since we have an exact solution, the H-theorem holds

$$(39) \quad \frac{d}{dt} \sum_{i=1}^4 f_i(t) \log(f_i(t)) \leq 0.$$

Since at each step of the splitting technique (38) we solve exactly the equation, the method is unconditionally stable in time and verifies (35), that is, we have the decay of entropy.

**Remark 2** As in computations the function  $B(t) = \frac{1}{\rho}(1 - e^{-\rho C t})$  could be too expensive to evaluate, it would be possible to replace it by an approximation  $h(t)$  which has the same behaviour:

- $h(0) = 0$  and  $0 < h(t) < \frac{1}{\rho}$
- $h'(0) = C$  and  $h'(t) \geq 0$
- $\lim_{t \rightarrow \infty} h(t) = \frac{1}{\rho}$

this leads to the approximation of the weights  $f_1, \dots, f_4$

$$f_i(t) = f_i^0 \pm (f_3^0 f_4^0 - f_1^0 f_2^0) h(t).$$

We have then

$$(40) \quad \frac{d}{dt} \sum_{i=1}^4 f_i(t) \log f_i(t) = \frac{h'(t)}{(1 - \rho h(t))} (f_3(t)f_4(t) - f_1(t)f_2(t)) \log\left(\frac{f_1(t)f_2(t)}{f_3(t)f_4(t)}\right)$$

and since the term  $\frac{h'(t)}{(1 - \rho h(t))}$  remains positive the decay of the entropy still holds for  $t \in [0, \infty[$ . In the case of an explicit time discretization for solving the Boltzmann equation for the

four velocities system, which correspond to the approximation of  $B(t)$  by  $Ct$ , we are sure that the decay of entropy hold only for  $t \in [0, \frac{1}{C\rho}]$ . Another problem with this time discretization is that the relaxation to the Maxwellians is too fast. Another natural approximate is

$$h(t) = \frac{(1 - \frac{1}{1+\rho Ct})}{\rho}$$

which corresponds to the approximation of  $e^{-x}$  by  $\frac{1}{1+x}$  but now the relaxation to the maxwellians is too slow. A solution could be to tabulate the function  $e^{-x}$  and to do linear interpolation between two consecutive values.

We now propose two methods using splitting for reduction to 4-velocities systems.

The first one (method B) is derived from the sublattice method. As we have seen before, in the sublattice method at each time step we take the following approximation of  $Q(\bar{f}^n, \bar{f}^n)$

$$Q_{a,b}(\bar{f}^n, \bar{f}^n) = \sum_{S \in S_V} a_S Q_S(\bar{f}^n, \bar{f}^n)$$

where  $Q_S(\bar{f}, \bar{f})$  is given by (36). We make then a new approximation for each  $Q_S(\bar{f}^n, \bar{f}^n)$ . Let  $(c_1, \dots, c_{\text{Card}(S)})$  be the elements of  $S$ , which are pairs of velocities. Given a random permutation  $\pi$  of the indices of  $c$  we form the pair  $(c_{\pi(2p)}, c_{\pi(2p+1)})$ , we take the following approximation of  $Q_S(\bar{f}^n, \bar{f}^n)$ :

$$\tilde{Q}_S(\bar{f}^n, \bar{f}^n) = \left( \frac{C_S \text{Card}(S)}{\text{Card}(S) + 1 - \epsilon} \right) \sum_{p=1}^{\frac{\text{Card}(S)}{2}} E_{i(c_{\pi(2p)}), j(c_{\pi(2p)}), k(c_{\pi(2p+1)}), l(c_{\pi(2p+1)})}(\bar{f}^n, \bar{f}^n)$$

where  $\epsilon = 0$  or  $1$ ,  $\text{Card}(S) \equiv \epsilon \pmod{2}$ . The expectation of this random approximation is indeed  $Q_S^n$ . Now we use the splitting method (38) in the following way: we pick the sphere  $S$  for which the center is in  $L_{a,b}$  in a randomized fashion and for each  $S$  we solve all of the equations

$$\frac{d\bar{f}(t)}{dt} = a_S \left( \frac{C_S \text{Card}(S)}{\text{Card}(S) + 1 - \epsilon} \right) E_{i(c_{\pi(2p)}), j(c_{\pi(2p)}), k(c_{\pi(2p+1)}), l(c_{\pi(2p+1)})}(\bar{f}, \bar{f})$$

which is equivalent in fact to solve exactly the system

$$\frac{d\bar{f}(t)}{dt} = a_S \tilde{Q}_S(\bar{f}, \bar{f}),$$

because we have cut the set of velocities of  $S$  in systems of four velocities which are not coupled.

The second method (method C), using the splitting technique (38), starts by using formula (37) for the discrete collision kernel  $Q(\bar{f}, \bar{f})$ . We use a Monte Carlo quadrature formula with technique of importance sampling for obtaining an approximation of  $Q(\bar{f}^n, \bar{f}^n)$ . We suppose  $\mu$  so that  $\sum_{i=1}^N \mu_i = 1$ . We choose  $M$  velocities  $v_{jm}$  according to the probability law  $\mu$  and we take

$$Q(\bar{f}^n, \bar{f}^n) \simeq \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^N \sum_{\{v_k, v_l\} \in \tilde{S}_{ijm}} \frac{C_{\tilde{S}_{ijm}} E_{i,jm,k,l}(\bar{f}^n, \bar{f}^n)}{\text{Card}(\tilde{S}_{ijm})(\mu_i + \mu_{jm} + \mu_k + \mu_l)}$$



and for each sum over the sphere defined by  $v_i$  and  $v_{j_m}$  we choose randomly a pair  $(k_{i,j_m}, l_{i,j_m})$ . This gives the new approximation

$$Q(\bar{f}^n, \bar{f}^n) \simeq \tilde{Q}(\bar{f}^n, \bar{f}^n) = \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^N \frac{C_{\tilde{S}_{ijm}} E_{i,j_m,k_{i,j_m},l_{i,j_m}}(\bar{f}^n, \bar{f}^n)}{\mu_i + \mu_{j_m} + \mu_{k_{i,j_m}} + \mu_{l_{i,j_m}}}.$$

In order to have the best approximation possible, the probability law  $\mu$  must be chosen to minimize the variance of the process. In practice we choose  $\mu$  close as possible to  $\frac{f_j^n}{\sum_{k=1}^N f_k^n}$ .

When we are far away from thermal equilibrium, we suggest to take  $\mu = \frac{\bar{f}^n}{\sum_{k=1}^N f_k^n}$ . Near equilibrium we suggest to take the Maxwellian which has the same five first moments that  $\frac{f_j^n}{\sum_{k=1}^N f_k^n}$ .

Now we apply the splitting method (38) with  $\tilde{Q}(\bar{f}, \bar{f})$  defining by

$$\tilde{Q}(\bar{f}, \bar{f}) = \frac{1}{M} \sum_{i=1}^N \sum_{m=1}^M \frac{C_{\tilde{S}_{ijm}}}{\mu_i + \mu_{j_m} + \mu_{k_{i,j_m}} + \mu_{l_{i,j_m}}} E_{i,j_m,k_{i,j_m},l_{i,j_m}}(\bar{f}, \bar{f}).$$

For the permutation  $\pi$  in the process (38) we take a random permutation.

**Remark 3** *In the absence of boundary it can be noticed that the methods described in this paragraph, since they verify (35) for all space cells, give, when they are combined with the first order finite volume method for the transport phase, a discrete analogue of (12).*

## 6 Numerical results

### 6.1 Shock wave

We compare the results obtained with our discrete velocity method (DVM) with those obtained with a direct simulation Monte Carlo (DSMC) code in the case of a shock wave for a hard sphere gas. This DSMC code uses the Bird method without time counter for the collision phase.

The Mach number of the shock wave is approximatively 6.2. For the comparisons the calculations are made in an unsteady fashion by using a classical procedure to produce a shock. At the beginning the flow is uniform with a velocity  $u_\infty = -8$  and a temperature such that  $RT_\infty = 1.85$  and at  $x=0$  we put a specular wall. The domain of the computation is  $[0, 25\lambda_\infty]$ , where  $\lambda_\infty$  is the mean free path at infinity. We look for the solution when the shock arrived at a distance  $15\lambda_\infty$  of the wall which correspond at  $t = 1.14$ .

For the DSMC computation the time step is  $\Delta t = .006 = 0.8\tau_\infty$ , where  $\tau_\infty$  is the mean free time at infinity and we take a uniform grid with  $\Delta x = 0.5\lambda_\infty$ . The number of particles at the initial time is 1200 in each cell of space. The DSMC computation takes 300 seconds on a Cray YMP.

The DVM computations are made with the same grid. The time step is  $\Delta t = 0.8\tau_\infty$  and for a variant without splitting between collisions and transport  $\Delta t = 0.533\tau_\infty$ .

For the DVM we take  $V = B(0, 12)$  and  $\Delta v = 2.4$  and  $\Delta v = 1.2$  which correspond to  $N = 515$  and  $N = 4169$  velocities. We use the sublattice acceleration (method A) and the values of the parameters  $b$  are:

$b = 3$  for  $N = 515$ ,  $b = 3$  and  $6$  for  $N = 4169$ .

On the Cray YMP the method with  $N = 515$  and  $b = 3$  takes 10 seconds. With 4169 velocities and  $b = 6$  the computation takes 86 seconds.

The comparisons are made on the following profiles:

- Profiles of the density and the temperature.

- Rayleigh line:  $p_{xx}$  as a function of  $\frac{1}{\rho}$ . One can see that through the shock we have the relation  $A\frac{1}{\rho} + p_{xx} = B$  where  $A$  and  $B$  are two constants (see [3]) and then the points  $(\frac{1}{\rho}, p_{xx})$  are on a straight line called Rayleigh line [5].

As one can see on figures (1) and (2), the results are good for both choices of the number of velocities. For the Rayleigh line the best results are obtained with the second order scheme for the transport phase. We note that with a smaller  $\Delta v$ ,  $\Delta v = .8$  which correspond to approximatively 14000 velocities we do not improve the result corresponding to  $N=4169$ .

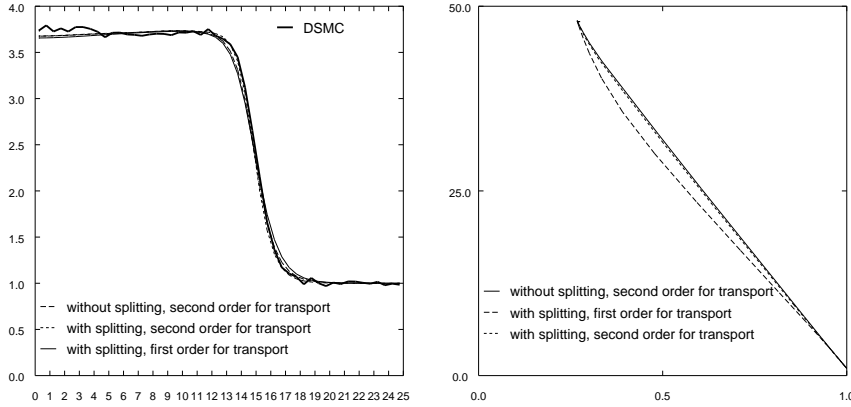


Figure 1: Density and Rayleigh line with  $N = 515$  and  $b = 3$ .

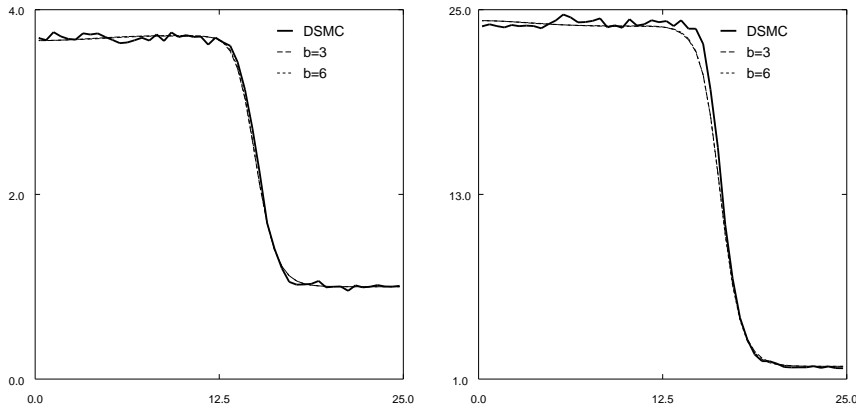


Figure 2: Variation of the parameter  $b$ . Density and  $RT$  with  $N = 4169$  and second order for the transport step.

## 6.2 Two-dimensionnal results: compression ramp

We compare the results obtained with the same DSMC method used for monodimensionnal problems, with those obtained with our DVM for a compression ramp placed in a supersonic flow. The geometry is a flat plate of 5 cm followed by a ramp of  $10^\circ$ . The characteristics for the flow at infinity are those for Mach 4 and Mach 20 as in [12]. Because we consider a monoatomic gas the Mach numbers are in fact 3.67 and 18.8.

### At mach 3,67

- $v_\infty = 669,3 \text{ m/s}$
- $n_\infty = 2,769 \cdot 10^{21} / \text{m}^3$
- $T_\infty = 69,76 \text{ K}$
- molecular mass:  $4,815 \cdot 10^{-26} \text{ kg}$
- temperature of wall:  $T_w = 336 \text{ K}$
- the mean free path at infinity:  $\lambda_\infty = 2,348 \cdot 10^{-4} \text{ m}$
- the mean free path at the wall:  $\lambda_\infty = 2,158 \cdot 10^{-4} \text{ m}$
- Knudsen number at infinity:  $Kn_\infty = 0,0047$

### At mach 18,8

- $v_\infty = 1503 \text{ m/s}$
- $n_\infty = 3,716 \cdot 10^{20} / \text{m}^3$
- $T_\infty = 13,32 \text{ K}$
- molecular mass:  $4,651 \cdot 10^{-26} \text{ kg}$
- temperature of wall:  $T_w = 290 \text{ K}$
- the mean free path at infinity:  $\lambda_\infty = 2,35 \cdot 10^{-3} \text{ m}$
- the mean free path at the wall:  $\lambda_\infty = 1,03 \cdot 10^{-3} \text{ m}$
- Knudsen number at infinity:  $Kn_\infty = 0,047$

The cross section in the two cases is of the VHS type. For the expression of the cross section, the mean free path and the values of parameters used in the VHS model see [12]. In the two cases we have a perfect accommodation at the wall. For DSMC and our scheme we used the same grid. For the flow at Mach 3,67 we used a nonuniform grid of 5250 quadrangulars elements. The size of the mesh in the direction perpendicular to  $v_\infty$  at the beginning of the flat plate and the corner are of the order of the mean free path near the wall. At Mach 18,8 we used a uniform grid with 3589 quadrangulars elements. The parameters for the DVM are the following:

-Mach 3.67: sublattice (method B) with  $N = 515$ ,  $b=3$ ,  $\Delta t = 5 \cdot 10^{-7} \text{ s}$ , and the numbers of iterations is 900.

-Mach 18.8: method C with  $N = 3405$ ,  $\Delta t = 3.89 \cdot 10^{-6} \text{ s}$ , and the numbers of iterations is

58. Since the flow is far away from thermal equilibrium, we take  $\mu = \frac{f_j^n}{\sum_{k=1}^N f_k^n}$ .

For the DVM, we initialized the computations with Maxwellians such that they give the exact density mean velocity and temperature after projection on the velocity grid. The DSMC runs take 120 minutes on a CRAY YMP. At Mach 3,67 the number of samples is 1500 with an average of 20 particles in space cell. At Mach 18,8 the number of sample is 700 again with an average of 20 particles in space cell. At Mach 3,67 the DVM takes 120 minutes CPU (approximatively 80 per cent of the time for the transport phase) with the same computer as for DSMC. For Mach 18,8 the discrete velocity method takes 60 minutes (approximatively 80 per cent of the time for the collision phase).

On pictures (3) to (10) the density, temperature, velocity in the y direction and Mach number are plotted for the two methods. The DSMC calculations, as we have an incorrect account of the boundary condition at the downstream boundary (particles can leave the domain but none can enter through this boundary) the results are bad on a small region forward the boundary and near the wall but they are not affected in the rest of the domain. For the two Mach numbers, the results of the DVM are in good accordance with the results of the DSMC method (for the four quantities shown, the isolines have the same level) and are much less noisy than DSMC results.

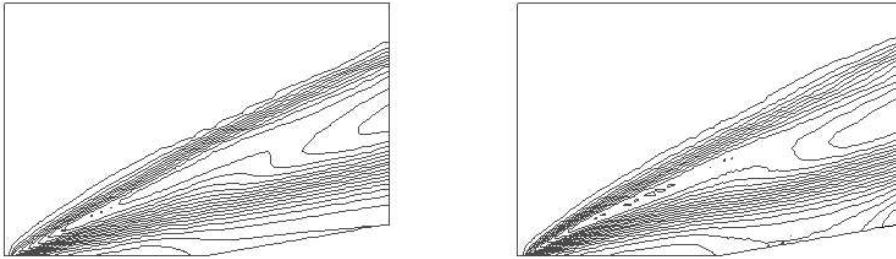


Figure 3: compression ramp at Mach 3.67, **density**, left DVM, right DSMC method.

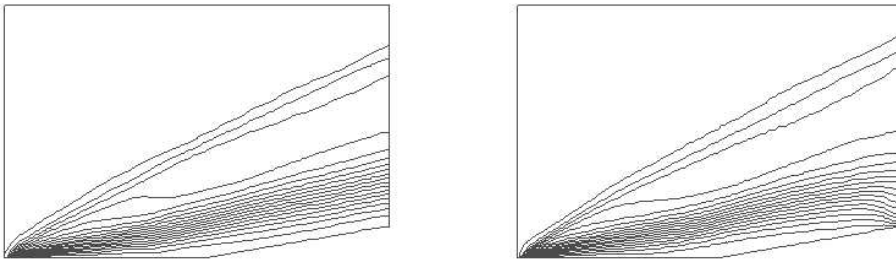


Figure 4: compression ramp at Mach 3.67, **temperature**, left DVM, right DSMC method.

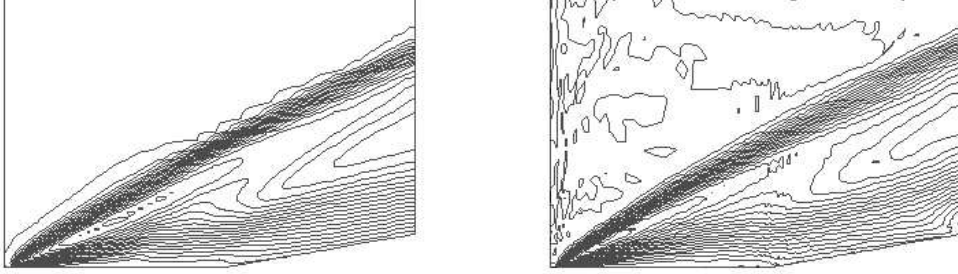


Figure 5: compression ramp at Mach 3.67,  $v_y$ , left DVM, right DSMC method.

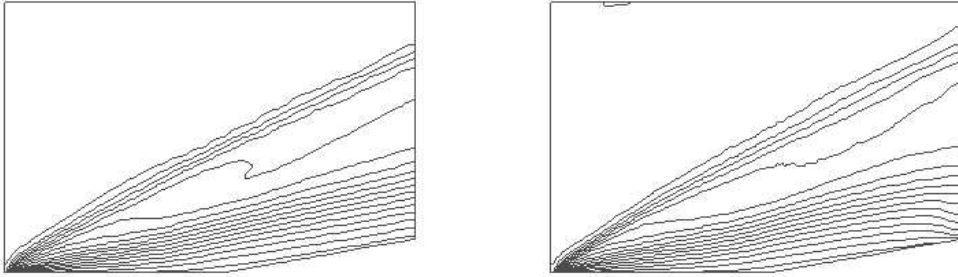


Figure 6: compression ramp at Mach 3.67, **Mach number**, left DVM, right DSMC method.

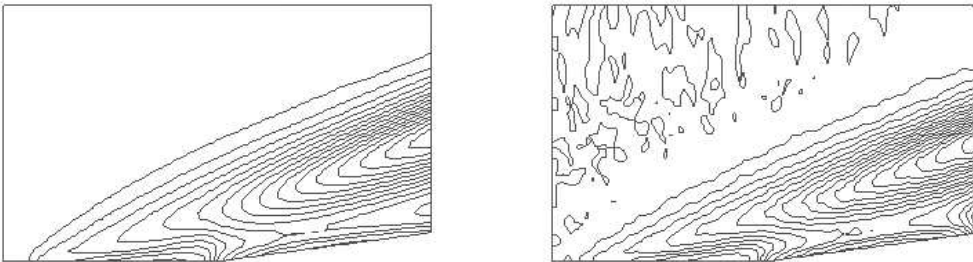


Figure 7: compression ramp at Mach 18.8, **density**, left DVM, right DSMC method.

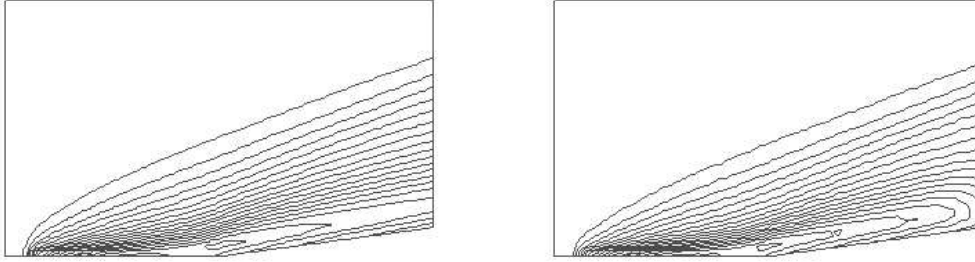


Figure 8: compression ramp at Mach 18.8, **temperature**, left DVM, right DSMC method.

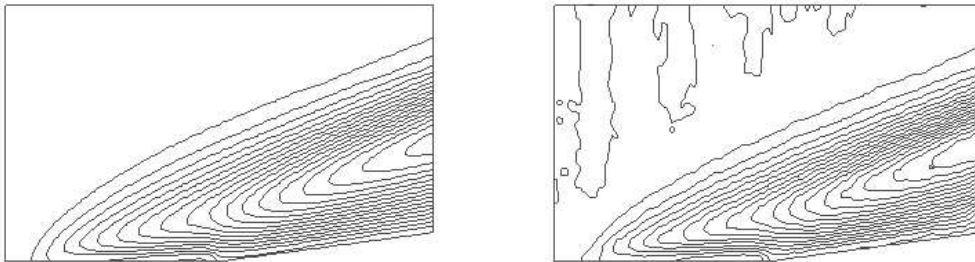


Figure 9: compression ramp at Mach 18.8,  **$v_y$** , left DVM, right DSMC method.

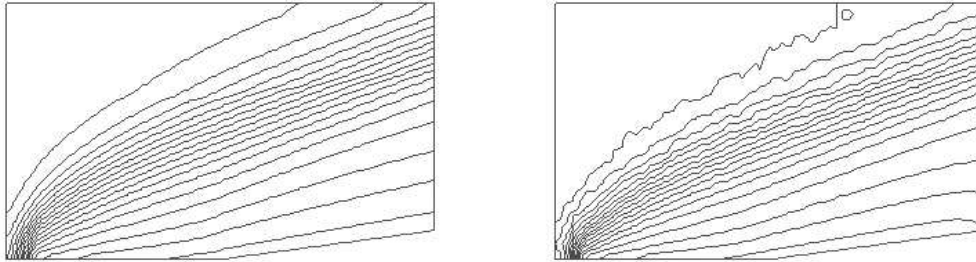


Figure 10: compression ramp at Mach 18.8, **Mach number**, left DVM, right DSMC method.

## 7 Conclusions

The Boltzmann equation for the discrete velocity model that we used seems to give good results in rarefied gas dynamics for monoatomic species as one can see with numerical results or in [11]. Acceleration procedures, like those we described in this paper, must be employed to give acceptable computational time. If employed, these acceleration techniques make the DVM an interesting alternative to the DSMC method in aerodynamics applications. Despite the fact that these acceleration procedures are of Monte Carlo type, the results remain good and seem to be almost free of noise. As one can see in [6] and by the use of our acceleration techniques, we think that we are able to extend this method to gas with internal degrees of freedom or to gas mixtures.

### Acknowledgments

I am indebted to Pr P. DEGOND for many helpful discussions.

## References

- [1] G. A. BIRD, “*Molecular Gas Dynamics*”, Clarendon Press, Oxford, (1976).
- [2] C. BUET, *Résolution déterministe de l’équation de Boltzmann*, note interne CEA, (1994).
- [3] C. CERCIGNANI, *The Boltzmann Equation and Its Applications*, Springer, New York, (1988).
- [4] R. GATIGNOL, *Théorie cinétique des gaz à répartitions discrètes de vitesses*, Springer, New York, (1975).
- [5] D. GOLDSTEIN, B. STURTEVANT and J. E. BROADWELL, *Investigations of the Motion of Discrete-Velocity Gases*, in “Rarefied Gas Dynamics: Theoretical and Computational Techniques”, E. P. Muntz, D. P. Weaver and D. H. Campbell (eds), Progress in Astronautics and Aeronautics, Vol.118, AIAA, Washington DC, (1989).
- [6] D. B. GOLDSTEIN, *Discrete-Velocity collision dynamics for polyatomic molecules*, Phys. Fluids A4 pp 1831-1839, (1992).
- [7] F. GROPENGIESSER, H. NEUNZERT, J. STRUCKMEIER *Computational methods for the Boltzmann equation*. Venice 1989: The state of Art in Appl. and Industrial math., eds. R. Spigler, Kluwer Acad. Publ., (1990).

- [8] G.H. HARDY and E.M. WRIGHT, *An introduction to the number theory*, Clarendon Press, Oxford, (1938).
- [9] R. ILLNER and W. WAGNER, *A random discrete velocity model and approximation of the Boltzmann equation*, Journal Stat. Phys. 70 (3/4) A2 pp 773-792, (1993).
- [10] R. ILLNER and W. WAGNER, *random discrete velocity models and approximation of the Boltzmann equation. Conservation of momentum and energy*, Transp. Th. Stat. Phys. 23 (1-3) A2 pp 27-38, (1994).
- [11] T. INAMURO and B. STURTEVANT, *Numerical Study of Discrete-Velocity Gases*, Phys. Fluids A2 pp 2196-2203, (1990).
- [12] J.C. LENGRAND, K.S. HEFFNER, A. CHPOUN, *RC 90-8 Etude 1 Rampe de compression en gaz rarefies, travaux dans le domaine de l'hypersonique GDR Hypersonique Rapport final de convention DRET(DGA) N°89.34.080.00.47075.01*, (1990).
- [13] K. NANBU, *Direct simulation schemes derived from the Boltzmann equation*, J. Phys, Japan 49 p. 2042, (1980).
- [14] F. ROGIER and J. SCHNEIDER, *A direct method for solving the Boltzmann Equation*, Transp. Th. Stat. Phys, (1994).
- [15] J. SCHNEIDER, *Une méthode déterministe pour la résolution de l'équation de Boltzmann*, Ph.D thesis, University Paris 6, (1993).
- [16] B. VAN LEER, *Towards the ultimate conservative difference scheme. V, A second order sequel of Godunov's method*, J. Comput. Phys., Vol 32, (1979).



# CONSERVATIVE AND ENTROPY SCHEMES FOR THE BOLTZMANN COLLISION OPERATOR OF POLYATOMIC GASES

C. BUET

*CEA-CELV 94195 Villeneuve Saint Georges CEDEX, France.*

We propose two discrete velocity models derived from the Boltzmann equation of Larsen-Borgnakke type for polyatomic gases. These two models are natural extensions of previously discussed discrete velocity models used for monoatomic gases. These two models have the same properties as the continuous one, which are conservation of mass, momentum and energy, discrete Maxwellians as equilibrium states and H-theorems.

## 1. Introduction

numerical methods for the Boltzmann equation of monoatomic gases, (see [5, 10, 20, 18]), are based on the kinetic theory of gases with a discrete velocity repartition [9]. These methods have been developed in the case of monoatomic gases. Only very few extensions to polyatomic gases have been made. Goldstein [11] gives the dynamics of collisions for a discrete polyatomic gas; Nanbu [17] proposes a discrete Boltzmann equation which is not related with the continuous Larsen-Borgnakke model. We present the natural extension of the discrete velocity model given in [5] to the polyatomic case. Our model is based on the Larsen-Borgnakke model [3] where the internal energy is assumed to take continuous values. Two discrete velocity models will be proposed, which both share the same properties as the continuous Larsen-Borgnakke model. The derivation of these two discrete models is a first step to obtain conservative and entropy decreasing numerical schemes for the Boltzmann equation of polyatomic gases. Space and time discretization and numerical results will be the subject of a forthcoming paper.

## 2. The Larsen-Borgnakke Model

A gas with  $\delta$  " internal degrees of freedom", associated with a polytropic constant  $\gamma = \frac{5+\delta}{3+\delta}$ , is described by a distribution function  $f(x, v, I, t)$ , where  $(x, v, I, t) \in \mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{R}^+$ ,  $x$  being the position variable of the molecule,  $v$  its velocity,  $I^2$  its internal energy and  $t$  the time. We refer to [6] for general considerations on distribution functions and the Boltzmann equation. The number density  $n(x, t)$  of parti-

cles, the momentum density  $n(x, t)U(x, t)$  and the total energy density  $n(x, t)E(x, t)$  are defined respectively by

$$\begin{pmatrix} n(x, t) \\ n(x, t)U(x, t) \\ n(x, t)E(x, t) \end{pmatrix} = \int_{\mathbb{R}^3 \times \mathbb{R}^+} \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} + I^2 \end{pmatrix} f(x, v, I, t) dv I^{\delta-1} dI. \quad (2.1)$$

Following [4], the collision operator for  $f(x, v, I, t)$  is defined by:

$$Q_\delta(f, f) = \int_{\Delta} B(f' f'_* - f f_*) dv_* I_*^{\delta-1} dI_* d\eta (r(1-r))^{\frac{\delta}{2}-1} dr R^2 (1-R^2)^{\delta-1} dR \quad (2.2)$$

with

$$\begin{aligned} g &= \frac{v - v_*}{2} = \text{relative velocity}, \\ E^2 &= |g|^2 + I^2 + I_*^2 = \text{total energy}, \\ (v_*, I_*, \eta, r, R) &\in \Delta = \mathbb{R}^3 \times \mathbb{R}^+ \times S^{2,+} \times [0, 1]^2, \\ B &:= B(E, |Rg|, |Rg_*|, I^2 r(1-R^2), I_*^2 (1-r)(1-R^2)) > 0, \\ f &= f(x, v, I, t), \quad f_* = f(x, v_*, I_*, t), \quad f' = f(x, v', I', t), \quad f'_* = f(x, v'_*, I'_*, t), \end{aligned}$$

and the collision process is defined by

$$\begin{cases} v + v_* = v' + v'_* \\ g' = \frac{RE}{|g|} \{g - 2(g \cdot \eta) \eta\} \\ I' = \sqrt{r(1-R^2)} E \\ I'_* = \sqrt{(1-r)(1-R^2)} E \end{cases} \quad (2.3)$$

$S^2$  is the unit sphere of  $\mathbb{R}^3$ . The corresponding Boltzmann equation reads:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q_\delta(f, f). \quad (2.4)$$

The properties (see [4]) of the Boltzmann collision operator (2.2) are conservation of mass, momentum and energy, and dissipation of entropy: let  $\varphi(v, I)$  be any smooth test function, we consider a weak formulation of the collision operator (2.2) which can be symmetrized as follows

$$\begin{aligned} & \int_{\mathbb{R}^3 \times \mathbb{R}^+} Q_\delta(f, f) \varphi dv I^{\delta-1} dI \\ &= -\frac{1}{4} \int_{\mathbb{R}^3 \times \mathbb{R}^+} Q_\delta(\varphi' + \varphi'_* - \varphi - \varphi_*) dv I^{\delta-1} dI. \end{aligned} \quad (2.5)$$

Then the conservation of mass, momentum and energy can be written with  $\varphi = 1, v, \frac{|v|^2}{2} + I^2$

$$\int_{\mathbb{R}^3 \times \mathbb{R}^+} Q_\delta(f, f) \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} + I^2 \end{pmatrix} dv I^{\delta-1} dI = 0 \quad (2.6)$$

and the dissipation of entropy reads

$$\int_{\mathbb{R}^3 \times \mathbb{R}^+} Q_\delta(f, f) \log(f) dv I^{\delta-1} dI \leq 0. \quad (2.7)$$

Any equilibrium distribution function,  $f$  satisfying  $Q_\delta(f, f) = 0$ , is a Maxwellian

$$f(v) = C_\delta \frac{\rho}{(RT)^{(3+\delta)/2}} \exp\left(-\frac{|v-u|^2 + 2I^2}{2RT}\right), \quad (2.8)$$

where  $\rho, T \in \mathbb{R}$ ,  $\rho > 0, T > 0$ , and  $u \in \mathbb{R}^3$ .  $(\rho, u, T)$  are the density, mean velocity and temperature of the gas and  $C_\delta$  is a constant of normalization.

In the homogeneous case, i.e. when the distribution function is independent of the space variable  $x$ , the  $H$ -theorem follows from (2.7):

$$\begin{aligned} & \frac{d}{dt} \int_{\mathbb{R}^3 \times \mathbb{R}^+} f(v, I, t) \log f(v, I, t) dv I^{\delta-1} dI \\ &= \int_{\mathbb{R}^3 \times \mathbb{R}^+} Q_\delta(f, f) (1 + \log f) dv I^{\delta-1} dI \leq 0. \end{aligned} \quad (2.9)$$

Furthermore  $H(f)$  can only be minimal if  $f$  is an equilibrium distribution (i. e. if  $f$  is a Maxwellian).

For the classical variable hard sphere (VHS) model  $B$  is simply given by:

$$B = CR^{1-2\alpha} |g|^{-2\alpha} |g \cdot \eta| \quad (2.10)$$

with  $\alpha \in [0, \frac{1}{2}]$ .

At the first order in the collision dominated limit (i.e. when the Knudsen number tends to zero), the fluid limit of the Boltzmann equation (2.4) yields the Euler equations for a polytropic gas with  $\gamma \in [1, \frac{5}{3}]$ .

Another formulation of the Boltzmann equation for polyatomic gases is (see [7])

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = I^{\alpha-\delta} Q_\alpha(f, f), \quad (2.11)$$

where  $\alpha \geq 1$ . and the moments are still defined by (2.1). In [7]  $\alpha$  is taken equal to 2. This is a manner to simplify the numerical treatment of the collision operator for different values of  $\delta$ .

### 3. Another Forms of the Collision Operator $Q_\delta$

For our purpose the form of the collision operator (2.2) is not well adapted. We change the formulation of the collision operator to make the derivation of discrete velocity and internal energy models easier. In (2.4) we make the change of variables

$$\eta \rightarrow \omega = \frac{g}{|g|} - 2\left(\frac{g}{|g|} \cdot \eta\right) \eta$$

with  $\omega \in S^2$  for which the jacobian is  $(\frac{|g|}{4|\eta \cdot g|})^{-1}$ . The collision process for  $g$  is now defined by

$$g' = RE\omega$$

We also make the following change of variables  $(a, b) \rightarrow (R, r)$  defined by

$$R = \cos a, \quad r = \cos^2 b$$

for  $(a, b) \in [0, \pi/2]^2$ . The whole collision process is now defined by

$$\begin{cases} v' + v'_* = v + v_* \\ g' = (\cos a)E\omega \\ I' = (\sin a \cos b)E \\ I'_* = (\sin a \sin b)E \end{cases} \quad (3.12)$$

and the collision operator can be written

$$Q_\delta(f, f) = \int_{\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times [0, \pi/2]^2} B(f' f'_* - f f_*) d\sigma \quad (3.13)$$

and

$$d\sigma = (\cos b \sin a)^{\delta-1} (I_* \sin b \sin a)^{\delta-1} dv_* dI_* d\omega \sin a \cos^2 a da db$$

$$B := B(E, |g|, |g'|, |g \cdot g'|, II', I_* I'_*) > 0$$

For example, in the case of the VHS model  $B$  is of the form

$$B = C |Rg|^{1-2\alpha}.$$

We introduce the following notations:

$$\Omega = \begin{pmatrix} \omega \cos a \\ \cos b \sin a \\ \sin b \sin a \end{pmatrix}, \quad V = \begin{pmatrix} g \\ I \\ I_* \end{pmatrix},$$

where  $(\omega, a, b) \in S^2 \times [0, \pi/2]^2$ . The superficial measure on the quarter  $S_+^4$  of the sphere  $S^4$  defined by  $S_+^4 = \{\Omega = (\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5) \in \mathbb{R}^5 / |\Omega_1|^2 + |\Omega_2|^2 + |\Omega_3|^2 + |\Omega_4|^2 + |\Omega_5|^2 = 1, \Omega_4 > 0, \Omega_5 > 0\}$  is  $d\Omega = d\omega db \sin a \cos^2 a da$ . Then we have

$$Q_\delta(f, f) = \int_{(v_*, I_*) \in \mathbb{R}^3 \times \mathbb{R}_+, \Omega \in S_+^4} \mathcal{B} \cdot (f' f'_* - f f_*) dv_* dI_* d\Omega \quad (3.14)$$

with the collision process defined by

$$V' = |V| \cdot \Omega, \quad \Omega' = \frac{V}{|V|} \quad (3.15)$$

and with:

$$f = f(x, v, I, t), \quad f_* = f(x, v_*, I_*, t)$$

$$f' = f(x, \frac{v+v_*}{2} + g', I', t), \quad f'_* = f(x, \frac{v+v_*}{2} - g', I'_*, t)$$

$$\mathcal{B} = \Omega_4^{\delta-1} I_*^{\delta-1} \Omega_5^{\delta-1} B.$$

The proof of the H-theorem with this form of the collision operator (3.14) can be easily obtained using the property (2.5). We can remark that

$$dI_* dv_* dI dv d\Omega = d(\frac{v+v_*}{2}) dV d\Omega$$

let us define the transformations

$$\begin{pmatrix} g \\ I \\ I_* \\ \Omega_1 \\ \Omega_2 \\ \Omega_3 \\ \Omega_4 \\ \Omega_5 \end{pmatrix} \rightarrow \begin{pmatrix} V' \\ \Omega' \end{pmatrix} = \begin{pmatrix} -g \\ I_* \\ I \\ \Omega_1 \\ \Omega_2 \\ \Omega_3 \\ \Omega_5 \\ \Omega_4 \end{pmatrix}$$

and

$$\begin{pmatrix} V \\ \Omega \end{pmatrix} \rightarrow \begin{pmatrix} V' \\ \Omega' \end{pmatrix} = \begin{pmatrix} |V|.\Omega \\ \frac{V}{|V|} \end{pmatrix}.$$

These transformations are both involutives and their jacobian is equal to unity. Therefore they preserve the measure  $dV d\Omega$ , i.e.  $dV d\Omega = dV' d\Omega'$ . Using the invariance of

$$B.I^{\delta-1}\Omega_4^{\delta-1}I_*^{\delta-1}\Omega_5^{\delta-1},$$

by these transformations we recover property (2.5) for the collision operator (3.14) expressed in the new variables, by exchanging  $(v, I)$  and  $(v_*, I_*)$  or  $(v, I, v_*, I_*)$  and  $(v', I', v'_*, I'_*)$  respectively. Starting from (2.2) and (2.3), we can also rewrite the collision operator in a more classical form by using the change of variables

$$\begin{cases} e = I^2 \\ \omega = \frac{g}{|g|} - 2(\frac{g}{|g|}.\eta)\eta. \end{cases}$$

Now  $E$  is given by  $|g|^2 + e + e_*$ , the collision process is defined by

$$\begin{cases} v + v_* = v' + v'_* \\ g' = R\sqrt{E}\omega \\ e' = r(1 - R^2)E \\ e'_* = (1 - r)(1 - R^2)E \end{cases} \quad (3.16)$$

and the collision operator can be written

$$Q_\delta(f, f) = \int_{\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times [0,1]^2} B(f' f'_* - f f_*) d\sigma \quad (3.17)$$

with

$$d\sigma = dv_* e_*^{\frac{\delta}{2}-1} de_* d\omega R^2 (1-R^2)^{\delta-1} dR [r(1-r)]^{\frac{\delta}{2}-1} dr,$$

$$B := B(E, |g| \cdot |g'|, |g \cdot g'|, ee', e_* e'_*) > 0.$$

#### 4. Discrete Velocity and Energy Models

We start from (3.14) and (3.17) to derive discrete Boltzmann equations. We obtain two models which differ by the discretization of the internal energy  $e$ . In the first one we discretize uniformly  $\sqrt{e}$  while in the second model the discretization is uniform in  $e$ . With these two models, it is straightforward to derive a discrete version of the equation (2.11).

##### 4.1. A first discrete velocity and energy model

We consider the space homogeneous problem

$$\begin{cases} \frac{df}{dt} = Q_\delta(f, f) \\ f|_{t=0} = f_0(v, I) \end{cases} \quad (4.18)$$

where  $Q_\delta(f, f)$  is given by (3.14).

###### 4.1.1. Discretization

We take a regular discretization of  $\mathbb{R}^3 \times \mathbb{R}_+$ :

let us introduce  $\Delta v > 0$ ,  $\{e_1, e_2, e_3, e_4\}$  the canonical base of  $\mathbb{Z}^4$ ,  $z_i = (v_i, I_i) = (i + \frac{1}{2}e_4)\Delta v$ ,  $i = (i_1, i_2, i_3, i_4) \in L = \mathbb{Z}^3 \times \mathbb{N}$ , and an approximation  $f_i(t)$  of  $(\Delta v)^4 f(z_i, t)$ . We introduce a particle approximation of the problem (4.18). Given any test function  $\varphi(v, I)$  we have

$$\begin{aligned} & \int_{\mathbb{R}^3 \times \mathbb{R}_+} \frac{df}{dt} \varphi(v, I) I^{\delta-1} dv dI \\ &= \frac{1}{4} \int_{\mathbb{R}^3 \times \mathbb{R}_+} Q_\delta(\varphi(v, I) + \varphi(v_*, I_*) - \varphi(v'_*, I'_*) - \varphi(v', I')) I^{\delta-1} dv dI \end{aligned} \quad (4.19)$$

and

$$\begin{aligned} & \frac{1}{4} \int_{\mathbb{R}^3 \times \mathbb{R}_+} Q_\delta(\varphi + \varphi_* - \varphi' - \varphi'_*) I^{\delta-1} dv dI \\ &= \int_{\mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}^3 \times \mathbb{R}_+ \times S_+^4} \mathcal{C}(\varphi + \varphi_* - \varphi' - \varphi'_*)(f' f'_* - f f_*) d\Omega dv_* dI_* dv dI \end{aligned} \quad (4.20)$$

where we have set for simplicity

$$\begin{aligned} \varphi &= \varphi(v, I), \quad \varphi_* = \varphi(v_*, I_*), \\ \varphi' &= \varphi\left(\frac{v+v_*}{2} + g', I'\right), \quad \varphi'_* = \varphi\left(\frac{v+v_*}{2} - g', I'_*\right), \end{aligned}$$

$$C = \frac{1}{4} \mathcal{B} I^{\delta-1} = \frac{1}{4} I^{\delta-1} \Omega_4^{\delta-1} I_*^{\delta-1} \Omega_5^{\delta-1} B := \mathcal{C}(e, |g| \cdot |g'|, |g \cdot g'|, II', I_* I'_*).$$

We obtain an approximation of the two terms of the equality (4.19) by using quadrature formulae for the integrals with respect to  $(v_*, I_*)$ ,  $(v, I)$ , the quadrature points of which are the lattice points of  $\Delta v L$ . First, we get an approximation of the left hand side, of the form

$$\int_{\mathbb{R}^3 \times \mathbb{R}_+} \frac{df}{dt} \varphi(v, I) I^{\delta-1} dv dI \simeq (\Delta v)^4 \sum_{i \in L} \frac{df_i}{dt} \varphi(z_i) I_i^{\delta-1} \quad (4.21)$$

For the right hand side we make some transformations. We set

$$U = \frac{v + v_*}{2}$$

we have

$$dv dv_* dI dI_* = 8 dU dV$$

and, therefore,

$$\begin{aligned} & \int_{\mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}^3 \times \mathbb{R}_+ \times S_+^4} \mathcal{C}(\varphi + \varphi_* - \varphi' - \varphi'_*)(f' f'_* - f f_*) dv_* dI_* dv dI d\Omega \\ &= 8 \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}_+^2 \times S_+^4} \mathcal{C}(\varphi + \varphi_* - \varphi' - \varphi'_*)(f' f'_* - f f_*) dU dV d\Omega. \end{aligned} \quad (4.22)$$

We discretize first in the variable  $U$ . Since  $z_i \in \Delta v L$  then for  $U$  the quadrature points  $U_m$  are the elements of  $\frac{\Delta v}{2} \mathbb{Z}^3$ . We obtain a quadrature formula for the collision operator of the form

$$\begin{aligned} & \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}_+^2 \times S_+^4} \mathcal{C}(\varphi + \varphi_* - \varphi' - \varphi'_*)(f' f'_* - f f_*) d\Omega dU dV \\ & \simeq \Delta v^3 \sum_{m \in \mathbb{Z}^3} \int_{\mathbb{R}^3 \times \mathbb{R}_+^2 \times S_+^4} \mathcal{C}(\varphi(U_m + g, I) + \varphi(U_m - g, I_*) \\ & \quad - \varphi(U_m + g', I') - \varphi(U_m - g', I'_*)) \\ & \quad (f(U_m + g', I') f(U_m - g', I'_*) - f(U_m + g, I) f(U_m - g, I_*)) dV d\Omega. \end{aligned} \quad (4.23)$$

We now discretize in the variable  $V$ . For a fixed  $U_m$  we can write  $U_m = \Delta v(\frac{\varepsilon}{2} + p)$  where  $p \in \mathbb{Z}^3$  and  $\varepsilon = (\varepsilon_1, \varepsilon_1, \varepsilon_1) \in \{0, 1\}^3$ . Since the points  $(v, I)$  lie in  $L$ , the quadrature points for  $V$  are then necessarily in  $\Delta v((\frac{\varepsilon}{2} + \mathbb{Z}^3) \times (\mathbb{N} + \frac{1}{2})^2)$ . Using the same type of quadrature formula as for  $U$ , we have the approximation

$$\begin{aligned} & \sum_{m \in \mathbb{Z}^3} \int_{\mathbb{R}^3 \times \mathbb{R}_+^2 \times S_+^4} \mathcal{C}(\varphi(U_m + g, I) + \varphi(U_m - g, I_*) - \varphi(U_m + g', I') \\ & \quad - \varphi(U_m - g', I'_*)) (f(U_m + g', I') f(U_m - g', I'_*) - f(U_m + g, I) f(U_m - g, I_*)) \end{aligned}$$

$$\begin{aligned}
& f(U_m - g, I_*) dV d\Omega \simeq \Delta v^8 \sum_{i,j \in L \times L} \int_{S_+^4} \mathcal{C}(V = V_{ij}) \\
& (\varphi(z_i) + \varphi(z_j) - \varphi(U_{ij} + g'_{ij}, I'_{ij}) - \varphi(U_{ij} - g'_{ij}, I'_{*ij})) \\
& (f(U_{ij} + g'_{ij}, I'_{ij}) f(U_{ij} - g'_{ij}, I'_{*ij}) - f(z_i) f(z_j)) d\Omega, \quad (4.24)
\end{aligned}$$

where we have set

$$\begin{aligned}
V_{ij} &= \begin{pmatrix} \frac{z_{i_1} - z_{j_1}}{2} \\ \frac{z_{i_2} - z_{j_2}}{2} \\ \frac{z_{i_3} - z_{j_3}}{2} \\ z_{i_4} \\ z_{j_4} \end{pmatrix} = \begin{pmatrix} g_{ij} \\ I_i \\ I_j \end{pmatrix}, \\
U_{ij} &= \begin{pmatrix} \frac{z_{i_1} + z_{j_1}}{2} \\ \frac{z_{i_2} + z_{j_2}}{2} \\ \frac{z_{i_3} + z_{j_3}}{2} \end{pmatrix},
\end{aligned}$$

and  $V'_{ij} = (g'_{ij}, I'_{ij}, I'_{*ij})$  is the image of  $V_{ij}$  by the transformation (3.15). Formula (3.15) shows that, when  $\Omega$  varies in  $S_+^4$ ,  $V'_{ij}$  varies on the sphere of radius  $|V_{ij}|$  and centered at the point  $U_{ij}$ . As we have seen we can write  $U_m = \Delta v(\frac{\varepsilon_{ij}}{2} + p_{ij})$ . So we define

$$\begin{aligned}
S_{ij} &= \{(k, l) \in L^2 \text{ and such that } U_{ij} = U_{kl} \text{ and } |V_{ij}| = |V_{kl}|\} \\
&= \{(z_k, z_l) = ((U_{ij} + g_{kl}, (V_{kl})_4), (U_{ij} - g_{kl}, (V_{kl})_5)) \text{ with} \\
&V_{kl} \in \Delta v(\frac{\varepsilon_{ij}}{2} + \mathbb{Z}^3) \times (\mathbb{N} + \frac{1}{2})^2 \text{ such that } |V_{ij}| = |V_{kl}|\}. \quad (4.25)
\end{aligned}$$

For  $(k, l) \in S_{ij}$ , we can define a unique  $\Omega_{ij}^{kl} \in S_+^4$ , such that the first of formula (3.15) holds for  $(v, I) = z_i, (v_*, I_*) = z_j, (v', I') = z_k, (v'_*, I'_*) = z_l$ . The sets  $S_{ij}$  are not empty. For the integral with respect to  $\Omega$  which appears at the right hand side of (4.24), we use a quadrature formula using the  $\Omega_{ij}^{kl}$  as quadrature points. We make the assumption that the points  $\Omega_{ij}^{kl}$  are well distributed over  $S_+^4$ . By definition of  $\mathcal{C}$  we have

$$\begin{aligned}
& \int_{S_+^4} \mathcal{C}(V = V_{ij}) \cdot (\varphi(z_i) + \varphi(z_j) - \varphi(U_{ij} + g'_{ij}, I'_{ij}) - \varphi(U_{ij} - g'_{ij}, I'_{*ij})) \\
& (f(U_{ij} + g'_{ij}, I'_{ij}) f(U_{ij} - g'_{ij}, I'_{*ij}) - f(z_i) f(z_j)) d\Omega \\
&= \frac{I_i^{\delta-1} I_j^{\delta-1}}{4} \int_{S_+^4} B(V = V_{ij}) \cdot (\varphi(z_i) + \varphi(z_j) - \varphi(U_{ij} + g'_{ij}, I'_{ij}) - \varphi(U_{ij} - g'_{ij}, I'_{*ij})) \\
& (f(U_{ij} + g'_{ij}, I'_{ij}) f(U_{ij} - g'_{ij}, I'_{*ij}) - f(z_i) f(z_j)) \Omega_4^{\delta-1} \Omega_5^{\delta-1} d\Omega.
\end{aligned}$$



We can remark that

$$\int_{S_+^4} \Omega_4^{\delta-1} \Omega_5^{\delta-1} d\Omega = C_\delta \quad (4.26)$$

where  $C_\delta$  is a constant just depending of  $\delta$ . The use of the above mentioned quadrature formule yields

$$\begin{aligned} & \int_{S_+^4} B(V = V_{ij})(\varphi(z_i) + \varphi(z_j) - \varphi(U_{ij} + g'_{ij}, I'_{ij}) - \varphi(U_{ij} - g'_{ij}, I'_{*ij})) \\ & \quad (f(U_{ij} + g'_{ij}, I'_{ij})f(U_{ij} - g'_{ij}, I'_{*ij}) - f(z_i)f(z_j))\Omega_4^{\delta-1}\Omega_5^{\delta-1}d\Omega \\ & \simeq \sum_{(k,l) \in S_{ij}} \frac{\mathcal{M}(S_+^4)}{\text{Card}(S_{ij})} B(V = V_{ij}, V' = V_{kl}) \left(\frac{I_k I_l}{|V_{ij}|}\right)^{\delta-1} \\ & \quad (\varphi(z_i) + \varphi(z_j) - \varphi(z_k) - \varphi(z_l))(f(z_k)f(z_l) - f(z_i)f(z_j)) \end{aligned} \quad (4.27)$$

and

$$\int_{S_+^4} \Omega_4^{\delta-1} \Omega_5^{\delta-1} d\Omega \simeq \sum_{(k,l) \in S_{ij}} \frac{\mathcal{M}(S_+^4)}{\text{Card}(S_{ij})} \left(\frac{I_k I_l}{|V_{ij}|}\right)^{\delta-1}. \quad (4.28)$$

It is thus legitimate to identify the right hand side of equation (4.28) with  $C_\delta$  according to (4.26), and to insert the resulting identity into (4.27). This yields:

$$\begin{aligned} & \int_{S_+^4} \mathcal{C}(V = V_{ij}) \cdot (\varphi(z_i) + \varphi(z_j) - \varphi(U_{ij} + g'_{ij}, I'_{ij}) - \varphi(U_{ij} - g'_{ij}, I'_{*ij})) \\ & \quad (f(U_{ij} + g'_{ij}, I'_{ij})f(U_{ij} - g'_{ij}, I'_{*ij}) - f(z_i)f(z_j))d\Omega \\ & \simeq \frac{I_i^{\delta-1} I_j^{\delta-1}}{4} \sum_{(k,l) \in S_{ij}} p_{kl} B_{ij}^{kl} (\varphi(z_i) + \varphi(z_j) \\ & \quad - \varphi(z_k) - \varphi(z_l))(f(z_k)f(z_l) - f(z_i)f(z_j)) \end{aligned} \quad (4.29)$$

with

$$B_{ij}^{kl} = C_\delta B(V = V_{ij}, V' = V_{kl})$$

and

$$p_{kl} = \frac{I_k^{\delta-1} I_l^{\delta-1}}{\sum_{(k,l) \in S_{ij}} I_k^{\delta-1} I_l^{\delta-1}}.$$

For the VHS model in which

$$B = C \left( \frac{|g||g'|}{|V|} \right)^{1-2\alpha} = \left( |g| \sqrt{\Omega_1^2 + \Omega_2^2 + \Omega_3^2} \right)^{1-2\alpha}$$

we can make another approximation for the integral with respect to  $\Omega$  which appears in the left hand side of (4.29) which gives a better approximation of the collision frequencies. By noticing that in this case

$$\int_{S_+^4} \left( \frac{|g'|}{|V|} \right)^{1-2\alpha} \Omega_4^{\delta-1} \Omega_5^{\delta-1} d\Omega = D_{\delta,\alpha} \quad (4.30)$$

where  $D_{\delta,\alpha}$  is a constant just depending on  $\alpha$  and  $\delta$ , we can also make the following approximations of (4.30)

$$D_{\delta,\alpha} \simeq \sum_{(k,l) \in S_{ij}} \frac{\mathcal{M}(S_+^4)}{\text{Card}(S_{ij})} \left( \frac{|g_{kl}|}{|V_{ij}|} \right)^{1-2\alpha} \left( \frac{I_k I_l}{|V_{ij}|} \right)^{\delta-1}. \quad (4.31)$$

We obtain the same type of approximation (4.29) with

$$B_{ij}^{kl} = D_{\delta,\alpha} |g_{ij}|^{1-2\alpha}$$

and

$$p_{kl} = \frac{|g_{kl}|^{1-2\alpha} I_k^{\delta-1} I_l^{\delta-1}}{\sum_{(k,l) \in S_{ij}} |g_{kl}|^{1-2\alpha} I_k^{\delta-1} I_l^{\delta-1}}$$

This approximation for the VHS model forces the discrete model to give us the good collision frequency between  $z_i$  and  $z_j$ .

No error estimates for the quadrature formula (4.29) is available up to now. One problem is that the number of quadrature points  $\Omega_{ij}^{kl}$  on  $S_+^4$  is a non monotone function of  $|V_{ij}|$ . Another one is that very few things can be said about the location of the points  $\Omega_{ij}^{kl}$  on  $S_+^4$ . The only known result about the well distribution of these points is for the class of spheres which have their centers lying exactly on the velocity lattice (see for example [2]). The problem is then equivalent to find the distribution over the unit sphere of all the decompositions of an integer number into a sum of  $n$  square of integer numbers. In this case and by eliminating some values of  $n$  it can be proved that the set of the decompositions is well distributed over the unit sphere (see [8, 15]). When the center is not in the velocity lattice, the problem is now equivalent to the well distribution of special subsets of the decompositions of an integer number into a sum of  $n$  square of integer numbers, for which no results seems to be known. For the discrete monoatomic model ([5, 10]), this result allows, by eliminating the sphere that have not their centers of the velocity lattice, to prove the consistency of the discrete model with the continuous one (see [2]).

We shall estimate the number of the points  $\Omega_{ij}^{kl}$  by adapting a very classical theorem of number theory, about the number of decompositions of an integer number into a sum of squares of integers numbers (see [13]). This estimate will show that the discrete sphere are non empty and that the number of points  $\Omega_{ij}^{kl}$  "tends" to infinity with the diameter. We give the result in the general case. We suppose that the dimension of space is  $d \geq 1$ . We set  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d) \in \{0, 1\}^d$ . For  $i \in \mathbb{Z}^d$ , we have  $|i - \frac{\varepsilon}{2}|^2 \in \mathbb{N} + |\frac{\varepsilon}{2}|^2$  where, for  $i \in \mathbb{Z}^d$ ,  $|i|^2 = \sum_{p=1}^d i_p^2$ . Let

$$r_{\varepsilon,d}(n) = \text{Card} \left( \left\{ i \in \mathbb{Z}^d \mid \left| i - \frac{\varepsilon}{2} \right|^2 = n + \left| \frac{\varepsilon}{2} \right|^2 \right\} \right), \quad \text{for } n \in \mathbb{N}$$

the number of points of  $\mathbb{Z}^d$  on the sphere having the center  $\frac{\varepsilon}{2}$  and the radius

$(n + \left|\frac{\varepsilon}{2}\right|^2)^{\frac{1}{2}}$ . We write  $\mathcal{M}(E)$  for the Lebesgue measure of a measurable subset  $E$  of  $\mathbb{R}^n$ . We have the

**Lemma 4.1** For  $d \geq 2$ ,

$$\sum_{k=0}^n r_{\varepsilon,d}(k) = \mathcal{M}(S^{d-1})n^{\frac{d}{2}} + O(n^{\frac{d-1}{2}})$$

and  $r_{\varepsilon,d}(n) = O(n^{\frac{d-2}{2}+\delta})$  for all  $\delta > 0$ , or equivalently,  $r_{\varepsilon,d}(n) = o(n^{\frac{d-2}{2}+\delta})$  for all  $\delta > 0$ .

**Proof.** By setting  $E_n = \left\{i \in \mathbb{Z}^d \mid \left|i - \frac{\varepsilon}{2}\right|^2 \leq n + \left|\frac{\varepsilon}{2}\right|^2\right\}$ , we have

$$\sum_{k=0}^n r_{\varepsilon,d}(k) = \text{Card}(E_n).$$

At each point  $i$  we associate the cube  $C_i$  having  $i + \alpha$  for vertices with  $\alpha = (\alpha_1, \dots, \alpha_d) \in \{0, 1\}^d$ . We have  $\mathcal{M}(C_i) = 1$  and therefore,

$$\text{Card}(E_n) = \sum_{i \in E_n} \mathcal{M}(C_i) = \mathcal{M}\left(\bigcup_{i \in E_n} C_i\right).$$

It is clear that

$$\bigcup_{i \in E_n} C_i \subset B\left(\frac{\varepsilon}{2}, \sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} + \sqrt{d}\right).$$

For  $x \in B\left(\frac{\varepsilon}{2}, \sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} - \sqrt{d}\right)$  we have

$$\sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} - \sqrt{d} \geq \left|x - \frac{\varepsilon}{2}\right| = \left|[x] + \{x\} - \frac{\varepsilon}{2}\right| \geq \left|[x] - \frac{\varepsilon}{2}\right| - |\{x\}|$$

where  $[x]$  is the integer part of  $x$  and  $\{x\} = x - [x]$ . Hence,  $[x] \in E_n$  and in consequence we have then

$$B\left(\frac{\varepsilon}{2}, \sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} - \sqrt{d}\right) \subset \bigcup_{i \in E_n} C_i$$

and consequently

$$\mathcal{M}(S^{d-1})\left(\sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} - \sqrt{d}\right)^d \leq \sum_{k=0}^n r_{\varepsilon,d}(k) \leq \mathcal{M}(S^{d-1})\left(\sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} + \sqrt{d}\right)^d.$$

Since

$$\mathcal{M}(S^{d-1})\left(\sqrt{n + \left|\frac{\varepsilon}{2}\right|^2} \pm \sqrt{d}\right)^d$$

is clearly  $\mathcal{M}(S^{d-1})n^{\frac{d}{2}} + O(n^{\frac{d-1}{2}})$  we have proved the first part of the lemma. The second assertion is in fact a consequence of the following lemma (see [13]):

**Lemma 4.2** *If we call  $s(n)$  the number of decompositions of  $n$  in a sum of two squares of integer numbers then  $s(n) = O(n^\delta)$  for all  $\delta > 0$  or, equivalently,  $s(n) = o(n^\delta)$  for all  $\delta > 0$ .*

We prove the second assertion by recursion on the dimension of the space. The lemma yields the result for  $d = 2$ . We shall prove that the result holds for  $d > 2$ .

Assume that the result holds for one  $d \geq 2$ . Let  $i$  such that  $\sum_{p=1}^{d+1} i_p^2 = n$ . We have then  $i_{d+1}^2 \leq n$  which give  $|i_{d+1}| \leq \lfloor \sqrt{n} \rfloor$  and now

$$r_{0,d+1}(n) \leq \sum_{i_{d+1}=0}^{\lfloor \sqrt{n} \rfloor} r_{0,d}(n - i_{d+1}^2)$$

Using the assumption for  $d$ , we have

$$r_{0,d+1}(n) = O\left(\sum_{i_{d+1}=0}^{\lfloor \sqrt{n} \rfloor} (n - i_{d+1}^2)^{\delta + \frac{d-2}{2}}\right)$$

For simplicity we set  $\alpha = \delta + \frac{d-2}{2}$ . By the precedent equality, we obtain

$$r_{0,d+1}(n) = O\left((\lfloor \sqrt{n} \rfloor + 1)^{2\alpha+1} \cdot \frac{1}{\lfloor \sqrt{n} \rfloor + 1} \sum_{i_{d+1}=0}^{\lfloor \sqrt{n} \rfloor} \left(\frac{n}{(\lfloor \sqrt{n} \rfloor + 1)^2} - \left(\frac{i_{d+1}}{\lfloor \sqrt{n} \rfloor + 1}\right)^2\right)^\alpha\right).$$

The function  $x \rightarrow (x - a)^\alpha$  is increasing in  $x$  for all  $\alpha > 0$  and then

$$r_{0,d+1}(n) = O\left((\lfloor \sqrt{n} \rfloor + 1)^{2\alpha+1} \cdot \frac{1}{\lfloor \sqrt{n} \rfloor + 1} \sum_{i_{d+1}=0}^{\lfloor \sqrt{n} \rfloor} \left(1 - \left(\frac{i_{d+1}}{\lfloor \sqrt{n} \rfloor + 1}\right)^2\right)^\alpha\right)$$

But

$$\frac{1}{\lfloor \sqrt{n} \rfloor + 1} \sum_{i_{d+1}=0}^{\lfloor \sqrt{n} \rfloor} \left(1 - \left(\frac{i_{d+1}}{\lfloor \sqrt{n} \rfloor + 1}\right)^2\right)^\alpha = O\left(\int_0^1 (1 - y^2)^\alpha dy\right)$$

and

$$(\lfloor \sqrt{n} \rfloor + 1)^{2\alpha+1} = O(n^{\frac{1}{2}+\alpha})$$

Then we get the desired estimate for  $r_{0,d+1}(n)$ :

$$r_{0,d+1}(n) = O(n^{\frac{1}{2}+\alpha}) = O(n^{\frac{d+1-2}{2}+\delta})$$

In the case of  $\varepsilon \neq 0$  we can go back to the case  $\varepsilon = 0$  by noting that if  $i$  is such that

$$\left|i - \frac{\varepsilon}{2}\right|^2 = n + \left|\frac{\varepsilon}{2}\right|^2$$

or, equivalently

$$|2i - \varepsilon|^2 = 4n + |\varepsilon|^2$$

this implies

$$2i - \varepsilon \in \{j \mid |j|^2 = 4n + |\varepsilon|^2\}$$

and we have then the following inequality

$$r_{\varepsilon,d}(n) \leq r_{0,d}(4n + |\varepsilon|^2).$$

Using the result for  $r_{0,d}(n)$  and the fact that  $|\varepsilon|^2 \leq d$  we have

$$r_{\varepsilon,d}(n) = O\left\{(4n + |\varepsilon|^2)^{\frac{d-2}{2}} + \delta\right\} = O\left\{n^{\frac{d-2}{2}} + \delta\right\}$$

This ends the proof.  $\square$

This result shows that in the sense of the Cesaro mean value,  $\text{Card}(S_{ij})$ , which is  $\frac{1}{4}r_{\varepsilon,5}(|V|^2)$  with  $\varepsilon = (\eta_1, \eta_2, \eta_3, 1, 1)$  and  $(i - j) \equiv \eta \pmod{2}$ , behaves like  $|V|^3$ . A similar weak result can be easily obtained for the well distribution of the points of  $S_{ij}$  on the corresponding continuous quarter of sphere. Given a function  $g$  defined and continuous on the unit sphere  $S^4$  we define  $G$  on  $\mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}_+$  by:

$$G(x) = g\left(\frac{x}{|x|}\right).$$

A classical result on the well distribution of the points of a regular lattice  $\Delta v \mathbb{Z}^n$  which lye in a regular domain of  $\mathbb{R}^n$ , states that

$$\lim_{\Delta v \rightarrow 0} \frac{\sum_{i \Delta v \in D \cap \Delta v \mathbb{Z}^n} \phi(x)}{\sum_{i \Delta v \in D} 1} = \int_D \phi(x) dx$$

for all smooth test functions  $\phi$ . This gives us by setting  $N = \frac{1}{\Delta v}$ ,  $x_i = (i + \frac{\varepsilon}{2})\Delta v$  with  $i$  in  $\mathbb{Z}^n$  and by taking  $n = 5$ , the following identity:

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{\int_{B(\frac{\varepsilon}{2}\Delta v, 1) \cap \mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}_+} dx}{\sum_{j=1}^N \frac{1}{4}r_{\varepsilon,5}(j)} \sum_{j=1}^N \sum_{i \in \mathcal{I}_j} G(x_i) \\ = \int_{B(\frac{\varepsilon}{2}\Delta v, 1) \cap \mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}_+} G(x) dx. \end{aligned}$$

with

$$\mathcal{I}_j = \{i \text{ such that } x_i \in \mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}_+ \text{ and } |x_i - \frac{\varepsilon}{2}\Delta v|^2 = j\Delta v^2\}.$$

But

$$\int_{B(\frac{\varepsilon}{2}\Delta v, 1) \cap \mathbb{R}^3 \times \mathbb{R}_+ \times \mathbb{R}_+} G(x) dx = \frac{1}{5} \int_{S_+^4} g(\omega) d\omega$$

so in the sense of the Cesaro mean value, the points of  $\{|x_i - \frac{\varepsilon}{2}\Delta v|^2 = j\Delta v^2\}$  are well distributed that is

$$\lim_{j \rightarrow \infty} \frac{r_{\varepsilon,5}(j)}{4} \int_{S_+^4} d\omega \sum_{|x_i - \frac{\varepsilon}{2}\Delta v|^2 = j\Delta v^2} G(x_i) = \int_{S_+^4} g(\omega) d\omega$$

These two results tends to show that the approximation (4.29) is "reasonably accurate".

The overall approximation of the right hand side of (4.19) is now of the form

$$\begin{aligned} \frac{1}{4} \int_{\mathbb{R}^3 \times \mathbb{R}_+} Q_\delta(\varphi(v, I) + \varphi(v_*, I_*) - \varphi(v'_*, I'_*) - \varphi(v', I')) I^{\delta-1} dv dI \\ \simeq \frac{1}{4} \sum_{i,j \in L \times L} \sum_{k,l \in S_{i,j}} (\varphi(z_i) + \varphi(z_j) \\ - \varphi(z_k) - \varphi(z_l)(f_k f_l - f_i f_j) p_{kl} B_{ij}^{kl} I_i^{\delta-1} I_j^{\delta-1}). \end{aligned} \quad (4.32)$$

By setting  $\bar{f} = f_i, i \in L$ , and noticing that

$$p_{kl} B_{ij}^{kl} I_i^{\delta-1} I_j^{\delta-1} = p_{kl} B_{ji}^{kl} I_i^{\delta-1} I_j^{\delta-1} = p_{lk} B_{ij}^{lk} I_i^{\delta-1} I_j^{\delta-1} = p_{ij} B_{kl}^{ij} I_k^{\delta-1} I_l^{\delta-1}$$

we obtain the following approximation of the continuous homogeneous problem

$$\frac{df_i}{dt} = \bar{Q}_\delta(\bar{f}, \bar{f})_i \quad (4.33)$$

with

$$\bar{Q}_\delta(\bar{f}, \bar{f})_i = \sum_{j \in L} \sum_{(k,l) \in S_{ij}} (A_{kl}^{ij} f_k f_l - A_{ij}^{kl} f_i f_j) \quad (4.34)$$

or

$$\bar{Q}_\delta(\bar{f}, \bar{f})_i = \sum_{(j,k,l) \in (L)^3} (A_{k,l}^{i,j} f_k f_l - A_{ij}^{kl} f_i f_j), \quad (4.35)$$

where we have extended the definition of  $A_{ij}^{kl}$  and for further purposes, we also introduce the tensor  $\mathcal{A}_{ij}^{kl}$  by:

$$A_{ij}^{kl} = \frac{\mathcal{A}_{ij}^{kl}}{I_i^{\delta-1}} = \begin{cases} I_j^{\delta-1} p_{kl} B_{ij}^{kl} & \text{if } (k, l) \in S_{ij} \\ 0 & \text{otherwise.} \end{cases} \quad (4.36)$$

We consider now a bounded velocity domain. The issue is to replace the Boltzmann equation in the whole velocity space domain, by a bounded space one, for which the algebraic properties displayed in section 3.1 still hold. We proceed as in [18, 19]. Let  $\mathcal{D}_{v,I}$  be a bounded domain of  $\mathbb{R}^3 \times \mathbb{R}_+$ , and let  $\chi((v, I), (v_*, I_*), (v', I'), (v'_*, I'_*))$  be the following characteristic function

$$\chi(a, b, c, d) = \begin{cases} 1 & \text{if } (a, b, c, d) \in \mathcal{D}_{v,I} \\ 0 & \text{otherwise.} \end{cases} \quad (4.37)$$

Now, let us consider the Boltzmann operator

$$Q_\delta(f, f)(v) = \int_{\mathcal{D}_{v,I}} \int_{S_+^4} \chi((v, I), (v_*, I_*), (v', I'), (v'_*, I'_*)) \mathcal{B} \\ (f' f'_* - f f_*) dv_* I_*^{\delta-1} dI_* d\Omega. \quad (4.38)$$

For  $v \in \mathcal{D}_{v,I}$ , it is easy to show that properties (2.6) to (2.9) still hold, with the only difference that the coefficient of  $\frac{|v|^2}{2} + I^2$  in (2.8) is no more necessarily positive. Indeed, its positivity for the full space case follows from integrability requirements on the Maxwellian, which can no more be used because of the boundedness of the domain. The particle approximation of problem (4.18) is now restricted to approximations  $f_i$  of  $(\Delta v)^4 f(z_i)$  for  $z_i \in \Delta v L \cap \mathcal{D}_{v,I}$ . Let  $E$  be the set of indices, which is included in  $L$ , for which  $z_i \in \mathcal{D}_{v,I}$ . The discrete homogeneous problem for  $i \in E$  is now written

$$\frac{df_i(t)}{dt} = \bar{Q}_\delta(\bar{f}, \bar{f})_i = \sum_{j \in E} \sum_{(k,l) \in \tilde{S}_{ij}} (A_{k,l}^{i,j} f_k f_l - A_{ij}^{kl} f_i f_j)$$

with

$$\tilde{S}_{ij} = \{(k, l) \in S_{ij} \text{ such that } k, l \in E\} \quad (4.39)$$

where

$$A_{ij}^{kl} = \frac{\mathcal{A}_{ij}^{kl}}{I_i^{\delta-1}} = \begin{cases} I_j^{\delta-1} \tilde{p}_{kl} B_{ij}^{kl} & \text{if } z_i, z_j \in \mathcal{D}_{v,I} \text{ and } (k, l) \in \tilde{S}_{ij} \\ 0 & \text{otherwise.} \end{cases} \quad (4.40)$$

with

$$\tilde{p}_{kl} = \frac{I_k^{\delta-1} I_l^{\delta-1}}{\sum_{(k,l) \in \tilde{S}_{ij}} I_k^{\delta-1} I_l^{\delta-1}}.$$

For example, in the VHS model case, we can take

$$\tilde{p}_{kl} = \frac{|g_{kl}|^{1-2\alpha} I_k^{\delta-1} I_l^{\delta-1}}{\sum_{(k,l) \in \tilde{S}_{ij}} |g_{kl}|^{1-2\alpha} I_k^{\delta-1} I_l^{\delta-1}}.$$

#### 4.1.2. Properties of the discrete collision operator

It is easy to check that the tensor  $\mathcal{A}_{ij}^{kl}$  is non negative and satisfies the following symmetry properties

$$\mathcal{A}_{ij}^{kl} = \mathcal{A}_{ji}^{kl} = \mathcal{A}_{ij}^{lk} \quad (4.41)$$

and also the microreversibility property

$$\mathcal{A}_{ij}^{kl} = \mathcal{A}_{kl}^{ij} \quad (4.42)$$

Therefore, see (4.32), we have the discrete analogue of identity (2.5): let  $\bar{\varphi} = (\varphi_i)_{i \in E}$  be a test sequence, then

$$\sum_{i \in E} \bar{Q}_\delta(\bar{f}, \bar{f})_i \varphi_i I_i^{\delta-1} = \frac{1}{4} \sum_{(i,j,k,l) \in (E)^4} \mathcal{A}_{ij}^{kl} (f_k f_l - f_i f_j) (\varphi_i + \varphi_j - \varphi_k - \varphi_l) \quad (4.43)$$

Using the definition of tensor  $\mathcal{A}_{ij}^{kl}$  and equality (4.43) it is easy to show that the discrete analogue of conservation of mass, momentum and energy

$$\sum_{i \in E} \bar{Q}_\delta(\bar{f}, \bar{f})_i I_i^{\delta-1} \begin{pmatrix} 1 \\ v_i \\ \frac{|v_i|^2}{2} + I_i^2 \end{pmatrix} = 0, \quad (4.44)$$

holds. Also using (4.43) and the classical inequality  $(y - x)(\log x - \log y) \leq 0$  for any  $x, y > 0$ , a discrete H-theorem holds that yields dissipation of entropy  $H(t) = \sum_{i \in E} f_i(t) \log f_i(t) I_i^{\delta-1}$

$$\begin{aligned} \frac{dH(t)}{dt} &= \sum_{i \in E} \bar{Q}_\delta(\bar{f}, \bar{f})_i I_i^{\delta-1} \log(f_i) \\ &= \frac{1}{4} \sum_{(i,j,k,l) \in (E)^4} \mathcal{A}_{ij}^{kl} (f_k f_l - f_i f_j) (\log(f_i) + \log(f_j) - \log(f_k) - \log(f_l)) \\ &= \frac{1}{4} \sum_{(i,j,k,l) \in (E)^4} \mathcal{A}_{ij}^{kl} (f_k f_l - f_i f_j) (\log(f_i f_j) - \log(f_k f_l)) \leq 0. \end{aligned} \quad (4.45)$$

As in the monoatomic case, the equilibrium states,  $\bar{f}^\infty$ , for which  $\frac{dH(t)}{dt} = 0$ , are characterized by (see [9]):

**Proposition 4.1** *The following properties are equivalent*

1.  $\sum_{i \in E} \bar{Q}_\delta(\bar{f}^\infty, \bar{f}^\infty)_i \log(f_i^\infty) = 0$
2.  $\bar{Q}_\delta(\bar{f}^\infty, \bar{f}^\infty)_i = 0 \ \forall i \in E$
3.  $\overline{\log f^\infty} = (\log f_i^\infty)_{i \in E}$  is an invariant of collision that is  $\overline{\log f^\infty} \in \{\bar{\varphi} \text{ such that } \varphi_i + \varphi_j - \varphi_k - \varphi_l = 0 \text{ for all } i, j, k, l \text{ such that } A_{ij}^{kl} \neq 0\}$
4.  $f_i^\infty f_j^\infty - f_k^\infty f_l^\infty = 0$  if  $A_{ij}^{kl} \neq 0$

To see that these properties are equivalent one must recall that

$$(y - x)(\log x - \log y) = 0 \Leftrightarrow x = y$$

and use (4.43). Now, for the specific model given by (4.36) or by (4.40), it is noticeable that the reciprocal of (4.44) holds, like in the continuous case and so the only equilibrium states are discrete Maxwellians.



For a bounded domain we suppose that  $\mathcal{D}_{v,I}$  is of the form  $\mathcal{D}_{v,I} = B(0, R) \times [0, S]$  or  $\mathcal{D}_{v,I} = [-R, R]^3 \times [0, S]$ . For such bounded domains, the set  $E$  is

$$E = \{i \in L, i_1^2 + i_2^2 + i_3^2 \leq M_1, i_4 \leq N\}$$

or

$$E = \{i \in L, i \sup_{p=1,3} i_p \leq M_2, i_4 \leq N\},$$

where  $M_1, M_2$  and  $N$  are integers and we assume that  $M_1 \geq 3$ ,  $M_2 \geq 1$  and  $N \leq M_2 - 2$ . In the case of an unbounded domain for  $(v, I)$ , we get  $E = L$ . We have

**Lemma 4.3** *For the model given by (4.36) or by (4.40) the invariants of collisions are given by*

$$\varphi_i = A(v_i^2 + 2I_i^2) + \langle B, v_i \rangle + C$$

with  $A$  and  $C \in \mathbb{R}$  and  $B \in \mathbb{R}^3$  and the equilibrium states  $\bar{f}^\infty$  have the form

$$f_i^\infty = \exp(A(v_i^2 + 2I_i^2) + \langle B, v_i \rangle + C)$$

**Proof.** For  $\varphi = (\varphi_i)_{i \in E}$ , we let  $\varphi(m)$  be the restriction of  $\varphi$  to the subset  $E_m = \{i \in E, i_4 = m\}$  that is,  $\varphi(m)$  is the restriction of  $\varphi$  at the level  $m + \frac{1}{2}$ . We note  $\bar{i} = (i_1, i_2, i_3)$  for any element  $i$  of  $E_m$ . For  $\varphi(m)$  we have the result

$$\varphi_i(m) = A(m)|\bar{i}|^2 + \langle B(m), \bar{i} \rangle + C(m)$$

with  $A(m)$  and  $C(m) \in \mathbb{R}$  and  $B(m) \in \mathbb{R}^3$ . We consider only elastic collisions between two elements of  $E_m$ . We consider first the case of  $E_m = \mathbb{Z}^3$ . We say that  $(\bar{i}, \bar{j}) \rightarrow (\bar{k}, \bar{l})$  is a possible collision if  $\mathcal{A}_{(\bar{i}, m), (\bar{j}, m)}^{(\bar{k}, m), (\bar{l}, m)} \neq 0$  that implies by the symmetry of  $\mathbb{Z}^3$  that the collision  $(-\bar{i}, -\bar{j}) \rightarrow (-\bar{k}, -\bar{l})$  is also admissible. Let  $e_1 = (1, 0, 0)$ ,  $e_2 = (0, 1, 0)$ ,  $e_3 = (0, 0, 1)$  be the canonical basis of  $\mathbb{Z}^3$ . We search  $\varphi(m)$  such that

$$\varphi_{\bar{i}}(m) + \varphi_{\bar{j}}(m) - \varphi_{\bar{k}}(m) - \varphi_{\bar{l}}(m) = 0 \quad \text{for all } \mathcal{A}_{(\bar{i}, m), (\bar{j}, m)}^{(\bar{k}, m), (\bar{l}, m)} \neq 0$$

We set

$$a_{\bar{i}}(m) = \varphi_{\bar{i}}(m) + \varphi_{-\bar{i}}(m) \text{ and } b_{\bar{i}}(m) = \varphi_{\bar{i}}(m) - \varphi_{-\bar{i}}(m)$$

By construction we have  $a_{-\bar{i}}(m) = a_{\bar{i}}(m)$ ,  $b_{-\bar{i}}(m) = -b_{\bar{i}}(m)$  and then  $b_0(m) = 0$ .

We show recursively on  $L = |\bar{i}|_\infty = \max_{n=1}^3 |i_n|$  that

$$a_{\bar{i}} - a_0 = |\bar{i}|^2 \cdot (a_{e_1} - a_0), \quad b_{\bar{i}} = i_1 b_{e_1} + i_2 b_{e_2} + i_3 b_{e_3}$$

This is evidently true for  $a_{\bar{i}}$  when  $|\bar{i}|_\infty = 1$  because  $(e_\alpha, -e_\alpha) \rightarrow (e_\beta, -e_\beta)$  is an admissible collision. For  $b_{\bar{i}}$  this is trivial. We suppose now that it is true until rank  $L$ . Let  $\bar{i} \in \mathbb{Z}^3$  such that  $|\bar{i}|_\infty = L + 1$ . If  $(\bar{i}, \bar{j}) \rightarrow (\bar{k}, \bar{l})$  is a possible collision then,

by construction,  $a_{\bar{i}} + a_{\bar{j}} = a_{\bar{k}} + a_{\bar{l}}$  and  $b_{\bar{i}} + b_{\bar{j}} = b_{\bar{k}} + b_{\bar{l}}$ . Since the following collisions are admissible

$$(\bar{i}, 0) \rightarrow (i_1 e_1 + i_2 e_2, i_3 e_3), \quad (i_1 e_1 + i_2 e_2, 0) \rightarrow (i_1 e_1, i_2 e_2)$$

we have

$$a_{\bar{i}} + a_0 = a_{i_1 e_1 + i_2 e_2} + a_{i_3 e_3} = a_{i_1 e_1} + a_{i_2 e_2} + a_{i_3 e_3} - a_0$$

and then

$$a_{\bar{i}} - a_0 = (a_{i_1 e_1} - a_0) + (a_{i_2 e_2} - a_0) + (a_{i_3 e_3} - a_0)$$

and for  $b_{\bar{i}}$

$$b_{\bar{i}} = b_{i_1 e_1} + b_{i_2 e_2} + b_{i_3 e_3}.$$

It suffices then to verify that

$$a_{(L+1)e_\alpha} - a_0 = (L+1)^2(a_{e_1} - a_0), \quad b_{(L+1)e_\alpha} = (L+1)b_{e_\alpha}. \quad (4.46)$$

We define  $u = (L-1)e_\alpha$ ,  $v = Le_\alpha + e_\beta$  and  $w = Le_\alpha - e_\beta$ . Since  $|u|_\infty = L-1$  and  $|v|_\infty = |w|_\infty = L$  the assumption holds for  $u, v, w$ . Since the following collision is possible

$$\left( (L+1)e_\alpha, u \right) \rightarrow \left( v, w \right),$$

(4.46) is true: indeed we have for  $a_{\bar{i}}$

$$a_{(L+1)e_\alpha} + a_u = a_v + a_w$$

and then

$$\begin{aligned} a_{(L+1)e_\alpha} - a_0 &= a_w - a_0 + a_v - a_0 - (a_u - a_0) \\ &= (|w|^2 + |v|^2 - |u|^2)(a_{e_1} - a_0) \\ &= (L^2 + 1 + L^2 + 1 - (L-1)^2)(a_{e_1} - a_0) \\ &= (L+1)^2(a_{e_1} - a_0). \end{aligned}$$

and for  $b_{\bar{i}}$

$$b_{(L+1)e_\alpha} = b_v + b_w - b_u = Lb_{e_\alpha} + b_{e_\beta} + Lb_{e_\alpha} - b_{e_\beta} - (L-1)b_{e_\alpha} = (L+1)b_{e_\alpha}$$

Since  $\varphi_{\bar{i}} = \frac{a_{\bar{i}} + b_{\bar{i}}}{2}$  we have the result for  $\varphi_{\bar{i}}$  with

$$A(m) = \frac{a_{e_1} - a_0}{2}, \quad C(m) = \frac{a_0}{2}, \quad B(m) = \left( \frac{b_{e_1}}{2}, \frac{b_{e_2}}{2}, \frac{b_{e_3}}{2} \right).$$

Recall that  $E_m = \{\bar{i} \in \mathbb{Z}^3 / i_1^2 + i_2^2 + i_3^2 \leq M_1\}$  or  $E_m = \{\bar{i} \in \mathbb{Z}^3 / \sup_{k=1}^3 i_k \leq M_2\}$ . It is then easy to check, from the above analysis, that the result for the form of  $\varphi(m)$  remains valid provided  $M_1 \geq 3$  and  $M_2 \geq 1$ .

Let us consider the case of inelastic collisions. For a given  $m$  such that  $m \leq M_1 - 2$  or  $m \leq M_2 - 2$  in the case of bounded domain, we consider the following collisions

$$\begin{aligned} z_i = \Delta v(0, 0, 0, m + \frac{1}{2}), z_j = \Delta v(1, 0, 0, \frac{1}{2}) &\rightarrow z_k = \Delta v(m + 1, 0, 0, \frac{1}{2}), \\ z_l = \Delta v(-m, 0, 0, \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} z_i = \Delta v(1, 0, 0, m + \frac{1}{2}), z_j = \Delta v(0, 0, 0, \frac{1}{2}) &\rightarrow z_k = \Delta v(m + 1, 0, 0, \frac{1}{2}), \\ z_l = \Delta v(-m, 0, 0, \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} z_i = \Delta v(0, 1, 0, m + \frac{1}{2}), z_j = \Delta v(0, 0, 0, \frac{1}{2}) &\rightarrow z_k = \Delta v(m + 1, 0, 0, \frac{1}{2}), \\ z_l = \Delta v(-m, 0, 0, \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} z_i = \Delta v(0, 0, 1, m + \frac{1}{2}), z_j = \Delta v(0, 0, 0, \frac{1}{2}) &\rightarrow z_k = \Delta v(m + 1, 0, 0, \frac{1}{2}), \\ z_l = \Delta v(-m, 0, 0, \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} z_i = \Delta v(2, 0, 0, m + \frac{1}{2}), z_j = \Delta v(0, 0, 0, \frac{1}{2}) &\rightarrow z_k = \Delta v(m + 2, 0, 0, \frac{1}{2}), \\ z_l = \Delta v(1 - m, 0, 0, \frac{1}{2}) \end{aligned}$$

For these collisions we have  $\mathcal{A}_{ij}^{kl} \neq 0$  and then for  $\varphi$  we must have

$$\varphi_i + \varphi_j = \varphi_k + \varphi_l$$

By using the result for  $\varphi$  at the levels  $\frac{1}{2}$  and  $m + \frac{1}{2}$  of internal energy we derive the following set of equations

$$\begin{cases} C(m) = 2A(0)((m + \frac{1}{2})^2 - \frac{1}{4}) + C(0) \\ A(m) + B_1(m) = A(0) + B_1(0) \\ A(m) + B_2(m) = A(0) + B_2(0) \\ A(m) + B_3(m) = A(0) + B_3(0) \\ 4A(m) + 2B_1(m) = 4A(0) + 2B_1(0) \end{cases}$$

for which the solution is

$$\begin{cases} C(m) = 2A(0)((m + \frac{1}{2})^2 - \frac{1}{4}) + C(0) \\ A(m) = A(0) \\ B_1(m) = B_1(0) \\ B_2(m) = B_2(0) \\ B_3(m) = B_3(0) \end{cases}$$

and then for each  $m$

$$\varphi_{\bar{i}}(m) = A(0)|\bar{i}|^2 + \langle B(0), \bar{i} \rangle + 2A(0)((m + \frac{1}{2})^2 - \frac{1}{4}) + C(0)$$

We have indeed for any  $i \in E$ , using the definition of  $z_i = (v_i, I_i)$

$$\varphi_i = A(|v_i|^2 + 2I_i^2) + \langle B, v_i \rangle + C.$$

Since to say that  $\bar{f}^\infty$  is an equilibrium state is equivalent to say that  $\overline{\log f^\infty}$  is an invariant of collisions we have indeed

$$\bar{f}_i^\infty = \exp(A(|v_i|^2 + 2I_i^2) + \langle B, v_i \rangle + C),$$

which ends the proof.  $\square$

Since the only invariants of collisions are  $(1)_{i \in E}, (v_i)_{i \in E}, (|v_i|^2 + 2I_i^2)_{i \in E}$  the constants  $A, B, C$  are functions of the density, mean velocity, and temperature of  $\bar{f}$ . All these properties show that our discrete collision operator for polyatomic gases behaves like the continuous one as we have claimed.

#### 4.2. A second discrete velocity and energy model

We propose a discrete collision operator which mimics the discrete-velocity collision dynamics described in [11] for polyatomic gases. We show that this discrete collision operator, which uses a finer discretization of the internal energy than the previously described one, has also the same properties as the continuous one. Now we use the expression of  $Q_\delta$  given by (3.17) and for simplicity, we restrict ourselves to the case of VHS models.

##### 4.2.1. Discretization

We again take a regular discretization of  $\mathbb{R}^3 \times \mathbb{R}^+$ : Let  $\Delta v > 0$ ,

$$(v_i, e_i) = (i_1 \Delta v, i_2 \Delta v, i_3 \Delta v, (i_4 + \frac{1}{2}) \Delta v^2), \quad (4.47)$$

with  $i = (i_1, i_2, i_3, i_4) \in L = \mathbb{Z}^3 \times \mathbb{N}$ , and we let  $f_i(t) \simeq (\Delta v)^5 f(v_i, e_i, t)$ . We introduce a particle approximation of the problem (4.18). Given any test function  $\varphi(v, e)$  we have

$$\begin{aligned} & \int_{\mathbb{R}^3 \times \mathbb{R}^+} \frac{df}{dt} \varphi(v, e) e^{\frac{\delta}{2}-1} dv de \\ &= \frac{1}{4} \int_{\mathbb{R}^3 \times \mathbb{R}^+} Q_\delta(\varphi(v, e) + \varphi(v_*, e_*) - \varphi(v'_*, e'_*) - \varphi(v', e')) e^{\frac{\delta}{2}-1} dv de \\ &= \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{R}^+ \times S^2 \times [0,1]^2} B(\varphi + \varphi_* - \varphi' - \varphi'_*)(f' f'_* - f f_*) d\sigma \end{aligned} \quad (4.48)$$

with the measure and  $B$  of the form

$$d\sigma = dv_* dv e^{\frac{\delta}{2}-1} de_* e^{\frac{\delta}{2}-1} d\omega R^2 (1 - R^2)^{\delta-1} dR [r(1-r)]^{\frac{\delta}{2}-1} dr$$

$$B := B(E, |g|, |g'|, |g \cdot g'|, ee', e_* e'_*) > 0$$

We derive an approximation of the two terms of the equality (4.48) by using quadrature formulae for the integrals. For the left hand side, we obtain

$$\int_{\mathbb{R}^3 \times \mathbb{R}^+} \frac{df(v, e)}{dt} \varphi(v, e) e^{\frac{\delta}{2}-1} dv de \simeq \sum_{i \in L} \frac{df_i}{dt} \varphi(v_i, e_i) e_i^{\frac{\delta}{2}-1}. \quad (4.49)$$

For the right hand side term of (4.48) we make some transformations. We set

$$U = \frac{v + v_*}{2}$$

and

$$dv dv_* = 8dU dg.$$

Let us define the domain of admissible parameters

$$D_{g, e, e_*} = \{(g', e') \in \mathbb{R}^3 \times \mathbb{R}^+ \text{ such that } |g'|^2 + e' \leq E(g, e, e_*) = |g|^2 + e + e_*\}$$

and we use the following change of variables for fixed  $(g, e, e_*)$

$$\begin{cases} \omega = \frac{g'}{|g'|} \\ R = \frac{|g'|}{\sqrt{E(g, e, e_*)}} \\ r = \frac{e'}{E(g, e, e_*) - |g'|^2} \end{cases}$$

where  $(g', e') \in D_{g, e, e_*}$ . We have

$$d\omega [r(1-r)]^{\frac{\delta}{2}-1} dr R^2 (1-R^2)^{\delta-1} dR = \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E(g, e, e_*)^{\delta+\frac{1}{2}}} dg' de'$$

and with

$$e'_* = E(g, e, e_*) - |g'|^2 - e'$$

The right hand side of (4.48) can be written

$$\begin{aligned} & \frac{1}{4} \int_{\mathbb{R}^4} Q_\delta(\varphi + \varphi_* - \varphi' - \varphi'_*) e^{\frac{\delta}{2}-1} dv de \\ &= 2 \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{R}^+} \left\{ \int_{D_{g, e, e_*}} B(\varphi + \varphi_* - \varphi' - \varphi'_*) (f' f'_* - f f_*) \right. \\ & \quad \left. \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E(g, e, e_*)^{\delta+\frac{1}{2}}} dg' de' \right\} dU dg de de_*. \quad (4.50) \end{aligned}$$

We first discretize with respect to  $U$ . The quadrature points are the elements of  $\frac{\Delta v}{2} \cdot \mathbb{Z}^3$ . We have

$$\begin{aligned} & \frac{1}{4} \int_{\mathbb{R}^4} Q_\delta(\varphi + \varphi_* - \varphi' - \varphi'_*) e^{\frac{\delta}{2}-1} dv de \\ & \simeq \frac{\Delta v^3}{4} \sum_{U_m \in \frac{\Delta v}{2} \cdot \mathbb{Z}^3} \int_{\mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{R}^+} \left\{ \int_{D_{g,e,e_*}} B(\varphi(U_m + g, e) + \varphi(U_m - g, e_*)) \right. \\ & \quad - \varphi(U_m + g', e') - \varphi(U_m - g', e'_*) (f(U_m + g', e') f(U_m - g', e'_*) \\ & \quad \left. - f(U_m + g, e) f(U_m - g, e_*)) \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E(g, e, e_*)^{\delta+\frac{1}{2}}} dg' de' \right\} dg de de_*. \quad (4.51) \end{aligned}$$

By discretizing now with respect to  $g, e, e_*$  with the quadrature points in  $(U_m + \Delta v \mathbb{Z}^3) \times (\Delta v^2 (\mathbb{N} + \frac{1}{2}))^2$ , and by setting

$$U_{ij} = \frac{v_i + v_j}{2}, \quad g_{ij} = \frac{v_i - v_j}{2}, \quad D_{ij} = D_{g_{ij}, e_i, e_j}, \quad E_{ij} = E_{g_{ij}, e_i, e_j}$$

we obtain the approximation

$$\begin{aligned} & \frac{1}{4} \int_{\mathbb{R}^4} Q_\delta(\varphi + \varphi_* - \varphi' - \varphi'_*) e^{\frac{\delta}{2}-1} dv de \\ & \simeq \frac{\Delta v^{10}}{4} \sum_{i,j \in L} e_i^{\frac{\delta}{2}-1} e_j^{\frac{\delta}{2}-1} \left\{ \int_{D_{ij}} B(\varphi(v_i, e_i) + \varphi(v_j, e_j) - \varphi(U_{ij} + g', e') \right. \\ & \quad - \varphi(U_{ij} - g', e'_*)) (f(U_{ij} + g', e') f(U_{ij} - g', e'_*) \\ & \quad \left. - f(v_i, e_i) f(v_j, e_j)) \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E_{ij}^{\delta+\frac{1}{2}}} dg' de' \right\} \quad (4.52) \end{aligned}$$

We must now discretize the collision integral with respect to  $g'$  and  $e'$ . For fixed  $i$  and  $j$  we take the quadrature points for  $g', e'$  in  $L_{ij} = (U_{ij} + \Delta v \mathbb{Z}^3) \times \Delta v^2 (\mathbb{N} + \frac{1}{2})$ .

One can verify that for  $(g', e') \in L_{ij} \cap D_{ij}$  we have  $e'_* \in \Delta v^2 (\mathbb{N} + \frac{1}{2})$  and  $U_{ij} \pm g' \in \Delta v \cdot \mathbb{Z}^3$ . We use the same kind of quadrature formula as for the other variables  $U, g, e, e_*$ . First we remark that

$$\int_{D_{ij}} B \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E_{ij}^{\delta+\frac{1}{2}}} dg' de' = C_{\alpha\delta} |g_{ij}|^{1-2\alpha} \quad (4.53)$$

where  $C_{\alpha\delta}$  is a constant just depending of  $\delta$  and  $\alpha$ . In the same manner as for the first discrete model, by defining

$$S_{ij} = \{k, l \in L \text{ such that } g_{kl} \in L_{ij} \cap D_{ij} \text{ and } U_{kl} = U_{ij}\}$$

we obtain the following approximation

$$C_{\alpha\delta}|g_{ij}|^{1-2\alpha} \simeq \Delta v^5 \sum_{k,l \in S_{ij}} \frac{C}{E_{ij}^{1+\delta-\alpha}} (e_k e_l)^{\frac{\delta}{2}-1} |g_{kl}|^{1-2\alpha} |g_{ij}|^{1-2\alpha} \quad (4.54)$$

and

$$\begin{aligned} & \int_{D_{ij}} B(\varphi(v_i, e_i) + \varphi(v_j, e_j) - \varphi(U_{ij} + g', e') - \varphi(U_{ij} - g', e'_*)) \\ & (f(U_{ij} + g', e')f(U_{ij} - g', e'_*) - f(v_i, e_i)f(v_j, e_j)) \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E_{ij}^{\delta+\frac{1}{2}}} dg' de' \\ & \simeq \Delta v^5 \sum_{k,l \in S_{ij}} \frac{C}{E_{ij}^{1+\delta-\alpha}} (e_k e_l)^{\frac{\delta}{2}-1} |g_{kl}|^{1-2\alpha} |g_{ij}|^{1-2\alpha} \\ & (\varphi(v_i, e_i) + \varphi(v_j, e_j) - \varphi(v_k, e_k) - \varphi(v_l, e_l)) \\ & (f(v_k, e_k)f(v_l, e_l) - f(v_i, e_i)f(v_j, e_j)), \end{aligned} \quad (4.55)$$

where the constant  $C$  comes from the definition of the VHS cross section (2.10). We can remark that for  $(k, l) \in S_{ij}$  we have  $S_{kl} = S_{ij}$ . (4.53), (4.54) and (4.55) gives the approximation

$$\begin{aligned} & \int_{D_{ij}} B(\varphi(v_i, e_i) + \varphi(v_j, e_j) - \varphi(U_{ij} + g', e') - \varphi(U_{ij} - g', e'_*)) \\ & (f(U_{ij} + g', e')f(U_{ij} - g', e'_*) - f(v_i, e_i)f(v_j, e_j)) \frac{(e' e'_*)^{\frac{\delta}{2}-1}}{E_{ij}^{\delta+\frac{1}{2}}} dg' de' \\ & \simeq C_{\alpha\delta}|g_{ij}|^{1-2\alpha} \sum_{k,l \in S_{ij}} p_{kl} (\varphi(v_i, e_i) + \varphi(v_j, e_j) - \varphi(v_k, e_k) - \varphi(v_l, e_l)) \\ & (f(v_k, e_k)f(v_l, e_l) - f(v_i, e_i)f(v_j, e_j)) \end{aligned} \quad (4.56)$$

with the weight  $p_{kl}$  defined by

$$p_{kl} = \frac{(e_k e_l)^{\frac{\delta}{2}-1} |g_{kl}|^{1-2\alpha}}{\sum_{k,l \in S_{ij}} (e_k e_l)^{\frac{\delta}{2}-1} |g_{kl}|^{1-2\alpha}}.$$

We set  $\bar{f} = (f_i)_{i \in L}$ . Using (4.49), (4.52), (4.56) and the definition of the  $f_i$  we deduce a particle approximation of the continuous homogeneous problem of the form

$$\sum_{i \in L} \frac{df_i}{dt} e_i^{\frac{\delta}{2}-1} \delta(v - v_i) \otimes \delta(e - e_i) = \sum_{i \in L} \bar{Q}_\delta(\bar{f}, \bar{f})_i e_i^{\frac{\delta}{2}-1} \delta(v - v_i) \otimes \delta(e - e_i) \quad (4.57)$$

with

$$\bar{Q}_\delta(\bar{f}, \bar{f})_i = \sum_{j \in L} \sum_{(k,l) \in S_{ij}} (A_{kl}^{ij} f_k f_l - A_{ij}^{kl} f_i f_j) \quad (4.58)$$

or

$$\bar{Q}_\delta(\bar{f}, \bar{f})_i = \sum_{(j,k,l) \in (L)^3} (A_{k,l}^{i,j} f_k f_l - A_{ij}^{kl} f_i f_j), \quad (4.59)$$

with

$$A_{ij}^{kl} = \frac{\mathcal{A}_{ij}^{kl}}{e_i^{\frac{\delta}{2}-1}} = \begin{cases} C_{\alpha\delta} e_j^{\frac{\delta}{2}-1} p_{kl} |g_{ij}|^{1-2\alpha} & \text{if } (k,l) \in S_{ij} \\ 0 & \text{otherwise.} \end{cases} \quad (4.60)$$

#### 4.2.2. Properties of the discrete collision operator

As for the first model, by construction the tensor  $\mathcal{A}_{ij}^{kl}$  is non negative, and satisfies the symmetry properties

$$\mathcal{A}_{ij}^{kl} = \mathcal{A}_{ji}^{kl} = \mathcal{A}_{ij}^{lk} \quad (4.61)$$

and the so called microreversibility property

$$\mathcal{A}_{ij}^{kl} = \mathcal{A}_{kl}^{ij}. \quad (4.62)$$

We can write a discrete analogue of identity (2.5): let  $\bar{\varphi} = (\varphi_i)_{i \in L}$  be a test sequence, then

$$\sum_{i \in L} \bar{Q}_\delta(\bar{f}, \bar{f})_i \varphi_i e_i^{\frac{\delta}{2}-1} = \frac{1}{4} \sum_{(i,j,k,l) \in (L)^4} \mathcal{A}_{ij}^{kl} (f_k f_l - f_i f_j) (\varphi_i + \varphi_j - \varphi_k - \varphi_l). \quad (4.63)$$

Using the definition of tensor  $\mathcal{A}_{ij}^{kl}$  and equality (4.63) we write the discrete analogue of conservation of mass, momentum and energy according to

$$\sum_{i \in L} \bar{Q}_\delta(\bar{f}, \bar{f})_i e_i^{\frac{\delta}{2}-1} \left( \frac{1}{\frac{|v_i|^2}{2} + e_i} \right) = 0. \quad (4.64)$$

For a global positive solution the discrete H-theorem, that is the dissipation of entropy

$$H(t) = \sum_{i \in L} f_i(t) \log(f_i(t)) e_i^{\frac{\delta}{2}-1},$$

hold:

$$\frac{dH(t)}{dt} = \frac{1}{4} \sum_{(i,j,k,l) \in (L)^4} \mathcal{A}_{ij}^{kl} (f_k f_l - f_i f_j) (\log(f_i f_j) - \log(f_k f_l)) \leq 0. \quad (4.65)$$

As in the first model, the equilibrium states  $\bar{f}^\infty$ , are characterized by the same properties (see proposition 1). For this model the equilibrium states are still discrete Maxwellians:



**Lemma 4.4** For the discrete collision operator defining by (4.58) and (4.60) the invariants of collisions are given by

$$\varphi_i = A(v_i^2 + 2e_i) + \langle B, v_i \rangle + C$$

with  $A$  and  $C \in \mathbb{R}$  and  $B \in \mathbb{R}^3$  and the equilibrium states  $\bar{f}^\infty$  have the form

$$f_i^\infty = \exp(A(v_i^2 + 2e_i) + \langle B, v_i \rangle + C)$$

**Proof.** For  $\varphi = (\varphi_i)_{i \in L}$ , we let  $\varphi(m)$  be the restriction of  $\varphi$  to the subset  $L_m = \{i \in L, i_4 = m\}$  that is,  $\varphi(m)$  is the restriction of  $\varphi$  at the level  $m + \frac{1}{2}$  of internal energy. We set  $\bar{i} = (i_1, i_2, i_3)$  for an element  $i$  of  $L_m$ . For  $\varphi(m)$  we have already proved, (see proof of lemma 3), that

$$\varphi_i(m) = A(m)|\bar{i}|^2 + \langle B(m), \bar{i} \rangle + C(m)$$

with  $A(m)$  and  $C(m) \in \mathbb{R}$  and  $B(m) \in \mathbb{R}^3$ . We set  $A(0) = A$ ,  $C(0) = C$ ,  $B(0) = B$ . We have now to show that for any  $m$ , we have

$$A(m) = A, B(m) = B, C(m) = 2mA + C. \quad (4.66)$$

It is readily seen that (4.66) implies lemma 4 from the equivalence properties given in proposition 1. The idea to prove (4.66) is to show that all the levels of internal energy are sufficiently coupled. We prove (4.66) inductively on  $m$ . Assume the result holds for  $m \leq m_0$ . Let  $m$  be an integer in  $[0, m_0 + 2]$  such that  $m$  can be written as a sum of three squares of integers  $a, b, c$  that is  $m = a^2 + b^2 + c^2$ . In all the cases the largest possible  $m$  is greater or equal to 2. Consequently for any  $m_0$  we have  $m_0 + 2 - m \leq m_0$ . The inelastic collisions can be characterized by a relation on the multiindices  $(i, j)$  and  $(k, l)$  for the velocities and internal energies as defined in (4.47). Among the many collisions processes, we shall consider the following cases:

$$i = (0, 0, 0, m_0 + 1), j = (2, 0, 0, 0) \rightarrow k = (a + 1, b, c, m_0 + 2 - m), l = (1 - a, -b, -c, 0)$$

$$i = (2, 0, 0, m_0 + 1), j = (0, 0, 0, 0) \rightarrow k = (a + 1, b, c, m_0 + 2 - m), l = (1 - a, -b, -c, 0)$$

$$i = (0, 2, 0, m_0 + 1), j = (0, 0, 0, 0) \rightarrow k = (a, b + 1, c, m_0 + 2 - m), l = (-a, 1 - b, -c, 0)$$

$$i = (0, 0, 2, m_0 + 1), j = (0, 0, 0, 0) \rightarrow k = (a, b, c + 1, m_0 + 2 - m), l = (-a, -b, 1 - c, 0)$$

$$i = (4, 0, 0, m_0 + 1), j = (2, 0, 0, 0) \rightarrow k = (a + 3, b, c, m_0 + 2 - m), l = (3 - a, -b, -c, 0)$$

One can verify that, for these collision processes, we have  $\mathcal{A}_{ij}^{kl} \neq 0$ , which implies that the invariants of collision satisfy

$$\varphi_i(m_0 + 1) + \varphi_j(0) = \varphi_k(m_0 + 2 - m) + \varphi_l(0).$$

By the use of the result (4.66) for  $m_0 + 2 - m$ , we have then the set of equations

$$\begin{cases} C(m_0 + 1) = 2(m_0 + 1)A + C \\ 4A(m_0 + 1) + 2B_1(m_0 + 1) = 4A + 2B_1 \\ 4A(m_0 + 1) + 2B_2(m_0 + 1) = 4A + 2B_2 \\ 4A(m_0 + 1) + 2B_3(m_0 + 1) = 4A + 2B_3 \\ 16A(m_0 + 1) + 4B_1(m_0 + 1) = 16A + 4B_1 \end{cases}$$

for which the unique solution is

$$\begin{cases} C(m_0 + 1) = 2(m_0 + 1)A + C \\ A(m_0 + 1) = A \\ B_1(m_0 + 1) = B_1 \\ B_2(m_0 + 1) = B_2 \\ B_3(m_0 + 1) = B_3. \end{cases}$$

This proves (4.66) at order  $m_0 + 1$ . Using the integrability requirements, that is

$$\sum_{i \in L} f_i^\infty e_i^{\frac{\delta}{2}-1} < +\infty,$$

we necessarily have  $A < 0$ .  $\square$

As for the first model, we can bound the velocity-energy domain by a straightforward adaptation of the analysis presented for the first model. We mention that if the resulting set for the multiindices  $i$  is of the form

$$\{i \in L, i_1^2 + i_2^2 + i_3^2 \leq M, i_4 \leq N\}$$

or

$$\{i \in L, i \sup_{p=1,3} i_p \leq M, i_4 \leq N\},$$

with  $M \geq 4$ , then all the properties of the collision operators are still satisfied except for the sign of the coefficient of  $A$  in the equilibrium state (see the discussion of this problem in the first case).

## References

1. G. and A. Bird, *Molecular Gas Dynamics*, (Clarendon Press, 1976).
2. A.V. Bobylev, A. Palczewski and J. Schneider, *On approximation of the Boltzmann equation by discrete velocity models*, C.R. Acad. Sci. serie I, (1995) 639-644.
3. C. Borgnakke, P.S. Larsen *Statistical model for Monte-Carlo simulation of polyatomic gas mixtures*, J. Comp. Phys., 18 (1975) 405-420.
4. J. F. Bourgat, L. Desvillettes, P. Le Tallec, B. Perthame, *Microreversible collisions for polyatomic gases and Boltzmann's theorem*, Eur. J. Mech. B fluids, (1994).
5. C. Buet, *A discrete-velocity scheme for the boltzmann operator of rarefied gas dynamics*, Trans. Th. Stat. Phys., 25 (1996) 33-60.
6. C. Cercignani, *The Boltzmann Equation and its Applications*, (Springer, 1988).
7. L. Desvillettes, *Sur un modele de type Borgnakke-Larsen conduisant a des lois d'energie non-linéaires en température pour les gaz parfaits polyatomiques*, Actes du workshop du GDR SPARCH, (1995).

8. W.Duke, *Hyperbolic distribution problems and half integerweight mass forms*, Invent. Math., 92 (1988).
9. R. GATIGNOL, *Théorie cinétique des gaz à répartitions discrètes de vitesses*, Lect. Notes in Phys. (Springer, 1975), Vol 36.
10. D. Goldstein, B. Sturtevant and J. E. Broadwell, *Investigations of the Motion of Discrete-Velocity Gases*, in "Rarefied Gas Dynamics: Theoretical and Computational Techniques", E. P. Muntz, D. P. Weaver and D. H. Campbell (eds), Progress in Astronautics and Aeronautics, 118, AIAA, Washington DC, (1989).
11. D. B. Goldstein, *Discrete-Velocity collision dynamics for polyatomic molecules*, Phys. Fluids A4, (1992) 1831-1839.
12. F. Gropengiesser, H. Neunzert, J. Struckmeier, *Computational methods for the Boltzmann equation*. Venice 1989: The state of Art in Appl. and Industrial math., eds. R. Spigler, Kluwer acad. publ., (1990).
13. G.H. Hardy and E.M. Wright, *An introduction to the number theory*, (Clarendon Press, 1938).
14. T. Inamuro and B. Sturtevant, *Numerical Study of Discrete-Velocity Gases*, Phys. Fluids A2, (1990) 2196-2203.
15. H. Iwaniec, *Fourier coefficients of modular forms of half integral weight*, Invent. Math., 87, (1987).
16. K. Nanbu, *Direct simulation schemes derived from the Boltzmann equation*, J.Phys., Japan 49, (1980).
17. K. Nanbu, *Model kinetic equation for the distribution of discretized internal energy*, Math. Mod. Meth. Applied Sci., (1992).
18. F. Rogier and J. Schneider, *A direct Method for solving the Boltzmann Equation*, Transp. Th. Stat. Phys., (1994).
19. J. Schneider, *Une méthode déterministe pour la résolution de l'équation de Boltzmann*, Ph.D thesis, University Paris 6, (1993).
20. Z.Tan and P.L.Varghese *The  $\Delta - \varepsilon$  method for the Boltzmann equation*, J.Comp. Phys., 110 (1994).

# Fast Algorithms for Numerical, Conservative, and Entropy Approximations of the Fokker–Planck–Landau Equation<sup>1</sup>

C. Buet,\* S. Cordier,† P. Degond,‡ and M. Lemou‡

\*CEA-CEL-V 94195 Villeneuve Saint Georges Cedex, France; †Laboratoire d'Analyse Numérique, CNRS-URA 189, Université de Paris VI, 4, Place Jussieu 75252 Paris cedex 05, France; ‡MIP, CNRS UMR 5640, UFR MIG Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex, France

Received June 26, 1996; revised February 5, 1997

We present fast numerical algorithms to solve the nonlinear Fokker–Planck–Landau equation in 3D velocity space. The discretization of the collision operator preserves the properties required by the physical nature of the Fokker–Planck–Landau equation, such as the conservation of mass, momentum, and energy, the decay of the entropy, and the fact that the steady states are Maxwellians. At the end of this paper, we give numerical results illustrating the efficiency of these fast algorithms in terms of accuracy and CPU time. © 1997 Academic Press

## 1. INTRODUCTION: THE FOKKER–PLANCK–LANDAU EQUATION

The Fokker–Planck–Landau (FPL) equation is used for the description of binary collisions between charged particles, for which the interaction potential is the long-range Coulomb interaction.

We restrict ourselves to a single-species plasma since the methods can easily be extended to the multispecies case (see Remark 3.3). The present algorithms are based on the discretization of the FPL operator given in [5] the main features of which are summarized in Section 2.

We denote by  $f(x, v, t)$  the distribution function, a solution of the scaled FPL equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q(f, f), \quad (1.1)$$

where  $Q(f, f)$  is the FPL collision operator:

$$Q(f, f) = \nabla_v \cdot \left( \int_{\mathbb{R}^3} \Phi(v - v_*) ((\nabla_v f) f_* - (\nabla_{v_*} f) f) dv_* \right) \quad (1.2)$$

with

$$f = f(x, v, t), \quad f_* = f(x, v_*, t), \\ \nabla_v f = \nabla_v f(x, v, t), \quad \nabla_{v_*} f = \nabla_{v_*} f(x, v_*, t)$$

and  $\Phi(v)$  is the  $3 \times 3$  matrix:

$$\Phi(v) = |v|^{\gamma+2} S(v), \quad S(v) = I_3 - \frac{v \otimes v}{|v|^2}. \quad (1.3)$$

$S(v)$  is the orthogonal projector onto the plane orthogonal to  $v$ ;  $\gamma$  is a real parameter which leads to the usual classification in hard potentials ( $\gamma > 0$ ), Maxwellian molecules ( $\gamma = 0$ ), or soft potentials ( $\gamma < 0$ ). This latter case involves the Coulombian case itself (i.e.,  $\gamma = -3$ ). The present numerical analysis is concerned with the physically interesting case (the Coulombian one) and the Maxwellian case ( $\gamma = 0$ ) which enable us to compare the numerical results with exact solutions [6].

As is well known in the physics literature and is mathematically established by the work of Arsenev and Buryac [13] and Desvillettes [14], the FPL collision operator is the limit of the Boltzmann operator for a sequence of a scattering cross section which converges in a convenient sense to a delta function at zero scattering-angle. In the case of a Coulomb interaction, Degond and Lucquin-Desreux obtained the FPL collision operator as the leading term of the cutoff Boltzmann operator when the parameter of the cutoff tends to zero [16]. Concerning the existence of solutions, Arsenev and Peskov have established the existence of weak-solutions for a short time in the case of the spatially homogeneous FPL equation for the Coulomb potential.

The algebraic structure of the FPL operator is similar to that of the Boltzmann operator. This leads to well-known physical properties such as the decay of the entropy, the conservation of mass, momentum, and energy, and the characterization of the equilibrium states by Maxwellians.

<sup>1</sup> This work was partially completed with CEA-CEL-V under Contract W 003 772/216.

Indeed, these properties can easily be shown on the weak form of the FPL operator,

$$\begin{aligned} & \int_{\mathbb{R}^3} Q(f, f)(v) \psi(v) dv \\ &= -\frac{1}{2} \int \int_{\mathbb{R}^3 \times \mathbb{R}^3} f f_* (\nabla_v \psi - \nabla_{v_*} \psi)^T \Phi(v - v_*) (\nabla_v (\ln f) \\ & \quad - \nabla_{v_*} (\ln f)) dv dv_* \end{aligned} \quad (1.4)$$

for any smooth test function  $\psi$ . From this duality relation, it is an easy matter to check that the only functions  $\psi$  such that for all  $f$ ,  $\int Q(f, f) \psi dv = 0$ , are linear combinations of 1,  $v$ , and  $|v|^2$  (conservation of mass, momentum, and energy).

Furthermore, letting  $\psi = \ln(f)$  in (1.4) leads to the entropy inequality (H-theorem):

$$\int_{\mathbb{R}^3} Q(f, f)(v) \ln(f(v)) dv \leq 0. \quad (1.5)$$

The equilibrium distribution functions, i.e., the functions  $f$  such that  $Q(f, f) = 0$  are Maxwellians,

$$M_{\mathcal{N}, u, T}(v) = \frac{\mathcal{N}}{(2\pi v_{\text{th}}^2)^{3/2}} \exp\left(-\frac{|v - u|^2}{2v_{\text{th}}^2}\right), \quad (1.6)$$

where  $\mathcal{N}$  is the density of particles and  $v_{\text{th}}$  is the thermal velocity of the gas which depends on its temperature  $T$  through the relation:  $v_{\text{th}} = \sqrt{kT/m}$ , where  $k$  is the Boltzmann constant and  $m$  is the mass of the particle.

In this paper, we are concerned with numerical approximations of the spatially homogeneous FPL equation in the whole 3D velocity space. The starting point of this work is the discretization of the FPL operator given in [5]. Alternate methods are finite difference schemes that have been investigated in [9] in the isotropic case and in [12, 8, 10] for the cylindrically symmetric problems. We also refer to Larroche [17] for a mass conserving finite volume scheme. Recently, conservative and entropic discretizations of axisymmetric FPL operators are investigated in [18]. In [4, 5] a numerical discretization in three-dimensional velocity space that satisfies discrete analogues of the above-mentioned properties is presented and is summarized in the next section. A bibliography on previous works on such methods can be found in [4, 5].

The discretization [4, 5] in three-dimensional velocity space satisfies all properties required by the physical nature of the problem. Unfortunately a direct numerical implementation of this method is very expensive. Its cost is of order  $N^2$  when  $N$  is the number of the discrete velocity points. Our first approach for reducing this cost is the use of the sublattices method following the works by Buet [1–3] on the Boltzmann equation. This leads to a cost  $N^2/a^3$ ,

where  $a$  is the sublattice mesh size. The second strategy is an adaptation of multigrid methods to the FPL equation. It leads to a cost of the order of  $N \ln N$ .

The outline of the paper is as follows: In Section 2, we review the basis of the conservative discretizations introduced in [5]. Section 3 is devoted to a symmetrized version of the method [5] while Section 4 is concerned with the fast algorithms that are themselves the sublattices and the multigrid algorithms. In Section 5, we give numerical results.

## 2. A CLASS OF ENTROPY-DECREASING SCHEMES

By a standard splitting algorithm, we may restrict ourselves to the space-homogeneous FPL equation

$$\frac{\partial f}{\partial t} = Q(f, f), \quad f|_{t=0} = f_0(v), \quad (2.1)$$

where  $Q(f, f)$  is given by (1.2) and  $f_0$  is the initial data. We introduce a regular discretization of  $\mathbb{R}^3$ ,  $v_i = i\Delta v$ ,  $i = (i^1, i^2, i^3) \in \mathbb{Z}^3$ , and denote by  $\bar{f}_i$  an approximation of  $f(v_i)$ . Let  $D$  be a finite-difference operator that approximates the usual gradient operator  $\nabla$  at least up to the first order, and let  $D^*$  be its formal adjoint. For any “test sequence,”  $\bar{\psi} = (\bar{\psi}_i)_{i \in \mathbb{Z}^3}$ ,  $D\bar{\psi}$  is a sequence  $(D\bar{\psi})_{i \in \mathbb{Z}^3}$  of vectors of  $\mathbb{R}^3$ ,

$$(D\bar{\psi})_i = ((D^1\bar{\psi})_i, (D^2\bar{\psi})_i, (D^3\bar{\psi})_i) \in \mathbb{R}^3, \quad (2.2)$$

where the components  $(D^s\bar{\psi})_i$ ,  $s = 1, 2, 3$ , approximates the partial derivatives  $(\partial\bar{\psi}/\partial x_s)(v_i)$ . Such an operator is of the form

$$(D\bar{\psi})_i = \sum_{k \in \mathbb{Z}^3} a_k \bar{\psi}_{i+k}, \quad (2.3)$$

where the vectors  $a_k = (a_{k,1}, a_{k,2}, a_{k,3}) \in \mathbb{R}^3$  satisfy

$$\sum_{k \in \mathbb{Z}^3} a_k = 0, \quad \sum_{k \in \mathbb{Z}^3} a_{k,s} k_r \Delta v = \delta_{sr}, \quad (2.4)$$

$k_r$ , being the  $r$ th component of  $k$ . Conditions (2.4) state that  $D$  coincides with the exact gradient for constant or linear functions, or equivalently, that  $D$  is an approximation of  $\nabla$  at least up to the first order. The formal adjoint  $D^*$  or  $D$  is given by

$$(D^*\bar{\psi})_i = \sum_{k \in \mathbb{Z}^3} a_k^* \bar{\psi}_{i+k} \quad (2.5)$$

with

$$a_k^* = a_{-k} \quad \forall k \in \mathbb{Z}^3. \quad (2.6)$$

Note that  $D^*$  is an approximation of  $-\nabla$ .

The approximation  $\bar{Q}(\bar{f}, \bar{f})_i$  of  $Q(f, f)(v_i)$  is defined for any test sequence  $\psi$  by

$$\sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i \psi_i = -\frac{1}{2} \sum_{(i,j) \in \mathbb{Z}^3 \times \mathbb{Z}^3} \bar{f}_i \bar{f}_j ((D\bar{\psi})_i - (D\bar{\psi})_j)^T \Phi(v_i - v_j) ((D(\ln \bar{f}))_i - (D(\ln \bar{f}))_j) \Delta v^3. \quad (2.7)$$

The scheme defined by (2.7) decays the entropy:

$$\sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i (\ln \bar{f})_i \leq 0. \quad (2.8)$$

A collisional invariant is defined as a sequence  $\bar{\psi}_i$  such that

$$\sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i \bar{\psi}_i = 0 \quad \forall (\bar{f}_i)_{i \in \mathbb{Z}^3}, \quad (2.9)$$

or equivalently from (2.7) such that

$$(D\bar{\psi})_i - (D\bar{\psi})_j \in \text{Ker}(\Phi(v_i - v_j)) \quad \forall i, j \in \mathbb{Z}^3. \quad (2.10)$$

A discrete equilibrium distribution function (i.e., a function  $\bar{f}_i$  such that  $\bar{Q}(\bar{f}, \bar{f})_i = 0$ ) is clearly, from (2.7), such that  $(\ln \bar{f})_i$  is a collisional invariant (i.e., satisfies (2.10)). It is proved in [5] that (2.10) is equivalent to the existence of a real number  $\lambda$ , independent of  $i$  and  $j$ :

$$(D\bar{\psi})_i - (D\bar{\psi})_j = \lambda(i - j) \Delta v \quad \forall i, j \in \mathbb{Z}^3. \quad (2.11)$$

However, nothing more can be said unless specifying the discrete differential operator  $D$ . This is shown on two simplest cases (we refer to [5] for details).

Case 1, the right (resp. left) uncensored operator  $D = D_+$  (resp  $D = D_-$ ) defined by

$$(D_+^s \bar{\psi})_i = \frac{\bar{\psi}_{i+e_s} - \bar{\psi}_i}{\Delta v}, \quad s = 1, 2, 3, \quad (2.12)$$

$$\left( \text{resp. } (D_-^s \bar{\psi})_i = \frac{\bar{\psi}_i - \bar{\psi}_{i-e_s}}{\Delta v}, \quad s = 1, 2, 3 \right),$$

where  $e_s$  is the  $s$ th vector of the canonical basis of  $\mathbb{R}^3$ .

Case 2, the centered operator  $D = D_c$  defined by

$$(D_c^s \bar{\psi})_i = \frac{\bar{\psi}_{i+e_s} - \bar{\psi}_{i-e_s}}{2\Delta v}, \quad s = 1, 2, 3. \quad (2.13)$$

The operators  $D_+$  and  $D_-$  are clearly first order, while  $D_c$  is second order. For these two cases, we have the following results.

LEMMA 2.1 (Uncentered case). (i) *The collisional invariants (i.e., the sequences  $\bar{\psi}_i$  such that (2.11) holds) are linear combinations of 1,  $v_i$ , and  $|v_i|^2$ .*

(ii) *The equilibrium distribution functions are the discrete Maxwellians.*

(iii) *Conservation of mass, momentum, and energy hold.*

LEMMA 2.2 (Centered case). (i) *The collisional invariants are linear combinations of  $v_i$ ,  $|v_i|^2$ , and of the following eight sequences  $\bar{\chi}_{i_*}$ , labelled by  $i_* \in \{0, 1\}^3$ , defined by*

$$(\bar{\chi}_{i_*})_i = \begin{cases} 1 & \text{if } i^k \equiv i_*^k \pmod{2} \quad \forall k \in \{1, 2, 3\}, \\ 0 & \text{otherwise,} \end{cases} \quad (2.14)$$

where  $i^k$  denotes the  $k$ th component of  $i \in \mathbb{Z}^3$ .

(ii) *The discrete equilibrium functions are exponentials of the above described collisional invariants.*

(iii) *Conservation of mass, momentum, and energy hold. But seven other independent spurious conservation laws hold associated with  $(\bar{\chi}_{i_*})_{i_* \in \{0, 1\}^3}$ ,*

$$\sum_{i \in \mathbb{Z}^3} \bar{Q}(\bar{f}, \bar{f})_i (\bar{\chi}_{i_*})_i = 0 \quad \forall i_* \in \{0, 1\}^3. \quad (2.15)$$

(Note that  $1 = \sum_{i_* \in \{0, 1\}^3} \bar{\chi}_{i_*}$ , so that conservation of mass can be deduced from the eight conservation laws (2.15).)

The use of the centered discrete difference operator leads to nonphysical equilibrium states (i.e., non-Maxwellian functions). On the other hand, the use of the uncensored discrete operator destroys the symmetry of the problem and does not give satisfactory results. To overcome this problem, we introduce a symmetrization of the discrete FPL operator, based on the averaging of the uncensored discretizations in the various directions of coordinates.

### 3. SYMMETRIZATION OF THE UNCENTERED DISCRETE DIFFERENTIATION

By combining “upwind” and “downwind” uncensored differences in the various direction of coordinates, we define eight uncensored difference operators denoted by  $D_\varepsilon$ , for  $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \{-1, 1\}^3$  as

$$(D_\varepsilon f)_i = \frac{1}{\Delta v} \begin{pmatrix} \varepsilon_1(f_{i+\varepsilon_1 e_1} - f_i) \\ \varepsilon_2(f_{i+\varepsilon_2 e_2} - f_i) \\ \varepsilon_3(f_{i+\varepsilon_3 e_3} - f_i) \end{pmatrix}. \quad (3.1)$$

At variance, there is only one centered operator denoted by  $D_c$ :

$$(D_c f)_i = \frac{1}{2\Delta v} \begin{pmatrix} f_{i+e_1} - f_{i-e_1} \\ f_{i+e_2} - f_{i-e_2} \\ f_{i+e_3} - f_{i-e_3} \end{pmatrix}. \quad (3.2)$$

We denote by  $Q^\varepsilon$  and  $Q^c$  the corresponding discretized FPL operator defined by the duality relation (2.7). In this section we introduce the symmetrization of the FPL operator  $Q^{us}$  obtained by taking the average of the operators  $Q^\varepsilon$  over all  $\varepsilon \in \{-1, 1\}^3$ :

$$Q^{us} = \frac{1}{8} \sum_{\varepsilon \in \{-1, 1\}^3} Q^\varepsilon. \quad (3.3)$$

We also introduce a new difference operator  $\Delta$  as

$$(\Delta f)_i = \frac{1}{(\Delta v)^2} \begin{pmatrix} f_{i+e_1} + f_{i-e_1} - 2f_i \\ f_{i+e_2} + f_{i-e_2} - 2f_i \\ f_{i+e_3} + f_{i-e_3} - 2f_i \end{pmatrix} \quad (3.4)$$

and denote by  $Q^\Delta$  the operator defined by the duality relation

$$\sum_{i \in \mathbb{Z}^3} Q_i^\Delta \psi_i = -\frac{1}{2} \sum_{(i,j) \in \mathbb{Z}^3 \times \mathbb{Z}^3} \bar{f}_i \bar{f}_j ((\Delta \bar{\psi})_i - (\Delta \bar{\psi})_j)^T \Phi^\Delta(v_i - v_j) \quad (3.5)$$

$$((\Delta(\ln \bar{f}))_i - (\Delta(\ln \bar{f}))_j) \Delta v^3,$$

where we have set

$$\Phi^\Delta(v) = |v| \text{Diag}(|v|^2 - v_1^2, |v|^2 - v_2^2, |v|^2 - v_3^2) \quad (3.6)$$

and  $\text{Diag}(x, y, z)$  denotes the diagonal matrix whose diagonal elements are  $x$ ,  $y$ , and  $z$ .

With these notations, a simple calculation yields the following result.

**PROPOSITION 3.1.** *We have*

$$\sum_{i \in \mathbb{Z}^3} Q_i^{us} \psi_i = \sum_{i \in \mathbb{Z}^3} Q_i^c \psi_i + \frac{\Delta v^2}{4} \sum_{i \in \mathbb{Z}^3} Q_i^\Delta \psi_i. \quad (3.7)$$

The operator  $Q^{us}$  is the sum of two contributions: the first one is the centered operator  $Q^c$  and the second one is a sort of viscosity term which serves to eliminate all spurious collisional invariants that may be generated by

the first contribution (see Lemma 2.2)). The numerical implementation of  $Q^{us}$  in the form (3.7) is clearly less expensive than the one in the form (3.3).

On the other hand, it is possible to reduce the computational cost of the viscosity term (3.5) by replacing the sum over  $(i, j) \in \mathbb{Z}^3 \times \mathbb{Z}^3$  in (3.5) by a sum over  $(i, j) \in \mathbb{Z}^3 \times \mathbb{Z}^3$  with  $|i - j| \leq \sqrt{2}$ . In the following proposition we show that this reduction does not affect the conservation properties and does not generate any spurious collisional invariant. We denote by  $Q^{usr}$  the FPL operator using this reduction procedure, i.e.,

$$\sum_{i \in \mathbb{Z}^3} Q_i^{usr} \psi_i = \sum_{i \in \mathbb{Z}^3} Q_i^c \psi_i + \frac{\Delta v^2}{4} \sum_{i \in \mathbb{Z}^3} Q_i^{\Delta r} \psi_i \quad (3.8)$$

with

$$\sum_{i \in \mathbb{Z}^3} Q_i^{\Delta r} \psi_i = -\frac{1}{2} \sum_{\substack{i, j \in \mathbb{Z}^3 \times \mathbb{Z}^3 \\ |i-j| \leq \sqrt{2}}} \bar{f}_i \bar{f}_j \times ((\Delta \bar{\psi})_i - (\Delta \bar{\psi})_j)^T \Phi^\Delta(v_i - v_j) \quad (3.9)$$

$$((\Delta(\ln \bar{f}))_i - (\Delta(\ln \bar{f}))_j) \Delta v^3$$

and we have

**PROPOSITION 3.2.** *For the discrete FPL operator defined by formulas (3.8) and (3.9), the collisional invariants are of the form*

$$\psi_i = A|v_i|^2 + \langle B, v_i \rangle + C. \quad (3.10)$$

*Proof.* By Lemma 2.2 the collisional invariants for the centered difference operator have the form

$$\psi_i = A|v_i|^2 + \langle B, v_i \rangle + \sum_{k \in \{0, 1\}^3} C_k (\chi_k)_i, \quad (3.11)$$

where  $C_k = C_{k_1 k_2 k_3}$  are arbitrary coefficients and  $\chi_k$  is defined in Lemma 2.2. But the viscosity term (3.9) gives the following additional relations:

$$(\Delta \bar{\psi})_i - (\Delta \bar{\psi})_j \in \text{Ker } \Phi^\Delta(i - j) \quad (3.12)$$

for all  $i, j$  such that  $|i - j| \leq \sqrt{2}$  which implies

$$[|i - j|^2 - (i_k - j_k)^2][C_{\bar{i}+e_k} - C_{\bar{j}+e_k} - (C_{\bar{i}} - C_{\bar{j}})] = 0 \quad (3.13)$$

for all  $k \in \{1, 2, 3\}$  and for all  $i, j$  such that  $|i - j| \leq \sqrt{2}$ . The notation  $\bar{i}$  denotes the class of  $i$  modulo 2, that is the vector whose components are the classes of  $i$ 's components modulo 2 ( $\bar{i} \in \{0, 1\}^3$ ). To simplify we suppose that  $i = (0, 0, 0)$  and denote by  $(e_1, e_2, e_3)$  the canonical basis of  $\mathbb{R}^3$ .

For each fixed  $k \in \{1, 2, 3\}$  we choose  $l \in \{1, 2, 3\}$  such that  $k \neq l$  and take  $j = e_l$ , we get by (3.13)

$$C_{e_k} - C_{\overline{e_l + e_k}} - (C_{000} - C_{e_l}) = 0 \quad (3.14)$$

for all  $k \neq l$ . Now, taking  $j = e_k + e_l$  with  $k \neq l$  we obtain

$$C_{e_k} - C_{e_l} - (C_{000} - C_{\overline{e_k + e_l}}) = 0. \quad (3.15)$$

Combining these last two relations, we get  $C_{e_k} = C_{000}$  for all  $k \in \{1, 2, 3\}$ . Inserting this in (3.15) we have also  $C_{\overline{e_k + e_l}} = C_{000}$  for all  $k, l \in \{1, 2, 3\}$ .

Finally, by writing (3.13) for  $k = 1$  and  $j = e_2 + e_3$ , we obtain  $C_{111} = C_{000}$  which concludes the proof thanks to relation  $1 = \sum_{k \in \{0,1\}^3} \chi_k$ . ■

*Remark 3.3.* Extension to the multispecies Fokker-Planck equation. We extend the above method to two species of particles denoted by indices  $a$  and  $b$ , respectively. The distribution functions of these two species satisfy the following system of homogeneous FPL equations:

$$\frac{\partial f_a}{\partial t} = Q_a(v_a), \quad \frac{\partial f_b}{\partial t} = Q_b(v_b), \quad (3.16)$$

where we have set (with  $\alpha = a$  or  $b$  and  $\beta = b$  or  $a$ , respectively):

$$Q_\alpha(v_\alpha) = \frac{1}{m_\alpha} \nabla_{v_\alpha} \cdot \int_{\mathbb{R}^3} \Phi(v_\alpha - v_\beta) \left( \frac{1}{m_\alpha} \nabla_{v_\alpha} f_\alpha f_\beta - \frac{1}{m_\beta} \nabla_{v_\beta} f_\beta f_\alpha \right) dv_\beta. \quad (3.17)$$

Let  $\psi_\alpha = \psi_\alpha(v_\alpha)$  be two test functions (for  $\alpha = a$  and  $b$ ). We have

$$\begin{aligned} & \int_{\mathbb{R}^3} Q_a(v_a) \psi_a dv_a + \int_{\mathbb{R}^3} Q_b(v_b) \psi_b dv_b \\ &= - \int_{\mathbb{R}^3 \times \mathbb{R}^3} f_a f_b \times \left( \frac{1}{m_a} \nabla_a \psi_a - \frac{1}{m_b} \nabla_b \psi_b \right)^T \\ & \quad \Phi(v_a - v_b) \left( \frac{1}{m_a} \nabla_a \ln f_a - \frac{1}{m_b} \nabla_b \ln f_b \right). \end{aligned} \quad (3.18)$$

We consider different mesh sizes for the two species and define two regular discretizations of  $\mathbb{R}^3$  according to:

$$v_i^a = i \Delta v_a, v_i^b = i \Delta v_b \quad \text{for } i \in \mathbb{Z}^3, \quad (3.19)$$

and the associated finite difference operators as:

$$(D_\alpha \psi)_i = \frac{1}{\Delta v_\alpha} \sum_{k \in \mathbb{Z}^3} a_k \psi_{i+k}. \quad (3.20)$$

We also consider the operators  $D_{\alpha,\varepsilon}$  and  $D_{\alpha,c}$  defined by (3.1) and (3.2) with the mesh size  $\Delta v_\alpha$ . With these notations, it is an easy matter to see that we have

**PROPOSITION 3.4.** *For the centered difference operators  $D_{\alpha,c}$ , we have the conservation of mass, momentum, and energy. However, for the uncentered difference operators  $D_{\alpha,\varepsilon}$  the conservation of energy holds if and only if  $\Delta v_a = \Delta v_b$ .*

## 4. FAST ALGORITHMS

### 4.1. Deterministic Schemes: Sublattices Methods

The computational complexity is of order  $N^2$  which is much too big for a practical use of the discrete FPL operators. To reduce this cost, a first strategy is to use sublattices as it was done for the Boltzmann collision operator [1]. We present a brief description of the method and show how to preserve the physical properties. We also show how to design the algorithm in order to avoid spurious collisional invariants. In this section we deal with the uncentered discrete difference operator although the following results remain valid if we use the symmetrized operator given by (3.3).

For  $a \in \mathbb{Z}$ ,  $a \geq 2$ , we define the discrete operator  $Q_i[a]$  by the duality relation

$$\sum_{i \in \mathbb{Z}^3} Q_i[a] \psi_i = -\frac{1}{2} \sum_{i=j[a]} \bar{f}_i \bar{f}_j \quad (4.1)$$

$$((D\bar{\psi})_i - (D\bar{\psi})_j)^T \Phi(v_i - v_j) ((D(\ln \bar{f}))_i - (D(\ln \bar{f}))_j) \Delta v^3,$$

where the definition of  $i \equiv j[a]$  means that  $i - j$  is a multiple of  $a$ . The following result makes the collisional invariants for this discretization precise.

**PROPOSITION 4.1.** *Let  $A$  and  $B$  real constants, and let  $C_i$  only depending on the class of  $i$  modulo  $a$  (i.e.,  $C_i = C_{\bar{i}}$  with  $\bar{i}$  the class of  $i$  modulo  $a$ ); then*

$$\bar{\psi}_i = A|v_i|^2 + \langle B, v_i \rangle + C_{\bar{i}}$$

are collisional invariants generated by the discrete operator  $Q_i[a]$  defined by (4.1).

*Proof.*  $\bar{\psi}$  is a collisional invariant for the operator  $Q_i[a]$  if and only if

$$(D\bar{\psi})_i - (D\bar{\psi})_j \in \text{Ker } \Phi(i - j) \quad (4.2)$$



for all  $i, j$  such that  $i \equiv j[a]$ . Replacing  $\bar{\psi}_i$  successively by  $|v_i|^2$ ,  $v_i$ , and  $C_i$ , we easily obtain the desired result. ■

Now we shall modify this method in order to preserve the Maxwellians as unique possible equilibrium states: Let  $a$  and  $b$  two mutually prime integers, i.e., such that  $a \wedge b = 1$ , where  $a \wedge b$  is the greatest common divisor of  $a$  and  $b$ . We consider the two corresponding operators  $Q_i[a]$  and  $Q_i[b]$  defined by the duality formula (4.1).

We set

$$Q_i[a, b] = \frac{1}{2}(Q_i[a] + Q_i[b]) \quad \text{for all } i \in \mathbb{Z}^3. \quad (4.3)$$

We have the following result.

**PROPOSITION 4.2.** *If we use the uncentered difference operator, then the collisional invariants of the discrete operator given by formula (4.3) are the linear combinations of mass, momentum, and energy.*

*Proof.* A collisional invariant of the discrete operator  $Q_i[a, b]$  must be a collisional invariant for both  $Q_i[a]$  and  $Q_i[b]$ . Therefore, if  $\bar{\psi}$  is a collisional invariant generated by  $Q_i[a, b]$  then

$$(D\bar{\psi})_i - (D\bar{\psi})_j \in \text{Ker } \Phi(i - j) \quad (4.4)$$

for all  $i, j$  such that  $i \equiv j[a]$  or  $i \equiv j[b]$ .

Therefore, for  $i$  and  $k \in \mathbb{Z}^3$ , we have

$$(D\bar{\psi})_{i+ak} - (D\bar{\psi})_i = \lambda(i, k)ak \quad (4.5)$$

$$(D\bar{\psi})_{i+bk} - (D\bar{\psi})_i = \mu(i, k)bk \quad (4.6)$$

with  $\lambda(i, k), \mu(i, k) \in \mathbb{R}$ .

Thus, from (4.5), we can write two relations,

$$(D\bar{\psi})_{i+ak} - (D\bar{\psi})_{i+al} = \lambda(i + la, k - l)a(k - l)$$

$$(D\bar{\psi})_{i+ak} - (D\bar{\psi})_{i+al} = \lambda(i, k)ak - \lambda(i, l)al,$$

for all  $i, k, l \in \mathbb{Z}^3$  and easily obtain:  $\lambda(i, k) = \lambda(i, l)$  for all  $i, k, l \in \mathbb{Z}^3$ , which means that  $\lambda(i, k)$  is independent of  $k$ , and the same is true for  $\mu(i, k)$ . Let  $\lambda(i, k) = \lambda_i$ , and  $\mu(i, k) = \mu_i$ . Now by writing

$$(D\bar{\psi})_{i+abk} - (D\bar{\psi})_i = \lambda_i abk = \mu_i abk \quad \forall i, k \in \mathbb{Z}^3, \quad (4.7)$$

we obtain  $\lambda_i = \mu_i = \alpha_i$  for all  $i \in \mathbb{Z}^3$ . On the other hand, we have

$$\begin{aligned} 2a\alpha_i k &= (D\bar{\psi})_{i+2ak} - (D\bar{\psi})_i \\ &= (D\bar{\psi})_{i+2ak} - (D\bar{\psi})_{i+ak} + (D\bar{\psi})_{i+ak} - (D\bar{\psi})_i \\ &= a\alpha_{i+ak}k + a\alpha_i k \end{aligned}$$

which gives

$$\alpha_{i+ak} = \alpha_i \quad \forall i, k \in \mathbb{Z}^3$$

and this is also true for  $b$ :  $\alpha_{i+bk} = \alpha_i \quad \forall i, k \in \mathbb{Z}^3$ .

Now let  $i$  and  $j$  two arbitrary elements of  $\mathbb{Z}^3$ , since  $a \wedge b = 1$ , the Bezout identity gives the existence of two triples  $q, r \in \mathbb{Z}^3$  such that

$$i - j = aq + br. \quad (4.8)$$

Then we have

$$\alpha_j = \alpha_{j+aq} = \alpha_{j+aq+br} = \alpha_i$$

and, finally, we deduce that  $\alpha_i$  does not depend on  $i$  and then

$$(D\bar{\psi})_i - (D\bar{\psi})_j = \alpha(i - j) \quad (4.9)$$

for all  $i, j \in \mathbb{Z}^3$ . In the case of the uncentered difference operator  $D$ , this classically implies (see [5]) that  $\psi_i$  is a linear combination of the mass, the momentum, and the energy, and concludes the proof of Proposition 4.2. ■

## 4.2. Random Methods: Multigrid Algorithms

In this section we compute the discrete FPL collision operator by using a multigrid method with numerical integration of Monte-Carlo type. The computational complexity of this simulation is of order  $N \ln N$ , where  $N$  is the number of discrete velocity points. This approach takes its inspiration from the method of Greengard and Rokhlin [19]. In a subsequent work [7], a new method goes further in the adaptation of [19] to the FPL equation.

### 4.2.1. Description of the Method

To simplify the notations, we set

$$\begin{aligned} H(v, w) &= -\frac{1}{2}f(v)f(w)[\nabla_v \psi - \nabla_w \psi]^T \\ &\quad \Phi(v - w)[\nabla_v(\ln f) - \nabla_w(\ln f)]. \end{aligned} \quad (4.10)$$

We assume that the discrete velocity domain is a cube  $C_0$  of length 1 which contains  $N = (2^n)^3 = 8^n$  discrete points lying on a regular cubic lattice. The algorithm is the following:

- *Step 0.* We just write the FPL operator in a weak form:

$$\int_{C_0} Q(f, f)(v)\psi(v) dv = \int_{C_0 \times C_0} H(v, w) dv dw. \quad (4.11)$$

• *Step 1.* We split the cube  $C_0$  (the parent) into 8 regular boxes  $C_1^r$  (the children),  $r \in \{0, 1\}^3$ . Each box  $C_1^r$  is of length  $\frac{1}{2}$  and its center is

$$O_1^r = \left( \frac{1}{2^2} + \frac{r_1}{2}, \frac{1}{2^2} + \frac{r_2}{2}, \frac{1}{2^2} + \frac{r_3}{2} \right) \quad (4.12)$$

with  $r = (r_1, r_2, r_3) \in \{0, 1\}^3$ . we set  $I_1 = \{0, 1\}^3$ . Again, we do not do any numerical approximation; we only write

$$\int_{C_0} Q(f, f)(v) \psi(v) dv = \sum_{(r, r') \in I_1^2} \int_{C_1^r \times C_1^{r'}} H(v, w) dv dw. \quad (4.13)$$

• *Step  $k$  ( $k \geq 2$ ).* We denote by  $C_k^r$  the boxes of level  $k$  obtained after splitting each of those of level  $k-1$  (the parents) into eight regular boxes (the children). More precisely  $C_k^r$  is the box of length  $1/2^k$  and whose center is

$$O_k^r = \left( \frac{1}{2^{k+1}} + \frac{r_1}{2^k}, \frac{1}{2^{k+1}} + \frac{r_2}{2^k}, \frac{1}{2^{k+1}} + \frac{r_3}{2^k} \right) \quad (4.14)$$

with  $r = (r_1, r_2, r_3) \in I_k = \{0, 1, \dots, 2^k - 1\}^3$ . If  $C_{k-1}^R$  is the father of  $C_k^r$  then it is easy to see that

$$O_{k-1}^R = \left( \frac{1}{2^k} + \frac{R_1}{2^{k-1}}, \frac{1}{2^k} + \frac{R_2}{2^{k-1}}, \frac{1}{2^k} + \frac{R_3}{2^{k-1}} \right)$$

with

$$\begin{aligned} R_i &= \frac{r_i}{2} & \text{if } r_i \text{ is even} \\ R_i &= \frac{r_i - 1}{2} & \text{if } r_i \text{ is odd.} \end{aligned} \quad (4.15)$$

*Remark 4.3.* To obtain the children of  $C_k^r$ , we add to the center  $O_k^r$  the quantities

$$\frac{1}{2^{k+1}} (\varepsilon_1, \varepsilon_2, \varepsilon_3),$$

where

$$(\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \{-1, 1\}^3.$$

*Remark 4.4.* To obtain the nearest neighbours of  $C_k^r$ , we add to the center the quantities  $(1/2^k) (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ , where  $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \{-1, 0, 1\}^3$  (27 neighbours).

The box  $C_k^r$  will be said to be “well separated” from  $C_k^{r'}$  if and only if  $C_k^{r'}$  is a child of one of the nearest

neighbours of the parent of  $C_k^r$  and is not a nearest neighbour of  $C_k^r$ .

The notation *ws* means “well separated” and *nws* means “not well separated,” i.e., neighbours.

To simplify the algorithm we only describe the method for level 2:

At level 2, we split each box  $C_1^r$  of level 1 into 8 boxes and write

$$\begin{aligned} \int_{C_0} Q(f, f)(v) \psi(v) dv &= \sum_{\substack{(r, r') \in I_2^2 \\ r \text{ ws } r'}} \int_{C_2^r \times C_2^{r'}} H(v, w) dv dw \\ &+ \sum_{\substack{(r, r') \in I_2^2 \\ r \text{ nws } r'}} \int_{C_2^r \times C_2^{r'}} H(v, w) dv dw. \end{aligned} \quad (4.16)$$

If  $C_2^r$  is “well separated” from  $C_2^{r'}$ , then we replace the corresponding integral by a numerical approximation of Monte-Carlo type. In the second case we do not do any numerical approximation and pass to level 3. We repeat this process until step  $n$ , where we perform a Monte-Carlo approximation not only for the “well-separated” boxes but also for the nearest neighbours.

#### 4.2.2. Numerical Integration of Monte-Carlo Type

We assume that we are at a fixed level  $k$  and we want to approximate the expression

$$\int_{C_k^r \times C_k^{r'}} H(v, w) dv dw \quad (4.17)$$

when  $C_k^r$  and  $C_k^{r'}$  are “well separated.”

A direct approximation requires  $8^{2(n-k)}$  evaluations. But this leads to an amount of work proportional to  $N^2$ . Therefore, in order to have a cost of order  $N \ln N$  we must use only  $n_k = 8^{n-k}$  evaluations to approximate (4.17) such that after  $n_k$  iterations all the pairs  $(i, j) \in C_k^r \times C_k^{r'}$  were chosen. One way is the following.

Let  $\{1, 2, \dots, n_k\}$  be a numbering of the  $n_k$  elements of  $C_k^r$  or  $C_k^{r'}$  and let  $\pi$  be a randomly chosen permutation of  $\{1, 2, \dots, n_k\}$ . In the first time step we approximate (5.21) by a Monte Carlo quadrature formula using pairs  $(l, \pi(l)) \in C_k^r \times C_k^{r'}$ . In the second time step we use pairs  $(r, \pi^2(r))$ , etc. until covering the maximum number of possible pairs  $(l, l') \in C_k^r \times C_k^{r'}$ , i.e., until the number of iterations reaches the order of  $\pi$  in the group  $\mathcal{S}_{n_k}$  of permutations of the set  $\{1, 2, \dots, n_k\}$ . Therefore, for all the pairs  $(l, l') \in C_k^r \times C_k^{r'}$  to be chosen, the permutation  $\pi$  must be of order  $n_k$ . If such a choice of  $\pi$  is possible then after  $n_k$  iterations in time the integral will be well approximated since all pairs  $(l, l') \in C_k^r \times C_k^{r'}$  will have been chosen. For the next  $n_k$

time steps, we change randomly the permutation  $\pi$  and repeat the same process.

Finally, when  $C_k^r$  and  $C_k^{r'}$  are “well separated,” the Monte-Carlo approximation of (4.17) is given by

$$H_k^{r,r'} \stackrel{\text{def}}{=} 8^{n-k} (\Delta v)^6 \sum_{(i,\pi(i)) \in C_k^r \times C_k^{r'}} H(v_i, v_{\pi(i)}). \quad (4.18)$$

This way of choosing random collision pairs  $(v_i, v_{\pi(i)})$  was first suggested by Babovsky in [20] as an improvement of Nanbu’s scheme of the Boltzmann equation. In this work, we can also find a similar (Monte Carlo) quadrature formula which is proved to be consistent with the corresponding integral. This method can be applied to prove the consistency with (4.18) of the integral:

$$\int_{C_k^r \times C_k^{r'}} H(v, w) dv dw.$$

However, the use of the multigrid methodology introduces additional consistency errors that have not been analysed precisely. The method is probably more precise in the Coulomb case, where the collision cross section decays as the relative velocity increases than for other interactions forces with nondecreasing cross sections. Moreover, rigorous proofs of those statements are beyond the scope of this paper and will be the subject of future works.

Finally, we point out that the conservation properties and the decrease of the entropy are still satisfied. Indeed the expression (4.10) of  $H(v, w)$  vanishes when we replace  $\psi$  by 1,  $v_i$ , or  $v_i^2$  and becomes negative when we replace  $\psi$  by  $\ln f$ . On the other hand, a rigorous treatment of spurious collisional invariants is not clear and is not addressed here. It seems, however, that the multigrid method does not generate spurious collisional invariants (when we use the uncentered or the symmetrized uncentered difference operators), and the numerical tests confirm clearly this assertion.

## 5. NUMERICAL RESULTS

We present numerical tests of the above two methods (sublattices and multigrids) on two cases: the Maxwellian case ( $\gamma = 0$ ) and the Coulombian case ( $\gamma = -3$ ). In all these tests, we use a regular grid of size  $\Delta v$  in the velocity space which contains  $N = (2^n)^3$  points and  $n$  takes the value  $n = 4$  (grid  $16 \times 16 \times 16$ ) or the value  $n = 5$  (grid  $32 \times 32 \times 32$ ). The length of this grid is denoted by  $v_{\max}$  and the number of points of one edge is  $2^n$ . The discrete velocity domain is then the set of points  $v_i = (i^1 \Delta v, i^2 \Delta v, i^3 \Delta v)$  with  $i = (i^1, i^2, i^3)$  ( $0 \leq i^k \leq 2^n - 1$ ,  $k = 1, 2, 3$ ). We also consider the center of the domain  $v_0 = (v_{\max}/2, v_{\max}/2, v_{\max}/2)$ . If  $f$  is the distribution function, we set:  $f_i = f(v_i)$ .

Finally, we make precise that all the following numerical tests are performed within the uncentered and symmetrized FPL operator. In these tests, we consider the evolution in time of the following quantities:

- Discrete kinetic entropy,

$$H_d(t) = \sum_{i \in \mathbb{Z}^3} f_i(t) \log f_i(t) \Delta v^3. \quad (5.1)$$

- Discrete moment of order 4,

$$\mathcal{M}_d^{(4)}(t) = \sum_{i \in \mathbb{Z}^3} (|i| \Delta v)^4 f_i(t) \Delta v^3. \quad (5.2)$$

- Discrete temperatures,

$$T_x(t) = \sum_{i \in \mathbb{Z}^3} (i^1 \Delta v - u_0^1)^2 f_i(t) \Delta v^3 \quad (5.3)$$

$$T_y(t) = \sum_{i \in \mathbb{Z}^3} (i^2 \Delta v - u_0^2)^2 f_i(t) \Delta v^3 \quad (5.4)$$

$$T_z(t) = \sum_{i \in \mathbb{Z}^3} (i^3 \Delta v - u_0^3)^2 f_i(t) \Delta v^3 \quad (5.5)$$

$$T(t) = \frac{1}{3} [T_x(t) + T_y(t) + T_z(t)], \quad (5.6)$$

where  $i = (i^1, i^2, i^3)$ ,  $u_0 = (u_0^1, u_0^2, u_0^3) = (1/\mathcal{N}) \int_{\mathbb{R}^3} v f(v) dv$ , and  $\mathcal{N} = \int_{\mathbb{R}^3} f(v) dv$ .

- Quadratic error: if  $f^{\text{exact}}$  is an exact solution in the Maxwellian case [6], and  $f$  is the approximate solution by sublattices or multigrids schemes corresponding to the initial data  $f_0(v) = f^{\text{exact}}(0, v)$ , then we define the quadratic error as

$$EQ(t) = \sum_{i \in \mathbb{Z}^3} |f_i(t) - f_i^{\text{exact}}(t)|^2 (\Delta v)^3. \quad (5.7)$$

The numerical tests are performed with:  $v_{\max} = 6$ ,  $n = 4$ , or 5 (i.e.,  $N = 16^3$ , or  $32^3$ ). The two methods (sublattices and multigrid) are tested on two different types of initial datas:

*Test 1.* The Maxwellian case: the initial data is chosen in the class of known exact isotropic solutions [6]. Our numerical results are compared with the simplest element of this class of exact solutions,

$$f^{\text{exact}}(v, t) = M_{\mathcal{N}, v_0, T}(v) (1 + c_2 Q_2[(v - v_0)/v_{th}] \exp(-8 \mathcal{N} t)), \quad (5.8)$$

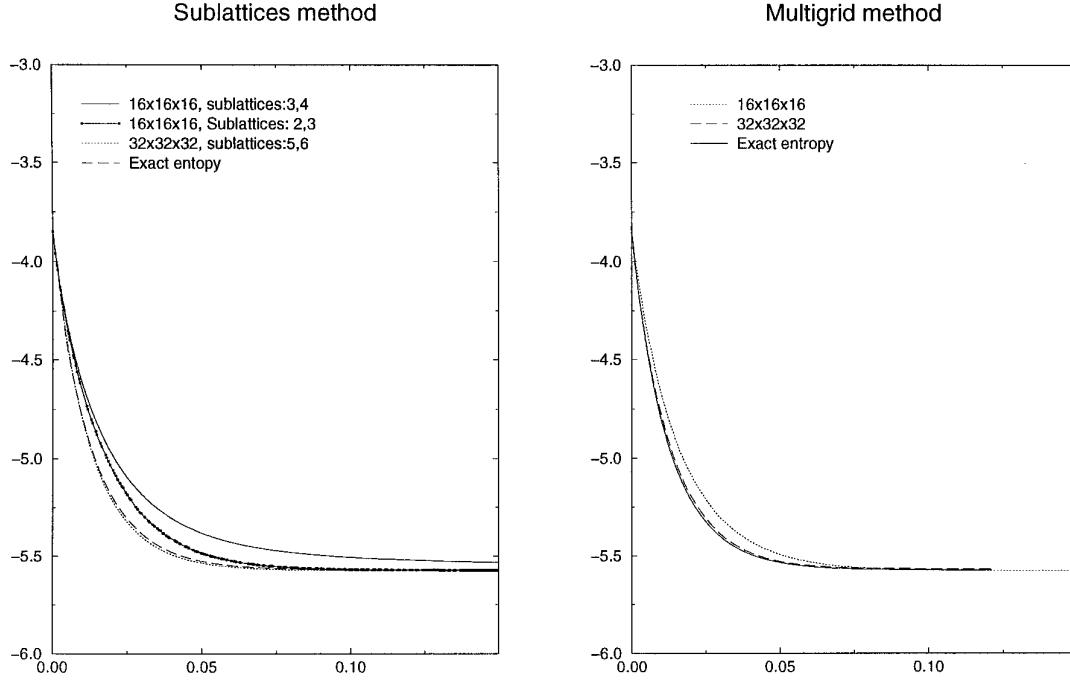


FIG. 1. Kinetic entropy for Maxwellian case.

where:

$$Q_2(v) = \frac{1}{120}(v^4 - 10v^2 + 15) \quad (5.9)$$

is a Sonine polynomial. We recall that  $v_0 = (3., 3., 3.)$  is the center of the domain and choose  $v_{th} = 0.6$ ,  $\mathcal{N} = 5$ . We note that, in this case, the temperatures in various directions  $T_x$ ,  $T_y$ , and  $T_z$  are equal because of the isotropy of the solution and of the isotropy properties of the FPL operator. The evolutions of the entropy and the order 4 moment are compared with their exact evolution in time (Figs. 1 and 3). The evolution of the entropy induced by the multigrid scheme is a little more accurate than the one induced by the sublattice algorithm. To reach the same accuracy, it is necessary to decrease the sublattice size, and then to increase the computational cost. On the other hand, oscillations arise in the time evolution of the moment of order 4 for the multigrid scheme, while the evolution is smooth for the sublattices algorithm. Notice, however, that the relative variations of the moment of order 4 and, thus, of these oscillations, are small. The quadratic error between the distribution function obtained by the numerical schemes and the exact solution of the FPL equation is plotted (Fig. 6) and shows the efficiency of the random-multigrid method in terms of accuracy.

*Test 2.* The Coulombian case: the initial data is now chosen to be bi-Maxwellian i.e., a sum of two Maxwellian functions,

$$f_0(v) = \frac{1}{2}(M_{\mathcal{N}v_{01},T}(v) + M_{\mathcal{N}v_{02},T}(v)), \quad (5.10)$$

where  $M_{\mathcal{N}v,T}$  is given by (1.6), and

$$v_{01} = (2., 3., 3.), \quad v_{02} = (4., 3., 3.).$$

We finally choose  $v_{th} = 0.45$ ,  $\mathcal{N} = 5$ .

The evolutions in time of the entropy and of the temperatures are now compared with the results of exact schemes 2.7 which have a quadratic complexity. The curves of Figs. 2, 5 show that the multigrid algorithm is a little more accurate than the sublattices method. The curves of Fig. 5 give the relaxations in time of the temperatures in various directions of velocity coordinates to their final values and confirm the accuracy of these algorithms. These temperatures are constant if we choose an isotropic initial distribution (as in Test 1 for the Maxwellian case). In the result given by Fig. 4, however, we again observe oscillations of the moment of order 4 with the multigrid algorithm, while the sublattices method gives smoother results. As in Test 1, we note that these oscillations have small relative values.

The curves given in Fig. 7 illustrate the fact that two sublattice sizes are necessary to avoid non-Maxwellian steady states as it is shown in Proposition 4.2. Indeed, a simple use of only one sublattice size ( $a = 5$  for  $32 \times 32$  grid) leads to a final distribution function which is far from the realistic equilibrium state. This is shown

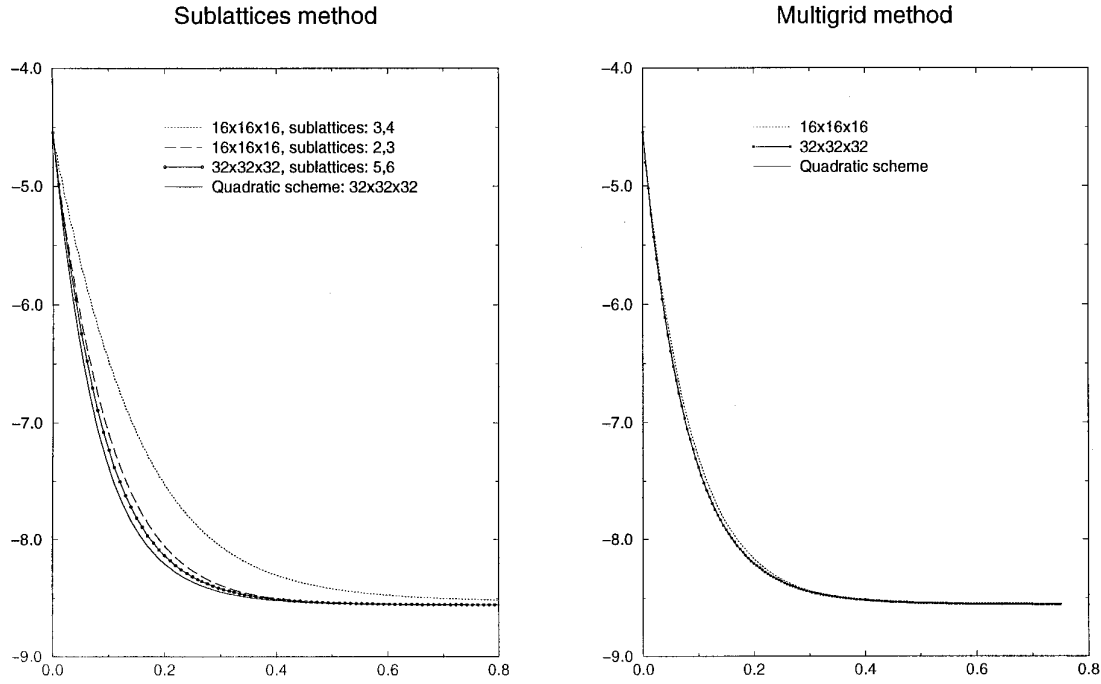


FIG. 2. Kinetic entropy for Coulombian case.

