Chapter 4.
Accurate numerical methods for the Boltzmann equation

Francis Filbet[1] and Giovanni Russo[2]

[1] Mathématiques et Applications, Physique Mathématique d'Orléans (MAPMO), CNRS-Université d'Orléans, B.P. 6759, 45067 Orléans, France.
filbet@labomath.univ-orleans.fr

[2] Università di Catania, Viale Andrea Doria 6 95125 Catania, Italia.
russo@dmi.unict.it

Summary. We present accurate methods for the numerical solution of the Boltzmann equation of rarefied gas. The methods are based on a time splitting technique. On the one hand, the transport is solved by a third order accurate (in space) Positive and Flux conservative (PFC) method. On the other hand, the collision step is treated by a Fourier approximation of the collision integral, which guarantees spectral accuracy in velocity, coupled with high order integrators in time preserving stationary states. Several space dependent numerical tests in 2D and 3D illustrate the accuracy and robustness of the methods.

1 Introduction.

In a microscopic description of rarefied neutral gas, the gas particles move by a constant velocity until they undergo binary collisions. In a kinetic picture, the properties of the gas are described by a density function in phase space, $f(t, x, v)$, called the *distribution function*, which gives the number of particles per unit volume in phase space at time t. The distribution function satisfies the Boltzmann equation, an integro-differential equation, which describes the effect of the free flow and binary collisions between the particles. In absence of external forces the time evolution of a single component mono atomic gas, the Boltzmann equation reads to (Cf.[6, 27])

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{k_n} Q(f, f), \quad x, v \in \mathbb{R}^d \tag{1}$$

where $d > 1$ denotes the dimension of the velocity space, the number $k_n > 0$ is called Knudsen number and is proportional to the mean free path between collisions. In the right hand side, $Q(f, f)$ is the so-called *collision operator* given by

$$Q(f,f)(v) = Q^+(f,f) - L[f]f \tag{2}$$

with

$$Q^+(f,f) = \int_{\mathbb{R}^d} \int_{S^{d-1}} B(|v-v_*|,\theta) f(v') f(v'_*)\, d\omega\, dv_*, \tag{3}$$

$$L[f] = \int_{\mathbb{R}^d} \int_{S^{d-1}} B(|v-v_*|,\theta) f(v_*)\, d\omega\, dv_*. \tag{4}$$

In the above integrals, v and v_* are the velocities after the collision of two particles which had the velocities v' and v'_* before the encounter. The deflection angle θ is the angle between $v - v_*$ and $v' - v'_*$.

Here the pre-collision velocities are parameterized by

$$v' = \frac{1}{2}(v + v_* + |v - v_*|\omega), \quad v'_* = \frac{1}{2}(v + v_* - |v - v_*|\omega), \tag{5}$$

where ω is a unit vector of the sphere S^{d-1}.

The quantities $Q^+(f,f)$ and $L[f]f$ are the gain and loss term, respectively. The precise form of the kernel B, which characterizes the details of the binary interactions, depends on the physical properties of the gas. In the case of inverse k-th power forces between particles, the kernel has the form

$$B(|v-v_*|,\theta) = b_\alpha(\theta)|v - v_*|^\alpha, \tag{6}$$

where $\alpha = (k-5)/(k-1)$. In particular, we will consider the variable hard sphere (VHS) model [2] i.e. $b_\alpha(\theta) = C_\alpha$ where C_α is a positive constant. The case $\alpha = 0$ is referred to as Maxwellian gas whereas the case $\alpha = 1$ yields the Hard Sphere gas. Note that in the case of Maxwellian gas the coefficient of the loss term, $L[f]$, does not depend on v. Boltzmann's collision operator has the fundamental properties of conserving mass, momentum and energy

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

and satisfies the well-known Boltzmann's H-theorem

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

Boltzmann H-theorem implies that any equilibrium distribution function, i.e. any function f for which $Q(f,f) = 0$, has the form of a locally Maxwellian distribution

$$M(\rho, u, T)(v) = \frac{\rho}{(2\pi T)^{d/2}} \exp(-\frac{|u-v|^2}{2T}), \tag{9}$$

where ρ, u, T are the density, mean velocity and temperature of the gas

$$\rho = \int_{\mathbb{R}^d} f(v) dv, \quad u = \frac{1}{\rho}\int_{\mathbb{R}^d} v\, f(v) dv, \quad T = \frac{1}{3\rho}\int_{\mathbb{R}^d} |u-v|^2\, f(v) dv. \tag{10}$$

Among the different approaches for the approximation of the Boltzmann equation, we may distinguish between deterministic and Monte Carlo methods. The first usually provide accurate oscillations-free solutions, but they are much more expensive than Monte Carlo methods with the same number of discrete degrees of freedom. For example, if we denote by n the number of parameters which characterize the density with respect to the velocity variables in a space homogeneous calculation, the computational cost of a conventional deterministic method for the evaluation of the collisional integral is much larger than n^2.

As a consequence most numerical computations are based on probabilistic Monte-Carlo techniques at different levels. Examples are the *Direct Simulation Monte Carlo method* (DSMC) by Bird [2] and the modified Monte Carlo method by Nanbu and Babovsky [14, 1]. For a detailed description of such methods we refer to previous chapters of this book.

Probabilistic particle methods present different advantages: the computational cost is strongly reduced and approximatively can be considered of the order of the number of points n. Moreover, the computer memory requirement is highly reduced, since the particles concentrate where the function is not small, and memory is not wasted representing a function which is virtually zero in most phase space. For these reasons, particle methods have no competitor for situations very far from thermodynamical equilibrium.

However, deterministic methods can be much more accurate, and can be competitive with Monte Carlo methods for problems in which the solution is not very far from thermodynamical equilibrium, and high accuracy is required. In the framework of deterministic approximations, the most popular class of methods is based on the so called *discrete velocity models* (DVM) of the Boltzmann equation. All these methods [4, 26, 12, 23] make use of regular discretizations on hypercubes in the velocity field and construct a discrete collision mechanics on the nodes of the hypercube in order to preserve the main physical properties. Although the numerical results have shown that these schemes are able to avoid fluctuations, their computational cost is high (in general $O(n_a n^2)$, where n_a is the number of parameters used for the angular integration, typically in such methods $n_a \approx O(n^{1/3})$) and, due to the particular choice of the integration points imposed by the conservation properties, the order of accuracy is lower than that of a standard quadrature formula applied directly to the collision operator. Hence we observe that the requirement of maintaining at a discrete level the main physical properties of the continuous equation makes it extremely difficult to obtain high order accuracy. Moreover, even if conservation properties are not imposed from the beginning, an accurate scheme would provide an accurate approximation of the conserved quantities.

In [35], Pareschi and Perthame developed a discretization of the collision operator based on expanding in Fourier series the distribution function with respect to the velocity variable. The resulting spectral approximation can be evaluated with a computational cost of $O(n^2)$ which is lower than that of previous deterministic methods. Bobylev and Rjasanow [3] used a Fourier transform approximation of the distribution function, and they were able to obtain exact conservation by a suitable modification of the evolution equations for the Fourier coefficients. The method proposed is second order accurate. On the other hand, Pareschi and Russo [20] developed a scheme based on the approximation of the distribution function by a periodic function in phase space, and its discretization by truncated Fourier series. Evolution equations for the Fourier modes are explicitly derived for the Variable Hard Sphere (VHS) model. The method provides spectral accuracy in the velocity domain, which is the highest accuracy achieved by a numerical method for the Boltzmann equation, and the computational complexity of the collisional operator is $O(n^2)$. The method preserves mass, and approximates with spectral accuracy momentum and energy. For a more detailed description of the spectral approach to the Boltzmann equation and to other kinetic equations see for example [17].

Here, we are interested in the construction of an accurate method for the space non homogeneous Boltzmann equation [10]. The discretization of the transport step has to be done carefully because it induces physical oscillations in the velocity space. In this chapter we construct a fractional step deterministic scheme for the time dependent Boltzmann equation, which is based on five main ingredients

Fractional step in time allows to treat separately the transport and the collision.

Fourier-Spectral method for the evolution of the collision step allows a very accurate discretization in velocity domain, at a reasonable computational cost [20].

Positive and Flux Conservative (PFC) finite volume method for the free transport [8] provides a third order (in space) accurate scheme for the evolution of distribution function during the transport step. The scheme is conservative, and preserves positivity. It is much less dissipative than Essentially Non Oscillatory (ENO) and Weighted Essentially Non Oscillatory (WENO) schemes usually used for hyperbolic systems of conservation laws [9, 24]. We also refer to [7] for the implementation of different boundary conditions.

Positive time discretization. A suitable time discretization of the collisional equation is used, which allows a large stability time step, even for problems with considerably small Knudsen number. The time discretization method for the collision step is based on a modified Time Relaxed scheme [21].

Multiple resolution. A different resolution will be used in velocity space in the transport and in the collision step. Considering that the collision step is more expensive, and more accurate (spectral accuracy) than the transport step, it is convenient to use more points in velocity space during the transport step.

