Journal of Computational Physics **179,** 43–67 (2002)
doi:10.1006/jcph.2002.7043

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].

0021-9991/02 $35.00

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^∞ of f or εf 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

$$\frac{\partial f}{\partial t} = \frac{1}{\sqrt{\varepsilon}} \frac{\partial}{\partial \varepsilon} \int_0^{\varepsilon_0} f f' \left(\frac{\partial}{\partial \varepsilon} \ln f - \frac{\partial}{\partial \varepsilon'} \ln f' \right) k(\varepsilon, \varepsilon')\, d\varepsilon', \tag{1.1}$$

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,

$$\int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon}\, d\varepsilon = -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} f f' \left(\frac{\partial}{\partial \varepsilon} \phi - \frac{\partial}{\partial \varepsilon'} \phi' \right) \left(\frac{\partial}{\partial \varepsilon} \ln f - \frac{\partial}{\partial \varepsilon'} \ln f' \right) k\, d\varepsilon'\, d\varepsilon, \tag{1.2}$$

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))

$$\rho = \int_0^{\varepsilon_0} f(\varepsilon) \sqrt{\varepsilon}\, d\varepsilon, \quad \rho E = \int_0^{\varepsilon_0} f(\varepsilon) \varepsilon^{3/2}\, d\varepsilon. \tag{1.3}$$

The mathematical (or negative) entropy H, defined by

$$H = \int_0^{\varepsilon_0} f(\varepsilon) \ln(f(\varepsilon))\, d\varepsilon, \tag{1.4}$$

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

$$\int_0^{\varepsilon_0} \frac{\partial f}{\partial t} \phi \sqrt{\varepsilon}\, d\varepsilon = -\frac{1}{2} \int_0^{\varepsilon_0} \int_0^{\varepsilon_0} \left(\frac{\partial}{\partial \varepsilon} \phi - \frac{\partial}{\partial \varepsilon'} \phi' \right) \left(f' \frac{\partial}{\partial \varepsilon} f - f \frac{\partial}{\partial \varepsilon'} f' \right) k\, d\varepsilon'\, d\varepsilon. \tag{1.5}$$

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 $(\varepsilon_i)_{i=1\ldots N} = (i-1)\Delta\varepsilon$ such that $\varepsilon_N = \varepsilon_0$. The ε-derivatives are approximated according to the simplest choice of finite difference operator; namely, we define for any discretized function $(\phi_i)_{i=1\dots N}$ $\mathrm{D}\phi_i = (\phi_{i+1} - \phi_i)/\Delta\varepsilon$, $i = 1\dots N-1$. 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 ϕ. By using the trapezoidal quadrature formula with respect to the measure $\sqrt{\varepsilon}\,d\varepsilon$, we approximate it by

$$\int_0^{\varepsilon_0}\phi(\varepsilon)\sqrt{\varepsilon}\,d\varepsilon = \sum_{i=1}^{N-1}\int_{\varepsilon_i}^{\varepsilon_{i+1}}\phi(\varepsilon)\sqrt{\varepsilon}\,d\varepsilon \simeq \sum_{i=1}^{N-1}\frac{1}{2}(\phi_i + \phi_{i+1})v_{i+1/2}\Delta\varepsilon \overset{def}{=} \sum_{i=1}^{N} c_i\phi_i, \tag{2.1}$$

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),

$$(\text{r.h.s.}) = -\frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\int_{\varepsilon_i}^{\varepsilon_{i+1}}\int_{\varepsilon_j}^{\varepsilon_{j+1}} ff'\left(\frac{\partial}{\partial\varepsilon}\phi - \frac{\partial}{\partial\varepsilon'}\phi'\right)\left(\frac{\partial}{\partial\varepsilon}\ln f - \frac{\partial}{\partial\varepsilon'}\ln f'\right)k\,d\varepsilon'\,d\varepsilon. \tag{2.2}$$