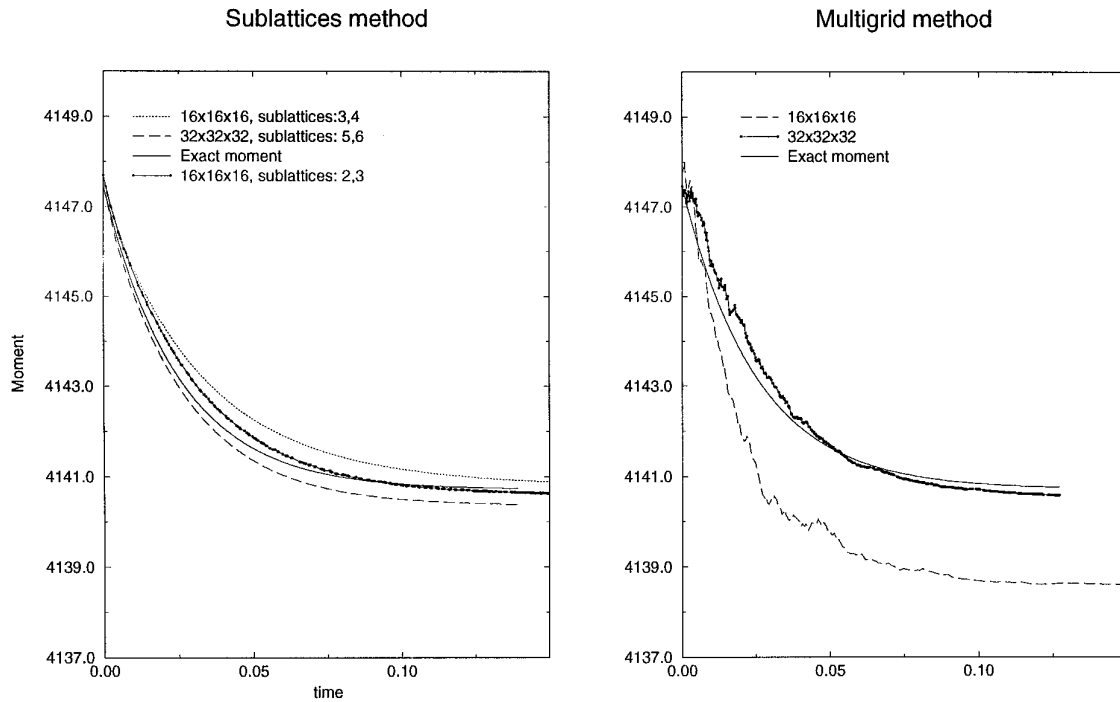


FIG. 3. Moment of order 4 for Maxwellian case.

by plotting the relaxation of the value of the distribution function at the center of the grid (Fig. 7). A difference between the two relaxations given by the use of one ( $a = 5$ ) or two ( $a = 5, b = 6$ ) sublattice sizes is observed (Fig. 7).

These simulations were carried on a *DEC AlphaServer 2100 4/275 OSF/1* monoprocessor, and the CPU times per iteration in time for the two algorithms are listed on the following table (in units of seconds (s) or minutes (min)):

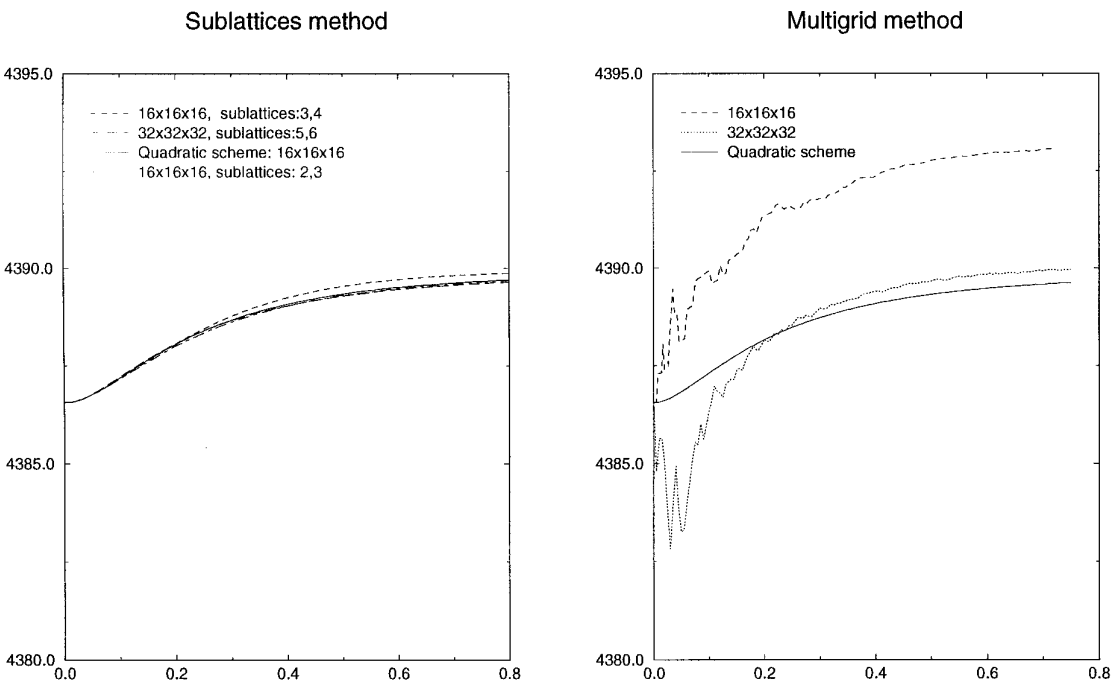


FIG. 4. Moment of order 4 for Coulombian case.

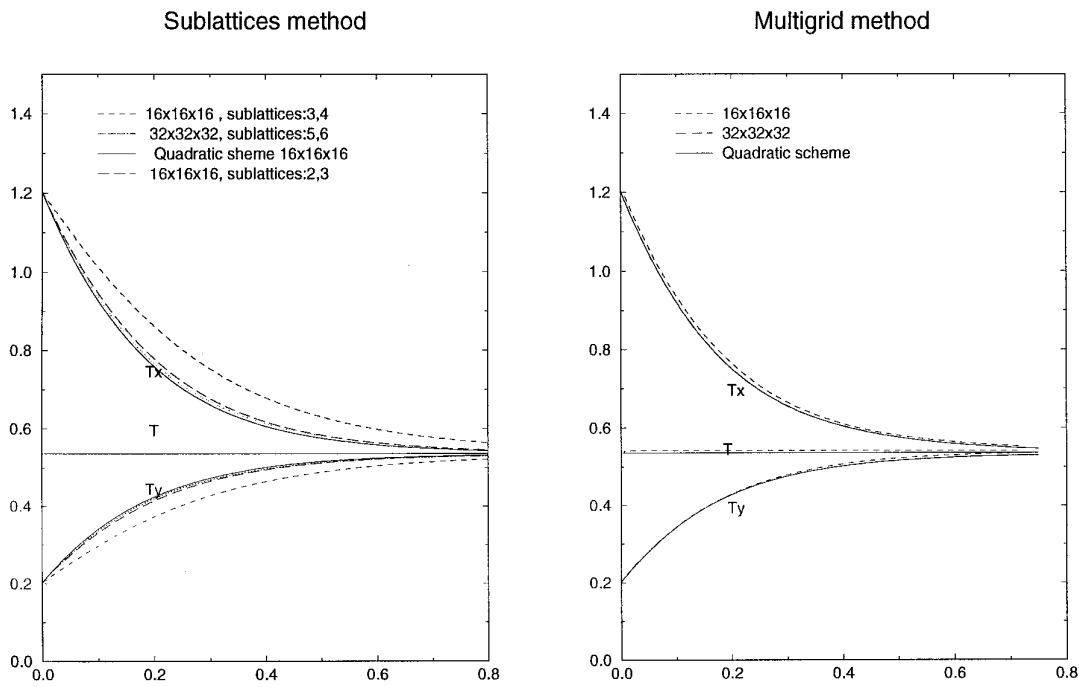


FIG. 5. Températures  $T_x$ ,  $T_y$ , and  $T$  for Coulombian case.

	16 × 16 × 16		32 × 32 × 32	
Sublattice sizes	2,3	3,4	5,6	7,8
Sublattices	3 s	1.3 s	30 s	14 s
Multigrids	0.4 s		8 s	
Quadratic schemes	53 s		60 min	

6. CONCLUSIONS

We have implemented two methods to decrease the computational time required for the evaluation of the discrete FPL operator 2.7. The first one is a sublattices method,

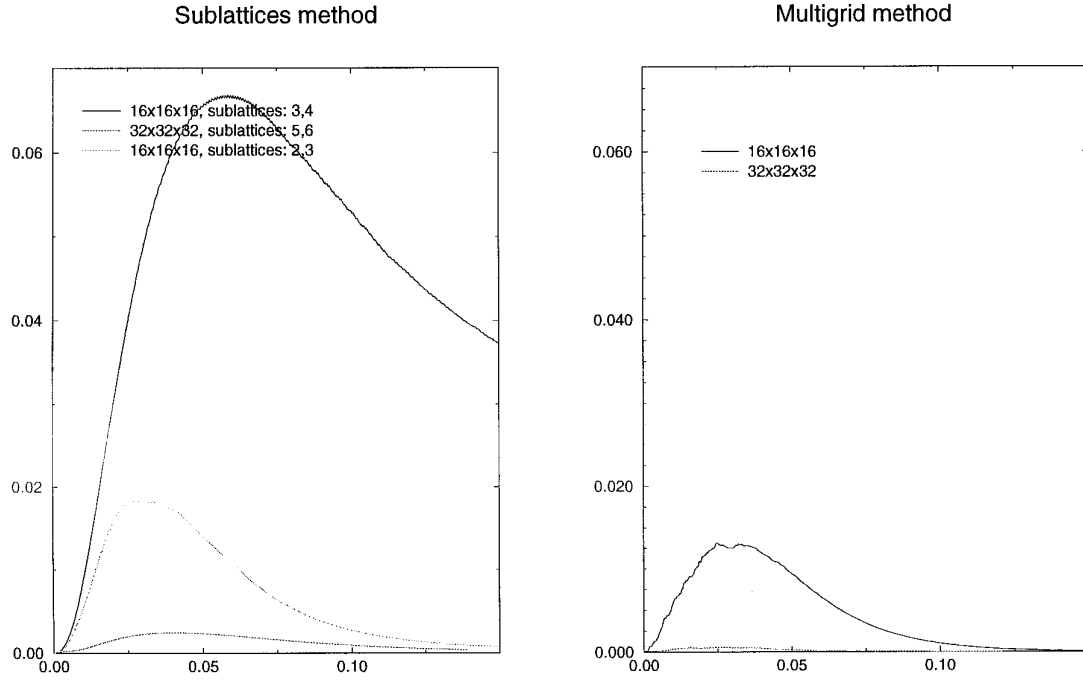


FIG. 6. Quadratic error for Maxwellian case.

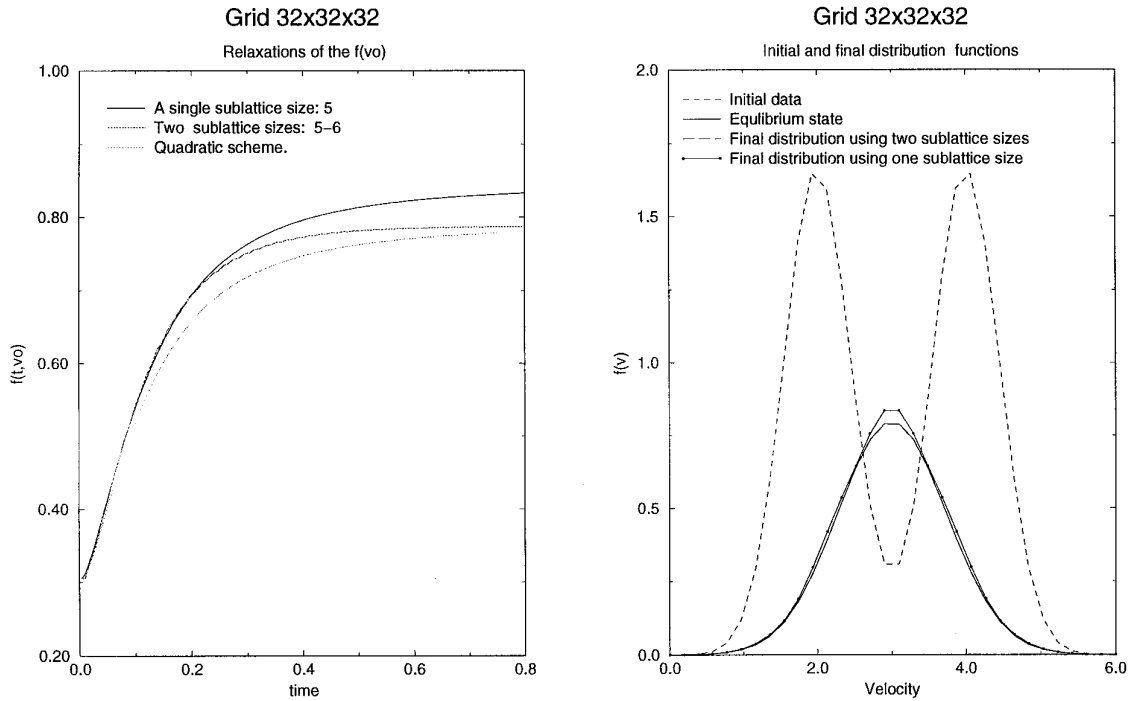


FIG. 7. Coulombian case: Relaxations and Equilibrium states using one and two sublattice sizes. The numerical equilibrium state obtained by using two sublattices sizes coincides with the realistic (the Maxwellian) final distribution.

while the second one is a multigrid method with random evaluations. We have noticed that both algorithms are conservative and decrease the kinetic entropy. Their computational cost is highly reduced, compared with the original quadratic scheme (divided by a factor of order 40 for a  $16 \times 16 \times 16$  grid and by a factor of order 200 for a  $32 \times 32 \times 32$  grid). The sublattices algorithm does not give satisfactory results when the sublattice sizes are too big (for instance, 9, 10 for a grid  $32 \times 32 \times 32$ ), since the number of collisional invariants to be suppressed is large. The numerical tests show that the multigrid algorithm is a little bit more accurate than the sublattice scheme and the CPU times are better. On the other hand, the sublattices method (which is deterministic) provides smoother results in some circumstances. To remove these oscillations, a deterministic version of the multigrid method is proposed in a forthcoming paper [7]. This alternative approach mainly consists on replacing the Monte Carlo integrations by multipole expansions in the spirit of the work by Greengard and Rokhlin [19] and allows us to control the error by the simple choice of the order of these multipole expansions.

## REFERENCES

1. C. Buet, Résolution déterministe de l'équation de Boltzmann, note interne CEA, 1993.
2. C. Buet, A discrete velocity scheme for the Boltzmann operator of rarefied gas dynamics, *Transp. Theory Stat. Phys.*, to appear.
3. C. Buet, Discrete velocity and internal energy models derived from the Boltzmann operator for polyatomic gases with the Larsen-Borgnakke model of energy exchange, *Math. models methods Appl. Sci.*, to appear.
4. B. Lucquin-Desreux, Discrétisation de l'opérateur de Fokker-Planck dans le cas homogène, *C.R. Acad. Sci. Paris A Sér. I*, **314**, 407 (1992).
5. P. Degond and B. Lucquin-Desreux. An entropy scheme for the Fokker-Planck collision of plasma kinetic theory. *Numer. Math.* **68**, 239 (1994).
6. M. Lemou. Exact solutions of the Fokker-Planck equation. *C.R. Acad. Sci. Ser. I*, **319**, 579 (1994).
7. M. Lemou, Multipole expansions for the Fokker-Planck equation, *Numerische Mathematik*, to appear.
8. I. F. Potapenko and V. A. Chuyanov, A completely conservative difference scheme for the two-dimensional Landau equation, *U.S.S.R. Comput. Maths. Math. Phys.* **20**(2), 249 (1980).
9. A. V. Bobilev, I. F. Potapenko, and V. A. Chuyanov, Kinetic equations of the Landau type as a model of the Boltzmann equation and completely conservative difference schemes, *U.S.S.R. Comput. Maths. Math. Phys.* **20**(4), 190 (1981).
10. J. C. Whitney, Finite Difference Methods for the Fokker-Planck Equation, *J. Comput. Phys.* **6**, 483 (1970).
11. Y. A. Berezin, V. N. Khudik and M. S. Pekker, Conservative finite-difference schemes for the Fokker-Planck Equation not violating the law of an increasing entropy, *J. Comput. Phys.* **69**, 163 (1987).
12. M. S. Pekker and V. N. Khudik, Conservative Difference Schemes for the Fokker-Planck equation, *U.S.S.R. Comput. Maths. Math. Phys.* **24**(3), 206 (1984).
13. A. A. Arsenev and O. E. Buryac, On the connection between a solution of the Boltzmann equation and a solution of the Landau-Fokker-Planck equation, *Math. USSR Sb.* **69**(2), 465 (1991).
14. L. Desvillettes, On asymptotics of the Boltzmann equation when the collisions become grazing. *Trans. Theory and Stat. Phys.*, **21**(3), 259 (1992).
15. A. A. Arsenev and N. V. Peskov, On the existence of a generalized solution of Landau's equation, *USSR Comput. Maths math. Phys* **17**, 241 (1977).
16. P. Degond and B. Lucquin-Desreux, The Fokker-Planck asymptotics of the Boltzmann collision operator in the Coulomb case, *Math. Models and Methods in Appl. Sci.* **2**(2), 167 (1992).
17. O. Larroche, Kinetic Simulations of a plasma collision experiment, *Phys. Fluids B*, **5**(8) (1993).
18. E. Frenod and B. Lucquin-Desreux, On conservative and entropic discrete axisymmetric Fokker-Planck operators, in preparation (unpublished).
19. L. Greengard and V. Rokhlin, A fast algorithm for a particle simulation. *J. Comput. Phys.* **73**, 325 (1987).
20. H. Babovsky, A convergence proof for Nanbu's Boltzmann simulation scheme. *Eur. J. Mech. B/Fluids* **8**(1), 41 (1989).



# Regularized Boltzmann operators

Christophe Buet<sup>1</sup>, Stéphane Cordier<sup>2</sup>, Pierre Degond<sup>3</sup>.

November 29, 2005

(1) Centre de Limeil Valenton,  
C. E. A. /C. E. L. -V.  
94195 Villeneuve Saint Georges Cedex, France  
email: buet@limeil. cea. fr

(2) Laboratoire d'Analyse Numérique, URA CNRS 189,  
Université Pierre et Marie Curie, tour 55-65, 5<sup>ème</sup> étage,  
4 Place Jussieu, 75 252 Paris cedex 05, France  
email: cordier@ann. jussieu. fr

(3) Mathématiques pour l'Industrie et la Physique, UMR CNRS 9974,  
UFR MIG, Université Paul Sabatier  
118, route de Narbonne 31062 Toulouse Cedex, France  
email: degond@mip. ups-tlse. fr

## Abstract

In this paper, we propose two regularization approaches for the Boltzmann collision operator. The constructed operators preserve the mass, momentum and energy; their equilibrium states are Maxwellians and they satisfy the H theorem. In the first approach, the regularization consists in allowing microscopic collisions which do not exactly preserve energy and momentum. However, the limit of the mollified operator when the cut-off parameter tends to 0 is not the usual Boltzmann operator unless a certain condition on the distribution function is satisfied. In the second approach, the regularization relies on a smoothing of the masses of the particles and leads to a regularized operator which formally tends to the Boltzmann operator for any arbitrary distribution function, when the cut-off parameter tends to zero.

# 1 Introduction

In the recent past, various new numerical methods for solving the Boltzmann equation have been investigated, in particular, those based on discrete velocity models (DVMs) (see [11], [5], [22] and [19]). In such methods, the velocities lie on a fixed lattice  $L$  of  $\mathbb{R}^3$ . The consistency of these DVMs are closely related to the repartition of integer roots of the equation  $x^2 + y^2 + z^2 = n$ . So far, partial consistency results have been obtained via number theory (for example [2], [5]). From the numerical and practical point of view, the main difficulty with DVMs is the small number of pairs of discrete post collisional velocities for a given pair of pre-collisional velocities. Indeed, the number of intersection points between the collision sphere and the lattice  $L$  of discrete velocities may be very small [12]. In such circumstances, the grid needs to be refined and the cost becomes prohibitive. To waive this difficulty, a smoothing of the collision sphere is necessary. This is the first motivation of the present work.

The second motivation is to propose a basis for the approximation of the Boltzmann collision operator by means of the particle methods [8, 17, 21]. Monte Carlo methods allow to treat both the transport and collision operators in a fairly natural and easy way[1]. However, the Monte Carlo treatment of the collision integral generates a fairly high level of noise. It would therefore be desirable to design a particle method which would allow a flexible treatment of the transport term, together with a deterministic treatment of the collision integral in order to decrease the noise level. This goal has been achieved in linear transport theory [8, 17, 21] but has for long time been stopped by the inability of finding a regularization of the microscopic collision process which still yields a macroscopically conservative Boltzmann operator. For short, the treatment of collisions by a deterministic particle method would essentially reduce to find a 4-velocity model for velocities which are not exactly co-spheric (in the center of the mass frame), but such that, after integration over all the possible 4-velocity configurations, the collision operator would still be conservative (in mass, momentum and energy). This rather imprecise statement will be made clearer later on. The goal of the present paper is precisely to investigate whether it is possible to perform such regularization.

In this paper, two possible strategies to address these problems are presented. The first strategy consists in mollifying the collision sphere and in considering the collisional velocities on a spherical shell rather than on a sphere. Therefore, momentum and energy are not preserved at the level of the microscopic collision. However, they may be conserved at the macroscopic level if the scattering cross section of these unphysical collisions is carefully chosen. Indeed, in section 2 we shall construct a mollified Boltzmann operator such that:

- conservations of mass, momentum and energy holds;
- the steady state solutions are Maxwellians;
- the H-theorem holds.

These three properties are required for having a correct convergence of the distribution function towards the Maxwellian distribution at large times, which is of primary importance. We prove the following results:

- The construction of a mollified operator for which either energy or both momentum and energy conservations are relaxed at a microscopic level but such that they are satisfied at the macroscopic level is always possible (see sections 2.1, 2.2 and 2.4). In section 2.3, we show that the associated cross sections can be chosen in a simple form i. e. as a piecewise constant function.
- This collision operator may be chosen in such a way that, when the regularization parameter tends to 0, it converges to the usual Boltzmann operator with a modified scattering cross section. We prove that the limit cross sections coincide with the usual one if and only if the moments of the distribution function satisfy some constraint (see section 2.5).

A distinctive feature of this mollified collision operator is that its associated scattering cross section depends on the distribution function itself. This heavily complicates the structure of the collision operator and therefore, its implementation.

The second strategy is based on a modification of the masses of the particles during the collision and will be developed in section 3. Once again, the microscopic collision is modified but the operator is constructed in such a way that it preserves the above three properties at the microscopic level. The main advantage of this method is to be "macroscopic" compared with the previous one in that the associated scattering cross section essentially depends on the first three moments of the distribution function instead of the microscopic details of it. Moreover, this mollified operator tends, at least formally, when the cut-off parameter tends to 0, to the usual Boltzmann operator.

We recall some classical features of the Boltzmann collision operator in the following paragraphs. We shall restrict ourselves to a monoatomic gas and consider the Boltzmann collision operator of the form

$$Q[f](v) = \int_{\mathbb{R}^3 \times S_+^2} \sigma \left( |v - v_1|, \frac{(v - v_1, \Omega)}{|v - v_1|} \right) |v - v_1| (f' f'_1 - f f_1) d\Omega dv_1, \quad (1.1)$$

where  $\sigma$  is the differential scattering cross section and  $f' = f(v')$ ,  $f'_1 = f(v'_1)$ ,  $f = f(v)$ ,  $f_1 = f(v_1)$ ,  $v$  and  $v_1$  (resp.  $v'$  and  $v'_1$ ) are the particle velocities before (resp. after) the collision and are given by

$$v' = v - (v - v_1, \Omega) \Omega, \quad v'_1 = v_1 + (v - v_1, \Omega) \Omega. \quad (1.2)$$

$(x, y)$  denotes the dot product of the vectors  $x$  and  $y$  of  $\mathbb{R}^3$  and  $\Omega$  is an arbitrary angle  $\Omega \in S^2$ , where  $S^2$  is the unit sphere of  $\mathbb{R}^3$ . These relations express the conservation of momentum and energy during a collision (the conservation of the number of particles is obviously satisfied)

$$v + v_1 = v' + v'_1, \quad (1.3)$$

$$|v|^2 + |v_1|^2 = |v'|^2 + |v'_1|^2. \quad (1.4)$$

The standard properties of the Boltzmann operator can easily be presented on the weak formulation of the Boltzmann operator. Let  $\Psi$  be a test function, we have:

$$\text{P1 (Conservations): } \int_{v \in \mathbb{R}^3} Q[f](v) \Psi(v) dv = 0, \quad \forall f, \quad (1.5)$$

if and only if there exists  $a \in \mathbb{R}$ ,  $b \in \mathbb{R}^3$  and  $c \in \mathbb{R}$  such that  $\Psi(v) = a + b \cdot v + c|v|^2$ .

$$\text{P2 (Maxwellians): } \int_{\mathbb{R}^3} Q[f](v) \Psi(v) dv = 0, \quad \forall \Psi, \quad (1.6)$$

if and only if the distribution function is a Maxwellian

$$M_{\rho, u, T} = \frac{\rho}{(2\pi T)^{3/2}} \exp\left(\frac{-|v - u|^2}{2T}\right). \quad (1.7)$$

Finally, the entropy decay reads:

$$\text{P3 (H-theorem): } \int_{\mathbb{R}^3} Q[f](v) \ln(f(v)) dv \leq 0, \quad \forall f. \quad (1.8)$$

These properties hold true for sufficiently smooth and fast decaying distribution functions.

## 2 Mollifying the collision sphere.

Note first that the Boltzmann operator (1.1) can be equivalently written in the form:

$$Q[f](v) = \int_{(\mathbb{R}^3)^4} c(v, v_1, v', v'_1) (f' f'_1 - f f_1) dv' dv'_1 dv_1, \quad (2.1)$$

where the integration is now taken upon the velocities  $dv' dv'_1 dv_1$  and

$$c(v, v_1, v', v'_1) = \delta_0(v + v_1 - v' - v'_1) \delta_0(|v - v_1|^2 - |v' - v'_1|^2) C\left(|v - v_1|, \frac{(v - v_1, \Omega)}{|v - v_1|}\right), \quad (2.2)$$

$$C\left(|v - v_1|, \frac{(v - v_1, \Omega)}{|v - v_1|}\right) = \sigma\left(|v - v_1|, \frac{(v - v_1, \Omega)}{|v - v_1|}\right) |v - v_1| |v' - v'_1|, \quad (2.3)$$

where  $\delta_0$  represents the delta measure located at  $x = 0$ . A straightforward calculation gives (from  $\Omega = \frac{(v' - v'_1) - (v - v_1)}{|(v' - v'_1) - (v - v_1)|}$ ):

$$\left| \frac{(v - v_1, \Omega)}{|v - v_1|} \right| = \frac{1}{2} \frac{|(v' - v'_1) - (v - v_1)|}{|v - v_1|}. \quad (2.4)$$

Note that the functions  $c$  and  $C$  are only defined for velocities  $v$ ,  $v_1$ ,  $v'$  and  $v'_1$  satisfying the conservation relations (1.3)-(1.4). The conservation of momentum and energy are now taken care of by the delta measures in (2.2). Using this formulation, which can be found for instance in Cercignani [7], a natural generalization of the Boltzmann operator, consists in mollifying these measures in order to increase the number of possible post-collisional velocities.

Let  $\Psi$  be a test function. We present the construction of the mollified Boltzmann operator on the following standard symmetrized weak formulation of (2.1):

$$\begin{aligned} (Q[f], \Psi) &= \int_{\mathbb{R}^3} Q[f](v) \Psi(v) dv, \\ &= \frac{-1}{4} \int_{(\mathbb{R}^3)^4} c(v, v_1, v', v'_1) (\Psi' + \Psi'_1 - \Psi - \Psi_1) (f' f'_1 - f f_1) dv dv' dv'_1 dv_1, \end{aligned} \quad (2.5)$$

where  $\Psi'$ ,  $\Psi'_1$ ,  $\Psi$  and  $\Psi_1$  again denote  $\Psi(v')$ ,  $\Psi(v'_1)$ ,  $\Psi(v)$  and  $\Psi(v_1)$ , respectively. Our regularization procedure consists in smoothing the delta measures  $\delta$  in the definition (2.2) of the cross section  $c$  into a positive function  $\delta^\epsilon$  depending on a smoothing parameter  $\epsilon$ . One goal is to achieve that

$$\delta^\epsilon \left( \frac{v+v_1}{2}, \frac{v-v_1}{2}, \frac{v'+v'_1}{2}, \frac{v'-v'_1}{2} \right) \xrightarrow{\epsilon \rightarrow 0} \delta_0(v+v_1-v'-v'_1) \delta_0(|v-v_1|^2 - |v'-v'_1|^2). \quad (2.6)$$

while  $\delta^\epsilon$  being designed such that the conservation properties are not lost. We shall see that we cannot always achieve this aim. We drop the dependence upon  $\epsilon$  for the moment i. e.  $\delta^\epsilon = \delta$ .

The mollified Boltzmann operator  $\tilde{Q}$  is written in the following symmetrized weak form:

$$\begin{aligned} (\tilde{Q}[f], \Psi) = & \frac{-1}{4} \int_{(\mathbb{R}^3)^4} \tilde{C} \left( \frac{v-v_1}{2}, \frac{v'-v'_1}{2} \right) (\Psi' + \Psi'_1 - \Psi - \Psi_1) \\ & \left( f' f_1 \delta \left( \frac{v+v_1}{2}, \frac{v-v_1}{2}, \frac{v'+v'_1}{2}, \frac{v'-v'_1}{2} \right) \right. \\ & \left. - f f_1 \delta \left( \frac{v'+v'_1}{2}, \frac{v'-v'_1}{2}, \frac{v+v_1}{2}, \frac{v-v_1}{2} \right) \right) dv dv' dv'_1 dv_1, \end{aligned} \quad (2.7)$$

where  $\tilde{C}$  is defined (from (2.4) and (2.3)) as

$$\tilde{C}(z, z') = C \left( \overline{|z|}, \frac{|z-z'|}{2\overline{|z|}} \right), \quad (2.8)$$

where  $\overline{|z|}$  stands for some averaged value of  $|z|$  and  $|z'|$ , like, for example

$$\overline{|z|} = \sqrt{|z||z'|} \quad \text{or} \quad \overline{|z|} = \frac{(|z| + |z'|)}{2}.$$

The only required properties for  $\overline{|z|}$  is to coincide with  $|z|$  for  $z = z'$ , to be smooth and symmetric (when exchanging  $z$  and  $z'$ ) in order to ensure the consistency of the construction. In this section, we first obtain necessary conditions on the function  $\delta$  such that properties (P1), (P2) and (P3) are satisfied by  $\tilde{Q}$  (subsection 2.1). Then, in subsection 2.2 we prove the existence of the mollified operators. When both energy and momentum conservations are smoothed we prove that the regularized operator can be chosen in a simple form (using a piecewise constant correction function  $S$ ) under some conditions on the distribution function  $f$ . In subsection 2.4, we show an explicit construction of the mollified operator without any conditions on the distribution function, when only the energy conservation relation is smoothed, and in subsection 2.5 we study its convergence to the Boltzmann operator when the smoothing parameter tends to 0.

## 2.1 Construction of the regularization functions.

### 2.1.1 Change of variables.

We use the change of variables from the velocities  $v$ ,  $v_1$  (resp.  $v'$  and  $v'_1$ ) to the velocities of the center of mass frame (denoted by  $w$ , resp.  $w'$ ) and relative velocity

$z$  (resp.  $z'$ ). More precisely, we set:

$$\begin{cases} w = \frac{v + v_1}{2}, & z = \frac{v - v_1}{2}, \\ w' = \frac{v' + v'_1}{2}, & z' = \frac{v' - v'_1}{2}. \end{cases} \quad (2.9)$$

We recall that conservations of momentum and energy at the binary collision level are written in these variables:  $w = w'$  and  $|z|^2 = |z'|^2$ . The Jacobian of this transformation is  $(\frac{1}{2})^3$ . Thus, the Boltzmann operator  $\tilde{Q}$  can be written in a weak formulation:

$$\begin{aligned} (\tilde{Q}[f], \Psi) &= -2 \int_{(\mathbb{R}^3)^4} \tilde{C}(z, z') (\Psi' + \Psi'_1 - \Psi - \Psi_1) (f' f'_1 \delta(w, z, w', z') \\ &\quad - f f_1 \delta(w', z', w, z)) dz dz' dw' dw, \end{aligned} \quad (2.10)$$

where the functions ( $f$  or  $\Psi$ ) are evaluated at the points  $v, v_1, v', v'_1$  depending on  $w, w', z$  and  $z'$  as defined in (2.9).

### 2.1.2 Maxwellian steady states i. e. (P2).

The collision operator has to vanish for Maxwellian distribution functions. This property (P2) cannot be achieved if  $\delta$  is independent of the distribution function. Actually, we require a bit more indeed, that the integrand in (2.10) identically vanishes when the distribution function is a Maxwellian (as defined in (1.6)); this is the usual distinction between the global and detailed balance properties. Here the detailed balance property is written:

$$M(w' + z') M(w' - z') \delta(w, z, w', z') = M(w + z) M(w - z) \delta(w', z', w, z) \quad (2.11)$$

for any Maxwellian distribution function defined by (1.7). Note that we have, for any Maxwellian  $M_{\rho, u, T}$  and any vectors  $w$  and  $z$

$$M_{\rho, u, T}(w + z) M_{\rho, u, T}(w - z) = M_{\rho, u, \frac{T}{2}}(w) M_{\rho, 0, \frac{T}{2}}(z). \quad (2.12)$$

Therefore, (2.11) is achieved if and only if  $\delta$  is such that

$$\delta(w, z, w', z') = M_{\rho, u, \frac{T}{2}}(w) M_{\rho, 0, \frac{T}{2}}(z) S(w, z, w', z'), \quad (2.13)$$

where  $S$  is a positive symmetric function of the pairs  $(w, z)$  and  $(w', z')$ :

$$S(w, z, w', z') = S(w', z', w, z), \quad \forall (w, z, w', z') \in (\mathbb{R}^3)^4. \quad (2.14)$$

In the remainder, we shall restrict to functions  $S$  which satisfy

$$\exists \alpha \mid S(w, z, w', z') > 0, \quad |w - w'| \leq \alpha, \quad |z - z'| \leq \alpha.$$

Let us now define the Maxwellian distribution function  $M^f$  which has the same first three moments as  $f$  i. e. ,  $M^f = M_{\rho_f, u_f, T_f}$  such that

$$\begin{pmatrix} \rho_f \\ \rho_f u_f \\ \frac{1}{2} \rho_f (u_f^2 + 3T_f) \end{pmatrix} = \int_{\mathbb{R}^3} f(v) \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} dv \quad (2.15)$$

### 2.1.3 Conservations laws i. e. (P1).

Then, we consider conservations of momentum and energy, since conservation of mass is obvious (i. e.  $(\tilde{Q}, \Psi) = 0$  for  $\Psi(v) = 1$ ). The conservation of momentum is written:

$$(\tilde{Q}[f], v) = 0 \Leftrightarrow \int_{(\mathbb{R}^3)^4} \tilde{C}(z, z') (w - w') \left( f'_1 f' M^f M_1^f - f f_1 (M^f)' (M_1^f)' \right) S(w, z, w', z') dz dz' dw' dw,$$

and the conservation of energy:

$$(\tilde{Q}[f], |v|^2) = 0 \Leftrightarrow \int_{(\mathbb{R}^3)^4} \tilde{C}(z, z') (|w|^2 - |w'|^2 + |z|^2 - |z'|^2) \left( f'_1 f' M^f M_1^f - f f_1 (M^f)' (M_1^f)' \right) S(w, z, w', z') dz dz' dw' dw.$$

Let us denote  $X = (w, z) \in (\mathbb{R}^3)^2$  (and  $X' = (w', z')$ ). The conservation relations can then be written:

$$\int_{\mathbb{R}^{12}} \tilde{S}(X, X') \left( \frac{(w - w')}{|X|^2 - |X'|^2} \right) (F(X')H(X) - F(X)H(X')) dX dX' = 0, \quad (2.16)$$

with

$$\begin{aligned} F(X) &= f(w + z)f(w - z), \\ H(X) &= M^f(w + z)M^f(w - z), \\ \tilde{S}(X, X') &= \tilde{C}(z, z')S(w, z, w', z'), \\ |X|^2 &= |w|^2 + |z|^2. \end{aligned} \quad (2.17)$$

Note that  $\tilde{S}$  is a nonnegative symmetric function of its argument which is positive in a strip:

$$\tilde{S}(X, X') \geq 0, \quad \tilde{S}(X, X') = \tilde{S}(X', X), \quad \forall (X, X') \in (\mathbb{R}^6)^2, \quad (2.18)$$

$$\tilde{S}(X, X') > 0, \quad \forall X, X' \quad |X - X'| < \alpha. \quad (2.19)$$

Therefore,  $\tilde{S} \neq 0$  on a non negligible set (with respect to the Lebesgue measure  $dX dX'$ ). It is easy to check that the functions  $F$  and  $H$  satisfy (from relations (2.15) and with  $d = 6$ ):

$$\int_{\mathbb{R}^d} \left( \frac{1}{X_i} \right) F(X) dX = \int_{\mathbb{R}^d} \left( \frac{1}{X_i} \right) H(X) dX, \quad \forall i = 1 \dots 3. \quad (2.20)$$

Note that the conservation relations (2.16) are symmetric with respect to the exchange of  $X$  and  $X'$ . This will be used to reduce the integration domains.

## 2.2 Existence of $\tilde{S}$ in the general case.

We now construct the function  $\tilde{S}$  from  $\mathbb{R}^{12}$  to  $\mathbb{R}^+$  such that conservation of momentum and energy holds. These conservation laws can be written in the form

$$\int_{(\mathbb{R}^3)^4} G_i(X, Y) \tilde{S}(X, Y) dX dY = 0, \quad i = 0 \dots 3. \quad (2.21)$$

where the functions  $G_i$  are defined by

$$G_0(X, Y) = (F(X)H(Y) - F(Y)H(X))(|X|^2 - |Y|^2), \quad (2.22)$$

$$G_i(X, Y) = (F(X)H(Y) - F(Y)H(X))(X_i - Y_i), \quad i \in [1..3] \quad (2.23)$$

The existence of such a function  $\tilde{S}$  (or equivalently  $S$  since  $\tilde{C}$  has been already chosen) in the cone of positive functions requires the following necessary and sufficient condition

### Proposition 2.1

*There exists a positive and non vanishing function  $\tilde{S}$  satisfying (2.21) if and only if there exists no positive linear combination of the functions  $G_i$ .*

**Proof :** The proof of the direct implication is obvious, by contradiction. We shall now prove that if the intersection of the space  $V$  generated by the functions  $(G_i)_{i=1\dots 4}$  and the cone  $C$  of positive functions reduces to the null function, then there exists a positive function  $\tilde{S}$  satisfying (2.21) i. e. orthogonal to the functions  $(G_i)_{i=1\dots 4}$ .

The sets  $C$  and  $V$  are considered as subsets of the Banach space  $H = L^2$  and they are non empty, convex and disjoint. Moreover,  $C$  is closed and it generates  $H$  in the sense that  $C + (-C) = H$  and  $C \cap (-C) = \{0\}$ . If Hahn Banach theorem applies, these sets can be separated by an hyperplane i. e. there exists a non vanishing function  $\tilde{S} \in H$  such that

$$\langle \tilde{S}, C \rangle \subset \mathbb{R}^+, \quad \langle \tilde{S}, V \rangle = 0, \quad (2.24)$$

which gives the result. The difficulty relies on the fact that none of these sets are open. However, we still obtain the result. Let us consider  $f_0 \neq 0$  in  $C$  and define  $C_0 = \{f \geq f_0\}$ . We show that the distance between  $V$  and  $C_0$  is such that  $\text{dist}(V, C_0) = 2r > 0$ : by contradiction, if  $\exists (f_n) \in C$  and  $(x_n)$  in  $V$  such that

$$|f_n + f_0 - x_n| \rightarrow 0$$

Then, if  $x_n$  is bounded, it converges toward  $x_0 \in V$  (since  $V$  is closed) up to the extraction of a subsequence and  $f_n + f_0 \rightarrow x_0$ . Therefore  $f_n$  converges to  $(x_0 - f_0) \in C$  (since  $C$  is closed) and hence,  $x_0 = 0$ . This implies that  $f_0 \in (-C)$  and contradicts  $f_0 \neq 0$ . When  $x_n$  is not bounded, one can consider the sequence  $\frac{x_n}{|x_n|}$  and obtain again a contradiction.

Let us now define the following mollification of  $V$ :

$$V_r \stackrel{\text{def}}{=} \{x \in H \mid \exists y \in V, \quad x \in B(y, r)\}.$$



$V_r$  is now an open convex set which is disjoint from  $C_0$ , since  $r = \text{dist}(V, C_0)/2$ . We can now apply the Hahn Banach theorem which gives the existence of a non vanishing function  $\tilde{S} \in H$  and of a real number  $\alpha$  such that

$$\langle \tilde{S}, x \rangle \geq \alpha, \forall x \in C_0, \quad \langle \tilde{S}, x \rangle \leq \alpha, \forall x \in V_r,$$

The second condition implies that  $\langle \tilde{S}, x \rangle \leq \alpha, \forall x \in V$  and therefore (since  $V$  is a vector space)  $\langle \tilde{S}, G_i \rangle = 0, \forall i \in [1 \cdots 4]$  and  $\alpha = 0$ . The first condition reads  $\langle \tilde{S}, th + f_0 \rangle \geq 0$  for all  $h \in C$  and for all  $t \geq 0$ . When  $t \rightarrow \infty$ , we obtain

$$\langle \tilde{S}, h \rangle \geq 0, \forall h \in C.$$

This implies that  $\tilde{S} \in C$  and ends the proof.  $\square$

**Remark 2.2** *The same proof holds when the functions  $G_i$  are in  $L^p(\mathbb{R}^n)$  for  $p < \infty$  and insures the existence of a function  $f \in L^{p'}$ . Since the functions  $G_i$  of interest are naturally in  $L^1$ , we have the existence of  $\tilde{S} \in L^\infty$  (provided  $V \cap C = \{0\}$  using the above notations).*

**Remark 2.3** *It can be proved that the function  $S$  can be chosen as an analytical function and everywhere strictly positive ([18]).*

**Remark 2.4** *Note that the construction of positive functions orthogonal to some subspaces of  $C^2[0, T]$  is described in ([14]).*

We now prove that the non existence of a positive combination of the  $G_i$ 's is satisfied provided the distribution function  $f$  is not almost everywhere equal to a Maxwellian.

**Lemma 2.5**

*Suppose that there exists  $(\lambda_i)_{i=0,\dots,3}$  such that the function defined by*

$$G = \sum_{i=0}^3 \lambda_i G_i$$

*is positive on a non negligible set. Then, we have*

$$F = H, \text{ a.e.}$$

**Proof :** We prove this by contradiction: we suppose that  $F \neq H$  on a non negligible set and we assume the existence of  $(\lambda_i)_{i=0,\dots,3}$  such that  $G(X, Y) \geq 0$  a. e.  $(X, Y) \in \mathbb{R}^{2d}$  and such that  $G > 0$  on a non negligible set of  $\mathbb{R}^{2d}$ .

By the definition of the function  $G_i$  we have:

$$G(X, Y) = (F(X)H(Y) - H(X)F(Y))(P(X) - P(Y))$$

with  $P(X) = \alpha_0 |X|^2 + \sum_{i=1}^3 \alpha_i X_i$ . By integrating over  $Y$ , we obtain

$$\int_{\mathbb{R}^d} (F(X)H(Y) - H(X)F(Y))(P(X) - P(Y)) dY > 0, \text{ a.e. } X \in \mathbb{R}^d. \quad (2.25)$$

With the use of (2.20), we shall assume, without any loss of generality, that

$$\int_{\mathbb{R}^6} \left( \frac{1}{P(X)} \right) F(X) dX = \int_{\mathbb{R}^6} \left( \frac{1}{P(X)} \right) H(X) dX = \left( \frac{1}{\alpha} \right). \quad (2.26)$$

Then, (2.25) yields

$$(P(X) - \alpha)(F(X) - H(X)) > 0. \quad (2.27)$$

This gives

$$\begin{aligned} \int_{P(X) > \alpha} P(X)(F(X) - H(X))dX &\geq \alpha \int_{P(X) > \alpha} (F(X) - H(X))dX, \\ &= -\alpha \int_{P(X) < \alpha} (F(X) - H(X))dX, \\ &= \alpha \int_{P(X) < \alpha} (H(X) - F(X))dX, \\ &\geq \int_{P(X) < \alpha} (H(X) - F(X))P(X)dX, \end{aligned}$$

and one of these inequalities at least is strict by the fact that  $G > 0$  on a non negligible set of  $\mathbb{R}^{2d}$ . This is in contradiction with

$$\int (H(X) - F(X))P(X)dX = 0,$$

and ends the proof.  $\square$

With Lemma 2.5 we prove the existence of a regularized operator  $\tilde{Q}$  which satisfies properties (P1)-(P3):

**Theorem 2.6**

*There exists a positive function  $S$  such that the collision operator  $\tilde{Q}$  based on the above constuction satisfies properties (P1), (P2) and (P3).*

**Proof :** Either  $f = M^f$  a.e and then  $\tilde{Q} = 0$  or  $f \neq M^f$  on a non negligible set, then  $F \neq H$  on a non negligible set and there exists a function  $\tilde{S}$  from Proposition 2.1. The verification of the remaining property (P3) follows from the positivity of  $\tilde{S}$ . Details are left to the reader.  $\square$

### 2.3 Existence of a piecewise constant function S.

In this section, we make an hypothesis which is a little bit restrictive but which has two main advantages: first, this assumption is easy to check numerically; second, it allows us to prove that the function  $S$  can be chosen in a very simple form, namely as a piecewise constant function. This choice is particularly convenient from the computational point of view.

First let us define the sets  $\Omega_i^\pm$ ,  $i = 0, \dots, 3$ , according to the following lemma: we denote  $\text{meas}(A)$  the measure of a set  $A$  with respect to  $dXdY$ . We have the lemmas:

**Lemma 2.7**

Assume  $F(X) \neq H(X)$  on a non negligible set of  $\mathbb{R}^d$  and define:

$$\Omega_0^\pm = \left\{ (X, Y) \in \mathbb{R}^{2d} \text{ s.t. } \pm G_0(X, Y) > 0 \right\}, \quad (2.28)$$

$$\Omega_i^\pm = \left\{ (X, Y) \in \mathbb{R}^{2d} \text{ s.t. } \pm G_i(X, Y) > 0 \right\}. \quad (2.29)$$

Then, we have  $\text{meas}(\Omega_i^\pm) > 0$ , for all  $i = 0 \dots 3$ .

The proof of lemma 2.7 follows exactly the same lines as that of lemma 2.5 and is therefore omitted. Next, we have:

**Lemma 2.8**

$\forall i = 0 \dots 3$ , at least one of the sets  $\Omega_i^+$  or  $\Omega_i^-$  defined by (2.28)-(2.29) with  $F$  and  $H$  given by (2.17) is negligible (and then, both  $\Omega_i^+$  and  $\Omega_i^-$  are negligible) or equivalently  $F = H$  a. e. if and only if  $f = M^f$ , for a. e.  $v \in \mathbb{R}^3$ .

**Proof :** Assume one of the sets  $\Omega_i^\pm$  or  $\Omega_i^\mp$  is negligible. From Lemma 2.7, we have that  $F$  is equal to  $H$ , a. e. . This can be written in the form:

$$f(w+z)f(w-z) = M^f(w+z)M^f(w-z), \quad \text{a.e. } (w, z) \in \mathbb{R}^6,$$

which implies

$$\int_{\mathbb{R}^3} f(v_1)f(v)dv_1 = \int_{\mathbb{R}^3} M^f(v_1)M^f(v)dv_1, \quad \text{a.e. } v \in \mathbb{R}^3,$$

and yields

$$f(v) = M^f(v), \quad \text{a.e. } v \in \mathbb{R}^3.$$

□

Now, we assume the following: any intersections of the 8 sets  $\Omega_i^\pm$  are non negligible i. e.

$$\text{meas}\left(\cap_{j=0}^3 \Omega_j^{\alpha_j}\right) > 0, \quad \forall (\alpha_j)_{j=0,\dots,3} \in \{\pm 1\}^4. \quad (2.30)$$

We prove that this condition allows to construct the function  $S$  such that (2.21) holds, as a product of characteristic functions. More precisely, we set

$$S(X, X') = \prod_{i=0}^4 \chi_i^\pm(X, X') \quad (2.31)$$

where the functions  $\chi_i^\pm$  are defined for  $i = 0 \dots 3$  by

$$\chi_i^\pm(X, X') = \begin{cases} a_i^+, & \forall (X, X') \in \Omega_i^+, \\ a_i^-, & \text{elsewhere} \end{cases} \quad (2.32)$$

with 8 positive real numbers  $a_i^\pm$  to be determined. Note that  $\Omega_i^- \subset (\Omega_i^+)^c$ ,  $\forall i$ , by construction. We define the following 64 constants, which depend on the distribution function  $f$  and of the choice of  $\tilde{C}$ :

$$I_i^\alpha = \int_{\Omega_0^{\alpha_0} \cap \Omega_1^{\alpha_1} \cap \Omega_2^{\alpha_2} \cap \Omega_3^{\alpha_3}} \tilde{C}(z, z') (F(X')H(X) - F(X)H(X')) (X_i - X'_i) dX dX', \quad (2.33)$$

$\forall i \in [0 \cdots 3]$  and  $\forall \alpha = (\alpha_i)_{i=0 \cdots 3} \in \{+1, -1\}^4$  and we use the convention  $X_0 = |X|^2$ , for the sake of simplicity. Using (2.30), we have the following sign properties:

$$\alpha_i I_i^\alpha > 0. \quad (2.34)$$

where  $\alpha = (\alpha_i)_{i=1 \cdots 4} \in (\{+1, -1\})^4$ . With these notations and for function  $S$  of the form (2.31), the conservation of momentum (for  $i = 1, 2, 3$ ) and energy (for  $i = 0$ ) can be written as the following system of equations in the variables  $a_i^\pm$  with  $n = 4$ :

$$\sum_{(\alpha) \in (\{+1, -1\})^n} I_i^\alpha \prod_{k=0}^{n-1} a_k^{\alpha_k} = 0, \quad i = 0 \cdots n-1. \quad (2.35)$$

We shall now prove that systems of  $n$  equations of this type for  $2n$  unknowns  $a_i^\pm$  have a non trivial positive solution (with the  $n$  supplementary following constraints  $a_i^- = 1$ )

**Proposition 2.9**

Let  $n \in \mathbb{N}$  and  $(I_i^\alpha)_{i=0 \cdots n-1, \alpha \in (\{+1, -1\})^n} \in (\mathbb{R})^{n2^n}$  such that (2.34) holds, be given. The system (2.35) has a non vanishing solution  $(a_i^\pm)_{i=0 \cdots n-1} \in (\mathbb{R}^{+*})^{2n}$ , not necessarily unique.

**Proof :** Let us fix  $a_i^- = 1$  and construct sequences  $((a_i)^n)_{n \in \mathbb{N}}$  which tend to  $a_i^+$  when  $n \rightarrow \infty$  as follow:

Initialize the sequences  $((a_i)^n)_{n \in \mathbb{N}}$  (for  $i = 1, \cdots, n$ ) by 1 i. e.  $a_i^+ = (a_i^0)^0 = 1$ . Assume  $a_i^m$  are known (in  $(\mathbb{R}^+)^n$ ), we compute  $a_i^{m+1}$  using the  $i^{th}$  equation in (2.35) where the  $a_j^+$  for  $j \neq i$  are equal to  $a_j^m$  and  $a_j^- = 1$  for all  $j = 0 \cdots n-1$ . We have

$$a_i^{m+1} = - \frac{\sum_{(\alpha)|\alpha_i=-1} I_i^\alpha \prod_{k=0}^{n-1} a_k^{\alpha_k}}{\sum_{(\alpha)|\alpha_i=+1} I_i^\alpha \prod_{k=0, k \neq i}^{n-1} a_k^{\alpha_k}}. \quad (2.36)$$

Note that by construction we have, for all  $m \in \mathbb{N}$   $a_i^{m+1} > 0$  (since  $\alpha_i I_i^\alpha > 0$  for any  $\alpha$  and the numerator contains  $\prod_{k=0}^{n-1} a_k^- = 1$ ). We skip the term  $a_i^- = 1$  in the numerator product. Then, for any of the  $2^{n-1}$  terms (associated with a particular  $\alpha^0$ ) in the numerator sum, we have its equivalent (i. e.  $\alpha$  with the same  $\alpha_j = \alpha_j^0$ ,  $j \neq i$ ) in the denominator, therefore:

$$\frac{I_i^{\alpha^0} \prod_{k=0, k \neq i}^{n-1} a_k^{\alpha_k^0}}{\sum_{(\alpha)|\alpha_i=+1} I_i^\alpha \prod_{k=0, k \neq i}^{n-1} a_k^{\alpha_k}} \leq \frac{\max_{(\alpha), i=0 \cdots n-1} |I_i^\alpha|}{\min_{(\alpha), i=0 \cdots n-1} |I_i^\alpha|} \stackrel{def}{=} R. \quad (2.37)$$

Then, by adding these  $2^{n-1}$  inequalities, we get an a priori upper bound for  $a_i^{m+1}$  defined by (2.36):

$$a_i^{m+1} < 2^{n-1} R. \quad (2.38)$$

Since the sum in the numerator includes a term  $\prod_{k=0}^{n-1} a_k^-$  equal to unity, we have also a lower bound:

$$a_i^{m+1} > \frac{R^{-1}}{(2^{n-1}R)^{n-1}}. \quad (2.39)$$

Hence, the above constructed sequence lies in a compact set of  $\mathbb{R}^n$  and, thus, up to an extraction, has a limit point which gives a solution for system (2.35) which is non trivial and positive thanks to (2.39). This ends the proof.  $\square$

The convergence of the whole sequence requires the uniqueness of the possible limit, which actually seems to hold, from the numerical point of view. This result may be obtained using algebra technics since (2.35) is a system of polynomial equations of degree  $n$  for the variables  $a_i^+$  and of degree 1 with respect to each of  $a_i^+$  separately. This remains to be proved.

We apply proposition 2.9 with  $I_i^\alpha$  given by (2.33) and  $n = 4$ . Then, we can also satisfy a normalization constraint of the form

$$\sum_{(\alpha) \in (\{+1, -1\})^4} \left( \int_{\cap_{i=0}^3 \Omega_i^{\alpha_i}} \tilde{C}(X, X') F(X) H(X') dX dX' \right) \prod_{k=0}^3 a_k^{\alpha_k} = N_{coll} > 0, \quad (2.40)$$

where the right hand side is the given number of collisions for the exact Boltzmann operator defined by:

$$N_{coll} \stackrel{def}{=} \int_{(\mathbb{R}^3)^3} \tilde{C}(z, z) (f(w+z)f(w-z)) dz dw = \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times S_2} C f f_1 dv dv_1 d\Omega > 0, \quad (2.41)$$

Therefore, it suffices to consider an arbitrary solution given by proposition 2.9 and to multiply the pair  $(a_1^+, a_1^-)$  by a constant in order to satisfy the normalization requirement (2.40).

Note that the coefficients  $I_i^\alpha$  are defined from the distribution function as integrals over  $\Omega_j^\pm$  (also defined from the distribution function). Thus, the assumption (2.30) on the intersections of  $\Omega_i^\pm$  naturally depends on the details of the distribution function and has to be checked. Efficient algorithms to achieve this computation practically still need to be designed.

## 2.4 Description of the energy mollified operator

In this section, we prove that a generalized Boltzmann operator can be constructed in the particular case in which only the microscopic energy conservation is relaxed. More precisely, it consists in assuming that  $\delta(w, z, w', z')$  in (2.10) is of the form

$$\delta(w, z, w', z') = \delta_z(|z|^2, |z'|^2) \delta_0(w - w'), \quad (2.42)$$

where  $\delta_0$  is the delta function. Then, the conservation of momentum is automatically verified. The detailed balanced property (2.11) can be simplified using  $(w = w')$  and leads to

$$\exp\left(\frac{-|z|^2}{T_f}\right) \delta_z(z, z') = \exp\left(\frac{-|z'|^2}{T_f}\right) \delta_z(z', z) \quad (2.43)$$

where  $T_f$  is the temperature of  $f$  defined by (2.15). This condition (2.43) reads

$$\delta_z(z, z') = \exp\left(\frac{|z|^2 - |z'|^2}{2T_f}\right) S(z, z'), \quad (2.44)$$

with  $S(z, z')$  symmetric and positive. The conservation of energy can be written in the following form:

$$\begin{aligned} \int_{(\mathbf{R}^3)^3} \tilde{C}(z, z') (|z|^2 - |z'|^2) & \left( f(w+z)f(w-z) \exp\left(\frac{|z|^2 - |z'|^2}{2T_f}\right) - \right. \\ & \left. - f(w+z')f(w-z') \exp\left(\frac{|z'|^2 - |z|^2}{2T_f}\right) \right) S(z, z') dz dz' dw = 0, \end{aligned} \quad (2.45)$$

and, after integration with respect to the variable  $w$ ,

$$\begin{aligned} \int_{(\mathbf{R}^3)^2} \tilde{C}(z, z') (|z|^2 - |z'|^2) & \left( \alpha_f(z) \exp\left(\frac{-|z'|^2}{T_f}\right) - \alpha_f(z') \exp\left(\frac{-|z|^2}{T_f}\right) \right) \\ & \exp\left(\frac{(|z'|^2 + |z|^2)}{2T_f}\right) S(z, z') dz dz' = 0. \end{aligned} \quad (2.46)$$

with

$$\alpha_f(z) = \int_{\mathbf{R}^3} f(w+z)f(w-z)dw \quad (2.47)$$

A corollary of lemma 2.7 proves that the function inside the integral in (2.46)

$$G(z, z') = (|z|^2 - |z'|^2) \left( \alpha_f(z) \exp\left(\frac{|z|^2 - |z'|^2}{2T_f}\right) - \alpha_f(z') \exp\left(\frac{|z'|^2 - |z|^2}{2T_f}\right) \right),$$

is strictly positive (resp. negative) on a non negligible set denoted by  $\Omega^+$  (resp  $\Omega^-$ ) if and only if  $f$  is not almost everywhere equal to the Maxwellian  $M^f$ . Then, a positive function  $S$  such that the energy conservation holds can be found. It can be chosen constant on the set  $\Omega^\pm$  in the spirit of the section 2.3 without any supplementary hypothesis on the distribution function in this case. Then we have

**Theorem 2.10**

*There exists a positive function  $S$  such that the operator  $\tilde{Q}$  of the form (2.7) with  $\tilde{C}$  given by (2.8),  $\delta$  given by (2.42) and  $\delta_z$  by (2.44) verifies properties (P1)-(P2)-(P3).*

## 2.5 Convergence when $\varepsilon \rightarrow 0$ .

In this section, we intend to see if we can make the regularization depend on a (small) smoothing parameter  $\varepsilon$  such that (2.6) holds when  $\varepsilon \rightarrow 0$ . We restrict to the energy regularization investigated in the previous section. We now consider scattering cross sections of the form

$$S_\varepsilon(z, z') = S_\varepsilon(|z|^2, |z'|^2) = \xi_\varepsilon(|z|^2 - |z'|^2) \cdot \chi_\varepsilon(|z|^2, |z'|^2), \quad (2.48)$$

with  $\xi$  an even and positive function such that  $\xi_\varepsilon(t) = \frac{1}{\varepsilon} \xi\left(\frac{t}{\varepsilon}\right)$  and  $\int \xi(t) dt = 1$ . We shall now prove the

**Proposition 2.11**

A necessary condition for the energy operator  $\tilde{Q}^\varepsilon$  with  $S^\varepsilon$  given by (2.48) to be such that

$$\lim_{\varepsilon \rightarrow 0} \tilde{Q}^\varepsilon = Q(f, f), \text{ in } \mathcal{D}$$

(for smooth distribution function  $f$ ) is that

$$\int_0^\infty s c\left(\frac{s}{2}, \frac{s}{2}\right) \mu_f(s) ds = 0 \quad (2.49)$$

where  $\mu_f(s)$  a particular moment of  $f$  defined below. In the case where  $\tilde{c}$  is a constant, (Variable Hard Sphere models), condition (2.49) is written:

$$\int_{w,z} C(|z|) \left(|z| - \frac{T_f}{|z|}\right) f(w+z) f(w-z) dw dz = 0 \quad (2.50)$$

**Proof :** The conservation of energy can be written (after angular integration of the variables  $z$  and  $z'$ )

$$\begin{aligned} 0 = & \int_0^\infty \int_0^\infty \tilde{c}(|z|^2, |z'|^2) \left(|z|^2 - |z'|^2\right) \left( \bar{\alpha}_f(|z|^2) \exp\left(\frac{|z'|^2 - |z|^2}{2T_f}\right) - \right. \\ & \left. \bar{\alpha}_f(|z'|^2) \exp\left(\frac{|z|^2 - |z'|^2}{2T_f}\right) \right) S_\varepsilon(|z'|^2, |z|^2) |z|^2 dz |z'|^2 dz', \end{aligned} \quad (2.51)$$

where we assume that the scattering cross section verifies the simplifying assumption:

$$\tilde{C}(|z|\beta, |z'|\beta') = \tilde{c}(|z|^2, |z'|^2) \tilde{c}(\beta, \beta'), \quad (2.52)$$

with  $z = |z|\beta$ ,  $\beta \in S^2$ ,  $z' = |z'|\beta'$ ,  $\beta' \in S^2$  are the expressions of  $z$  and  $z'$  in spherical coordinates and where  $\bar{\alpha}_f$  is defined by

$$\bar{\alpha}_f(|z|^2) = \int_{\beta \in S^2} \int_{\beta' \in S^2} \int_{w \in \mathbb{R}^3} \tilde{c}(\beta, \beta') f(w + |z|\beta) f(w - |z|\beta) d\beta d\beta' dw. \quad (2.53)$$

We use the following change of variable  $u = |z|^2$ ,  $u' = |z'|^2$  in (2.51):

$$\begin{aligned} 0 = & \int_0^\infty \int_0^\infty \tilde{c}(u, u') (u - u') \left( \bar{\alpha}_f(u) \exp\left(\frac{u' - u}{2T_f}\right) - \bar{\alpha}_f(u') \exp\left(\frac{u - u'}{2T_f}\right) \right) \\ & \xi_\varepsilon(u - u') \chi_\varepsilon(u, u') \sqrt{u} du \sqrt{u'} du'. \end{aligned} \quad (2.54)$$

We set  $s = u + u'$  and  $t = u' - u$  and define

$$\bar{G}_0(s, t) = \sqrt{s^2 - t^2} \tilde{c}\left(\frac{s-t}{2}, \frac{s+t}{2}\right) \left( \bar{\alpha}_f\left(\frac{s-t}{2}\right) \exp\left(\frac{t}{2T_f}\right) - \bar{\alpha}_f\left(\frac{s+t}{2}\right) \exp\left(\frac{-t}{2T_f}\right) \right), \quad (2.55)$$

for all  $(s, t) \in \{s > 0, |t| \leq s\} \stackrel{\text{def}}{=} \Delta$ . The energy conservation now reads

$$\int_\Delta \bar{G}_0(s, t) \xi_\varepsilon(t) \chi_\varepsilon\left(\frac{s-t}{2}, \frac{s+t}{2}\right) ds dt = 0. \quad (2.56)$$

When  $\varepsilon \rightarrow 0$ , the function  $\xi_\varepsilon$  tends to a delta measure, and, by a classical result about smoothing kernels, the above constructed collision operator converges to the usual Boltzmann one, at least formally, provided first that

$$\bar{c}\left(\frac{s}{2}, \frac{s}{2}\right) = c(s) \quad (2.57)$$

where  $c(s)$  is the physical scattering cross section defined at (2.2) and second, that

$$\lim_{t \rightarrow 0} \chi_\varepsilon\left(\frac{s-t}{2}, \frac{s+t}{2}\right) = 1, \quad \forall s > 0, \quad (2.58)$$

uniformly for  $\varepsilon > 0$ . We now investigate under what condition it is possible to find a family of functions  $\chi_\varepsilon$  satisfying (2.58) which guarantee that the energy conservation (2.56) is satisfied. Let us suppose that (2.56) holds. Then, when  $t \rightarrow 0$ , we have, assuming enough regularity on the distribution function,

$$\lim_{t \rightarrow 0} \bar{G}_0(s, t) = \frac{1}{2} s t^2 \bar{c}\left(\frac{s}{2}, \frac{s}{2}\right) \mu_f(s) \quad (2.59)$$

with

$$\mu_f(s) = \bar{\alpha}_f'\left(\frac{s}{2}\right) + \frac{\bar{\alpha}_f\left(\frac{s}{2}\right)}{T_f} \quad (2.60)$$

Therefore, at the leading order when  $\varepsilon \rightarrow 0$ , (2.56) leads to

$$\int_0^\infty s \bar{c}\left(\frac{s}{2}, \frac{s}{2}\right) \mu_f(s) \int_{-s}^s \xi_\varepsilon(t) \chi_\varepsilon\left(\frac{s-t}{2}, \frac{s+t}{2}\right) t^2 dt ds = 0. \quad (2.61)$$

But, by (2.58), we can write

$$\chi_\varepsilon\left(\frac{s-t}{2}, \frac{s+t}{2}\right) = 1 + \eta_\varepsilon(s, t), \quad (2.62)$$

with for all  $s$ ,

$$\lim_{t \rightarrow 0} \eta_\varepsilon(s, t) = 0, \quad (2.63)$$

uniformly when  $\varepsilon \rightarrow 0$ . Inserting (2.62) in (2.61), we get

$$\int_0^\infty s \bar{c}\left(\frac{s}{2}, \frac{s}{2}\right) \mu_f(s) \left( \int_{-s}^s t^2 \xi_\varepsilon(t) dt \right) \left( 1 + \int_{-s}^s \eta_\varepsilon(s, t) d\nu_\varepsilon(t) \right) ds = 0, \quad (2.64)$$

with

$$d\nu_{\varepsilon, s}(t) = \frac{t^2 \xi_\varepsilon(t) dt}{\int_{-s}^s t^2 \xi_\varepsilon(t) dt}. \quad (2.65)$$

It is easy to see that

$$d\nu_{\varepsilon, s}(t) \rightarrow \delta(t), \quad \text{as } \varepsilon \rightarrow 0, \quad \forall s > 0, \quad (2.66)$$

vaguely. Therefore, with (2.63), we have

$$\int_{-s}^s \eta_\varepsilon(s, t) d\nu_{\varepsilon, s}(t) \rightarrow 0, \quad \text{as } \varepsilon \rightarrow 0. \quad (2.67)$$



$\forall s > 0$ , which by means of the Lebesgue theorem, implies:

$$\int_0^\infty s \bar{c}\left(\frac{s}{2}, \frac{s}{2}\right) \mu_f(s) ds = 0. \quad (2.68)$$

It is easily seen that, by the same argument as in lemma 2.7,  $\mu_f(s)$  is not of constant sign. Therefore, a function  $\bar{c}$  such that (2.68) is satisfied can always be found. However, here, we require (2.57) i. e.  $c = \bar{c}$  where  $c$  is the physical scattering cross section defined by (2.2). Therefore, (2.68) is no more a constraint on  $\bar{c}$ , but rather a constraint on  $f$ . And, obviously, this constraint is in general not satisfied.  $\square$

**Remark 2.12** *Note that  $\mu_f = 0$  if and only if  $f = M^f$  (at least when  $\tilde{C}$  is constant which is the case with Variable Hard Sphere (VHS) models). The constructed collision operator for the energy relaxation case is thus a good approximation of the Boltzmann operator for distribution functions close to equilibrium.*

### 3 Mollifying the masses during the collision.

The second regularization approach is based on the introduction of artificial masses belonging to a certain interval around the physical mass. The scattering cross section for collisions of particle with unphysical masses is designed to provide the properties (P1) – (P3). The main advantage of this method is that this generalized scattering cross section only depends on the first 3 moments of the distribution function (the number density, the mean velocity and the temperature). This is very interesting from the practical point of view since, for example in the homogeneous case, these corrections are computed once for all at the beginning. The principle of the mass regularization, is to act as if the particles do not have the same masses before and after the collision. For a similar approach devoted to multispecies flows, the reader can also refer to [6].

#### 3.1 Description of the collision process.

In this section, we describe the collision process. Let  $\vec{v}$  and  $\vec{v}_1$  be two incident velocities (in  $\mathbb{R}^3$ ) corresponding to two particles of masses  $m$  and  $m_1$ . We consider post collisional velocities  $\vec{v}'$  and  $\vec{v}'_1$  associated with particles of masses  $m'$  and  $m'_1$  such that conservation of momentum and energy hold for this particular collision (see (1.3) and (1.4), for comparison):

$$m\vec{v} + m_1\vec{v}_1 = m'\vec{v}' + m'_1\vec{v}'_1, \quad (3.1)$$

$$m|\vec{v}|^2 + m_1|\vec{v}_1|^2 = m'|\vec{v}'|^2 + m'_1|\vec{v}'_1|^2. \quad (3.2)$$

The case  $m = m'$  and  $m_1 = m'_1$  may be physically interpreted as a collision between particles of different species (and different masses:  $A + B \rightarrow A' + B'$ ), whereas the general case could be interpreted as a collision with chemical reactions ( $A + B \rightarrow C + D$ ), with a conservation of the total mass. However, in this paper, we shall distinguish the physical masses of the particles (which are the same) and the artificial masses (which serve us to generalize the collision operator). Although there is no

physical justification, this procedure allows to build a collision operator satisfying the properties (P1) – (P3) defined in section 1. We denote:

$$x = \frac{m}{m + m_1}, \quad 1 - x = \frac{m_1}{m + m_1}, \quad (3.3)$$

$$y = \frac{m'}{m' + m'_1}, \quad 1 - y = \frac{m'_1}{m' + m'_1}, \quad (3.4)$$

with  $(x, y) \in ]0, 1[^2$ . The conservation relations (3.1) and (3.2) can be written:

$$x\vec{v} + (1 - x)\vec{v}_1 = y\vec{v}' + (1 - y)\vec{v}'_1, \quad (3.5)$$

$$x|\vec{v}|^2 + (1 - x)|\vec{v}_1|^2 = y|\vec{v}'|^2 + (1 - y)|\vec{v}'_1|^2. \quad (3.6)$$

We search now to parametrize the velocities  $(\vec{v}', \vec{v}'_1)$  for a fixed pair  $(\vec{v}, \vec{v}_1)$  and fixed parameters  $x$  and  $y$ . We define the velocities of the pre- and post- collisional center of mass frames:

$$\vec{V}(x) = x\vec{v} + (1 - x)\vec{v}_1, \quad (3.7)$$

$$\vec{V}'(y) = y\vec{v}' + (1 - y)\vec{v}'_1. \quad (3.8)$$

To ensure momentum conservation it is enough to take (remember that  $\vec{V}'(y) = \vec{V}(x)$ ):

$$\vec{v}' = \vec{V}(x) + (1 - y)r\vec{\omega}, \quad (3.9)$$

$$\vec{v}'_1 = \vec{V}(x) - yr\vec{\omega}, \quad (3.10)$$

with  $\vec{\omega} \in S^2$  and  $r > 0$ ,  $S^2$  being the unit sphere of  $\mathbb{R}^3$  and  $r$  to be determined later on. We now impose the conservation of energy (3.6). We calculate the two quantities  $|\vec{v}'|^2$  and  $|\vec{v}'_1|^2$ , with (3.9) and (3.10):

$$|\vec{v}'|^2 = |\vec{V}(x)|^2 + 2r(1 - y)(\vec{\omega}, \vec{V}(x)) + (1 - y)^2r^2, \quad (3.11)$$

$$|\vec{v}'_1|^2 = |\vec{V}(x)|^2 + 2ry(\vec{\omega}, \vec{V}(x)) + y^2r^2, \quad (3.12)$$

with  $(., .)$  the scalar product of two vectors of  $\mathbb{R}^3$ . We deduce that:

$$y|\vec{v}'|^2 + (1 - y)|\vec{v}'_1|^2 = |\vec{V}(x)|^2 + y(1 - y)r^2. \quad (3.13)$$

We shall compute the quantity  $r$  using the conservation of energy, in terms of  $\vec{v}_1$ ,  $\vec{v}_3$ ,  $x$  and  $y$ :

$$|\vec{V}(x)|^2 + y(1 - y)r^2 = |\vec{V}(x)|^2 + x(1 - x)|\vec{v} - \vec{v}_1|^2. \quad (3.14)$$

This equation has a unique solution  $r > 0$  given by

$$r \stackrel{def}{=} \left( \frac{x(1 - x)}{y(1 - y)} \right)^{\frac{1}{2}} |\vec{v} - \vec{v}_1|. \quad (3.15)$$

### 3.2 Microreversibility and invariant measures.

We define the transformation  $\mathcal{T}$  as follows (the exponent  $T$  denotes the transpose of a vector):

**Definition 3.1**

To a given  $(\vec{v}, \vec{v}_1) \in (\mathbb{R}^3)^2$ ,  $\vec{\omega} \in S^2$ ,  $(x, y) \in ]0, 1[^2$ , the transformation  $\mathcal{T}$  associates  $(\vec{v}', \vec{v}_1') \in (\mathbb{R}^3)^2$ ,  $\vec{\omega}' \in S^2$ ,  $(x', y') \in ]0, 1[^2$  such that

$$\mathcal{T}(\vec{v}, \vec{v}_1, x, y, \omega)^T \stackrel{\text{def}}{=} (\vec{v}', \vec{v}_1', y, x, \frac{\vec{v} - \vec{v}_1}{|\vec{v} - \vec{v}_1|})^T \quad (3.16)$$

where  $(\vec{v}', \vec{v}_1')$  are defined by (3.9)-(3.10). This transformation describes the collision process.

We also define the space of collisional parameters by

$$\mathcal{E} = \{(\vec{v}, \vec{v}_1, x, y, \vec{\omega})^T \in (\mathbb{R}^3)^2 \times S^2 \times ]0, 1[^2\}. \quad (3.17)$$

This transformation can be decomposed in the center of mass frame according to  $\mathcal{T} = \Phi^{-1} \circ \mathcal{C} \circ \Phi$  with  $\mathcal{C}$  the collision process in the center of mass frame and  $\Phi$  being defined by

**Definition 3.2**

$\forall (\vec{v}, \vec{v}_1, x, y, \vec{\omega})^T \in \mathcal{E}$ , we define

$$\Phi(\vec{v}, \vec{v}_1, x, y, \vec{\omega})^T \stackrel{\text{def}}{=} (V(\vec{x}) = x\vec{v} + (1-x)\vec{v}_1, \vec{g} = \vec{v} - \vec{v}_1, x, y, \vec{\omega})^T. \quad (3.18)$$

Its Jacobian determinant verifies:

$$\det(\partial\mathcal{T}) = \det(\partial\Phi^{-1}) \circ \det(\partial\mathcal{C}) \circ \det(\partial\Phi).$$

We calculate the Jacobian of the transformation  $\Phi$ :

$$|\det(\partial\Phi)| = 1.$$

The Jacobian of  $\mathcal{T}$  can be written:

$$|\det(\partial\mathcal{T}(\mathcal{V}))| = \sqrt{\frac{|g|^2}{g_y^2 + g_z^2}} \sin(\theta). \quad (3.19)$$

with  $\vec{g} = (g_x, g_y, g_z)$  and with the classical parametrization of  $\vec{\omega}$

$$\vec{\omega} = (\cos(\theta), \sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi)). \quad (3.20)$$

Parametrizing  $\vec{\omega}' = \frac{v - v_1}{|v - v_1|}$  by

$$\vec{\omega}' = (\cos(\theta'), \sin(\theta')\cos(\phi'), \sin(\theta')\sin(\phi')). \quad (3.21)$$

and remarking that

$$\sqrt{\frac{g_y^2 + g_z^2}{|g|^2}} = \sin(\theta'),$$

we can put the Jacobian of  $\mathcal{T}$  in the form

$$\det(\partial\mathcal{T}(\mathcal{U})) = \frac{J(\mathcal{U})}{J(\mathcal{U}')}$$

with  $\mathcal{U} = (v, v_1, x, y, \omega)$ ,  $\mathcal{U}' = \mathcal{T}(\mathcal{U})$  and

$$J(\mathcal{U}) = (x(1-x))^{\frac{3}{2}} \sin(\theta).$$

Now we set, for simplicity

$$p(x) = 32(x-x^2)^{\frac{3}{2}} \quad (3.22)$$

Note that the factor 32 is chosen such that  $p(\frac{1}{2}) = 1$  since it will allow to recover the standard Boltzmann operator in a following section. We deduce that

**Proposition 3.3**

*The following measure*

$$p(x) \sin(\theta) dx dy dv_1 dv_3 d\vec{\omega}, \quad (3.23)$$

*is invariant under the transformation  $\mathcal{T}$ .*

**Remark 3.4** *Let  $\mathcal{T}$  be an involutive differentiable application of  $\mathcal{E}$ . Then, the measure  $\sqrt{|\det(\partial\mathcal{T}(X))|} dX$  is invariant under the transformation  $\mathcal{T}$ . Indeed, by the use of the property  $\mathcal{T} = \mathcal{T}^{-1}$ , we deduce*

$$\partial\mathcal{T}(X) = \left(\partial\mathcal{T}^{-1}(\mathcal{T}(X))\right)^{-1}. \quad (3.24)$$

*By taking the determinant of each member and by putting  $X' = \mathcal{T}(X)$*

$$\det(\partial\mathcal{T}(X)) = \frac{1}{\det(\partial\mathcal{T}(X'))}, \quad (3.25)$$

*then we have*

$$dX' = d(\mathcal{T}(X)) = \det(\partial\mathcal{T}(X)) dX = \sqrt{\frac{|\det(\partial\mathcal{T}(X))|}{|\det(\partial\mathcal{T}(X'))|}} dX. \quad (3.26)$$

*Invariants quantities by the transformation  $\mathcal{T}$  can be easily built: let  $h$  be defined on  $\mathcal{E}$  and taking values in  $\mathbb{R}^d$  and  $g$  be a real function. For example, we can consider the quantities of the form:*

$$g(\langle h(X), h(\mathcal{T}(X)) \rangle),$$

*where  $\langle x, y \rangle$  stands for the inner product in  $\mathbb{R}^d$ .*

### 3.3 Definition of the mass regularized operator.

We consider a sub interval  $\mathcal{I}$  of  $]0, 1[$  having the form  $\mathcal{I} = ]\eta, 1 - \eta[$ , with  $\frac{1}{2} > \eta > 0$  and 'cut off' functions of the form

$$h_\varepsilon(x - \frac{1}{2}) = \frac{1}{\varepsilon} \xi\left(\frac{x - \frac{1}{2}}{\varepsilon}\right), \quad (3.27)$$

with  $\xi(z)$  an even, positive and sufficiently smooth function verifying

$$\int_{z \in \mathcal{I}} \xi(z - \frac{1}{2}) dz = 1.$$

We consider also the function  $\chi$  defined by

$$\chi(a, b, c, d, x, y) = \begin{cases} 1 & \text{if } (x \ln(a) + (1-x) \ln(b) - y \ln(c) - (1-y) \ln(d)) (cd - ab) \leq 0 \\ 0 & \text{elsewhere} \end{cases} \quad (3.28)$$

#### Definition 3.5

For all  $\varepsilon > 0$  and for any given distribution function  $f$ , we define

$$\begin{aligned} \mathcal{C}_\varepsilon(f, f) &= \int_{\mathbb{R}^3} \int_{S^2} \int_{(x,y) \in \mathcal{I}^2} q\left(\frac{|\vec{v} - \vec{v}_1| + |\vec{v}' - \vec{v}'_1|}{2}, \vec{\omega}\right) \\ &\quad \chi\left(\frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y\right) \\ &\quad \left(\frac{M_1 M f' f'_1 - M'_1 M' f f_1}{\sqrt{M_1 M M'_1 M'}}\right) \\ &\quad h_\varepsilon(x - \frac{1}{2}) h_\varepsilon(y - \frac{1}{2}) d\vec{v}_1 d\vec{\omega} 2x p(x) dx dy, \end{aligned}$$

where  $(\vec{v}', \vec{v}'_1, \vec{\omega}', x', y') = \mathcal{T}(\vec{v}, \vec{v}_1, \vec{\omega}, x, y)$  ( $\mathcal{T}$  being defined in definition 3.1,  $q(u, \vec{\omega}) = u \sigma(u, \vec{\omega})$  and  $\sigma$  is the differential scattering and  $M = M^f$  is the Maxwellian with the same first five moments as  $f$ ).

We formally have:

#### Proposition 3.6

The limit when  $\varepsilon \rightarrow 0$  of  $\mathcal{C}_\varepsilon(f, f)$  given by definition 3.5 is the usual Boltzmann operator

$$\lim_{\varepsilon \rightarrow 0} \mathcal{C}_\varepsilon(f, f) = \int_{\mathbb{R}^3} \int_{S^2} (f' f'_1 - f f_1) q(|\vec{v} - \vec{v}_1|, \vec{\omega}) d\vec{v}_1 d\vec{\omega}, \text{ in } \mathcal{D}. \quad (3.29)$$

**Proof :** In the limit  $\varepsilon \rightarrow 0$ , we have, in the distributional sense

$$\lim_{\varepsilon \rightarrow 0} h_\varepsilon(x - \frac{1}{2}) = \delta_{\frac{1}{2}}, \quad (3.30)$$

where  $\delta_{\frac{1}{2}}$  is the delta measure located at  $x = \frac{1}{2}$  and, for  $x = y = \frac{1}{2}$  we have  $p(\frac{1}{2}) = 1$  and  $\chi(a, b, c, d, \frac{1}{2}, \frac{1}{2}) \equiv 1$ : indeed,  $\forall (a, b, c, d) \in \mathbb{R}_+^4$ , we verify:

$$\lim_{x \rightarrow \frac{1}{2}, y \rightarrow \frac{1}{2}} \chi(a, b, c, d, x, y) = 1. \quad (3.31)$$

This is obvious in the case  $ab = cd$  and, in the converse case,  $ab \neq cd$  and for  $(x, y)$  close enough of  $(\frac{1}{2}, \frac{1}{2})$ , we have  $\chi = 1$  by monotony of the logarithm:

$$(a - b)(\ln(b) - \ln(a)) \leq 0 \quad \forall a, b \geq 0.$$

Finally, for any Maxwellian, any 4-tuple of velocities  $(\vec{v}, \vec{v}', \vec{v}_1, \vec{v}_1')$  satisfying the conservation of momentum and energy (3.1)-(3.2) and for  $x = y = \frac{1}{2}$ , we have

$$M_1 M = M' M'_1.$$

This ends the proof.  $\square$

### 3.4 Properties of the operator $\mathcal{C}_\varepsilon$ .

We now check the properties (P1) – (P2) – (P3) for  $\mathcal{C}_\varepsilon(f, f)$  given by definition (3.5). For that purpose, we introduce the weak formulation of this operator and we symmetrize it as follows: by definition 3.5, momentum and energy conservation relation (3.3) (3.4) and proposition 3.3 on invariant measures, we have, for all test functions  $\phi$ :

$$\begin{aligned} \int_{\mathbf{R}^3} \mathcal{C}_\varepsilon(f, f) \phi(\vec{v}) d\vec{v} &= \frac{1}{2} \int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \int_{S^2} \int_{(x, y) \in \mathcal{I}^2} q\left(\frac{|\vec{v} - \vec{v}_1| + |\vec{v} - \vec{v}_1'|}{2}, \vec{\omega}\right) \\ &\quad \chi\left(\frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y\right) \\ &\quad \left(\frac{M_1 M f' f'_1 - M'_1 M' f f_1}{\sqrt{M_1 M M'_1 M'}}\right) (x\phi + (1-x)\phi_1 - y\phi' - (1-y)\phi'_1) \\ &\quad h_\varepsilon\left(x - \frac{1}{2}\right) h_\varepsilon\left(y - \frac{1}{2}\right) d\vec{v}_1 d\vec{\omega} d\vec{v}' p(x) dx dy. \end{aligned}$$

#### 3.4.1 Conservation laws i. e. (P1).

##### Proposition 3.7

*Conservations of mass, momentum and energy hold true*

$$\int_{\vec{v} \in \mathbf{R}^3} \begin{pmatrix} 1 \\ \vec{v} \\ |\vec{v}|^2 \end{pmatrix} \mathcal{C}_\varepsilon(f, f) d\vec{v} = \vec{0}, \quad (3.32)$$

**Proof :** By taking successively  $\phi = 1$ ,  $\vec{v}$  and  $|\vec{v}|^2$  in the above weak formulation, we can easily verify from (3.5) and (3.6) that (3.32) holds. This ends the proof.  $\square$

We note that the term

$$\chi\left(\frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y\right),$$

in the definition of the collision operator, does not play any role for the establishment of the conservation laws.

### 3.4.2 Entropy dissipation i. e. (P3).

#### Proposition 3.8

We have the following inequality

$$\int_{\mathbf{R}^3} \mathcal{C}_\varepsilon(f, f) \ln(f) d\vec{v} \leq 0. \quad (3.33)$$

**Proof :** Using the weak formulation of  $\mathcal{C}_\varepsilon(f, f)$  and replacing  $\phi$  by  $\ln(f)$  in it, we obtain

$$\begin{aligned} \int_{\mathbf{R}^3} \mathcal{C}_\varepsilon(f, f) \ln(f) d\vec{v} &= \frac{1}{2} \int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \int_{S^2} \int_{(x,y) \in \mathcal{I}^2} q\left(\frac{|\vec{v} - \vec{v}_1| + |\vec{v} - \vec{v}_1'|}{2}, \vec{\omega}\right) \\ &\quad \chi\left(\frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y\right) \\ &\quad \left(\frac{M_1 M f' f'_1 - M'_1 M' f f_1}{\sqrt{M_1 M M'_1 M'}}\right) \\ &\quad (x \ln(f) + (1-x) \ln(f_1) - y \ln(f') - (1-y) \ln(f'_1)) \\ &\quad h_\varepsilon\left(x - \frac{1}{2}\right) h_\varepsilon\left(y - \frac{1}{2}\right) d\vec{v}_1 d\vec{v} d\vec{\omega} p(x) dx dy. \end{aligned} \quad (3.34)$$

By the use of the conservations of mass, momentum and energy, we have that  $\ln(M)$  is an invariant for the collision operator, for any Maxwellian  $M$ . Therefore, (3.34) can be written as:

$$\begin{aligned} \int_{\mathbf{R}^3} \mathcal{C}_\varepsilon(f, f) \ln(f) d\vec{v} &= \frac{1}{2} \int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \int_{S^2} \int_{(x,y) \in \mathcal{I}^2} q\left(\frac{|\vec{v} - \vec{v}_1| + |\vec{v} - \vec{v}_1'|}{2}, \vec{\omega}\right) \\ &\quad \chi\left(\frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y\right) \\ &\quad \left(\frac{f' f'_1}{M' M'_1} - \frac{f f_1}{M M_1}\right) \sqrt{M_1 M M'_1 M'} \\ &\quad \left(x \ln\left(\frac{f}{M}\right) + (1-x) \ln\left(\frac{f_1}{M_1}\right) - y \ln\left(\frac{f'}{M'}\right) - (1-y) \ln\left(\frac{f'_1}{M'_1}\right)\right) \\ &\quad h_\varepsilon\left(x - \frac{1}{2}\right) h_\varepsilon\left(y - \frac{1}{2}\right) d\vec{v}_1 d\vec{v} d\vec{\omega} p(x) dx dy. \end{aligned}$$

The definition (3.28) of  $\chi\left(\frac{f}{M}, \frac{f_1}{M_1}, \frac{f'}{M'}, \frac{f'_1}{M'_1}, x, y\right)$  ensures artificially the positivity of the above expression and ends the proof.  $\square$

### 3.4.3 Maxwellian steady states i. e. (P2).

By construction, we easily verify

#### Proposition 3.9

$$f(\vec{v}) = M_{\rho, u, T}(\vec{v}) \Rightarrow \mathcal{C}_\varepsilon(f, f) = 0. \quad (3.35)$$

The converse implication is not clear, because the term  $\chi$  can vanish for some velocities even if the distribution function is not identically a Maxwellian. The only equilibrium states of an operator of the form

$$Q_{\alpha,\varepsilon}(f, f) = (1 - \alpha)\mathcal{C}_\varepsilon(f, f) + \alpha Q(f, f), \quad (3.36)$$

with  $\alpha \in ]0, 1[$ ,  $\mathcal{C}_\varepsilon(f, f)$  given by definition 3.5 and  $Q(f, f)$  the standard Boltzmann operator given by (3.29) are the Maxwellians. Indeed, since the only steady states of  $Q(f, f)$  are the Maxwellians, we have proved that  $Q_{\alpha,\varepsilon}$  satisfies properties (P1) – (P2) – (P3). The modification of  $\mathcal{C}_\varepsilon$  into  $Q_{\alpha,\varepsilon}$  is not unnatural from the numerical point of view. It can be seen as a splitting of the collision operator between the regularized operator which allows much more collisions and the standard operator which eliminates spurious steady states.

#### Acknowledgements.

The authors thank T. Lachand-Robert and P. Mironescu (Laboratory of numerical analysis, University of Paris 6) for their ideas concerning the Proposition 2.1.

## References

- [1] G. A. BIRD, Molecular Gas Dynamics and the direct simulation of gas flows, Clarendon Press, Oxford, 1994.
- [2] A. BOBYLEV, A. PALCZEWSKI, J. SCHNEIDER, *A consistency result for a discrete-velocity model of the Boltzmann equation*, Institute of Applied Mathematics, Warsaw University, 1995.
- [3] C. BORGNACKE, PS. LARSEN, *Statistical model for Monte-Carlo simulation of polyatomic gas mixtures*, Journal of Comp Phys, **18** p. 405-420, (1975).
- [4] J. F. BOURGAT, L. DESVILLETES, P. LE TALLEC, B. PERTHAME, *Microreversible collisions for polyatomic gases and Boltzmann's theorem*, Eur Journal of Mech, B fluids, (1994).
- [5] C. BUET, *A discrete-velocity scheme for the boltzmann operator of rarefied gas dynamics*, to be published in Transp. Theory. Stat. Phys.
- [6] C. BUET, S. CORDIER, *Discrete Velocity Methods for Boltzmann equation for multispecies flows*, in preparation.
- [7] C. CERCIGNANI, The Boltzmann Equation and its Applications, Springer, New York, (1988).
- [8] P. DEGOND, S. MAS GALLIC, *The weighted particle method for convection diffusion equations*, Math. Comp, **53**, pp 485-507 (part 1) and pp 509-525 (part 2), (1989)
- [9] L. DESVILLETES, *Sur un modele de type Borgnakke-Larsen conduisant a des lois d'energie non-linéaires en température pour les gaz parfaits polyatomiques*, Actes du workshop du GDR SPARCH, (1995).
- [10] R. GATIGNOL, Théorie cinétique des gaz á répartitions discrètes de vitesses, Springer, New York, (1975).



- [11] D. GOLDSTEIN, B. STURTEVANT, J. E. BROADWELL, *Investigations of the Motion of Discrete-Velocity Gases*, in "Rarefied Gas Dynamics: Theoretical and Computational Techniques", E. P. Muntz, D. P. Weaver and D. H. Campbell (eds), Progress in Astronautics and Aeronautics, 118, AIAA, Washington DC, (1989).
- [12] D. B. GOLDSTEIN, *Discrete-Velocity collision dynamics for polyatomic molecules*, Phys. Fluids A4, pp 1831-1839, (1992).
- [13] F. GROPENGIESSER, H. NEUNZERT, J. STRUCKMEIER, *Computational methods for the Boltzmann equation*. Venice 1989: The state of Art in Appl. and Industrial math., eds. R. Spigler, Kluwer acad. publ., (1990)
- [14] A. HARAUX, V. KOMORNIK, *Oscillations of Anharmonic Fourier Series and the Waves Equation*, Revista Matematica, Iberoamericana, 1 n°4, pp 57-77, (1985).
- [15] G. H. HARDY and E. M. WRIGHT, An introduction to the number theory, Clarendon Press, Oxford, (1938).
- [16] T. INAMURO, B. STURTEVANT, *Numerical Study of Discrete-Velocity Gases*, Phys. Fluids A2 pp 2196-2203, (1990).
- [17] S. MAS GALLIC, F. POUPAUD, *A deterministic particle method for the linearized Boltzmann operator*, Trans. Theory Stat. Phys. 17, 311-345., 4 (1987).
- [18] P. MIRONESCU and T. LACHAND-ROBERT, personal communication, unpublished.
- [19] K. NANBU, *Direct simulation schemes derived from the Boltzmann equation*, J. Phys, Japan 49 p. 2042, (1980).
- [20] K. NANBU, *Model kinetic equation for the distribution of discretized internal energy*, Math Methods and Models in the Applied Sci, (1992).
- [21] B. NICLOT, P. DEGOND, F. POUPAUD, *Deterministic particles simulations of the Boltzmann transport equation of semiconductors*, J. Comp. Phys., 78, pp 313-345, (1988).
- [22] F. ROGIER, J. SCHNEIDER, *A direct Method for solving the Boltzmann Equation*, Transp. Theory. Stat. Phys, 23, pp 313-338 (1994).
- [23] J. SCHNEIDER, *Une méthode déterministe pour la résolution de l'équation de Boltzmann*, Ph. D thesis, University Paris 6, (1993).
- [24] Z. TAN, P. L. VARGHESE, *The  $\Delta-\epsilon$  method for the Boltzmann equation*, J. Comput. Phys., 110, (1994).

# Conservative and Entropy Decaying Numerical Scheme for the Isotropic Fokker–Planck–Landau Equation

C. Buet\* and S. Cordier†,<sup>1</sup>

\*C.E.A., Bruyères le Châtel, France; and †Laboratoire d'Analyse Numérique, CNRS- URA 189,  
Université Paris VI, 75252 Paris, France  
E-mail: buet@bruyeres cea.fr and cordier@ann.jussieu.fr

Received July 31, 1997; revised April 6, 1998

---

Homogeneous Fokker–Planck–Landau equation denoted by FPLE is studied for Coulombian and isotropic distribution function, i.e. when the distribution function depends only on time and on the modulus of the velocity. We derive a new conservative and entropy decaying semi-discretized FPLE for which we prove the existence of global in time, positive. For the time-discretized equation, we give upper bound for the time step which guarantes positivity and entropy decay of the numerical solution. © 1998 Academic Press

**Key Words:** kinetic models; Fokker–Planck–Landau equation; system of ordinary differential equations; Cauchy problem; numerical scheme; entropy.

---

## 1. INTRODUCTION

The FPLE is commonly used in plasma physics when studying kinetical effects between charged particles under Coulomb interaction. The homogeneous isotropic FPLE describes thermalization processes of the plasma in isotropic situations for the velocity variable and independent of the space variable. Another interest of the FPLE is to produce precise solutions in order to study numerical schemes in the 3D velocity space [4, 5, 8, 16–18] or in the 2D axisymmetric case [15]. Indeed, no explicit solutions are known for the Coulomb potential case  $\gamma = -3$ , defined in Section 2, contrary to the Maxwellian case  $\gamma = 0$  [13]. There are also applications in the astrophysics field, where the FPLE is used for star cluster modelling [6, 7].

Existence results for the continuous FPLE can be found in [9, 10, 1]. These results can certainly be extended for the isotropic equation considered here.

<sup>1</sup> Corresponding author.

A conservative and entropy scheme for the (spherical and homogeneous) FPLE was first proposed by Berezin, Khudick, and Pekker [2]. They give an upper bound for the time step to ensure the decay of entropy without a complete proof of their assertion. Entropy decay is physically relevant and seems to prevent oscillations (as shown in the sequel on numerical examples and proved for the linear case in [4]). At the continuous level and for obvious physical reason, the solution remains positive at any time. Thus, the discretization must preserve this property and this does not appear clearly in [2]. See [4] for an example of a conservative discretization which does not preserve positivity for all positive initial data. In this work, we prove the positivity of the solution for the semi-discretized and time-discretized solution for arbitrarily large time.

The aim of this paper is to propose a new conservative and entropy decaying scheme for FPLE for which, first, we prove the existence of a unique and global in time solution for the semi-discretized problem and, second, for the time-discretized equation we exhibit an upper bound on the time step to ensure the positivity and the decay of the entropy. Moreover, we show that the cost of the numerical evaluation of this operator is proportional to the number of discretization points despite its quadratic structure. Let us also mention that this scheme can be considered on an arbitrary mesh, contrary to the discretization considered in [4, 5, 14]. This last property permits us to refine the mesh size for small velocity and, thus, to obtain more accurate solutions. However, some questions remain open like the long-time behaviour of the semi-discretized or time-discretized solution, although it is expected that the distribution function converges to the discretized Maxwellian.

## 2. THE HOMOGENEOUS AND ISOTROPIC FPLE

We denote by  $F(\mathbf{v}, t)$  the distribution function solution of the scaled integro-differential equation

$$\frac{\partial F}{\partial t} = Q(F, F) = \nabla_{\mathbf{v}} \cdot \left( \int_{\mathbb{R}^3} \Phi(\mathbf{v} - \mathbf{v}_*) ((\nabla_{\mathbf{v}} F) F_* - (\nabla_{\mathbf{v}_*} F) F) d\mathbf{v}_* \right), \quad (2.1)$$

where  $Q(F, F)$  is the Fokker-Planck collision operator written in the so called Landau form with the standard notations (for example  $F_* = F(\mathbf{v}_*, t)$ ) and  $\Phi(\mathbf{v})$  is the following  $3 \times 3$  matrix:

$$\Phi(\mathbf{v}) = |\mathbf{v}|^{\gamma+2} S(\mathbf{v}), \quad S(\mathbf{v}) = I_3 - \frac{\mathbf{v} \otimes \mathbf{v}}{|\mathbf{v}|^2}. \quad (2.2)$$

$S(\mathbf{v})$  is the orthogonal projector onto the plane orthogonal to  $\mathbf{v}$ .  $\gamma$  is a real parameter which leads to the usual classification in hard potentials ( $\gamma > 0$ ), maxwellian molecules ( $\gamma = 0$ ) or soft potentials ( $\gamma < 0$ ). This latter case involves the Coulomb case (i.e.,  $\gamma = -3$ ) which is of primary importance for plasma applications. The well-known physical properties of (2.1) are similar to that of the Boltzmann operator such as the decay of the entropy, the conservation of mass, momentum, and energy, and the characterization of the equilibrium states by Maxwellians. We refer to [8, 16] for a detailed presentation of this equation.

It can be easily check that isotropic initial data leads to an isotropic solution for the classical nonlinear FPLE. In other words, if the distribution function  $F(\mathbf{v}, t)$  depends only of the modulus of the velocity  $v = \|\mathbf{v}\|$  at time  $t = 0$ , then this holds for any arbitrary time  $t$ ; i.e., there exists a function  $f$  such that  $F(\mathbf{v}, t) = f(v, t)$  (see [2, 17, 18]). In the Coulomb case,

such isotropic distribution function  $f(\varepsilon, t)$ , where  $\varepsilon = v^2$  is the energy variable, satisfies a dimensionless equation of the form:

$$\frac{\partial f}{\partial t} = \frac{1}{\sqrt{\varepsilon}} \frac{\partial}{\partial \varepsilon} \int_0^\infty f(\varepsilon) f(\varepsilon') \left( \frac{\partial}{\partial \varepsilon} \ln f(\varepsilon) - \frac{\partial}{\partial \varepsilon} \ln f(\varepsilon') \right) k(\varepsilon, \varepsilon') d\varepsilon'. \quad (2.3)$$

For numerical simulations, we reduce the integration domain in FPLe to a bounded domain in the variable  $\varepsilon$  as in [2] :

$$\frac{\partial f}{\partial t} = \frac{1}{\sqrt{\varepsilon}} \frac{\partial}{\partial \varepsilon} \int_0^{\varepsilon_0} f(\varepsilon) f(\varepsilon') \left( \frac{\partial}{\partial \varepsilon} \ln f(\varepsilon) - \frac{\partial}{\partial \varepsilon} \ln f(\varepsilon') \right) k(\varepsilon, \varepsilon') d\varepsilon', \quad (2.4)$$

where  $k(\varepsilon, \varepsilon') = \inf(\varepsilon^{3/2}, (\varepsilon')^{3/2})$  and  $\varepsilon_0$  is choosen such that the distribution function is near zero outside the ball of radius  $\varepsilon_0$ . Physically,  $\varepsilon_0$  is choosen larger than the typical scaled energy. We refer to [2] for a physical justification of this scaling. This operator can be equivalently written in the following weak form: for any sufficiently smooth and decaying test function  $\phi(\varepsilon)$ ,

$$\begin{aligned} \int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon} d\varepsilon &= -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} f(\varepsilon) f(\varepsilon') \left( \frac{\partial \phi(\varepsilon)}{\partial \varepsilon} - \frac{\partial \phi(\varepsilon')}{\partial \varepsilon} \right) \\ &\quad \times \left( \frac{\partial \ln f(\varepsilon)}{\partial \varepsilon} - \frac{\partial \ln f(\varepsilon')}{\partial \varepsilon} \right) k(\varepsilon, \varepsilon') d\varepsilon' d\varepsilon. \end{aligned} \quad (2.5)$$

This operator satisfies the conservation of mass (resp. energy) by choosing  $\phi = 1$  (resp.  $\phi = \varepsilon$  in (2.5))

$$\rho = \int_0^{\varepsilon_0} f(\varepsilon) \sqrt{\varepsilon} d\varepsilon, \quad (2.6)$$

$$\rho E = \int_0^{\varepsilon_0} f(\varepsilon) \varepsilon^{3/2} d\varepsilon. \quad (2.7)$$

The entropy defined by

$$H = \int_0^{\varepsilon_0} f(\varepsilon) \ln(f(\varepsilon)) \sqrt{\varepsilon} d\varepsilon \quad (2.8)$$

decays with time (by letting  $\phi = \ln(f)$  in the weak formulation of FPLe) and satisfies the classical H theorem

$$\partial_t H = 0 \Leftrightarrow f = \exp(-A\varepsilon + B).$$

### 3. THE SEMI-DISCRETIZED FPLe

Let us introduce the discretization  $f_i = f(\varepsilon_i)$ , where  $(\varepsilon_i)_{i=1 \dots N}$  is an increasing sequence such that  $\varepsilon_1 = 0$ ,  $\varepsilon_N = \varepsilon_0$ , and  $(\Delta \varepsilon_i = (\varepsilon_{i+1} - \varepsilon_i))_{i=1 \dots N-1}$ , is also increasing. The  $\varepsilon$ -derivative are approximated according to the simplest choice of finite difference operator namely, we define for any discretized function  $(\phi_i)_{i=1 \dots N}$

$$D\phi_i = \frac{(\phi_{i+1} - \phi_i)}{\Delta \varepsilon_i}, \quad i = 1 \dots N-1.$$

Let us introduce some notations. We define  $\varepsilon_{i+1/2} = (\varepsilon_{i+1} + \varepsilon_i)/2$  and  $v_{i+1/2}$  as the mean value of the velocity on  $[\varepsilon_i, \varepsilon_{i+1}]$ , i.e.

$$v_{i+1/2} = \frac{1}{\Delta\varepsilon_i} \int_{\varepsilon_i}^{\varepsilon_{i+1}} \sqrt{\varepsilon} d\varepsilon = \frac{2}{3\Delta\varepsilon_i} (\varepsilon_{i+1}^{3/2} - \varepsilon_i^{3/2}).$$

Let us consider first the discretization of the expression  $\int_0^{\varepsilon_0} \phi \sqrt{\varepsilon} d\varepsilon$  for any function  $\phi$ . By writing

$$\int_0^{\varepsilon_0} \phi \sqrt{\varepsilon} d\varepsilon = \sum_{i=1}^{N-1} \int_{\varepsilon_i}^{\varepsilon_{i+1}} \phi \sqrt{\varepsilon} d\varepsilon$$

and using the trapezoidal quadrature formula with respect to the measure  $\sqrt{\varepsilon} d\varepsilon$ , we approximate it by

$$\sum_{i=1}^{N-1} \frac{1}{2} (\phi_i + \phi_{i+1}) v_{i+1/2} \Delta\varepsilon_i.$$

By factorizing the terms  $\phi_i$  in the above expression, we obtain

$$\frac{1}{2} \phi_1 v_{3/2} \Delta\varepsilon_1 + \frac{1}{2} \sum_{i=2}^{N-1} \phi_i (v_{i+1/2} \Delta\varepsilon_i + v_{i-1/2} \Delta\varepsilon_{i-1}) + \frac{1}{2} \phi_N v_{N-1/2} \Delta\varepsilon_N \stackrel{\text{def}}{=} \sum_{i=1}^N c_i \phi_i, \quad (3.1)$$

where  $c_i$  are such that  $c_1 = v_{3/2} \Delta\varepsilon_1/2 = \frac{1}{3} \varepsilon_2^{3/2}$ ,

$$c_i = \frac{1}{2} (v_{i+1/2} \Delta\varepsilon_i + v_{i-1/2} \Delta\varepsilon_{i-1}) = \frac{1}{3} (\varepsilon_{i+1}^{3/2} - \varepsilon_{i-1}^{3/2}),$$

for  $i = 2 \dots N-1$  and  $c_N = v_{N-1/2} \Delta\varepsilon_{N-1}/2 = \frac{1}{3} (\varepsilon_N^{3/2} - \varepsilon_{N-1}^{3/2})$ . Once applied to the left-hand side of (2.5) with  $(\partial f / \partial t) \phi$ , we obtain the discretization of  $\int_0^{\varepsilon_0} (\partial f / \partial t) \phi \sqrt{\varepsilon} d\varepsilon$  as  $\sum_{i=1}^N c_i (\partial f_i / \partial t) \phi_i$ . We now turn to the discretization of the right-hand side of (2.5),

$$\begin{aligned} (\text{r.h.s.}) &= -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \int_{\varepsilon_i}^{\varepsilon_{i+1}} \int_{\varepsilon_j}^{\varepsilon_{j+1}} f(\varepsilon) f(\varepsilon') \left( \frac{\partial}{\partial \varepsilon} \phi(\varepsilon) - \frac{\partial}{\partial \varepsilon} \phi(\varepsilon') \right) \\ &\quad \times \left( \frac{\partial}{\partial \varepsilon} \ln f(\varepsilon) - \frac{\partial}{\partial \varepsilon} \ln f(\varepsilon') \right) k(\varepsilon, \varepsilon') d\varepsilon' d\varepsilon. \end{aligned} \quad (3.2)$$

Using for each integrals of (3.2) a midpoint quadrature formula, we approximate (3.2) by

$$-\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_i g_j k_{i,j} \Delta\varepsilon_i \Delta\varepsilon_j (D\phi_i - D\phi_j) (D(\ln f)_i - D(\ln f)_j) \quad (3.3)$$

with

$$k_{i,j} = k(\varepsilon_{i+1/2}^{3/2}, \varepsilon_{j+1/2}^{3/2}),$$

and the terms  $g_i$  stand for a second-order approximation of the distribution function at the center of the interval  $[\varepsilon_i, \varepsilon_{i+1}]$ . In the paper of Berezin *et al.* [2], the terms  $g_i$  are taken

as an arithmetic mean of  $f_i$  and  $f_{i+1}$ . This yields a discrete model for which it cannot be proved that the distribution function remains positive as it must be. We take a second-order approximation as the harmonic mean; that is,

$$g_i \stackrel{\text{def}}{=} \frac{2}{1/f_i + 1/f_{i+1}} = \frac{2f_i f_{i+1}}{f_i + f_{i+1}}. \quad (3.4)$$

Such an approximation has been already used by the authors (see [4]) for the linear and 3D nonlinear cases of the Fokker–Planck–Landau equation and the resulting discretized models for which the existence of a global positive solution is proved. Note that  $D\phi_i$  is also a second-order approximation of the derivative in the center of the cell  $[\varepsilon_i, \varepsilon_{i+1}]$ . We shall denote by  $D_{i,j}$  the terms  $(D(\ln f)_i - D(\ln f)_j)$  for simplifying the notations. Hence, the weak formulation of the semi-discretized model reads

$$\sum_{i=1}^N c_i \frac{\partial f_i}{\partial t} \phi_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_i \Delta \varepsilon_j (D\phi_i - D\phi_j) D_{i,j}. \quad (3.5)$$

By factorizing the terms  $\phi_i$  in the right-hand side of (3.5), we get

$$(\text{r.h.s.}) = \sum_{i=2}^{N-1} \phi_i (p_i - p_{i-1}) + \phi_1 p_1 - \phi_N p_{N-1}$$

for all  $i = 1 \cdots N-1$ :

$$p_i \stackrel{\text{def}}{=} \sum_{j=1}^{N-1} g_i g_j k_{i,j} D_{i,j} \Delta \varepsilon_j. \quad (3.6)$$

Finally, by identifying the terms involving  $\phi_i$  in (3.5), we obtain the system of ordinary differential equations (which is of the same form as in the 3D case presented in [4, 5]),

$$\frac{\partial f_i}{\partial t} = \text{FP}_i^s, \quad i = 1 \cdots N, \quad (3.7)$$

with  $\text{FP}_1^s = p_1/c_1$ ,  $\text{FP}_i^s = (p_i - p_{i-1})/c_i$  for  $i = 2 \cdots N-1$ , and  $\text{FP}_N^s = -p_{N-1}/c_{N-1}$ . The conservation laws imply that the discretized analogous of mass (2.6) and energy (2.7) defined as

$$\rho = \sum_{j=1}^N c_j f_j \quad (\text{mass}), \quad \rho E = \sum_{j=1}^N c_j f_j \varepsilon_j \quad (\text{energy}),$$

are conserved through the evolution of the system. These conservation properties can be easily checked by taking  $\phi_i = 1$  and  $\phi_i = \varepsilon_i$  in (3.5). Moreover, the entropy decays using the discretized definition of the entropy (in the spherical case)

$$H = H(f_i) \stackrel{\text{def}}{=} \sum_{j=1}^N c_j f_j \ln(f_j). \quad (3.8)$$

The verification is straightforward using the weak discretized formulation (3.5) with test function  $\phi_i = \ln(f_i)$ . Note that in the present case, the conservations and entropy decay

hold whatever the discretization grid is uniform or not, which is not the case in the 3D case [4, 5].

The existence of a positive global in time solution for this system follows exactly the same line as the one of the full 3D system [4].

**THEOREM 3.1.** *The Cauchy problem for the differential equation (3.7) with a strictly positive initial data admits a unique, positive and entropy solution for any time.*

*Proof.* The existence and unicity of the solution for small time is obtained using classical Cauchy Lipschitz theorem. Indeed, there is no singularity in this system in the logarithmic terms, using  $f_i^0 > 0$ . Thus, the existence of a solution global in time holds, provided that the solution cannot vanish in finite time at some points. We follow exactly the same lines as for the full 3D system [4]. Using mass conservation, showing that the  $f_i$ 's cannot vanish in finite time is equivalent to checking that the function

$$K = \sup_{i=1}^{N-1} \left( \left| \frac{f_i}{f_{i+1}} \right|, \left| \frac{f_{i+1}}{f_i} \right| \right) \quad (3.9)$$

remains bounded in finite time. This function is convenient since these ratios actually appear in the  $D(\ln f)_i$  terms. We have the following estimates (which are no more true with the arithmetic average instead of (3.4)):

$$0 \leq g_i \leq 2f_{i+1} \quad \text{or} \quad 2f_i \quad \forall i = 1 \cdots N-1. \quad (3.10)$$

Using (3.9) and the mass conservation, we have the estimate for the terms  $p_i$ ,

$$|p_i| \leq C g_i \ln(K), \quad (3.11)$$

where  $C$  is a generic constant throughout the rest of the proof, depending on the number of grid points  $N$ , domain size  $\varepsilon_0$ , the grid  $\varepsilon_i$ , and the initial data  $(f_i^0)_{i \in I}$ . Indeed, we have the following upper bounds:

$$|D_{i,j}| \leq 2 \sup_{i=1 \cdots N-1} |D(\ln f)_i| \leq 2 \ln(K), \quad i = 1 \cdots N-1,$$

$$\sum_{j=1}^{N-1} g_j k_{i,j} \Delta \varepsilon_j \leq \sum_{j=1}^{N-1} g_j \varepsilon_j^{3/2} \Delta \varepsilon_j \leq 2 \sum_{j=1}^{N-1} g_j \varepsilon_{j+1} c_j \leq 4 \sum_{j=1}^{N-1} f_{j+1} \varepsilon_{j+1} c_{j+1} \leq 4\rho E.$$

Note that the inequality

$$\varepsilon_j^{3/2} \Delta \varepsilon_j \leq 2\varepsilon_{j+1} c_j, \quad (3.12)$$

has been used which is equivalent to

$$\frac{3\varepsilon_j^{3/2}(\varepsilon_{j+1} - \varepsilon_j)}{\varepsilon_{j+1}(\varepsilon_{j+1}^{3/2} - \varepsilon_j^{3/2})} \leq 3 \sup_{x \in [0,1]} \frac{x^{3/2}(1-x)}{1-x^{3/2}} = 2.$$

Then, using (3.10), we have

$$\left| \frac{p_i - p_{i-1}}{c_i} \right| \leq C \ln(K) f_i. \quad (3.13)$$

Thus, we have for any  $i = 1 \dots N - 1$

$$\frac{\partial(f_i/f_{i+1})}{\partial t} = \frac{1}{f_{i+1}} \frac{\partial f_i}{\partial t} - \frac{f_i}{f_{i+1}^2} \frac{\partial f_{i+1}}{\partial t}.$$

Finally, using (3.13), we have

$$\left| \frac{\partial K}{\partial t} \right| \leq C K \ln(K), \quad (3.14)$$

which implies  $K(t) \leq K(0) \exp(\exp(Ct))$  and this ends the proof.  $\blacksquare$

*Remark 3.2.* Taking an arithmetic mean for  $g_i$  terms, that is  $g_i = (f_i + f_{i+1})/2$  (like in the work of Berezin *et al.*; see [2]) leads to the function  $K$  (see [4]) for an estimate of the form

$$\left| \frac{\partial K}{\partial t} \right| \leq C K^2 \ln(K). \quad (3.15)$$

Since this differential equation has no global solution in time, it cannot be proved that the semi-discretization described in [2] has a global positive solution.

*Remark 3.3.* An alternative proof can be given following the ideas presented in next section (see Proposition 4.1). Indeed, we show that the discrete collision term can always be written as

$$\text{FP}_i^s = G_i + K_i f_i,$$

where  $G_i$  is positive (gain term) and  $K_i$  is bounded by some constant  $C$ . So that for all  $i$  we have

$$\frac{df_i}{dt} \geq -C f_i.$$

Such inequality implies that the weights  $f_i$  cannot vanish in finite time.

#### 4. THE TIME-DISCRETIZED FPLe

In this section, the bars denote the various quantities (like  $f_i$ ) at time  $t_{n+1} = t_n + \Delta t$  defined recursively. Let us introduce the following time explicit scheme

$$\bar{f}_i = f_i + \Delta t \text{FP}_i^s, \quad (4.1)$$

where  $\text{FP}_i^s$  is defined by (3.7), of the form  $p_i - p_{i-1}/c_i$  for  $i = 2, \dots, N - 1$ , and  $p_i$  can be written in the form

$$p_i = g_i (D(\ln f)_i A_i - B_i) \quad \forall i = 1 \dots N - 1,$$

with

$$A_i = \sum_{j=1}^{N-1} g_j k_{i,j} \Delta \varepsilon_j \quad \text{and} \quad B_i = \sum_{j=1}^{N-1} g_j D(\ln f)_j k_{i,j} \Delta \varepsilon_j. \quad (4.2)$$



#### 4.1. Cost and Implementation of the Algorithm

The particular form of the discrete function  $k_{i,j}$ ,

$$k_{i,j} = \begin{cases} \varepsilon_{i+1/2}^{3/2} & \forall i < j, \\ \varepsilon_{j+1/2}^{3/2} & \forall j \geq i, \end{cases}$$

permits us to evaluate all the  $N$  terms  $A_i$  and  $B_i$  in  $O(N)$  operations. Indeed, we have, using the definition of  $k_{i,j}$ ,

$$A_i = \varepsilon_{i+1/2}^{3/2} \sum_{N-1 \geq j > i} g_j \Delta \varepsilon_j + \sum_{1 \leq j \leq i} g_j \varepsilon_{j+1/2}^{3/2} \Delta \varepsilon_j \quad (4.3)$$

and

$$B_i = \varepsilon_{i+1/2}^{3/2} \sum_{N-1 \geq j > i} g_j D(\ln f) \Delta \varepsilon_j + \sum_{1 \leq j \leq i} g_j D(\ln f) \varepsilon_{j+1/2}^{3/2} \Delta \varepsilon_j. \quad (4.4)$$

Obviously,  $A_i$  and  $B_i$  can be evaluated using three loops. The detailed algorithm for the computation of all the terms  $p_i$  reads :

ALGORITHM 4.1.

```

 $\alpha_1 = g_1 * \varepsilon_{1+1/2}^{3/2} * \Delta \varepsilon_1;$ 
 $\gamma_1 := g_1 * D(\ln f)_1 * \varepsilon_{1+1/2}^{3/2} * \Delta \varepsilon_1;$ 
for  $i := 2$  to  $N - 1$  do
   $\alpha_i := \alpha_{i-1} + g_i * \varepsilon_{i+1/2}^{3/2} * \Delta \varepsilon_i;$ 
   $\gamma_i := \gamma_{i-1} + g_i * D(\ln f)_i * \varepsilon_{i+1/2}^{3/2} * \Delta \varepsilon_i;$ 
end for

 $\beta_{N-1} := g_{N-1} * \Delta \varepsilon_{N-1};$ 
 $\delta_{N-1} := g_{N-1} * D(\ln f)_{N-1} * \Delta \varepsilon_{N-1};$ 
for  $i := N - 2$  to  $1$  do
   $\beta_i := \beta_{i+1} + g_i * \Delta \varepsilon_i;$ 
   $\delta_i := \delta_{i+1} + g_i * D(\ln f)_i * \Delta \varepsilon_i;$ 
end for

for  $i := 1$  to  $N - 1$  do
   $A_i := \varepsilon_{i+1/2}^{3/2} * \beta_i + \alpha_i;$ 
   $B_i := \varepsilon_{i+1/2}^{3/2} * \delta_i + \gamma_i;$ 
   $p_i := g_i * (D(\ln f)_i * A_i + B_i);$ 
end for

```

#### 4.2. Time Step Restriction for Positivity and Entropy Decay

The main questions about the time explicit scheme (4.1) concern positivity and entropy decay property. By positivity of the scheme, we mean that the terms  $\bar{f}_i$  are positive if the terms  $f_i$  are positive and by entropy decay, the property  $H(\bar{f}_i) \leq H(f_i)$ , where the discretized entropy  $H$  is defined by (3.8).

We obtain a time step limitation in order that the scheme remains positive and the entropy decays. The last question is related to the series of time steps; its divergence provides a

positive and entropy decaying time-discretized solution for any arbitrary large time. We perform the analysis for two natural grids which are used for the numerical examples presented later.

The first one is a **uniform grid in velocity**, where the nodes of the grid are defined by the sequence  $\varepsilon_i = (i - 1)^2 \Delta v^2$  with  $\Delta v = \sqrt{\varepsilon_0}/(N - 1)$ ,  $N$  being the number of grid points. For this choice, the geometric quantities used in the definition of the scheme are

- $\Delta \varepsilon_i = (\Delta v)^2 (2i - 1)$ .
- $c_i = (3(i - 1)^2 + 1) \Delta v^3 / 3$  except for  $i = N$ , where  $c_N = (3N^2 - 9N + 7) \Delta v^3 / 3$ .
- $\varepsilon_{i+1/2} = (2i^2 - 2i + 1) \Delta v^2 / 2$ .

The second type is a **uniform grid in energy**, i.e.  $\varepsilon_i = (i - 1) \Delta \varepsilon$  with  $\Delta \varepsilon = \varepsilon_0 / (N - 1)$  and the geometric quantities used reads now:

- $\Delta \varepsilon_i = \Delta \varepsilon$ .
- $c_i = (i^{3/2} - (i - 2)^{3/2}) \Delta \varepsilon^{3/2} / 3$  except for  $i = 1$  and for  $i = N$  for which  $c_1 = \Delta \varepsilon^{3/2}$  and  $c_N = ((N - 1)^{3/2} - (N - 2)^{3/2}) \Delta \varepsilon^{3/2}$  respectively. It is easy to check that we have the following lower bound for  $c_i$  which will be usefull later:  $c_i \geq \Delta \varepsilon^{3/2} ((i - 1) + \sqrt{i(i - 1)}) / (3\sqrt{N})$  for  $i = 2 \cdots N - 1$ , and for  $i = N$ ,  $c_N \geq \Delta \varepsilon^{3/2} (N - 3/2 + \sqrt{(N - 1)(N - 2)}) / (3\sqrt{N})$ .
- $\varepsilon_{i+1/2} = (2i - 1) \Delta \varepsilon / 2$ .

For these two grids, we obtain sufficient conditions for the time step in order to ensure positivity and entropy decay. We summarize this result as

**PROPOSITION 4.1.** *For each grid considered above, there exists a constant  $C$  which depends only on the density  $\rho$ , the entropy  $H$ , and the length  $\varepsilon_0$ , such that the scheme (4.1) is positive and entropy decaying under a time step restriction of the form  $\Delta t \leq C \Delta v^2$  for the **uniform grid in velocity** or  $\Delta t \leq C \Delta \varepsilon^2$  for the **uniform grid in energy**.*

*Proof.* Let us first exhibit a sufficient condition on the time step to guarantee entropy decay. Suppose that there exists a time step  $\Delta t^0$  such that for all  $\Delta t \in [0, \Delta t^0[$  all the terms  $\bar{f}_j$  are positive. Then, using the definition (3.8), the entropy associated with the scheme (4.1) is

$$\bar{H} = H(\Delta t) = \sum_{j=1}^N c_j \bar{f}_j \ln(\bar{f}_j) \geq \sum_{j=1}^N c_j (f_j + \Delta t \text{FP}_j^s) \ln(f_j + \Delta t \text{FP}_j^s). \quad (4.5)$$

Then, we have, using the inequality  $\ln(1 + x) < x \quad \forall x > -1$  and the conservation of the mass,

$$H(\Delta t) \leq H(0) + \Delta t \sum_{i=1}^N c_i \text{FP}_i^s \ln(f_i) + \Delta t^2 \sum_{i=1}^N c_i (\text{FP}_i^s)^2 / f_i \stackrel{\text{def}}{=} \tilde{H}(\Delta t)$$

for all  $\Delta t \in [0, \Delta t^0[$ . Thus, a sufficient condition for the entropy decay is to choose  $\Delta t$  such that  $\tilde{H}(\Delta t) \leq \tilde{H}(0) = H(0)$ , or equivalently,

$$\Delta t \leq \frac{-\sum_{i=1}^N c_i \text{FP}_i^s \ln(f_i)}{\sum_{i=1}^N c_i (\text{FP}_i^s)^2 / f_i}.$$

By construction, we have

$$-\sum_{i=1}^N c_i \text{FP}_i^s \ln(f_i) = \frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_i \Delta \varepsilon_j D_{i,j}^2 \geq 0.$$

On the other hand, we have

$$\begin{aligned} \sum_{i=1}^N c_i (\text{FP}_i^s)^2 / f_i &\leq \frac{p_1^2}{c_1 f_1} + \frac{p_{N-1}^2}{f_N c_{N-1}} + \sum_{i=2}^{N-1} \frac{1}{c_i f_i} (p_i - p_{i-1})^2 \\ &\leq \frac{p_1^2}{c_1 f_1} + \frac{p_{N-1}^2}{f_N c_{N-1}} + \sum_{i=2}^{N-1} \frac{2}{c_i f_i} (p_i^2 + p_{i-1}^2). \end{aligned}$$

Using the definition of the term  $p_i$ , we have

$$p_i^2 = \left( \sum_{i=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_j D_{i,j} \right)^2 \leq \left( \sum_{i=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_j \right) \left( \sum_{i=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_j D_{i,j}^2 \right),$$

and using that  $g_i \leq 2f_i$ , we obtain

$$\begin{aligned} \frac{p_i^2}{f_i c_i} &\leq \frac{2}{c_i \Delta \varepsilon_i} \left( \sum_{i=1}^{N-1} g_j k_{i,j} \Delta \varepsilon_j \right) \left( \sum_{i=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_j \Delta \varepsilon_i D_{i,j}^2 \right) \\ &\leq \sup_{i=1 \dots N} \left( \frac{2}{c_i \Delta \varepsilon_i} \sum_{i=1}^{N-1} k_{i,j} g_j \Delta \varepsilon_j \right) \left( \sum_{i=1}^{N-1} g_i g_j k_{i,j} \Delta \varepsilon_j \Delta \varepsilon_i D_{i,j}^2 \right). \end{aligned}$$

The same estimate can be obtained for  $p_{i-1}/f_i c_i$ , since the sequence  $\Delta \varepsilon_i$  is increasing. By summing these inequalities, one obtains

$$\sum_{i=1}^N c_i (\text{FP}_i^s)^2 / f_i \leq \sup_{i=1 \dots N} \left( \frac{16}{c_i \Delta \varepsilon_i} \sum_{j=1}^{N-1} k_{i,j} g_j \Delta \varepsilon_j \right) \left( - \sum_{i=1}^N c_i \text{FP}_i^s \ln(f_i) \right).$$

Finally, the time step has to satisfy

$$\Delta t \cdot \sup_{i=1}^N \left( \frac{16}{c_i \Delta \varepsilon_i} \sum_{j=1}^{N-1} k_{i,j} g_j \Delta \varepsilon_j \right) \leq 1. \quad (4.6)$$

Equation (4.6) gives the time step limitation used for the numerical examples. We must now find an upper bound of the denominator of (4.6). We detail the majoration for the uniform grid in velocity, since for the uniform grid in energy it follows the same lines. The problem is to estimate the above terms independently of  $i$ . For the uniform grid in the velocity variable, using the expressions of  $\Delta \varepsilon_i$  and  $c_i$  and since  $\varepsilon_{i+1/2}^{3/2} \geq k_{i,j}$ , we have for the denominator of (4.6)

$$A_i = \left( \sum_{j=1}^{N-1} k_{i,j} g_j \Delta \varepsilon_j \right) \leq \varepsilon_{i+1/2}^{3/2} \left( \sum_{j=1}^{N-1} g_j \Delta \varepsilon_j \right). \quad (4.7)$$

Using the definitions of  $c_i$ ,  $\Delta\varepsilon_i$ , and  $\varepsilon_{i+1/2}$ , one also has

$$\frac{\varepsilon_{i+1/2}^{3/2}}{c_i \Delta\varepsilon_i} = \frac{3}{2\sqrt{2}} \left( \frac{(2i^2 - 2i + 1)^{3/2}}{(3(i-1)^2 + 1)(2i-1)} \right) \frac{1}{\Delta v^2}.$$

Since the term depending on  $i$  is bounded, we have a time step of the form

$$\Delta t \leq C \left( \sum_{j=1}^{N-1} g_j \Delta\varepsilon_j \right) (\Delta v)^2,$$

where  $C$  is a constant independent of the data. Let us now bound the term  $\sum_{j=1}^{N-1} g_j \Delta\varepsilon_j$ . By the Cauchy–Schwarz inequality, we have

$$\sum_{j=1}^{N-1} g_j \Delta\varepsilon_j \leq \sqrt{\sum_{j=1}^{N-1} g_j^2 c_j} \sqrt{\sum_{j=1}^{N-1} (\Delta\varepsilon_j^2 / c_j)}. \quad (4.8)$$

By replacing  $\Delta\varepsilon_j$  and  $c_j$  by their values, it is easy to check that  $\sqrt{\sum_{j=1}^{N-1} (\Delta\varepsilon_j^2 / c_j)}$  is bounded by a constant which depends only on the length of the domain  $\varepsilon_0$ . On the other hand, one defines the discrete  $L_2$  norm

$$\frac{1}{2} \sqrt{\sum_{j=1}^{N-1} g_j^2 c_j} \leq \sqrt{\sum_{j=1}^N f_j^2 c_j} \stackrel{\text{def}}{=} \|f\|_2.$$

Finally, the scheme is entropy decaying under a condition for the time step of the form

$$\Delta t \leq C(\varepsilon_0, \|f\|_2) \Delta v^2.$$

For the uniform grid in energy, as indicated above, the majoration can be carried out using the same techniques. This gives the inequalities

$$A_i = \left( \sum_{j=1}^{N-1} k_{i,j} g_j \Delta\varepsilon_j \right) \leq \varepsilon_{i+1/2} \left( \sum_{j=1}^{N-1} \varepsilon_{i+1/2}^{1/2} g_j \Delta\varepsilon_j \right), \quad (4.9)$$

instead of (4.7). It is also necessary to use the lower bound on  $c_i$ . We obtain the same kind of time step restriction for entropy decaying as for the uniform grid in velocity, but with  $\Delta\varepsilon^2$  instead of  $\Delta v^2$ .

Let us now exhibit a sufficient condition on the time step to guarantee the positivity of the scheme (4.1). Using the notation defined above, we can write the terms  $\text{FP}_i^s$  as a sum of a positive term  $G_i$  and a pseudo-loss term (which is not necessarily negative) of the form  $K_i f_i$ , with bounded coefficients  $K_i$ . Indeed,  $\text{FP}_i^s$  reads

$$\frac{1}{c_i} \left( \frac{A_i g_i}{\Delta\varepsilon_i} \ln \left( \frac{f_{i+1}}{f_i} \right) + \frac{A_{i-1} g_{i-1}}{\Delta\varepsilon_{i-1}} \ln \left( \frac{f_{i-1}}{f_i} \right) - g_i B_i - g_{i-1} B_{i-1} \right),$$

where  $A_i$  and  $B_i$  are defined by (4.3) and (4.2), respectively. First, it is easy to check that all the terms  $B_i$  are bounded; then, using (3.10),  $B_i g_i / (f_i c_i)$  are bounded and are taking into account in  $K_i$ . The same result holds for  $B_{i-1} g_{i-1} / (f_{i-1} c_{i-1})$ .

Consider now the term containing  $A_i$ . If  $f_{i+1} \geq f_i$ , this term is positive, then it is taken into account in the gain term  $G_i$ . On the contrary, since  $A_i \geq 0$  and  $g_i \geq 0$ , this term is negative and in such case, we have

$$\left| g_i \ln \left( \frac{f_{i+1}}{f_i} \right) \right| \leq 2 \sup_{1 \geq x \geq 0} |x \ln(x)| f_i = 2e^{-1} f_i \leq 2f_i,$$

$$\left| \frac{1}{c_i f_i \Delta \varepsilon_i} A_i g_i \ln \left( \frac{f_{i+1}}{f_i} \right) \right| \leq \frac{2A_i}{c_i \Delta \varepsilon_i}.$$

This term taken into account in  $K_i$ . It is straightforward to show the same result for

$$\frac{1}{c_i \Delta \varepsilon_{i-1}} A_{i-1} g_{i-1} \ln \left( \frac{f_{i-1}}{f_i} \right).$$

Therefore, we have

$$\text{FP}_i^s = G_i + K_i f_i$$

with  $G_i \geq 0$  and  $K_i$  bounded. Then,  $\bar{f}_i = f_i + \text{FP}_i^s$  is positive provided that

$$\Delta t \leq \left( \max_i |K_i| \right)^{-1},$$

and it is easy to check that

$$\max_i |K_i| \leq 2 \left( \max_i \left| \frac{A_i}{c_i \Delta \varepsilon_i} \right| + \max_i \left| \frac{B_i}{c_i} \right| \right).$$

Then, under the condition

$$\Delta t \leq \frac{1}{2} \left( \max_i \left| \frac{A_i}{c_i \Delta \varepsilon_i} \right| + \max_i \left| \frac{B_i}{c_i} \right| \right)^{-1}, \quad (4.10)$$

the scheme is positive. Let us now detail for the uniform grid in  $v$  such a time restriction. Recall the inequality obtained from (4.7)

$$\frac{A_i}{c_i \Delta \varepsilon_i} \leq C(\varepsilon_0, \|f\|_2) / (\Delta v)^2. \quad (4.11)$$

For the terms  $|B_i/c_i|$  we use

$$\left| \frac{B_i}{c_i} \right| \frac{\varepsilon_{i+1/2}}{c_i} \sum_{j=1}^{N-1} \varepsilon_{j+1/2}^{1/2} g_j |(D \ln f)_j| \Delta \varepsilon_j \leq \frac{C}{\Delta v} \sum_{j=1}^{N-1} \varepsilon_{j+1/2}^{1/2} g_j |(D \ln f)_j| \Delta \varepsilon_j. \quad (4.12)$$

Using (3.10) and the fact that the sequence  $\Delta \varepsilon_j$  is increasing, we have

$$\begin{aligned} \sum_{j=1}^{N-1} \varepsilon_{j+1/2}^{1/2} g_j |(D \ln f)_j| \Delta \varepsilon_j &\leq \sum_{j=1}^{N-1} \frac{\varepsilon_{j+1/2}^{1/2}}{\Delta \varepsilon_j} g_j (|\ln f_j| + |\ln f_{j+1}|) \Delta \varepsilon_j \\ &\leq 4 \sum_{j=1}^{N-1} \frac{\varepsilon_{j+1/2}^{1/2}}{\Delta \varepsilon_j} f_j |\ln f_j| \Delta \varepsilon_j + 2 f_N |\ln(f_N)| \varepsilon_{N-1/2}^{1/2} \\ &\leq \frac{C'}{\Delta v} \sum_{j=1}^N f_j |\ln f_j| \Delta \varepsilon_j. \end{aligned}$$

Then, using the Cauchy–Schwarz inequality as for (4.8), we obtain

$$\sum_{j=1}^{N-1} \Delta \varepsilon_j f_j |\ln f|_j \leq \sqrt{\sum_{j=1}^{N-1} f_j^2 |\ln f_j|^2 c_j} \sqrt{\sum_{j=1}^{N-1} \frac{\Delta \varepsilon_j^2}{c_j}} \leq C(\varepsilon_0, \|f \ln(f)\|_2). \quad (4.13)$$

Collecting all the results, we show that, in the case of an uniform grid in  $v$ , there exists a constant  $C$  such that for any time step satisfying

$$\Delta t \leq C(\varepsilon_0, \|f \ln(f)\|_2, \|f\|_2) \Delta v^2 \quad (4.14)$$

the scheme is positive. For the case of an uniform grid in  $\varepsilon$ , we do not detail the calculations. One uses an estimate of the form (4.7) instead of (4.11) to obtain an upper bound of  $A_i/(c_i \Delta \varepsilon_i)$ . For  $B_i/(c_i)$ , one proceeds exactly as for the uniform grid in velocity. Finally, we obtain (4.14) with  $\Delta \varepsilon^2$  instead of  $\Delta v^2$ .

For each type of grid, by taking  $C = \min(C_1, C_2)$ , we obtain the desired result. ■

*Remark 4.2.* On the numerical examples,  $\max_i(f_i)$  (and consequently the  $L_2$  norm  $\|f\|_2$ ) appears to be bounded not only uniformly in time (for a fixed  $N$ ), which can be proved using the mass conservation, but also independently of the mesh size. This remains to be proved in order to approach the problem of the convergence.

*Remark 4.3.* As we will see in the next section on a numerical example, preserving the positivity only, by taking a time step of the form  $\Delta t = \alpha \Delta t^0$ , where  $\Delta t^0$  is the maximum allowable time step satisfying

$$f_i + \Delta t^0 \text{FP}_i^s \geq 0 \quad \forall i \in I$$

with the CFL factor  $\alpha$  equal to 0.5, for example, does not permit to avoid oscillations. However, (4.6) and (4.10) yield to a nonoscillatory scheme even if there is no maximum principle for the nonlinear FPLe.

## 5. NUMERICAL TEST FOR THE FPLe

The numerical test presented now is extracted from the work of Rosenbluth *et al.* [19] and has been used by Larroche *et al.* [12] and Frenod and Lucquin [11] to test numerical methods for the Fokker–Planck–Landau equation. The initial data is given by

$$f^0(\varepsilon) = 0.01 \exp(-10[(\sqrt{\varepsilon} - 0.3)/0.3]^2). \quad (5.1)$$

We will show the entropy, the Linnick functionnal and the distribution function at time  $t \in \{9, 36, 81, 144, 225, 324, 441, 576, 729, 900\}$ . The Linnick functionnal is defined by

$$L(t) = \int_{v \in \mathbb{R}^3} \frac{(\nabla_v f)^2}{f} dv \quad (5.2)$$

and for an isotropic function this reduces to

$$L(t) = \int_{\varepsilon \geq 0} \left( \frac{\partial f}{\partial \varepsilon} \right)^2 \frac{\varepsilon^{3/2}}{f} d\varepsilon. \quad (5.3)$$

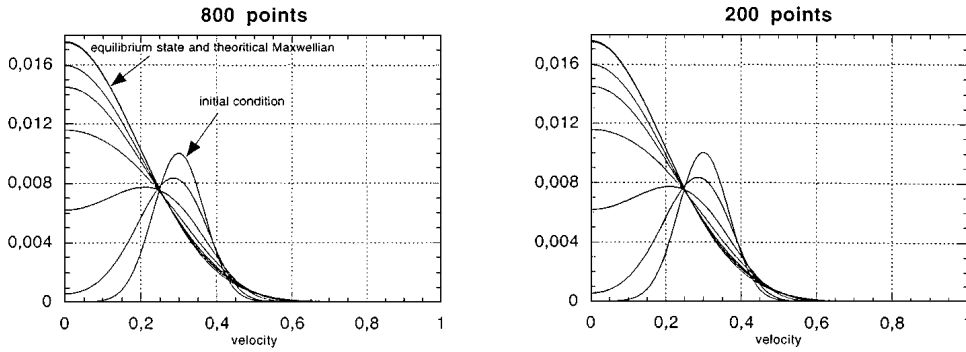


FIG. 1. Distribution function for the uniform grid in velocity.

The Linnick functional is known to be decreasing in time for the Boltzmann equation and the linear Fokker–Planck equation [3, 20, 9, 10]. Since the nonlinear FPLE is the so-called grazing collisions limit of the Boltzmann equation, one can expect that it is also decreasing for FPLE. For initial data (5.1) the Linnick functional is actually time-decreasing. Moreover, this functional illustrates very well the instabilities due to a nonentropy decaying scheme.

The tests run with two types of meshes, a uniform grid in the modulus of the velocity  $v = \sqrt{\varepsilon}$  and the other in the energy variable  $\varepsilon$  already described in the preceding section. For the two type of grids we take either 200 or 800 points of discretization and  $\varepsilon_0 = 1$ . The computations were carried out with a global time step equal to 1 and using subcycling inside each time step in order to preserve the positivity of the solution and to respect the entropy condition (4.6) with a CFL factor equal to 4.

The tests have been performed on a personal Apple computer, with a 160 Mhz PPC 603 chip. The cost of evaluating the solution during a time step for the uniform grid in energy (resp. in velocity) is about 0.09 s (resp. 0.05 s) for 200 cells and 7.46 s (resp. 3.71 s) for 800 cells; 900 time steps are performed. Note that the increase of the computational time is in good agreement with the theoretical estimate, since it is around a cubic function of the number  $N$  of points ( $\Delta t \leq C/N^2$  and linear cost  $O(N)$  of the algorithm).

Figures 1 and 2 show the time relaxation of the distribution function at various times: initial condition, time  $t \in \{9, 36, 81, 144, 225, 900\}$ , and the equilibrium state. Figures 3 and 4 show for this thermalization experiment the time relaxation of the  $H$  and Linnick

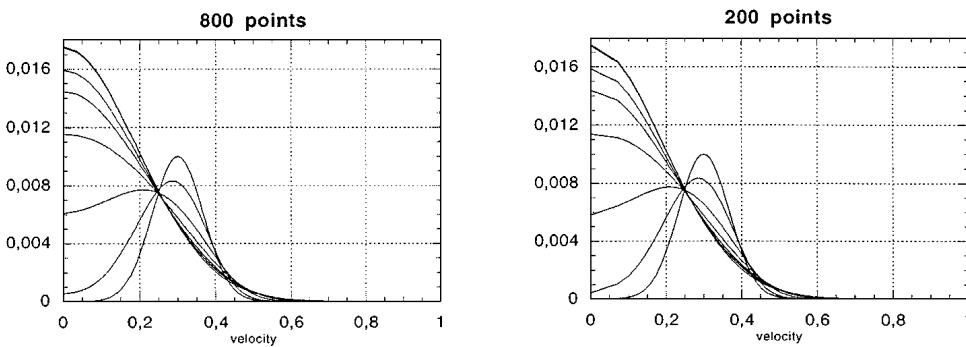


FIG. 2. Distribution function for the uniform grid in energy.

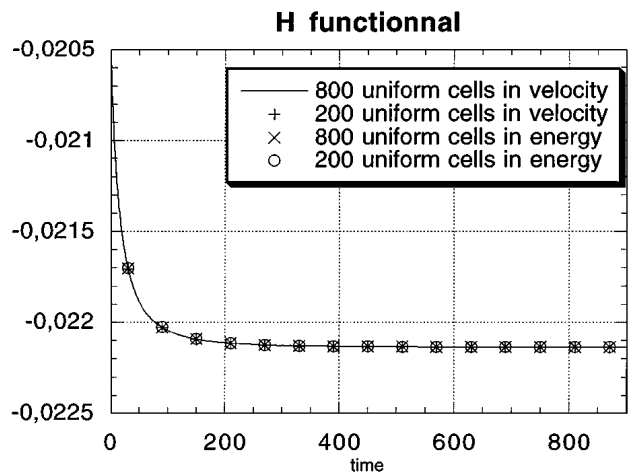


FIG. 3. Entropy.

functionnals. One can observe for these quantities that the results are very close to each other.

We also run a simulation using 200 cells uniformly distributed in energy and by only imposing the positivity of the solution using at each iteration half of the maximal allowable time step that guarantees the positivity. The run is only three times faster than the ones made with a CFL equal to 4 for the entropy condition 4.6.

Figures 5, 6, and 7 show the relaxation of the H and Linnick functionnals and the distribution function at  $t = 81$ . For the entropy, Fig. 5 compares the result with the entropy obtained for the same grid and the entropy-decaying scheme. The noisy curve corresponds to the “nonentropy-decaying” computation. For the first time steps, the two curves are very close. For large time, it is clear that the “nonentropy-decaying” computation has some difficulties reaching the equilibrium state, but the result does not seem too bad. On the Fig. 7 we plot the distribution function obtained with the two schemes at  $t = 81$ . The results are qualitatively the same for other times except for the small ones. The domain of oscillations

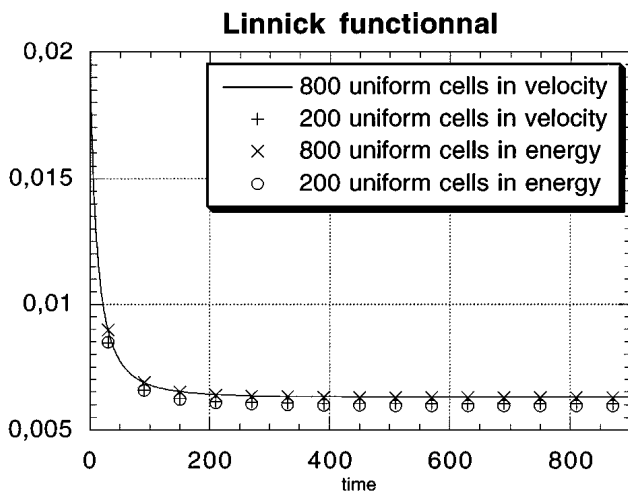


FIG. 4. Linnick fonctionnal.



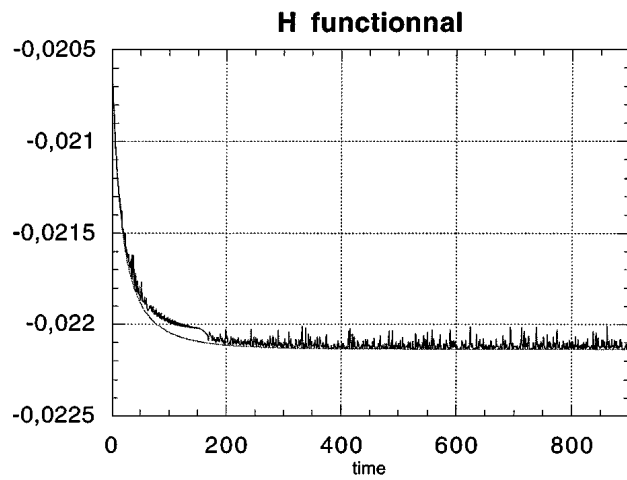


FIG. 5. Entropy: Comparison between entropy and nonentropy computations.

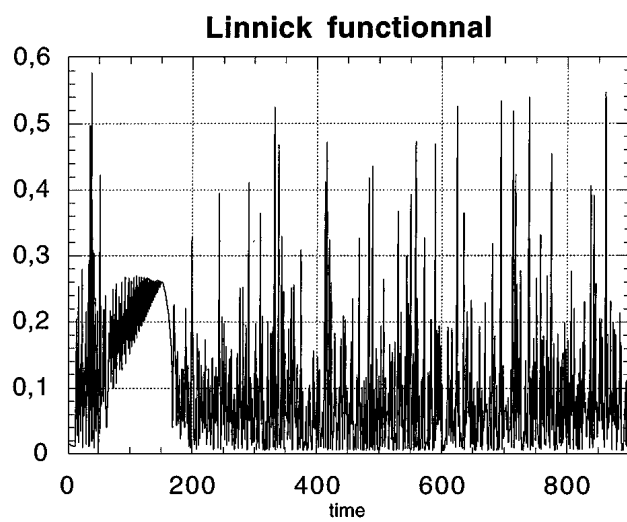


FIG. 6. Nonentropy computation: Linnick fonctionnal.

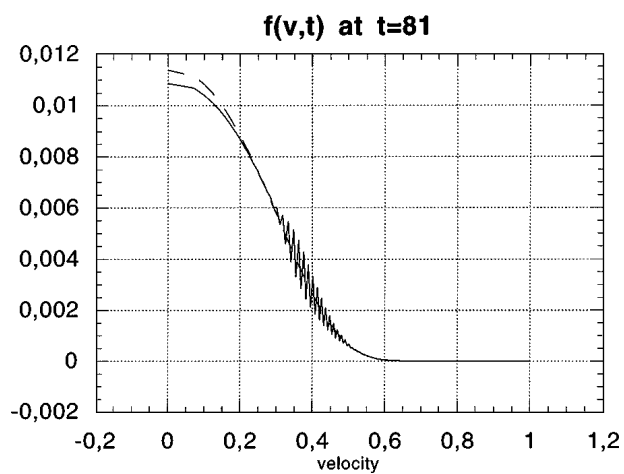


FIG. 7. Comparison between entropy and nonentropy computation for the distribution function at  $t = 81$ .

is independent of time and of the number of grid points. At large time, these oscillations persist but are damped. The main difference can be seen on the Linnick functional for which the results appear totally randomized, with no relationship to the "exact" (entropy decaying one) behaviour. Note that the same type of computation with a uniform grid in velocity produces a totally different behaviour. The distribution function is uniquely noisy near  $v = 0$  which explains, by recalling that the measure of integration is  $\sqrt{\varepsilon} d\varepsilon$ , that functionals of the distribution function have a correct time relaxation.

## 6. CONCLUSIONS

We provide for the simplest case of the isotropic, homogeneous, and Coulomb Fokker–Planck–Landau equation, a complete analysis of a conservative and entropy decaying numerical scheme. This scheme is very close to the scheme proposed in [2] since the modification consists in taking the harmonic average, instead of the arithmetic one for the evaluation of  $g_i$ . The main advantage of this scheme is to provide rigorously for the first time the existence of solutions for the semi-discretized model and time step restrictions to ensure positivity and entropy decaying of the scheme. We show that relaxing these time step conditions provides suspicious numerical results of FPLE for any time (see plots of Linnick functional).

We refer to [4] for similar analysis for the linear and nonlinear FPLE given by (2.1).

## REFERENCES

1. A. A. Arsenev and N. V. Peskov, On the existence of a generalized solution of Landau's equation, *USSR Comput. Maths Math. Phys.* **17**, 241 (1977).
2. Yu. A. Berezin, V. N. Khudick, and M. S. Pekker, Conservative finite difference schemes for the Fokker–Planck equation not violating the law of an increasing entropy, *J. Comput. Phys.* **69**, 163 (1987).
3. A. V. Bobylev and G. Toscani, On the generalization of the Boltzmann H-theorem for a spatially homogeneous Maxwell gas, *J. Math. Phys.* **33**, 7 (1992).
4. C. Buet and S. Cordier, Numerical analysis of conservatives and entropy schemes for the Fokker–Planck–Landau equation, *SIAM J. Numer. Anal.*, submitted.
5. C. Buet, S. Cordier, P. Degond, and M. Lemou, Fast algorithms for the the Fokker–Planck equation, *J. Comput. Phys.* **133**, 310 (1997).
6. H. Cohn, Numerical integration of the Fokker–Planck equation and the evolution of stars clusters, *Astrophys. J.* **234**, 1036 (1979).
7. H. Cohn, Late core collapse in star clusters and the gravothermal instability, *Astrophys. J.* **242**, 765 (1980).
8. P. Degond and B. Lucquin-Desreux, An entropy scheme for the Fokker–Planck collision of plasma kinetic theory, *Numer. Math.* **68**, 239 (1994).
9. L. Desvillettes and C. Villani, On the spatially homogeneous Landau equation for hard potentials. Part I. Existence, uniqueness, and smoothness, DMI, ENS de Paris, preprint.
10. L. Desvillettes and C. Villani, On the spatially homogeneous Landau equation for hard potentials. Part II. H-theorem and applications, DMI, ENS de Paris, preprint.
11. E. Frenod and B. Lucquin-Desreux, On conservative and entropic discrete axisymmetric Fokker–Planck operators, in preparation.
12. O. Larroche, Kinetic Simulations of a plasma collision experiment, *Phys. Fluids B* **5**, No. 8 (1993).
13. M. Lemou, Exact solutions of the Fokker–Planck equation, *C.R. Acad. Sci. Ser. I* **319**, 579 (1994).
14. M. Lemou, Multipole expansions for the Fokker–Planck equation, in preparation.
15. M. Lemou, Fast algorithm for the axisymmetric Fokker–Planck equation, in preparation.

16. B. Lucquin-Desreux, Discrétisation de l'opérateur de Fokker–Planck dans le cas homogène, *C.R. Acad. Sci. Paris Sér. I, A* **314**, 407 (1992).
17. M. S. Pekker and V. N. Khudik, Conservative difference schemes for the Fokker–Planck equation, *U.S.S.R. Comput. Math. Math. Phys.* **24**(3), 206 (1984).
18. I. F. Potapenko and V. A. Chuyanov, A completely conservative difference scheme for the two-dimensional Landau equation, *U.S.S.R. Comput. Math. Math. Phys.* **20**(2), 249 (1980).
19. M. N. Rosenbluth, W. Macdonald, and D. L. Judd, Fokker–Planck equation for an inverse-square force, *Phys. Rev.* **107**(1), 1 (1957).
20. G. Toscani, Entropy production and the rate of convergence to equilibrium for the Fokker–Planck equation, Univ. of Pavia, Dept. of Math., preprint.

## NUMERICAL ANALYSIS OF CONSERVATIVE AND ENTROPY SCHEMES FOR THE FOKKER–PLANCK–LANDAU EQUATION\*

C. BUET<sup>†</sup> AND S. CORDIER<sup>‡</sup>

**Abstract.** Conservatives and entropy schemes for the Fokker–Planck–Landau (FPL) equation are studied. We prove the existence of a unique positive and global in time solution for the homogeneous linear and nonlinear discretized (either in the velocity space or both in the velocity space and in time) FPL equation. The stability analysis of these schemes leads to sufficient conditions on the time-step that guarantee positivity and entropy decay.

**Key words.** kinetic models, Fokker–Planck–Landau equation, system of ordinary differential equations, Cauchy problem, numerical schemes

**AMS subject classifications.** 34A10, 65L05, 65M05, 82A45

**PII.** S0036142997322102

**1. Introduction.** The Fokker–Planck–Landau (FPL) equation describes the binary collisional effects (through long range Coulombian interaction) in a plasma [20] and can be derived from the Boltzmann equation in the grazing collision limit [11, 13, 1, 2]. Other applications may be found in astrophysics, for example, in the study of star cluster models [8, 9]. The linear FPL equation can also be used for diphasic flow modeling [10] and in space plasma physics for polar outflow modeling of a minor ion [23, 30]. Note that it is generally necessary to solve the nonlinear FPL equation to obtain quantitative agreement with experimental data [24]. The resulting prohibitive numerical cost forces many authors to consider simplified models, although rapid algorithms can now be used [7, 22].

In plasma physics, conservative schemes are of great importance since nonconservative designed schemes generally produce artificial heating or cooling of the plasma, as mentioned in [18] for the Boltzmann equation. Therefore, such methods require a great number of discretization points and thus, a prohibitive CPU time. We refer to [20] for such discretizations, which are based on the so-called Rosenbluth form of the FPL equation. In this paper, the major preoccupation is to reduce the cost of the algorithm using FFT method. In other words, entropy decay is important to ensure the thermalization of the plasma to the physically relevant temperature. Moreover, the methods considered here can be easily extended to multispecies plasma, as explained in [7], which is one of the major difficulties with the nonconservative schemes. Numerically, it can also be observed that entropy decay and positivity generally entail a nonoscillatory scheme. This property will be proved in the linear case. We refer to [6] for numerical evidence of these oscillations for the spherical Fokker–Planck equation when the entropy decay is not satisfied. Conservatives and entropy schemes for the nonlinear FPL equation can be only designed using the Landau and the so-called log formulation (which will be defined later on) of the equation, as proved by many

---

\*Received by the editors May 27, 1997; accepted for publication (in revised form) April 13, 1998; published electronically May 5, 1999.

<http://www.siam.org/journals/sinum/36-3/32210.html>

<sup>†</sup>C.E.A., Bruyères le Châtel, France (buet@bruyeres cea.fr).

<sup>‡</sup>Laboratoire d'Analyse Numérique, CNRS-URA 189, Université Paris VI, 75252 Paris, France (cordier@ann.jussieu.fr).

authors [12, 3, 27, 5, 25, 17, 26]. Indeed, we construct in the appendix a conservative scheme which is based on the Landau form but in the nonlog formulation for which a positive initial distribution function does not remain positive after an arbitrarily small time.

One open problem that is addressed here concerns existence of positive solutions for such discretizations. Another point of interest is the entropy decay of the fully discretized equation (i.e., also in time). Indeed, the entropy decay is guaranteed in the log Landau form only for the semidiscretized problem (i.e., discretized in the velocity space and continuous in time, which leads to a system of coupled nonlinear differential equations), but not for the fully discretized problem which is actually implemented in numerical simulations. Berezin, Khudick, and Pekker [3] provide an insight into these questions but their response remains incomplete; they give no rigorous proof either for the decay of the entropy for their time discretized scheme or for the existence of a positive (time-discretized or not) solution. In the present paper, we will prove some existence and stability results for the fully discretized and semidiscretized FPL equations in the homogeneous case, i.e., when there is no  $x$ -space dependence, using the Landau log formulation in both linear and nonlinear cases.

In section 2, we deal with the linear FPL equation, that is, a simpler situation in which it is easy to understand what happens when the Landau log formulation of the equation is discretized. We study the semidiscretized problem which leads to a system of ordinary differential equations (ODEs). These equations are nonlinear since the discretization is done on the log formulation in order to satisfy the entropy decay, as explained in the rest of this paper. We prove that this system still verifies a maximum principle and has a global in time solution with a uniform nonnegative lower bound. We show that the distribution function converges toward the discrete Maxwellian when  $t \rightarrow \infty$ . Then, we consider the time-discretized problem and we prove that the proposed explicit scheme has the same properties as the semidiscretized problem (and as the continuous one). More precisely, we construct a piecewise constant in time approximation of the velocity-discretized distribution function. This discretized solution remains positive, entropy decaying, nonoscillatory, and tends to the discrete Maxwellian for large time. These properties are achieved under some time-step restriction that provides a CFL-like condition for the scheme.

In section 3, we consider the three-dimensional nonlinear FPL equation, which is more interesting for plasma physics applications. We introduce a new conservative and entropy decaying discretization of the Fokker–Planck operator that guarantees the existence of a positive solution. More precisely, we prove that the semidiscretized problem has a global in time, positive, and entropy solution. However, we have no uniform, strictly positive lower bound for the solution in arbitrary large time and, for this reason, we cannot prove that the solution converges toward the velocity-discretized Maxwellian. Then, we consider the time-discretized problem. We prove that the sequence of time steps for which the solution remains positive at each iteration, forms a divergent series. Using this fact, we construct an approximate entropy solution for an arbitrarily large time. We also exhibit a lower bound for the time-step that ensures the decay of the entropy. Although this scheme is entropy decaying, the convergence of the discretized distribution function toward the discrete Maxwellian (or equivalently a lower-bound uniform in time on its weights) remains an open problem.

Finally, we point out that these first existence results of an entropy positive discretized solution can also be viewed as a first step toward proving any convergence or consistency result for these methods.

## 2. The linear FPL operator.

**2.1. The continuous homogeneous linear FPL equation.** Let us consider the linear FPL equation in dimension  $d$ , which is a second-order differential equation. This convection-diffusion equation can equivalently be written in the following forms (for strictly positive distribution functions):

$$\begin{aligned} (2.1) \quad \frac{\partial f}{\partial t} &= \nabla \cdot \left( (\vec{v} - \vec{u})f + T\vec{\nabla}_v f \right) \\ (2.2) \quad &= \nabla \cdot \left( Tf\vec{\nabla}_v \log(f/M_f) \right) \\ &= \nabla \cdot \left( Tf \left( \frac{(\vec{v} - \vec{u})}{T} + \vec{\nabla}_v \log f \right) \right) = \nabla \cdot \left( TM_f \vec{\nabla}_v (f/M_f) \right), \end{aligned}$$

where  $\nabla \cdot$  and  $\vec{\nabla}_v$  are the divergence and gradient operator, respectively,  $f(v, t)$  is the distribution function (with  $\vec{v} \in \mathbb{R}^d$ ,  $t > 0$ ).  $T$  represents a constant temperature,  $\vec{u}$  a constant velocity, and  $M_f$  the Maxwellian with the same first moment  $n_f$  as  $f$ , defined by

$$\begin{aligned} (2.3) \quad M_f(\vec{v}) &= \frac{n_f}{(2\pi T)^{d/2}} \exp\left(-\frac{\|\vec{v} - \vec{u}\|^2}{2T}\right), \\ n_f(t) &= \int_{\mathbb{R}^d} f(v, t) dv \quad (\text{mass}), \\ \vec{u}_f(t) &= \frac{1}{n_f(t)} \int_{\mathbb{R}^d} \vec{v} f(v, t) dv \quad (\text{velocity}), \\ T_f(t) &= \frac{1}{d n_f(t)} \int_{\mathbb{R}^d} \|\vec{v} - \vec{u}_f(t)\|^2 f(v, t) dv \quad (\text{temperature}). \end{aligned}$$

Equation (2.2) is called Landau-log formulation of the linear FPL equation. In the following, we assume  $T = 1$  for the sake of simplicity. For any test-function  $\phi(v)$ , we have, integrating (2.2) by part

$$(2.4) \quad \int_{\mathbb{R}^d} \frac{\partial f}{\partial t} \phi(v) dv = - \int_{\mathbb{R}^d} f \vec{\nabla}_v \phi^T \cdot \vec{\nabla}_v \log(f/M_f) dv,$$

where  $x^T$  denotes the transpose vector. By taking  $\phi \equiv 1$ , we obtain the mass conservation which is the only conservation for this equation, i.e.,  $n_f(t) = n_f(0) = n$ . Thus, the Maxwellian is constant in time. Note that, in general, this equation does not preserve momentum and energy. Indeed, letting  $\phi = v$  in (2.4), we get

$$\frac{d}{dt} (\vec{u}_f(t) - \vec{u}) = -(\vec{u}_f(t) - \vec{u}),$$

where  $u$  is the constant velocity. Thus, we have

$$\vec{u}_f(t) = \vec{u}_f(0) \exp(-t) + \vec{u} (1 - \exp(-t)).$$

Therefore, the velocity is not constant except if  $\vec{u}_f(0) = \vec{u}$ . Let us now choose

$$\phi(\vec{v}) = \frac{1}{dn} \|\vec{v} - \vec{u}_f(t)\|^2,$$

in (2.4) or equivalently in the weak form of (2.1), we have ( $T = 1$ )

$$\begin{aligned} \frac{d}{dt} (T_f(t) - 1) &= \frac{-1}{dn} \int_{\mathbb{R}^d} \vec{\nabla}_v (\|\vec{v} - \vec{u}_f(t)\|^2)^T \cdot ((\vec{v} - \vec{u}_f)f + \vec{\nabla}_v f) dv \\ &= 2 - \frac{2}{dn} \int f(\vec{v} - \vec{u}_f(t))^T \cdot (\vec{v} - \vec{u}) dv = 2(1 - T_f(t)), \end{aligned}$$

since  $\vec{\nabla}_v (\|\vec{v} - \vec{u}_f(t)\|^2) = 2(\vec{v} - \vec{u}_f(t))$  and  $\int f(\vec{v} - \vec{u}_f(t)) dv = \vec{0}$ . Thus, we have

$$T_f(t) = T_f(0) \exp(-2t) + (1 - \exp(-2t)),$$

i.e., the temperature tends to the equilibrium temperature  $T = 1$ . Using  $\phi = \log(f/M_f)$ , we obtain the H-theorem

$$H = \int_{\mathbb{R}^d} f \log(f/M_f) dv, \quad \frac{dH}{dt} = - \int_{\mathbb{R}^d} f \|\vec{\nabla}_v \log(f/M_f)\|^2 dv \leq 0$$

and  $dH/dt = 0 \Leftrightarrow f = M_f$ . The existence of a solution for the Cauchy problem associated with (2.1) is classical. It can be easily shown that the solution remains strictly positive for nonnegative initial data. The remainder of this section is devoted to the existence proof of solution for the discretized problem using the log formulation, i.e., (2.2). We obtain a discrete analogous of the H-theorem and prove that the solution tends to the discretized Maxwellian for large time.