In the next section we give a general setup to solve kinetic equations in non homogeneous situations. Then, we describe the PFC method for the free transport and the spectral method for the evolution of the collision step. Several numerical issues are discussed and time dependent and stationary problems are proposed. Finally, in the last section we draw conclusions.

2 The general framework.

Let us consider the initial-boundary value problem for the Boltzmann transport equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{k_n} Q(f,f) \tag{11}$$

$$f(0,x,v) = f_0(x,v)$$

where $x \in \Omega \subset \mathbb{R}^d$, $v \in \mathbb{R}^d$, $t \in [0,T]$. Boundary conditions will be specified in the section on numerical results and we refer to [7] for their implementation. We discretize time into discrete values t^n, and we denote by $f^n(x,v)$ an approximation of the distribution function $f(t^n,x,v)$. As it is usually done for a kinetic equation like (11), a simple first order time splitting is obtained considering, in a small time interval $\Delta t = [t^n, t^{n+1}]$, the numerical solution of the transport step

$$\begin{cases} \dfrac{\partial f^*}{\partial t} + v \cdot \nabla_x f^* = 0, \\ f^*(0,x,v) = f^n(x,v), \end{cases} \tag{12}$$

and the space homogeneous collision step

$$\begin{cases} \dfrac{\partial f^{**}}{\partial t} = \dfrac{1}{k_n} Q(f^{**}, f^{**}), \\ f^{**}(0,x,v) = f^*(\Delta t, x, v), \end{cases} \tag{13}$$

We shall denote by $S_1(\Delta t)$ and $S_2(\Delta t)$ the solution operators corresponding respectively to the transport and collision step, i.e. we can write

$$f^*(\Delta t, x, v) = S_1(\Delta t) f^n(x,v),$$

$$f^{**}(\Delta t, x, v) = S_2(\Delta t) f^*(\Delta t, x, v).$$

The approximated value at time t^{n+1} is then given by

$$f^{n+1}(x,v) = f^{**}(\Delta t, x, v) = S_2(\Delta t) S_1(\Delta t) f^n(x,v). \tag{14}$$

We assume that S_1 and S_2 represent either exact or at least second order evolution operators in time of transport and collision step, respectively.

A second order scheme for non stiff problems can be easily derived simply by symmetrizing the first order scheme [48]

$$f^{n+1} = S_1(\Delta t/2)S_2(\Delta t)S_1(\Delta t/2)f^n, \tag{15}$$

provided every step is solved with a method at least second order accurate in time [16]. Although higher order splitting strategies are available, in practice they are seldomly used because of stability problems. We remind that second order accuracy for such complex problems is considered "high order" in this field.

In the next two sections we discuss transport and collision steps. As we shall see, the grid step size in time, space and velocity are not directly related by strict stability requirements, and therefore one can benefit from high order accuracy whenever possible.

3 Discretization of the transport step.

In this section, we discuss the numerical resolution of the Vlasov equation which characterizes the transport step (12)

$$\partial_t f + \nabla_x (v\, f) = 0, \quad \forall (t,x) \in \mathbb{R}^+ \times \mathbb{R}^d. \tag{16}$$

Then, the solution of the transport equation at time t^{n+1} reads

$$f(t^{n+1}, x) = f(t^n, x - v\,\Delta t), \quad \forall x \in \mathbb{R}^d.$$

For simplicity, let us restrict ourselves to a one dimensional problem. We introduce a uniform mesh, characterized by a finite set of mesh points $\{x_{i+1/2}\}_{i\in I}$ on the computational domain. We will use the notation $\Delta x = x_{i+1/2} - x_{i-1/2}$, $C_i = [x_{i-1/2}, x_{i+1/2}]$ and x_i the center of C_i. Assuming the values of the distribution function are known at time $t^n = n\,\Delta t$ on cells C_i, we compute the new values at time t^{n+1} by integration of the distribution function on each sub-interval. Thus, using the explicit expression of the solution, we have

$$\int_{x_{i-1/2}}^{x_{i+1/2}} f(t^{n+1}, x)dx = \int_{x_{i-1/2}-v\,\Delta t}^{x_{i+1/2}-v\,\Delta t} f(t^n, x)dx,$$

then, setting

$$\Phi_{i+1/2}(t^n) = \int_{x_{i+1/2}-v\,\Delta t}^{x_{i+1/2}} f(t^n, x)dx,$$

we obtain the conservative form

$$\int_{x_{i-1/2}}^{x_{i+1/2}} f(t^{n+1}, x)dx = \int_{x_{i-1/2}}^{x_{i+1/2}} f(t^n, x)dx \,+\, \Phi_{i-1/2}(t^n) \,-\, \Phi_{i+1/2}(t^n). \tag{17}$$

The evaluation of the average of the solution over $[x_{i-1/2}, x_{i+1/2}]$ allows to ignore fine details of the exact solution which may be costly to compute. The main step is now to choose an efficient method to reconstruct the distribution function from the cell average on each cell C_i. We will consider a reconstruction via primitive function preserving positivity and maximum values of f [8]. Let $F(t^n, x)$ be a primitive of the distribution function $f(t^n, x)$, if we denote by

$$f_i^n = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} f(t^n, x) dx,$$

then $F(t^n, x_{i+1/2}) - F(t^n, x_{i-1/2}) = \Delta x\, f_i^n$ and

$$F(t^n, x_{i+1/2}) = \Delta x \sum_{k=0}^{i} f_k^n =: w_i^n.$$

First we construct an approximation of the primitive on the small interval $[x_{i-1/2}, x_{i+1/2}]$ using the stencil $\{x_{i-3/2}, x_{i-1/2}, x_{i+1/2}, x_{i+3/2}\}$

$$\begin{aligned}\tilde{F}_h(t^n, x) = w_{i-1}^n + (x - x_{i-1/2}) f_i^n + \frac{1}{2\Delta x}(x - x_{i-1/2})(x - x_{i+1/2})[f_{i+1}^n - f_i^n] \\ + \frac{1}{6\Delta x^2}(x - x_{i-1/2})(x - x_{i+1/2})(x - x_{i+3/2})[f_{i+1}^n - 2\, f_i^n + f_{i-1}^n],\end{aligned}$$

where we use the relation $w_i^n - w_{i-1}^n = \Delta x\, f_i^n$. Thus, by differentiation, we obtain a third order accurate approximation of the distribution function on the interval $[x_{i-1/2}, x_{i+1/2}]$

$$\begin{aligned}&\tilde{f}_h(t^n, x) = \frac{\partial \tilde{F}_h}{\partial x}(t^n, x) = f_i^n + \\ &+ \frac{1}{6\,\Delta x^2}\Big[2\,(x - x_i)(x - x_{i-3/2}) + (x - x_{i-1/2})(x - x_{i+1/2})\Big](f_{i+1}^n - f_i^n) \\ &- \frac{1}{6\,\Delta x^2}\Big[2\,(x - x_i)(x - x_{i+3/2}) + (x - x_{i-1/2})(x - x_{i+1/2})\Big](f_i^n - f_{i-1}^n).\end{aligned}$$

Unfortunately, this approximation does not preserve positivity of the distribution function f. Then, in order to satisfy a maximum principle and to avoid spurious oscillations we introduce slope correctors

$$\begin{aligned}&f_h(t^n, x) = f_i^n + \qquad\qquad (18)\\ &+ \frac{\epsilon_i^+}{6\,\Delta x^2}\Big[2\,(x - x_i)(x - x_{i-3/2}) + (x - x_{i-1/2})(x - x_{i+1/2})\Big](f_{i+1}^n - f_i^n) \\ &- \frac{\epsilon_i^-}{6\,\Delta x^2}\Big[2\,(x - x_i)(x - x_{i+3/2}) + (x - x_{i-1/2})(x - x_{i+1/2})\Big](f_i^n - f_{i-1}^n),\end{aligned}$$

with

$$\epsilon_i^{\pm} = \begin{cases} \min\Big(1; 2\, f_i^n/(f_{i\pm1}^n - f_i^n)\Big) & \text{if } f_{i\pm1}^n - f_i^n > 0, \\ \min\Big(1; -2\,(f_\infty - f_i^n)/(f_{i\pm1}^n - f_i^n)\Big) & \text{if } f_{i\pm1}^n - f_i^n < 0, \end{cases} \qquad (19)$$

where $f_\infty = \max_{j\in I}\{f_j^n\}$ is a local maximum.

The theoretical properties of this reconstruction can be summarized by the following

Proposition 1. *The approximation of the distribution function $f_h(x)$, defined by (18)-(19), satisfies*

- *The conservation of the average: for all $i \in I$,* $\int_{x_{i-1/2}}^{x_{i+1/2}} f_h(x)dx = \Delta x\, f_i$.
- *The maximum principle: for all $x \in (x_{min}, x_{max})$,* $0 \le f_h(x) \le f_\infty$.