Using for each integral in (2.2) a midpoint quadrature formula, we approximate (2.2) by

$$-\frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} g_{i,j}k_{i,j}\Delta\varepsilon\Delta\varepsilon(\mathrm{D}\phi_i - \mathrm{D}\phi_j)(\mathrm{D}(\ln f)_i - \mathrm{D}(\ln f)_j), \tag{2.3}$$

with $k_{i,j} = k(\varepsilon_{i+1/2}^{3/2}, \varepsilon_{j+1/2}^{3/2})$ 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 $g_{i,j}$

$$g_{i,j} \overset{def}{=} \frac{f_i\mathrm{D}f_j - f_j\mathrm{D}f_i}{\mathrm{D}(\ln f)_j - \mathrm{D}(\ln f)_i}, \quad \text{if } \mathrm{D}(\ln f)_j \neq \mathrm{D}(\ln f)_i, \tag{2.4}$$

and $g_{i,j} = f_i f_j$ when $\mathrm{D}(\ln f)_j = \mathrm{D}(\ln f)_i$ 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

$$-\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} k_{i,j} \Delta\varepsilon \Delta\varepsilon (\mathrm{D}\phi_i - \mathrm{D}\phi_j)(f_j \mathrm{D} f_i - f_i \mathrm{D} f_j). \tag{2.5}$$

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 $\mathrm{D}\phi_i$ 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

$$\sum_{i=1}^{N} c_i \frac{\partial f_i}{\partial t} \phi_i = -\frac{1}{2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} \Delta\varepsilon \Delta\varepsilon (\mathrm{D}\phi_i - \mathrm{D}\phi_j)(\mathrm{D}(\ln f)_i - \mathrm{D}(\ln f)_j), \tag{2.6}$$

or equivalently, using the definition of $g_{i,j}$,

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

By factorizing the terms ϕ_i in (2.6) as explained in [6], we obtain a system of ordinary differential equations of the form

$$\frac{df_i}{dt} = \mathrm{FP}_i, \quad i = 1 \ldots N, \tag{2.7}$$

with $\mathrm{FP}_1 = p_1/c_1$, $\mathrm{FP}_i = (p_i - p_{i-1})/c_i$, for $i = 2 \ldots N-1$, and $\mathrm{FP}_N = -p_{N-1}/c_{N-1}$, and for all $i = 1 \ldots N-1$,

$$p_i \stackrel{def}{=} \sum_{j=1}^{N-1} g_{i,j} k_{i,j} D_{i,j} \Delta\varepsilon, \tag{2.8}$$

where $D_{i,j}$ stands for $(\mathrm{D}(\ln f)_i - \mathrm{D}(\ln f)_j)$. 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 $i = 2 \ldots N-1$,

$$\frac{df_i}{dt} = \frac{1}{c_i}\left(\sum_{j=1}^{n-1} k_{i,j} f_j f_{i+1} + \sum_{j=1}^{n-1} k_{i-1,j} f_{j+1} f_{i-1} - \left(\sum_{j=1}^{n-1} k_{i,j} f_{j+1} + k_{i-1,j} f_j\right) f_i\right)$$

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*,

$$L_1 \le \frac{3}{\sqrt{2}} \quad \text{and} \quad L_2 \le \frac{3}{\sqrt{2}}.$$

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 ε_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

$$H = H(f) \stackrel{def}{=} \sum_{j=1}^{N} c_j f_j \ln(f_j), \tag{2.9}$$

decays in time. This can be easily checked using the weak discretized formulation (2.6) with test function $\phi_i = \ln(f_i)$,

$$\frac{dH}{dt} = \sum_{i=1}^{N} c_i \frac{df_i}{dt} \ln(f_i) + \sum_{i=1}^{N} c_i \frac{d\ln(f_i)}{dt} f_i = -\frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} g_{i,j} k_{i,j} \Delta\varepsilon\Delta\varepsilon D_{i,j}^2 \le 0,$$

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

$$\frac{dH}{dt} = -\frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} k_{i,j}(f_{i+1}f_j - f_{j+1}f_i)(\ln(f_{i+1}f_j) - \ln(f_{i+1}f_j)). \tag{2.10}$$