**2.2. The semidiscretized linear FPL equation.** We shall consider the problem discretized in the velocity space. Let us first discuss the discretization of the gradient, i.e., the choice of the finite difference operators.

**2.2.1. Finite difference operator.** Let  $D$  be a finite difference operator that approximates the usual gradient operator  $\vec{\nabla}_v$  at least up to the first order defined as follows: for any test sequence  $\psi = (\psi_i)_{i \in \mathbb{Z}^d}$ ,  $D\psi$  is a sequence  $(D\psi)_i \in \mathbb{Z}^d$  of vectors of  $\mathbb{R}^d$  defined by

$$(D\psi)_i = ((D^s \psi)_i)_{s=1, \dots, d} \in \mathbb{R}^d,$$

where the components  $(D^s \psi)_i$  for  $s = 1, \dots, d$  approximate the partial derivatives  $\partial \psi / \partial x_s(v_i)$ . The  $s$ -component of such finite difference operator is of the form

$$(2.5) \quad (D^s \psi)_i = \sum_{k \in N^s} \alpha_k^s \psi_{i+k},$$

where  $N^s$  is a finite set of indices  $k \in \mathbb{Z}^d$  which contains the neighbors of the points involved for the  $s$ -component by the finite difference operator. We also set throughout the rest of the paper  $N = \cup_{s=1, \dots, d} N^s$ . The vectors  $\alpha_k = (\alpha_k^s)_{s=1, \dots, d} \in \mathbb{R}^d$  satisfy the following symmetry properties:

$$(2.6) \quad \sum_{k \in N^s} \alpha_k^s = 0, \quad \sum_{k \in N^s} \alpha_k^s k^r \Delta v = \delta_{sr}, \quad s = 1, \dots, d, \quad r = 1, \dots, d,$$

where  $k^r$  is the  $r$ th component of the vector  $k \in \mathbb{Z}^d$  and  $\delta_{sr}$  is the Kronecker symbol. Condition (2.6) states that  $D$  is the exact gradient for constant and linear functions or equivalently, that  $D$  is an approximation of  $\vec{\nabla}_v$  at least up to the first order. The formal adjoint  $D^*$  of  $D$  is given, for each component, by

$(D^{*s}\psi)_i = \sum_{-k \in N^s} \alpha_k^{*s} \psi_{i+k}$ , with  $\alpha_k^{*s} = \alpha_{-k}^s \quad \forall (-k) \in N^s$ . There are two simple cases of such discretized operator. On the one hand, the  $2^d$  uncentered difference operators denoted by  $D_\varepsilon$ , for  $\varepsilon = (\varepsilon_j)_{j=1,\dots,d} \in \{-1, 1\}^d$  defined by

$$(D_\varepsilon^s \psi)_i = \frac{1}{\Delta v} (\varepsilon_s (\psi_{i+\varepsilon_s e_s} - \psi_i)),$$

for such uncentered, we have  $(D_\varepsilon)^* = -(D_{-\varepsilon})$ . On the other hand, the centered operator  $D_c$  defined by  $(D_c^s \psi)_i = (\psi_{i+e_s} - \psi_{i-e_s})/2\Delta v$  for  $s = 1, \dots, d$ . The operators  $D_\varepsilon$  are clearly first-order approximations while  $D_c$  is a second-order one. For a given uncentered operator  $D_\varepsilon$ , the sets  $N^s$  are  $N^s = \{i, i + \varepsilon_s e_s\}$ . For the sake of simplicity, we restrict ourselves in the linear case (section 2) to discrete gradient  $D = D_\varepsilon$  with any choice of  $\varepsilon$ . We shall also introduce the discretized divergence operator as  $(D \cdot \psi)_i = \sum_{s=1}^d (D^s \psi)_i = \frac{1}{\Delta v} \sum_{s=1}^d (\varepsilon_s (\psi_{i+\varepsilon_s e_s} - \psi_i))$ .

**2.2.2. The semidiscretized operator.** The distribution function is approximated by a piecewise function on a fixed regular mesh of the form  $v_i = i\Delta v$ , where  $i \in \mathbb{Z}^d$ . Let us denote by  $f_i(t)$  the value of the approximated distribution function at velocity  $v_i$  and time  $t$ . The evolution in time of these functions are governed by a coupled system of nonlinear equations which is the discretized version of (2.2):

$$(2.7) \quad \frac{df_i}{dt} = F P_i^L = (D^* \cdot p)_i, \quad p_i^s = g_i^s (D^s \log(f/M))_i,$$

where the terms  $g_i^s$  have now to be defined,  $M$  is the discretized Maxwellian with the same mass  $n_f$  as  $f$ , velocity  $u$  and temperature  $T = 1$ , as in (2.3). This Maxwellian is thus chosen such that its mass is the same as the mass of the initial distribution function:

$$\sum_i M_i = \sum_i f_i^0.$$

Note that the simplest and natural choice  $g_i^s = f_i$  corresponds to the semidiscretization proposed by Degond and Lucquin in [25, 12] for the nonlinear FPL equation, described in the next section. In this paper, we shall use a modification of the scheme by defining, in a general way

$$(2.8) \quad g_i^s = \frac{(\sharp N^s) \prod_{k \in N^s} f_{i+k}}{\sum_{k' \in N^s} \left( \prod_{k \in N^s - \{k'\}} f_{i+k} \right)}, \quad i \in \mathbb{Z}^d,$$

where  $(\sharp N)$  is the cardinal of any finite subset  $N$  of  $\mathbb{Z}^d$ . This term  $g_i^s$  is a rather good approximation of  $f_i$  when the distribution function is smooth but it is a rough one when the distribution function has some “hole,” i.e., takes very small values at some velocity because all neighbors of these velocities with vanishing weight will be associated with vanishing  $g_i^s$ . We have the following estimates:

$$(2.9) \quad 0 \leq g_i^s \leq (\sharp N^s) f_{i+k} \quad \forall i \in \mathbb{Z}^d \text{ and } \forall k \in N^s.$$

Using the choice  $D = D_\varepsilon$  of discrete gradient with  $(\sharp N^s) = 2$  and the correspondent definition of the set  $N^s$ , the scheme can be simplified further:

$$p_i^s = \frac{\varepsilon_s}{\Delta v} g_i^s \left( \log \left( \frac{f_{i+\varepsilon_s e_s}}{M_{i+\varepsilon_s e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right)$$



for  $s = 1, \dots, d$ . Then, the system reads, surprisingly, in a form independent of the direction  $\varepsilon$  used to calculate the discrete gradient,

$$(2.10) \quad \frac{df_i}{dt} = F P_i^L = \frac{1}{\Delta v^2} \sum_{\mu \in \{-1, 1\}} \sum_{s=1 \dots d} g_{i, i+\mu e_s} \left( \log \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right),$$

with the terms  $g_{i,j}$ , which are given by

$$(2.11) \quad g_{i,j} = \frac{2f_i f_j}{f_i + f_j} \text{ with } g_i^s = g_{i, i+\varepsilon_s e_s} = \frac{2f_i f_{i+\varepsilon_s e_s}}{f_i + f_{i+\varepsilon_s e_s}}, \quad g_{i-\varepsilon_s e_s}^s = g_{i, i-\varepsilon_s e_s}.$$

Hence, this modification ( $f_i \mapsto g_i$ ) allows to symmetrize the formulation of the discretized linear Fokker–Planck equation.

**2.2.3. Finite volume interpretation.** Let us show that the scheme (2.10) can be derived using classical finite volume approach. Starting again from the FPL linear equation in its log form, i.e., (2.2) (again with  $T = 1$ )

$$\frac{\partial f}{\partial t} = \vec{\nabla}_v \cdot \left( f \vec{\nabla}_v \log(f/M_f) \right),$$

integrating it over the cell  $C_i$  defined as the centered cubic cell surrounding the point  $v_i$  and using the Green formula, one obtains

$$(2.12) \quad \int_{C_i} \frac{\partial f}{\partial t} dv = \sum_{\mu \in \{-1, 1\}} \sum_{s=1, \dots, d} \int_{\partial C_{(i, i+\mu e_s)}} \mu f \frac{\partial \log f / M_f}{\partial v^s} d\sigma,$$

where  $\partial C_{(i, i+\mu e_s)}$  stands for the interface between cells  $C_i$  and  $C_{i+\mu e_s}$  (with a normal vector  $\mu \vec{e}_s$  and, thus, a normal derivative  $\mu \frac{\partial}{\partial v^s}$ ) and  $d\sigma$  the superficial measure on this interface. By taking midpoint quadrature formula for each side of this equation, we have

$$(2.13) \quad \frac{df_i}{dt} = \frac{\Delta v^{d-1}}{\Delta v^d} \left( \sum_{\mu \in \{-1, 1\}} \sum_{s=1, \dots, d} f \left( \frac{v_i + v_{i+\mu e_s}}{2} \right) \mu \frac{\partial \log \left( \frac{f}{M_f} \right)}{\partial v^s} \left( \frac{v_i + v_{i+\mu e_s}}{2} \right) + O(\Delta v^2) \right).$$

Now, approximating  $\partial \log f / M_f / \partial v^s (v_i + v_{i+\mu e_s} / 2)$  with the second-order approximation

$$\frac{\mu}{\Delta v} \left( \log \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right),$$

and using a classical, second-order approximation designed for the numerical treatment of such diffusive equation [16] which consists of the harmonic averaged of the diffusion coefficients

$$f \left( v = \frac{v_i + v_{i+\mu e_s}}{2} \right) \approx \frac{2f_i f_{i+\mu e_s}}{f_i + f_{i+\mu e_s}},$$

one recovers the scheme (2.10). Note that such approximation of the diffusion coefficients guarantees the continuity of the fluxes at the interface.

**2.2.4. Reduction to a bounded velocity domain.** From a numerical point of view, it is necessary to consider a bounded velocity domain, i.e., to assume that the index  $i$  belongs to a finite set of integer which is assumed of the form  $\{1 \cdots n\}^d \stackrel{\text{def}}{=} I$  for the sake of simplicity. Let us define the “interior” set  $\mathcal{I}$  such that  $i \in \mathcal{I}$  if and only if  $\forall k \in N$ ,  $(i+k) \in I$ . The definition (2.7) of  $FP_i^L$  is modified as follows on the discretized weak formulation:

$$(2.14) \quad \sum_{i \in I} \phi_i FP_i^L = - \sum_{i \in \mathcal{I}} \sum_{s=1, \dots, d} g_i^s (D\phi)_i \cdot (D \log(f/M_f))_i.$$

Then,  $FP_i^L$  is unchanged for any “interior” point  $i \in \mathcal{I}$ . For the “frontier” points, it leads to suppress the terms of the form

$$\left( \log \left( \frac{f_{i \pm \varepsilon_s e_s}}{M_{i \pm \varepsilon_s e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right)$$

for any index such that  $i \pm \varepsilon_s e_s \notin I$ . Finally,  $FP_i^L$ , on a bounded domain  $I$  is of the form  $\forall i \in I$

$$(2.15) \quad FP_i^L = \sum_{\mu \in \{-1, 1\}} \sum_{s=1, \dots, d} a_{i, i+\mu e_s} g_{i, i+\mu e_s} \left( \log \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right),$$

where the constant coefficients  $a_{i, i+\mu e_s}$  were defined as  $\frac{1}{\Delta v^2}$  if the point  $i + \mu e_s$  lies in  $I$ , or 0 if not and  $g_{i, i+\mu e_s}$  is given by (2.11).

**2.2.5. The semidiscrete H-theorem.** The H-theorem is satisfied with the following discrete entropy functional:

$$(2.16) \quad H = \sum_{i \in I} f_i \log(f_i/M_i).$$

Indeed, using the weak formulation (2.14) with  $\phi_i = \log(f_i/M_i)$ , we have using the mass conservation

$$\frac{dH}{dt} = - \sum_{i \in \mathcal{I}} \sum_{\mu \in \{-1, 1\}} \sum_{s=1, \dots, d} a_{i, i+\mu e_s} g_{i, i+\mu e_s} \left( \log \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right)^2 \leq 0.$$

Moreover, we have (if the terms  $g_i^s$  are strictly positive):

$$(2.17) \quad \frac{dH}{dt} = 0 \Leftrightarrow \forall (i, j) \in I^2, f_i/M_i = f_j/M_j.$$

Finally, using the mass conservation, we have  $f_i = M_i \forall i \in I$ .

**2.3. Existence for the semidiscretized linear FPL equation.** This section is devoted to the proof of the following theorem.

**THEOREM 2.1.** *Let  $(f_i^0) \in \mathbb{R}^{(\#I)}$  such that  $f_i^0 > 0 \forall i \in I$ . Then, the Cauchy problem for the system  $(FP^L)$*

$$\frac{df_i}{dt} = FP_i^L, \quad f_i(t=0) = f_i^0 \quad \forall i \in I,$$

with  $FP_i^L$  defined by (2.15) is well posed. Moreover, there exists a global in time solution  $(f_i(t))_{i \in I}$  such that

$$(2.18) \quad \forall i \in I, \lim_{t \rightarrow \infty} f_i(t) = M_i,$$

where  $M_i$  is the associated discrete Maxwellian, defined by (2.3).

*Proof.* The local existence is obtained using standard Lipschitz property of the equations. Thus, there exists a solution at least for small time. The difficulty arises when some of the  $f_i$  tend to 0 due to the singularity of the log terms.

Once a positive lower bound is obtained for the  $f_i$ , one has an upper bound using the conservation of mass. Hence, there are only two alternatives: either there exists a global positive solution or the solution has a maximal lifetime  $t_0 < \infty$  such that  $f_i(t) \rightarrow 0$  when  $t \rightarrow t_0$  for some  $i \in I$ . We shall now prove that the solution is global using a maximum principle. Let us recall the proof although it can probably be obtained using a more general theory, since it provides us an explicit lower bound for the distribution function. Define  $h_S$  and  $h_I$  as follows:

$$(2.19) \quad h_S(t) = \sup_{i \in I} \frac{f_i(t)}{M_i}, \quad h_I(t) = \inf_{i \in I} \frac{f_i(t)}{M_i}.$$

Note that  $h_I$  and  $h_S$  are continuous functions of  $t$  (for sufficiently small time  $t$  such that the solution exists) such that  $h_S(0) \geq 1$  and  $0 < h_I(0) \leq 1$  since the Maxwellian  $M_i$  has the same first moment as  $f_i$ . Let us prove that  $h_S$  is a decreasing function by contradiction. Assume there exists  $t_1$  and  $t_2 > t_1$  such that  $h_S(t_2) > h_S(t_1)$ . Then, there exists  $t_3 \in [t_1, t_2]$  such that  $h_S$  is maximal. Let  $I_S$  be the set of indexes  $i$  for which  $h_S(t_3) = f_i(t_3)/M_i$ . At such point  $i$ , we have

$$\frac{d(f_i/M_i)}{dt}(t_3) \leq 0.$$

Indeed, the time evolution of  $f_i$  is governed by (2.15), and, setting  $h_i = (f_i/M_i)$  is a sum of terms of the form  $\log(h_{i+k}/h_i)(t_3)$ , with  $k \in N$ , multiplied by positive terms  $a_{i,i+k}g_{i,i+k}$ . Thus,  $h_i$  being maximal ( $i \in I_S$ ), these terms are negative. The minima cannot be isolated since in this case, we get  $d(f_i/M_i)(t_3)/dt < 0$ , which contradicts  $i \in I_S$ . Step by step, we obtain that  $h_i = 1$ , i.e.,  $f_i = M_i \forall i \in I$ . Finally, we have proved that  $h_S$  is decreasing. A similar proof holds for  $h_I$  increasing. This naturally provides a uniform bound for the weights  $f_i$  and for any time  $t$

$$h_S(0) \sup_{i \in I} M_i \geq f_i(t) \geq h_I(0) \inf_{i \in I} M_i \quad \forall i \in I, \quad \forall t \geq 0.$$

Let us now prove that  $f_i \rightarrow M_i$  when  $t \rightarrow \infty$ . Indeed, the following “weighted  $L^2$ ” distance between  $f_i$  and  $M_i$

$$\tilde{D} = \sum_{i \in I} M_i \left( \frac{f_i}{M_i} - 1 \right)^2,$$

is decreasing, using  $\phi_i = (f_i/M_i - 1)$  in the weak formulation (2.14),

$$\begin{aligned} \frac{d\tilde{D}}{dt} &= 2 \sum_{i \in I} \frac{df_i}{dt} \left( \frac{f_i}{M_i} - 1 \right) = -2 \sum_{i \in I} \sum_{\mu \in \{-1,1\}} \sum_{s=1 \dots d} a_{i,i+\mu e_s} g_{i,i+\mu e_s} \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} - \frac{f_i}{M_i} \right) \\ &\quad \times \left( \log \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} \right) - \log \left( \frac{f_i}{M_i} \right) \right) \leq 0, \end{aligned}$$

using the standard inequalities  $(x-y)(\log(x) - \log(y)) \geq 0$ . The distance is decreasing and tends to 0 for any sequence of time. More precisely, the maximum principle, proved above, ensures the existence of  $C > 0$  (depending only on the initial data and particularly on the initial lower bound of the weights) such that

$$\frac{d\tilde{D}}{dt} \leq C \sum_{i \in \mathcal{I}} \sum_{\mu \in \{-1,1\}} \sum_{s=1\dots d} \left( \frac{f_{i+\mu e_s}}{M_{i+\mu e_s}} - \frac{f_i}{M_i} \right)^2 \leq \tilde{C}\tilde{D}.$$

Thus, the functions  $f_i(t)$  tend exponentially to  $M_i \forall i \in I$  when  $t \rightarrow \infty$  using a Grönwall lemma. This last result can be related to a recent paper of Toscani [28] concerned with the continuous linear Fokker-Planck equation for which he obtains similar exponential decay toward the Maxwellian.

The convergence toward the equilibrium can also be proved using the Csiszar-Kullback inequality [19] for the continuous problem (see [28]). Since the terms  $f_i$  are bounded below for  $t \in [0, \infty[$  then necessarily  $\lim_{t \rightarrow \infty} H(t) = 0$ . Applying the Csiszar-Kullback inequality which states that for any couple of strictly positive functions  $F$  and  $G$  and any probability measure  $dP$  we have

$$\|F - G\|_{L^1(dP)}^2 \leq 2 \int F \ln \left( \frac{F}{G} \right) dP;$$

then

$$0 \leq \left( \sum_i |f_i - M_i| \right)^2 \leq 2H(f),$$

which implies that

$$\lim_{t \rightarrow \infty} \sum_i |f_i(t) - M_i| = 0. \quad \square$$

Note that this theorem can be proved even in the case  $g_i = f_i$ . However, the following cannot be extended without the modification and the use of the estimate (2.9) on  $g_{i,j}$  we have proposed.

**2.4. The time-discretized linear FPL equation.** Let us consider now the explicit time discretization of the preceding problem. The distribution function is assumed known at time  $t$  and equal to  $(f_i)_{i \in I}$ . Let us denote by  $\bar{v}$  the computed value of any variable  $v$  at time  $t + \Delta t$ . The scheme is of the form

$$(2.20) \quad \bar{f}_i = f_i + \Delta t F P_i^L,$$

where  $F P_i^L$  is defined by (2.15). The main properties of this explicit in time scheme (2.20) are summarized in the following proposition.

**PROPOSITION 2.2.** *Let  $(f_i^0) \in \mathbb{R}^{(\sharp I)}$  such that  $f_i^0 > 0 \forall i \in I$ . Then there exists a time step of the form  $C/\Delta v^2$  with the constant  $C$  depending only on the initial condition for which the scheme (2.20) is positive and conservative and decays entropy. Furthermore, the scheme leads to a solution which verifies*

$$(2.21) \quad \forall i \in I, \lim_{t \rightarrow \infty} f_i(t) = M_i,$$

where  $M_i$  is the associated discrete Maxwellian, defined by (2.3).

The proof is postponed until the appendix. Note that the scheme leads to a nonoscillatory solution since the maximum principle prevents the creation of any supplementary local extremum (since local maximum decreases and local infimum increases).

**3. The nonlinear FPL equation.** We shall now consider the tridimensional nonlinear FPL equation in the homogeneous case for any interaction potential. We restrict ourselves to a single-species plasma since the methods can easily be extended to the multispecies case as explained in [7].

**3.1. The continuous nonlinear FPL equation.** We denote by  $f(v, t)$  the distribution function, a solution of the following scaled FPL equation:

$$(3.1) \quad \frac{\partial f}{\partial t} = Q(f, f) = \nabla_v \cdot \left( \int_{\mathbb{R}^3} \Phi(v - v_*) ((\vec{\nabla}_v f) f_* - (\vec{\nabla}_{v_*} f) f) dv_* \right),$$

where  $Q(f, f)$  is the FPL collision operator written in the so-called Landau form with the standard notations (for example,  $f_* = f(x, v_*, t)$ ) and  $\Phi(v)$  is the following  $3 \times 3$  matrix:

$$(3.2) \quad \Phi(v) = |v|^{\gamma+2} S(v), \quad S(v) = I_3 - \frac{v \otimes v}{|v|^2}.$$

$S(v)$  is the orthogonal projector onto the plane orthogonal to  $v$ .  $\gamma$  is a real parameter which leads to the usual classification in hard potentials ( $\gamma > 0$ ), Maxwellian molecules ( $\gamma = 0$ ) or soft potentials ( $\gamma < 0$ ). This latter case involves the Coulombian case (i.e.,  $\gamma = -3$ ) which is of primary importance for plasma applications. The well-known physical properties of the FPL operator are similar to that of the Boltzmann operator such as the decay of the entropy, the conservation of mass, momentum, and energy, and the characterization of the equilibrium states by Maxwellians. These properties can easily be shown on the weak form of the FPL operator which can be found in [11, 7, 25, 12].

**3.2. The semidiscretized nonlinear FPL equation.** The distribution function is approximated by piecewise constant functions on a fixed regular mesh of the form  $v_i = i\Delta v$ , where  $i = (i_1, i_2, i_3)$  belongs, as in the linear case, to a finite set of integer which is assumed of the form  $\{1 \cdots n\}^3 \stackrel{\text{def}}{=} I$  for the sake of simplicity. We refer to [11, 25, 12] for a detailed derivation of this discretization. Denote by  $f_i(t)$  the value of the approximated distribution function at velocity  $v_i$  and time  $t$ . The time evolution of this discretized function is then governed by a coupled system of nonlinear equations of the form (for  $i \in I$ )

$$(3.3) \quad \frac{\partial f_i}{\partial t} = (D^* \cdot p)_i, \quad p_i = \Delta v^3 \sum_{j \in I} f_i f_j \Phi((i - j)\Delta v) ((D \log f)_i - (D \log f)_j),$$

where  $D$  is again a finite difference operator that approximates the usual gradient operator  $\nabla$  at least up to the first order,  $D^*$  its formal adjoint, as defined in section 2.2.1. The reduction to a bounded domain follows the same lines as in the linear case presented before. In the weak formulation, this system of differential equations can be written equivalently for any test sequence  $(\psi_i)_{i \in I}$  as

$$(3.4) \quad \sum_{i \in I} \frac{\partial f_i}{\partial t} \psi_i \Delta v^3 = -\frac{1}{2} \Delta v^6 \sum_{(i,j) \in I^2} f_i f_j ((D\psi)_i - (D\psi)_j)^T \Phi(v_i - v_j) ((D(\log f))_i - (D(\log f))_j).$$

As shown in [11, 12, 25], this discrete model is conservative and decays entropy (provided that there exists a solution, which is the aim of the present paper). The only

equilibrium states are the discrete Maxwellians for decentered approximation  $D_\varepsilon$  of the gradient but introduce spurious collisional invariant for the centered  $D_c$  one (previously defined in the linear case). The good choice for the discretized Fokker–Planck operator is the arithmetic average over all (eight) decentered operators  $D_\varepsilon$  which is equal to the centered operator plus a diffusive perturbation (see [7] for details).

From the positivity of the matrix  $\Phi$ , it is easy to check that such a scheme decays the entropy using the weak form (3.4) with  $\psi_i = \log f_i$

$$(3.5) \quad H(t) \stackrel{\text{def}}{=} \Delta v^3 \sum_{i \in I} f_i (\log f_i) \leq \Delta v^3 \sum_{i \in I} f_i^0 (\log f_i^0) = H(0).$$

Note that this property is not achieved if one discretizes the Fokker–Planck operator in the “nonlog” form (see [12] for a precise statement). This was the first reason for using the “log”-discretization we presented here. There is another reason which is also of main importance: there exists some positive initial condition for which the “nonlog” discretized FPL equations leads to a solution with negative values of the distribution function in arbitrary small time (see appendix for details).

**3.3. Existence for a semidiscretized nonlinear FPL equation.** We shall now turn to the proof of existence of positive solutions for the Cauchy problem for the system defined by (3.3) with positive initial conditions  $f_i(t=0) = f_i^0 > 0 \forall i \in I$ . We also refer to [7] for methods such as multigrid or sublattice for reducing the numerical cost in evaluating this quadratic operator. These methods are based on a simplified discrete operator of the form

$$(3.6) \quad \Delta v^6 \sum_{i \in \mathbb{Z}^3} F P_i \psi_i \Delta v^3 \\ = -\frac{1}{2} \sum_{(i,j) \in I^2} a_i a_j ((D\psi)_i - (D\psi)_j)^T \Phi(v_i - v_j) ((D \log f)_i - (D \log f)_j),$$

where the coefficients  $a_i$  depend on the distribution function, some of them can be null, and their sum over all the indices is bounded. The number of nonnull terms determines the cost of the method. For the original discrete operator there are  $(\#I)^2$  nonnull terms so the method is quadratic. In multigrid methods, since Monte Carlo integration is used [7], only  $\#I \log(\#I)$  terms are nonzero at each time step. Therefore the cost is only of the order of  $\#I \log(\#I)$ . The following analysis also applies for such rapid methods. The existence of solutions for small time can be easily obtained using classical Cauchy–Lipschitz theorem. Indeed, there is no singularity in this system neither in the log terms, using  $f_i^0 > 0$ , nor in the matrix  $\Phi$ , since for  $i \neq j$  we have  $\|v_i - v_j\| \geq \Delta v$ .

Then, these solutions had to remain positive due to the log terms and they are global in time provided that none of the  $f_i$  vanishes. Indeed, we get for free an upper bound (uniform in time, depending on  $\Delta v$ ) for  $f_i$  using the conservation of mass  $\sum_{i \in I} f_i \Delta v^3 = \sum_{i \in I} f_i^0 \Delta v^3$ . Finally, we have the only two following alternatives: either there exists a maximal time  $t_0 > 0$  of existence, i.e., such that for at least some index  $i_0 \in I$ ,  $\lim_{t \rightarrow t_0} f_{i_0}(t) = 0$ , or there exists a solution global in time. In other words, existence of global solution for the differential equations is related to the fact that  $f_i$  can only vanish in infinite time.

Assuming that there exists only one index  $i_0$  for which the distribution tends to 0 when  $t \rightarrow 0$ , one separates in the various terms involved in  $(D^* \cdot p)_i$  the ones containing

$\log(f_{i_0})$  which blows up. Then, it is easy to check that this term is multiplied by a negative constant and therefore, the leading order term in  $df_{i_0}/dt$  is positive which provides the contradiction. However, this method cannot be extended, at least to our knowledge, to the general case where the distribution function vanishes simultaneously at several locations.

Showing that the weights  $f_i$  cannot vanish in finite time is equivalent to showing that the function

$$(3.7) \quad K = \sup_{i \in \mathcal{I}, k \in N} \left| \frac{f_i}{f_{i+k}} \right|$$

remains bounded in finite time as in the linear case. This function is convenient since these ratios actually appear in the  $D(\log f)_i$  terms. A direct calculation, which is left to the reader and related to the one given in the next section, gives the following estimates:

$$(3.8) \quad \left| \frac{dK}{dt} \right| \leq CK^2 \log(K)$$

for some constant  $C > 0$ . This estimate obviously does not give an upper bound for the value of  $K$  for arbitrary large time and therefore, no lower bound for the  $f_i$  in finite time. Other methods for proving the existence of positive solution for the system of ODEs (3.3) rely on functional approaches (i.e., find a function of  $f_i$  which tends to  $\infty$  if one of the  $f_i \rightarrow 0$  and proves that it remains bounded) like, for example, the Linnick functional [4]. Note that for the continuous Boltzmann or FPL equations, the distribution function can be bounded below by some Maxwellian [29, 14, 15].

To ensure the existence of global positive solutions, we introduce a slightly different discretized operator which has exactly the same properties (conservation, entropy) as the one defined in [11, 7, 25, 12] but for which the proof can be complete using the first direct method. The trick consists of modifying the terms  $f_i$  and  $f_j$ , respectively, in the formulae (3.3) by some approximations  $g_i$  and  $g_j$ , respectively, defined as in the linear case by

$$g_i = \begin{cases} \frac{(\sharp N) \prod_{k \in N} f_{i+k}}{\sum_{k' \in N} \left( \prod_{k \in N - \{k'\}} f_{i+k} \right)}, & i \in \mathcal{I}, \\ 0 & \text{if } i \in I - \mathcal{I}, \end{cases}$$

where  $N$  is a finite subset of  $\mathbb{Z}^3$  and  $(\sharp N)$  is the cardinal of  $N$ . Let us recall that  $g_i$  is a rather good approximation of  $f_i$  when the distribution function is smooth but is rough when the distribution function is very peaked at some velocity. The Cauchy problem associated with the modified ODEs, i.e., the semidiscretized FPL equation (3.3) reads for  $i \in I$

$$(3.9) \quad \frac{\partial f_i}{\partial t} = FP_i = (D^* \cdot p)_i, \quad p_i = \Delta v^3 \sum_{j \in \mathcal{I}} g_i g_j \Phi((i-j)\Delta v) ((D \log f)_i - (D \log f)_j),$$

We have the following result.

**THEOREM 3.1.** *Let  $(f_i^0) \in \mathbb{R}^{(\sharp I)}$  such that  $f_i^0 > 0 \forall i \in I$ . The Cauchy problem with initial conditions  $f_i(t=0) = f_i^0$  for the differential system (3.9) has a unique positive entropy solution for arbitrary large time.*

*Proof.* The existence and unicity of the solution of system (3.9) with strictly positive initial data for small time is obtained using classical Cauchy–Lipschitz theorem.

We will now prove that the solution cannot vanish at some velocities in finite time and, thus, the solution is global in time. The discrete H-theorem is given by (3.5) and can be proved using the weak formulation. We recall the following estimates:

$$(3.10) \quad 0 \leq g_i \leq (\sharp N) f_{i+k} \quad \forall i \in \mathcal{I} \text{ and } \forall k \in N.$$

Using definition of  $K$  and  $p_i$ , i.e., (3.7) and (3.9), we have the following estimate for the vectors  $p_i$ :

$$(3.11) \quad \|p_i\| \leq C g_i \log(K),$$

where  $C$  is a generic constant throughout the rest of the paper, depending on the number of grid points  $(\sharp I)$ , the potential parameter  $\gamma$ , the velocity mesh size  $\Delta v$ , the initial condition  $(f_i^0)_{i \in I}$ , the coefficients  $a_k$ , and the cardinal of the set  $N$ . Indeed, we have the following upper bounds:

$$\begin{aligned} \sup_{(i,j) \in I^2} \|\Phi(v_i - v_j)\| &\leq C, \quad |f_i| \leq C \quad \forall i \in I, \\ \|(D \log f)_i\| &\leq C \quad \forall i \in \mathcal{I}. \end{aligned}$$

Then, using (3.10), we have

$$(3.12) \quad |(D^* \cdot p)_i| \leq C \log(K) \sup_{k \in N} g_{i-k} \leq C \log(K) f_i.$$

Then, we have for any  $i \in \mathcal{I}$  and any  $k \in N$

$$\frac{d(f_i/f_{i+k})}{dt} = \frac{1}{f_{i+k}} \frac{df_i}{dt} - \frac{f_i}{f_{i+k}^2} \frac{df_{i+k}}{dt}.$$

Finally, using (3.12) and since  $\frac{df_i}{dt} = (D^* \cdot p)_i$ , we have

$$(3.13) \quad \left| \frac{dK}{dt} \right| \leq CK \log(K),$$

which implies  $K(t) \leq K(0) \exp(\exp(Ct))$  and this concludes the proof.  $\square$

This proof can be carried out for the diffusive perturbation defined in [7] by again changing the  $f_i$  into  $g_i$  as explained above. We do not detail the proof since it follows exactly the same lines. The key point is that the modified scheme (with  $g_i$  defined by (2.8)) satisfies (3.10), which implies (3.12), while the original scheme [11, 7, 25, 12] (with  $g_i = f_i$ ) leads to

$$(3.14) \quad |(D^* \cdot p)_i| \leq CK \log(K) f_i,$$

and thus (3.13) becomes (3.8).

Note that for large time, the distribution function should tend at least formally to the discretized Maxwellian since its entropy decays (see Csiszar–Kullback inequality). Therefore, there must exist some constant  $K_0 > 1$  such that  $K(t) \leq K_0$  for all time. However, the convergence toward the discrete Maxwellian or equivalently the existence of a lower bound for the weights  $f_i$  uniform in time remains an open problem up to now for the discretized nonlinear FPL equation. This point is currently being investigated.



**3.4. The time-discretized nonlinear FPL equation.** Let us consider an explicit time discretization as in the linear case. The distribution function is assumed known at time  $t$  and equal to  $f_i$  and its value at time  $\bar{t} = t + \Delta t$  denoted by  $\bar{f}_i$  is given by the following explicit scheme:

$$(3.15) \quad \bar{f}_i = f_i + \Delta t F P_i,$$

where  $F P_i$  is defined by (3.9). We shall determine conditions on the time step  $\Delta t$  under which the scheme gives positive and entropy solution for an arbitrary large time. We resume our results in the following proposition.

**PROPOSITION 3.2.** *There exists a time-step sequence  $\Delta t_n$  such that the scheme (3.15) defines recursively a positive and entropy solution at any time (i.e.,  $\sum \Delta t_n = \infty$ ).*

*Proof.* First, note that (3.12) gives a positive constant  $C > 0$  such that

$$|F P_i| = |(D^* . p)_i| \leq C \log(K) f_i,$$

where  $K = \max_{i \in \mathcal{I}, k \in N} f_i / f_{i+k}$ . Let us define  $\Delta t_1 = 1/C \log K$  and choose  $\Delta t = \alpha \Delta t_1$ , with  $0 < \alpha < 1$ . Then, we have using (3.15) and (3.12),

$$\bar{K} = \max_{i \in \mathcal{I}, k \in N} \frac{\bar{f}_i}{\bar{f}_{i+k}} \leq \frac{K(1+\alpha)}{(1-\alpha)} = \beta K,$$

with  $\beta = (1+\alpha)/(1-\alpha) > 1$ . Note the difference with the linear case where the maximum principle insures that the function  $K$  decreases at each iteration. Thus, we have by recursion at iteration  $n$

$$\log(K_n) \leq n \log(\beta) + \log(K_0),$$

with  $K_0 = K(0) = \max_{i \in \mathcal{I}, k \in N} f_i^0 / f_{i+k}^0 < \infty$ . Therefore, for the time step  $\Delta t_n$  at iteration  $n$ , which is defined recursively by  $\Delta t_n = \alpha / C \log(K_n)$  and for which the solution is positive, we have the following estimate:

$$\Delta t_n \geq \frac{\alpha}{C(n \log(\beta) + \log(K_0))}.$$

The right-hand side of this inequality leads to a divergent sum and thus the solution remains positive and can be constructed for any arbitrary large time, i.e., after  $n$  iterations the time is equal to

$$t_n = \sum_{k \leq n} \Delta t_k \rightarrow \infty, \quad n \rightarrow \infty.$$

We have now to check that the entropy decays. Defining the discrete entropy  $H$  by (3.5), we obtain, as in the linear case (see appendix A-1), that it is decreasing provided that the time-step is smaller than

$$\Delta \tilde{t}_1 \stackrel{\text{def}}{=} \min \left( \alpha \Delta t_1, \frac{-\Delta v^3 \sum_{i \in I} F P_i \log(f_i)}{\Delta v^3 \sum_{i \in I} \frac{(F P_i)^2}{f_i}} \right).$$

Using the weak formulation, we have

$$(3.16) \quad \left| \Delta v^3 \sum_{i \in I} F P_i \log(f_i) \right| = \Delta v^6 \sum_{(i,j) \in \mathcal{I}^2} g_i g_j X_{i,j}^T \Phi(v_i - v_j) X_{i,j},$$

where  $X_{i,j}$  is a  $d$  vector defined by  $X_{i,j} = (D \log f)_i - (D \log f)_j$ . On the other hand, using the convention where terms of the form  $a_{i+k}$  are set to zero if  $i+k \notin I$ , we have for any  $i \in I$

$$(FP_i)^2 = (D^* \cdot p)_i^2 \leq \frac{C'}{\Delta v^2} \sum_{k \in N} \|p_{i+k}\|^2,$$

where  $C'$  is a constant which depends only on the cardinal of the set  $N$  and

$$\|p_{i+k}\|^2 = \left\| g_{i+k} \Delta v^3 \sum_{j \in \mathcal{I}} g_j \Phi(v_{i+k} - v_j) X_{i+k,j} \right\|^2.$$

Let us introduce the sequences  $U_j$  and  $V_j$  as

$$U_j = \left( \sqrt{\frac{g_{i+k} g_j \Delta v^3}{\|v_{i+k} - v_j\|^s}} \right)_j, \quad V_j = \left( \sqrt{\frac{g_{i+k} g_j \Delta v^3}{\|v_{i+k} - v_j\|^s}} \|S(v_{i+k} - v_j) X_{i+k,j}\| \right)_j.$$

The definition of  $p_{i+k}$  leads, using the Cauchy-Schwarz inequality and definition of the matrix  $S(v)$  (with  $s = 1$  in the Coulombian case), to

$$\|p_{i+k}\|^2 \leq \|U\|^2 \|V\|^2,$$

with the norm

$$\|U\|^2 = \sum_{j \in \mathcal{I}} \frac{g_{i+k} g_j \Delta v^3}{\|v_{i+k} - v_j\|^s}, \quad \|V\|^2 = \Delta v^3 \sum_{j \in \mathcal{I}} g_{i+k} g_j \|v_{i+k} - v_j\|^s \|\Phi(v_{i+k} - v_j) X_{i+k,j}\|^2.$$

Thus, we have, using twice  $g_i \leq (\#N) f_i \forall i$ ,

$$\begin{aligned} & \|p_{i+k}\|^2 / f_i \\ & \leq (\#N)^2 \left( \sup_{i \in I} \sum_{j \in \mathcal{I}} \frac{f_j \Delta v^3}{\|v_i - v_j\|^s} \right) \Delta v^3 \sum_{j \in \mathcal{I}} g_{i+k} g_j \|v_{i+k} - v_j\|^s \|\Phi((i+k-j)\Delta v) X_{i+k,j}\|^2, \end{aligned}$$

and adding these inequalities, we obtain

$$\begin{aligned} & \sum_{i \in I} \frac{(FP_i)^2}{f_i} \Delta v^3 \\ & \leq \frac{C'}{\Delta v^2} (\#N)^2 \left( \sup_{i \in I} \sum_{j \in \mathcal{I}} \frac{f_j \Delta v^3}{\|v_i - v_j\|^s} \right) \Delta v^6 \sum_{(i,j) \in \mathcal{I}^2} g_i g_j \|v_i - v_j\|^s \|\Phi(v_i - v_j) X_{i,j}\|^2. \end{aligned}$$

Moreover, we have for all vectors  $v$  and  $X$

$$X^T S(v) X = \|S(v) X\|^2 \Rightarrow X^T \Phi(v) X = \|v\|^s \|\Phi(v) X\|^2,$$

and then, using (3.16), we have proved that

$$\sum_{i \in I} \frac{(FP_i)^2 \Delta v^3}{f_i} \leq \frac{C'}{\Delta v^2} (\#N)^2 \left( \sup_{i \in I} \sum_{j \in \mathcal{I}} \frac{f_j \Delta v^3}{\|v_i - v_j\|^s} \right) \times \Delta v^3 \sum_{i \in I} FP_i \log(f_i),$$

and thus

$$\Delta t \stackrel{\text{def}}{=} \min \left( \alpha \Delta t_1, \frac{\Delta v^2}{C'(\sharp N)^2} \times \left( \sup_{i \in I} \sum_{j \in \mathcal{I}} \frac{f_j \Delta v^3}{\|v_{i+k} - v_j\|^s} \right)^{-1} \right)$$

yields a positive and entropy scheme.  $\square$

For the simplest uncentered discrete gradient, a simple calculus gives us  $C' = 4$  and  $(\sharp N) = 4$ . For other choices of the discrete gradient the calculus of these constants can be achieved by straightforward calculations. In numerical simulations, the time step that guarantees positivity and entropy decay of the solution tends to a constant one (and not to zero as the time-step sequence constructed in the proof).

Moreover, the condition on the entropy is more restrictive than the positivity one and insures a nonoscillatory solution. This differs from the linear case. If the condition on the entropy is relaxed, it is readily seen on numerical simulations (see, for example, [6]) that unphysical oscillations appear.

**4. Conclusions.** The existence of a global, positive, conservative, and entropy solution for the semidiscretized nonlinear FPL equation as proposed (on Landau-log form) in [12, 3, 27, 5, 26] cannot be proved, at least to our knowledge, without the modification proposed here (i.e.,  $f_i \mapsto g_i$ ). Moreover, when such velocity discretized models are also discretized in time, it is not proved that they still preserve entropy decay and positivity of the solution even though these properties are satisfied by the semidiscretized models [11].

We show that the constructed discrete model for FPL equation gives global positive solution and its time-discretized version preserves all the properties of the semidiscretized one. In the linear case, the modification that we propose is related to a classical approximation of the diffusion coefficients. For the nonlinear equation, we give sufficient, CFL-like conditions on the time step that guarantees entropy decay and positivity.

We refer to [7, 22] for fast algorithms for solving the nonlinear Fokker–Planck equation in Landau form. As explained in the present paper, the proofs apply to the multigrid and sublattice methods and can probably be extended to multipole methods using the modification we propose. Let us also mention a forthcoming paper concerned with the isotropic distribution function [6], for which the results presented here can be improved. It is possible in this particular case to use a nonuniform grid, to obtain a uniform estimate on the time step (instead of the divergent series we have constructed), and to reduce the quadratic cost to a linear cost without any supplementary approximation. The isotropic case can be used to compute reference solutions—or benchmarks—since the linear cost of this one-dimensional problem makes it possible to compute very accurate solutions even in the Coulombian case ( $\gamma = -3$ ), for which there is no known explicit solution unlike the Maxwellian case ( $\gamma = 0$ , see [21]).

The next step is to prove the convergence of the constructed sequences of approximated solutions toward the solution of the continuous FPL equation. First, one shall study the convergence of the time-discretized solution (using the explicit scheme presented in section 3.4) toward the solution of the system of ODE corresponding to the semidiscretized FPL equation presented in section 3.2 when  $\Delta t \rightarrow 0$ . Second, we shall prove convergence of the solution of the semidiscretized FPL equation to the solution of the continuous equation when  $\Delta v \rightarrow 0$ . This result is much more difficult to prove.

**Appendix. A1. Proof of Proposition 2.2.** Let us define  $h_i = f_i/M_i$ . We have from (2.20)

$$\bar{h}_i = h_i + \frac{FP_i^L}{M_i} \Delta t h_i \left( 1 + \Delta t \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s} \frac{g_{i,i+\mu e_s}}{f_i} \log \left( \frac{h_{i+\mu e_s}}{h_i} \right) \right).$$

Recall that  $h_I, h_S$  defined in the preceding proof can be equivalently written as

$$h_I = \inf_{i \in I, k \in N[(i+k) \in I]} h_{i+k}, \quad h_S = \sup_{i \in I, k \in N[(i+k) \in I]} h_{i+k}.$$

We also define  $K = h_S/h_I > 1$ . Let us prove first that there exists a uniform lower bound on the time step  $\Delta t$  for which  $h_S$  is decreasing and  $h_I$  is increasing. Assume

$$\Delta t \leq \Delta t_1 \stackrel{\text{def}}{=} \frac{1}{M} \inf_{x \in ]1, K]} \frac{(x-1)}{x \log x} < \infty,$$

with (using the definitions of  $a_{i,i+\mu e_s}$ )

$$M = \frac{8d}{\Delta v^2} \geq 4 \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s}.$$

Using (2.9) we have  $g_{i,i+\mu e_s} \leq 2f_i$  by construction and therefore

$$h_i + 2\Delta t \left( \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s} \right) h_i \log \left( \frac{h_I}{h_i} \right) \leq \bar{h}_i,$$

and

$$\bar{h}_i \leq h_i + 2\Delta t \left( \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s} \right) h_i \log \left( \frac{h_S}{h_i} \right).$$

Moreover, for  $\Delta t \leq \Delta t_1$ , we have

$$4\Delta t \left( \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s} \right) x \log x \leq x - 1 \quad \forall x \in ]1, K].$$

This implies (for  $x = h_i/h_I$  and  $x = h_S/h_i$ ) that

$$h_I \leq (h_I + h_i)/2 \leq \bar{h}_i \leq (h_i + h_S)/2 \leq h_S.$$

Thus,  $\sup_{i \in I} h_i$  decreases,  $\inf_{i \in I} h_i$  increases, and  $\bar{K} \leq K$ . This discrete maximum principle insures that the scheme is positive. These bounds for the time-steps are not optimal. Since the coefficients  $a_{i,i+k}$  depend on the velocity mesh size as  $\Delta v^{-2}$ , the stability condition we get, can be written in the form

$$\Delta t \leq \Delta t_1 = C \Delta v^2, \quad C = \frac{1}{8d} \inf_{x \in ]1, K(0)]} \frac{(x-1)}{x \log x}$$

as is usual for such convection-diffusion operator.

We shall determine another condition for the scheme being entropy decaying. Let us assume the first condition is satisfied. Thus, the terms  $f_i$  and also  $g_{i,i+\mu e_s}$  have a uniform lower and upper bound. The entropy at time  $t + \Delta t$  is of the form

$$\bar{H} = \sum_{i \in I} (f_i + \Delta t F P_i^L) \log((f_i + \Delta t F P_i^L)/M_i).$$

Then, using  $\log(1+h) \leq h \ \forall h > -1$ , we have (with  $h = \frac{\Delta f}{f} \geq -1$  since the scheme is positive)

$$\begin{aligned} (f + \Delta f) \log\left(\frac{f + \Delta f}{M}\right) &= (f + \Delta f) \log\left(\frac{f}{M} \left(1 + \frac{\Delta f}{f}\right)\right) \\ &\leq (f + \Delta f) \left(\log\left(\frac{f}{M}\right) + \frac{\Delta f}{f}\right) \\ &\leq f \log\left(\frac{f}{M}\right) + \Delta f \log\left(\frac{f}{M}\right) + \Delta f + \frac{\Delta f^2}{f}. \end{aligned}$$

Summing this inequality (with  $f = f_i$  and  $\Delta f = \Delta t F P_i^L$ ) over  $i \in I$ , the first term of the right-hand side gives  $H$ , the third vanishes by mass conservation, and one obtains

$$\bar{H} \leq H + \Delta t \sum_{i \in I} F P_i^L \log(f_i/M_i) + (\Delta t)^2 (F P_i^L)^2 / f_i \stackrel{def}{=} \tilde{H}.$$

Note that, as previously shown, we have if  $f_i \neq M_i$  for at least one index  $i \in I$

$$\sum_{i \in I} F P_i^L \log(f_i/M_i) < 0.$$

Therefore, the scheme is entropy decaying provided that the time step verifies

$$\Delta t \leq \Delta t_2 \stackrel{def}{=} \min\left(\Delta t_1, \frac{-\sum_{i \in I} F P_i^L (\log(f_i/M_i))}{\sum_{i \in I} (F P_i^L)^2 / f_i}\right) < \infty.$$

By the definition (2.15) of  $F P_i^L$ , and applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned} (F P_i^L)^2 &\leq \left( \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s} g_{i,i+\mu e_s} \right) \\ &\times \left( \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s} g_{i,i+\mu e_s} \log^2\left(\frac{h_{i+\mu e_s}}{h_i}\right) \right). \end{aligned}$$

The first term of the right-hand side is less than  $2f_i \sum_{\mu \in \{-1,1\}} \sum_{s=1,\dots,d} a_{i,i+\mu e_s}$  using again (2.9). The second term already appears in the computation of  $\frac{dH}{dt}$ . Dividing this inequality by  $f_i$ , using the definition  $a_{i,i+\mu e_s}$ , and summing over the indices  $i$  give

$$\sum_{i \in I} \frac{(F P_i^L)^2}{f_i} \leq \frac{4d}{\Delta v^2} \left( - \sum_{i \in I} F P_i^L (\log(f_i/M_i)) \right).$$

We set  $\delta t = \frac{\Delta v^2}{4d}$ . Note that  $\delta t \geq 2\Delta t_1$  since

$$\inf_{x \in ]1, K(0)]} \frac{(x-1)}{x \log x} \leq 1.$$

Then, any  $\Delta t$  smaller than  $\Delta t_1$  yields to a positive, entropy, and nonoscillatory scheme. Note that the entropy condition is less restrictive than the positivity one in the linear case. The scheme (2.20) defines a sequence of  $\{f_i^n\}_{n \in \mathbb{N}}$ , which serves us to define a piecewise-constant in time function and  $v$  by

$$f_i(t) = f_i^n \quad \forall t \in [t_n, t_{n+1}[,$$

with  $t_0 = 0$ , and  $t_{n+1} = t_n + \Delta t$  and the  $f_i^n$  defined above by recursion. We have

$$K^{-1}M_i \leq f_i \leq KM_i \quad \forall i \in I,$$

with  $K = K(0)$ , defined above. We have now to prove that this discrete solution tends to the Maxwellian. First, note that the associated entropy  $H$  defined by

$$H(t) = \sum_{i \in I} f_i^n \log(f_i^n / M_i) \quad \forall t \in [t_n, t_{n+1}[$$

is decreasing and has a lower bound. Thus, it converges and the Cauchy criteria implies that  $(H_{n+1} - H_n) \rightarrow 0$  as  $n \rightarrow \infty$  where we have set  $H_n = H(t_n)$ . Define  $\tilde{H}$  for  $t \in [t_n, t_{n+1}[$  as previously by

$$\tilde{H}(t) = H(t_n) + (t - t_n) \sum_{i \in I} FP_i^L \log(f_i^n / M_i) + (t - t_n)^2 (FP_i^L)^2 / f_i^n.$$

Then,  $\tilde{H}$  satisfies

$$\tilde{H}(t_n) = H(t_n), H(t) - H_n \leq \tilde{H}(t) - H_n \leq 0 \quad \forall t \in [t_n, t_{n+1}[.$$

Thus, we have for  $\Delta t = t_{n+1} - t_n \leq \Delta t_1 \leq \delta t / 2$  using  $(H_{n+1} - H_n) \rightarrow 0$ :

$$\Delta t_1 \sum_{i \in I} (FP_i^L)^2 / f_i + \sum_{i \in I} FP_i^L \log(f_i^n / M_i) \rightarrow 0,$$

and, since we have proved

$$\left( - \sum_{i \in I} FP_i^L \log(f_i^n / M_i) \right) \geq (\delta t) \left( \sum_{i \in I} (FP_i^L)^2 / f_i^n \right),$$

we obtain  $\sum_{i \in I} FP_i^L \log(f_i^n / M_i) \rightarrow 0$  when  $n \rightarrow \infty$  which is equivalent to  $f_i^n \rightarrow M_i \quad \forall i \in I$  since the  $g_i$  have a uniform lower bound. As for the semidiscretized model, the convergence of the sequence  $f_i^k$  toward the equilibrium can also be proved using the Csiszar-Kullback inequality and the fact that under the time-step restrictions to ensure the decay of the entropy and the positivity of the distribution function we have necessarily  $\lim_{k \rightarrow \infty} H^k = 0$ .

This concludes the proof.  $\square$

**A2. Negative solution for the nonlog form of nonlinear FPL equation.**

The nonlog formulation of the discretized nonlinear FPL equation (3.3) can be written as

$$\frac{\partial f_i}{\partial t} = (D^* \cdot p)_i, \quad p_i = \Delta v^3 \sum_{j \in \mathcal{I}} \Phi((i-j)\Delta v) (f_j(Df)_i - f_i(Df)_j), \quad i \in I.$$

Let us consider, for example,  $n = 8$ , and the following initial data:  $f_i = \mu \forall i \in I$  except for two indices  $i_0 = (2, 2, 2)$  and  $j_0 = (6, 6, 6)$  where  $f_{i_0} = f_{j_0} = 1$ . We choose for the log form the simplest centered operator, i.e.,  $D_c$ . First, let  $\mu = 0$ . The system defined above gives a negative value for the derivative of  $f_k$  with respect to  $t$  at time  $t = 0$  for  $k = (7, 6, 7)$ . A numerical computation of the terms  $p_i$  gives

$$\frac{d}{dt} f_k(t=0) = -1.457165e^{-04}.$$

Therefore, a positive initial condition gives a solution such that  $\exists k \in I, f_k(t) < 0$  for any arbitrary small value of  $t$ . Finally, the continuity of the solution of the nonlog formulation of FPL equation with respect to the initial data implies that, for  $\mu$  small enough, there exists  $t_\mu$  such that  $f_k(t_\mu) < 0$  for some indices  $k$ . Thus, a strictly positive initial data becomes negative after finite time ( $t_\mu$ ) when using the nonlog form of the FPL equation. On the contrary, this proof does not apply to the log form which is singular when some  $f_k$  tends to 0 and, thus, its solutions do not depend continuously on the initial data near such situations (with vanishing weights). Numerical simulations illustrating this property are available by sending e-mails to the authors.

## REFERENCES

- [1] A.A. ARSENE'V AND O.E. BURYAC, *On the connection between a solution of the Boltzmann equation and a solution of the Landau-Fokker-Planck equation*, Math. USSR Sbornik, 69 (1991), pp. 465–478.
- [2] A.A. ARSENE'V AND N.V. PESKOV, *On the existence of a generalized solution of Landau's equation*, USSR Comput. Math. Math. Phys., 17 (1977), pp. 241–246.
- [3] YU.A. BEREZIN, V.N. KHUDICK, AND M.S. PEKKER, *Conservative finite difference schemes for the Fokker-Planck equation not violating the law of an increasing entropy*, J. Comput. Phys., 69 (1987), pp. 163–174.
- [4] A.V. BOBYLEV AND G. TOSCANI, *On the generalization of the Boltzmann H-theorem for a spatially homogeneous Maxwell gas*, J. Math. Phys., 33 (1992), p. 7.
- [5] A.V. BOBYLEV, I.F. POTAPENKO, AND V.A. CHUYANOV, *Kinetic equations of the Landau-type as a model of the Boltzmann equation and completely conservative difference schemes*, U.S.S.R. Comput. Math. Math. Phys., 20, 4 (1981), pp. 190–201.
- [6] C. BUET AND S. CORDIER, *Numerical Scheme for the Spherical Fokker-Planck-Landau operator*, J. Comput. Phys., preprint in L.A.N. (Paris 6) 97032, 1997; submitted.
- [7] C. BUET, S. CORDIER, P. DEGOND, AND M. LEMOU, *Fast algorithms for numerical conservative and entropy approximations of the Fokker-Planck-Landau operator*, J. Comput. Phys., 133 (1997), pp. 1036–1053.
- [8] H. COHN, *Numerical integration of the Fokker-Planck equation and the evolution of star clusters*, The Astrophysical Journal, 234 (1979), pp. 1036–1053.
- [9] H. COHN, *Late core collapse in star clusters and the gravothermal instability*, The Astrophysical Journal, 242 (1980), pp. 765–771.
- [10] J.F. CLOUET AND K. DOMELEVO, *Solution of a kinetic stochastic equation modelling a spray in a turbulence gas flow*, R.I. 330, CMAP, École Polytechnique, Palaiseau, France, 1996; <http://www.cmap.polytechnique.fr>.
- [11] P. DEGOND AND B. LUCQUIN-DESREUX, *The Fokker-Planck asymptotics of the Boltzmann collision operator in the Coulomb case*, Math. Models Methods Appl. Sci., 2 (1992), pp. 167–182.

- [12] P. DEGOND AND B. LUCQUIN-DESREUX, *An entropy scheme for the Fokker-Planck collision of plasma kinetic theory*, Numer. Math., 68 (1994), pp. 239–262.
- [13] L. DESVILLETES, *On asymptotics of the Boltzmann equation when the collisions become grazing*, Transport Theory and Statist. Phys., 21 (1992), pp. 259–276.
- [14] L. DESVILLETES AND C. VILLANI, *On the Spatially Homogeneous Landau Equation for Hard Potentials. Part I: Existence, Uniqueness and Smoothness*, Preprint du DMI, ENS de Paris, Paris, France, 1998.
- [15] L. DESVILLETES AND C. VILLANI, *On the Spatially Homogeneous Landau Equation for Hard Potentials. Part II: H-theorem and Applications*, Preprint du DMI, ENS de Paris, Paris, France, 1998.
- [16] R. EYMARD, T. GALLOUET, AND R. HERBIN, *Schémas de type volume finis*, École cea-edf-inria, problèmes nonlinéaires appliqués, Paris, 1992.
- [17] E. FRENOD AND B. LUCQUIN-DESREUX, *On conservative and entropy discrete axisymmetric Fokker-Planck operators*, RAIRO Modél. Math. Anal. Numér., 33 (1998), pp. 307–339.
- [18] R. ILLNER AND S. RJASANOW, *Random discrete velocity method: possible bridges between the Boltzmann equation, discrete velocity models and particle simulation?* Nonlinear Kinetic Theories and Mathematical Aspects of Hyperbolic Systems, in Proceedings, Rapallo, 1992.
- [19] S. KULLBACK, *A lower bound for discrimination information interms of variation*, IEEE Trans. Inform. Theory, 4 (1967), pp. 126–127.
- [20] O. LARROCHE, *Kinetic simulations of a plasma collision experiment*, Phys. Fluids B, 5 (1993), pp. 2816–2840.
- [21] M. LEMOU, *Exact solutions of the Fokker-Planck equation*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 579–583.
- [22] M. LEMOU, *Multipole expansions for the Fokker-Planck equation*, Numer. Math., 78 (1998), pp. 597–618.
- [23] O. LIE-SVENDSEN AND M.H. REES, *An improved kinetic model for the polar outflow of a minor ion*, J. Geophys. Res., (1995).
- [24] S. LIVI AND E. MARSCH, *Generation of solar wind proton tails and double beams by coulomb collisions*, J. Geophys. Res., 92 (1987), pp. 7255–7261.
- [25] B. LUCQUIN-DESREUX, *Discrétisation de l'opérateur de Fokker-Planck dans le cas homogène*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 407–411.
- [26] M.S. PEKKER AND V.N. KHUDIK, *Conservative difference schemes for the Fokker-Planck equation*, U.S.S.R. Comput. Math. Math. Phys., 24 (1984), pp. 206–210.
- [27] I.F. POTAPENKO AND V.A. CHUYANOV, *A completely conservative difference scheme for the two-dimensional Landau equation*, U.S.S.R. Comput. Math. Math. Phys., 20 (1980), pp. 249–253.
- [28] G. TOSCANI, *Entropy Production and the Rate of Convergence to Equilibrium for the Fokker-Planck Equation*, preprint, Dept. of Math., Univ. Pavia, Pavia, Italy, 1997.
- [29] B. WENNBORG AND A. PULVIRENTI, *A Maxwellian Lower Bound for the Boltzmann Equation*, preprint, 1996.
- [30] G.R. WILSON, *Semikinetic modeling of the outflow of ionospheric plasma through the topside collisional to collisionless transition region*, J. Geophys. Res., 97 (1992), pp. 10551.



## NUMERICAL SOLUTION OF AN IONIC FOKKER–PLANCK EQUATION WITH ELECTRONIC TEMPERATURE\*

C. BUET<sup>†</sup>, S. DELLACHERIE<sup>‡</sup>, AND R. SENTIS<sup>†</sup>

**Abstract.** We describe a numerical scheme for dealing with an ion/electron collision operator of the Fokker–Planck type; for that purpose, we introduce the notion of the *entropic average* of two positive quantities. This scheme has the property to be entropic in the sense of Boltzmann’s H-theorem under a *CFL* criteria. Moreover, we prove that the solution of the semidiscrete scheme converges towards a unique Maxwellian equilibrium state when the time grows. Numerical applications are given and show that our scheme is more precise than the classical Chang–Cooper one.

**Key words.** kinetics model, Fokker–Planck–Landau equation, plasma physics, numerical scheme

**AMS subject classifications.** 65M06, 65M12, 82C40, 82D10

**PII.** S0036142999359669

**Introduction.** In hot plasmas such as the plasmas of microballs in the inertial confinement fusion framework (see [1]), the characteristic length of variation of hydrodynamic quantities may be of the order of a micrometer, but when the temperature arises up to some KeV, the mean free path of the ions may be greater than a micrometer. Thus, to simulate the behavior of such a plasma, it is necessary to study the kinetic models of the evolution of ions coupled with an electronic population which is assumed to be Maxwellian. Before writing the relevant kinetic model, we recall the most simple fluid model for hot plasmas which is called *two-temperatures Euler equations* (in what follows, we assume that there is only one ionic species whose atomic mass is  $m$ , the ionization level being  $Z$ ):

$$(0.1) \quad \frac{\partial}{\partial t} N + \nabla_x \cdot (N \vec{U}) = 0,$$

$$(0.2) \quad \frac{\partial}{\partial t} (mN \vec{U}) + \nabla_x \cdot (mN \vec{U} \otimes \vec{U}) + \nabla_x (NT + P_e) = \vec{0},$$

$$(0.3) \quad \frac{\partial}{\partial t} \left( \frac{3}{2} NT \right) + \nabla_x \cdot \left( \frac{3}{2} NT \vec{U} \right) + NT \nabla_x \cdot \vec{U} = 3\Omega N (T_e - T),$$

$$(0.4) \quad \frac{\partial}{\partial t} \mathcal{E}_e(T_e) + \nabla_x \cdot [\mathcal{E}_e(T_e) \vec{U}] + P_e \nabla_x \cdot \vec{U} = 3\Omega N (T - T_e),$$

where  $N$ ,  $\vec{U}$ ,  $T$ , and  $T_e$  are, respectively, the ionic density, the ionic macroscopic velocity, the ionic temperatures, and electronic temperatures.  $\mathcal{E}_e(T_e) = \frac{3}{2} ZNT_e$  and  $P_e = ZNT_e$  are the internal energy and the pressure of the electrons.  $\Omega < 0$ , defined with (4.1), is the collision frequency and is of the form  $\Omega = N \cdot \Omega_0$ , where  $\Omega_0$  depends continuously on  $N$  and  $T_e^{-3/2}$  (see [2], [3], or [4]). For the numerical treatment of this macroscopic model, see [5] and [6]; for the physical analysis of this model and the link with a two fluid model, see, for example, [3].

---

\*Received by the editors July 30, 1999; accepted for publication (in revised form) March 15, 2001; published electronically September 19, 2001.

<http://www.siam.org/journals/sinum/39-4/35966.html>

<sup>†</sup>Commissariat à l’Énergie Atomique, 91680 Bruyères-le-Châtel, France (buet@bruyeres cea.fr, sentis@bruyeres cea.fr).

<sup>‡</sup>Commissariat à l’Énergie Atomique, 91191 Gif sur Yvette, France (stephane.dellacherie@cea.fr).

**The kinetic model.** Now we are concerned with the kinetic model. The ionic distribution  $f = f(t, x, v)$  ( $x \in \mathbf{R}^3$  and  $\vec{v} \in \mathbf{R}^3$ ) and the electronic temperature  $T_e(t, x)$  are solutions of

$$(0.5) \quad \frac{\partial}{\partial t} f + \vec{v} \cdot \nabla_x f - \frac{\nabla_x P_e}{Nm} \cdot \nabla_v f = B(f) + S(f),$$

$$(0.6) \quad \frac{\partial}{\partial t} \mathcal{E}_e(T_e) + \nabla_x \cdot [\mathcal{E}_e(T_e) \vec{U}] + P_e \nabla_x \cdot \vec{U} = -\frac{m}{2} \langle v^2 S(f) \rangle.$$

We have set  $\langle \bullet \rangle = \int \bullet d\vec{v}$  and we define  $N, \vec{U}$  by

$$N = \langle f \rangle, \quad N \vec{U} = \langle f \vec{v} \rangle.$$

The Fokker–Planck operator  $S(f)$  is a model describing the collisions of the ions against the electrons and is defined with

$$(0.7) \quad S(f)(\vec{v}) = \Omega \nabla_v \cdot \left[ (\vec{v} - \vec{U}) f + \frac{T_e}{m} \nabla_v f \right].$$

$B(f)$  is the classical quadratic Landau operator (see [2], [3], and [4]). On the origin of the system (0.5) and (0.6), see [7] and the references therein; for a mathematical approach, see [12] and [13]. For some general features related to its numerical treatment, see [7] and [9] (in the stationary case).

The numerical solution of the overall system (0.5) and (0.6) can be done with a finite difference method in the phase space  $(x, \vec{v})$  with a splitting in five stages:

(1) resolution of  $\frac{\partial}{\partial t} f + \vec{v} \cdot \nabla_x f = 0$  with an upwind scheme with respect to the  $x$  variable;

(2) resolution of the ion/ion Fokker–Planck operator, i.e., we solve  $\frac{\partial}{\partial t} f = B(f)$ . For example, see [11], [14], [16], and [18] for a conservative and entropic scheme (and [17] in the isotropic case);

(3) resolution of  $\frac{\partial}{\partial t} f = \frac{\nabla_x P_e}{Nm} \cdot \nabla_v f$  with an upwind scheme with respect to the  $\vec{v}$  variable;

(4) resolution of the Fokker–Planck operator  $S(f)$ , i.e., we solve

$$(0.8) \quad \begin{cases} \frac{\partial}{\partial t} f = S(f), \\ \frac{\partial}{\partial t} \mathcal{E}_e = -\frac{m}{2} \langle v^2 S(f) \rangle; \end{cases}$$

(5) resolution of the remaining part of the electronic energy equation.

In this paper, we describe only the fourth stage, which is the most technical. Thus, we introduce the notion of *entropic average* which allows us to build a numerical scheme with very strong convergence and stability properties. To our knowledge, even the classical Chang–Cooper [8] numerical scheme does not realize all these properties (see also [9] and [10]).

**Plan of the paper.** In section 1, we give the main properties of the kinetic system: We check that there exists an entropy which is the sum of an ionic part equal to  $\langle f \log f \rangle$  and an electronic part. This entropy is decreasing with time, and we check that if  $f$  is a Maxwellian function, the first three moments of (0.5) yield

the two-temperatures Euler equations. Section 2 is devoted to the introduction of the *entropic average* and to the analysis of a semidiscretization of the system (0.8) with respect to the velocity variable: the *entropic average* allows us to build a scheme whose discretized distribution  $f$  converges in large time  $t$  to the projection on the velocity grid of a Maxwellian distribution, properties which are not shown with other schemes as those described in [8] and [10].

In section 3, we describe the full discretization of the system (0.8): first, we build a positive and entropic explicit conservative scheme under a classical *CFL* criteria, and second, we build a semi-implicit conservative scheme which preserves the thermodynamical equilibrium. Finally, in section 4, we give numerical results which show that our scheme is more precise than the classical Chang-Cooper one (cf. [8]).

**1. Preliminaries.** Let us remark that even with a constant  $\Omega$ , the operator  $S(f)$  is not a linear operator with respect to  $f$ . Indeed,  $\vec{U}$  depends on  $f$  and we can write

$$S(f)(\vec{v}) = \Omega_0 \nabla_v \cdot \left[ \int (\vec{v} - \vec{v}_*) f(\vec{v}_*) f(\vec{v}) d\vec{v}_* + \int f(\vec{v}_*) d\vec{v}_* \frac{T_e}{m} \nabla_v f \right].$$

Let us define the Maxwellian

$$\mathbf{M}_{\vec{U}, T}(\vec{v}) = \frac{N}{(2\pi T/m)^{3/2}} \exp \left[ -\frac{m(\vec{v} - \vec{U})^2}{2T} \right].$$

We can write the operator  $S(f)$  in the Landau form

$$(1.1) \quad S(f) = \Omega \frac{T_e}{m} \nabla_v \cdot \left[ f \nabla_v \log \left( f / \mathbf{M}_{\vec{U}, T_e} \right) \right].$$

We do not emphasize in this paper the domain of the operator  $S(f)$ , but we assume in the following that  $f$  is a positive function belonging to  $L^1[(1 + |\vec{v}|^2)dv]$  and that  $|\vec{v}|^2 f(v) \rightarrow 0$  and  $|\vec{v}|^2 \nabla_v f \rightarrow 0$  when  $|\vec{v}| \rightarrow +\infty$ .

For any  $f$ , we introduce an ionic temperature defined by

$$(1.2) \quad 3NT = m \langle (\vec{v} - \vec{U})^2 f \rangle.$$

Using the properties

$$\begin{aligned} \int \int (\vec{v} - \vec{v}_*) f(\vec{v}_*) f(\vec{v}) d\vec{v}_* d\vec{v} &= \vec{0}, \\ m \int \int \vec{v} (\vec{v} - \vec{v}_*) f(\vec{v}_*) f(\vec{v}) d\vec{v}_* d\vec{v} + NT_e \int \nabla_v f d\vec{v} &= 3N(T - T_e), \end{aligned}$$

we can check that the operator  $S(f)$  satisfies

$$(1.3) \quad \langle S(f) \rangle = 0, \quad \langle S(f) \vec{v} \rangle = \vec{0}, \quad \frac{m}{2} \langle S(f) v^2 \rangle = 3\Omega N(T_e - T).$$

Moreover, by using the Landau form and by assuming that  $f = 0$  when  $|\vec{v}| \rightarrow +\infty$ , we easily see that

$$\left\langle \log \left( f / \mathbf{M}_{\vec{U}, T_e} \right) S(f) \right\rangle = -\Omega \frac{T_e}{m} \int f \left[ \nabla_v \log \left( f / \mathbf{M}_{\vec{U}, T_e} \right) \right]^2 d\vec{v},$$

and we get the following lemma.

LEMMA 1.1. *For all  $f > 0$  and  $T_e > 0$ , we have*

$$(1.4) \quad \left\langle S(f) \log \left( f / \mathbf{M}_{\vec{U}, T_e} \right) \right\rangle \leq 0.$$

Moreover,  $\langle S(f) \log(f / \mathbf{M}_{\vec{U}, T_e}) \rangle = 0$  if and only if  $f = \mathbf{M}_{\vec{U}, T_e}$ .

We know that the operator  $B(f)$  conserves the mass, the momentum, and the energy and we have  $\langle B(f) \log f \rangle \leq 0$ . Thus

$$(1.5) \quad \left\langle B(f) \log \left( f / \mathbf{M}_{\vec{U}, T_e} \right) \right\rangle \leq 0.$$

Lemma 1.1 with (1.5) gives the following proposition.

PROPOSITION 1.2. *Let  $f$  and  $T_e$  be solutions of (0.5) and (0.6) with  $T_e$  smooth enough with respect to the variable  $x$ . Then we have the relation of the decay of the entropy*

$$(1.6) \quad \frac{\partial}{\partial t} (\langle f \log f \rangle + \mathcal{H}_e) + \nabla_x \cdot (\langle \vec{v} f \log f \rangle + \vec{U} \mathcal{H}_e) \leq 0, \quad \text{where } \mathcal{H}_e = ZN \log(NT_e^{-3/2}).$$

*Proof of Proposition 1.2.* Let us denote  $\overline{\overline{P_i}} = m \langle f(\vec{v} - \vec{U}) \otimes (\vec{v} - \vec{U}) \rangle$ . By taking the two first moments of the kinetic equation (0.5), a classical calculus yields

$$(1.7) \quad \frac{\partial}{\partial t} (mN\vec{U}) + \nabla_x \cdot (mN\vec{U} \otimes \vec{U}) + \nabla_x (ZNT_e) + \nabla_x \cdot \overline{\overline{P_i}} = \vec{0}.$$

Now, using the classical relation

$$(1.8) \quad \frac{\partial}{\partial t} (Nw) + \nabla_x \cdot (Nw\vec{U}) = N \left( \frac{\partial}{\partial t} w + \vec{U} \cdot \nabla_x w \right),$$

which is valid for any  $w$ , we get

$$(1.9) \quad N \left( \frac{\partial}{\partial t} (N^{-1}) + \vec{U} \cdot \nabla_x (N^{-1}) \right) - \nabla_x \cdot \vec{U} = 0,$$

$$mN \left( \frac{\partial}{\partial t} \vec{U} + \vec{U} \cdot \nabla_x \vec{U} \right) + \nabla_x (ZNT_e) + \nabla_x \cdot \overline{\overline{P_i}} = \vec{0}.$$

If we multiply the kinetic equation (0.5) with  $mv^2/2$ , we get

$$(1.10) \quad m \frac{\partial}{\partial t} \left\langle \frac{v^2}{2} f \right\rangle + m \nabla_x \cdot \left\langle \vec{v} \frac{v^2}{2} f \right\rangle + \vec{U} \cdot \nabla_x (ZNT_e) = 3\Omega N(T_e - T).$$

(i) *The electronic entropy.* According to (0.4) and (1.8), the electronic energy balance equation may be written as

$$\frac{3}{2} \left( \frac{\partial T_e}{\partial t} + \vec{U} \cdot \nabla_x T_e \right) + T_e \nabla_x \cdot \vec{U} = 3 \frac{\Omega}{Z} (T - T_e).$$

Thus

$$\frac{\partial}{\partial t} \log(NT_e^{-3/2}) + \vec{U} \cdot \nabla_x \log(NT_e^{-3/2}) = 3 \frac{\Omega}{Z} \frac{(T_e - T)}{T_e};$$

that is to say

$$(1.11) \quad \frac{\partial \mathcal{H}_e}{\partial t} + \nabla_x \cdot (\vec{U} \mathcal{H}_e) = 3\Omega N \frac{(T_e - T)}{T_e}.$$

(ii) *The ionic entropy.* According to Lemma 1.1, the inequality (1.5), the relation  $\langle \partial_t f + \vec{v} \cdot \nabla_x f \rangle = 0$ , and the relation  $\langle \nabla_v f \rangle = \vec{0}$ , we see that

$$\left\langle \left( \log f + \frac{m(\vec{v} - \vec{U})^2}{2T_e} \right) \frac{\partial}{\partial t} f \right\rangle + \left\langle \left( \log f + \frac{m(\vec{v} - \vec{U})^2}{2T_e} \right) \nabla_x \cdot (f \vec{v}) \right\rangle \leq 0.$$

Since  $\langle f \frac{\partial}{\partial t} \log f \rangle + \langle f \vec{v} \cdot \nabla_x \log f \rangle = 0$ , we get

$$\begin{aligned} \frac{\partial}{\partial t} \langle f \log f \rangle + \nabla_x \cdot \langle \vec{v} f \log f \rangle &\leq \frac{m}{T_e} \left[ - \left\langle \frac{\partial}{\partial t} \frac{v^2}{2} f \right\rangle + \left\langle \vec{v} \cdot \vec{U} \frac{\partial}{\partial t} f \right\rangle - \nabla_x \cdot \left\langle \vec{v} \frac{v^2}{2} f \right\rangle \right. \\ &\quad \left. + \langle (\vec{v} \cdot \vec{U}) \nabla_x \cdot (f \vec{v}) \rangle \right]. \end{aligned}$$

On the other hand, we can notice that

$$\begin{aligned} m \left( \left\langle \vec{v} \cdot \vec{U} \frac{\partial}{\partial t} f \right\rangle + \langle (\vec{v} \cdot \vec{U}) \nabla_x \cdot (f \vec{v}) \rangle \right) &= m \left[ \vec{U} \cdot \partial_t (N \vec{U}) + \vec{U} \cdot \langle \nabla_x \cdot (f \vec{v} \otimes \vec{v}) \rangle \right] \\ &= \vec{U} \cdot \left[ \partial_t (mN \vec{U}) + \nabla_x \cdot (mN \vec{U} \otimes \vec{U}) \right] \\ &\quad + \vec{U} \cdot \nabla_x \bar{P}_i. \end{aligned}$$

However, according to (1.7), this last expression is equal to

$$-\vec{U} \cdot \nabla_x \bar{P}_i - \vec{U} \cdot \nabla_x (ZNT_e) + \vec{U} \cdot \nabla_x \bar{P}_i = -\vec{U} \cdot \nabla_x (ZNT_e).$$

By gathering the previous relations and (1.10), we get

$$\frac{\partial}{\partial t} \langle f \log f \rangle + \nabla_x \cdot \langle \vec{v} f \log f \rangle \leq 3\Omega N \frac{(T - T_e)}{T_e}$$

which, by adding (1.11), gives the result.  $\square$

*Remark.* If the ionic distribution function is Maxwellian, we have  $\bar{P}_i = \bar{1}NT$  and, according to (1.10), we obtain

$$\left( \frac{\partial}{\partial t} + \nabla_x \cdot (\vec{U} \cdot) \right) \left( \frac{3}{2} NT + N \frac{m}{2} |\vec{U}|^2 \right) + \vec{U} \cdot \nabla_x (ZNT_e) + \nabla_x \cdot (\vec{U} NT) = 3\Omega N (T_e - T).$$

Thus, according to (1.7), we see that (0.3) is satisfied. Then,  $N$ ,  $U$ ,  $T$ , and  $T_e$  satisfy the fluid system (0.1), (0.2), (0.3), and (0.4).

**2. Semidiscretized scheme for the ion/electron Fokker–Planck operator.** For the sake of simplicity, we shall consider only the monodimensional Cartesian case; that is,  $f$  depends only of  $t \in \mathbf{R}^+$  and  $v \in \mathbf{R}$ . Then,  $\mathcal{E}_e(T_e) = \frac{Z}{2}NT_e$  and  $3NT$  has to be replaced by  $NT$  in (1.2) and the system (0.8) becomes

$$(2.1) \quad \begin{cases} \frac{\partial}{\partial t} f = S(f), \\ \frac{Z}{2} \frac{\partial}{\partial t} (NT_e) = \Omega N(T - T_e) \end{cases}$$

with

$$S(f) = \Omega \partial_v \left[ (v - U)f + \frac{T_e}{m} \partial_v f \right].$$

We will study discretization in velocity only (semidiscretized model). We use a discretization of  $\mathbf{R}$  by a finite difference grid  $\{v_j\}$  ( $j \in \{1, \dots, j_{max}\}$ ) and  $\Delta v = v_{j+1} - v_j$  is constant. Let us note  $f_j(t)$ , the evaluation of  $f(t)$  at the point  $v_j$ . From now on, we set

$$\langle \Psi \rangle = \Delta v \sum_j \Psi_j.$$

Let us recall that, in the continuous case, the collision operator  $S(f)$  verifies the conservation properties (1.3), which implies in the homogeneous case that  $\partial_t N = 0$  and  $\partial_t U = 0$ ; these properties will have to be verified in the discretized case (which will be verified; see Proposition 2.2).

**2.1. Definition of the semidiscretized scheme and preliminary results.**  
**Boundary conditions in the general continuous case for a compact velocity domain.** Since we will consider compact velocity domain  $\mathcal{V}$  for the numerical applications, we briefly study the boundary conditions that we have to impose in the general continuous case to verify again the conservation properties (1.3) when the velocity domain is compact. The first consequence is that we cannot say that  $f$  and  $\nabla_v f$  are equal to zero on the frontier  $\delta\mathcal{V}$  of the compact velocity domain and, then, we must give new boundary conditions even for the general continuous system; for this one, in order to ensure the mass conservation, the best boundary conditions are the Robin ones

$$(2.2) \quad (\vec{v} - \vec{U})f + \frac{T_e}{m} \nabla_v f = \vec{0} \quad \text{when} \quad \vec{v} \in \delta\mathcal{V}.$$

And, to ensure the momentum conservation, we have to define the macroscopic velocity  $\vec{U}$  with

$$(2.3) \quad \vec{U} = \frac{\int_{\mathcal{V}} \vec{v} f(v) dv}{\int_{\mathcal{V}} f(v) dv} + \delta \vec{u},$$

where

$$(2.4) \quad \delta \vec{u} = \frac{T_e}{m \int_{\mathcal{V}} f(v) dv} \int_{\delta\mathcal{V}} f(v) d\vec{S},$$

$d\vec{S}$  being the measure of  $\delta\mathcal{V}$ . And, due to the corrective term  $\delta \vec{u}$ , it is easy to verify that

$$(2.5) \quad \int_{\mathcal{V}} m \frac{v^2}{2} S(f) dv = d\Omega \cdot (T_e - T) \cdot \int_{\mathcal{V}} f(v) dv$$

( $d$  is the dimension of the velocity space  $\mathcal{V}$ ) with

$$(2.6) \quad \begin{cases} T = \frac{m}{d \int_{\mathcal{V}} f(v) dv} \int_{\mathcal{V}} (\vec{v} - \vec{U})^2 f(v) dv + \delta t, \\ \delta t = \frac{T_e}{d \int_{\mathcal{V}} f(v) dv} \int_{\delta \mathcal{V}} (v - U) f(v) d\vec{S}. \end{cases}$$

We will use this remark at the discretized level when  $j = 1$  and  $j = j_{max}$ . Indeed, we define the following numerical scheme with the following discrete boundary conditions.

**Statement of the semidiscrete scheme.** For initial conditions  $f^0$  and  $T_e^0$ , we consider the following scheme:

$$(2.7) \quad \begin{cases} \partial_t f_j = S(f)_j & \forall j \in \{1, \dots, j_{max}\} : f_j(0) = f_j^0, \\ \frac{Z}{2} \partial_t (NT_e) = \Omega(\widetilde{NT} - NT_e), & T_e(0) = T_e^0, \end{cases}$$

with

$$(2.8) \quad S(f)_j = \frac{\Omega}{\Delta v} \left[ (v_{j+1/2} - \widetilde{U}) \widetilde{f}_{j+1/2} - (v_{j-1/2} - \widetilde{U}) \widetilde{f}_{j-1/2} \right] + \frac{\Omega T_e}{m \Delta v^2} (a_j f_{j+1} - b_j f_j + c_j f_{j-1})$$

and

$$N = \langle f \rangle,$$

$$(2.9) \quad \begin{cases} \widetilde{N} = \sum_j \widetilde{f}_{j+1/2} \Delta v, \\ \widetilde{U} = \sum_j v_{j+1/2} \widetilde{f}_{j+1/2} \Delta v / \widetilde{N} + \delta \widetilde{u}, \\ \widetilde{NT} = \sum_j m (v_{j+1/2} - \widetilde{U})^2 \widetilde{f}_{j+1/2} \Delta v + \widetilde{N} \cdot \delta \widetilde{t}. \end{cases}$$

$\widetilde{f}_{j+1/2}(t)$  is an appropriate average of  $f_j(t)$  and  $f_{j+1}(t)$ : we will see that we will use the *entropic average* (see Definition 2.1) to define this average.

$\delta \widetilde{u}$  and  $\delta \widetilde{t}$  are corrective terms necessary for the conservation of momentum and energy, knowing that the distribution  $f$  is not always equal to zero on the boundary of  $[v_1, v_{j_{max}}]$ . These terms are defined in the following way:

$$(2.10) \quad \begin{cases} \delta \widetilde{u} = \frac{T_e}{m \widetilde{N}} (f_{j_{max}} - f_1), \\ \delta \widetilde{t} = \frac{T_e}{\widetilde{N}} \left[ f_{j_{max}} (v_{j_{max}+1/2} - \widetilde{U}) + f_1 (\widetilde{U} - v_{1/2}) \right]. \end{cases}$$

Let us remark that (2.9) and (2.10) are the discrete versions of the relations (2.3), (2.4), and (2.6) (see the proof of Proposition 2.2 for a justification of these formulas).

From a practical point of view, we can say that when  $|v_1|$  and  $|v_{j_{max}}|$  will be large enough, these corrective terms will be very small since the distributions  $\{f_j\}$  converge to the projection of a Maxwellian distribution on the velocity mesh  $\{v_j\}$

when  $t$  goes to infinity; see Theorem 2.8. Then, it would be possible to forget them for the numerical applications.

**Discrete boundary conditions.** To take into account the Robin's boundary conditions (2.2) at the boundary of the velocity domain  $\mathcal{V} = [v_1, v_{j_{\max}}]$ , we set

$$(2.11) \quad \begin{cases} a_j = 1 \text{ if } j \neq j_{\max}, \\ b_j = 2 \text{ if } j \in \{2, \dots, j_{\max} - 1\}, \\ c_j = 1 \text{ if } j \neq 1, \\ b_1 = b_{j_{\max}} = 1 \text{ and } a_{j_{\max}} = c_1 = 0 \end{cases}$$

and

$$(2.12) \quad \tilde{f}_{1/2} = \tilde{f}_{j_{\max}+1/2} \equiv 0.$$

Now, we introduce the *entropic average*.

DEFINITION 2.1. *The entropic average  $\tilde{f}$  of two strictly positive quantities  $x$  and  $y$  is defined by*

$$\tilde{f} = \begin{cases} \frac{x-y}{\log x - \log y} & \text{if } x \neq y, \\ x & \text{otherwise.} \end{cases}$$

By continuity, we extend this definition by setting  $\tilde{f} = 0$  if  $x = 0$  or  $y = 0$ .

Notice that for any smooth and positive function  $f$ , we have

$$f(v + \Delta v/2) = \frac{f(v + \Delta v) - f(v)}{\log f(v + \Delta v) - \log f(v)} + O(\Delta v^2),$$

which makes the discretized operator (2.8) of second order in velocity space.

**Preliminary results.**

PROPOSITION 2.2. *The conservation laws*

$$\begin{cases} \langle S(f) \rangle = 0, \\ \langle vS(f) \rangle = 0, \\ m\langle \frac{v^2}{2} S(f) \rangle = \Omega(NT_e - \widetilde{NT}) \end{cases}$$

are verified. Thus,  $N \equiv \langle f \rangle$ ,  $NU \equiv \langle vf \rangle$ , and  $m\langle \frac{(v-U)^2}{2} f \rangle + \frac{Z}{2} NT_e$  do not depend on  $t$ .

*Proof of Proposition 2.2.* The first relation comes from

$$\begin{aligned} \langle S(f) \rangle &= \frac{\Omega}{\Delta v} \left[ (v_{j_{\max}+1/2} - \tilde{U}) \tilde{f}_{j_{\max}+1/2} - (v_{1/2} - \tilde{U}) \tilde{f}_{1/2} \right] \\ &+ \frac{\Omega T_e}{m \Delta v^2} \left[ \sum_{j=2}^{j_{\max}} f_j - \sum_{j=1}^{j_{\max}-1} f_j - \sum_{j=2}^{j_{\max}} f_j + \sum_{j=1}^{j_{\max}-1} f_j \right] = 0. \end{aligned}$$

We also have

$$\begin{aligned} \langle vS(f) \rangle &= -\Omega \sum_{j=1}^{j_{\max}} (v_{j+1/2} - \tilde{U}) \tilde{f}_{j+1/2} \Delta v + \frac{\Omega T_e}{m} (f_1 - f_{j_{\max}}) \\ &= \Omega \delta u \sum_{j=1}^{j_{\max}} \tilde{f}_{j+1/2} \Delta v + \frac{\Omega T_e}{m} (f_1 - f_{j_{\max}}) = 0. \end{aligned}$$



For the variation of the ionic energy, we obtain

$$\begin{aligned}
 \Omega^{-1} m \left\langle \frac{v^2}{2} S(f) \right\rangle &= - \sum_{j=1}^{j_{\max}} m(v_{j+1/2} - \tilde{U}) v_{j+1/2} \tilde{f}_{j+1/2} \Delta v \\
 &\quad - T_e \left( \sum_{j=2}^{j_{\max}} f_j v_{j-1/2} - \sum_{j=1}^{j_{\max}-1} f_j v_{j+1/2} \right) \\
 &= - \sum_{j=1}^{j_{\max}} m(v_{j+1/2} - \tilde{U}) v_{j+1/2} \tilde{f}_{j+1/2} \Delta v \\
 &\quad - T_e \left( f_{j_{\max}} v_{j_{\max}-1/2} - f_1 v_{3/2} - \sum_{j=2}^{j_{\max}-1} f_j \Delta v \right) \\
 &= - \sum_{j=1}^{j_{\max}} m(v_{j+1/2} - \tilde{U})^2 \tilde{f}_{j+1/2} \Delta v + m \tilde{N} \tilde{U} \delta \tilde{u} \\
 &\quad - T_e (f_{j_{\max}} v_{j_{\max}+1/2} - f_1 v_{1/2} - N) \\
 &= (NT_e - \widetilde{NT}). \quad \square
 \end{aligned}$$

Now, we introduce the numerical entropy

$$H(f, T_e) = \langle f \log f \rangle - \frac{ZN}{2} \log T_e$$

and the Maxwellian

$$\mathbf{M}_{\tilde{U}, T_e}(v, t) = \frac{N}{\sqrt{2\pi T_e/m}} \exp \left[ -\frac{m(v - \tilde{U})^2}{2T_e} \right]$$

associated with  $(f, T_e)$ . Using the fact that

$$(2.13) \quad \langle v^2 S(f) \rangle = \langle (v - \tilde{U})^2 S(f) \rangle,$$

we can show that

$$(2.14) \quad \frac{\partial}{\partial t} H(f, T_e) = \langle S(f) \log f \rangle - \frac{\Omega(\widetilde{NT} - NT_e)}{T_e} = \left\langle S(f) \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}} \right) \right\rangle.$$

This is still true by defining  $\tilde{f}_{j+1/2}(t)$  in another way, but the entropic average allows us to have the following results.

LEMMA 2.3. *For all strictly positive  $f_j(t)$  and  $T_e(t)$ , we have*

$$\begin{aligned}
 (2.15) \quad S(f) &= \frac{\Omega T_e}{m \Delta v^2} \left\{ \tilde{f}_{j+1/2} \left[ \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}} \right)_{j+1} - \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}} \right)_j \right] \right. \\
 &\quad \left. - \tilde{f}_{j-1/2} \left[ \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}} \right)_j - \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}} \right)_{j-1} \right] \right\}
 \end{aligned}$$

and

$$(2.16) \quad \frac{\partial}{\partial t} H(f, T_e) = -\frac{\Omega T_e}{m \Delta v^2} \sum \tilde{f}_{j+1/2} \left[ \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}^{\infty}} \right)_{j+1} - \log \left( \frac{f}{\mathbf{M}_{\tilde{U}, T_e}^{\infty}} \right)_j \right]^2 \leq 0.$$

This lemma shows that there is equivalence between the convection-diffusion form (0.7) and the Landau form (1.1) of the discretized Fokker-Planck collision operator and that the semidiscretized scheme is entropic: these strong properties, due to the definition of the *entropic average*, are essential to obtain the convergence results in the next paragraph.

*Proof of Lemma 2.3.* The first point comes from

$$(2.17) \quad \log(\mathbf{M}_{\tilde{U}, T_e}^{\infty})_{j+1} - \log(\mathbf{M}_{\tilde{U}, T_e}^{\infty})_j = -\frac{m \Delta v}{T_e} (v_{j+1/2} - \tilde{U}).$$

The second one comes from the first point and from (2.14).  $\square$

And now we can show the following result.

**PROPOSITION 2.4.** *For all strictly positive initial conditions, the semidiscretized scheme defined by (2.7) and (2.8) has a global positive solution and*

$$\inf_{t \in [0, +\infty[} T_e(t) > 0.$$

The proof of Proposition 2.4 is in the appendix.

**COROLLARY 2.5.** *There exists a constant  $H^\infty$  such that*

$$\lim_{t \rightarrow +\infty} H(f, T_e) = H^\infty.$$

*Proof of Corollary 2.5.* A direct consequence of Proposition 2.4 is that  $H(f, T_e)$  is well defined on  $[0, +\infty[$ . By applying Lemma 2.3, we see that  $H(f, T_e)$  is a decreasing function. Since the scheme conserves the energy,  $T_e(t)$  is bounded from above, which implies that  $H(f, T_e)$  is bounded from below on  $[0, +\infty[$  since  $x \mapsto x \log x$  is bounded from below on  $[0, +\infty[$ . Therefore,  $H(f, T_e)$  has a limit  $H^\infty > -\infty$  when  $t$  goes to infinity.  $\square$

**2.2. Convergence of the semidiscretized scheme toward an unique equilibrium.** Now we define the numerical equilibrium state  $f_j^\infty$  which is a projection of a Maxwellian distribution on the discretized velocity grid  $\{v_j\}$ .

**DEFINITION 2.6.** *The numerical equilibrium state  $f_j^\infty$  is defined with*

$$(2.18) \quad f_j^\infty = N \cdot \frac{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty}(v_j)}{\langle \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty} \rangle},$$

where

$$(2.19) \quad \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty}(v) = \frac{N}{\sqrt{2\pi T_e^\infty/m}} \exp \left[ -\frac{m(v - \tilde{U}^\infty)^2}{2T_e^\infty} \right],$$

knowing that  $(\tilde{U}^\infty, T_e^\infty)$  is a solution of the nonlinear system

$$(2.20) \quad \begin{cases} \frac{N}{\langle \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty} \rangle} \cdot \langle v \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty} \rangle = \langle v f^0 \rangle, \\ \frac{N}{\langle \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty} \rangle} \cdot \langle (v - U^0)^2 \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}^{\infty} \rangle + \frac{ZN}{m} T_e^\infty = \langle (v - U^0)^2 f^0 \rangle + \frac{ZN}{m} T_e^0. \end{cases}$$

Let us notice that  $\langle \mathbf{M}_{U,T} \rangle = N \sum_j \exp[-\frac{m(v-U)^2}{2T}] \frac{\Delta v}{\sqrt{2\pi T/m}}$  is not exactly equal to  $N$  according to the discretization errors.

We can now state the following result.

PROPOSITION 2.7. *For all strictly positive initial conditions,*

(i) *the semidiscretized scheme defined by (2.7) and (2.8) ensures that there exists a subsequence  $(t_k)$  such that the functions  $f_j(t_k)$  converge to the equilibrium state  $f_j^\infty$  given by (2.18);*

(ii) *any solution  $(\tilde{U}^\infty, T_e^\infty)$  of (2.20) satisfies*

$$\forall t : H(f^\infty, T_e^\infty) \leq H(f, T_e);$$

(iii) *the system (2.20) admits a unique solution.*

The proof of Proposition 2.7 is in the appendix.

Now we write the main result of this section.

THEOREM 2.8. *For all strictly positive initial conditions, the semidiscretized scheme defined by (2.7) and (2.8) ensures that*

(i)

$$\lim_{t \rightarrow +\infty} f_j(t) = f_j^\infty,$$

(ii)

$$\lim_{t \rightarrow +\infty} T_e(t) = T_e^\infty,$$

where  $f_j^\infty$  and  $T_e^\infty$  are given by the unique equilibrium state defined with (2.18), (2.19), and (2.20).

Let us remark that

$$\tilde{U}^\infty \simeq U^0 = \frac{\langle v f^0 \rangle}{N}$$

and that

$$T_e^\infty \simeq \frac{T^0 + Z T_e^0}{1 + Z},$$

where  $N T^0 = \langle (v - U^0)^2 f^0 \rangle$ .

*Proof of Theorem 2.8.* We know that  $H(f, T_e)$  has a limit (see Corollary 2.5) and that  $H(f, T_e)$  is bounded from below by  $H(f^\infty, T_e^\infty)$  (see point (ii) of Proposition 2.7). Then we can write

$$\lim_{t \rightarrow +\infty} H(f, T_e) - H(f^\infty, T_e^\infty) = a \geq 0.$$

On the other hand, using point (i) of Proposition 2.7, we can say that there exists a sequence  $(t_k)$  such that

$$\lim_{t_k \rightarrow +\infty} f_j(t_k) = f_j^\infty.$$

Then, for the sequence  $(t_k)$ , we also have

$$\lim_{t_k \rightarrow +\infty} H(f, T_e)(t_k) - H(f^\infty, T_e^\infty) = 0.$$

Then,  $a = 0$ ; that is,

$$(2.21) \quad \lim_{t \rightarrow +\infty} H(f, T_e)(t) - H(f^\infty, T_e^\infty) = 0.$$

However, the Csiszar–Kullback inequality (cf. [21] and [22]) allows us to write that

$$\|f_j - f_j^\infty\|_{l_1}^2 \leq 2 \sum_j f_j \log \left( \frac{f_j}{f_j^\infty} \right) \Delta v$$

and

$$\left\| \mathbf{M}_{\tilde{U}^\infty, T_e} - \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \right\|_{L_1}^2 \leq 2 \int \mathbf{M}_{\tilde{U}^\infty, T_e} \log \left( \frac{\mathbf{M}_{\tilde{U}^\infty, T_e}}{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}} \right) dv.$$

Then, by applying Lemma A.1 (see the appendix) with the limit (2.21), we obtain

$$\lim_{t \rightarrow +\infty} \|f_j - f_j^\infty\|_{l_1} = 0$$

and

$$\lim_{t \rightarrow +\infty} \left\| \mathbf{M}_{\tilde{U}^\infty, T_e} - \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \right\|_{L_1} = 0,$$

this last limit showing that  $\lim_{t \rightarrow +\infty} T_e(t) = T_e^\infty$ .  $\square$

**3. Fully discretized scheme.** Let us denote by a superscript  $n$  the values of the various variables at the time  $t^n$  and let us define the time step  $\Delta t = t^{n+1} - t^n$ .

**3.1. Time explicit scheme.** We define the following explicit scheme:

$$(3.1) \quad \begin{cases} \frac{1}{\Delta t} (f_j^{n+1} - f_j^n) = S(f^n)_j, \\ \frac{1}{\Delta t} \left[ \frac{Z}{2} (NT_e)^{n+1} - \frac{Z}{2} (NT_e)^n \right] = \Omega^n (\widetilde{NT}^n - N^n T_e^n) \end{cases}$$

with

$$(3.2) \quad \begin{aligned} S(f^n)_j &= \frac{\Omega^n}{\Delta v} \left[ (v_{j+1/2} - \tilde{U}^n) \tilde{f}_{j+1/2}^n - (v_{j-1/2} - \tilde{U}^n) \tilde{f}_{j-1/2}^n \right] \\ &+ \frac{\Omega^n T_e^n}{m \Delta v^2} (a_j f_{j+1}^n - b_j f_j^n + c_j f_{j-1}^n). \end{aligned}$$

$\tilde{f}_{j+1/2}^n$  is the entropic average of  $f_j^n$  and  $f_{j+1}^n$  (see Definition 2.1);  $a_j$ ,  $b_j$ , and  $c_j$  are given by (2.11); and we set  $\tilde{f}_{1/2}^n = \tilde{f}_{j_{\max}+1/2}^n \equiv 0$ .  $\tilde{U}^n$  and  $\widetilde{NT}^n$  are defined by (2.9); and  $\Omega^n$  is evaluated with the values known at time  $t^n$ . We obtain the following conservation relations.

**PROPOSITION 3.1.** *The conservation laws*

$$\begin{cases} \langle f^{n+1} \rangle = \langle f^n \rangle, \\ \langle v f^{n+1} \rangle = \langle v f^n \rangle, \\ m \langle \frac{v^2}{2} f^{n+1} \rangle - m \langle \frac{v^2}{2} f^n \rangle = \Delta t \Omega^n (N^n T_e^n - \widetilde{NT}^n) \end{cases}$$

are verified. Thus,  $N^n \equiv \langle f^n \rangle \equiv N$ ,  $N^n U^n \equiv \langle v f^n \rangle \equiv NU$ , and  $m \langle \frac{(v-U^n)^2}{2} f^n \rangle + \frac{Z}{2} N^n T_e^n \equiv \frac{N}{2} (T^0 + Z T_e^0)$  do not depend on  $n$ .

That is, the scheme conserves the mass, the momentum, and the energy, which is consistent with (1.3). Let us note that the equation for the electronic energy can be written in the following way:

$$(3.3) \quad T_e^{n+1} = T_e^n \left( 1 - \Delta t \frac{2\Omega^n}{Z} \right) + 2\Delta t \Omega^n \frac{\widetilde{NT}^n}{ZN}.$$

**Positivity and conservation of the equilibrium state.** Now we show that the time explicit scheme defined by (3.1) and (3.2) is positive under a CFL criteria and verifies a discrete version of the H-theorem. We set

$$(3.4) \quad \Delta t_1^n = \frac{m}{4\Omega^n T_e^n} \cdot \frac{\Delta v^2}{\mathfrak{M}^n} \quad \text{and} \quad \Delta t_2^n = \frac{Z}{2\Omega^n}$$

with

$$\mathfrak{M}^n = \max_j \left[ \frac{\mathbf{M}_{\widetilde{U}^n, T_e^n, j \pm 1}}{\mathbf{M}_{\widetilde{U}^n, T_e^n, j}} \right]$$

(where  $j$  and  $j \pm 1 \in \{1, \dots, j_{\max}\}$ ). Let us recall that

$$\mathbf{M}_{\widetilde{U}^n, T_e^n, j} = \frac{N}{\sqrt{2\pi T_e^n/m}} \exp \left[ -\frac{m(v_j - \widetilde{U}^n)^2}{2T_e^n} \right],$$

that  $f^\infty$  is defined by (2.18), and that  $(\widetilde{U}^\infty, T_e^\infty)$  is the unique solution of the system (2.20).

We have the following result.

**PROPOSITION 3.2.** *For all strictly positive initial conditions, the explicit scheme defined by (3.1) and (3.2) preserves the positivity of  $f_j^{n+1}$  and  $T_e^{n+1}$  and verifies a discrete version of the H-theorem*

$$\langle S(f^n) \log(f^n / \mathbf{M}_{\widetilde{U}^n, T_e^n}) \rangle \leq 0$$

under the CFL criteria

$$(3.5) \quad \Delta t < \min(\Delta t_1^n, \Delta t_2^n).$$

Moreover, we have

$$f^n = f^\infty \quad \text{and} \quad T_e^n = T_e^\infty > 0 \quad \Longleftrightarrow \quad f^{n+1} = f^n \quad \text{and} \quad \forall j : f_j^n > 0, \quad T_e^n > 0.$$

The proof of Proposition 3.2 is in the appendix.

*The decay of the entropy.* Now we show that the scheme is entropic, that  $(T_e^n)$  is bounded from below, and that the time step  $\Delta t$  does not vanish in finite time under a very weak hypothesis. We set

$$\Delta t_3^n = \Delta t_1^n \cdot \frac{h_{\min}^n}{h_{\max}^n} \cdot \frac{1}{1 + \alpha^n} \quad \text{and} \quad \Delta t_4^n = \frac{1}{2} \Delta t_2^n,$$

$$\begin{cases} h_{\max}^n = \max_k \left( \frac{f^n}{\widetilde{M}_{U^n, T_e^n}} \right)_k, \\ h_{\min}^n = \min_k \left( \frac{f^n}{\widetilde{M}_{U^n, T_e^n}} \right)_k, \\ \alpha^n = \frac{1}{Z} \cdot \frac{\max_k (v_k - \widetilde{U}^n)^4}{(T_e^n/m)^2}, \end{cases}$$

and

$$H^n = \langle f^n \log f^n \rangle - \frac{ZN}{2} \log T_e^n.$$

**PROPOSITION 3.3.** *For all strictly positive initial conditions, the scheme defined by (3.1) and (3.2) verifies the inequality*

$$H(f^\infty, T_e^\infty) \leq H^{n+1} \leq H^n$$

if we have

$$(3.6) \quad \Delta t < \min(\Delta t_3^n, \Delta t_4^n).$$

The proof of Proposition 3.3 is in the appendix.

**COROLLARY 3.4.** *For all strictly positive initial conditions, and under the CFL condition (3.6), we have*

$$\inf_n T_e^n > 0,$$

and there exists a constant  $H^\infty$  such that

$$\lim_{n \rightarrow +\infty} H^n = H^\infty.$$

Since  $\inf_n T_e^n > 0$  (see Corollary 3.4), we also have  $\inf_n \Delta t_4^n > 0$ . However, we do not have  $\inf_n \Delta t_3^n > 0$  because this property is related to the property  $\inf_{j,n} (f_j^n) > 0$  which seems to be difficult to obtain. (This lower bound would prove that the explicit scheme converges to the equilibrium state  $f_j^\infty$  given by (2.18).) However, we have to check on the numerical experiments that the coefficient  $\Delta t_3^n$  does not vanish.

**3.2. Time semi-implicit scheme.** Despite its good behavior, the explicit scheme is expensive since the CFL condition is in  $\Delta v^2$ . This leads us to implicate the diffusive term. Then, we define the semi-implicit scheme as

$$(3.7) \quad \begin{cases} \frac{1}{\Delta t} (f_j^{n+1} - f_j^n) = S(f^n, f^{n+1})_j, \\ \frac{1}{\Delta t} \left[ \frac{1}{2} (N_e T_e)^{n+1} - \frac{1}{2} (N_e T_e)^n \right] = \Omega^n (\widetilde{NT}^{n+1/2} - NT_e^n) \end{cases}$$

with

$$(3.8) \quad \begin{aligned} S(f^n, f^{n+1})_j &= \frac{\Omega^n}{\Delta v} \left[ (v_{j+1/2} - \widetilde{U}^{n+1/2}) \widetilde{f}_{j+1/2}^n - (v_{j-1/2} - \widetilde{U}^{n+1/2}) \widetilde{f}_{j-1/2}^n \right] \\ &+ \frac{\Omega^n T_e^n}{m \Delta v^2} (a_j f_{j+1}^{n+1} - b_j f_j^{n+1} + c_j f_{j-1}^{n+1}). \end{aligned}$$

As in the explicit case,  $\tilde{f}_{j+1/2}^n$  is the entropic average of  $f_j^n$  and  $f_{j+1}^n$ ;  $a_j$ ,  $b_j$ , and  $c_j$  are given by (2.11) and we set  $\tilde{f}_{1/2}^n = \tilde{f}_{j_{\max}+1/2}^n \equiv 0$ . In (3.7) and (3.8),  $\tilde{U}^{n+1/2}$  and  $\widetilde{NT}^{n+1/2}$  (defined with (2.9)) are semi-implicit only through the corrective terms  $\delta\tilde{u}^{n+1/2}$  and  $\delta\tilde{t}^{n+1/2}$  defined with

$$\begin{cases} \delta\tilde{u}^{n+1/2} = \frac{T_e^n}{mN^n} (f_{j_{\max}}^{n+1} - f_1^{n+1}), \\ \delta\tilde{t}^{n+1/2} = \frac{T_e^n}{N^n} \left[ f_{j_{\max}}^{n+1} (v_{j_{\max}+1/2} - \tilde{U}^{n+1/2}) + f_1^{n+1} (\tilde{U}^{n+1/2} - v_{1/2}) \right]. \end{cases}$$

However, for and only for the numerical applications, we will neglect the corrective terms  $\delta\tilde{u}^{n+1/2}$  and  $\delta\tilde{t}^{n+1/2}$ , and then  $\tilde{U}^{n+1/2}$  and  $\widetilde{NT}^{n+1/2}$  will be completely explicit.

We obtain the following conservation laws.

PROPOSITION 3.5. *The conservation laws*

$$(3.9) \quad \begin{cases} \langle f^{n+1} \rangle = \langle f^n \rangle, \\ \langle v f^{n+1} \rangle = \langle v f^n \rangle, \\ m \langle \frac{v^2}{2} f^{n+1} \rangle - m \langle \frac{v^2}{2} f^n \rangle = \Delta t \Omega^n (N^n T_e^n - \widetilde{NT}^{n+1/2}) \end{cases}$$

are verified. Thus,  $N^n \equiv \langle f^n \rangle \equiv N$ ,  $N^n U^n \equiv \langle v f^n \rangle \equiv NU$ , and  $m \langle \frac{(v-U^n)^2}{2} f^n \rangle + \frac{Z}{2} N^n T_e^n \equiv \frac{N}{2} (T^0 + Z T_e^0)$  do not depend on  $n$ .

That is, the semi-implicit scheme is also totally conservative and consistent with (1.3), and the discrete temperature equation for the electrons has the same form as the explicit scheme (3.3); that is,

$$T_e^{n+1} = T_e^n \left( 1 - \Delta t \frac{2\Omega^n}{Z} \right) + 2\Delta t \Omega^n \frac{\widetilde{NT}^{n+1/2}}{N^n}.$$

*Conservation of the equilibrium state by the semi-implicit scheme.* For this scheme, we cannot prove an H-theorem since we cannot obtain a formulation of  $S(f^n, f^{n+1})_j$  equivalent to (A.9). However, we have the following result.

PROPOSITION 3.6. *The scheme defined by (3.7) and (3.8) preserves the equilibrium state; that is,*

$$f^n = f^\infty \quad \text{and} \quad T_e^n = T_e^\infty > 0 \quad \Longleftrightarrow \quad f^{n+1} = f^n \quad \text{and} \quad \forall j : f_j^n > 0, \quad T_e^n > 0.$$

We recall that  $f^\infty$  is defined by the relation (2.18).

*Proof of Proposition 3.6.* Let us assume that

$$f^n = C \cdot \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \quad \text{and} \quad T_e^n = T_e^\infty.$$

Then, as for the proof of the relation (2.15), we get

$$\begin{aligned} & \forall \Delta t \quad (f^{n+1} - C \mathbf{M}_{\tilde{U}^\infty, T_e^\infty})_j \\ &= \Delta t \frac{\Omega^n T_e^n}{m \Delta v^2} \left[ a_j (f^{n+1} - C \mathbf{M}_{\tilde{U}^\infty, T_e^\infty})_{j+1} - b_j (f^{n+1} - C \mathbf{M}_{\tilde{U}^\infty, T_e^\infty})_j \right. \\ & \quad \left. + c_j (f^{n+1} - C \mathbf{M}_{\tilde{U}^\infty, T_e^\infty})_{j-1} \right] \end{aligned}$$

with  $C = \frac{N}{\langle \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \rangle}$ . Since the diffusive matrix  $\mathcal{D}$  (see (2.11)) is diagonally dominant,  $\mathcal{I} + \Delta t \cdot \mathcal{D}$  is definite positive, which necessarily shows that

$$f^{n+1} = C \cdot \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} = f^n.$$

Now, by assuming that  $f^{n+1} = f^n$ , we can apply the equality (A.11) and we get

$$\begin{aligned} & \sum_j S(f^n, f^n)_j \log(f / \mathbf{M}_{\tilde{U}, T_e})_j^n \\ &= -\frac{\Omega^n T_e^n}{m \Delta v^2} \sum_j \tilde{f}_{j+1/2}^n \left[ \log(f / \mathbf{M}_{\tilde{U}, T_e})_{j+1}^n - \log(f / \mathbf{M}_{\tilde{U}, T_e})_j^n \right]^2 = 0, \end{aligned}$$

which shows that there exists  $C$  such that

$$f^n = C \cdot \mathbf{M}_{\tilde{U}^n, T_e^n}$$

since, by hypothesis,  $\tilde{f}_{j+1/2}^n > 0$  and  $\Omega^n T_e^n > 0$ . To conclude, we have to say only that the solution of the system (2.20) is unique (see point (iii) of Proposition 2.7) and that  $\mathbf{M}_{\tilde{U}^n, T_e^n}$  is a solution of the system (2.20) since the scheme is conservative.  $\square$

*On the positivity of the semi-implicit scheme.* We can remark that the semi-implicit scheme realizes a splitting between the convective and the diffusion parts of the operator. Then, by neglecting the corrective term  $\delta \tilde{u}^{n+1/2}$ , we can formulate it as

$$\frac{1}{\Delta t} (f_j^{n+1/2} - f_j^n) = \frac{\Omega^n}{\Delta v} \left[ (v_{j+1/2} - \tilde{U}^n) \tilde{f}_{j+1/2}^n - (v_{j-1/2} - \tilde{U}^n) \tilde{f}_{j-1/2}^n \right],$$

discretization of

$$(3.10) \quad \partial_t f = \Omega \partial_v [(v - U)f],$$

and

$$\frac{1}{\Delta t} (f_j^{n+1} - f_j^{n+1/2}) = \frac{\Omega^n T_e^n}{m \Delta v^2} (a_j f_{j+1}^{n+1} - b_j f_j^{n+1} + c_j f_{j-1}^{n+1}),$$

discretization of

$$(3.11) \quad \partial_t f = \frac{\Omega T_e}{m} \partial_{v^2} f.$$

It is clear that  $f_j^{n+1/2} > 0 \implies f_j^{n+1} > 0$  for any time step  $\Delta t$  since the inverse of a diagonal dominant matrix is a positive matrix. The difficulty arises for the convective part since we have to find a criteria  $\Delta t \leq \Delta t^n$  such that  $f_j^n > 0 \implies f_j^{n+1/2} > 0$ . Such a criteria with  $\Delta t^n$  not going toward 0 can be found only if  $\inf_{j,n}(f_j^n)$  is strictly positive. In our case, we can expect only that we would have  $\inf_{j,n}(f_j^n) > 0$  thanks to the smoothing effect of the diffusion operator, knowing that the solution of (3.10) is a classical Dirac measure when  $t$  goes to infinity; the numerical results seem to confirm this fact.



**4. Numerical results.** We test the numerical semi-implicit scheme. In the following test cases, the collision frequency  $\Omega$  is given by (see [2], [3], or [4])

$$(4.1) \quad \Omega = \frac{4}{3} \sqrt{2\pi} \frac{\sqrt{m_e} N Z^3 e^4 \log \Lambda}{m T_e^{3/2}},$$

where  $\log \Lambda$  is the Coulombian logarithm which is supposed to be a constant and equal to 10. We recall that the numerical entropy is defined by

$$H(t) = \langle f \log f \rangle - \frac{ZN}{2} \log T_e.$$

We introduce a  $CFL$  number equal to

$$CFL = \Delta t / \frac{m \Delta v^2}{4 \Omega^0 T_e^0}.$$

It is related to the discretization of the diffusion operator and then to the theoretical time step  $\Delta t_1^n$  given by (3.4). Moreover, we take

$$v_{j_{\max}} = -v_{1/2} = 5 \sqrt{\frac{T_e^0}{m}} \quad \text{with} \quad T_e^0 = 2 \text{ KeV}$$

and

$$\Delta v = \frac{1}{10} \cdot \sqrt{\frac{T_e^0}{m}};$$

that is,  $j_{\max} = 100$  (except for the test case 3 also tested with  $j_{\max} = 10$ ). Then, we can rewrite the variable  $CFL$  as (when  $j_{\max} = 100$ )

$$\Delta t = \frac{CFL}{400} \cdot \frac{1}{\Omega^0}.$$

For the three first test cases, the results of our explicit and semi-implicit schemes (i.e., by using the entropic average) are quasi-identical because the  $CFL$  number is close to 1. In the fourth test case, we show that the semi-implicit scheme allows us to use very large time steps.

*Test case 1.* We consider the following initial conditions:

$$\begin{cases} f^0 = \text{bi-Maxwellian centered and of temperature } T^0, \\ T^0 = 1 \text{ KeV}, \\ T_e^0 = 2 \text{ KeV}, \\ U^0 = 0 \end{cases}$$

with

$$\begin{cases} \rho/m = 10^{22} \text{ cm}^{-3}, \\ m = 2.5m_p \text{ (where } m_p \text{ is the proton mass)}, \\ Z = 2, \end{cases}$$

and we take  $CFL = 1.7$ . We can remark on Figures 1, 2, and 3 that the relaxation has a good behavior. We can also remark that when  $j_{\max} = 100$ , the entropic average and the arithmetic average give very close results.

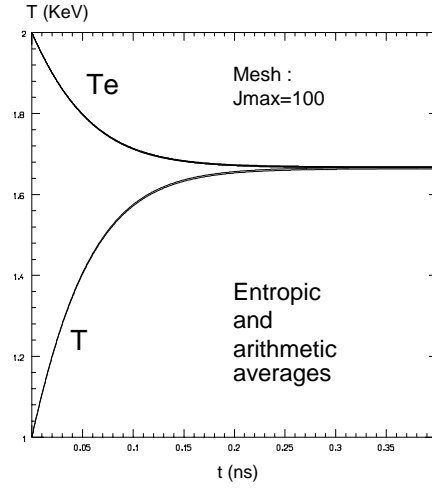


FIG. 1.

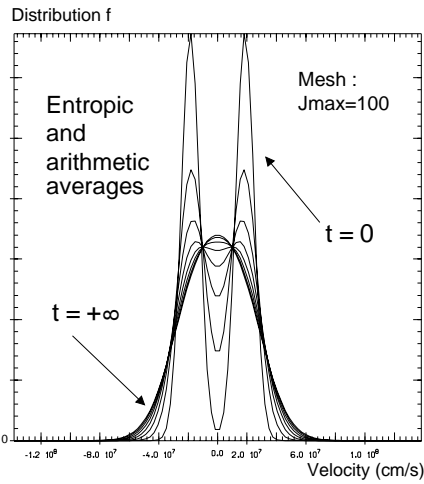


FIG. 2.

*Test case 2.* To study the conservation of the equilibrium state, we now consider a plasma whose initial conditions are

$$\begin{cases} f^0 = \text{centered Maxwellian with temperature } T^0, \\ T^0 = 1 \text{ KeV}, \\ T_e^0 = 1 \text{ KeV}, \\ U^0 = 0, \end{cases}$$

and we study the behavior of the scheme when  $\tilde{f}$  is defined by the entropic average, the classical Chang–Cooper average (see [8]), the arithmetic average, and the harmonic

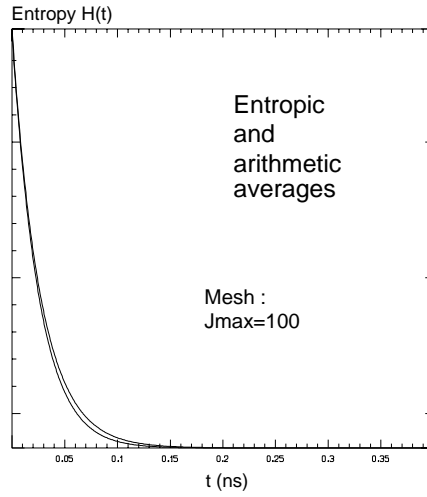


FIG. 3.

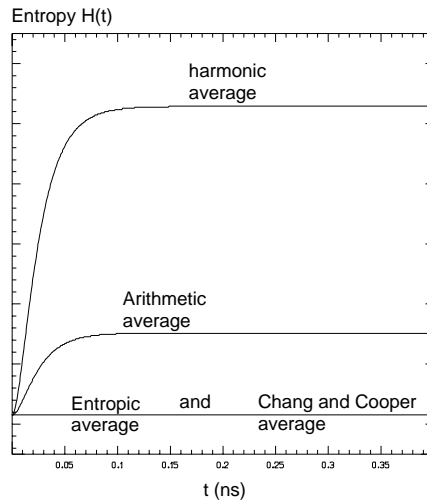


FIG. 4.

average (used in [17]) with  $j_{\max} = 100$ . Figure 4 shows that only the two first choices of  $\tilde{f}$  preserve the equilibrium state.

*Test case 3.* To study more finely the difference between all choices of  $\tilde{f}$ , we consider the following initial conditions:

$$\left\{ \begin{array}{l} f^0 = \text{centered Maxwellian with temperature } T^0, \\ T^0 = 1 \text{ KeV}, \\ T_e^0 = 2 \text{ KeV}, \\ U^0 = 0 \end{array} \right.$$

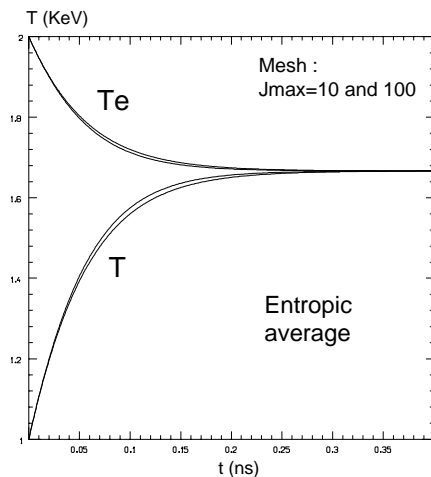


FIG. 5.

with

$$\begin{cases} \rho/m = 10^{22} \text{ cm}^{-3}, \\ m = 2.5m_p \text{ (where } m_p \text{ is the mass of a proton)}, \\ Z = 2, \end{cases}$$

and we take  $CFL = 1.7$ . We also remark on Figures 5–13 that these four averages give similar results on a fine grid, but only the entropic and the Chang–Cooper ones give good relaxation that trends to the right equilibrium state and positivity of the solution. For  $j_{\max} = 10$ , we also verify that the scheme is entropic only for the entropic (see Figure 11) and Chang–Cooper (see Figure 12) averages. Figures 5, 7, 11, and 12 show also that the entropic averages give results that are better than the ones given by the Chang–Cooper average on a small grid.

*Test case 4: Study of the time step.* We can see in Figure 14 (see test case 3 for the initial conditions) that the semi-implicit scheme gives a good relaxation of the temperature even when the  $CFL$  factor is large (we have taken  $CFL = 108$ ). However, we can also notice that for this test case, the value of the  $CFL$  factor to ensure the decay of the entropy is about of an order of 20 (with  $j_{\max} = 100$ ). For the explicit scheme running on the same test case, the maximum  $CFL$  factor is about 1.7, independently of the velocity step  $\Delta v$ .

All the results of these test cases show us that the explicit and semi-implicit schemes using the entropic average allow us to simulate the ion/electron collision operator well even on a grid with very large  $\Delta v$ . Moreover, the semi-implicit scheme accepts very large time steps.

**5. Conclusion.** To solve the ion/electron Fokker–Planck equation in Cartesian geometry, we have proposed a numerical scheme of finite difference type based on the *entropic average*, an average which is introduced in this paper (see Definition 2.1). The semi-implicit scheme has been used in [19, Part 2, Chapter 4] for solving the full kinetic system (0.5) and (0.6). This scheme may obviously be extended to the dimension 2 or 3, and we have shown that it is a good alternative to the Chang–Cooper

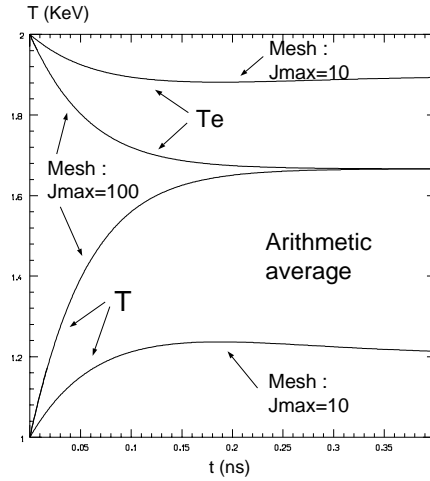


FIG. 6.

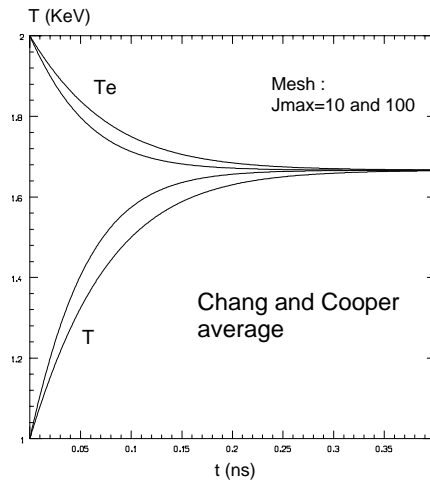


FIG. 7.

average: indeed, it always gives positive and precise solutions even on very coarse grids (see also [19, Part 2, Chapter 4]) for hard numerical applications coming from the inertial confinement fusion field) and it is simpler. Moreover, it is possible to recover the results of this paper in axisymmetrical or spherical geometries (see [19, Part 1, Chapter 4]), and the entropic average may also be used for the numerical treatment of other kinetic operators of Fokker-Planck type (see [15] and [20])—for instance, for the classical quadratic Fokker-Planck operator  $B(f)$  in spherical geometry which is

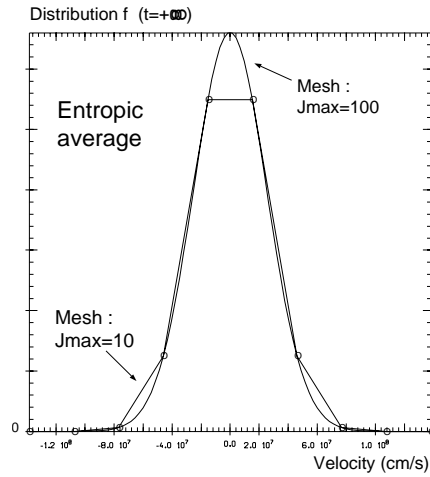


FIG. 8.

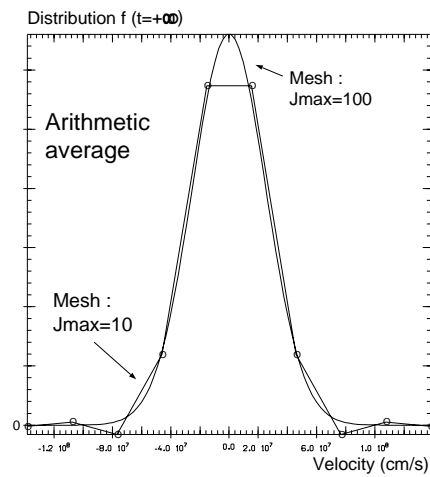


FIG. 9.

defined by (see [11])

$$B(f)(w) = w^{-1/2} \frac{\partial}{\partial w} \int_{\mathbf{R}^+} \left( f(w') \frac{\partial f}{\partial w}(w) - f(w) \frac{\partial f}{\partial w'}(w') \right) \min(w^{3/2}, w'^{3/2}) dw'.$$

Here,  $w$  is the square of the modulus of the velocity. In this framework, the *entropic average* defines a scheme which is closely related to one of the schemes proposed in [11] (see [15]).

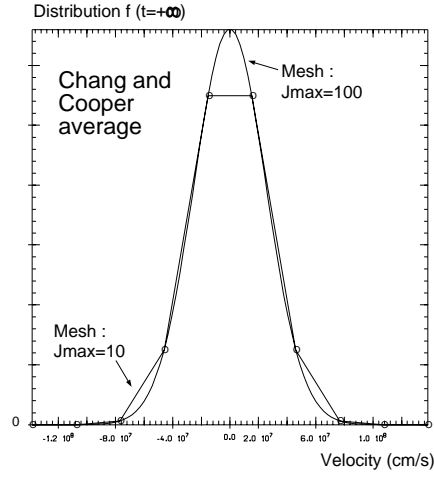


FIG. 10.

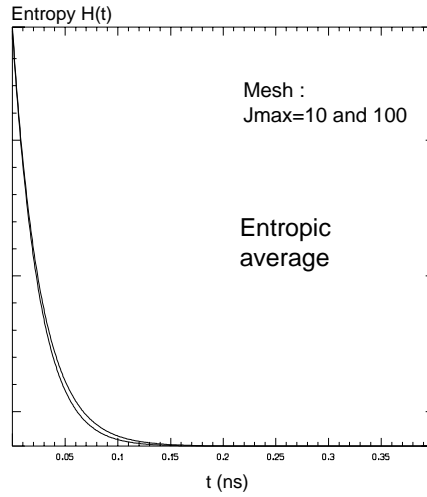


FIG. 11.

### Appendix. Proof of Propositions 2.4, 2.7, 3.2, and 3.3.

**A.1. Proof of Proposition 2.4.** Let us remark that by applying the Cauchy–Peano theorem (or the Cauchy–Lipschitz theorem), we can say that there exists a maximum interval  $[0, T[$  on which the semidiscretized scheme (2.7) and (2.8) admits a strictly positive solution. (We recall that the initial conditions are strictly positive.) Let us note that  $H(f, T_e)(t)$  is defined on  $[0, T[$  and that, due to Lemma 2.3, we have  $H(f, T_e)(t) \leq H(f, T_e)(0) < +\infty$  on  $[0, T[$ ; then, since  $x \mapsto x \log x$  is bounded from below on  $[0, +\infty[$ , we have  $\inf_{t \in [0, T[} T_e(t) > 0$  (otherwise,  $\lim_{t \rightarrow T} H(f, T_e)(t) = +\infty$ ) and, consequently,  $\inf_{t \in [0, T[} \Omega(t) > 0$ .

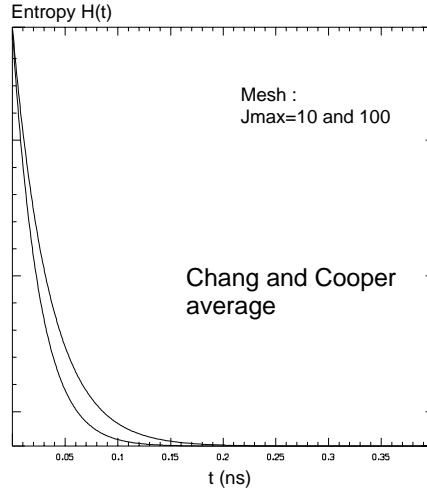


FIG. 12.

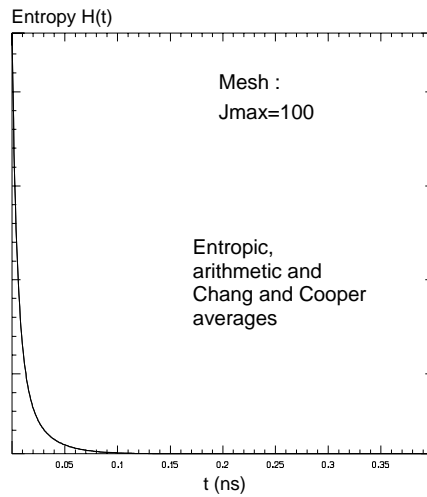


FIG. 13.

If  $T < +\infty$ , necessarily,  $\exists j_0 / \lim_{t \rightarrow T} f_{j_0}(t) = 0$  which implies that  $\lim_{t \rightarrow T} \tilde{f}_{j_0 \pm 1/2}(t) = 0$ . When  $\lim_{t \rightarrow T} \tilde{N}(t) > 0$ , we have  $\sup_{t \in [0, T[} |\tilde{U}(t)| < +\infty$ ; and when  $\lim_{t \rightarrow T} \tilde{N}(t) = 0$  (i.e.,  $\forall j : \lim_{t \rightarrow T} \tilde{f}_{j+1/2}(t) = 0$  which does not mean that  $\forall j : \lim_{t \rightarrow T} f_j(t) = 0$ ), by continuity, we obtain  $\lim_{t \rightarrow T} \tilde{U}(t) = 0$ . Then, in each case, when  $T < +\infty$ , the convective part of (2.8) goes to zero when  $t$  goes to  $T$ , which shows that

$$\lim_{t \rightarrow T} \partial_t f_{j_0}(t) = \lim_{t \rightarrow T} \frac{\Omega(t) T_e(t)}{m \Delta v^2} [f_{j_0+1}(t) + f_{j_0-1}(t)] \geq 0.$$



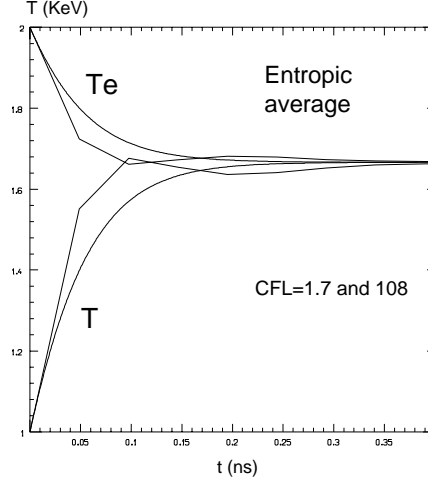


FIG. 14.

On the other hand, since  $f_{j_0}(t)$  is strictly positive for  $t < T$ , necessarily, we must have

$$\lim_{t \rightarrow T} \partial_t f_{j_0}(t) \leq 0.$$

Thus,  $\lim_{t \rightarrow T} \partial_t f_{j_0}(t) = 0$ , and then  $\lim_{t \rightarrow T} f_{j_0+1}(t) = \lim_{t \rightarrow T} f_{j_0-1}(t) = 0$  since  $\inf_{t \in [0, T[} T_e(t) > 0$  and  $\inf_{t \in [0, T[} \Omega(t) > 0$ . By continuation, we obtain that

$$T < +\infty \implies \forall j : \lim_{t \rightarrow T} f_j(t) = 0,$$

which is impossible since we know that for all  $t \in [0, T[ : \sum_j f_j(t) \Delta v = N(t) = N(0) > 0$  (see Proposition 2.2).

Then, the only possibility is that the maximal interval is equal to  $[0, +\infty[$ . And, finally, we can write that  $\inf_{t \in [0, +\infty[} T_e(t) > 0$ .

**A.2. Proof of Proposition 2.7.** Let us recall the notation

$$\mathbf{M}_{U,T}(v, t) = \frac{N}{\sqrt{2\pi T/m}} \exp \left[ -\frac{m(v - U)^2}{2T} \right].$$

**A.2.1. Preliminary results: Lemmas A.1, A.2, and A.3.** The proof of Proposition 2.7 uses the following lemma.

LEMMA A.1. *The relation*

$$H(f, T_e) - H(f^\infty, T_e^\infty) = \sum_j f_j \log \left( \frac{f_j}{f_j^\infty} \right) \Delta v + Z \int \mathbf{M}_{\tilde{U}^\infty, T_e} \log \left( \frac{\mathbf{M}_{\tilde{U}^\infty, T_e}}{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}} \right) dv$$

is verified.

LEMMA A.2. *For all sequence  $(t_k)$  such that*

$$(A.1) \quad \lim_{t_k \rightarrow +\infty} \max_j \left( \tilde{f}_{j+1/2}(t_k) \left[ \log(f/\mathbf{M}_{\tilde{U}, T_e})_{j+1}(t_k) - \log(f/\mathbf{M}_{\tilde{U}, T_e})_j(t_k) \right]^2 \right) = 0,$$

it exists a subsequence, still denoted  $(t_k)$ , such that

(i)

$$\inf_k \left| \tilde{N}(t_k) \right| > 0, \quad \sup_k \left| \tilde{U}(t_k) \right| < +\infty \quad \text{and} \quad \sup_k |\delta \tilde{u}(t_k)| < +\infty,$$

(ii)

$$(A.2) \quad \inf_{k,j} f_j(t_k) > 0.$$

The proof of Lemma A.2 is based on the following lemma.

LEMMA A.3. *Let us define two sequences  $x_k$  and  $y_k$  of positives reals,  $\tilde{f}_k$  their entropic average, and  $z_k$  a real sequence. And let us suppose that*

$$\forall k : \tilde{f}_k [\log x_k - \log y_k - z_k]^2 \leq C.$$

Then, the two following properties are verified:

(i) *If for one  $C'$  positive we have for all  $k$*

$$z_k > -C',$$

then

$$x_k \rightarrow 0 \implies y_k \rightarrow 0.$$

(ii) *If  $z_k$  is bounded, then*

$$\tilde{f}_k \rightarrow 0 \implies x_k \rightarrow 0 \text{ and } y_k \rightarrow 0.$$

*Proof of Lemma A.1.* We can write

$$\begin{aligned} H(f, T_e) - H(f^\infty, T_e^\infty) &= \sum_j f_j \log f_j \Delta v + Z \int \mathbf{M}_{\tilde{U}^\infty, T_e} \log \mathbf{M}_{\tilde{U}^\infty, T_e} dv \\ &\quad - \sum_j f_j^\infty \log f_j^\infty \Delta v - Z \int \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \log \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} dv \\ &= \sum_j f_j \log \left( \frac{f_j}{f_j^\infty} \right) \Delta v + Z \int \mathbf{M}_{\tilde{U}^\infty, T_e} \log \left( \frac{\mathbf{M}_{\tilde{U}^\infty, T_e}}{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}} \right) dv \\ &\quad + \left[ \sum_j (f_j - f_j^\infty) \log f_j^\infty \Delta v + Z \int \left( \mathbf{M}_{\tilde{U}^\infty, T_e} - \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \right) \log \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} dv \right]. \end{aligned}$$

Since this scheme conserves the mass, we have

$$\sum_j (f_j - f_j^\infty) \log f_j^\infty \Delta v = \frac{1}{T_e^\infty} \sum_j \frac{m}{2} \left( v_j - \tilde{U}^\infty \right)^2 (f_j^\infty - f_j) \Delta v.$$

Then

$$\sum_j (f_j - f_j^\infty) \log f_j^\infty \Delta v = \frac{1}{T_e^\infty} \sum_j \frac{m}{2} (v_j - U^0)^2 (f_j^\infty - f_j) \Delta v$$

$$+ \frac{m(U^0 - \tilde{U}^\infty)}{2T_e^\infty} \left[ (U^0 - \tilde{U}^\infty) \sum_j (f_j^\infty - f_j) \Delta v + 2 \sum_j (v_j - U^0) (f_j^\infty - f_j) \Delta v \right].$$

However, we have  $\sum_j (f_j^\infty - f_j) \Delta v = 0$  and  $\sum_j (v_j - U^0) (f_j^\infty - f_j) \Delta v = 0$  since the scheme conserves the mass and the momentum, and since  $\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}$  is a solution of the system (2.20). Then we can write

$$\sum_j (f_j - f_j^\infty) \log f_j^\infty \Delta v = \frac{1}{T_e^\infty} \sum_j \frac{m}{2} (v_j - U^0)^2 (f_j^\infty - f_j) \Delta v.$$

And since

$$\int (\mathbf{M}_{\tilde{U}^\infty, T_e} - \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}) \log \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} dv = -\frac{N}{2T_e^\infty} (T_e - T_e^\infty),$$

in conclusion we have

$$H(f, T_e) - H(f^\infty, T_e^\infty) = \sum_j f_j \log \left( \frac{f_j}{f_j^\infty} \right) \Delta v + Z \int \mathbf{M}_{\tilde{U}^\infty, T_e} \log \left( \frac{\mathbf{M}_{\tilde{U}^\infty, T_e}}{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}} \right) dv$$

$$+ \frac{1}{T_e^\infty} \left[ \sum_j \frac{m}{2} (v_j - U^0)^2 (f_j^\infty - f_j) \Delta v + \frac{ZN}{2} (T_e^\infty - T_e) \right].$$

However, the last term of the right-hand side is equal to zero since the scheme conserves the energy and since  $\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}$  is a solution of the system (2.20). Then, we obtain the result.  $\square$

*Proof of Lemma A.2.* (i) If there exists a subsequence of  $(t_k)$  such that

$$(A.3) \quad \lim_{t_k \rightarrow +\infty} \tilde{N}(t_k) = 0,$$

then, for any  $j$ , we get

$$\lim_{t_k \rightarrow +\infty} \tilde{f}_{j+1/2}(t_k) = 0.$$

Since  $N(t_k) = N(0)$ , there is at least one  $j$  such that (up to an extraction)

$$(A.4) \quad \inf_k f_j(t_k) > 0.$$

Moreover, up to an extraction,  $\tilde{U}(t_k)$  is bounded from below or bounded from above; let us suppose that  $\tilde{U}(t_k)$  is bounded from below. (Proof is similar if we suppose that  $\tilde{U}(t_k)$  is bounded from above.) Since  $\tilde{f}_{j+1/2}(t_k)$  converges to 0 and since  $\inf_k f_j(t_k) > 0$ , we deduce that  $f_{j+1}(t_k)$  converges to 0. The relation (A.1) can be also written as

$$\lim_{t_k \rightarrow +\infty} \tilde{f}_{j+1/2}(t_k) [\log f_{j+1}(t_k) - \log f_j(t_k) - z_k]^2 = 0 \quad \text{with}$$

$$(A.5) \quad z_k = -\frac{m\Delta v}{T_e(t)} (v_{j+1/2} - \tilde{U}(t_k)).$$

And since  $\inf_k T_e(t_k) > 0$  (see Proposition 2.4), we can see that there exists  $C' > 0$  such that  $z_k > -C'$ ; the hypothesis of Lemma A.3(i) is verified with  $x_k = f_{j+1}(t_k)$  and  $y_k = f_j(t_k)$ . Then,  $\lim_{t_k \rightarrow +\infty} f_j(t_k) = 0$ , which is in contradiction with (A.4). Thus, (A.3) is false and we have

$$\inf_k \left| \tilde{N}(t_k) \right| > 0.$$

Knowing that

$$v_{1/2} + \frac{T_e(t)}{m} \cdot \frac{(f_{j_{\max}}(t) - f_1(t))}{\tilde{N}(t)} \leq \tilde{U}(t) \leq v_{j_{\max}+1/2} + \frac{T_e(t)}{m} \cdot \frac{(f_{j_{\max}}(t) - f_1(t))}{\tilde{N}(t)},$$

we also obtain

$$\sup_k \left| \tilde{U}(t_k) \right| < +\infty.$$

( $\Delta v$  is fixed and  $N(t) < +\infty$ ; then, for all  $j, t \geq 0 : f_j(t) < +\infty$ .) Using the definition of  $\delta \tilde{u}(t)$ , we also have

$$\sup_k |\delta \tilde{u}(t_k)| < +\infty.$$

(ii) If (A.2) is false, there exists a  $j_0$  such that, up to an extraction,  $\lim_{t_k \rightarrow +\infty} f_{j_0}(t_k) = 0$ . Then, we have  $\lim_{t_k \rightarrow +\infty} \tilde{f}_{j_0 \pm 1/2}(t_k) = 0$ . Since the relation (A.5) is true with  $z_k$  bounded, point (ii) of Lemma A.3 claims that also

$$\lim_{t_k \rightarrow +\infty} f_{j_0-1}(t_k) = 0 \quad \text{and} \quad \lim_{t_k \rightarrow +\infty} f_{j_0+1}(t_k) = 0,$$

which gives

$$\lim_{t_k \rightarrow +\infty} \tilde{f}_{j_0-3/2}(t_k) = 0 \quad \text{and} \quad \lim_{t_k \rightarrow +\infty} \tilde{f}_{j_0+3/2}(t_k) = 0.$$

By continuation, we deduce that for any  $j$ ,  $\lim_{t_k \rightarrow +\infty} f_j(t_k) = 0$  and  $\lim_{t_k \rightarrow +\infty} N(t_k) = 0$  which is in contradiction with the conservation of the mass.  $\square$

*Proof of Lemma A.3.* (i) We suppose that there exists a subsequence of  $y_k$ , also called  $y_k$ , which is bounded below by a constant  $\alpha > 0$ . Then, there is a contradiction since

$$\begin{aligned} \tilde{f}_k [\log x_k - \log y_k - z_k]^2 &= (x_k - y_k) [\log x_k - \log y_k] + \tilde{f}_k z_k^2 - 2z_k(x_k - y_k) \\ &\geq -y_k [\log x_k - \log y_k] + 2z_k(y_k - x_k) + o(1) \rightarrow +\infty. \end{aligned}$$

(ii) This point is a consequence of the first point since  $x_k$  and  $y_k$  play symmetrical roles.  $\square$

**A.2.2. Proof of point (i) of Proposition 2.7.** Since  $H(f, T_e)$  is continuous, decreasing, and bounded from below (see Lemma 2.3 and Corollary 2.5), and since  $\inf_{t \in [0, +\infty[} T_e(t) > 0$  (see Proposition 2.4), the relation (2.16) shows that there exists a sequence  $(t_k)$  going to  $+\infty$  and such that

$$(A.6) \quad \lim_{t_k \rightarrow +\infty} \max_j \left( \tilde{f}_{j+1/2}(t_k) \left[ \log(f/\mathbf{M}_{\tilde{U}, T_e})_{j+1}(t_k) - \log(f/\mathbf{M}_{\tilde{U}, T_e})_j(t_k) \right]^2 \right) = 0.$$

By using point (ii) of Lemma A.2, we know that up to an extraction of a subsequence

$$\inf_{k,j} \tilde{f}_{j+1/2}(t_k) > 0$$

(except for  $j = 0$  and  $j = j_{\max}$  because of the boundary limits (2.12)). And using (A.6), we obtain

$$\lim_{t_k \rightarrow +\infty} \max_j \left[ \log(f/\mathbf{M}_{\tilde{U}, T_e})_{j+1}(t_k) - \log(f/\mathbf{M}_{\tilde{U}, T_e})_j(t_k) \right] = 0,$$

which implies that there exists  $C > 0$  such that

$$\lim_{t_k \rightarrow +\infty} \frac{f_j(t_k)}{\mathbf{M}_{\tilde{U}, T_e, j}(t_k)} = C.$$

By applying point (i) of Lemma A.2, we see that  $\tilde{U}(t_k)$  is bounded. Moreover,  $T_e(t)$  is also bounded since the scheme conserves the energy. Then, we can state that there exists  $T_e^\infty > 0$  (since  $T_e(t)$  is also bounded from below; see Proposition 2.4),  $\tilde{U}^\infty$ , and a subsequence of  $(t_k)$  still noted  $(t_k)$  such that

$$\lim_{t_k \rightarrow +\infty} T_e(t_k) = T_e^\infty > 0 \quad \text{and} \quad \lim_{t_k \rightarrow +\infty} \tilde{U}(t_k) = \tilde{U}^\infty,$$

which gives

$$\lim_{t_k \rightarrow +\infty} \mathbf{M}_{\tilde{U}, T_e, j}(t_k) = \mathbf{M}_{\tilde{U}^\infty, T_e^\infty, j}$$

and then

$$\lim_{t_k \rightarrow +\infty} f_j(t_k) = C \cdot \mathbf{M}_{\tilde{U}^\infty, T_e^\infty, j}.$$

Since the scheme conserves the mass and the energy,  $(\tilde{U}^\infty, T_e^\infty)$  is a solution of (2.20) and we have

$$C = N / \langle \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \rangle.$$

**A.2.3. Proof of point (ii) of Proposition 2.7.** Let  $(\tilde{U}^\infty, T_e^\infty)$  be a solution of the system (2.20) and let us recall Jensen's inequality (see [21] and [22]) which says that

$$\int g[w(v)] d\mu(v) \geq \int \mu(v) dv \cdot g \left[ \frac{\int w(v) d\mu(v)}{\int \mu(v) dv} \right]$$

for each convex function  $g(w)$  and finite positive measure  $d\mu(v)$ . Then, by applying this inequality with the function  $g(w) = w \log w$  and the measure  $\sum_j f_j^\infty \delta(v_j)$  ( $\delta$  being the classical Dirac measure) or the measure  $\mathbf{M}_{\tilde{U}^\infty, T_e^\infty} dv$ , we get

$$\sum_j f_j \log \left( \frac{f_j}{f_j^\infty} \right) \Delta v > 0 \quad \text{and} \quad \int \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \log \left( \frac{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}}{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}} \right) dv > 0$$

since

$$\frac{\sum_j f_j \Delta v}{\sum_j f_j^\infty \Delta v} = 1 \quad \text{and} \quad \frac{\int \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}(v) dv}{\int \mathbf{M}_{\tilde{U}^\infty, T_e^\infty}(v) dv} = 1.$$

To conclude, we have to use only Lemma A.1.

**A.2.4. Proof of point (iii) of Proposition 2.7.** To obtain the unicity, we now apply the Csiszar–Kullback inequality (cf. [21] and [22]), which gives a better result for the convergence toward the equilibrium state

$$\|f_j - f_j^\infty\|_{l_1}^2 \leq 2 \sum_j f_j \log \left( \frac{f_j}{f_j^\infty} \right) \Delta v$$

and

$$\left\| \mathbf{M}_{\tilde{U}^\infty, T_e} - \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \right\|_{L_1}^2 \leq 2 \int \mathbf{M}_{\tilde{U}^\infty, T_e} \log \left( \frac{\mathbf{M}_{\tilde{U}^\infty, T_e}}{\mathbf{M}_{\tilde{U}^\infty, T_e^\infty}} \right) dv.$$

Then

$$\|f_j - f_j^\infty\|_{l_1}^2 + Z \left\| \mathbf{M}_{\tilde{U}^\infty, T_e} - \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \right\|_{L_1}^2 \leq 2[H(f, T_e) - H(f^\infty, T_e^\infty)]$$

by using Lemma A.1. Thus, if the system (2.20) has two solutions  $\mathbf{M}_{\tilde{U}^\infty, 1, T_e^{\infty, 1}}$  and  $\mathbf{M}_{\tilde{U}^\infty, 2, T_e^{\infty, 2}}$  with

$$\mathbf{M}_{\tilde{U}^\infty, k, T_e^{\infty, l}}(v, t) = \frac{N}{\sqrt{2\pi T_e^{\infty, l}/m}} \exp \left[ -\frac{m(v - \tilde{U}^{\infty, k})^2}{2T_e^{\infty, l}} \right]$$

( $k, l \in \{1, 2\}$ ), we have

$$H(f^{\infty, 2}, T_e^{\infty, 2}) - H(f^{\infty, 1}, T_e^{\infty, 1}) \geq 0 \quad \text{and} \quad H(f^{\infty, 1}, T_e^{\infty, 1}) - H(f^{\infty, 2}, T_e^{\infty, 2}) \geq 0,$$

which gives

$$H(f^{\infty, 2}, T_e^{\infty, 2}) - H(f^{\infty, 1}, T_e^{\infty, 1}) = 0.$$

Then, we have

$$(A.7) \quad \left\| f_j^{\infty, 2} - f_j^{\infty, 1} \right\|_{l_1}^2 = 0 \quad \text{and} \quad \left\| \mathbf{M}_{\tilde{U}^\infty, k, T_e^{\infty, l}} - \mathbf{M}_{\tilde{U}^\infty, k, T_e^{\infty, k}} \right\|_{L_1}^2 = 0.$$

And the first part of (A.7) gives

$$(A.8) \quad \forall j : f_j^{\infty, 2} = f_j^{\infty, 1}.$$

The equality (A.8) shows that  $\tilde{U}^{\infty, 2} = \tilde{U}^{\infty, 1}$  which implies that  $\mathbf{M}_{\tilde{U}^\infty, k, T_e^{\infty, l}} = \mathbf{M}_{\tilde{U}^\infty, l, T_e^{\infty, l}}$ . Then, using the second part of (A.7), we obtain  $\mathbf{M}_{\tilde{U}^\infty, 1, T_e^{\infty, 1}} = \mathbf{M}_{\tilde{U}^\infty, 2, T_e^{\infty, 2}}$  and then  $T_e^{\infty, 2} = T_e^{\infty, 1}$  which gives us the unicity of the solution of the system (2.20).

**A.3. Proof of Proposition 3.2.** • As for the semidiscretized problem, we show that

$$(A.9) \quad S(f^n)_j = \frac{\Omega^n T_e^n}{m \Delta v^2} (k_{j+1/2} + k_{j-1/2})$$

with

$$k_{j+1/2} = \tilde{f}_{j+1/2}^n \left[ \log(f_{j+1}^n / f_j^n) - \log(\mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j}) \right].$$

If  $k_{j+1/2}$  and  $k_{j-1/2}$  are positive,  $f_j^{n+1}$  will be positive. We restrict the study for the more restrictive situation which corresponds to

$$k_{j+1/2} < 0 \quad \text{and} \quad k_{j-1/2} < 0,$$

which is equivalent to

$$f_{j+1}^n/f_j^n < \mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j} \quad \text{and} \quad f_{j-1}^n/f_j^n < \mathbf{M}_{\tilde{U}^n, T_e^n, j-1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j}.$$

We now suppose that  $f_{j+1}^n \neq f_j^n$ . Then, we can write

$$k_{j+1/2} = f_j^n \left[ (f_{j+1}^n/f_j^n - 1) - \frac{(f_{j+1}^n/f_j^n - 1)}{\log(f_{j+1}^n/f_j^n)} \cdot \log(\mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j}) \right].$$

Since  $x \mapsto \frac{x-1}{\log x}$  is continuous, positive, and increasing on  $\mathbf{R}^+$  and since  $f_{j+1}^n/f_j^n < \mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j}$ , we get

$$\begin{aligned} |k_{j+1/2}| &\leq f_j^n \left[ |f_{j+1}^n/f_j^n - 1| + \left| \mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j} - 1 \right| \right] \\ &\leq f_j^n \left[ \max(1, f_{j+1}^n/f_j^n) + \max\left(1, \mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j}\right) \right] \\ &\leq 2f_j^n \max\left(1, \mathbf{M}_{\tilde{U}^n, T_e^n, j+1} / \mathbf{M}_{\tilde{U}^n, T_e^n, j}\right) \\ &\leq 2f_j^n \max_i \left( \mathbf{M}_{\tilde{U}^n, T_e^n, i\pm 1} / \mathbf{M}_{\tilde{U}^n, T_e^n, i} \right). \end{aligned}$$

This is still true when  $f_{j+1}^n = f_j^n$ . Now, according to (A.9), we obtain that

$$(A.10) \quad f_j^{n+1} > f_j^n \left[ 1 - \frac{4\Omega^n T_e^n \Delta t}{m\Delta v^2} \max_i \left( \mathbf{M}_{\tilde{U}^n, T_e^n, i\pm 1} / \mathbf{M}_{\tilde{U}^n, T_e^n, i} \right) \right].$$

This is still true if  $k_{j+1/2}$  or  $k_{j-1/2}$  is positive. Therefore

$$\Delta t < \Delta t_1^n \implies \forall j, f_j^{n+1} > 0.$$

• The positivity of  $T_e^{n+1}$  when  $\Delta t < \Delta t_2^n$  is clear if we use the expression (3.3) and the definition of  $\widetilde{NT}^n$ .

• To prove the H-theorem, we use the relation (A.9) which allows us to write

$$(A.11) \quad \sum_j S(f^n)_j \log(f/\mathbf{M}_{\tilde{U}, T_e})_j^n = -\frac{\Omega^n T_e^n}{m\Delta v^2} \sum_j \tilde{f}_{j+1/2}^n [\log(f/\mathbf{M}_{\tilde{U}, T_e})_{j+1}^n - \log(f/\mathbf{M}_{\tilde{U}, T_e})_j^n]^2 \leq 0.$$

• Let us now show the conservation of the equilibrium state. We suppose that

$$f^n = \frac{N}{\langle \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \rangle} \mathbf{M}_{\tilde{U}^\infty, T_e^\infty} \quad \text{and} \quad T_e^n = T_e^\infty.$$

By construction, we have (see the system (2.20))

$$\langle f^n \rangle = N^0, \quad \langle (v - U^0) f^n \rangle = 0$$

and

$$\left\langle \left\{ \left[ \frac{1}{2} m(v - U^0)^2 + Z \frac{T_e^n}{2} \right] - \left[ \frac{T^0}{2} + Z \frac{T_e^0}{2} \right] \right\} \cdot f^n \right\rangle = 0$$

since  $T_e^n = T_e^\infty$ . Using the boundary conditions  $\tilde{f}_{1/2}^n = \tilde{f}_{j_{\max}+1/2}^n \equiv 0$ , we can also verify that  $\tilde{U}^n = \tilde{U}^\infty$ . Then,

$$\exists C > 0 / f^n = C \cdot \mathbf{M}_{\tilde{U}^n, T_e^n}.$$

And, according to (A.9), we get for all  $j : S(f^n)_j = 0$ ; that is to say  $f^{n+1} = f^n$ . For the converse, if for all  $j : S(f^n)_j = 0$ , then, according to (A.11), we get

$$\forall j : \log(f / \mathbf{M}_{\tilde{U}, T_e}^n)_{j+1}^n = \log(f / \mathbf{M}_{\tilde{U}, T_e}^n)_j^n$$

since  $\tilde{f}_{j+1/2}^n > 0$  (for  $j \in \{2, \dots, j_{\max} - 1\}$ ) and  $\Omega^n T_e^n > 0$  by hypothesis. This shows that

$$\exists C > 0 / f^n = C \cdot \mathbf{M}_{\tilde{U}^n, T_e^n}.$$

To conclude, we have to say only that the solution of the system (2.20) is unique (see point (iii) of Proposition 2.7) and that  $\mathbf{M}_{\tilde{U}^n, T_e^n}$  is a solution of the system (2.20) since the scheme is conservative.

#### A.4. Proof of Proposition 3.3.

**A.4.1. Preliminary result: Lemma A.4.** The proof of Proposition 3.3 uses the following lemma.

LEMMA A.4. *When  $\tilde{f}_{j+1/2}^n$  is the entropic average of  $f_j^n$  and of  $f_{j+1}^n$ , the inequality*

$$\sum_j \frac{S(f^n)_j^2}{f_j^n} \leq -\frac{4\Omega^n T_e^n}{m\Delta v^2} \cdot \mathfrak{M}^n \frac{h_{\max}^n}{h_{\min}^n} \sum_j S(f^n)_j \log \left( \frac{f^n}{\mathbf{M}_{\tilde{U}^n, T_e^n}} \right)_j$$

is verified.

*Proof of Lemma A.4.* By applying Schwarz's inequality, we obtain

$$S(f^n)_j^2 \leq \frac{\Omega^n T_e^n}{m\Delta v^2} (\tilde{f}_{j+1/2}^n + \tilde{f}_{j-1/2}^n) \cdot \frac{\Omega^n T_e^n}{m\Delta v^2} \left\{ \tilde{f}_{j+1/2}^n \left[ \log \left( h_{j+1}^n / h_j^n \right) \right]^2 + \tilde{f}_{j-1/2}^n \left[ \log \left( h_{j-1}^n / h_j^n \right) \right]^2 \right\},$$

where  $h_j^n = f_j^n / \mathbf{M}_{\tilde{U}^n, T_e^n, j}$ . Moreover, we can verify that

$$\frac{f_{j\pm 1}^n}{f_j^n} \leq \frac{\mathbf{M}_{\tilde{U}^n, T_e^n, j\pm 1}}{\mathbf{M}_{\tilde{U}^n, T_e^n, j}} \cdot \frac{h_{\max}^n}{h_{\min}^n} \leq \mathfrak{M}^n \cdot \frac{h_{\max}^n}{h_{\min}^n}.$$

Then

$$(A.12) \quad \max_j \left( \frac{\tilde{f}_{j+1/2}^n + \tilde{f}_{j-1/2}^n}{f_j^n} \right) \leq \frac{2(\mathfrak{M}^n h_{\max}^n / h_{\min}^n - 1)}{\log(\mathfrak{M}^n h_{\max}^n / h_{\min}^n)}$$



since  $\frac{x-1}{\log x}$  is an increasing function when  $x > 0$ . Finally, we obtain

$$\sum_j \frac{S(f^n)_j^2}{f_j^n} \leq \frac{\Omega^n T_e^n}{m \Delta v^2} \cdot \frac{2(\mathfrak{M}^n h_{\max}^n / h_{\min}^n - 1)}{\log(\mathfrak{M}^n h_{\max}^n / h_{\min}^n)} \cdot \frac{2\Omega^n T_e^n}{m \Delta v^2} \sum_j \tilde{f}_{j+1/2}^n [\log(h_{j+1}^n / h_j^n)]^2.$$

On the other hand, due to the equalities (2.14) and (2.16), we know that

$$\sum_j S(f^n)_j \log(f^n / \mathbf{M}_{\tilde{U}^n, T_e^n})_j = -\frac{\Omega^n T_e^n}{m \Delta v^2} \sum_j \tilde{f}_{j+1/2}^n [\log(h_{j+1}^n / h_j^n)]^2,$$

which give us

$$\sum_j \frac{S(f^n)_j^2}{f_j^n} \leq -\frac{4\Omega^n T_e^n}{m \Delta v^2} \cdot \frac{\mathfrak{M}^n h_{\max}^n / h_{\min}^n - 1}{\log(\mathfrak{M}^n h_{\max}^n / h_{\min}^n)} \sum_j S(f^n)_j \log\left(\frac{f^n}{\mathbf{M}_{\tilde{U}^n, T_e^n}}\right)_j.$$

Then, if we remark that  $\frac{x-1}{\log x} \leq \max(x, 1)$  when  $x > 0$ , we get the result.  $\square$

**A.4.2. Proof of Proposition 3.3.** Using the inequality for all  $x : \log(1+x) \leq x$ , we obtain

$$\sum_j [f^{n+1} \log(f^{n+1})]_j \leq \sum_j [f^n \log(f^n)]_j + \Delta t \sum_j \left[ S(f^n) \log(f^n) + \Delta t \frac{S(f^n)^2}{f^n} \right]_j;$$

that is,

$$(A.13) \quad H^{n+1} \leq \langle f^n \log(f^n) \rangle + \Delta t \left\langle S(f^n) \log(f^n) + \Delta t \frac{S(f^n)^2}{f^n} \right\rangle - \frac{ZN^n}{2} \log(T_e^{n+1}).$$

On the other side, we have

$$(A.14) \quad T_e^{n+1} = T_e^n \left( 1 - \frac{\Delta t}{ZN^n T_e^n} m \langle v_j^2 S(f^n) \rangle \right).$$

Using (2.13), we verify that

$$\langle S(f^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n}) \rangle = -\frac{m}{2T_e^n} \langle v^2 S(f^n) \rangle,$$

which allows us to claim that

$$(A.15) \quad T_e^{n+1} = T_e^n \left( 1 + \frac{2\Delta t}{ZN^n} \left\langle S(f^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n}) \right\rangle \right)$$

using (A.14). Then, by putting (A.15) in (A.13), we find

$$\begin{aligned} H^{n+1} &\leq H^n + \Delta t \left\langle S(f^n) \log(f^n) + \Delta t \frac{S(f^n)^2}{f^n} \right\rangle \\ &\quad - \frac{ZN^n}{2} \log \left[ 1 + \frac{2\Delta t}{ZN^n} \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v \right]. \end{aligned}$$

We remark that

$$\Delta t < \Delta t_2^n \implies T_e^{n+1} > 0 \implies \frac{2\Delta t}{ZN^n} \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v > -1.$$

We deduce that

$$2\Delta t < \Delta t_2^n \implies \frac{2(2\Delta t)}{ZN^n} \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v > -1;$$

that is,

$$\Delta t < \Delta t_4^n = \Delta t_2^n / 2 \implies T_e^{n+1} > 0 \quad \text{and} \quad \frac{2\Delta t}{ZN^n} \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v > -\frac{1}{2}.$$

On the other side, we easily verify that

$$\forall x > -\frac{1}{2} : \log\left(\frac{1}{1+x}\right) < x(2x-1).$$

And, by setting  $x = \frac{2\Delta t}{ZN^n} \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v$ , we obtain that

$$\begin{aligned} H^{n+1} &\leq H^n + \Delta t \sum_j S(f_j^n) \log\left(\frac{f_j^n}{\mathbf{M}_{\tilde{U}^n, T_e^n}}\right)_j \Delta v + \Delta t \frac{S(f_j^n)^2}{f_j^n} \Delta v \\ &\quad + \frac{4\Delta t^2}{ZN^n} \left[ \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v \right]^2. \end{aligned}$$

Using Schwarz's inequality, we can also write

$$\begin{aligned} \frac{4\Delta t^2}{ZN^n} \left[ \sum_j S(f_j^n) \log(\mathbf{M}_{\tilde{U}^n, T_e^n})_j \Delta v \right]^2 &= \frac{4\Delta t^2}{ZN^n} \left[ \sum_j S(f_j^n) \frac{m(v_j - \tilde{U}^n)^2}{2T_e^n} \Delta v \right]^2 \\ &\leq \frac{4\Delta t^2}{ZN^n} \sum_j \frac{S(f_j^n)^2}{f_j^n} \Delta v \cdot \sum_j f_j^n \frac{(v_j - \tilde{U}^n)^4}{4(T_e^n/m)^2} \Delta v \\ &\leq \Delta t^2 \sum_j \frac{S(f_j^n)^2}{f_j^n} \Delta v \cdot \frac{\max_k (v_k - \tilde{U}^n)^4}{Z(T_e^n/m)^2}. \end{aligned}$$

Then

$$H^{n+1} \leq H^n + \Delta t \sum_j \left[ S(f_j^n) \log\left(\frac{f_j^n}{\mathbf{M}_{\tilde{U}^n, T_e^n}}\right)_j + \Delta t(1 + \alpha^n) \frac{S(f_j^n)^2}{f_j^n} \right] \Delta v.$$

And, by applying Lemma A.4, we obtain

$$H^{n+1} \leq H^n + \Delta t \left[ 1 - \frac{4\Delta t \Omega^n T_e^n}{m \Delta v^2} \cdot \mathfrak{M}^n \frac{h_{\max}^n}{h_{\min}^n} (1 + \alpha^n) \right] \cdot \sum_j S(f_j^n) \log\left(\frac{f_j^n}{\mathbf{M}_{\tilde{U}^n, T_e^n}}\right)_j \Delta v.$$

We conclude the proof by using the equality (2.14) and the inequality (2.16) which allow us to write that when  $\Delta t < \Delta t_3^n$ , we have  $H^{n+1} \leq H^n$ . To show that  $H^n \geq H(f^\infty, T_e^\infty)$ , we use the first point of Proposition 2.7, which does not depend on the time discretization.