Moreover, if we assume the Total Variation of the distribution function $f(x)$ is bounded, then we obtain the global estimate:

$$\int_{x_{min}}^{x_{max}} |f_h(x) - \tilde{f}_h(x)|\, dx \le 4\, TV(f)\, \Delta x,$$

where $\tilde{f}_h$ denotes the third order approximation of f without slope corrector.

Proof. Let us consider $x \in C_i = [x_{i-1/2}, x_{i+1/2}]$ and denote by

$$\alpha(x) = \frac{1}{\Delta x^2}\Big[2\,(x - x_i)(x - x_{i-3/2}) + (x - x_{i-1/2})(x - x_{i+1/2})\Big],$$
$$\beta(x) = \frac{1}{\Delta x^2}\Big[2\,(x - x_i)(x - x_{i+3/2}) + (x - x_{i-1/2})(x - x_{i+1/2})\Big].$$

It is easy to check that

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \alpha(x)dx = \int_{x_{i-1/2}}^{x_{i+1/2}} \beta(x)dx = 0,$$

then the conservation of the average immediately follows. To obtain the preservation of positivity, assuming the values f_j are positive, we observe that in the cell C_i, the function $\alpha(x)$ is increasing whereas $\beta(x)$ decreases and $\alpha(x)$, $\beta(x) \in [-1, 2]$. Then, we split $f_h(x)$ as the sum of $h(x)$ and $g(x)$ with

$$h(x) = \frac{1}{3}\left[f_i + \frac{\alpha(x)}{2}\epsilon_i^+(f_{i+1} - f_i)\right], \quad g(x) = \frac{1}{3}\left[2\,f_i - \frac{\beta(x)}{2}\epsilon_i^-(f_i - f_{i-1})\right].$$

The function $h(x)$ (*resp.* $g(x)$) is only a combination of f_i and f_{i+1} (*resp.* f_{i-1} and f_i), then from the value of ϵ_i^+ (*resp.* ϵ_i^-), it is easy to prove that $h(x)$ (*resp.* $g(x)$) is positive. Using a similar decomposition, we also prove that $f_h(x)$ is bounded by f_∞.

Now, we prove the global estimate on the positive reconstruction:

$$
\begin{aligned}
&\int_{x_{min}}^{x_{max}} |f_h(x) - \tilde{f}_h(x)|dx \\
&= \sum_i \int_{x_{i-1/2}}^{x_{i+1/2}} \mid \alpha(x)\,(1-\epsilon_i^+)[f_{i+1} - f_i] + \beta(x)\,(1-\epsilon_i^-)[f_i - f_{i-1}] \mid dx \\
&\leq 2\,\Delta x \sum_i (1-\epsilon_i^+)|f_{i+1} - f_i| + 2\,\Delta x \sum_i (1-\epsilon_i^-)|f_i - f_{i-1}| \\
&\leq 4\,\Delta x \sum_i |f_{i+1} - f_i| \leq 4\,\Delta x\, TV(f).
\end{aligned}
$$

□

Remark 1. If the solution is smooth, we can check numerically that the scheme is third order. But, the numerical analysis of such a nonlinear scheme is really difficult to perform.

4 Spectral approximation of the collision operator.

We consider now the space homogeneous Boltzmann equation in each cell,

$$\frac{\partial f}{\partial t} = Q^+(f,f) - L[f]f \tag{20}$$

with Q^+ and L given by equations (3) and (4). To keep notation simple, we have fixed $k_n = 1$. A simple change of variables permits to write

$$Q^+(f,f) = \int_{\mathbb{R}^d} \int_{S^{d-1}} B(|g|,\theta) f(v') f(v'_*)\, d\omega\, dg, \tag{21}$$

$$L(f) = \int_{\mathbb{R}^d} \int_{S^{d-1}} B(|g|,\theta) f(v-g)\, d\omega\, dg, \tag{22}$$

where $g = v - v_*$ and then

$$v' = v - \frac{1}{2}(g - |g|\omega), \quad v'_* = v - \frac{1}{2}(g + |g|\omega). \tag{23}$$

First, from the conservation of the momentum and the total energy, $(v'_*)^2 + (v')^2 = v_*^2 + v^2$, we get the following result [35]:

Lemma 1. *Let* $\mathrm{Supp}(f(v)) \subset \mathcal{B}(0,R)$ *then*

i) $\mathrm{Supp}(Q(f,f)(v)) \subset \mathcal{B}(0,\sqrt{2}R)$*,*

ii)

$$Q(f,f)(v) = \int_{\mathcal{B}(0,2R)} \int_{S^{d-1}} B(|g|,\theta)[f(v')f(v'_1) - f(v)f(v-g)]\, d\omega\, dg,$$

with $v', v'_*, v - g \in \mathcal{B}(0,(2+\sqrt{2})R)$.

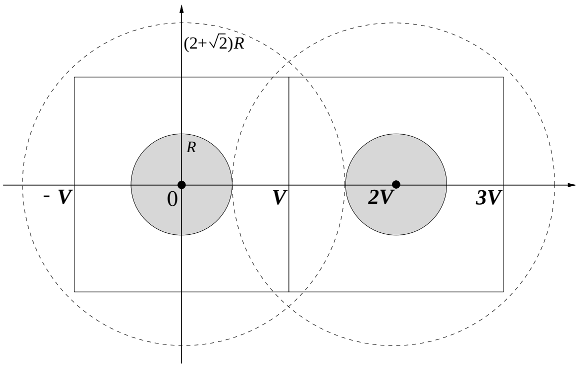


Fig. 1. Restriction of the distribution function on the periodic box $[-\pi,\pi]\times[-\pi,\pi]$ in two dimensions.

As a consequence of the above lemma, in order to write a spectral approximation to (20) we can consider the distribution function $f(v)$ restricted on $[-V,V]^d$ with $V \geq (2+\sqrt{2})R)$, assuming $f(v)=0$ on $[-V,V]^d \setminus \mathcal{B}(0,R)$, and extend it by periodicity to a periodic function on $[-V,V]^d$. In view of Fig. 1, the shortest period can be restricted to $[-V,V]$, with $V \geq (3+\sqrt{2})R/2$.

If the distribution function is well approximated by a function of compact support in velocity space, then the above approximation will provide an accurate evaluation of the collision integral.

To simplify the notation let us take $V=\pi$ and hence $R=\lambda\pi$ with $\lambda = 2/(3+\sqrt{2})$. Hereafter, we used just one index to denote the tree-dimensional sums with respect to the vector $k=(k_1,..,k_d)\in\mathbb{Z}^d$, hence we set

$$\sum_{k=-N}^{N} = \sum_{k_1,..,k_d=-N}^{N}.$$

The approximate function f_N is represented as the truncated Fourier series

$$f_N(v) = \sum_{k=-N}^{N} \hat{f}_k e^{ik\cdot v}, \tag{24}$$

$$\hat{f}_k = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} f(v) e^{-ik\cdot v}\,dv.$$

In a Fourier-Galerkin method the fundamental unknowns are the coefficients $\hat{f}_k,\ k=-N,\dots,N$. We obtain a set of ODEs for the coefficients $\hat{f}_k$ by requiring that the residual of (20) be orthogonal to all trigonometric polynomials of degree $\leq N$. Hence for $k=-N,\dots,N$

$$\int_{[-\pi,\pi]^d}\left(\frac{\partial f_N}{\partial t} + f_N\,L(f_N) - Q^+(f_N,f_N)\right) e^{-ik\cdot v}\,dv = 0. \tag{25}$$

By substituting expression (24) in (22) and (21) we get respectively

$$f_N\,L(f_N) = \sum_{l=-N}^{N}\sum_{m=-N}^{N} \hat{f}_l\,\hat{f}_m \hat{B}(m,m) e^{i(l+m)\cdot v},$$

and

$$Q^+(f_N, f_N) = \sum_{l=-N}^{N}\sum_{m=-N}^{N} \hat{f}_l\,\hat{f}_m \hat{B}(l,m) e^{i(l+m)\cdot v},$$

where the *kernel modes* $\hat{B}(l,m)$ are given by

$$\hat{B}(l,m) = \int_{\mathcal{B}(0,2\lambda\pi)}\int_{S^{d-1}} B(|g|,\theta) e^{-ig\cdot\frac{(l+m)}{2} - i|g|\omega\cdot\frac{(m-l)}{2}}\,d\omega\,dg. \tag{26}$$

Using the orthogonality property we get from (25)

$$\frac{\partial \hat{f}_k}{\partial t} = \sum_{m=k-N}^{N} \hat{f}_{k-m}\,\hat{f}_m(\hat{B}(k-m,m) - \hat{B}(m,m)), \tag{27}$$

with the initial condition

$$\hat{f}_k(0) = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} f_0(v) e^{-ik\cdot v}\,dv. \tag{28}$$

The evaluation of the right hand side of (71) requires exactly $\mathcal{O}(N^{2d})$ operations. We emphasize that the usual cost for a method based on N^d parameters for f in the velocity space is $\mathcal{O}(N^{2d}M)$ where M is the numbers of angle discretizations. The loss term on the right hand side is a convolution sum and thus transform methods allow this term to be evaluated only in $\mathcal{O}(N^d \log N)$ operations. Hence the most expensive part of the computation is represented by the gain term.