We shall prove that $\frac{dH}{dt} = 0$ is equivalent to $f_i = M_i$, where M_i is the discrete Maxwellian

$$M_i = n_0 \exp(\alpha\varepsilon_i), \tag{2.11}$$

where n_0 and α 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 (ρ, E being the data, α, n_0 the unknowns) can be reduced to the following equation for the parameter α

$$E = \frac{\sum_{j=1}^{N} c_j\varepsilon_j \exp(\alpha\varepsilon_j)}{\sum_{j=1}^{N} c_j \exp(\alpha\varepsilon_j)}.$$

It is proved that this defines a unique α, which is negative when ε_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.

$$\sup_i K_i(f) \leq \frac{9\rho(f)}{(\Delta\varepsilon)^2}\sqrt{3T + \frac{\Delta\varepsilon}{2}}. \tag{2.12}$$

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

$$K_i(f) \leq \frac{2}{\Delta\varepsilon c_i}\sum_{j=1}^{N} k_{i,j} f_j \Delta\varepsilon \leq \frac{2\sqrt{\varepsilon_{i+1/2}}}{\Delta\varepsilon c_i}\sum_{j=1}^{N} \varepsilon_{j+1/2} f_j \Delta\varepsilon$$

$$\leq \frac{2\sqrt{\varepsilon_{i+1/2}}}{\Delta\varepsilon c_i}\sum_{j=1}^{N} \sqrt{\varepsilon_{j+1/2}} f_j c_j \frac{\Delta\varepsilon\sqrt{\varepsilon_{j+1/2}}}{c_j} \leq \frac{6\sqrt{2}\rho(f)}{(\Delta\varepsilon)^2}\sqrt{3T + \frac{\Delta\varepsilon}{2}}.$$

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

$$\tau = \frac{(\Delta\varepsilon)^2}{9\rho(f)\sqrt{E(f) + \frac{\Delta\varepsilon}{2}}}. \tag{2.13}$$

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,

$$f_i^{t+\Delta t} = f_i^t + \Delta t FP_i^t \geq f_i^t(1 - \Delta t/\tau) > 0,$$

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*

$$\forall i, \quad \lim_{t\to\infty} f_i = M_i.$$

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_∞.

We have shown that $H_\infty = 0$ necessarily by contradiction. Else, there exists an increasing sequence t_k such that $t_k \to \infty$ when $k \to \infty$ and

$$f_i(t_k) \to f_i^\infty, \quad \frac{dH(f \mid M)(t_k)}{dt} \to 0;$$

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 $t \to \infty$ (not only for a sequence of increasing time). Then, we use the Czizar–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 $f_i \to M_i$. ∎

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

$$f_1(t) = f_1^0 + F(t)/c_1, \quad f_2(t) = f_2^0 - F(t)/c_2,$$
$$f_3(t) = f_3^0 - F(t)/c_3, \quad f_4(t) = f_4^0 + F(t)/c_4, \tag{3.3}$$

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

$$A = \frac{f_1^0}{c_4} + \frac{f_4^0}{c_1} + \frac{f_3^0}{c_2} + \frac{f_2^0}{c_3}, \quad B = \left(\frac{1}{c_2c_3} - \frac{1}{c_1c_4}\right)\left(f_2^0 f_3^0 - f_1^0 f_4^0\right)$$
$$\Delta = A^2 - 4B, \quad D = \frac{2\left(f_2^0 f_3^0 - f_1^0 f_4^0\right)}{A + \sqrt{\Delta}}, \quad \tilde{D} = \left(\frac{1}{c_2c_3} - \frac{1}{c_1c_4}\right)D/\Delta.$$

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,

$$f^{n+1} = f^n + \Delta t FP(f^n).$$

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 τ is defined by (2.13) and depends only on ρE, ρ, 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 $\Delta f = \Delta t FP(f^n)_i = \Delta t FP_i^n = \Delta t \frac{1}{c_i}(p_i^n - p_{i-1}^n)$. Adding these inequalities, and using the conservation of mass and definition of the discretized entropy, we obtain

$$H^{n+1} \leq H^n + \Delta t \sum_i FP_i^n \ln\left(f_i^n\right) + (\Delta t)^2 \sum_i \left(FP_i^n\right)^2 / f_i^n.$$