## REFERENCES

- [1] J. J. DUDERSTADT AND G. A. MOSES, *Inertial Confinement Fusion*, Wiley-Interscience, New York, 1982.
- [2] S. I. BRAGINSKII, *Transport processes in a plasma*, in Reviews of Plasma Physics, Vol. 1, Consultants Bureau, New York, 1965, pp. 205–311.
- [3] A. DECOSTER, *Fluid equations and transport coefficients of plasmas*, in Modeling of Collisions, P. A. Raviart, ed., Masson, Paris, 1998.
- [4] I. P. SHKAROVSKY, T. W. JOHNSTON, AND M. P. BACHYNSKI, *The Particle Kinetics of Plasmas*, Addison-Wesley, Reading, MA, 1966.
- [5] R. L. BROWERS AND J. R. WILSON, *Numerical Modeling in Applied Physics and Astrophysics*, Jones-Bartlett, Boston, 1991.
- [6] B. DESPRÉS, *Inégalité entropique pour un Solveur conservatif du système de la dynamique des gaz en coordonnées de Lagrange*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1301–1306.
- [7] M. CASANOVA, O. LARROCHE, AND J. P. MATTE, *Kinetic simulation of a collisional shock wave in a plasma*, Phys. Rev. Lett., 67 (1991), pp. 2143–2146.
- [8] J. S. CHANG AND G. COOPER, *A practical difference scheme for Fokker-Planck equations*, J. Comput. Phys., 6 (1970), pp. 1–16.
- [9] V. A. MOUSSEAU AND D. A. KNOLL, *Fully implicit kinetic solution of collisional plasmas*, J. Comput. Phys., 136 (1997), pp. 308–323.
- [10] E. M. EPPERLEIN, *Implicit and conservative difference scheme for the Fokker-Planck equation*, J. Comput. Phys., 112 (1994), pp. 291–297.
- [11] YU. A. BEREZIN, V. N. KHUDICK, AND M. S. PEKKER, *Conservative finite difference schemes for the Fokker-Planck equation not violating the law of an increasing entropy*, J. Comput. Phys., 69 (1987), pp. 163–174.
- [12] P. DEGOND AND B. LUCQUIN-DESREUX, *The Fokker-Planck asymptotics of the Boltzmann collision operator in the Coulomb case*, Math. Models Methods Appl. Sci., 2 (1992), pp. 167–182.
- [13] P. DEGOND AND B. LUCQUIN-DESREUX, *The asymptotics of collision operators for two species of particles of disparate masses*, Math. Models Methods Appl. Sci., 6 (1996), pp. 405–436.
- [14] P. DEGOND AND B. LUCQUIN-DESREUX, *An entropy scheme for the Fokker-Planck collision operator of plasma kinetic theory*, Numer. Math., 68 (1994), pp. 239–262.
- [15] C. BUET AND S. CORDIER, *Numerical analysis of the isotropic Fokker-Planck-Landau equation*, in preparation.
- [16] C. BUET AND S. CORDIER, *Numerical analysis of conservative and entropy schemes for the Fokker-Planck-Landau equation*, SIAM J. Numer. Anal., 36 (1999), pp. 953–973.
- [17] C. BUET, S. CORDIER, *Conservative and entropy decaying numerical scheme for the isotropic Fokker-Planck-Landau equation*, J. Comput. Phys., 145 (1998), pp. 228–245.
- [18] C. BUET, S. CORDIER, P. DEGOND, AND M. LEMOU, *Fast algorithms for numerical, conservative and entropy approximations of the Fokker-Planck-Landau equation*, J. Comput. Phys., 133 (1997), pp. 310–322.
- [19] S. DELLACHERIE, *Contribution à l'analyse et à la simulation numériques des équations cinétiques décrivant un plasma chaud*, Ph.D. Thesis, Université Denis Diderot Paris VII, Paris, 1998.
- [20] S. DELLACHERIE, *Sur un schéma numérique semi-discret appliqué à une équation de Fokker-Planck isotrope*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 1219–1224.
- [21] S. KULLBACK, *A lower bound for discrimination information in terms of variation*, IEEE Trans. Inform. Theory, 13 (1966), pp. 126–127.
- [22] G. TOSCANI, *Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation*, Quart. Appl. Math., 57 (1999), pp. 521–541.

# The grazing collision limit for the Boltzmann–Lorentz model

C. Buet<sup>a</sup>, S. Cordier<sup>b</sup> and B. Lucquin-Desreux<sup>b</sup>

<sup>a</sup>CEA/DRIR, B.P. 12, 91 680 Bruyères-Le-Châtel, France

E-mail: buet@bruyeres.cea.fr

<sup>b</sup>Laboratoire d'Analyse Numérique, UMR CNRS 7598, Université Pierre et Marie Curie, Boîte courrier 187, 75 252 Paris cedex 05, France

E-mail: cordier@ann.jussieu.fr; lucquin@ann.jussieu.fr

**Abstract.** The Lorentz operators are derived from either Boltzmann or Fokker–Planck collisions operators when considering a mixture of species with disparate masses [8]. The Fokker–Planck operator is the so called “grazing collision limit” of the Boltzmann operator as proved in [1,12,7]. In our simpler case, we improve the results by proving uniform in time convergence and by controlling the speed of the trend to equilibrium. The results are based on a spectral analysis of the operators which share the same basis of eigenfunctions.

## 1. Introduction

The Fokker–Planck collision operator is usually considered as an approximation of the Boltzmann collision operator when the collisions become grazing. This has been proved in a series of papers, starting from Arsen'ev and Buryak who show in [1] this convergence result under restrictive assumptions on the scattering cross section and on the initial data. A mathematical framework has also been given by Desvillettes for more physical situations, but which still exclude the Coulombian case; in [12], the scattering cross-section is smooth and depends upon a small parameter  $\varepsilon$  which tends to zero. This parameter however does not have any precise physical meaning. A quite different asymptotics is used in [7] to treat the Coulombian case: here, the scattering cross section has a non integrable singularity when the relative velocity of the colliding particles tends to zero. Moreover, the small parameter involved in this asymptotics has an actual physical meaning: it is clearly identified to the plasma parameter. Recently, Villani [19] obtained a complete rigorous proof of this asymptotic problem in the space homogeneous situation, and for potentials which are not “too soft”.

In this paper, we are concerned with a simplified collision operator which is known as the Boltzmann–Lorentz model in plasma physics. It is used to describe the effects of collisions of electrons with neutral particles. In first approximation, the electrons are assumed to diffuse with a stationary equilibrium distribution of target particles. The simplified Boltzmann–Lorentz model is then derived from the Boltzmann equation in the limit of small electron mass with respect to the mass of atoms; this asymptotics has been completely justified from the theory of kinetic collisional operators in [8]. The Boltzmann–Lorentz model can also be found in the framework of semi-conductors (see [4]). More sophisticated models of the same form have also been recently studied in the context of wave-particle modelling [10], cometary plasma [9], ionic plasma thurster. . .

In the same way, we can define a simplified Fokker–Planck–Lorentz model which can be derived from the Fokker–Planck–Landau equation in the limit of small electron mass with respect to the mass of ions (see again [8]). Physical situations actually exist for which this operator appears as the leading order collision term (see [14] for an example in the context of plasmas). From a probabilistic point of view, the Fokker–Planck–Lorentz operator describes a random walk of the particles on any sphere of constant energy.

In this paper, we are interested with the limit of the Boltzmann–Lorentz operator towards the Fokker–Planck–Lorentz model in the so-called grazing collision limit. Two situations are considered: a first one, denoted by “case 1” which corresponds to the asymptotics developed by Desvillettes in [12]; a second one, called “case 2”, is the asymptotics introduced in [7] to treat the Coulombian case. In both cases, we show the convergence of the operators, but also of the solutions of the Cauchy problems associated with these operators: for the last, the convergence is uniform with respect to time, on any finite time interval. But, thanks to a precise spectral analysis, we can go further. First, we can show a uniform convergence result when time goes to infinity. Second, we have a precise knowledge of the convergence speed towards equilibrium. This spectral analysis also presents a great interest as a numerical point of view: it gives in particular exact solutions which allow the validation of numerical codes. Numerical experiments are the object of a forthcoming paper [6].

Our paper is organized in the following way. In part 2, we present the operators under consideration. We then introduce the grazing collision asymptotics and show that the Boltzmann–Lorentz operator tends towards the Fokker–Planck–Lorentz operator, for very regular distribution functions. Less regular situations are considered in part 3. A spectral analysis is then carried out. The key point relies on the fact that the two operators have the same eigenfunctions. Convergence results are then established for the eigenvalues associated with these eigenfunctions, either in case 1, or in the Coulombian case. The proofs are detailed in the three dimensional case, which is the real physical case; main results are also valid in two dimensions (see Remark 3.4). We show the convergence of the solutions of the Cauchy problem associated with the Boltzmann–Lorentz operator towards the solutions of the Cauchy problem associated with the Fokker–Planck–Lorentz operator. In part 4, we generalize this result in presence of an exterior magnetic field. The dependance with respect to the energy variable is also restored.

## 2. The Lorentz models

### 2.1. Definition

The “Boltzmann–Lorentz” model is the elastic-collision operator which has the following expression in  $d$  velocity-dimension [8]:

$$Q(f)(\omega) = \int_{S^{d-1}} B(\omega - \omega') [f(\omega') - f(\omega)] d\omega', \quad (2.1)$$

where  $S^{d-1}$  denotes the unit sphere in  $\mathbb{R}^d$ . The cross section  $B$  is a positive function which expression is directly connected to the type of interacting potential between the particles. More precisely,  $B(\omega - \omega')$  only depends on  $\|\omega - \omega'\|$ .

Physical cases are mainly in three dimensions; from now on, we suppose that  $d = 3$ . Let us precise the notations. We introduce the following local orthonormal basis  $(e_1, e_2, \omega)$ . Using spherical coordinates, we can write

$$\omega' = \sin \theta (\cos \phi e_1 + \sin \phi e_2) + \cos \theta \omega, \quad (2.2)$$

where  $\phi \in [0, 2\pi]$  and  $\theta \in [0, \pi]$ . As a physical point of view, the angle  $\theta$  represents the so called “scattering angle”, i.e., the angle of deviation undergone during a collision by a particle having  $\omega$  as initial velocity and  $\omega'$  as post-collisional velocity. The integrand  $d\omega'$  denotes the elementary area of  $S^2$ ; with the above notations, we have  $d\omega' = \sin \theta d\theta d\phi$ . Now, a simple computation shows that  $\|\omega - \omega'\|^2 = 2(1 - \cos(\theta))$ , so that  $B(\omega - \omega')$  only depends on the scattering angle  $\theta$  (and not on  $\phi$ ), or, equivalently, on the scalar product  $\omega \cdot \omega'$ .

The Lorentz–Fokker–Planck operator is nothing but the classical Laplace–Beltrami operator on the unit sphere  $S^2$ . Still using spherical coordinates, it is defined by:

$$P(f) = \Delta_\omega f = \frac{1}{\sin \theta} \left[ \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin \theta} \frac{\partial^2 f}{\partial \phi^2} \right]. \quad (2.3)$$

Let us recall that this operator in fact appears quite naturally, when considering for example collisions of heavy charged particles against light ones. We write the velocities  $v$  in spherical coordinates (i.e.,  $v = |v|\omega$ ). The leading order term of the Fokker–Planck–Landau collision operator is then precisely, up to a multiplicative function of the modulus of the velocity, the operator  $P(f)$ . We refer to [14] for a precise description of this.

## 2.2. Cross sections for grazing collisions

In the grazing collision asymptotics, the cross section  $B$  in the Boltzmann–Lorentz operator (2.1) is supposed to depend on a small parameter  $\varepsilon$ ; we denote it by  $B^\varepsilon$ . The associated Boltzmann–Lorentz operator then writes

$$Q^\varepsilon(f)(\omega) = \int_{S^{d-1}} B^\varepsilon(\omega - \omega') [f(\omega') - f(\omega)] d\omega'. \quad (2.4)$$

Let us now precise the definition of the cross section  $B^\varepsilon$ . As we have seen previously, the kernel  $B^\varepsilon$  only depends on one angle  $\theta \in [0, \pi]$ , so that we can write  $B^\varepsilon(\omega - \omega') d\omega' = \overline{B}^\varepsilon(\theta) d\theta d\phi$ . Now for most potentials,  $\overline{B}^\varepsilon$  is of the form [12]

$$\overline{B}^\varepsilon(\theta) = \frac{1}{\varepsilon^3} \overline{B}\left(\frac{\theta}{\varepsilon}\right), \quad (2.5)$$

with  $\overline{B}(\theta) = \sigma(\theta) \sin(\theta) \chi_{[0, \pi]}(\theta)$ ; we denote by  $\chi_{[a, b]}$  the characteristic function of the interval  $[a, b]$  and  $\sigma$  is a positive function. We then have:

$$\overline{B}^\varepsilon(\theta) = \frac{1}{\varepsilon^3} \sigma\left(\frac{\theta}{\varepsilon}\right) \sin\left(\frac{\theta}{\varepsilon}\right) \chi_{[0, \varepsilon\pi]}(\theta). \quad (2.6)$$

Unfortunately, this asymptotics does not allow to treat the Coulombian case, which is the most relevant physical case. In the last, the kernel writes

$$\overline{B}^\varepsilon(\theta) = \sigma(\theta) \frac{1}{\text{Log}(1/\sin \frac{\theta}{2})} \frac{\sin \theta}{[\sin \frac{\theta}{2}]^4} \chi_{[\varepsilon, \pi]}(\theta), \quad (2.7)$$

where the positive function  $\sigma$  is such that  $\sigma(0) \neq 0$ , and the parameter  $\varepsilon$  has a physical meaning: it is what physicists call the “plasma parameter” [7]. The logarithm factor represents the so called “Coulombian logarithm”.

From now on, we denote by “case 1” the non-Coulombian case which corresponds to cross-sections of the form (2.6), and by “case 2” the Coulombian one; in the last, the cross sections are given by (2.7).

### 2.3. The grazing limit for smooth distribution functions

In this part, we present a formal justification of the so called grazing collision limit, which can be justified in the context of regular distribution functions. In the next paragraph, we shall perform a spectral analysis of the operators  $Q^\varepsilon$  and  $P$ . This will allow to get a more precise result, for less regular distribution functions; but it will also give a precise result concerning the trend to equilibrium.

**Proposition 2.1.** *Let  $f \in C^3(S^2)$ . We consider the Boltzmann–Lorentz operator (2.4) with cross sections of the form (2.6), i.e., case 1. We suppose moreover that the cross section  $\sigma$  is such that:*

$$C = \int_0^\pi \mu^2 \sigma(\mu) \sin \mu \, d\mu < +\infty. \quad (2.8)$$

Then, for all  $\omega \in S^2$ , we have:

$$\lim_{\varepsilon \rightarrow 0} Q^\varepsilon(f)(\omega) = C \frac{\pi}{2} \Delta_\omega f(\omega).$$

**Proof.** The function  $f$ , defined on the sphere, is first extended on the whole space in such a way that the third derivative of  $f$  remains bounded. For example, one can use the following extension of  $f$ :  $\tilde{f}(\omega) = f(\omega/\|\omega\|)\Psi(\|\omega\|)$ , where  $\omega \in \mathbb{R}^d$  and  $\Psi$  is a  $C^3$  function which is equal to 1 in a neighborhood of 1. We then use the following Taylor expansion

$$f(\omega') = f(\omega) + (\omega' - \omega)_i \partial_i \tilde{f}(\omega) + \frac{(\omega' - \omega)_{i,j}^2}{2} \partial_{i,j}^2 \tilde{f}(\omega) + R(\omega, \omega'),$$

where the integral remainder is such that  $|R(\omega, \omega')| \leq M \|\omega - \omega'\|^3$ , due to the regularity assumptions on  $f$ . Note that now  $\omega$  are vectors in  $\mathbb{R}^3$  and  $\partial_i$  denote the partial derivatives of  $f$  in the  $i$ th direction and  $x_{i,j}$  is the tensor product  $x_i x_j$ . The index  $i$  and  $j$  varie in  $\{1, 2, 3\}$  and we use the Einstein convention (summation of repeated index).

We write  $\omega'$  in the basis (2.2) linked to  $\omega$ . Recall that in this basis  $\omega' - \omega = \sin \theta (\cos \phi e_1 + \sin \phi e_2) + (\cos \theta - 1)\omega$ , since  $e_3 = \omega$ . The first term in (2.4) corresponding to derivatives of order 1 vanishes using the evenness of the kernel. It remains:

$$Q^\varepsilon(f)(\omega) = \frac{1}{2} \partial_{i,j}^2 \tilde{f}(\omega) \int_{S^2} \overline{B}^\varepsilon(\theta) (\omega' - \omega)_{i,j} \, d\theta \, d\phi + R^\varepsilon(\omega),$$

where  $\theta, \phi$  are the spherical coordinate of  $\omega'$  in the basis (2.2), while  $R^\varepsilon$  is defined by:  $R^\varepsilon(\omega) = \int_{S^2} \overline{B}^\varepsilon(\theta) R(\omega - \omega') d\theta d\phi$ . Let us now consider the first integral term, i.e., the matrix defined by:

$$\Phi_{i,j}^\varepsilon(\omega) \stackrel{\text{def}}{=} \int_{S^2} \overline{B}^\varepsilon(\theta) (\omega' - \omega)_{i,j} d\theta d\phi.$$

This matrix is diagonal since it vanishes for each  $i \neq j$  by evenness. Then, for  $i = j = 3$ , we have, integrating in  $\phi$  and using (2.6)

$$\Phi_{3,3}^\varepsilon(\omega) = 2\pi \int_0^\pi (\cos \theta - 1)^2 \overline{B}^\varepsilon(\theta) d\theta = \frac{8\pi}{\varepsilon^2} \int_0^\pi \sin^4 \frac{\varepsilon\mu}{2} \sigma(\mu) \sin \mu d\mu \rightarrow 0, \quad \text{as } \varepsilon \rightarrow 0,$$

thanks to the assumption (2.8). By integrating over  $\phi$ , we also have

$$\Phi_{1,1}^\varepsilon(\omega) = \Phi_{2,2}^\varepsilon(\omega) = \pi \int_0^\pi \sin^2(\theta) \overline{B}^\varepsilon(\theta) d\theta = \frac{\pi}{\varepsilon^2} \int_0^\pi \sin^2(\varepsilon\mu) \sigma(\mu) \sin \mu d\mu \rightarrow \pi C, \quad \text{as } \varepsilon \rightarrow 0,$$

by using the assumption (2.8) and the definition of  $C$ . Therefore,  $\Phi^\varepsilon \rightarrow \Phi$ , when  $\varepsilon \rightarrow 0$ , where  $\Phi$  is the projection matrix onto the plane perpendicular to  $\omega$  multiplied by  $\pi C$ , that is to say  $\Phi(\omega) = \pi C(\text{Id} - (\omega \times \omega)/\|\omega\|^2)$ . Moreover, we have:  $\partial_{i,j} \tilde{f}(\omega) \cdot \Phi_{i,j}(\omega) = \nabla \cdot (\Phi(\omega) \nabla \tilde{f}) = \Delta_\omega \tilde{f}(\omega)$ , for  $\omega \in S^2$ . Using the following upper bound  $\|\omega' - \omega\|^3 \leq 8 \sin^3(\theta/2)$ , we get an estimate for the remainder term

$$|R^\varepsilon(\omega)| \leq \frac{16\pi M}{\varepsilon^2} \int_0^\pi \sin^3 \frac{\varepsilon\mu}{2} \sigma(\mu) \sin \mu d\mu,$$

which shows that this term vanishes when  $\varepsilon \rightarrow 0$ . This ends the proof.  $\square$

Using the semigroup theory, we deduce from Proposition 2.1 that  $T^\varepsilon$ , the continuous semigroup associated with  $Q^\varepsilon$ , converges towards the semigroup  $T$  associated with the Laplace–Beltrami operator. Then, the Pazy theorem [16] allows to obtain for any arbitrary large time  $T > 0$  and any initial data  $f_0 \in L^p$ , the following estimate:

$$\lim_{\varepsilon \rightarrow 0} \sup_{t \in [0, T]} \|T^\varepsilon(t)(f_0) - T(t)(f_0)\|_{L^p(S^2)} = 0.$$

The convergence is thus uniform in time but on bounded intervals of the type  $[0, T]$ . This also implies convergence results in some  $L^\infty(0, \infty, L^p(S^2))$  weak (like in the work of Desvillettes). We shall not detail the proof of this result since it does not give us informations on the trend to equilibrium, i.e., on the large time behaviour of the solution. It is however expected that, for large time, the solution homogenizes, i.e., tends towards a constant state (the average of the initial distribution function on the unit sphere). But, it is not clear, using such techniques, if the convergence is uniform in time; the trend towards this constant state can be as slow as the grazing collision parameter  $\varepsilon$  tends to 0. In the next part, we shall answer to this point using spectral analysis of the operator which gives more precise (explicit, uniform in time) result on the existence of solution and on its grazing collision limit.



### 3. Spectral analysis of the Boltzmann–Lorentz model

#### 3.1. The spectral analysis

The spherical harmonics  $Y_{lm}$  defined by [15]

$$Y_{lm}(\theta, \phi) = (-1)^m i^l \left[ \frac{l+1/2}{2\pi} \frac{(l-m)!}{(l+m)!} \right]^{1/2} e^{im\phi} (\sin \theta)^m \left( \frac{d}{dx} \right)^m P_l(\cos \theta),$$

where  $P_l$  denotes the Legendre polynomials, are eigenfunctions of the Laplace–Beltrami operator. More precisely, we have, for  $l \geq 0$  and  $-l \leq m \leq l$  [15]:

$$-\Delta_\omega Y_{lm} = \nu_l Y_{lm}, \quad \text{with } \nu_l = l(l+1).$$

These functions form an orthonormal basis of the space  $L^2(S^2)$ . Moreover, they are also eigenfunctions of the operator  $Q^\varepsilon$  [3,11], and we have:

$$-Q^\varepsilon(Y_{lm}) = \nu_l^\varepsilon Y_{lm}, \quad \text{with } \nu_l^\varepsilon = 2\pi \int_0^\pi [1 - P_l(\cos \theta)] \overline{B}^\varepsilon(\theta) d\theta. \quad (3.1)$$

These eigenvalues first satisfy  $\nu_l^\varepsilon \geq 0$ ,  $\nu_0^\varepsilon = 0$ , as a direct consequence of the fact that [13]:  $\forall x \in [-1, 1]$ ,  $-1 \leq P_l(x) \leq 1$ ,  $P_0(x) = 1$ . In order to get a uniform bound, in terms of  $\varepsilon$ , of the eigenvalues, we first show the following preparatory lemma:

**Lemma 3.1.** *Let us suppose that:*

$$\int_0^\pi \sigma(\theta) \sin \theta d\theta < +\infty, \quad \text{in case 1,} \quad (3.2)$$

$$\sup_{\theta \in [0, \pi]} \sigma(\theta) < +\infty, \quad \text{in case 2.} \quad (3.3)$$

Then, we have:

$$\exists C_1(\sigma) > 0 \text{ such that } \forall \varepsilon > 0, \int_0^\pi \theta^2 \overline{B}^\varepsilon(\theta) d\theta \leq C_1(\sigma). \quad (3.4)$$

Moreover, there exists a positive constant  $C(\sigma)$ , independent of  $\varepsilon$ , such that:

$$\text{for all } l \geq 0, \quad \nu_l^\varepsilon \leq C(\sigma) \nu_l. \quad (3.5)$$

**Proof.** We first prove (3.4). The first case results from a simple change of variables. In the Coulombian case, we have:

$$\int_0^\pi \theta^2 \overline{B}^\varepsilon(\theta) d\theta = \frac{2}{\text{Log}(1/\sin \frac{\varepsilon}{2})} \int_\varepsilon^\pi \theta^2 \sigma(\theta) \frac{\cos \frac{\theta}{2}}{[\sin \frac{\theta}{2}]^3} d\theta \leq 16 \sup_{[0,1]} \left( \frac{\arcsin x}{x} \right)^2 \sup_{\theta \in [0, \pi]} \sigma(\theta),$$

which gives the expected estimate. Let us now show (3.5). Since [13]  $P_l(1) = 1$ ,  $\sup_{t \in [-1,1]} |P'_l(t)| \leq l(l+1)/2$ , we have:

$$|1 - P_l(1+x)| = \left| \int_0^x P'_l(1+t) dt \right| \leq |x| \frac{l(l+1)}{2}.$$

Applying this to  $x = \cos(\theta) - 1$ , we get, since  $|x| \leq \theta^2/2$ :

$$\nu_l^\varepsilon \leq \pi \frac{l(l+1)}{2} \int_0^\pi \theta^2 \overline{B}^\varepsilon(\theta) d\theta,$$

which gives (3.5) for  $C(\sigma) = (\pi/2)C_1(\sigma)$ .

### 3.2. The Cauchy problem

We are interested with the following Cauchy problem

$$\frac{\partial f^\varepsilon}{\partial t} = Q^\varepsilon(f^\varepsilon), \quad f^\varepsilon(t=0) = f_0, \quad (3.6)$$

where  $f_0$  is a given function in  $L^2(S^2)$ . We split this function in the orthonormal basis formed by the spherical harmonics, i.e., we write:

$$f_0 = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} Y_{lm}, \quad \text{with } \|f_0\|_{L^2(S^2)}^2 = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2 < +\infty. \quad (3.7)$$

From the above spectral analysis, we easily get:

**Proposition 3.2.** *The function  $f^\varepsilon$  defined by*

$$f^\varepsilon(t) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} \exp(-\nu_l^\varepsilon t) Y_{lm} \quad (3.8)$$

*is a weak solution, in the space  $L^\infty((0, \infty); L^2(S^2))$ , of (3.6). Moreover, if the initial data is such that*

$$\sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2 \nu_l < +\infty, \quad (3.9)$$

*then the above solution is the unique one such that  $f^\varepsilon \in L^\infty(0, \infty; H^1(S^2))$  and  $Q^\varepsilon(f^\varepsilon) \in L^\infty(0, \infty; H^{-1}(S^2))$ .*

**Proof.** The first point results from (3.1) and the non negativity of the  $\nu_l^\varepsilon$ . Now condition (3.9) implies that  $Q^\varepsilon(f^\varepsilon) \in L^\infty((0, \infty); H^{-1}(S^2))$ . In fact any  $\varphi \in H^1(S^2)$  writes

$$\varphi = \sum_{l \geq 0} \sum_{m=-l}^{m=l} b_{lm} Y_{lm}, \quad \text{with } \sum_{l \geq 0} \sum_{m=-l}^{m=l} b_{lm}^2 \nu_l < +\infty,$$

so that by Cauchy–Schwartz inequality, we have, for any fixed  $L > 0$ ,

$$\begin{aligned} \left| \int_{S^2} \left( \sum_{l \leq L} \sum_{m=-l}^{m=l} a_{lm} \nu_l^\varepsilon Y_{lm} \right) \varphi \, d\omega \right| &= \left| \sum_{l \leq L} \sum_{m=-l}^{m=l} a_{lm} b_{lm} \nu_l^\varepsilon \right| \\ &\leq C(\sigma) \sqrt{\sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2 \nu_l} \sqrt{\sum_{l \geq 0} \sum_{m=-l}^{m=l} b_{lm}^2 \nu_l}, \end{aligned}$$

thanks to the estimate (3.5). The Banach–Steinhaus theorem shows that  $\sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} \nu_l^\varepsilon Y_{lm} \in H^{-1}(S^2)$ , so that  $Q^\varepsilon(f^\varepsilon) \in L^\infty((0, \infty); H^{-1}(S^2))$ . We also have, for all  $t > 0$ ,

$$\langle Q^\varepsilon(f^\varepsilon)(t), f^\varepsilon(t) \rangle_{H^{-1}, H^1} = - \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2 \nu_l^\varepsilon \exp(-2\nu_l^\varepsilon t) \leq 0,$$

which gives

$$\int_{S^2} (f^\varepsilon(t, \omega))^2 \, d\omega \leq \int_{S^2} (f_0(\omega))^2 \, d\omega,$$

and shows the uniqueness.  $\square$

In the same way, the function  $f$  defined by

$$f(t) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} \exp(-\nu_l t) Y_{lm} \quad (3.10)$$

is a weak solution, in the space  $L^\infty((0, \infty); L^2(S^2))$  of the Cauchy problem:

$$\frac{\partial f}{\partial t} = \Delta_\omega f, \quad f(t=0) = f_0. \quad (3.11)$$

Moreover, if the initial condition satisfies (3.9), then the above solution is the unique one such that  $f \in L^\infty((0, \infty); H^1(S^2))$  and  $\Delta_\omega f \in L^\infty((0, \infty); H^{-1}(S^2))$ .

### 3.3. The grazing collision limit

Let us now examine the convergence of the eigenvalues when  $\varepsilon \rightarrow 0$ .

**Lemma 3.3.** *Let us suppose that:*

$$\int_0^\pi \sigma \theta \sin \theta \, d\theta = \frac{2}{\pi}, \quad \text{in case 1,} \quad (3.12)$$

$$\sigma(0) = \frac{1}{2\pi}, \quad \text{in case 2.} \quad (3.13)$$

Let  $\varphi = \varphi(\theta)$  be any function of class  $C^2$  such that  $\varphi(0) = \varphi'(0) = 0$ . Then:

$$\int_{S^2} B^\varepsilon(\omega) \varphi(\theta) d\omega \rightarrow 2\varphi''(0), \quad \text{when } \varepsilon \rightarrow 0. \quad (3.14)$$

**Proof.** In both cases, we use a change of variables followed by a Taylor expansion around point 0. We get in the first case:

$$\int_{S^2} B^\varepsilon(\omega) \varphi(\omega) d\omega \rightarrow \pi \varphi''(0) \left( \int_0^\pi \sigma x \sin x x^2 dx \right), \quad \text{when } \varepsilon \rightarrow 0.$$

In the Coulombian case, we have:

$$\int_{S^2} B^\varepsilon(\omega) \varphi(\omega) d\omega = \frac{4\pi}{\text{Log}(1/\sin \frac{\varepsilon}{2})} \int_\varepsilon^\pi (\varphi\sigma)(\theta) \frac{\cos(\frac{\theta}{2})}{[\sin \frac{\theta}{2}]^3} d\theta \rightarrow 4\pi\sigma(0)\varphi''(0), \quad \text{when } \varepsilon \rightarrow 0,$$

because  $\varphi(0) = \varphi'(0) = 0$ .

We deduce from this the:

**Proposition 3.4.** *Under the hypotheses of Lemma 3.3, we have:*

$$\text{for all } l \geq 0, \quad \nu_l^\varepsilon \rightarrow \nu_l = l(l+1), \quad \text{when } \varepsilon \rightarrow 0. \quad (3.15)$$

For any finite time interval  $[0, T]$ , we have:

$$\sup_{[0, T]} \|f^\varepsilon(t) - f(t)\|_{L^2(S^2)} \rightarrow 0, \quad \text{when } \varepsilon \rightarrow 0. \quad (3.16)$$

**Proof.** We first apply (3.14) with  $\varphi(\theta) = 1 - P_l(\cos \theta)$ . The point (3.15) is then a simple consequence of the fact that:  $\varphi''(0) = P_l'(1) = l(l+1)/2$ . We deduce from (3.8) and (3.10) that

$$f^\varepsilon(t) - f(t) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} Y_{lm} [\exp(-\nu_l^\varepsilon t) - \exp(-\nu_l t)], \quad (3.17)$$

with  $\sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2 < +\infty$ , so that we can write, for any  $t \in [0, T]$ :

$$\|f^\varepsilon(t) - f(t)\|_{L^2(S^2)}^2 \leq 2 \sum_{l \geq L} \sum_{m=-l}^{m=l} a_{lm}^2 + \sum_{l=0}^{L-1} \sum_{m=-l}^{m=l} a_{lm}^2 |\nu_l^\varepsilon - \nu_l| T.$$

The first sum is the remainder of a convergent series, which can be arbitrary small for a sufficiently large index  $L$ . The second part is then a finite sum of terms which all vanish when  $\varepsilon \rightarrow 0$ , which shows (3.16) and ends the proof.  $\square$

### 3.4. The large time behaviour

We are now interested with the large time behaviour of the solutions (3.8) and (3.10). In order to do this, we need a positive lower bound for the opposite of the eigenvalues  $\nu_l^\varepsilon$  and  $\nu_l$ , which means we have to “eliminate” the zero eigenvalue which corresponds to the equilibrium states. Since the kernels of the Lorentz operators are formed of constant functions, we easily get following result (we state it here for  $Q^\varepsilon$ , but this result also trivially holds for  $P$ ):

**Lemma 3.5.** *Let  $f^\varepsilon$  be given by (3.8). The function  $g^\varepsilon$  defined by  $g^\varepsilon = f^\varepsilon - I(f_0)$ , with  $I(f_0) = \frac{1}{4\pi} \int_{S^2} f_0 d\omega$ , is of zero mean value over the unit sphere and it satisfies the following Cauchy problem:*

$$\frac{\partial g^\varepsilon}{\partial t} = Q^\varepsilon(g^\varepsilon), \quad g^\varepsilon(t=0) = g_0, \quad \text{with } g_0 = f_0 - I(f_0). \quad (3.18)$$

Moreover, using the notation (3.7), we have:

$$g_0 = \sum_{l>0} \sum_{m=-l}^{m=l} a_{lm} Y_{lm}, \quad g^\varepsilon(t) = \sum_{l>0} \sum_{m=-l}^{m=l} a_{lm} \exp(-\nu_l^\varepsilon t) Y_{lm}. \quad (3.19)$$

We now analyze the convergence towards equilibrium. First, since  $g = f - I(f_0)$ , with  $f$  given by (3.10), is of zero mean value, we have the following ellipticity relation, where  $\lambda_1$  denotes the first non zero eigenvalue of the Laplace–Beltrami operator [2]:

$$\|\nabla_\omega g\|_{L^2(S^2)}^2 \geq \frac{1}{\lambda_1} \|g\|_{L^2(S^2)}^2;$$

this gives  $\int_{S^2} g^2 d\omega \leq \exp(-\frac{1}{\lambda_1} t) \int_{S^2} g_0^2 d\omega$ , which shows that  $g$  converges exponentially fast towards zero, when time goes to infinity. Now, if there exists a positive constant  $C$  such that, for all  $\varepsilon$  we have  $B^\varepsilon \geq C$ , then the same behaviour holds for  $g^\varepsilon$ . Our aim is to generalize this result for situations where the above assumption on the kernel is not satisfied. In order to do so, we need a uniform lower bound for the  $\nu_l^\varepsilon$ .

**Lemma 3.6.** *We suppose the assumption (3.2) fulfilled, with  $\sigma$  non identically equal to zero. In the Coulombian case, we suppose that there exists positive constants  $\theta_0$  and  $\sigma_0$ ,  $\theta_0 < \pi/2$ , such that:  $\inf_{\theta \in [0, \theta_0]} \sigma(\theta) \geq \sigma_0$ . Then, we have:*

$$\exists \varepsilon_0 > 0, C_2(\sigma) > 0 \text{ such that } \forall \varepsilon \in ]0, \varepsilon_0[, \quad \int_0^{\frac{\pi}{2}} \theta^2 \overline{B^\varepsilon}(\theta) d\theta \geq C_2(\sigma). \quad (3.20)$$

Moreover, there exists positive constants  $\nu^0$  and  $\varepsilon_0$ , such that:

$$\text{for all } l \geq 1 \text{ and all } \varepsilon \in [0, \varepsilon_0], \quad \nu_l^\varepsilon \geq \nu^0. \quad (3.21)$$

**Proof.** We first prove (3.20). Like in Lemma 3.1, the first case results from a simple change of variables; we find  $\varepsilon_0 = 1/2$  and  $C_2(\sigma) = \int_0^\pi \sigma(u) \sin uu^2 du$ . In the Coulombian case, we have, choosing  $\varepsilon_0 = \theta_0$ ,

$$\int_0^{\frac{\pi}{2}} \theta^2 \overline{B}^\varepsilon(\theta) d\theta \geq \frac{8\sqrt{2}}{\text{Log}(1/\sin \frac{\varepsilon}{2})} \sigma_0 \text{Log}\left(\frac{\theta_0}{\varepsilon}\right),$$

which gives the following estimate for  $\varepsilon_0$  sufficiently small:  $\int_0^{\frac{\pi}{2}} \theta^2 \overline{B}^\varepsilon(\theta) d\theta \geq 4\sqrt{2}\sigma_0$ . Let us now get a uniform lowerbound for  $\nu_l^\varepsilon$ . We recall the Laplace formula [13] ( $i^2 = -1$ )

$$P_l(x) = \frac{1}{\pi} \int_0^\pi (x + i\sqrt{1-x^2} \cos \phi)^l d\phi, \quad -1 \leq x \leq 1,$$

which gives:  $|P_l(x)| \leq \frac{1}{\pi} \int_0^\pi (x^2 + (1-x^2)\cos^2 \phi)^{l/2} d\phi$ . The quantity inside the integral being less than 1, we deduce that, for any  $l \geq 2$ , we have  $|P_l(x)| \leq \frac{1}{\pi} \int_0^\pi (x^2 + (1-x^2)\cos^2 \phi) d\phi$ , so that:  $|P_l(x)| \leq (1+x^2)/2$ . Now this last relation is also valid for  $l = 1$ , since  $P_1(x) = x$ , which finally gives:  $1 - P_l(x) \geq (1-x^2)/2$ , for  $l \geq 1$ . Using the inequality  $\sin(\theta) \geq (2/\pi)\theta$ , for  $\theta \in [0, \pi/2]$ , we deduce the following lower bound, for any  $l \geq 1$ :

$$\nu_l^\varepsilon \geq \pi \int_0^\pi \sin^2 \theta \overline{B}^\varepsilon(\theta) d\theta \geq \frac{4}{\pi} \int_0^{\frac{\pi}{2}} \theta^2 \overline{B}^\varepsilon(\theta) d\theta,$$

which gives the expected result, once condition (3.20) is fulfilled.

We can now derive the asymptotic behaviour when time goes to infinity.

**Theorem 3.7.** *There exists a positive constant  $\varepsilon_0$  such that:*

$$\sup_{\varepsilon \in [0, \varepsilon_0]} \|f^\varepsilon(t) - I(f_0)\|_{L^2(S^2)} \rightarrow 0, \quad \text{when } t \rightarrow +\infty. \quad (3.22)$$

More precisely, using the notations of Lemma 3.6, we have:

$$\sup_{\varepsilon \in [0, \varepsilon_0]} \|f^\varepsilon(t) - I(f_0)\|_{L^2(\mathbb{R}^3 \times S^2)} \leq \exp(-\nu^0 t) \|f_0\|_{L^2(\mathbb{R}^3 \times S^2)}. \quad (3.23)$$

**Proof.** In order to prove the uniform convergence result (3.22), we in fact only need a weaker form of the estimate given in Lemma 3.6. We here suppose that:

$$\text{for all } l \geq 1, \exists \nu_l^0 > 0, \varepsilon_l^0 > 0 \text{ such that for all } \varepsilon \in [0, \varepsilon_l^0] \quad \nu_l^\varepsilon \geq \nu_l^0. \quad (3.24)$$

Now by (3.19), we have, for all  $t > 0$ ,

$$\|g^\varepsilon(t)\|_{L^2(S^2)}^2 \leq \sum_{0 < l \leq L-1} \sum_{m=-l}^{m=l} a_{lm}^2 \exp(-2\nu_l^\varepsilon t) + \sum_{l \geq L} \sum_{m=-l}^{m=l} a_{lm}^2,$$

because the  $\nu_l^\varepsilon$  are all non negative. Now, since  $g_0 \in L^2(S^2)$ , the series  $\sum_{l>0} \sum_{m=-l}^{m=l} a_{lm}^2$  is convergent, so that the second sum in the above expression (which is independent of  $\varepsilon$ ) can be arbitrary small for  $L$  large enough. On the other hand, the first term is a finite sum of terms that all converge exponentially fast towards 0 when time goes to infinity, on account of the above assumption  $\nu_l^\varepsilon \geq \nu_l^0 > 0$ . More precisely, there exists a positive constant  $\varepsilon_0$  such that, for all  $\varepsilon \in [0, \varepsilon_0]$ , we have:

$$\sum_{0 < l \leq L-1} \sum_{m=-l}^{m=l} a_{lm}^2 \exp(-2\nu_l^\varepsilon t) \leq \sum_{0 < l \leq L-1} \sum_{m=-l}^{m=l} a_{lm}^2 \exp(-2\nu_l^0 t).$$

When  $t$  goes to infinity, this sum tends then to zero uniformly with respect to  $\varepsilon$ , which finally gives (3.22).

In the three dimensional case (which is the case under consideration here), we have a more precise result. In fact, thanks to the uniform lower bound (3.21), we get the estimate (3.23): the decrease, when time goes to infinity, is thus exponential.

**Remark 1.** In the two dimensional case, we obtain the same uniform convergence result (3.22). In case 1, the decrease, when time goes to infinity, is exponential, because we can find for the opposite of the eigenvalues a uniform lower bound with respect to  $\varepsilon$ , such as in Lemma 3.6. But in the Coulombian case, we could only manage to find the weaker estimate (3.24).  $\square$

#### 4. Some complements

##### 4.1. The dependence with respect to the modulus of the velocity variable

The distribution function in fact depends on the whole velocity variable  $v = \rho\omega$ ,  $\rho = |v|$ , although the differential operator only acts on the variable  $\omega \in S^2$ . If we recover the whole velocity variable, the Cauchy problems (3.6) and (3.11) respectively write [14]

$$\frac{\partial f^\varepsilon}{\partial t} = A^\varepsilon(f^\varepsilon) = |v|^\gamma Q^\varepsilon(f^\varepsilon), \quad f^\varepsilon(t=0) = f_0, \quad (4.1)$$

$$\frac{\partial f}{\partial t} = A(f) = |v|^\gamma \Delta_\omega f, \quad f(t=0) = f_0. \quad (4.2)$$

The power  $\gamma$  is directly connected to the type of intercatating potential between the particles. The case  $\gamma > 0$  corresponds to what is usually called “hard” potentials,  $\gamma < 0$  to “soft” potentials, while  $\gamma = 0$  is the particular case of Maxwellians molecules.

The unknown is  $f = f(t, \rho, \omega)$ , with  $\rho \in \mathbb{R}^+$ ,  $\omega \in S^2$ . We denote by  $L_W^2(0, \infty)$  and  $X$  the following weighted spaces

$$L_W^2(0, \infty) = \left\{ \psi = \psi(\rho), \int_0^{+\infty} \psi(\rho) \rho^2 d\rho < +\infty \right\},$$

$$X = L_W^2((0, \infty); L^2(S^2)) = \left\{ \psi = \psi(\rho, \omega), \int_{S^2} \int_0^{+\infty} \psi^2(\rho, \omega) \rho^2 d\rho d\omega < +\infty \right\}.$$

We split the square integrable initial data  $f_0 = f_0(\rho, \omega)$  ( $\rho = |v|$ ) in the following way

$$f_0(\rho, \omega) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}(\rho) Y_{lm}(\omega),$$

with  $\|f_0\|_X^2 = \int_0^\infty \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2(\rho) \rho^2 d\rho < +\infty$ . With a proof similar to that of Proposition 3.2, we easily get:

**Proposition 4.1.** *The function  $f^\varepsilon$  defined by*

$$f^\varepsilon(t, \rho, \omega) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}(\rho) \exp(-\nu_l^\varepsilon \rho^\gamma t) Y_{lm}(\omega) \quad (4.3)$$

*is a weak solution, in the space  $L^\infty(0, \infty; X)$ , of (4.1). Moreover, if the initial data is such that*

$$\int_{\mathbb{R}^3} \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm}^2(\rho) \nu_l \rho^2 d\rho < +\infty,$$

*then the above solution is the unique one such that*

$$f^\varepsilon \in L^\infty(0, \infty; L_W^2((0, \infty); H^1(S^2))) \quad \text{and} \quad Q^\varepsilon(f^\varepsilon) \in L^\infty(0, \infty; L_W^2((0, \infty); H^{-1}(S^2))).$$

Concerning the large time behaviour, we will not obtain in general (i.e., for any type of potentials) an exponential decrease when time goes to infinity. With the notations of Lemma 3.5, the function  $g^\varepsilon$  defined by

$$g^\varepsilon(t, x, \omega) = \sum_{l > 0} \sum_{m=-l}^{m=l} a_{lm}(\rho) \exp(-\nu_l^\varepsilon \rho^\gamma t) Y_{lm}(\omega)$$

satisfies the following Cauchy problem  $\partial g^\varepsilon / \partial t = A^\varepsilon(g^\varepsilon)$ ,  $g^\varepsilon(t = 0) = f_0 - I(f_0)$ . Thanks to the estimate (3.21), we have, for  $\varepsilon$  small enough,

$$\|g^\varepsilon(t)\|_{L^2(\mathbb{R}^3 \times S^2)}^2 \leq \int_{\mathbb{R}^3} \sum_{l > 0} \sum_{m=-l}^{m=l} a_{lm}^2(\rho) \exp(-2\nu^0 \rho^\gamma t) \rho^2 d\rho$$

so that this quantity tends to zero when time goes to infinity, but it tends exponentially fast towards zero only in the case  $\gamma = 0$ . For  $\gamma > 0$ , this exponential decrease holds outside any ball centered at zero with arbitrary small radius, while for  $\gamma < 0$ , it happens inside any ball centered at zero with arbitrary large radius. Let us introduce the  $\rho$  dependent function defined by:  $I(f_0)(\rho) = \frac{1}{4\pi} \int_{S^2} f_0(\rho, \omega) d\omega$ . Gathering all the cases, we have shown the following result:

**Theorem 4.2.** *There exists a positive constant  $\varepsilon_0$  such that:*

$$\sup_{\varepsilon \in [0, \varepsilon_0]} \|f^\varepsilon(t) - I(f_0)\|_X \rightarrow 0, \quad \text{when } t \rightarrow +\infty. \quad (4.4)$$



#### 4.2. The presence of an external magnetic field

The action of a magnetic field  $B$  on the spherical harmonics is well known [11]. In particular, when we take the direction of the vector  $B$  as the axis of the spherical coordinates, one has the following identity, for any  $\omega \in S^2$ :  $(\omega \wedge B) \cdot \nabla_\omega Y_{lm}(\omega) = -im\|B\|Y_{lm}(\omega)$ . This allows an explicit computation of the solution of the following Cauchy problem

$$\frac{\partial f^\varepsilon}{\partial t} + (\omega \wedge B) \cdot \nabla_\omega f^\varepsilon = Q^\varepsilon(f^\varepsilon), \quad f^\varepsilon(t=0) = f_0, \quad (4.5)$$

with  $f_0$  in  $L^2(S^2)$ . In fact, still using the expansion (3.7) of  $f_0$ , we have:

**Proposition 4.3.** *Let  $\nu_{l,m}^\varepsilon = \nu_l^\varepsilon - im\|B\|$ . The function  $f^\varepsilon$  defined by*

$$f^\varepsilon(t) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} \exp(-\nu_{l,m}^\varepsilon t) Y_{lm} \quad (4.6)$$

is a weak solution, in the space  $L^\infty((0, \infty); L^2(S^2))$ , of (4.5). Moreover, if the initial data is such that (3.9) holds, then the above solution is the unique one such that

$$f^\varepsilon \in L^\infty((0, \infty); H^1(S^2)) \quad \text{and} \quad Q^\varepsilon(f^\varepsilon) \in L^\infty((0, \infty); H^{-1}(S^2)).$$

In the same way, the function  $f$  defined by

$$f(t) = \sum_{l \geq 0} \sum_{m=-l}^{m=l} a_{lm} \exp(-\nu_{l,m} t) Y_{lm},$$

with  $\nu_{l,m} = \nu_l - im\|B\|$  satisfies the Cauchy problem

$$\frac{\partial f}{\partial t} + (\omega \wedge B) \cdot \nabla_\omega f = \Delta_\omega f, \quad f(t=0) = f_0, \quad (4.7)$$

with the same data  $f_0$ . Finally, we also keep the following exponential decay when time goes to infinity:

$$\sup_{\varepsilon \in [0, \varepsilon_0]} \|f^\varepsilon(t) - I(f_0)\|_{L^2(\mathbb{R}^3 \times S^2)} \leq \exp(-\nu^0 t) \|f_0\|_{L^2(\mathbb{R}^3 \times S^2)}. \quad (4.8)$$

#### Acknowledgements

The authors acknowledge support from the GdR SPARCH of the Centre National de la Recherche Scientifique (France) and from the TMR project ‘‘Asymptotic methods in kinetic theory’’ (TMR number: ERB FMRX CT97 0157), run by the European Community. They also would like to thank the referee for helpfull remarks.

## References

- [1] A.A. Arsenev and O.E. Buryak, On the connection between a solution of the Boltzmann equation and a solution of the Landau–Fokker–Planck equation, *Math. USSR Sbornik* **69**(2) (1991), 465–478.
- [2] T. Aubin, *Nonlinear Analysis on Manifolds. Monge–Ampère Equations*, Springer, 1982.
- [3] M. Bayet, J.L. Delcroix, and J.F. Denisse, Théorie cinétique des plasmas homogènes faiblement ionisés, II, *J. Phys. Rad.* **16** (1955), 274–280.
- [4] N. Ben Abdallah and P. Degond, On a hierarchy of macroscopic models for semiconductors, *J. Math. Phys.* **37**(7) (1996), 3306–3333.
- [5] H. Brézis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, 1992.
- [6] S. Cordier, B. Lucquin-Desreux and A. Sabry, Numerical approximation of the Vlasov–Fokker–Planck–Lorentz model, in: *ESAIM Proceedings of CEMRACS 1999*, to appear. Available at <http://www.emath.fr/proc/>
- [7] P. Degond and B. Lucquin-Desreux, The Fokker–Planck asymptotics of the Boltzmann collision operator in the Coulomb case, *Math. Models Methods Appl. Sci.* **2**(2) (1992), 167–182.
- [8] P. Degond and B. Lucquin-Desreux, The asymptotics of collision operators for two species of particles of disparate masses, *Math. Models Methods Appl. Sci.* **6**(3) (1996), 405–436.
- [9] P. Degond and P.F. Peyrard, Un modèle de collisions onde-particules en physique des plasmas; application à la dynamique des gaz, *C. R. Acad. Sci. Paris Sér. I* **323**(2) (1996), 209–214.
- [10] P. Degond, J. Lopez and P.F. Peyrard, On the Macroscopic dynamics induced by a model wave-particle collision operator, *Contin. Mech. Thermodyn.* **10**(3) (1998), 153–178.
- [11] J.L. Delcroix and A. Beers, *Introduction à la Physique des Plasmas*, Editions du CNRS, 1994.
- [12] L. Desvillettes, On asymptotics of the Boltzmann equation when the collisions become grazing, *Transp. Theory Statist. Phys.* **21**(3) (1992), 259–276.
- [13] *Handbook of Mathematical Functions*, eds M. Abramowitz and I.A. Stegun, Dover, New York, 1972.
- [14] B. Lucquin-Desreux, Diffusion of electrons by multicharged ions, *Math. Models Methods Appl. Sci.* **10**(3) (2000), 409–440.
- [15] J.C. Nedelec, Ondes acoustiques et électromagnétiques, Équations intégrales, cours de DEA, Édition 1996, École Polytechnique.
- [16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Applied Mathematical Sciences, Vol. 44, Springer, 1983.
- [17] G. Sansone, *Orthogonal Functions*, Pure and Applied Mathematics, Vol. 9, Interscience Publ., 1959.
- [18] G. Toscani, Entropy production and the rate of convergence to equilibrium for the Fokker–Planck equation, Preprint of Univ. Pavia, Dept. of Math., 1997.
- [19] C. Villani, On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations, *Arch. Rat. Mech. Anal.* **143**(3) (1998), 273–307.

# Numerical Analysis of the Isotropic Fokker–Planck–Landau Equation

C. Buet\* and S. Cordier†

\*Commissariat à l'Énergie Atomique, 91680 Bruyères-le-Châtel, France; and †MAPMO, UMR 6628, Université d'Orléans, B.P. 6759, 45072 Orléans, France  
E-mail: Stephane.Cordier@univ-orleans.fr, Christophe.Buet@cea.fr

Received February 1, 2000; revised October 15, 2001

---

The homogeneous Fokker–Planck–Landau equation is investigated for Coulombic potential and isotropic distribution function, i.e., when the distribution function depends only on time and on the modulus of the velocity. We derive a conservative and entropy decaying semidiscretized Landau equation for which we prove the existence of global in-time positive solutions. This scheme is not based on the so-called “Landau–Log” formulation of the operator and ensures the physically relevant long-time behavior of the solution. © 2002 Elsevier Science (USA)

*Key Words:* kinetic models; Fokker–Planck–Landau equation; system of ordinary differential equations; Cauchy problem; plasma physics; numerical schemes.

---

## INTRODUCTION

The Fokker–Planck–Landau equation (FPLE in the remainder) is commonly used in plasma physics when kinetic effects between charged particles under Coulomb interaction are studied.

The isotropic FPLE is generally used in the modeling of inertial controlled fusion. More precisely, it is used to describe electronic energy transport phenomena in a plasma produced by a laser. Under some conditions, it is well known that the fluid theory, for which the hydrodynamics equations are closed using the law for the thermal fluxes proposed by Spitzer–Harm [27], is not valid [14, 15]. A more accurate solution is to use a model based on the expansion of the FPLE in spherical harmonics, by only retaining the first two terms and the isotropic FPL operator is the leading order of the collision operator [14, 15]. Expansion of such ideas to the relativistic case can be found in [26] and references therein. In this paper, the author emphasize the care that must be taken in the numerical treatment of the classical FPLE. There are other applications, for example, in astrophysics where the FPLE is used for star cluster modeling [10, 11].

A conservative and entropy scheme for the spherical and homogeneous FPLe was first proposed in [3]. The authors give an upper bound for the time step to ensure the decay of the mathematical entropy without a complete proof of their assertion. Entropy decay is very important since it is physically relevant and seems to prevent oscillations, as shown in numerical examples in [6] and proved for the linear case in [5]. At the continuous level and for obvious physical reasons, the solution remains positive at any time, as proved by Desvillettes and Villani in the general 3D case [13]. Thus, the discretization must preserve this property and this does not appear clearly in [3]. See [5] for an example of a conservative discretization which does not preserve positivity for all positive initial data. Such schemes have been studied in [5] and references therein and these schemes rely on the so-called “Landau–Log” formulation of the operator, to be defined in the next section. In 1987, Berezin *et al.* announced that, in the isotropic case, the main properties can be achieved on the “nonlog form” of the FPLe [3]. One aim of this paper is to provide a proof of this assertion and to obtain some insight into the long-time behavior of the solutions for the semidiscretized FPLe.

Indeed, it has been proved recently in [6] that the existence of a unique, conservative, entropy decaying and global in-time solution holds for the semidiscretized FPLe. However, some questions were still open such as the long-time behavior of the semidiscretized or time discretized solution for which it is expected that the distribution function converges toward the discretized Maxwellian. We shall prove this property. Let us point out that this is the first result to our knowledge of the long-time behavior of the solution of the discretized FPLe.

This paper is organized as follows: in the first part, we recall briefly the continuous FPLe in the homogeneous and isotropic case, and we refer to [6] for more details. Then, we present the non-log discretization and we prove the properties of conservation, H-theorem, and trend to equilibrium. In the third section, we prove the existence of a global positive solution using a classical upper bound of the loss term as usual for the Boltzmann equation and that this solution tends to the Maxwellian. The last section is devoted to the time discretization approximation of the FPLe. For the time explicit discretization we prove that under a time step restriction involving the  $L^\infty$  of  $f$  or  $\varepsilon f$  the scheme is positive and entropic. We prove also that second-order time discretization defines a positive scheme. The derivation of an implicit scheme is also considered.

The isotropic FPLe could also be used to produce reference solutions to study numerical schemes proposed in the 3D velocity space [5, 7] or in the 2D axisymmetric case [16, 22] since no analytical solutions are known in the Coulombic case. The extension of this non-log form for the full tridimensional case, which is of physical interest for plasma physics, is not straightforward. Indeed, the simplest way to discretize the non-log form is not entropy decaying and provides a negative distribution function after arbitrary short time as shown in [5]. The study of the convergence of the constructed solutions when the mesh size  $\Delta\varepsilon$  goes to 0 is beyond the scope of this paper.

## 1. THE HOMOGENEOUS AND ISOTROPIC FPLe

We present the homogeneous nonlinear FPLe in the isotropic case where the distribution function  $f(\vec{x}, \vec{v}, t)$  depends only on the modulus of the velocity  $v = \|\vec{v}\|$  and on the time  $t$ ; i.e.,  $f(\vec{x}, \vec{v}, t) = f(v, t)$ . We shall consider  $f$  as a function of  $\varepsilon = v^2$ , which is the energy variable. For isotropic distribution functions, the FPLe for Coulombic potentials can be

written (see [3, 6] for more details), on a bounded domain  $\varepsilon \in [0, \varepsilon_0]$ , in the form

$$\frac{\partial f}{\partial t} = \frac{1}{\sqrt{\varepsilon}} \frac{\partial}{\partial \varepsilon} \int_0^{\varepsilon_0} f f' \left( \frac{\partial}{\partial \varepsilon} \ln f - \frac{\partial}{\partial \varepsilon'} \ln f' \right) k(\varepsilon, \varepsilon') d\varepsilon', \quad (1.1)$$

where we define  $k = k(\varepsilon, \varepsilon') = \inf(\varepsilon^{3/2}, (\varepsilon')^{3/2})$  and  $f$  (resp.  $f'$ ) denotes  $f(\varepsilon, t)$  (resp.  $f(\varepsilon', t)$ ) to simplify the notations.

This operator can be equivalently written in the following weak form (let  $\phi(\varepsilon)$  be any function time independent test (smooth and decaying)) by integrating (1.1) by parts,

$$\int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon} d\varepsilon = -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} f f' \left( \frac{\partial}{\partial \varepsilon} \phi - \frac{\partial}{\partial \varepsilon'} \phi' \right) \left( \frac{\partial}{\partial \varepsilon} \ln f - \frac{\partial}{\partial \varepsilon'} \ln f' \right) k d\varepsilon' d\varepsilon, \quad (1.2)$$

where we assume that  $\frac{\partial}{\partial \varepsilon} \phi(\varepsilon_0) = 0$  and also that  $k(0, \varepsilon) = 0$  to get rid of the boundary terms in the integration by parts. Let us recall that FPLe satisfies the conservation of mass (resp. energy) (by choosing  $\phi = 1$  (resp.  $\phi = \varepsilon$ ) in (1.2))

$$\rho = \int_0^{\varepsilon_0} f(\varepsilon) \sqrt{\varepsilon} d\varepsilon, \quad \rho E = \int_0^{\varepsilon_0} f(\varepsilon) \varepsilon^{3/2} d\varepsilon. \quad (1.3)$$

The mathematical (or negative) entropy  $H$ , defined by

$$H = \int_0^{\varepsilon_0} f(\varepsilon) \ln(f(\varepsilon)) d\varepsilon, \quad (1.4)$$

is decreasing in time, by letting  $\phi = \ln(f)$  in the weak formulation of FPLe and using the mass conservation, and satisfies the H-theorem  $\partial_t H = 0 \Leftrightarrow f = \exp(-A\varepsilon + B)$ . Note that the FPLe can be equivalently written in the so-called non-log weak form

$$\int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon} d\varepsilon = -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} \left( \frac{\partial}{\partial \varepsilon} \phi - \frac{\partial}{\partial \varepsilon'} \phi' \right) \left( f' \frac{\partial}{\partial \varepsilon} f - f \frac{\partial}{\partial \varepsilon'} f' \right) k d\varepsilon' d\varepsilon. \quad (1.5)$$

In previous works [12], the discretization was performed on the log form (1.2) of the FPLe to prove the decay of entropy. In this paper, we prove that this property can be achieved on a discretization on the non-log form (1.5). Note that, at the continuous level, the two formulations are equivalent but this is not the case after the discretizations we shall now present.

## 2. THE SEMIDISCRETIZED PROBLEM

The discretization of the FPLe follows exactly the same lines as the discretization described in [6]. We briefly recall the notations which are used in the remainder of the paper.

### 2.1. Discretization in Velocity Space

Let us introduce the uniform discretization  $f_i = f(\varepsilon_i)$ , where  $(\varepsilon_i)_{i=1\dots N} = (i-1)\Delta\varepsilon$  such that  $\varepsilon_N = \varepsilon_0$ . The  $\varepsilon$ -derivatives are approximated according to the simplest choice of

finite difference operator; namely, we define for any discretized function  $(\phi_i)_{i=1\dots N}$   $D\phi_i = (\phi_{i+1} - \phi_i)/\Delta\varepsilon$ ,  $i = 1 \dots N - 1$ . We note  $\varepsilon_{i+1/2} = (\varepsilon_{i+1} + \varepsilon_i)/2$  and  $v_{i+1/2}$  as the mean value of the velocity on  $[\varepsilon_i, \varepsilon_{i+1}]$ ; i.e.,  $v_{i+1/2} = \frac{1}{\Delta\varepsilon} \int_{\varepsilon_i}^{\varepsilon_{i+1}} \sqrt{\varepsilon} d\varepsilon = \frac{2}{3\Delta\varepsilon} (\varepsilon_{i+1}^{3/2} - \varepsilon_i^{3/2})$ . Let us consider first the discretization of the expression  $\int_0^{\varepsilon_0} \phi(\varepsilon) \sqrt{\varepsilon} d\varepsilon$  for any function  $\phi$ . By using the trapezoidal quadrature formula with respect to the measure  $\sqrt{\varepsilon} d\varepsilon$ , we approximate it by

$$\int_0^{\varepsilon_0} \phi(\varepsilon) \sqrt{\varepsilon} d\varepsilon = \sum_{i=1}^{N-1} \int_{\varepsilon_i}^{\varepsilon_{i+1}} \phi(\varepsilon) \sqrt{\varepsilon} d\varepsilon \simeq \sum_{i=1}^{N-1} \frac{1}{2} (\phi_i + \phi_{i+1}) v_{i+1/2} \Delta\varepsilon \stackrel{\text{def}}{=} \sum_{i=1}^N c_i \phi_i, \quad (2.1)$$

with  $c_i$  defined by the above formula such that  $c_1 = v_{3/2} \Delta\varepsilon = \frac{1}{3} \varepsilon_2^{3/2}$ ,  $c_i = \frac{1}{2} (v_{i+1/2} \Delta\varepsilon + v_{i-1/2} \Delta\varepsilon) = \frac{1}{3} (\varepsilon_{i+1}^{3/2} - \varepsilon_{i-1}^{3/2})$ , for  $i = 2 \dots N - 1$ , and  $c_N = v_{N-1/2} \Delta\varepsilon = \frac{1}{3} (\varepsilon_N^{3/2} - \varepsilon_{N-1}^{3/2})$ . Once applied to the left-hand side of (1.5) with  $\frac{\partial f}{\partial t} \phi$ , we obtain the discretization of

$$\int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon} d\varepsilon \quad \text{as} \quad \sum_{i=1}^N c_i \frac{\partial f_i}{\partial t} \phi_i.$$

We now turn to the discretization of the right-hand side of (1.5),

$$(\text{r.h.s.}) = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \int_{\varepsilon_i}^{\varepsilon_{i+1}} \int_{\varepsilon_j}^{\varepsilon_{j+1}} f f' \left( \frac{\partial}{\partial \varepsilon} \phi - \frac{\partial}{\partial \varepsilon'} \phi' \right) \left( \frac{\partial}{\partial \varepsilon} \ln f - \frac{\partial}{\partial \varepsilon'} \ln f' \right) k d\varepsilon' d\varepsilon. \quad (2.2)$$

Using for each integral in (2.2) a midpoint quadrature formula, we approximate (2.2) by

$$-\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} \Delta\varepsilon \Delta\varepsilon (D\phi_i - D\phi_j) (D(\ln f)_i - D(\ln f)_j), \quad (2.3)$$

with  $k_{i,j} = k(\varepsilon_{i+1/2}, \varepsilon_{j+1/2})$  and the terms  $g_{i,j}$  standing for an approximation of the distribution function product  $f_i f_j$  at the center of the interval  $[\varepsilon_i, \varepsilon_{i+1}] \times [\varepsilon_j, \varepsilon_{j+1}]$ , which are now to be defined.

## 2.2. Choice of the Functions $g_{i,j}$

In [3], the terms  $g_{i,j}$  are of the form  $g_i g_j$ , where the  $g_i$  are taken as an arithmetic mean of  $f_i$  and  $f_{i+1}$ . This yields a discrete model for which it cannot be proved that the distribution function remains positive, as it must be. In [6], we consider the harmonic average; that is,  $(2f_i f_{i+1})/(f_i + f_{i+1})$ . This approximation was already used in [5], for the linear and 3D nonlinear cases of the Fokker–Planck–Landau equation and the resulting discrete model for which the existence of a global positive solution can be proved using the estimate  $g_i \leq 2 \min(f_i, f_{i+1})$ . In this paper, we consider the expression for  $g_{i,j}$

$$g_{i,j} \stackrel{\text{def}}{=} \frac{f_i Df_j - f_j Df_i}{D(\ln f)_j - D(\ln f)_i}, \quad \text{if } D(\ln f)_j \neq D(\ln f)_i, \quad (2.4)$$

and  $g_{i,j} = f_i f_j$  when  $D(\ln f)_j = D(\ln f)_i$  but the corresponding contribution in the sum vanishes. Indeed, for a uniform grid and only in this case the above expression can be simplified into

$$g_{i,j} = \frac{f_i f_{j+1} - f_j f_{i+1}}{\ln(f_{j+1} f_i) - \ln(f_{i+1} f_j)}.$$

Using the mean value theorem for the  $\ln$  function, we have

$$\min(f_i f_{j+1}, f_j f_{i+1}) \leq g_{i,j} \leq \max(f_i f_{j+1}, f_j f_{i+1}).$$

Note that this approximation is of second order, for a uniform grid. Using this expression of  $g_{i,j}$ , (2.3) becomes

$$-\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} k_{i,j} \Delta \varepsilon \Delta \varepsilon (D\phi_i - D\phi_j) (f_j Df_i - f_i Df_j). \quad (2.5)$$

One recovers the scheme proposed in [3], which can be obtained directly from the non-log form (1.5) of the FPLe. We prefer to derive it from the log form because this helps to check easily the main properties of the operator, conservation, H-theorem, which were given without any proof in [3].

Note that  $D\phi_i$  is also a second-order approximation of the derivative at the center of the cell  $[\varepsilon_i, \varepsilon_{i+1}]$ . Thus, if there exists a smooth solution of FPLe, the discretization error will be of second order. This is some kind of consistency result for the scheme.

Note that such average (2.4), in the case the uniform grid and for the linear Fokker-Planck equation, has already been used in [8] and is called the entropic average.

### 2.3. The System of ODE Associated to the Semidiscretized FLPE

From (2.1) and (2.5), the weak semidiscretized formulation of FPLe reads

$$\sum_{i=1}^N c_i \frac{\partial f_i}{\partial t} \phi_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} \Delta \varepsilon \Delta \varepsilon (D\phi_i - D\phi_j) (D(\ln f)_i - D(\ln f)_j), \quad (2.6)$$

or equivalently, using the definition of  $g_{i,j}$ ,

$$\sum_{i=1}^N c_i \frac{\partial f_i}{\partial t} \phi_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} k_{i,j} \Delta \varepsilon \Delta \varepsilon (D\phi_i - D\phi_j) (f_j Df_i - f_i Df_j).$$

By factorizing the terms  $\phi_i$  in (2.6) as explained in [6], we obtain a system of ordinary differential equations of the form

$$\frac{df_i}{dt} = \text{FP}_i, \quad i = 1 \dots N, \quad (2.7)$$

with  $\text{FP}_1 = p_1/c_1$ ,  $\text{FP}_i = (p_i - p_{i-1})/c_i$ , for  $i = 2 \dots N-1$ , and  $\text{FP}_N = -p_{N-1}/c_{N-1}$ , and for all  $i = 1 \dots N-1$ ,

$$p_i \stackrel{\text{def}}{=} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} D_{i,j} \Delta \varepsilon, \quad (2.8)$$

where  $D_{i,j}$  stands for  $(D(\ln f)_i - D(\ln f)_j)$ . One can also write the equivalent non-log form using the definition of  $g_{i,j}$  (2.4).

$$p_i = \sum_{j=1}^{N-1} k_{i,j}(f_j f_{i+1} - f_i f_{j+1}),$$

and system (2.7) becomes, for  $i = 2 \dots N-1$ ,

$$\frac{df_i}{dt} = \frac{1}{c_i} \left( \sum_{j=1}^{n-1} k_{i,j} f_j f_{i+1} + \sum_{j=1}^{n-1} k_{i-1,j} f_{j+1} f_{i-1} - \left( \sum_{j=1}^{n-1} k_{i,j} f_{j+1} + k_{i-1,j} f_j \right) f_i \right)$$

and can be written in the form of gain and loss as usual for the Boltzmann type operator,

$$\frac{df_i}{dt} = K_i(f) - L_i(f) f_i.$$

Note the three-diagonal structure of this nonlinear system of ordinary differential equations. Let us end the description of the discrete FPLe by a useful result for the following sections:

**LEMMA 2.1.** *If we set  $L_1 = \sup_i (\Delta \varepsilon \sqrt{\varepsilon_{i+1/2}}) / c_i$  and  $L_2 = \sup_i (\Delta \varepsilon \sqrt{\varepsilon_{i+1/2}}) / c_{i+1}$  and if  $N$  is sufficiently large then  $L_1$  and  $L_2$  are uniformly bounded in  $N$ ; that is,*

$$L_1 \leq \frac{3}{\sqrt{2}} \quad \text{and} \quad L_2 \leq \frac{3}{\sqrt{2}}.$$

The basic proof relies on the explicit definition of the sequences  $\varepsilon_{i+1/2}$  and  $c_i$ .

## 2.4. Properties of the Semidiscretized FPLe

One can now check the conservation of mass and energy (1.3) at the discretized level,

$$\rho = \sum_{j=1}^N c_j f_j \text{ (mass)}, \quad \rho E = \sum_{j=1}^N c_j f_j \varepsilon_j \text{ (energy)},$$

where the sequence  $c_i$  defined by (2.1) corresponds to the measure associated with the choice of  $\varepsilon_i$ . Let us assume for the moment that there exists a (vector) solution  $f(t)$  of system (2.7) that is global, strictly positive, and smooth in time. The two quantities defined above for this solution  $f$  are conserved through the evolution of the system by taking  $\phi_i = 1$  and  $\phi_i = \varepsilon_i$  in (2.6). Moreover, the discretized entropy defined by

$$H = H(f) \stackrel{\text{def}}{=} \sum_{j=1}^N c_j f_j \ln(f_j), \tag{2.9}$$

decays in time. This can be easily checked using the weak discretized formulation (2.6) with test function  $\phi_i = \ln(f_i)$ ,

$$\frac{dH}{dt} = \sum_{i=1}^N c_i \frac{df_i}{dt} \ln(f_i) + \sum_{i=1}^N c_i \frac{d \ln(f_i)}{dt} f_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} \Delta \varepsilon \Delta \varepsilon D_{i,j}^2 \leq 0,$$



since the second sum vanishes using mass conservation. Note that the property can also be verified directly on the non-log form using the  $(x - y)(\ln x - \ln y) \geq 0$  property (as usual for the Boltzmann equation). Indeed, one has

$$\frac{dH}{dt} = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} k_{i,j} (f_{i+1} f_j - f_{j+1} f_i) (\ln(f_{i+1} f_j) - \ln(f_{i+1} f_i)). \quad (2.10)$$

We shall prove that  $\frac{dH}{dt} = 0$  is equivalent to  $f_i = M_i$ , where  $M_i$  is the discrete Maxwellian

$$M_i = n_0 \exp(\alpha \varepsilon_i), \quad (2.11)$$

where  $n_0$  and  $\alpha$  are such that mass and energy are the same as for the initial data,

$$\rho = n_0 \sum_{j=1}^N c_j \exp(\alpha \varepsilon_j), \quad \rho E = n_0 \sum_{j=1}^N c_j \varepsilon_j \exp(\alpha \varepsilon_j).$$

This system of two equations ( $\rho, E$  being the data,  $\alpha, n_0$  the unknowns) can be reduced to the following equation for the parameter  $\alpha$

$$E = \frac{\sum_{j=1}^N c_j \varepsilon_j \exp(\alpha \varepsilon_j)}{\sum_{j=1}^N c_j \exp(\alpha \varepsilon_j)}.$$

It is proved that this defines a unique  $\alpha$ , which is negative when  $\varepsilon_0$  is large enough. The existence of such an equilibrium state is discussed in Appendix A.

The converse implication ( $f = M \Rightarrow dH/dt = 0$ ) is obvious, since all the terms in the sum vanish. We can prove the direct implication easily, which is usually not easy to prove for other collision operators. Indeed, the term in the sum (2.6) vanishes for any discrete test function if and only if

$$f_{j+1} f_i = f_{i+1} f_j, \quad \forall (i, j) \in [1, N-1]^2.$$

Therefore, the ratio  $f_{i+1}/f_i$  is constant and thus the sequence  $f_i$  is geometric, i.e., equal to  $M_i$  given by (2.11).

## 2.5. Existence of a Global Solution for the Semidiscretized FPLe

The existence of a positive global-in-time solution for this system is based on the upper bound of the loss term like in the proof for the Boltzmann equation [17]. We have the following upper bound for the loss term  $K_i(f)$ .

LEMMA 2.2.

$$\sup_i K_i(f) \leq \frac{9\rho(f)}{(\Delta\varepsilon)^2} \sqrt{3T + \frac{\Delta\varepsilon}{2}}. \quad (2.12)$$

*Proof.* Let us first examine the situation for the interior points, that is, for  $i = 2, \dots, N - 1$ . In this case, let us recall that the gain terms are

$$K_i(f) = \frac{1}{\Delta \varepsilon c_i} \left( \sum_{j=2}^N k_{i,j-1} f_j \Delta \varepsilon + \sum_{j=1}^{N-1} k_{i-1,j} f_j \Delta \varepsilon \right)$$

and we recall that  $k_{ij} = \min(\varepsilon_{i+1/2}^{3/2}, \varepsilon_{j+1/2}^{3/2})$ . Using the inequality  $\min(a^{3/2}, b^{3/2}) \leq \sqrt{a} \min(a, b)$ , the fact that  $k_{ij}$  is an increasing sequence in  $i$  and  $j$ , Lemma 2.1, and the Cauchy–Schwartz inequality, we have

$$\begin{aligned} K_i(f) &\leq \frac{2}{\Delta \varepsilon c_i} \sum_{j=1}^N k_{i,j} f_j \Delta \varepsilon \leq \frac{2\sqrt{\varepsilon_{i+1/2}}}{\Delta \varepsilon c_i} \sum_{j=1}^N \varepsilon_{j+1/2} f_j \Delta \varepsilon \\ &\leq \frac{2\sqrt{\varepsilon_{i+1/2}}}{\Delta \varepsilon c_i} \sum_{j=1}^N \sqrt{\varepsilon_{j+1/2}} f_j c_j \frac{\Delta \varepsilon \sqrt{\varepsilon_{j+1/2}}}{c_j} \leq \frac{6\sqrt{2}\rho(f)}{(\Delta \varepsilon)^2} \sqrt{3T + \frac{\Delta \varepsilon}{2}}. \end{aligned}$$

Let us examine now the situation at the boundary. For  $i = 1$  we have

$$K_1(f) = \frac{1}{\Delta \varepsilon c_1} \sum_{j=2}^N k_{1,j-1} f_j \Delta \varepsilon \leq \frac{\sqrt{\varepsilon_{3/2}}}{\Delta \varepsilon c_1} \sum_{j=1}^N \varepsilon_{j+1/2} f_j \Delta \varepsilon \leq \frac{9}{2(\Delta \varepsilon)^2} \rho(f) \sqrt{E(f) + \frac{\Delta \varepsilon}{2}}.$$

The case  $i = N$  gives the same upper bound. ■

We define

$$\tau = \frac{(\Delta \varepsilon)^2}{9\rho(f)\sqrt{E(f) + \frac{\Delta \varepsilon}{2}}}. \quad (2.13)$$

**PROPOSITION 2.3.** *The Cauchy problem for the differential equation (2.7) with strictly positive initial data admits a unique positive entropic global in-time solution.*

*Proof.* The existence and uniqueness of the solution for short times are obtained using the classical Cauchy–Lipschitz theorem. If we prove that the solution remains positive for any time, then mass conservation gives an upper bound for the weights. Therefore, the solutions cannot blow up in finite time and we have a positive solution for arbitrary long times. We shall use the upper bound of the loss terms  $K_i$  of Lemma 2.2.

Equation (2.12) implies that for all  $i$ , we have

$$\frac{df_i}{dt} \geq -\frac{1}{\tau} f_i \Rightarrow f_i(t) \geq f_i(t=0) \exp(-t/\tau).$$

Such inequality implies that the weights  $f_i$  cannot vanish in finite time. ■

Note that using an explicit time discretization, this estimate provides a time step limitation for positivity,

$$f_i^{t+\Delta t} = f_i^t + \Delta t F P_i^t \geq f_i^t (1 - \Delta t/\tau) > 0,$$

if the time step is such that  $\Delta t < \tau$ . We prove the following result concerning the long-time behavior.

LEMMA 2.4. *For all strictly positive initial conditions, the solution of (2.7) verifies*

$$\forall i, \quad \lim_{t \rightarrow \infty} f_i = M_i.$$

*Proof.* We first prove that

$$H(f) \geq H(M),$$

where  $M$  is the Maxwellian with the same moment as  $f$ . This is sometimes called the Gibbs lemma and the proof relies on the Jensen inequality. At the discrete level, we have

$$\begin{aligned} H(f \parallel M) &= H(f) - H(M) = \sum_{i=1}^N c_i f_i \ln(f_i) - \sum_{i=1}^N c_i M_i \ln(M_i) \\ &= \sum_{i=1}^N c_i f_i \ln(f_i/M_i) + \sum_{i=1}^N c_i (f_i - M_i) \ln(M_i); \end{aligned}$$

the second sum vanishes using conservation laws and the first sum is positive using the convexity of the function  $x \mapsto x \ln(x)$  and the Jensen inequality. Thus,  $H(f \parallel M)(t)$  is decreasing in time and positive. It converges to some value  $H_\infty$ .

We have shown that  $H_\infty = 0$  necessarily by contradiction. Else, there exists an increasing sequence  $t_k$  such that  $t_k \rightarrow \infty$  when  $k \rightarrow \infty$  and

$$f_i(t_k) \rightarrow f_i^\infty, \quad \frac{dH(f \parallel M)(t_k)}{dt} \rightarrow 0;$$

indeed the weights lie in a compact set and are in finite number. It has to be proved that  $\frac{dH(f \parallel M)}{dt}$  is continuous in time. This is clear since this is a functional defined using smooth functions from the weights  $f_i$  which are at least  $C^1$  functions of the time and we have shown that  $\frac{dH}{dt} = 0 \Rightarrow f = M$ .

Thus,  $f_i^\infty = M_i$ . Using the monotonicity of  $H$ , one has the convergence of  $H(f \parallel M)$  when  $t \rightarrow \infty$  (not only for a sequence of increasing time). Then, we use the Czizár-Kullback inequality (see [18])

$$\|f - M\|_{L^1}^2 \leq 2H(f \parallel M);$$

i.e.,

$$\left( \sum_{i=1}^N c_i \|f_i - M_i\| \right)^2 \leq 2 \sum_{i=1}^N c_i f_i \ln(f_i/M_i).$$

This latter inequality proves that  $f_i \rightarrow M_i$ . ■

This also provides a uniform in-time, strictly positive lower bound for the  $f_i$ . Indeed, there exists  $t_*$  such that  $\forall t > t_*$ ,  $H(f \parallel M) \leq \min_i M_i^2/4$ . Moreover,  $f_i$  is strictly positive on the interval  $[0, t_*]$ , and thus, there exists a minimum  $f^{\min}$  for the finite number of  $f_i$  on  $[0, t_*]$ . The obtained lower bound is not explicit.

### 3. TIME DISCRETIZATION

In this section, we shall investigate different methods for discretizing in time the system of ordinary differential equations (2.7)–(2.8). First, we shall consider explicit schemes and afterward implicit schemes. In both cases we discuss the properties and the cost of the scheme.

#### 3.1. First-Order Explicit Scheme

Let us now consider explicit schemes.

First, note that the system (2.7) that determines the evolution of the distribution function  $f$  can be written as a sum of four-velocities, a so-called Broadwell system (see [17]). Moreover, the entropy function  $x \ln(x)$  is convex and decays provided it decays for each Broadwell system. We shall take advantage of this particular structure to describe the first scheme.

More precisely, system (2.7) can be written in the form

$$\frac{df}{dt} = \sum_{i,j} B_{i,j}(f) + \sum_{i,j} \tilde{B}_{i,j}(f),$$

with  $(B_{i,j}(f))_k = 0$  if  $k \notin \{i, i+1, j, j+1\}$ . The sum  $\tilde{B}$  will have exactly the same structure for the index  $\{i-1, i, j, j-1\}$ . Thus, each term  $B_{i,j}$  only modifies four components of  $f$ . We denote  $f_1, f_2, f_3$ , and  $f_4$  (or for the coefficients  $c_k$ ) for  $f_i, f_{i+1}, f_j$ , and  $f_{j+1}$ , respectively. The evolution of these functions due to the term  $B_{i,j}$  is given by

$$\begin{aligned} \frac{df_1}{dt} &= \frac{C}{c_1}(f_2 f_3 - f_1 f_4), \\ \frac{df_2}{dt} &= \frac{-C}{c_2}(f_2 f_3 - f_1 f_4), \\ \frac{df_3}{dt} &= \frac{-C}{c_3}(f_2 f_3 - f_1 f_4), \\ \frac{df_4}{dt} &= \frac{C}{c_4}(f_2 f_3 - f_1 f_4), \end{aligned} \tag{3.1}$$

if  $i \neq j+1$  or

$$\begin{aligned} \frac{df_1}{dt} &= \frac{2C}{c_1}(f_2 f_3 - f_1 f_1), \\ \frac{df_2}{dt} &= \frac{-C}{c_2}(f_2 f_3 - f_1 f_1), \\ \frac{df_3}{dt} &= \frac{-C}{c_3}(f_2 f_3 - f_1 f_1), \end{aligned} \tag{3.2}$$

if  $i = j+1$  and with  $C = k_{i,j}$  (or  $C = k_{i-1,j}$  for the term  $\tilde{B}_{i,j}$ ). Note that the evolution of one particular index  $i_0$  involves  $N$  generalized Broadwell systems. Note that for the special case  $j+1 = i$ , system (3.2) is of the form (3.1) with  $c_4 = c_1$  and  $f_1(0) = f_4(0)$ . The exact solution for the Cauchy problem associated to such a Broadwell system (3.1) can

be computed explicitly as

$$\begin{aligned} f_1(t) &= f_1^0 + F(t)/c_1, & f_2(t) &= f_2^0 - F(t)/c_2, \\ f_3(t) &= f_3^0 - F(t)/c_3, & f_4(t) &= f_4^0 + F(t)/c_4, \end{aligned} \quad (3.3)$$

where  $F$  is given by

$$F(t) = D - \frac{D \exp(-C\sqrt{\Delta}t)}{1 + \tilde{D}(1 - \exp(-C\sqrt{\Delta}t))},$$

with

$$\begin{aligned} A &= \frac{f_1^0}{c_4} + \frac{f_4^0}{c_1} + \frac{f_3^0}{c_2} + \frac{f_2^0}{c_3}, & B &= \left( \frac{1}{c_2 c_3} - \frac{1}{c_1 c_4} \right) (f_2^0 f_3^0 - f_1^0 f_4^0) \\ \Delta &= A^2 - 4B, & D &= \frac{2(f_2^0 f_3^0 - f_1^0 f_4^0)}{A + \sqrt{\Delta}}, & \tilde{D} &= \left( \frac{1}{c_2 c_3} - \frac{1}{c_1 c_4} \right) D / \Delta. \end{aligned}$$

The (partial) entropy  $\sum_{k=1}^4 c_k \ln(f_k(t)) f_k(t)$  decays in time. The solution remains always positive.

Let us now consider the full coupled system as a linear system, which is obviously not the case, and take a superposition of the solution of the elementary Broadwell system. More precisely, let us define  $f_{i,j}$  as the exact solution, defined previously, of the Cauchy problem for the system

$$\frac{df_{i,j}}{dt} = 2N^2 B_{i,j}, \quad f_{i,j}(t=0) = f^0,$$

and  $\tilde{f}_{i,j}$  as the solution for the terms  $\tilde{B}_{i,j}$  with the same initial data. Then

$$f = \frac{1}{2N^2} \sum_{i,j} f_{i,j} + \tilde{f}_{i,j}$$

is a first-order approximation of the solution of (2.7) which preserves positivity, decays the entropy, since it decays the entropy for any Broadwell system, and conserves mass and energy for all time. The cost of this method is  $O(N^2)$  for one time step.

The second method is based on a complete explicit scheme,

$$f^{n+1} = f^n + \Delta t F P(f^n).$$

First, let us exhibit a condition such that the scheme remains positive. As explained in the proof of Theorem 1, this property holds provided that  $\Delta t < \tau$ , where  $\tau$  is defined by (2.13) and depends only on  $\rho E$ ,  $\rho$ , and  $c_1 = \min_i c_i$ .

The main advantage of this method is that the cost is linear. Indeed, due to the definition of  $k_{i,j} = \min(\varepsilon_i^{3/2}, \varepsilon_j^{3/2})$ , the evaluation of the coefficients of the matrix  $D$  defined before can be performed in  $O(N)$  operations as explained in [6].

For the entropy, we shall use the same ideas as those in [5]. We have

$$(f + \Delta f) \ln(f + \Delta f) \leq f \ln(f) + \Delta f \ln f + \Delta f + (\Delta f)^2,$$

with  $f = f_i^n$  and  $\Delta f = \Delta t F P(f^n)_i = \Delta t F P_i^n = \Delta t \frac{1}{c_i} (p_i^n - p_{i-1}^n)$ . Adding these inequalities, and using the conservation of mass and definition of the discretized entropy, we obtain

$$H^{n+1} \leq H^n + \Delta t \sum_i F P_i^n \ln(f_i^n) + (\Delta t)^2 \sum_i (F P_i^n)^2 / f_i^n.$$

We take

$$\Delta t = \min \left( \tau, \frac{-\sum_i F P_i^n \ln(f_i^n)}{\sum_i (F P_i^n)^2 / f_i^n} \right).$$

Another estimate can be obtained from the fact that the system is a sum of generalized Broadwell systems (3.1). In fact, we can split the sum over all the  $O(N^2)$  possible quadruplets in  $\tilde{N}$  subsets such that each integer appears at most once in a given subset. See Appendix B for such a partition. Then the system reads

$$\frac{df}{dt} = \sum_{p=1}^{\tilde{N}} \sum_{(i,j) \in \Theta_p} B_{i,j}(f),$$

where  $\Theta_p$  corresponds to one subset (see Appendix B) and  $B_{i,j}$  is one of the generalized Broadwell systems. We use the same splitting ideas as before, with the exact solution replaced by the explicit scheme. Then the explicit scheme can be written

$$f^{n+1} - f^n = \Delta t \sum_{p=1}^{\tilde{N}} \sum_{(i,j) \in \Theta_p} B_{i,j}(f^n).$$

Define

$$f^p = f^n + \Delta t \tilde{N} \sum_{(i,j) \in \Theta_p} B_{i,j}(f^n).$$

We have  $f^{n+1} = \frac{1}{\tilde{N}} \sum_{p=1}^{\tilde{N}} f^p$ . Since the entropy is convex, we have

$$H^{n+1} \leq \frac{1}{\tilde{N}} \sum_{p=1}^{\tilde{N}} \sum_i c_i f_i^p \ln(f_i^p).$$

For any fixed  $p$ , the generalized Broadwell systems involved in  $\Theta_p$  are distinct. Thus, the entropy decays provided that it decays for any system in  $\Theta_p$  where  $C$  is multiplied by  $\tilde{N}$ . It remains to compute the time step  $\Delta t$  such that the explicit scheme for such a generalized Broadwell system decays the entropy.

LEMMA 3.1. *There exists a constant  $C$  such that for each Broadwell model the time explicit scheme with time step  $t$  is positive and entropic under the time step restriction*

$$t \leq \frac{C(\Delta \varepsilon)^2}{\sup_i (\varepsilon_i f_i)}.$$

*Proof.* We consider two indices  $i$  and  $j$  such that  $i, i+1, j, j+1 \in \{1, \dots, N\}$  are distinct and the Broadwell model associated with these points coming from the splitting of the full system in  $\tilde{N}$  operators of an independent Broadwell model is

$$\frac{df_i}{dt} = -\frac{C_{ij}}{c_i}Q, \quad \frac{df_{j+1}}{dt} = -\frac{C_{ij}}{c_{j+1}}Q, \quad \frac{df_{i+1}}{dt} = \frac{C_{ij}}{c_{i+1}}Q, \quad \frac{df_j}{dt} = \frac{C_{ij}}{c_i}Q,$$

with  $Q = f_{j+1}f_i - f_{i+1}f_j$  and  $C_{ij} = \tilde{N} \min(\varepsilon_{i+1/2}^{3/2}, \varepsilon_{j+1/2}^{3/2})$ . A time explicit discretization of such a differential equation reads

$$\begin{aligned} f_i(t) &= g_i - t \frac{C_{ij}}{c_i}Q, & f_{j+1}(t) &= g_{j+1} - t \frac{C_{ij}}{c_{j+1}}Q, \\ f_{i+1}(t) &= g_{i+1} + t \frac{C_{ij}}{c_{i+1}}Q, & f_j(t) &= g_j + t \frac{C_{ij}}{c_i}Q, \end{aligned}$$

where  $g_i, g_j, g_{i+1}, g_{j+1}$  are the initial conditions and indeed  $Q = g_{i+1}g_j - g_{j+1}g_i$ . Using Lemma 2.1 and the bound for  $\tilde{N}$  (see Appendix B) such a scheme is positive provided that  $t \leq \tau_1 = \frac{\sqrt{2}\Delta\varepsilon^2}{3C_N\varepsilon_0 \sup_{k=i+1, j, j+1}(\varepsilon_{k+1/2}g_k)}$ .

The numerical entropy associated with this scheme is

$$\begin{aligned} H(t) &= c_i f_i(t) \log(f_i(t)) + c_{i+1} f_{i+1}(t) \log(f_{i+1}(t)) \\ &\quad + c_j f_j(t) \log(f_j(t)) + c_{j+1} f_{j+1}(t) \log(f_{j+1}(t)). \end{aligned}$$

We want to choose  $t$  such that  $H(t) \leq H(0)$ . Now for the sake of simplicity we set  $C = C_{ij}$ . One can easily verify that

$$H'(t) = CQ \log \left( \frac{g_j + \frac{Ct}{c_j}Qg_{i+1} + \frac{Ct}{c_{i+1}}Q}{g_i + \frac{Ct}{c_i}Qg_{j+1} + \frac{Ct}{c_{j+1}}Q} \right).$$

By construction, we have  $H'(0) \leq 0$ . We exclude the case  $H'(0) = 0$ , which corresponds to  $Q = 0$  and for which indeed for all time  $H(t) \leq H(0)$ , so that we assume  $Q \neq 0$ .  $H(t)$  is a  $C^1$  function of the time, and decreasing in the neighborhood of the origin. By defining  $\tau$  as the first time for which  $H'(\tau) = 0$ , for all  $t \in [0, \tau]$  we will have  $H(t) \leq H(0)$ . Let us now find an upper bound for  $\tau$ . First we must have  $\tau \leq \tau_1$ . Since we have supposed  $Q \neq 0$ , one can easily verify that  $H'(t) = 0$  reads

$$C^2Q \left( \frac{1}{c_{i+1}c_j} - \frac{1}{c_{j+1}c_i} \right) t^2 + C \left( \frac{g_j}{c_{i+1}} \frac{g_{i+1}}{c_j} - \frac{g_i}{c_{j+1}} \frac{g_{j+1}}{c_i} \right) t - 1 = 0.$$

$\tau$  is the solution of the second-order equation  $At^2 + Bt - 1 = 0$  with

$$A = C^2Q \left( \frac{1}{c_{i+1}c_j} - \frac{1}{c_{j+1}c_i} \right) \quad \text{and} \quad B = C \left( \frac{g_j}{c_{i+1}} \frac{g_{i+1}}{c_j} - \frac{g_i}{c_{j+1}} \frac{g_{j+1}}{c_i} \right).$$

If the discriminant is negative, then the entropy still decreases on  $[0, \tau_1]$  or else there are two

real roots  $(-B \mp \sqrt{B^2 + 4A})/2A$ . Now in all the cases  $\tau$  is given by  $\tau = (-B + \sqrt{B^2 + 4A})/2A = 2/(B + \sqrt{B^2 + 4A})$ . Thus an upper bound for  $\tau$  is given by  $\tau_2 = 1/(|B| + \sqrt{|A|})$ . It is easy to find an upper bound for  $\sqrt{|A|}$  and  $|B|$ . We have using Lemma 2.1

$$\begin{aligned} |B| &\leq \tilde{N} \left( \max \left| \frac{C g_{i+1}}{c_j} + \frac{C g_j}{c_{i+1}} \right|, \left| \frac{C g_{j+1}}{c_i} + \frac{C g_i}{c_{j+1}} \right| \right) \\ &\leq 2\tilde{N} \sup_{k=i,i+1,j,j+1} (\varepsilon_{k+1/2} g_k) \max_{k=i,i+1,j,j+1} \left| \frac{\sqrt{\varepsilon_{k+1/2}}}{c_k} \right| \\ &\leq \frac{1}{\Delta\varepsilon} 3\sqrt{2}\tilde{N} \sup_{k=i,i+1,j,j+1} (\varepsilon_{k+1/2} g_k) \end{aligned}$$

and for  $\sqrt{|A|}$

$$\begin{aligned} |A| &\leq \tilde{N} \left| \varepsilon_{i+1/2} f_i \varepsilon_{j+1/2} f_{j+1} - \varepsilon_{j+1/2} f_j \varepsilon_{i+1/2} f_{i+1} \right| \left| \frac{\sqrt{\varepsilon_{i+1/2} \varepsilon_{j+1/2}}}{c_{i+1} c_j} - \frac{\sqrt{\varepsilon_{i+1/2} \varepsilon_{j+1/2}}}{c_{j+1} c_i} \right| \\ &\leq \frac{9}{2} \tilde{N}^2 \sup_{k=i,i+1,j,j+1} (\varepsilon_{k+1/2} g_k) \end{aligned}$$

so that an upper bound for  $\tau_2$  is given by

$$\tau_2 \geq \frac{\Delta\varepsilon^2}{\frac{9}{\sqrt{2}} C_N \varepsilon_0 \sup_{k=i,i+1,j,j+1} (\varepsilon_{k+1/2} g_k)} = \tau_3.$$

We must now consider the special case  $j+1 = i$ . The Broadwell model is now

$$\frac{df_i}{dt} = -2 \frac{C_i}{c_i} Q, \quad \frac{df_{i-1}}{dt} = \frac{C_i}{c_{i-1}} Q, \quad \frac{df_{i+1}}{dt} = \frac{C_i}{c_{i+1}} Q,$$

with  $Q = f_{i+1} f_{i-1} - f_i^2$  and now  $C_i = C_{i,i-1} = \tilde{N} \min(\varepsilon_{i+1/2}^{3/2}, \varepsilon_{i-1/2}^{3/2})$ . The time explicit discretization of such a differential equation reads

$$f_i(t) = g_i - 2t \frac{C_i}{c_i} Q, \quad f_{i+1}(t) = g_{i+1} + t \frac{C_i}{c_{i+1}} Q, \quad f_{i-1}(t) = g_{i-1} + t \frac{C_i}{c_{i-1}} Q,$$

where  $g_i, g_{i-1}, g_{i+1}$  is the initial condition and indeed  $Q = g_{i+1} g_{i-1} - g_i^2$ .

The numerical entropy associated with this scheme is

$$H(t) = c_i f_i(t) \log(f_i(t)) + c_{i-1} f_{i-1}(t) \log(f_{i-1}(t)) + c_{i+1} f_{i+1}(t) \log(f_{i+1}(t)).$$

We could do the same analysis as for the Broadwell model for four distinct velocities and one finds that under the time step restriction  $t \leq \frac{1}{2} \tau_3$  the explicit scheme for such a Broadwell model is positive and entropic. ■

As a consequence if the time step for the explicit scheme for the FPLE equation satisfies

$$\Delta t \leq \frac{\Delta\varepsilon^2}{\frac{18}{\sqrt{2}} C_N \varepsilon_0 \sup_{i=1}^{N-1} (\varepsilon_{i+1/2} \sup(f_i, f_{i+1}))} = \Delta t_e, \quad (3.4)$$

then the scheme is positive and entropic. Let us now analyze the dependence of  $\Delta t_e$  through  $\sup_i (\varepsilon_{i+1/2} \sup(f_i, f_{i+1}))$ . We can first remark that we can replace  $\sup_i (\varepsilon_{i+1/2} \sup(f_i, f_{i+1}))$



by  $\varepsilon_0 \sup_i (f_i)$  using the definition of  $\varepsilon_{i+1/2}$ . The second remark is that  $\sup_i (\varepsilon_{i+1/2} \sup (f_i, f_{i+1})) \leq \frac{3}{2} \sup_i (\varepsilon_i f_i)$ . The third remark is that  $\Delta t_e$  could never vanish thanks to the conservation of the mass and the temperature. It would be interesting to have also an estimate of these norms for the continuous problem and at equilibrium, that is, when  $f = \frac{\rho}{(2\pi kT)^{3/2}} \exp^{-\varepsilon/2kT}$ , where  $\rho$  is the density and  $T$  is the real temperature. In this case

$$\|\varepsilon f\|_\infty = \frac{\rho}{(\pi)^{3/2} (2kT)^{1/2}} \exp(1)^{-1} \quad \text{and} \quad \|f\|_\infty = \frac{\rho}{(2\pi kT)^{3/2}}.$$

To our knowledge there is no result about the  $L_\infty$  norm of  $f$  or  $\varepsilon f$  for the continuous FPLE. The only known result is for the Boltzmann equation and it has been obtained by Arkeryd [1]. We notice at this point that for discrete velocity methods [4] for the Boltzmann equation, it is possible to use the above method to find a time step restriction to ensure the decay of the entropy using a time explicit discretization.

Numerical examples show that during the time evolution these norms remain bounded by the corresponding norms for the initial condition and the equilibrium state.

### 3.2. Second-Order Explicit Scheme

Let us now consider second-order time discretization. We have made the choice of the Runge-Kutta of order 2. Let

$$f^{n+1/2} = f^n + \Delta t F P(f^n).$$

The scheme is defined by

$$f^{n+1} = f^n + \frac{\Delta t}{2} (F P(f^n) + F P(f^{n+1/2})). \quad (3.5)$$

We can notice that this scheme is indeed conservative in mass and energy and preserves the equilibrium state.

Let us now define  $g_0, g_1, g_2, g_3$  by

$$\begin{aligned} g_0 &= f^n, \quad g_1 = g_0 + \frac{1}{\mu} F P(g_0), \\ g_2 &= \frac{1}{2} \left( g_0 + g_1 + \frac{1}{\mu} (F P(g_0, g_1) + F P(g_1, g_0)) \right), \quad g_3 = g_1 + \frac{1}{\mu} F P(g_1), \end{aligned}$$

where  $\mu$  is a positive parameter and  $F P(f, g) + F P(g, f)$  is the polar form associated to the quadratic operator  $F P(f)$ .

If we set now  $x = \mu \Delta t$ , (3.5) can be rewritten as

$$f^{n+1} = (1 - x + x^2/2)g_0 + \frac{x}{2}(2 - 3x + x^2)g_1 + (x^2 - x^3)g_2 + \frac{x^3}{2}g_3. \quad (3.6)$$

The implementation of such a scheme is made in the form (3.5) so that the cost is double the cost of the first-order scheme. But to analyze the positivity and the entropic properties of this scheme, the form (3.6) is more suitable. Let us remark that  $f^{n+1}$  is a positive linear convex combination of  $g_0, g_1, g_2$ , and  $g_3$  if and only if  $x \leq 1$ ; that is,  $t \leq 1/\mu$ .

Let us analyze the positivity of such a scheme. It suffices to choose  $\mu$  such that all the  $g_i$ 's are positive. Using the analysis made for the first-order explicit scheme it is clear that if  $\mu$  verifies the CFL condition  $\frac{1}{\mu} \leq \tau$ , where  $\tau$  is defined by (2.13), then  $g_0$ ,  $g_1$ , and  $g_3$  are positive. It is also true for  $g_2$  since

$$g_2 = \frac{1}{2} \left( g_0 \left( 1 - \frac{1}{\mu} K(g_1) \right) + g_1 \left( 1 - \frac{1}{\mu} K(g_0) \right) + \frac{1}{\mu} (G(g_0, g_1) + G(g_1, g_0)) \right),$$

where  $G(f, g) + G(g, f)$  is the polar form associated to the positive and quadratic operator  $G(f)$  and then it is also a positive operator. Since  $g_0$  and  $g_1$  have the same mass and energy it is clear using Lemma 2.2 that if  $\frac{1}{\mu} \leq \tau$  then  $g_2$  is positive. The same holds for  $g_3$ .

Let us study the decay of the entropy. We want that

$$H(f^{n+1}) \leq H(f^n).$$

Using the convexity of the function  $y \rightarrow y \log(y)$ , it is sufficient to find  $x = \mu \Delta t$  such that

$$H(g_1) \leq H(g_0), \quad H(g_2) \leq H(g_0), \quad H(g_3) \leq H(g_0).$$

Using the result obtained for the first-order scheme, it is not possible to find a time step restriction of the form  $\Delta t \leq C(\Delta \varepsilon)^2$ , since the constant  $C$  depends on the  $L^\infty$  norm of  $f$  or  $\varepsilon f$  for which we have no results concerning their evolution and since  $g_i$  depends on the  $g_k$ 's for  $k = 1, \dots, i-1$ .

### 3.3. Implicit Schemes

The full implicit scheme for the FPLe can be written as

$$f = g + tFP(f), \tag{3.7}$$

where  $f, g$  denote  $N$ -dimensional vectors and the collision operator  $FP$  corresponds to system (2.7).

The existence of a solution for the implicit scheme is ensured by the Brouwer fixed point theorem. We set  $\rho$  as the mass of  $g$  and  $C > 0$  such that  $C\rho f + FP(f)$  is a positive operator for all positive  $f$  and the mass of  $f$  is less than or equal to  $\rho$ . Then (3.7) can be rewritten as

$$f(1 + \rho Ct) = g + \rho Ct \left( f + \frac{FP(f)}{\rho C} \right). \tag{3.8}$$

The mapping

$$T(f) = \frac{1}{1 + \rho Ct} g + \frac{\rho Ct}{1 + \rho Ct} \left( f + \frac{FP(f)}{\rho C} \right) \tag{3.9}$$

is continuous from the convex compact set

$$E = \{f > 0 \text{ such that mass of } f \text{ is less or equal to } \rho\}$$

into itself. Thus the Brouwer fixed point theorem ensures the existence of an element  $f^*$  of  $E$  such that  $f^* = T(f^*)$  and  $f^*$  necessarily has the same mass and energy as those of  $g$ . The main problem of this result is that this is not a constructive procedure.

Let us recall that the implicit scheme is automatically entropic. Indeed, we have

$$H(f) - H(g) = \int g \ln(f/g) + \Delta t \int FP(f) \ln(f).$$

Using the classical inequality  $x \ln(y/x) \leq y - x$ , the mass conservation and  $\int FP(f) \ln(f) \leq 0$  we have the desired result. Note that this classical result holds, the sum being discrete or not.

In practice, one should find an iterative method to solve (3.7) such that the sequence of the approximated solutions to (3.7) converges toward a fixed point for sufficiently small time step. Generally, such methods never compute exactly the solution of the implicit scheme (since the iterative procedure is stopped at some point) and this could introduce a large error in energy (see [14]) if the iterative procedure did not conserve this quantity. Moreover, such a method could be very expensive.

The difficulty of defining an iterative method to solve (3.7) comes from the fact that  $FP(f)$  can be written as  $D(f) \cdot f + C(f) \cdot f$ , where  $D(f)$  and  $C(f)$  are tridiagonal matrices ( $D(f)$  is a M-matrix,  $D(f) \cdot f$  represents the diffusive part of the operator, and  $C(f) \cdot f$  is the convective part), but  $D(f)$  and  $C(f)$  highly depend on  $f$  and do not conserve the energy separately. Moreover  $C(f) \cdot f$  does not correspond to an upwind discretization of the convective part.

An interesting constructive procedure to find  $f^*$  is the one based on the proof of the existence of a solution for the Boltzmann equation in the homogeneous case due to Arkeryd [2]. The aim of the method is to choose  $C$  sufficiently large such that  $C\rho(f)f + FP(f)$ , with  $\rho(f)$  the mass of  $f$ , is a positive and monotone operator. This is always possible in our case since it is a quadratic operator building a monotone sequence of approximation which then converges toward a limiting value in the same space. Equation (3.7) can be rewritten as

$$f(1 + \rho Ct) = g + \rho(f)Ct \left( f + \frac{FP(f)}{\rho(f)C} \right).$$

By setting

$$T(f) = \frac{1}{1 + \rho Ct} g + \frac{\rho(f)Ct}{1 + \rho Ct} \left( f + \frac{FP(f)}{\rho(f)C} \right),$$

the iterate procedure is defined by

$$f_{p+1} = T(f_p)$$

starting from  $f_0 = 0$ . One can easily verify that such a procedure defines an increasing sequence of  $f_p$  which converges toward a limiting value  $f^*$  for each  $t$  such that each of the  $f_p$  have the same energy as  $g$ .

But one can also easily verify that  $f^*$  is such that  $\rho(f^*) = \min(\rho, 1/Ct)$ ; that is,  $f^*$  is a solution of (3.7) if and only if the time step verifies  $\rho Ct \leq 1$ , which is an explicit time step restriction. Thus such a method is not suitable for an implicit scheme. Moreover the convergence is very slow.

A more efficient solution to obtain an implicit scheme for the FPLE has been proposed by Epperlein in [14] and it is based on the linearization of the collision operator. More

precisely, he writes (see Eq. (16) in [14])

$$FP(f^{n+1}) = FP(f^n) + \frac{\partial FP(f^n)}{\partial f}(f^{n+1} - f^n) + O(t^2).$$

In our case, we have

$$\frac{\partial FP(f^n)}{\partial f} f^n = 2FP(f^n),$$

since  $FP$  is a quadratic form in  $f$ . Retaining only the linearized operator, the implicit scheme reduces to the system

$$\left( I_d - \Delta t \frac{\partial FP(f^n)}{\partial f} \right) f^{n+1} = \left( I_d - \frac{\Delta t}{2} \frac{\partial FP(f^n)}{\partial f} \right) f^n.$$

Thus, the solution can be obtained directly (without an iterative procedure). But, one needs to solve a full linear system and the cost is  $O(N^3)$  as for the explicit scheme (a linear cost of one evaluation of the collision term and time step restriction is  $\Delta t \leq C(\Delta \varepsilon)^2$ ) for simulating the same time interval. This method is conservative and preserves the Maxwellian state but it is not proved at least to our knowledge that the solution remains positive and that the entropy decays for any time step. Moreover the equilibrium state cannot be achieved in one step; subcycling is needed. We refer to [14] for more details on this method.

In conclusion, it seems impossible to find an iterative procedure to compute the implicit solution which is conservative in mass and energy and entropic at each step for a cost lower than the cost of the explicit scheme, which is  $O(N^3)$ .

To treat high densities or equivalently small mean free path zones, we suggest using a subcycling method until one has attained a time simulation not too large compared with the time collision. Afterward we suggest continuing the simulation in one step by using the method based on Wild sums proposed by Pareschi *et al.* in [24] (the aim of this method consists of replacing the kinetic equation by the BGK equation near the equilibrium) or replacing the FPLE equation by the linear Fokker–Planck equation near the equilibrium state, since for the linear Fokker–Planck equation it is possible to have a low-cost implicit scheme.

#### 4. EXTENSIONS

In this last section, we review some possible extensions. One of the main advantages of this method is its natural generalization to a multispecies case preserving all the properties (conservation, entropy, etc.).

##### 4.1. Nonuniform Grid

It would be useful to extend this non-log discretization of the FPLE on nonuniform meshes, like a uniform mesh in velocity that has been considered in [6]. Unfortunately this is not straightforward on the non-log form of the FPLE if we want to preserve all of the properties of (2.7). A direct discretization of the non-log form (1.2) of the FPLE as in [3] gives a conservative scheme but does not preserve the positivity and the equilibrium state unless the grid is uniform as shown before. Using the Chang and Cooper formulae (see [9, 14]) permits us to preserve the equilibrium state but nothing can be said about the decay

of the entropy and the positivity of the scheme. A way to achieve this goal could be to discretize the log form as in [6] and to use the same kind of average of the product  $g_{i,j}$  as in Section 2.2 to recover a scheme not involving a log term.

#### 4.2. Multispecies

First, let us write the isotropic collision operator with interspecies collision (denote by  $a$  and  $b$  the two species)

$$\begin{aligned}\partial_t f_a &= \frac{\mu_{ab}^2}{m_a} \frac{1}{\sqrt{\varepsilon_a}} \frac{\partial}{\partial \varepsilon_a} \int_0^{\varepsilon_a} f_a(\varepsilon_a) f_b(\varepsilon_b) \left( \frac{1}{m_a} \frac{\partial}{\partial \varepsilon_a} \ln f_a(\varepsilon_a) - \frac{1}{m_b} \frac{\partial}{\partial \varepsilon_b} \ln f_b(\varepsilon_b) \right) \\ &\quad \times k(\varepsilon_a, \varepsilon_b) d\varepsilon_b \\ \partial_t f_b &= \frac{\mu_{ab}^2}{m_b} \frac{1}{\sqrt{\varepsilon_b}} \frac{\partial}{\partial \varepsilon_b} \int_0^{\varepsilon_b} f_a(\varepsilon_a) f_b(\varepsilon_b) \left( \frac{1}{m_b} \frac{\partial}{\partial \varepsilon_b} \ln f_b(\varepsilon_b) - \frac{1}{m_a} \frac{\partial}{\partial \varepsilon_a} \ln f_a(\varepsilon_a) \right) \\ &\quad \times k(\varepsilon_a, \varepsilon_b) d\varepsilon_a,\end{aligned}$$

where  $\mu_{ab} = \frac{m_a m_b}{m_a + m_b}$  is the reduced mass and as for the one species operator  $k(x, y) = \min(x^{3/2}, y^{3/2})$ .

Using the change of variables  $E_a = \varepsilon_a m_a$  and  $E_b = \varepsilon_b m_b$  the system leads to

$$\begin{aligned}\partial_t f_a &= \frac{\sqrt{m_a} \mu_{ab}^2}{m_b} \frac{1}{\sqrt{E_a}} \frac{\partial}{\partial E_a} \int_0^{E_a} f_a(E_a) f_b(E_b) \left( \frac{\partial}{\partial E_a} \ln f_a(E_a) - \frac{\partial}{\partial E_b} \ln f_b(E_b) \right) \\ &\quad \times k(E_a, E_b) dE_b \\ \partial_t f_b &= \frac{\sqrt{m_b} \mu_{ab}^2}{m_a} \frac{1}{\sqrt{E_b}} \frac{\partial}{\partial E_b} \int_0^{E_b} f_a(E_a) f_b(E_b) \left( \frac{\partial}{\partial E_b} \ln f_b(E_b) - \frac{\partial}{\partial E_a} \ln f_a(E_a) \right) \\ &\quad \times k(E_a, E_b) dE_a.\end{aligned}$$

It is straightforward to extend the discretization (2.7) for the multispecies FPLe, using a uniform grid for the two species with  $\Delta E_a = \Delta E_b$ . We refer to [8] for such an analysis for a mixture of electrons and ions.

#### 5. NUMERICAL RESULTS

*The classical Rosenbluth test.* The numerical test presented now is inspired from the work of Rosenbluth *et al.* [25] and has been used by Larroche [19] and Frenod and Lucquin-Desreux [16] to test numerical methods for the Fokker-Planck-Landau equation. The initial data are given by

$$f^0(\varepsilon) = 0.01 \exp(-10[\sqrt{\varepsilon} - 0.3/0.3]^2). \quad (5.1)$$

We take a uniform grid of 50 meshes and  $\varepsilon_0 = 2$ . All the quantities are normalized. We will show the entropy and the distribution function at time  $t = 0, 9, 36, 81, 144, 225, 324, 441, 576, 729$ , and 900 for the first-order scheme (Fig. 1). The same tests have been performed with the second-order scheme and give similar results (the errors are compared below).

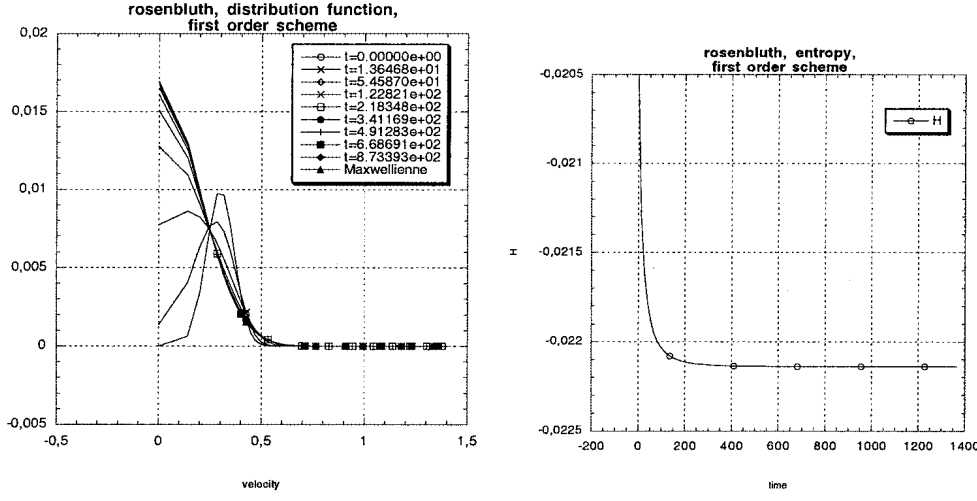
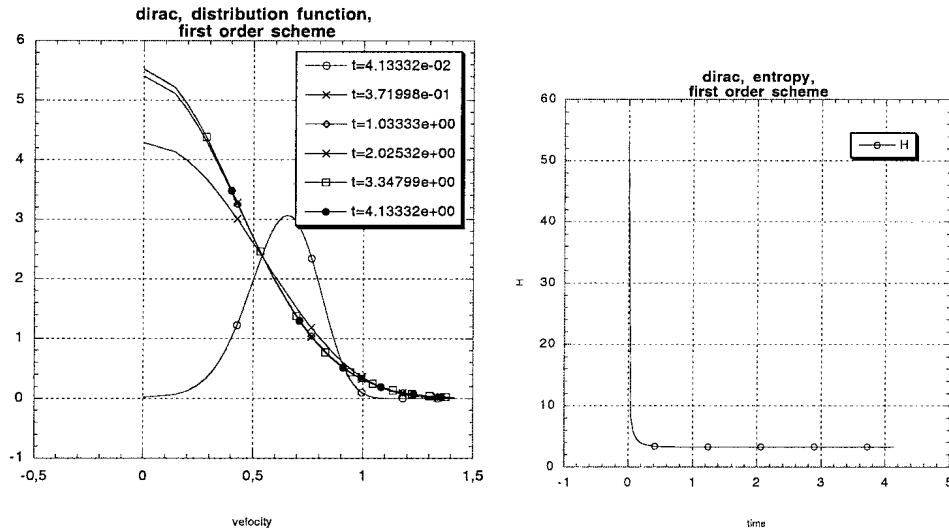


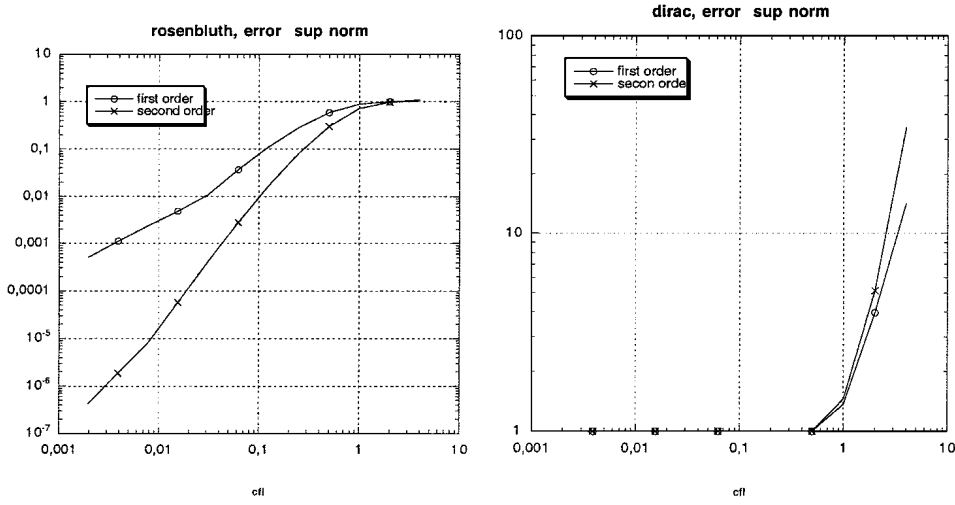
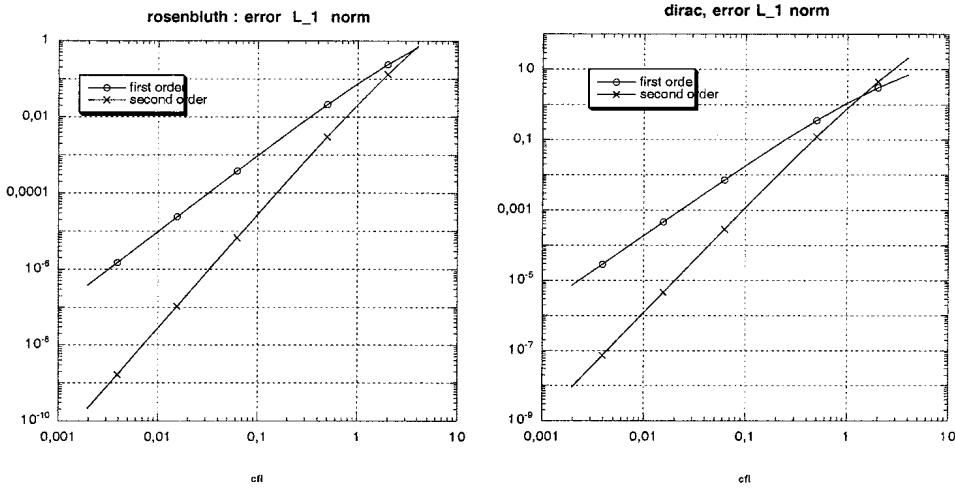
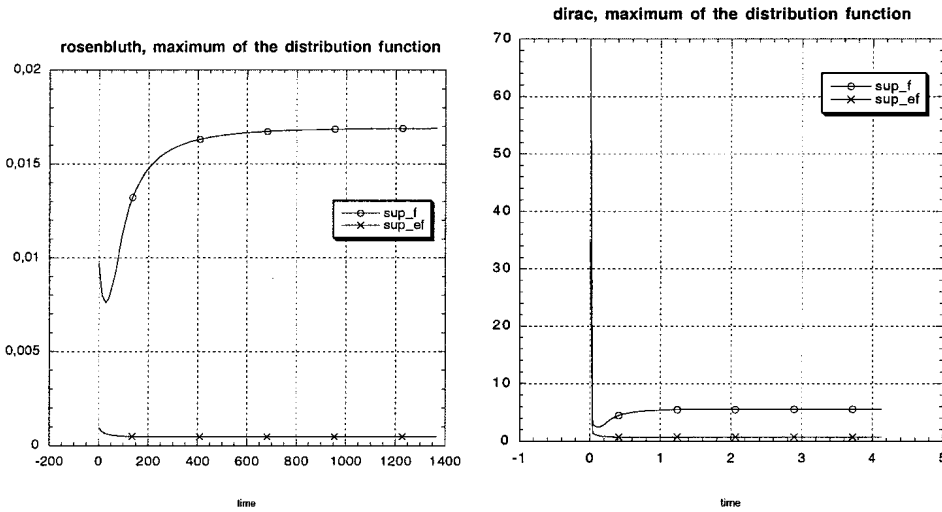
FIG. 1. Rosenbluth test, first-order scheme.

*Second test: Dirac initial distribution.* We choose a Dirac measure in energy that is a spherical shell in the tridimensional velocity space. This typical test cannot be performed with the log scheme. We use the same grid as that for the Rosenbluth test. We will show the entropy and the distribution function at different times between  $t = 0$  and 100 collision times for the first-order scheme (Fig. 2). Once again, the same tests have been performed with the second-order scheme and give similar results.

*Time discretization error.* We show the error due to the time discretization using the first- and second-order scheme on one time step starting from the Rosenbluth initial condition or from a  $\delta$  function. We show the error in  $L_\infty$  norm (Fig. 3) and also in  $L_1$  norm (Fig. 4).

*$L^\infty$  norm for  $f$  and  $\varepsilon f$ .* For the two test cases presented here we show also the time evolution of the  $L_\infty$  norm for  $f$  and  $\varepsilon f$  in Fig. 5. For the two examples the  $L^\infty$  norm of  $\varepsilon f$

FIG. 2.  $\delta$  function test, first-order scheme.

FIG. 3. Error for the  $L_\infty$  norm.FIG. 4. Error for the  $L_1$  norm.FIG. 5. Time evolution of  $L_\infty$  norm for  $f$  and  $\varepsilon f$ .

is much smaller than for  $f$ . We can also remark that these two norms remain bounded in time and seem to depend only on the initial condition and the equilibrium state. The errors are shown using log display for the axis.

## 6. CONCLUSION

Numerical methods for the FPLE not involving the use of the log of the distribution function have practical interest since one can use them for distribution functions that are null in some portion of the numerical velocity space (e.g., a Dirac initial distribution). The scheme based on the non-log form of the Landau equation in the isotropic case has a very simple structure like the discrete model of the Boltzmann equation. We have shown that this scheme can be rendered entropic under time CFL criteria involving the  $L^\infty$  norm of the solution. In this explicit form, this scheme has a cost comparable to the existing implicit schemes for this equation and has all the properties of the continuous model. But implicit time discretization of this scheme is not straightforward as claimed in [3]. Moreover, this scheme has good properties only for uniform meshes in energy.

## APPENDIX A

### Existence and Uniqueness of an Equilibrium Steady State

We shall now prove the existence and uniqueness of a steady-state equilibrium function for a discretized one-dimensional distribution function.

The uniqueness is needed to prove the reverse implication in the H-theorem. Indeed, for some sequence, we have

$$d_t H(f_i(t_k)) \rightarrow 0, \quad f_i(t_k) \rightarrow f_i^\infty.$$

This implies that  $d_t H(f_i^\infty) = 0$  by continuity but we need to prove that  $M_i$  is the unique solution for this system of equations.

More precisely, for any discretized positive distribution function  $f_i^0$ , there exists a unique  $M_{T^0}$  such that

$$\rho(f^0) = \rho(M_{T^0}), \quad E(f^0) = E(M_{T^0}),$$

where the discretized density and energy are defined by

$$\rho(g) = \sum_{i \in I} c_i g_i, \quad E(g) = \sum_{i \in I} c_i \varepsilon_i g_i,$$

and the equilibrium function is of the form

$$M_{T^0}(i) = n_0 \exp(-\varepsilon_i/T^0).$$

Moreover, the temperature is positive if and only if

$$E(f^0)/\rho(f^0) < E_\infty \stackrel{\text{def}}{=} \frac{\sum_{i \in I} c_i \varepsilon_i}{\sum_{i \in I} c_i}.$$



Let us recall that  $c_i$  are defined by (2.1) that can be simplified for a uniform grid in

$$c_i = \frac{1}{2}(v_{i+1/2} + v_{i-1/2})\Delta\varepsilon;$$

i.e.,  $c_i \approx \sqrt{\varepsilon_i}\Delta\varepsilon$ .

One can assume that  $\rho(f^0) = \sum_{i \in I} c_i f_i^0 = 1$  by choosing the density  $n_0$  in the definition of  $M_{T^0}$ .

Then, we have to determine  $T$  such that

$$E(T) \stackrel{\text{def}}{=} \frac{\sum_{i \in I} c_i \varepsilon_i \exp(-\varepsilon_i/T)}{\sum_{i \in I} c_i \exp(-\varepsilon_i/T)} = E^0,$$

with  $E^0 = \sum_{i \in I} c_i \varepsilon_i f_i^0$ . Let us first consider the case  $T > 0$ . The function  $E(T)$  is smooth and continuous on  $]0, \infty[$ . Straightforward calculations give

$$\lim_{T \rightarrow +\infty} E(T) = \frac{\sum_{i \in I} c_i \varepsilon_i}{\sum_{i \in I} c_i}$$

and

$$\lim_{T \rightarrow 0^+} E(T) = \varepsilon_0 = 0.$$

The derivative of  $E(T)$  with respect to  $T$  is given by

$$\frac{dE}{dT} = \frac{1}{T^2} \frac{(\sum_{i \in I} c_i \exp(-\varepsilon_i/T))(\sum_{i \in I} c_i \varepsilon_i^2 \exp(-\varepsilon_i/T)) - (\sum_{i \in I} c_i \varepsilon_i \exp(-\varepsilon_i/T))^2}{(\sum_{i \in I} c_i \exp(-\varepsilon_i/T))^2}.$$

Then, the Cauchy-Schwartz inequality ensures that  $E$  is decreasing with  $T$ . Therefore, when  $E^0 \in ]0, E_\infty[$ , there exists a unique  $T > 0$  such that  $E(T) = E^0$ .

Let us now turn to the (unphysical) case of negative temperature. The function  $E(T)$  is again continuous (in fact, the only point of discontinuity is 0). We have

$$\lim_{T \rightarrow -\infty} E(T) = \varepsilon_N$$

and

$$\lim_{T \rightarrow 0^-} E(T) = E_\infty$$

and the function is decreasing. Therefore, there exists a unique negative  $T$  if and only if  $f^0$  is such that  $E(f^0) \in ]E_\infty, \varepsilon_N[$ .

In the case of negative temperature, this means that the initial distribution is not well represented on the grid and the maximal energy  $\varepsilon_N$  should be increased.

## APPENDIX B

### Partitions into $O(N)$ Independent Subsets

The evolution of  $f_i$  is governed by a system which is the sum of a four-velocities system involving integers in the set

$$\Theta = \{(i, i+1, j, j+1), s.t. i > j, i = 1, \dots, N-1\}.$$

We shall construct a partition of  $\Theta$  into  $O(N)$  subsets involving only distinct integers.

First, note that each quadruplet in  $\Theta$  is determined by the couple  $(i, j)$ . Let us split  $\Theta$  into subsets according to the value of  $k = i - j$ . The case  $k = 1$  is particular since the corresponding subset can be split into three classes according to the value of  $i \bmod 3$  since  $(i, i + 1 = j, j + 1)$  are consecutive integers in this case. For any  $k > 1$  fixed, the quadruplets are characterized by the value of  $i$  (which are either odd or even). If  $k$  is even, then the subset is divided in two parts: the integers such that  $i \bmod 2k \in [0, k[$  and the others (such that  $i \bmod 2k \in [k, 2k - 1[$ ). In the case where  $k$  is odd, the subsets are separated into three parts according to the value of  $i \bmod 2k$  being in  $[0, k - 1]$ , in  $[k + 1, 2k - 1[$ , or in  $[k, k + 1]$ .

Let us introduce the notation  $\Theta = \cup_{i=1, \dots, \tilde{N}} \Theta_i$ . It is easy to see that the partition described above is such that one integer is at most in one of the quadruplet of a given subset. Moreover, there are  $O(N)$  subsets. In fact, the number of subsets  $\tilde{N}$  in the partition is bounded by  $C_N N$ , where  $C_N$  is close to 5 (four subsets for each  $k$  even and six for each odd  $k$ ) and is bounded by 6.

## REFERENCES

1. L. Arkeryd, On the Boltzmann equation, *Arch. Ration. Mech. Anal.* **34**, 1 (1972).
2. L. Arkeryd,  $L^\infty$  estimates for the space homogeneous Boltzmann equation, *J. Stat. Phys.* **31**, 347 (1982).
3. Yu. A., Berezin, V. N. Khudick, and M. S. Pekker, Conservative finite difference schemes for the Fokker–Planck equation not violating the law of an increasing entropy, *J. Comput. Phys.* **69**, 163 (1987).
4. C. Buet, A discrete-velocity scheme for the Boltzmann operator of rarefied gas dynamics, *Transport Theory Stat. Phys.* **25**, 33 (1996).
5. C. Buet and S. Cordier, Numerical analysis of conservatives and entropy schemes for the FPLE, *SIAM J. Numer. Anal.* **36**, 953 (1999).
6. C. Buet and S. Cordier, Conservative and entropy schemes for the isotropic FPLE, *J. Comput. Phys.* **145**, 228 (1998).
7. C. Buet, S. Cordier, P. Degond, and M. Lemou, Fast algorithms for the Fokker–Planck equation, *J. Comput. Phys.* **133**, 310 (1997).
8. C. Buet, S. Dellacherie, and R. Sentis, Numerical solution of an ionic Fokker–Planck equation with electronic temperature, *SIAM J. Numer. Anal.* **39**(4), 1219–1253 (2001).
9. J. S. Chang and G. Cooper, A practical difference scheme for Fokker–Planck equations, *J. Comput. Phys.* **6**, 1 (1970).
10. H. Cohn, Numerical integration of the Fokker–Planck equation and the evolution of stars clusters, *Astrophys. J.* **234**, 1036 (1979).
11. H. Cohn, Late core collapse in star clusters and the gravothermal instability, *Astrophys. J.* **242**, 765 (1980).
12. P. Degond and B. Lucquin-Desreux, The Fokker–Planck asymptotics of the Boltzmann collision operator in the Coulomb case, *Math. Models Meth. Appl. Sci.* **2**(2), 167 (1992).
13. L. Desvillettes and C. Villani, On the spatially homogeneous Landau equation for hard potentials. I. Existence, uniqueness and smoothness. *Comm. Partial Differential Equations* **25**(1–2), 179 (2000), II.  $H$ -theorem and applications, *Comm. Partial Differential Equations* **25**(1–2), 261 (2000).
14. E. M. Epperlein, Fokker–Planck modelling of electrons transport in laser-produced plasmas, *Laser Particle Beams* **2**(2), 257 (1994).
15. E. M. Epperlein, Implicit and conservative difference schemes for the Fokker–Planck equation, *J. Comput. Phys.* **112**, 291 (1994).
16. E. Frenod and B. Lucquin-Desreux, On conservative and entropic discrete axisymmetric Fokker–Planck operators, *RAIRO, Model. Math. Anal. Num.* **132**, 307–339 (1998).
17. R. Gatignol, *Théorie cinétique des gaz à répartitions discrètes de vitesses* (Springer-Verlag, New York, 1975).
18. S. Kullback, A lower bound for discrimination information in terms of variation, *IEE Trans. Inform. Theory* **4**, 126 (1967).

19. O. Larroche, Kinetic simulations of a plasma collision experiment, *Phys. Fluids B* **5**(8), 2816–2840 (1993).
20. M. Lemou, Exact solutions of the Fokker–Planck equation, *C.R. Acad. Sci. Sér. I* **319**, 579 (1994).
21. M. Lemou, Multipole expansions for the Fokker–Planck–Landau operator, *Numer. Math.* **78**(4), 597 (1998).
22. M. Lemou, Numerical algorithms for axisymmetric Fokker–Planck–Landau operators, *J. Comput. Phys.* **157**, 762 (2000).
23. W. M. Macdonald, M. N. Rosenbluth, and W. Chuck, Relaxation of a system of particles with Coulomb interactions, *Phys. Rev.* **107**(2), 350 (1957).
24. E. Gabetta, L. Pareschi, and G. Toscani, Relaxation schemes for nonlinear kinetic equations, *SIAM J. Numer. Anal.* **34**, 2168 (1997).
25. M. N. Rosenbluth, W. Macdonald, and D. L. Judd, Fokker–Planck equation for an inverse-square force, *Phys. Rev.* **107**(1), 1–6 (1957).
26. I. Shkarofsky, Expansion of the relativistic Fokker–Planck equation including nonlinear terms and a non-Maxwellian background, *Phys. Plasmas* **4**, 2464 (1997).
27. L. Spitzer and R. Harm, *Phys. Rev.* **89**, 977 (1953).

## MULTIFLUID IONIZATION MODELS

C. BUET<sup>1</sup>, S. CORDIER<sup>2</sup> and P.A. RAVIART<sup>3</sup>

<sup>1</sup> *Commissariat à l'Énergie Atomique, BP 12, 91680 Bruyères-le-Châtel, France.  
 Christophe.Buet@cea.fr*

<sup>2</sup> *MAPMO, UMR 6628, Université d'Orléans, France.  
 Stephane.Cordier@univ-orleans.fr*

<sup>3</sup> *Centre de Mathématiques Appliquées, Ecole Polytechnique, 91218 Palaiseau cedex, France.  
 raviart@cmapx.polytechnique.fr*

(Leave 1 inch blank space for publisher.)

In this paper, we present a multi-fluid ionization model. We prove that this stationary, mono-dimensional model has a maximal solution which is not global at variance with the mono-species case and we present a numerical method for solving this highly singular system of ordinary differential equations. Numerical results are compared with those obtained for other models.

**Keywords:** Singular system of ordinary differential equations, Cauchy problem, Fluid models, Numerical schemes, Ionization.

### 1. Introduction

In many problems encountered in plasma physics, ionization processes play an essential role. This is the case when considering the extraction of an ion beam from a neutral plasma.

In a reaction chamber, electrons ionize a gas through fairly complicated ionization processes: the ion beam is extracted from the generated plasma through a small aperture facing an electrode carried to an electric potential which is strongly negative with respect to the plasma potential. Indeed, simple ionization models are able to predict the density and the current density of the ion beam at the extraction. Moreover, these quantities, which can be computed analytically, are independent of the ionization process which implies that one does not need to model the plasma from which the ion beam is extracted. For these results, we refer for instance to <sup>10</sup>.

These simple models suppose that only a single species of ions exists in the plasma. This hypothesis is not realistic in the applications : a real plasma contains several species of ions. The purpose of this paper is to study a multispecies ionization fluid model and to determine the densities and current densities at the extraction of the beam. Although an analytic approach is no longer available in the multispecies case, we will be able to give precise qualitative answers: the densities and the current densities of each ion species at the extraction now depend on the ionization process but we indicate a simple way of computing them numerically.

The paper is organized as follows. In Section 1, we introduce a multispecies one-dimensional stationary fluid model that can be viewed as a singular perturbation problem for a multifluid Euler-Poisson type system. The paper is devoted to the study of the formal limit problem called plasma approximation which models the neutral plasma from which the ion beam is extracted. In Section 2, we show that solving this limit model amounts to solve a highly singular Cauchy problem for a nonlinear differential system. We state an

existence result and describe qualitative properties of the solution of this Cauchy problem. In particular, the solution is shown to exist on a maximal interval  $[0, x_0)$ :  $x_0$  characterizes the limit of the neutral plasma at which the ion beam is extracted. In Section 3, we establish a priori bounds for the solution of the model. Using these estimates, we give in Section 4 the proof of the result of Section 2. In Section 5, we present a numerical method of solution of the plasma approximation and we compare the results obtained for this model with those obtained for a multispecies kinetic model and for a one velocity model.

## 2. The model.

We consider a simple device (cf. <sup>10</sup>): an *unmagnetized* plasma is generated between two parallel plane absorbing electrodes at the same potential. The device is thus symmetric with respect to the plane  $X = 0$ , the electrodes being located at  $X = \pm a$ . Moreover, we can use a one-dimensional plane modeling. We suppose that the plasma consists of electrons with charge  $-e$  and of  $p$  species of ions, indexed by  $\alpha = 1, \dots, p$ , with mass  $m_\alpha$  and charge  $q_\alpha = Z_\alpha e$ . We assume that the electrons behave as an isothermal fluid with temperature  $T_e$ . Then by neglecting the inertia of electrons in the electron momentum conservation equation, we obtain that the electron density  $N_e$  is related to the electric potential  $\Phi$  by the Maxwell - Boltzmann relation

$$N_e = C \exp\left(\frac{e\Phi}{kT_e}\right)$$

where  $C$  is a constant. Denoting by  $N_0$  the electron density and setting  $\Phi = 0$  at the plasma center  $X = 0$  we obtain

$$N_e = N_0 \exp\left(\frac{e\Phi}{kT_e}\right). \quad (2.1)$$

On the other hand, we assume that the ions of the species  $\alpha$  are formed at rest with an ionization rate  $G_\alpha = G_\alpha(N_e)$  which depends only on the electron density. Typically, we have

$$G_\alpha(N_e) = \nu_\alpha N_0 \left(\frac{N_e}{N_0}\right)^{\gamma_\alpha}$$

where  $\nu_\alpha$  is a collision frequency and  $\gamma_\alpha \geq 0$  is a constant which characterizes the ionization process ( $\gamma_\alpha = 0, 1, 2$  in practice). In addition, we suppose that the ions are non collisional and that we can neglect the temperature of each species. Assuming stationarity, we obtain that the density  $N_\alpha$  and the velocity  $U_\alpha$  of the ions of the species  $\alpha$  satisfy the equations

$$\frac{d}{dX} (N_\alpha U_\alpha) = G_\alpha(N_e), \quad (2.2)$$

$$m_\alpha \frac{d}{dX} (N_\alpha U_\alpha^2) = -Z_\alpha e N_\alpha \frac{d\Phi}{dX}. \quad (2.3)$$

Finally, the electric potential  $\Phi$  satisfies the Poisson equation

$$-\frac{d^2\Phi}{dX^2} = \frac{e}{\varepsilon_0} \left( \sum_{\alpha=1}^p Z_\alpha N_\alpha - N_e \right). \quad (2.4)$$

We supplement the above equations (2.1)-(2.4) with the following boundary conditions which reflect the symmetry of the device

$$U_\alpha(0) = 0, \quad 1 \leq \alpha \leq p, \quad (2.5)$$

$$\Phi(0) = \frac{d\Phi}{dX}(0) = 0. \quad (2.6)$$

The problem (2.1)-(2.6) has indeed to be viewed as a singular perturbation problem. This becomes clear when performing a scaling of the above equations. We introduce characteristic quantities:

- a length  $L = \frac{1}{\nu} \sqrt{\frac{kT_e}{m}}$  where  $\nu$  is a typical ionization frequency and  $m$  a typical ion mass,
- an electric potential  $\bar{\Phi} = -\frac{kT_e}{e}$ ,
- an electron density  $\bar{N}_e = N_0$ ,

and for each ion's species  $\alpha$

- a density  $\bar{N}_\alpha = \frac{N_0}{Z_\alpha}$
- a velocity  $\bar{U}_\alpha = \sqrt{\frac{Z_\alpha kT_e}{m_\alpha}}$ ,
- an ionization rate  $\bar{G}_\alpha = \frac{1}{L} \sqrt{\frac{kT_e}{Z_\alpha m_\alpha}} \bar{N}_e = \nu \sqrt{\frac{Z_\alpha m}{m_\alpha}} \bar{N}_e$ .

If we define the dimensionless quantities  $x, \varphi, n_\alpha, n_e, u_\alpha$  by

$$X = Lx, \Phi = \bar{\Phi}\varphi, N_e = \bar{N}_e n_e, N_\alpha = \bar{N}_\alpha n_\alpha, U_\alpha = \bar{U}_\alpha u_\alpha$$

and we set

$$g_\alpha(n_e) = \frac{1}{\bar{G}_\alpha} G_\alpha(\bar{N}_e n_e),$$

the equations (2.1)-(2.4) become respectively

$$n_e = \exp(-\varphi), \quad (2.7)$$

$$\frac{d}{dx} (n_\alpha u_\alpha) = g_\alpha(n_e), \quad 1 \leq \alpha \leq p, \quad (2.8)$$

$$\frac{d}{dx} (n_\alpha u_\alpha^2) = n_\alpha \frac{d\varphi}{dx}, \quad 1 \leq \alpha \leq p, \quad (2.9)$$

$$\varepsilon^2 \frac{d^2 \varphi}{dx^2} = \sum_{\alpha=1}^p n_\alpha - n_e \quad (2.10)$$

where

$$\varepsilon = \frac{\nu}{\omega_p} = \frac{\lambda_D}{L}, \quad \lambda_D = \sqrt{\frac{\varepsilon_0 kT_e}{N_0 e^2}}, \quad \omega_p^2 = \frac{N_0 e^2}{\varepsilon_0 m},$$

while the boundary conditions (2.5),(2.6) give

$$u_\alpha(0) = 0, \quad 1 \leq \alpha \leq p, \quad (2.11)$$

$$\varphi(0) = \frac{d\varphi}{dx}(0) = 0. \quad (2.12)$$

For the physical validity of the model, we refer to <sup>14,10</sup> in which the case of a single ion species is considered. This model is in fact the basis for the numerical simulation of ion extraction (cf <sup>15,16</sup>)

In most of the physical situations the Debye length  $\lambda_D$  is far smaller than the ionization characteristic length  $L$  so that  $\varepsilon > 0$  is a small parameter and the initial value problem (2.7)-(2.12) is indeed a singular perturbation problem. In fact, this problem involves two essential mathematical difficulties (see for example <sup>5,9</sup>):

(i) It presents a singularity at the origin and the authors are not aware of any existence result of a solution at least in the multispecies case  $p \geq 2$  although extensive numerical computations have shown that for any  $\varepsilon > 0$  such a solution exists and is uniquely defined on the whole half line  $x \geq 0$ . Note however that an existence result has been proved in the case  $p = 1$  (cf. <sup>1, 12</sup>);

(ii) As we shall see it in Section 2, the formal limit problem corresponding to  $\varepsilon = 0$  has a solution which is defined only on a finite interval  $[0, x_0)$ . As a consequence, the asymptotic analysis of the problem (2.7)-(2.12) appears to be far from being standard and seems to need new ideas and techniques.

This paper is devoted to the study of this formal limit problem or *plasma approximation* in the physicists' terminology. It may be considered as a first step towards the mathematical analysis of the singular perturbation problem (2.7)-(2.12). This formal limit consists in setting  $\varepsilon = 0$  in (2.10); we obtain the condition

$$n_e = \sum_{\alpha=1}^p n_{\alpha} \quad (2.13)$$

which expresses that the plasma is locally electrically neutral. For deriving the limit model, we eliminate the electric potential: using (2.7), we obtain

$$\frac{d\varphi}{dx} = -\frac{1}{n_e} \frac{dn_e}{dx},$$

and

$$\varphi(0) = 0 \iff n_e(0) = 1.$$

Hence the plasma approximation amounts to find  $\{(n_{\alpha}, u_{\alpha}); 1 \leq \alpha \leq p\}$ , solutions of the differential equations

$$\frac{d}{dx} (n_{\alpha} u_{\alpha}) = g_{\alpha}(n_e), \quad 1 \leq \alpha \leq p, \quad (2.14)$$

$$\frac{d}{dx} (n_{\alpha} u_{\alpha}^2) + \frac{n_{\alpha}}{n_e} \frac{dn_e}{dx} = 0, \quad 1 \leq \alpha \leq p, \quad (2.15)$$

with the initial conditions

$$n_e(0) = 1, \quad (2.16)$$

$$u_{\alpha}(0) = 0, \quad 1 \leq \alpha \leq p, \quad (2.17)$$

and the neutrality condition (2.13).

### 3. The plasma approximation.

Before establishing the existence of a solution of the limit model (2.13)-(2.17) and studying its properties, we need to put this problem in a more convenient form. We set

$$j_\alpha = n_\alpha u_\alpha, \quad k_\alpha = n_\alpha u_\alpha^2 \quad (3.1)$$

where  $j_\alpha$  and  $k_\alpha$  represent respectively the scaled current and kinetic energy of the  $\alpha$  species, so that the equations (2.14), (2.15) are respectively written

$$\frac{dj_\alpha}{dx} = g_\alpha(n_e), \quad (3.2)$$

$$\frac{dk_\alpha}{dx} + \frac{n_\alpha}{n_e} \frac{dn_e}{dx} = 0. \quad (3.3)$$

By replacing  $k_\alpha$  by  $\frac{j_\alpha^2}{n_\alpha}$  and using (2.13) and (3.2), Eq.(3.3) becomes

$$-u_\alpha^2 \frac{dn_\alpha}{dx} + \frac{n_\alpha}{n_e} \sum_{\beta=1}^p \frac{dn_\beta}{dx} = -2u_\alpha g_\alpha(n_e).$$

Next, we introduce the vectors of dimension  $2p$

$$U = \begin{pmatrix} n_1 \\ \vdots \\ n_p \\ j_1 \\ \vdots \\ j_p \end{pmatrix}, \quad G(U) = \begin{pmatrix} -2u_1 g_1(n_e) \\ \vdots \\ -2u_p g_p(n_e) \\ g_1(n_e) \\ \vdots \\ g_p(n_e) \end{pmatrix}$$

and the  $2p \times 2p$  matrix

$$A(U) = \left( \begin{array}{c|c} B(U) & \mathbf{0} \\ \hline \mathbf{0} & I \end{array} \right)$$

where  $B(U)$  is the  $p \times p$  matrix

$$B(U) = \begin{pmatrix} \frac{n_1}{n_e} - u_1^2 & \frac{n_1}{n_e} & \dots & \frac{n_1}{n_e} \\ \frac{n_2}{n_e} & \frac{n_2}{n_e} - u_2^2 & \dots & \frac{n_2}{n_e} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{n_p}{n_e} & \frac{n_p}{n_e} & \dots & \frac{n_p}{n_e} - u_p^2 \end{pmatrix}$$

so that differential system (2.14), (2.15) may be equivalently written

$$A(U) \frac{dU}{dx} = G(U) \quad (3.4)$$



with again the neutrality condition (2.13). The initial conditions (2.16) and (2.17) give the following  $p + 1$  constraints

$$\begin{cases} n_e(0) = 1 \\ j_\alpha(0) = 0, \quad 1 \leq \alpha \leq p \end{cases} \quad (3.5)$$

and do not specify the whole initial condition  $U(0)$  in the multispecies case  $p \geq 2$ . On the other hand, the matrix  $\mathbf{A}(U)$  may be singular. The following result shows in particular that this is indeed the case at  $x = 0$ .

**Lemma 1.** We have

$$\det(\mathbf{A}(U)) = \frac{(-1)^{p-1}}{n_e} \left( \prod_{\alpha=1}^p u_\alpha^2 \right) \left( \sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2} - n_e \right). \quad (3.6)$$

**Proof.** We have obviously

$$\det(\mathbf{A}(U)) = \det(\mathbf{B}(U)).$$

Let us prove by induction that  $\det(\mathbf{B}(U))$  is given by the right hand side of (3.6). The result is clearly true for  $p = 1$ . Assume that it holds for  $p - 1$ . By developing  $\det(\mathbf{B}(U))$  along its first line, we find

$$\begin{aligned} \det(\mathbf{B}(U)) &= \left( \frac{n_1}{n_e} - u_1^2 \right) \begin{vmatrix} \frac{n_2}{n_e} - u_2^2 & \dots & \frac{n_2}{n_e} \\ \vdots & & \vdots \\ \frac{n_p}{n_e} & \dots & \frac{n_p}{n_e} - u_p^2 \end{vmatrix} - \\ &\quad - \frac{n_1}{n_e} \begin{vmatrix} \frac{n_2}{n_e} & \frac{n_2}{n_e} & \dots & \frac{n_2}{n_e} \\ \frac{n_3}{n_e} & \frac{n_3}{n_e} - u_3^2 & \dots & \frac{n_3}{n_e} \\ \vdots & \vdots & & \vdots \\ \frac{n_p}{n_e} & \frac{n_p}{n_e} & \dots & \frac{n_p}{n_e} - u_p^2 \end{vmatrix} \\ &\quad + \dots + (-1)^{p-1} \frac{n_1}{n_e} \begin{vmatrix} \frac{n_2}{n_e} & \frac{n_2}{n_e} - u_2^2 & \dots & \frac{n_2}{n_e} \\ \vdots & \vdots & & \vdots \\ \frac{n_{p-1}}{n_e} & \frac{n_{p-1}}{n_e} & \dots & \frac{n_{p-1}}{n_e} - u_{p-1}^2 \\ \frac{n_p}{n_e} & \frac{n_p}{n_e} & \dots & \frac{n_p}{n_e} \end{vmatrix}. \end{aligned}$$

Now, we use the induction hypothesis for evaluating the first determinant and we compute the remaining determinants by subtracting the first column for the others. We obtain

$$\begin{aligned} \det(\mathbf{B}(U)) &= (-1)^{p-2} \left( \frac{n_1}{n_e} - u_1^2 \right) \frac{1}{n_e} \left( \prod_{\alpha=2}^p u_\alpha^2 \right) \left( \sum_{\alpha=2}^p \frac{n_\alpha}{u_\alpha^2} - n_e \right) + \\ &\quad + (-1)^{p-1} \left\{ \frac{n_1 n_2}{n_e^2} u_3^2 \dots u_p^2 + \dots + \frac{n_1 n_p}{n_e^2} u_2^2 \dots u_{p-1}^2 \right\} = \\ &= \frac{(-1)^{p-1}}{n_e^2} \left( \prod_{\alpha=1}^p u_\alpha^2 \right) \left\{ \left( -\frac{n_1}{u_1^2} + n_e \right) \left( \sum_{\alpha=2}^p \frac{n_\alpha}{u_\alpha^2} - n_e \right) + \frac{n_1}{u_1^2} \left( \sum_{\alpha=2}^p \frac{n_\alpha}{u_\alpha^2} \right) \right\} \end{aligned}$$

and (3.6) follows.  $\square$

As a consequence of (3.6), we obtain that the matrix  $\mathbf{A}(\mathbf{U})$  is singular if and only if one of the two following situations holds:

- (i)  $p \geq 2$  and  $u_\alpha = 0$  for at least two indices  $\alpha$ ;
- (ii) we have

$$\sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2} = n_e. \quad (3.7)$$

In particular, in the multispecies case  $p \geq 2$ , the initial conditions (2.17) imply that the matrix  $\mathbf{A}(\mathbf{U}(0))$  is singular whatever the  $n_\alpha(0)$ 's may be. In fact, since  $\mathbf{B}(\mathbf{U}(0))$  is a matrix of rank 1, we find that 0 is an eigenvalue of  $\mathbf{A}(\mathbf{U}(0))$  of multiplicity  $p - 1$  with a corresponding eigenspace of dimension 1.

In all the sequel, we will call (physically admissible) solution of the initial value problem (2.13)-(2.17) or (3.4), (3.5) any  $C^1$  function  $\mathbf{U}$  from an interval  $I = [0, x_1]$  into  $\mathbf{R}^{2p}$  which satisfies the equations (3.4), (3.5) and the physical constraints

$$n_\alpha(x) > 0, \quad \alpha \leq 1 \leq p. \quad (3.8)$$

We will also assume the following hypothesis.

$$\left\{ \begin{array}{l} \text{each (nondimensional) ionization rate } g_\alpha : \mathbf{R}_+ \rightarrow \mathbf{R}_+ \text{ is an increasing} \\ C^1 \text{ function which satisfies } g_\alpha(n_e) > 0 \quad \forall n_e > 0 \quad (\dagger). \end{array} \right. \quad (3.9)$$

We set

$$g(n_e) = \sum_{\alpha=1}^p g_\alpha(n_e). \quad (3.10)$$

Then, we can state the main result of this paper:

**Theorem 2.** Assume the hypothesis (3.9). The plasma approximation model (2.13)-(2.17) has a (physically admissible) solution defined in a maximal interval  $[0, x_0)$  with  $x_0 < +\infty$ , which possesses the following properties:

- (i)  $n_e \in C^2([0, x_0))$  is a strictly decreasing function;
- (ii) we have at  $x = 0$

$$n_\alpha(0) = \frac{g_\alpha(1)}{g(1)}, \quad \frac{dn_\alpha}{dx}(0) = 0, \quad 1 \leq \alpha \leq p, \quad (3.11)$$

$$\frac{dn_e}{dx}(0) = 0, \quad \frac{d^2 n_e}{dx^2}(0) = -2g(1)^2; \quad (3.12)$$

(iii)  $\mathbf{U}(x_0) = \lim_{x \rightarrow x_0} \mathbf{U}(x)$  exists;

(iv) the matrix  $\mathbf{A}(\mathbf{U}(x_0))$  is singular and we have at  $x = x_0$

$$\left\{ \begin{array}{l} \sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2} = n_e \\ \sum_{\alpha=1}^p n_\alpha u_\alpha^2 = 1 - n_e \end{array} \right. \quad (3.13)$$

\*This hypothesis, which is physically realistic, can be slightly weakened mathematically.

and

$$n_e(x_0) \leq \frac{1}{2}; \quad (3.14)$$

(v) we have

$$\lim_{x \rightarrow x_0} \frac{dn_\alpha}{dx}(x) = -\infty, \quad \lim_{x \rightarrow x_0} \frac{du_\alpha}{dx}(x) = +\infty, \quad 1 \leq \alpha \leq p. \quad (3.15)$$

We conjecture that this solution is unique (cf. Remark 3). Let us illustrate the above result by considering two simple but useful examples.

**Example 1.** *The case of a single ion species.* If  $p = 1$ , the system (2.13)-(2.17) reduces to

$$\begin{cases} \frac{d}{dx}(nu) = g(n), \\ \frac{d}{dx}(nu^2 + n) = 0, \\ n(0) = 1, \quad u(0) = 0, \end{cases} \quad (3.16)$$

where  $n$  denotes the density of both electrons and ions and  $u$  is the ion velocity. This system (3.16) can be easily solved. Indeed, using the second equation (3.16) and the initial conditions, we obtain

$$nu^2 + n = 1$$

and therefore

$$u^2 = \frac{1-n}{n}$$

which yields  $n \leq 1$ . On the other hand, the first equation (3.16) together with the initial conditions yield

$$(nu)(x) = \int_0^x g(n(y))dy \geq 0.$$

Hence  $u$  is  $\geq 0$  and we have

$$u = \sqrt{\frac{1-n}{n}}.$$

By replacing  $u$  by its value in the first equation (3.16), we find

$$\frac{d}{dx} \sqrt{n(1-n)} = g(n)$$

so that (3.16) amounts to solve the equation

$$\int_1^{n(x)} \frac{(1-2m)dm}{2g(m)\sqrt{m(1-m)}} = x. \quad (3.17)$$

Next, we notice that the function

$$f(n) = \int_1^n \frac{(1-2m)dm}{2g(m)\sqrt{m(1-m)}}$$

has in the interval  $[0, 1]$  a unique maximum  $x_0 = f(\frac{1}{2})$  at  $n = \frac{1}{2}$  and is increasing in  $[0, \frac{1}{2}]$  then decreasing in  $[\frac{1}{2}, 1]$ . Therefore the equation (3.17) has a solution only for  $x \leq x_0$ . Since  $n(0) = 1$ , we obtain that the function  $x \rightarrow n(x)$  decreases from 1 to  $\frac{1}{2}$  as  $x$  increases from 0 to  $x_0$ . Moreover, we have clearly

$$\lim_{x \rightarrow x_0} \frac{dn}{dx}(x) = -\infty.$$

We thus get conclusions of the Theorem 2 with  $n(x_0) = \frac{1}{2}$ . Moreover we have

$$u(x_0) = 1, \quad j(x_0) = (nu)(x_0) = \frac{1}{2}. \quad \square \quad (3.18)$$

Note that for a constant ionization rate  $g = g_0$ , the exact solution is given by

$$n(x) = \frac{1 + \sqrt{1 - 4g_0^2 x^2}}{2}.$$

**Example 2.** *A particular case of multispecies ionization.* We next consider the case -which will be useful in the proof of the Theorem- where the ionization rates are proportional, i.e.,

$$g_\alpha(n_e) = a_\alpha g(n_e), \quad 1 \leq \alpha \leq p, \quad (3.19)$$

for some constants  $a_\alpha > 0$ ,  $1 \leq \alpha \leq p$ , with  $\sum_{\alpha=1}^p a_\alpha = 1$ . Then, denoting by  $(n, u)$  the solution of (3.16), it is an obvious matter to check that

$$n_\alpha = a_\alpha n, \quad u_\alpha = u, \quad 1 \leq \alpha \leq p, \quad (3.20)$$

is solution of the system (2.13)-(2.17). Let us prove that this is the only solution. When (3.19) holds, (2.14) is written:

$$\frac{d}{dx}(n_\alpha u_\alpha) = a_\alpha g(n_e),$$

which yields

$$n_\alpha u_\alpha = a_\alpha j, \quad j(x) = \int_0^x g(n_e(y)) dy.$$

Replacing in (2.15)  $n_\alpha$  by  $\frac{a_\alpha j}{u_\alpha}$  gives

$$\frac{d}{dx}(j u_\alpha) + \frac{1}{u_\alpha} \frac{1}{n_e} \frac{dn_e}{dx} = 0,$$

and therefore

$$\frac{d}{dx}(ju_\alpha)^2 = -\frac{2j}{n_e} \frac{dn_e}{dx}.$$

Since  $(ju_\alpha)(0) = 0$ , we obtain that  $u_\alpha$  is independent of  $\alpha$ . Hence we have for some function  $u$

$$u_\alpha = u, \quad n_\alpha = a_\alpha n, \quad n = \frac{j}{u}.$$

Moreover the pair  $(n, u)$  is easily seen to be the solution of (3.16) which proves our assertion.  $\square$

Let us conclude this section by giving a physical picture of the results of Theorem 1. Going back to the device considered in Section 1, we obtain that the plasma approximation is valid in a slab of maximal width  $2Lx_0$ . Then two cases are to be considered.

(i) If  $a \leq X_0 = Lx_0$ , the plasma is quasineutral in the whole interelectrode domain at the exception of thin layers located at each electrode  $X = \pm a$ . We refer to <sup>4, 3, 13</sup> for an analysis of the corresponding boundary layer problem in the absence of the ionization source terms.

(ii) If  $a > X_0$ , the plasma is quasi neutral in the slab  $|X| \leq X_0$ . At  $X_0$ , the electric neutrality breaks down: ions are extracted from the plasma while the electrons remains confined in the neutral plasma zone. We refer to <sup>11,10</sup> for a physical description of ion extraction from a plasma and to <sup>2,13</sup> for a mathematical discussion of a related but simpler situation.

In the case of one single ion species, the velocity  $U(X_0)$  and the current density  $J(X_0) = Ze(NU)(X_0)$  of the ions leaving the neutral plasma at  $X_0$  are independent of the ionization rate:

$$U(X_0) = \sqrt{\frac{kT_e}{m}}, \quad J(X_0) = \frac{N_0 Ze}{2} \sqrt{\frac{kT_e}{m}}.$$

These expressions have been indeed obtained by the physicists (see <sup>10</sup>) and are known respectively as the ion acoustic velocity and the Bohm current.