4.1 Analysis of the kernel modes.

In this section we study the main characteristics of the kernel modes and in particular we give an explicit representation of them for the VHS model.

Let us start from equation (26). In the VHS model, the kernel does not depend on the angle θ: $B = C_\alpha |g|^\alpha$. One has

$$\hat{B}(l,m) = C_\alpha \int_{\mathcal{B}(0,2\lambda\pi)} |g|^\alpha \exp\left(-ig\cdot\frac{l+m}{2}\right) I_2(|g|, l-m), dg \tag{29}$$

where

$$I_2(|g|, l-m) = \int_{S^{d-1}} \exp\left(i|g|\omega\cdot\frac{l-m}{2}\right)\,d\omega. \tag{30}$$

We shall consider separately 3D and 2D collision model. The 3D case is the important one for practical application. However, a two dimensional collisional model will be considered for test problems.

2D case.

For the computation in 2D we start from (29) and (30). We will consider only the VHS model.

In this case it is

$$I_2 = \int_S \exp(i|q| \cdot \omega)\, d\omega = \int_0^{2\pi} \exp(ir \cos\theta)\, d\theta$$
$$= 2\int_0^{\pi} \cos(r\cos\theta)\, d\theta = 2\pi J_0(r), \tag{31}$$

where $r = |q| = |g||l-m|/2$, and J_0 is the Bessel function of order 0. By inserting the result in the expression (29) for $\hat{B}(l,m)$, one obtains [20]

$$\hat{B}(l,m) = C_\alpha 2\pi \int_{\mathcal{B}(0,2\lambda\pi)} |g|^\alpha \exp(-ig\cdot(l+m)/2) J_0(|l-m||g|/2)\, dg.$$

Making use of polar coordinates, the expression for the coefficients becomes

$$\hat{B}(l,m) = C_\alpha 2\pi \int_0^{2\pi\lambda} \rho^{1+\alpha} \left(\int_0^{2\pi} \cos(|l+m|\rho/2)\cos\theta\, d\theta \right) J_0(|l-m|\rho/2)\, d\rho$$
$$= C_\alpha 4\pi^2 \int_0^{2\pi\lambda} \rho^{1+\alpha} J_0(|l+m|\rho/2) J_0(|l-m|\rho/2)\, d\rho$$
$$= C_\alpha 4\pi^2 (2\pi\lambda)^{2+\alpha} \int_0^1 r^{1+\alpha} J_0(\xi r) J_0(\eta r)\, dr \tag{32}$$

where $\xi = |l+m|\lambda\pi$, $\eta = |l-m|\lambda\pi$. Taking now $C_\alpha = (4\pi^2(2\pi\lambda)^{2+\alpha})^{-1}$, the expression of $\hat{B}(l,m)$ becomes

$$\hat{B}(l,m) = F_\alpha(\xi,\eta)$$

with

$$F_\alpha(\xi,\eta) = \int_0^1 r^{1+\alpha} J_0(\xi r) J_0(\eta r)\, dr. \tag{33}$$

From (33) it is easy to prove that an analogue of proposition 2 holds also in the two dimensional case. Note that also in this case each kernel mode can be computed as a 1-D integral and stored in an array.

3D case.

Let $q = |g|(l-m)/2$. Then I_2 is computed as follows

$$I_2(|g|, l-m) = \int_{S^2} e^{iq\cdot\omega}\, d\omega = 2\pi \int_0^{\pi} e^{i|q|\cos\theta} \sin\theta\, d\theta$$
$$= 2\pi \int_{-1}^{1} e^{i|q|\mu}\, d\mu = 4\pi\,\mathrm{Sinc}(|q|)$$
$$= 4\pi\,\mathrm{Sinc}\left(\frac{|g||l-m|}{2}\right) \tag{34}$$

where

$$\mathrm{Sinc}(x) \equiv \frac{\sin x}{x}.$$

Let $p = (l+m)/2$. Then, taking into account the previous result, one obtains [20]

$$\hat{B}(l,m) = C_\alpha 4\pi \int_{\mathcal{B}(0,2\lambda\pi)} |g|^\alpha \,\mathrm{Sinc}(|l-m||g|/2) \exp(-ip \cdot g)\, dg$$

Making use of spherical coordinates, with $\rho = |g|$, one has

$$\begin{aligned} \hat{B}(l,m) &= C_\alpha 8\pi^2 \int_0^{2\pi\lambda} \rho^{2+\alpha} \,\mathrm{Sinc}(|l-m|\rho/2)\, d\rho \int_0^\pi \exp(-i|p|\rho\cos\theta)\sin\theta\, d\theta \\ &= C_\alpha 16\pi^2 \int_0^{2\pi\lambda} \rho^{2+\alpha} \,\mathrm{Sinc}(|l-m|\rho/2)\,\mathrm{Sinc}(|l+m|\rho/2)\, d\rho. \end{aligned} \tag{35}$$

With the change of variables $\rho = 2\lambda\pi r$ the coefficient $\hat{B}(l,m)$ can be written as

$$\hat{B}(l,m) = C_\alpha 16\pi^2 (2\lambda\pi)^{3+\alpha} \int_0^1 r^{2+\alpha} \,\mathrm{Sinc}(\xi r)\,\mathrm{Sinc}(\eta r)\, dr$$

where $\xi = |l+m|\lambda\pi$, $\eta = |l-m|\lambda\pi$. To simplify notations let us assume that

$$C_\alpha = (16\pi^2(2\lambda\pi)^{3+\alpha})^{-1}.$$

In this case the coefficient can be written as

$$\hat{B}(l,m) = F_\alpha(\xi,\eta)$$

where

$$F_\alpha(\xi,\eta) = \int_0^1 r^{2+\alpha} \,\mathrm{Sinc}(\xi r)\,\mathrm{Sinc}(\eta r)\, dr. \tag{36}$$

From (36) it is easy to prove the following

Proposition 2. *Let $F_\alpha(\xi,\eta)$ be defined by (36) then*

i) $F_\alpha(\xi,\eta) = F_\alpha(\eta,\xi)$,
ii) if $\alpha > -3$ then $|F_\alpha(\xi,\eta)| \le F_\alpha(0,0) = (3+\alpha)^{-1}$,
iii) if $\alpha > -1$ then $|F_\alpha(\xi,\eta)| \le [\xi\eta(1+\alpha)]^{-1}$.

For integer values of α, F_α has an explicit analytical expression. We give here the expressions for $\alpha = 0$ (Maxwellian gas) and $\alpha = 1$ (Hard Sphere gas)

$$F_0(\xi,\eta) = \frac{p\sin(q) - q\sin(p)}{2\xi\eta pq} \tag{37}$$

$$F_1(\xi,\eta) = \frac{p^2(q\sin(q)+\cos(q)) - q^2(p\sin(p)+\cos(p)) - 4\xi\eta}{2\xi\eta p^2 q^2} \tag{38}$$

where $p = (\xi+\eta)$, $q = (\xi-\eta)$.

Storage of Fourier coefficients.

Note that, since the five-fold integral (26) which defines the $\hat{B}(l,m)$ has been reduced to a one-dimensional integral (36), for non-integer values of α, the value of the coefficients can be easily computed numerically by an accurate quadrature formula, and stored in an array at the beginning of the calculation.

In $3D$, the storage of coefficients $\hat{B}(l,m)$ is of order $O(n^6)$, where n is the number of half modes for each direction. But, it can be easily reduced to $O(n^4)$. Indeed, the matrix $\hat{B}(l,m)$ only depends on $|k| = |l+m|$ and $|l-m|$, it is then replaced by the smaller matrix $\hat{\mathbf{B}}(i,j)$, where integers i, j are given by

$$0 \le i = |k|^2 \le 3n^2, \quad 0 \le j = |l-m|^2 \le 12n^2.$$

4.2 Properties of the spectral method.

We state here the main theoretical results of the Fourier-Spectral method, concerning consistency and spectral accuracy. For any function $f(v)$, let $f_N(v)$ denote the truncated Fourier series of f, and let $\mathcal{P}_N : L^2([-\pi,\pi]^d) \to I\!P^N$ denote the projection operator, with

$$I\!P^N = span\left\{e^{ik\cdot v} \mid -N \le k_j \le N,\, j = 1,..,d\right\}.$$

Then the following results hold [20]

Proposition 3. *Let $f \in L^2([-\pi,\pi]^d)$, $f \ge 0$, $\forall\, v \in [-\pi,\pi]^d$, and let us define*

$$\begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix} := \int_{[\pi,\pi]^d} f \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv. \qquad (39)$$

and let us denote by ρ_N, ρu_N, and ρe_N the moments of f_N, then the following relations hold

$$\rho = \rho_N,$$
$$|\rho u - \rho u_N| \le \frac{C_1}{N}||f||_2,$$
$$|\rho e - \rho e_N| \le \frac{C_2}{N^2}||f||_2.$$

The estimates given above can be strongly improved if f is smooth. If $f \in H^r_p([-\pi,\pi]^d)$, where $r \ge 0$ is an integer and $H^r_p([-\pi,\pi]^d)$ is the subspace of the Sobolev space $H^r([-\pi,\pi]^d)$, which consists of periodic functions [5], for each $\varphi \in L^2([-\pi,\pi]^d)$ we have

$$|< f,\varphi > - < f,\varphi_N > | \le ||\varphi||_2\, ||f - f_N||_2 \le \frac{C}{N^r}||\varphi||_2||f||_{H^r_p},$$

where $||\cdot||_{H^r_p}$ denotes the norm in $H^r_p([-\pi,\pi]^d)$, and $< f, g >$ denotes the scalar product in L^2_p. This inequality shows that the projection error on the moments decay faster than algebraically when the solution is infinitely smooth.