We take

$$\Delta t = \min\left(\tau, \frac{-\sum_i FP_i^n \ln\left(f_i^n\right)}{\sum_i \left(FP_i^n\right)^2 / f_i^n}\right).$$

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 Θ_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

$$H^{n+1} \leq \frac{1}{\tilde{N}} \sum_{p=1}^{\tilde{N}} \sum_i c_i f_i^p \ln\left(f_i^p\right).$$

For any fixed p, the generalized Broadwell systems involved in Θ_p are distinct. Thus, the entropy decays provided that it decays for any system in Θ_p where C is multiplied by $\tilde{N}$. It remains to compute the time step Δt 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

$$f_i(t) = g_i - t\frac{C_{ij}}{c_i}Q, \quad f_{j+1}(t) = g_{j+1} - t\frac{C_{ij}}{c_{j+1}}Q,$$
$$f_{i+1}(t) = g_{i+1} + t\frac{C_{ij}}{c_{i+1}}Q, \quad f_j(t) = g_j + t\frac{C_{ij}}{c_i}Q,$$

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 $t \le \tau_1 = \frac{\sqrt{2}\Delta\varepsilon^2}{3C_N\varepsilon_0 \sup_{k=i,i+1,j,j+1}(\varepsilon_{k+1/2}g_k)}$.

The numerical entropy associated with this scheme is

$$H(t) = c_i f_i(t)\log(f_i(t)) + c_{i+1}f_{i+1}(t)\log(f_{i+1}(t))$$
$$+ c_j f_j(t)\log(f_j(t)) + c_{j+1}f_{j+1}(t)\log(f_{j+1}(t)).$$

We want to choose t such that $H(t) \le H(0)$. 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 $H'(0) \le 0$. We exclude the case $H'(0) = 0$, which corresponds to $Q = 0$ and for which indeed for all time $H(t) \le H(0)$, so that we assume $Q \ne 0$. $H(t)$ is a C^1 function of the time, and decreasing in the neighborhood of the origin. By defining τ as the first time for which $H'(\tau) = 0$, for all $t \in [0, \tau]$ we will have $H(t) \le H(0)$. Let us now find an upper bound for τ. First we must have $\tau \le \tau_1$. Since we have supposed $Q \ne 0$, one can easily verify that $H'(t) = 0$ reads

$$C^2Q\left(\frac{1}{c_{i+1}c_j} - \frac{1}{c_{j+1}c_i}\right)t^2 + C\left(\frac{g_j}{c_{i+1}}\frac{g_{i+1}}{c_j} - \frac{g_i}{c_{j+1}}\frac{g_{j+1}}{c_i}\right)t - 1 = 0.$$

τ 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 τ is given by $\tau = (-B+\sqrt{B^2+4A})/2A = 2/(B+\sqrt{B^2+4A})$. Thus an upper bound for τ 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

$$
\begin{aligned}
|B| &\leq \tilde{N}\left(\max\left|\frac{Cg_{i+1}}{c_j}+\frac{Cg_j}{c_{i+1}}\right|, \left|\frac{Cg_{j+1}}{c_i}+\frac{Cg_i}{c_{j+1}}\right|\right) \\
&\leq 2\tilde{N} \sup_{k=i,i+1,j,j+1}\left(\varepsilon_{k+1/2}g_k\right) \max_{k=i,i+1,j,j+1}\left|\frac{\sqrt{\varepsilon_{k+1/2}}}{c_k}\right| \\
&\leq \frac{1}{\Delta\varepsilon} 3\sqrt{2}\tilde{N} \sup_{k=i,i+1,j,j+1}\left(\varepsilon_{k+1/2}g_k\right)
\end{aligned}
$$