The picture is not so simple in the multispecies case: the velocity  $U_\alpha(X_0)$  and the current density  $J_\alpha(X_0) = Z_\alpha e(N_\alpha U_\alpha)(X_0)$  of the ions of the species  $\alpha$  leaving the plasma at  $X_0$  depend now on the ionization rates. However as we will see it in Section 5, extensive numerical simulations indicate that we have in any case

$$u_\alpha(x_0) \simeq 1,$$

and therefore by (3.13)

$$j(x_0) \simeq \frac{1}{2}.$$

This yields

$$U_\alpha(X_0) \simeq \sqrt{\frac{Z_\alpha kT_e}{m_\alpha}}, \quad \sum_{\alpha=1}^p \frac{J_\alpha(X_0)}{\sqrt{\frac{Z_\alpha kT_e}{m_\alpha}}} \simeq \frac{eN_0}{2},$$

where  $J_\alpha = Z_\alpha e N_\alpha U_\alpha$ . However, we have no further a priori information on the partial current densities  $J_\alpha$  which are of practical importance for the physicists.

#### 4. A priori estimates.

As a preliminary step for proving Theorem 2, we first derive a priori bounds for any solution of the problem (2.13)-(2.17) (or (3.4),(3.5)). Using (3.2) and (2.17), we have

$$j_\alpha(x) = \int_0^x g_\alpha(n_e(y)) dy. \quad (4.1)$$

Then it follows from the hypothesis (3.9) that

$$j_\alpha(x) > 0 \quad \text{for all } x > 0, \quad 1 \leq \alpha \leq p. \quad (4.2)$$

Since

$$k_\alpha = \frac{j_\alpha^2}{n_\alpha}$$

we find by (3.8)

$$k_\alpha(x) > 0 \quad \text{for all } x > 0, \quad 1 \leq \alpha \leq p. \quad (4.3)$$

The following result will play an essential role in all the sequel.

**Lemma 3.** Any solution of the system (2.13)-(2.17) satisfies the inequality

$$\frac{dn_e}{dx}(x) < 0 \quad (4.4)$$

at all point  $x > 0$  where this solution exists.

**Proof.** By multiplying (3.3) by  $k_\alpha$ , we obtain

$$\frac{1}{2} \frac{d}{dx} k_\alpha^2 = -\frac{n_\alpha k_\alpha}{n_e} \frac{dn_e}{dx} = -\frac{j_\alpha^2}{n_e} \frac{dn_e}{dx}$$

and since  $k_\alpha(0) = 0$

$$k_\alpha^2(x) = -2 \int_0^x \left( \frac{j_\alpha^2}{n_e} \frac{dn_e}{dx} \right)(y) dy. \quad (4.5)$$

Then it follows from (4.3) that

$$\int_0^x \frac{j_\alpha^2}{n_e} \frac{dn_e}{dx}(y) dy < 0 \quad \text{for all } x > 0.$$

Together with (4.2), this implies that (4.4) holds for  $x > 0$  small enough. In order to prove (4.4) for any  $x > 0$ , it is enough to check that  $\frac{dn_e}{dx}$  cannot vanish. Indeed, using (2.13), we can write

$$\frac{dn_e}{dx} = \frac{d}{dx} \left( \sum_{\alpha=1}^p n_\alpha \right) = \frac{d}{dx} \left( \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha} \right) = \sum_{\alpha=1}^p \left( 2 \frac{j_\alpha}{k_\alpha} \frac{dj_\alpha}{dx} - \frac{j_\alpha^2}{k_\alpha^2} \frac{dk_\alpha}{dx} \right)$$

so that by (3.2),(3.3)

$$\frac{dn_e}{dx} = \sum_{\alpha=1}^p \left( 2 \frac{j_\alpha}{k_\alpha} g_\alpha(n_e) + \frac{j_\alpha^2}{k_\alpha^2} \frac{n_\alpha}{n_e} \frac{dn_e}{dx} \right).$$

Hence we obtain

$$\left( 1 - \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha^2} \frac{n_\alpha}{n_e} \right) \frac{dn_e}{dx} = 2 \sum_{\alpha=1}^p \frac{j_\alpha}{k_\alpha} g_\alpha(n_e). \quad (4.6)$$

Since by (3.9), (4.2), (4.3), we have

$$\frac{j_\alpha}{k_\alpha} g_\alpha(n_e) > 0 \quad \text{for all } x > 0,$$

we find

$$\left( 1 - \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha^2} \frac{n_\alpha}{n_e} \right) \frac{dn_e}{dx} > 0 \quad \text{for all } x > 0 \quad (4.7)$$

which proves our assertion.  $\square$

**Remark 3.** The inequality (4.6) can be also written

$$\left( n_e - \sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2} \right) \frac{dn_e}{dx} > 0 \quad \text{for all } x > 0. \quad (4.8)$$

Then it follows from (3.6) that the matrix  $\mathbf{A}(\mathbf{U}(x))$  is never singular at a point  $x > 0$  where the solution  $\mathbf{U}$  of (3.4), (3.5) is defined. Hence, at such a point  $x > 0$ , (3.4) becomes

$$\frac{d\mathbf{U}}{dx} = \mathbf{F}(\mathbf{U}), \quad \mathbf{F}(\mathbf{U}) = \mathbf{A}(\mathbf{U})^{-1} \mathbf{G}(\mathbf{U})$$

and  $\mathbf{F}$  is a  $C^1$  function. Therefore the uniqueness of the solution of (2.13)-(2.17) holds as soon as one can prove the uniqueness of the solution in a neighborhood of  $x = 0$ .  $\square$

**Lemma 4.** We have:

$$n_\alpha \leq n_e \leq 1, \quad 1 \leq \alpha \leq p \quad (4.9)$$

and

$$j = \sum_{\alpha=1}^p j_\alpha \leq \frac{1}{2} \quad (4.10)$$

**Proof.** The bounds (4.9) follow at once from Lemma 3 and the initial condition (2.16). Let us check (4.10). By summing (3.3) with respect to  $\alpha$  and using (2.13), we obtain the conservation law

$$\frac{d}{dx} \left( \sum_{\alpha=1}^p k_\alpha + n_e \right) = 0$$

and by integration from 0 to  $x$

$$\sum_{\alpha=1}^p k_{\alpha} = 1 - n_e. \quad (4.11)$$

On the other hand, the neutrality condition (2.13) can be equivalently written

$$\sum_{\alpha=1}^p \frac{j_{\alpha}^2}{k_{\alpha}} = n_e. \quad (4.12)$$

Then summing (4.11) and (4.12) gives

$$\sum_{\alpha=1}^p j_{\alpha} \left( \frac{j_{\alpha}}{k_{\alpha}} + \frac{k_{\alpha}}{j_{\alpha}} \right) = 1.$$

Since by (4.2) and (4.3)

$$\frac{j_{\alpha}}{k_{\alpha}} > 0 \quad \text{for } x > 0$$

and

$$a + \frac{1}{a} \geq 2 \quad \text{for all } a > 0,$$

this implies the bound (4.10).  $\square$

Next, we want to make more precise the behavior of  $n_e$  in a neighborhood of  $x = 0$ . In all the sequel the following function

$$h(x) = -2 \int_0^x y^2 \left( \frac{1}{n_e} \frac{dn_e}{dx} \right) (y) dy \quad (4.13)$$

will be frequently used. We can then state

**Lemma 5.** For all  $x \geq 0$  such that  $n_e(x) \geq \frac{1}{2}$ , we have

$$1 - 4g(1)^2 x^2 \leq n_e(x) \leq 1 - g\left(\frac{1}{2}\right) \left\{ \max_{1 \leq \alpha \leq p} \frac{g_{\alpha}(\frac{1}{2})^2}{g_{\alpha}(1)} \right\} x^2 \quad (4.14)$$

and

$$g_{\alpha}\left(\frac{1}{2}\right) \left\{ \max_{1 \leq \beta \leq p} \frac{g_{\beta}(\frac{1}{2})^2}{g_{\beta}(1)} \right\} x^2 \leq k_{\alpha}(x) \leq 4g_{\alpha}(1)g(1)x^2. \quad (4.15)$$

**Proof.** Since  $g_{\alpha}$  is an increasing function, we have for  $n_e \in [\frac{1}{2}, 1]$

$$g_{\alpha}\left(\frac{1}{2}\right) \leq g_{\alpha}(n_e) \leq g_{\alpha}(1)$$



and by (4.1)

$$g_\alpha\left(\frac{1}{2}\right)x \leq j_\alpha(x) \leq g_\alpha(1)x. \quad (4.16)$$

Then, using the bounds (4.4) and (4.16) together with (4.13), (4.5) yields

$$g_\alpha\left(\frac{1}{2}\right)\sqrt{h} \leq k_\alpha \leq g_\alpha(1)\sqrt{h} \quad (4.17)$$

so that we obtain by (4.11)

$$1 - g(1)\sqrt{h} \leq n_e \leq 1 - g\left(\frac{1}{2}\right)\sqrt{h}. \quad (4.18)$$

It remains to derive upper and lower bounds for  $h$ . We begin with the upper bound. Assuming  $n_e(x) \geq \frac{1}{2}$ , we have

$$h(x) \leq -4 \int_0^x y^2 \frac{dn_e}{dx}(y) dy$$

and therefore,

$$h(x) \leq 4(1 - n_e(x))x^2. \quad (4.19)$$

Together with the first inequality (4.18), this implies the first inequality (4.14) and therefore

$$h(x) \leq 16 g(1)^2 x^4. \quad (4.20)$$

For obtaining a lower bound for  $h$ , we start from

$$n_e = \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha} \leq 1$$

which yields

$$k_\alpha \geq j_\alpha^2, \quad 1 \leq \alpha \leq p.$$

Using again (4.5) and the estimates (4.16), (4.17), we obtain

$$g_\alpha(1)^2 h \geq k_\alpha^2 \geq j_\alpha^4 \geq g_\alpha\left(\frac{1}{2}\right)^4 x^4, \quad 1 \leq \alpha \leq p$$

which yields

$$h(x) \geq \left\{ \max_{1 \leq \alpha \leq p} \frac{g_\alpha\left(\frac{1}{2}\right)^4}{g_\alpha(1)^2} \right\} x^4. \quad (4.21)$$

Now the second bound (4.14) and the bounds (4.15) follow from (4.17), (4.18), (4.20) and (4.21).  $\square$

Let us next estimate the first and second derivatives of  $n_e$  in a neighborhood of  $x = 0$ . **Lemma 6.** There exists a constant  $C > 0$  depending only on  $g_\alpha\left(\frac{1}{2}\right)$ ,  $g_\alpha(1)$ ,  $1 \leq \alpha \leq p$ , such that we have for  $x \geq 0$  small enough

$$-\frac{dn_e}{dx}(x) \leq Cx \quad (4.22)$$

**Proof.** Observe that (4.6) may be equivalently written

$$-\frac{dn_e}{dx} = \frac{A}{B} \quad (4.23)$$

with

$$A = 2n_e \sum_{\alpha=1}^p \frac{j_\alpha}{k_\alpha} g_\alpha(n_e), \quad (4.24)$$

$$B = \sum_{\alpha=1}^p \frac{j_\alpha^4}{k_\alpha} - n_e. \quad (4.25)$$

Denoting by  $C_i > 0$ ,  $i = 1, 2, \dots$ , various constants depending only on  $g_\alpha(\frac{1}{2})$ ,  $g_\alpha(1)$ ,  $1 \leq \alpha \leq p$ , and using the bounds (4.9), (4.15) and (4.16) we obtain on one hand

$$A \leq \frac{C_1}{x}. \quad (4.26)$$

On the other hand, we find

$$\sum_{\alpha=1}^p \frac{j_\alpha^4}{k_\alpha^3} \geq \frac{C_2}{x^2},$$

and for  $x$  small enough

$$B \geq \frac{C_3}{x^2}. \quad (4.27)$$

The estimate (4.22) follows.  $\square$

Note that all the previous bounds hold if the  $g'_\alpha$ s are  $C^0$  functions. In order to obtain an estimate for  $\frac{d^2 n_e}{dx^2}$ , we need to assume more regularity on the functions  $g_\alpha$ : we thus suppose as in (3.9) that each  $g_\alpha$  is a  $C^1$  function but it would be enough to assume  $g_\alpha \in W^{1,\infty}(0, 1)$ .

**Lemma 7.** We have for  $x \geq 0$  small enough

$$\left| \frac{d^2 n_e}{dx^2}(x) \right| \leq C \quad (4.28)$$

where  $C > 0$  is a constant which depends only on  $g_\alpha(\frac{1}{2})$ ,  $g_\alpha(1)$  and  $\|g'_\alpha\|_{L^\infty(\frac{1}{2}, 1)}$ ,  $1 \leq \alpha \leq p$ .

**Proof.** We proceed as in the previous lemma. Differentiating (4.23) gives

$$-\frac{d^2 n_e}{dx^2} = \frac{1}{B} \frac{dA}{dx} - \frac{A}{B^2} \frac{dB}{dx}. \quad (4.29)$$

Hence we need only to estimate  $\left| \frac{dA}{dx} \right|$  and  $\left| \frac{dB}{dx} \right|$ . We have

$$\begin{aligned} \frac{dA}{dx} &= 2 \frac{dn_e}{dx} \sum_{\alpha=1}^p \frac{j_\alpha g_\alpha}{k_\alpha} + 2n_e \sum_{\alpha=1}^p \left\{ \frac{g_\alpha}{k_\alpha} \frac{dj_\alpha}{dx} + \frac{j_\alpha}{k_\alpha} g'_\alpha \frac{dn_e}{dx} - \frac{j_\alpha g_\alpha}{k_\alpha^2} \frac{dk_\alpha}{dx} \right\} \\ &= 2 \frac{dn_e}{dx} \sum_{\alpha=1}^p \frac{j_\alpha g_\alpha}{k_\alpha} + 2n_e \sum_{\alpha=1}^p \left\{ \frac{g_\alpha^2}{k_\alpha} + \frac{j_\alpha}{k_\alpha} g'_\alpha \frac{dn_e}{dx} + \frac{j_\alpha^3 g_\alpha}{k_\alpha^3} \frac{1}{n_e} \frac{dn_e}{dx} \right\} \end{aligned}$$

which yields

$$\left| \frac{dA}{dx} \right| \leq \frac{C_1}{x^2}. \quad (4.30)$$

Similarly, we can write

$$\begin{aligned} \frac{dB}{dx} &= \sum_{\alpha=1}^p \left\{ \frac{4j_\alpha^3}{k_\alpha^3} \frac{dj_\alpha}{dx} - \frac{3j_\alpha^4}{k_\alpha^4} \frac{dk_\alpha}{dx} \right\} - \frac{dn_e}{dx} = \\ &= \sum_{\alpha=1}^p \left\{ \frac{4j_\alpha^3 g_\alpha}{k_\alpha^3} + \frac{3j_\alpha^6}{k_\alpha^5} \frac{1}{n_e} \frac{dn_e}{dx} \right\} - \frac{dn_e}{dx} \end{aligned}$$

which yields

$$\left| \frac{dB}{dx} \right| \leq \frac{C_2}{x^3}. \quad (4.31)$$

The desired bound (4.28) follows from (4.29) and the inequalities (4.26), (4.27) and (4.30), (4.31).  $\square$

## 5. Proof of Theorem 2.

We begin by characterizing  $n_e$  as the solution of a nonlinear integro-differential equation. Using (4.5), we have

$$k_\alpha = \sqrt{-2 \int_0^x \left( \frac{j_\alpha^2}{n_e} \frac{dn_e}{dx} \right)(y) dy}. \quad (5.1)$$

Hence, writing the neutrality condition (2.13) as

$$n_e = \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha},$$

we obtain that  $n_e$  is solution of the equation

$$n_e = \sum_{\alpha=1}^p \frac{j_\alpha^2}{\sqrt{-2 \int_0^x \left( \frac{j_\alpha^2}{n_e} \frac{dn_e}{dx} \right)(y) dy}} \quad (5.2)$$

where  $j_\alpha$  is given in terms of  $n_e$  by (4.1).

Now, the proof of Theorem 1 consists of two main steps. In the first step, we show that the conclusions of the theorem hold as soon as one knows a (physically admissible) solution of the problem defined in a neighborhood of  $x = 0$ . The second step is essentially devoted to the existence of such a local solution.

Let us thus assume that there exists a solution of (2.13)-(2.17) in a neighborhood of  $x = 0$ . Denote by  $[0, x_0)$  the maximal interval of existence of the solution and let us check that we have  $x_0 < +\infty$ . We proceed by contradiction. Assume in the contrary  $x_0 = +\infty$ . We observe that, by (3.2) and (3.9), each function  $j_\alpha$  is strictly increasing. Moreover using in addition (4.4), we have

$$\frac{d^2 j_\alpha}{dx} = g'_\alpha(n_e) \frac{dn_e}{dx} \leq 0$$

so that  $j_\alpha$  is also concave. Then, (4.10) yields

$$\lim_{x \rightarrow \infty} j(x) \leq \frac{1}{2}.$$

Let us next show that this is indeed impossible. We first consider the case where  $g(0) > 0$ . We obtain

$$j(x) = \int_0^x g(n_e(y)) dy \geq g(0)x$$

so that  $j(x) \geq \frac{1}{2}$  for  $x$  large enough which is excluded. We pass to the case where  $g(0) = 0$ . The function  $j$  being strictly increasing and concave, we have necessarily

$$\lim_{x \rightarrow \infty} \frac{dj}{dx}(x) = 0,$$

and therefore

$$\lim_{x \rightarrow \infty} g(n_e(x)) = 0.$$

Using the hypothesis (3.9), we obtain

$$\lim_{x \rightarrow \infty} n_e(x) = 0.$$

On the other hand, since  $j_\alpha$  is increasing and bounded above, we have

$$\lim_{x \rightarrow \infty} j_\alpha(x) = j_\alpha(\infty) > 0.$$

Similarly, each function  $k_\alpha$  is a strictly increasing function since by (3.3), (3.8) and (4.4)

$$\frac{dk_\alpha}{dx}(x) > 0 \quad \text{for all } x > 0.$$

Moreover, it follows from (4.11) that  $k_\alpha$  is bounded above so that

$$\lim_{x \rightarrow \infty} k_\alpha(x) = k_\alpha(\infty) > 0.$$

Hence we obtain

$$\lim_{x \rightarrow \infty} n_e(x) = \sum_{\alpha=1}^p \frac{j_\alpha^2(\infty)}{k_\alpha(\infty)} > 0$$

which leads again to a contradiction. Thus, we have necessarily  $x_0 < +\infty$ .

Let us next check that

$$U(x_0) = \lim_{x \rightarrow x_0} U(x)$$

exists. Using again the fact that  $j_\alpha$  and  $k_\alpha$  are increasing functions which are bounded above, we have

$$\lim_{x \rightarrow x_0} j_\alpha(x) = j_\alpha(x_0) > 0, \quad \lim_{x \rightarrow x_0} k_\alpha(x) = k_\alpha(x_0) > 0$$

and therefore

$$\lim_{x \rightarrow x_0} n_\alpha(x) = n_\alpha(x_0) = \frac{j_\alpha^2(x_0)}{k_\alpha(x_0)} > 0$$

which proves our assertion.

The matrix  $A(U(x_0))$  is necessarily singular. Otherwise, from the initial condition  $U(x_0)$  at  $x_0$ , one could extend the solution of (3.4) beyond  $x_0$ , which is excluded since  $[0, x_0)$  is the maximal interval of existence of the solution. Then it follows from (3.6) that

$$\sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2}(x_0) = n_e(x_0). \quad (5.3)$$

On the other hand, we find by passing to the limit in (4.11)

$$\sum_{\alpha=1}^p k_\alpha(x_0) = \sum_{\alpha=1}^p (n_\alpha u_\alpha^2)(x_0) = 1 - n_e(x_0).$$

This proves (3.13). As a consequence we have at  $x_0$

$$\sum_{\alpha=1}^p n_\alpha(u_\alpha^2 + \frac{1}{u_\alpha^2}) = 1$$

which yields the inequality (3.14).

Let us then prove (3.15). By passing to the limit in (4.8) and using (4.4) together with (5.3), we first obtain

$$\lim_{x \rightarrow x_0} \frac{dn_e}{dx}(x) = -\infty$$

and by (3.3)

$$\lim_{x \rightarrow x_0} \frac{dk_\alpha}{dx}(x) = -\frac{n_\alpha(x_0)}{n_e} \lim_{x \rightarrow x_0} \frac{dn_e}{dx}(x) = +\infty.$$

Since by (3.2)

$$\frac{dk_\alpha}{dx} = \frac{d}{dx}(u_\alpha j_\alpha) = u_\alpha \frac{dj_\alpha}{dx} + j_\alpha \frac{du_\alpha}{dx} = u_\alpha g_\alpha(n_e) + j_\alpha \frac{du_\alpha}{dx}$$

we get

$$\lim_{x \rightarrow x_0} \frac{du_\alpha}{dx}(x) = \frac{1}{j_\alpha(x_0)} \left\{ \lim_{x \rightarrow x_0} \frac{dk_\alpha}{dx} - u_\alpha(x_0) g_\alpha(n_e(x_0)) \right\} = +\infty.$$

On the other hand, since

$$\frac{dj_\alpha}{dx} = u_\alpha \frac{dn_\alpha}{dx} + n_\alpha \frac{du_\alpha}{dx} = g_\alpha(n_e),$$

we have

$$\lim_{x \rightarrow x_0} \frac{dn_\alpha}{dx}(x) = \frac{1}{u_\alpha(x_0)} \left\{ g_\alpha(n_e(x_0)) - n_\alpha(x_0) \lim_{x \rightarrow x_0} \frac{du_\alpha}{dx}(x) \right\} = -\infty.$$

We thus have proved the properties (iii)-(v) of the theorem.

We pass to the second step of the proof. It remains to show the existence of a (physically admissible) solution in a neighborhood of  $x = 0$  which satisfies the properties (i) and (ii) of the theorem. This is far from being obvious since the differential system (3.4) is singular at  $x = 0$ . We begin by constructing an approximate solution of (2.13)-(2.17) in the following way. For any  $\eta > 0$  arbitrarily small and for  $1 \leq \alpha \leq p$ , we introduce an approximation  $g_\alpha^\eta$  of  $g_\alpha$  satisfying the properties

$$g_\alpha^\eta(n_e) = g_\alpha(1), \quad 1 - \eta \leq n_e \leq 1, \quad (5.4)$$

$$g_\alpha^\eta \text{ is bounded in } W^{1,\infty}(0, 1), \quad (5.5)$$

$$g_\alpha^\eta \rightarrow g_\alpha \text{ in } C^0([0, 1]) \text{ as } \eta \rightarrow 0. \quad (5.6)$$

This is indeed possible since we have assumed  $g_\alpha \in C^1([0, 1])$ . We then solve the problem (2.13)-(2.17) with  $g_\alpha$  replaced by  $g_\alpha^\eta$ . We know from Example 2 that such a problem has a unique solution as long as the corresponding electron density  $n_e^\eta$  satisfies  $n_e^\eta \geq 1 - \eta$ . At the point  $a^\eta$  such that

$$n_e^\eta(a^\eta) = 1 - \eta,$$

the first part of the proof ensures that we can extend the solution *up* to a point  $x_0^\eta$  with.

$$n_e^\eta(x_0^\eta) \leq \frac{1}{2}.$$

Let  $U^\eta$  be the solution thus obtained; we want to prove that  $U^\eta$ , or at least a subsequence, converges towards a solution of (3.4), (3.5), or (2.13)-(2.17), as  $\eta$  tends to zero.

First, we show that  $U^\eta$  is defined in a fixed interval  $[0, x_1]$  independent of  $\eta$ . Indeed, using the first inequality (4.14), we have

$$n_e^\eta(x) \geq 1 - 4g^\eta(1)^2 x^2 = 1 - 4g(1)^2 x^2.$$

Hence for

$$1 - 4g(1)^2 x^2 \geq \frac{1}{2},$$

i.e., for

$$x \leq \frac{1}{2\sqrt{2}g(1)},$$

we have

$$n_e^\eta(x) \geq \frac{1}{2}.$$

The approximate solution  $U^\eta$  is thus defined in an interval containing  $\left[0, \frac{1}{2\sqrt{2}g(1)}\right]$  and

we can choose for  $x_1$  any number less than  $\frac{1}{2\sqrt{2}g(1)}$ .

Let us next show that one can extract from  $(n_e^\eta)$  a subsequence which converges towards a solution of (5.2). Choosing  $x_1$  small enough, we deduce from Lemmas 4, 6 and 7 that  $n_e^\eta$  remains in a bounded set of  $W^{2,\infty}(0, x_1)$  as  $\eta$  tends to zero. Hence we can extract a subsequence still denoted by  $(n_e^\eta)$  such that

$$n_e^\eta \longrightarrow n_e \quad \text{in } C^1([0, x_1]).$$

By passing to the limit, we obtain that  $n_e$  is a decreasing function solution of (5.2) which satisfies

$$\begin{aligned} n_e(0) &= 1, \\ \frac{1}{2} &\leq n_e(x) \leq 1 \quad \text{for all } x \in [0, x_1] \end{aligned}$$

together with the estimates (4.14) and (4.22). In addition,  $j_\alpha$  defined from  $n_e$  by (4.1) satisfies the estimates (4.16) while  $k_\alpha$  defined from  $n_e$  by (5.1) satisfies the estimates (4.15). Setting

$$n_\alpha = \frac{j_\alpha^2}{k_\alpha} = \frac{j_\alpha^2}{\sqrt{-2 \int_0^x (j_\alpha^2 \frac{dn_e}{dx})(y) dy}}.$$

we find easily that  $U$  is a solution of the problem (3.4), (3.5) in  $[0, x_1]$  or equivalently  $\{(n_\alpha, u_\alpha); 1 \leq \alpha \leq p\}$  is a solution of the equations (2.13)-(2.17).

Let us then check that  $U$  is a physically admissible solution. Since  $g_\alpha$  is assumed to be a  $C^1$  function,  $j_\alpha$  a  $C^2$  function. Moreover we have

$$j_\alpha(x) > 0 \quad \text{for all } x > 0.$$

Next it follows from (5.1) and (4.15) that  $k_\alpha$  is a  $C^1$  function which satisfies

$$\begin{cases} k_\alpha(0) = \frac{dk_\alpha}{dx}(0) = 0, \\ k_\alpha(x) > 0 \quad \text{for all } x > 0. \end{cases}$$

Then  $n_\alpha$  is clearly a  $C^1$  function for  $x > 0$  and

$$n_\alpha(x) > 0 \quad \text{for all } x > 0.$$

It remains to analyze the behavior of  $n_\alpha$  at  $x = 0$ . We have by (5.5)

$$|g_\alpha(n_e) - g_\alpha(1)| \leq C_1 |n_e - 1| \leq C_2 x^2$$

and therefore

$$j_\alpha(x) = g_\alpha(1)x + O(x^3).$$

Next, using (5.1), we have

$$\begin{aligned} k_\alpha^2(x) &= -2 \int_0^x \{g_\alpha(1)^2 y^2 + O(y^4)\} \left( \frac{1}{n_e} \frac{dn_e}{dx} \right)(y) dy = \\ &= g_\alpha(1)^2 h(x) + O(x^6). \end{aligned}$$

which gives since  $h(x) = O(x^4)$

$$k_\alpha = g_\alpha(1)\sqrt{h} + O(x^4).$$

We thus obtain

$$n_\alpha = \frac{j_\alpha^2}{k_\alpha} = \frac{g_\alpha(1)^2 x^2 + O(x^4)}{g_\alpha(1)\sqrt{h} + O(x^4)} = g_\alpha(1) \frac{x^2}{\sqrt{h}} + O(x^2)$$

and

$$n_e = \sum_{\alpha=1}^p n_\alpha = g(1) \frac{x^2}{\sqrt{h}} + O(x^2) = 1 + O(x^2).$$

This yields

$$\sqrt{h} = g(1)x^2 + O(x^4) \tag{5.7}$$

and

$$n_\alpha = \frac{g_\alpha(1)}{g(1)} + O(x^2).$$

Hence  $n_\alpha$  is continuously differentiable at  $x = 0$  and we have (3.11).

It remains only to prove that  $n_e$  is a  $C^2$  function. We use (4.23)-(4.25). Since  $A$  and  $B$  are  $C^1$  functions for  $x > 0$  and

$$B = \sum_{\alpha=1}^p \frac{j_\alpha^4}{k_\alpha^3} - n_e = \sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2} - n_e$$

is positive in  $[0, x_1]$  by Lemma 1, we obtain that  $\frac{dn_e}{dx}$  is a  $C^1$  function for  $x > 0$ . On the other hand, using (4.11) and (5.7) we have

$$1 - n_e = \sum_{\alpha=1}^p g_\alpha(1)\sqrt{h} + O(x^4) = g(1)^2 x^2 + O(x^4).$$

Hence  $n_e$  is twice continuously differentiable at  $x = 0$  and we have (3.12). This concludes the proof of Theorem 2.  $\square$



## 6. Numerical Results

We present in this section numerical simulations corresponding to various ionization rates and various numbers of ion species. We also compare the results provided by a one-velocity fluid model and a kinetic model.

We first describe the numerical method of solution of the plasma approximation equations (2.14)-(2.17). It consists in solving the plasma equation (5.2). Since  $n_e(x)$  is a strictly decreasing function on  $[0, x_0[$ , we can look equivalently for the inverse function  $x(n_e)$  for  $n_e \leq 1$ . We write

$$n_e = \sum_{\alpha=1}^p \frac{j_\alpha^2(n_e)}{k_\alpha(n_e)}. \quad (6.1)$$

As  $n_e(0) = 1$ , we have by (4.1)

$$j_\alpha(n_e) = - \int_{n_e}^1 g_\alpha(n) \frac{dx(n)}{dn_e} dn, \quad (6.2)$$

and by (5.1)

$$k_\alpha(n_e) = \sqrt{2 \int_{n_e}^1 \frac{j_\alpha^2(n)}{n} dn}. \quad (6.3)$$

Note also that the Theorem 2 yields the following behaviors for  $n_e$ ,  $j_\alpha$  and  $k_\alpha$  as  $x \rightarrow 0$ :

$$n_e(x) \approx 1 - g(1)^2 x^2, \quad j_\alpha(x) \approx g_\alpha(1)x, \quad k_\alpha(x) \approx g_\alpha(1)g(1)x^2.$$

Hence, we obtain as  $n_e \rightarrow 1$

$$x(n_e) \approx \frac{1}{g(1)} \sqrt{1 - n_e}, \quad j_\alpha(n_e) \approx \frac{g_\alpha(1)}{g(1)} \sqrt{1 - n_e}, \quad k_\alpha(n_e) \approx g(1)g_\alpha(1)(1 - n_e). \quad (6.4)$$

In order to solve the equations (6.1)-(6.3), we introduce a mesh in the variable  $n_e$

$$n_e^0 = 1 > n_e^1 > n_e^2 > \dots > n_e^i > n_e^{i+1} > \dots$$

and we set

$$\Delta n^i = n_e^i - n_e^{i+1}.$$

We denote by  $x^i$ ,  $j_\alpha^i$  and  $k_\alpha^i$  the approximate values of  $x(n_e^i)$ ,  $j_\alpha(n_e^i)$  and  $k_\alpha(n_e^i)$  respectively. These sequences are initialized by

$$n^0 = 1, \quad x^0 = 0, \quad j_\alpha^0 = 0, \quad k_\alpha^0 = 0.$$

and for the first index  $i = 1$ , using the expansion (6.4) of the solution near the origin, we take

$$x^1 = \frac{1}{g(1)} \sqrt{\Delta n^0}, \quad j_\alpha^1 = \frac{g_\alpha(1)}{g(1)} \sqrt{\Delta n^0}, \quad k_\alpha^1(x) = g_\alpha(1)g(1)\Delta n^0. \quad (6.5)$$

For  $i \geq 1$ , we compute  $x^{i+1}$ ,  $j_\alpha^{i+1}$  and  $k_\alpha^{i+1}$  from  $x^i$ ,  $j_\alpha^i$  and  $k_\alpha^i$  by

$$\begin{cases} x^{i+1} = x^i + \Delta n^i z^i, \\ j_\alpha^{i+1} = j_\alpha^i - \Delta n^i g_\alpha(n_e^i) z^i \\ k_\alpha^{i+1} = \sqrt{(k_\alpha^i)^2 + \frac{2\Delta n^i}{n_e^i} (j_\alpha^i)^2} \end{cases} \quad (6.6)$$

where  $z^i \simeq x'(n_e^i)$  is solution of the equation

$$n_e^{i+1} = \sum_{\alpha=1}^p \frac{(j_\alpha^{i+1})^2}{k_\alpha^{i+1}} \quad (6.7)$$

Eq. (6.7) is indeed an equation of the second degree in  $z^i$ . Let us now analyze its solvability. For the sake of simplicity, we drop the indices  $i$  and we put a tilde on the quantities corresponding to the index  $i+1$ . Then, taking into account the 2nd equation (6.6), we obtain that  $z$  satisfies

$$(\Delta n)^2 \sum_{\alpha=1}^p \frac{g_\alpha^2(n_e)}{k_\alpha} z^2 - 2\Delta n \sum_{\alpha=1}^p \frac{j_\alpha g_\alpha(n_e)}{k_\alpha} z + \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha} - n_e + \Delta n = 0 \quad (6.8)$$

Since  $x'(n_e) < 0$ , we look for the negative root (if it exists) of (6.8). The sum of the roots being positive, such a negative root exists if and only if

$$\sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha} - n_e + \Delta n < 0.$$

Since

$$n_e = \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha},$$

the above condition reads

$$\Delta n < H(\Delta n) \quad (6.9)$$

where

$$H(\xi) = \sum_{\alpha=1}^p \frac{j_\alpha^2}{k_\alpha} (1 - (1 + 2 \frac{j_\alpha^2 \xi}{n_e k_\alpha^2})^{-1/2}). \quad (6.10)$$

Note that  $H(0) = 0$  and  $H'$  is a strictly decreasing function on  $\mathbb{R}^+$  while  $H''$  is a strictly increasing one. Hence,

$$\Delta n H'(0) + \frac{(\Delta n)^2}{2} H''(0) \leq H(\Delta n) \leq \Delta n H'(0). \quad (6.11)$$

First, using the 2nd inequality of (6.11), we obtain that

$$H'(0) \geq 1,$$

is a necessary condition for (6.9) to hold. This condition reads

$$\frac{1}{n_e} \sum_{\alpha=1}^p \frac{j_\alpha^4}{k_\alpha^3} \geq 1,$$

or equivalently

$$\sum_{\alpha=1}^p \frac{n_\alpha}{u_\alpha^2} \geq n_e.$$

This means that  $x(n_e)$  belongs indeed to the interval of existence  $[0, x_0]$  of the plasma approximation. Next, using the first inequality of (6.11), we find that

$$\Delta n \leq \Delta n H'(0) + \frac{(\Delta n)^2}{2} H''(0),$$

is a sufficient condition for (6.9) to hold. It reads

$$\Delta n \leq \frac{2n_e}{3} \frac{\sum_{\alpha=1}^p \frac{j_\alpha^4}{k_\alpha^3} - n_e}{\sum_{\alpha=1}^p \frac{j_\alpha^6}{k_\alpha^5}} \quad (6.12)$$

and provides an upper bound for  $\Delta n$  at each step of the computation. The discriminant of the second order equation 6.8 reads

$$\Delta = 4(\Delta n)^2 \left\{ \left( \sum_{\alpha=1}^p \frac{j_\alpha g_\alpha}{\sqrt{k_\alpha}} \right)^2 - \left( \sum_{\alpha=1}^p \frac{g_\alpha^2}{\sqrt{k_\alpha}} \right) \left( \sum_{\alpha=1}^p \frac{j_\alpha^2}{\sqrt{k_\alpha}} - n_e + \Delta n \right) \right\}$$

and it is obviously positive according to (6.12). The other physical quantities (partial density and velocity for the  $\alpha$ -th species) are then obtained easily by

$$n_\alpha^i = (j_\alpha^i)^2 / k_\alpha^i, \quad u_\alpha^i = k_\alpha^i / j_\alpha^i. \quad (6.13)$$

Let us first present numerical results corresponding to a hydrogen-helium plasma, in equal part, with two ion species  $H_2^+$  and  $H_e^+$ . We assume that the ionization process is driven by primary electrons of high energy ( $1keV$ ) and by secondary electrons of low energy ( $16eV$ ) which have been produced through the ionization of neutrals by primary electrons and then heated. The case  $\gamma = 0$  corresponds to ionization by primary electrons while the case  $\gamma = 1$  corresponds to ionization by the secondary electrons as in the definition of the ionization rates  $g$ . The ionization coefficients taken from <sup>7</sup> are given by

$$\begin{aligned} H_2^+ &: 3.85 \cdot 10^{-8} cm^3 s^{-1} (primary), \quad 1.66 \cdot 10^{-8} cm^3 s^{-1} (secondary) \\ H_e^+ &: 2.4 \cdot 10^{-8} cm^3 s^{-1} (primary), \quad 3.6 \cdot 10^{-9} cm^3 s^{-1} (secondary) \end{aligned}$$

and are related to the cross section  $\sigma$  and the velocity  $V$  of the electrons (defined from their energy  $1keV$  or  $16eV$ ) by

$$S = \sigma V.$$

Since the densities  $n_a$  of the target particle in the ionization process (hydrogen and helium) are supposed constants and equal, the collision frequencies  $\nu$  in the definition of  $g$  are proportionnal to the above ionization coefficients  $S$

$$\nu = n_a S.$$

Thus, the *scaled* ionization rates are of the form:

case 1:  $p = 2$ ,  $g_1(n_e) = 2.4 + 0.36n_e$ ,  $g_2(n_e) = 3.85 + 1.66n_e$ .

We have plotted the following quantities: ion densities  $n_\alpha$  and electron density  $n_e$  (Fig. 1), ion velocities  $u_\alpha$  (Fig. 2) and current densities  $j_\alpha$  and  $j$  (Fig. 3).

In fact, we are interested more specifically by the values of the physical quantities at the point  $x_0$  (characterized in Theorem 2) since these values are to be taken as injection parameters for an ion beam extracted from the neutral plasma. We obtain

$$\begin{aligned} n_1 &= 0.1695, & j_1 &= 0.1690, & u_1 &= 0.9974, \\ n_2 &= 0.3304, & j_2 &= 0.3309, & u_2 &= 1.0013, \\ n_e &= 0.499996, & j &= 0.4999987. \end{aligned}$$

Observe that the nondimensional velocity  $u_\alpha$  of each species is close to 1 and moreover

$$n_e(x_0) \approx \frac{1}{2}, \quad j(x_0) \approx \frac{1}{2}. \quad (6.14)$$

Hence, the total extracted current density is approximatively given by Bohm's criterion as in the single ion species case considered in Example 1. Somewhat surprisingly, this seems to be a general feature of this model, i.e., (6.14) holds true. For illustrating this unexplained property, we have considered three cases (which have not necessarily a physical meaning) corresponding to highly different numbers of species or ionization rates:

- (i) case 2:  $p = 2$ ,  $g_1(n) = 1$ ,  $g_2(n) = n$ .
- (ii) case 3:  $p = 2$ ,  $g_1(n) = 1$ ,  $g_2(n) = 10^6 n^2$ ,
- (iii) case 4:  $p = 10$ ,  $g_\alpha(n) = C_\alpha n^\alpha$ ,  $\alpha = 1, \dots, 10$ , with the following arbitrary values for the  $C_\alpha$ :  
1, 1, 0.5, 3, 1, 0.25, 8, 0.5, 3, 100.

We give the values of the (scaled) electronic density and of the extracted current at  $x_0$

Case	$x_0$	$n_e$ $u_1$	$j$ $u_2$	$j_1$ $n_1$	$j_2$ $n_2$
2	$2.6514 \cdot 10^{-01}$	$4.998 \cdot 10^{-01}$ $9.8766 \cdot 10^{-01}$	$4.9995 \cdot 10^{-01}$ $1.01490 \cdot 10^{+00}$	$2.6514 \cdot 10^{-01}$ $2.6845 \cdot 10^{-01}$	$2.3481 \cdot 10^{-01}$ $2.31363 \cdot 10^{-01}$
3	$6.7611 \cdot 10^{-07}$	$4.9002 \cdot 10^{-01}$ $9.3237 \cdot 10^{-01}$	$4.9002 \cdot 10^{-01}$ $1.0000 \cdot 10^{+00}$	$6.7611 \cdot 10^{-07}$ $7.2515 \cdot 10^{-07}$	$4.90021 \cdot 10^{-01}$ $4.9002 \cdot 10^{-01}$
4	$1.9419 \cdot 10^{-02}$	$4.9086 \cdot 10^{-01}$ $7.6288 \cdot 10^{-01}$	$4.9803 \cdot 10^{-01}$ $8.0728 \cdot 10^{-01}$	$1.9419 \cdot 10^{-02}$ $2.5455 \cdot 10^{-02}$	$1.44308 \cdot 10^{-03}$ $1.7875 \cdot 10^{-02}$

In the case 4, we give below the partial currents, velocities and densities for the ten species

$$j_\alpha \quad (1.9420 \cdot 10^{-02}, 1.4430 \cdot 10^{-02}, 5.5456 \cdot 10^{-03}, 2.640 \cdot 10^{-02}, 7.1979 \cdot 10^{-03}, \\ 1.5110 \cdot 10^{-03}, 4.1559 \cdot 10^{-02}, 2.2775 \cdot 10^{-03}, 1.2187 \cdot 10^{-02}, 3.6749 \cdot 10^{-01}),$$

$$u_\alpha \quad (7.6288 \cdot 10^{-01}, 8.0728 \cdot 10^{-01}, 8.5101 \cdot 10^{-01}, 8.9260 \cdot 10^{-01}, 9.3089 \cdot 10^{-01}, \\ 9.6513 \cdot 10^{-01}, 9.9504 \cdot 10^{-01}, 1.0207 \cdot 10^{+00}, 1.0424 \cdot 10^{+00}, 1.0607 \cdot 10^{+00})$$

$$n_\alpha \quad (2.5455 \cdot 10^{-02}, 1.7875 \cdot 10^{-02}, 6.5165 \cdot 10^{-03}, 2.9587 \cdot 10^{-02}, 7.7322 \cdot 10^{-03}, \\ 1.5656 \cdot 10^{-03}, 4.1766 \cdot 10^{-02}, 2.2313 \cdot 10^{-03}, 1.1691 \cdot 10^{-02}, 3.4644 \cdot 10^{-01}).$$

Let us now compare the results obtained by using this fluid model with those obtained using two other fluid and kinetic related models. As a reference case, we will choose the case 2 considered above.

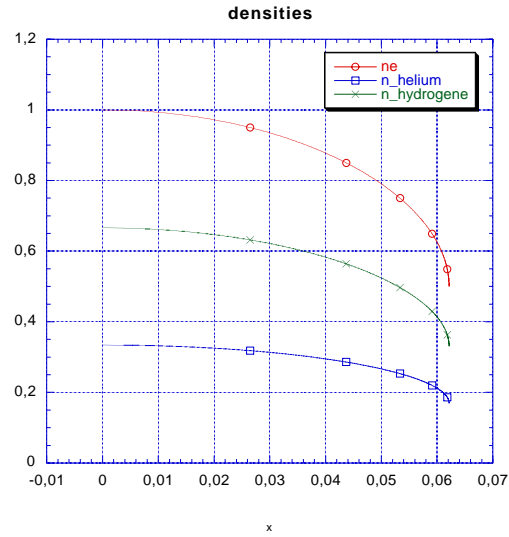


Fig. 1. Case  $H^+ - H_e$  - densities

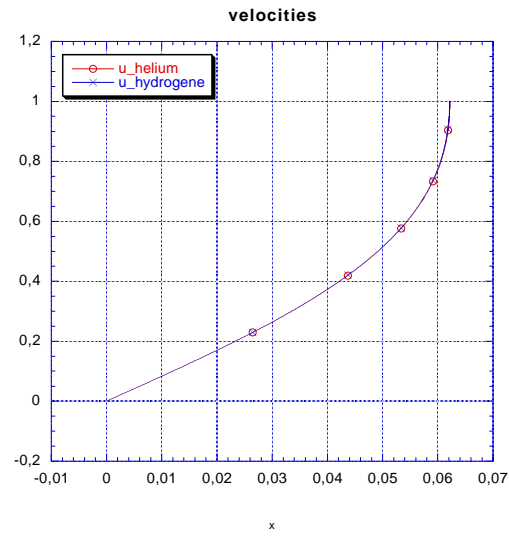
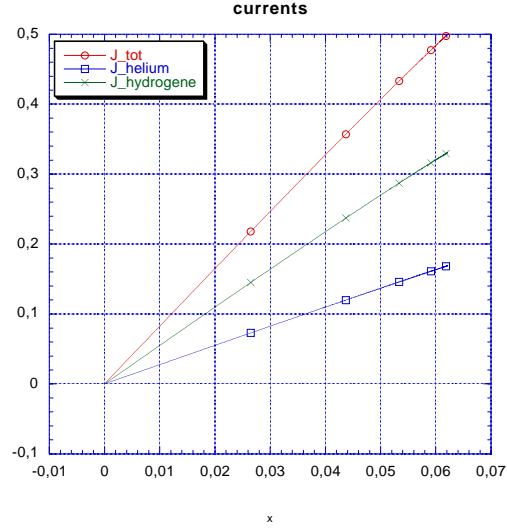


Fig. 2. Case  $H^+ - H_e$  - velocities

Fig. 3. Case  $H^+ - H_e$  - currents

a) *A one-velocity fluid model.* We have already observed that the *scaled* velocities  $u_\alpha$  are approximatively equal. Hence, it makes sense to consider the following one-velocity plane model under the same physical hypotheses (stationarity, cold ions)

$$\begin{cases} \frac{d(n_e u)}{dx} = g_\alpha(n_e), \\ \frac{d(n_e u^2)}{dx} + \frac{n_e}{n_e} \frac{dn_e}{dx} = 0, \\ n_e(0) = 1, u(0) = 1. \end{cases} \quad (6.15)$$

We note that the pair  $(n_e, u)$  is indeed the solution of the model (2.16) considered in Example 1. Thus, we have:

$$n_e(x_0) = 1/2, \quad j(x_0) = 1/2.$$

In the case 2, i.e.,  $p = 2$ ,  $g_1 = 1$  and  $g_2 = n$ , the function  $x(n)$  can be determined explicitly

$$x(n) = \frac{\pi}{2} - \sin^{-1}(2n - 1) + \frac{3\sqrt{2}}{4} \left( \sin^{-1} \left( \frac{3n - 1}{n + 1} \right) - \frac{\pi}{2} \right)$$

which gives the following values

$$\begin{cases} n_1 = j_1 = \int_0^{x_0} g_1(n(y)) dy = x_0 = x(1/2) = \frac{\pi}{2} + \frac{3\sqrt{2}}{4} \left( \sin^{-1} \left( \frac{1}{3} \right) - \frac{\pi}{2} \right) \approx 0.265..., \\ n_2 = j_2 = j - j_1 \approx 0.235. \end{cases}$$

which are very close to those provided by the multifluid model.

<sup>†</sup>Note however that the physical velocities  $U_\alpha = \sqrt{\frac{Z_\alpha k_B T_e}{m_\alpha}} u_\alpha$  are all distinct.

b) *A multi-species kinetic model.* Let us now compare the above results with the numerical values given by a kinetic model. In the multispecies kinetic model introduced in <sup>6</sup>, the fluid equations (1.8)-(1.9) are replaced by the Vlasov equation with ionization source term

$$v \frac{\partial f_\alpha}{\partial x} + \frac{d\varphi}{dx} \frac{\partial f_\alpha}{\partial v} = g_\alpha(n_e) \delta(v), \quad x > 0, \quad v \in \mathbf{R} \quad (6.16)$$

where  $f_\alpha$  is the distribution function of the ion species  $\alpha$  and the boundary conditions (1.11) by

$$f_\alpha(0, v) = 0, \quad v > 0. \quad (6.17)$$

Then, we look for functions  $f_\alpha$ ,  $1 \leq \alpha \leq p$  and  $\varphi$  solutions of (1.7)-(6.16)-(1.10) with the boundary conditions (6.17) and (1.12) where in the Poisson equation (1.10) we have

$$n_\alpha = \int_{-\infty}^{\infty} f_\alpha dv.$$

Now, it is a simple matter to generalize the results of <sup>1</sup> which consider the case of a single ion species to the case of several ion species as explained in <sup>6</sup>. Assuming that the function  $\varphi$  is monotone increasing, we obtain

$$n_\alpha(x) = \int_0^x \frac{g(\exp(-\varphi(y))) dy}{\sqrt{2(\varphi(x) - \varphi(y))}},$$

so that the plasma approximation  $\sum_{\alpha=1}^p n_\alpha = n_e = \exp(-\varphi)$  amounts to find  $\varphi$  solution of

$$\int_0^x \frac{g(\exp(\varphi(y)))}{\sqrt{2(\varphi(x) - \varphi(y))}} dy = \exp(-\varphi(x)), \quad \varphi(0) = 0. \quad (6.18)$$

Again, the plasma equation (6.18) has a solution  $\varphi$  defined in a maximal interval  $[0, x_0)$  with  $x_0 < +\infty$  and we find

$$\varphi(x_0) \approx 0.854, \quad n(x_0) \approx 0.426, \quad j(x_0) \approx 0.486.$$

independently of the ionization rates  $g_\alpha$ . Moreover, in the plasma approximation, choosing  $(\varphi, v)$  instead of  $(x, v)$  as the set of independent variables, we obtain

$$f_\alpha(\varphi, v) = \begin{cases} \frac{\sqrt{2}}{\pi} \frac{g_\alpha(\exp(\frac{v^2}{2} - \varphi))}{g \exp((\frac{v^2}{2} - \varphi))} \left[ \frac{1}{\sqrt{\varphi - \frac{v^2}{2}}} - 2F(\sqrt{\varphi - \frac{v^2}{2}}) \right], & 0 < v < \sqrt{2\varphi} \\ 0 & \text{otherwise.} \end{cases} \quad (6.19)$$

where  $F$  is given by  $F(x) = \exp(-x^2) \int_0^x \exp(t^2) dt$ . In addition, the function  $x(\varphi)$  can be computed by integrating the following ordinary differential equation

$$x'(\varphi) = \frac{2\sqrt{2}}{\pi g(\exp(-\varphi))} \frac{d}{d\varphi} F(\sqrt{\varphi}).$$

In the case 2 where  $g_1 = 1$ ,  $g_2 = n$ , we find that the partial extracted currents are

$$j_1 = x_0 = 0.261116, \quad j_2 = j_\infty - j_1 = 0.225995.$$

Note also that the scaled temperature is equal to 0.046. This small value of the temperature of the solution of the kinetic model is a justification of the hypothesis of cold ions used in the other models.

Let us summarize the density and current at the end of the plasma sheath for each model:

Model	$n(x_0)$	$j(x_0)$	$j_1 = x_0$	$j_2 = j(x_0) - j_1$
1: multifluid (2.13)-(2.17)	0.499	0.499	0.265	0.235
2: fluid, one velocity (6.15)	$0.5^*$	$0.5^*$	0.265	0.235
3: kinetic ()	$0.426^*$	$0.486^*$	0.261	0.226

The quantities with  $*$  are independent of the ionization rates. Note that the results are similar. In particular, the ratio  $j_1/j_2$  is close to 1.1 in all the models.

## 7. Extensions to more complex models

Let us mention some possible extensions of the present analysis. First, it is possible to take into account additional source or friction terms (section 6.1). It should be also interesting to perform a asymptotic analysis of the quasineutral system ( $\varepsilon \rightarrow 0$ ) i.e. to justify from the mathematical point of view the convergence of solutions of system (1.7)-(1.10) toward solutions of system (1.14)-(1.17) we have constructed. Finally, let us point out that the hyperbolicity of the associated evolution problem imposes that the ion velocities remains equal (section 6.2).

### 7.1. Friction terms

We can include friction terms in our quasineutral model without changing the conclusions of Theorem 3. Consider for simplicity the case of a single ion species studied in Example 2. If we replace the second equation (2.16) by

$$\frac{d}{dx}(nu^2 + n) = -f(n, u),$$

where  $f(n, u)$  stands for a friction term which satisfies:

$$f(n, u) \geq 0, \quad \forall n \geq 0, \quad \forall u \in \mathbf{R},$$

we obtain again that the modified system has a unique solution defined on a maximal interval  $[0, x_0)$ ,  $x_0 < +\infty$ , and moreover, we have

$$u(x_0) = 1, \quad n(x_0) \leq 1/2.$$

The proof goes along the same lines, but in that simple case in a much simpler way than the proof of Theorem 2.

### 7.2. Evolution problems.

Consider the nonstationary problem associated with the quasineutral model (1.13)–(1.15). It can be written in the form

$$\begin{cases} \frac{dn_\alpha}{dt} + \frac{d(n_\alpha u_\alpha)}{dx} = g_\alpha(n_e), & n_e = \sum_{\alpha=1}^p n_\alpha, \quad 1 \leq \alpha \leq p, \\ \frac{du_\alpha}{dt} + u_\alpha \frac{du_\alpha}{dx} + \frac{T_e}{n_e} \frac{dn_e}{dx} = -g_\alpha(n_e)u_\alpha/n_\alpha, & 1 \leq \alpha \leq p. \end{cases} \quad (7.1)$$



This first order nonlinear system is clearly hyperbolic for  $p = 1$ . In the case  $p = 2$ , extending the results of <sup>8</sup>, one can show that the system is hyperbolic if and only if  $u_1 = u_2$ . This suggests to pay attention to the one-velocity fluid model (6.15). **8. Conclusions**

As we have seen, despite the fact that the model from the physical point of view is very rough, it is complicated from the mathematical point of view. From the physical point of view, the main approximation in this model is the Maxwell-Boltzmann assumption for the electrons and the mathematical validity of this assumption is an open problem.

#### Acknowledgement

We thank A. Compant La Fontaine for helpful discussions about the hydrogen-helium test.

#### References

1. N. ben Abdallah and S. Mas-Gallic, personal communication.
2. N. ben Abdallah, S. Mas-Gallic and P.A. Raviart, *Analysis and asymptotics of a one-dimensional ion extraction model*, Asymptotic Anal. **10** (1998), pp. 1-28.
3. N. ben Abdallah, S. Mas-Gallic and P.A. Raviart, *A mathematical analysis of electric probe models*, Trans. Th. Stat. Phys. **25** (1996), pp. 263-281.
4. A. Ambroso, F., Mehats and P.A. Raviart, *On singular perturbation problems for the nonlinear Poisson equation*, Research Report , Centre de mathématiques appliquées, Ecole Polytechnique(2000).
5. Y. Brenier and E. Grenier, *Limite singulière du système de Vlasov-Poisson dans le régime de quasi neutralité: Le cas indépendant du temps. (Singular limit of the Vlasov-Poisson system in the quasi-neutral regime: The time independent case)*, C. R. Acad. Sci., Paris, Ser. I **318**, No.2 (1994), pp. 121-124 .
6. C. Buet, S. Cordier and P.A. Raviart, *Multispecies kinetic ionization problems*, Research report , Laboratoire d'analyse numérique, Universit Paris 6 (1999).
7. H. W. Drawin, Z. Physik, vol. 164 (1961), p. 513 .
8. S. Cordier, *Hyperbolicity of the hydrodynamical model of plasmas under the quasineutrality hypothesis*, Math. Models. Appl. Sc. (M2AS), **18** (1995), pp. 627-647.
9. S. Cordier and E. Grenier *Quasineutral limit of an Euler-Poisson system arising from plasma physics*, Commun. Partial Differ. Equations **25**, No.5-6 (2000), pp. 1099-1113.
10. A.F. Forrester, **Large Ions Beams: Fundamentals of Generation and Propagation**, (Wiley,1988).
11. C. Lejeune, *Extraction of high-intensity ion beams from plasma source: theoretical and experimental treatments*, in *Applied Charged Particle Optics, Part C: Very High Density Beams*, A. Septier Ed. (Academic Press,1983), pp. 207-293,.
12. J. Rappaz, personal communication.
13. P.A. Raviart, *On singular perturbation problems for the nonlinear Poisson equation or a mathematical approach to electrostatic sheaths and plasma erosion*, Proceeding of the 4th SPARCH Summer School (1997).
14. L. Tonks and I. Langmuir, *General theory of the plasma of an arc*, Phys. Rev. **34**, 876 (1929).
15. J.H. Whealton, E.F. Jaeger and J.C. Whitson, *Optics of ion beams of arbitrary perveance extracted from a plasma*, J. Comput. Physics **27** (1978), pp. 32-41 .
16. J.H. Whealton, *Ion optics arithmetic and its implication for the positive ion CTR program*, I.E.E.E Trans Nucl. Sc., NS-28,n 2 (1978), pp. 1358-1361 .

## DIFFUSION LIMIT OF THE LORENTZ MODEL: ASYMPTOTIC PRESERVING SCHEMES \*

CHRISTOPHE BUET<sup>1</sup>, STÉPHANE CORDIER<sup>2</sup>, BRIGITTE LUCQUIN-DESREUX<sup>3</sup> AND SIMONA MANCINI<sup>3</sup>

**Abstract.** This paper deals with the diffusion limit of a kinetic equation where the collisions are modeled by a Lorentz type operator. The main aim is to construct a discrete scheme to approximate this equation which gives for any value of the Knudsen number, and in particular at the diffusive limit, the right discrete diffusion equation with the same value of the diffusion coefficient as in the continuous case. We are also naturally interested with a discretization which can be used with few velocity discretization points, in order to reduce the cost of computation.

**Mathematics Subject Classification.** 82C70, 35B40, 65N06.

Received: December 12, 2001. Revised: March 1, 2002.

### 1. INTRODUCTION

In this article, we study the numerical schemes for a kinetic equation in the diffusive regime:

$$\varepsilon \partial_t f + \cos \theta \partial_x f = \frac{1}{\varepsilon} \mathcal{L}(f). \quad (1.1)$$

The problem is one-dimensional in the space variable  $x$  and bi-dimensional in the velocity variable  $v = (\cos \theta, \sin \theta)$ . The unknown distribution function  $f = f(x, \theta, t)$  is a function of position  $x$  ( $x \in \mathbb{R}$ : in this article we shall not consider boundary conditions), of the velocity angle  $\theta \in [-\pi, \pi]$  and of time  $t > 0$ . The operator  $\mathcal{L}(f)$  is a linear collision operator of Lorentz type. Lorentz operators appear for example when considering elastic collisions of heavy particles (*e.g.* ions) against light ones (*e.g.* electrons); it is the first order term of the inter-species collision operator representing the collisions of the heavy particles on the light ones, when doing an asymptotic expansion in terms of the small mass ratio (see [8, 35]). This operator does not depend on the energy variable, *i.e.* on the modulus of the velocity. It is defined in the Boltzmann case by (for simplicity,

---

*Keywords and phrases.* Hilbert expansion, diffusion limit.

\* The authors acknowledge support from the TMR project “Asymptotic methods in kinetic theory” (TMR number: ERB FMRX CT97 0157), run by the European Community.

<sup>1</sup> CEA/DAM Ile de France, BP 12, 91680 Bruyères-Le-Châtel, France. e-mail: [Christophe.Buet@cea.fr](mailto:Christophe.Buet@cea.fr)

<sup>2</sup> Laboratoire MAPMO, UMR 6628, Université d'Orléans, 45067 Orléans, France. e-mail: [Stephane.Cordier@univ-orleans.fr](mailto:Stephane.Cordier@univ-orleans.fr)

<sup>3</sup> Laboratoire d'Analyse Numérique, UMR 7598, Université Pierre et Marie Curie, BP 187, 75252 Paris Cedex 05, France. e-mail: [lucquin@ann.jussieu.fr](mailto:lucquin@ann.jussieu.fr) & [smancini@ann.jussieu.fr](mailto:smancini@ann.jussieu.fr)

we drop the dependence with respect to  $x$  and  $t$ , since the collision operator only acts on  $\theta$ ):

$$\mathcal{L}(f)(\theta) = \int_{S^1} K(\theta' - \theta)[f(\theta') - f(\theta)] d\theta' \quad (1.2)$$

and in the Fokker-Planck case by:

$$\mathcal{L}(f)(\theta) = \partial_{\theta\theta}^2 f. \quad (1.3)$$

We recall that the Boltzmann-Lorentz operator (1.2) converges (up to multiplication by the second order moment of the scattering cross-section  $K$ ) to the Fokker-Planck-Lorentz one (1.3) when the cross-section concentrates, *i.e.* when the scattering angle during a collision,  $\theta' - \theta$ , is very small. This is the so-called “grazing collision limit” [3, 6, 36], which is also valid in the non-linear case [7, 9].

It is well known in literature, that for  $\varepsilon \ll 1$  the solution of (1.1) converges to the solution of a diffusion problem, with respect to the space variable. The diffusion coefficient of this last equation may be computed by means of a Hilbert expansion method (see Sect. 2). Our goal is to derive a numerical scheme that is relevant for any value of  $\varepsilon$ . For simplicity, we will consider the Fokker-Planck-Lorentz collision operator defined by (1.3). Nevertheless, we remark that our results can be extended to the Boltzmann-Lorentz collision operator defined by (1.2).

This paper is inspired by a series of articles written by Jin and Levermore about diffusive limit of the isotropic Boltzmann-Lorentz operator, where the cross-section is constant with respect to the velocity variable. This fact does not allow to pass to the grazing collision limit, excluding thus the Fokker-Planck case. The authors consider discrete ordinate methods in the velocity variable, or equivalently in the cosine of the angle (the distribution function being isotropic in the others directions). More precisely, they determine the quadrature points in the integral with respect to velocity such that, in the diffusion limit, one recovers a heat equation with the correct diffusion coefficient. A first paper, see [22], is devoted to the discretization in velocities, the distribution function depending continuously of the space and time variables. The authors construct quadrature sets corresponding to a small number of discrete velocities such that the diffusion coefficient and also the boundary conditions are compatible with the diffusive limit. In a second paper, see [23], they investigate the fully discrete case (both on the velocity and the space variables), the problem being stationary. This can be seen as the problem for one time step iteration using an implicit scheme. In another paper with Golse, see [13], they study the convergence of these schemes.

Our aim is to extend this analysis to other elastic collision operators with an arbitrary cross-section (*i.e.* not necessarily constant) and for example in the case of the grazing collision limit, *i.e.* for the Fokker-Planck Lorentz operator (1.3).

There is a huge literature on related topics taking different names. We claim that the following keywords are used to qualify very close problematic although the goal and the methods used are different.

- **Diffusive ( $a = 1$ ) or hydrodynamic ( $a = 0$ ) limit**

$$\varepsilon^a \partial_t f + \cos \theta \partial_x f = \frac{1}{\varepsilon} \mathcal{L}(f).$$

The aim is to derive numerical schemes which can be used either in rarefied (where  $\varepsilon$  is of order 1) or in dense regions ( $\varepsilon \ll 1$ ). Depending on the collision operator  $\mathcal{L}$ , one expects the solution to converge towards an equilibrium state (typically a Maxwellian distribution) and that the conserved quantities are solution of the Euler equations (for  $a = 0$ ) or Navier-Stokes equations (for  $a = 1$ ), see [26, 29, 43, 44].

A lot of works have been devoted to the coupling of rarefied and hydro-dynamical domains. For this purpose, numerical schemes compatible with the fluid limit, called *Asymptotic preserving* schemes, are an alternative to the matching of boundary conditions, these last being sometimes hard to design. In fact, these schemes can work uniformly with respect to the relaxation parameter (see [27] and its introduction for more details).

• **Stiff source term for hyperbolic systems**

A widely studied topic deals with hyperbolic systems of conservation laws (*e.g.* Euler equations) with stiff source terms (see [4, 5, 28]) due, for example, to the modeling of rapid chemical reactions. These systems read:

$$\partial_t U + \partial_x F(U) = \frac{1}{\varepsilon} S(f)$$

and are related to the so called *relaxation methods* (see [24, 28, 40, 41]) that consists in replacing a nonlinear problem of the above form by a relaxed linear system of the following form:

$$\partial_t U + \partial_x V = 0, \quad \partial_t V + a \partial_x U = -\frac{1}{\varepsilon} (V - F(U))$$

where  $\varepsilon$  is called relaxation rate and  $a$  is an arbitrary velocity (with  $-\sqrt{a} < F'(U) < \sqrt{a}$ ).

We remark that there is another area in the hyperbolic systems field that is related to such asymptotic namely the *kinetic schemes*: in this direction, one replaces the hyperbolic system of interest (*e.g.* the Euler equation) by a kinetic formulation in the hydro-dynamical limit (*e.g.* a B.G.K type equation), see [34, 42]. Then, one derives a scheme on the kinetic formulation and takes its hydro-dynamical limit. These methods have stability, accuracy and efficiency advantages. Some results on convergence are also available, see [2].

Let us also mention another recently way to design Asymptotic Preserving Schemes, the *Well Balanced* schemes, see [14, 16, 17, 33], designed to capture stationary fluxes for hyperbolic systems with source terms. In [14–16] such a scheme is used for Goldstein–Taylor type models (two characteristic speeds). A Chapman–Enskog expansion shows that the asymptotic limit of the scheme is the good one for Goldstein–Taylor type models, but on systems that have more than two characteristic speeds the asymptotic seems hard to generalize.

There is a large number of applications based on transport equations having a diffusive asymptotic. Let us mention, for instance, neutron transport, radiative transfer in the “optically thick limit” (see [1, 31, 32, 38, 39] and the references therein) and semiconductor modeling (see [25, 29, 37]). However, many of the previous works deal with the so-called telegrapher equation, or equivalently Goldstein–Taylor equation which is a kinetic equation where the distribution function is localized on two opposite velocities (see [4, 20, 21, 26, 27, 40]). On the other hand, in this paper, we shall not separate the particle density function with respect to the sign of the velocity, usually called parity method or even-odd decomposition. We refer more precisely to the introduction of Section 4.4 for a detailed explanation of this fact and a comparison with previous works. Let us finally mention that the space discretization scheme we propose could be also apply to the Boltzmann–Lorentz operator with an arbitrary cross-section.

This work is divided as follows. In Section 2, we briefly recall the Hilbert expansion method at the continuous level in order to derive the diffusion model. Then, we consider successively the discretization with respect to the velocity angle  $\theta$ , to the space variable  $x$ , and finally to the time  $t$ .

In Section 3, following Jin–Levermore (see [22, 23], for what regards the isotropic Boltzmann–Lorentz collision operator), we consider the problem discretized only with respect to the velocity variable. We look for a choice of velocity discretization points such that the diffusion coefficient tends towards the value obtained in the continuous case. Our goal is also to use a small number (denoted by  $N_\theta$ ) of discretization points with respect to velocity angle, in order to avoid too much expensive computations. We prove that if we consider a uniform grid, we obtain the right diffusion coefficient only in the limit  $N_\theta \rightarrow \infty$  (with an accuracy of order 2). In order to use a kinetic description with a small number of discretization points and to preserve the right asymptotic when  $\varepsilon \rightarrow 0$ , it is thus necessary to consider a non-uniform (but symmetric) grid. In particular, we treat the case of a grid with 4 or 8 discretization points.

In Section 4, we consider the space discretization (the velocity discretization is assumed to be known). More precisely, we study the discretization of the transport term  $v \partial_x f$ . The boundary layer problem is not taken

into account, and will be studied in futures works. The first approach consists of discretizing this term by means of an upwind scheme, but this method leads to a parasite solvability condition, *i.e.* we get an infinite diffusion coefficient. On the contrary, the centered scheme converges to a discretization of the Laplacian (with respect to  $x$ ), but the discretization acts on a double mesh, the *even* points of the mesh being decoupled from the *odd* points, yielding to some spurious modes (which gives numerical oscillations in the computations). We then consider a  $\theta$ -scheme ( $\varepsilon$  part for the upwind scheme and  $(1 - \varepsilon)$  for the centered one): the first part effectively *re-couples* the discretization points, but it introduces an error of order  $\Delta x$  in the value of the diffusion coefficient. Finally, we take into account the *modified Jin-Levermore* scheme which is obtained following the strategy proposed by Jin and Levermore (see [24, 28]). It consists in writing a finite volume type scheme and in computing the fluxes at the interfaces using the leading order term (or equivalently, the steady state equation) in the upwind scheme for the half mesh. We can also give an interpretation of this scheme as a finite discontinuous elements  $P^1$  scheme (see [45]).

Section 5 deals with the time discretization. First, we remark that the usual methods of splitting transport and collisions are not suitable in the limit  $\varepsilon \rightarrow 0$ . In fact, both the collision and the transport parts yield to projecting onto the constant states (with respect to velocity or position), and in only two time iterations one may get a constant function. Moreover, to avoid the CFL condition, which would lead to a very small time step when  $\varepsilon \rightarrow 0$ , we must use implicit schemes. Numerical results are given in Section 6. Some final results and comments are then presented as a conclusion in Section 7.

## 2. THE DIFFUSION LIMIT AT THE CONTINUOUS LEVEL

Let us first recall the derivation of the diffusion equation from the kinetic one by means of the Hilbert expansion method. As explained in the introduction, we consider the following kinetic equation of unknown  $f = f(x, \theta, t)$  in one space variable  $x \in \mathbb{R}$ , the angular velocity being  $\theta \in (0, 2\pi)$  with periodic condition on  $\theta$ , at the diffusion scale  $(t/\varepsilon^2, x/\varepsilon)$ :

$$\varepsilon \partial_t f + \cos \theta \partial_x f = \frac{1}{\varepsilon} \frac{\partial^2 f}{\partial \theta^2}, \quad (2.1)$$

where the collision term is the Fokker–Planck–Lorentz operator and  $\varepsilon \ll 1$  denotes the Knudsen number. We now use a classical Hilbert method by expanding  $f$  in terms of  $\varepsilon$ :

$$f = f^0 + \varepsilon f^1 + \varepsilon^2 f^2 + \dots,$$

and identifying terms of equal powers in (2.1). We successively obtain:

$$\frac{\partial^2 f^0}{\partial \theta^2} = 0, \quad (2.2)$$

$$\cos \theta \partial_x f^0 = \frac{\partial^2 f^1}{\partial \theta^2}, \quad (2.3)$$

$$\partial_t f^0 + \cos \theta \partial_x f^1 = \frac{\partial^2 f^2}{\partial \theta^2}. \quad (2.4)$$

From the first equation (2.2) and periodicity, we deduce that  $f^0$  is independent of the velocity angle  $\theta$ . The solvability condition for the second equation writes:

$$\int_0^{2\pi} \cos \theta \partial_x f^0 d\theta = 0, \quad (2.5)$$

and this condition is actually satisfied, because  $f^0$  does not depend on  $\theta$  and  $\int_0^{2\pi} \cos \theta \, d\theta = 0$ . Moreover, we have (up to the addition of a function which only depends on the space variable):

$$f^1 = -\cos \theta \, \partial_x f^0. \quad (2.6)$$

Finally, the solvability condition for equation (2.4) writes:  $2\pi \partial_t f^0 + \int_0^{2\pi} \cos \theta \, \partial_x f^1 \, d\theta = 0$ , which, on account of (2.6) and the fact that  $\int_0^{2\pi} \cos^2 \theta \, d\theta = \pi$  gives:

$$\partial_t f^0 - \frac{1}{2} \partial_{xx}^2 f^0 = 0. \quad (2.7)$$

This is the limit diffusion equation we obtain at the continuous level with the diffusion coefficient  $\nu = 1/2$ .

### 3. THE VELOCITY DISCRETIZATION

In this part, we consider the discretization with respect to the velocity variable, the time and space variables remaining continuous. This is the so-called discrete ordinate method. We shall perform the same analysis as in the continuous case (see Sect. 2). Our goal is to choose a small number of discretization points such that we obtain a discrete version of the diffusion equation (2.7) with still the right diffusion coefficient, *i.e.*  $\nu = 1/2$ . Let us first consider a uniform grid.

#### 3.1. Uniform grid

Let us define the uniform discretization of  $S^1$  as the angle  $\theta_j = j\Delta\theta$  with  $\Delta\theta = 2\pi/N_\theta$  and  $j = 0 \dots N_\theta - 1$ , where the indices  $j$  are denoted modulo  $N_\theta$ : *i.e.*  $j = N_\theta$  corresponds to the same velocity as  $j = 0$ . The discretized version of (2.1) is a finite system of transport equations coupled by the collision term:

$$\varepsilon \partial_t f_j + \cos \theta_j \, \partial_x f_j = \frac{1}{\varepsilon} (Lf)_j \quad (3.1)$$

where  $f_j(x, t) = f(x, \theta_j, t)$  is the value of the discretized distribution function  $f$  at time  $t$ , position  $x$  and velocity  $(\cos \theta_j, \sin \theta_j)$ . The operator  $L$  (with entries  $L_{ij}$ ) in the right hand side of equation (3.1) stands for the discretized version of the Laplacian operator with respect to the variable  $\theta$ . Using standard finite difference techniques and taking into account periodic boundary conditions, we obtain:

$$L_{ij} = \frac{1}{\Delta\theta^2} \begin{cases} -2 & \text{if } i = j \\ 1 & \text{if } |i - j| \equiv 1[N]. \end{cases}$$

Let us recall some useful spectral properties of the matrix  $L$ :

- The kernel of the matrix  $L$  is generated by the vector  $\mathbf{1} = (1, 1, 1, \dots, 1)$ .
- The vector  $(\cos \theta_j)_{j=0 \dots N_\theta-1}$  is an eigenvector of matrix  $L$ ; the associated eigenvalue is  $-\lambda$  where  $\lambda$  is the positive number given by:

$$\lambda = 2 \frac{1 - \cos(\Delta\theta)}{(\Delta\theta)^2}. \quad (3.2)$$

Performing the Hilbert expansion on the discretized problem (3.1), *i.e.*:  $f_j = f_j^0 + \varepsilon f_j^1 + \varepsilon^2 f_j^2 + \dots$  and identifying terms of equal powers of  $\varepsilon$ , we get:

$$(Lf^0)_j = 0, \quad (3.3)$$

$$\cos \theta_j \partial_x f^0 = (Lf^1)_j, \quad (3.4)$$

$$\partial_t f^0 + \cos \theta_j \partial_x f^1 = (Lf^2)_j. \quad (3.5)$$

At the order  $\varepsilon^{-1}$ , we obtain that  $f_j^0$  does not depend on  $\theta_j$ , since the kernel of  $L$  consists of constant vectors; we simply set:  $f_j^0 = f^0$ . Now, the solvability condition for equation (3.4) reads:  $\sum_{j=0}^{N_\theta-1} \cos \theta_j \partial_x f^0 = 0$ , which, because  $f^0$  is independent of  $\theta_j$ , leads to:

$$\sum_{j=0}^{N_\theta-1} \cos \theta_j = 0. \quad (3.6)$$

We can interpret relation (3.6) as a symmetry condition. This condition is satisfied by our choice of discretization points (with  $N_\theta$  even), but it does not hold if we translate the discretization points into  $\theta_0 + 2\pi j/N_\theta$  with  $\theta_0 \neq 0$  for example. We look for a solution  $f^1$  of (3.4) orthogonal to the kernel of  $L$  ( $\text{Ker}(L)^\perp = \{f : \langle f, \mathbf{1} \rangle = 0\}$ ), *i.e.* such that:

$$\sum_{j=0}^{N_\theta-1} f_j^1 = 0. \quad (3.7)$$

Then, as  $f^0$  is independent of  $\theta_j$  and  $(\theta_j)$  satisfies (3.6), the unique solution of (3.4) reads:

$$f_j^1 = -\frac{\cos \theta_j}{\lambda} \partial_x f^0. \quad (3.8)$$

Finally, considering (3.8), equation (3.5) becomes:  $\partial_t f^0 - \frac{\cos^2 \theta_j}{\lambda} \partial_{xx}^2 f^0 = (Lf^2)_j$ . The solvability condition for this equation ( $\sum_j (Lf^2)_j = 0$ ) writes:

$$\partial_t f^0 - \frac{1}{2\lambda} \partial_{xx}^2 f^0 = 0, \quad (3.9)$$

because we have:  $(1/N_\theta) \sum_j \cos^2 \theta_j = 1/2$ . We then get a diffusion equation for  $f^0$ , but with a diffusion coefficient  $\nu = \frac{1}{2\lambda}$ : in fact, this coefficient is never equal to the expected value  $1/2$  (since  $\lambda \neq 1$ , see (3.2)), but it converges towards it when  $N_\theta$  tends to  $\infty$  (with an accuracy of order  $(\Delta\theta)^2$ ). Nevertheless, if we want to deal with small number of points in  $\theta$ , for lowering the computational cost, we have to investigate non-uniform grids. When using a large number of points, one can use a uniform discretization.

### 3.2. Non-uniform grid

First, let us briefly recall the discretization of the second derivative with respect to the angle using non-uniform grids, in order to introduce the notations. Given a discretization of  $S^1$ ,  $\theta_j$  for  $j = 0 \dots N_\theta - 1$ , let us denote  $\Delta_j^+ = (\theta_{j+1} - \theta_j)$  and  $\Delta_j^- = (\theta_j - \theta_{j-1})$ . Applying a Taylor expansion with respect to the point  $\theta_j$  we obtain:

$$\begin{aligned} f(\theta_j + \Delta_j^+) &= f_j + \Delta_j^+ f'_j + \frac{\Delta_j^{+2}}{2} f''_j + o(\Delta_j^{+2}), \\ f(\theta_j - \Delta_j^-) &= f_j - \Delta_j^- f'_j + \frac{\Delta_j^{-2}}{2} f''_j + o(\Delta_j^{-2}). \end{aligned}$$

By combining these Taylor expansions, we get the following approximation of the second derivative at  $\theta_j$ :

$$f''_j \simeq 2 \frac{f(\theta_j + \Delta_j^+)}{\Delta_j^+ (\Delta_j^+ + \Delta_j^-)} + 2 \frac{f(\theta_j - \Delta_j^-)}{\Delta_j^- (\Delta_j^+ + \Delta_j^-)} - 2 \frac{f_j}{\Delta_j^+ \Delta_j^-}. \quad (3.10)$$

We shall discretize the Laplace operator using this formula when  $N_\theta = 4$  or  $8$ . We remark that in the general case the vector with components  $\cos \theta_j$  is no more an eigenvector of  $L$ , as it is easily seen replacing  $f_j$  by  $\cos \theta_j$  in (3.10).

Let us perform the same discretized Hilbert expansion as presented in the uniform grid case (Sect. 3.1). At first order,  $f_j^0 = f^0$  is independent on  $\theta_j$ . Let us define  $Y \in \mathbb{R}^{N_\theta}$  as the unique solution of

$$(LY)_j = \cos \theta_j, \quad \text{for all } j \in \{0, \dots, N_\theta - 1\}, \quad (3.11)$$

such that  $\sum_{j=0}^{N_\theta-1} Y_j = 0$ . Then, the unique solution  $f^1$  of equation (3.4) satisfying (3.7) is given by:

$$f^1 = Y \partial_x f^0.$$

Equation (3.9) now reads:  $\partial_t f^0 + \cos \theta_j Y_j \partial_{xx}^2 f^0 = (Lf^2)_j$ , and the solvability condition for this equation of unknown  $f^2$  still gives a diffusion equation for  $f^0$  where the diffusion coefficient  $\nu$  is now given by:

$$\nu = -\frac{1}{N_\theta} \sum_{j=0}^{N_\theta-1} Y_j \cos \theta_j = -\frac{1}{N_\theta} LY \cdot Y. \quad (3.12)$$

We shall now show that for  $N_\theta = 4$  or  $N_\theta = 8$ , one can construct symmetric sets of discretization points that give the good diffusion coefficient. This can be related to the analysis presented in [22,23] for the Boltzmann–Lorentz operator.

#### 4 points of discretization

Let us first consider the case  $N_\theta = 4$ , *i.e.* 4 points of discretization in velocity:  $\theta_0, \theta_1, \theta_2, \theta_3$ . We still consider a symmetric discretization, *i.e.* such that (3.6) holds.

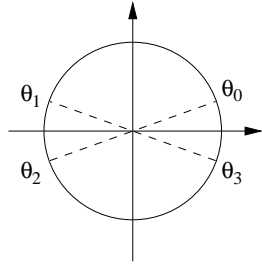


FIGURE 1. 4 discretization points in velocity.

More precisely, we choose  $\theta_0, \theta_1 = \pi - \theta_0, \theta_2 = \pi + \theta_0, \theta_3 = 2\pi - \theta_0$  *i.e.*  $\cos \theta_0 = -\cos \theta_1 = -\cos \theta_2 = \cos \theta_3$  (see Fig. 1). The goal is to determine  $\theta_0$  in such a way that the associated diffusion coefficient  $\nu$  is exactly equal to  $1/2$ .



The discrete Laplacian matrix corresponding to this nonuniform grid reads:

$$L = 2 \begin{pmatrix} \frac{-1}{\Delta_0 \Delta_3} & \frac{1}{\Delta_0(\Delta_0 + \Delta_3)} & 0 & \frac{1}{\Delta_3(\Delta_0 + \Delta_3)} \\ \frac{1}{\Delta_0(\Delta_0 + \Delta_1)} & \frac{-1}{\Delta_0 \Delta_1} & \frac{1}{\Delta_1(\Delta_0 + \Delta_1)} & 0 \\ 0 & \frac{1}{\Delta_1(\Delta_1 + \Delta_2)} & \frac{-1}{\Delta_1 \Delta_2} & \frac{1}{\Delta_2(\Delta_1 + \Delta_2)} \\ \frac{1}{\Delta_3(\Delta_2 + \Delta_3)} & 0 & \frac{1}{\Delta_2(\Delta_2 + \Delta_3)} & \frac{-1}{\Delta_2 \Delta_3} \end{pmatrix}$$

where we denote by  $\Delta_0 = \theta_1 - \theta_0$ ,  $\Delta_1 = \theta_2 - \theta_1$ ,  $\Delta_2 = \theta_3 - \theta_2$ ,  $\Delta_3 = \theta_0 - \theta_3$  the discretization steps.

**Proposition 3.1.** *Given  $N_\theta = 4$ , there exists a unique value of  $\theta_0$  such that we have the right diffusion coefficient, i.e.  $\nu = 1/2$ .*

*Proof.* We have to find  $Y = (Y_0, Y_1, Y_2, Y_3) \in \mathbb{R}^4$  such that:  $LY = (\cos \theta)$  with  $\sum_{j=0}^3 Y_j = 0$ , where the notation  $(\cos \theta)$  denotes the vector with components  $(\alpha, -\alpha, -\alpha, \alpha)$ , with  $\alpha = \cos \theta_0$ . After some easy computations, we find

$$Y = -B(\cos \theta), \quad \text{with } B = \pi(\pi - 2\theta)/4,$$

and the diffusion coefficient is given by  $\nu(\theta_0) = -\frac{1}{4}Y \cdot (\cos \theta) = B\alpha^2$ . In order to find the right diffusion coefficient, we now have to look for a value  $\theta_0 \in [0, \pi/2]$  such that  $\nu(\theta_0) = 1/2$ , where

$$\nu(\theta) = \frac{\pi}{4}(\pi - 2\theta) \cos^2 \theta.$$

We have  $\nu(0) = \pi^2/4 > 1/2$ ,  $\nu(\pi/2) = 0 < 1/2$  and  $\nu'(\theta_0) < 0$ . Thus, there exists a unique  $\theta_0 \in ]0, \pi/2[$  such that  $\nu(\theta_0) = 1/2$ .  $\square$

Surprisingly, for the non uniform (but symmetric) discrete point, the vector with components  $\cos \theta_j$  is an eigenvector, associated with the eigenvalue  $-B^{-1}$ , for the matrix  $L$ . An approximated value of  $\theta_0$ , namely  $\theta_0 = 0.8462$  has been computed using Maple<sup>®</sup>. This value differs from  $\pi/4$ , which is the case of uniform grid, but it is close to it.

### 8 points of discretization

We now consider the case  $N_\theta = 8$ , i.e. 8 points of discretization in velocity. Let us denote by  $\theta_j$  with  $j = 0, \dots, 7$  a general 8-points discretization of  $S^1$ .

We choose a symmetric discretization, i.e. such that condition (3.6) holds, as follows:  $\theta_0$  and  $\theta_1$  belonging to  $[0, \pi/2]$  (the others angles being induced by this choice (see Fig. 2)). We now have two angles ( $\theta_0$  and  $\theta_1$ ) that we shall determine in such a way that first the diffusion coefficient is the right one, and secondly that  $Y$  is an eigenvector (this second condition is arbitrary; we add it in order to determine the two angles in a unique way and by analogy with previous cases). The discretization step, which we recall is not fixed, is given by:

$$\begin{aligned} \Delta_0 &= \theta_1 - \theta_0, \quad \Delta_1 = \theta_2 - \theta_1 = \pi - 2\theta_2, \quad \Delta_2 = \theta_3 - \theta_2 = \theta_1 - \theta_0, \\ \Delta_3 &= \theta_4 - \theta_3 = 2\theta_0, \quad \Delta_4 = \theta_5 - \theta_4 = \theta_1 - \theta_0, \quad \Delta_5 = \theta_6 - \theta_5 = \pi - 2\theta_1, \\ \Delta_6 &= \theta_7 - \theta_6 = \theta_1 - \theta_0, \quad \Delta_7 = \theta_0 - \theta_7 = 2\theta_0. \end{aligned}$$

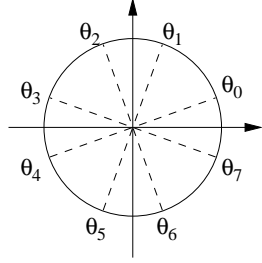


FIGURE 2. 8 discretization points in velocity.

We also impose that  $\theta_1 = m\theta_0$  for  $0 < \theta_0 < \pi/2$  and  $1 < m < \frac{\pi}{2\theta_0}$ . With the same method applied in the case of 4 velocities, we can derive the discretized version of the Laplacian  $L$ :

$$L = 2 \begin{pmatrix} \frac{-1}{\Delta_0 \Delta_7} & \frac{1}{\Delta_0(\Delta_0 + \Delta_7)} & 0 & \dots & 0 & \frac{1}{\Delta_7(\Delta_0 + \Delta_7)} \\ \frac{1}{\Delta_0(\Delta_0 + \Delta_1)} & \frac{-1}{\Delta_0 \Delta_1} & \frac{1}{\Delta_1(\Delta_0 + \Delta_1)} & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{1}{\Delta_5(\Delta_5 + \Delta_6)} & \frac{-1}{\Delta_5 \Delta_6} & \frac{1}{\Delta_6(\Delta_5 + \Delta_6)} \\ 1 & 0 & \dots & 0 & \frac{1}{\Delta_6(\Delta_6 + \Delta_7)} & \frac{-1}{\Delta_6 \Delta_7} \end{pmatrix}$$

$$= 2 \begin{pmatrix} B & B_1 & 0 & 0 & 0 & 0 & 0 & B_2 \\ A_1 & A & A_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_2 & A & A_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & B_1 & B & B_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & B_2 & B & B_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_1 & A & A_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_2 & A & A_1 \\ B_2 & 0 & 0 & 0 & 0 & 0 & B_1 & B \end{pmatrix}$$

where we have set

$$\begin{aligned} A &= -1/\alpha\gamma, & A_1 &= 1/\gamma(\gamma + \alpha), & A_2 &= 1/\alpha(\gamma + \alpha), \\ B &= -1/\beta\gamma, & B_1 &= 1/\gamma(\gamma + \beta), & B_2 &= 1/\beta(\gamma + \beta), \end{aligned}$$

with:  $\alpha = \pi - 2m\theta_0$ ,  $\beta = 2\theta_0$  and  $\gamma = \theta_0(m - 1)$ . We still look for a vector  $Y \in \mathbb{R}^8$  solution of (3.11) and such that it is an eigenvector for the matrix  $L$ , *i.e.*:

$$LY = (\cos \theta) = -\lambda Y. \quad (3.13)$$

We remark that now the vector  $(\cos \theta)$  is given by  $(\cos \theta) = (\bar{\alpha}, \bar{\beta}, -\bar{\beta}, -\bar{\alpha}, -\bar{\alpha}, -\bar{\beta}, \bar{\beta}, \bar{\alpha})$ , where  $\bar{\alpha} = \cos \theta_0$  and  $\bar{\beta} = \cos \theta_1 = \cos(m\theta_0)$ . From system (3.13) we get the following system:

$$\begin{aligned} 2B_1(\bar{\beta} - \bar{\alpha}) &= -\lambda\bar{\alpha}, \\ -2A_1(-\bar{\alpha} + \bar{\beta}) - 4\bar{\beta}A_2 &= -\lambda\bar{\beta}. \end{aligned} \quad (3.14)$$

We recall that  $\alpha$ ,  $\beta$  and  $\gamma$  depend on  $(m, \theta_0)$  so that  $A, A_1, A_2, B, B_1$  and  $B_2$  also depend  $(m, \theta_0)$ . Thus, equation (3.14) yields to

$$(\pi - 2m\theta)(\pi - m\theta - \theta)(\cos(m\theta) - \cos(\theta))\cos(m\theta) = (m\theta + \theta)\cos(\theta)[(\pi - 2m\theta)\cos(\theta) - (\pi - 2\theta)\cos(m\theta)]. \quad (3.15)$$

It remains to determine  $\theta_0$  and  $m$  such that the diffusion coefficient  $\nu$  is equal to  $(1/2)$  *i.e.* such that:

$$-2(\cos(m\theta_0) - \cos \theta_0) = (m^2 - 1)\theta_0^2(\cos^2 \theta_0 + \cos^2(m\theta_0))\cos \theta_0. \quad (3.16)$$

Solving (3.15, 3.16) using Maple, we get :  $\theta_0 = 0.4318$  and  $m = 2.764$ .

#### 4. THE SPACE DISCRETIZATION

We use a finite difference discretization (see [11]) in order to approximate the transport problem (2.1). We recall that the aim is to derive a discrete model which at the limit  $\varepsilon \rightarrow 0$  gives a good approximation of the diffusion equation (2.7). The velocity discretization is done using the finite difference scheme described in Section 3 (in particular, we shall set  $N_\theta = 4$  in the simulations), and we recall that  $L$  is the discrete operator approximating the Laplacian  $\frac{\partial^2}{\partial \theta^2}$ .

We now describe different schemes for the convective part and study the asymptotic properties of the global space-velocity scheme. In all the sequel, we denote by  $f_{i,j}$  an approximation of  $f(x_i, \theta_j, \cdot)$ , where  $x_i = i\Delta x$  are the points of discretization in space variable (for simplicity we set  $h = \Delta x$ ), and  $v_j = (\cos \theta_j, \sin(\theta_j))$ ,  $j \in \{0, \dots, N_\theta - 1\}$  are those with respect to the velocity variable. We recall in particular that these points  $\theta_j$  are symmetric in the sense that they satisfy the symmetry condition (3.6), which is also the discrete analogous of the continuous solvability condition (2.5). We also recall that these points are not necessarily equi-distributed, and in particular that for  $N_\theta$  small ( $N_\theta = 4$  for example) we have shown that, in order to recover the good diffusion coefficient, these points are necessarily distributed in a non uniform way (we still refer to Sect. 3). Furthermore, denoting by  $Y$  the vector in  $\mathbb{R}^{N_\theta}$  such that  $(LY)_j = \cos \theta_j$ , for all  $j \in \{0, \dots, N_\theta - 1\}$  with  $\sum Y_j = 0$ , the discrete diffusion coefficient is then given by (3.12) and at least in the simple cases  $N_\theta = 4$  or 8 there exists a unique choice of such discrete points  $\theta_j$  such that  $\nu$  is the good diffusion coefficient, *i.e.* such that:

$$\nu = \frac{1}{N_\theta} \sum_j Y_j \cos \theta_j = \frac{1}{2}.$$

In what follows, the space-velocity discretized kinetic equation reads, for all  $i, j$ :

$$\varepsilon \partial_t f_{i,j} + \cos \theta_j (Df)_{i,j} = \frac{1}{\varepsilon} (Lf_{i,\cdot})_j, \quad (4.1)$$

where  $D$  denotes a finite difference operator approximating the first derivative with respect to the space variable.

#### 4.1. The upwind scheme

The more natural way to discretize the first derivative  $\partial_x f$  is the upwind scheme  $D^u$ , which is defined by:

$$D^u f_{i,j} = \begin{cases} \frac{f_{i,j} - f_{i-1,j}}{h}, & \text{if } \cos \theta_j > 0, \\ \frac{f_{i+1,j} - f_{i,j}}{h}, & \text{if } \cos \theta_j < 0. \end{cases}$$

We consider the discrete equation (4.1) with  $D = D^u$  and use an Hilbert expansion, for all  $i, j$

$$f_{i,j} = f_{i,j}^0 + \varepsilon f_{i,j}^1 + \varepsilon^2 f_{i,j}^2 + \dots \quad (4.2)$$

Identifying terms of equal powers, we get, for all  $i, j$ :

$$(L f_{i,\cdot}^0)_j = 0, \quad (4.3)$$

$$\cos \theta_j (D^u f^0)_{i,j} = (L f_{i,\cdot}^1)_j, \quad (4.4)$$

$$\partial_t f_{i,j}^0 + \cos \theta_j (D^u f^1)_{i,j} = (L f_{i,\cdot}^2)_j. \quad (4.5)$$

Equation (4.3) shows that  $f^0$  does not depend on  $j$ . Now, equation (4.4) is solvable only if the following solvability condition is fulfilled:

$$\sum_{j=0}^{N_\theta-1} \cos \theta_j (D^u f^0)_{i,j} = 0.$$

But, on account of condition (3.6), we deduce that this is equivalent to:

$$f_{i+1}^0 - 2f_i^0 + f_{i-1}^0 = 0,$$

which is the discrete approximation of a stationary diffusion equation. This is not the good diffusion limit we expected, since it gives an infinite diffusion coefficient  $\nu = \infty$ . Hence, this scheme is not *asymptotic preserving*, since it gives a wrong diffusion equation at the limit  $\varepsilon \rightarrow 0$ .

#### 4.2. The centered scheme

Let us now examine the centered scheme which we shall denote by  $D^c$ :

$$D^c f_{i,j} = \frac{f_{i+1,j} - f_{i-1,j}}{2h}.$$

The discrete problem is then (4.1) with  $D = D^c$ . Identifying again the terms of equal powers in the Hilbert expansion (4.2) we get, for all  $i, j$ :

$$(L f_{i,\cdot}^0)_j = 0, \quad (4.6)$$

$$\cos \theta_j (D^c f^0)_{i,j} = (L f_{i,\cdot}^1)_j, \quad (4.7)$$

$$\partial_t f_{i,j}^0 + \cos(\theta_j) (D^c f^1)_{i,j} = (L f_{i,\cdot}^2)_j. \quad (4.8)$$

Equation (4.6) shows that  $f^0$  still does not depend on  $j$ . The solvability condition of equation (4.7) is then automatically satisfied, on account of the symmetry property (3.6). Moreover, we have:

$$f_{i,j}^1 = Y_j (D^c f^0)_i, \quad (4.9)$$

where  $Y$  is given by (3.11). The solvability condition for equation (4.8) writes:

$$N_\theta \partial_t f_i^0 + \sum_{j=0}^{N_\theta-1} \cos \theta_j (D^c f_{\cdot,j}^1)_i = 0,$$

which gives, on account of (4.9) and of the definition of  $\nu$ , the following discrete diffusion equation:

$$\partial_t f_i^0 - \nu (\Delta_{2h} f^0)_i = 0, \quad (4.10)$$

where we have used the notation:

$$(\Delta_{2h} \phi)_i = \frac{\phi_{i+2} - 2\phi_i + \phi_{i-2}}{(2h)^2}.$$

In the same way, we shall denote by  $\Delta_h$  the classical “three points scheme”:

$$(\Delta_h \phi)_i = \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{h^2}.$$

Although the discrete equation (4.10) is a consistent approximation of the continuous diffusion equation (2.7), we remark that this scheme will generate numerical oscillations, due to spurious modes: the discrete points corresponding to an even index  $i$  do not influence those corresponding to an odd one. We have actually observed this phenomena, by doing numerical experiments with an initial data of Dirac type (see Fig. 3). Since this scheme generates numerical oscillations, we drop it.

#### 4.3. The $\varepsilon$ scheme $D^\varepsilon$

An interesting idea however is to combine this scheme to the upwind one (since the last scheme gives a diffusion operator of type  $\Delta_h$  instead of  $\Delta_{2h}$ ), in order to avoid this phenomena. But we have seen that the upwind scheme does not give the right diffusion equation, on account of the solvability condition obtained when identifying constant terms. So a good compromise consist in studying the following convex combination:

$$D^\varepsilon = (1 - \varepsilon)D^c + \varepsilon D^u;$$

we shall denote it by “ $\varepsilon$ - scheme”.

The discrete problem is here (4.1) with  $D = D^\varepsilon$  and the Hilbert expansion method gives, for all  $i, j$ :

$$(L f_{i,\cdot}^0)_j = 0, \quad (4.11)$$

$$\cos \theta_j (D^c f^0)_{i,j} = (L f_{i,\cdot}^1)_j, \quad (4.12)$$

$$\partial_t f_{i,j}^0 + \cos \theta_j \left[ (D^c f_{\cdot,j}^1)_i - (D^c f_{\cdot,j}^0)_i + (D^u f^0)_{i,j} \right] = (L f_{i,\cdot}^2)_j. \quad (4.13)$$

From equation (4.11), we still have that  $f^0$  independent of  $j$ , and the solvability condition for equation (4.12) is fulfilled, on account of (3.6). Moreover, as (4.12) coincide with (4.7), we have (4.9). The solvability condition for equation (4.13) is then:

$$N_\theta \partial_t f_i^0 + \sum_{j=0}^{N_\theta-1} \cos \theta_j (D^c f_{\cdot,j}^1)_i + \sum_{j=0}^{N_\theta-1} \cos \theta_j (D^u f^0)_{i,j} = 0,$$

which gives, on account of expression (4.9), of the definition of  $\nu$  and of the condition (3.6), the following discrete diffusion equation:

$$\partial_t f_i^0 - \nu (\Delta_{2h} f^0)_i - \nu' (\Delta_h f^0)_i = 0,$$

where we have set:

$$\nu' = \frac{h}{2N_\theta} \sum_{j=0}^{N_\theta-1} |\cos \theta_j|.$$

This equation is still a consistent approximation of the diffusion equation (2.7) and it links the points with even index to those with an odd one. Thus, we can expect this scheme to suppress the oscillations given by the pure centered scheme. This has been effectively observed numerically, but for an initial data which support contains at least two grid points (in space variable). For a Dirac type initial data however, the oscillations still persist (see Fig. 4). This scheme thus seems to be a relatively good one, although the diffusion coefficient is not exactly the good one (up to an  $O(h)$  error). We now examine a last scheme which seems to have all the expected properties.

#### 4.4. The modified Jin-Levermore scheme

This new scheme is inspired by the scheme proposed by Jin and Levermore for the telegraph problem in [24] (which is based on the use of the steady states approximation method, see for example [10, 47], ...). In this work, Jin and Levermore studied semi-discrete numerical schemes for hyperbolic systems with stiff relaxation terms that have a long time behavior governed by reduced systems of parabolic type, but the diffusion terms are in fact corrective terms of order  $\varepsilon$ . A similar problem, but under diffusive scaling, called the Goldstein–Taylor model (see [12, 46]), has been numerically considered by [26]: here, both the convective terms and the relaxation terms are stiff (like for our problem), which gives additional difficulties. One of the mean ideas consists in writing the stiff terms as source terms, then to use a splitting in time algorithm, treating the relaxation terms by an implicit scheme and the new non-stiff convective terms by classical upwind schemes (or more accurate second order schemes with slope limiters). Special care has been taken to assure that the scheme possess the corrective diffusive limit. An extension of this work to more general source terms have been recently considered, in [26] for a bi-dimensional Boltzmann–Lorentz type operator, and for the linear Boltzmann equation in [25]. Both works are concerned with the actual diffusive regime and are both based on a splitting of the distribution into its even and odd part, giving then a system which is treated with similar discretization than the Goldstein–Taylor model previously described. A new difficulty arises in the Boltzmann case to solve the collision step implicitly: one needs in fact to invert an integral operator, which is not easy to do in an efficient way. A velocity discretization using Hermite polynomials is then performed.

It is in fact possible to adapt this method here to our problem, by first splitting  $f$  into an even and an odd part and then using the discrete algorithm proposed in [25]. Moreover, let us point out that the implicit treatment of the collision part is here far simpler for our model than for the Boltzmann case considered in [25]. For the continuous in time scheme, we obtain the good diffusion limit with the right coefficient of diffusion. For the full discrete problem, we obtain the right diffusion coefficient, up to an  $O(h)$  error, like for the  $\varepsilon$  scheme proposed in Section 4.3.

It is one of the reasons why we decided to construct a new scheme, still based on Jin-Levermore’s approach, but where it would be possible to directly discretize  $f$  (without needing to split it into an even and an odd part) which would be far simpler. Moreover, the splitting method proposed in [22] seems interesting in so far the discretization with respect to the space variable is “unconnected” from the velocity one, but as the stiff convective terms appear in the right hand side of the relaxation part, it seems that both meshes are coupled, giving this scheme less attractive. Finally, we point out that the scheme proposed below can be also applied to the Boltzmann–Lorentz operator with an arbitrary cross-section.

The idea for the construction of our new scheme is based on an evaluation of the fluxes at the interface between the two cells  $[x_i - h/2, x_i + h/2]$  and  $[x_{i+1} - h/2, x_{i+1} + h/2]$ . The scheme writes (4.1), with  $D = D^{1/2}$  defined by:

$$(D^{1/2}f)_{i,j} = \frac{f_{i+1/2,j} - f_{i-1/2,j}}{h}.$$

The idea for the evaluation of the interface values  $f_{i+1/2,j}$  (or  $f_{i-1/2,j}$ ) is first to write (up to an  $O(h^2)$  error) the Taylor expansion of  $f_{i+1/2,j}$ :

$$\begin{aligned} f_{i,j} &\simeq f_{i+1/2,j} - \frac{h}{2}(\partial_x f_{.,j})_{i+1/2}, & \text{if } \cos \theta_j > 0, \\ f_{i+1,j} &\simeq f_{i+1/2,j} + \frac{h}{2}(\partial_x f_{.,j})_{i+1/2}, & \text{if } \cos \theta_j < 0, \end{aligned}$$

which corresponds to a classical upwind scheme, according to the sign of the velocity. Now, in order to compute the space gradients, we consider the following continuous equation, with respect to the space variable only, *i.e.*

$$\varepsilon \partial_t f + \cos \theta \partial_x f = \frac{1}{\varepsilon} Lf,$$

in which we neglect the lowest order terms, with respect to  $\varepsilon$ , *i.e.* the  $O(\varepsilon)$  term. This gives the following system (for simplicity, we suppose that, from now on, we never have  $\cos \theta_j = 0$ ):

$$f_{i,j} = f_{i+1/2,j} - \frac{h}{2\varepsilon \cos \theta_j} (Lf_{i+1/2,.})_j, \quad \text{if } \cos \theta_j > 0, \quad (4.14)$$

$$f_{i+1,j} = f_{i+1/2,j} + \frac{h}{2\varepsilon \cos \theta_j} (Lf_{i+1/2,.})_j, \quad \text{if } \cos \theta_j < 0. \quad (4.15)$$

In other words, the interface values satisfy:

$$\left[ \left( Id - \frac{h}{2\varepsilon |\cos \theta_j|} L \right) f_{i+1/2,.} \right]_j = \begin{cases} f_{i,j}, & \text{if } \cos \theta_j > 0, \\ f_{i+1,j}, & \text{if } \cos \theta_j < 0. \end{cases}$$

We now use the Hilbert method, expanding  $f_{i,j}$  according to (4.2) and identifying terms of equal powers in the kinetic equation (4.1); we get, for all  $i, j$ :

$$(Lf_{i,.}^0)_j = 0, \quad (4.16)$$

$$\cos \theta_j \left( D^{1/2} f^0 \right)_{i,j} = (Lf_{i,.}^1)_j, \quad (4.17)$$

$$\partial_t f_{i,j}^0 + \cos \theta_j \left( D^{1/2} f^1 \right)_{i,j} = (Lf_{i,.}^2)_j. \quad (4.18)$$

The first equation shows, as usual, that  $f_{i,j}^0$  does not depend on  $j$ . Now, in order to solve the other equations, we also need to expand  $f_{i+1/2,j}$  in terms of  $\varepsilon$ . We set for all  $i, j$

$$f_{i+1/2,j} = f_{i+1/2,j}^0 + \varepsilon f_{i+1/2,j}^1 + \varepsilon^2 f_{i+1/2,j}^2 + \dots$$

and identify terms of equal powers in equations (4.14, 4.15), this gives:

$$\left(Lf_{i+1/2,\cdot}^0\right)_j = 0, \quad (4.19)$$

$$\left(Lf_{i+1/2,\cdot}^1\right)_j = -\frac{f_i^0 - f_{i+1/2,j}^0}{h/2} \cos \theta_j, \quad \text{if } \cos \theta_j > 0, \quad (4.20)$$

$$\left(Lf_{i+1/2,\cdot}^1\right)_j = \frac{f_{i+1}^0 - f_{i+1/2,j}^0}{h/2} \cos \theta_j, \quad \text{if } \cos \theta_j < 0. \quad (4.21)$$

We first deduce from (4.19) that  $f_{i+1/2,j}^0$  does not depend on  $j$ ; we simply denote it by  $f_{i+1/2}^0$ . To compute  $f_{i+1/2,\cdot}^1$  from equations (4.20, 4.21), there appears a necessary condition of solvability which writes:

$$-\left(\sum_{j, \cos \theta_j > 0} \cos \theta_j\right) \frac{f_i^0 - f_{i+1/2}^0}{\frac{h}{2}} + \left(\sum_{j, \cos \theta_j < 0} \cos \theta_j\right) \frac{f_{i+1}^0 - f_{i+1/2}^0}{\frac{h}{2}} = 0,$$

*i.e.*  $f_i^0 - f_{i+1/2}^0 + f_{i+1}^0 - f_{i+1/2}^0 = 0$ , on account of (3.6). We deduce that  $f_{i+1/2}^0$  is the mean value of  $f^0$  at points  $x_i$  and  $x_{i+1}$ , *i.e.* if we have

$$f_{i+1/2,j}^0 = f_{i+1/2}^0 = \frac{f_i^0 + f_{i+1}^0}{2}.$$

Injecting this expression in (4.20, 4.21), we get, for all  $j$  such that  $\cos \theta_j \neq 0$ :

$$\left(Lf_{i+1/2,\cdot}^1\right)_j = \frac{f_{i+1}^0 - f_i^0}{h} \cos \theta_j,$$

which gives:

$$f_{i+1/2,j}^1 = Y_j \frac{f_{i+1}^0 - f_i^0}{h}. \quad (4.22)$$

We now turn back to equations (4.17, 4.18). As  $(D^{1/2}f^0)_{i,j}$  is independent of  $j$ , the solvability condition of equation (4.17) is satisfied. On account of expression (4.22), we have  $(D^{1/2}f^1)_{i,j} = Y_j(\Delta_h f^0)_i$ , so that the solvability condition of equation (4.18) simply writes:

$$\partial_t f_i^0 - \nu(\Delta_h f^0)_i = 0,$$

which is exactly the discrete diffusion equation we expected, with the good coefficient of diffusion  $\nu$ : this scheme seems to be the best one, which has been confirmed by numerical tests (see Figs. 5 and 6).

#### 4.5. A finite volume interpretation of the new scheme

This new scheme can be interpreted as a generalization of the  $P^1$  discontinuous finite elements scheme proposed by G. Samba (see [45]) for a stationary kinetic equation of Boltzmann–Lorentz type (see also [30, 39]). But it can also be interpreted in terms of a classical finite volume scheme expressed on a refined mesh of size  $h/2$ . We now detail this interpretation.

The aim is to explain the computation (4.14, 4.15) of the numerical flux. We first set  $i' = i - 1/2$  and introduce the new discretization points  $x_{i'} = i'h$  and  $x_{i'+k/4} = (i' + k/4)h$ , for  $k \in \mathbb{Z}$ . We set:

$$f_{i+1/2,j} = \frac{1}{2} [f_{i'+3/4,j} + f_{i'+5/4,j}],$$



where  $f_{i'+3/4,j}$  and  $f_{i'+5/4,j}$  solve the following equations:

$$\varepsilon \partial_t f_{i'+3/4,j} + \cos \theta_j \frac{F_{i'+1,j} - F_{i'+1/2,j}}{h/2} = \frac{1}{\varepsilon} (L f_{i'+3/4,\cdot})_j, \quad (4.23)$$

$$\varepsilon \partial_t f_{i'+5/4,j} + \cos \theta_j \frac{F_{i'+3/2,j} - F_{i'+1,j}}{h/2} = \frac{1}{\varepsilon} (L f_{i'+5/4,\cdot})_j, \quad (4.24)$$

with:

$$F_{i'+1/2,j} = \frac{1}{2} [f_{i'+1/4,j} + f_{i'+3/4,j}],$$

$$F_{i'+1,j} = \begin{cases} f_{i'+3/4,j} & \text{if } \cos \theta_j > 0, \\ f_{i'+5/4,j} & \text{if } \cos \theta_j < 0. \end{cases}$$

In the same way, we also naturally have:

$$\varepsilon \partial_t f_{i'+1/4,j} + \cos \theta_j \frac{F_{i'+1/2,j} - F_{i',j}}{h/2} = \frac{1}{\varepsilon} (L f_{i'+1/4,\cdot})_j. \quad (4.25)$$

Now equations (4.23) and (4.24) give by addition:

$$\varepsilon \partial_t f_{i+1/2,j} + \cos \theta_j \frac{F_{i'+3/2,j} - F_{i'+1/2,j}}{h} = \frac{1}{\varepsilon} (L f_{i+1/2,\cdot})_j,$$

or equivalently:

$$\varepsilon \partial_t f_{i+1/2,j} + \cos \theta_j \frac{F_{i+1,j} - F_{i,j}}{h} = \frac{1}{\varepsilon} (L f_{i+1/2,\cdot})_j.$$

We here recover the equation of evolution (4.1) expressed at the new spatial grid point  $x_{i+1/2}$ . It remains to find an equation for the flux  $F_{i,j}$  at the interface of the refined mesh. This one is obtained by summing equations (4.23–4.25); we get:

$$\varepsilon \partial_t F_{i,j} + \cos \theta_j \frac{F_{i'+1,j} - F_{i',j}}{h} = \frac{1}{\varepsilon} (L F_{i,\cdot})_j. \quad (4.26)$$

A simple computation first shows that:

$$\frac{F_{i'+1,j} - F_{i',j}}{h} = \begin{cases} \frac{2}{h} [F_{i,j} - f_{i-1/2,j}] & \text{for } \cos \theta_j > 0, \\ \frac{2}{h} [f_{i+1/2,j} - F_{i,j}] & \text{for } \cos \theta_j < 0. \end{cases}$$

Now, neglecting the lowest order term (with respect to  $\varepsilon$ ) in (4.26), we get:

$$\frac{2 \cos \theta_j}{h} [F_{i,j} - f_{i-1/2,j}] = \frac{1}{\varepsilon} (L F_{i,\cdot})_j, \quad \text{if } \cos \theta_j > 0,$$

$$\frac{2 \cos \theta_j}{h} [f_{i+1/2,j} - F_{i,j}] = \frac{1}{\varepsilon} (L F_{i,\cdot})_j, \quad \text{if } \cos \theta_j < 0,$$

or equivalently, for  $i' = i + 1/2$ :

$$\begin{aligned} f_{i',j} &= F_{i'+1/2,j} - \frac{h}{2\varepsilon \cos \theta_j} (LF_{i'+1/2,\cdot})_j, & \text{if } \cos \theta_j > 0, \\ f_{i'+1,j} &= F_{i'+1/2,j} + \frac{h}{2\varepsilon \cos \theta_j} (LF_{i'+1/2,\cdot})_j, & \text{if } \cos \theta_j < 0, \end{aligned}$$

*i.e.* we exactly recover formulae (4.14, 4.15) at the new grid point  $x_{i'}$ .

## 5. THE TIME DISCRETIZATION

We now present the discretization in time. We first remark that it is of no use to apply a splitting in time method. In fact, for the first time step, the collision part

$$\varepsilon \partial_t f = \frac{1}{\varepsilon} \mathcal{L} f$$

would project on the constants with respect to the velocity angle  $\theta$ , and the transport part

$$\varepsilon \partial_t f + \cos \theta \partial_x f = 0$$

would project on the constant with respect to the position variable  $x$ . Thus, in two time steps we will find a constant function (and not the good diffusion equation). The fact that splitting also fails in the diffusion limit is not surprising and we also refer to the Section A.2 where a similar problem occurs for the directional splitting in the multi-dimensional case.

On the other hand, when using an explicit scheme, we get time step restrictions. Due to the diffusive scaling of the equation (1.1), the time step stability condition associated with the transport term is of the form

$$\Delta t \leq \varepsilon \Delta x, \quad (5.1)$$

since the velocities are of modulus smaller than 1, whereas the time step condition for the collision part reads:

$$\Delta t \leq \tau \varepsilon^2 (\Delta v)^2, \quad (5.2)$$

where  $\tau$  is the collision time (equal 1 in this paper). Since the aim of the studied schemes is to be used for arbitrary small values of  $\varepsilon$ , the cost of such explicit schemes will be prohibitive especially due to the collision part.

Therefore, we shall always use implicit schemes for the collision part. For the transport part, we can use either fully explicit scheme (with the restriction given by (5.1)), semi-implicit scheme (for example using the  $\varepsilon$ -scheme described in Sect. 4.3) or a fully implicit scheme. More precisely, in the semi-implicit scheme, we implicit the collision and  $(1 - \varepsilon)D^c$ , the first part of the transport, and treat explicitly the upwind transport part  $\varepsilon D^u$ . This leads us to a stability condition of the form  $\Delta t \leq \Delta x$ .

From numerical point of view, the semi and fully implicit scheme requires to invert  $N_x \times N_\theta$  matrices where  $N_x$  (respectively  $N_\theta$ ) is the number of point in the discretization with respect to  $x$  (respectively  $v$ ).

## 6. NUMERICAL RESULTS

The numerical tests are devoted to verify that the proposed discretization gives the right diffusion coefficient. We consider an initial data equal to a Dirac measure in  $x$  and uniform in  $\theta$  or concentrated in  $\theta$  (but this latter choice does not change the solutions when  $\varepsilon \rightarrow 0$ ):

$$f(x, \theta, t = 0) = \delta_{x=0} \quad \text{or} \quad f(x, \theta, t = 0) = \delta_{x=0} \delta_{\theta=0}.$$

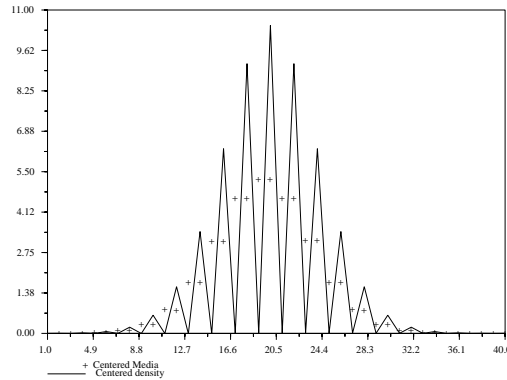


FIGURE 3. Centered scheme.

For such initial data, the solution  $f^\varepsilon$  of problem (1.1) behaves when  $\varepsilon \rightarrow 0$  as the solution  $f^0$  of the heat equation, which is given by:

$$f^0(x, t) = \frac{1}{\sqrt{4\pi\nu t}} \exp\left(-\frac{x^2}{4\nu t}\right).$$

In particular, the second moment of the solution is close to that of a Gaussian

$$\int f^\varepsilon(t, x, \theta) x^2 dx d\theta \propto 2\nu t.$$

In other words, the evolution of the second moment of  $f^\varepsilon$  becomes linear in time as  $\varepsilon \rightarrow 0$  with a slope related to the diffusion coefficient of the limiting diffusive equation.

In the results presented here, we use the four velocities case presented in Section 3.2 for a given angle  $\theta_0$ . Here, and also in the following figures,  $\varepsilon = 0.001$ , the numerical scheme is implicit in time,  $\Delta x = 40$ ,  $\Delta\theta = 4$ ,  $\Delta t = 0.01$  and time equal to 1 (*i.e.* 100 iterations).

In Figure 3, we plot the particle density  $n(x, t)$ ,

$$n(x, t) = \int f(x, \theta, t) d\theta,$$

for the centered scheme. In particular, there are underlined the numerical oscillations due to the non-coupling between the odd and even meshes. Finally, the dotted line is a post-treated curve giving the mean values of  $n(x, t)$  between two successive meshes.

The density  $n(x, t)$  for the  $\varepsilon$ -scheme is plotted in Figure 4. We remark that the numerical oscillations in Figure 3 are smoothened by the effect of the upwind scheme. The difference between the diffusion coefficient in the  $\varepsilon$ -scheme and the centered scheme is not remarkable due to the fact that  $h = \Delta x$  is small. The continuous line corresponds to the “optimal” angle computed for the nonuniform grid with four points, while the dotted line corresponds to the density  $n(x, t)$  computed for the angle  $\theta_0 = \pi/6$ .

In Figure 5, we plot the density  $n(x, t)$  for the modified Jin-Levermore scheme. The numerical oscillations have disappeared, and the diffusion coefficient is exactly the same that in the centered scheme. The continuous line corresponds to the “optimal” angle computed for the nonuniform grid with four points, while the dotted line corresponds to an angle equal to  $\pi/6$ .

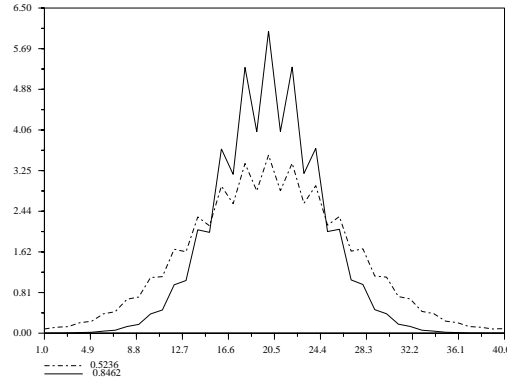
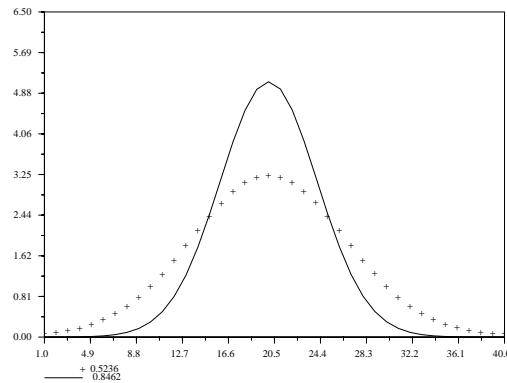

 FIGURE 4.  $\varepsilon$ -scheme.


FIGURE 5. Modified Jin-Levermore scheme.

In Figure 6, we compare the two densities in the case of the centered scheme (the means value curve) and in the case of the modified Jin-Levermore scheme. We note that the Dirac mass has diffused in the same way, as it was attended from the previous analysis.

In Figure 7, we plot the curves of the diffusion coefficient  $\nu(\theta)$  computed from the second moment, when  $\theta$  is varying. We compare the result found for the exact solution of the heat equation with the one obtained by means of the modified Jin-Levermore scheme. Both the curves intersect the value  $1/2$  for the optimal value of the angle  $\theta_0 = 0.864$ .

In Figure 8 we plot the density  $n(x, t)$  for four different values of  $\varepsilon$  at time 1. For small values of  $\varepsilon$  (0.01 and 0.00001) the curves are indistinguishable. For “large” values of  $\varepsilon$  (larger than 1), the solution is very close to the initial data. For intermediate values ( $\varepsilon = 0.5$ ), one observes the splitting of the initial delta measure into two peaks: one moving to the right the other to the left, as expected in such kinetic model.

Finally, in Figure 9 we plot the diffusion coefficient curves in function of  $\varepsilon = 0.01, \dots, 0.1$  for three different values of the discretization angle  $\theta_0 = \pi/6, \pi/3, 0.864$ . We note that for  $\theta_0 = 0.864$  (which is the angle computed by Maple for the non-uniform four points discretization) the diffusion coefficient  $\nu$  is close to  $1/2$  when  $\varepsilon = 0.01$ .

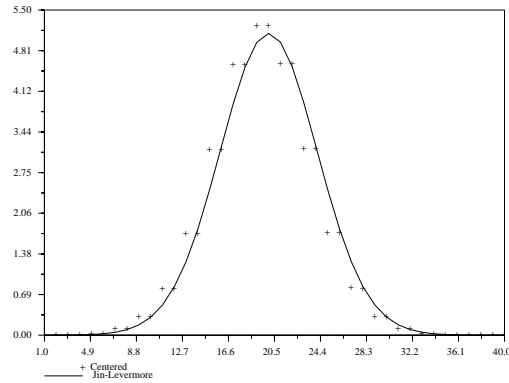
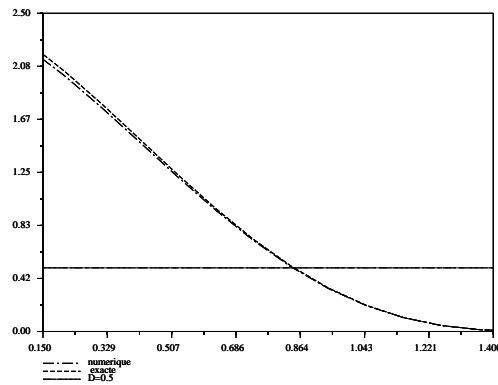


FIGURE 6. Modified Jin Levermore-centered scheme.

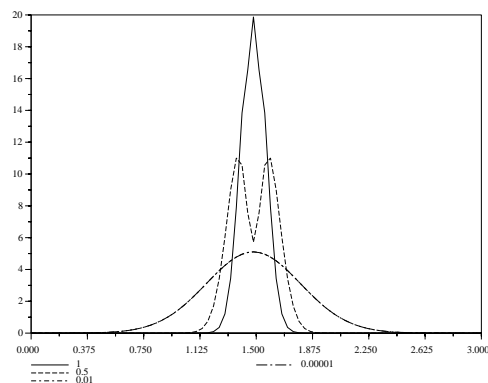
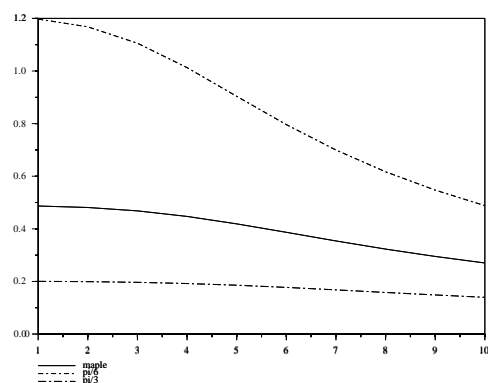
FIGURE 7. Diffusion coefficient w.r.t.  $\theta$ .

## 7. CONCLUSIONS

We first remark that in the  $\varepsilon$  scheme proposed in Section 4.3, it is possible to replace the upwind scheme by the modified Jin-Levermore scheme of Section 4.4, that has all the required properties for  $\varepsilon \leq h$ . On the other hand, if  $\varepsilon \geq h$ , then the upwind scheme seems sufficient. Thus, the ultimate choice for the discretization in space is, for example,

$$D = \max(0, 1 - \varepsilon/h)D^{1/2} + \min(1, \varepsilon/h)D^u. \quad (7.1)$$

Let us emphasize that every method leading to the good diffusion equation (see for example [15, 24–26]), requires the inversion of a huge (but sparse) matrix, of size  $N_x \times N_\theta$  where  $N_x$  is the number of discretization points in space and  $N_\theta$  is the number of discretization points in velocity. In fact, one needs implicit schemes in order to avoid too restrictive time step condition in terms of  $\varepsilon$ , and a splitting of space and velocity is not possible. Indeed, for all these methods in order to compute the fluxes we must solve a stationary problem (see for example


 FIGURE 8. Density for three values of  $\varepsilon$ .

 FIGURE 9. Diffusion coefficient  $\nu$  versus  $\varepsilon$ .

equation (2.3)) of the type:

$$v \cdot \nabla_x f^0 = \mathcal{L}(f^1).$$

Such a problem is necessarily multi-dimensional, *i.e.* one cannot split space directions  $x$  and velocity variables  $v$ . We also remark that there is no non-trivial hydro-dynamical limit (see Sect. A.1). Moreover, the bi-dimensional case with respect to the space variable can be treated by a similar approach but the directional splitting does not give the good diffusion coefficient (see Sect. A.2). Nevertheless, the finite volume approach described in Section 4.5, could be a way to design a fully multi-dimensional asymptotic preserving scheme.

Some questions remain open:

- Take into account boundary condition (or equivalently bounded region in  $x$ ): as in [13, 22–24], the solution has a boundary layer as  $\varepsilon \rightarrow 0$  which is treated using extrapolation length.
- Generalize to a non uniform grid in space and analyze the two dimensional case (in space) with unstructured meshes, or other geometries with for example spherical or axial symmetry.

- Consider inelastic collisions or force fields that couples the kinetic energy level *i.e.* consider a distribution function depending of the kinetic energy.
- Numerical test with regions where transport dominates ( $\varepsilon \gg 1$ ) and other highly collisional ( $\varepsilon \ll 1$ ) are needed in order to validate the choice proposed in (7.1).
- Study the positiveness and other properties of the solution when using Hilbert expansion method.

## A. APPENDIX: SOME FINAL REMARKS

### A.1. Hydro-dynamical limit

Consider the Boltzmann kinetic equation at a hydrodynamic regime:

$$\partial_t f + \cos \theta \partial_x f = \frac{1}{\varepsilon} \mathcal{L}(f).$$

Is it possible to make the same analysis as for the diffusion regime. We begin by the continuous level with  $\mathcal{L}(f) = \partial_{\theta\theta}^2 f$ . Expanding  $f$  in powers of  $\varepsilon$ , and identifying the equal orders, we get:

$$0 = \mathcal{L}(f^0), \quad (\text{A.1})$$

$$\partial_t f^0 + \cos \theta \partial_x f^0 = \mathcal{L}(f^1), \quad (\text{A.2})$$

$$\partial_t f^1 + \cos \theta \partial_x f^1 = \mathcal{L}(f^2). \quad (\text{A.3})$$

Then solving equation (A.1) gives that  $f^0$  is independent on the  $\theta$  variable:  $f^0 = f^0(x, t)$ . The solvability condition for equation (A.2) reads:

$$2\pi \partial_t f^0 + \int_0^{2\pi} \cos \theta \partial_x f^0 d\theta = 0,$$

which implies  $f^0$  independent on  $t$ , too. Thus,  $f^0 = f^0(x)$  and  $f^1 = f^1(x, \theta) = -\cos \theta \partial_x f^0$ . Finally, the solvability condition for equation (A.3) is given by:

$$\int_0^{2\pi} -(\cos \theta)^2 \partial_x^2 f^0 d\theta = 0,$$

which first gives  $f^0 = C'x + C$  but, thanks to the behavior at infinity, we finally get  $f^0 = C$ ; we also deduce that  $f^1 = 0$ . Thus, it shows that the hydro-dynamical limit is trivial.

### A.2. Bi-dimensional case

Let us consider a two dimensional case. We shall show that a directional splitting is not suitable since it leads to a wrong diffusion coefficient. We shall present the analysis on the simplest, centered case.

We denote by  $i$  the index for the  $x$  position variable, by  $k$  the index for the  $y$  position variable. For simplicity, we shall consider only the discretization with respect to the position variables  $x$  and  $y$ , keeping the velocity  $\theta$  and the time  $t$  at a continuous level. Applying a centered scheme to the kinetic equation, leads to:

$$\varepsilon \partial_t f_{i,k} + \cos \theta D^1 f_{i,k} + \sin \theta D^2 f_{i,k} = \frac{1}{\varepsilon} \mathcal{L} f_{i,k}, \quad (\text{A.4})$$

where

$$D^1 f_{i,k} = \left( \frac{f_{i+1,k} - f_{i-1,k}}{2\Delta x} \right)$$

and

$$D^2 f_{i,k} = \left( \frac{f_{i,k+1} - f_{i,k-1}}{2\Delta y} \right).$$

Performing the Hilbert expansion method with

$$f_{i,k} = f_{i,k}^0 + \varepsilon f_{i,k}^1 + \varepsilon^2 f_{i,k}^2,$$

we get at the order  $\varepsilon^{-1}$ :

$$f_{i,k} = f_{i,k}^0 = \rho_{i,k}, \quad \text{independent on } \theta.$$

Then at the order  $\varepsilon^0$  we obtain:

$$\cos \theta D^1 f_{i,k}^0 + \sin \theta D^2 f_{i,k}^0 = \mathcal{L} f_{i,k}^1, \quad (\text{A.5})$$

which gives:

$$f_{i,k}^1 = -\cos \theta D^1 f_{i,k}^0 - \sin \theta D^2 f_{i,k}^0.$$

Finally, at the order  $\varepsilon^1$ , we have:

$$\partial_t f_{i,k}^0 + \cos \theta D^1 f_{i,k}^1 + \sin \theta D^2 f_{i,k}^1 = \mathcal{L} f_{i,k}^2$$

and integrating with respect to  $\theta$  and replacing  $f^1$  by (A.5), this yields to the diffusion equation on a double mesh:

$$\partial_t \rho_{i,k} - \frac{1}{2} \left( \frac{\rho_{i+2,k} + \rho_{i-2,k} - 2\rho_{i,k}}{(2\Delta x)^2} \right) - \frac{1}{2} \left( \frac{\rho_{i,k+2} + \rho_{i,k-2} - 2\rho_{i,k}}{(2\Delta y)^2} \right) = 0.$$

Still this discretization has the same problems of the one in the one-dimensional case: it does not couple the odd and even grid points of the mesh.

We remark also that the same analysis of the modified Jin-Levermore scheme can be performed when using a rectangular structured mesh but it is much more complicated with a general mesh, see [18, 19, 45] for works in this direction. Finally, it seems that a directional splitting (alternative directions) for the modified Jin-Levermore scheme leads to a wrong diffusion coefficient. Indeed, (A.4) can be splitted in two parts:

$$\varepsilon \partial_t f_{i,k} = \left( -\cos \theta D^1 f_{i,k} + \frac{1}{2\varepsilon} \mathcal{L} f_{i,k} \right) + \left( -\sin \theta D^2 f_{i,k} + \frac{1}{2\varepsilon} \mathcal{L} f_{i,k} \right).$$

The first bracket corresponds to the derivative with respect to  $x$  and half of the collision, the second bracket corresponds to the derivative with respect to  $y$  and half of the collision. Then, using a splitting, one solves the first part which gives when  $\varepsilon$  tends to zero, the diffusion equation in  $x$  with coefficient  $\nu_x = 1/4$ , instead of  $1/2$ . Thus, in order to recover the right diffusion coefficient one may not use a directional splitting.

## REFERENCES

- [1] M.L. Adams, Subcell balance methods for radiative transfer on arbitrary grids. *Transport Theory Statist. Phys.* **27** (1997) 385–431.
- [2] R. Botchorishvili, B. Perthame and A. Vasseur, Equilibrium schemes for scalar conservation laws with stiff sources. *Inria report RR-3891* (2000), <http://www.inria.fr/RRRT/RR-3891.html>



- [3] C. Buet, S. Cordier and B. Lucquin-Desreux, The grazing collision limit for the Boltzmann–Lorentz model. *Asymptot. Anal.* **25** (2001) 93–107.
- [4] R.E. Caflisch, S. Jin and G. Russo, Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM J. Numer. Anal.* **34** (1997) 246–281.
- [5] G.Q. Chen, C.D. Levermore and T.P. Liu, Hyperbolic conservation laws with stiff relaxation terms and entropy. *Comm. Pure Appl. Math.* **47** (1994) 187–830.
- [6] S. Cordier, B. Lucquin-Desreux and A. Sabry, Numerical approximation of the Vlasov–Fokker–Planck–Lorentz model. *ESAIM: Proceed. CEMRACS 1999* (2001), <http://www.emath.fr/Maths/Proc/Vol.10>
- [7] P. Degond and B. Lucquin-Desreux, The Fokker–Planck asymptotics of the Boltzmann collision operator in the Coulomb case. *Math. Models Methods Appl. Sci.* **2** (1992) 167–182.
- [8] P. Degond and B. Lucquin-Desreux, The asymptotics of collision operators for two species of particles of disparate masses. *Math. Models Methods Appl. Sci.* **6** (1996) 405–436.
- [9] L. Desvillettes, On asymptotics of the Boltzmann equation when the collisions become grazing. *Transport Theory Statist. Phys.* **21** (1992) 259–276.
- [10] J. Glimm, G. Marshall and B.J. Plohr, A generalized Riemann problem for quasi one dimensional gas flows. *Adv. in Appl. Math.* **5** (1984) 1–30.
- [11] E. Godlewski and P.A. Raviart, Numerical approximations of hyperbolic systems of conservation laws. Springer-Verlag, New York, *Appl. Math. Sci.* **118** (1996).
- [12] S. Goldstein, On diffusion by discontinuous movements, and on the telegraph equation. *Quart. J. Mech. Appl. Math.* **4** (1951) 129–156.
- [13] F. Golse, S. Jin and C.D. Levermore, The convergence of numerical transfer schemes in diffusive regimes I: discrete-ordinate method. *SIAM J. Numer. Anal.* **36** (1999) 1333–1369.
- [14] L. Gosse, A priori error estimate for a well-balanced scheme designed for inhomogeneous scalar conservation laws. *C. R. Acad. Sci. Paris Sér. I Math.* **327** (1998) 467–472.
- [15] L. Gosse, A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. *Math. Models Methods Appl. Sci.* **11** (2001) 339–365.
- [16] L. Gosse and A.Y. Leroux, A well-balanced scheme designed for inhomogeneous scalar conservation laws. *C. R. Acad. Sci. Paris Sér. I Math.* **323** (1996) 543–546.
- [17] J.M. Greenberg and A.Y. Leroux, A well balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.* **33** (1996) 1–16.
- [18] F. Hermeline, A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.* **160** (2000) 481–499.
- [19] F. Hermeline, Two coupled particle-finite volume methods using Delaunay–Voronoi meshes for the approximation of Vlasov–Poisson and Vlasov–Maxwell equations. *J. Comput. Phys.* **106** (1993).
- [20] S. Jin, Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.* **21** (1999) 441–454.
- [21] S. Jin, Numerical integrations of systems of conservation laws of mixed type. *SIAM J. Appl. Math.* **55** (1995) 1536–1551.
- [22] S. Jin and C.D. Levermore, The discrete-ordinate method in diffusive regimes. *Transport Theory Statist. Phys.* **20** (1991) 413–439.
- [23] S. Jin and C.D. Levermore, Fully-discrete numerical transfer in diffusive regimes. *Transport Theory Statist. Phys.* **22** (1993) 739–791.
- [24] S. Jin and C.D. Levermore, Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. *J. Comput. Phys.* **126** (1996) 449–467.
- [25] S. Jin and L. Pareschi, Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes. *J. Comput. Phys.* **161** (2000) 312–330.
- [26] S. Jin, L. Pareschi and G. Toscani, Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. *SIAM J. Numer. Anal.* **35** (1998) 2405–2439.
- [27] S. Jin, L. Pareschi and G. Toscani, Uniformly accurate diffusive relaxation schemes for multiscale transport equations. *SIAM J. Numer. Anal.* (2000).
- [28] S. Jin and Z. Xin, The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm. Pure Appl. Math.* **XLVIII** (1995) 235–276.
- [29] A. Klar, An asymptotic-induced scheme for non stationary transport equations in the diffusive limit. *SIAM J. Numer. Anal.* **35** (1998) 1073–1094.
- [30] E.W. Larsen, The asymptotic diffusion limit of discretized transport problems. *Nuclear Sci. Eng.* **112** (1992) 336–346.
- [31] E.W. Larsen and J.E. Morel, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. II. *J. Comput. Phys.* **83** (1989) 212–236.
- [32] E.W. Larsen, J.E. Morel and W.F. Miller Jr., Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *J. Comput. Phys.* **69** (1987) 283–324.

- [33] R.J. LeVeque, Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *J. Comput. Phys.* **146** (1998) 346–365.
- [34] P.L. Lions, B. Perthame and P.E. Souganidis, Existence of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates. *Comm. Pure Appl. Math.* **49** (1996) 599–638.
- [35] B. Lucquin-Desreux, Diffusion of electrons by multicharged ions. *Math. Models Methods Appl. Sci.* **10** (2000) 409–440.
- [36] B. Lucquin-Desreux and S. Mancini, A finite element approximation of grazing collisions (submitted).
- [37] P.A. Markowich, C. Ringhoffer and C. Schmeiser, *Semiconductor equations*. Springer-Verlag (1994).
- [38] W.F. Miller Jr. and T. Noh, Finite differences *versus* finite elements in slab geometry, even-parity transport theory. *Transport Theory Statist. Phys.* **22** (1993) 247–270.
- [39] J.E. Morel, T.A. Wareing and K. Smith, A linear-discontinuous spatial differencing scheme for  $S_n$  radiative transfer calculations. *J. Comput. Phys.* **128** (1996) 445–462.
- [40] G. Naldi and L. Pareschi, Numerical schemes for kinetic equations in diffusive regimes. *Appl. Math. Lett.* **11** (1998) 29–55.
- [41] L. Pareschi, Central differencing based numerical schemes for hyperbolic conservation laws with relaxation terms. *J. Num. Anal.* (to appear).
- [42] B. Perthame, *An introduction to kinetic schemes for gas dynamics. An introduction to recent developments in theory and numerics for conservation laws*. L.N. in Computational Sc. and Eng., 5, D. Kroner, M. Ohlberger and C. Rohde Eds., Springer (1998).
- [43] K.H. Prendergast and K. Xu, Numerical hydrodynamics for gas-kinetic theory. *J. Comput. Phys.* **109** (1993) 53–66.
- [44] K.H. Prendergast and K. Xu, Numerical Navier-Stokes solutions from gas kinetic theory. *J. Comput. Phys.* **114** (1994) 9–17.
- [45] G. Samba, Limite asymptotique d'un schéma d'éléments finis linéaires discontinus lumpés en régime diffusion. *Rapport CEA* (to appear).
- [46] G.I. Taylor, Diffusion by continuous movements. *Proc. London Math. Soc.* **20** (1921) 196–212.
- [47] B. Vanleer, On the relation between the upwind differencing schemes of Engquist-Osher, Godunov and Roe. *SIAM J. Sci. Stat. Comp.* **5** (1984) 1–20.

# Numerical method for the Compton scattering operator.

C. Buet<sup>1</sup>, S. Cordier<sup>2</sup>

<sup>1</sup>Commissariat à l'Énergie Atomique, 91680 Bruyères-le-Châtel  
email `christophe.buet@cea.fr`

<sup>2</sup>Laboratoire MAPMO, UMR 6628,  
Université Orléans, 45067 Orléans, France  
email `Stephane.Cordier@univ-orleans.fr`

**Abstract** In this paper, we present a new discretization of the Kompaneets equation. This equation can be derived from the Quantum Boltzmann equation for cross sections localized around small changes in the energy. The two collision operators share the same properties (conservation, entropy and steady states). The numerical methods are designed in order to preserve these properties which insures the correct long time behaviour. A supplementary difficulty arises for initial data with a density larger than a critical one, associated with a Planck distribution. In this case, the equilibrium state are the sum of a smooth Planck distribution plus a Delta measure on zero energy. The scheme is able to deal with this singularity.

## 1 Introduction

We are interested in the dynamics of a low energy, homogeneous and isotropic photon gas that interacts via Compton scattering with a low energy electron gas in thermodynamical equilibrium.

In this paper, we first review the main properties of the Quantum Boltzmann equation (QBE) and its "grazing" collision limit that is called the Kompaneets equation (which is a Fokker-Planck like equation). We will recall recent result of Escobedo and Mischler [EM, EM2] about the equilibrium state of such equation that behaves under some conditions as a "concentration near the origin" (section 2).

In section 3.1, we present a numerical scheme for the QBE that preserves the properties of the equation described above.

Then, we present a numerical scheme for the Kompaneets equations that has the same properties i.e. are compatible with all the properties of the continuous equation. This allows us to observe the "concentration" phenomena. We present two ways for deriving this scheme. The first way is based on the asymptotic that transform the QBE into the Kompaneets equation at the discrete level. The second one is based on the same ideas as the Chang-Cooper paper [CC]. These two methods lead surprisingly to the same scheme up to a multiplicative constant (that goes to 1 when the mesh is refined) and this fact allows us to use either method usually devoted to Boltzmann like equation (e.g. entropy decay) or to Fokker-Planck type equation (Maximum principle).

We shall also deal with the difficulty due to the mass concentration phenomena. For this, we shall define generalized Bose-Einstein distribution function (with negative value of  $\nu$ ) that converge (when the mesh is refined) toward the Planck distribution plus a Dirac measure at origin.

We then illustrate the schemes on the following result by Escobedo et al. [E3] : even if the initial density is smaller than the Planck one, the solution of the Kompaneets equation (with the boundary conditions) is not always global in time (on the contrary, the solutions of the QBE are well defined for any time) provided that the initial density (arbitrary small) is close enough to the origin. This is due to a balance between the "Burgers" term (with a negative velocity that push the particle toward the origin and the diffusion term that spreads the solution. We shall observe this "blow-up" in finite time.

This paper is related to previous works of the authors on the Boltzmann or Fokker-Planck-Landau operator [BCDL, BC]. In these papers, numerical discretizations are designed in order to be compatible with relevant physical properties such as conservation laws and entropy decay using Discrete Velocities Method. These methods allow to insure the correct large time behaviour i.e. the trend to thermodynamical equilibrium. The numerical efficiency of the proposed algorithms have been tested also for the non homogeneous case (where the distribution function depend on the space variable too) in a paper by the authors and F. Filbet [BCF].

## 2 Quantum Boltzmann and Kompaneets Equation

In this section, we briefly present the equation we are interested in, namely the Quantum Boltzmann equation and its "grazing" limit the so called Kom-

paneets equation. We refer to [K] for the original paper on these equations and to [EM, EM2] for a more recent and mathematical presentation.

## 2.1 The Quantum Boltzmann Equation

We consider an isotropic and homogeneous photon gas scattered by cold electrons at thermodynamical equilibrium. The distribution function of the photons  $f$  depends on time and on the energy variable  $k$  :  $f(k, t)$  represents the number of photons that have energy  $k \geq 0$  at time  $t \geq 0$ . This function  $f$  obeys to the following scaled quantum Boltzmann equation (QBE)

$$k^2 \frac{\partial f}{\partial t} = \int_0^\infty b(k', k)(f'(1 + f) \exp(-k) - f(1 + f') \exp(-k')) dk', \quad (2.1)$$

where we omit the variable  $k$  and  $t$  for simplicity and  $f'$  as usual denotes  $f(k', t)$ . The so called cross section  $b$  - positive and symmetric- is related to the probability for a given particle at energy level  $k$  to be scattered to the energy level  $k'$ . The exponential terms represents the distribution function of electrons. Defining  $g = k^2 f$ , equation (2.1) can be rewritten as

$$\frac{\partial g}{\partial t} = \int_0^\infty b(k', k)(g'(k^2 + g) \exp(-k) - g(k'^2 + g') \exp(-k')) dk', \quad (2.2)$$

Let us define the function  $h$  of two variable as

$$h(g, k) \stackrel{def}{=} \frac{g}{k^2 + g}. \quad (2.3)$$

In the reminder, we shall also note  $h = h(g, k) = h(g(k), k)$ . Using this notation, the QBE equation (2.2) reads in the more symmetric form

$$\frac{\partial g}{\partial t} = \int_0^\infty b(k', k) e^k e^{k'} (k^2 + g)(k'^2 + g')(h(g', k') e^{k'} - h(g, k) e^k) dk', \quad (2.4)$$

In the reminder of the paper, we write  $QB$  for the Quantum Boltzmann operator defined above as the right hand side in (2.4).

Multiplying (2.4) by a test function  $\Psi(k)$ , integrating over  $k$  and replacing  $k$  by  $k'$ , we have the weak symmetrized formulation of the QB operator

$$\begin{aligned} \int_0^\infty \frac{\partial g}{\partial t} \Psi(k) dk &= \frac{1}{2} \int_0^\infty \int_0^\infty (\Psi(k) - \Psi(k')) b(k', k) e^{-k} e^{-k'} \\ &\quad (k^2 + g)(k'^2 + g')(h(g', k') e^{k'} - h(g, k) e^k) dk' dk. \end{aligned} \quad (2.5)$$

Using this weak formulation, one obtains easily the unique law of conservation for the QBE, that is the conservation of the total density

$$N \stackrel{\text{def}}{=} N(g) = \int_0^\infty g dk, \quad \frac{d}{dt} N(g) = 0. \quad (2.6)$$

The second main property of the QBE is the decay of entropy. We first define the function  $s(x, k)$  of two variables as

$$s(x, k) \stackrel{\text{def}}{=} x \log(x) + k^2 \log(k^2) + kx - (k^2 + x) \log(k^2 + x),$$

and the functional entropy,  $H$  as

$$H(g) \stackrel{\text{def}}{=} \int_0^\infty s(g(k), k) dk \quad (2.7)$$

Then,  $H$  decays in time

$$\frac{d}{dt} H \leq 0.$$

This property, so called  $H$ -theorem can be easily checked, at least formally, on the weak formulation of the QBE. Indeed, we have

$$\frac{d}{dt} s(g(k), k) = \log\left(\frac{g \exp(k)}{k^2 + g}\right) \frac{d}{dt} g(k).$$

Hence, the result is obtained by choosing

$$\Psi(k) = \log(h(g, k) \exp(k)) = \log\left(\frac{g \exp(k)}{k^2 + g}\right).$$

in the weak form (2.5) using that  $b \geq 0$ .

## 2.2 The Kompaneets equation

When the cross section  $b$  concentrates on small modification of the energy (this asymptotic can be related to the grazing collision limit of the classical Boltzmann equation, see [DV]), one obtains a Fokker-Planck type equation that has first been derived by Kompaneets in [K]. This equation shares the same properties (total density,  $N$ , conservation and entropy decay) with the QBE. Let us refer to [EM2] for recent mathematical work on this equation. It is proved that solution of the QBE tends to the solution of the Kompaneets equation which we shall now describe.

In this "grazing" collision limit, which is detailed in the above reference and presented at a discrete level in section 3.2.1, the QBE becomes

$$\partial_t f(k, t) = k^{-2} \partial_k (k^4 (f + f^2 + \partial_k f)), \quad (2.8)$$

i.e. the sum of a convective term, a Burgers like nonlinear term and a diffusive term with weights  $k^2$  and  $k^4$ . This so called Kompaneets equation (denoted by K equation in this article) can be equivalently written as

$$\partial_t f(k, t) = k^{-2} \partial_k \left( k^4 (1 + f)^2 \left( \frac{f}{1 + f} + \partial_k \frac{f}{1 + f} \right) \right). \quad (2.9)$$

A third expression of the Kompaneets equation reads

$$\partial_t g = \partial_k \left( \exp -(k + \mu) (g + k^2)^2 \partial_k \frac{h(g, k)}{h(B_\mu, k)} \right), \quad (2.10)$$

where we use  $g = k^2 f$ ,  $h$  defined by (2.3) and the so called Bose-Einstein state (see section 2.3)

$$B_\mu(k) \stackrel{def}{=} \frac{k^2}{\exp(k + \mu) - 1}, \quad (2.11)$$

for  $\mu \geq 0$ . This is easily checked using the identity  $h(B_\mu(k), k) = \exp -(k + \mu)$ .

The properties of this equation can be verified on its weak formulation. Multiplying (2.10) by a test function  $\Psi(k)$  and integrating over  $k$ , we have the weak formulation of the Kompaneets equation

$$\int_0^\infty \frac{\partial g}{\partial t} \Psi(k) dk = \int_0^\infty \partial_k \Psi(k) \exp -(k + \mu) (g + k^2)^2 \partial_k \frac{h(g, k)}{h(B_\mu, k)} dk. \quad (2.12)$$

The  $H$  functional entropy (resp.  $N$  total density) defined by (2.7) (resp. by (2.6)) is decaying (resp. constant) as it can be checked using  $\Psi = \log(h(g, k) \exp(k))$  (resp.  $\Psi = 1$ ) in the above weak form.

As usual (see [BC, BCF, BCDL] ), the formulations are (formally) equivalent at the continuous level. This is no more the case once discretized : the properties of the discretized equation will depend of the form which is discretized.

## 2.3 Entropy and equilibrium states

Let us now turn to the equilibrium states or stationary solutions of QB or K equations. Such functions should minimize the entropy  $H$  for a fixed total density  $N$ . The function  $s$  being concave with respect to its first variable, it is easy to find the minimum as zero of its derivative (with respect to  $x$ ) for any fixed  $k$ . This gives the so-called Bose-Einstein distribution  $B_\mu(k)$ , defined by (2.11), with  $\mu$ . The coefficient  $\mu$  has to be positive in order that the density is finite i.e. the function  $B_\mu$  is integrable. The limit case  $\mu = 0$  is related to the Planck distribution. This parameter  $\mu$  is a function of the density - we denote by  $N_\mu = N(B_\mu)$ -

and this function is decreasing since the Bose-Einstein states are ordered : if  $\mu < \nu$  then  $B_\mu > B_\nu$ . The application  $\mu \mapsto N_\mu$  maps  $[0, \infty)$  into  $]0, N(B_0)]$ .

Thus, there is a critical or maximal density  $N_0 = N(B_0)$  that corresponds to the Planck distribution  $\nu = 0$ . For any initial density with a density greater than  $N_0$ , there is no classical equilibrium state. Caffisch and Levermore prove in [CD] that, in this case, the minimum of entropy is realized by the Planck distribution  $B_0$  plus a Dirac measure located at  $k = 0$ ,  $\delta_0$ , that does not change the value of the entropy  $H$  i.e.  $H(B_0 + \alpha\delta_0) = H(B_0)$ .

We summarize these results in the following H-theorem for the QB or the K equation:

**Theorem 2.1** *For any positive weak solution of QB or K equation with initial data  $g^0$ , one has  $\frac{d}{dt}H(g) \leq 0$  with equality if and only if  $g = M$ , the equilibrium state  $M$  being defined by*

- if  $N(g^0) \leq N_0$  then  $M = B_\mu$  with  $\mu \geq 0$  such that  $N(g^0) = N(B_\mu)$
- if  $N(g^0) > N_0$  then  $M = B_0 + (N(g^0) - N_0)\delta_0$

For the details and the proof of this result we refer to [CD, EM, EM2] .

Escobedo and Mischler study the evolution problem and prove the convergence in large time of the (weak) solution of QBE toward these equilibrium states starting from smooth initial data. More precisely, they prove that, in the second case i.e. for an initial density larger than the critical one, the regular part of the solution converges to the Planck distribution (in  $L^1(\varepsilon, \infty)$  for any  $\varepsilon > 0$ ) and the reminding part of the density condensates near the origin.

Let us precise the result for the Cauchy problem associated with the QB or the K equation with an initial data  $g^0 \in S$  defined by

$$S = \{g \in L^1_{[0, \infty[} \cap L^\infty_{[0, \infty[}, kg \in L^1_{[0, \infty[} g \log(g) \in L^1_{[0, \infty[}\}.$$

The existence and uniqueness for the QBE in this space is proved in [EM, EM2] . However, this hypothesis for  $g^0$  is not sufficient to guarantee the existence of a global positive solution for the K equation. Blow up in finite time can occur even if  $N(g^0) \leq N(P)$ . A sufficient condition to ensure existence of global in time, positive solution is that  $g^0 \leq P$ .



We also use another form of the entropy, so called relative entropy. We note  $H(g|M) = H(g) - H(M)$  with  $M$  the equilibrium state associated with  $g^0 \in S$ , defined in the theorem 2.1. A simple calculation gives

$$H(g|M) = \int (g \log(g/M) - (k^2 + g) \log((k^2 + g)/(k^2 + M))) dk. \quad (2.13)$$

## 2.4 A Maximum principle and a Czizar-Kullback like inequality

We first prove a maximum principle for positive and global solutions  $g \in C^1([0, \infty[, S)$  for the QB or the K equation. Our result can be stated as

**Lemma 2.1** *Let  $g \in C^1([0, \infty[, S)$  such that the  $H$ -theorem holds i.e. the functional defined by (2.7) decays*

- *if there exists  $\mu$  such that  $\alpha \stackrel{\text{def}}{=} \sup \frac{h(g^0, k)}{h(B_\mu, k)} < +\infty$  then  $\sup \frac{h(g, k)}{h(B_\mu, k)} \leq \alpha$ .*
- *if there exists  $\mu$  such that  $\frac{h(g^0, k)}{h(B_\mu, k)} \geq 1$  then  $\frac{h(g, k)}{h(B_\mu, k)} \geq 1$ .*

Using that for any  $\mu \geq 0$ ,  $h(B_\mu, k) = \exp(-(k + \mu))$ , we can write the QBE (2.5) as

$$\begin{aligned} \int_0^\infty \frac{\partial g}{\partial t} \Psi(k) dk &= \frac{1}{2} \int_0^\infty \int_0^\infty (\Psi(k) - \Psi(k')) b(k', k) \exp(-(k + \mu)) \exp(-(k' + \mu)) \\ &\quad (k^2 + g)(k'^2 + g') \left( \frac{h(g', k')}{h(B_\mu, k')} - \frac{h(g, k)}{h(B_\mu, k)} \right) dk' dk. \end{aligned} \quad (2.14)$$

and the K equation as (2.12) These weak formulations are useful to prove the decay of the relative entropy. The proof of this Lemma is postponed in Appendix A.

Note that in the first case (  $\sup \frac{h(g^0, k)}{h(B_\mu, k)} < +\infty$  ) the solution behaves as  $g \sim Ck^2 \exp(-k)$  for sufficiently large  $k$ .

**Remark 2.2** *This result gives an upper bound of the solution only for  $k > \max(0, \log(\alpha))$ .*

For some special cases we deduce the following corollary

**Corollary 2.3** *If  $g \in C^1([0, \infty[, S)$  and if there exist  $\mu_1$  and  $\mu_2$  such that  $B_{\mu_2} \leq g^0 \leq B_{\mu_1}$  then  $B_{\mu_2} \leq g \leq B_{\mu_1}$ .*

*Proof.* We have just to prove the second inequality since the first is exactly the second part of Lemma 2.1. Since the function  $h(g, k)$  is monotone in  $g$  for fixed  $k$ , we have

$$\frac{h(g^0, k)}{h(B_{\mu_1}, k)} = \alpha \leq 1.$$

Lemma 2.1 gives  $h(g, k) \leq h(B_{\mu_1}, k)$  and using again the fact that  $h(g, k)$  is increasing in the variable  $g$  the result follows.  $\square$

Concerning the trend to equilibrium, we prove the following Czar-Kullback like inequality

**Lemma 2.2** *If  $g \in C^1([0, \infty[, S)$  and  $\sup \frac{h(g^0, k)}{h(B_0, k)} < +\infty$  then for any  $k_0 > 0$  there exists a constant  $C$  depending only on  $k_0, N(g^0)$  and  $\sup \frac{h(g^0, k)}{h(B_0, k)}$  such that*

$$\|g - M\|_{L^1_{[k_0, \infty[}} \leq CH(g|M)^{\frac{1}{2}}.$$

*Proof.* The proof is divided in two parts, one part for "small"  $k$  and the second one for "large"  $k$ . The entropy  $H(g|M)$  is defined by (2.13). Using the identities  $1 - h(g, k) = \frac{k^2}{g+k^2}$  and  $\frac{1-h(M, k)}{1-h(g, k)} = \frac{k^2+g}{k^2+M}$ , we can rewrite  $H$  as

$$H(g|M) = \int (k^2 + g) \left( h(g, k) \log \frac{h(g, k)}{h(M, k)} + (1 - h(g, k)) \log \frac{1 - h(g, k)}{1 - h(M, k)} \right) dk.$$

Then, using the inequality, see [kullback,]

$$p_1 \log \left( \frac{p_1}{p_2} \right) + (1 - p_1) \log \left( \frac{1 - p_1}{1 - p_2} \right) \geq 2(p_1 - p_2)^2 + \frac{3}{4}(p_1 - p_2)^4 \geq 2(p_1 - p_2)^2, \quad (2.15)$$

which holds for any  $0 \leq p_1 \leq 1$  and  $0 \leq p_2 \leq 1$ , for  $p_1 = h(g, k)$  and  $p_2 = h(M, k)$ , we obtain

$$H(g|M) \geq 2 \int (k^2 + g) \left( \frac{g}{k^2 + g} - \frac{M}{k^2 + M} \right)^2 dk.$$

Note that

$$(k^2 + g) \left( \frac{g}{k^2 + g} - \frac{M}{k^2 + M} \right)^2 = \frac{k^4 |g - M|^2}{(k^2 + g)(k^2 + M)^2}.$$

Then, for any  $k_0 > 0$  and  $R > k_0$ , we obtain using Cauchy-Schwarz inequality

$$\int_{k_0}^R |g - M| dk \leq \left( \int_{k_0}^R \frac{(k^2 + g)(k^2 + M)^2}{k^4} dk \right)^{1/2} \left( \int_{k_0}^R \frac{k^4 |g - M|^2}{(k^2 + g)(k^2 + M)^2} dk \right)^{1/2}.$$

It is easy to check that there exists  $C_1 > 0$  such that

$$\left( \int_{k_0}^R \frac{(k^2 + g)(k^2 + M)^2}{k^4} dk \right)^{1/2} \leq C_1(k_0, R, N(g_0)),$$

with  $C_1(k_0, R)$  which goes to  $+\infty$  as  $k_0 \rightarrow 0$  or  $R \rightarrow \infty$ . Thus, we have

$$\int_{k_0}^R |g - M| dk \leq C_1(k_0, R, N(g_0)) H(g|m). \quad (2.16)$$

We now consider the case of "large"  $k$ . Define  $y(t)$  as

$$y(t) = H(M^t), \quad M^t \stackrel{\text{def}}{=} M + t(g - M).$$

Thus

$$y'(t) = \int_{k \geq 0} (g - M) \partial_g s(M^t) dk,$$

and using  $M^{t=0} = M$ , which is the minimum of  $s$ , we have  $y'(t=0) = 0$ ,

$$y''(t) = \int_{k \geq 0} (g - M)^2 \partial_{gg} s(M^t) dk = \int_{k \geq 0} \frac{k^2 (g - M)^2}{M^t (M^t + k^2)} dk.$$

Using now the Taylor's formula with integral reminder gives us

$$y(t) = H(M) + \frac{1}{2} \int_0^1 (1 - z) \int_{k \geq 0} \frac{k^2 (g - M)^2}{M^z (M^z + k^2)} dk dz$$

Using  $M^t = M + t(g - M) \leq M + g$ , we obtain

$$y(t) \geq H(M) + \frac{1}{2} \int_{k \geq 0} \frac{k^2 (g - M)^2}{(M + g)(M + g + k^2)} dk$$

that is, by taking  $t = 1$  i.e.  $M^{t=1} = g$  and  $y(t=1) = H(g)$

$$\int_{k \geq 0} \frac{k^2 (g - M)^2}{(M + g)(M + g + k^2)} dk \leq 2H(g|M).$$

Using again Cauchy-Schwarz inequality, for all  $R \geq 0$

$$\int_{k \geq R} |g - M| dk \leq \left( \int_{k \geq R} \frac{(M + g)(k^2 + M + g)}{k^2} dk \right)^{1/2} \left( \int_{k \geq R} \frac{k^2 |g - M|^2}{(M + g)(g + k^2 + M)} dk \right)^{1/2},$$

and

$$\int_{k \geq R} \frac{(M + g)(k^2 + M + g)}{k^2} dk = \int_{k \geq R} (g + M) dk + \int_{k \geq R} \frac{(g + M)^2}{k^2},$$

the first term of the r.h.s. is bounded by  $2N(g^0)$ . For the second term, since we have assumed that  $\alpha = \sup \frac{h(g^0, k)}{h(B_0, k)} < +\infty$ , and using lemma 2.1, we have  $\sup \frac{h(g, k)}{h(B_0, k)} < \alpha$  and, for  $k > \max(\log(\alpha), 0)$ , we have  $g \leq \frac{\alpha k^2 \exp(-k)}{1 - \alpha \exp(-k)}$ . Choosing  $R > \log(\alpha)$  there exists a constant  $C_2 = C_2(N(g^0), R)$  such that

$$\int_{k \geq R} |g - M| dk \leq C_2 \left( \int_{k \geq R} \frac{k^2 |g - M|^2}{(M + g)(g + k^2 + M)^2} dk \right)^{1/2},$$

which gives us

$$\int_{k \geq R} |g - M| dk \leq C_2 H(g|M)^{1/2}. \quad (2.17)$$

Using  $R = \alpha$  in (2.16) and in (2.17), we have

$$\|g - M\|_{L^1_{[k_0, \infty[}} \leq C H(g|M)^{1/2},$$

with  $C = C(k_0, N(g^0), \sup \frac{h(g^0, k)}{h(P, k)}) = \max(C_1, C_2)$  as we claimed.  $\square$

**Remark 2.4** *It can be checked that the constant  $C = C(k_0, N(g^0), \sup \frac{h(g^0, k)}{h(P, k)})$  blows up as  $k_0$  tends to 0. However if there exists  $\mu$  such that  $g^0 \leq B_\mu$  and  $\mu > 0$ , then by means of the corollary 2.3,  $g \leq B_\mu$ . This implies that  $\frac{g+M}{k^2}$  is bounded by  $\frac{1}{\exp(k+\mu)-1}$  which is  $L^1$  and this is still true with  $k_0 = 0$  and thus the constant only depends on the initial data  $C = C(g^0)$ .*

In some case, the entropy permits to control the trend of  $g$  toward  $M$ .

**Lemma 2.3** *For initial data such that  $g^0 < B_0$ , the solution of the K equation verifies  $\lim_{t \rightarrow +\infty} H(g|M) = 0$*

*Proof.* We know that for such initial data that there exists a global in time positive solution, [EM, EM2]. Thus,  $H(g|M)$  is well defined at any time,  $C^1$  in  $t$ , monotone decreasing and positive. Thus, there exists  $a \geq 0$  such that  $\lim_{t \rightarrow +\infty} H(g|M) = a$  and an increasing sequence  $t_n \rightarrow \infty$  such that  $H'(g_n|M)(t_n) \rightarrow 0$ , with  $g_n = g(t_n)$ , as  $n \rightarrow \infty$ .