We state a consistency result in the L^2-norm for the approximation of the collision operator $Q(f,f)$ with $Q_N(f_N,f_N)$ [20, 43],

Theorem 1. *Let $f \in L^2([-\pi,\pi]^d)$, and $B(u,\theta) = C_\alpha |u|^\alpha$, with $\alpha > 0$ then*

$$||Q(f,f) - Q_N(f_N,f_N)||_2 \leq C \left(||f - f_N||_2 + \frac{||Q(f_N,f_N)||_{H_p^r}}{N^r} \right), \quad \forall\, r \geq 0, \tag{40}$$

where C depends on $||f||_2$.

Proof. First, let us split the error in two parts

$$\begin{aligned} ||Q(f,f) - Q_N(f_N,f_N)||_2 \leq\ & ||Q(f,f) - Q(f_N,f_N)||_2 \\ & + ||Q(f_N,f_N) - Q_N(f_N,f_N)||_2 \end{aligned}$$

On the one hand, observing that $Q(f_N,f_N) \in I\!P^{2N}$ and hence $Q(f_N,f_N)$ is periodic and infinitely smooth [5]

$$||Q(f_N,f_N) - Q_N(f_N,f_N)||_2 \leq \frac{C_r}{N^r} ||Q(f_N,f_N)||_2, \quad \forall r \geq 0. \tag{41}$$

On the other hand using the symmetry of the Boltzmann operator, we get

$$Q(f,f) - Q(f_N,f_N) = Q(f + f_N, f - f_N). \tag{42}$$

Now, let us prove that

$$||Q(f + f_N, f - f_N)||_2 \leq C ||f + f_N||_1 \, ||f - f_N||_2. \tag{43}$$

We use a duality argument : for each function $\varphi \in L^2([-\pi,\pi]^d)$, we get form the Hölder inequality

$$\begin{aligned} \left| \int Q(f,g)\varphi \, dv \right| &= C_\alpha \left| \int \int |v - v_*|^\alpha g(v_*)\, f(v)\, \left(\varphi(v') - \varphi(v)\right) d\omega dv_\star dv \right| \\ &\leq C_\alpha ||g||_1 \sup_{v_* \in [-\pi,\pi]^d} \int |v - v_*|^\alpha f(v) \left| \int_{S^{d-1}} (\varphi(v') - \varphi(v)) d\omega \right| dv. \end{aligned}$$

Let us fix $v_\star$, then from the Cauchy-Schwartz inequality we obtain

$$\begin{aligned} & \int |v - v_*|^\alpha f(v) \left| \int_{S^{d-1}} (\varphi(v') - \varphi(v)) d\omega \right| dv \\ & \leq ||f||_2 \left\| |v - v_*|^\alpha \int_{S^{d-1}} (\varphi(v') - \varphi(v)) d\omega \right\|_2 . \end{aligned}$$

Using the invariance by translation, it is enough to prove this estimate for $v_* = 0$. Moreover, the function $v \to |v|^\alpha$ defined in $[-\pi,\pi]^d$ is bounded, and thus there exists a constant C, independant of N, such that

$$\left\| |v|^\alpha \int_{S^{d-1}} (\varphi(v') - \varphi(v)) d\omega \right\|_{L^2} \leq C \, ||\varphi||_2. \tag{44}$$

Finally, we have shown that there exists a constant C, independant of N, such that for each smooth function φ

$$\left| \int Q(f,g)\varphi \, dv \right| \leq C \, ||g||_{L^1} ||f||_2 ||\varphi||_2,$$

which proves the inequality (43) with $g = f + f_N$ and $f = f - f_N$. Gathering inequalities (41) and (43), we conclude the proof. □

The previous estimate states that the rate of convergence in the L^2-norm of $Q_N(f_N)$ to $Q(f)$ depends only on the speed of convergence of f_N to f. Hence if f_N is spectrally accurate so it is $Q_N(f_N)$. The following corollary states the spectral accuracy of the approximation of the collision operator

Corollary 1. *Let $f \in H^r([-\pi,\pi]^d)$, $r \geq 0$ then*

$$||Q(f) - Q_N(f_N)||_2 \leq \frac{C}{N^r} \left(||f||_{H^r} + ||Q(f_N)||_{H^r} \right). \tag{45}$$

4.3 Time discretization of the collision operator.

Here we focus on the time evolution of the collision step. Let Δt denote the time step of the transport phase. The goal is to solve, in each cell, the space homogeneous Boltzmann equation

$$\frac{\partial f^*}{\partial t} = \frac{1}{k_n} Q(f^*, f^*)$$

$$f^*(0,v) = f^n(v)$$

where, for simplicity, we drop the space dependence. One could use any second order time discretization, such as a Runge-Kutta method, with the same time step, Δt, used for the convection step, for the ordinary differential system of the Fourier modes, (71). If the time step is too large (for accuracy or stability reasons), then a smaller time step, $\Delta t_c < \Delta t$, can be used during this phase. Since each cell is independent, Δt_c may depend on the cell. If $\Delta t_c << \Delta t$, then a multi-step scheme can be used to improve efficiency and accuracy. With standard methods such as Runge-Kutta or multi-step, it is difficult to control positivity of the solution. Here we propose a time discretization which provides essential positivity of the distribution function, and allows the use of rather large time steps, even in regimes in which the Knudsen number is quite small. The schemes that we use are based on a variation of time relaxed (TR) schemes [21], which have been effectively used in the development of Monte Carlo methods suitable for a very wide range of Knudsen number, and for the space non-homogeneous Boltzmann equation [10]. We briefly recall here the idea behind the TR schemes.

Let us consider an equation of the form

$$\frac{\partial f}{\partial t} = \frac{1}{k_n}\left[P(f,f) - \mu f\right], \qquad (46)$$
$$f(0,v) = f_0(v),$$

where $\mu \neq 0$ is a constant and P a positive bilinear operator.

The Boltzmann equation for Maxwell molecules has the above form, with $Q^+(f,f) = P(f,f)$, and $L[f] = \mu$.

Let us replace the time variable t and the function $f = f(t,v)$ using the equations

$$\tau = (1 - e^{-\mu t/k_n}), \qquad F(\tau,v) = f(t,v)e^{\mu t/k_n}. \qquad (47)$$

Then F is easily shown to satisfy

$$\frac{\partial F}{\partial \tau} = \frac{1}{\mu}P(F,F), \qquad (48)$$

with $F(\tau = 0, v) = f_0(v)$.

Now, the solution to the Cauchy problem for (48) can be sought in the form of a power series

$$F(\tau,v) = \sum_{k=0}^{\infty} \tau^k f_k(v), \qquad f_{k=0}(v) = f_0(v), \qquad (49)$$

where the functions f_k are given by the recurrence formula

$$f_{k+1}(v) = \frac{1}{k+1}\sum_{h=0}^{k}\frac{1}{\mu}P(f_h, f_{k-h}), \quad k = 0, 1, \ldots \qquad (50)$$

Making use of the original variables we obtain the following formal representation of the solution to the Cauchy problem (20), called *Wild sum expansion* [29].

$$f(t,v) = e^{-\mu t/k_n}\sum_{k=0}^{\infty}\left(1 - e^{-\mu t/k_n}\right)^k f_k(v). \qquad (51)$$

The coefficients f_k have the property that

$$\lim_{k\to\infty} f_k(v) = M(v), \qquad (52)$$

where $M(v)$ is the Maxwellian, satisfying

$$Q(M,M) = 0.$$

Representation (51) and property (52) suggest the use of a truncation of series (51) as a numerical scheme for time discretization, namely

$$f^{n+1}(v) = (1-\tau)\sum_{k=0}^{m} \tau^k f_k(v) + \tau^{m+1} M(v), \tag{53}$$

with $f_k(v)$ computed from $f^n(v)$. Such scheme is of order m in $\mu\Delta t/k_n$, and has the following properties [21]