and for $\sqrt{|A|}$

$$
\begin{aligned}
|A| &\leq \tilde{N}\left|\varepsilon_{i+1/2}f_i\varepsilon_{j+1/2}f_{j+1} - \varepsilon_{j+1/2}f_j\varepsilon_{i+1/2}f_{i+1}\right|\left|\frac{\sqrt{\varepsilon_{i+1/2}\varepsilon_{j+1/2}}}{c_{i+1}c_j} - \frac{\sqrt{\varepsilon_{i+1/2}\varepsilon_{j+1/2}}}{c_{j+1}c_i}\right| \\
&\leq \frac{9}{2}\tilde{N}^2 \sup_{k=i,i+1,j,j+1}\left(\varepsilon_{k+1/2}g_k\right)
\end{aligned}
$$

so that an upper bound for τ_2 is given by

$$
\tau_2 \geq \frac{\Delta\varepsilon^2}{\frac{9}{\sqrt{2}}C_N\varepsilon_0 \sup_{k=i,i+1,j,j+1}\left(\varepsilon_{k+1/2}g_k\right)} = \tau_3.
$$

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

$$
\Delta t \leq \frac{\Delta\varepsilon^2}{\frac{18}{\sqrt{2}}C_N\varepsilon_0 \sup_{i=1}^{N-1}\left(\varepsilon_{i+1/2}\sup(f_i, f_{i+1})\right)} = \Delta t_e, \tag{3.4}
$$

then the scheme is positive and entropic. Let us now analyze the dependence of Δt_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 Δt_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 ρ 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_∞ norm of f or εf 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

$$f^{n+1/2} = f^n + \Delta t FP(f^n).$$

The scheme is defined by

$$f^{n+1} = f^n + \frac{\Delta t}{2}\big(FP(f^n) + FP\big(f^{n+1/2}\big)\big). \tag{3.5}$$

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

$$g_0 = f^n, \quad g_1 = g_0 + \frac{1}{\mu}FP(g_0),$$
$$g_2 = \frac{1}{2}\left(g_0 + g_1 + \frac{1}{\mu}(FP(g_0, g_1) + FP(g_1, g_0))\right), \quad g_3 = g_1 + \frac{1}{\mu}FP(g_1),$$

where μ is a positive parameter and $FP(f, g) + FP(g, f)$ is the polar form associated to the quadratic operator $FP(f)$.

If we set now $x = \mu\Delta t$, (3.5) can be rewritten as

$$f^{n+1} = (1 - x + x^2/2)g_0 + \frac{x}{2}(2 - 3x + x^2)g_1 + (x^2 - x^3)g_2 + \frac{x^3}{2}g_3. \tag{3.6}$$

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 μ such that all the g_i's are positive. Using the analysis made for the first-order explicit scheme it is clear that if μ verifies the CFL condition $\frac{1}{\mu} \leq \tau$, where τ 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 $y \to y\log(y)$, 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^∞ norm of f or εf for which we have no results concerning their evolution and since g_i depends on the g_k's for $k = 1, \ldots, i-1$.

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 ρ 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 ρ. 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 $\rho Ct <= 1$, 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)

$$\begin{aligned}
\partial_t f_a &= \frac{\mu_{ab}^2}{m_a}\frac{1}{\sqrt{\varepsilon_a}}\frac{\partial}{\partial\varepsilon_a}\int_0^{\varepsilon_a} f_a(\varepsilon_a)f_b(\varepsilon_b)\left(\frac{1}{m_a}\frac{\partial}{\partial\varepsilon_a}\ln f_a(\varepsilon_a) - \frac{1}{m_b}\frac{\partial}{\partial\varepsilon_b}\ln f_b(\varepsilon_b)\right)\\
&\quad\times k(\varepsilon_a,\varepsilon_b)\,d\varepsilon_b\\
\partial_t f_b &= \frac{\mu_{ab}^2}{m_b}\frac{1}{\sqrt{\varepsilon}}\frac{\partial}{\partial\varepsilon_b}\int_0^{\varepsilon_a} f_a(\varepsilon_a)f_b(\varepsilon_b)\left(\frac{1}{m_b}\frac{\partial}{\partial\varepsilon_b}\ln f_b(\varepsilon_b) - \frac{1}{m_a}\frac{\partial}{\partial\varepsilon_a}\ln f_a(\varepsilon_a)\right)\\
&\quad\times k(\varepsilon_a,\varepsilon_b)\,d\varepsilon_a,
\end{aligned}$$