Since for all  $n$ ,  $g_n \leq B_0$ , using Dunford-Pettis theorem, up to an extraction  $g_n \rightarrow \bar{g}$ , in the sense of measures and  $\bar{g}$  is such that  $H'(\bar{g}|M) = 0$ , that is  $\bar{g} = M$ . We can also prove that, up to an extraction,  $\{g_n\}$  is bounded in  $W^{1,1}_{[k_0, \infty[}$ , for any  $k_0 > 0$ . The K equation can be written as

$$\frac{\partial g}{\partial t} = \frac{\partial F}{\partial k}$$

with  $F$  defined by  $F = (k^2 - 2k)g + g^2 + k^2 \frac{\partial g}{\partial k}$ . Thus

$$\frac{\partial g}{\partial k} = \frac{1}{k^2} (F - (k^2 - 2k)g - g^2),$$

and then

$$\int_{k_0}^{+\infty} \left| \frac{\partial g}{\partial k} \right| dk \leq \int_{k_0}^{+\infty} \frac{1}{k^2} |F| dk + \int_{k_0}^{+\infty} \frac{1}{k^2} |(k^2 - 2k)g + g^2| dk. \quad (2.18)$$

Since  $g \leq B_0$ , the second term of the r.h.s. of the above inequality (2.18) is clearly bounded by a constant which does not depend on  $k_0$ . Using Cauchy-Schwarz inequality, we have

$$\int_{k_0}^{+\infty} \frac{1}{k^2} |F| dk \leq \left( \int_{k_0}^{+\infty} \frac{1}{k^2} \frac{g(k^2 + g)}{k^4} dk \right)^{1/2} \left( \int_{k_0}^{+\infty} \frac{1}{k^2} \frac{F^2}{g(k^2 + g)} dk \right)^{1/2}.$$

Let us now consider the production of entropy

$$\frac{dH}{dt} = \int_k \log \left( \frac{h(g)}{h(B_0)} \right) \frac{dg}{dt} dk.$$

Then, using  $\frac{\partial g}{\partial t} = \frac{\partial F}{\partial k}$  and integrating by parts, one get

$$\frac{dH}{dt} = \int_k \partial_k \left( \log \left( \frac{h(g)}{h(B_0)} \right) \right) F dk,$$

or, using the definition (2.3) of  $h$ ,

$$\frac{dH}{dt} = - \int_0^{+\infty} \frac{1}{k^2} \frac{F^2}{g(k^2 + g)} dk,$$

and then

$$\int_{k_0}^{+\infty} \frac{1}{k^2} |F| dk \leq \left( \int_{k_0}^{+\infty} \frac{g(k^2 + g)}{k^4} dk \right)^{1/2} \left\| \frac{dH}{dt} \right\|^{1/2}.$$

Using one more time  $g \leq B_0$ , one can also verify that

$$\int_{k_0}^{+\infty} \frac{1}{k^2} \frac{g(k^2 + g)}{k^4} dk,$$

is bounded by a constant which depends only of  $k_0$ . The result can be summarized as, for each  $k_0$  there exist two constants  $C_1$  and  $C_2(k_0)$  such that

$$\int_{k_0}^{+\infty} \left| \frac{\partial g}{\partial k} \right| dk \leq C_1 + C_2(k_0) \left\| \frac{dH}{dt} \right\|^{1/2}.$$

By construction, the sequence  $g_n$  is such that  $H'(g)$  tends to zero. Thus, the sequence  $H'(g_n)$  is bounded and thus for any  $k_0 > 0$ , the sequence  $g_n$  is bounded in  $W_{[k_0, \infty[}^{1,1}$ .

Now by means of this estimate, Helly theorem and diagonal extraction, up to an extraction we have  $g_n \rightarrow M$  a.e.. One can also easily check that if  $0 \leq g \leq B_0$  then  $|s(g, k)| \leq |s(B_0, k)|$  and  $s(B_0, k)$  is indeed in  $L^1([0, \infty[)$ . Thus, using Lebesgue dominated convergence theorem

$$\lim_{t_n \rightarrow +\infty} H(g_n) = H(M),$$

that is  $a = 0$ . □

### 3 Semidiscretization

Let us now turn to the discretization in the energy variable of the QB and of the K equation. We consider an uniform grid in  $k$  denoted by

$$k_i = ih$$

for  $i = 0 \cdots n$ , with  $k_0 = nh$ .

#### 3.1 Semidiscretization for the QBE

We shall detail the scheme for two asymptotic cross sections, the uniform one and the grazing or concentrated one.

Let us recall the QBE, once restricted to a bounded domain

$$\frac{\partial g}{\partial t} = \int_0^{k_0} b(k, k')((k^2 + g)e^{-k}g' - (k'^2 + g')e^{-k'}g)dk'. \quad (3.1)$$

and the associated symmetrized weak formulation reads

$$\int_0^{k_0} \Psi \frac{\partial g}{\partial t} = \frac{1}{2} \int_0^{k_0} \int_0^{k_0} b((k^2 + g)e^{-k}g' - (k'^2 + g')e^{-k'}g)(\Psi - \Psi')dk'dk, \quad (3.2)$$

for any test function  $\Psi$ . The discretization of (3.1) is based on a standard quadrature formula for the above integral

$$\frac{\partial g_i}{\partial t} = h \sum_{j=0}^N b(k_i, k_j) (g_j(k_i^2 + g_i)e^{-k_i} - g_i(k_j^2 + g_j)e^{-k_j}). \quad (3.3)$$

This gives a numerical method that is conservative and entropy decaying. Indeed, from (3.3) we have the discrete analogue of (3.2)

$$\sum_{i=0}^N \Psi_i \frac{\partial g_i}{\partial t} = \frac{h^2}{2} \sum_{i=0}^N \sum_{j=0}^N b(k_i, k_j) g_i g_j \left( \frac{(k_i^2 + g_i) e^{-k_i}}{g_i} - \frac{(k_j^2 + g_j) e^{-k_j}}{g_j} \right) (\Psi_i - \Psi_j). \quad (3.4)$$

Then, using the above weak discretized formulation, we obtain the discrete version of the H-theorem for the QBE with the discrete entropy defined by  $H(g) = h \sum_{i=0}^N s(g_i, k_i)$ , that is  $\frac{dH}{dt} \leq 0$ . By construction Bose-Einstein functions are equilibrium points of the discrete QBE. The relations defining the discrete equilibrium states,  $\frac{dH}{dt} = 0$ , imply that (with  $P_i = (B_0)_i$ )

- $\frac{h(g_i, k_i)}{h(P_i, k_i)} = \frac{h(g_j, k_j)}{h(P_j, k_j)}$  for all  $i, j > 0$ .
- $g_0 (g_j - \exp(-k_j)(k_j^2 + g_j)) = 0$ , for all  $j$ .

If  $g_0(t = 0) = 0$ , one can check on (3.3) that  $g_0(t) = 0$  for all  $t$  and, in this case, the equilibrium states  $M$  associated to  $g$  are Bose-Einstein function  $B_\mu$  eventually with  $\mu$  negative. This fact happens, as we will see also for the K equation, in the case of  $N(g^0) > N(B_0)$ .

If  $g_0(0) \neq 0$ , the situation is quite different: if  $N(g^0) \leq N(B_0)$  then necessarily  $M = B_\mu$  with a positive  $\mu$ , which implies that  $\lim_{t \rightarrow +\infty} g(k = 0, t) = 0$ , and, if  $N(g^0) > N(B_0)$ , then  $M = B_0 + \alpha \delta_0$  with  $\alpha = N(g^0) - N(B_0)$  (this is not proved because  $M_0 = 0$  and  $M = B_\mu$  with negative  $\mu$  is also a possible equilibrium state for the differential system).

These nonlinear ordinary differential systems can be very simplified in the two particular cases of uniform and concentrated cross sections.

### 3.1.1 Uniform cross section

When  $b(k, k') = \bar{b}(k)\bar{b}(k')$ , the evaluation of the double integral (3.2) reduces to two (simple) moments :

$$\frac{\partial g}{\partial t} = \left( \int_0^{k_0} \bar{b}(k') g' dk' \right) \bar{b}(k)(k^2 + g)e^{-k} - \left( \int_0^{k_0} \bar{b}(k')(k'^2 + g')e^{-k'} dk' \right) \bar{b}(k)g. \quad (3.5)$$

Note that this expression is compatible with the measure values solution of the form  $g = g_{reg} + \alpha \delta_0$  considered in [EM2] (which satisfied a system of equation for both  $g_{reg}(k, t)$  and  $\alpha(t)$ , described in [EM] ). One replaces the integral by a

discrete sum. Assume  $\bar{b}(k) = 1$  for example (as in [EM2]). Let us note  $M_0$  the discrete density and define the first moment as

$$M_1 = h \sum_{i=1}^n (k_i^2 + g_i) e^{-k_i}.$$

We consider the following explicit scheme

$$\frac{g_i^{n+1} - g_i^n}{\Delta t_n} = M_0 k_i^2 e^{-k_i} + (M_0 e^{-k_i} - M_1^n) g_i^n. \quad (3.6)$$

The conservation of density can be verified by summing the r.h.s. in the above equations. Positivity is preserved at any iteration, provided that

$$\Delta t_n \leq \frac{1}{M_1^n}.$$

Indeed, multiplying (3.6) by  $e^{k_i}$  and summing over  $i$  gives

$$\begin{aligned} \frac{M_1^{n+1} - M_1^n}{\Delta t_n} &= M_0 \left( \sum k_i^2 g_i^n e^{-2k_i} + e^{-2k_i} \right) h - M_1^n h \sum e^{k_i} g_i^n \\ &\leq M_0 M_1^n - M_1^n (M_1^n - \sum k_i^2 e^{k_i} h) \\ &\leq C M_1^n - (M_1^n)^2, \end{aligned}$$

with  $C = M_0 + \sum k_i^2 e^{k_i} h$ . Then,

$$M_1^{n+1} \leq M_1^n + C \Delta t_n M_1^n - \Delta t_n (M_1^n)^2,$$

and using  $\Delta t_n = \frac{1}{M_1^n}$ , we have by induction that  $M_1^n \leq C$  and the time steps are bounded from below.

## 3.2 Semidiscretization for the Kompaneets equation

For the discretization of the K equation we proceed in two ways. The first one is based on the scheme for the QBE and the fact that for concentrated cross section around  $k = k'$ , the K equation is the asymptotic limit of the QBE, see [EM]. The second one is adapted from the method proposed by Chang and Cooper, see [CC], for linear equation of Fokker-Planck type.

### 3.2.1 Concentrated cross section

One consider a cross section sequences on the form

$$b^\varepsilon(k, k') = \frac{1}{\varepsilon^3} \tilde{b} \left( \frac{k - k'}{\varepsilon} \right),$$



where  $\tilde{b}$  is a positive and even function. This gives the K equation (when  $\tilde{b}(k, k') = e^{k/2}e^{k'/2}$ ).

If we assume, for simplicity, that  $\tilde{b}$  is compactly supported with  $\text{supp}(\tilde{b}) = ]-3/2, 3/2[$  and choose  $\varepsilon = h$ . The cut-off/smoothing parameter  $\varepsilon$  is equal to the mesh size. Then, in the double sum (3.4), only the terms such that  $(i - j) = \pm 1$  do not vanish. We obtain the following scheme

$$\begin{aligned} \frac{\partial g_i}{\partial t} = & \frac{\tilde{b}(1)}{h^2} \sum_{i=0}^n g_{i+1}(k_i^2 + g_i)e^{h/2} - g_i(k_{i+1}^2 + g_{i+1})e^{-h/2} + \\ & + g_{i-1}(k_i^2 + g_i)e^{-h/2} - g_i(k_{i-1}^2 + g_{i-1})e^{h/2}. \end{aligned} \quad (3.7)$$

Note that this system (with  $\tilde{b}(1) = 1$ ) can be also written as

$$\begin{aligned} \frac{\partial g_i}{\partial t} = & \frac{1}{h^2}(g_{i+1}k_i^2e^{h/2} + g_{i-1}k_i^2e^{-h/2} - (k_{i+1}^2e^{-h/2} + k_{i-1}^2e^{h/2})g_i) + \\ & + \frac{1}{h^2}(e^{h/2} - e^{-h/2})(g_i g_{i+1} - g_i g_{i-1}). \end{aligned} \quad (3.8)$$

The first part corresponds to a tridiagonal linear system and the second part to the non linear Burgers term  $f^2$ . One can write this system in the form

$$\frac{\partial g_i}{\partial t} = \frac{1}{h} \left( F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} \right), \quad (3.9)$$

with  $F_{-1/2} = F_{N+1/2} = 0$  and for  $i = 0, \dots, n-1$

$$\begin{aligned} F_{i+\frac{1}{2}} &= \frac{1}{h} \left( \exp\left(\frac{h}{2}\right) k_i^2 g_{i+1} - \exp\left(-\frac{h}{2}\right) k_{i+1}^2 g_i \right) + \left( \exp\left(\frac{h}{2}\right) - \exp\left(-\frac{h}{2}\right) \right) g_i g_{i+1} \\ &= \frac{1}{h} \exp\left(-\frac{k_i + k_{i+1}}{2}\right) (g_i + k_i^2)(g_{i+1} + k_{i+1}^2) \left( \frac{h(g_{i+1}, k_{i+1})}{h(B_{i+1}, k_{i+1})} - \frac{h(g_i, k_i)}{h(B_i, k_i)} \right) \\ &= \frac{1}{h} h \left( B_\mu\left(\frac{k_i + k_{i+1}}{2}\right), \frac{k_i + k_{i+1}}{2} \right) \left( \frac{h(g_{i+1}, k_{i+1})}{h(B_{i+1}, k_{i+1})} - \frac{h(g_i, k_i)}{h(B_i, k_i)} \right), \end{aligned}$$

for any bose -Einstein state  $B_\mu$ , since  $h(B_\mu, k) = \exp(k + \mu)$ . This is the consistency of this scheme for the Kompaneets equation. Indeed, the flux of the K equation (2.10) is

$$\mathcal{F}(g, k) = (g + k^2)^2 \exp(-k) \frac{\partial}{\partial k} \left( \frac{h(g, k)}{h(P, k)} \right).$$

Using a Taylor expansion around  $k_{i+\frac{1}{2}} = \frac{k_i + k_{i+1}}{2}$  we obtain

$$F_{i+\frac{1}{2}} = \mathcal{F}(g_{i+\frac{1}{2}}, k_{i+\frac{1}{2}}) + O(h^2).$$

### 3.2.2 Chang and Cooper method

We now follow the method proposed by Chang and Cooper [CC] . We use from (2.9) which is close to the linear Fokker-Planck expression and discretize the fluxes at the interface  $F_{i+\frac{1}{2}}$  between  $k_i$  and  $k_{i+1}$  in a standard centered finite difference for the diffusion part

$$\left(\partial_k \frac{f}{1+f}\right)_{i+\frac{1}{2}} = \frac{1}{h} \left( \frac{f_{i+1}}{1+f_{i+1}} - \frac{f_i}{1+f_i} \right) = \frac{f_{i+1} - f_i}{h(1+f_{i+1})(1+f_i)},$$

a linear combination for the convective part :

$$\left(\frac{f}{1+f}\right)_{i+\frac{1}{2}} = (1 - \delta_{i+\frac{1}{2}}) \frac{f_{i+1}}{1+f_{i+1}} - \delta_{i+\frac{1}{2}} \frac{f_i}{1+f_i},$$

where  $0 \leq \delta_{i+\frac{1}{2}} \leq 1$  has to be define further, and, using an harmonic average for the "diffusion coefficients"

$$(k^4(1+f)^2)_{i+\frac{1}{2}} = k_i^2 k_{i+1}^2 (1+f_i)(1+f_{i+1}).$$

This last choice permits to simplify the denominator in the diffusion discretization which is now linear in  $f$ . Then, the flux can be written after some simple calculus as (with  $g_i = k_i^2 f_i$ )

$$F_{i+\frac{1}{2}} = \delta_{i+\frac{1}{2}} g_i (g_{i+1} + k_{i+1}^2) + (1 - \delta_{i+\frac{1}{2}}) g_{i+1} (g_i + k_i^2) + \frac{1}{h} (k_i^2 g_{i+1} - k_{i+1}^2 g_i). \quad (3.10)$$

The coefficients  $\delta_{i+\frac{1}{2}}$  are now to be determined. We shall choose them in order that the Bose-Einstein equilibrium states are preserved by this scheme. We impose that for  $g_i = M_i = \frac{k_i^2}{e^{k_i + \alpha} - 1}$  with any real value of  $\alpha$ , the flux vanishes. This gives

$$\delta_{i+\frac{1}{2}} = \frac{1}{h} - \frac{1}{e^h - 1}. \quad (3.11)$$

Note that this coefficients are independent of  $i$  for such a uniform grid and have the following Taylor expansion for small  $h$ :

$$\delta_{i+\frac{1}{2}} = \frac{1}{2} - \frac{1}{12}h + \frac{1}{720}h^3 - \frac{1}{30240}h^5 + O(h^6).$$

This means that when  $h \rightarrow 0$ , the scheme becomes a symmetrized approximation of the convective part.

Using this choice (3.11), the formulae (3.10) can be simplified into

$$F_{i+\frac{1}{2}} = g_{i+1} (g_i + k_i^2) + \frac{1}{e^h - 1} (k_i^2 g_{i+1} - k_{i+1}^2 g_i). \quad (3.12)$$

The same expression can be found for the flux  $F_{i-\frac{1}{2}}$

$$F_{i-\frac{1}{2}} = g_i(g_{i-1} + k_{i-1}^2) + \frac{1}{e^h - 1}(k_{i-1}^2 g_i - k_i^2 g_{i-1}).$$

Then, we obtain the same semidiscretization (using the notation  $F_{-1/2} = F_{n+1/2} = 0$ ) as in (3.9) up to a multiplicative factor that goes to 1 as  $h \rightarrow 0$

$$\frac{e^{h/2} - e^{-h/2}}{h} = 1 - \frac{1}{24}h^2 + \frac{7}{5760}h^4 + O(h^5).$$

**Remark 3.1** Note that, for a non uniform grid, the same method applies, replacing  $h$  by its local value  $\Delta k_{i+\frac{1}{2}} = (k_{i+1} - k_i)$ . However, the two schemes are no more proportional although the ratio converges to 1 (when the mesh is refined i.e.  $\max_i \Delta k_{i+\frac{1}{2}} \rightarrow 0$ ).

### 3.2.3 Properties of the two schemes

A discrete integration by parts gives the weak form for the two above schemes:

$$\sum_{i=0}^n \Psi_i \frac{\partial g_i}{\partial t} = -\frac{1}{2} \sum_{i=0}^{n-1} (\Psi_{i+1} - \Psi_i) F_{i+\frac{1}{2}}. \quad (3.13)$$

Thus, if we define the discrete density  $N(g)$  as  $\sum_{i=0}^n h g_i$  and the discrete entropy as  $H(g) = \sum_{i=0}^n h s(g_i, k_i)$  one verifies the conservation of density ( $\frac{d}{dt} N(g) = 0$ ) and the decay of entropy.

Indeed we have

$$\frac{d}{dt} H(g) = \sum_{i=0}^N h \partial_g s(g_i, k_i) \frac{dg_i}{dt} = \sum_{i=0}^N h \log \left( \frac{h(g_i, k_i)}{h(P_i, k_i)} \right) \frac{dg_i}{dt},$$

thus using (3.10)

$$\frac{d}{dt} H(g) = - \sum_{i=0}^N h C (k_i^2 + g_i) (k_{i+1}^2 + g_{i+1}) (\log(\lambda_{i+1}) - \log(\lambda_i)) (\lambda_{i+1} - \lambda_i), \quad (3.14)$$

with  $\lambda_{i+1} = \frac{h(g_{i+1}, k_{i+1})}{h(P_{i+1}, k_{i+1})}$ ,  $C = \exp(-\frac{k_i + k_{i+1}}{2})$  for the flux (3.10) or  $C = \frac{\exp(-(k_i + \mu))}{\exp(h) - 1}$  for the flux (3.12), and  $C$  is strictly positive. Using the classical inequality  $(x - y)(\log(x) - \log(y)) \geq 0$ , we have

$$\frac{d}{dt} H(g) \leq 0.$$

### 3.2.4 Existence of global positive solution.

We write system (3.8) by factorizing the gain term  $G(g)$  and the loss term  $L(g)$  as usual for Boltzmann equation

$$\frac{dg_i}{dt} = \frac{1}{h^2} (G(g)_i - L(g)_i g_i), \quad (3.15)$$

with  $G(g)_i \geq 0$  and  $L(g)_i \geq 0$ . We have the existence local in time using the Cauchy-Lipschitz theorem starting from  $g_i^0 > 0$  for all  $i > 0$ . Then, the loss term being bounded (for a given initial density  $N$ ), we have that the solution remains positive. Finally, using the conservation of density, we have an upper bound for the semidiscretized solution  $g_i \leq N/h$  and the solution is global in time. Note that this upper bound does not prevent concentration.

### 3.2.5 Discrete equilibrium state

We shall prove that the discrete equilibrium states are the generalized Bose-Einstein distribution functions. We restrict ourselves to the case  $f^0(0) < \infty$  that is  $g^0(0) = 0$ . It is easy to verify that if  $g_0(t=0) = 0$  then  $g_0(t) = 0$ . Thus, the distribution function is discretized on  $[h, k_0]$  i.e. for  $i = 1 \cdots n$ . From (3.14), it is easy to verify that  $\frac{d}{dt}H(g) = 0$  if and only if  $g = B_\mu$  with  $\mu$  eventually negative.

Let us consider, for any fixed density  $N$ , the discrete Bose-Einstein distribution

$$B_\alpha^h(k) = \frac{k^2}{e^{k+\alpha} - 1} \chi_{k \geq h},$$

where  $\alpha$  is such that

$$N^h \stackrel{\text{def}}{=} \sum_{i=1}^n B_\alpha^h(k_i) h = N.$$

Note that  $N_h$  is decreasing with  $\alpha$  and is one to one from  $] -h, \infty[$  to  $[0, \infty[$ . Thus, any arbitrary density  $N$  can be associated with a generalized discrete Bose-Einstein state provided that the mesh is refined enough such that  $N > N^0$  or  $Nh < k_0$ . When  $n \rightarrow \infty$  or  $h \rightarrow 0$ , one has for  $N \leq N^0$

$$N_h \rightarrow N(B_\mu),$$

and for  $N > N^0$ ,  $N_h \rightarrow B_0 + (N - N^0)\delta_0$ . Indeed, in the second case, one has  $-h \leq \alpha \leq 0$  and thus  $\alpha \rightarrow 0$  as  $h \rightarrow 0$ . One can check easily that if  $N_h(g^0) > N_h(B_0)$ , the equilibrium state associated to  $g^0$  is a Bose-Einstein state  $B_\alpha^h$  with  $\alpha \in [-h, 0[$ . Let us now precise the relation between such equilibrium state and  $B_0 + (N(g^0) - N(B_0))\delta_0$ . For simplicity, we consider the problem in continuous variable  $k$ , thus  $N_h$  is now

$$N_h \stackrel{\text{def}}{=} \int_{k \geq h} M_h(k) dk.$$

We shall prove the following result

**Proposition 3.2** *For any  $N > 0$ ,  $B_\alpha^h$  converge when  $h \rightarrow 0$  in the sense of distribution toward the continuous equilibrium functions that is, when  $N \leq N_0$ , there exists  $\beta > 0$  such that  $B_\alpha^h \rightarrow B_\beta$  i.e.  $\alpha \rightarrow \beta$  as  $h \rightarrow 0$  and when  $N > N^0$*

$$B_\alpha^h \rightarrow B_0 + (N - N^0)\delta_0$$

where  $B_0$  is the Planck distribution and  $N^0$  its density. Moreover, the associated entropy converges i.e.  $H(M_h) \rightarrow H(B_0)$  when  $N > N^0$  and  $H(B_\alpha^h) \rightarrow H(B_\beta)$  when  $N \leq N^0$ .

The proof is postponed in Appendix B and can be applied for a finite domain  $[0, k_0]$  or with discrete measure (sums instead of integrals).

### 3.2.6 Maximum principle for the discrete equation Kompaneets

Let us prove a stronger result that the existence of a positive solution when the initial data is in between two Bose-Einstein functions. Indeed, as for the linear Fokker-Planck operator, the Kompaneets equation satisfies a Maximum principle that is verified on its discretized version.

**Proposition 3.3** *Let  $g$  a solution of (3.8). If there exists  $\mu$  such that  $g(t = 0, k) \leq B_\mu(k)$  (resp.  $g(t = 0, k) \geq B_\mu(k)$ ), then, for all  $t > 0$ , one has  $g(t, k) \leq B_\mu(k)$  (resp.  $g(t, k) \geq B_\mu(k)$ ).*

*Proof.* First, one checks easily that the scheme (3.8) can be written in the form

$$\frac{dg_i}{dt} = C_{i+\frac{1}{2}} D \left( \frac{h(g, k)}{h(B, k)} \right)_{i+\frac{1}{2}} - C_{i-\frac{1}{2}} D \left( \frac{h(g, k)}{h(B, k)} \right)_{i-\frac{1}{2}} \quad (3.16)$$

where we note  $h(g, k) = \frac{g(k)}{k^2 + g(k)}$  and  $D$  is a finite difference operator defined as

$$D\phi_{i-\frac{1}{2}} = \phi_{i+1} - \phi_i,$$

and  $C_{i+\frac{1}{2}}$  are non negative coefficients (with the notation  $C_{-1/2} = C_{N+1/2} = 0$ ) that depend on the function  $g$  and of  $k$ .

Note that for any Bose-Einstein function  $B_\mu$ ,

$$h(B_0, k) = \frac{B_0}{k^2 + B_0} = \exp(-k) = h(B_\mu, k) \exp(-\mu)$$

Therefore, one can change  $B_0$  into any Bose-Einstein  $B_\mu$  in formula (3.16) with coefficient multiplied by  $\exp(-\mu)$ . Moreover,  $h(g, k)$  is a increasing function of  $g$  since

$$\frac{\partial h(g, k)}{\partial g} = \frac{k^2}{(k^2 + g^2)}.$$

We denote by

$$\lambda_i = \frac{h(g_i, k_i)}{h(B_\mu(k_i), k_i)},$$

then the  $(\lambda_i)$  satisfy a system of the same form i.e.

$$\frac{d\lambda_i}{dt} = E_{i+\frac{1}{2}} D(\lambda)_{i+\frac{1}{2}} - E_{i-\frac{1}{2}} D(\lambda)_{i-\frac{1}{2}},$$

with non negative  $E_{i+\frac{1}{2}}$ .

Define  $\lambda_S = \max_{i=1 \dots N} \lambda_i$ . The function  $\lambda_S$  is a piecewise  $C^1$  function of  $t$ . By definition of  $\lambda_S$ , we have,  $\forall t > 0$ , there exists a subset  $I_S(t)$  of  $\{1, \dots, N\}$ , such that  $\lambda_i(t) = \lambda_S(t)$  for all  $i \in I_S(t)$ . Thus,  $\forall i \in I_S(t), \forall j \in \{1, \dots, N\}$ ,  $\lambda_i \geq \lambda_j$  and using the positivity of the coefficients  $E$ , we obtain that  $\frac{d\lambda_S}{dt} \leq 0$ . The same idea proves that the minimum of  $\lambda_i$  increases with time.  $\square$

**Remark 3.4** Assume that for some  $\mu$ ,  $g(t=0) \leq B_\mu$ , then  $h(g(t=0), k) \leq h(B_\mu, k)$  or equivalently  $\lambda_S(0) \leq 1$  then for all  $t$ ,  $\lambda_S(t) \leq 1$  i.e.  $g(t, k) \leq B_\mu(k)$ . In the same way, if for some  $\mu$ ,  $g(t=0) \geq B_\mu$ , then  $g(t, k) \geq B_\mu(k)$ .

Thus, using the Kullback like inequality (2.2), which is also valid for discrete measures, we can prove the following lemma

**Lemma 3.1** When  $g(t=0) \leq B_0$ , the distribution function converges toward its equilibrium.

*Proof.* Using  $H(g|M) = H(g) - H(M)$ , where  $M$  is the discrete equilibrium associated to  $g$ , is decreasing in time and is positive. Thus, there exists a increasing and diverging sequence  $t_k$  such that  $H'(t_k) \rightarrow 0$ . The zero of the derivative of  $H$  are the equilibrium  $M$ . Hence,  $g(t_k)$  goes to  $M$  and  $H(g(t_k)|M) \rightarrow H(M|M) = 0$ . Since  $H(g|M)$  is decreasing, we have necessary

$$\lim_{t \rightarrow \infty} H(g|M) = 0,$$

and, by lemma 2.2, we obtain, the convergence of  $g$  toward its equilibrium  $M$ .  $\square$

**Remark 3.5** The results of this section are valid for the discrete QBE, the proofs being analogous.

## 4 Time discretization for the Kompaneets equation

We eliminate explicit time discretization since, when Bose condensation occurs, time step to ensure positivity would be in  $O(h^3)$  to compare with  $\Delta t = O(h^2)$  for classical parabolic problems. Therefore, we shall only consider implicit scheme. We shall see that a fully implicit scheme has good properties but seems hard to implement at a reasonable cost. We propose an alternative implicit scheme with a low cost but for which we cannot prove all the features of the fully implicit scheme. As illustrated by numerical examples, this scheme works well.

### 4.1 Fully implicit scheme

Let us consider an implicit scheme of the form ( $g$  represents  $g^n$  at iteration  $n$  and  $\bar{g}$  denotes  $g^{n+1}$ )

$$\bar{g} = g + tQ(\bar{g}). \quad (4.17)$$

Assume that the scheme is positive i.e.  $g^n > 0 \Rightarrow g^{n+1} > 0$ , we shall prove that it is automatically entropy decaying i.e.  $H^{n+1} = H(g^{n+1}) < H(g^n) = H^n$ .

Indeed,  $H(g) = \int s(g, k)$  defined by (2.13) and using a second order Taylor expansion, at point  $\bar{g}$  with integral reminder, we have

$$H(\bar{g}) - H(g) = \partial_g H(\bar{g})(tQ(\bar{g})) + t^2 \int_0^1 (1-z)Q(\bar{g})^T \partial_{gg}^2 H(\bar{g} + z(g - \bar{g}))Q(\bar{g})dz,$$

where  $\partial_g H$  denotes the functional derivative

$$\partial_g H = \int \partial_g(s(g, k))dk = \int \log\left(\frac{g \exp(k)}{k^2 + g}\right) dk,$$

and we have  $\partial_g H(\bar{g})(tQ(\bar{g})) \leq 0$ . Moreover, the second derivative with respect to  $g$  is negative

$$\partial_{gg}^2 s(g, k) = \frac{1}{g} - \frac{1}{g + k^2} \leq 0,$$

and this concludes the proof.

The existence of a positive solution for the implicit scheme is ensured by the Brouwer fixed point theorem. We choose  $C > 0$  such that  $CN(g)f + Q(f)$  is a positive operator for all positive  $f$  such that the density of  $f$  less or equal to  $N(g)$ . Then (4.17) can be rewritten as ( $f$  denotes  $g^{n+1}$  and  $g$  denotes  $g^n$ ):

$$f(1 + N(g)Ct) = g + N(g)Ct \left( f + \frac{Q(f)}{N(g)C} \right). \quad (4.18)$$

The mapping  $f \mapsto T(f)$

$$T(f) = \frac{1}{1 + N(g)Ct}g + \frac{N(g)Ct}{1 + N(g)Ct} \left( f + \frac{Q(f)}{N(g)C} \right), \quad (4.19)$$

is continuous from the convex compact set

$$E = \{f > 0 \text{ such that density of } f \text{ is less or equal to } N(g)\}$$

into itself thus the Brouwer fixed point theorem insure the existence of an element  $f^*$  of  $E$  such that  $f^* = T(f^*)$  and necessarily  $f^*$  has the same density and energy that  $g$ .

Despite its good properties, since the implicit scheme is non linear an iterative procedure is needed and have to be stopped before exact convergence.

## 4.2 Semi-implicit scheme

The method we suggest is to treat the linear part implicitly and the non linear part semi-implicitly but in such way that the properties of the semidiscretized system are preserved. As we have seen the differential system (3.8) can be written as

$$\frac{d}{dt}g_i = \frac{1}{h} \left( F(g)_{i+\frac{1}{2}} - F(g)_{i-\frac{1}{2}} \right) \quad (4.20)$$

with the fluxes  $F(g)_{i+\frac{1}{2}}$  defined by (3.10) and have the structure

$$F(g)_{i+\frac{1}{2}} = \bar{F}L_{i+\frac{1}{2}}(g) + \bar{F}B_{i+\frac{1}{2}}(g)$$

with  $\bar{F}L_{i+\frac{1}{2}}(g) = a_{i+\frac{1}{2}}g_{i+1} - b_{i+\frac{1}{2}}g_i$ ,  $\bar{F}B_{i+\frac{1}{2}}(g) = c_{i+\frac{1}{2}}g_i g_{i+1}$ , and  $a_{i+\frac{1}{2}}$ ,  $b_{i+\frac{1}{2}}$  and  $c_{i+\frac{1}{2}}$  are non negative,  $\bar{F}B_{i+\frac{1}{2}}(g)$  is the Burgers flux and  $\bar{F}L_{i+\frac{1}{2}}(g)$  is the flux for the linear Kompaneets equation.

The semi implicit scheme consists in treating all the fluxes  $\bar{F}L_{i+\frac{1}{2}}(g)$  implicitly and to implicit only the term  $g_{i+1}$  in the Burgers fluxes  $\bar{F}B_{i+\frac{1}{2}}(g)$ . That is, if we assume that  $g$  is known at time  $t$ , we compute  $\bar{g}$  at time  $t + \Delta t$  as:

$$\bar{g}_i = g_i + \frac{\Delta t}{h} \left( \bar{F}_{i+\frac{1}{2}}(\bar{g}) - \bar{F}_{i-\frac{1}{2}}(\bar{g}) + \bar{B}_{i+\frac{1}{2}}(g, \bar{g}) - \bar{B}_{i-\frac{1}{2}}(g, \bar{g}) \right) \quad (4.21)$$

and  $\bar{B}_{i+\frac{1}{2}}(g, \bar{g}) = c_{i+\frac{1}{2}}g_i \bar{g}_{i+1}$ . One can verify that the density is preserved i.e.  $\sum g_i = \sum \bar{g}_i$ .

This system can be written in the form

$$(Id - \Delta t M(g))\bar{g} = g \quad (4.22)$$



where  $M(g)$  is a tridiagonal matrix and one can check easily that  $M(g)$  is also a so-called  $L$  matrix, that depends on  $g$  i.e.  $M(g)_{ii} > 0$  and  $M(g)_{ij} \leq 0$  for all  $i \neq j$ ).

The main property of this semi-implicit scheme concerns its positivity:

**Lemma 4.1** *If  $g$  is positive, then  $\bar{g}$  defined by (4.21) is positive for all time steps  $\Delta t$*

*Proof.* Due to the special structure of  $M(g)$ , it is easy to check that one can always construct  $X > 0$  such that  $X \in \ker M(g)$  which is equivalent to find  $X > 0$  such that

$$\bar{F}_{i+\frac{1}{2}}(X) + B_{i+\frac{1}{2}}(g, X) = 0 \quad (4.23)$$

for all  $i$ : start from one index  $i_0$  by setting  $X_{i_0}$  arbitrary strictly positive and in virtue of the positivity of  $a_{i+\frac{1}{2}}$ ,  $b_{i+\frac{1}{2}}$ ,  $c_{i+\frac{1}{2}}$  relation (4.23) generates positive strictly  $X_i$ .

Then there exists  $X > 0$  such that  $(Id - \Delta t M(g))X = X$ , that is if we set  $D$  such that  $D_{i,j} = \delta_{i,j} X_i$ , thus  $(Id - \Delta t M)D$  is a diagonal dominant matrix that is  $(Id - \Delta t M)$  is a generalized diagonal dominant matrix, which is equivalent to the fact that  $M$  is an  $M$ -matrix or in other words it has a positive inverse positive inverse, [berman - plemons] . This means that the scheme is unconditionally positive (whatever the condensation occurs).  $\square$

Note that  $X$  is related to equilibrium state in the prrof. Concerning the equilibrium states we have also the following result

**Lemma 4.2** *The scheme (4.21) preserves the equilibrium state.*

*Proof.* If we write  $Q(g)$  the operator of the right hand side of (4.20), by construction  $Q(g) = 0$  if  $g$  is an equilibrium state. Since  $M(g)g = Q(g)$ , (4.22) reads

$$(Id - \Delta t M(g))(\bar{g} - g) = Q(g) = 0.$$

Since the matrix  $M(g)$  is a M-matrix , then we have  $g = \bar{g}$  as we claim it.  $\square$

One should choose time step of the form  $C_1 h$  where  $C_1$  is such that the entropy decays. We are not able to exhibit a condition on the time step to ensure the decay of the discrete entropy. But, as we will see on the numerical examples, using a time step corresponding to a the convective equation that is of the form  $\Delta t < C \Delta k$  leads to a satisfactory behaviour of  $H$  even for singular initial data.

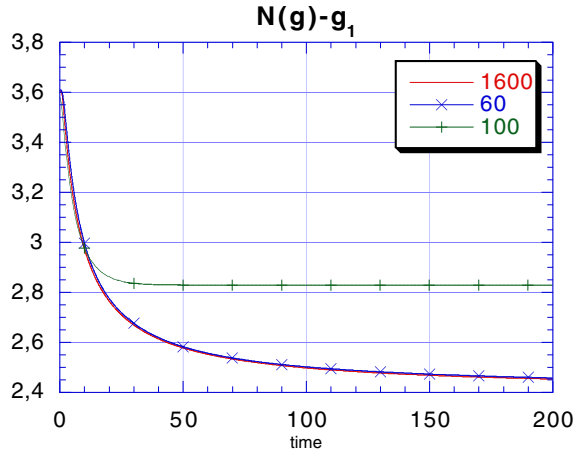


Figure 1: Density in case 1.

## 5 Numerical tests for the Kompaneets schemes

We will illustrate the scheme on the two following examples. The first one corresponds to a Planck distribution multiplied by  $3/2$ . In this case, concentration near origin occurs after few iterations since the initial density is greater than the critical one. On the second case, we consider a initial data with a lower density but we still observe a concentration when the initial density is close enough to 0, as expected after the analysis in [EHV] .

### 5.1 Relaxation of $\alpha B_0$ with $\alpha = 3/2$

We plot the evolution in time of three macroscopic quantities, density for non zero energy, energy and entropy. The runs corresponds to the following three cases

- label "1600" is a reference computation with 1600 points of discretizations and  $g_0(0) = 0$ .
- label "100" corresponds to 100 points using the same method and with  $g_0(0) = 0$ .
- label "60" denotes the same method but with only 60 points and with an initial data  $g_0(0) \neq 0$  but very small ( $10^{-12}$ ).

More precisely, in Fig1.1 we plot the quantity  $\sum_{i \geq 1} g_i h$  (the indices start at 1) i.e. the discrete version of  $\int_h^{k_0} g dk$ . This quantities is constant when the solution

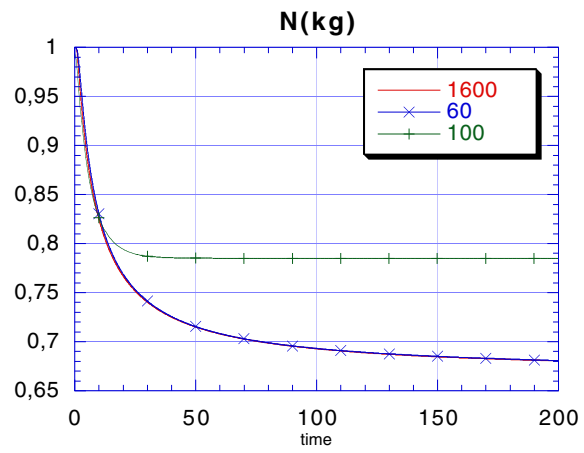


Figure 2: Energy in case 1.

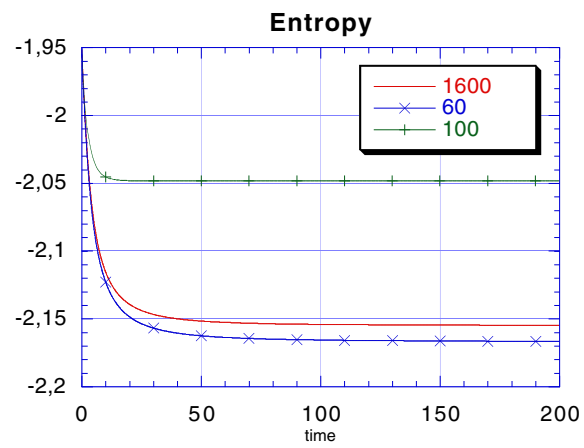


Figure 3: Entropy in case 1.

remains smooth. It decays when concentration occurs. Fig. 2 (resp. 3) illustrates the evolution of the energy (resp. entropy). The differences of the three runs are in the number of discretization points and in the initial data (at zero energy). Indeed, as explained before (section 3.2.5) there are two ways to discretize the singularity. Either,  $g_0(0)$  then,  $g_0(t) = 0$ ,  $\forall t > 0$  and the distribution function converges to a generalized Bose-Einstein equilibrium (with  $\mu < 0$ ) that converges in a distributional sense toward the Planck distribution ( $B_0$ ) plus a dirac measure at  $k = 0$  or  $g_0(0) \neq 0$  and the discretized distribution function converges for  $i \geq 0$  toward the Planck distribution, and the remaining density concentrates on  $k_0 = 0$ .

The presented results show that it is much better to discretize with  $g_0(0) \neq 0$ , in particular in terms of entropy (the entropy is lower with only 60 points and  $g_0(0) \neq 0$  than with 1600 points and  $g_0(0) = 0$ , i.e. generalized Bose-Einstein equilibrium).

## 5.2 Initial gaussian distribution.

In this case, we consider an initial gaussian distribution with a sub-critical density  $N < N^0$ . In the Compton case (or QBE), the solution goes toward its Bose Einstein equilibrium. It has been proved in [EHV] that, for the Kompaneets equation, the solution may not be global in time provided that the initial density is close enough to the origin. Indeed, there is a balance between the Burgers term and the diffusion term : the Burgers part is a convective toward the origin that leads to concentration, whereas the diffusion spreads off the distribution that becomes smoother.

In Fig 1.4, we plot the distribution for different time when the initial data is a Gaussian centred at  $k = 2$  with 3200 points of discretization. In this case, the diffusion dominates and the distribution function goes to its Bose-Einstein distribution (with  $\nu > 0$ ). In Fig 1.5, we plot the same quantities (distribution function versus  $k$  at different time) but with an gaussian initial data (same total density, centered at  $k = 1$ ). In this case, the Burgers term is stronger and the distribution function becomes singular near origin after a finite time  $T^*$ . Clearly, the solution is no more valid for  $t > T^*$  but the simulation can be continued since the Kompaneets discretization can be interpreted as the discretization of QBE with a small parameter  $\varepsilon$  linked to the discretization step  $h$  (see section section 3.2.1) for which such a concentration is possible.

On Fig 1.6, we plot the evolution in time of the density for non zero energy and we observe the concentration at time  $T^*$  : some part of the density goes into the zero energy part of the distribution function. On Fig 1.7 (resp. 1.8), we plot the energy (resp. entropy) versus time.

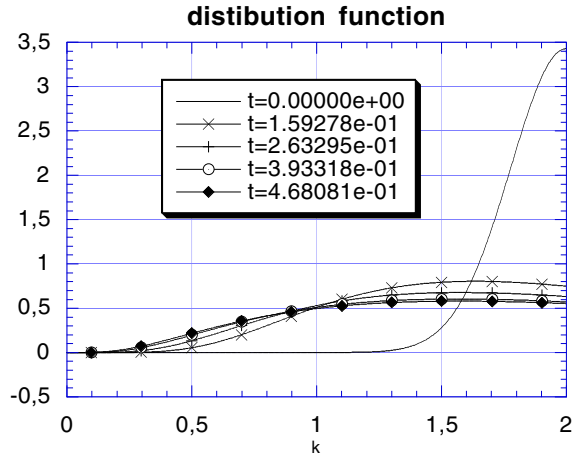


Figure 4: Evolution of the distribution function in case 2,  $k = 2$ .

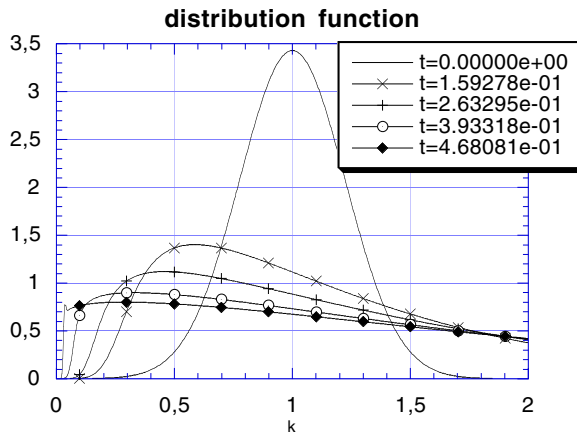


Figure 5: Evolution of the distribution function in case 2,  $k = 1$ .

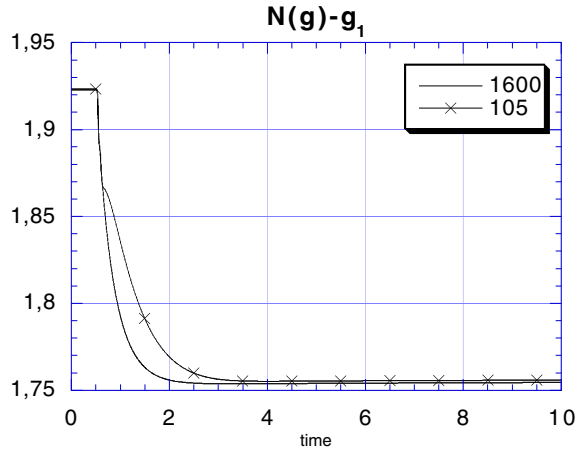


Figure 6: Density versus time in case 2 ,  $k = 1$ .

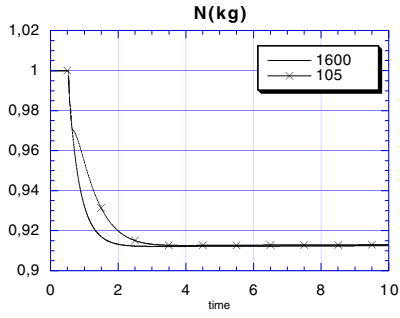


Figure 7: *energy*

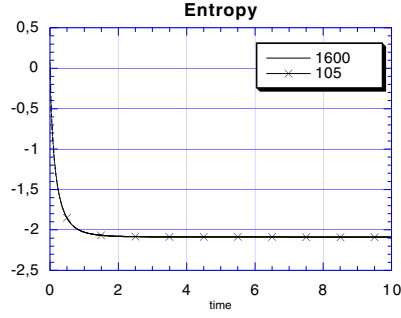


Figure 8: *entropy*

## 6 Conclusions

The main improvements compared with the previous method investigated in [P] are

- This method does not require to construct a equilibrium state that usually required an iterative method (which is not always conservative).
- The semi-discretized system preserve all the properties : decay of entropy, conservation of density
- The semi-implicit scheme is unconditionally positive
- Scheme 1 is of order 2 in  $k$

Moreover, we are able to deal with condensation near the origin when the initial data has a density larger than the Planck one or is too close to the origin, both of these phenomena are illustrated by numerical results. There are two ways to describe this concentration at the discrete level : either  $g_0(0) = 0$  and then, this remains equal 0 at all time - in this case, the equilibrium state are generalized Bose-Einstein state with  $\mu < 0$  that converge in a distribution sense to the Planck distribution plus a Dirac at  $k = 0$ , see section 3.2.5 and appendix B - or  $g_0(0) \neq 0$  - in this case, the smooth part converges toward the Planck distribution and the remaining density concentrates on  $k = 0$ . The preliminary numerical results indicate that the second case is better and more precisely, that the first case is unstable : when the initial data at  $k = 0$  is not equal zero, the scheme converges toward the second representation of the singularity.

Some questions remain to be attend in further works :

- Is the semi-implicit scheme entropic ?
- How to construct a full implicit scheme (without iterative method) ?
- Can we insure the same properties for general cross sections with the (QBE).

## Appendix A : proof of Lemma 2.2

*Proof.* We begin by the first part of the lemma. It is then easy to check that it suffice to prove the result only in the case  $\mu = 0$ , i.e. for  $B_0$ .

For  $k \in [0, \log^+(\alpha)]$  with  $\log^+(x) = \max(0, \log(x))$ , since the solution is non negative and

$$h(g, k) \leq 1 \leq \alpha \exp(-k) = \alpha h(B_0, k).$$

For  $k \in ]\log(\alpha), \infty[$ , consider any function  $G(x)$  such that  $G(x) \geq 0$  for  $x \geq 0$ ,  $0 < G'(x) < \infty$  for  $x \geq 0$  and  $G(x) = 0$  for  $x \leq 0$ . Then, define  $H_\alpha(x, k)$  as

$$H_\alpha(x, k) = \int_{x_0}^x G\left(\frac{h(s, k)}{h(B_0, k)} - \alpha\right) ds$$

for  $k \in ]\log^+(\alpha), \infty[$  with  $x_0$  the unique positive solution of  $h(x_0, k) = \alpha h(B_0, k)$  (note that  $x_0 = \frac{\alpha k^2 \exp(-k)}{1 - \alpha \exp(-k)} > 0$ ) and  $H_\alpha(x, k) = 0$  for  $k \leq \log^+(\alpha)$ .

Set now

$$E(t) = \int_{k \geq 0} H_\alpha(g(k), k) dk.$$

It can be checked that by construction  $E(t) \geq 0$ ,  $E(t)$  is continuous and by the hypothesis  $E(0) = 0$ . Moreover,  $E(t)$  is  $C^1$  and we have

$$\frac{d}{dt}E = \int_{k \geq 0} G \left( \frac{h(g, k)}{h(B_0, k)} - \alpha \right) \frac{dg}{dt} dk.$$

Now using the weakform (2.5) of the QBE, it is easy to see that, since  $G \left( \frac{h(g, k)}{h(B_0, k)} - \alpha \right)$  is monotone increasing in the variable  $g$  for each  $k$ . Thus, using the inequality  $(x - y)(\phi(x) - \phi(y)) \geq 0$ , for any monotone increasing function  $\phi$ ,  $\frac{d}{dt}E \leq 0$ .

For the K equation, using the weak form (2.12), it is also clear that  $\frac{d}{dt}E \leq 0$ . Thus, for both equation,  $E(t) = 0$  for all time  $t$ , which is equivalent to  $\frac{h(g, k)}{h(B_0, k)} \leq \alpha$  almost everywhere in  $k \in ]\log^+(\alpha), \infty[$ .

For the proof of the second part of the lemma we proceed as for the first part. With the function  $G$  defined above, we set  $H$  as

$$H(x, k) = \int_{x_1}^x G \left( \frac{1 - h(s, k)}{h(B_0, k)} \right) ds,$$

if  $g \leq x_1$ . and  $H(x, k) = 0$  for  $g \geq x_1$  and define

$$E(t) = \int_{k \geq 0} H(g, k) dk.$$

Proceeding as for the first part we obtain that  $E(t)$  is monotone decreasing, positive and  $E(0) = 0$ . Thus for any time  $E(t) = 0$  that is  $h(g, k) \geq h(B_0, k)$  a.e. as we have claimed it.  $\square$

## Appendix B : proof of Proposition 3.3

*Proof.* As explained above  $\alpha$  is a function of  $N$  and of  $h$ .

Define  $N(\alpha, h)$  as (for any  $h \geq 0$  and  $\alpha \geq -h$ )

$$N(\alpha, h) = \int_h^\infty \frac{k^2}{\exp(k + \alpha) - 1} dk.$$

The equation  $N(\alpha, h) = N$  determines implicitly  $\alpha$  as a function of  $N$  and  $h$ . If  $N > N^0$ , since  $N^0 = N(0, 0)$ ,  $\alpha(N, h) \in [-h, 0[$  thus  $\alpha \rightarrow 0$  as  $h \rightarrow 0$ . Let  $\phi$  a test function, we have

$$L(h) = \int_0^\infty \phi(k)(M_h(k) - B_0(k)) dk = \int_h^\infty \frac{k^2 \phi(k)}{\exp(k + \alpha(h)) - 1} dk - \int_0^\infty \frac{k^2 \phi(k)}{\exp(k) - 1} dk.$$



We write  $\phi(k) = \phi(0) + k\psi(k)$  where  $\psi$  is a  $C^\infty(0, \infty)$ . We know that  $\alpha(h) \rightarrow 0$ . Thus, we have

$$\begin{aligned} L(h) &= \int_h^\infty k^2 \phi(0) \left( \frac{1}{\exp(k + \alpha(h)) - 1} - \frac{1}{\exp(k) - 1} \right) dk + \\ &\quad + \int_h^\infty k^3 \psi(k) \left( \frac{1}{\exp(k + \alpha(h)) - 1} - \frac{1}{\exp(k) - 1} \right) dk - \int_0^h \frac{k^2 \phi(k)}{\exp(k) - 1} dk. \end{aligned}$$

The first term gives

$$\int_h^\infty k^2 \left( \frac{1}{\exp(k + \alpha(h)) - 1} - \frac{1}{\exp(k) - 1} \right) dk = N - \int_h^\infty \frac{k^2}{\exp(k) - 1} dk \rightarrow N - N^0.$$

- the second can be written as

$$(1 - \exp(\alpha(h))) \int_h^\infty \frac{k^3 \psi(k) dk}{(\exp(k + \alpha(h)) - 1)(\exp(k) - 1)} dk \rightarrow 0,$$

and the third term goes to 0. Thus, we proved that

$$L(h) \rightarrow (N - N^0)\phi(0).$$

This is equivalent to the convergence of  $B_\alpha^h$  toward  $B_0 + (N - N_0)\delta_0$  in the distribution sense.

Let us now consider the case  $N < N^0$ . There exists  $\beta > 0$  such that  $N(\beta, 0) = N$ . We want to prove that  $\alpha(h)$  ( $N$  being fixed) goes to  $\beta$  as  $h \rightarrow 0$ . The function  $\alpha$  being decreasing with  $h$  and being positive for  $h$  small enough, one has  $0 < \alpha(h) < \beta$ . Moreover,  $\alpha$  is continuous, then  $\lim_{h \rightarrow 0} \alpha(h) = \alpha(0) = \beta$ . Then by dominated convergence theorem, we have that  $B_\alpha^h$  tends to  $B_\beta$  in  $L^1(0, \infty)$  (and in the distributional sense).

Let us now prove the convergence of the entropy. One has

$$H(g) = \int_0^\infty \left( g \log\left(\frac{g}{k^2 + g}\right) + k^2 \log\left(\frac{k^2}{k^2 + g}\right) + kg \right) dk.$$

For the function  $B_\alpha^h$ , one easily checks that

$$\frac{B_\alpha^h}{k^2 + B_\alpha^h} = e^{-(k+\alpha)}, \quad \frac{k^2}{k^2 + B_\alpha^h} = 1 - e^{-(k+\alpha)}.$$

Then, using the expression of  $H$  and  $B_\alpha^h$ , one obtains

$$H(B_\alpha^h) = -\alpha \int_0^\infty B_\alpha^h + \int_h^\infty k^2 \log(1 - \exp(-(k + \alpha))) dk.$$

The first term is equal to  $-\alpha N$  that goes to 0 because  $\alpha \rightarrow 0$ . Let us note  $g_h(k) = \log(1 - \exp(-(k + \alpha)))$ . We know that  $g_h$  converges to  $g_0$  almost everywhere and is integrable on  $[0, \infty[$ . Note that  $\forall k > \varepsilon$  since  $\alpha \leq 0$ ,  $(g_0 - g_h)(k) \geq 0$ .

$$(g_0 - g_h)(k) = \log\left(\frac{1 - \exp(-k)}{1 - \exp(-(k + \alpha))}\right) \leq \exp(\alpha)(\exp(-\alpha) - 1) \int_h^\infty \frac{k^2 dk}{\exp(-(k + \alpha)) - 1}$$

using  $\log(1 + h) \leq h$ . This proves that

$$\int_h^\infty (g_0 - g_h)(k) dk \rightarrow 0.$$

But,

$$H(B_\alpha^h) - H(B_0) = -\alpha N + \int_h^\infty (g_h - g_0)(k) dk - \int_0^h g_0(k) dk.$$

The last term tends to 0 since  $g_0$  is integrable. This proves the convergence. In the second case  $N < N^0$  we know that  $\alpha(h)$  is a decreasing function of  $h$  that converges toward  $\beta > 0$  such that

$$\int_0^\infty \frac{k^2 dk}{\exp(k + \beta) - 1} = N$$

The proof given for  $N > N^0$  can be applied in this case. □

## Bibliography

- [BP] Berman, A., Plemmons R.J., Nonnegative matrices in the mathematical sciences. Classics in Applied Mathematics, 1994
- [B] S. N. Bose, Plancks Gesetz und Lichtquantenhypothese, *Z. Phys.* **26** (1924), 178–181.
- [BCF] Buet, Christophe, Cordier, Stéphane and Filbet, Francis, Comparison of numerical schemes for Fokker-Planck-Landau equation. *ESAIM Proc.*, vol 10, p. 161-181, (2001).
- [BC] Buet, C. and Cordier, S., Numerical analysis of conservative and entropy schemes for the Fokker-Planck-Landau equation., *SIAM J. Numer. Anal.*, vol 36, No 3, p. 953-973, (1998).
- [BCDL] Buet, C.; Cordier, S.; Degond, P.; Lemou, M. Fast algorithms for numerical, conservative, and entropy approximations of the Fokker-Planck-Landau equation. *J. Comput. Phys.* 133, No.2, p. 310-322, (1997).

- [CL] R.E. Caflisch, C.D. Levermore, Equilibrium for radiation in a homogeneous plasma, *Phys. Fluids* **29**, 748-752 (1986)
- [CC] Chang J.S., Cooper G., A practical difference scheme for Fokker-Planck equations. *J. Comput. Phys.* 6, 1-16 (1970).
- [C] G. Cooper, Compton Fokker-Planck equation for hot plasmas, *Phys. Rev. D* **3**, 2312-2316 (1974)
- [DV] L. Desvillettes, C. Villani, On the Spatially Homogeneous Landau Equation for Hard Potentials. Part I: Existence, Uniqueness and Smoothness, *CPDE* vol. 25, n. 1-2, (2000), pp. 179-259 and Part II: H-Theorem and Applications , *CPDE*, vol. 25, n. 1-2, (2000), pp. 261-298.
- [EHV] M. Escobedo, M.A. Herrero, J.J.L. Velazquez, A nonlinear Fokker-Planck equation modeling the approach to thermal equilibrium in a homogeneous plasma, *Trans. Amer. Math. Soc.* **350**, 3837-3901 (1998)
- [EM1] M. Escobedo, S. Mischler, Equation de Boltzmann quantique homogène: existence et comportement asymptotique, *Note au C. R. Acad. Sci. Paris* **329** Série I, 593-598 (1999)
- [EM2] M. Escobedo, S. Mischler: On a quantum Boltzmann equation for a gas of photons, *J. Math. Pures Appl.* **80**, 5 (2001), pp. 471-515.
- [EM3] M. Escobedo, S. Mischler, M.A. Valle, On Boltzmann equation for a gas of quantum (and relativistic) particles, *work in preparation*
- [EMV] M. Escobedo, S. Mischler, J.J.L. Velazquez, Specified long time behavior for the Boltzmann-Compton equation, *work in preparation*
- [Ka] O. Kavian, Remarks on the Kompaneets Equation, a Simplified Model of the Fokker Planck equation, to appear in *Séminaire J.L. Lions - Collège de France - série rouge* (Pitman-Longman-Wesley)
- [K] A.S. Kompaneets, The establishment of thermal equilibrium between quanta and electrons, *Soviet Physics JETP*, (1957)
- [Ku] Kullback, S. , On the convergence of discrimination information. *IEEE Trans. Information Theory* IT-14 1968 765-766
- [LLPS] Larsen E.W., Levermore C.D., Pomraning G.C., Sanderson J.G. Discretization methods for one-dimensional Fokker-Planck operators. *J. Comput. Phys.* 61, 359-390 (1985)



# Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics

Christophe Buet<sup>a,\*</sup>, Bruno Despres<sup>a,b</sup>

<sup>a</sup>Commissariat à l'énergie Atomique, DSSI, BP 12, Bruyères le Chatel 91680, France

<sup>b</sup>Laboratoire JLL, 175 rue du Chevaleret, 75013 Paris, France

Received 26 February 2003; accepted 23 May 2003

## Abstract

This work addresses some asymptotic regimes for the coupling of radiation and hydrodynamics, and is inspired by the still non-answered need of high resolution and robust schemes for the numerical solutions of these problems. Using a simple characterization of the isotropy of the scattering in the comobile reference frame, we derive various asymptotic regimes. Among them is the non-equilibrium regime. Then we prove that the method of moments is compatible with the non-equilibrium regime. We also study the Rankine–Hugoniot relations.

© 2003 Elsevier Ltd. All rights reserved.

**Keywords:** Radiative hydrodynamics; Non-equilibrium diffusion regime; Method of moments

## 1. Introduction

This work addresses some asymptotic regimes for the coupling of radiation and hydrodynamics, and is inspired by the still non-answered need of high resolution and robust schemes for the numerical solutions of these problems. Following Lowrie et al. [1] and Lowrie and Morel [2], we consider that numerical progress should be possible using *Eulerian* conservative high-order Godunov-type scheme.

But it raises many difficulties: many models are written in the comobile or Lagrangian reference frame which moves with the fluid, see [3–9] and references therein. For mathematical aspects of the radiative transfer equation and related issues see for instance [10]. Actually, one can distinguish at least three approaches: a purely Lagrangian approach where everything is calculated in the moving reference frame; a comobile approach where some quantities are calculated in Eulerian reference frame and others are calculated in the moving reference frame (see for example [8]); the last approach

\* Corresponding author.

E-mail addresses: [christophe.buet@cea.fr](mailto:christophe.buet@cea.fr) (C. Buet), [despres@ann.jussieu.fr](mailto:despres@ann.jussieu.fr), [bruno.despres@cea.fr](mailto:bruno.despres@cea.fr) (B. Despres).

is purely Eulerian, everything is calculated in the Eulerian reference frame. This last approach, purely Eulerian, is the one we address here. The main idea of this work is that these Eulerian models should at least contain non-equilibrium models, that is models where the temperature of the fluid is different from the temperature of the radiation  $T \neq T_r$ . To our knowledge, a direct derivation of non-equilibrium model in the Eulerian frame was still missing. This is one of the contribution of this work. The method used is a simplified asymptotic analysis expansion, very similar to a *Chapman–Enskog* or *Hilbert* expansion, with an appropriate scaling of the flow where the scattering source term is of course relativistic as prescribed in [3] and others. For the sake of simplicity of mathematical developments the analysis is made in the context of the gray hypothesis.

An important and original feature of our analysis is the compatibility of all the expansions with the mathematical structure of the scattering. A consequence is that we really need to do the analysis using  $v$  and  $\vec{n}$ , and thus we cannot pre-integrate along  $v$  as it can be done with absorption only.

The main trick is a formula which allows to simplify the algebraic complexity of the expansion: this formula is a simple characterization in the lab frame of the isotropy of the scattering in the comobile frame. The consequence is the appearance of the famous non-conservative  $p_r \nabla \cdot \vec{v}$  term in the radiation equation. Note that this point was already emphasized in [3] for example, but in a Lagrangian frame and with a lot physical intuition, and discussion of the various scales of the problem (a modern presentation is [11]). The proof we give is more mathematically based and uses invariance relations true for this class of Lorentz models as a main tool to simplify the analysis: in particular the invariance of the measure  $v dv d\vec{n} = v_0 dv_0 d\vec{n}_0$ .

In a second part we study another class of models, the so-called moment models. We consider the most standard and simplest one with two unknowns which are  $E_r$  the energy of the radiation and  $\vec{F}_r$  the radiation flux. Then, we prove that this moment model contains the non-equilibrium limit with the  $p_r \nabla \cdot \vec{v}$  for *smooth solutions*, using once more a Chapman–Enskog expansion. To our surprise it appears this is not true for *discontinuous solutions*, such as shocks and contact discontinuities: discontinuous solutions of the moment model do not tend to discontinuous solutions of the non-equilibrium model. To fix this question we propose a modification of the moment model, which is based on the choice of new unknowns which are  $S_r$  the entropy of the radiation and  $\vec{F}_r$  the radiation flux. We prove that this modified moment model contains the non-equilibrium limit for both smooth solutions *and* discontinuous solutions. Since numerical methods based on Godunov methods and the Riemann problem need these discontinuous solutions, it is an indication that the modified moment model with  $(S_r, \vec{F}_r)$  is better suited for the development of conservative and Eulerian numerical schemes than the classical moment model with  $(E_r, \vec{F}_r)$ . Numerical results will be presented in a future work.

## 2. Models

### 2.1. Relativistic gas dynamics

The Euler system of inviscid gas dynamics with full Lorentz invariance is [3,6,12]

$$\frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) = 0,$$

$$\begin{aligned} \frac{\partial}{\partial t} \left( \gamma \left( 1 + \frac{h}{c^2} \right) \rho \vec{v} \right) + \nabla \cdot \left( \gamma \left( 1 + \frac{h}{c^2} \right) \rho \vec{v} \otimes \vec{v} + p \mathbf{I} \right) &= 0, \\ \frac{\partial}{\partial t} \left( \gamma \rho \left( 1 + \frac{h}{c^2} \right) - \frac{p}{c^2} \right) + \nabla \cdot \left( \gamma \rho \left( 1 + \frac{h}{c^2} \right) \vec{v} \right) &= 0. \end{aligned} \quad (1)$$

For this kind of Lorentz invariant models, one must recall that there is a distinction between the Eulerian reference frame also referred to as the lab frame, and the comobile reference frame which moves with the fluid also referred to as the Lagrangian frame. The density  $\rho = 1/\tau$  in the lab frame is different from the density calculated in the comobile frame  $\rho_0 = 1/\tau_0$ . In what follows the subscript 0 will designate any quantity measured in the comobile frame. One has  $\rho = \gamma \rho_0$  where  $\gamma$  is defined by

$$\gamma = \frac{1}{\sqrt{1 - |\vec{v}|^2/c^2}} \quad (2)$$

where  $c$  is the velocity of light. In (1),  $h$  is the enthalpy of the fluid calculated in the comobile frame  $h = e + p\tau_0$ , where  $p$  is the pressure. If one assumes for simplicity a perfect gas pressure law

$$p = \Gamma \frac{e}{\tau_0}, \quad \Gamma > 0, \quad (3)$$

then the enthalpy is simply  $h = (\Gamma + 1)e$ . Here,  $e$  is the internal energy of the fluid calculated in the comobile frame. Multiplying the last equation of (1) with  $c^2$  and subtracting the first one multiplied by  $c^2$  for the sake of convenience, we rewrite it as

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} \left( \gamma \left( 1 + \frac{h}{c^2} \right) \rho \vec{v} \right) + \nabla \cdot \left( \gamma \left( 1 + \frac{h}{c^2} \right) \rho \vec{v} \otimes \vec{v} + p \mathbf{I} \right) &= 0, \\ \frac{\partial}{\partial t} \left( c^2 \gamma \rho \left( 1 + \frac{e}{c^2} + \frac{|\vec{v}|^2}{c^2} \frac{p\tau_0}{c^2} \right) - c^2 \rho \right) \\ + \nabla \cdot \left( c^2 \left( \gamma \rho \left( 1 + \frac{e}{c^2} + \frac{|\vec{v}|^2}{c^2} \frac{p\tau_0}{c^2} \right) - c^2 \rho \right) \vec{v} + p \vec{v} \right) &= 0. \end{aligned} \quad (4)$$

For the sake of simplicity of notations, we define

$$\vec{v}_2 = \gamma \left( 1 + \frac{h}{c^2} \right) \vec{v} \quad \text{and} \quad E_2 = c^2 \gamma \left( 1 + \frac{e}{c^2} + \frac{u^2}{c^2} \frac{p\tau_0}{c^2} \right) - c^2.$$

With these notations (4) is equivalent to

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}_2) + \nabla \cdot (\rho \vec{v}_2 \otimes \vec{v} + p \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t}(\rho E_2) + \nabla \cdot (\rho E_2 \vec{v} + p \vec{v}) &= 0. \end{aligned} \quad (5)$$

## 2.2. Galilean gas dynamics

The classic Euler system of inviscid gas dynamics with Galilean invariance is recovered as the limit of (6) when  $|\vec{v}|/c \rightarrow 0$ . We consider the regime

$$\varepsilon = \frac{|\vec{v}|}{c}, \quad \frac{e}{c^2} = O(\varepsilon^2). \quad (6)$$

Indeed one has

$$\vec{v}_2 = \vec{v} + O(\varepsilon^2) \quad \text{and} \quad E_2 = e + \frac{1}{2} |\vec{v}|^2 + O(\varepsilon^2). \quad (7)$$

Let us define the classical total energy  $E = e + \frac{1}{2} |\vec{v}|^2$ . Then the classic Euler system of inviscid gas dynamics

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t}(\rho E) + \nabla \cdot (\rho E \vec{v} + p \vec{v}) &= 0, \end{aligned} \quad (8)$$

is recovered as the  $O(\varepsilon^2)$  approximation of (1). Thus, it is possible to base the coupling of radiation and hydrodynamics on (8) even if relaxation source terms must be Lorentz invariant for the final model to be correct. In order to simplify the presentation and since we are interested mainly in flows moving at moderate velocities, we use (8) instead of (1) in the rest of this paper.

## 2.3. Transfer equation for photons

The transfer equation for photons is

$$\frac{1}{c} \frac{\partial}{\partial t} I + \vec{n} \cdot \nabla I = S_t(v, \vec{n}), \quad (9)$$

where  $I(t, x : v, n)$  is the intensity of the radiation,  $v$  the frequency and  $\vec{n}$  the direction of the photons. The source term is  $S_t(v, \vec{n})$ . It is well known that the source term has to be Lorentz invariant for the total coupled system to be accurate [3] (but the derivation of the non-equilibrium diffusion limit that we give in this work is another proof that  $S_t(v, \vec{n})$  must be Lorentz invariant even for small velocities). In this work we follow [1] and consider a simplified source term where  $S_t = S_a + S_s$  is the sum of two contributions. The first one takes into account the absorption/re-emission of photons by the matter

$$S_a(v, \vec{n}) = \frac{v_0}{v} \sigma_a(v_0) \left[ \left( \frac{v}{v_0} \right)^3 B(v_0, T) - I \right]. \quad (10)$$

Here  $B(v_0, T)$  is the Planckian

$$B(v_0, T) = \frac{2h v_0^3}{c^2} (e^{h v_0 / k T} - 1)^{-1} \quad (11)$$

and  $\sigma_a(v_0) \geq 0$  is the absorption coefficient. In definitions (10)–(17), one has to use the frequency and direction of the photon calculated in the comobile frame

$$v_0 = \gamma v \left( 1 - \frac{\vec{n} \cdot \vec{v}}{c} \right) \quad \text{and} \quad \vec{n}_0 = \left( \frac{v}{v_0} \right) \left[ \vec{n} - \frac{\gamma}{c} \vec{v} \left( 1 - \frac{\vec{n} \cdot \vec{v}}{c} \left( \frac{\gamma}{\gamma + 1} \right) \right) \right]. \quad (12)$$

Another important invariance relation [3,6,8] between the intensity of the radiation in the lab frame and the intensity of the radiation in the comobile frame is

$$\frac{I}{v^3} = \frac{I_0}{v_0^3}. \quad (13)$$

Defining also the Planckian measured in the comobile frame as  $B(v_0, T)/v^3 = B_0(v_0, T)/v_0^3$ , one gets another expression of the absorption/re-emission contribution (10)

$$S_a(v, \vec{n}) = \frac{v^2}{v_0^2} \sigma_a(v_0) [B_0(v_0, T) - I_0(v_0, \omega_0)]. \quad (14)$$

The second term takes into account the scattering of photons by the matter [3]

$$S_s(v, \vec{n}) = \frac{v^2}{v_0^2} (S_s)_0(v_0, \vec{n}_0), \quad (15)$$

where the scattering measured in the comobile frame is

$$(S_s)_0(v_0, \vec{n}_0) = \sigma_s(v_0) \left[ \frac{1}{4\pi} \int I(v_0, \vec{n}'_0) d\vec{n}'_0 - I_0(v_0, \vec{n}_0) \right]. \quad (16)$$

Here,  $\sigma_s(v_0) \geq 0$  is the scattering coefficient. Since the scattering (16) is clearly isotropic in the comobile frame one gets

$$\int \int (S_s)_0(v_0, \vec{n}_0) dv_0 d\vec{n}_0 = 0.$$

Another possibility for the scattering is [1]

$$S_s(v, \vec{n}) = \frac{v_0}{v} \sigma_s(v_0) \left[ \left( \frac{v}{v_0} \right)^3 \frac{1}{4\pi} \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}' - I \right]. \quad (17)$$

With [1] we use the following definition for  $v'$  which enters in the scattering contribution:

$$v' = v \frac{1 - \vec{n} \cdot \vec{v}/c}{1 - \vec{n}' \cdot \vec{v}/c}. \quad (18)$$

In Appendix A we prove that (17) is equal to (15). An important consequence of (15) is

**Lemma 1** (Characterization of the isotropy of the scattering). *The isotropy of the scattering in the comobile frame is characterized by*

$$\int \int \frac{v_0}{v} S_s(v, \vec{n}) dv d\vec{n} = 0. \quad (19)$$



This is due to

$$\begin{aligned} 0 &= \int \int (S_s)_0(v_0, \vec{n}_0) dv_0 d\vec{n}_0, \\ &= \int \int \frac{v_0^2}{v^2} (S_s)_0(v_0, \vec{n}_0) dv_0 d\vec{n}_0 = \int \int \frac{v_0}{v} (S_s)(v_0, \vec{n}_0) dv d\vec{n}, \end{aligned}$$

where we use the invariance of the integration measure  $v dv d\vec{n} = v_0 dv_0 d\vec{n}_0$ , see [3]. Eq. (19) means that even if the scattering is isotropic in the comobile frame, then the scattering is non-isotropic in the lab frame.

#### 2.4. The full system for the coupling of radiation and hydrodynamics

In order to couple radiation and hydrodynamics we need the influence of radiation on the matter. So we define

$$S_E = \int \int S_t dv d\vec{n} \quad \text{and} \quad \vec{S}_F = \frac{1}{c} \int \int \vec{n} S_t dv d\vec{n}. \quad (20)$$

Here,  $S_E(\vec{S}_F)$  characterizes the energy (resp. impulse) exchange between the radiation and the matter. Following [3] we modify (8) and get the system that couples gas dynamics and the radiation

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I}) &= -\vec{S}_F, \\ \frac{\partial}{\partial t}(\rho E) + \nabla \cdot (\rho E \vec{v} + p \vec{v}) &= -S_E, \\ \frac{1}{c} \frac{\partial}{\partial t} I(v, \vec{n}) + \vec{n} \cdot \nabla I(v, \vec{n}) &= S_t(v, \vec{n}) \quad \forall v, \vec{n}. \end{aligned} \quad (21)$$

Note that a consequence of (19) is that the integrated scattering on the right-hand side of the system is  $\int \int S_s dv d\vec{n}$  and is generally non-zero. Thus, this term contributes to the right-hand side for the impulse and the energy equations. It is only the integrated scattered radiation in the comobile frame which is zero  $\int \int (S_s)_0 dv_0 d\vec{n}_0 = 0$ . Concerning the integrated scattered radiation in the lab frame we only have (19): this formula will play an important role in the asymptotic analysis of the system (21) in order to derive the non-equilibrium diffusion model.

#### 2.5. Thermodynamic compatibility of the model

The physically correct radiative entropy (see [13]) is

$$S_r = -\frac{2k}{c^3} \int \int v^2 [n \log n - (n+1) \log(n+1)] dv d\vec{n}, \quad (22)$$

where by definition

$$n = \frac{c^2}{2h} \frac{I}{v^3} = \frac{c^2}{2h} \frac{I_0}{v_0^3} \quad (23)$$

is a relativistic invariant. First and second variations of  $S_r$  with respect to  $I$  are given in

$$dS_r = -\frac{k}{ch} \int \int \frac{1}{v} \log\left(\frac{n}{n+1}\right) dI dv d\vec{n} \quad (24)$$

and

$$d^2S_r = -\frac{c^2k}{2h^2} \int \int \frac{1}{v^4} \frac{1}{n(n+1)} dI dI' dv d\vec{n}. \quad (25)$$

From (25) it is clear that the entropy is strictly concave with respect to  $I$ . On the other hand, it is clear that  $S_r$  given in (22) is not Lorentz invariant (only  $n$  and  $v dv d\vec{n}$  are invariant). Let us define the radiative entropy flux  $\vec{Q}_r$ :

$$\vec{Q}_r = -\frac{2k}{c^2} \int \int v^2 [n \log n - (n+1) \log(n+1)] \vec{n} dv d\vec{n}. \quad (26)$$

Consider a smooth solution of (21) where the right-hand side is  $S_t = S_a + S_s$  given in (10)–(17). Then one has

$$\partial_t S_r + \nabla \vec{Q}_r = -\frac{k}{h} \int \int \frac{1}{v} \log\left(\frac{n}{n+1}\right) (S_a + S_s) dv d\vec{n}. \quad (27)$$

Be careful that the notations for the source terms,  $S_t$ ,  $S_a$  and  $S_s$ , is closed to the notations for the entropies,  $S_r$  and  $S$ . The entropy of the fluid is  $S$  with the fundamental law of thermodynamic

$$T dS = d\left(E - \frac{1}{2} \vec{v} \cdot \vec{v}\right) + p d\frac{1}{\rho} = dE - \vec{v} \cdot d\vec{v} + p d\frac{1}{\rho}. \quad (28)$$

A standard consequence of (21) and (28) is

$$\partial_t(\rho S) + \nabla(\rho \vec{v} S) = \frac{1}{T}(-S_E + \vec{v} \cdot \vec{S}_F), \quad (29)$$

that is

$$\begin{aligned} \partial_t(\rho S) + \nabla(\rho \vec{v} S) &= -\frac{1}{cT} \int \int (S_a + S_s) dv d\vec{n} + \frac{\vec{v}}{cT} \int \int (S_a + S_s) \vec{n} dv d\vec{n} \\ &= -\frac{1}{T} \int \int \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) (S_a + S_s) dv d\vec{n}. \end{aligned} \quad (30)$$

So the total entropy production is

$$\partial_t(\rho S + S_r) + \nabla \cdot (\rho \vec{v} S + \vec{Q}_r) = \vec{Q}_a + \vec{Q}_s, \quad (31)$$

where

$$\vec{Q}_a = -\frac{k}{h} \int \int \frac{1}{v} \log\left(\frac{n}{n+1}\right) S_a dv d\vec{n} - \frac{1}{T} \int \int \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) S_a dv d\vec{n} \quad (32)$$

and

$$\vec{Q}_s = -\frac{k}{h} \int \int \frac{1}{v} \log\left(\frac{n}{n+1}\right) S_s dv d\vec{n} - \frac{1}{T} \int \int \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) S_s dv d\vec{n}. \quad (33)$$

**Lemma 2** (Thermodynamic compatibility of the scattering). *The fluid entropy production due to the scattering is always zero  $\int \int (1 - \vec{v} \cdot \vec{n}/c) S_s \, dv \, d\vec{n} = 0$ . The scattering entropy production  $\vec{Q}_s$  is always non-negative  $\vec{Q}_s \geq 0$ :  $\vec{Q}_s = 0$  if and only if the radiation is isotropic in the comobile frame.*

One has  $\vec{Q}_s = \vec{Q}_s^R + \vec{Q}_s^F$  where  $\vec{Q}_s^R$  is the contribution of the transfer equation to the entropy production and  $\vec{Q}_s^F$  is the fluid entropy production. First

$$\begin{aligned}\vec{Q}_s^F &= -\frac{1}{T} \int \int \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) (S_s)_0 \, dv \, d\vec{n} = -\frac{1}{T} \int \int \frac{v^2}{v_0^2} \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) (S_s)_0 \, dv \, d\vec{n} \\ &= \frac{1}{\gamma T} \int \int \frac{v}{v_0} (S_s)_0 \, dv \, d\vec{n} \quad (\text{see (12)}) \\ &= \frac{1}{\gamma T} \int \int (S_s)_0 \, dv \, d\vec{n} \quad (\text{invariance of the measure}).\end{aligned}$$

Due to the isotropy of the scattering in the comobile frame  $\vec{Q}_s^F = 0$ . The other term

$$\vec{Q}_s^R = -\frac{k}{h} \int \int \frac{1}{v} \log\left(\frac{n}{n+1}\right) S_s \, dv \, d\vec{n} \quad (34)$$

with  $S_s$  is given by (15). So  $\vec{Q}_s^R = -k/h \int \int (1/v_0) \log(n/(n+1)) (S_s)_0 \, dv_0 \, d\vec{n}_0$  due to the Lorentz invariance of the measure  $v \, dv \, d\vec{n} = v_0 \, dv_0 \, d\vec{n}_0$ . Then (16) implies  $\vec{Q}_s^R = \int \int (1/v_0) \sigma_s(v_0) q_s \, dv_0 \, d\vec{n}_0$ , where

$$q_s = \left( f\left(\frac{1}{4\pi} \int I(v_0, \vec{n}_0) \, d\vec{n}_0\right) - f(I_0) \right) \times \left( \frac{1}{4\pi} \int I(v_0, \vec{n}_0) \, d\vec{n}_0 - I_0 \right).$$

Here  $f$  denotes the function

$$f(x) = \frac{k}{h} \log\left(\frac{\frac{c^2}{2h} \frac{x}{v_0^3}}{\frac{c^2}{2h} \frac{x}{v_0^3} + 1}\right) \quad (35)$$

such that  $f(I_0) = (k/h) \log(n/(n+1))$ , see Eq. (102). Since  $f$  is strictly increasing for non-negative  $x$ , then

$$q_s = f' \left( (1 - \alpha) I_0 + \alpha \frac{1}{4\pi} \int I(v_0, \vec{n}_0) \, d\vec{n}_0 \right) \left[ \frac{1}{4\pi} \int I(v_0, \vec{n}_0) \, d\vec{n}_0 - I_0 \right]^2.$$

So  $q_s \geq 0$  and is zero if and only if  $(1/4\pi) \int I(v_0, \vec{n}_0) \, d\vec{n}_0 - I_0 = 0$ . Now the proof ends.

In order to state a similar result for the absorption–emission right-hand side  $S_a$ , a little difficulty arises. Indeed, we have replaced the Lorentz invariant Euler gas dynamics equation (4) by Galileo invariant gas dynamics equation (8), and this is a good approximation as soon as  $|\vec{v}|/c$  is small. However, in complete rigor the source term in (21) should have been placed on the right-hand side of the relativistic gas dynamic system (4) but not on the right-hand side of Galilean gas dynamic system (8). The difference produces a small discrepancy which might give as well an eventually negative term in the entropy production. However, this term is proportional to  $(|\vec{v}|/c)^2$  so it is meaningless for non-relativistic gases.

**Lemma 3** (Thermodynamic compatibility of the absorption–emission). *The absorption–emission entropy production  $\vec{Q}_a$  is the sum of two contributions  $\vec{Q}_a = P_a + \delta P_a$ . The first one is non-negative  $P_a \geq 0$  and is zero if and only if the radiation intensity is equal to the Planckian in the comobile frame  $I_0 = B_0(v_0, T)$ . The correction  $\delta P_a$  is due to the approximation of the relativistic gas dynamics system by the Galilean gas dynamic system:  $\delta P_a = O(|\vec{v}|^2/c^2)$  for smooth solutions.*

Let us go back to the definition of the entropy production  $\vec{Q}_a$ . We pose  $\vec{Q}_a = \vec{Q}_a^R + \vec{Q}_a^F$  where  $\vec{Q}_a^R$  (resp.  $\vec{Q}_a^F$ ) is the first (resp. second) term in (32) and is the contribution of the radiation (resp. fluid) to the total entropy production.

From (31) one has

$$\begin{aligned}\vec{Q}_a^F &= \partial_i(\rho S) + \nabla(\rho \vec{v} S) = -\frac{1}{T}(S_E + \vec{S}_F \cdot \vec{v}) \\ &= -\frac{\gamma}{T}(S_E + \vec{S}_F \cdot \vec{v}) + \frac{\gamma-1}{T}(S_E + \vec{S}_F \cdot \vec{v}) \\ &= -\frac{1}{T} \int \int \sigma_a S \gamma \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) dv d\vec{n} + \delta P_a,\end{aligned}\quad (36)$$

where by definition

$$\delta P_a = (\gamma - 1) \frac{S_E + \vec{S}_F \cdot \vec{v}}{T} = (\gamma - 1)[\partial_i(\rho S) + \nabla(\rho \vec{v} S)]. \quad (37)$$

$\delta P_a$  is clearly second order since  $\gamma - 1 = O(|\vec{v}|^2/c^2)$ . Due to (14) and (13), one has

$$\vec{Q}_a^F = -\frac{1}{T} \int \int \sigma_a \frac{v^2}{v_0^2} (B_0(v_0, T) - I_0) \gamma \left(1 - \frac{\vec{v} \cdot \vec{n}}{c}\right) dv d\vec{n} + \delta P_a,$$

that is

$$\vec{Q}_a^F = -\frac{1}{T} \int \int \sigma_a (B_0(v_0, T) - I_0) dv_0 d\vec{n}_0 + \delta P_a \quad (38)$$

due to (12) and the invariance of the measure  $v dv d\vec{n} = v_0 dv_0 d\vec{n}_0$ . On the other hand, we have

$$\vec{Q}_a^R = -\frac{k}{h} \int \int \frac{\sigma_a}{v_0} \log\left(\frac{n}{n+1}\right) \frac{v^2}{v_0^2} (B_0(v_0, T) - I_0) dv d\vec{n},$$

that is

$$\vec{Q}_a^R = -\frac{k}{h} \int \int \frac{\sigma_a}{v_0} \log\left(\frac{n}{n+1}\right) (B_0(v_0, T) - I_0) dv_0 d\vec{n}_0. \quad (39)$$

Combining (38) and (39) we get  $\vec{Q}_a = \vec{Q}_a^R + \vec{Q}_a^F = P_a + \delta P_a$  where by definition

$$\begin{aligned}P_a &= - \int \int \frac{\sigma_a}{v_0} \left( \frac{k}{h} \log\left(\frac{n}{n+1}\right) - \frac{v_0}{T} \right) (B_0(v_0, T) - I_0) dv_0 d\vec{n}_0 \\ &= \int \int \frac{\sigma_a}{v_0} (f(B_0(v_0, T)) - f(I_0))(B_0(v_0, T) - I_0) dv_0 d\vec{n}_0.\end{aligned}$$

The function  $f$  is defined in (35). So

$$P_a = \int \int \frac{\sigma_a(v_0)}{v_0} f'((1-\alpha)I_0 + \alpha B_0(v_0, T)) [B_0(v_0, T) - I_0]^2 dv_0 d\vec{n}_0 \geq 0$$

and is zero if and only if  $I_0 = B(v_0, T)$ . Now, the proof is complete.

### 3. Simplified models

Following [1] we study some asymptotic regimes of the full system (21) by means of non-dimensional variables. First, we assume that  $\sigma_a$  and  $\sigma_s$  are independent of the frequency

$$\sigma_a(v_0) = \sigma_a, \quad \sigma_s(v_0) = \sigma_s. \quad (40)$$

This hypothesis is called the gray hypothesis. This is of course a very crude approximation, but is motivated by the mathematical analysis.

Second, we introduce some hydrodynamics scales where  $a_\infty$  is a characteristic value of the fluid velocity and so on

$$\begin{aligned} x &= \hat{x}l, \quad t = \hat{t}l/a_\infty, \quad \rho = \hat{\rho}\rho_\infty, \quad v = \hat{v}a_\infty, \\ p &= \hat{p}\rho_\infty a_\infty^2, \quad T = \hat{T}T_\infty, \quad v = \hat{v}kT_\infty/h, \quad I = \hat{I}hca_r T_\infty^3/k, \\ \sigma_a &= \hat{\sigma}_a/\lambda_a, \quad \sigma_s = \hat{\sigma}_s/\lambda_s. \end{aligned} \quad (41)$$

A caret denotes a non-dimensional quantity, and

$$a_r = \frac{8\pi^5 k^4}{15c^3 h^3}. \quad (42)$$

We also define two non-dimensional parameters

$$\mathcal{C} = \frac{c}{a_\infty} \quad \text{and} \quad \mathcal{P} = \frac{a_r T_\infty^4}{\rho_\infty a_\infty^2}. \quad (43)$$

The first parameter,  $\mathcal{C}$ , is always large parameter for a flow non-relativistic. The second parameter,  $\mathcal{P}$ , measures the ratio of the radiative energy over the internal energy. It is possible to simplify by taking  $\mathcal{P} = 1$  in many cases, but the sake of compatibility with the notations of [1] we keep  $\mathcal{P}$ . We refer the reader to the paper [1] where it is shown that the non-dimensional equations derived from (21) are

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I}) &= -\mathcal{P} \vec{S}_F, \\ \frac{\partial}{\partial t}(\rho E) + \nabla \cdot (\rho E \vec{v} + p \vec{v}) &= -\mathcal{P} \mathcal{C} S_E, \\ \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} I + \vec{n} \cdot \nabla I &= S_I(v, \vec{n}). \end{aligned} \quad (44)$$

The interaction is now characterized by  $S_t = S_a + S_s$  with

$$S_a(v, \vec{n}) = \mathcal{L} \frac{v_0}{v} \sigma_a(v_0) \left[ \left( \frac{v}{v_0} \right)^3 B(v_0, T) - I \right] \quad (45)$$

and

$$S_s(v, \vec{n}) = \mathcal{L} \mathcal{L}_s \frac{v_0}{v} \sigma_s(v_0) \left[ \left( \frac{v}{v_0} \right)^3 \frac{1}{4\pi} \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}' - I \right], \quad (46)$$

where

$$\mathcal{L} = \frac{1}{\lambda_a}, \quad \mathcal{L}_s = \frac{\lambda_a}{\lambda_s} \quad (47)$$

and

$$B(v_0, T) = \frac{15v_0^3}{4\pi^5} (e^{v_0/T} - 1)^{-1}, \quad (48)$$

$$v_0 = v\gamma_L(1 - \vec{n} \cdot \vec{v}/\mathcal{C}), \quad v' = v \frac{1 - \vec{n} \cdot \vec{v}/\mathcal{C}}{1 - \vec{n}' \cdot \vec{v}/\mathcal{C}}, \quad (49)$$

$$\gamma_L = 1/\sqrt{1 - |\vec{v}|^2/\mathcal{C}^2}. \quad (50)$$

Since  $\partial_t E_r + \mathcal{C} \vec{F}_r = \mathcal{C} S_E$  we may prefer to use

$$\begin{aligned} \frac{\partial}{\partial t} (\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} \left( \rho \vec{v} + \frac{\mathcal{P}}{\mathcal{C}} \vec{F}_r \right) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I} + \mathcal{P} \mathbf{P}_r) &= 0, \\ \frac{\partial}{\partial t} (\rho E + \mathcal{P} E_r) \nabla \cdot (\rho E \vec{v} + p \vec{v} + \mathcal{P} \mathcal{C} \vec{F}_r) &= 0, \\ \frac{1}{\mathcal{C}} f \frac{\partial}{\partial t} I + \vec{n} \cdot \nabla I &= S_t(v, \vec{n}), \end{aligned} \quad (51)$$

instead of (44). Here,  $E_r$  (resp.  $\vec{F}_r \mathbf{P}_r$ ) is simply the total integrated intensity (resp. flux, pressure tensor) of the radiation

$$E_r = \int \int I dv d\vec{n}, \quad \vec{F}_r = \int \int \vec{n} I dv d\vec{n}, \quad \mathbf{P}_r = \int \int \vec{n} \otimes \vec{n} I dv d\vec{n}. \quad (52)$$

The non-dimensional entropy of the radiation is obtained by computing the ratio of (22) over  $\mathcal{P} \rho_\infty a_\infty^2 / T_\infty$ . Thus, we get the non-dimensional entropy of the radiation

$$S_r = -\frac{15}{4\pi^5} \int \int v^2 (n \log n - (n+1) \log(n+1)) dv d\vec{n}. \quad (53)$$

The non-dimensional entropy flux of the radiation is

$$\vec{Q}_r = -\frac{15}{4\pi^5} \int \int v^2 (n \log n - (n+1) \log(n+1)) \vec{n} dv d\vec{n}. \quad (54)$$

Here,  $n$  is already a non-dimensional variable that satisfies

$$n = \frac{4\pi^5}{15} \frac{I}{v^3}. \quad (55)$$

The non-dimensional entropy equation for smooth solutions of (44) or (51) with a zero right-hand side is

$$\frac{\partial}{\partial t} (\rho S + \mathcal{P} S_r) + \nabla \cdot (\rho S \vec{v} + \mathcal{P} \mathcal{C} \vec{Q}_r) = 0. \quad (56)$$

In the sequel we shall study mainly two different regimes. The first regime is the *equilibrium diffusion regime*

$$\mathcal{P} = O(1), \quad \mathcal{C} = O(\varepsilon^{-1}), \quad \mathcal{L}_s = O(\varepsilon^2), \quad \mathcal{L} = O(\varepsilon^{-1}),$$

the second regime is the *non-equilibrium diffusion regime*

$$\mathcal{P} = O(1), \quad \mathcal{C} = O(\varepsilon^{-1}), \quad \mathcal{L}_s = O(\varepsilon^{-2}), \quad \mathcal{L} = O(\varepsilon^1).$$

We also define some moment models using the variable Eddington factor approach.

### 3.1. Equilibrium diffusion

In this section we perform a formal Chapman–Enskog expansion for the non-dimensional system (44).

**Lemma 4.** Assume  $\mathcal{P} = 1$ ,  $\mathcal{C} = \varepsilon^{-1}$ ,  $\mathcal{L}_s = \varepsilon^2$ ,  $\mathcal{L} = \varepsilon^{-1}$  and assume the gray hypothesis (40). Then a first-order approximation of system (51) is

$$\begin{aligned} \frac{\partial}{\partial t} (\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} (\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + (p + p_r) \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t} (\rho E + E_r) + \nabla \cdot ((\rho E + E_r) \vec{v} + (p + p_r) \vec{v}) &= \nabla \cdot \left( \frac{1}{3\sigma_a} \nabla T^4 \right), \end{aligned} \quad (57)$$

where

$$E_r = T^4, \quad p_r = \frac{1}{3} T^4.$$

This system is often referred to as the equilibrium-diffusion limit [3,6].

The Chapman–Enskog expansion may be split in three steps

step 1: Let us first begin with the non-dimensional transfer equation rewritten as

$$\partial_t I + \frac{1}{\varepsilon} \vec{n} \cdot \nabla I = \frac{1}{\varepsilon^2} \frac{v_0}{v} \sigma_a \left[ \left( \frac{v}{v_0} \right)^3 B(v_0, T) - I \right] + O(1). \quad (58)$$

Here, only the absorption–emission contribution is taken into account, the scattering contribution is  $O(1)$ . Since we desire to equate all  $O(\varepsilon^{-2})$  and  $O(\varepsilon^{-1})$  terms in the equation, we have to expand the right-hand side. So (58) is rewritten as

$$\partial_t I + \frac{1}{\varepsilon} \vec{n} \cdot \nabla I = \frac{1}{\varepsilon^2} A_{-2} + \frac{1}{\varepsilon} A_{-1} + O(1). \quad (59)$$

With straightforward notations, one has

$$\begin{aligned} T &= T_0 + \varepsilon T_1 + O(\varepsilon^2), \\ v &= v_0 + v_0 \varepsilon \vec{n} \cdot \vec{v} + O(\varepsilon^2) \quad (\text{due to (49)}), \\ B(v_0, T) &= B(v, T) - \varepsilon v \vec{n} \cdot \vec{v} \frac{\partial}{\partial v} B(v, T) + O(\varepsilon^2), \\ B(v, T) &= B(v, T_0) + \varepsilon \frac{\partial}{\partial T} B(v, T_0) T_1 + O(\varepsilon^2), \\ I(v, \vec{n}) &= I^0 + \varepsilon I^1(v, \vec{n}) + O(\varepsilon^2). \end{aligned} \quad (60)$$

Be careful that the intensity in the comobile frame  $I_0$  has little to do with first-order term in the expansion of the intensity in the lab frame  $I_0$ . Expansion of the right-hand side (58) and (59) gives  $A_{-2} = B(v, T_0) - I^0(v, \vec{n})$  and

$$\begin{aligned} A_{-1} &= -\vec{n} \cdot \vec{v} (B(v, T_0) - I^0(v, \vec{n})) \\ &\quad + \left( \frac{\partial}{\partial T} B(v, T_0) T_1 + \vec{n} \cdot \vec{v} (3 - v \partial_v) B(v, T_0) - I^1(v, \vec{n}) \right). \end{aligned}$$

Equating all negative powers of  $\varepsilon$  in (59), one gets

$$\begin{aligned} 0 &= B(v, T_0) - I^0(v, \vec{n}), \\ \vec{n} \cdot \nabla I^0 &= \sigma_a \left[ \frac{\partial}{\partial T} B(v, T_0) T_1 + \vec{n} \cdot \vec{v} (3 - v \partial_v) B(v, T_0) - I^1(v, \vec{n}) \right]. \end{aligned} \quad (61)$$

*step 2:* Next stage is to use these relations in order to expand  $E_r$ ,  $\vec{F}_r$  and  $\mathbf{P}_r$  given by (52). For the radiation energy one obtains

$$\begin{aligned} E_r &= \int \int I^0(v, \vec{n}) dv d\vec{n} + O(\varepsilon) = \int \int B(v, T_0) dv d\vec{n} + O(\varepsilon) \\ &= \frac{15}{4\pi^5} \int \int \frac{v^3}{e^{v/T} - 1} dv d\vec{n} + O(\varepsilon) = T_0^4 + O(\varepsilon). \end{aligned} \quad (62)$$

since  $\int \int [v^3/(e^v - 1)] dv d\vec{n} = 4\pi^5/15$ . It remains to study the radiation flux using (61)

$$\begin{aligned} \vec{F}_r &= \int \int (I^0(v, \vec{n}) + \varepsilon I^1(v, \vec{n})) \vec{n} dv d\vec{n} + O(\varepsilon^2) \\ &= \int \int B(v, T_0) \vec{n} dv d\vec{n} + \varepsilon \int \int \left( \frac{\partial}{\partial T} B(v, T_0) T_1 + \vec{n} \cdot \vec{v} (3 - v \partial_v) B(v, T_0) - \frac{1}{\sigma_a} \vec{n} \cdot \nabla I^0 \right) \\ &\quad \times \vec{n} dv d\vec{n} + O(\varepsilon^2). \end{aligned} \quad (63)$$



In this expression integrals of functions linear with respect to  $\vec{n}$  disappear. An elementary integration by parts gives  $\int v \partial_v B(v, T_0) dv = - \int B(v, T_0) dv$ . Since  $\int \vec{n} \otimes \vec{n} dv dn = \frac{1}{3} I_d$  where  $I_d$  is the identity matrix then

$$\vec{F}_r = \varepsilon \left[ \vec{v} \frac{4}{3} T_0^4 - \frac{1}{3\sigma_a} \nabla T_0^4 \right] + O(\varepsilon^2). \quad (64)$$

Finally,

$$\mathbf{P}_r = \int \int (I^0(v, \vec{n}) + \varepsilon I^1(v, \vec{n})) \vec{n} \otimes \vec{n} dv d\vec{n} + O(\varepsilon^2) = \frac{1}{3} T_0^4 \mathbf{I} + O(\varepsilon). \quad (65)$$

step 3: We expand the first three equations of (51) using (62)–(65). Thus, one obtains the system, exact to  $O(\varepsilon)$ ,

$$\begin{aligned} \frac{\partial}{\partial t}(\rho^0) + \nabla \cdot (\rho^0 \vec{v}^0) &= 0, \\ \frac{\partial}{\partial t}(\rho^0 \vec{v}^0) + \nabla \cdot (\rho^0 \vec{v}^0 \otimes \vec{v}^0 + (p_0 + \frac{1}{3} T_0^4) \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t}(\rho^0 E^0 + T_0^4) + \nabla \cdot \left( (\rho^0 E^0 + T_0^4) \vec{v}^0 + (p_0 + \frac{1}{3} T_0^4) \vec{v}^0 - \frac{1}{3\sigma_a} \nabla T_0^4 \right) &= 0. \end{aligned} \quad (66)$$

The proof now ends.

### 3.2. Non-equilibrium diffusion

In this section, we study the non-equilibrium diffusion limit of the model. One has

**Lemma 5.** Assume that  $\mathcal{P} = 1$ ,  $\mathcal{C} = \varepsilon^{-1}$ ,  $\mathcal{L}_s = \varepsilon^{-2}$ ,  $\mathcal{L} = \varepsilon^1$  and assume the gray hypothesis (40). Then a first-order approximation of system (51) is

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t}(\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + (p + p_r) \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \nabla \cdot ((\rho E + E_r) \vec{v} + (p + p_r) \vec{v}) &= \nabla \cdot \left( \frac{1}{3\sigma_s} \nabla T_r^4 \right), \\ \frac{\partial}{\partial t} E_r + \nabla \cdot (\vec{v} E_r) + p_r \nabla \cdot \vec{v} &= \nabla \cdot \left( \frac{1}{3\sigma_s} \nabla T_r^4 \right) + \sigma_a (T^4 - T_r^4), \end{aligned} \quad (67)$$

where

$$E_r = T_r^4, \quad p_r = \frac{1}{3} T_r^4.$$

This system is often referred to as the non-equilibrium-diffusion limit [3,6].

The scaling means that the scattering is the dominant contribution in the source term. Note that (67) has been already derived by various authors, but in Lagrangian coordinates and using more

physical intuition together with a deep understanding of the various scales of the problem than a rigorous mathematical approach. As for us we use a direct Chapman–Enskog expansion, divided in three steps. The key point is the use of formula (19) which allows to simplify the algebraic complexity of the expansion.

*step 1:* Following the method used in the previous section, we begin with the transfer equation in (44) rewritten as

$$\begin{aligned} \partial_t I + \frac{1}{\varepsilon} \vec{n} \cdot \nabla I = \frac{v_0}{v} \sigma_a(v_0) \left[ \left( \frac{v}{v_0} \right)^3 B(v_0, T) - I \right] \\ + \frac{1}{\varepsilon^2} \frac{v_0}{v} \sigma_s(v_0) \left[ \left( \frac{v}{v_0} \right)^3 \frac{1}{4\pi} \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}' - I \right]. \end{aligned} \quad (68)$$

We want to rewrite this expression as

$$\partial_t I + \frac{1}{\varepsilon} \vec{n} \cdot \nabla I = \frac{1}{\varepsilon^2} A_{-2} + \frac{1}{\varepsilon} A_{-1} + A_0 + O(\varepsilon). \quad (69)$$

So we use (60) and expand the scattering

$$\left[ \left( \frac{v}{v_0} \right)^3 \frac{1}{4\pi} \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}' - I \right] = C_{-2} + \varepsilon C_{-1} + \varepsilon^2 C_0 + O(\varepsilon^3). \quad (70)$$

Then

$$\begin{aligned} A_{-2} &= \sigma_s C_{-2} = \sigma_s \left( \frac{1}{4\pi} \int I^0(v, \vec{n}) d\vec{n} - I^0 \right), \\ A_{-1} &= \sigma_s C_{-1} = \sigma_s \left( -\frac{1}{4\pi} \int \vec{n} \cdot \vec{v} I^0(v, \vec{n}) d\vec{n} + 3\vec{n} \cdot \vec{v} \frac{1}{4\pi} \int I^0(v, \vec{n}) d\vec{n} \right. \\ &\quad \left. + \frac{1}{4\pi} \int v(\vec{n}' \cdot \vec{v} - \vec{n} \cdot \vec{v}) \frac{\partial I^0(v, \vec{n}')}{\partial v} d\vec{n}' + \frac{1}{4\pi} \int I^1(v, \vec{n}) d\vec{n} - I^1 - \vec{v} \cdot \vec{n} C_{-2} \right), \end{aligned}$$

and

$$A_0 = \sigma_s C_0 + \sigma_a(B(v, T) - I^0). \quad (71)$$

The contribution  $\int v(\vec{n}' \cdot \vec{v} - \vec{n} \cdot \vec{v})(\partial I^0(v, \vec{n}')/\partial v) d\vec{n}'$  is the consequence of

$$\begin{aligned} I(v', \vec{n}') &= I \left( v \frac{1 - \varepsilon \vec{n} \cdot \vec{v}}{1 - \varepsilon \vec{n}' \cdot \vec{v}}, \vec{n}' \right) \\ &= I(v, \vec{n}') + \varepsilon v(\vec{n}' \cdot \vec{v} - \vec{n} \cdot \vec{v}) \frac{\partial I(v, \vec{n}')}{\partial v} + O(\varepsilon^2). \end{aligned}$$

Now equating the  $\varepsilon^{-2}$  terms in Eq. (68), we get

$$C_{-2} = \frac{1}{4\pi} \int I^0(v, \vec{n}) d\vec{n} - I^0 = 0,$$

which means that the radiation is isotropic  $I^0(v, \vec{n}) = I^0(v)$ . Equating the  $\varepsilon^{-1}$  terms in Eq. (68) and using  $\int I^0(v, \vec{n}) \vec{n} d\vec{n} = 0$  plus  $\int v(\vec{n}' \cdot \vec{v} - \vec{n} \cdot \vec{v})(\partial I^0(v, \vec{n}')/\partial v) d\vec{n}' = -\vec{n} \cdot \vec{v} \int v(\partial I^0(v, \vec{n})/\partial v) d\vec{n}$ , we get

$$\sigma_s C_{-1} = \vec{n} \cdot \nabla I^0 = \sigma_s \left( 3\vec{n} \cdot \vec{v} \frac{1}{4\pi} \int I^0(v, \vec{n}) d\vec{n} - \frac{1}{4\pi} \vec{n} \cdot \vec{v} \int v \frac{\partial I^0(v, \vec{n})}{\partial v} d\vec{n} + \frac{1}{4\pi} \int I^1(v, \vec{n}) d\vec{n} - I^1 \right). \quad (72)$$

It remains to express the  $\varepsilon^0$  terms in expression (68). In order to simplify the analysis which can be very cumbersome, we use the property (19) which expresses the fact that the scattering contribution is isotropic in the comobile reference frame. So

$$\int \int \frac{v_o}{v} (C_{-2} + \varepsilon C_{-1} + \varepsilon^2 C_0 + O(\varepsilon^3)) dv d\vec{n} = 0.$$

Since  $v_o/v = (1 - \varepsilon \vec{n} \cdot \vec{v})/\gamma$ , then we are able to expand in power of  $\varepsilon$ . We get three formulas. The first one is  $\int \int C_{-2} dv d\vec{n} = 0$ , the second one is  $\int \int C_{-1} dv d\vec{n} = \int \int \vec{n} \cdot \vec{v} C_{-2} dv d\vec{n}$  and the third one is

$$\int \int C_0 dv d\vec{n} = \int \int \vec{n} \cdot \vec{v} C_{-1} dv d\vec{n}. \quad (73)$$

Combining (73) with (72), it gives

$$\sigma_s \int \int C_0 dv d\vec{n} = \int \int (\vec{n} \cdot \vec{v})(\vec{n} \cdot \nabla I^0) dv d\vec{n} = \frac{1}{3} \vec{v} \cdot \int \nabla I^0 dv. \quad (74)$$

Here we have used the isotropy of  $I_0$  and

$$\int \int \vec{n} \otimes \vec{n} d\vec{n} = \frac{1}{3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \frac{1}{3} \mathbf{I}. \quad (75)$$

This expression will be used to simplify the integrated form of (71).

*step 2:* Now we use all this informations in order to calculate  $E_r$  and  $F_r$ .

First, we define  $T_r^4 = \int \int I^0(v, \vec{n}) dv d\vec{n}$ , so

$$E_r = \int \int I^0(v, \vec{n}) dv d\vec{n} + O(\varepsilon) = T_r^4 + O(\varepsilon). \quad (76)$$

Eq. (74) can be also written as

$$\sigma_s \int \int C_0 dv d\vec{n} = \vec{v} \cdot \nabla p_r, \quad (77)$$

where by definition  $p_r = \frac{1}{3} T_r^4$ . Then we compute  $\vec{F}_r$

$$\begin{aligned} \vec{F}_r &= \int \int (I^0(v, \vec{n}) + \varepsilon I^1(v, \vec{n})) \vec{n} dv d\vec{n} + O(\varepsilon^2), \\ \vec{F}_r &= \varepsilon \int \int \left[ \vec{n} \cdot \vec{v} \frac{1}{4\pi} \int \left( 3 - v \frac{\partial}{\partial v} \right) I^0(v, \vec{n}) d\vec{n} \right. \\ &\quad \left. + \frac{1}{4\pi} \int I^1(v, \vec{n}) d\vec{n} - \frac{1}{\sigma_s} \vec{n} \cdot \nabla I^0 \right] \vec{n} dv d\vec{n} + O(\varepsilon^2). \end{aligned}$$

Using an integration by part to get rid of the  $v\partial/\partial v$  we get

$$\vec{F}_r = \varepsilon \left[ \vec{v} \frac{4}{3} T_r^4 - \frac{1}{3\sigma_s} \nabla T_r^4 \right] + O(\varepsilon^2). \quad (78)$$

This formula is the same as (64), except that the temperature of the fluid  $T_0$  is replaced by what will be referred to as the temperature of the radiation  $T_r$ . Similarly to (65) one has

$$\mathbf{P}_r = \int \int (I^0(v, \vec{n}) + \varepsilon I^1(v, \vec{n})) \vec{n} \otimes \vec{n} dv d\vec{n} + O(\varepsilon^2) = \frac{1}{3} T_r^4 \mathbf{I} + O(\varepsilon). \quad (79)$$

step 3: It remains to expand (44) using (76)–(79). One gets exactly the same result as in the equilibrium regime, except that  $T_0$  is replaced by  $T_r$ . Thus, one obtains the system, exact to  $O(\varepsilon)$ ,

$$\begin{aligned} \frac{\partial}{\partial t}(\rho^0) + \nabla \cdot (\rho^0 \vec{v}^0) &= 0, \\ \frac{\partial}{\partial t}(\rho^0 \vec{v}^0) + \nabla \cdot (\rho^0 \vec{v}^0 \otimes \vec{v}^0 + (p_0 + \frac{1}{3} T_r^4) \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t}(\rho^0 E^0 + T_r^4) + \nabla \cdot \left( (\rho^0 E^0 + T_r^4) \vec{v}^0 + (p_0 + \frac{1}{3} T_r^4) \vec{v}^0 - \frac{1}{3\sigma_s} \nabla T_r^4 \right) &= 0. \end{aligned} \quad (80)$$

Finally, we need an equation for  $T_r$  in order to close (80). For this we integrate the equation of transfer and consider only the  $O(1)$  contribution. It gives

$$\partial_t \int \int I^0 dv d\vec{n} + \nabla \cdot \int \int I \vec{n} dv d\vec{n} = \int \int A_0 dv d\vec{n} + O(\varepsilon). \quad (81)$$

Since  $A_0$  is given by (71), one has

$$\begin{aligned} \int \int A_0 dv d\vec{n} &= \sigma_a \int \int (B(v, T) - I_0) dv d\vec{n} + \sigma_s \int \int C_0 dv d\vec{n} + O(\varepsilon) \\ &= \sigma_a (T_0^4 - T_r^4) + \vec{v} \cdot \nabla p_r + O(\varepsilon). \end{aligned}$$

Using the value of  $\nabla \cdot \int \int I \vec{n} dv d\vec{n}$  given in (78), one gets

$$\partial_t E_r + \nabla \cdot \left( \vec{v} \frac{4}{3} T_r^4 - \frac{1}{3\sigma_s} \nabla T_r^4 \right) = \sigma_a (T_0^4 - T_r^4) + \vec{v} \cdot \nabla p_r + O(\varepsilon)$$

finally rewritten as

$$\partial_t E_r + \nabla \cdot (\vec{v} E_r) + p_r \nabla \cdot \vec{v} = \sigma_a (T_0^4 - T_r^4) + \nabla \cdot \left( \frac{1}{3\sigma_s} \nabla T_r^4 \right) + O(\varepsilon). \quad (82)$$

In conjunction with (80), we have the proof.

**Corollary 1.** *If one uses a non-relativistic source term in the right-hand side of the transfer equation (51), then one misses the  $p_r \nabla \cdot \vec{v}$  in the non-equilibrium diffusion model (67).*

Non-relativistic means that one equates  $v = v_0$  in the definition of the source term, even for  $\vec{v} \neq 0$ . So a non-relativistic scattering source term will be isotropic in the lab frame, and not in the comobile

frame as in (19). Since  $p_r \nabla \cdot \vec{v}$  is a direct consequence of formula (19), the proof of the corollary ends.

**Corollary 2.** Consider the non-equilibrium diffusion model (67). Let us define

$$\bar{S}_r = \frac{4}{3} T_r^3. \quad (83)$$

Then (67) may be rewritten as

$$\begin{aligned} \frac{\partial}{\partial t} (\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} (\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + (p + p_r) \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t} (\rho E + E_r) + \nabla \cdot ((\rho E + E_r) \vec{v} + (p + p_r) \vec{v}) &= \nabla \cdot \left( \frac{1}{3\sigma_s} \nabla T_r^4 \right), \\ \frac{\partial}{\partial t} \bar{S}_r + \nabla \cdot (\vec{v} \bar{S}_r) &= \frac{1}{T_r} \nabla \cdot \left( \frac{1}{3\sigma_s} \nabla T_r^4 \right) + \sigma_a \frac{T^4 - T_r^4}{T_r}. \end{aligned} \quad (84)$$

The quantity  $\bar{S}_r$  is formerly the radiative entropy at equilibrium. But since  $S_r$  has been already defined in (22) then  $\bar{S}_r$  is for the moment different from  $S_r$ . We will see in Lemma 7 that these quantities are in some sense equal. The proof of corollary 2 is just a matter of direct calculation. We just divide the last equation of (67) by  $T_r$  and rearrange all terms due to

$$\begin{aligned} \frac{1}{T_r} \left( \frac{\partial}{\partial t} E_r + \nabla \cdot (\vec{v} E_r) + p_r \nabla \cdot \vec{v} \right) &= \frac{1}{T_r} \left( \frac{\partial}{\partial t} T_r^4 + \nabla T_r^4 \cdot \vec{v} + T_r^4 \nabla \cdot \vec{v} + \frac{1}{3} T_r^4 \nabla \cdot \vec{v} \right) \\ &= \frac{\partial}{\partial t} \left( \frac{4}{3} T_r^3 \right) + \nabla \cdot \left( \frac{4}{3} T_r^3 \right) \cdot \vec{v} + \frac{4}{3} T_r^3 \nabla \cdot \vec{v} = \frac{\partial}{\partial t} \bar{S}_r + \nabla \cdot (\vec{v} \bar{S}_r). \end{aligned}$$

An important advantage of (84) against (67) is that it admits a natural conservative limit for

$$\sigma_s \approx +\infty \quad \text{and} \quad \sigma_a \approx 0. \quad (85)$$

Indeed the limit is the system of conservation laws

$$\begin{aligned} \frac{\partial}{\partial t} (\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} (\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + (p + p_r) \mathbf{I}) &= 0, \\ \frac{\partial}{\partial t} (\rho E + E_r) + \nabla \cdot ((\rho E + E_r) \vec{v} + (p + p_r) \vec{v}) &= 0, \\ \frac{\partial}{\partial t} \bar{S}_r + \nabla \cdot (\vec{v} \bar{S}_r) &= 0. \end{aligned} \quad (86)$$

This system admits discontinuous solutions (*à la* Rankine–Hugoniot), which is not the case for (67) due to the non-conservative term  $p_r \nabla \cdot \vec{v}$ . This remark will be used in our last section.

### 3.3. $P^1$ moments models

Let us now recall briefly another type of approximate transport models, the  $P^1$  moment models. By taking the moments of the radiative transfer equation in (44) against 1 and  $\vec{n}$ , one obtain the general form of  $P^1$  moment models

$$\frac{1}{\mathcal{C}} \frac{\partial}{\partial t} E_r + \nabla \vec{F}_r = S_E, \quad (87)$$

$$\frac{1}{\mathcal{C}} \frac{\partial}{\partial t} \vec{F}_r + \nabla \mathbf{P}_r = \vec{S}_F. \quad (88)$$

It is expected that such moment models should give a good approximation of the solution of the radiative transfer equation in (44) in the case of small anisotropy (that is  $\|\vec{F}_r/E_r\|$  small). But they are widely used in the full range of anisotropy (that is for  $\|\vec{F}_r/E_r\| \leq 1$ ). However, one must close the system (87) and (88). This is done with a formula where  $\mathbf{P}_r$  is given in terms of  $E_r$  and  $\vec{F}_r$ .

A popular closure is called the *Variable Eddington Factor*. In this method one takes  $\mathbf{P}_r$  as

$$\mathbf{P}_r = E_r D_r = E_r \left( \frac{1-\chi}{2} \mathbf{I} + \frac{3\chi-1}{2} f \otimes f \right), \quad (89)$$

where  $\vec{f} = \vec{F}_r/E_r$ :  $\chi = \chi(\|\vec{f}\|)$  is the *Eddington factor* and is usually chosen as a function of  $\|\vec{f}\|$ . Then the job consist to specify the Eddington factor. As we can notice in the literature, there is a lot of propositions. Considering the solution of the transfer equation, it seems natural to choose an Eddington factor which satisfies the constraints  $\|\vec{f}\|^2 \leq \chi(\|\vec{f}\|) \leq 1$ ,  $\chi(0) = 1/3$ ,  $\chi(1) = 1$  (we refer to [14]).

The Eddington approximation, which correspond to a constant Eddington factor  $\chi = \frac{1}{3}$  does not fulfill all these requirements: it leads to solutions for which  $\|\vec{f}\| > 1$ , see [15]. We refer to [16] or [14] for different types of closure. To derive a  $P^1$  approximation of the relativistic transfer equation in (44), we will use a method based on maximizing the entropy under constraints and which was already used in the non-relativistic case by several authors [13,14,17,18]. As we will see in the next, this approach gives a well-known Eddington factor, see [7,14,18].

$$\chi(x) = \frac{3 + 4x^2}{5 + 2\sqrt{4 - 3x^2}}. \quad (90)$$

Another difficulty with the variable Eddington factor approach is that one must deal with the relativistic source terms and approximate these terms by simpler ones, essentially by expanding them using Taylor expansion in power of  $|v|/c$ . All the discussion is about which term must be kept to give the good equilibrium state.

As for us we consider that such approximate models must contain the equilibrium diffusion approximation. Our contribution in this paper is to show how to design  $P^1$  approximate models which permit also to recover the non-equilibrium diffusion approximation. To our understanding this is not the case for some previous works, see for example [3,13]. For example, the  $P^1$  model used by Feugeas and Dubroca [13] coupled with the Euler equations, for the matter, cannot give the  $p_r \nabla \cdot \vec{v}$  term in the non-equilibrium diffusion approximation since the source term have been treated in a non-relativistic way (it is an application of the Corollary 1).

#### 4. Method of moments

The method of moments applied to systems of conservation laws (we refer to [14,19] in a particular context [15], and references therein) amounts to the derivation of reduced and well-posed systems of conservations laws: reduced means that the number of equations of the reduced system is smaller than for the original system; well posed means that the reduced system is still hyperbolic if the original system was hyperbolic. Applications of the method of moments to the transfer equation may be founded in [13,17].

##### 4.1. Generality about the method of moments for radiation hydrodynamics

Let us consider the non-dimensional radiative entropy  $S_r$  defined by (53).

**Lemma 6.** *Let  $(E_r, \vec{F}_r) \in \mathbb{R} \times \mathbb{R}^3$  with  $E_r > 0$  and  $|\vec{F}_r| \leq E_r$ . The minimum of the radiative entropy  $S_r$  with constraints*

$$\int \int I \, dv \, d\vec{n} = E_r \quad \text{and} \quad \int \int I \vec{n} \, dv \, d\vec{n} = \vec{F}_r \quad (91)$$

is given by

$$n = \frac{1}{e^{(v/\Theta_r) + v\vec{b} \cdot \vec{n}/\Theta_r} - 1} \quad (92)$$

where  $\Theta_r > 0$  and  $|\vec{b}| < 1$ .

Then the pressure is given by the well-known formula (Levermore [14], see also in the appendix)

$$\mathbf{P}_r = E_r D_r = E_r \left( \frac{1-\chi}{2} \mathbf{I} + \frac{3\chi-1}{2} \vec{f} \otimes \vec{f} \right) \quad (93)$$

with  $\vec{f} = \vec{F}_r/E_r$  and the Eddington factor  $\chi$  given by

$$\chi = \frac{3 + 4\|\vec{f}\|^2}{5 + 2\sqrt{4 - 3\|\vec{f}\|^2}}. \quad (94)$$

Since  $S_r$  is strictly concave with respect to  $I$ , it is sufficient to check the optimality conditions, see [13,17] for more details. We construct the Lagrangian with Lagrange multipliers  $1/\Theta_r \in \mathbb{R}$  and  $\vec{b}/\Theta_r \in \mathbb{R}^3$

$$L = S_r - \frac{1}{\Theta_r} E_r - \frac{\vec{b}}{\Theta_r} \cdot \vec{F}_r.$$

The optimality conditions are simply

$$0 = dL = dS_r - \frac{1}{\Theta_r} dE_r - \frac{\vec{b}}{\Theta_r} d\vec{F}_r. \quad (95)$$

Due to (52) and  $dS_r = \int_v \int_{\vec{n}} (1/v) \log(n/(n+1)) dI dv d\vec{n}$ , one gets

$$0 = - \int_v \int_{\vec{n}} \left[ \frac{1}{v} \log\left(\frac{n}{n+1}\right) + \frac{1}{\Theta_r} + \frac{\vec{b}}{\Theta_r} \cdot \vec{n} \right] dI dv d\vec{n}.$$

Thus, one has  $\log(n/(n+1)) + v/\Theta_r + (v\vec{b}/\Theta_r) \cdot \vec{n} = 0$ , that is  $n/(n+1) = e^{-(v/\Theta_r + v\vec{b} \cdot \vec{n}/\Theta_r)}$  which is equivalent to (92). Since  $n \geq 0$  for all  $(v, \vec{n})$  then we need the compatibility condition  $\Theta_r > 0$  and  $|\vec{b}| < 1$ . The proof now ends.

Then one defines the reduced and closed system

$$\begin{aligned} \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} E_r + \nabla \vec{F}_r &= S_E, \\ \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} \vec{F}_r + \nabla \mathbf{P}_r &= \vec{S}_F. \end{aligned} \quad (96)$$

These four equations (one for  $E_r$  and three for  $\vec{F}_r$ ) are the moments of (17) against  $(1, \vec{n}_1, \vec{n}_2, \vec{n}_3) = (1, \vec{n})$ . System (96) is closed in the sense that (96) contains four equations while the intensity is given by four degrees of freedom (see (98)). The complete moment model deduced from (96) and (51) is then

$$\begin{aligned} \frac{\partial}{\partial t} (\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} \left( \rho \vec{v} + \frac{\mathcal{P}}{\mathcal{C}} \vec{F}_r \right) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I} + \mathcal{P} \mathbf{P}_r) &= 0, \\ \frac{\partial}{\partial t} (\rho E + \mathcal{P} E_r) + \nabla \cdot (\rho E \vec{v} + p \vec{v} + \mathcal{P} \vec{F}_r) &= 0, \\ \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} E_r + \nabla \vec{F}_r &= S_E, \\ \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} \vec{F}_r + \nabla \cdot \mathbf{P}_r &= \vec{S}_F, \end{aligned} \quad (97)$$

together with

$$\frac{4\pi^5}{15} \frac{I}{v^3} = \frac{1}{e^{(v/\Theta_r) + v\vec{b} \cdot \vec{n}/\Theta_r} - 1}. \quad (98)$$

#### 4.2. Chapman–Enskog expansion of the moment model

In this section we prove that the moment model (97) contains the non-equilibrium diffusion in the sense that (67) is recovered as the Chapman–Enskog expansion of the moment model.

**Lemma 7.** Assume that  $\mathcal{P} = 1$ ,  $\mathcal{C} = \varepsilon^{-1}$ ,  $\mathcal{L}_s = \varepsilon^{-2}$ ,  $\mathcal{L} = \varepsilon^1$  and assume the gray hypothesis (40). Then a first-order approximation of system (97) is the non-equilibrium diffusion model (67). The



coefficients of the radiation (see (98)) are

$$\Theta_r = T_r + O(\varepsilon), \quad \vec{b} = O(\varepsilon). \quad (99)$$

The second equation  $\vec{b} = O(\varepsilon)$  means that the radiation is isotropic at the limit. We also get  $S_r = \bar{S}_r + O(\varepsilon)$  where  $\bar{S}_r = \frac{4}{3} T_r^3$  (83).

This result means that we have not lost too much informations by taking the first two moments of the transfer equation. At least we are able to recover the non-equilibrium diffusion limit. It also justifies the use of  $\bar{S}_r$ , which is the radiative entropy, in (84). One must be convinced that this result is not trivial. Indeed, one never assumes in the analysis of the non-equilibrium diffusion limit that radiation is closed to a Planckian with a radiative temperature  $T_r$ . So the exact shape of the radiation  $I$  is not addressed in the non-equilibrium diffusion limit, even if the model behaves just as if the radiation intensity is closed to a Planckian around  $T_r$ . On the other hand, the moment model assumes such a representation for the intensity, which is a generalized Planckian. So one might loose too much information with the moment model compared with the non-equilibrium diffusion model. The lemma shows it is not the case. The proof essentially uses the same method as in the proof of Lemma 5.

Step 1: Of course we expand

$$E_r = E_{r0} + \varepsilon E_{r1} + O(\varepsilon^2), \quad \vec{F}_r = \mathbf{F}_{r0} + \varepsilon \mathbf{F}_{r1} + O(\varepsilon^2)$$

and

$$\Theta_r = \Theta_{r0} + O(\varepsilon), \quad \vec{b} = \vec{b}_0 + O(\varepsilon).$$

Thus, the intensity of radiation is  $I = I^0 + \varepsilon I^1 + o(\varepsilon)$  where

$$I^0 = \frac{15}{4\pi^5} v^3 \frac{1}{e^{v/(\Theta_{r0} + v\vec{b}_0 \cdot \vec{n})/\Theta_{r0}} - 1}. \quad (100)$$

Comparing the system (97) and (69) and using the same scaling, we deduce some equalities

$$\int \int C_{-2} dv d\vec{n} = 0, \quad \int \int C_{-2} \vec{n} dv d\vec{n} = 0 \quad (\text{order } \varepsilon^{-2}), \quad (101)$$

$$\int \int \vec{n} \cdot \nabla I_0 dv d\vec{n} = \sigma_s \int \int C_{-1} dv d\vec{n} = 0 \quad (102)$$

and

$$\int \int \vec{n} \otimes \vec{n} \nabla I_0 dv d\vec{n} = \sigma_s \int \int C_{-1} \vec{n} dv d\vec{n} = 0 \quad (\text{order } \varepsilon^{-1}), \quad (103)$$

where  $C_{-2}$  and  $C_{-1}$  are defined in (69). The first equation of (101) is of course always true. The second equation of (101) implies that

$$-\int \int \vec{n} I^0 d\vec{n} = \int \int \vec{n} \left( \frac{1}{4\pi} \int I^0(v, \vec{n}) d\vec{n} - I^0 \right) = 0.$$

It is equivalent (see the appendix) to

$$0 = \frac{4}{3 + |\vec{b}_0|^2} \left( \int \int I^0 d\vec{n} \right) \vec{b}_0$$

which in turn implies  $\vec{b}_0 = 0$ : the first-order term of the radiation is isotropic. Eqs. (102) and (103) are the integrated counterpart of (72). Since (73) is still true and the first-order term of the radiation is isotropic we also deduce (74).

Steps 2 and 3 are identical to steps (2) and (3) in the proof of Lemma 5.

Step 4: We have to check (99): since  $\vec{b}_0 = 0$  we already have  $\vec{b} = O(\varepsilon)$ . Due to (100) one has  $\int \int I d\vec{n} = \Theta_r^4 + O(\varepsilon)$ . Compared with the definition of  $T_r$  (76) we get  $\Theta_r = T_r + O(\varepsilon)$ . The proof now ends.

#### 4.3. An alternative moment model

In this section we use an elementary algebraic relation in order to rewrite the moment model (97) using  $(S_r, \vec{F}_r)$  instead of  $(E_r, \vec{F}_r)$ . Of course it is possible to use the general theory of the method of moment to prove the result but the proof presented here has the advantage to be self contained.

**Lemma 8.** *Smooth solutions of (97) are also smooth solutions of*

$$\begin{aligned} \frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial}{\partial t} \left( \rho \vec{v} + \frac{\mathcal{P}}{\mathcal{C}} \vec{F}_r \right) + \nabla \cdot (\rho \vec{v} \otimes \vec{v} + p \mathbf{I} + \mathcal{P} \mathbf{P}_r) &= 0, \\ \frac{\partial}{\partial t} (\rho E + \mathcal{P} E_r) + \nabla \cdot (\rho E \vec{v} + p \vec{v} + \mathcal{P} \mathcal{C} \vec{F}_r) &= 0, \\ \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} S_r + \nabla \cdot \vec{Q}_r &= \frac{1}{\Theta_r} (S_E + \vec{b} \cdot \vec{S}_F), \\ \frac{1}{\mathcal{C}} \frac{\partial}{\partial t} \vec{F}_r + \nabla \cdot \mathbf{P}_r &= \vec{S}_F. \end{aligned} \quad (104)$$

Due to (95) one has

$$\partial_t S_r = \frac{1}{\Theta_r} \partial_t E_r + \frac{\vec{b}}{\Theta_r} \cdot \partial_t \vec{F}_r. \quad (105)$$

It is easy to get a similar formula for the entropy flux. The non-dimensional entropy flux is  $\vec{Q}_r = -(15/4\pi^5) \int \int v^2 (n \log n - (n+1) \log(n+1)) \vec{n} dv d\vec{n}$ . Thus,

$$\begin{aligned} d\vec{Q}_r &= \int \int \frac{1}{v} \log \left( \frac{n}{n+1} \right) dI \vec{n} dv d\vec{n} \\ &= \int \int \frac{1}{v} \left( \frac{1}{\Theta_r} + \frac{\vec{b} \cdot \vec{n}}{\Theta_r} \right) dI \vec{n} dv d\vec{n} = \frac{1}{\Theta_r} d\vec{F}_r + d\mathbf{P}_r \frac{\vec{b}}{\Theta_r}. \end{aligned}$$

So we have for partial derivatives in space

$$\nabla \cdot \vec{Q}_r = \sum_{j=1}^3 \partial_j \vec{Q}_r^j = \sum_{j=1}^3 \left( \frac{1}{\Theta_r} \partial_j \vec{F}_r^j + \sum_{k=1}^3 \partial_j \mathbf{P}_r^{jk} \frac{\vec{b}^k}{\Theta_r} \right). \quad (106)$$

Since the pressure tensor  $\mathbf{P}_r$  is symmetric, then  $\mathbf{P}_r^{jk} = \mathbf{P}_r^{kj}$ . Thus, (106) is equal to

$$\nabla \cdot \vec{Q}_r = \frac{1}{\Theta_r} \nabla \cdot F_r + \frac{\vec{b}}{\Theta_r} \cdot (\nabla \cdot \mathbf{P}_r). \quad (107)$$

Combining (97) and (105)–(107), the result of the lemma is now straightforward. In the rest of this paper (104a) is referred to as the modified moment model, as opposed to the moment model (97).

## 5. Rankine–Hugoniot relations

In this section we advocate that, in some regimes, (104) is probably physically more relevant than (97). We base the discussion on *discontinuous solutions* of (97) and (104). From the general theory of hyperbolic systems of conservations laws we already know that discontinuous solutions of (97) and (104) are different. But to what amount? If the difference tends to zero in the non-equilibrium limit, we have to conclude that the difference is non-essential in this non-equilibrium regime, and that both (97) and (104) can be used. The main result of this section is that the difference is large, and that the correct approximation is the modified moment model (104).

Let  $\vec{d}$  be the normal derivative on a line of discontinuity of the solution and let  $\sigma$  be the velocity of the line of discontinuity. In the following NE (resp. MM, mMM) stands for Non-Equilibrium (86) model (resp. Moment (97) Model, modified Moment (104) Model). The Rankine–Hugoniot relations for these models are

$$(NE) \begin{cases} -\sigma[\rho] + \vec{d} \cdot [\rho \vec{v}] = 0, \\ -\sigma[\rho \vec{v}] + \vec{d} \cdot [\rho \vec{v} \otimes \vec{v} + (p + p_r)\mathbf{I}] = 0, \\ -\sigma[\rho E + E_r] + \vec{d} \cdot [(\rho E + E_r)\vec{v} + (p + p_r)\vec{v}] = 0, \\ -\sigma[S_r] + \vec{d} \cdot [\vec{v} S_r] = 0, \end{cases} \quad (108)$$

$$(MM) \begin{cases} -\sigma[\rho] + d \cdot [\rho \vec{v}] = 0, \\ -\sigma \left[ \rho \vec{v} + \frac{\mathcal{P}}{\mathcal{C}} \vec{F}_r \right] + \vec{d} \cdot [\rho \vec{v} \otimes \vec{v} + p\mathbf{I} + \mathcal{P}\mathbf{P}_r] = 0, \\ -\sigma[\rho E + \mathcal{P}E_r] + \vec{d} \cdot [\rho E \vec{v} + p\vec{v} + \mathcal{P}\mathcal{C}\vec{F}_r] = 0, \\ -\sigma[E_r] + \vec{d} \cdot [\mathcal{C}\vec{F}_r] = 0, \\ -\sigma[\vec{F}_r] + \vec{d} \cdot [\mathcal{C}\mathbf{P}_r] = 0, \end{cases} \quad (109)$$

and

$$(\text{mMM}) \begin{cases} -\sigma[\rho] + d.[\rho\vec{v}] = 0, \\ -\sigma \left[ \rho\vec{v} + \frac{\mathcal{P}}{\mathcal{C}} \vec{F}_r \right] + \vec{d}.[\rho\vec{v} \otimes \vec{v} + p\mathbf{I} + \mathcal{P}\mathbf{P}_r] = 0, \\ -\sigma[\rho E + \mathcal{P}E_r] + \vec{d}.[\rho E\vec{v} + p\vec{v} + \mathcal{P}\mathcal{C}\vec{F}_r] = 0, \\ -\sigma[S_r] + \vec{d}.[\mathcal{C}\vec{Q}_r] = 0, \\ -\sigma[\vec{F}_r] + \vec{d}.[\mathcal{C}\mathbf{P}_r] = 0, \end{cases} \quad (110)$$

In all these expressions  $[f]$  stands for the difference of the left and right state across the discontinuity line:  $[f] = f_R - f_L$ . All these systems of Rankine–Hugoniot relations must be supplemented by entropy inequalities. We observe that

**Lemma 9** (Compatibility of the modified moment model with the non-equilibrium model). *Let us consider the Rankine–Hugoniot relations of the modified moment model (110) between a left and a right state for which the regime is  $\mathcal{P} = 1$ ,  $\mathcal{C} = \varepsilon^{-1}$ ,  $\mathcal{L}_s = \varepsilon^{-4}$ ,  $\mathcal{L} = \varepsilon^2$ : note that this regime is compatible with the assumption (85) already encountered in the study of the non-equilibrium limit. We consider a shock velocity  $\sigma = O(1)$  which means that we are interested only in shocks at moderate velocities  $O(1)$  in the lab frame. Then*

(a) *An  $O(\varepsilon)$  approximation of (110) is (108). The equation for the radiative temperature at discontinuities is*

$$-\sigma[T_r^3] + \vec{d}.[\vec{v}T_r^3] = 0. \quad (111)$$

(b) *The system (108) is not an  $O(\varepsilon)$  of (109), since the limit equation for the temperature is*

$$-\sigma[T_r^4] + \vec{d}.\left[\frac{4}{3}\vec{v}T_r^4\right] = 0. \quad (112)$$

The method we use is of course based on three Chapman–Enskog expansions: a Chapman–Enskog expansion for the left state; another for the right state: the last one for the Rankine–Hugoniot relations. In order to simplify the discussion we remark that the regime  $\mathcal{P} = 1$ ,  $\mathcal{C} = \varepsilon^{-1}$ ,  $\mathcal{L}_s = \varepsilon^{-4}$ ,  $\mathcal{L} = \varepsilon^2$  is equivalent to  $\mathcal{P} = 1$ ,  $\mathcal{C} = \varepsilon^{-1}$ ,  $\mathcal{L}_s = \varepsilon^{-2}$ ,  $\mathcal{L} = \varepsilon^1$  plus  $\sigma_s = O(\varepsilon^{-1})$  and  $\sigma_a = O(\varepsilon)$ . So we are able to reuse the analysis of the non-equilibrium diffusion model, but with elimination of the diffusion and absorption due to  $\sigma_s = O(\varepsilon^{-1})$  and  $\sigma_a = O(\varepsilon)$ . Of course, the first three equations of the non-equilibrium limit are contained in both the moment model and modified moment model so the real difficulty is the last equation, i.e.  $E_r$  (resp.  $S_r$ ) equation in the moment (resp. modified moment) model.

*Left state:* A consequence of (78) together with the hypothesis  $\sigma_s = O(\varepsilon^{-1})$  is

$$\mathbf{F}_{rL} = \varepsilon \vec{v}_L \left( \frac{4}{3} T_{rL}^4 \right) + O(\varepsilon^2). \quad (113)$$

Since we also know by a direct calculation that

$$\mathbf{F}_{rL} = -\frac{4T_{rL}^4}{3 + |\vec{b}_{L0} + \varepsilon \vec{b}_{L1} + O(\varepsilon^2)|^2} (\vec{b}_{L0} + \varepsilon \vec{b}_{L1} + O(\varepsilon^2))$$

it means that  $\vec{b}_{L0} = 0$  and  $\vec{b}_{L1} = -\vec{v}$ . It implies that  $n$  given in (98) is also  $n = \frac{1}{e^{(v/\Theta_r)(1-\varepsilon\vec{v}\cdot\vec{n}+O(\varepsilon^2))}-1}$ , that is

$$n = \frac{1}{e^{(v_0/\Theta_r)(1+O(\varepsilon^2))}-1}. \quad (114)$$

We need an expansion to the first order of the entropy flux  $\mathbf{Q}_{rL}$ . Since the non-dimensional entropy flux is

$$\vec{Q}_{rL} = - \int \int v^2 [n_L \log n_L - (n_L + 1) \log(n_L + 1)] \vec{n} \, dv \, d\vec{n},$$

then  $\vec{Q}_{rL}$  is also

$$\begin{aligned} & - \int \int (v_0 \vec{n}_0 + \varepsilon v_0 \vec{v} + o(\varepsilon)) [n_L \log n_L - (n_L + 1) \log(n_L + 1)] v_0 \, dv_0 \, d\vec{n}_0 \\ & = - \int \int v_0^2 \vec{n}_0 [n_L \log n_L - (n_L + 1) \log(n_L + 1)] \, dv_0 \, d\vec{n}_0 \\ & \quad - \varepsilon \vec{v}_L \int \int v_0^2 [n_L \log n_L - (n_L + 1) \log(n_L + 1)] \, dv_0 \, d\vec{n}_0 + O(\varepsilon^2). \end{aligned}$$

By (114)

$$- \int \int v_0^2 \vec{n}_0 [n_L \log n_L - (n_L + 1) \log(n_L + 1)] \, dv_0 \, d\vec{n}_0 = O(\varepsilon^2).$$

We also have directly that

$$\int \int v_0^2 [n_L \log n_L - (n_L + 1) \log(n_L + 1)] \, dv_0 \, d\vec{n}_0 = S_{rL} + O(\varepsilon)$$

so

$$\vec{Q}_{rL} = \varepsilon \vec{v}_L S_{rL} + O(\varepsilon). \quad (115)$$

*Right state:* Similarly  $\mathbf{F}_{rR} = \varepsilon \vec{v}_R (\frac{4}{3} T_{rR}^4) + O(\varepsilon^2)$  and  $\vec{Q}_{rR} = \varepsilon \vec{v}_R S_{rR} + O(\varepsilon)$ .

*Discussion of Rankine–Hugoniot relations:* It is now an easy matter to deduce  $O(\varepsilon)$  Rankie–Hugoniot approximations of (109) and (110). We expand the shock velocity  $\sigma = \sigma_0 + O(\varepsilon)$ . One has

$$- \sigma_0 (T_{rR}^4 - T_{rL}^4) + (\frac{4}{3} \vec{v}_{R0} T_{rR}^4 - \frac{4}{3} \vec{v}_{L0} T_{rL}^4) = 0 \quad (116)$$

and

$$- \sigma_0 (T_{rR}^3 - T_{rL}^3) + (\vec{v}_{R0} T_{rR}^3 - \vec{v}_{L0} T_{rL}^3) = 0. \quad (117)$$

It is then clear that these Rankine–Hugoniot relations (116) and (117) are different, and that (117) is a correct approximations to the Rankine–Hugoniot relations of the Non-equilibrium limit. On the other hand, (116) is not correct. The proof now ends.

**Corollary 3.** *Let us specialize the previous discussion for contact discontinuities, that is  $\sigma = \vec{d} \cdot \vec{v}_L = \vec{d} \cdot \vec{v}_R$ . Then Eq. (112) implies that the left and right radiative temperatures are the same,  $T_{rR} = T_{rL}$ . On the other hand (111) degenerates in the sense that  $T_{rR}$  and  $T_{rL}$  are arbitrary for this equation, which is one more time compatible with the non-equilibrium model.*

The proof is straightforward. This result means that the moment model does not contain classical contact discontinuities. On the other hand, the modified moment model has these classical contact discontinuity profiles where all equations degenerate.

## 6. Conclusion and numerical issues

So we have justified mathematically by means of rigorous asymptotic expansions the non-equilibrium diffusion limit, already proposed in [13]. We proved that the extra term  $p_r \nabla \cdot \vec{v}$  is a consequence of the fact that the scattering is, when isotropic, isotropic only in the comobile frame. We also prove that discontinuous solutions of standard moment model  $(E_r, \mathbf{F}_r)$  are not compatible with discontinuous solution of the non-equilibrium diffusion limit. On the other hand, discontinuous solutions of the modified moment model  $(S_r, \mathbf{F}_r)$  are compatible with discontinuous solution of the non-equilibrium diffusion limit.

Since modern numerical methods for the solutions of systems of conservation laws with source terms are based on Riemann solvers and shocks solutions, it is reasonable to think that the study we made about the modified moment model should help in the design of more accurate and robust conservative Eulerian schemes. But this needs to be confirmed more firmly, both theoretically and numerically.

A still open issue is the generalization of this work to moment models with frequency groups, relaxing the gray hypothesis.

## Appendix A. Equivalence between the form (17) and (15) for the scattering operator $S_s$

We want to show that

$$\frac{v^2}{v_0^2} \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) = \frac{v_0}{v} \left( \frac{v^3}{v_0^3} \frac{1}{4\pi} \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}' - I \right)$$

with  $v'$  defined by

$$v' = \frac{1 - \frac{\vec{n} \cdot \vec{v}}{c}}{1 - \frac{\vec{n}' \cdot \vec{v}}{c}} v.$$

One can easily realize that the question resume to show that

$$\oint I_0 d\vec{n}_0 = \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}'.$$

This can be easily verified using the fact that  $v dv d\vec{n} = v_0 dv_0 d\vec{n}_0$  and by regularizing  $\oint I_0 d\vec{n}_0$ . First consider a positive and function  $\phi(x)$  such that  $\int_{\mathbb{R}} \phi = 1$ . As a result of the theory of distribution it is well known that the sequence  $\phi^\varepsilon = \varepsilon \phi(x/\varepsilon)$  converge toward the Dirac function when  $\varepsilon \rightarrow 0$ . Thus,

$$\oint I_0 = \lim_{\varepsilon \rightarrow 0} \int I_0(\vec{v}, \vec{n}_0) \phi^\varepsilon(\vec{v}_0 - v_0) d\vec{n}_0 d\vec{v}_0.$$

Since the measure  $v dv d\vec{n}$  is invariant under Lorentz transform we have

$$\begin{aligned} & \int I_0(\bar{v}_0, \vec{n}_0) \phi^\varepsilon(\bar{v}_0 - v_0) d\vec{n}_0 d\bar{v}_0 \\ &= \int \frac{v'}{\bar{v}_0} I_0(\bar{v}_0(v', \vec{n}'), \vec{n}_0(v', \vec{n}')) \phi^\varepsilon(\bar{v}_0(v', \vec{n}') - v_0) dv' d\vec{n}' \end{aligned}$$

and using the invariance relation (13) for the radiative intensity we have

$$\int I_0(\bar{v}, \vec{n}_0) \phi^\varepsilon(\bar{v}_0 - v_0) d\vec{n}_0 d\bar{v}_0 = \int \frac{\bar{v}_0^2}{v'^2} I(v', \vec{n}') \phi^\varepsilon(\bar{v}_0(v', \vec{n}') - v_0) dv' d\vec{n}'$$

with the relation  $\bar{v}_0(v', \vec{n}') = \gamma v'(1 - \vec{n}' \cdot \vec{v}/c)$ . Thus, making now the change of variables

$$\bar{v}_0 \rightarrow v' = \frac{\bar{v}_0}{\gamma(1 - \frac{\vec{n}' \cdot \vec{v}}{c})}$$

for fixed  $\vec{n}'$  we have

$$\begin{aligned} & \int I_0(\bar{v}_0, \vec{n}_0) \phi^\varepsilon(\bar{v}_0 - v_0) d\vec{n}_0 d\bar{v}_0 \\ &= \int \left( \int \frac{1}{1 - \frac{\vec{n}' \cdot \vec{v}}{c}} \frac{\bar{v}_0^2}{v'^2} I(v'(\bar{v}_0, \vec{n}'), \vec{n}') \phi^\varepsilon(\bar{v}_0 - v_0) dv' \right) d\vec{n}'. \end{aligned}$$

Thus, using the fact that, in the weak sense,  $\phi^\varepsilon \rightarrow \delta_0$

$$\begin{aligned} \oint I_0 &= \lim_{\varepsilon \rightarrow 0} \int \left( \int \frac{1}{1 - \frac{\vec{n}' \cdot \vec{v}}{c}} \frac{\bar{v}_0^2}{v'^2} I(v'(\bar{v}_0, \vec{n}'), \vec{n}') \phi^\varepsilon(\bar{v}_0 - v_0) dv' \right) d\vec{n}' \\ &= \int \lim_{\varepsilon \rightarrow 0} \left( \int \frac{1}{1 - \frac{\vec{n}' \cdot \vec{v}}{c}} \frac{\bar{v}_0^2}{v'^2} I(v'(\bar{v}_0, \vec{n}'), \vec{n}') \phi^\varepsilon(\bar{v}_0 - v_0) dv' \right) d\vec{n}' \\ &= \int \left( 1 - \frac{\vec{n}' \cdot \vec{v}}{c} \right) I(v', \vec{n}') d\vec{n}' = \int \frac{v_0}{v'} I(v', \vec{n}') d\vec{n}' \end{aligned}$$

with  $v_0 = \gamma v'(1 - \vec{v} \cdot \vec{n}'/c)$ . Since we have also  $v_0 = \gamma v(1 - \vec{v} \cdot \vec{n}/c)$  thus

$$v' = v \frac{1 - \frac{\vec{v} \cdot \vec{n}}{c}}{1 - \frac{\vec{v}' \cdot \vec{n}'}{c}}.$$

## Appendix B. Useful formulas and details of each terms of our $P^1$ model

We give here some very useful formulas for some moments integrals of a generalized Planck function. For a generalized Planck function  $(15/4\pi^5) v^3 / \exp((v/x)(1 + \vec{y} \cdot \vec{n})) - 1$  we define  $M_1(x, \vec{y})$ ,  $\vec{M}_2(x, \vec{y})$ ,  $\mathbf{M}_3(x, \vec{y})$  as its moments against 1,  $\vec{n}$  and  $\vec{n} \otimes \vec{n}$  respectively. We recall that elementary

calculations give, see [13,17] for example:

$$M_1(x, \vec{y}) = \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} I(v, \vec{n}) d\vec{n} dv = x^4 \frac{3 + \|\vec{y}\|^2}{3(1 - \|\vec{y}\|^2)^3} \quad (B.1)$$

$$\vec{M}_2(x, \vec{y}) = \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} \vec{n} I(v, \vec{n}) d\vec{n} dv = -\frac{4x^4 \vec{y}}{3(1 - \|\vec{y}\|^2)^3} \quad (B.2)$$

by setting  $\vec{f} = \vec{M}_2(x, \vec{y})/M_1(x, \vec{y})$  one finds that

$$\begin{aligned} \vec{y} &= \left( \frac{\sqrt{4 - 3\|\vec{f}\|^2} - 2}{\|\vec{f}\|^2} \right) \vec{f}, \\ \mathbf{M}_3(x, \vec{y}) &= \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} \vec{n} \otimes \vec{n} I(v, \vec{n}) d\vec{n} dv \\ &= \left( \frac{1 - \|\vec{y}\|^2}{3 + \|\vec{y}\|^2} \mathbf{I} + \frac{3 + \|\vec{y}\|^2}{4} \vec{f} \otimes \vec{f} \right) M_1(x, \vec{y}). \end{aligned} \quad (B.3)$$

In our  $P^1$  model we suppose that

$$I = (15/4\pi^5) v^3 / \exp\left(\frac{v}{\Theta_r} + \frac{v\vec{b} \cdot \vec{n}}{\Theta_r}\right) - 1.$$

Thus, the first three moments  $E_r$ ,  $\vec{F}_r$  and  $\mathbf{P}_r$  are given by

$$E_r = M_1(\Theta_r, \vec{b}), \quad (B.4)$$

$$\vec{F}_r = \vec{M}_2(\Theta_r, \vec{b}), \quad (B.5)$$

$$\mathbf{P}_r = \mathbf{M}_3(\Theta_r, \vec{b}). \quad (B.6)$$

We detail now the expression for the entropy  $S_r$  and the associated flux of entropy  $\vec{Q}_r$ . First we deal with  $S_r$ :

$$\begin{aligned} S_r &= -\frac{15}{4\pi^5} \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} v^2 (n \log n - (n+1) \log(n+1)) dv d\vec{n} \\ &= -\frac{15K}{\pi^4} \Theta_r^3 \int_{\vec{n}} \frac{1}{(1 + \vec{b} \cdot \vec{n})^3} d\vec{n} \int_0^{+\infty} z^2 (m \log(m) - (m+1) \log(m+1)) dz \end{aligned} \quad (B.7)$$

with  $m = 1/(\exp(z) - 1)$ . Easy computations give

$$\int_{\vec{n}} \frac{1}{1 + \vec{b} \cdot \vec{n}} d\vec{n} = \frac{4\pi}{(1 - \|\vec{b}\|^2)^2}.$$

Now we define the function  $g(\alpha)$  by

$$g(\alpha) = \int v^2 (M \log M - (M+1) \log(M+1)) dv d\vec{n}$$



with  $M = (\exp(v/\alpha) - 1)^{-1}$ . Since  $M/(M + 1) = \exp(-v/\alpha)$  one has

$$g'(\alpha) = -\alpha^2 \int \frac{v^4 \exp(v)}{(\exp(v) - 1)^2} dv$$

and by integrating by parts

$$g'(\alpha) = -\alpha^2 \int 4v^3 \frac{1}{\exp(v) - 1} = -4\alpha^2 \frac{4\pi^5}{15} \frac{M_1(1, 0)}{4\pi} = -4\alpha^2 \frac{4\pi^5}{15} \frac{1}{4\pi}$$

thus

$$g(\alpha) = -\frac{4}{3} \alpha^3 \frac{4\pi^5}{15} \frac{1}{4\pi}$$

and

$$\int_0^{+\infty} z^2 (m \log(m) - (m + 1) \log(m + 1)) dz = g(1) = -\frac{4}{3} \frac{4\pi^5}{15} \frac{1}{4\pi},$$

which give finally

$$S_r = \frac{4}{3} \frac{\Theta_r^3}{(1 - \|\vec{b}\|^2)^2}.$$

Now we compute the entropy flux. We have

$$\begin{aligned} \vec{Q}_r &= -\frac{15}{4\pi^5} \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} v^2 (n \log n - (n + 1) \log(n + 1)) \vec{n} dv d\vec{n} \\ &= -\frac{15}{4\pi^5} \Theta_r^3 g(1) \int \frac{\vec{n}}{(1 + \vec{b} \cdot \vec{n})^3} d\vec{n}. \end{aligned} \quad (\text{B.8})$$

One has now to compute  $\vec{V} = \int [\vec{n}/(1 + \vec{b} \cdot \vec{n})^3] d\vec{n}$ . We show that  $\vec{V}$  is collinear to  $\vec{b}$ , that is there exist a real  $\lambda$  such that  $\vec{V} = \lambda \vec{b}$ . First, we show that for every vector  $\vec{b}^\perp$  perpendicular to  $\vec{b}$  we have  $\vec{V} \cdot \vec{b}^\perp = 0$ : up to a rotation, let us choose a reference such that  $\vec{b} = (\|\vec{b}\|, 0, 0)$  and  $\vec{b}^\perp = (0, x, y)$ . Then we have

$$\begin{aligned} \vec{V} \cdot \vec{b}^\perp &= \int \frac{\vec{n} \cdot \vec{b}^\perp}{(1 + \vec{b} \cdot \vec{n})^3} d\vec{n} \\ &= \int_{\theta \in [0, \pi], \phi \in [0, 2\pi]} \frac{x \sin \theta \cos \phi + y \sin \theta \sin \phi}{(1 + \cos \theta \|\vec{b}\|)^3} \sin \theta d\theta d\phi = 0 \end{aligned}$$

thus there exist  $\lambda$  real such that  $\vec{V} = \lambda \vec{b}$ . It remains now to compute  $\lambda$ . We have

$$\begin{aligned} \lambda \|\vec{b}\|^2 &= \int \frac{\vec{n} \cdot \vec{b}}{(1 + \vec{b} \cdot \vec{n})^3} d\vec{n} \\ &= \int_{\theta \in [0, \pi], \phi \in [0, 2\pi]} \frac{\|\vec{b}\| \cos \theta \sin \theta}{(1 + \cos \theta \|\vec{b}\|)^3} d\theta = 2\pi \int_{-1}^1 \frac{\|\vec{b}\| x}{(1 + x \|\vec{b}\|)^3} dx \end{aligned}$$

thus

$$\vec{V} = 2\pi \frac{\vec{b}}{\|\vec{b}\|} \int_{-1}^1 \frac{\|\vec{b}\| x}{(1 + x \|\vec{b}\|)^3} dx$$

and one has

$$\int_{-1}^1 \frac{\|\vec{b}\|x}{(1+x\|\vec{b}\|)^3} dx = \frac{1}{\|\vec{b}\|} \int_{-1}^1 \left( \frac{1}{(1+x\|\vec{b}\|)^2} - \frac{1}{(1+x\|\vec{b}\|)^3} \right) dx = -\frac{2\|\vec{b}\|}{(1-\|\vec{b}\|^2)^2}$$

thus

$$\vec{Q}_r = \frac{15}{4\pi^5} \Theta_r^3 g(1) 2\pi \frac{\vec{b}}{\|\vec{b}\|} \frac{2\|\vec{b}\|}{(1-\|\vec{b}\|^2)^2}$$

and since by the same type of calculus one obtain

$$S_r = \frac{15}{4\pi^5} \Theta_r^3 g(1) \frac{4\pi}{(1-\|\vec{b}\|^2)^2}$$

one has

$$\vec{Q}_r = -\vec{b} S_r.$$

We compute now the relaxation terms for the  $P^1$  model. The source terms reads

$$S_E = \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} S_a(v, \vec{n}) + S_s(v, \vec{n}) dv d\vec{n} = S_E^a + S_E^s, \quad (\text{B.9})$$

$$\vec{S}_F = \int_{v \in [0, +\infty[} \int_{\vec{n} \in S^2} \vec{n} (S_a(v, \vec{n}) + S_s(v, \vec{n})) dv d\vec{n} = S_F^a + S_F^s. \quad (\text{B.10})$$

Let us detail each of these terms by categories. For the emission–absorption effect:

$$\begin{aligned} S_E^a &= \int \int \frac{v^2}{v_0^2} B(v_0, T) - \frac{v_0}{v} I dv d\vec{n} \\ &= \frac{15}{4\pi^5} \int \int v^2 v_0 \left( \frac{1}{\exp\left(\frac{v_0}{T}\right) - 1} - \frac{1}{\exp\left(\frac{v}{\Theta_r}(1 + \vec{n} \cdot \vec{b})\right) - 1} \right) dv d\vec{n} \end{aligned}$$

and since

$$\begin{aligned} v_0 &= \gamma v (1 - \varepsilon \vec{n} \cdot \vec{v}) \\ &= \frac{15}{4\pi^5} \times \int \int \gamma v^3 (1 - \varepsilon \vec{n} \cdot \vec{v}) \left( \frac{1}{\exp\left(\frac{v}{T} \gamma (1 - \varepsilon \vec{n} \cdot \vec{v})\right) - 1} - \frac{1}{\exp\left(\frac{v}{\Theta_r}(1 + \vec{n} \cdot \vec{b})\right) - 1} \right) dv d\vec{n} \end{aligned}$$

and now using (B.4) and (B.5), we obtain

$$S_E^a = \gamma (M_1(T/\gamma, -\varepsilon \vec{v}) - \varepsilon \vec{v} \cdot \vec{M}_2(T/\gamma, -\varepsilon \vec{v}) - E_r + \varepsilon \vec{v} \cdot \vec{F}_r), \quad (\text{B.11})$$

$$S_F^a = \frac{15}{4\pi^5} \int \int \vec{n} v^2 v_0 \left( \frac{1}{\exp\left(\frac{v_0}{T}\right) - 1} - \frac{1}{\exp\left(\frac{v}{\Theta_r}(1 + \vec{n} \cdot \vec{b})\right) - 1} \right) dv d\vec{n}$$

and using (B.5) and (B.6), we obtain

$$S_F^a = \gamma (\vec{M}_2(T/\gamma, -\varepsilon \vec{v}) - \varepsilon \mathbf{M}_3(T/\gamma, -\varepsilon \vec{v}) \vec{v} - \vec{F}_r + \varepsilon \mathbf{P}_r \vec{v}). \quad (\text{B.12})$$

We consider now the relaxation term due to the scattering. We recall that the scattering term we consider can be written as

$$S_s = \frac{v^2}{v_0^2} \left( \frac{1}{4\pi} \oint I_0 - I_0 \right)$$

thus

$$S_E^s = \int S_s \, dv \, d\vec{n} = \int \frac{v^2}{v_0^2} \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) dv \, d\vec{n}.$$

and since  $v \, dv \, d\vec{n}$  is Lorentz-invariant measure,

$$S_E^s = \int S_s \, dv \, d\vec{n} = \int \frac{v}{v_0} \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) dv_0 \, d\vec{n}_0.$$

Thus, using  $v = \gamma v_0 (1 + \varepsilon \vec{v} \cdot \vec{n}_0)$  we have

$$S_E^s = \gamma \left( \int I_0 \, dv_0 \, d\vec{n}_0 - \int (1 + \varepsilon \vec{v} \cdot \vec{n}_0) I_0 \, dv_0 \, d\vec{n}_0 \right).$$

But

$$\frac{v}{\Theta_r} (1 + \vec{b} \cdot \vec{n}) = \frac{\gamma v_0 (1 + \varepsilon \vec{v} \cdot \vec{n}_0)}{\Theta_r} + \frac{\vec{b}}{\Theta_r} v \vec{n}. \quad (\text{B.13})$$

Now, for  $v \vec{n}$  we call that we have

$$\begin{aligned} v_0 \vec{n}_0 &= v \left( n - \gamma \varepsilon \vec{v} \left( 1 - \varepsilon \frac{\gamma}{\gamma + 1} \vec{n} \cdot \vec{v} \right) \right) \\ &= (-\gamma \varepsilon \vec{v}) v + \left( 1 + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) v \vec{n} \end{aligned}$$

or conversely

$$v \vec{n} = (\gamma \varepsilon \vec{v}) v_0 + \left( 1 + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) v_0 \vec{n}_0. \quad (\text{B.14})$$

Using (B.14) in the right-hand side of (B.13), expanding, and rearranging the term one obtains

$$\frac{v}{\Theta_r} (1 + \vec{b} \cdot \vec{n}) = \frac{v_0}{\Theta_{r,0}} (1 + \vec{b}_0 \cdot \vec{n}_0) \quad (\text{B.15})$$

with  $\Theta_{r,0}$  and  $\vec{b}_0$  defining by

$$\Theta_{r,0} = \frac{\Theta_r}{\gamma(1 + \varepsilon \vec{b} \cdot \vec{v})}$$

and

$$\vec{b}_0 = \frac{1}{1 + \varepsilon \vec{b} \cdot \vec{v}} \left( \varepsilon \vec{v} + \left( \frac{1}{\gamma} \mathbf{I} + \frac{\varepsilon \gamma}{\gamma + 1} \vec{v} \otimes \vec{v} \right) \vec{b} \right)$$

which give the expression of the radiative intensity in the comobile frame

$$I_0 = \frac{15}{4\pi^5} \frac{v_0^3}{\exp(\frac{v_0}{\Theta_{r,0}}(1 + \vec{b}_0 \cdot \vec{n}_0)) - 1}.$$

Now it is simple to compute the relaxation terms due to the scattering

$$\begin{aligned} S_E^s &= \gamma(M_1(\Theta_{r,0}, \vec{b}_0) - M_1(\Theta_{r,0}, \vec{b}_0) - \varepsilon \vec{v} \cdot \vec{M}_2(\Theta_{r,0}, \vec{b}_0)) \\ &= -\gamma \varepsilon \vec{v} \cdot \vec{M}_2(\Theta_{r,0}, \vec{b}_0) \end{aligned} \quad (\text{B.16})$$

and

$$\begin{aligned} S_F^s &= \int S_s \vec{n} \, d\vec{n} = \int \frac{v^2}{v_0^2} \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) \vec{n} \, d\vec{n} \\ &= \int \frac{v}{v_0} \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) \vec{n}_0 \, dv_0 \, d\vec{n}_0 \end{aligned}$$

and using one more time the relation (B.14) for  $v\vec{n}$  we obtain

$$\begin{aligned} S_F^s &= \int \left( \gamma \varepsilon \vec{v} + \left( 1 + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) \vec{n}_0 \right) \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) \vec{n} \, dv_0 \, d\vec{n}_0 \\ &= \left( \mathbf{I} + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) \int \vec{n}_0 \left( \frac{1}{4\pi} \oint I_0 - I_0 \right) \vec{n} \, dv_0 \, d\vec{n}_0. \end{aligned}$$

Finally, using (B.2) we find that

$$S_F^s = - \left( \mathbf{I} + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) \vec{M}_2(\Theta_{r,0}, \vec{b}_0). \quad (\text{B.17})$$

We are also able to compute the relaxation term for  $S_r$  we recall that we have

$$S_{S_r} = \frac{S_E + \vec{b} \cdot S_F}{\Theta_r}.$$

By example for the scattering this gives

$$S_{S_r}^s = -\frac{1}{\Theta_r} \left( \vec{M}_2(\Theta_{r,0}, \vec{b}_0) \cdot \left( \gamma \varepsilon \vec{v} + \left( \mathbf{I} + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) \vec{b} \right) \right)$$

but

$$\left( \gamma \varepsilon \vec{v} + \left( \mathbf{I} + \frac{\varepsilon \gamma^2}{\gamma + 1} \vec{v} \otimes \vec{v} \right) \vec{b} \right) = \gamma(1 + \varepsilon \vec{b} \cdot \vec{v}) \vec{b}_0$$

thus

$$S_{S_r}^s = -\frac{\gamma(1 + \varepsilon \vec{b} \cdot \vec{v})}{\Theta_r} (\vec{b}_0 \cdot \vec{M}_2(\Theta_{r,0}, \vec{b}_0)).$$

and using (B.2) and the definition of  $\Theta_{r,0}$  we have finally

$$S_{S_r}^s = \frac{4}{3} \frac{\|\vec{b}_0\|^2}{(1 - \|\vec{b}_0\|^2)^2} \Theta_{r,0}^3. \quad (\text{B.18})$$

## References

- [1] Lowrie RB, Morel JE, Hittinger JA. The coupling of radiation and hydrodynamics. *Astrophys J* 1999;521:432–50.
- [2] Lowrie RB, Morel JE. Issues with high-resolution Godunov methods for radiation hydrodynamics. *JQSRT* 2001;69(4):475–89.
- [3] Mihalas D, Mihalas BW. Foundations of radiation hydrodynamics. Oxford: Oxford University Press, 1984.
- [4] Auer LH, Mihalas D. On laboratory-frame radiation hydrodynamics. *JQSRT* 2001;71(1):61–97.
- [5] Lowrie RB, Mihalas D, Morel JE. Comoving-frame radiation transport for non-relativistic fluid velocities. *JQSRT* 2001;69(3):291–304.
- [6] Pomraning GC. The equations of radiation hydrodynamics. Oxford: Pergamon Press, 1973.
- [7] Castor JI. Lectures on Radiation Hydrodynamics. UCRL-JC-134209.
- [8] Munier A, Weaver R. *Comput Phys Rep* 1986;3:125–64.
- [9] Tassart J. Transfert de Rayonnement. In: Dautray R, Watteau JP, editors. *La Fusion Thermonucléaire inertielle par laser*. Eyrolles; 1993.
- [10] Bardos C, Golse F, Perthame B. The radiative transfer equations: existence of solutions and diffusion approximation under accretivity assumptions—a survey. *Transp Theory Stat Phys* 1987;16:637–52.
- [11] Mihalas D. Computational methods for astrophysical fluid flow, Saas-Fee advanced course, lectures notes. Berlin: Springer, 1997.
- [12] Lifshitz LD, Landau E. *Mécanique des fluides*. Paris: MIR.
- [13] Feugeas JL, Dubroca B. Entropy moment closure hierarchy for the radiative transfer equation. Private communication.
- [14] Levermore CD. Relating Eddington factors to flux limiters. *JQSRT* 1984;31(2):149–60.
- [15] Buet C, Cordier S. in preparation.
- [16] Olson GL, Auer LH, Hall ML. Diffusion, P1, and other approximate forms of radiation transport. Los Alamos report LA-UR-99-471.
- [17] Müller I, Ruggeri T. Rational extended thermodynamics. Springer Tracts in Natural Philosophy, vol. 37, 2nd ed. New York, NY: Springer, 1998.
- [18] Minerbo GN. Maximum entropy Eddington factors. *JQSRT* 1978;20:541–5.
- [19] Després B. Hyperbolic systems of conservation laws with equality or convex constraints and entropy. Rapport du laboratoire d'analyse numérique, R 00026, 2000.