Proposition 4. *The Time-Relaxed scheme given by (53) satisfies*

i) conservation: *if $P(f,g)$ is a non negative bilinear operator such that there exist some functions $\phi(v)$ with the following property*

$$\int_{\mathbb{R}^d} P(f,f)\phi(v)\,dv = \mu \int_{\mathbb{R}^d} f\phi(v)\,dv, \tag{54}$$

and the initial condition f^0 is a non negative function, then f^{n+1} is nonnegative for any $\mu\Delta t/k_n$, and satisfies

$$\int_{\mathbb{R}^d} f^{n+1}\phi(v)\,dv = \int_{\mathbb{R}^d} f^n\phi(v)\,dv; \tag{55}$$

ii) asymptotic preservation (AP):
for any $m \geq 1$, we have

$$\lim_{\mu\Delta t/k_n \to \infty} f^{n+1} = M(v). \tag{56}$$

Proof. The result is straightforward using the construction of the scheme. □

The above time discretization can be generalized using different weight functions to combine the influence of the high order coefficients appearing in the Wild sum (51). In general such schemes can be written as

$$f^{n+1} = \sum_{k=0}^{m} A_k f_k + A_{m+1} M, \tag{57}$$

where the coefficients f_k are given by (50) using $f = f^n(v)$. The weights $A_k = A_k(\tau)$ are nonnegative functions that satisfy some consistency condition.

Proposition 5. *If $A_k(\tau) \geq 0$ satisfy*

i) consistency:

$$\lim_{\tau\to 0} A_1(\tau)/\tau = 1, \quad \lim_{\tau\to 0} A_k(\tau)/\tau = 0, \quad k = 2, \ldots, m+1, \tag{58}$$

ii) conservation:

$$\sum_{k=0}^{m+1} A_k = 1 \quad \tau \in [0,1], \tag{59}$$

iii) asymptotic preservation (AP):

$$\lim_{\tau \to 1} A_k(\tau) = 0, \quad k = 0, \dots, m, \tag{60}$$

then (57) is a consistent discretization of problem (46) that satisfies proposition 4.

A choice of functions which satisfies the previous requirements is given by

$$A_k = (1-\tau)\tau^k, \; k = 0, \dots, m, \quad A_{m+1} = \tau^{m+1}, \tag{61}$$

which correspond to the scheme (53). A better choice of parameters is [19]

$$A_k = (1-\tau)\tau^k, \; k = 0, \dots, m-1, \quad A_m = 1 - \sum_{k=0}^{m} A_k - A_{m+1}, \quad A_{m+1} = \tau^{m+2}, \tag{62}$$

which corresponds to take $f_{m+1} = f_m$, $f_k = M$, $k \geq m+2$ in (51). However, other choices are possible and it is an open problem the determination of the *optimal* set of functions A_k that satisfies the previous requirements and guarantees the most accurate approximation.

The Boltzmann equation for Maxwell molecules has the form (46), with $P(f,f) = Q^+(f,f)$. In order to apply the same discretization to a more general B.E., we can proceed as follows. We write the Boltzmann equation in the form (46), with

$$P(f,g) = Q^+(f,g) + \frac{1}{2}\left(\mu(f+g) - L[f]g - L[g]f\right),$$

where the operator P is written in a symmetric form. If we choose

$$\mu \geq L[f](v) \; \forall v \in \mathbb{R}^d, \tag{63}$$

then $P(f,f)$ is a positive symmetric operator. However, in general $L(v)$ is an unbounded function, and therefore a constant μ satisfying (63) does not exist. Even if we consider that the discrete velocities lie in a bounded domain $\Omega_v = [-V,V]^3$, a choice of μ satisfying (63) may lead to excessive numerical viscosity, as is evident from standard truncation analysis.

Ideally, one should choose the smallest value of μ that guarantees positivity of operator P. This should be obtained, for a given function $f(v)$, by imposing that

$$\min_{v \in \Omega_v} [Q^+(v) + (\mu - L(v))f(v)] = 0. \tag{64}$$

A first order (non AP) TR scheme has the structure

$$f^{n+1}(v) = A_0(\tau)f^n(v) + A_1(\tau)f_1(v),$$

with $f_1 = P(f^n, f^n)/\mu$. Because of positivity of the coefficients $A_0(\tau)$ and $A_1(\tau)$, if $P(f,f)$ is positive, then the scheme is positive.

Condition (64) is not practical, since the region of phase space Ω_v near the edge is not physically representative, because of the approximation of the distribution function by a periodic function in velocity. A better choice is obtained by computing a critical constant μ_c as

$$\mu_c = \max_{v \in \Omega_c} \left(L(v) - \frac{Q_+}{f} \right), \tag{65}$$

where $\Omega_c \subset \Omega_v$ is a smaller region, for example, $\Omega_c = [-V/2, V/2]^3$. Then the constant μ is computed as $\mu = C\mu_c$, where C is a safety factor of order one. In all the calculations we used the value $C = 3/2$, which we found a good compromise between numerical positivity and numerical viscosity.

These practical criteria deserve further analysis. We have to keep in mind that the spectral scheme itself does not preserve positivity rigorously [20]. However, the lack of positivity is very small, and can be neglected for all practical purposes. Even if positive spectral scheme can be obtained, as shown in [20] and [19], they are not practical, because of the lack of accuracy and excessive smoothing.

Accuracy requires a small value of $\mu\Delta t/k_n$. If μ is kept constant, independently of k_n, then the time step becomes exceedingly small for small values of k_n, and the method becomes inefficient. The approach that we outlined above allows rather large time steps, even for small values of k_n. The reason for this is that when k_n is small, then gain and loss terms balance each other, and therefore the quantity μ computed as above becomes small. It appears in fact that for small values of k_n, μ scales with k_n, and the ratio μ/k_n remains bounded.

With these considerations in mind, we maintain the same criterion for the evaluation of the optimal μ, even for higher order TR schemes.

A second order (non AP) TR scheme has the structure

$$f^{n+1}(v) = A_0(\tau) f^m(v) + A_1(\tau) f_1(v) + A_2(\tau) f_2(v),$$

with $f_2(v) = P(f^n, f_1)$. The constant μ is computed as above, $\mu = 3/2\mu_c$, with μ_c given by (65). For all practical purpose the scheme can be considered positive, although it is not rigorously positive.

To conclude this section, let us mention that in [10], we present two algorithms to reduce the computational cost and to improve accuracy. On the one hand, a parallel algorithm based on the fractional step approach is given. On the other hand, a simple multi-resolution is proposed to solve the transport and collision steps using different grids.

5 Numerical tests.

In this section, we present test cases showing the effectiveness of the spectral-PFC method to get an accurate solution of the Boltzmann equation. We first give two simple numerical tests in the 3D homogeneous case (without x) in order to illustrate the spectral accuracy of the method. In space dependent tests we used 2D and 3D models of the Boltzmann equation in velocity space and present results to compare the scheme with the well known Monte-Carlo method for the Boltzmann equation. We refer to [10] for more numerical results (Riemann problem and stationary shock waves for a model Boltzmann equation 2D in velocity). Finally, we present a comparison with the Monte Carlo method for the evalutation of a stationary shock.

5.1 3D space homogeneous case: spectral accuracy.

We consider 3D Maxwellian molecules (i.e. $\alpha = 0$), with $C_0 = 1/(4\pi)$. This problem has an exact solution given by

$$f(t,v) = \frac{\exp(-v^2/2S)}{2S\,(2\pi\sigma)^{3/2}}\left[5S - 3 + \frac{1-S}{S}v^2\right],$$

where $S = 1 - \exp(-(t+t_0)/6)$, $t \geq 0$, $t_0 = 5.6 > 6\,\log(5/2)$. This test is used to check spectral accuracy, by comparing the error at a given time, when using $n = 8$, 16, and 32 Fourier modes for each dimension, to check the accuracy in the conservation of energy and to observe the evolution of the fourth moment. Because of the symmetry of the problem, the moments of order 1 and 3 are conserved, within round-off error. In figure 2 we report the L^1 relative error vs time, for different number of modes, and the fourth order moment.