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

$$\begin{aligned}
\partial_t f_a &= \frac{\sqrt{m_a}\mu_{ab}^2}{m_b}\frac{1}{\sqrt{E_a}}\frac{\partial}{\partial E_a}\int_0^{E_a} f_a(E_a)f_b(E_b)\left(\frac{\partial}{\partial E_a}\ln f_a(E_a) - \frac{\partial}{\partial E_b}\ln f_b(E_b)\right)\\
&\quad\times k(E_a,E_b)\,dE_b\\
\partial_t f_b &= \frac{\sqrt{m_b}\mu_{ab}^2}{m_b}\frac{1}{\sqrt{E_b}}\frac{\partial}{\partial E_b}\int_0^{E_a} f_a(E_a)f_b(E_b)\left(\frac{\partial}{\partial E_b}\ln f_b(E_b) - \frac{\partial}{\partial E_a}\ln f_a(E_a)\right)\\
&\quad\times k(E_a,E_b)\,dE_a.
\end{aligned}$$

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

$$f^0(\varepsilon) = 0.01\exp(-10[\sqrt{\varepsilon} - 0.3/0.3]^2). \tag{5.1}$$

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).

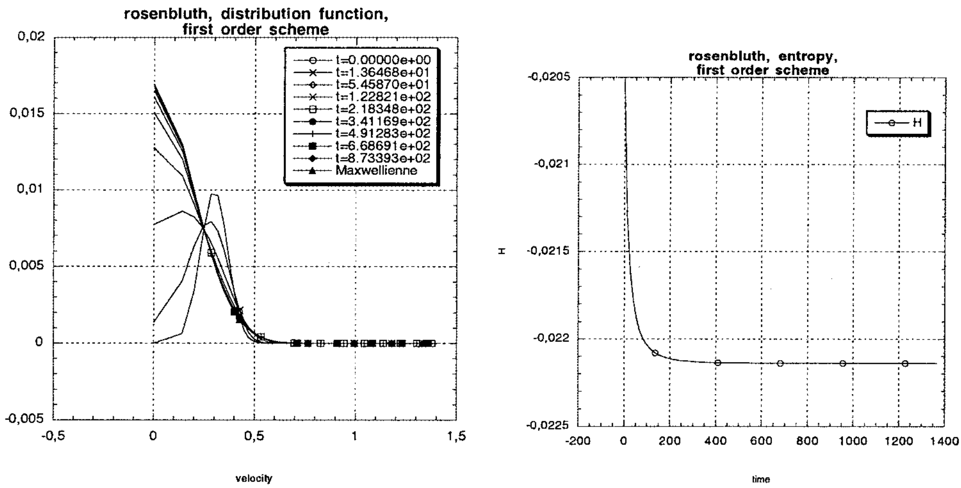

FIG. 1. Rosenbluth test, first-order scheme.

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 δ function. We show the error in L_∞ norm (Fig. 3) and also in L_1 norm (Fig. 4).

L^∞ norm for f and εf. For the two test cases presented here we show also the time evolution of the L_∞ norm for f and εf in Fig. 5. For the two examples the L^∞ norm of εf