## Further reading

- Levermore CD. Moment closure hierarchies for kinetic theories. *J Stat Phys* 1996;83(5–6):1021–65.
- Sentis R. Sur les équations de diffusion multigroupe en transfert radiatif. Note CEA 2603, 1989.

## A Conservative and Entropy Scheme for a Simplified Model of Granular Media

C. Buet,<sup>1,\*</sup> S. Cordier,<sup>2</sup> and V. Dos Santos<sup>3</sup>

<sup>1</sup>Departement des Sciences de la Simulation et de l'Information, CEA,  
Bruyères le Chatel, France

<sup>2</sup>Laboratoire MAPMO, UMR 6628, Université d'Orléans, Orléans,  
France

<sup>3</sup>Laboratoire Vision et Robotique, Université d'Orléans, IUT de Bourges,  
Bourges, France

### ABSTRACT

In this paper, we present a numerical scheme for a non linear Fokker–Planck equation of one-dimensional granular medium. We consider a kinetic description of a system of particles undergoing nearly elastic particles and interacting with a thermal bath. We construct a numerical method which preserve all the properties of the continuous model, conservation laws, and decay of the entropy. Moreover, the discretization is such that, on a fixed grid, we deal with arbitrary small temperatures for the bath. Explicit and implicit time discretization are analyzed.

---

\*Correspondence: C. Buet, Departement des Sciences de la Simulation et de l'Information, CEA, 91680 Bruyères le Chatel, France; E-mail: christophe.buet@cea.fr.



*Key Words:* Entropy scheme; Granular media; Discretization; Kinetic models; Inelastic collisions; Fokker–Planck equation.

## 1. INTRODUCTION

The model we consider has been proposed by MacNamara and Young (1992; 1993), and has been studied from mathematical point of view (Benedetto et al. 1997; 1998; 1999). We also refer to Baldassarri et al. (2002), Esteban and Perthame (1991) for kinetic modelisations of inelastic collisions. The model is derived from a system of particles moving in one dimension and that undergo inelastic collision. The unknown of the kinetic model is the distribution function that represents the number of particles with velocity  $v \in \mathbb{R}$ , at time  $t$ . The collisions are modeled by a Boltzmann type collision operator, for inelastic collision with an hard sphere cross section. The collisions being inelastic, they lead to a decay of energy and the distribution function concentrates for large time on zero velocity. When there is a lot ( $N \gg 1$ ) of such collision, weakly inelastic, the Boltzmann equation reduces to a Fokker–Planck type equation, (MacNamara and Young, 1994; Toscani, 2000).

In this work, we consider such a system of particles immersed in a thermal bath at constant temperature  $\sigma$ , which counterbalances the loss of energy due to inelastic collisions. When  $\sigma = 0$ , the distribution function converges to a Dirac mass at origin in the sense of weak convergence of measure (Benedetto et al., 1997). In the case  $\sigma > 0$ , the equilibrium states behave like  $\exp(-|v|^3)$  for large velocities (Benedetto et al., 1998), the equilibrium state are not Maxwellian.

Let us also mention that some works has been devoted to the hydrodynamical limit of this system: one obtains a system of two conservation laws for the density  $\rho$  and the momentum  $\rho u$ , that can be written as the Euler system for isotropic gases, with a pressure law  $P(\rho) = \rho^\gamma$ ,  $\gamma = 1/3$  (the exponent  $1/3$  yields to mathematical difficulties, since the obtained system lies in between classical gas dynamics,  $\gamma > 1$ , and pressure less gas,  $\gamma = 0$ , that have very different behavior).

The aim of this paper is to present a deterministic discretization, that is, compatible with the known properties of the operator. Moreover, the method is such that we can treat any temperature of the bath with the same grid, in other words, as  $\sigma = 0$  goes to zero the scheme degenerate to a conservative and entropy discretization of the pure granular case. The method is close to the one developed for the Fokker–Planck–Landau equation or the Kompaneets equation (Cf. Buet and Cordier, 1997; 1998a; 2003). For generalized discrete models of kinetic equations, we refer to (Görsch, 2002). Let us emphasize that the proposed method has a linear computational cost. Explicit





and Implicit time discretization are detailed. Implicit schemes appear to be more efficient.

Some numerical results are presented. In particular, the non-Maxwellian equilibrium states are obtained.

## 2. THE ONE DIMENSIONAL MODEL OF GRANULAR MEDIA

We consider a simplified 1D model of granular media described by example in Benedetto et al. (1998). The particles are described by their distribution function  $f(v, t)$  of velocity  $v$  and time  $t$ . This function obeys

$$\partial_t f = \partial_v (Ff + \sigma \partial_v f) \quad (1)$$

where  $Ff$  is the pure granular term,

$$F(v) = \int_R |v - v'| (v - v') f(v') dv' \quad (2)$$

and  $\sigma$  is an arbitrary positive constant related to the temperature of the bath. All the more,  $\sigma \partial_v^2 f$  represents the thermal reservoir, where  $\sigma$  is linked to the temperature. Let us define  $\rho$  and  $u_f$  as the mass and the mean velocity, respectively

$$\rho = \int_R f(v') dv', \quad u_f = \frac{1}{\rho} \int_R f(v') v' dv' \quad (3)$$

### 2.1. Properties of the Continuous Model

The properties of this model are the conservation of mass and momentum, the decay of the energy for  $(\sigma = 0)$  and of the entropy:

#### Definition 2.1

Let us define the temperature and the entropy by, respectively,

$$T(f) = \int_{v'} |v - u_f|^2 f(v) dv \quad (4)$$

$$\mathcal{E} = \int \sigma [\ln(f(v))] f(v) dv + \frac{1}{6} \int_v \int_{v'} |v - v'|^3 f(v') f(v) dv' dv \quad (5)$$

and let be  $E_\alpha(f)$  and  $H_\alpha(v)$ , defined by

$$E_\alpha(f) = \iint |v - v'|^\alpha f(v') f(v) dv' dv \quad (6)$$

$$H_{f,\alpha}(v) = \int_{v'} |v - v'|^\alpha f(v') dv' \quad (7)$$





We shall omit the dependency of  $H$  with respect to  $f$  when there are no ambiguity.

**Properties 2.1.1**

Then, we get

$$\partial_f E_3(f) = 2 \int H_3(f, v) dv \quad (8)$$

$$\partial_v F = 2 \int |v - v'| f(v') dv' = 2H_1(v) \quad (9)$$

$$3\partial_v H_3 = F \quad (10)$$

These properties can be easily checked, by splitting the integrals into  $v' > v$  and  $v > v'$ .

The existence and uniqueness of an equilibrium state are given by Benedetto et al. (1998). The strict convexity of  $\mathcal{E}$ , follows from the following result that will be also useful to prove the entropy decay for an implicit scheme:

**Lemma 2.2**

For any function  $f$  in

$$\tilde{\mathcal{P}} = \{f: (1 + |v|^4)f \in L_1(R), \int f = 0 \text{ and } \int v f(v) dv = 0\}$$

$$E_3(f) \geq 0.$$

*Proof.*

Let  $f \in \mathcal{C}_0^\infty$ , and integrable. We have:

$$H_3(f, v) = \int f(v') |v - v'|^3 dv' = - \int f(v - v') |v'|^3 dv'$$

and, differentiating with respect to  $v$

$$\partial_v H_3(f, v) = 3 \int f(v') |v - v'| (v - v') dv' = - \int \partial_v f(v - v') |v'|^3 dv'$$

which gives the following identity

$$\partial_v^2 H_3(f, v) = 6H_3(f, v) = -H_3(\partial_v^2 f, v)$$

Then,

$$\begin{aligned} E_1(f) &= \int f(v) H_1(f, v) dv = \frac{1}{6} \int f(v) \partial_v^2 H_3(f, v) dv \\ &= -\frac{1}{6} \int \partial_v f(v) \partial_v H_3(f, v) dv = -\frac{1}{6} E_3(\partial_v f(v)) \end{aligned}$$





Moreover,  $\partial_v^2 H_1(f, v) = 2f(v)$ ,  $E_1(\partial_v f) = - \int f^2 dv < 0$ , and then

$$E_1(\partial_v f) = -\frac{1}{6} \int \partial_v^2 f H_3(\partial_v^2 f) dv = -\frac{1}{6} E_3(\partial_v^2 f)$$

so we have  $E_3(\partial_v^2 f) = \int f^2(v) dv > 0$  and  $\partial_v^2 f \in \phi$ .

Since for any  $g \in C^\infty$ , with compact support and such that  $\int g = 0$  and  $\int v g(v) dv = 0$ , we can find a function  $f \in C^\infty$ , integrable and such that  $g = \partial_v^2 f$  thus,  $E_3(g) = 6 \int f^2(v) dv$ . By density, the result hold for any function of  $\phi$ . ■

For sake of simplicity, we omit the index and set  $E = E_3$  and  $H = H_3$  in the reminder.

## 2.2. Pure Granular

We suppose  $\sigma = 0$  in (1), and let  $f = f(t, v)$  be a function so that:

$$\partial_t f = \partial_v(Ff) \quad (11)$$

By the definition of  $F$ , we have

$$\int_v Ff dv = 0 \quad (12)$$

We can write the weak formulation:

### Proposition 2.3

Let  $f$  be a solution of Eq. (11). Then,  $f$  satisfies for all test functions  $\phi$ :

$$\begin{aligned} \int \partial_t f \phi &= - \int \partial_v \phi Ff = - \int \int \partial_v \phi |v - v'| (v - v') f(v) f(v') dv dv' \\ &= -\frac{1}{2} \int \int (\partial_v \phi - \partial_{v'} \phi') |v - v'| (v - v') f(v) f(v') dv dv' \end{aligned} \quad (13)$$

by symmetry of  $v$  and  $v'$ . We can obtain also the weak form

$$\int \partial_t f \phi = -\frac{1}{2\rho} \int (\partial_v(\phi) - \partial_{v'}(\phi')) (F - F') f f' dv dv' \quad (14)$$

These equations verify mass and momentum conservation and entropy decay.



*Proof.*

Taking respectively  $\phi = 1, v, v^2$  in Eq. (13), we get respectively mass and momentum conservation and the temperature decay. For  $\phi = v^2$  and  $\rho = 1$

$$\begin{aligned}\partial_t \int f \phi &= -\frac{1}{2} \iint (2v - 2v') |v - v'| f(v) f(v') dv dv' \\ &= - \iint |v - v'| (v - v')^2 f(v) f(v') dv dv' = -E(f)\end{aligned}$$

So  $d_t \int v^2 f \leq 0$  implies the temperature decay, and the equilibrium corresponds to  $E = 0$ .

Moreover, using  $3\partial_t H = F$  and  $d_t E = 2 \int H$ , we obtain for  $\phi = 3H(v)$ :

$$\begin{aligned}d_t E &= 3 \int \partial_t f H \\ &= -\frac{1}{2} \iint (F(v) - F(v')) |v - v'| (v - v') f(v) f(v') dv dv'\end{aligned}\quad (15)$$

Note that  $F$  is strictly increasing:

$$\partial_v F = \int 2|v - v'| f(v') dv' > 0$$

Therefore,  $(F(v) - F(v'))(v - v') \geq 0$  and  $d_t E \leq 0$ . Thus we prove the decay of entropy. If we suppose that there is conservation, i.e.,  $d_t E = 0$ , as  $F$  is strictly increasing, we get  $f(v) = 0$  and by the conservation of mass,  $f(v)$  is of the form  $\rho \cdot \delta(v - u)$ . ■

### 2.3. Case of the Granular Immersed in a Thermal Bath

Let us now consider the case  $\sigma \neq 0$  in Eq. (1). Let  $M$  define as follows

$$\partial_v M = \frac{-F}{\sigma} M \quad (16)$$

i.e.,

$$M(v) = C \exp\left(\frac{-1}{3\sigma} H(v)\right) \quad (17)$$

We rewrite the Eq. (1)

$$\partial_t f = \partial_v \left( \sigma M \partial_v \left( \frac{f}{M} \right) \right) = \frac{\sigma}{\rho} \partial_v \left( \int_{v'} f M \partial_v \left( \frac{f}{M} \right) - f M' \partial_{v'} \left( \frac{f}{M} \right)' dv' \right)$$





where  $\rho = \int f(v) dv$  denotes the mass and because the second term is zero by integrating by part and using Eqs. (16) and (12):

$$\int M' \partial_v \left( \frac{f}{M} \right)' dv' = - \int \partial_v M' \left( \frac{f}{M} \right)' dv' = - \frac{1}{6\sigma} \int F(v') f(v') dv' = 0$$

Then, we can write the following weak symmetrized form of Eq. (1):

$$\begin{aligned} \partial_t \int f \phi dv = & \frac{-\sigma}{2\rho} \int (\partial_v \phi - \partial_{v'} \phi') f f' \\ & \times \left( \partial_v \log \left( \frac{f}{M} \right) - \partial_{v'} \log \left( \frac{f}{M} \right)' \right) dv dv' \end{aligned} \quad (18)$$

The mass and momentum conservation, as the decay of entropy, are obvious, by choosing  $\phi = 1$ ,  $v$ , and  $\log(f/M)$ , respectively.

Note that the above formulation has a structure close to the Fokker–Planck–Landau one's, in the so called log form, that has been studied by Benedetto and Caglioti (1999) and Buet and Cordier (1997). The only terms in the above expression that can be modified without changing the properties of the operator (conservation of mass, momentum, and the decay of entropy) is the product  $ff'$ .

### 3. DISCRETIZATION

In this section, we introduce a discrete version of Eq. (1), that degenerates correctly to the pure granular equation, when  $\sigma \rightarrow 0$ , i.e., “an asymptotic preserving scheme.” Moreover, this scheme must conserve the properties of conservation of the mass, and momentum, as the decay of entropy.

We first deal with the pure granular case, which brings up the following problem: find a scheme that preserve the decay of entropy.

Initially designed for the Fokker–Planck linear equations, the method of Chang–Cooper (cf. Chang and Cooper, 1970) allowed us to build such a discretization.

For all the different schemes, we consider an uniform grid in velocity:

$$v_i = i\Delta v$$

for  $i = 1, N$ . The macroscopic quantities are defined using standard quadrature formula: for the mass

$$\rho = \sum_i f_i \Delta v$$



the momentum

$$\rho u_f = \sum_i f_i v_i \Delta v$$

and the temperature

$$T = \sum_i f_i (v - u_f)^2 \Delta v$$

### 3.1. Case of Pure Granular: $\sigma = 0$

This section is devoted to the limiting case  $\sigma = 0$ . We present two different methods.

#### 3.1.1. First Method

Let us first consider the more natural discretized version of the granular equation (11). The term  $F$  is discretized using a standard quadrature formula

$$F_i = \sum_j |v_i - v_j| (v_i - v_j) f(v_j) \Delta v \quad (19)$$

This satisfies

$$\sum_i F_i f_i = 0 \quad (20)$$

by symmetry and  $F_{i+1} - F_i \geq (\Delta v)^2 \rho$  and thus it is strictly increasing. Indeed, as the grid is uniform, one gets:

$$v_{i+1} - v_i = \Delta v, \quad |v_{i+1} - v_j| > |v_i - v_j| - \Delta v$$

thus

$$\begin{aligned} F_{i+1} - F_i &\geq \Delta v \left( \sum_j |v_{i+1} - v_j| f_j \Delta v \right) - \Delta v \sum_j (v_i - v_j) f_j \Delta v \\ &\geq \Delta v \sum_j (|v_{i+1} - v_j| - (v_i - v_j)) f_j \Delta v = \Delta v \sum_j K_{i,j} f_j \Delta v \end{aligned}$$

1. if  $v_{i+1} > v_j$  then  $K_{i,j} = \Delta v > 0$ ,
2. If  $v_{i+1} < v_j$  then  $K_{i,j} = 2v_j - v_{i+1} - v_i = 2(v_j - v_{i+1}) + \Delta v > \Delta v$

So we get the inequality  $F_{i+1} - F_i \geq (\Delta v)^2 \rho$ .





We consider the following upwind scheme for this transport equation (in the velocity variable):

$$\partial_t \sum_i f_i \phi_i = -\frac{1}{\Delta v} \sum_i ((\phi_{i+1} - \phi_i) F_i^- + (\phi_i - \phi_{i-1}) F_i^+) f_i \quad (21)$$

Let us mention that the  $N$  terms  $F_i$  can be computed in  $O(N)$  operations using a splitting of the sum into the  $j$  before and after  $v_i$  that reduce the complexity of their evaluation. Note also that there exists some index  $i_0$  such that  $F_i < 0$  for all  $i < i_0$  and  $F_i > 0$  for all  $i > i_0$  (this index can move in time). Moreover, we have  $F_1 < 0$  and  $F_N > 0$  and thus, no boundary condition are needed: we do not have to prescribe the value of  $f_0$  of  $f_{N+1}$ .

The mass and momentum are conserved taking  $\phi_i = 1$ ,  $v_i$  in Eq. (21) and using the condition (20). The evolution of the temperature is

$$\frac{dT}{dt} = -6E - (\Delta v) \sum_i (F_i^- - F_i^+) f_i$$

The second term of the right hand side is positive and one cannot conclude about the decay of the temperature. For the discrete entropy,  $E$ , the same conclusion holds.

However, discrete steady state for this scheme exists: consider state such that  $i_0$  with  $f_i = 0$  for all  $i > i_0 + 1$  or  $i < i_0$ . The values of the distribution function on the two non vanishing points and the value of  $i_0$  are determined from mass and momentum conservation:

$$f_{i_0+1} + f_{i_0} = \rho, \quad v_{i_0+1} f_{i_0+1} + v_{i_0} f_{i_0} = \rho u_f \quad (22)$$

### 3.1.2. Second Method

We define the discrete analog of  $H(v)$  and the entropy  $E$  (5) by

$$H_i = \frac{1}{3} \sum_j |v_i - v_j|^3 f_j \Delta v$$

and

$$E = \frac{1}{2} \sum_i H_i f_i \Delta v$$

A second discretization can be obtained in the pure granular equation. By integration by parts, we get:

$$\int \phi \partial_v (Ff) dv = - \int \int \partial_v \phi(v) |v - v'| (v - v') f(v) f(v') dv' dv$$





The analogous discrete weak formulation is

$$\partial_t \sum_i f_i \phi_i \Delta v = -(\Delta v)^2 \sum_{i,j} D\phi_i^j f_i f_j |v_i - v_j| (v_i - v_j) \quad (23)$$

where the finite difference operator at point  $i$  is uncentered in a direction depending on the point  $j$  as

$$D\phi_i^j = \frac{\phi_{i+1} - \phi_i}{\Delta v} 1_{i < j} + \frac{\phi_i - \phi_{i-1}}{\Delta v} 1_{i > j} \quad (24)$$

This particular choice can be physically interpreted going back to a Boltzmann quasi-elastic monodimensional model.

First, a symmetry can be output of Eq. (23), to obtain

$$\partial_t \sum_i f_i \phi_i \Delta v = -\frac{1}{2} (\Delta v)^2 \sum_{i,j} (D\phi_i^j - D\phi_j^i) f_i f_j |v_i - v_j| (v_i - v_j) \quad (25)$$

Let us consider two particles with velocity  $v_i$  and  $v_j$  (with  $i < j$  for example). The only elastic collision between these two particles consist in swapping the both, but this does not change the distribution function. If one allows slightly non elastic collisions but preserving the momentum, i.e., the postcollisional velocities have the same average (the particle have the same mass) as before the collision. Therefore, the less inelastic post-collisional velocities are  $(v_{i+1}, v_{j-1})$  and  $(v_{i-1}, v_{j+1})$ .

To decrease the energy of the system is equivalent to decrease the relative velocity, so the velocities must get closer. The latter corresponds to a increasing of energy and is not physically relevant. The choice which corresponds to a minimal decrease of energy, is thus  $(v_i, v_j) \mapsto (v_{i+1}, v_{j-1})$  (for  $i < j$ ).

### Proposition 3.1

The scheme (23) and (24) preserves mass, and momentum, and the decay of energy and of entropy  $E = \frac{1}{2} \sum_i H_i f_i dv$  with  $H_i = \frac{1}{3} \sum_j |v_i - v_j|^3 f_j \Delta v$ .

*Proof.*

Standing  $(\phi = 1)$  in Eq. (24), one gets mass conservation.





For momentum, let  $(\phi_i = v_i)$ , (24) write as  $1_{i < j} + 1_{i > j}$ , and (23) is null. Using Eq. (25) with  $(\phi_i = v_i^2)$  one gives:

$$\begin{aligned} 2 \frac{dT}{dt} = & -(\Delta v)^2 \sum_{i < j} (v_{i+1}^2 + v_{j-1}^2 - v_i^2 - v_j^2) f_i f_j |v_i - v_j| (v_i - v_j) \\ & + (\Delta v)^2 \sum_{i > j} (v_{i-1}^2 + v_{j+1}^2 - v_i^2 - v_j^2) f_i f_j |v_i - v_j| (v_i - v_j) \leq 0 \end{aligned} \quad (26)$$

For the entropy, one obtains:

$$\begin{aligned} 2 \frac{d}{dt} E = & -(\Delta v)^2 \sum_{i < j} (H_{i+1} - H_i - (H_j - H_{j-1})) f_i f_j |v_i - v_j| (v_i - v_j) \\ & + (\Delta v)^2 \sum_{i > j} (H_{i-1} + H_{j+1} - H_i - H_j) f_i f_j |v_i - v_j| (v_i - v_j) \end{aligned}$$

Using the convexity of  $w \rightarrow |w - a|^3$ , one find that the sequence  $H_{i+1} - H_i$  is increasing in  $i$  or equivalently in  $v_i$  and then

$$\frac{d}{dt} E \leq 0 \quad \blacksquare$$

### Lemma 3.2

The system (23) has a global positive solution, for any positive initial data  $(f_i^0 > 0, \text{ for all } i)$ .

*Proof.*

The existence of a solution for this semi-discretized system can be easily obtained using a Cauchy–Lipschitz theorem for small time. This solution is global in time using a minoration of the solution by retaining only the loss term and the mass conservation (that provides an upperbound) Indeed, if  $f$  is solution, the corresponding discrete velocities system for each  $i$  is

$$\begin{aligned} \partial_t f_i = & (\Delta v) [f_{i-1} \sum_{j < i} f_j (v_{i-1} - v_j)^2 + f_{i+1} \sum_{j > i} f_j (v_{i+1} - v_j)^2] \\ & - (\Delta v) f_i \sum_j f_j (v_i - v_j)^2 \end{aligned}$$

Moreover,

$$(\Delta v) \sum_j f_j (v_i - v_j)^2 < K$$







where  $K$  is a constant, which depends of the length of the domain, and of the conserved quantities (mass and momentum). Thus, with  $\sum_i f_i = \rho = \text{constant}$ , we get

$$\rho \geq f_i(t) \geq f_i^0 \exp(-Kt)$$

Therefore, the maximal solution is global in time. ■

Equilibrium state for this scheme:

### Lemma 3.3

$dT/dt = 0$  if and only if  $f_i = 0$  for  $i < i_0$  and  $i > i_0 + 1$ ,  $i_0$  and  $f_{i_0}, f_{i_0+1}$  are determined using mass and momentum conservation.

*Proof.*

We start from Eq. (26). The converse is obviously true. For the direct implication, let  $f$  be such that the right hand side of Eq. (26) is null, but with a fixed mass and momentum. Since the mass is not null, they necessary exist an index  $i_0$  such that  $f_{i_0} > 0$ . Since in the r.h.s. of Eq. (26), all the terms in factor of  $f_i f_j$  are non positive except for  $j = i - 1$  or  $j = i + 1$ , this leads to  $f_i = 0$  for  $i < i_0 - 1$  and  $i > i_0 + 1$ . Now, since the term in factor of  $f_{i_0-1} f_{i_0+1}$  is also negative thus one of  $f_{i_0-1}$  and  $f_{i_0+1}$  is null. Equation (22) determines  $i_0$  and the value of  $f_{i_0} f_{i_0+1}$ . ■

If one compares the two schemes proposed here, the first one is uncentered at a “macroscopic level” that depends only on the integrated value of  $F$  whereas, the second consists in a “microscopic” uncentered scheme where only the physically relevant collision are allowed.

### 3.2. Chang–Cooper Method for a Fokker–Planck Linear Equation

Before using the Chang–Cooper method for the granular media, we briefly recall this method on a more simple linear Fokker–Planck equation: set  $F = v$  in Eq. (1),

$$\partial_t f = \partial_v(vf) + \sigma \partial_v^2 f \quad (27)$$

that can be put under the form

$$\partial_t f = \sigma \partial_v \left( M \partial_v \left( \frac{f}{M} \right) \right)$$

where  $M$  is a Maxwellian, and  $M(v) = \exp(-|v|^2/2\sigma)$





Originally, the method of Chang–Cooper was designed to preserve equilibrium state of Fokker–Planck equation. Another interesting feature of this method is its ability to degenerate correctly to an upwind scheme, for the equation of convection, when  $\sigma \rightarrow 0$ , with velocity grid fixed. Moreover, it is an entropy decaying method and not just an equilibrium state preserving method.

We set  $M_i = M(v_i)$ . This method can be written as

$$\partial_t f_i = \frac{F_{i+1/2} - F_{i-1/2}}{\Delta v}$$

with  $F_{i+1/2} = \sigma/(\Delta v) \tilde{M}_{i+1/2} (f_{i+1}/M_{i+1} - f_i/M_i)$ , where

$$\tilde{M}_{i+1/2} = \frac{M_i M_{i+1}}{M_{i+1} - M_i} (\ln M_{i+1} - \ln M_i)$$

thus

$$F_{i+1/2} = \frac{\sigma}{\Delta v} (f_{i+1} - f_i) + \frac{\sigma}{\Delta v} \left( -1 + \frac{\tilde{M}_{i+1/2}}{M_{i+1}} \right) f_{i+1} + \frac{\sigma}{\Delta v} \left( 1 - \frac{\tilde{M}_{i+1/2}}{M_i} \right) f_i$$

Analyze each term:

$$\begin{aligned} \frac{\sigma}{\Delta v} \left( 1 - \frac{\tilde{M}_{i+1/2}}{M_i} \right) &= \frac{\sigma}{\Delta v} \left( 1 - \frac{M_{i+1}}{M_{i+1} - M_i} (\ln M_{i+1} - \ln M_i) \right) \\ &= \frac{\sigma}{\Delta v} \ln \frac{M_{i+1}}{M_i} \left( \frac{1}{\ln M_{i+1} - \ln M_i} - \frac{M_{i+1}}{M_{i+1} - M_i} \right) \\ &= -\frac{\sigma}{\Delta v} \ln \frac{M_{i+1}}{M_i} \left( \frac{1}{\ln M_i/M_{i+1}} - \frac{1}{M_i/M_{i+1} - 1} \right) \end{aligned}$$

Setting  $w = \ln(M_i/M_{i+1})$ , then we get:

$$\frac{\sigma}{\Delta v} \left( 1 - \frac{\tilde{M}_{i+1/2}}{M_i} \right) = \frac{\sigma}{\Delta v} w \left( \frac{1}{w} - \frac{1}{\exp w - 1} \right) = \frac{\sigma}{\Delta v} w h(w) = \frac{\sigma}{\Delta v} w \theta$$

where  $\theta = h(w)$  and  $h$  is the function

$$h(x) = \frac{1}{x} - \frac{1}{e^x - 1}$$

Now,  $h$  is positive on  $\mathbb{R}$ , decreasing and varies between 0 and 1. Similarly, the second term reads:

$$\frac{\sigma}{\Delta v} \left( -1 + \frac{\tilde{M}_{i+1/2}}{M_{i+1}} \right) = \frac{\sigma}{\Delta v} w h(-w)$$



and  $h(-w) = -h(w) + 1$ , thus

$$\frac{\sigma}{\Delta v} \left( -1 + \frac{\tilde{M}_{i+1/2}}{M_{i+1}} \right) = -\frac{\sigma}{\Delta v} w(\theta - 1)$$

Then

$$\begin{aligned} F_{i+1/2} &= \frac{\sigma}{\Delta v} (f_{i+1} - f_i) + \frac{\sigma}{\Delta v} (-w(\theta - 1)) f_{i+1} + \frac{\sigma}{\Delta v} (w\theta) f_i \\ &= \frac{\sigma}{\Delta v} (f_{i+1} - f_i) + \frac{\sigma w}{\Delta v} (\theta f_i + (1 - \theta) f_{i+1}) \end{aligned}$$

We get an “upwind” scheme, mixed with a “ $\theta$ -scheme:” this is the Chang–Cooper method.

It is easy to verify that this discretization correspond to the discretization of the weak formulation of the FPL equation

$$\partial_t f_i \phi_i = \sum_i \sigma(D\phi)_{i+1/2} \left( M_{i+1/2} D \left( \frac{f}{M} \right)_{i+1/2} \right) \quad (28)$$

where  $Dg$  stands for the centered finite difference, i.e.,  $D(g)_{i+1/2} = (g_{i+1} - g_i)/\Delta v$  and the coefficients  $M_{i+1/2}$  is an average of the value between  $M_i$  and  $M_{i+1}$  to be defined. Such type of scheme is by construction entropy decaying and provide the good equilibrium state, see Benedetto and Caglioti (1999). By taking  $M_{i+1/2}$  as

$$M_{i+1/2} = \frac{M_i M_{i+1}}{M_{i+1} - M_i} \log \left( \frac{M_{i+1}}{M_i} \right) \quad (29)$$

this gives the Chang–Cooper scheme.

This scheme degenerates toward an upwind scheme for  $\partial_t f = \partial_v(vf)$  when  $\sigma \rightarrow 0$ . Indeed, the scheme (28) with the choice (29) reads:

$$\sum_i \partial_t f_i \phi_i = \sum_i 2\sigma(D^*\phi)_{i+1/2} \frac{\log(M_{i+1}/M_i)}{M_{i+1} - M_i} (f_{i+1} M_i - M_{i+1} f_i)$$

using

$$\sigma \log \left( \frac{M_{i+1}}{M_i} \right) = \frac{-v_{i+1}^2 + v_i^2}{2} = -\Delta v v_{i+1/2}$$

with  $v_{i+1/2} = (v_{i+1} + v_i)/2$ . Moreover, for  $i$  such that  $v_i > 0$

$$\frac{M_i}{M_{i+1} - M_i} \rightarrow -1$$





as  $\sigma \rightarrow 0$  whereas it tends to 0 if  $v_i < 0$ . Thus the limit of the scheme as  $\sigma \rightarrow 0$  is

$$\begin{aligned}\partial_t f_i &= v_{i+1/2}^+(f_{i+1} - f_i) + v_{i-1/2}^-(f_i - f_{i-1}) \\ \partial_t f_i &= 2v_{i+1/2}(f_{i+1} - f_i)1_{v_i > 0} + 2v_{i-1/2}(f_i - f_{i-1})1_{v_i < 0}\end{aligned}$$

denotes the positive/negative part of  $x$ , i.e.  $x^\pm = (x \pm |x|)/2$ .

### 3.3. Discretization with $\sigma \neq 0$

Taking the granular Eq. (1), we look for a discretization similar to the FPL one's, based on the Chang–Cooper method. Moreover, this scheme must preserve the properties of conservation and of decay, and degenerate to the equation of pure granular, when  $\sigma \rightarrow 0$ .

#### 3.3.1. A Discretization Based on the Symmetric Form

We define the discrete analog of  $H(v)$  and the entropy  $\mathcal{E}$  (5) by

$$H_i = \frac{1}{3} \sum_j |v_i - v_j|^3 f_j \Delta v$$

and

$$\mathcal{E} = \sigma \sum_i f_i \ln(f_i) \Delta v + \frac{1}{2} \sum_i H_i f_i \Delta v$$

We define also a discrete version of (17) by

$$M_i = \exp\left(-\frac{H_i}{\sigma}\right)$$

From the  $H_i$ s we define discrete of  $F$  by

$$F_{i+1/2} = \frac{H_{i+1} - H_i}{\Delta v}$$

and using the convexity of the function  $w \rightarrow |w - a|^3$  one can verify easily that  $F_{i+1/2} \geq F_{i-1/2}$ , thus the sequence  $F_{i+1/2}$  is increasing in  $i$ .

One can remark that by splitting the sum in two parts  $\sum_{j \geq i}$  and  $\sum_{j \leq i}$  than the  $H_i$ 's can be evaluated in  $O(n)$  operations as explained by Buct and Cordier (1998b), (2002) for a similar equation, the isotropic Fokker–Planck–Landau equation.



In order to preserve mass momentum and the decay of the entropy we discretize the weak-symmetrized form (18),

$$\partial_t \int f \phi dv = \frac{-\sigma}{2\rho} \iint (\partial_v \phi - \partial_{v'} \phi') f f' \left( \partial_v \log \left( \frac{f}{M} \right) - \partial_{v'} \log \left( \frac{f}{M} \right)' \right) dv dv'$$

as follows

$$\begin{aligned} & \frac{-\sigma}{2\rho} \sum_i \sum_j (D\phi_{i+1/2} - D\phi_{j+1/2}) f_{i+1/2} f_{j+1/2} \\ & \times \left( D \log \left( \frac{f}{M} \right)_{i+1/2} - D \log \left( \frac{f}{M} \right)_{j+1/2} \right) \end{aligned} \quad (30)$$

where  $Dg_{i+1/2} = (g_{i+1} - g_i)/\Delta v$  and the averaged value of the product  $f_{i+1/2} f_{j+1/2}$  has to be defined.

On this form conservation of mass and mean velocity and the decay of the discrete entropy  $\mathcal{E}$  are easily verified choosing  $\phi = 1, v, H$ , respectively.

We shall now define the product  $f_{i+1/2} f_{j+1/2}$  in such a way that, we will have the much simplest scheme as possible, that degenerate correctly when  $\sigma$  go to 0, and for which the collision term can be evaluated at the lower cost as possible, that is, in  $O(n)$ . Moreover, approximating  $f_{i+1/2} f_{j+1/2}$  by any positive formula allows to insure conservation of the mass, the mean velocity, and the decaying of the entropy. The first idea is to use the so called entropic average, see (Buet et al., 2001), that permits to get rid of the log term

$$\begin{aligned} f_{i+1/2} f_{j+1/2} &= \left( \frac{(f/M)_j D(f/M)_{i+1/2} - (f/M)_i D(f/M)_{j+1/2}}{D \log(f/M)_{i+1/2} - D \log(f/M)_{j+1/2}} \right) \\ &\times M_{i+1/2} M_{j+1/2} \end{aligned} \quad (31)$$

Using this choice, one get

$$\begin{aligned} & \frac{-\sigma}{2\rho} \sum_i \sum_j (D\phi_{i+1/2} - D\phi_{j+1/2}) M_{i+1/2} M_{j+1/2} \\ & \times \left( \left( \frac{f}{M} \right)_j D \left( \frac{f}{M} \right)_{i+1/2} - \left( \frac{f}{M} \right)_i D \left( \frac{f}{M} \right)_{j+1/2} \right) \end{aligned} \quad (32)$$

that can be related to the non-log form of the Fokker-Planck equation:

$$\left( \frac{f}{M} \right)_j D \left( \frac{f}{M} \right)_{i+1/2} - \left( \frac{f}{M} \right)_i D \left( \frac{f}{M} \right)_{j+1/2} = \left( \frac{f}{M} \right)_j \left( \frac{f}{M} \right)_{i+1} - \left( \frac{f}{M} \right)_i \left( \frac{f}{M} \right)_{j+1}$$





We define the product  $M_{i+1/2}M_{j+1/2}$  in such way that the scheme reduces to a “good” scheme when  $\sigma \rightarrow 0$ , and using the analysis done for the linear Fokker–Planck equation, we choose

$$C_{ij} = M_{i+1/2}M_{j+1/2} = \frac{M_{i+1}M_jM_iM_{j+1}}{M_{i+1}M_j - M_iM_{j+1}} \log\left(\frac{M_{i+1}M_j}{M_iM_{j+1}}\right) \quad (33)$$

The ODE system has the following structure

$$\begin{aligned} \partial_t f_i = & \sigma \sum_j C_{ij} \left( \frac{f_{i+1}f_j}{M_{i+1}M_j} - \frac{f_{j+1}f_i}{M_{j+1}M_i} \right) \\ & + C_{i-1,j} \left( \frac{f_{i-1}f_j}{M_{i-1}M_j} - \frac{f_{j-1}f_i}{M_{j-1}M_i} \right) \end{aligned} \quad (34)$$

with  $C_{ij}$  given by Eq. (33).

On this form, the cost to evaluate the coefficients of the differential system is quadratic.

In order to recover a linear cost, let us simplify the expression (33) for  $M_{i+1}M_j > M_{j+1}M_i$ , i.e., with  $z_j = H_{j+i} - H_j + H_i - H_{i+1} > 0$ . One approximates

$$\frac{1}{M_{j+1}M_i - M_{i+1}M_j} = \frac{1}{M_{j+1}M_i(1 - \exp(-z_j/\sigma))} \approx \frac{1}{M_{j+1}M_i} \left( \frac{z_j}{\sigma} \right)$$

Then,

$$C_{ij} \approx M_jM_{i+1} \left( 1 + \frac{\sigma}{z_j} \right) \frac{z_j}{\sigma} \quad (35)$$

and doing the same type of approximation for  $j \leq i$  one obtain the discrete weak formulation

$$\begin{aligned} \sum_{i=0}^n \partial_t f_i \phi_i = & -\frac{1}{2\rho\Delta v} \sum_{i,j=0}^{n-1} (D\phi_{i+1/2} - D\phi_{j+1/2}) \left( \frac{f_{i+1}f_i}{M_{i+1}M_j} - \frac{f_{j+1}f_i}{M_{j+1}M_i} \right) \\ & \times (M_{j+1}M_i(\sigma + z_j) \cdot 1_{\{j>i\}} + M_{i+1}M_j(\sigma - z_j) \cdot 1_{\{j<i\}}) \end{aligned} \quad (36)$$

and as

$$F_{j+1/2} - F_{i+1/2} := \frac{H_{j+1} - H_j}{\Delta v} - \frac{H_{i+1} - H_i}{\Delta v} = \frac{z_j}{\Delta v}$$



we deduce

$$\begin{aligned} \sum_{i=0}^n \partial_t f_i \phi_i &= -\frac{\sigma}{\rho \Delta v} \sum_{i,j=0}^{n-1} D \phi_{i+1/2} (M_{j+1} M_i \cdot 1_{\{j>i\}} + M_{i+1} M_j \cdot 1_{\{j<i\}}) \\ &\quad \times \left( \frac{f_{i+1} f_j}{M_{i+1} M_j} - \frac{f_{j+1} f_i}{M_{j+1} M_i} \right) - \frac{1}{\rho} \sum_{i,j=0}^{n-1} D \phi_{i+1/2} (F_{j+1/2} - F_{i+1/2}) \\ &\quad \times (M_{j+1} M_i \cdot 1_{\{j>i\}} - M_{i+1} M_j \cdot 1_{\{j<i\}}) \left( \frac{f_{i+1} f_j}{M_{i+1} M_j} - \frac{f_{j+1} f_i}{M_{j+1} M_i} \right) \end{aligned} \quad (37)$$

Using the choice (35) for the coefficients  $C_{ij}$ , one gets a tridiagonal matrix: for  $i = 1, \dots, n-1$

$$\partial_t f_i = a_{i+1} f_{i+1} - (a_i + b_i) f_i + b_{i-1} f_{i-1} \quad (38)$$

and the boundary terms are,

$$\partial_t f_0 = a_1 f_1 - b_0 f_0 \quad (39)$$

$$\partial_t f_n = -a_n f_n + b_{n-1} f_{n-1} \quad (40)$$

where the coefficients  $a_i$ , and  $b_i$  are all positives and defined by the following relations, with, for  $i = 1, \dots, n$ :

$$\begin{aligned} a_i &= \frac{1}{\rho \Delta v} \left[ \frac{M_{i-1}}{M_i} \left( \frac{\sigma}{\Delta v} - F_{i-1/2} \right) \sum_{j=i-1}^{n-1} \frac{M_{j+1}}{M_j} f_j + \left( F_{i-1/2} + \frac{\sigma}{\Delta v} \right) \sum_{j=0}^{i-1} f_j \right. \\ &\quad \left. + \frac{M_{i-1}}{M_i} \sum_{j=i-1}^{n-1} \frac{M_{j+1}}{M_j} f_j F_{j+1/2} - \sum_{j=0}^{i-1} f_j F_{j+1/2} \right] \end{aligned} \quad (41)$$

and for  $i = 0, \dots, n-1$ :

$$\begin{aligned} b_i &= \frac{1}{\rho \Delta v} \left[ \left( \frac{\sigma}{\Delta v} - F_{i+1/2} \right) \sum_{j=i}^{n-1} f_{j+1} + \frac{M_{i+1}}{M_i} \left( F_{i+1/2} + \frac{\sigma}{\Delta v} \right) \sum_{j=0}^i \frac{M_j}{M_{j+1}} f_{j+1} \right. \\ &\quad \left. + \sum_{j=i}^{n-1} f_{j+1} F_{j+1/2} - \frac{M_{i+1}}{M_i} \sum_{j=0}^i \frac{M_j}{M_{j+1}} f_{j+1} F_{j+1/2} \right] \end{aligned} \quad (42)$$

Note that the coefficients  $a_i$  and  $b_i$ , can be evaluated in  $O(n)$  steps.

From Eqs. (38)–(40), we note that the system can be written as a difference of flows:

$$\partial_t f_i = G_{i+1/2} - G_{i-1/2}$$

where  $G_{i+1/2} = a_{i+1} f_{i+1} - b_i f_i$ , the conservation of the mass is obvious.



### Properties 3.3.2

This scheme conserves mass and momentum ( $\phi_i = v_i$ ) for the system (37), and decays of the discrete entropy  $\mathcal{E}$ .

*Proof.*

By construction, the approximation of  $f_{i+1/2}f_{j+1/2}$  is positive. Thus starting from Eq. (30) if we set  $\phi_i = 1$  and  $v_i$  respectively, we obtain the conservation of the mass  $\rho$  and the mean velocity  $u$ .

For the entropy, as for the continuous model we have

$$\frac{d}{dt}\mathcal{E} = \sum_i \frac{df_i}{dt} \left( \ln f_i + \frac{H_i}{\sigma} \right) = \sum_i \frac{df_i}{dt} \ln \left( \frac{f_i}{M_i} \right)$$

Thus by taken  $\phi_i = \ln(f_i/M_i)$  in Eq. (30) we have evidently  $(d/dt)\mathcal{E} = 0$ . ■

If  $f$  is such that  $f_i = M_i$  thus we have  $(d/dt)\mathcal{E} = 0$ . For the existence of such equilibrium, we refer to Benedetto et al. (1998), their proof remains valid if we have a discrete measure instead of the Lebesgue measure.

### 3.3.3. Limit When $\sigma \rightarrow 0$ : a Third Discretization

If one considers the limit  $\sigma \rightarrow 0$  in the scheme (38)–(42), one gets the form based on the weak formulation  $ff'(F' - F)$  instead of  $|v - v'|(v - v')$ .

This scheme degenerates correctly when  $\sigma \rightarrow 0$ , indeed,  $z_j$  is positive, defined by

$$\exp \frac{-z_j}{\sigma} = \frac{M_{j+1}M_i}{M_jM_{i+1}}$$

So, when  $\sigma \rightarrow 0$ ,  $\exp(-z_j/\sigma) \rightarrow 0$  and the scheme becomes

$$\begin{aligned} \sum_{i=0}^n \partial_t f_i \phi_i \Delta v &= -\frac{1}{2\rho} \sum_{i,j=0}^{n-1} (D\phi_{i+1/2} - D\phi_{j+1/2}) \cdot (F_{i+1/2} - F_{j+1/2}) \\ &\quad \times (f_{j+1/2} \cdot 1_{\{j>i\}} + f_{i+1/2} \cdot 1_{\{j<i\}})(\Delta v)^2 \end{aligned} \quad (43)$$

which is a discrete form of the weak symmetrized form (14) of the pure granular operator.

### Properties 3.3.4

The system limit conserves mass and momentum, and we have the decay of energy and of the discrete entropy. Equilibrium states are such that  $f_i = 0$  for  $i < i_0$  and  $i > i_0 + 1$  and  $i_0, f_{i_0}, f_{i_0+1}$  are determined from mass and momentum, see Eq. (22).







*Proof.*

By taking  $\phi = 1$ ,  $v$  in Eq. (43), mass and momentum conservation are obtained. The sequence  $F_{i+1/2}$  is increasing in  $i$  and taking  $\phi = v^2$  we have the decay of the temperature

$$\begin{aligned} \partial_t T = & -\frac{\Delta v}{\rho} \sum_{i,j=0}^{n-1} (v_{i+1/2} - v_{j+1/2}) \cdot (F_{i+1/2} - F_{j+1/2}) \\ & \times (f_{j+1} f_i \cdot 1_{\{j>i\}} + f_{i+1} f_j \cdot 1_{\{j<i\}}) (\Delta v)^2 \end{aligned} \quad (44)$$

where  $v_{i+1/2} = (v_i + v_{i+1})/2$ . The sequence  $v_{i+1/2}$  is also increasing in  $i$  and then all the terms  $(v_{i+1/2} - v_{j+1/2})$  and  $(F_{i+1/2} - F_{j+1/2})$  are non negative, which gives the result.

For the entropy we have

$$\begin{aligned} \partial_t E = & -\frac{1}{\rho} \sum_{i,j=0}^{n-1} (F_{i+1/2} - F_{j+1/2})^2 (f_{j+1} f_i \cdot 1_{\{j>i\}} + f_{i+1} f_j \cdot 1_{\{j<i\}}) \\ & \times (\Delta v)^2 \end{aligned} \quad (45)$$

Concerning the equilibrium, we start from the entropy production term (45). First let us verify that the sequence  $F_{i+1/2}$  is a strictly increasing sequence in  $i$ . By the definition of  $F_{i+1/2}$  we have

$$\begin{aligned} F_{i+1/2} - F_{i-1/2} &= \frac{1}{\Delta v} (H_{i+1} + H_{i-1} - 2H_i) \\ &= \frac{1}{\Delta v} \sum_j f_j (|v_{i+1} - v_j|^3 + |v_{i-1} - v_j|^3 - 2|v_i - v_j|^3) \end{aligned}$$

One can verify easily that we have the lowerbound

$$(|v_{i+1} - v_j|^3 + |v_{i-1} - v_j|^3 - 2|v_i - v_j|^3) \geq (\Delta v)^2$$

so that

$$F_{i+1/2} - F_{i-1/2} \geq (\nabla v)^2 \rho \quad (46)$$

Consider  $f$  such that the right hand side of Eq. (45) is null. Since the mass  $\rho$  of  $f$  is not null there exist  $i_0$  such that  $f_{i_0} > 0$ . All the terms involving in the right hand side of Eq. (45) are null since they have all the same sign. Using then Eq. (46), we obtain that  $f_j = 0$  for  $j > i_0 + 1$  and  $j < i_0 - 1$ . And  $i = i_0 - 1$  and  $j = i_0$  gives  $f_{i_0+1} f_{i_0-1} = 0$ . In other words, one of the both terms is null. That shows that  $f$  is as we have claimed it: the values of  $i_0, f_{i_0}$ , and  $f_{i_0+1}$  are uniquely determined by the values of the mass and the momentum. ■



### 3.4. Time Discretization

We shall present the time discretization of system in the form (38)–(40).

#### 3.4.1. Explicit Scheme

In the explicit case, the system can be written:

$$f_i^{n+1} - f_i^n = \Delta t (a_{i+1}^n f_{i+1}^n - c_i^n f_i^n + b_{i-1}^n f_{i-1}^n) f_i^{n+1} - f_i^n = \Delta t Q_i^n f_i^n$$

i.e.,

$$f^{n+1} = \left[ Id + \Delta t \begin{pmatrix} -c_0^n & a_1^n & 0 & \cdots & \cdots & 0 \\ b_0^n & -c_1^n & a_2^n & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & b_{n-2}^n & -c_{n-1}^n & a_n^n \\ 0 & \cdots & \cdots & 0 & b_{n-1}^n & -c_n^n \end{pmatrix} \right] f^n$$

where  $c_i^n = a_i^n + b_i^n$ .

A sufficient condition to have a positive matrix is to take  $\Delta t$  satisfying  $1 - c_i \Delta t > 0$  for all  $i$ . Therefore, we choose  $\Delta t_{\max} = (\sup c_i)^{-1}$ .

Notice that  $\Delta t_{\max}$  behave like  $O(\Delta v^2 / \sigma + (\Delta v) \sup_i |F_{i+1/2}|)$ .

The properties of conservation of mass and momentum, of the system can be checked easily ( $\phi_i = 1, v_i$ ).

#### 3.4.2. Implicit Scheme

The implicit scheme reads:

$$f_i^{n+1} - f_i^n = \Delta t a_{i+1}^{n+1} f_{i+1}^{n+1} - \Delta t c_i^{n+1} f_i^{n+1} + \Delta t b_{i-1}^{n+1} f_{i-1}^{n+1}$$

i.e.,

$$f^n = (Id - \Delta t Q(f^{n+1})) f^{n+1}$$

where  $Q(f^{n+1})$  is the previous matrix, at range  $n + 1$ , instead of  $n$ . We use an iterative method:

$$f^n = (Id - \Delta t Q(f^n)) f^{n+l}$$



and we define the process as:

$$(Id - \Delta t Q(g^n))g^{n+1} = f^n$$

$$g^0 = f^n$$

to resolve it, where  $g^n$  is the  $n^{ieme}$  iterate.

At each iteration, mass and positivity are preserved. The conservation of mass follows from structure of the matrix  $Q$ , which corresponds to a conservative scheme. The last point follows from a classical algebra result (Berman and Plemmons, 1994).

Let  $N$  be a matrix  $\in \mathbb{R}^n \times \mathbb{R}^n$ , tridiagonal, whose diagonal coefficients are positives, the others negatives. Then  $N^{-1}$  is positive, if and only if, there exists a diagonal positive matrix  $D$ , such that  $D^{-1}ND$  is diagonal dominant.  $N = (Id - \Delta t Q(g^n))$ , take  $G \in \text{Ker}(Q)$  positive which is always possible, then  $NG = \tilde{G}$ , where  $\tilde{G}$  is the diagonal positive matrix formed with the coefficients of  $G$ . Thus  $\tilde{G}^{-1}NG = Id$  which is a diagonal dominant matrix. This concludes the proof.

One can also notice that if  $f^n$  is an equilibrium then  $g^n = f^n$ .

#### Lemma 3.4

For the implicit scheme the entropy is decaying.

*Proof.*

The entropy  $\mathcal{E}$  can be written as:

$$\mathcal{E}(f) = \sigma \int (f \ln f + \frac{1}{6} \int |v - v'|^3 f(v) f'(v) dv) dv = \sigma \int (\ln f - \ln N) f dv$$

where  $N(v) = \exp(-(1/6\sigma) \int |v - v'| f(v') dv') = \exp(-3H(v)/2\sigma)$ . We get

$$\begin{aligned} \mathcal{E}^{n+1} - \mathcal{E}^n &= \sigma \left( \int f^{n+1} \ln \left( \frac{f}{N} \right)^{n+1} dv - \int f^n \ln \left( \frac{f}{N} \right)^n dv \right) \\ &= \sigma \left( \int f^n \left( \ln \left( \frac{f}{N} \right)^{n+1} - \ln \left( \frac{f}{N} \right)^n \right) dv \right. \\ &\quad \left. + \Delta t \int Q^{n+1} f^{n+1} \ln \left( \frac{f}{N} \right)^{n+1} dv \right) \end{aligned}$$





with  $Q^{n+1} = Q(f^{n+1})$ ,

$$\begin{aligned} \mathcal{E}^{n+1} - \mathcal{E}^n &\leq \sigma \left( \int f^n \ln \left( \frac{f^{n+1}}{f^n} \right) - \int f^n \ln \left( \frac{N^{n+1}}{N^n} \right) dv \right. \\ &\quad \left. + \Delta t \int Q^{n+1} f^{n+1} \ln(N)^{n+1} dv \right) \end{aligned}$$

Using the entropy decay on the weak form (18) with  $\phi = \ln(f/M)$  ( $M = N^2$ ), one gets

$$\begin{aligned} \mathcal{E}^{n+1} - \mathcal{E}^n &\leq \sigma \left( \int_v f^n \ln(N)^n dv + \int_v f^{n+1} \ln(N)^{n+1} dv \right. \\ &\quad \left. - 2 \int_v f^n \ln(N)^{n+1} dv \right) \\ &\leq \sigma \left( \int_v f^n (\ln(N)^n - \ln(N)^{n+1}) dv + \int_v (f^{n+1} - f^n) \ln(N)^{n+1} dv \right) \end{aligned}$$

By definition of  $N$ , we have:

$$\begin{aligned} \mathcal{E}^{n+1} - \mathcal{E}^n &\leq \frac{1}{2} \int_v \int_{v'} |v - v'|^3 [f^n f^{n+1} - f^n] - f^{n+1} (f^{n+1} - f^n) dv dv' \\ &\leq -\frac{1}{2} \int_v \int_{v'} |v - v'|^3 [(f^{n+1} - f^n)(f^{n+1} - f^n)] dv dv' \end{aligned}$$

Since from lemma (2.2) the application

$$g \rightarrow \int \int g(v)g(v')|v - v'|^3 dv dv'$$

is positive for functions  $g$  such that  $\int g = \int vg = 0$ , we have the decay of entropy.

The proof for the discrete formulation follow the same lines:

$$\begin{aligned} \mathcal{E}^{n+1} - \mathcal{E}^n &\leq \sigma \left( \sum_i f_i^n (\ln(N_i)^n - \ln(N_i)^{n+1}) \Delta v \right. \\ &\quad \left. + \sum_i (f_i^{n+1} - f_i^n) \ln(N_i)^{n+1} \Delta v \right) \end{aligned}$$





The definition of  $N$  gives:

$$\mathcal{E}^{n+1} - \mathcal{E}^n \leq -\frac{1}{2} \sum_i \sum_j |v_i - v_j|^3 [(f_j^{n+1} - f_j^n)(f_i^{n+1} - f_i^n)] (\Delta v)^2 \quad (47)$$

By a discrete version of the lemma (2.2), which can be obtained by regularizing the sum of Dirac mass, we conclude to the decay of entropy. ■

#### 4. NUMERICAL RESULTS

We illustrate the method on two tests, a Dirac and a Gaussian for initial data. For the two tests the velocity domains is  $[0, 12.5]$ , the total mass is one and the two initial data are centered at  $v = 4$ . For the gaussian the variance is 1. For all, the run of the time step is taken equals to 0.1. We use two uniform grids of 50 and 200 points. For the explicit scheme, we use sub cycling technique to respect the CFL constraint that guarantee positivity of the scheme. For the implicit scheme, the iterative process is stopped when the error on the relative velocity is smaller than  $10^{-12}$ .

##### 4.1. Computational Cost

Let us first compare the computational cost between explicit and implicit. The costs correspond to  $\sigma = 1$ .

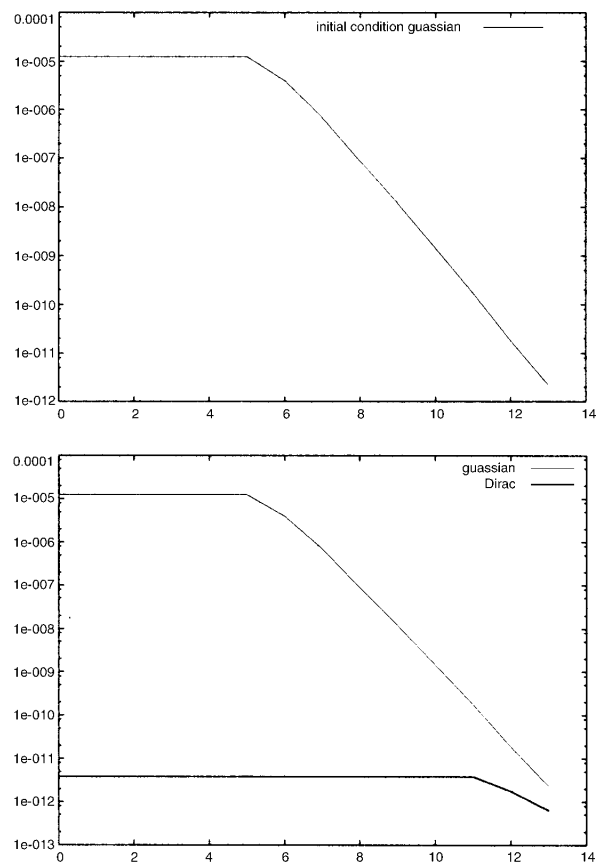
For the implicit scheme the number of iterates varies from 12, at the first time step, to 1 after few time steps when starting from the gaussian, and from 5 to 1 when starting from the Dirac, and this with the two grids. The number of iterates seems to does not depend of the number of points. For the explicit scheme, the number of subcycles is nearly constant during the computation. Starting from the gaussian or the Dirac we need 60 subcycles with 50 points and 283 subcycles with 200 points. The cost of the iterations is the same for the implicit and the explicit.

More the grid is fine, larger is the cost ratio explicit/implicit. For very coarse grid (10 points) the cost of the explicit scheme when starting from the gaussian is smaller than the cost of the implicit but only for the first-time steps of the relaxation.

One interesting feature of the implicit scheme concerns its ability to reach an equilibrium state, despite the fact that the iterative process is stopped, when the error on the relative velocity is smaller than  $O(10^{-n})$ . This implementation of the implicit scheme, strictly speaking, is not conservative for the mean velocity, but we observe that for any value of the allowed error, in long time, the



scheme converge toward an equilibrium state for which, the mean velocity is nearly those of the initial condition upto an error of  $O(10^{-n})$ . We plot the difference between the average velocity after total relaxation and the initial one, in function of the error, for the two types of initial condition, a Dirac mass and a Gaussian, see Fig. 1.



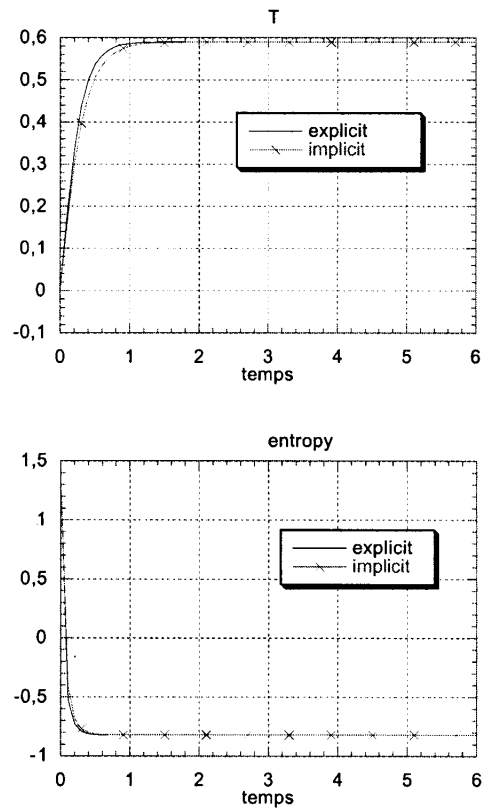
**Figure 1.** Gaussian and Dirac initial condition: difference between the average velocity after total relaxation and the initial one in function of the error, in log scale.



#### 4.2. Graphics

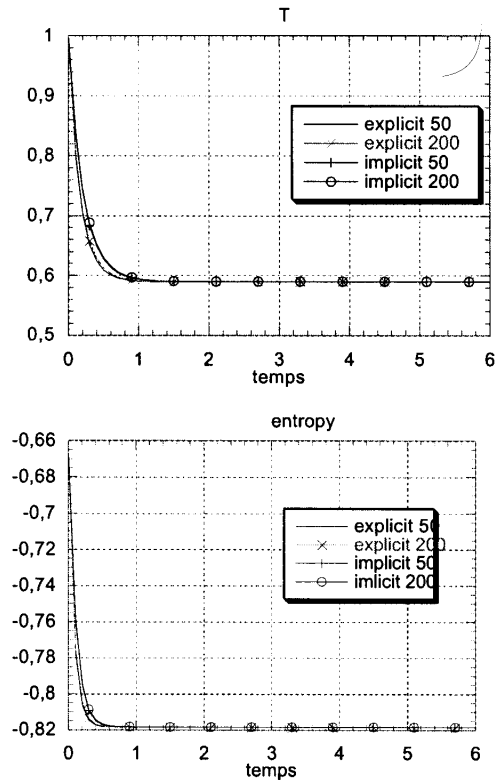
Figure 2 shows the evolution in time of the temperature starting from the Dirac mass, with 50 meshes,  $\sigma = 1$  and explicit and implicit scheme.

Figure 3 shows the evolution of the temperature starting from the Gaussian, with 50 and 200 grid points and again with  $\sigma = 1$  and using explicit and implicit version of the scheme.



**Figure 2.** Initial condition: Dirac mass. Evolution of the temperature and the entropy, 50 points,  $\sigma = 1$ .





**Figure 3.** Initial condition: Gaussian. Evolution of the temperature and the entropy, 50 and 200 points,  $\sigma = 1$ .

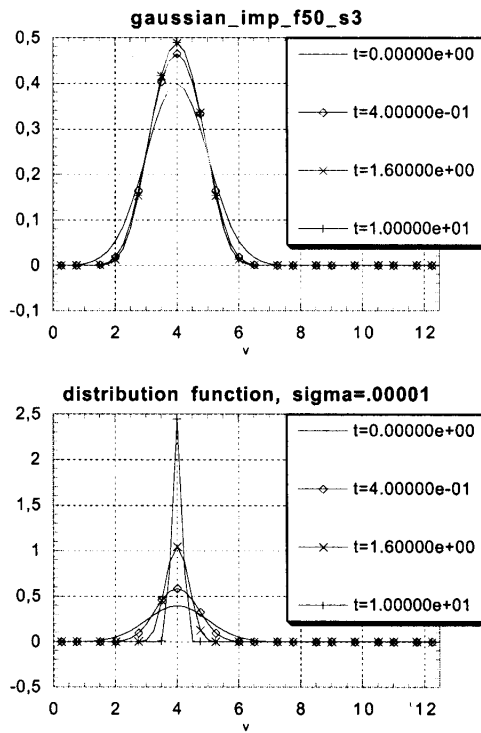
Starting from the Gaussian, we can compare for  $\sigma = 1$  and  $\sigma = 0.00001$  the evolution of the distribution function in Fig. 4 and the relaxation of the temperature Fig. 5. Implicit scheme is used with 50 grid points.

We can also show in Fig. 6, the relaxation of the temperature for the two initial condition when  $\sigma = 1$ , using 50 grid points and implicit scheme.

To finish let us show the behavior of the equilibrium state  $f_\infty$ : we compare the Gaussian  $\exp(-v^2/2)$  with the equilibrium state obtained with this initial data and corresponding to  $\sigma = 1$  and which behave for large  $|v|$  as







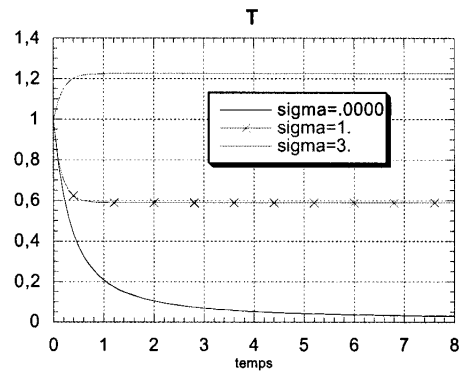
**Figure 4.** Initial condition: Gaussian. Evolution of the distribution function, 50 points,  $\sigma = 1$  and  $\sigma = 10^{-5}$ .

$\exp(-a|v|^3)$ . We plot  $-\log f$  in function of  $v$ , in logarithmic scale, see Fig. 7. It is easily seen in Fig. 7 that for large velocity, the distribution function has the expected behavior, the slope of  $-\log f(v)$  tends to 3 for the steady states of granular media, and to 3 for Maxwellian.

## 5. CONCLUSION

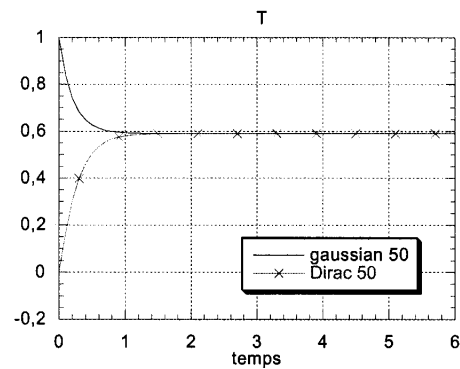
In 1D, without friction term, the Chang–Cooper method allows us to construct a discretization for the granular equation.



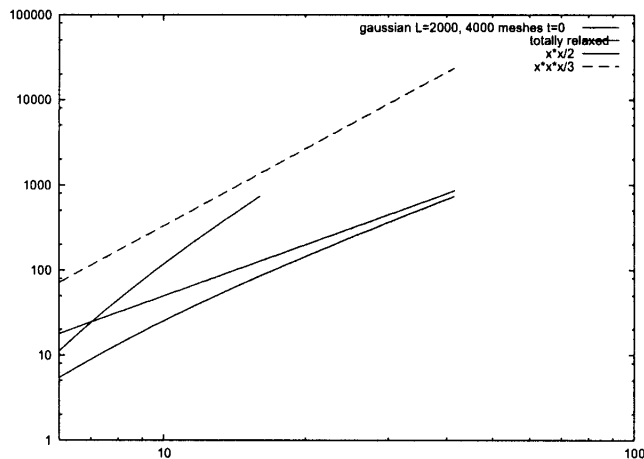


**Figure 5.** Evolution of the temperature starting from Gaussian with different values of the temperature of the thermal bath,  $\sigma = 1$  and  $\sigma = 10^{-5}$ , 50 points.

Moreover, the discretization is performed at a linear cost in the number of grid points. An implicit scheme has been used. It converges in large time to the equilibrium state described by Benedetto et al., (1998) and it preserves the properties of the equation. Moreover, the cost of such implicit scheme is less than the explicit one. This implicit method could also be used for the



**Figure 6.** Evolution of the temperature starting from Dirac mass and Gaussian,  $\sigma = 1$ , 50 points.



**Figure 7.** The distribution function in log scale for a gaussian and the equilibrium state associated for  $\sigma = 1$ .

numerical methods described by Buet and Cordier (1998b; 2002) for the isotropic Fokker–Planck–Landau equation.

In this study, the friction term was ignored. Since the discretization's process is applied to a symmetric form of the equation, the friction term cannot be included within the presented framework.

However, it should be possible to take it into account using splitting technics.

## REFERENCES

- Baldassarri, A., Puglisi, A., Marconi, U. (2002). Kinetics models of inelastic gases. *Math. Models Meth. Appl. Sci.* 12:965–984.
- Benedetto, D., Caglioti, E. (1999). The collapse phenomenon in one-dimensional inelastic point particle systems. *Physica D* 132:457.
- Benedetto, D., Caglioti, E., Pulvirenti, M. (1997). A kinetic equation for granular media. *Math. Mod. Num. Anal. M2AN* 31(5):615.
- Benedetto, D., Caglioti, E., Carillo, J. A., Pulvirenti, M. (1998). A non-Maxwellian equilibrium distribution for the one-dimensional granular media. *J. Stat. Phys.* 91(5/6):979.





- Benedetto, D., Caglioti, E., Golse, F., Pulvirenti, M. (1999). A hydrodynamical model arising in the context of granular media. *Comput. Math. Appl.* 38(7–8):121–131.
- Berman, A., Plemmons, R. J. (1994). Nonnegative matrices in the mathematical sciences. *Classics Appl. Math.* 9.
- Buet, C., Cordier, S. (1998a). Numerical analysis of conservative and entropy schemes for the Fokker-Planck-Landau equation. *SIAM J. Numer. Anal.* 36(3):953–973.
- Buet, C., Cordier, S. (1998b). Conservative and entropy decaying numerical scheme for the isotropic Fokker-Planck-Landau equation. *J. Comput. Phys.* 145(1):228–245.
- Buet, C., Cordier, S. (2002). Numerical analysis of the isotropic Fokker-Planck-Landau equation. (*English*) *J. Comput. Phys.* 179(1):43–67.
- Buet, C., Cordier, S. (2003). Bellomo, N., Gatignol, R., Eds. *Numerical Method for the Compton Scattering Operator*. Lecture Notes on the Discretization of the Boltzmann Equation World Scientific (to appear).
- Buet, C., Cordier, S., Degond, P., Lemou, M. (1997). Fast algorithms for numerical, conservative, and entropy approximations of the Fokker-Planck-Landau equation. *J. Comput. Phys.* 133(2):310–322.
- Buet, C., Dellacherie, S., Sentis, R. (2001). Numerical solution of an ionic Fokker-Planck equation with electronic temperature. *SIAM J. Numer. Anal.* 39(4):1219–1253.
- Chang, J. S., Cooper, G. (1970). A practical difference scheme for Fokker-Planck equations. *J. Comput. Phys.* 6:1–16.
- Esteban, M., Perthame, B. (1991). On the modified Enskog equation for elastic and inelastic collisions. *Models XIth Spin. Ann. Inst. Poincaré* 8:289–308.
- Görsch, D. (2002). Generalized discrete velocity models. *Math. Models Meth. Appl. Sci.* 12:49–76.
- MacNamara, S., Young, W. R. (1992). Inelastic collapse and clumping in a one-dimensional granular medium. *Phys. Fluids*, A4 3:496.
- MacNamara, S., Young, W. R. (1993). Kinetics of a one-dimensional granular medium in the quasi-elastic limit. *Phys. Fluids* A5 1:1619.
- MacNamara, S., Young, W. R. (1994). Inelastic collapse in two dimension. *Phys. Rev. E* 50:R28.
- Toscani, G. (2000). One dimensional kinetic models of granular flows. *M2 AN* 34:6.

Received June 26, 2002

Revised February 17, 2003

Accepted June 4, 2003

Copyright © Marcel Dekker, Inc. All rights reserved.



## Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique

Christophe BUET, Stéphane DELLACHERIE, Rémi SENTIS

CEA, Bruyères-le-Châtel, service MLS, B.P.12, 91680 Bruyères-le-Châtel, France

(Reçu le 9 mai 1998, accepté le 25 mai 1998)

---

**Résumé.** On décrit un schéma numérique pour le traitement d'un opérateur de collision ion/électron de type Fokker-Planck; pour cela on introduit la notion de *moyenne entropique* de deux quantités positives. Ce schéma a la propriété d'être entropique au sens du théorème H de Boltzmann sous un critère de type CFL. On montre de plus que la solution converge en temps grand vers un unique état d'équilibre maxwellien.  
© Académie des Sciences/Elsevier, Paris

### *Numerical solution of a ionic Fokker-Planck equation with electronic temperature*

**Abstract.** We describe a numerical scheme for dealing with an ion electron/collision operator of the Fokker-Planck type; for that purpose, we introduce the notion of entropic average of two positive quantities. This scheme has the property to be entropic in the sense of Boltzmann's H theorem under a CFL criteria. Moreover, we prove that the solution converges when the time grows, towards a unique Maxwellian equilibrium state. © Académie des Sciences/Elsevier, Paris

---

### *Abridged English Version*

We are concerned with the modelization of a hot plasma with only one species whose atomic mass in  $m$ , the ionization level being  $Z$ . The ionic distribution  $f = f(t, x, v)$  ( $x \in \mathbf{R}^3$  and  $\vec{v} \in \mathbf{R}^3$ ) and the electronic temperature  $T_e(t, x)$  are solutions of (1), (2), where  $\mathcal{E}_e(T_e) = \frac{3}{2} Z N T_e$ ,  $P_e = Z N T_e$  are the internal energy and the pressure of the electrons. Moreover,  $N, \vec{U}$  are the density and the macroscopic velocity of the ions and  $\langle \cdot \rangle = \int \cdot d\vec{v}$ . The Fokker-Planck operator  $S$  (see (3)) describes the ion/electron collisions and the classical quadratic Fokker-Planck operator  $B$  (see [7]) describes the ion/ion collisions.

---

Note présentée par Jacques-Louis LIONS.

The numerical solution of the overall system (1) and (2) can be done with a finite difference method in the phase space  $(\vec{x}, \vec{v})$  with a five stages splitting:

- 1) Resolution of  $\frac{\partial}{\partial t}f + \vec{v} \cdot \nabla_x f = 0$  with an upwind (with respect to the  $x$  variable) scheme.
- 2) Resolution of the ion/ion Fokker–Planck operator, i.e. we solve  $\frac{\partial}{\partial t}f = B(f)$ . (See [7] for example for a conservative and entropic scheme.)
- 3) Resolution of  $\frac{\partial}{\partial t}f = \frac{\nabla_x P_e}{Nm} \nabla_v f$  with an upwind (with respect to the  $\vec{v}$  variable) scheme.
- 4) Resolution of the Fokker–Planck operator  $S$ , i.e. we solve  $\frac{\partial}{\partial t}f = S(f)$  coupled with  $\frac{\partial}{\partial t}\mathcal{E}_e = \frac{m}{2}\langle v^2 S(f) \rangle$ .
- 5) Resolution of the remaining part of the electronic energy equation.

In this Note, we describe only the fourth stage and we introduce the semi-discrete and the fully discrete scheme with (9), (10) and (16) knowing that the distribution  $f$  at the interfaces of the mesh is defined with the *entropic average*: see the definition (8). The main results of this Note are summarized in the following propositions (where  $\langle g \rangle = \sum_j = g_j \Delta v$ ).

**PROPOSITION 3.** – *For all initial conditions  $\{(f_j^0)_j, T_e^0\}$  strictly positive, the scheme (9) and (10) is such that  $f_j(t)$  and  $T_e(t)$  are strictly positive. Moreover, the numerical entropy  $\langle f \log f \rangle + ZN \log(NT_e^{-3/2})$  decreases and, when  $t \rightarrow \infty$ , we have:*

$$T_e(t) \rightarrow T_e^\infty, \quad f_j(t) \rightarrow f_j^\infty = N \frac{\mathcal{M}_{U^\infty, T_e^\infty}(v_j)}{\langle \mathcal{M}_{U^\infty, T_e^\infty} \rangle}$$

knowing that  $(U^\infty, T_e^\infty)$  is the unique solution of the system (14)-(15).

**PROPOSITION 4.** – *For all initial conditions  $\{(f_j^0)_j, T_e^0\}$  strictly positive, the scheme (16) is such that  $f^n > 0$ ,  $\inf_n(T_e^n) > 0$ , and we have:*

$$H^\infty \leq H^{n+1} \leq H^n \quad (\text{Gibbs lemma})$$

when  $\Delta t < \min(\Delta t_1^n, \Delta t_2^n)$ . Moreover, the thermodynamical equilibrium is preserved, i.e.

$$f_j^n = \frac{N}{\langle \mathcal{M}_{U^\infty, T_e^\infty} \rangle} \mathcal{M}_{U^\infty, T_e^\infty}(v_j) \text{ and } T_e^n = T_e^\infty \iff f^{n+1} = f^n.$$

See §2 for the definition of  $H^\infty$ ,  $\Delta t_1^n$ ,  $\Delta t_2^n$ , and the Maxwellian equilibrium  $\mathcal{M}_\cdot$ ;  $H^n$  is the value of the numerical entropy at time  $t^n$ . Numerical results show that this scheme describes well the ion/electron collisions, even on a coarse velocity grid.

L'évolution d'une population  $f = f(t, x, v)$  d'ions (de masse  $m$  et de charge  $Z$ ) et de la température électronique  $T_e = T_e(t, x)$  (où  $x \in \mathbf{R}^3$  et  $\vec{v} \in \mathbf{R}^3$ ) est régie par le système :

$$\begin{cases} \frac{\partial}{\partial t}f + \vec{v} \cdot \nabla_x f - \frac{\nabla_x P_e}{Nm} \cdot \nabla_v f = B(f) + S(f), \\ \frac{\partial}{\partial t}\mathcal{E}_e(T_e) + \nabla_x \cdot (\mathcal{E}_e(T_e)\vec{U}) + P_e \nabla_x \vec{U} = -\frac{m}{2}\langle v^2 S(f) \rangle, \end{cases} \quad (1)$$

où  $\mathcal{E}_e(T_e) = \frac{3}{2} ZNT_e$ ,  $P_e = ZNT_e$ ,  $N = \langle f \rangle$ ,  $N\vec{U} = \langle f\vec{v} \rangle$  et

$$S(f)(\vec{v}) = \Omega \nabla_v \cdot \left[ (\vec{v} - \vec{U}) f(v) + \frac{T_e}{m} \nabla_v f \right]; \quad (3)$$

de plus,  $B$  est l'opérateur de Fokker–Planck quadratique classique (voir [4]) et  $\langle \cdot \rangle \equiv \int_{\mathbf{R}^3} \cdot d\vec{v}$ . Le coefficient  $\Omega$  dépend continûment de  $N$  et  $T_e$  (voir [6]). Sur l'origine de ce système, voir [3] ou [7]. Pour les détails de cette Note, voir [2].

## 1. Généralités

On suppose que les fonctions  $f$  sont positives et à décroissance rapide vers 0 quand  $|\vec{v}| \rightarrow \infty$ . On sait que l'opérateur  $B$  est conservatif en masse, impulsion et énergie ; de plus, il fait décroître la quantité  $\langle f \log f \rangle$  qui est l'entropie ionique (car  $\langle B(f) \log f \rangle \leq 0$ ). D'autre part, en introduisant une température ionique  $T$  définie par  $3NT = m \langle (\vec{v} - \vec{U})^2 f \rangle$ , l'opérateur  $S$  vérifie :

$$\langle S(f) \rangle = 0, \quad \langle S(f) \vec{v} \rangle = 0, \quad \frac{m}{2} \langle S(f) v^2 \rangle = 3\Omega N (T_e - T).$$

Définissons la maxwellienne  $\mathbf{M}_{N, \vec{U}, T}(\vec{v}) = \frac{N}{(2\pi T/m)^{3/2}} \exp\left[-\frac{m(\vec{v} - \vec{U})^2}{2T}\right]$ . En supposant  $f$  maxwellienne et en prenant les trois premiers moments de (1), on obtient :

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \nabla_x \vec{U}\right) N &= 0, \quad \left(\frac{\partial}{\partial t} + \nabla_x \vec{U}\right) (N\vec{U}) + \nabla_x (NT + P_e) = 0, \\ \left(\frac{\partial}{\partial t} + \nabla_x \vec{U}\right) \left(\frac{3}{2} NT\right) + NT \nabla_x \vec{U} &= 3\Omega N (T_e - T), \end{aligned}$$

ce qui, couplé à (2), donne le système Euler bitempérature classique (voir par exemple [5]).

PROPOSITION 1. – Pour tout  $f > 0$  et  $T_e > 0$ , on a :

$$\left\langle S(f) \log \left( f / \mathbf{M}_{N, \vec{U}, T_e} \right) \right\rangle \leq 0.$$

De plus,  $\left\langle S(f) \log \left( \mathbf{M}_{N, \vec{U}, T_e} \right) \right\rangle = 0$  si et seulement si  $f = \mathbf{M}_{N, \vec{U}, T_e}$ .

PROPOSITION 2. – Soient  $f$  et  $T_e$  solutions de (1) avec  $T_e$  régulière en  $x$ . Alors on a la relation de décroissance de l'entropie

$$\frac{\partial}{\partial t} (\langle f \log f \rangle + \mathcal{H}_e) + \nabla_x \cdot (\langle \vec{v} f \log f \rangle + \vec{U} \mathcal{H}_e) \leq 0, \text{ où } \mathcal{H}_e = ZN \log(NT_e^{-3/2}). \quad (5)$$

La résolution numérique du système complet se fait par une méthode de différences finies en espace et en vitesse grâce à un splitting en cinq phases indiqué dans la version anglaise. On décrit ci-dessous un schéma numérique pour la phase 4.

## 2. Schéma numérique semi-discret

Plaçons-nous dans un cadre monodimensionnel en  $\vec{v}$  pour simplifier ;  $\mathcal{E}_e(T_e)$  devient alors  $\frac{1}{2} ZNT_e$ , et

$$m \langle (v - U)^2 f \rangle = NT, \quad \frac{m}{2} \langle v^2 S(f) \rangle = \Omega (NT - NT_e).$$

Il s'agit de discrétiser en vitesse le système non linéaire suivant :

$$\frac{\partial}{\partial t} f = S(f), \quad f(0) = f^0, \quad (6)$$

$$\frac{\partial}{\partial t} \mathcal{E}_e(T_e) = \Omega (m \langle (v - U)^2 f \rangle - NT_e), \quad T_e(0) = T_e^0. \quad (7)$$

On utilise une discrétisation de  $\mathbf{R}$  grâce à une grille de différences finies  $\{v_j\}_{1, \dots, j_{\max}}$  (où  $v_{j+1} - v_j = \Delta v$ ), et on note  $f_j = f_j(t)$  l'évaluation de  $f(t)$  en  $v_j$ . Comme on tronque le domaine en vitesse, il convient d'imposer la condition au bord du domaine :  $(v - U)f + \frac{T_e}{m} \partial_v f = 0$ . Dorénavant,  $\langle g \rangle$  désigne  $\sum_{j=1}^{j_{\max}} g_j \Delta v$ .

C. Buet et al.

Pour deux quantités strictement positives  $x$  et  $y$ , on définit leur *moyenne entropique*  $\tilde{f}_{x,y}$  par :

$$\tilde{f}_{x,y} = \frac{x-y}{\log x - \log y} \text{ si } x \neq y, \quad \tilde{f}_{x,y} = x \text{ sinon} \quad (8)$$

et l'on pose  $\tilde{f}_{x,y} = 0$  si  $xy = 0$ . Par ailleurs, notons  $(Af)_i = a_j(f_{j+1} - f_j) - b_j(f_j - f_{j-1})$  avec  $a_j = 1$  (sauf  $a_{j_{\max}} = 0$ ),  $b_j = 1$  (sauf  $b_1 = 0$ ) et considérons le système semi-discrétisé satisfait par  $\{f_j(t), T_e(t)\}$  :

$$\frac{\partial}{\partial t} f_j = S(f)_j, \quad S(f)_j = \frac{\Omega}{\Delta v} \left[ (v_{j+1/2} - \tilde{U}) \tilde{f}_{j+1/2} - (v_{j-1/2} - \tilde{U}) \tilde{f}_{j-1/2} \right] + \frac{\Omega T_e}{m \Delta v^2} (Af)_i, \quad (9)$$

$$\frac{\partial}{\partial t} \mathcal{E}_e(T_e) + \frac{m}{2} \langle v_j^2 S(f) \rangle = 0 \quad (10)$$

avec les conditions initiales naturelles et

$$\tilde{U} = \sum_j v_{j+1/2} \tilde{f}_{j+1/2} \left( \sum_j \tilde{f}_{j+1/2} \right)^{-1}, \quad (11)$$

$\tilde{f}_{j+1/2}$  étant la moyenne entropique de  $f_j$  et de  $f_{j+1}$ ; on pose d'autre part  $\tilde{f}_{1/2} = \tilde{f}_{j_{\max}+1/2} \equiv 0$ . On obtient les propriétés de bilan

$$\langle S(f) \rangle = 0, \quad \langle v_j S(f) \rangle = 0. \quad (12)$$

En toute rigueur, la dernière relation n'est vraie que si  $f_1 = 0$  et  $f_{j_{\max}} = 0$  (voir [2] pour une analyse rigoureuse). De plus, la formule (10) est consistante avec (7) car on a :

$$\frac{m}{2} \langle v^2 S(f) \rangle = \Omega (N T_e - \widetilde{N T}), \quad \text{où } \widetilde{N T} = \frac{m \Delta v}{2} \sum_j (v_{j+1/2} - \tilde{U})^2 \tilde{f}_{j+1/2}.$$

Pour des  $f_j$  et  $T_e$  positifs, on introduit l'entropie numérique  $H(t) = \langle f \log f \rangle - \frac{ZN}{2} \log T_e$ . Notons  $\mathcal{M}_{u,y} = \left( \frac{m}{2\pi y} \right)^{1/2} \exp \left[ -\frac{m(v-u)^2}{2y} \right]$ .

LEMME. – Pour tous  $f_j(t)$  et  $T_e(t)$  positifs, en posant  $g = f / \mathcal{M}_{\tilde{U}, T_e}$ , on a :

$$\begin{aligned} S(f) \frac{m \Delta v^2}{\Omega T_e} &= \tilde{f}_{j+1/2} (\log(g_{j+1}) - \log(g_j)) \\ &\quad - \tilde{f}_{j-1/2} (\log(g_j) - \log(g_{j-1})) \\ \frac{\partial}{\partial t} H &= -\frac{\Omega T_e}{m \Delta v^2} \sum_j \tilde{f}_{j+1/2} [\log(g_{j+1}) - \log(g_j)]^2 \leq 0. \end{aligned}$$

PROPOSITION 3. – Pour toute condition initiale  $\{(f_j^0)_j, T_e^0\}$  strictement positive, le schéma (9) et (10) est tel que  $f_j(t) > 0$  et  $\inf_{t \in [0, +\infty[} T_e(t) > 0$ . L'entropie  $H(t)$  décroît et quant  $t \rightarrow \infty$  :

$$T_e(t) \rightarrow T_e^\infty, \quad f_j(t) \rightarrow f_j^\infty = \hat{N} \mathcal{M}_{U^\infty, T_e^\infty}(v_j) \quad \text{où } \hat{N} = N \langle \mathcal{M}_{U^\infty, T_e^\infty} \rangle^{-1}$$

sachant que  $(U^\infty, T_e^\infty)$  est l'unique solution du système :

$$\langle v \mathcal{M}_{U^\infty, T_e^\infty} \rangle \hat{N} = \langle v f^0 \rangle, \quad (14)$$

$$\langle (v - U^0)^2 \mathcal{M}_{U^\infty, T_e^\infty} \rangle \hat{N} + \frac{ZN}{m} T_e^\infty = \langle (v - U^0)^2 f^0 \rangle + \frac{ZN}{m} T_e^0. \quad (15)$$



### 3. Schéma numérique totalement discret, explicite

En utilisant la moyenne entropique  $\tilde{f}_{j+1/2}^n$  de  $f_j^n$  et de  $f_{j+1}^n$  et les conditions aux limites comme dans le schéma semi-discret, la discrétisation explicite de (9), (10) s'écrit :

$$f_j^{n+1} - f_j^n = \Delta t S(f^n, T_e^n)_j, \quad \frac{ZN}{2} (T_e^{n+1} - T_e^n) = \Delta t \Omega^n (\widetilde{NT}^n - N^n T_e^n). \quad (16)$$

On obtient les propriétés de conservation :

$$\begin{cases} \langle f^{n+1} \rangle = \langle f^n \rangle, & \langle v f^{n+1} \rangle = \langle v f^n \rangle, \\ \left\langle \frac{v^2}{2} f^{n+1} \right\rangle - \left\langle \frac{v^2}{2} f^n \right\rangle = \frac{\Delta t}{m} \Omega^n (N^n T_e^n - \widetilde{NT}^n). \end{cases}$$

Notons  $H^\infty = \langle f^\infty \log f^\infty \rangle - \frac{ZN}{2} \log T_e^\infty$  et  $\Delta t_1^n = \frac{m \Delta v^2}{\beta \Omega^n T_e^n \mathfrak{M}^n} \min_k \left( \frac{f^n}{\mathcal{M}^n} \right)_k \left[ \max_k \left( \frac{f^n}{\mathcal{M}^n} \right)_k \right]^{-1}$  avec :

$$\mathfrak{M}^n = \max_k \frac{\mathcal{M}_{k\pm 1}^n}{\mathcal{M}_k^n}, \quad \mathcal{M}_k^n = e^{-\frac{m(v_k - \tilde{v}^n)^2}{2T_e^n}}, \quad \beta = 4 + 4 \frac{\max_k (v_k)^4}{Z (T_e^n/m)^2}.$$

PROPOSITION 4. – Pour toute condition initiale  $\{(f_j^0)_j, T_e^0\}$  strictement positive, le schéma (16) est tel que  $f^n > 0$  et  $\inf_n (T_e^n) > 0$ ; de plus, on a :

$$H^\infty \leq H^{n+1} \leq H^n \text{ (lemme de Gibbs)}$$

sous le critère CFL  $\Delta t < \Delta t_1^* = \min(\Delta t_1^n, \frac{Z}{4\Omega^n})$ . D'autre part, on a :

$$f_j^n = f_j^\infty \text{ et } T_e^n = T_e^\infty \iff f^{n+1} = f^n.$$

Sous l'hypothèse  $\Omega = C^{\text{ste}} \frac{N}{T_e^{3/2}}$  (voir [3]), on montre que  $\Delta t_1^*$  ne tend pas vers 0 quand  $n$  augmente.

### 4. Résultats numériques

On considère le problème (6), (7) (où  $v$  est monodimensionnelle), on prend  $v_{j_{\max}} = -v_1 = 5\sqrt{\frac{T_e^0}{m}}$  avec  $j_{\max} = 100$  et  $Z = 2$ ,  $m = 2.5m_{\text{proton}}$ . Les conditions initiales sont les suivantes :

$$\begin{cases} f^0 = \text{maxwellienne centrée de température } T^0, \\ N = 10^{22} \text{ cm}^{-3}, \quad U = 0, \quad T^0 = 1 \text{ Kev}, \quad T_e^0 = 2 \text{ Kev}. \end{cases}$$

On compare sur les figures 1 à 4 les résultats obtenus avec les schémas utilisant la moyenne entropique, la moyenne de Chang et Cooper (voir [4]) et la moyenne arithmétique sur un maillage fin et sur un maillage grossier. (Si la température est faible en début de calcul et est très forte en fin de calcul, il faut avoir un  $\Delta v$  adapté, ce qui conduit à une discrétisation sur quelques mailles seulement de la fonction de répartition initiale  $f^0$ .) Lorsque le maillage est grossier, on constate que la moyenne entropique donne un résultat plus précis que celle de Chang et Cooper et que la moyenne arithmétique ne préserve plus la positivité de la distribution  $f$  (voir figure 4).

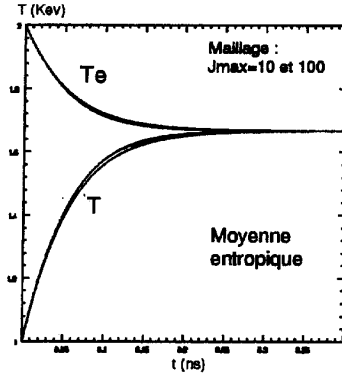


Figure 1. – Températures.

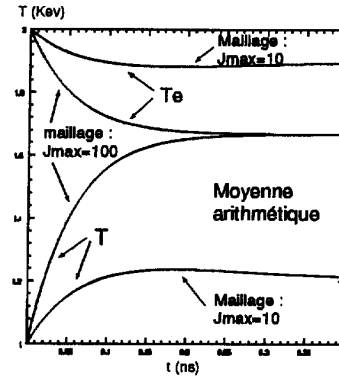


Figure 2. – Températures.

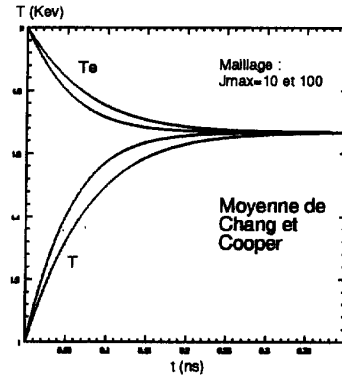


Figure 3. – Températures.

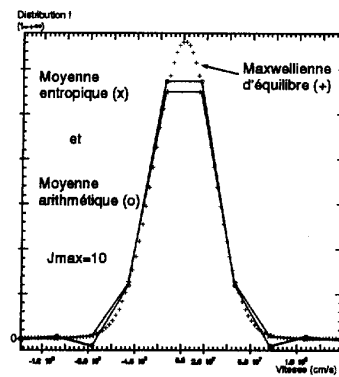


Figure 4. – Distribution  $f$  à  $t = +\infty$ .

Le schéma proposé ici s'étend en géométrie bidimensionnelle axisymétrique sans difficulté. Par ailleurs, on peut impliciter la partie diffusion de (16) en remplaçant  $S(f^n, T_e^n)_j$  par

$$\frac{\Omega^n}{\Delta v} \left[ (v_{j+1/2} - \tilde{U}) \tilde{f}_{j+1/2}^n - (v_{j-1/2} - \tilde{U}) \tilde{f}_{j-1/2}^n \right] + \frac{\Omega^n T_e^n}{m \Delta v^2} A f^{n+1},$$

et l'équilibre thermodynamique  $f^\infty$  est encore solution stationnaire avec ce schéma semi-implicite.

### Références bibliographiques

- [1] Buet C., Cordier St., Numerical Analysis of Conservative and Entropy Scheme for the Fokker-Planck-Landau Equation, SIAM J. Numer. Anal. (à paraître).
- [2] Buet C., Dellacherie S., Sentis R., Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique, Rapport CEA, 1998 et Dellacherie S., Thèse Univ. Paris VII, 1998.
- [3] Casanova M., Larroche O., Matte J.P., Kinetic Simulation of a Collisional Shock Wave in a Plasma, Phys. Rev. Lett. 67 (1991) 2143.
- [4] Chang J.S., Cooper G., A practical Difference Scheme for Fokker-Planck Equations, J. Comput. Phys. 6 (1970) 1-16.
- [5] Cordier S., Degond P., Markovitch D., Schmeiser C., Asymp. Anal. 11 (1995) 209-240 et Browers R., Wilson J.R., Numerical Modelling in Applied Physics, Jones-Bartlett, 1991.
- [6] Decoster A., Fluid Equation and Transport Coefficients of Plasmas, in: Modelling of collisions, P.A. Raviart (Ed.), Masson, 1998.
- [7] Degond P., Lucquin-Desreux B., Transport coefficient of plasmas and disparate mass binary gases, Transport Theory Statist. Phys. 25 (1996) 595.

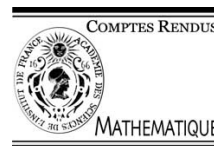


ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®

C. R. Acad. Sci. Paris, Ser. I 338 (2004) 951–956



Numerical Analysis

## Asymptotic preserving scheme and numerical methods for radiative hydrodynamic models

Christophe Buet<sup>a</sup>, Stéphane Cordier<sup>b</sup>

<sup>a</sup> *Département sciences de la simulation et de l'information, Commissariat à l'énergie atomique, BP 12, 91680 Bruyères le Chatel, France*

<sup>b</sup> *UMR MAPMO, CNRS 6628, Université d'Orléans, BP 6759, 45067 Orléans, France*

Received 30 March 2004; accepted 6 April 2004

Available online 7 May 2004

Presented by Philippe G. Ciarlet

---

### Abstract

In this Note, we present a scheme for nonlinear radiative systems which are compatible with diffusive asymptotics. The scheme is based on a splitting: firstly we use a relaxation step to change the problem into 2 identical systems of linear transport systems and, secondly, we use a so-called 'well balanced' scheme for each of the 2 systems. The main advantages of our scheme is that it is fully implicit and compatible with physical properties (positivity); it can be used with a nonconstant cross section and for nonuniform mesh. *To cite this article: C. Buet, S. Cordier, C. R. Acad. Sci. Paris, Ser. I 338 (2004).*

© 2004 Published by Elsevier SAS on behalf of Académie des sciences.

### Résumé

**Analyse asymptotique et méthodes numériques pour les méthodes de moments en hydrodynamique radiative.** Dans cette Note, nous présentons un schéma pour un modèle non linéaire de transfert radiatif, qui soit compatible avec la limite diffusion. Ce schéma est composé de deux étapes : une étape de relaxation qui transforme le système non linéaire en deux systèmes d'équation de transport linéaire identiques et un schéma « équilibre » pour chacun de ces systèmes. L'intérêt principal de notre schéma est d'être totalement implicite, de préserver les propriétés physiques (positivité) et d'être utilisable avec une section efficace variable et un maillage non uniforme. *Pour citer cet article : C. Buet, S. Cordier, C. R. Acad. Sci. Paris, Ser. I 338 (2004).*

© 2004 Published by Elsevier SAS on behalf of Académie des sciences.

---

### Version française abrégée

L'objectif de cette Note est de présenter une discrétisation d'un système hyperboliques de lois de conservation qui soit compatible avec le régime asymptotique de diffusion. On s'intéresse à un système de la forme (1) où  $\varepsilon$  est un petit paramètre.

En effet, lorsque  $\varepsilon \rightarrow 0$ , le système (1) se comporte comme une équation de diffusion dont le coefficient de diffusion est  $1/3\sigma$ ,  $\sigma$  étant la section efficace. Le modèle limite dont les vitesses de propagation sont infinies pose

---

*E-mail addresses:* [Christophe.Buet@cea.fr](mailto:Christophe.Buet@cea.fr) (C. Buet), [Stephane.Cordier@univ-orleans.fr](mailto:Stephane.Cordier@univ-orleans.fr) (S. Cordier).

de sérieux problèmes notamment au niveau numérique et de nombreux travaux ont été réalisés pour lever ces difficultés (limites de flux, facteurs d'Eddington variables...). Cette problématique est très étroitement liée aux études sur les termes sources raides pour les systèmes hyperboliques, les méthodes de relaxation, les schémas dits 'asymptotic preserving'....

La méthode que nous présentons ici est constituée de deux étapes : la première consiste à transformer le système de deux équations non linéaires en un double système (7) de deux équations, linéaires et découplées connus sous le nom d'équations du télégraphe. Il s'agit d'une application des méthodes de relaxation [6] et cela conduit à doubler le nombre d'inconnues. Pour chacun des systèmes obtenus, on utilise ensuite la discrétisation proposée dans [4] que l'on généralise pour une section efficace non constante et un maillage non uniforme.

Le schéma ainsi obtenu est totalement implicite et il a toutes les propriétés requises : consistance lorsque  $\Delta x \rightarrow 0$ , comportement asymptotique lorsque  $\varepsilon \rightarrow 0$ , préservation du domaine invariant (ce qui revient dans les nouvelles variables à garantir la positivité des solutions). Un test numérique avec  $\sigma$  variant de 0 (au centre) à 100 (sur les bords du domaine) illustre cette discrétisation.

L'avantage de la méthode présentée ici est qu'elle peut être utilisée avec une section efficace  $\sigma$  variable, ce qui est très important en pratique car de tels problèmes de transfert radiatif sont couplés avec un modèle hydrodynamique qui va déterminer la valeur de  $\sigma$ . Celle-ci sera d'ordre 1 dans les zones denses ou opaques et pourra être très faible dans les zones dites transparentes. Pour pouvoir utiliser un schéma sans restriction sur les valeurs de  $\sigma$ , il est donc indispensable que la discrétisation ait un bon comportement y compris dans les zones transparentes.

Nous nous sommes également attachés à présenter une méthode utilisable avec un maillage non uniforme car les codes de calcul utilisent des techniques de raffinement de maillage automatique et il est donc important de pouvoir traiter de tels maillages. Par ailleurs, l'extension au cas multidimensionnel et à la prise en compte de terme correctifs, par exemple pour tenir compte d'effet relativiste ou des échanges d'énergie entre les photons et la matière, sera présentée dans [1].

## 1. Introduction

In this Note, we are interested in systems arising from radiative hydrodynamic problems of the following form

$$\varepsilon \partial_t U + \partial_x F(U) = \frac{1}{\varepsilon} R(U), \quad (1)$$

where  $U = (\rho, j)$ ,  $F(U) = (j, \rho h(j/\rho))$ ,  $R(U) = (0, -\sigma j)$ ,  $\sigma(x) > 0$  is the cross section,  $\varepsilon$  is a small parameter and  $h$  is an odd, positive, convex function. The function  $h$  represents, in the radiative transfert problem, the so called Eddington factor. There exists many such functions. Let us mention, for example (see [7] and the reference therein for other examples and more details)

$$h(u) = \frac{3 + 4u^2}{5 + 2\sqrt{4 - 3u^2}}.$$

This function satisfies the following properties:

$$h(0) = \frac{1}{3}, \quad u^2 \leq h(u) \leq 1. \quad (2)$$

Under hypothesis (2), it can be proven that the following properties are preserved under time evolution:

$$\rho \geq 0, \quad \|j\| \leq \rho. \quad (3)$$

This invariant property has a physical interpretation, since  $\rho$  and  $j$  are the two first moments of the distribution function  $\rho = \int f d\omega$ ,  $j = \int f \omega d\omega$ . From the numerical point of view, this property means that the fluxes are the so-called limited fluxes.

In the limit as  $\varepsilon \rightarrow 0$ , a formal asymptotic limit implies that  $j = O(\varepsilon)$  due to the collision term and, at first order in  $\varepsilon$ , using the second equation of (1), we get

$$j = -\frac{\varepsilon}{\sigma} \partial_x (h(0)\rho). \quad (4)$$

Then, using  $h(0) = 1/3$  and dropping this ansatz into the first equation, we obtain the following diffusion approximation

$$\partial_t \rho - h(0) \partial_x \left( \frac{1}{\sigma} \partial_x \rho \right) = 0. \quad (5)$$

Note that the solution for  $\rho$  of the limit heat equation (5) and  $j$  given by (4) does not satisfy automatically condition (3), because the gradient  $\partial_x \rho$  can be arbitrarily large, e.g. if the initial data is discontinuous in  $\rho$ .

The goal of this Note is to present a scheme that is compatible with the limit  $\varepsilon \rightarrow 0$  and with the invariant property (3) and is implicit. It is based on a time splitting, in two steps. The first step is based on a relaxation method and the second on well-balanced schemes [4].

Various methods have been proposed to get rid of these difficulties, such as variable Eddington factors for the so-called *P1*-approximation, or flux limiters for the diffusion approximation (see [7] and references therein). This is also related to a series of papers about asymptotic preserving schemes for kinetic problems, well balanced schemes, stiff source terms and relaxation methods, in the context of hyperbolic systems [4–6,8].

## 2. The asymptotic preserving scheme

Let us now describe one iteration which is decomposed into two steps, first a relaxation step which replaced the nonlinear system by two uncoupled linear systems and, second, the resolution of these systems using a well balanced scheme. This method permits the change of the nonlinear system into a system of two linear systems and makes it possible to use direct implicit solvers.

### 2.1. The relaxed scheme

Let us describe one iteration from  $t = 0$  to  $t = \Delta t$ . We follow the method proposed in [6] which led to the introduction of an artificial vector valued variable  $(z, w)$  and to the expression of a linear system for the variables  $(\rho, z, w, j)$ . The first step is the so called zero relaxation limit, and it can be interpreted as a projection step onto the equilibrium states:

$$z = j, \quad w = \rho h \left( \frac{j}{\rho} \right). \quad (6)$$

The second step consists in solving, during  $\Delta t$ , the transport part:

$$\begin{cases} \partial_t \rho + \frac{1}{\varepsilon} \partial_x z = 0, \\ \partial_t z + \frac{a}{\varepsilon} \partial_x \rho + \frac{\sigma}{\varepsilon^2} z = 0, \\ \partial_t w + \frac{a}{\varepsilon} \partial_x j = 0, \\ \partial_t j + \frac{1}{\varepsilon} \partial_x w + \frac{\sigma}{\varepsilon^2} j = 0. \end{cases} \quad (7)$$

In the relaxation part, the original variables  $(\rho$  and  $j)$  are unchanged. The coefficient  $a$  is constant in space but has to be chosen at each time step in order to recover the correct diffusion coefficient and to insure the stability condition (see Proposition 2.1 in [6]).

Let us emphasize that (7) is just two linear and identical systems, which are uncoupled, one for the quantities  $(\rho$  and  $z)$ , the second for  $(w$  and a new variable  $\bar{j} = aj)$  which are in the form (1) but with  $h \equiv a$ .

Note that this system i.e. (1) with  $h \equiv a$ , once diagonalized, is the well-know Goldstein–Taylor or Telegraph equation with speed  $\pm\sqrt{a}$ :

$$\begin{cases} \partial_t u + \frac{\sqrt{a}}{\varepsilon} \partial_x u = \frac{\sigma}{2\varepsilon^2} (v - u), \\ \partial_t v - \frac{\sqrt{a}}{\varepsilon} \partial_x v = \frac{\sigma}{2\varepsilon^2} (u - v). \end{cases} \quad (8)$$

Setting  $U = \sqrt{a}\rho + z + w + \sqrt{a}j$ ,  $V = \sqrt{a}\rho - z + w - \sqrt{a}j$ ,  $\bar{U} = \sqrt{a}\rho + z - w - \sqrt{a}j$ ,  $\bar{V} = \sqrt{a}\rho - z - w + \sqrt{a}j$ ,  $(U, V)$  and  $(\bar{U}, \bar{V})$  satisfy (8), and we show that for the transport part, the invariant domain (3) comes from the positivity of  $U, V, \bar{U}, \bar{V}$  for a sufficiently large value of  $a$ , [1]. We prove that the choice

$$a = h \left( \max_{x \in \mathbb{R}} \|u(x)\| \right),$$

ensures the positivity of the initial data for  $U, V, \bar{U}, \bar{V}$  and then of the solution  $U, V, \bar{U}, \bar{V}$  of system (7). This choice also provides stability for the relaxed system. In the diffusive limit ( $\varepsilon \rightarrow 0$ ), or for large time behaviour, we expect that  $\max_{x \in \mathbb{R}} \|u(x)\| \rightarrow 0$  and therefore,  $a$  is close to  $1/3$ , i.e. we obtain the right asymptotic (5).

## 2.2. The well-balanced scheme for the transport part

We have now to discretize systems (8) for  $(\rho, z)$  and similarly for  $(w, \sqrt{a}j)$ . We introduce a non-uniform mesh: we note by  $x_i$ , the center of the cell of size  $\Delta x_i$  with  $i \in \mathbb{Z}$  and define  $\Delta x_{i+\frac{1}{2}} = (\Delta x_i + \Delta x_{i+1})/2$ . The proposed discretization is a so called well balanced scheme, which is an extension of the scheme described in [4].

$$\begin{cases} \frac{du_i}{dt} + M_{i-\frac{1}{2}} \frac{\sqrt{a}}{\varepsilon \Delta x_i} (u_i - u_{i-1}) = M_{i-\frac{1}{2}} \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} \frac{\sigma_{i-\frac{1}{2}}}{2\varepsilon^2} (v_i - u_i), \\ \frac{dv_i}{dt} - M_{i+\frac{1}{2}} \frac{\sqrt{a}}{\varepsilon \Delta x_i} (v_{i+1} - v_i) = M_{i+\frac{1}{2}} \frac{\Delta x_{i+\frac{1}{2}}}{\Delta x_i} \frac{\sigma_{i+\frac{1}{2}}}{2\varepsilon^2} (u_i - v_i), \end{cases} \quad (9)$$

where the coefficient  $M_{i+\frac{1}{2}}$  is defined by

$$M_{i+\frac{1}{2}} = \frac{2\sqrt{a}\varepsilon}{\sigma_{i+\frac{1}{2}} \Delta x_{i+\frac{1}{2}} + 2\sqrt{a}\varepsilon}, \quad (10)$$

and  $\sigma_{i+\frac{1}{2}}$  is an arbitrary average of  $\sigma$  at the interface (e.g. arithmetic, harmonic...). The above scheme corresponds to that proposed in [4] for a uniform mesh,  $\sigma = 2$  and a diffusion coefficient in the limit heat equation equal to  $\frac{1}{2}$ . Note that, in our case, the cross section is not assumed to be constant, which is the main interest from an applications point of view.

We can show that (9) is a monotone scheme and then (3) remains an invariant domain during the transport part. It is readily seen that, in the limit  $\max_i (\Delta x_i) \rightarrow 0$ , the coefficient  $M_{i+\frac{1}{2}}$  tends to 1, and the consistency of scheme (9) with the continuous system (8) is satisfied, provided that, in the limit, the mesh is smooth enough, i.e.  $\max_i (\frac{\Delta x_{i+1}}{\Delta x_i}) \rightarrow 0$ .

The scheme (9) can be written in the original variables  $(\rho, z)$  and the same for  $(w, \bar{j})$

$$\begin{cases} \frac{d\rho_i}{dt} + \frac{1}{\varepsilon \Delta x_i} (M_{i+\frac{1}{2}} z_{i+\frac{1}{2}} - M_{i-\frac{1}{2}} z_{i-\frac{1}{2}}) = 0, \\ \frac{dz_i}{dt} + \frac{a}{\varepsilon \Delta x_i} (M_{i+\frac{1}{2}} \rho_{i+\frac{1}{2}} - M_{i-\frac{1}{2}} \rho_{i-\frac{1}{2}}) = \frac{-\lambda_i}{2a\varepsilon^2} z_i + \frac{M_{i+\frac{1}{2}} - M_{i-\frac{1}{2}}}{\varepsilon \Delta x_i} (a\rho_i) \end{cases}$$

with

$$z_{i+\frac{1}{2}} = (z_i + z_{i+1} + \rho_{i+1} - \rho_i)/2, \quad \rho_{i+\frac{1}{2}} = (\rho_i + \rho_{i+1} + z_{i+1} - z_i)/2,$$

and

$$\lambda_i = \frac{\Delta x_{i+\frac{1}{2}}}{\Delta x_i} M_{i+\frac{1}{2}} \sigma_{i+\frac{1}{2}} + \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} M_{i-\frac{1}{2}} \sigma_{i-\frac{1}{2}}.$$

Note that the formulae can be simplified for a uniform mesh and constant cross section. In this case,  $\lambda = 2\sigma M$  and  $M$  is given by (10) with  $\sigma_i = \sigma$  and the mesh size being constant; the second term of the right-hand side vanishes since  $M_{i-\frac{1}{2}} = M_{i+\frac{1}{2}}$ . Then, the proposed scheme reduces to a classical Godunov scheme

$$\varepsilon \frac{dU_i}{dt} + M(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}})/\Delta x = \frac{M}{\varepsilon} R(U_i), \quad (11)$$

where the fluxes at the interfaces are given by

$$F_{i+\frac{1}{2}} = [-a_i F_{i+1} + a_{i+1} F_i - a_i a_{i+1} (U_{i+1} - U_i)]/(a_{i+1} - a_i), \quad (12)$$

which are just the upwind fluxes for (8) with  $a_{i+1} = -a_i = \sqrt{a}$  the characteristic speeds.

This form is useful because it is expressed in the original variable and this can be easily generalized to the nonlinear case. The discretization is a sum of a diffusive term and classical convective term. Thus, this discretization can be interpreted as a particular choice of adding a numerical viscosity term depending on  $\varepsilon$ .

Moreover, when  $\varepsilon \rightarrow 0$ , we show that the scheme is an approximation of the heat equation with a diffusion coefficient equal to  $1/3$  for radiative hydrodynamic applications.

### 2.3. Time discretization

We claim that a fully implicit time discretization is suitable. Indeed, a partial implicit time discretization as described in [4] led to a prohibitive parabolic CFL (time step restriction)  $\sigma \Delta t \leq (\Delta x)^2$  to ensure monotonicity. Note that the expected time step restriction with an explicit method for the transport part (CFL condition) is much more restrictive  $\frac{\Delta t}{\Delta x} \leq \varepsilon$ , and, similarly, the characteristic relaxation time is such that  $\Delta t \sigma \leq \varepsilon^2$ . The method proposed here allows us to use a fully implicit time discretization for the system of linear equations. Moreover, we prove that the properties of the invariant domain and asymptotic behaviour are the same as for time continuous discretization (9) (see [1] for more details).

## 3. Numerical results

We now illustrate our scheme on the following test. The domain is  $[0, 2]$ . The cross section is vanishing in the middle of the domain and is very large at the boundary.

$$\sigma(x) = 100(x-1)^4, \quad x \in ]0, 2[.$$

The initial data is a characteristic function, for  $\rho$  with support in  $[\frac{1}{2}, \frac{3}{2}]$ . The initial flux  $j$  is equal to 0. The simulated time is  $T = 0.1$  and the small parameter value takes the following values:  $\varepsilon = 0$  (diffusion),  $10^{-2,1,0}$ . The mesh is uniform with either 100 or 1000 points and the time step is chosen such that  $\Delta t/\Delta x = 0.05$  – see Fig. 1.

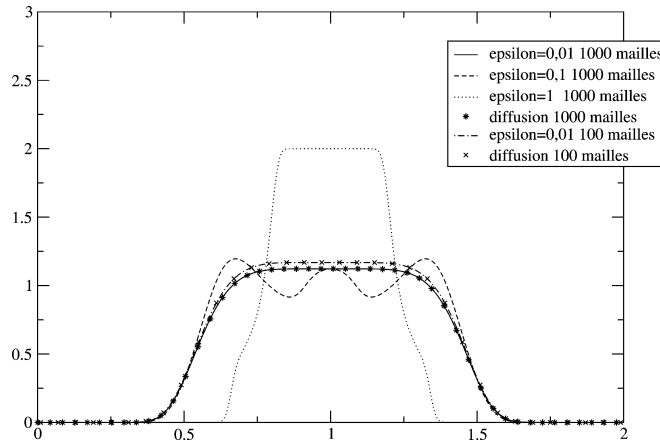


Fig. 1. Numerical example using test data.

Fig. 1. Exemple numérique avec des données d'essai.

#### 4. Conclusions

It is also possible to extend this approach to multi-dimensional problem (including Adaptive Mesh Refinement). Lastly, it is also possible to take into account more complex source terms, such as the terms coming from relativistic effects, presented in [2], or the coupling with a hydrodynamical model due to energy transfert.

Let us also mention that the above method can be used for collisional kinetic model in diffusive regimes (like the Lorentz model presented in [3]) at least in the case of constant cross section and uniform mesh in space.

#### References

- [1] C. Buet, S. Cordier, Asymptotic Preserving Scheme for radiative hydrodynamics model, 2004, in preparation.
- [2] C. Buet, B. Desprès, Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics, Preprint, HYKE2003-016, <http://hyke.org>, J. Quant. Spectrosc. Radiat. Transf. (2004), in press.
- [3] C. Buet, S. Cordier, L. Lucquin-Desreux, S. Mancini, Diffusion limits of the Lorentz model: asymptotic preserving schemes, Methods Math. Anal. Numer. 36 (4) (2002) 631–655.
- [4] L. Gosse, G. Toscani, An asymptotic preserving well-balanced scheme for the hyperbolic heat equation, C. R. Acad. Sci. Paris, Ser. I 334 (4) (2002) 337–342.
- [5] S. Jin, C.D. Levermore, Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation, J. Comput. Phys. 126 (1996) 449–467.
- [6] S. Jin, Z. Xin, The relaxation schemes for systems of conservation laws in arbitrary space dimensions, Comm. Pure Appl. Math. 48 (1995) 235–276.
- [7] C.D. Levermore, Relating Eddington factors to flux limiters, J. Quant. Spectrosc. Radiat. Transf. 31 (2) (1984) 149–160.
- [8] G. Naldi, L. Pareschi, Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation, SIAM J. Numer. Anal. 37 (2000) 1246–1270.

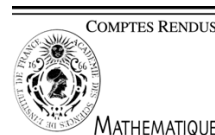




Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



C. R. Acad. Sci. Paris, Ser. I 340 (2005) 399–404



<http://france.elsevier.com/direct/CRASS1/>

Numerical Analysis

# On the non existence of monotone linear schema for some linear parabolic equations

Christophe Buet<sup>a</sup>, Stéphane Cordier<sup>b</sup>

<sup>a</sup> *Département sciences de la simulation et de l'information, Commissariat à l'énergie atomique, BP 12, 91680 Bruyères le Chatel, France*

<sup>b</sup> *UMR MAPMO – CNRS 6628, BP 6759, université d'Orléans, 45067 Orléans, France*

Received 26 November 2004; accepted 2 January 2005

Presented by Philippe G. Ciarlet

## Abstract

In this Note, we present a result concerning the non existence of linear monotone schema with fixed stencil on regular meshes for some linear parabolic equation in two dimensions. The parabolic equations of interest arise from non isotropic diffusion modelling. A corollary is that no linear monotone 9 points-schemes can be designed for the one-dimensional heat equation emerged in the plane with an arbitrary direction of diffusion. Some applications of this result are provided: for the Fokker–Planck–Lorentz model for electrons in the context of plasma physics; all linear monotone scheme for the one-dimensional hyperbolic heat equation treated as a two-dimensional problem are not consistent in the diffusion limit for an arbitrary direction of propagation. We also examine the case of the Landau equation. **To cite this article:** C. Buet, S. Cordier, *C. R. Acad. Sci. Paris, Ser. I 340 (2005)*.

© 2005 Académie des sciences. Published by Elsevier SAS. All rights reserved.

## Résumé

**Sur la non existence de schémas linéaires monotones pour certaines équations paraboliques linéaires.** Dans cette Note, nous présentons un résultat de non existence de schémas linéaires monotones avec un stencil fixé sur un maillage carré pour certaines équations paraboliques en dimension 2. Les équations paraboliques que l'on considère proviennent de modèles de diffusion anisotrope. Une conséquence du résultat est qu'il n'existe pas de schémas linéaires monotones à neuf points pour l'équation de la chaleur monodimensionnelle immergée dans le plan, avec une direction arbitraire. Nous présentons quelques applications : à l'équation de Fokker–Planck–Lorentz pour les électrons dans le contexte de la physique des plasmas ; Un schéma linéaire monotone pour l'équation de la chaleur hyperbolique monodimensionnelle et traité comme un problème bidimensionnel ne peut pas être consistant dans la limite de diffusion pour une direction arbitraire de propagation. On examine aussi le cas de l'équation de Landau. **Pour citer cet article :** C. Buet, S. Cordier, *C. R. Acad. Sci. Paris, Ser. I 340 (2005)*.

© 2005 Académie des sciences. Published by Elsevier SAS. All rights reserved.

*E-mail addresses:* [Christophe.Buet@cea.fr](mailto:Christophe.Buet@cea.fr) (C. Buet), [Stephane.Cordier@univ-orleans.fr](mailto:Stephane.Cordier@univ-orleans.fr) (S. Cordier).

1631-073X/\$ – see front matter © 2005 Académie des sciences. Published by Elsevier SAS. All rights reserved.  
doi:10.1016/j.crma.2005.01.020

### Version française abrégée

On s'intéresse à l'approximation numérique de l'équation parabolique de la forme (1).

Dans le cas où  $c = ab$ , cette équation représente l'équation de la chaleur monodimensionnelle immergée en dimension 2. La diffusion agit donc uniquement dans la direction  $(a, b)$ . Il est bien connu que cette équation vérifie un principe du maximum : soit deux données initiales telles que  $f_0 \geq g_0$  alors  $f \geq g$ .

On considère une grille de calcul cartésienne à mailles carrées de pas d'espace  $h$ .

Nous montrons alors (Proposition 2.2) que, pour un stencil fixe de la forme (5), il existe toujours des directions  $(a, b)$  de diffusion pour lesquelles on ne peut construire de schéma linéaire monotone. La preuve est basée sur l'analyse de l'erreur de consistance.

Ce résultat est encore valide pour une matrice de diffusion dépendant de la variable d'espace (Proposition 3.1) donc en particulier pour l'équation de Fokker–Planck–Lorentz (11) issue de la physique des plasmas et représentant un modèle simplifié des collisions avec les ions pour les électrons. Nous montrons de plus que, pour l'équation du télégraphe (13), il ne peut y avoir de schémas linéaires monotones consistant avec la limite de diffusion (1) avec  $c = ab$ , pour une direction arbitraire de propagation.

Nous indiquons enfin comment ce résultat pourrait signifier que pour l'équation de Fokker–Planck–Landau (14), il ne peut y avoir de schémas positifs pour une discrétisation de la forme naturelle c'est à dire de schémas quadratiques positifs, ce qui justifierait l'emploi d'algorithmes vraiment non linéaires par exemple basés sur la forme dite logarithmique (16), voir par exemple [3].

## 1. Introduction

In this Note, we are interested in two-dimensional linear parabolic equations of the following form

$$\partial_t f - \frac{1}{2} \nabla \cdot K \nabla f = 0, \quad (1)$$

in  $\mathbb{R}^2$  for the space variable i.e. with  $f(x, y)$ . The matrix  $K$  is supposed symmetric positive which ensure that the solutions of the Cauchy problem for Eq. (1) verify the principle of maximum

$$f_0 \geq g_0 \Rightarrow f \geq g. \quad (2)$$

For a constant matrix

$$K = \begin{pmatrix} a^2 & c \\ c & b^2 \end{pmatrix}, \quad (3)$$

with  $c = ab$  Eq. (1) is just a monodimensional heat equation with diffusion along the  $(a, b)$  direction.

## 2. Main result

Let us assume that the matrix  $K$  is constant positive i.e. the coefficients verify  $c^2 \leq a^2 b^2$ . We consider Cartesian mesh of  $\mathbb{R}^2$  of size  $h$ . The quadrature points are of the form  $(x, y)_i = ih$  with  $i = (i_1, i_2) \in \mathbb{Z}^2$  and we note  $f_i$  the approximated value of  $f$  at the points  $ih$ . We consider a generic linear scheme for Eq. (1) of the form

$$\frac{d}{dt} f_i = \frac{1}{h^2} \sum_{j \in S} a_j(h) f_{i+j} \quad (4)$$

with the function  $a_j(h)$  continuous and  $S$  represents the stencil,

$$S = \{j \text{ such that } -N \leq j_k \leq N, \ k = 1, 2\}, \quad (5)$$

$N$  is an integer which relies on the size of the stencil.  $N = 1$  represents 9-points scheme and  $N = 2$  a 25-points scheme.

All type of linear schemes for Eq. (1) such as finite volume, finite elements or finite differences schemes can be rewritten in the form (4).

We recall the definition of the consistency (see by example [8] or [6]):

**Definition 2.1.** Let us define the consistency error for the scheme (4) for a sufficiently smooth function  $f$  as

$$\begin{aligned} E(f, h) = & \frac{1}{h^2} \left( \sum_{j \in S} a_j(h) \right) f + \frac{1}{h} \left[ \left( \sum_{j \in S} j_1 a_j(h) \right) \partial_x f + \left( \sum_{j \in S} j_2 a_j(h) \right) \partial_y f \right] \\ & + \left[ \left( \sum_{j \in S} j_1^2 a_j(h) - a^2 \right) \partial_{xx} f + 2 \left( \sum_{j \in S} j_1 j_2 a_j(h) - c \right) \partial_{xy} f + \left( \sum_{j \in S} j_2^2 a_j(h) - b^2 \right) \partial_{yy} f \right] \\ & + O(h). \end{aligned} \quad (6)$$

A scheme is called consistent for Eq. (1) provided that the consistency error (6) is a  $o(1)$  i.e. tends to 0 as  $h \rightarrow 0$ .

Our main result concerning the monotonicity of the scheme (4) is the following:

**Proposition 2.2.** For every size of the stencil  $N$ , there is no linear monotone consistent scheme for a matrix  $K$  such that  $N \min(a^2, b^2) < |c| \leq |ab|$ .

**Proof.** We consider only the case  $0 \leq a \leq b$  and  $c > 0$ , the other cases are obtained by symmetry.

Using the definition of the consistency for some particular choice of  $f$ , the leading order term in (6) for constant function gives  $\sum_{j \in S} a_j(h) = o(h^2)$ . Then, using functions of the form  $Cx$  and  $Cy$  respectively, one gets

$$\sum_{j \in S} j_1 a_j(h) = o(h), \quad \sum_{j \in S} j_2 a_j(h) = o(h). \quad (7)$$

When  $h \rightarrow 0$ , the zeroth order term in the consistency error gives necessarily

$$\sum_{j \in S} j_1^2 a_j(0) = a^2, \quad (8)$$

$$\sum_{j \in S} j_1 j_2 a_j(0) = c, \quad (9)$$

$$\sum_{j \in S} j_2^2 a_j(0) = b^2. \quad (10)$$

Subtracting  $\frac{1}{N} \times (9)$  from (8) gives

$$\sum_{j \in S} j_1 \left( j_1 - \frac{j_2}{N} \right) a_j(0) = a^2 - \frac{c}{N}.$$

Note that all terms  $j_1(j_1 - \frac{j_2}{N})$  are positive and are not identically vanishing. On the other side, we have  $a^2 - \frac{c}{N} < 0$ . Thus, this implies that there is at least one index  $j_0 \neq (0, 0)$  such that  $a_{j_0}(0) < 0$ .

By continuity  $a_{j_0}(h) < 0$  for sufficiently small  $h < h_0$ . Thus, for all  $h < h_0$ , considering an initial data of the form  $f_i = 0$  for any index  $i$  except at the point  $i_0$  at which it is strictly positive, the solution of (4) is negative for  $t$  sufficiently small at point  $i_0 - j_0$ .

In conclusion, for  $h < h_0$  the scheme is not positive and, since it is linear, it is not monotone. This ends the proof.  $\square$

**Remark 1.** In the case  $c = ab$ , there is no linear monotone consistent 9-points ( $N = 1$ ) scheme except for a direction of diffusion aligned with the axis direction or with the principal diagonal i.e.  $(a, b) \in \{(\pm 1, 0), (0, \pm 1), (\pm 1, \pm 1)\}$ .

Always in the case  $c = ab$ , there is no 25-points linear monotone and consistent scheme if  $4 \min(a^2, b^2) < \max(a^2, b^2)$  and this is half of the possible directions. For the others cases, we have no response. Note that 25-points are rarely used for diffusion equations in 2-D.

**Remark 2.** Another way of to view this result is: if one wants to construct a monotone consistent scheme for an arbitrary matrix  $K$  one must consider a nonlinear scheme or use an infinite stencil, for example a stencil growing when refining the mesh size.

**Remark 3.** In the case  $c = ab$ , for the directions that are not concerned with Proposition 2.2, we can always construct a linear monotone and consistent scheme using points in the stencil which belongs to the direction  $(\pm 1, \pm N)$  and  $(\pm 1, \pm 1)$ .

### 3. Applications

#### 3.1. The Fokker–Planck–Lorentz equation for electrons

We consider now the case of  $K$  depending of the space variables  $x, y$ . A particular example of such an equation comes from the plasma physics: the Lorentz operator. Lorentz operators appear for example when considering elastic collisions of heavy particles (e.g. ions) against light ones (e.g. electrons). It is the first order term of the inter-species collision operator representing the collisions of the heavy particles on the light one. An asymptotic expansion in terms of the small mass ratio can be found in [4,5]. In this case, in 2-D, the matrix  $K$  is of the form

$$K(x, y) = \Psi\left(\frac{1}{x^2 + y^2}\right)(\text{Id} - (x, y)^t \otimes (x, y)) \quad (11)$$

where  $\Psi$  is a positive function. This can be interpreted as the heat equation over each sphere centered at 0 and is also called Laplace–Beltrami operator. The natural way to discretize such an equation would to consider a spherical mesh. The question we shall address here is what happens when a Cartesian grid is used.

The result of Proposition 2.2 is still valid for the matrix, depending on the space variables:

**Proposition 3.1.** *For a fixed stencil  $N$ , there is no linear monotone consistent scheme provided that there exists some open set in which the matrix  $K(x, y)$  is such that  $N \min(a^2, b^2) < |c| \leq |ab|$ .*

The result also holds for any mesh obtained using a smooth map from the uniform grid.

**Proof.** We consider only the case  $0 \leq a \leq b$ ; the other cases are obtained by symmetry. Without loss of generality, we consider a point  $(x_0, y_0)$  which belongs to all the meshes and such that  $\min(a^2, b^2) \leq N|c|$  at this point.

We write the diffusion operator in the following (expanded) form.

$$\nabla \cdot K \nabla f = a^2 \partial_{xx} f + 2c \partial_{xy} f + b^2 \partial_{yy} f + d \partial_x f + e \partial_y f$$

the truncature error is as in the constant case, except that now the functions  $a_j$  depends now on  $x$  and  $y$  and relations (7) are replaced by

$$\sum_{j \in S} j_1 a_j(h) - d = o(h), \quad \sum_{j \in S} j_2 a_j(h) - e = o(h) \quad (12)$$

but we still have relations (8)–(10). At point  $(x_0, y_0)$ , we conclude as for the constant case.  $\square$

The matrix  $K$  of Fokker–Planck–Lorentz model, see (11), satisfies the hypothesis of Proposition 3.1. Thus, for the Lorentz equation on a square mesh, there is no linear consistent monotone scheme.

### 3.2. Asymptotic preserving schemes for the hyperbolic heat equation

Let us consider the following hyperbolic heat equation in 2-D:

$$\begin{aligned}\varepsilon \partial_t u + a \partial_x u + b \partial_y u &= \frac{1}{\varepsilon} (v - u), \\ \varepsilon \partial_t v - a \partial_x v - b \partial_y v &= \frac{1}{\varepsilon} (u - v)\end{aligned}\quad (13)$$

with  $a$  and  $b$  constant such that  $a^2 + b^2 = 1$ . This is just the monodimensional hyperbolic heat equation along the direction  $(a, b)$ . The monodimensional system is also referred to as telegraph equations or Goldstein–Taylor equation.

It is well known that the asymptotic regime, when  $\varepsilon \rightarrow 0$  of this system is of the type (1) for the function  $f = u + v$  with the matrix  $K$ ,  $c = ab$ . In 1-D, or with a direction aligned with one of the axes of the meshes in 2-D, it is possible to construct a monotone linear scheme which is ‘asymptotic preserving’, that is, in the limit  $\varepsilon \ll 1$ , the scheme gives a linear monotone discretization of the 1-D heat equation; see, for example, [7].

Let us now assume that one uses a first order monotone finite volume method on a square mesh of size  $h$  to solve (13). Naturally, this scheme is a 9-point schemes. In the diffusive regime, that is for  $h \gg \varepsilon$  or for  $\frac{h}{\varepsilon}$  fixed and  $\varepsilon \ll 1$ , one obtains, at most, a 25-point scheme for  $f = u + v$ , and this scheme is also monotone.

However, Proposition 2.2 implies that there is no monotone linear consistent scheme for Eq. (1) with a direction of diffusion  $(a, b)$  such that  $4 \min(a^2, b^2) < \max(a^2, b^2)$ . Consequently, the scheme we used for the hyperbolic heat equation (13) cannot be asymptotic preserving for an arbitrary direction  $(a, b)$ : it leads to a non consistent scheme for  $f = u + v$  if  $4 \min(a^2, b^2) < \max(a^2, b^2)$ , or in other words, there is an isotropic numerical diffusion of size  $O(1)$  in these cases.

### 3.3. The Landau equation

Let us consider now the Fokker–Planck–Landau equation modeling self collisions for charged particles:

$$\frac{d}{dt} f = \nabla_v \cdot \int_{v'} \phi(v - v') (f' \nabla_v f - f \nabla_{v'} f') dv' \quad (14)$$

with  $f' = f(v')$ ,  $f = f(v)$ ,  $v, v' \in \mathbb{R}^3$  and

$$\phi(v - v') = \frac{1}{|v - v'|^3} (|v - v'|^2 \mathbf{Id} - (v - v') \otimes (v - v'))$$

is indeed symmetric and positive.

Solutions of (14) are expected to be positive since  $f$  is the function of distribution of the charged particles. This equation can also be written as

$$\frac{d}{dt} f = \nabla_v \cdot K(f) \nabla_v f + \nabla_v \cdot (C(f) f) \quad (15)$$

and the matrix  $K$  and the vector  $C$  (which are derivatives of the potentials of Rosenbluth), are defined by

$$\begin{aligned}K(f)(v) &= \int_{v'} \phi(v - v') f' dv', \\ C(f)(v) &= \int_{v'} \phi(v - v') \nabla_{v'} f' dv'\end{aligned}$$

and the matrix  $K$  is indeed symmetric positive. The collision term is quadratic, thus a natural discretization of (14) is quadratic that is, at each point of the discretization, the discretized Landau operator reads:

$$\sum_{j \in S} \left( \sum_{k \in \mathbb{Z}^3} a_{i+j,k}(h) f_k \right) f_{i+j}.$$

Proposition 3.1 suggests that such a scheme could never be a positive scheme, see [1] for such an example. This would justify the use of the fully nonlinear writing of the Landau equation, the so called Log form

$$\frac{d}{dt} f = \nabla_v \cdot \int_{v'} f f' \phi(v - v') (\nabla_v \log(f) - f \nabla_{v'} \log(f')) dv' \quad (16)$$

in order to derive a positive scheme (see [3,2]).

## References

- [1] C. Buet, S. Cordier, Numerical analysis of conservative and entropy schemes for the FPLE, *SIAM J. Numer. Anal.* 36 (1999) 953.
- [2] C. Buet, S. Cordier, P. Degond, M. Lemou, Fast algorithms for the Fokker–Planck equation, *J. Comput. Phys.* 133 (1997) 310–322.
- [3] P. Degond, B. Lucquin-Desreux, An entropy scheme for the Fokker–Planck collision operator of plasma kinetic theory, *Numer. Math.* 68 (2) (1994) 239–262.
- [4] P. Degond, B. Lucquin-Desreux, The Fokker–Planck asymptotics of the Boltzmann collision operator in the Coulomb case, *Math. Models Methods Appl. Sci.* 2 (2) (1992) 167–182.
- [5] P. Degond, B. Lucquin-Desreux, The asymptotics of collision operators for two species of particles of disparate masses, *Math. Models Methods Appl. Sci.* 6 (3) (1996) 405–436.
- [6] R. Eymard, T. Gallouet, R. Herbin, Finite Volume Methods, in: *Handbook of Numerical Analysis*, vol. VII, 2000, pp. 713–1020.
- [7] L. Gosse, G. Toscani, An asymptotic preserving well-balanced scheme for the hyperbolic heat equation, *C. R. Acad. Sci. Paris, Ser. I* 334 (2002) 1–6.
- [8] B. Lucquin-Desreux, *Equations aux dérivées partielles et leurs approximations*, Ellipses, 2004.

# On the Chang and Cooper numerical scheme applied to a linear Fokker-Planck equation

C. Buet<sup>1</sup> and S. Dellacherie<sup>2</sup>

Commissariat à l'Énergie Atomique

<sup>1</sup>Centre d'étude de Bruyres le Châtel

91680 Bruyres le Châtel, France

<sup>2</sup>Centre d'étude de Saclay

91191 Gif sur Yvette, France

e-mail: sdellacherie@cea.fr

## Abstract

In this article, we show that for a particular linear Fokker-Planck operator, the explicit Chang and Cooper scheme preserves the positivity of the distribution and allows the distribution to converge toward the thermodynamical equilibrium by preserving the decreasing of the entropy under a classical CFL criteria.

## Introduction

The Chang and Cooper scheme (cf. [1]) is a classical scheme (cf. [2], [3] and [4]) used to solve a kinetic equation of the type

$$\partial_t f = S(f)$$

where  $S(f)$  is a Fokker-Planck operator. The main property of the Chang and Cooper scheme is that, at least in the linear case, the numerical fluxes are equal to zero when the distribution  $f$  is equal to the equilibrium distribution: in other words, this scheme preserves the thermodynamical equilibrium when it is reached.

Although, up to now, it exists no convergence properties in large time toward this equilibrium. Then, we will show in that article that for a particular linear Fokker-Planck operator, the explicit Chang and Cooper scheme has good convergence properties.

The Fokker-Planck operator studied in that paper is defined by

$$S(f)(v) = \Omega \nabla_v \cdot \left[ (\vec{v} - \vec{U}_e) f(v) + \frac{T_e}{m} \nabla_v f \right] \quad (1)$$

for the convection-diffusion form;  $\vec{U}_e$ ,  $T_e > 0$  (which respectively are the velocity and the temperature of the medium),  $m > 0$  and  $\Omega > 0$  (which respectively are the atomic mass of the particle and the collision frequency

of the particle on the medium) are constants. Let us note that this operator is the linear version of the non linear ion-electron collision operator studied in [5] in which  $\Omega$ ,  $\vec{U}_e$  and  $T_e$  depends upon the time. By defining the equilibrium distribution

$$\mathcal{M}_{N, \vec{U}_e, T_e}(v) = \frac{N}{(2\pi T_e/m)^{3/2}} \exp \left[ -\frac{m(\vec{v} - \vec{U}_e)^2}{2T_e} \right],$$

we can also define  $S(f)$  with the two equivalent forms that is to say with

$$S(f) = \Omega \frac{T_e}{m} \nabla_v \cdot [f \nabla_v \log(f/\mathcal{M}_{N, \vec{U}_e, T_e})]$$

which is called the Landau form, and with

$$S(f) = \Omega \frac{T_e}{m} \nabla_v \cdot [\mathcal{M}_{N, \vec{U}_e, T_e} \nabla_v (f/\mathcal{M}_{N, \vec{U}_e, T_e})] \quad (2)$$

which is called the non logarithmic Landau form. The explicit Chang and Cooper scheme (cf. [1]) is built in order to make the discretised fluxes  $(\vec{v} - \vec{U}_e)f(v) + \frac{T_e}{m} \nabla_v f$  equal to zero when  $f = \mathcal{M}_{N, \vec{U}_e, T_e}$  which implies that this scheme preserves the thermodynamical equilibrium when it is reached.

In this article, we will show that by using the explicit Chang and Cooper numerical scheme to discretize the Fokker-Planck operator (1), the distribution  $f$  converges toward the thermodynamical equilibrium  $\mathcal{M}_{N, \vec{U}_e, T_e}$  in large time under a classical CFL criteria. For a seak of simplicity, we define the Fokker-Planck operator  $S(f)$  in cartesian geometry, the microscopic velocity having only one dimension. Then, we replace now  $(2\pi T_e/m)^{3/2}$  with  $\sqrt{2\pi T_e/m}$  and  $\nabla_v$  with  $\partial_v$ .

The velocity space is discretized with the series  $(v_j)$  where  $j \in \{1, \dots, j_{\max}\}$ ; the velocity step is constant and equal to  $\Delta v$  and finally, we define

$$<g> \equiv \sum_j g(v_j) \Delta v.$$

The time subscript is  $n$  and the time step is defined with  $\Delta t$ .

## 1 The explicit Chang and Cooper scheme

The explicit Chang and Cooper scheme is defined by

$$\frac{1}{\Delta t} (f_j^{n+1} - f_j^n) = S(f^n)_j \quad (3)$$

with

$$\begin{aligned} S(f^n)_j = & \frac{\Omega}{\Delta v} \left[ (v_{j+1/2} - U_e) \tilde{f}_{j+1/2}^n - (v_{j-1/2} - U_e) \tilde{f}_{j-1/2}^n \right] \\ & + \frac{\Omega T_e}{m \Delta v^2} (a_j f_{j+1}^n - b_j f_j^n + c_j f_{j-1}^n) \end{aligned} \quad (4)$$

where  $a_j = c_j = 1$  et  $b_j = 2$  except at the frontier of the velocity domain (see below).  $\tilde{f}_{j+1/2}^n$  is an approximation of  $f(t = t_n, v = v_{j+1/2})$  defined with the following definition (cf. [1]):



**Definition** The Chang and Cooper average  $\tilde{f}_{j+1/2}$  of the quantities  $f_j$  and  $f_{j+1}$  is defined by

$$\tilde{f}_{j+1/2} = \delta_{j+1/2} f_j + (1 - \delta_{j+1/2}) f_{j+1}$$

with

$$\delta_{j+1/2} = \frac{1}{w_{j+1/2}} - \frac{1}{\exp(w_{j+1/2}) - 1}$$

where

$$w_{j+1/2} = \frac{m\Delta v}{T_e} (v_{j+1/2} - U_e).$$

**Boundary conditions and the mass conservation** To make the scheme conservative, we have to impose boundary conditions of the type Robin that is to say we impose on the boundary velocity domain

$$(v - U_e)f + \frac{T_e}{m} \partial_v f = 0$$

which is equivalent to define for the numerical scheme

$$\begin{cases} a_j = 1 \text{ si } j \neq j_{\max}, \\ b_j = 2 \text{ si } j \in \{2, \dots, j_{\max} - 1\}, \\ c_j = 1 \text{ si } j \neq 1, \\ b_1 = b_{j_{\max}} = 1 \text{ and } a_{j_{\max}} = c_1 = 0 \end{cases} \quad (5)$$

and

$$\tilde{f}_{1/2} = \tilde{f}_{j_{\max}+1/2} \equiv 0. \quad (6)$$

We have the following conservation property:

**Property 1.1**

$$\langle f_j^{n+1} \rangle = \langle f_j^n \rangle.$$

Let us now introduce the following notations:

**Notation** We define

$$\widehat{\mathcal{M}}_{f^0, j+1/2} = \frac{\mathcal{M}_{f^0, j+1} \mathcal{M}_{f^0, j}}{\mathcal{M}_{f^0, j+1/2}}$$

where  $\widehat{\mathcal{M}}_{f^0, j+1/2}$  is the entropic average of  $\mathcal{M}_{f^0, j+1}$  and of  $\mathcal{M}_{f^0, j}$  that is to say

$$\widehat{\mathcal{M}}_{f^0, j+1/2} = \frac{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, j}}.$$

The entropic average was firstly introduce to discretize the ion-electron collision operator (see [5]) and the isotropic ion-ion collision operator (see [6]; see also the first part of [7] for more details). Now, we establish the following property and lemma:

**Property 1.2** When  $\tilde{f}_{j+1/2}^n$  is the Chang and Cooper average of  $f_j^n$  and of  $f_{j+1}^n$ , we can define  $\tilde{f}_{j+1/2}^n$  in the following way

$$\begin{aligned} \tilde{f}_{j+1/2}^n &= \frac{f_{j+1}^n - f_j^n}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, T_e, j}} \\ &+ \left( \frac{f_j^n}{\mathcal{M}_{f^0, j}} - \frac{f_{j+1}^n}{\mathcal{M}_{f^0, j+1}} \right) \cdot \frac{\mathcal{M}_{f^0, j+1} \mathcal{M}_{f^0, j}}{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}}. \end{aligned} \quad (7)$$

**Lemma 1.1** When  $\tilde{f}_{j+1/2}^n$  is the Chang and Cooper average of  $f_j^n$  and of  $f_{j+1}^n$ , the operator  $S(f^n)_j$  defined with (4) can be written with

$$S(f^n)_j = \frac{\Omega T_e}{m \Delta v^2} \left\{ \widehat{\mathcal{M}}_{f^0, j+1/2} [(f^n / \mathcal{M}_{f^0})_{j+1} - (f^n / \mathcal{M}_{f^0})_j] - \widehat{\mathcal{M}}_{f^0, j-1/2} [(f^n / \mathcal{M}_{f^0})_j - (f^n / \mathcal{M}_{f^0})_{j-1}] \right\} \quad (8)$$

where the boundary conditions (5) and (6) are replaced with the boundary conditions

$$f_0^n \equiv f_1^n \cdot \frac{\mathcal{M}_{f^0, 0}}{\mathcal{M}_{f^0, 1}} \quad \text{and} \quad f_{j_{\max}+1}^n \equiv f_{j_{\max}}^n \cdot \frac{\mathcal{M}_{f^0, j_{\max}+1}}{\mathcal{M}_{f^0, j_{\max}}}. \quad (9)$$

In other words, the Chang and Cooper average makes equivalent from a discretized point of view the convection-diffusion form (1) and the non logarithmic Landau form (2). More over, let us remark that the boundary conditions (9) are equivalent from a discretized point of view to the Robin type boundary conditions

$$\partial_v(f / \mathcal{M}_{f^0}) = 0.$$

**Proof of the property 1.2** We easily verify that

$$w_{j+1/2} \equiv \frac{m \Delta v}{T_e} (v_{j+1/2} - U_e) = -[(\log \mathcal{M}_{f^0})_{j+1} - (\log \mathcal{M}_{f^0})_j]. \quad (10)$$

Then

$$\delta_{j+1/2} = -\frac{1}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, j}} + \frac{\mathcal{M}_{f^0, j+1}}{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}}.$$

Then

$$\tilde{f}_{j+1/2}^n = \frac{f_{j+1}^n - f_j^n}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, j}} + \mathcal{M}_{f^0, T_e, j+1} \cdot \frac{f_j^n - f_{j+1}^n}{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}} + f_{j+1}^n$$

which shows that

$$\tilde{f}_{j+1/2}^n = \frac{f_{j+1}^n - f_j^n}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, j}} + \left( \frac{f_j^n}{\mathcal{M}_{f^0, j}} - \frac{f_{j+1}^n}{\mathcal{M}_{f^0, j+1}} \right) \cdot \frac{\mathcal{M}_{f^0, j+1} \mathcal{M}_{f^0, j}}{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}}.$$

□

**Proof of the lemma 1.1** By using (7), we obtain that when  $j \notin \{1, j_{\max}\}$

$$\begin{aligned} S(f^n)_j &= \frac{\Omega}{\Delta v} \left( \frac{f_j^n}{\mathcal{M}_{f^0, j}} - \frac{f_{j+1}^n}{\mathcal{M}_{f^0, j+1}} \right) \cdot \frac{\mathcal{M}_{f^0, j+1} \mathcal{M}_{f^0, j}}{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}} (v_{j+1/2} - U_e) \\ &\quad - \frac{\Omega}{\Delta v} \left( \frac{f_{j-1}^n}{\mathcal{M}_{f^0, j-1}} - \frac{f_j^n}{\mathcal{M}_{f^0, j}} \right) \cdot \frac{\mathcal{M}_{f^0, j} \mathcal{M}_{f^0, j-1}}{\mathcal{M}_{f^0, j} - \mathcal{M}_{f^0, j-1}} (v_{j-1/2} - U_e) \\ &+ \frac{\Omega}{\Delta v} \left[ \frac{f_{j+1}^n - f_j^n}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, j}} (v_{j+1/2} - U_e) - \frac{f_j^n - f_{j-1}^n}{\log \mathcal{M}_{f^0, j} - \log \mathcal{M}_{f^0, j-1}} (v_{j-1/2} - U_e) \right] \\ &\quad + \frac{\Omega T_e}{m \Delta v^2} (f_{j+1}^n - 2f_j^n + f_{j-1}^n). \end{aligned}$$

But, by taking into account (10), we can write that

$$\frac{\mathcal{M}_{f^0, j+1} \mathcal{M}_{f^0, j}}{\mathcal{M}_{f^0, j+1} - \mathcal{M}_{f^0, j}} (v_{j+1/2} - U_e) = -\frac{T_e}{m \Delta v} \widehat{\mathcal{M}}_{f^0, j+1/2}$$

and that

$$\frac{(v_{j+1/2} - U_e)}{\log \mathcal{M}_{f^0, j+1} - \log \mathcal{M}_{f^0, j}} = -\frac{T_e}{m\Delta v}.$$

Then, we have

$$\begin{aligned} S(f^n)_j &= \frac{\Omega T_e}{m\Delta v^2} \left[ \left( \frac{f_{j+1}^n}{\mathcal{M}_{f^0, j+1}} - \frac{f_j^n}{\mathcal{M}_{f^0, j}} \right) \cdot \widehat{\mathcal{M}}_{f^0, j+1/2} \right. \\ &\quad \left. - \left( \frac{f_j^n}{\mathcal{M}_{f^0, j}} - \frac{f_{j-1}^n}{\mathcal{M}_{f^0, j-1}} \right) \cdot \widehat{\mathcal{M}}_{f^0, j-1/2} \right] \\ &\quad - \frac{\Omega T_e}{m\Delta v^2} (f_{j+1}^n - 2f_j^n + f_{j-1}^n) \\ &\quad + \frac{\Omega T_e}{m\Delta v^2} (f_{j+1}^n - 2f_j^n + f_{j-1}^n) \end{aligned}$$

that is to say

$$\begin{aligned} S(f^n)_j &= \frac{\Omega T_e}{m\Delta v^2} \left\{ \widehat{\mathcal{M}}_{f^0, j+1/2} [(f^n / \mathcal{M}_{f^0})_{j+1} - (f^n / \mathcal{M}_{f^0})_j] \right. \\ &\quad \left. - \widehat{\mathcal{M}}_{f^0, j-1/2} [(f^n / \mathcal{M}_{f^0})_j - (f^n / \mathcal{M}_{f^0})_{j-1}] \right\}. \end{aligned}$$

If  $j \in \{1, j_{\max}\}$ , we easily verify that the equality (8) remains true when

$$f_0^n \equiv f_1^n \cdot \frac{\mathcal{M}_{f^0, 0}}{\mathcal{M}_{f^0, 1}} \quad \text{and} \quad f_{j_{\max}+1}^n \equiv f_{j_{\max}}^n \cdot \frac{\mathcal{M}_{f^0, j_{\max}+1}}{\mathcal{M}_{f^0, j_{\max}}}.$$

□

## 2 Positivity of the scheme

Let us introduce some new notations:

**Notation** Let us define

$$\begin{aligned} \Delta t^n &= \Delta t_1^0 \cdot \frac{h_{\min}^n}{h_{\max}^n} \quad \text{where} \quad \Delta t_1^0 = \frac{m}{4\Omega T_e} \cdot \frac{\Delta v^2}{M^0}, \\ &\quad \begin{cases} h_{\max}^n = \max_j \left( \frac{f_j^n}{\mathcal{M}_{f^0}} \right)_j, \\ h_{\min}^n = \min_j \left( \frac{f_j^n}{\mathcal{M}_{f^0}} \right)_j, \end{cases} \\ \mathcal{M}_{f^0} &= \frac{N^0}{\sqrt{2\pi T_e/m}} \exp\left[-\frac{m(v - U_e)^2}{2T_e}\right], \\ M^0 &= \max_j \left( \frac{\mathcal{M}_{f^0, j+1}}{\mathcal{M}_{f^0, j}} \right) \end{aligned}$$

and

$$H^n = \langle [f^n \log(f^n / \mathcal{M}_{f^0})]_j \rangle.$$

We have the following proposition:

**Proposition 2.1** *For all strictly positive initial condition, when  $\tilde{f}_{j+1/2}^n$  is the Chang and Cooper average of  $f_j^n$  and of  $f_{j+1}^n$ , the explicit scheme defined with (3) and with (4) verifies the inequality*

$$h_{\min}^n \leq h_{\min}^{n+1} \leq h_{\max}^{n+1} \leq h_{\max}^n$$

when

$$\Delta t < 2\Delta t_1^0.$$

More over, the time step  $\Delta t^n$  verifies

$$\Delta t^n \leq \Delta t^{n+1} \leq \Delta t_1^0.$$

Let us note that this proposition implies that

$$\inf_{j,n} f_j^n > 0$$

and allows to establish that the time step will never be equal to zero. Then, the numerical scheme preserves the positivity of the distribution under a CFL criteria, it exists  $h_{\min}^\infty$  and  $h_{\max}^\infty$  such that the series  $(h_{\min}^n)$  and  $(h_{\max}^n)$  admits the respective limits  $h_{\min}^\infty$  and  $h_{\max}^\infty$  when  $n$  goes to  $+\infty$ . But, we do not have still prove that  $h_{\min}^\infty = h_{\max}^\infty$ , equality which would imply that it would exist a constant  $C > 0$  such that

$$\lim_{n \rightarrow \infty} (f_j^n) = C \cdot (\mathcal{M}_{f^0,j}).$$

**Proof of the proposition 2.1** By defining  $h_j^n = f_j^n / \mathcal{M}_{f^0,j}$  and by using the lemma 1.1, we can say that

$$f_j^{n+1} = f_j^n + \frac{\Delta t \Omega T_e}{m \Delta v^2} \left[ \widehat{\mathcal{M}}_{f^0,j+1/2} (h_{j+1}^n - h_j^n) + \widehat{\mathcal{M}}_{f^0,j-1/2} (h_{j-1}^n - h_j^n) \right].$$

Then, it is obvious that

$$h_j^n - \frac{\Delta t \Omega T_e}{m \Delta v^2} \cdot \frac{\widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2}}{\mathcal{M}_{f^0,j}} (h_j^n - h_{\min}^n) \leq h_j^{n+1}$$

and that

$$h_j^{n+1} \leq h_j^n + \frac{\Delta t \Omega T_e}{m \Delta v^2} \cdot \frac{\widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2}}{\mathcal{M}_{f^0,j}} (h_{\max}^n - h_j^n).$$

Let us suppose that  $\Delta t$  is such that

$$\forall j, \Delta t \leq \frac{m \Delta v^2}{\Omega T_e} \cdot \frac{\mathcal{M}_{f^0,j}}{\widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2}}.$$

Then, we obtain that

$$\forall j : h_{\min}^n \leq h_j^{n+1} \leq h_{\max}^n.$$

More over, we have

$$\frac{\widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2}}{\mathcal{M}_{f^0,j}} = \frac{\log(\mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j+1})}{\mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j+1} - 1} + \frac{\log(\mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j-1})}{\mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j-1} - 1}.$$

Then

$$\frac{\widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2}}{\mathcal{M}_{f^0,j}} \leq \frac{1}{\min(1, \mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j+1})} + \frac{1}{\min(1, \mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j-1})}$$

since  $\forall x \geq 0 : \min(1, x) \leq (x - 1) / \log x$ . But

$$\begin{aligned} \frac{1}{\min(1, \mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j+1})} + \frac{1}{\min(1, \mathcal{M}_{f^0,j} / \mathcal{M}_{f^0,j-1})} &\leq \frac{2}{\min_k(\mathcal{M}_{f^0,k \pm 1} / \mathcal{M}_{f^0,k})} \\ &= 2 \max_k(\mathcal{M}_{f^0,k \pm 1} / \mathcal{M}_{f^0,k}) \\ &= 2M^0. \end{aligned}$$

Finally, we can then write that

$$\forall j : \frac{1}{2M^0} \leq \frac{\mathcal{M}_{f^0,j}}{\widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2}} \quad (11)$$

which allows to state that

$$\Delta t \leq 2\Delta t_1^0 \rightarrow h_{\min}^n \leq h_{\min}^{n+1} \leq h_{\max}^{n+1} \leq h_{\max}^n.$$

More over, it is obvious that

$$h_{\min}^n \leq h_{\min}^{n+1} \leq h_{\max}^{n+1} \leq h_{\max}^n \implies \Delta t^n \leq \Delta t^{n+1}.$$

□

### 3 Convergence toward the thermodynamical equilibrium

The following proposition shows that the distribution  $f^n$  converges toward the thermodynamical equilibrium:

**Proposition 3.1** *For all strictly positive initial condition, when  $\widetilde{f}_{j+1/2}^n$  is the Chang and Cooper average of  $f_j^n$  and of  $f_{j+1}^n$  and when*

$$\Delta t < \Delta t^n,$$

*the explicit scheme defined with (3) and with (4) verifies the entropic inequality*

$$H^{n+1} \leq H^n$$

and

$$\lim_{t^n \rightarrow +\infty} (f_j^n) = \frac{N^0}{\langle \mathcal{M}_{f^0,j} \rangle} \cdot (\mathcal{M}_{f^0,j})$$

Then, we have

$$\lim_{t^n \rightarrow +\infty} \Delta t^n = \Delta t_1^0.$$

**Proof of the proposition 3.1** Since  $\Delta t < \Delta t_1^0$ , we have  $f_j^{n+1} > 0$  by using the proposition 2.1: then, we can evaluate  $H^{n+1}$ . And, by using the inequality

$$\forall x > 0 : \log(x+1) < x,$$

we easily obtain that

$$H^{n+1} \leq H^n + \Delta t \sum_j \left[ S(f^n)_j \log \left( \frac{f^n}{\mathcal{M}_{f^0}} \right)_j + \Delta t \frac{S(f^n)_j^2}{f_j^n} \right] \Delta v. \quad (12)$$

More over, by applying the Schwarz's inequality, we obtain that

$$\begin{aligned} S(f^n)_j^2 &\leq \frac{\Omega T_e}{m \Delta v^2} \left( \widehat{\mathcal{M}}_{f^0,j+1/2} + \widehat{\mathcal{M}}_{f^0,j-1/2} \right) \cdot \\ &\quad \frac{\Omega T_e}{m \Delta v^2} \left[ \widehat{\mathcal{M}}_{f^0,j+1/2} (h_{j+1}^n - h_j^n)^2 + \widehat{\mathcal{M}}_{f^0,j-1/2} (h_{j-1}^n - h_j^n)^2 \right] \end{aligned}$$

where we have define  $h_j^n = f_j^n / \mathcal{M}_{f^0,j}$ . And, by using (11), we obtain

$$\sum_j \frac{S(f^n)_j^2}{f_j^n} \leq \frac{\Omega T_e}{m \Delta v^2} \cdot \frac{2M^0}{h_{\min}^n} \cdot \frac{2\Omega T_e}{m \Delta v^2} \sum_j \widehat{\mathcal{M}}_{f^0,j+1/2} (h_{j+1}^n - h_j^n)^2.$$

And since

$$\forall x \geq 0, \forall y \geq 0 : \frac{x-y}{\log(x/y)} \leq \max(x, y),$$

we have

$$\sum_j \frac{S(f^n)_j^2}{f_j^n} \leq \frac{\Omega T_e}{m \Delta v^2} \cdot \frac{2M^0}{h_{\min}^n} \cdot \frac{2\Omega T_e}{m \Delta v^2} \sum_j \widehat{\mathcal{M}}_{f^0, j+1/2} (h_{j+1}^n - h_j^n) \cdot (\log h_{j+1}^n - \log h_j^n) h_{\max}^n.$$

More over, we have

$$\begin{aligned} & \sum_j S(f^n)_j \cdot \log(f^n / \mathcal{M}_{f^0})_j = \\ & \frac{\Omega T_e}{m \Delta v^2} \sum_j \widehat{\mathcal{M}}_{f^0, j+1/2} [(f^n / \mathcal{M}_{f^0})_{j+1} - (f^n / \mathcal{M}_{f^0})_j] \log(f^n / \mathcal{M}_{f^0})_j \\ & - \frac{\Omega T_e}{m \Delta v^2} \sum_j \widehat{\mathcal{M}}_{f^0, j-1/2} [(f^n / \mathcal{M}_{f^0})_j - (f^n / \mathcal{M}_{f^0})_{j-1}] \log(f^n / \mathcal{M}_{f^0})_j \end{aligned}$$

that is to say

$$\begin{aligned} & \sum_j S(f^n)_j \cdot \log(f^n / \mathcal{M}_{f^0})_j = \tag{13} \\ & - \frac{\Omega T_e}{m \Delta v^2} \sum_j \widehat{\mathcal{M}}_{f^0, j+1/2} (h_{j+1} - h_j) \cdot (\log h_{j+1}^n - \log h_j^n) \leq 0 \end{aligned}$$

by using the convexity property of  $x \mapsto \log x$ . Then, we can write that

$$\sum_j \frac{S(f^n)_j^2}{f_j^n} \leq - \frac{4\Omega T_e}{m \Delta v^2} \cdot \frac{h_{\max}^n}{h_{\min}^n} M^0 \sum_j S(f^n)_j \log \left( \frac{f^n}{\mathcal{M}_{f^0}} \right)_j.$$

Finally, we obtain

$$H^{n+1} \leq H^n + \Delta t \left( 1 - \frac{4\Delta t \Omega T_e}{m \Delta v^2} \cdot M^0 h_{\max}^n / h_{\min}^n \right) \cdot \sum_j S(f^n)_j \log \left( \frac{f^n}{\mathcal{M}_{f^0}} \right)_j \Delta v.$$

Then, when  $\Delta t < \Delta t^n$ , by using the inequality (13), we obtain the inequality

$$H^{n+1} \leq H^n + \Delta t \left( 1 - \frac{\Delta t}{\Delta t^n} \right) \cdot \sum_j S(f^n)_j \log \left( \frac{f^n}{\mathcal{M}_{f^0}} \right)_j \Delta v \leq H^n. \tag{14}$$

(...) $\square$

## Conclusion

In that paper, we have shown that for a particular Fokker-Planck linear operator, the explicit Chang and Cooper scheme has very good properties: under a classical CFL criteria, it preserves the positivity of the distribution, the entropy decreases and the distribution converges toward the thermodynamical equilibrium.