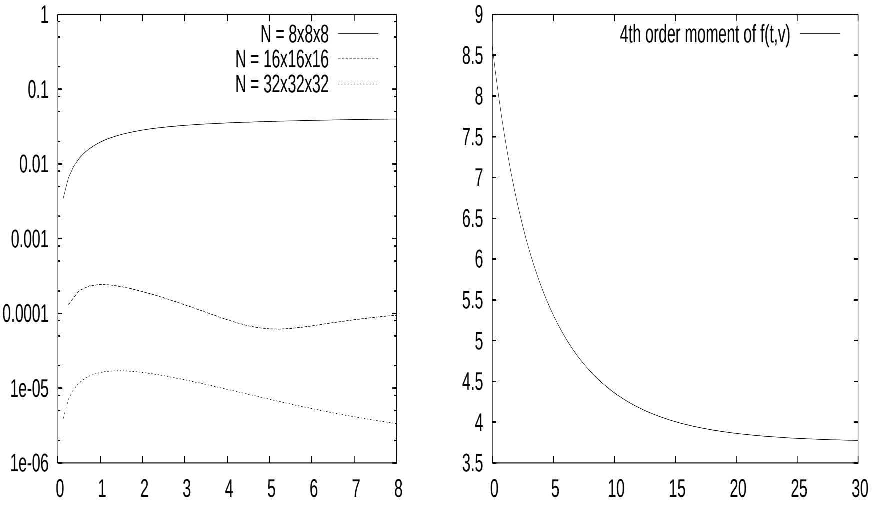


Fig. 2. 3D homogeneous case I: *evolution of the numerical L^1 relative error and the fourth order moment of $f(t,v)$.*

5.2 3D space homogeneous case: convergence to equilibrium.

We present a first result for the 3D Boltzmann equation without the transport part and consider Maxwellian ($\alpha = 0$) and hard-sphere ($\alpha = 1$) molecules, with $C_\alpha = 1/(2\pi)$. The initial condition is sum of two Gaussians

$$f(v,0) = \frac{1}{2(2\pi\sigma^2)^{3/2}} \left[\exp\left(-\frac{|v-v_1|^2}{2\sigma^2} \right) + \exp\left(-\frac{|v-v_2|^2}{2\sigma^2} \right) \right],$$

with $\sigma^2 = 0.2$, $v_1 = (1,1,1/4)$, $v_2 = (-1,-1,-1/4)$ and the final time is $t_{\text{max}} = 2$. This test is used to check the evolution of the distribution function and to observe the relaxation to equilibrium.

We first define the directional temperature

$$T_\alpha(t) = \frac{1}{\rho} \int_{\mathbb{R}^3} (v_\alpha - u_\alpha)^2 f(t,v) dv, \quad \alpha \in \{x,y,z\}$$

where ρ and u are given by (53) and the entropy

$$H(t) = \int_{\mathbb{R}^3} f(t,v) \, \log(f(t,v)) dv.$$

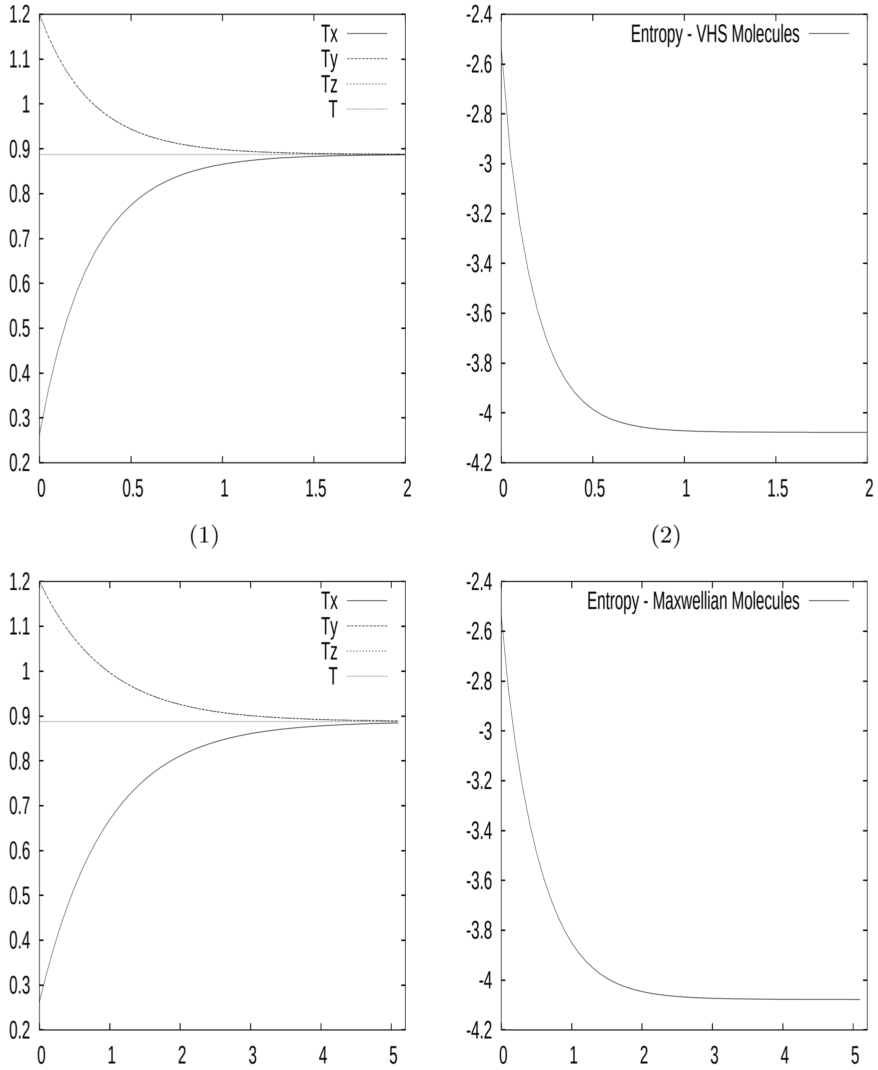
In Fig. 3, the relaxation of $T_\alpha(t)$, with $\alpha \in \{x,y,z\}$ and $H(t)$ for Hard-sphere and Maxwellian molecules are presented starting from the same initial data. Finally, the evolution of the distribution function f is given in Fig. 4.

5.3 Riemann problem: time dependent solutions.

This test deals with the numerical solution of the non homogeneous $1D \times 2D$ Boltzmann equation for hard sphere molecules ($\alpha = 1$). We present some results for one dimensional Riemann problem and compare them with the numerical solution obtained by the Monte-Carlo scheme. Let us note that the accuracy of the Monte Carlo solution is improved by performing averages of the solution itself by repeating the calculation several times with different seeds in the random number generator, and averaging the solution over the different runs. Then, we have computed an approximation for different Knudsen numbers, from rarefied regime up to the fluid limit. The solution in the hydrodynamic limit is also compared with the numerical solution of Euler system, which is obtained by Nessyahu-Tadmor scheme [15] using a large number of points (n_x=1600). The initial data is given by

$$\begin{cases} (\rho_l, u_l, T_l) = (1, 0, 1) & \text{if } 0 \le x \le 0.5, \\ (\rho_r, u_r, T_r) = (0.125, 0, 0.25) & \text{if } 0.5 < x \le 1, \end{cases}$$

In Fig. 3 we plot the results obtained in the rarefied regime (k_n=10^{-2}) using the Spectral-PFC scheme and the Time Relaxed Monte Carlo (TRMC)



(1) (2)

(3) (4)

Fig. 3. 3D homogeneous case II: *evolution of the temperature and the entropy for hard sphere molecules (1)-(2) and for Maxwellian molecules (3)-(4).*

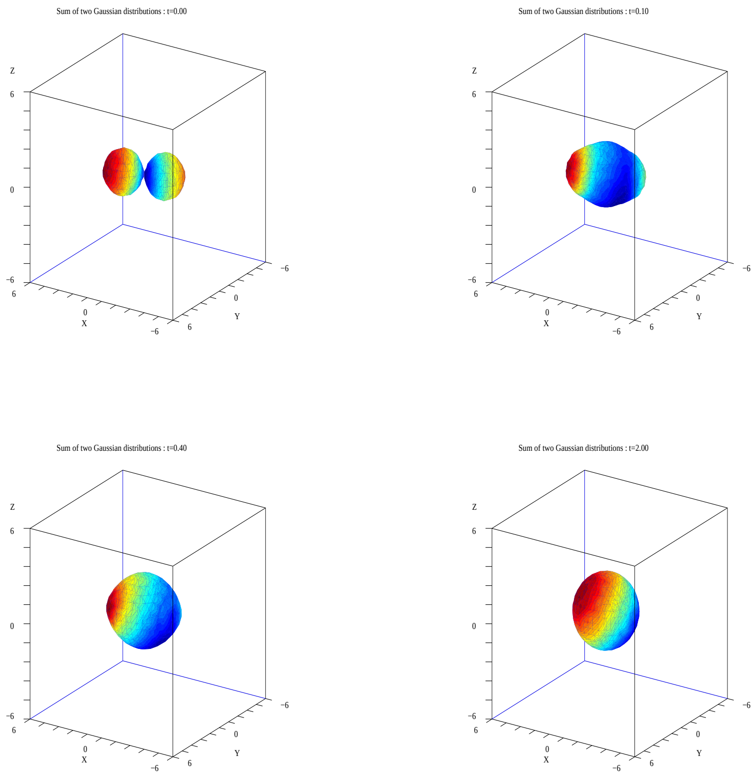


Fig. 4. 3D homogeneous case II: *evolution of the distribution function at time t=0, 0.1, 0.4 and 2 for hard sphere molecules (level set* $f(t, v) = 3.10^{-3}$*)*

method. The TRMC method is used with 100 cells in x containing 100 particles whereas the Spectral-PFC scheme is used with 64 points in x and the size of the velocity grid is 64×64 points for the transport and the total number of modes 32×32. We observe that the two solutions are in this case very comparable even if small oscillations, due to the statistical noise, persist. Concerning the computational time on one processor, the Spectral-PFC scheme is more efficient than Monte Carlo in this situation because the averaging highly increases the computational time (see Table 1). Let us note that in the two cases (Monte-Carlo and Spectral methods), the Time Relaxed scheme allows to use a large variety of Knudsen number ($k_n = 10^{-1}$, 10^{-2}) without increasing the computational cost. Finally, the computational time of the Spectral-PFC scheme can be highly reduced using the parallel algorithm presented before.