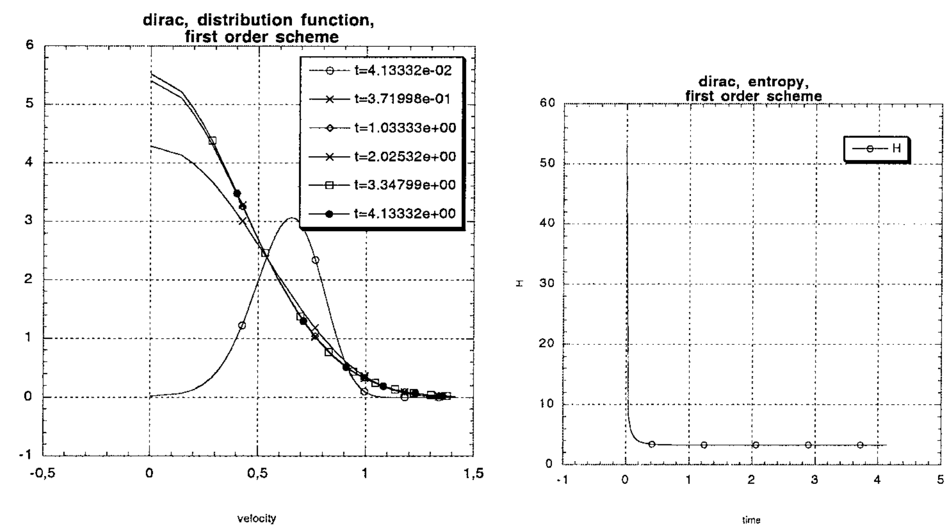

FIG. 2. δ function test, first-order scheme.

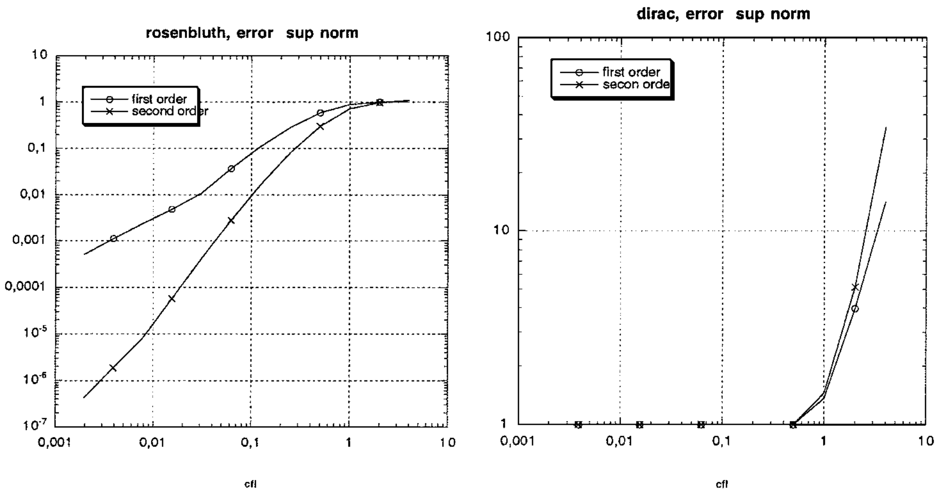

FIG. 3. Error for the L_∞ norm.

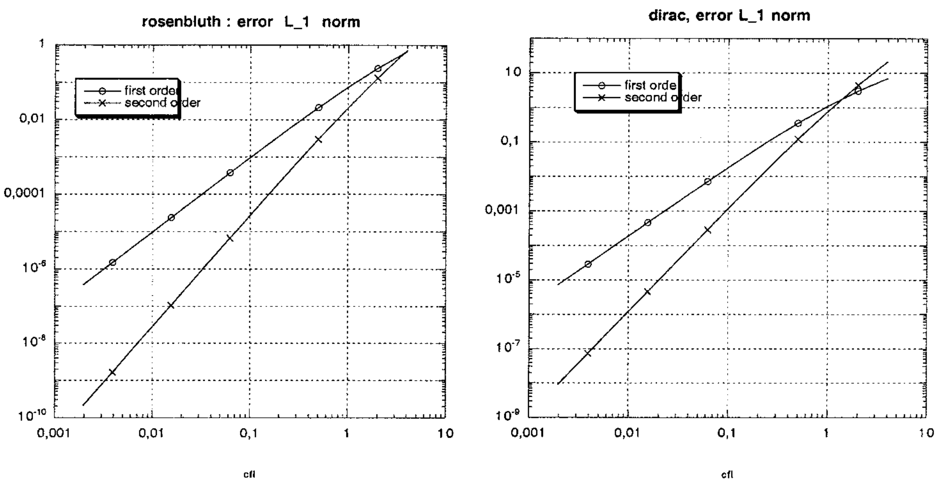

FIG. 4. Error for the L_1 norm.