This theoretical results are similar to those obtained with this linear Fokker-Planck operator by replacing the Chang and Cooper average with the entropic average introduced in [5]. But, we have to focus on the fact that all these results are obtained in the linear case.

Although, by using the entropic average, it is possible to obtain similar theoretical results in the non linear case (cf. [5]) which is not at all the case when we use the Chang and Cooper average (cf. [7], first part). This is underlined by the fact that when it is simulated ion-electron collisions under very hard conditions (as those met in the field of the Inertial Confinement Fusion: see the last chapter of the second part of [7]) with the non linear version of the linear Fokker-Planck operator discretized in that paper, it is possible to go to the end of the simulation without any computer error only by using the entropic average. Then, it seems that it will be very difficult to obtain good theoretical results in the non linear case for the Chang and Cooper average.

## References

- [1] **J. S. Chang and G. Cooper** - *A practical Difference Scheme for Fokker-Planck Equations* - Journal of Computational Physics, volume 6, p. 1-16, 1970.
- [2] **D. Deck et G. Samba** - *Le Code Procions* - Internal report CEA/DAM, n<sup>o</sup> 2561, 1994.
- [3] **E.M. Epperlein** - *Implicit and Conservative Difference Scheme for the Fokker-Planck Equation* - Journal of Computational Physics, volume 112, p. 291-297, 1994.
- [4] **V.A. Mousseau et D.A. Knoll** - *Fully implicit Kinetic Solution of Collisional Plasmas* - Journal of Computational Physics, volume 136, p. 308-323, 1997.
- [5] **C. Buet, S. Dellacherie et R. Sentis** - *Résolution numérique d'une équation de Fokker-Planck ionique avec température électronique* - C. R. Acad. Sci. Paris, t. 327, série I, p. 93-98, 1998.
- [6] **S. Dellacherie** - *Sur un schéma numérique semi-discret appliqué à un opérateur de Fokker-Planck isotrope* - C. R. Acad. Sci. Paris, t. 328, série I, p. 1219-1224, juin 1999.
- [7] **S. Dellacherie** - *Contribution à l'analyse et à la simulation numériques des équations cinétiques décrivant un plasma chaud* - Thèse de Doctorat de l'Université Denis Diderot (Paris VII), 1998.

# An asymptotic preserving scheme for Hydrodynamics Radiative Transfert Models

## Numerical Scheme for radiative transfert

Christophe Buet<sup>1</sup>, Stéphane Cordier<sup>2</sup>

<sup>1</sup> C.E.A. D.A.M.  
91 Bruyères le Chatel - France  
email : `Christophe.Buet@cea.fr`

<sup>2</sup> UMR MAPMO - CNRS 6628  
B.P. 6759 Université d'Orléans  
45067 Orléans cedex- France.  
email : `cordier@math.cnrs.fr`

Received: date / Revised version: date

**Summary** In this paper, we shall propose a numerical scheme consisting of two steps: the first based relaxation method and the second on the so called well balanced scheme. The derivation of the scheme relies on the resolution of the stationnary Riemann problem with source terms. The obtained scheme is compatible with the diffusive regime of hydrodynamics radiative transfert models. Some numerical results are shown.

## 1 Introduction

In this article , we are concerned with a model arising from radiative transfert modelling. It is well known that in very rarefied regions, the physically relevant model is a free transport one whereas in the dense regions, the radiative transfert becomes a diffusion equation. The aim of this work is to design a scheme for the two-moment systems that can be obtain, for example using maximum entropy technics. We refer to [23] for a recent presentation of the various closures.



The requirement for the scheme are to be used with a non uniform grid in space, to deal with varying scattering cross sections, to have the correct asymptotics behaviour in diffusive regimes, to be implicit in order to avoid too restrictive time steps limitations.

The solution we should described has been announced in [5] and it consists of two main steps : the first one is a relaxation method which permits to transform the nonlinear hyperbolic system into two independant linear systems, known as the Goldstein Taylor system or Telegraph equations; the second step is to use the so called well balanced scheme for each of the two systems. Moreover, we shall propose a new interpretation of the well balanced schemes as a Godunov scheme when dealing with source terms. This interpretation is more convenient in order to extend the well balanced schemes for multi dimensional problem.

The paper is organized as follows : in section 2, we recall the model of interest and its main properties, namely an invariant domain or equivalently the positivity of some quantities which are of great physical importance and the diffusive asymptotics i.e. regimes where the solutions behave like solution of a parabolic equation. In section 3, we describe the proposed numerical scheme and its derivation in two steps (relaxation method in subsection 3.1 and well balanced scheme in subsection 3.2). In section 4, we present two numerical results : the first one is concerned with a varying cross section and the second one to a coupled system with a heat equation for the temperature of material.

## 2 Radiative transfert hydrodynamical models

The models, we are interested in, arise from the radiation transport equation, which is a kinetic equation for the specific intensity of photons  $I(\Omega, \nu, r, t)$  after integration over the angular variable  $\Omega$  and the frequency  $\nu$ . We refer to [18, 6] for a detailed presentation.

In this paper, we are concerned with systems of conservation laws for the two first moments of the intensity, namely  $(\rho, \rho u)$  :

$$\begin{cases} \partial_t \rho + \nabla_x(\rho u) = 0 \\ \partial_t(\rho u) + \nabla_x \cdot P = 0, \end{cases} \quad (1)$$

where  $P$  is the pressure tensor of the form

$$P = \rho \left( h(|u|) I_d + (1 - 3h(|u|)) \frac{\mathbf{u} \otimes \mathbf{u}}{\|u\|^2} \right),$$

where  $h(x) = \frac{x}{g^{-1}(x)}$ ,  $x > 0$  with

$$g(x) = \coth(x) - \frac{1}{x} ,$$

the so called Langevin function.

### 2.1 Eddington factors

In monodimensional case (the velocity is parallel to the first axis) the above system reduces to

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) &= 0 \\ \partial_t(\rho u) + \partial_x \rho h(u) &= -\sigma \rho u , \end{cases} \quad (2)$$

where the function  $h$  has several forms, which are called Eddington factors. We refer to [18] or to [23] for a detailed presentation. For example, Kershaw suggests,

$$h(x) = \frac{1 + 2x^2}{3}. \quad (3)$$

Minerbo uses entropy arguments to obtain the Eddington factor ,

$$h(x) = 1 - \frac{2x}{g^{-1}(x)}, \quad (4)$$

which was called the maximum entropy Eddington factor.

Also, Minerbo [19] suggested that any intensity may be approximated as a linear function, so the Eddington factor Minerbo linear is:

$$h(x) = \begin{cases} \frac{1}{3}, x \in [0, \frac{1}{3}] , \\ \frac{1}{2}(1-x)^2 + x^2, x \in (\frac{1}{3}, 1] . \end{cases} \quad (5)$$

Using a Chapman-Enskog approach, we have the following Eddington factor suggested by Levermore :

$$h(x) = \coth x \left( \coth x - \frac{1}{x} \right) .. \quad (6)$$

Another Eddington factor is Levermore-Lorentz [19,18]:

$$h(x) = \frac{1}{3} + \frac{2x^2}{2 + \sqrt{4 - 3x^2}} . \quad (7)$$

Let us propose another function, with the same properties of Eddington factors, called Minerbo rational, which is defined by:

$$h(x) = \frac{1}{3} + \frac{2x^2}{5 - |x| - x^2} . \quad (8)$$

This Eddington factors is obtained by making some assumptions for the function

$$h(x) = \frac{1}{3} + \frac{a|x| + b}{cx^2 + d|x| + e} , \quad (9)$$

namely:

$$\begin{aligned} h(0) &= \frac{1}{3} , \\ h'(0) &= 0 , \\ h''(0) &= \frac{2}{5} . \end{aligned}$$

and

$$\begin{aligned} h(1) &= 1 , \\ h'(1) &= 2 . \end{aligned}$$

We remark that these Eddington factors are increasing and convex functions. Let us assume the following hypothesis on the  $h$  function :  $h$  is a increasing, convex function

$$u^2 < h(u) \leq 1 , h(0) = \frac{1}{3} , h(1) = 1 . \quad (10)$$

*Remark 1* Let us mention that for  $h \equiv 1$  , the system (2) is known as the Goldstein-Taylor equation or as the telegraph equation. This particular system will be used in the construction of the solution of the nonlinear system of interest.

*Remark 2* Note also that some two dimensional closure have been proposed [24] where the pressure is a function of  $\rho$  and  $j$  separately instead of a function of the form  $\rho h(j/\rho)$  as in (2).

## 2.2 Invariant domain

Let us now give some properties of the solutions of the nonlinear hyperbolic system (2).

First, for any physically admissible state  $(\rho, j)$  such that

$$\rho > 0, \quad \|j\| \leq \rho, \quad (11)$$

the system (2) is hyperbolic i.e. the matrix of transport coefficient is diagonalizable (see [27] for detailed definitions). Second, the solution

of the Riemann problem without source terms lies in the set of admissible state. Assuming that a Godunov type numerical scheme with a splitting of the source terms converges, the solution will remain in the admissible set is invariant.

From the physical point of view, the admissible state comes from the fact that  $\rho$  represents a number of photons which has to be non negative and that the mean velocity of the photons  $j/\rho$  is smaller than the speed of light (which is normalized to 1 in the choosen scaling).

Note that the proof of convergence of the scheme is behind the scope of this paper. We shall verify on the numerical results in the last section that this convergence is expected.

Let us rewrite the system in variable  $U = (\rho, u)$  with  $j = \rho u$ . The admissible states are characterized by  $\rho \geq 0$  and  $\|u\| \leq 1$ . The matrix of transport coefficient  $A(U)$  such that system (18) reads

$$\partial_t U + A(U) \partial_x U = R(U), \quad x \in \mathbb{R}, t > 0 \quad (12)$$

is given by

$$A(U) = \begin{pmatrix} u & \rho \\ \frac{h(u)-u^2}{\rho} & h'(u) - u \end{pmatrix}, \quad (13)$$

Its eigenvalues  $\lambda_{\pm}$  are given by

$$\lambda_{\pm} = \frac{1}{2} \left( h'(u) \pm \sqrt{(h' - 2u)^2 + 4(h - u^2)} \right) \quad (14)$$

which are both real using hypothesis (10) on the  $h$  function .

The invariant property of the approximated solution is based on a splitting argument between the transport part and the source terms. We claim that both of the two operators preserve the properties (11), then the composition of the two will have the same property. It remains to prove that the splitting procedure converges as the time step goes to zero using a argument related to Trotter formula

$$\exp(A + B)t = \lim_{n \rightarrow \infty} (\exp(At/n) \cdot \exp(Bt/n))^n$$

The invariant property is obvious on the source terms. Indeed,  $\rho$  remains constant whereas  $\|j\|$  decreases. For the transport part i.e. the system (2) with  $\sigma = 0$ , it can be checked that the properties hold true for the Riemann problem i.e. considering two states  $U_l$  and  $U_r$ , we verify that the 1-wave curve that contains  $U_l$  (and which consists of half shock curve and half rarefaction wave) intersects the 2-wave

curve that contains  $U_r$  in a so called intermediate state that is in the invariant domain.

More precisely, it can be proved that the 1-wave has the following behaviour in the  $\rho, u$  plane : the curve can be parametrized by  $u$  and the parametrization  $\rho(u)$  is decreasing and satisfies  $\rho(u) \rightarrow +\infty$  as  $u \rightarrow -1$ ,  $\rho(u) \leq 0, \forall u \in ]-1, 1[$  and by construction  $\rho(u_l) = u_l$ . Similarly, the 2-wave can also be parametrized by  $u$  and is increasing, positive and  $\lim_{u \rightarrow 1} \rho(u) = +\infty$ ,  $\rho(u_r) = \rho_r$ . Using the above properties of the 1 and 2-wave curves, one obtain the existence of an intermediate state  $(\rho_*, u_*)$  which satisfy the required properties (11). We refer to [3] for the details on the construction of the solution for the Riemann problem of system (2).

Then, if the Godunov type scheme converges, the invariant property will be satisfied for the continuous solution of the transport part.

*Remark 3* This invariant property (11) has a physical interpretation, since  $\rho$  and  $j$  the two first moments of the distribution function on the unit sphere  $\rho = \int f d\omega$ ,  $j = \int f \omega d\omega$ . From numerical point of view, this property means that the flux are so-called limited.

### 2.3 Asymptotic limit - diffusive regimes

Let us formally present the asymptotic limit of the system in so called diffusive regimes. In such scaling, the system (18) can be written in the form

$$\varepsilon \partial_t U + \partial_x F(U) = \frac{1}{\varepsilon} R(U), \quad (15)$$

where  $U = (\rho, j)$ ,  $F(U) = (j, \rho h(j/\rho))$ ,  $R(U) = (0, -\sigma j)$ ,  $\sigma(x) > 0$ , the cross section and  $\varepsilon$  is a small parameter.

In the limit  $\varepsilon \rightarrow 0$ , a formal asymptotic gives that  $j$  is  $O(\varepsilon)$  due to the collision term. At first order in  $\varepsilon$ , using the second equation of (15), we get

$$j = -\frac{\varepsilon}{\sigma} \partial_x (h(0)\rho). \quad (16)$$

This corresponds to suppress the time derivative term into the equation on  $j$ . Then, using  $h(0) = 1/3$  and using (16) into the first equation, we obtain the following diffusion approximation

$$\frac{\partial}{\partial t} \rho - \frac{\partial}{\partial x} \left( \frac{1}{3\sigma} \frac{\partial}{\partial x} \rho \right) = 0. \quad (17)$$

Note that the solution for  $\rho$  of the limit heat equation (17) and  $j$  given by (16) does not satisfy automatically the condition (11), because the gradient  $\partial_x \rho$  can be arbitrarily large e.g. if the initial data is discontinuous in  $\rho$ . Note also that  $j$  or  $\partial_x \rho$  are also solution of the same heat equation.

Our aim is now to design an implicit scheme compatible with the limit  $\varepsilon \rightarrow 0$  and with the invariant property (11).

Various methods have been proposed to get ride of these difficulties such as variable Eddington factors for the so-called  $P1$ -approximation or flux limiters for diffusion approximation (see [18] and ref. therein). This is also related to a series of papers about asymptotic preserving schemes for kinetic problems, well balanced schemes, stiff source terms and relaxation methods in the context of hyperbolic systems [12, 8].

### 3 Numerical method for radiative model

Our method is based on a time splitting in two steps. The first step is based on a relaxation method [12] and the second on a well-balanced schemes [8].

Our goal is to solve the nonlinear problem while preserving positivity. Linearizing the equation is not suitable since the invariance domain will not be preserved and it will give wrong results when using implicit method.

#### 3.1 The relaxed system.

Let us briefly recall the relaxation method according to [12]. Considering a system of the form

$$\partial_t u + \partial_x f(u) = 0,$$

it becomes in the relaxation limit  $\alpha \rightarrow 0$

$$\partial_t(u, v) + \partial_x(v, au) = (0, -(f(u) - v)/\alpha).$$

This can be written in the form

$$\partial_t U + A \partial_x U = R(U), \quad (18)$$

where  $A$  is a constant matrix of transport coefficient and  $R(U)$  is a , possibly nonlinear, source term.

Let us now apply this method to our system. The first step is to rewrite system (15) as the limit ( $\alpha \rightarrow 0$ ) of the following relaxation system

$$\begin{cases} \partial_t \rho + \frac{1}{\varepsilon} \partial_x z = 0 \\ \partial_t z + \frac{a}{\varepsilon} \partial_x \rho + \frac{\sigma}{\varepsilon^2} z = -\frac{1}{\alpha} (j - z) \\ \partial_t w + \frac{a}{\varepsilon} \partial_x j = \frac{1}{\alpha} (\rho h(\frac{j}{\rho}) - w) \\ \partial_t j + \frac{1}{\varepsilon} \partial_x w + \frac{\sigma}{\varepsilon^2} j = 0. \end{cases} \quad (19)$$

Note that, formally, the relaxed system is equivalent to (15) as the limit ( $\alpha \rightarrow 0$ ) since the equilibrium states are given by

$$z = j, \quad w = \rho h(\frac{j}{\rho}). \quad (20)$$

At this point, the coefficient  $a$  remains to be choosen.

Our method consists in splitting the transport part or left hand side of system (19) and the relaxation term i.e. the right hand side. In the relaxation part, the original variables ( $\rho$  and  $j$ ) are unchanged whereas the new variables ( $z$  and  $w$ ) converge to the equilibrium state given by (20) in the limit  $\alpha \rightarrow 0$ . Thus, the relaxation part reduces into a projection on equilibrium states. The coefficient  $a$  is constant in space but has to be choosen at each time step in order to recover the correct diffusion coefficient.

The transport part writes:

$$\begin{cases} \partial_t \rho + \frac{1}{\varepsilon} \partial_x z = 0 \\ \partial_t z + \frac{a}{\varepsilon} \partial_x \rho + \frac{\sigma}{\varepsilon^2} z = 0 \\ \partial_t w + \frac{a}{\varepsilon} \partial_x j = 0 \\ \partial_t j + \frac{1}{\varepsilon} \partial_x w + \frac{\sigma}{\varepsilon^2} j = 0. \end{cases} \quad (21)$$

Let us point out that the relaxation term on the equation over  $z$  is not classical. However, this term is very important in order to get the right asymptotic behaviour in diffusives regimes i.e. when  $\varepsilon \rightarrow 0$ .

Let us also emphasize that (21) is just two linear and identical systems. These systems are uncoupled : one for the quantities ( $\rho$  and  $z$ ), the second for ( $w$  and a new variable  $\tilde{j} = aj$ ) which are equivalent to the system (15) with  $h \equiv a$ .

Note that this system i.e. (15) with  $h \equiv a$  once diagonalized is a well-know Goldstein-Taylor or Telegraph equation with speed  $\pm\sqrt{a}$

i.e. two transport equations in opposite direction coupled by a relaxation term

$$\begin{cases} \partial_t u + \frac{\sqrt{a}}{\varepsilon} \partial_x u = \frac{\sigma}{2\varepsilon^2} (v - u) \\ \partial_t v - \frac{\sqrt{a}}{\varepsilon} \partial_x v = \frac{\sigma}{2\varepsilon^2} (u - v). \end{cases} \quad (22)$$

Let us mention that the invariant domain property (11) can be seen as the positivity of the transported quantities  $\rho \pm j > 0$ . In the above system, the transport quantities ( $u$  or  $v$ ) are  $\rho \pm z/\sqrt{a}$  (or  $j \pm w/\sqrt{a}$ ). We shall choose the coefficient  $a$  in such a way that the transport quantities remain positive.

Using the following change of variable

$$\begin{aligned} U &= \sqrt{a}\rho + z + w + \sqrt{a}j, \\ V &= \sqrt{a}\rho - z + w - \sqrt{a}j, \\ \bar{U} &= \sqrt{a}\rho + z - w - \sqrt{a}j, \\ \bar{V} &= \sqrt{a}\rho - z - w + \sqrt{a}j, \end{aligned} \quad (23)$$

we verify that  $(U, V)$  and  $(\bar{U}, \bar{V})$  satisfy a system of the form (22). We will show that for the transport part, the invariant domain (11) comes from the positivity of  $U, V, \bar{U}, \bar{V}$  for sufficiently large value of  $a$ .

The initial data for the new variables verify, using the equality of the projected variable at initial time  $z = j$  and  $w = \rho h(u)$ , become

$$\begin{aligned} U &= \rho(\sqrt{a} + h(u)) + j(\sqrt{a} + 1), \\ V &= \rho(\sqrt{a} + h(u)) - j(\sqrt{a} + 1), \\ \bar{U} &= \rho(\sqrt{a} - h(u)) - j(\sqrt{a} - 1), \\ \bar{V} &= \rho(\sqrt{a} - h(u)) + j(\sqrt{a} - 1). \end{aligned} \quad (24)$$

Then, using that  $j = \rho u$  and  $\rho \geq 0$ , we obtain that  $U$  at  $t = 0$  is positive provided that

$$\sqrt{a} + h(u) + (\sqrt{a} + 1)u \geq 0.$$

Let us assume that  $u$  (at initial time) lies in the interval  $[-b, b]$  with some  $b < 1$ . We have to choose  $a$ . Note that, for any  $a > b$  and any  $u \in [-b, b]$ ,

$$\sqrt{a} + h(u) + (\sqrt{a} + 1)u \geq \sqrt{a} + h(b) - (\sqrt{a} + 1)b.$$

Thus, the following value

$$\sqrt{a} = \frac{h(b) - b}{1 - b},$$

insures positivity. Any larger value of  $a$  will also be convenient, for example

$$a \stackrel{\text{def}}{=} h(b). \quad (25)$$



Indeed, it can be easily verify that this choice preserves positivity :

$$\sqrt{a}+h(u)+(\sqrt{a}+1)u \geq \sqrt{h(b)}+h(b)-(\sqrt{h(b)}+1)b = (\sqrt{h(b)}+1)(\sqrt{h(b)}-b) \geq 0,$$

since  $h(y) \geq y^2$ . The properties hold also the the other three quantities  $V, \bar{U}, \bar{V}$ .

Thus, we prove that the following choice for the coefficient  $a$

$$a = h(\max_{x \in \mathbb{R}} \|u(x)\|),$$

ensures the positivity of the initial data for  $U, V, \bar{U}, \bar{V}$  and then of the solution  $U, V, \bar{U}, \bar{V}$  of the system (21). The sketch of the proof for the positivity is similar to the one of section 2.2. We consider the two independant systems of the form (22). We construct the solution of such system as the limit of a approximated solution based on a splitting. For the transport part (as speed  $\pm\sqrt{a}$ ) the positivity of both  $u$  and  $v$  is obvious. For the source terms, we easily check that the solution of the relaxation

$$\partial_t u = \frac{\sigma}{2\varepsilon^2}(v - u), \quad \partial_t v = \frac{\sigma}{2\varepsilon^2}(v - u),$$

also preserves positivity. Thus, the approximation satisfy the properties and so does its limit. We assume the convergence of such splitting based algorithms, which is reasonable for any fixed  $\varepsilon$ .

In this case, we can prove directly the above result : For any solution of (22) with positive initial data, the solution remains positive. We define the positive and negative part of a fonction  $f$  and denote it by  $f^+$  and  $f^-$

$$f^+ = (f + |f|)/2, f^- = (f - |f|)/2.$$

We multiply the equations of (22) by  $u^-$  and  $v^-$  respectively and we integrate for  $x \in \mathbb{R}$

$$\begin{cases} \partial_t \int u^- u dx + \frac{\sqrt{a}}{\varepsilon} \int u^- \partial_x u = \int \frac{\sigma}{2\varepsilon^2} u^- (v - u) \\ \partial_t \int v^- v dx - \frac{\sqrt{a}}{\varepsilon} \int v^- \partial_x v = \int \frac{\sigma}{2\varepsilon^2} v^- (u - v). \end{cases}$$

We have  $u^- u = (u^-)^2$  and the integral of  $\partial_x (u^-)^2$  is zero assuming  $u(x, t) \rightarrow 0$  when  $x \rightarrow \pm\infty$ . Thus,

$$\partial_t \int (u^-)^2 + (v^-)^2 dx = \int \frac{\sigma}{2\varepsilon^2} (u^- (v - u) + v^- (u - v)) \leq 0.$$

Indeed,

$$u^-(v-u)+v^-(u-v) = u^-(v^++v^--u^-)+v^-(u^++u^--v^-) = u^-v^++u^+v^--(u^--v^-)^2.$$

The first two terms are non positive by definition of  $(u^-, v^+)$  and  $(u^+, v^-)$  respectively.

The initial data being positive implies  $u^-(t=0) = v^-(t=0) = 0$  and the above inequality proves that the  $L^2$  norm of  $u^-$  and  $v^-$  decay. This proves that  $u^-$  and  $v^-$  vanish for any time i.e. that the solution  $u$  and  $v$  remain positive.

*Remark 4* The proposed choice for  $a$  is not satisfactory since it depends on the whole solution  $u(x)$  for  $x \in \mathbb{R}$ . We refer to [6] for a variant of the solution proposed here that overcomes this difficulty.

In the diffusive limit ( $\varepsilon \rightarrow 0$ ) or for large time behaviour, we expect that  $\max_{x \in \mathbb{R}} \|u(x)\| \rightarrow 0$  and therefore,  $a$  will become close to  $1/3$  i.e. we obtain the right asymptotic (17).

*Remark 5* Despite its linear structure, the system (22) give raise to severe numerical problems. It can also be seen as a very simple model of kinetic theory of gases where particles can only have velocity  $\pm\sqrt{a}$ . In this context, the limit  $\varepsilon \rightarrow 0$  corresponds to a diffusion limit which a diffusion coefficient equal to  $\frac{a}{\sigma}$ .

### 3.2 An interpretation of the Well Balanced scheme

We shall now solve numerically the system (22) for  $(\rho, z)$  (and similarly for the other identical system in variable  $(w, aj)$ ).

We introduce a non-uniform mesh : we note  $x_i$ , the center of the cell of size  $\Delta x_i$  with  $i \in \mathbb{Z}$  and define  $\Delta x_{i+\frac{1}{2}} = (\Delta x_i + \Delta x_{i+1})/2$ .

In this part, we shall present an interpretation of the WB scheme for the so called telegraph equation and/or Goldstein-Taylor model

$$\partial_t u + \partial_x u = -\sigma(v - u), \partial_t v - \partial_x v = \sigma(v - u). \quad (26)$$

This system, with source term, can be rewritten in a more compact form

$$\partial_t U + A \partial_x U = R(U).$$

One possible way to recover the WB scheme is to approximate the source term using a quadrature formulae that localize the source at interface i.e.  $R(U)$  is replaced by

$$\sum_i \delta(x - x_{i+\frac{1}{2}}) \Delta x_{i+\frac{1}{2}} R(U_{i+\frac{1}{2}}).$$

Let us now introduce an extended non conservative hyperbolic system (without source term) for  $(U, id)$  where  $id$  represents the identity function (constant in time)

$$\begin{cases} \partial_t U + A \partial_x U = R(U) \partial_x id, \\ \partial_t id = 0, \end{cases} \quad (27)$$

with the matrix

$$B = \begin{pmatrix} A & -R \\ 0 & 0 \end{pmatrix}.$$

Note that since  $A$  is diagonalizable,  $B$  too and its spectrum consists of the eigenvalues of  $A$  and zero. Let us now assume that the matrix  $A$  is diagonal. Then, using a piecewise constant approximation of the auxillary function  $id(x)$  (and thus its derivative becomes sum of delta function localized at interfaces  $x_{i+\frac{1}{2}}$  with weights  $\Delta x_{i+\frac{1}{2}}$ ) yields to the quadrature formulae proposed above.

This way of introducing the localization of the source at interfaces permits to extend naturally this approach to non uniform mesh and to multi-dimensional problem.

We have now to solve the Riemann problem including the source term. Thus, we localize the analysis near the interface, assuming the time step small enough such that the waves will not interact from one cell to another. We have assumed  $A$  is diagonal, thus we can treat each component separately i.e. consider that  $U$  is a scalar and the matrix  $A$  reduces to a real number. Let us assume  $A > 0$  for example.

Let us now introduce a mollifier sequence  $\chi_\beta$  of the Dirac measure, for example, we can choose characteristic function with vanishing support  $\chi_\beta(x) = \frac{1}{\beta}$  for any  $\|x\| < \beta/2$  and 0 elsewhere. The regularized local Riemann problem with source reads for  $\|x\| < \beta/2$

$$\partial_t U_\beta + A \partial_x U_\beta = \frac{1}{\beta} R(U_\beta) \Delta x_{i+\frac{1}{2}},$$

and the initial data  $U(x, t=0) = U_L$  for  $x < 0$  and  $U(x, t=0) = U_R$  for  $x > 0$ . Since  $a > 0$  the transport propagates to the right, the solution of this Riemann problem satisfies

$$\begin{cases} U(x, t) = U_L, & \forall (x, t) \text{ s.t. } x < -\beta/2, \\ U(x, t) = U_R, & \forall (x, t) \text{ s.t. } x > At + \beta/2, \\ U(x, t) = U_*, & \forall (x, t) \text{ s.t. } \beta/2x < At + \beta/2, \end{cases}$$

where  $U_*$  is the outgoing value of  $U$  associated to the entering value  $U_L$  after crossing the interval  $[-\beta/2, \beta/2]$ . Note that we do not detail the solution inside the interval since this interval is vanishing in the limit  $\beta \rightarrow 0$ , but we need to compute the outgoing value  $U_*$ . When rescaling the interval  $y = x/\beta$ , the problem becomes

$$\beta \partial_t U(y, t) + A \partial_y U(y, t) = R(U) \Delta x_{i+\frac{1}{2}},$$

and formally, when  $\beta \rightarrow 0$  the problem becomes stationary. Thus, in the limit  $\beta \rightarrow 0$  the outgoing value  $U_*$  is given by the stationary solution.

Let us now solve the stationary problem for the telegraph equation of interest and an arbitrary mollifier. Setting  $\rho = (u + v)/2$  and  $j = (u - v)/2$ , the stationary equation (in variable  $y$ ) associated with (26) reads (let us assume  $\sigma = 1$  for simplifying the notations).

$$\partial_y \rho = -\chi(y)j, \quad \partial_y j = 0.$$

The current  $j$  is constant (equal to  $u_L - v_L$  and to  $u_R - v_R$ ) and the equation for  $\rho$  gives

$$\rho_R - \rho_L = j \int_{-1/2}^{1/2} \chi(y) dy = j.$$

Thus, the stationary solution is characterized by the equations

$$\rho_R - \rho_L = j = j_R = j_L.$$

These equations are independent of the choosen mollifier.

For the telegraph equation ( $A = 1$ ), the eigenvalues are  $\pm 1$  and the Riemann invariant are the functions  $u$  and  $v$  respectively. Then, the solution of the Riemann problem with a localized source terms is defined by

$$\begin{cases} U(x, t) = U_L = (u_L, v_L), & \forall (x, t) \text{ s.t. } x < -t, \\ U(x, t) = U_L^* = (u_L^*, v_L^*), & \forall (x, t) \text{ s.t. } -t \geq x > 0, \\ U(x, t) = U_R^* = (u_R^*, v_R^*), & \forall (x, t) \text{ s.t. } 0 \geq x < t, \\ U(x, t) = U_R = (u_R, v_R), & \forall (x, t) \text{ s.t. } x \geq t, \end{cases}$$

Indeed, the Riemann invariant  $u_L$  is constant through the left wave with speed  $-1$ . The intermediate states  $U_L^*$  and  $U_R^*$  are connected by a

stationnary wave and thus satisfy the above relations. Once expressed in the original variable  $\rho$  and  $j$ , the solution is uniquely defined by

$$\begin{cases} j_R^* = j_L^* = j_0, & \rho_R^* - \rho_L^* = -j_0 \Delta x, \\ \rho_L^* + j_0 = u_L, \\ \rho_R^* - j_0 = v_D. \end{cases}$$

The first equations are nothing but the relations for states connected by a stationnary wave. The last two equations come from the conservation of the  $u$  (resp.  $v$ ) through the wave of speed  $-1$  (resp.  $+1$ ). The solution of this linear system is

$$j_0 = \frac{u_L - u_R}{2 + \Delta x}, \quad \rho_L^* = u_L - j_0, \quad \rho_R^* = v_D + j_0.$$

Then, one can write a Godunov scheme using these exact solution for the Riemann problem.

When performing the same analysis i.e. solving the Riemann problem with a cross section  $\sigma/\varepsilon$  instead of  $\sigma$ ), we find

$$j_0 = \frac{u_L - u_R}{2 + \frac{\sigma \Delta x}{\varepsilon}},$$

and with velocity  $\pm\sqrt{a}$  (as in (22) instead of  $\pm 1$  (as in (26)), we have to replace  $\Delta x$  by  $\Delta x/\sqrt{a}$  and, thus, the coefficient becomes

$$j_0 = M(u_L - u_R), \quad M = \frac{2\varepsilon\sqrt{a}}{2\varepsilon\sqrt{a} + \sigma\Delta x}.$$

Last, we have to project onto piecewise constant solution at iteration  $n+1$ . The average value over the cell is given by

$$U_{i+\frac{1}{2}}^{n+1} = \frac{A\Delta t U_* + (\Delta x_{i+\frac{1}{2}} - A\Delta t)U_{i+1}}{\Delta x_{i+\frac{1}{2}}}.$$

We have a similar formula with  $U_i$  instead of  $U_{i+1}$  in the case  $A < 0$ .

Let us now integrate - in time and space - the obtained solution onto the cell without the thin boundary layer near interfaces i.e. we integrate the equation onto  $[x_{i-\frac{1}{2}} + \varepsilon/2, x_{i+\frac{1}{2}} - \varepsilon/2] \times [t^n, t^{n+1}]$ . By construction, the time derivative gives the difference  $U_i^{n+1} - U_i^n$  and the source term is identically zero. We obtain

$$U_i^{n+1} - U_i^n + \frac{1}{\Delta x_i} \int_0^{\Delta t} F(U_{i-\frac{1}{2}}^\varepsilon(s)) - F(U_{i-\frac{1}{2}}^\varepsilon(s)) ds = 0.$$

The obtained scheme is a direct extension of the, so called well balanced scheme described in [8] for telegraph equation with velocity  $\pm\sqrt{a}$ , variable cross section  $\sigma$  and non uniform mesh:

$$\begin{cases} \frac{\Delta u_i}{\Delta t} + M_{i-\frac{1}{2}} \frac{\sqrt{a}}{\varepsilon \Delta x_i} (u_i - u_{i-1}) = M_{i-\frac{1}{2}} \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} \frac{\sigma_{i-\frac{1}{2}}}{2\varepsilon^2} (v_i - u_i) \\ \frac{\Delta v_i}{\Delta t} - M_{i+\frac{1}{2}} \frac{\sqrt{a}}{\varepsilon \Delta x_i} (v_{i+1} - v_i) = M_{i+\frac{1}{2}} \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} \frac{\sigma_{i+\frac{1}{2}}}{2\varepsilon^2} (u_i - v_i). \end{cases} \quad (28)$$

where  $\frac{\Delta u_i}{\Delta t}$  denotes either the partial derivative of  $u_i$  with respect to  $t$  i.e. a semi-discretized system or a time discretization (e.g.  $\frac{u_i^{n+1} - u_i^n}{\Delta t}$ ) and the coefficient  $M_{i+\frac{1}{2}}$  defined by

$$M_{i+\frac{1}{2}} = \frac{2\sqrt{a}\varepsilon}{\sigma_{i+\frac{1}{2}} \Delta x_{i+\frac{1}{2}} + 2\sqrt{a}\varepsilon}, \quad (29)$$

with  $\sqrt{a}$  is the constant value of the limit diffusion coefficient as defined above and  $\sigma_{i+\frac{1}{2}}$  is an arbitrary average of  $\sigma$  at interface (e.g. arithmetic, harmonic...). The above scheme corresponds to the one proposed in [8] for a uniform mesh,  $\sigma = 2$  and a diffusion coefficient in the limit heat equation equal to  $\frac{1}{2}$ . Note that, in our case, the cross section is not assumed to be constant, which is of main interest from applications point of view.

We can show that (28) is a monotone scheme and then (11) remains an invariant domain during the transport part. It is readily seen that, in the limit  $\max_i(\Delta x_i) \rightarrow 0$ , the coefficient  $M_{i+\frac{1}{2}}$  tends to 1 and the consistency of the scheme (28) with the continuous system (22) is satisfied provided that, in the limit, the mesh is smooth enough i.e. that is locally an uniform mesh ( $\frac{\Delta x_{i+1}}{\Delta x_i} \rightarrow 1$  when  $\max_i(\Delta x_i) \rightarrow 0$ ).

Note that the proposed scheme (28) can be simplified for an uniform mesh and constant cross section. In this case, the coefficient  $M$  is also constant and the equation for  $\frac{\Delta u_i}{\Delta t}$  reads

$$\frac{\Delta u_i}{\Delta t} + M \frac{\sqrt{a}}{\varepsilon \Delta x} (u_i - u_{i-1}) = M \frac{\sigma}{2\varepsilon^2} (v_i - u_i)$$

can be equivalently written as

$$\frac{\Delta u_i}{\Delta t} + \frac{\sqrt{a}}{\Delta x \varepsilon} u_i = M \left( \frac{\sigma}{2\varepsilon^2} v_i + \frac{\sqrt{a}}{\varepsilon \Delta x} u_{i-1} \right). \quad (30)$$

This form is suitable for insuring the positivity of the solution or, equivalently, the condition  $\|j\| \leq \rho$  i.e.  $u_i(t=0) > 0 \forall i \in \mathbb{Z} \Rightarrow$

$u_i(t) > 0$ ,  $\forall i$ ,  $\forall t > 0$  either for the semi-discretized system (when  $\frac{\Delta u_i}{\Delta t}$  represents  $\frac{\partial u_i}{\partial t}$ ) or for semi-implicit schemes (when it represents  $\frac{u_i^{n+1} - u_i^n}{\Delta t}$ ). For semi-discretized system we have

$$u_i(t) \geq u_i(0) \exp(-\frac{\sqrt{a}}{\Delta x \varepsilon} t).$$

A similar argument holds for the positivity of  $v$ .

Let us mention a last equivalent forms of the scheme with uncen-tred source terms like in [1]

$$\frac{\Delta u_i}{\Delta t} + \frac{\sqrt{a}}{\Delta x \varepsilon} (u_i - v_i) = M \frac{\sqrt{a}}{\Delta x \varepsilon} (u_{i-1} - v_i). \quad (31)$$

These equivalences come from the equality

$$M(\frac{\sqrt{a}}{\Delta x \varepsilon} + \frac{\sigma}{2\varepsilon^2}) = \frac{\sqrt{a}}{\Delta x \varepsilon}.$$

### 3.3 Interpretation as HLL scheme

Let us now interpret the obtained scheme in terms of the so called Harten-Lax-Van Leer scheme described in [27].

Indeed, the scheme (28) can be written in the original variables  $(\rho, z)$  and the same for  $(w, j)$

$$\begin{cases} \frac{\partial \rho_i}{\partial t} + \frac{1}{\varepsilon \Delta x_i} (M_{i+\frac{1}{2}} z_{i+\frac{1}{2}} - M_{i-\frac{1}{2}} z_{i-\frac{1}{2}}) = 0, \\ \frac{\partial z_i}{\partial t} + \frac{a}{\varepsilon \Delta x_i} (M_{i+\frac{1}{2}} \rho_{i+\frac{1}{2}} - M_{i-\frac{1}{2}} \rho_{i-\frac{1}{2}}) = \frac{-\lambda_i}{2\varepsilon^2} z_i + \frac{M_{i+\frac{1}{2}} - M_{i-\frac{1}{2}}}{\varepsilon \Delta x_i} (a \rho_i), \end{cases} \quad (32)$$

with

$$z_{i+\frac{1}{2}} = (z_i + z_{i+1} + \rho_{i+1} - \rho_i)/2, \quad \rho_{i+\frac{1}{2}} = (\rho_i + \rho_{i+1} + z_{i+1} - z_i)/2,$$

and

$$\lambda_i = \frac{\Delta x_{i+\frac{1}{2}}}{\Delta x_i} M_{i+\frac{1}{2}} \sigma_{i+\frac{1}{2}} + \frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} M_{i-\frac{1}{2}} \sigma_{i-\frac{1}{2}}.$$

Once again, the consistency of the scheme (when  $\Delta x \rightarrow 0$  for fixed  $\varepsilon$ ) requires a asymptotically regular mesh. More precisely, this means that, when the mesh is refined, it becomes locally regular  $\frac{\Delta x_{i-\frac{1}{2}}}{\Delta x_i} \rightarrow 1$  as  $\max_i \Delta x_i \rightarrow 0$ .

One can also check the asymptotics  $\varepsilon \rightarrow 0$  (with fixed  $\Delta x_i$  presumably non uniform) using the above form. It is readily seen that  $z$

has to remain small and, more precisely, of the order of magnitude of  $O(\varepsilon)$ . On the other hand, the limit behaviour of  $M$  is given by

$$M_{i+\frac{1}{2}} \sim \frac{2\varepsilon\sqrt{a}}{\sigma_{i+\frac{1}{2}}\Delta x_{i+\frac{1}{2}}},$$

and for the  $\lambda$  coefficient

$$\lambda_i \sim \frac{4\varepsilon\sqrt{a}}{\Delta x_i}.$$

Moreover, the last term of the r.h.s. is small and the leading order of the equation for  $z$  is

$$\frac{a}{\varepsilon\Delta x_i} \left( \frac{2\varepsilon}{\sigma_{i+\frac{1}{2}}\Delta x_{i+\frac{1}{2}}} \rho_{i+\frac{1}{2}} - \frac{2\varepsilon}{\sigma_{i-\frac{1}{2}}\Delta x_{i-\frac{1}{2}}} \rho_{i-\frac{1}{2}} \right) = \frac{-4\varepsilon}{2\sqrt{a}\varepsilon^2\Delta x_i} z_i.$$

Furthermore, in the limit  $\varepsilon \rightarrow 0$ , we have (since  $z_i$  is of order  $\varepsilon$ )

$$z_{i+\frac{1}{2}} = (\rho_{i+1} - \rho_i)/2.$$

Then, reporting the last expressions in the equation for  $\rho_i$ , one obtain the discretization of the heat equation on a nonuniform grid. Indeed, retaining only the first order terms in the preceeding expressions, the equation for  $\rho_i$  in (32) becomes

$$\frac{\partial \rho_i}{\partial t} + \frac{\sqrt{a}}{\Delta x_i} \left( \frac{\rho_{i+1} - \rho_i}{\Delta x_{i+\frac{1}{2}}} - \frac{\rho_i - \rho_{i-1}}{\Delta x_{i-\frac{1}{2}}} \right) = 0.$$

Note that the formulae can be simplified for uniform mesh and constant cross section : in this case, we have  $\lambda = 2\sigma M$  and the second term of the right hand side vanishes. Then, the proposed scheme reduces to a classical Godunov scheme

$$\varepsilon \frac{\partial U_i}{\partial t} + M(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}})/\Delta x = \frac{M}{\varepsilon} R(U_i), \quad (33)$$

where the flux at interfaces are given by

$$F_{i+\frac{1}{2}} = [b_i F_{i+1} + b_{i+1} F_i + b_i b_{i+1} (U_{i+1} - U_i)] / (b_i + b_{i+1}), \quad (34)$$

which are just upwind fluxes for (22) with  $b_i = \sqrt{a}$ . The multiplicative coefficient  $M = \frac{2D\varepsilon}{\sigma\Delta x + 2D\varepsilon}$  comes from the well balanced scheme where  $D$  is the expected diffusion coefficient i.e.  $h(0)$ .

This form is useful because it is expressed in the original variable and this can be easily generalized to nonlinear case. The discretization is a sum of a diffusive term and classical convective term. Thus, this discretization can be interpreted as a particular choice of adding numerical viscosity depending of  $\varepsilon$ .



### 3.4 Comparison

Let us now detail the comparison with other proposed scheme for similar models.

For example, the scheme proposed in [15] can be written, using our notations, as an interpolated scheme between an upwind scheme and a centered one. Starting from the system in the form (2) with  $h = 1$  i.e. the telegraph equation, we set  $j = \varepsilon \bar{j}$

$$\begin{cases} \partial_t \rho + \partial_x \bar{j} &= 0 \\ \partial_t \bar{j} + (1 + (1 - \varepsilon^2)/\varepsilon^2) \partial_x \rho &= \frac{-\sigma}{\varepsilon^2} \bar{j}. \end{cases} \quad (35)$$

Let us split the system into a transport part with velocity  $\pm 1$  which is discretized using a upwind scheme (in variable  $u$  and  $v$ ) and the remaining part of the term  $\partial_x(\rho h)$  is taken as a source term, using a centered scheme. Once back into the original variable  $(\rho, \bar{j})$ , the semi-discretized system, on a uniform grid in space, reads

$$\begin{cases} \partial_t \rho_i + \frac{1}{2\Delta x} (\bar{j}_{i+1} - \bar{j}_{i-1} - (\rho_{i+1} + \rho_{i-1} - 2\rho_i)) = 0 \\ \partial_t \bar{j}_i + \frac{1}{2\Delta x} (\rho_{i+1} - \rho_{i-1} - (\bar{j}_{i+1} + \bar{j}_{i-1} - 2\bar{j}_i)) = \frac{-\sigma}{\varepsilon^2} \bar{j}_i \\ \quad + \frac{1}{\varepsilon^2} [\sigma \bar{j}_i + (1 - \varepsilon^2)(\rho_{i+1} - \rho_{i-1})/(2\Delta x)]. \end{cases} \quad (36)$$

The above discretization can be written equivalently in  $\rho, j$  as

$$\begin{cases} \partial_t \rho_i + \frac{1}{\varepsilon} \frac{1}{2\Delta x} (j_{i+1} - j_{i-1}) - \frac{1}{2\Delta x} (\rho_{i+1} + \rho_{i-1} - 2\rho_i) = 0 \\ \partial_t j_i + \frac{1}{\varepsilon} \frac{1}{2\Delta x} (\rho_{i+1} - \rho_{i-1}) - \frac{1}{2\Delta x} (j_{i+1} + j_{i-1} - 2j_i) = \frac{-\sigma}{\varepsilon^2} j_i \end{cases} \quad (37)$$

or,

$$\begin{cases} \partial_t \rho_i + \frac{1}{\varepsilon} \frac{1}{2\Delta x} [(1 - \varepsilon)(j_{i+1} - j_{i-1}) + \\ \quad + \varepsilon(j_{i+1} - j_{i-1} - (\rho_{i+1} + \rho_{i-1} - 2\rho_i))] = 0 \\ \partial_t j_i + \frac{1}{\varepsilon} \frac{1}{2\Delta x} [(1 - \varepsilon)(\rho_{i+1} - \rho_{i-1}) + \\ \quad + \varepsilon(\rho_{i+1} - \rho_{i-1} - (j_{i+1} + j_{i-1} - 2j_i))] = \frac{-\sigma}{\varepsilon^2} j_i \end{cases} \quad (38)$$

i.e. a centered scheme for the  $(1 - \varepsilon)$  part and the upwind scheme for the  $\varepsilon$  part.

The proposed scheme can also be related to scheme proposed recently in [1, 20] and compared with previous works like [11, 17, 15].

## 4 Numerical results

We shall now present 2 numerical tests. The first is a validation of our method with a strongly variable cross section. The second case is a more complex case with a coupling with heat equation for material. The scheme is implicit in time which leads to the solving of a band matrix system.

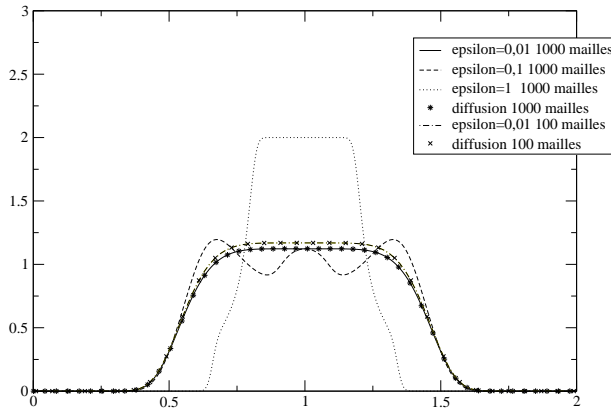
#### 4.1 Variable opacity coefficient

Let us consider the case

$$x \in [0, 2], \sigma(x) = 100(x - 1)^4.$$

This corresponds to a case of transparent media in the center ( $\sigma = 0$  for  $x = 1$ ) with opaque walls at boundary ( $x = 0$  and  $x = 2$ ).

The initial data is a characteristic function, for  $\rho$  with support in  $[\frac{1}{2}, \frac{3}{2}]$ . The initial flux  $j$  is equal 0. The simulated time is  $T = 0.1$  and the small parameter value takes the following values :  $\varepsilon = 0$  (i.e. the diffusion case ) and the following values  $10^{-2}, 0.1, 1$ . The mesh is uniform with either 100 or 1000 points and the time step is chosen such that  $\Delta t / \Delta x = 0.05$ . Note that the expected time step restriction for the transport part (C.F.L. condition) is much more restrictive  $\frac{\Delta t}{\Delta x} \leq \varepsilon$ , and, similarly, the characteristic relaxation time is such that  $\Delta t \sigma \leq \varepsilon^2$ .



The computation using various number of discretization points indicates that the solution converges when the mesh size (and thus the time step) goes to 0. This is some kind of numerical consistency result. On the other hand, when  $\varepsilon$  goes to 0, we obtain a solution that converges toward the one of a diffusion equation with the diffusion coefficient  $1/3$ . This illustrates that the proposed scheme is compatible with the diffusive asymptotics.

#### 4.2 Coupling with material

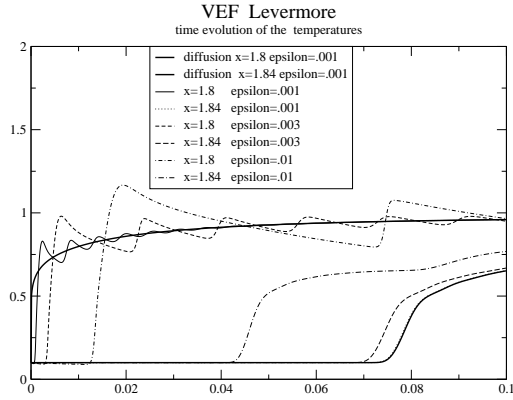
Let us consider the system coupled with an heat equation for some material. This test can be seen as a simplified model of laser-plasma interaction

$$\begin{cases} \varepsilon \partial_t \rho + \partial_x j = \frac{\tau}{\varepsilon} (K - \rho) \\ \varepsilon \partial_t j + \partial_x \rho h(j/\rho) = -\frac{\tau + \sigma}{\varepsilon} j \\ C_v \partial_t T = \frac{-\tau}{\varepsilon^2} (K - \rho), \end{cases} \quad (39)$$

$\sigma$  is the scattering coefficient for photons,  $\tau$  is the absorption coefficient of the material.  $K$  can be either given or equal to  $T^4$  following Stefan law. We have  $\tau = \frac{C_v}{T^3 \varepsilon^2}$ .

The choosen closure relation is the Levermore-Lorentz one [19,18] given by (7). We are using a time splitting :one time step for the moment systems in  $(\rho, j)$  with fixed temperature  $T$  and then, the system for  $(\rho, T)$ . The first test consists of a domain, scaled to  $[0, 2]$ , between 2 walls. Let us summarize the scaled parameters of the simulation

	$x < 0.1$	$0.1 < x < 0.2$	$0.2 < x < 1.8$	$1.8 < x < 1.9$	$x > 1.9$
$\sigma$	$+\infty$	0	0	0	$+\infty$
$C$	0	10	0	10	0
$\rho^0$	0	16	0	$10^{-4}$	0
$j^0$	0	0	0	0	0
$T^0$	0	2	0	$10^{-1}$	0



Initially, the left wall is hot and, then, by radiation, it warms the right one. We plot the evolution of temperature due to this heating

at the surface  $x = 1.8$ , and inside the right wall for  $x = 1.84$ . The computations were made with  $\Delta x = 1/500$  and  $\Delta t = 10^{-4}$ .

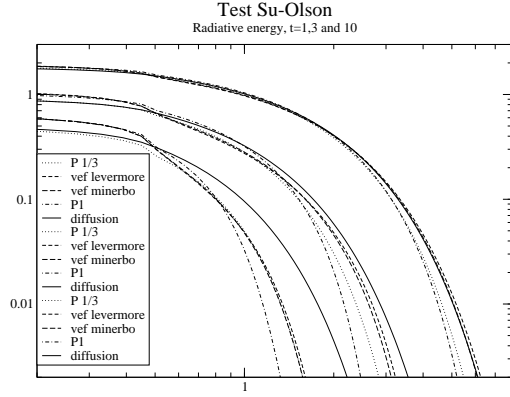
We observe on figure 2 that the solution converges toward those of the diffusion model when  $\varepsilon \rightarrow 0$  as expected. Note that the dependence with respect to  $\varepsilon$  is weaker in the material (at  $x = 1.84$ ) than at the surface (at  $x = 1.8$ ). The front of heating reaches the surface  $x = 1.8$  at time  $t = 0.012$  for  $\varepsilon = 0.01$  and this time goes to zero as  $\varepsilon \rightarrow 0$  as expected for a diffusive equation. The oscillations (in time) of the temperature at  $x = 1.8$  (surface of the material) are due to the heat wave that rebounds between the walls. The speed of the wave goes to infinity as  $\varepsilon \rightarrow 0$  and in the diffusive limit, the heat propagates instantaneously and, thus, the oscillations disappear. The front of heat is less sharp within the material and the delay to warm the material increases as  $\varepsilon \rightarrow 0$  (from  $t = 0.05$  at  $\varepsilon = 0.01$  to  $t = 0.075$  when  $\varepsilon \rightarrow 0$ ).

The second test is taken from [21], page 625, with a constant opacity but with  $C_v = \alpha T^3$  and a source term  $S = 1$  localized in  $\|x\| \leq 1/2$ . All coefficients in (39) are taken to one thus the system of equations to solve is

$$\begin{cases} \partial_t \rho + \partial_x j = (K - \rho) + S \\ \partial_t j + \partial_x \rho h(j/\rho) = -j \\ \partial_t K = -(K - \rho). \end{cases} \quad (40)$$

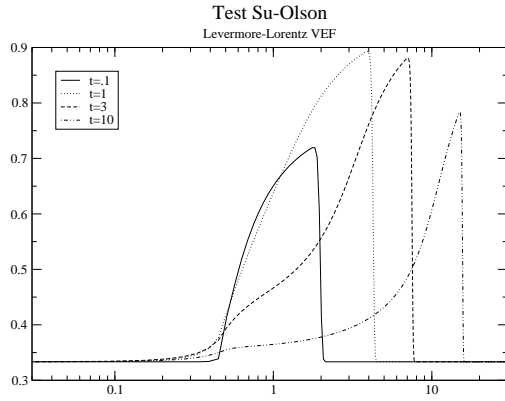
The computations were made with  $\Delta x = 3/50$  and  $\Delta t = 1/100$ . Solutions, with different Eddington factors are viewed at times  $t = .1, 1, 3$  and  $t = 10$ .

At early time, the discrepancies between various variable Eddington factors (VEF) and diffusion are important since diffusion is a less accurate model for short time. These differences, as expected, tend to reduce to zero as the time growing. At time  $t = 10$ , these discrepancies are negligible. We also show the results for the  $P1$  model with  $h = 1/3$  and for the, so called,  $P_{1/3}$  model derived from the  $P1$ , (see [21] for a detailed presentation), by taken  $h = 1$  and by multiplying the opacity by a factor 3. For these two constant Eddington factors, the discrepancies with diffusion results, still exist at time  $t = 10$ . This proves that is is important to use the correct closure relation or Eddington factor in order to recover the right behaviour of the solutions in particular at early time but also after rather long time.



Moreover, our results are very close to those obtained in [21], with a important difference : our scheme for variable Eddington factors gives no noisy solutions, in contrast with the corresponding results presented in [21]. This fact illustrates the robustness of our scheme.

We also show the profile of the Levermore-Lorentz Eddington factor at the same times.



for which we can give the same conclusion as for the precedent figure: no noisy solution for variables Eddington factors with our scheme. Thus, the instability or noise observed for the VEF calculations in [21] are certainly due to a bad discretization of nonlinear hyperbolic problem.

## 5 Conclusions

The proposed scheme has all the required properties announced in the introduction.

The scheme consists of two steps : first to replace the nonlinear into two independant and identical linear system of telegraph equations and, second, the use of a well balanced scheme for each of the two systems. The interpretation of the well balanced scheme as a Godunov scheme using the Riemann solution of hyperbolic system with source term (section 3.2) can be extended to more complex relaxation term.

A first example arises from relativistic effect as presented in the models described in [4]. The obtained equation is, in such case, a Burgers equation with diffusion instead of the heat equation in the case considered here. It can also be extended to discrete velocity models of kinetic equation in the diffusive regime with a linear collision operator of Lorentz type (see [2] for a detailed presentation). A third extension is to consider two dimensional (in space) model. The use of alternate direction (i.e. a splitting between the  $x$  and  $y$  direction) is, in diffusive regime, not suitable. Our method yields to choose distincts quadrature points for the transport in direction  $x$  and  $y$  respectively, or, in other words, it yields to localize the source terms at interfaces.

Moreover, the proposed approach can be combined with adaptive mesh refinement technics since it requires to evaluate the flux at interface by solving stationnary Riemann problem as explained in subsections 3.2 and 3.3. These are, in our opinion, promissing direction for forthcoming developpements of the proposed scheme.

The main drawback of the proposed scheme is that the choice of the velocity  $a$  for the relaxed system (2) is non local and thus, the diffusive regime is obtain only when the whole domain is in isotropic equilibrium. This is a rather severe limitation for the use of this scheme in complex situations. One possible solution is to use domain decompo-

sition and to use a different value of  $a$  in the different domains. We also mention a forthcoming paper [6] that proposed a variant of the proposed scheme for which the choice of the coefficient  $a$  is no more local.

## References

1. Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., Perthame, B. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comp.*, 25(6):2050–2065, 2004
2. Buet C., Cordier S., Lucquin-Desreux L. Mancini S. Diffusion limits of the Lorentz model: Asymptotic preserving schemes, *Meth. Math. Anal. Num.*, 36, 4, p. 631-655, 2002.
3. Christophe Buet, Stéphane Cordier, Laura Mereuta, The Riemann problem for hyperbolic system arising in radiatif transfert modelling. Manuscript 2002.
4. Buet C., Desprès B., Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics", *JQSRT*, Volume 85, Issues 3-4, 385-418 (2004) ..
5. Buet C., Cordier S., Asymptotic Preserving Scheme for radiative hydrodynamics model, *CRAS, Series I*, vol 338, 2004
6. Buet C., Desprès B., Splitted Approximations of Radiative Hydrodynamics and Asymptotic Preserving Schemes, in preparation, 2004. *titre ?????*
7. Caflish R.E., Jin S., Russo G., Uniformly Accurate Schemes for Hyperbolic Systems with Relaxation, *SIAM Journal on Num. Anal.*, 34, p. 246-281, 1997
8. Gosse, L., Toscani, G., An asymptotic preserving well-balanced scheme for the hyperbolic heat equation, *CRAS Série I*, 334, p. 1-6, 2002
9. Greenberg, J., Le Roux, A., A well balanced scheme for the numerical processing of source terms in hyperbolic systems, *SIAM J. Num. Anal.* 33, p. 1-16, 1996
10. Honingh, A., Anisotropic radiation fields : causality and quantum statistics, PhD dissertation, Institut for Theoretical Physics, university of Amsterdam, 2001.
11. Jin S., Levermore C.D., Numerical Schemes for Hyperbolic Systems of Conservation Laws with Stiff Diffusive Relaxation, *J.C.P.*, vol126, 1996
12. Jin S., Xin Z., The relaxation schemes for systems of conservation laws in arbitrary space dimensions, *Comm. Pure Appl. Math.*, 48, p. 235–276., 1995
13. Jin, S. Levermore, C.D. : Numerical Schemes for Hyperbolic Conservation Laws with Stiff Relaxation Terms, *Journal of computational physics* 126, 449-467, Artivle No.0149, 1996
14. Jin S., Pareschi, L. Toscani G., Diffusive relaxation schemes for multiscale discrete velocity kinetic equations, *SIAM J Numer. Anal.*, 35, 2405, 1998.
15. Jin S., Pareschi, L. Toscani G., Uniformly accurate diffusive relaxation schemes for multiscale transport equations, *SIAM J Numer. Anal.*, 38, 913, 2000.
16. Jin, S., Efficient asymptotic-preserving schemes (APS) for some multiscale kinetic equations, *SIAM J. Sci. Comput.* 21, p 441-454, 1999
17. R.J. LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods : the quasi-steady wave-propagation method. *Journal of Computational Physics*, 146:346-365, 1998.

18. Levermore, C.D., Relating Eddington factors to flux limiters, JQSRT, 31 (2), p. 149-160, 1984
19. Minerbo, G.N., Maximum entropy Eddington factors, JQSRT, 20, p. 541-545, 1978.
20. Naldi G., Pareschi L., Numerical Schemes for Hyperbolic Systems of Conservation Laws with Stiff Diffusive Relaxation, SIAM Journal on Num. Anal., 37, p. 1246-1270, 2000
21. G. L. Olson, L.H. Auer, M.L. Hall Diffusion, P1, and other approximate forms of radiation transport, JQRST 64, pp 619-634, 2000
22. G.C.Pomraning: The Equation of Radiation Hydrodynamics, (Pergamon Press.Oxford (1973))
23. J.M. Smit, L.J. van den Horn, S.A. Bludman, Closure in Flux Limited Neutrino Diffusion and Two Moment Transport, Astronomy and Astrophysics 356, 559, 2000.
24. J.M. Smit, J.Cernohorsky, and C.P. Dullemond, Hyperbolicity and critical points in two-moment approximate radiative transfer , Astrophys. 325, 203 ,1997.
25. B. Su and G.L. Olson, Benchmark results for the non-equilibrium marshak diffusion problem, J. Quant. Spectr. Rad. Transfer 56, 337-351,1996.
26. B. Su and G. L. Olson, Ananalytical benchmark for non-equilibrium radiative transfer in anisotropically scattering medium, Annals of Nuclear Energy, vol. 24, no. 13, pp. 1035?1055, 1997.
27. E.F.Toro : *Riemann Solvers and Numerical Methods for Fluid Dynamics* , Springer-Verlag, Second Edition, Chapter 10, 1999.



# Asymptotic Preserving and Positive Schemes for Radiation Hydrodynamics

Christophe Buet<sup>a</sup> Bruno Despres<sup>a</sup>

<sup>a</sup>*Département des Sciences de la Simulation et de l'Information, Commissariat à  
l'Énergie Atomique, 91680, Bruyères le Chatel, BP 12, France*

---

## Abstract

In view of radiation hydrodynamics computations, we propose an implicit and positive numerical scheme that captures the diffusion limit of the two-moments approximate model for the radiative transfer even on coarses grids. The positivity of the scheme is equivalent to say that the scheme preserves the limited flux property. Various test cases show the accuracy and robustness of the scheme.

*Key words:* radiation hydrodynamics, diffusion limit, flux limited and monotone scheme.

---

## 1 Introduction

The aim of this work is about accurate discretization of radiation hydrodynamics by means of Eulerian finite volume methods. In this direction we study and discretize a non linear two-moments model (2) for radiation in the Eulerian frame written in one space dimension. The model, called in the next  $M^1$ , derives from a maximum entropy principle as in [10,28,19,20,27,5,1]. This model and the related ones are stiff models, the stiffness parameter is  $\varepsilon$  small. For ICF (inertial confinement fusion applications)  $\varepsilon \approx 10^{-3}$ , for example see [31]. There are some regions of the space with very little interactions between the radiation and the matter and other regions with strong interactions with the matter. The case of little interactions is called the streaming regime. For strong interaction it is possible to derive diffusion approximations of (2). This is called the diffusion regime. Among the numerical methods that have been

---

*Email addresses:* `christophe.buet@cea.fr` (Christophe Buet),  
`bruno.despres@cea.fr` (Bruno Despres).

proposed for the discretization of simplified radiation models, we distinguish between discretization of diffusion approximations and direct discretization of two-moment models. Concerning numerical method for the discretization of diffusion approximations we quote [18] in which a technique of flux limiting is discussed in details. The technique of flux limiting somehow extends the domain of validity of these diffusion approximations to the streaming regime. We exhibit a new intermediate interaction regime, of pure hyperbolic type, for which the limit equation can be written in divergent form or in non divergent form. We privilege the divergence form of this equation. This is different from the work of [9] where the non divergent formulation is used for the radiation energy. This is important for the numerics since a non divergent discretization may converge to wrong discontinuous solutions [17]. It has been noticed since a long time, in the work of Mihalas [25] for example, that a direct discretization of a two-moments models can help to get a better discretization for both streaming and diffusion regimes. It is also possible that higher moments models could treat more efficiently the anisotropy of radiation in dimension greater than one.

In this direction, a preliminary question is how to discretize "correctly" a two-moments model for the radiation in the non relativistic case in dimension one, that is for  $\varepsilon$  small, and at the matter scale. That is, we desire to adapt the grid according to the length scale of the matter. From the point of view of photons the grid is coarse. Among the works dedicated to this specific problem we point out [22–24,2]. We show on a simple example that freezing naively the Eddington factor during the time step gives a non positive implicit linear discretization of the problem. This is based on the fact that the linear implicitation of the HLL solver is non positive if the equation is non-linear, that is if the Eddington factor is not constant. The new method we propose has the following features

**Positivity of the scheme:** the scheme guarantees the positivity in 1D

$$-E_r \leq F_r \leq E_r$$

where  $E_r$  is the energy of radiation and  $F_r$  is the radiation flux. This is equivalent to say the scheme is **flux limited**

$$\frac{|F_r|}{E_r} \leq 1. \tag{1}$$

The physical meaning of flux limitation is more natural with (1). This property is useful essentially in the streaming regime for which  $|F_r| \approx E_r$ . This flux limiting has nothing to do with the numerical tricks used for the discretization of diffusion approximations of two-moments models.

**Correct diffusion limits:** the equilibrium and non equilibrium diffusion limits of the scheme are correct even for a coarse mesh,

**Correct hyperbolic limit:** the numerical treatment of the purely hyperbolic limit (9) is compatible with the divergent form of the equation.

**Time step requirement:** the scheme is explicit for the hydrodynamic part and totally implicit for the radiation part. Therefore the only CFL limitation comes from the hydrodynamic part.

The first three properties are true at the PDE level for the basic two-moments model. Since the basic two-moments model considered in this paper is highly non-linear (see (2)), a direct implicit discretization of this model may be tricky if the material velocity  $v$  is non zero. We now describe the strategy we propose to obtain these properties at the numerical level. We think this approach is simpler.

The first ingredient to obtain all these properties is what we call a splitted approximation. Usually this kind of model is called a co-moving frame radiation hydrodynamics model, see [26,21,2]. The model consists in transporting the radiation entropy and the radiation entropy flux at velocity  $v$ , then to discretize with the matter at rest with  $v = 0$ . This splitted approximation has almost zero additional numerical cost. The maximal hyperbolic wave velocity of the splitted approximation is  $\frac{1}{\varepsilon} \pm v$  which slightly exceeds the physical maximal wave velocity  $\frac{1}{\varepsilon}$  which is the velocity of light in non-dimensional variables. This is shared with other co-moving frame radiation hydrodynamics model like [26,21,2]. We think that slight violation of the velocity of light in very rare cases is much better than diffusion approximations without flux limiting of the diffusion coefficient. Even with flux limiting of the diffusion coefficient it is difficult to guarantee a correct velocity for the front propagation in transparent materials. In this work we only consider an explicit discretization of the hydrodynamics part. On the other hand, the radiative step of the splitted model is discretized by an implicit method. The reason is the high speed of the radiation signal which is much greater than the mean velocity of the matter. Since we are interested in the time scale of the matter, the radiation must be treated implicitly.

The second ingredient lies in the discretization of the radiative step. The numerical scheme for the radiative step must be consistent with the diffusion approximation in opaque region and must preserve the flux limited property. It is well known, [14,15,29], that in opaque regions standard finite volume schemes lead to a wrong coefficient of diffusion for coarse grids. It is illuminating to consider the hyperbolic heat equation

$$\partial_t u + \frac{1}{\varepsilon} \partial_x v = 0, \quad \partial_t v + \frac{1}{\varepsilon} \partial_x u = -\frac{1}{\varepsilon^2} u.$$

The asymptotic regime  $\varepsilon \rightarrow 0^+$  is the heat equation  $\partial_t u - \partial_{xx} u = 0$ . The problem stressed in [14,15,29] is that a discretization with stable upwinded

schemes of the hyperbolic equation has the asymptotic limit

$$\partial_t u - \left(1 + \frac{C\Delta x}{\varepsilon}\right) \partial_{xx} u = 0, \quad C > 0.$$

Therefore the discretization of the hyperbolic equation on a coarse grid may have the limit

$$\partial_{xx} u = 0$$

which is wrong. Last decade "asymptotic preserving schemes" (AP scheme), [14,15,29,12], have been developed in order to recover the correct diffusion asymptotic limit. Thus AP schemes seems to be the good tool for radiative two-moment models. But actual AP schemes suffer of some limitations. Semi implicit AP scheme like those proposed in [15,14] can handle non-linear system but are, in the best case, only positive and stable under a parabolic CFL condition and this not acceptable for our purpose since we do not want such time step limitation. The AP scheme proposed by Gosse in [12] has a better stability property, but it is difficult to extend this scheme for non linear systems: Moreover the method used in [12] seems quite impossible to extend for a resonant problem like the one we study in this work. Therefore we have developed a new discretization of radiation part of the model. This implicit discretization is unconditionally positive, and has the correct diffusion limit. It is based on two basic schemes. The first one is a "relaxation" scheme, see for example [16], which has the advantage to transform the initial non linear problem into a linear one: this is why we solve two linear systems at each time step. The second one is the AP scheme described [12], in order to obtain the right asymptotic diffusion limit with no time step limitation for the positivity. This discretization of the radiative step is an improved version of the one proposed in [7], it can handle more efficiently opaque and transparent zones. Concerning the cost of such a discretization, let us mention that in one dimension the matrix can be inverted by the a low cost algorithm like the LU one.

The plan is as follows. In the second section we study the  $M^1$  model for the radiation and recall various diffusion approximations of this model. In the third section we derive what we call splitted approximations for the  $M^1$  model. Numerical schemes for each part of the splitting are proposed and analyzed in the fourth section. In the fifth section we show some numerical results to demonstrate the features of the scheme, with application to the full system of hydrodynamics coupled with radiation. The final section is the conclusion. We sketch future developments of the method.

## 2 Basic equations

### 2.1 The two-moments model $M^1$

The starting point of the analysis is the the following two-moments model for radiation in the Eulerian frame and written with non dimensional variables in one space dimension (see [1] for a full derivation)

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} F_r = \frac{\gamma \sigma_a}{\varepsilon^2} (T^4 - E_r + \varepsilon v F_r) - \frac{\gamma \sigma_s}{\varepsilon} v F_r^0, \\ \frac{\partial}{\partial t} F_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} P_r = -\frac{\gamma \sigma_a}{\varepsilon^2} (F_r - \varepsilon v (T^4 + P_r)) - \frac{\gamma \sigma_s}{\varepsilon^2} F_r^0, \\ P_r = \chi E_r, \quad \chi = \frac{3+4f^2}{5+2\sqrt{4-3f^2}}, \quad f = F_r/E_r. \end{cases} \quad (2)$$

In this model  $E_r$  is the energy of radiation,  $F_r$  is the flux of radiation,  $P_r$  is the pressure of radiation and  $\chi$  is called the Eddington factor.  $f$  measures the anisotropy of radiation. For physical reasons the modulus of  $f$  is bounded by one,  $|f| \leq 1$ . This Eddington factor was proposed first by Levermore [18] and can be recovered from a maximum entropy principle as in [28,19,20,27,5]. For a matter at rest, it can be checked, by means of the Godunov method, that solutions of the system (2) for a general Eddington factor  $\chi$  verify this flux limited property provided that the initial condition satisfy it and if the Eddington factor  $\chi$  is a even convex positive function satisfying  $\chi(\pm 1) = 1$ . The Eddington factor considered here have these properties <sup>1</sup>.

This system has two exterior parameters. The first one is  $T$  the temperature of the matter.  $\varepsilon$  is a small parameter that measures the velocity of light versus the sound velocity in the matter. For non relativistic plasmas the velocity of the matter can be assumed of the same magnitude than sound velocity in the matter, therefore is also small with respect to the velocity of light. The second exterior parameter is  $v$  the velocity of the matter. Therefore  $\gamma = 1/\sqrt{1 - \varepsilon^2 v^2} \geq 1$  is very close to one for non relativistic plasmas.  $F_r^0$  is the radiation flux energy measured in the co-mobile reference frame that moves with the matter  $F_r^0 = \gamma^2((1 + \varepsilon^2 v^2)F_r - \varepsilon v(E_r + P_r))$ . The particular form of the right hand side is due to the underlying compatibility of the relaxation source terms with the Lorentzian invariance of radiation. The absorption-emission opacity  $\sigma_a$  and the scattering opacity  $\sigma_s$  are mean opacities. In practice there are functions of  $T$  and the density of the matter. This model, (2), comes

---

<sup>1</sup> Let us mention that an approximation of (2) is called the  $P^1$  model, for which the Eddington factor is a constant function  $\chi = 1/3$ . This value comes from the choice  $f = 0$  in (2). Since the  $P^1$  model is linear, it is trivial to check that the solutions of this model do not verify the flux limited property.

from a maximum entropy principle with Lorentzian invariant source terms [1]. In particular one can take  $\gamma = 1$  in (2). Other simpler sources have been proposed in the literature [21]. The conclusions of the present study are easy to generalize to the source terms of [21] which are very close to the ones used in this work.

## 2.2 Hyperbolicity and entropy of the radiation

The  $M^1$  model (2) is strictly hyperbolic for  $0 \leq f^2 < 1$  and  $1 < f^2 < \frac{4}{3}$ . For  $f^2 = \frac{4}{3}$  is non more differentiable. For  $f = 1$  the Jacobian matrix of the flux is

$$\frac{\partial(F_r, P_r)}{\partial(E_r, F_r)} = \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix}. \text{ Therefore the system is only weakly hyperbolic at } f = \pm 1.$$

We notice that  $f = \pm 1$  corresponds to a strongly non isotropic radiation flux. Since the eigenvalues coincide for  $f = \pm 1$  it means the system is resonant for  $|F_r| = E_r$ .

The method of moment assumes the intensity of the radiation is the generalized Planckian  $\frac{4\pi^5}{15} \frac{I}{\nu^3} = \frac{1}{e^{\frac{\nu}{T_r} + \frac{\nu b n}{T_r}} - 1}$ . The energy of radiation and the radiation flux can also be computed with respect to  $T_r$  and  $b$ . One has  $E_r = \int \int I(\nu, n) d\nu dn$ ,  $F_r = \int \int n I(\nu, n) d\nu dn$  and  $P_r = \int \int n \otimes n I(\nu, n) d\nu dn$ . After some standard calculations [28,1] one gets

$$E_r = T_r^4 \frac{3 + b^2}{3(1 - b^2)^3}, \quad F_r = -T_r^4 \frac{4b}{3(1 - b^2)^3}. \quad (3)$$

Let  $S_r$  and  $Q_r$  be the entropy and entropy flux of the radiation defined by  $S_r = -\frac{15}{4\pi^5} \int \int \nu^2 (n \log n - (n+1) \log(n+1)) d\nu dn$  and  $Q_r = -\frac{15}{4\pi^5} \int \int \nu^2 (n \log n - (n+1) \log(n+1)) n d\nu dn$ . One gets the simplified expressions

$$S_r = \frac{4}{3} \frac{T_r^3}{(1 - |b|^2)^2}, \quad Q_r = -b S_r. \quad (4)$$

The mapping  $(E_r, F_r) \mapsto (T_r, b)$  is degenerate for  $b = \pm 1$ . Since  $\frac{F_r}{E_r} = f = -\frac{4b}{3+b^2}$ , then  $b = \pm 1$  is equivalent to  $f = \mp 1$ . It has been proved in [1] (anyway this is compatible with the theory of moment for hyperbolic system of conservation laws, see [18]) that an algebraic combination of the two equations of the system (2) gives the equation  $\frac{\partial}{\partial t} S_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} Q_r = \frac{1}{T_r} S_E + \frac{b}{T_r} S_F$ . This comes from the second principle of thermodynamic for the moment model, namely  $T_r dS_r = dE_r + b dF_r$ .

### 2.3 Equilibrium diffusion limit

Very classically [21,22] the equilibrium diffusion limit of (2) is recovered with the scaling  $\sigma_a = O(1)$  and  $\sigma_s = O(1)$ . Then of course the dominant contribution in the right hand side of (2) is due to  $\sigma_a$ . One gets  $E_r = T^4 + O(\varepsilon)$ ,  $F_r = -\varepsilon \frac{1}{\sigma_a + \sigma_s} \frac{\partial}{\partial x} P_r + \varepsilon \frac{4}{3} T^4 + O(\varepsilon^2)$  and  $P_r = \frac{1}{3} T^4 + O(\varepsilon)$ . Since  $F_r = O(\varepsilon)$  then  $\chi \approx \frac{1}{3}$ . Assuming  $v = 0$  a typical diffusion equilibrium model for the temperature is [21]

$$\frac{\partial}{\partial t}(T + T^4) = \partial_x \left( \frac{1}{3(\sigma_a + \sigma_s)} \partial_x T^4 \right). \quad (5)$$

This type of model is valid only in opaque materials, but is also used outside of its natural validity domain in transparent materials for which  $\sigma_a$  and  $\sigma_s$  are very small. To see what is the problem in transparent materials it is sufficient to notice that the model is degenerate parabolic and admits finite speed of propagation for the fronts  $T = 0$ . At the front the temperature can be expanded like  $T \approx C_1 |x - x_f(t)|^{\frac{1}{3}}$ . The speed of propagation of the front is  $x'_f(t) = \frac{C_2}{3(\sigma_a + \sigma_s)} \partial_x T^3$ . Provided  $\partial_x T^3 = O(1)$  the propagation of the front is highly non physical in transparent materials since the  $|x'_f(t)| \rightarrow \infty$ . In any cases  $|x'_f(t)|$  should be smaller than  $\frac{1}{\varepsilon}$ .

### 2.4 Non-equilibrium diffusion limit

It has recently been noticed that the scaling  $\sigma_a = O(\varepsilon)$  and  $\sigma_s = O(1)$  helps to get a non-equilibrium diffusion limit of the  $M^1$  model. Here non-equilibrium means that  $E_r \neq T^4$ . It explains why  $\sigma_a$  needs to be small. On the other hand  $\sigma_s$  is kept of order one, so that the  $M^1$  model admits a diffusion limit. Thus the model we presents hereafter is a simple non trivial model for the coupling of radiation and hydrodynamics. With the above scaling one gets  $E_r = T_r^4 + O(\varepsilon)$ ,  $F_r = -\varepsilon \frac{1}{\sigma_s} \frac{\partial}{\partial x} T_r^4 + \varepsilon \frac{4}{3} v T_r^4 + O(\varepsilon^2)$  and  $P_r = \frac{1}{3} T_r^4 + O(\varepsilon)$ . With these asymptotics and definitions, then the equation for  $E_r$  becomes

$$\frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \left( -\frac{1}{\sigma_s} \frac{\partial}{\partial x} E_r + \frac{4}{3} v E_r \right) = \sigma_a (T^4 - E_r) + v \frac{\partial}{\partial x} \left( \frac{1}{3} E_r \right). \quad (6)$$

One gets

$$\frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} (v E_r) + \frac{E_r}{3} \frac{\partial}{\partial x} v = \sigma_a (T^4 - E_r) + \frac{\partial}{\partial x} \left( \frac{1}{\sigma_s} \frac{\partial}{\partial x} E_r \right). \quad (7)$$

Thus the radiation energy  $E_r$  is the solution of a fluid like equation with pressure  $P_r = \frac{1}{3}E_r$ . In (7) the velocity is a given function of time and space  $v = v(t, x)$ :  $\frac{\partial}{\partial x}v \neq 0$  a priori. Therefore  $v$  may be discontinuous. Other approximations of the two-moments model exist in the literature. One commonly used consists in using (7) with another definition of the radiative pressure and with  $\sigma_t = \sigma_a + \sigma_s$  in the right hand side. One gets

$$\frac{\partial}{\partial t}E_r + \frac{\partial}{\partial x}(vE_r) + \frac{E_r}{3}\frac{\partial}{\partial x}v = \sigma_a(T^4 - E_r) + \frac{\partial}{\partial x}\left(\frac{1}{\sigma_a + \sigma_s}\frac{\partial}{\partial x}E_r\right). \quad (8)$$

It means that  $\sigma_a$  is no more considered as a very small quantity with respect to  $\sigma_s$ . We will use this model in the numerical section. The interest of such models is clear. Paraphrasing [26]: The radiation temperature  $T_r$  can be quite different from the material temperature. The spectrum of the radiation field can be non-Planckian.

The model (8) has the same limitation than equilibrium diffusion approximation like (5) in transparent materials, that is  $\sigma_a$  and  $\sigma_s$  small. However a new intermediate regime appears. Let us assume that  $\sigma_a$  is small and  $\sigma_s$  large. For example  $\sigma_a = O(\varepsilon)$  and  $\sigma_s = O(\varepsilon^{-1})$ . In this case the equation becomes

$$\frac{\partial}{\partial t}E_r^{\frac{3}{4}} + \frac{\partial}{\partial x}vE_r^{\frac{3}{4}} = 0 \iff \frac{\partial}{\partial t}E_r + \frac{\partial}{\partial x}vE_r + \frac{E_r}{3}\frac{\partial}{\partial x}v = 0. \quad (9)$$

The equation is written in divergence form on the left and in non divergence form on the right. The  $\frac{E_r}{3}$  term in the non divergence form is the thermodynamic pressure of radiation [26]. The divergence form of the equation is a conservation equation for  $E_r^{\frac{3}{4}}$  which is the number of photons at equilibrium. Assuming the matter encounters a shock, the velocity  $v$  is discontinuous. Therefore a continuous radiation energy  $E_r$  is required to give a non ambiguous value to the non conservative product  $\frac{E_r}{3}\frac{\partial}{\partial x}v$ . Unfortunately the hypothesis  $\sigma_s = O(\varepsilon^{-1})$  implies a vanishing diffusion and possible discontinuity of  $E_r$  at the limit. So the non conservative product  $\frac{E_r}{3}\frac{\partial}{\partial x}v$  can be ambiguous. On the other hand the divergence form of the equation is non ambiguously defined for a discontinuous velocity. The Rankine Hugoniot relations for the divergence formulation of (9) are

$$-D[T_r^3] + [vT_r^3] = 0, \quad E_r = T_r^4.$$

$D$  is the shock velocity. Assuming radiation is at equilibrium the radiation entropy is  $S_r = T_r^3$  and is also equal to the number of photons. Therefore the physical meaning of this RH relation is that the integrated number of photons  $\int T_r^3$  is conserved through the discontinuity. This is compatible with what is proposed at the numerical in [26], where no numerical viscosity is used for radiation. The situation is very closed to ion-electron fluid models, for which



the correct Rankine Hugoniot relation is obtained with the entropy of electron, see [36] and [8]. There is no entropy deposit on electrons or photons at shocks. This principle is compatible with the conservation of total energy, the total energy being the sum of the radiation energy and the fluid energy. For contact discontinuities  $D = v_L = v_R$  the Rankine Hugoniot vanishes.

### 2.5 Splitting method for the non-equilibrium diffusion limit

We begin with a preliminary remark about the discretization in time of the non equilibrium diffusion limit equation (7). The idea is to solve (7) with a splitting technique. This is we solve separately during the same time step

$$\begin{aligned} \partial_t S_r + \partial_x v S_r &= 0, \quad S_r = \frac{4}{3} T_r^3 \\ \text{followed by} & \\ \partial_t E_r &= \sigma_a (T^4 - E_r) + \frac{\partial}{\partial x} \left( \frac{1}{3\sigma_t} \frac{\partial}{\partial x} E_r \right), \quad E_r = T_r^4. \end{aligned} \tag{10}$$

The first equation in (10) is, for smooth solutions, algebraically equivalent to  $\partial_t E_r + \partial_x v E_r + \frac{1}{3} T_r^4 \partial_x v = 0$ . Therefore (10) is indeed a convenient way to split (7) into two separate parts. Note that the first equation in (10) is a first order conservation law, that can be easily discretized with a finite volume technique, while the second equation is of parabolic type and is easily discretized with a symmetric positive diffusion matrix. The implicit linear system can be solved with a conjugate gradient algorithm, which reveals very powerful (both in term of accuracy and speed) for this problem.

## 3 Splitted approximations of the $M^1$ model

Our aim is at the description of splitted approximation of the  $M^1$  model, in view of the discretization of (2) such that (7) is by construction an asymptotic limit of the scheme. We also take care of the natural reality condition that says that  $\frac{|F_r|}{E_r} \leq 1$ . This comes from  $\frac{|F_r|}{E_r} = \frac{|\int \int n I(\nu, n) d\nu dn|}{\int \int I(\nu, n) d\nu dn} \leq 1$  for a smooth smooth intensity  $I$ . Straightforward but lengthy calculations show that

**Proposition 1** *Assuming that  $T_r > 0$  then*

$$|b| < 1 \iff \frac{|Q_r|}{S_r} < 1 \iff \frac{|F_r|}{E_r} < 1. \tag{11}$$

From (3) and (4) it is sufficient to prove that :  $-1 < b < 1$  implies  $-1 < \frac{4b}{3+b^2} < 1$ . The range  $-1 < b < 1$  is the physical range in which the physical solution is searched. The inequalities are strict because the map  $(E_r, F_r) \rightarrow (S_r, Q_r)$  is singular at  $b = \pm 1$ . A convenient way to extend the range to  $-1 \leq b \leq 1$  is proposed in section 3.2.

### 3.1 Splitting method for the $M^1$ model

We generalize the splitting method of (10). The first part of the splitting is written as

$$\begin{cases} \frac{\partial}{\partial t} S_r + \frac{\partial}{\partial x} v S_r = 0, \\ \frac{\partial}{\partial t} Q_r + \frac{\partial}{\partial x} v Q_r = 0. \end{cases} \quad (12)$$

The first equation is natural, in the sense that it is the first equation of (10). The second equation is arbitrary : notice however that a consequence will be the compatibility with the realizability condition (11).

**Proposition 2** *Assume the velocity field  $v$  is smooth. Assume that  $\frac{|Q_r|}{S_r} < 1$  for all  $x$  at  $t = 0$ . Then  $\frac{|Q_r|}{S_r} < 1$  for all  $x$  and for all  $t > 0$ .*

A algebraic consequence of (12) is  $\frac{\partial}{\partial t} \frac{Q_r}{S_r} + v \frac{\partial}{\partial x} \frac{Q_r}{S_r} = 0$ . So the parameter  $b = -\frac{Q_r}{S_r}$  is just transported by the equation. It ends the proof.

**Proposition 3** *The system (12) is algebraically equivalent to*

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} v E_r + \frac{1}{3} E_r \frac{\partial}{\partial x} v = 0, \\ \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} v F_r + \frac{1}{3} F_r \frac{\partial}{\partial x} v = 0. \end{cases} \quad (13)$$

A proof is as follows. The system (12) is also  $\frac{\partial}{\partial t} S_r + v \frac{\partial}{\partial x} S_r + S_r \frac{\partial}{\partial x} v = 0$  and  $\frac{\partial}{\partial t} Q_r + v \frac{\partial}{\partial x} Q_r + Q_r \frac{\partial}{\partial x} v = 0$ . Since one has  $dE_r = \frac{\partial E_r}{\partial S_r} dS_r + \frac{\partial E_r}{\partial Q_r} dQ_r$  then

$$\frac{\partial}{\partial t} E_r + v \frac{\partial}{\partial x} E_r + \left( \frac{\partial E_r}{\partial S_r} S_r + \frac{\partial E_r}{\partial Q_r} Q_r \right) \frac{\partial}{\partial x} v = 0. \quad (14)$$

From its definition one has  $E_r = S_r^{\frac{4}{3}} \varphi(\frac{Q_r}{S_r})$ . Then

$$\frac{\partial E_r}{\partial S_r} S_r + \frac{\partial E_r}{\partial Q_r} Q_r = \left( \frac{4}{3} S_r^{\frac{1}{3}} \times \varphi\left(\frac{Q_r}{S_r}\right) - S_r^{\frac{4}{3}} \frac{Q_r}{S_r^2} \varphi'\left(\frac{Q_r}{S_r}\right) \right) S_r$$

$$+ \left( S_r^{\frac{4}{3}} \frac{1}{S_r} \varphi' \left( \frac{Q_r}{S_r} \right) \right) Q_r = \frac{4}{3} E_r.$$

Using this in (14) it gives the first equation of (13). Concerning the second equation in (13) we note that  $F_r = S_r^{\frac{4}{3}} \psi \left( \frac{Q_r}{S_r} \right)$ . Therefore a similar algebra ends the proof of the proposition.

Now we subtract (13) to the  $M^1$  model (2). This subtraction allows to reconstruct (2) using the splitting method. The subtraction is valid only for the derivatives in space, and not in time. We get

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \left( \frac{F_r}{\varepsilon} - v E_r \right) - \frac{E_r}{3} \frac{\partial}{\partial x} v = S_E, \\ \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} \left( \frac{P_r}{\varepsilon} - v F_r \right) - \frac{F_r}{3} \frac{\partial}{\partial x} v = S_F, \end{cases} \quad (15)$$

We claim that the non-equilibrium diffusion limit of (15) is (6) but with  $v \equiv 0$ . This property means that the splitting method has really subtract the drift effect due to the Lorentzian compatibility of the model.

**Proposition 4** *Whatever  $v$  is, the non-equilibrium diffusion limit ( $\sigma_a = O(\varepsilon^2)$  and  $\sigma_s = O(1)$ ) of the system (15) is the second equation of (10), that is*

$$\frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \left( -\frac{1}{\sigma_s} \frac{\partial}{\partial x} E_r \right) = \sigma_a (T^4 - E_r). \quad (16)$$

This property means that if we set  $v = 0$  in (15) then we respect the non-equilibrium diffusion limit of the  $M^1$  model, provided we use the splitting method. If  $v \equiv 0$  the system (15) is the standard radiation  $M^1$  model with the matter at rest.

The splitted model is made of (12) followed by (15) where we have set  $v = 0$ . This model is composed of (12) followed by

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \left( \frac{F_r}{\varepsilon} \right) = \frac{\sigma_a}{\varepsilon^2} (T^4 - E_r), \\ \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} \left( \frac{P_r}{\varepsilon} \right) = -\frac{\sigma_a + \sigma_s}{\varepsilon^2} F_r. \end{cases} \quad (17)$$

In the numerical section we will show how to derive accurate and robust schemes for the discretization of (12)-(17).

### 3.2 Elimination of the singularity near $|f| = 1$

The method of moment is singular near  $f = \pm 1$  since in this case  $T_r$  is zero for finite  $E_r$ , see formula (3). It simply means that, far from isotropy,  $T_r$  is not

a temperature and is only a parameter of the generalized Planckian. At the numerical level we prefer to use non singular equation near  $f = \pm 1$ . We think this is important to get accurate numerical results if  $f$  is close to 1, because singular equations are most of time difficult to discretize correctly.

We notice that it is sufficient to redefine  $S_r = E_r^{\frac{3}{4}}$  and  $Q_r = -bE_r^{\frac{3}{4}}$ , because (12) is equivalent to  $(\partial_t + v\partial_x)b = 0$  and  $\partial_t T_r^3 + \partial_x(vT_r^3) = 0$ . This transformation is a smooth map between  $(E_r, F_r)$  and  $(S_r, Q_r) = (E_r^{\frac{3}{4}}, -bE_r^{\frac{3}{4}})$ . From now on the splitted approximation we use is

$$\begin{cases} \frac{\partial}{\partial t} S_r + \frac{\partial}{\partial x} v S_r = 0, & S_r = E_r^{\frac{3}{4}} \\ \frac{\partial}{\partial t} Q_r + \frac{\partial}{\partial x} v Q_r = 0, & Q_r = -bE_r^{\frac{3}{4}}. \end{cases} \quad (18)$$

followed by

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \left( \frac{F_r}{\varepsilon} \right) = \frac{\sigma_a}{\varepsilon^2} (T^4 - E_r), \\ \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} \left( \frac{P_r}{\varepsilon} \right) = -\frac{\sigma_a + \sigma_s}{\varepsilon^2} F_r. \end{cases} \quad (19)$$

### 3.3 Extension for radiation hydrodynamics

Written in non dimensional variables the model we use is

$$\begin{cases} \frac{\partial}{\partial t} (\rho) + \frac{\partial}{\partial x} (\rho v) = 0, \\ \frac{\partial}{\partial t} (\rho v) + \frac{\partial}{\partial x} (\rho v^2 + p + \frac{E_r}{3}) = 0, \\ \frac{\partial}{\partial t} (\rho E + E_r) + \frac{\partial}{\partial x} (\rho E v + p v + \frac{E_r}{3} v + \frac{1}{\varepsilon} F_r) = 0, \\ \frac{1}{\varepsilon} \frac{\partial}{\partial t} E_r + \partial_x (v E_r) + \frac{E_r}{3} \partial_x v + \frac{\partial}{\partial x} F_r = \frac{\sigma_a}{\varepsilon^2} (T^4 - E_r), \\ \frac{1}{\varepsilon} \frac{\partial}{\partial t} F_r + \partial_x (v F_r) + \frac{F_r}{3} \partial_x v + \frac{\partial}{\partial x} P_r = -\frac{\sigma_a + \sigma_s}{\varepsilon^2} F_r, \end{cases} \quad (20)$$

where the source terms are  $O(\frac{1}{\varepsilon^2})$  in opaque materials and are negligible in transparent materials. In view of the previous analysis, we propose to split the radiative solver from the hydrodynamics solver. It means that (20) is solved

using

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho v^2 + p + \frac{E_r}{3}) = 0, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \frac{\partial}{\partial x}(\rho E v + p v + \frac{E_r}{3} v) = 0, \\ \frac{\partial}{\partial t} S_r + \frac{\partial}{\partial x}(u S_r) = 0, & S_r = E_r^{\frac{3}{4}}, \\ \frac{\partial}{\partial t} Q_r + \frac{\partial}{\partial x}(u Q_r) = 0, & Q_r = -b S_r, \end{cases} \quad (21)$$

with  $b$  defined by (3), and followed by

$$\begin{cases} \partial_t E_r + \frac{1}{\varepsilon} \partial_x F_r = \sigma_a \frac{T^4 - E_r}{\varepsilon^2} \\ \partial_t F_r + \frac{1}{\varepsilon} \partial_x P_r = -\sigma_t \frac{F_r}{\varepsilon^2}, \quad \sigma_t = \sigma_a + \sigma_s \\ \rho C_v \partial_t T = \sigma_a \frac{E_r - T^4}{\varepsilon^2}. \end{cases} \quad (22)$$

It is possible to treat directly the radiative moment system. We think our approach is simpler to capture the diffusion limit of the model.

#### 4 Derivation of the numerical schemes

The time step is  $\Delta t > 0$ , the mesh size is  $\Delta x > 0$ , the value of the radiative energy and flux in the cell  $[(j - \frac{1}{2})\Delta x, (j + \frac{1}{2})\Delta x]$  and at time step  $n$  is denoted as  $(E_r^{j,n}, F_r^{j,n})$ . Similarly the radiative entropy and radiative entropy flux are  $(S_r^{j,n}, Q_r^{j,n})$ . The velocity  $u_{j+\frac{1}{2}}^n$  is assumed to be known at each interface and at each time step.

##### 4.1 Discretization of (18)

The discretization is quite natural in the context of finite volume method. Many methods exist. For the sake of completeness we describe the most simple one. We assume that  $u_{j+\frac{1}{2}}^n \geq 0$  for all  $(j, n)$ , so that the up-winding is uniform. With this hypothesis, the scheme is

$$\begin{cases} \frac{S_r^{j,n+1} - S_r^{j,n}}{\Delta t} + \frac{u_{j+\frac{1}{2}}^n S_r^{j,n} - u_{j-\frac{1}{2}}^n S_r^{j-1,n}}{\Delta x} = 0, \\ \frac{Q_r^{j,n+1} - Q_r^{j,n}}{\Delta t} + \frac{u_{j+\frac{1}{2}}^n Q_r^{j,n} - u_{j-\frac{1}{2}}^n Q_r^{j-1,n}}{\Delta x} = 0. \end{cases} \quad (23)$$

**Proposition 5** Assume that  $\frac{|Q_r^{j,n}|}{S_r^{j,n}} \leq 1$  for all  $j$ . Assume the CFL condition  $\left(\max_j u_{j+\frac{1}{2}}^n\right) \Delta t \leq \Delta x$ . Then one has  $\frac{|Q_r^{j,n+1}|}{S_r^{j,n+1}} \leq 1$  for all  $j$ .

Let us use simplified notations for the sake of convenience :  $\alpha = S_r - Q_r$  and  $\beta = S_r + Q_r$ . Then  $\alpha$  and  $\beta$  satisfy the same equation which are positive under CFL.

#### 4.2 Discretization of the radiative step (22)

In this part we explain the main features of the scheme we propose for the discretization of the (22). First we show that a direct linear implicitation of the HLL like solvers leads to a non positive linear implicit problem. Second we modify the HLL solver and introduce a new scheme. We show that the modification fixes the positivity of the scheme. For expository purposes we take  $\sigma_a = \sigma_s = 0$  when dealing with the HLL solver and  $\sigma_a = 0, \sigma_s \geq 0$  for the new scheme. Finally we consider the general case  $\sigma_a \geq 0$  and  $\sigma_s \geq 0$ . The general method is to use the theory of  $M$ -matrices.

Since we describe and analyze implicit schemes, the boundary conditions are important because the boundary conditions modify the coefficients of the linear system. Therefore a bad discretization can pollute the  $M$ -matrix structure.

##### 4.2.1 A basic example of a non positive implicit solver

We begin with a simple basic example which shows that a direct implicitation is *a priori* non positive. Let us freeze the Eddington factor in (22) and take  $\sigma \equiv 0$

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{\partial}{\partial x} \left( \frac{F_r}{\varepsilon} \right) = 0, \\ \frac{\partial}{\partial t} F_r + \frac{\partial}{\partial x} \left( \frac{a^2 E_r}{\varepsilon} \right) = 0. \end{cases} \quad (24)$$

Here the coefficient  $a$  is the square root of the Eddington factor defined by  $a^2 = \frac{P_r}{E_r} \in [\frac{1}{3}, 1]$ . Then we discretize (24) with a standard approximate Godunov method, the so-called HHL scheme which is described in [35] page 298. Since the maximum wave speed is  $\frac{1}{\varepsilon}$  the scheme is

$$\begin{cases} \frac{E_r^{j,n+1} - E_r^{j,n}}{\Delta t} + \frac{1}{\varepsilon} \frac{F_r^{j+\frac{1}{2},n+1} - F_r^{j-\frac{1}{2},n+1}}{\Delta x} = 0, \\ \frac{F_r^{j,n+1} - F_r^{j,n}}{\Delta t} + \frac{1}{\varepsilon} \frac{(a^2 E)^{j+\frac{1}{2},n+1} - (a^2 E)^{j-\frac{1}{2},n+1}}{\Delta x} = 0. \end{cases} \quad (25)$$

The inter-facial flux is given by the first order finite volume approximation

$$\begin{cases} (a^2 E)^{j+\frac{1}{2},n+1} = \frac{(a_j^2 E_r^{j,n+1} + a_{j+1}^2 E_r^{j+1,n+1}) + F_r^{j,n+1} - F_r^{j+1,n+1}}{2}, \\ F_r^{j+\frac{1}{2},n+1} = \frac{F_r^{j,n+1} + F_r^{j+1,n+1} + E_r^{j,n+1} - E_r^{j+1,n+1}}{2}. \end{cases} \quad (26)$$

This is an implicit linear system that we solve with a direct solver in dimension one. A conjugate gradient solver can be used if needed in dimension two and more. Unfortunately this scheme does not necessarily respect the flux limiting property. The scheme may be **non positive**. To show this, we rely on the theory of  $M$  matrices [3]. This theory provide necessary and sufficient conditions such that the solution of a general linear system  $\mathcal{A}x = b$  gives a non negative solution  $x \geq 0$  for all  $b \geq 0$ . Here  $x \geq 0$  means that  $x_i \geq 0$  for all  $i$ . We retain the simple condition on  $\mathcal{A}$

$$\text{if } a_{ij} \leq 0 \ \forall i \neq j, \quad \text{and} \quad \left( \sum_j a_{ij} > 0 \ \forall i \text{ or } \sum_i a_{ij} > 0 \ \forall j \right)$$

then  $\mathcal{A}$  is an  $M$ -matrix with  $x = \mathcal{A}^{-1}b \geq 0$  for all  $b \geq 0$ . The flux limited condition  $|F_r| \leq E_r$  is equivalent to the positivity of  $u = E_r + F_r$  and  $v = E_r - F_r$ . Therefore we check the property of  $M$  matrix for the system (25) with the flux defined by (26) and for the unknown  $x = (u^{j,n+1}, v^{j,n+1})_j$  and the right hand side  $b = (u^{j,n}, v^{j,n})_j$ . For the sake of simplicity assume that  $\sigma \equiv 0$  so that  $M \equiv 1$ . The equation for  $v^{j,n+1}$  is

$$\begin{aligned} \frac{v^{j,n+1} - v^{j,n}}{\Delta t} + \frac{1}{4\varepsilon\Delta x} \left( (1 - a_{j+\frac{1}{2}}^2)u^{j+1,n+1} + (a_{j-\frac{1}{2}}^2 - 1)u^{j-1,n+1} \right. \\ \left. + (-3 - a_{j+\frac{1}{2}}^2)v^{j+1,n+1} + (-1 - a_{j-\frac{1}{2}}^2)v^{j-1,n+1} + 4v^{j,n+1} \right) = 0. \end{aligned}$$

If  $a_{j+\frac{1}{2}} < 1$  the extra-diagonal coefficient in front of  $u^{j+1,n+1}$  has the wrong sign. There is a similar property for the equation that gives  $u^{j,n+1}$ . Therefore the solution of the linear system can be non positive.

It is worthwhile to notice that the explicit scheme is positive. The equation becomes

$$\begin{aligned} \frac{v^{j,n+1} - v^{j,n}}{\Delta t} + \frac{1}{4\varepsilon\Delta x} \left( (1 - a_{j+\frac{1}{2}}^2)v^{j+1,n} + (a_{j-\frac{1}{2}}^2 - 1)v^{j-1,n} \right. \\ \left. + (-3 - a_{j+\frac{1}{2}}^2)v^{j+1,n} + (-1 - a_{j-\frac{1}{2}}^2)v^{j-1,n} + 4v^{j,n} \right) = 0. \end{aligned}$$

One can check that  $(1 - a_{j+\frac{1}{2}}^2)u^{j+1,n} + (a_{j-\frac{1}{2}}^2 - 1)u^{j-1,n} \leq 0$  because  $a_j^2 = a \left( \frac{F_r^{j,n}}{E_R^{j,n}} \right)^2 \geq 2 \left| \frac{F_r^{j,n}}{E_R^{j,n}} \right| - 1$ . Nevertheless the explicit scheme is restricted by the

explicit CFL condition  $\frac{\Delta t}{\varepsilon \Delta x} \leq 1$ . This is why we propose in the following a new linear implicit AP and unconditionally positive scheme.

#### 4.2.2 AP and positive discretization of (22) with $\sigma_a = 0$

The idea is to use a strategy that has been proposed and used in [6,7]. It consists in a linearization of the system combined with a well balanced scheme. Depending on the linearization, it is possible to design different numerical schemes with slightly different theoretical properties. In order to simplify the presentation, we consider (24). We will indicate later on how to deal with the temperature relaxation. Here we use a strategy that has been first proposed and used in [6,7]. It consists in the use of a relaxation method (see for example [16] for principle of relaxation method) to “linearize” the system and a “well-balanced” scheme. This combination gives stability and preservation of the asymptotic limit.

Let us first define the relaxation method. We introduce the intermediate relaxation unknowns  $X = \frac{F_r}{a}$  and  $Y = \frac{P_r}{a}$  and we consider the system

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a X = 0, \\ \frac{\partial}{\partial t} X + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a E_r = -\frac{\sigma}{\varepsilon^2} X + \frac{1}{\lambda} \left( \frac{F_r}{a} - X \right), \end{cases} \quad (27)$$

and

$$\begin{cases} \frac{\partial}{\partial t} Y + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a F_r = 0, \\ \frac{\partial}{\partial t} F_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a Y = -\frac{\sigma}{\varepsilon^2} F_r + \frac{1}{\lambda} \left( \frac{P_r}{a} - Y \right). \end{cases} \quad (28)$$

The coefficient  $a$  is the square root of the Eddington factor defined by  $a^2 = \frac{P_r}{E_r}$ . the additional variables are  $X \approx \frac{F_r}{a}$  and  $Y \approx \frac{P_r}{a}$ . More exactly if the relaxation coefficient  $\lambda$  tends to 0, then  $X \rightarrow \frac{F_r}{a}$  and  $Y \rightarrow \frac{P_r}{a}$ , and the asymptotic limit is indeed (19). In the original work [6,7], the relaxation variables were  $X = F_r$  and  $Y = P_r$ . Our choice has better properties.

As it is usually done, we consider the relaxed version of (27). This is no more that using a splitting technique to solve (27), by splitting the relaxation source from the rest and let formally  $\lambda$  tends to 0 in the relaxation step. This leads to the following transport-projection algorithm:

-Transport step: let  $a = \sqrt{\chi^n}$   $X^n = \frac{F_r^n}{a}$  and  $Y^n = \frac{P_r^n}{a}$  at the beginning of the



time step. Then we solve the systems

$$\begin{cases} \frac{\partial}{\partial t} E_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a X = 0, \\ \frac{\partial}{\partial t} X + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a E_r = -\frac{\sigma}{\varepsilon^2} X, \end{cases} \quad (29)$$

and

$$\begin{cases} \frac{\partial}{\partial t} Y + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a F_r = 0, \\ \frac{\partial}{\partial t} F_r + \frac{1}{\varepsilon} \frac{\partial}{\partial x} a Y = -\frac{\sigma}{\varepsilon^2} F_r, \end{cases} \quad (30)$$

during one time step. So we know the initial conditions at  $n\Delta t$  and we use (29-30) to compute all quantities at  $(n+1)\Delta t$  : that is  $E_r^{n+1}$ ,  $F_r^{n+1}$ ,  $X^{n+1}$  and  $Y^{n+1}$ .

-Projection step: we drop  $X^{n+1}$  and  $Y^{n+1}$  and reinitialize  $a = \sqrt{\chi^{n+1}}$ ,  $X^{n+1} = \frac{F_r^{n+1}}{a}$  and  $Y^{n+1} = \frac{P_r^{n+1}}{a}$ .

Therefore it remains to explain how we discretize (29-30) and what advantages come from this strategy. One can remark that (29) and (30) are decoupled.

The method we use for (29) or (30), comes from the work [11–13] about well-balanced schemes for hyperbolic systems with source terms. Originally the goal of this technique was to preserve stationary solutions. Essentially it consists in collapsing the source terms at interfaces and to write down the Godunov method. This step needs the computation of the solution of a Riemann problem with a Dirac source term at the interface. This is quite easy to for linear equations. We refer to [12] for the construction of the method, and to [7] for a simple explanation in the linear case. In what follows we only describe the result of the application of the method to our case.

We have decided to present the scheme with natural notations, such that the consistency of the scheme becomes clear. The coefficient  $M_{j+\frac{1}{2}}$  is defined on the interfaces by

$$M_{j+\frac{1}{2}} = \frac{2a_{j+\frac{1}{2}}\varepsilon}{\sigma_{j+\frac{1}{2}}\Delta x + 2a_{j+\frac{1}{2}}\varepsilon} \quad (31)$$

where quantities of the type  $q_{j+\frac{1}{2}}$  are mean value of  $q_j$  and  $q_{j+1}$ . Then we

discretize (29) with

$$\begin{cases} \frac{E_r^{j,n+1} - E_r^{j,n}}{\Delta t} + \frac{1}{\varepsilon} \frac{M_{j+\frac{1}{2}} a_{j+\frac{1}{2}} X^{j+\frac{1}{2},n+1} - M_{j-\frac{1}{2}} a_{j-\frac{1}{2}} X^{j-\frac{1}{2},n+1}}{\Delta x} = 0, \\ \frac{X^{j,n+1} - X^{j,n}}{\Delta t} + \frac{1}{\varepsilon} \frac{M_{j+\frac{1}{2}} a_{j+\frac{1}{2}} E^{j+\frac{1}{2},n+1} - M_{j-\frac{1}{2}} a_{j-\frac{1}{2}} E^{j-\frac{1}{2},n+1}}{\Delta x} \\ = -\frac{1}{2} \left( M_{j+\frac{1}{2}} \frac{a_j}{a_{j+\frac{1}{2}}} \frac{\sigma_{j+\frac{1}{2}}}{\varepsilon^2} + M_{j-\frac{1}{2}} \frac{a_j}{a_{j-\frac{1}{2}}} \frac{\sigma_{j-\frac{1}{2}}}{\varepsilon^2} \right) X_i^{n+1} \\ + \frac{M_{j+\frac{1}{2}} - M_{j-\frac{1}{2}}}{\varepsilon \Delta x} a_i E_i^{n+1}. \end{cases} \quad (32)$$

The initial value of the additional unknown  $X$  is  $X^{j,n} = \frac{F_r^{j,n}}{a_j}$ . The fluxes are

$$\begin{cases} a_{j+\frac{1}{2}} E^{j+\frac{1}{2},n+1} = \frac{a_j E^{j,n+1} + a_{j+1} E^{j+1,n+1}}{2} + \frac{a_j X^{j,n+1} - a_{j+1} X^{j+1,n+1}}{2}, \\ a_{j+\frac{1}{2}} X^{j+\frac{1}{2},n+1} = \frac{a_j X^{j,n+1} + a_{j+1} X^{j+1,n+1}}{2} + \frac{a_j E^{j,n+1} - a_{j+1} E^{j+1,n+1}}{2}. \end{cases} \quad (33)$$

The linear scheme (32-33) is implicit and can easily be solved in 1D. For the system (30), the scheme is exactly the same except that  $E_r$  is replaced by  $Y$  and that  $X$  is replaced by  $F_r$ . Nevertheless the initialization of the  $Y$  variable is  $Y^{j,n} = \frac{F_r^{j,n}}{a_j}$ .

**Proposition 6 The scheme is AP** Assume that  $\sigma = O(1)$  and  $\varepsilon$  is small. Then the diffusion limit of the scheme (25) is

$$\frac{E_r^{j,n+1} - E_r^{j,n}}{\Delta t} - \frac{\frac{1}{3\sigma_{j+\frac{1}{2}}}(E_r^{j+1,n+1} - E_r^{j,n+1}) - \frac{1}{3\sigma_{j-\frac{1}{2}}}(E_r^{j,n+1} - E_r^{j-1,n+1})}{\Delta x^2} = 0. \quad (34)$$

The second equation in (25) implies that  $F_r^{j,n+1}$  is small of order  $\varepsilon$  :  $F_r^{j,n+1} = O(\varepsilon)$ . Since this is true for all  $j$  and  $n$ , then the Eddington factor is close to one third  $a^2 = \frac{1}{3} + O(\varepsilon)$  for all  $j$  and  $n$ . Then

$$M_{j+\frac{1}{2}} = \frac{2\frac{1}{\sqrt{3}}\varepsilon}{\sigma_{j+\frac{1}{2}}\Delta x} + O(\varepsilon^{\frac{3}{2}}). \quad (35)$$

So the flux that matters in (25) is up to  $O(\varepsilon)$

$$\frac{M_{j+\frac{1}{2}} F_r^{j+\frac{1}{2},n+1}}{\varepsilon} = \frac{\frac{1}{\sqrt{3}} \frac{2\frac{1}{\sqrt{3}}\varepsilon}{\sigma_{j+\frac{1}{2}}\Delta x} (E_r^{j+1,n+1} - E_r^{j,n+1})}{2} = \frac{E_r^{j+1,n+1} - E_r^{j,n+1}}{3\sigma_{j+\frac{1}{2}}\Delta x}.$$

It ends the proof.

It is particularly convenient to use the diagonal form of the system. Let  $u = E_r + X$  and  $v = E_r - X$ . The equations can be written as a system in  $x = (u, v)$ .

$$\begin{cases} \frac{u^{j,n+1} - u^{j,n}}{\Delta t} + \frac{M_{j-\frac{1}{2}}}{\varepsilon} \frac{a_j u^{j,n+1} - a_{j-1} u^{j-1,n+1}}{\Delta x} \\ \quad = M_{j-\frac{1}{2}} \frac{a_j}{a_{j-\frac{1}{2}}} \frac{\sigma_{j-\frac{1}{2}}}{2\varepsilon^2} (v^{j,n+1} - u^{j,n+1}), \\ \frac{v^{j,n+1} - v^{j,n}}{\Delta t} - \frac{M_{j+\frac{1}{2}}}{\varepsilon} \frac{a_{j+1} v^{j+1,n+1} - a_j v^{j,n+1}}{\Delta x} \\ \quad = M_{j+\frac{1}{2}} \frac{a_j}{a_{j+\frac{1}{2}}} \frac{\sigma_{j+\frac{1}{2}}}{2\varepsilon^2} (u^{j,n+1} - v^{j,n+1}). \end{cases} \quad (36)$$

In view of (36) the natural discrete boundary conditions are

$$\begin{cases} j = 1 : a_{j-1} u^{j-i,n+1} = \alpha_1 a_j v^{j,n+1} + \alpha_2 \quad \text{with } \alpha_1, \alpha_2 \geq 0, \\ j = N : a_{j+1} v^{j+i,n+1} = \alpha_3 a_j u^{j,n+1} + \alpha_4 \quad \text{with } \alpha_3, \alpha_4 \geq 0. \end{cases} \quad (37)$$

Plugging the boundary conditions (37) inside the scheme (36) the linear system can be written as

$$\mathcal{A}x = b, \quad \mathcal{A} = I - \Delta t M^S - \Delta t M^{BC} \quad (38)$$

where  $M^S$  is the matrix of the scheme, and  $M^{BC}$  is for the boundary conditions. The  $\mathcal{A}$  is not symmetric. The right hand side is  $b = x^n + \Delta t(\alpha_2, 0, \dots, 0, \alpha_4)$ . By construction one has the property  $M_{ii}^S \leq 0$ ,  $M_{ij}^S \geq 0$  for  $i \neq j$  and  $\sum_i M_{ij}^S \leq 0$  and (this is trivial)  $M_{ii}^{BC} \leq 0$ ,  $M_{ij}^{BC} \geq 0$  for  $i \neq j$  and  $\sum_i M_{ij}^{BC} \leq 0$ . The inequality  $\sum_i M_{ij}^S \leq 0$  is an equality inside the domain :  $\sum_i M_{ij}^S = 0$  for  $2 \leq j \leq N-1$ . Therefore the diagonal of  $\mathcal{A}$  is strictly dominant with respect to the columns inside the domain. It is of particular interest to show that diagonal of  $\mathcal{A}$  is also dominant at the boundaries. It is the case for incoming boundary conditions for which  $\alpha_1 = \alpha_3 = 0$ . For transparent conditions one has also  $\alpha_2 = \alpha_4 = 0$ . For general boundary conditions the property is guaranteed if  $0 \leq \alpha_1 \leq 1$  and  $0 \leq \alpha_3 \leq 1$  : we note this property comes from the dissipative nature of the boundary conditions. From now on, we assume that

$$\mathcal{A}_{ij} \leq 0 \text{ for } i \neq j, \text{ and } \sum_i \mathcal{A}_{ij} > 0 \quad \forall j.$$

Therefore the matrix  $\mathcal{A}$  is an  $M$ -matrix. A general reference is [3]. That is  $\mathcal{A}$  is invertible and  $\mathcal{A}^{-1} \geq 0$ .

**Proposition 7 The scheme is positive** Assume that  $E_r^{j,n} \pm F_r^{j,n} \geq 0$  for all  $j$ . Then  $E_r^{j,n+1} \pm F_r^{j,n+1} \geq 0$  for all  $j$ .

This flux limited property is the reason why we prefer to use this relaxation method rather than the direct approach (26). The price to pay is that we need to solve two linear systems with twice the number of unknowns when compared with (26). The linear systems are absolutely identical. We divide the proof of proposition 7 in three steps.

- a) Claim that  $E_r^{j,n+1} \pm X^{j,n+1} \geq 0$**  The proof consists in writing the equation for  $u = E_r + X$  and  $v = E_r - X$ . This set of equations of linear equations can be rewritten as  $\mathcal{A}x = b$  where  $\mathcal{A}$  is the matrix of the system,  $x = (u^{j,n+1}, v^{j,n+1})$  is the unknown of the linear system and  $b = (u^{j,n}, v^{j,n})$  is the right hand side of the linear system. To prove the claim it is sufficient to prove that all component of  $x$  are non negative. First we assume that all component of  $b$  are non negative, which is due to the hypothesis :  $\frac{|X^{j,n}|}{E_r^{j,n}} = \frac{|F_r^{j,n}|}{a_j E_r^{j,n}} \leq 1$ . Second we use the  $M$ -matrix property. It proves the claim  $u^{j,n+1} = E_r^{j,n+1} + X^{j,n+1} \geq 0$  and  $v^{j,n+1} = E_r^{j,n+1} - X^{j,n+1} \geq 0$ .
- b) Claim that  $Y_r^{j,n+1} \pm F_r^{j,n+1} \geq 0$**  We use the same method. Let  $w = Y + F_r$  and  $z = Y - F_r$ . Then  $w^{j,n} = Y^{j,n} + F_r^{j,n} = \frac{P_r^{j,n}}{a_j} + F_r^{j,n} = a_j E_r^{j,n} \left( a_j + \frac{F_r^{j,n}}{E_r^{j,n}} \right)$ . Since  $a_j$  is a function of the non dimensional radiation flux  $\frac{F_r^{j,n}}{E_r^{j,n}}$ , then it is sufficient to study the function  $f \mapsto a = \sqrt{\chi}$ . Since  $\chi - f^2 = \frac{(1 - \sqrt{4 - 3f^2})^2}{3} \geq 0$ , then  $a \geq |f|$  and therefore  $w^{j,n} \geq 0$ . Similarly  $z^{j,n} \geq 0$ . The new values  $x' = (w^{j,n+1}, z^{j,n+1})$  are solutions of the same linear system,  $\mathcal{A}x' = b'$  where  $b' = (w^{j,n}, z^{j,n})$ . Since  $b' \geq 0$  then  $x'$  also. Therefore  $w^{j,n+1}, z^{j,n+1} \geq 0$ .
- c) Claim that  $E_r^{j,n+1} \pm F_r^{j,n+1} \geq 0$ .** Let us define  $r = E - Y + X - F$  and  $s = E - Y - (X - F)$ . Since  $E, X$  and  $Y, F$  are by definition solutions of the same linear system (32) but with a different initialization of course, then the same method can be used to study the non negativity of  $r^{j,n+1}, s^{j,n+1}$ . It is sufficient to check that  $r^{j,n}, s^{j,n}$  are non negative at the beginning of the time step. One has

$$\begin{cases} r^n = E_r^n - aE_r^n + \frac{F_r^n}{a} - F_r^n = (1 - a)(E_r^n + \frac{F_r^n}{a}) \geq 0, \\ s^n = E_r^n - aE_r^n - \frac{F_r^n}{a} + F_r^n = (1 - a)(E_r^n + \frac{F_r^n}{a}) \geq 0. \end{cases}$$

Therefore  $r^{n+1} \geq 0$  and  $s^{n+1} \geq 0$ . At the end of the time step

$$\begin{cases} 2(E_r^{n+1} - F_r^{n+1}) = r^{n+1} + z^{n+1} + v^{n+1} \geq 0, \\ 2(E_r^{n+1} + F_r^{n+1}) = s^{n+1} + w^{n+1} + u^{n+1} \geq 0. \end{cases}$$

### 4.3 Discretization of the full system (22)

Now we describe the scheme for the full system (22). The main tool is, indeed, the discretization described in the previous section (4.2.2). Let us define the quantities  $u, v, w, z$  as  $u = E_r + X$ ,  $v = E_r - X$ ,  $w = Y + F_r$  and  $z = Y - F_r$ . and let also  $\alpha_{j-\frac{1}{2}}(\sigma) = 2M_{j-\frac{1}{2}}(\sigma)\frac{a_j}{a_{j-\frac{1}{2}}}\frac{\sigma_{j-\frac{1}{2}}}{2\varepsilon^2}$  where  $M_{j-\frac{1}{2}}(\sigma)$  is defined in (31).

The transport part of the AP-relaxation scheme described in (4.2.2) can be written now, in its diagonalized form, as

$$\left\{ \begin{aligned} \rho_j C_v \frac{T^{j,n+1} - T^{j,n}}{\Delta t} &= \frac{\alpha_{j-\frac{1}{2}}(\sigma_a) + \alpha_{j+\frac{1}{2}}(\sigma_a)}{2} \left( \frac{u^{j,n+1} + v^{j,n+1}}{2} - (T^{j,n+1})^4 \right) \\ \frac{u^{j,n+1} - u^{j,n}}{\Delta t} + \frac{M_{j-\frac{1}{2}}(\sigma_t)}{\varepsilon} \frac{a_j u^{j,n+1} - a_{j-1} u^{j-1,n+1}}{\Delta x} \\ &= \alpha_{j-\frac{1}{2}}(\sigma_a) ((T^{j,n+1})^4 - u^{j,n+1}) + \left( \alpha_{j-\frac{1}{2}}(\sigma_t) - \alpha_{j-\frac{1}{2}}(\sigma_a) \right) \frac{v^{j,n+1} - u^{j,n+1}}{2}, \\ \frac{v^{j,n+1} - v^{j,n}}{\Delta t} - \frac{M_{j+\frac{1}{2}}(\sigma_t)}{\varepsilon} \frac{a_{j+1} v^{j+1,n+1} - a_j v^{j,n+1}}{\Delta x} \\ &= \alpha_{j+\frac{1}{2}}(\sigma_a) ((T^{j,n+1})^4 - v^{j,n+1}) + \left( \alpha_{j+\frac{1}{2}}(\sigma_t) - \alpha_{j+\frac{1}{2}}(\sigma_a) \right) \frac{u^{j,n+1} - v^{j,n+1}}{2}. \end{aligned} \right. \quad (39)$$

$$\left\{ \begin{aligned} \frac{w^{j,n+1} - w^{j,n}}{\Delta t} + \frac{M_{j-\frac{1}{2}}(\sigma_t)}{\varepsilon} \frac{a_j w^{j,n+1} - a_{j-1} w^{j-1,n+1}}{\Delta x} \\ &= \alpha_{j-\frac{1}{2}}(\sigma_a) (a_j (T^{j,n+1})^4 - w^{j,n+1}) + \left( \alpha_{j-\frac{1}{2}}(\sigma_t) - \alpha_{j-\frac{1}{2}}(\sigma_a) \right) \frac{z^{j,n+1} - w^{j,n+1}}{2}, \\ \frac{z^{j,n+1} - z^{j,n}}{\Delta t} - \frac{M_{j+\frac{1}{2}}(\sigma_t)}{\varepsilon} \frac{a_{j+1} z^{j+1,n+1} - a_j z^{j,n+1}}{\Delta x} \\ &= \alpha_{j+\frac{1}{2}}(\sigma_a) (a_j (T^{j,n+1})^4 - z^{j,n+1}) + \left( \alpha_{j+\frac{1}{2}}(\sigma_t) - \alpha_{j+\frac{1}{2}}(\sigma_a) \right) \frac{w^{j,n+1} - z^{j,n+1}}{2}. \end{aligned} \right. \quad (40)$$

This discretization remains AP, the proof is as in section (4.2.2). For an infinite domain, one can also easily check the conservation of the total energy  $\rho_j C_v T_j + E_{r,j}$ .

During the time step the systems (39) and (40) are, as before, decoupled. But now (39) is a non-linear system. The iterative procedure we use to solve system (39) is based on a fixed point procedure proposed first in [34] for non equilibrium diffusion model. The basic idea of this fixed point procedure is to use  $\Theta = T^4$  as unknown instead of  $T$ . In [34], it has been reported good convergence behavior of this method. For convenience the index  $n$  at the beginning of the time step has been dropped. Our iterative procedure for (39)

is

$$\left\{ \begin{aligned} \rho_j C_v \mu_j^q \frac{\Theta^{j,q+1} - \Theta^j}{\Delta t} &= \frac{\alpha_{j-\frac{1}{2}}(\sigma_a) + \alpha_{j+\frac{1}{2}}(\sigma_a)}{2} \left( \frac{u^{j,q+1} + v^{j,q+1}}{2} - \Theta^{j,q+1} \right) \\ \frac{u^{j,q+1} - u^j}{\Delta t} + \frac{M_{j-\frac{1}{2}}(\sigma_t)}{\varepsilon} \frac{a_j u^{j,q+1} - a_{j-1} u^{j-1,q+1}}{\Delta x} &= \alpha_{j-\frac{1}{2}}(\sigma_a) \left( \Theta^{j,q+1} - \frac{u^{j,q+1} + v^{j,q+1}}{2} \right) + \alpha_{j-\frac{1}{2}}(\sigma_t) \frac{v^{j,q+1} - u^{j,q+1}}{2}, \\ \frac{v^{j,q+1} - v^j}{\Delta t} - \frac{M_{j+\frac{1}{2}}(\sigma_t)}{\varepsilon} \frac{a_{j+1} v^{j+1,q+1} - a_j v^{j,q+1}}{\Delta x} &= \alpha_{j+\frac{1}{2}}(\sigma_a) \left( \Theta^{j,q+1} - \frac{u^{j,q+1} + v^{j,q+1}}{2} \right) + \alpha_{j+\frac{1}{2}}(\sigma_t) \frac{u^{j,q+1} - v^{j,q+1}}{2}. \end{aligned} \right. \quad (41)$$

The coefficient  $\mu_j^q$  is defined by  $\mu_j^q = \frac{T^{j,q} - T^j}{\Theta^{j,q} - \Theta^j}$ . By construction we have to solve at each iteration on  $q$  a linear system

$$\tilde{\mathcal{A}}^q \tilde{x}^q = \tilde{b}^q$$

where  $\tilde{x}^q = (\rho C_v \mu^q \Theta^{q+1}, u^{q+1}, v^{q+1})$  and  $\tilde{b}^q = (\rho C_v \mu^q \Theta, u, v)$ . The matrix  $\tilde{\mathcal{A}}^q$  is an  $M$ -matrix.

Therefore at each iteration  $q$  one has the positivity of the material temperature and the flux limited property. Next we show that the algorithm is strictly contracting. By computing the difference between two successive iterates of (41) one has the relation

$$\tilde{\mathcal{A}}^q \tilde{y}^q = \tilde{c}^q$$

where  $\tilde{y}^q = (\rho C_v \mu^q (\Theta^{q+1} - \theta^q), u^{q+1} - u^q, v^{q+1} - v^q)$  and  $\tilde{c}^q = (\rho C_v (\mu^q - \mu^{q-1}) \Theta^q, 0, 0)$ . The structure of  $\tilde{\mathcal{A}}^q$  is similar to the structure of  $\mathcal{A}$ , that is there exists a matrix  $M^q$  such that  $\tilde{\mathcal{A}}^q = I - \Delta t M^q$ . The important property of  $M^q$  is

$$M_{ij}^q \geq 0 \text{ for } i \neq j, \quad \sum_i M_{ij}^q \leq 0.$$

A consequence<sup>2</sup> is as follows: the matrix  $(\tilde{\mathcal{A}}^q)^{-1} \leq 1$  is contracting in  $l^1$ . Therefore

$$\sum_j \left( \rho C_v \mu_j^q |\Theta_j^{q+1} - \Theta_j^q| + \frac{1}{2} |u_j^{q+1} - u_j^q| + \frac{1}{2} |v_j^{q+1} - v_j^q| \right)$$

<sup>2</sup> Consider  $Ax = b$  where  $A = I - M$  and  $M$  is such that  $m_{ij} \geq 0$  for  $i \neq j$  and  $\sum_i m_{ij} \leq 0$ . The linear system is equivalent to  $(1 - m_{ii})x_i = b_i + \sum_{j \neq i} m_{ij}x_j$ . So  $(1 - m_{ii})|x_i| \leq |b_i| + \sum_{j \neq i} m_{ij}|x_j|$ . We get

$$\sum_i (1 - m_{ii})|x_i| \leq \sum_i |b_i| + \sum_i \sum_{j \neq i} m_{ij}|x_j| = \sum_i |b_i| + \sum_j \left( \sum_{i \neq j} m_{ij} \right) |x_j|.$$

So  $\sum_i |x_i| \leq \sum_i |b_i| + \sum_j (\sum_i m_{ij}) |x_j| \leq \sum_i |b_i|$ . The property is proved.

$$\leq \sum_j \left( \rho C_v \mu_j^{q-1} |\Theta_j^q - \Theta_j^{q-1}| |\beta_j^q| \right) \leq \left( \max_j |\beta_j^q| \right) \times \sum_j \left( \rho C_v \mu_j^{q-1} |\Theta_j^q - \Theta_j^{q-1}| \right).$$

where  $\beta_j^q = \frac{\mu_j^q - \mu_j^{q-1}}{\Theta_j^q - \Theta_j^{q-1}} \times \frac{\Theta_j^q}{\mu_j^{q-1}}$ . Using the definition of the quantities  $\mu_j^q$ , one can

easily check that, at convergence, this quantity scales like  $\beta_j^q \approx \frac{1}{4} \frac{3(T_j^q)^3 + 2(T_j^q)^2 T_{j,n} + T_j^q T_{j,n}^2}{(T_j^q)^3 + (T_j^q)^2 T_{j,n} + T_j^q T_{j,n}^2 + T_{j,n}^3}$ ,

so  $|\beta| < \frac{3}{4}$ . Since this number  $\beta$  is the asymptotic value of contraction ratio of the fixed point procedure, it explains why this algorithm is strictly contracting in any neighborhood of the solution. Moreover all steps of the fixed point procedure are positive (this property was already stressed in the work [34]). In our context it is also a consequence of the  $M$  matrix property.

## 5 Numerical results

All test cases have been performed with the algorithm proposed in this work. The boundary have been chosen so that the results are indepedant of the way the boundary conditions are enforced. In the Su-Olson test case, the problem is symmetric. For the streaming test problem we use free boundary conditions as described in equation (37). For the transfer test case we use infinite opacities near the boundaries. The boundary conditions for the other test cases have been treated with a combination of such boundary conditions.

### 5.1 Su-Olson test case

This test case come from [33]. It is a *basic non-equilibrium radiative transfer consisting of an initially cold, homogeneous, infinite, and isotropically scattering medium with an internal radiation source turned on at time zero* [30].

The opacity is taken constant but with  $C_v = \alpha T^3$ . The source term  $S = 1$  localized in  $|x| \leq 1/2$ . The system of equations to solve reduced to, see [30] for a precise derivation of the model,

$$\begin{cases} \partial_t E_r + \partial_x F_r = (T^4 - E_r) + S \\ \partial_t F_r + \partial_x P_r = -F_r \\ \partial_t T^4 = E_r - T^4. \end{cases} \quad (42)$$

The problem is symmetric. The solution has not the time to reach the exterior boundary. This test problem has been used to benchmark various strategies about the analytical definition  $(E_r, F_r) \mapsto P_r$ , diffusion approximations, and numerical algorithms to solve the problem. In [30] it has been claimed that

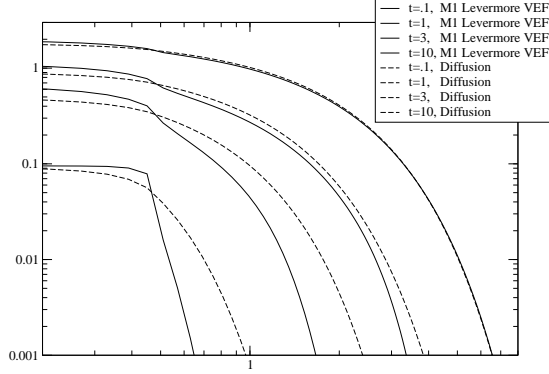


Fig. 1. Variable Eddington factor versus diffusion ( $F_r = -\frac{1}{3}\partial_x E_r$ ) at different times. Radiative energy (log-scaled) versus the optical depth (log-scaled)

the variable Eddington factor gives non smooth and oscillating solutions. On the contrary our results show that it is probably the numerical schemes used in [30] that were unstable for this test case, since our results are stable (i.e. non oscillating) and accurate.

In figure 1 we plot the result given by scheme for the solution of (42) versus the solution of the diffusion approximation of the same model, that is

$$\partial_t E_r - \frac{1}{3}\partial_{xx} E_r = (T^4 - E_r) + S.$$

The solution of the diffusion approximation has been computed with a standard parabolic implicit finite difference scheme. For small time, we see a difference between diffusion and variable Eddington factor. For large time, both methods give the same result, as predicted by the theory. In all cases the solution of the variable Eddington model (42) is free of oscillations. It illustrates the stability and robustness of the algorithm proposed in this work. For large time the variable Eddington model admits theoretically the diffusion model as asymptotic limit. At the numerical level this is also true because  $M_{j+\frac{1}{2}} = \frac{2}{2+\Delta x} \approx 1$  in the scheme, see equation (35).

In figure 2 we plot the Eddington factor  $f = \frac{F_r}{E_r}$  at different times. The result is free of oscillations. One can compare with the results of [30] page 629. Moreover  $-1 \leq f \leq 1$  as stated in proposition 7.



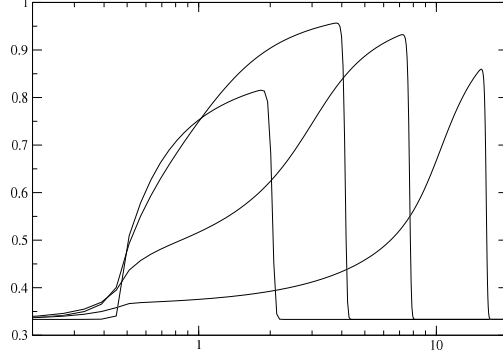


Fig. 2. Eddington factor  $\chi(f)$  where  $f = \frac{F_r}{E_r}$  versus the optical depth (log-scaled). Same times as in figure 1.

cells	100	200	400	800	1600
error in $L^1$ norm	0.3	0.21	0.15	0.109	0.077

Table 1

Error in  $L^1$  norm for the streaming test case. The order is approximatively 0.5.

## 5.2 Streaming

This numerical test case illustrates the ability of the scheme to be accurate and stable for streaming. We solve (2) with  $\sigma_a = \sigma_s = v = 0$  and  $\varepsilon = 1$ . The initial data is

$$E_1 = F_r = 1 \text{ if } 0.4 < x < 0.6, \quad E_1 = F_r = 0 \text{ otherwise.}$$

We use free boundary conditions, standard for advection problems. Since the velocity is one, the analytical solution at time  $t = 0.2$  is

$$t = 0.2 : E_1 = F_r = 1 \text{ if } 0.6 < x < 0.8, \quad E_1 = F_r = 0 \text{ otherwise.}$$

The error in  $L^1$  norm is given in table 5.2. One sees that the scheme is of order 0.5 in  $L^1$  norm. Since the scheme is implicit the price to pay is the extra numerical diffusion compared with an explicit discretization of the same model. However an implicit scheme is absolutely needed for real applications where the velocity of light is very large. The numerical solution of figure 3 has been computed with the same code as for the Sue-Olson test case, with exactly the same scheme. The time step is  $\Delta t = \frac{1}{2}\Delta x$ .

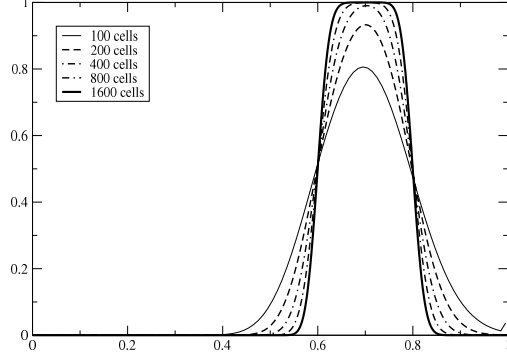


Fig. 3. Radiation energy versus  $x$

### 5.3 Stationary case

We solve  $\partial_t E_r + \frac{1}{\varepsilon} \partial_x F_r = \sigma_a(x) \frac{T^4 - E_r}{\varepsilon^2}$  and  $\partial_t F_r + \frac{1}{\varepsilon} \partial_x P_r = -\sigma_t(x) \frac{F_r}{\varepsilon^2}$ . The temperature  $T$  is prescribed  $T = 1$  if  $1.5 < x < 1.8$ ,  $T = 0$  otherwise. The opacity  $\sigma_t$  is  $\sigma_t = \bar{\sigma} = 1$  if  $1 < x < 1.5$ ,  $\sigma_t = 0$  otherwise. The opacity  $\sigma_a$  is  $\sigma_a = 1$  if  $1.5 < x < 2$ ,  $\sigma_a = 0$  otherwise. The light's velocity is  $\frac{1}{\varepsilon} = 10^3$ . The boundary condition on the right is a prescribed source term with an infinite opacity. On the left it is a free boundary condition. Moreover the support of the solution does not reach the right boundary.

For this problem it is possible to compute a quasi-analytical solution in the region  $0 < x < 1.5$ . For  $x = 1.5$  one has a boundary condition  $E_r = T_{1.5+} = 1$ . For  $x < 1.5$  the stationary solution of the system is given by  $\frac{1}{\varepsilon} \partial_x F_r = 0$  and  $\frac{1}{\varepsilon} \partial_x P_r = -\sigma_t \frac{F_r}{\varepsilon^2}$ . Therefore  $F_r$  is constant  $F_r \equiv F$ . On the other hand  $P_r(1.5) - P_r(1) = -d \bar{\sigma} \frac{F}{\varepsilon}$ ,  $d = 1.5 - 1$ . Since  $\sigma_a = 0$  for  $x < 1$  the stationary regime is an outgoing regime with  $F_r = -E_r$ ,  $x < 1$ . Therefore  $P_r(1.5) \approx \frac{E_r(1.5)}{3} = \frac{1}{3}$ ,  $P_r(1) = E_r(1)$ ,  $F = -E_r(1)$ . Note that  $P_r(1.5) \approx \frac{E_r(1.5)}{3}$  is a very accurate approximation because  $\varepsilon$  is small. Therefore one solves for  $E_r$  and gets  $E_r(1) = \frac{1}{3+3\bar{\sigma}d\varepsilon^{-1}} E_r(1.5) \approx 0$ . In between  $1 < x < 1.5$  one can solve for  $P_r$  which is affine with respect to  $x$ . Since  $P_r \approx \frac{E_r}{3}$  is a very good approximation, then a quasi-analytical solution is

$$E_r(1) = 0, E_r(1.5) = 1, E_r \text{ is affine in between.} \quad (43)$$

We plot the result of our algorithm in figure 4. One sees a very good agreement between the numerical solution of figure 4 and the quasi-analytical solution

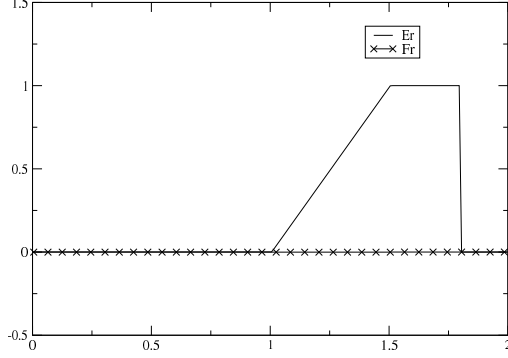


Fig. 4. Numerical solution for the stationary case.  $E_r$  and  $F_r$  versus  $x$ . The stationary source is for  $x \geq 1.5$

(43).

#### 5.4 Transfer

In this test case a transparent material is in between a hot opaque material on the left and a colder (but the temperature is non zero) opaque material on the right. The physical solution is a radiative flow that comes from the left to the right, so that the right material becomes hotter and the left one becomes colder due to the conservativity of energy. We solve (22). The opacity  $\sigma_a$  is a function of the material and of the temperature

$$\sigma_a = \frac{\sigma_0}{T^3} \text{ with } \sigma_0 = 0 \text{ for } 0.2 < x < 1.8 \text{ and } \sigma_0 = 1 \text{ otherwise.}$$

The opacity  $\sigma_t$  does not play a important role in the simulation, and is here only to mimic totally opaque boundary conditions

$$\sigma_t = 0 \text{ for } 0.1 < x < 1.9 \text{ and } \sigma_t = 10^4 \text{ otherwise.}$$

The initial data is  $E_r \equiv 0$ ,  $F_r \equiv 0$  and

$$T = 1 \text{ for } 0.1 < x < 0.2, T = 0 \text{ for } 0.2 < x < 1.8, T = 0.1 \text{ for } 1.8 < x < 1.9.$$

The computation is done with 1000 cells. We use infinite opacities near the boundaries. The result is plotted in figure 5. As noticed in the legend of figure 5, the results are the same with the diffusion approximation. The reason is that for  $t = 3$ , which is quite a large time for this test problem, the radiative

flux is almost zero in the transparent region. This is due to the multiple bounces of the radiative energy on both sides of the transparent region. To see a difference between the variable Eddington factor approach and the diffusion approximation, it is necessary to look at transitory regime around  $x = 1.8$ . This is done in figure 6 where we plot the history of temperature  $T$  (almost equal to the radiative temperature  $T \approx T_r$  because the material is opaque) at  $x = 1.8$  (in the first opaque cell) and at  $x = 1.84$  (in the opaque material). One sees a difference between the results depending on the model (which can be the diffusion approximation, the variable Eddington factor with  $\varepsilon = 0.01$  and the variable Eddington factor with  $\varepsilon = 0.001$ ) and on the point ( $x = 1.8$  or  $x = 1.84$ ). Essentially the variable Eddington factor with  $\varepsilon = 0.001$  is very close to the diffusion approximation. However the variable Eddington factor with  $\varepsilon = 0.01$  is slightly different. Oscillations have appeared that we interpret as the result of multiple radiative bounces between the opaque materials. For longer time, all models have the same history.

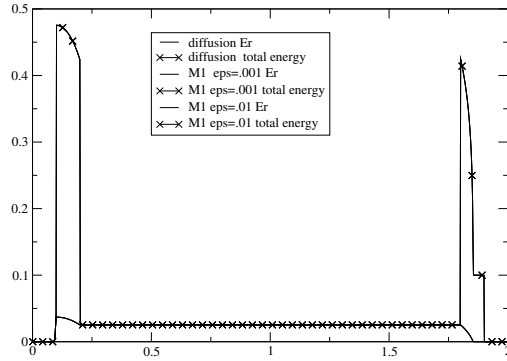


Fig. 5.  $E_r = T_r^4$  and the total energy  $T + E_r = T + T_r^4$  versus  $x$  at time  $t = 3$ , computed with the variable Eddington factor model and  $\varepsilon = 10^{-3}$ . For this test problem the result is the same with the diffusion approximation.

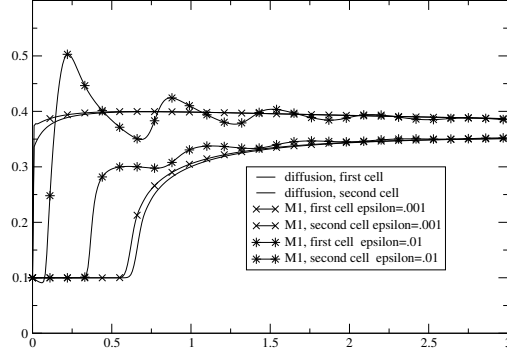


Fig. 6. Heating history in the opaque material : temperature  $T$  versus the time  $t$  for different models at  $x = 1.8$  or  $x = 1.84$ .

### 5.5 Coupling with hydrodynamics

#### 5.5.1 Radiative Riemann problem

If now we set  $\sigma_a = 0$  and  $\sigma_s = +\infty$  then we obtain the simplified hyperbolic set of equations

$$\begin{cases} \frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho v^2 + p + p_r) = 0, & p_r = \frac{E_r}{3}, \\ \frac{\partial}{\partial t}(\rho E + E_r) + \frac{\partial}{\partial x}(\rho E v + p v + p_r v) = 0, & E_r = T_r^4, \\ \frac{\partial}{\partial t} S_r + \frac{\partial}{\partial x}(u S_r) = 0, & S_r = T_r^3. \end{cases} \quad (44)$$

For this system we study a standard Riemann problem. The initial data are  $\rho = 1$ ,  $u = 0$ ,  $p = 1$ ,  $T_r = 1$  for  $x < 0.5$ , and  $\rho = .125$ ,  $u = 0$ ,  $p = 0.1$ ,  $T_r = 0.1$  for  $x > 0.5$ . The solution consists in a rarefaction fan, a contact discontinuity and a shock. Across the shock and the rarefaction fan  $\frac{T_r^3}{\rho}$  is preserved. A lot of hydrodynamic solvers can be used for the numerical resolution. We have used a standard Lagrange+remap solver with an acoustic flux. We use standard free boundary conditions well adapted for pure hyperbolic problems.

The results, which are exactly the same for the diffusion approximation and the variable Eddington factor model, are displayed in figure (7) for the density, velocity and total pressure, and in (8) for  $\frac{T_r^3}{\rho} = \frac{S_r}{\rho}$ , see [36,1].

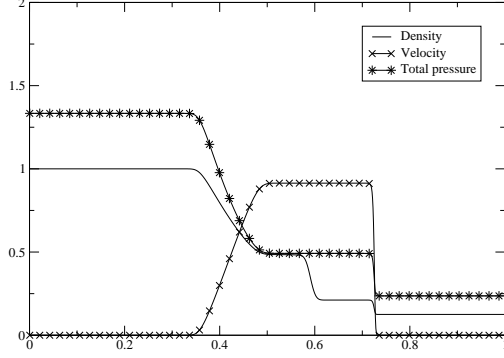


Fig. 7. Density, velocity and total pressure versus the position  $x$  at  $t = 0.1$  with 1000 cells : CFL=0.5.

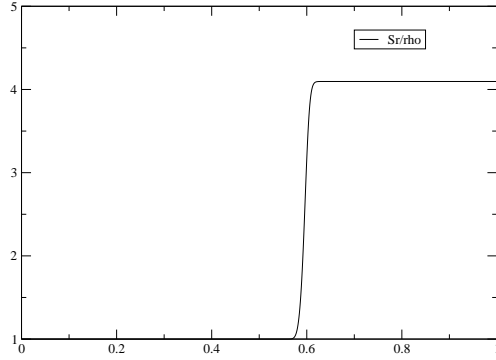


Fig. 8. Radiative entropy  $\frac{T_r^3}{\rho} = \frac{Sr}{\rho}$  versus  $x$  at  $t = 0.1$  with 1000 cells : CFL=0.5. One notices the exact preservation of this quantity across the shock and the rarefaction fan

### 5.5.2 Radiation hydrodynamics

For this test problem we solve the full radiation hydrodynamics (20) (using the splitting scheme (21) and (22)) and with the variable Eddington factor given in (2). The initial data are

$$x < .4 \text{ and } x > 1.6 : \rho = 0.1, u = 0, p_{\text{hydro}} = 0.002, T_r = T = \varepsilon/cv, F_r = 0,$$

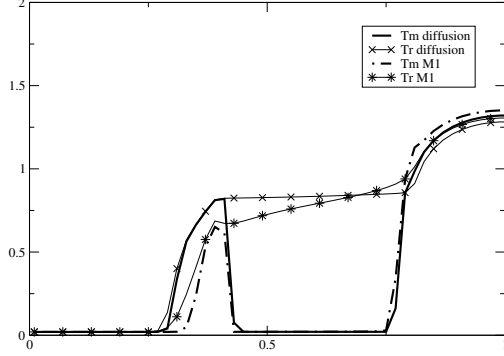


Fig. 9.  $T$  and  $T_r$  versus the position  $x$  for diffusion and the variable Eddington factor, 100 cells,  $T = 0.005$

$$\left. \begin{array}{l} .4 < x < .8 \text{ and} \\ 1.2 < x < 1.6 \end{array} \right\} \rho = 0.01, u = 0, p_{\text{hydro}} = 0.0002, T_r = T = \varepsilon/cv, F_r = 0,$$

$$.8 < x \text{ and } x < 1.2 : \rho = 0.1, u = 0, p_{\text{hydro}} = 0.001, T_r = 1., F_r = 0.$$

The opacities are

$\sigma \equiv 0, \tau = 0.2$  in the left, central and right regions, and  $\tau = 0.2 \times 10^{-7}$  otherwise.

An energy deposit is made at each time step in the central region

$$\varepsilon \leftarrow \varepsilon + 50000 \times \Delta t.$$

Since the central region is hotter, then a flow a radiative energy passes through the transparent intermediate regions, and arrived at the borders of the opaque extreme two regions. The time of obervation is such that the solution has not reach the exterior boundaries.

A zoom around  $x = 0.4$  is given in figure 9. One sees that diffusion is in advance with respect to the variable Eddington factor. An interpretation is that diffusion propagates instantaneously the radiation since the equation is global. Therefore the radiation is over-predicted with the diffusion model. On the other hand the variable Eddington factor is an hyperbolic model with finite speed of propagation. The radiation takes some time to arrive in the region, coming from the central region. Our results are stable.

We give a global view of the radiative quantities,  $E_r$ ,  $F_r$  and  $f = \frac{F_r}{E_r}$  in figure 10. The result is limited since  $|f| \leq 1$  everywhere. In figure (11) we plot the the

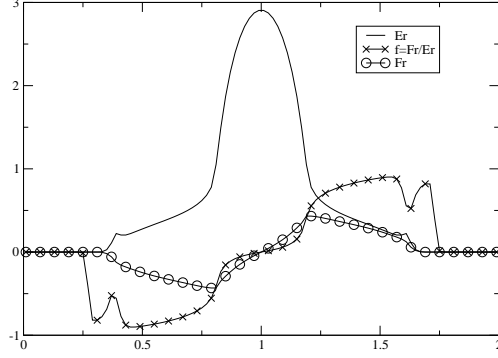


Fig. 10.  $E_r$ ,  $F_r$  and  $f = \frac{F_r}{E_r}$  versus the position  $x$  : 100 cells,  $T = 0.005$

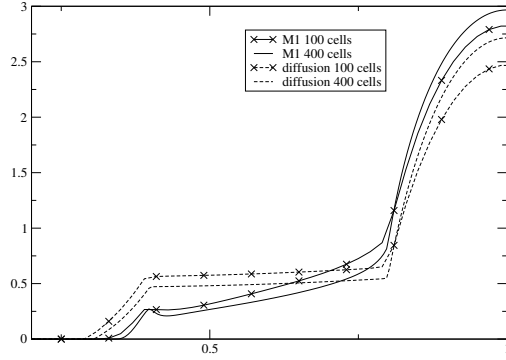


Fig. 11. Convergence study of the radiative energy plotted versus the position  $x$  : 100 and 400 cells.

radiative energy  $E_r$  in the interval  $0 < x < 1$  for 100 and 400 cells, to illustrate the stability and convergence of the diffusion approximation and the variable Eddington factor method. Both schemes are equally stable. The diffusion gives a flat profile due to the infinite propagation speed of this model.

In figure (12) we plot the velocity obtained with the diffusion approximation and with the variable Eddington factor method. The physics of the problem is the following : an energy deposit his made by the radiation at the interface opaque/transparent  $x = 0.4$ . Therefore the pressure increases and pushes the opaque material on the left ( $u < 0$  for  $x < 0.4$ ) and the transparent material



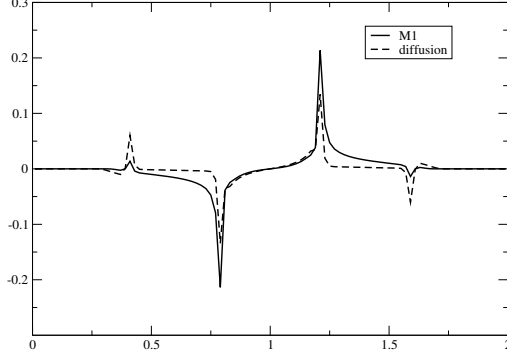


Fig. 12. Material velocity  $u$  versus the position  $x$  at a given time

on the right ( $u > 0$  for  $0.4 < x < 0.4 + \varepsilon$ ). Nevertheless there are some differences. Essentially the velocity profile is flat and equal to 0 for  $0.5 < x < 0.8$  computed with the diffusion approximation. A possible interpretation is the infinite speed of propagation of the diffusion approximation which enforces a constant pressure (therefore no acceleration in this region). On the other hand the propagation of speed is finite with the variable Eddington factor method : it explains the smooth acceleration on the left of the material in the region  $0.5 < x < 0.8$ . However the time scale is the time scale of radiation, which is much smaller than the hydrodynamics time scale. So the impact of these velocity fields is absolutely negligible on the density.

In the final test case, figure 13, we have zeroed the energy deposit. The velocity is  $v = 0$  and  $v = 200$ , everywhere at time  $t = 0$ . Therefore we expect the solution with  $v = 200$  to be equal to the solution with  $v = 0$ , but with a shift. The results show it is indeed the case. For this particular test problem, we found necessary to advect the opacity with an *exact* method. If not numerical diffusion smear the opacity profiles at the interfaces opaque/transparent which give bad results. Here the results are good.

## 6 Conclusion and perspectives

We think that the method proposed in this work gives good results in 1D. At least the method is stable and accurate for diffusion limits, the code is robust, and the numerical results are satisfactory for a large variety of test cases. One could also think about coupling the scheme proposed in this work with higher order schemes with a  $\theta$ -method. The interest could be a less dissipative scheme

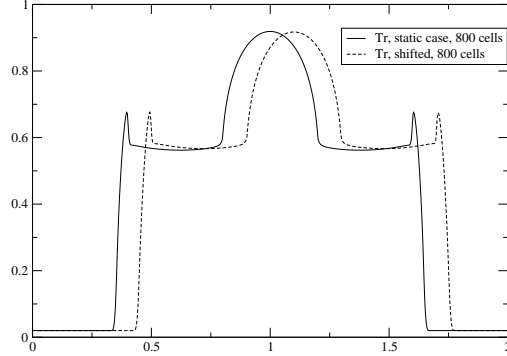


Fig. 13. Shifted versus non shifted calculations.  $E_r$  versus  $x$

in free streaming regions.

The method does not depend on the particular Eddington factor used for radiation, and therefore can be used in other areas, for neutrinos [32], electrons and all models with the same structure where the diffusion limit is somehow the isotropization of a more general system of partial differential equations [4].

Currently we work on the extension for multidimensional problems. The interest of moment models for multidimensional problems could be a better treatment of the anisotropy of radiation.

## References

- [1] C. Buet, B. Despres, "Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics", *Journal of Quantitative Spectroscopy and Radiative Transfer*, Volume 85, Issues 3-4, 385-418 (2004).
- [2] E. Audit, P. Charrier, J.-P. Chieze and B. Dubroca, A radiation-hydrodynamics scheme valid from the transport to the diffusion limit, [lanl.arxiv.org/abs/astro-ph/0206281](http://lanl.arxiv.org/abs/astro-ph/0206281), (2002)
- [3] A. Berman and R. J. Plemmons, *Non Negative Matrices in the Mathematical Sciences*, Classics in Applied Mathematics, SIAM, 1994.
- [4] F. Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws, and well-balanced schemes for sources*, *Frontiers in Mathematics series*, Birkhäuser, (2004).

- [5] T.A. Brunner and J.P. Holloway, One-dimensional Riemann solvers and the maximum entropy closure, *JQRST*, 69:543-566, 2001.
- [6] C. Buet, S. Cordier, "Asymptotic preserving scheme and numerical methods for radiative hydrodynamic models", *Comptes Rendus Mathematique*, Volume 338, Issue 12, 951-956 (2004).
- [7] C. Buet, S. Cordier, "An asymptotic preserving scheme for Hydrodynamics Radiative Transfer Models", submitted to *Num.Math.* (2005).
- [8] F. Coquel and C. Marmignon, Numerical methods for weakly ionized gas. *Astrophys. Space Sci.*, 260, No.1-2, 15-27. 1998.
- [9] W. Wenlong Dai, Paul R. Woodward, Numerical Simulations for Radiation Hydrodynamics: II. Transport Limit , *Journal of Computational Physics*, Volume 157, Issue 1, 1 January 2000, Pages 199-233.
- [10] J.L. Feugeas and B. Dubroca, Entropy moment closure hierarchy for the radiative transfer equation, *C. R. Acad. Sci., Paris, Sér. I, Math.* 329, No.10, 915-920 (1999).
- [11] L. Gosse, Localization effects and measure source terms in numerical schemes for Balance laws, *Math of Comp*, Vol 71, No 238, pp 553-582.
- [12] L. Gosse, G. Toscani, Space localization and well-balanced schemes for discrete kinetic models in diffusive regime. *SIAM J.NUMER Anal.* Vol 41, No 2641-658.
- [13] L. Gosse, A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. *Math. Mod. Meth. Appl. Sci.* 11 (2001) 339-365.
- [14] Jin, S. Levermore, C.D., Numerical Schemes for Hyperbolic Conservation Laws with Stiff Relaxation Terms, *Journal of computational physics* 126, 449-467, Article No.0149, 1996.
- [15] S. Jin, L. Pareschi et G. Toscani, Uniformly accurate diffusive relaxation schemes for multiscale transport equations, *SIAM J. Num. Anal.*, 38, 913, 2000.
- [16] Jin S., Xin Z., The relaxation schemes for systems of conservation laws in arbitrary space dimensions, *Comm. Pure Appl. Math.*, 48 , p. 235-276., 1995.
- [17] T.Y. Hou and P.G. LeFloch, Why nonconservative schemes converge to wrong solutions: error analysis, *Math. of Comput.* 62 (1994), 497-530.
- [18] C. D. Levermore, Relating Eddington factors to flux limiters, *J. Quant. Spectros. Radiat. Transfer*, Vol 31, No 2, pp. 149-160, 1984.
- [19] C. D. Levermore, Moment closure hierarchies for kinetic theories. *J. Stat. Phys.* 83, No.5-6, 1021-1065 1996.
- [20] C. D. Levermore, Moment closure hierarchies for kinetic theories. *J. Stat. Phys.* 83, No.5-6, 1021-1065 1996.

- [21] R.B. Lowrie, J.E. Morel and J.A. Hittinger, The coupling of radiation and hydrodynamics, the astrophysical journal, 521, 432–450 (1999).
- [22] R. B. Lowrie and J. E. Morel, Issues with high-resolution Godunov methods for radiation hydrodynamics. Journal of Quantitative Spectroscopy and Radiative Transfer, Volume 69, Issue 4, 15 May 2001, Pages 475-489.
- [23] Lowrie RB, Morel JE. Discontinuous Galerkin for hyperbolic systems with sti relaxation. In Discontinuous Galerkin Methods: Theory, Computation and Applications, Cockburn B, Karniadakis GE, Shu CW (eds). Springer: Berlin, 2000; 385.
- [24] R. B. Lowrie and J. E. Morel, Methods for hyperbolic systems with stiff relaxation, International Journal for Numerical Methods in Fluids, Volume 40, 2002, Pages 413-423.
- [25] D. Mihalas, Computational Methods for Astrophysical Fluid Flow, Saas-Fee Advanced Course, Lectures notes 1997, Springer.
- [26] D. Mihalas and B.W. Mihalas, Foundations of radiation hydrodynamics, Oxford press university (1984).
- [27] G. N. Minerbo, Maximum entropy Eddington factors, J. Quant. Spectros. Radiat. Transfer, Vol 20 pp541-545, 1978.
- [28] I. Müller and T. Ruggeri, Rational extended thermodynamics. 2nd ed. Springer Tracts in Natural Philosophy. 37. New York, NY: Springer.
- [29] G.Naldi, L. Pareschi, Numerical Schemes for Hyperbolic Systems of Conservation Law with Stiff Diffusive Relaxation, SIAM J. NUM. ANAL., Vol. 37, N.4 (2000), pp. 1246-1270.
- [30] G. L. Olson, L. H. Auer and M. L. Hall, Diffusion,  $P^1$ , and other approximate forms of radiation transport. Los Alamos report LA-UR-99-471.
- [31] Y. Saillard, Hydrodynamique de l'implosion d'une cible FCI . Comptes Rendus de l'Académie des Sciences - Series IV - Physics, Volume 1, Issue 6, August 2000, Pages 705-718.
- [32] J.M. Smit, J. Cernohorsky, C.P. Dullemond, hyperbolicity and critical points in two-moments approximate radiative transfer, Astron. Astrophys, 325, 203-211, (1997).
- [33] B. Su and G. L. Olson, Ann Nucl Energy, 24, 1035-55, (1997).
- [34] A.I. Shestakov, J.A. Greenough, L.H. Howell, Solving the radiation diffusion and energy balance equations using pseudo-transient continuation, Journal of Quantitative Spectroscopy and Radiative Transfer 90 (2005).
- [35] E. F. Toro, Riemann solvers and numerical methods for fluid dynamics apractical introduction, Springer1997.
- [36] Ya B. Zel'dovich, Yuri P. Raizer, Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena, Dover Publications, (2002).