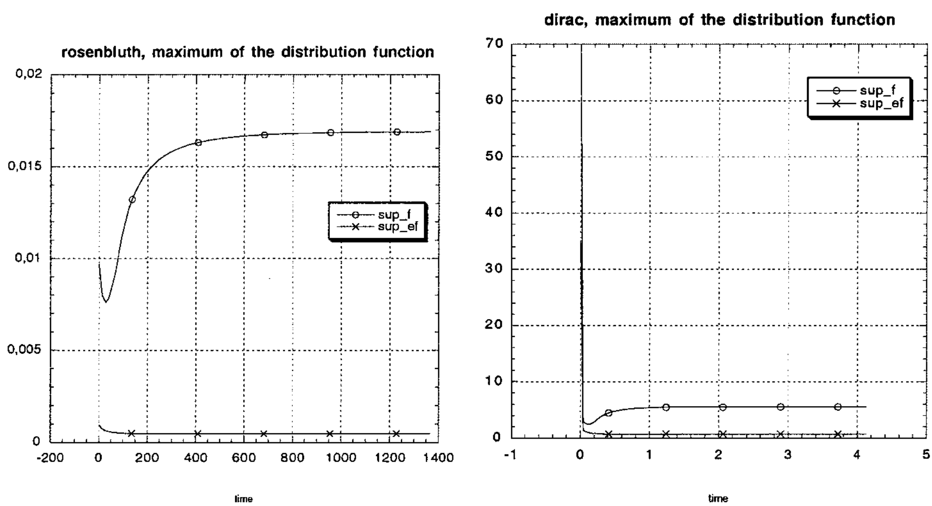

FIG. 5. Time evolution of L^∞ norm for f amd εf.

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^∞ 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

$$d_t H(f_i(t_k)) \to 0, \quad f_i(t_k) \to f_i^\infty.$$

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

$$E(f^0)/\rho(f^0) < E_\infty \stackrel{def}{=} \frac{\sum_{i\in I} c_i \varepsilon_i}{\sum_{i\in I} c_i}.$$

Let us recall that c_i are defined by (2.1) that can be simplified for a uniform grid in

$$c_i = \frac{1}{2}\left(v_{i+1/2} + v_{i-1/2}\right)\Delta\varepsilon;$$

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

$$E(T) \stackrel{def}{=} \frac{\sum_{i\in I} c_i\varepsilon_i \exp(-\varepsilon_i/T)}{\sum_{i\in I} c_i \exp(-\varepsilon_i/T)} = E^0,$$

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

$$\lim_{T\to+\infty} E(T) = \frac{\sum_{i\in I} c_i\varepsilon_i}{\sum_{i\in I} c_i}$$

and

$$\lim_{T\to 0^+} E(T) = \varepsilon_0 = 0.$$

The derivative of $E(T)$ with respect to T is given by

$$\frac{dE}{dT} = \frac{1}{T^2}\frac{\left(\sum_{i\in I} c_i \exp(-\varepsilon_i/T)\right)\left(\sum_{i\in I} c_i\varepsilon_i^2 \exp(-\varepsilon_i/T)\right) - \left(\sum_{i\in I} c_i\varepsilon_i \exp(-\varepsilon_i/T)\right)^2}{\left(\sum_{i\in I} c_i \exp(-\varepsilon_i/T)\right)^2}.$$

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

$$\lim_{T\to-\infty} E(T) = \varepsilon_N$$

and

$$\lim_{T\to 0^-} E(T) = E_\infty$$

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 that the initial distribution is not well represented on the grid and the maximal energy ε_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

$$\Theta = \{(i, i+1, j, j+1), s.t.\ i > j, i = 1, \ldots, N-1\}.$$

We shall construct a partition of Θ into $O(N)$ subsets involving only distinct integers.

First, note that each quadruplet in Θ is determined by the couple (i, j). Let us split Θ 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 ln$[0, k-1]$, in $[k+1, 2k-1[$, or in $[k, k+1]$.

Let us introduce the notation $\Theta = \cup_{i=1,\ldots\tilde{N}} \Theta_i$. 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^∞ 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 repartitions discrètes de vitesses* (Springer-Verlag, New York, 1975).
18. S. Kullback, A lower bound for discrimination information interms 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. 1* **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).