We also give the result of the computations close to the Euler limit ($k_n = 10^{-4}$) using 128 space cells for the Spectral-PFC method. In this case, a smaller time step ($\Delta t = 0.001$) is needed to keep good accuracy[3], which increases the computational time, while a small time step for the TRMC method does not influence the numerical solution due to the low order of the Monte-Carlo scheme (see Table 1). For this reason a large time step is used, which explains the lower computational cost of the TRMC scheme.

Finally, the profiles obtained with TRMC and Spectral-PFC methods are reported in Fig. 4. The use of first order scheme for the transport for the TRMC scheme is clearly not sufficient to give accurate results. On the opposite, using a small time step ($\Delta t = 0.001$), an accurate solution is obtained by the Spectral-PFC method, which is much less diffusive.

	TRMC	S-PFC
$k_n = 10^{-1}$	17 mn 25 sec	10 mn 50 sec
k_n =10^{-2}	17 mn 25 sec	10 mn 50 sec
k_n =10^{-4}	17 mn 25 sec	44 mn 20 sec

Table 1. Riemann problem: *the first column represents the value of Knudsen numbers k_n, the second one is the computational time obtained for the TRMC scheme and the third one is the computational time for the third order PFC scheme coupled with the spectral method for the collision operator.*

5.4 Shock profile: stationary solutions.

This test deals with the numerical solution of the non homogeneous $1D \times 3D$ Boltzmann equation for hard sphere molecules ($\alpha = 1$). We present numerical results for one dimensional stationary shock-profiles for different Knudsen number and compare the solution with one obtained by the Monte-Carlo method.

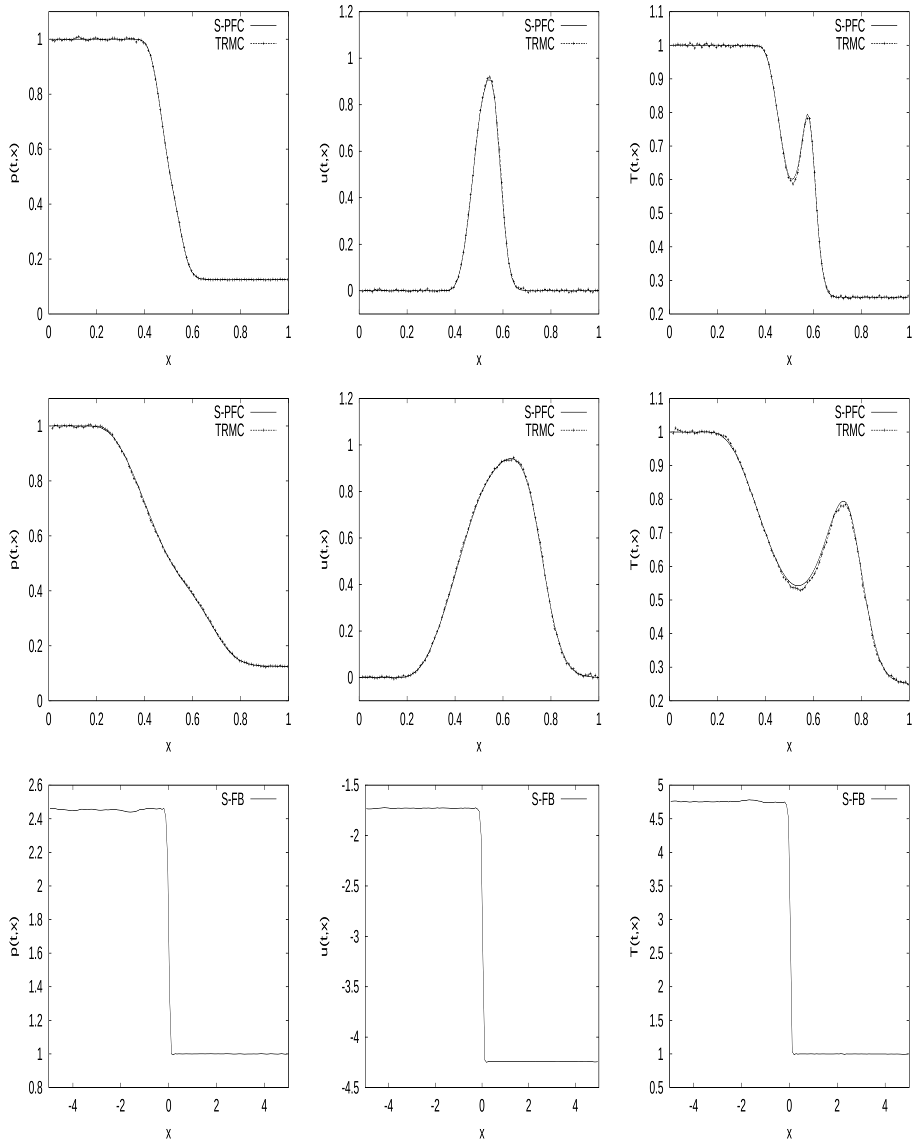
The gas is initially at the upstream equilibrium state in the left half-space and in the downstream equilibrium state in the right-half space. The upstream state are determined from downstream state using the Rankine-Hugoniot relations [28]. In the present calculations, the downstream state is characterized by

$$\rho^r = 1, \quad T^r = 1, \quad M = 2,$$

where M is the Mach number of the shock. The downstream mean velocity is then given by

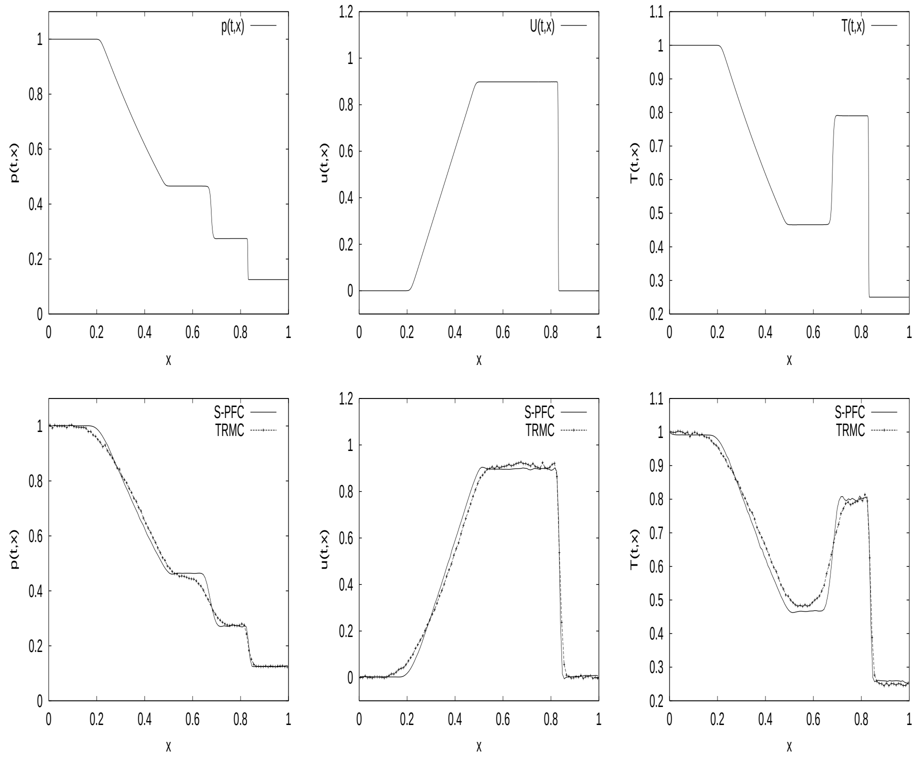
$$(u_x^r, u_y^r) = (-M\,\sqrt{\gamma T}, 0),$$

[3] The degradation of accuracy is typical of Strang splitting when one of the term is stiff (see [13]. Such degradation can be cured using a different approach for time discretization



(1) (2) (3)

Fig. 5. Riemann problem ($k_n = 10^{-2}$): *evolution of (1) the density* ρ*, (2) mean velocity* u *and (3) temperature* T *at* $t = 0.05$, 0.15, 0.20.

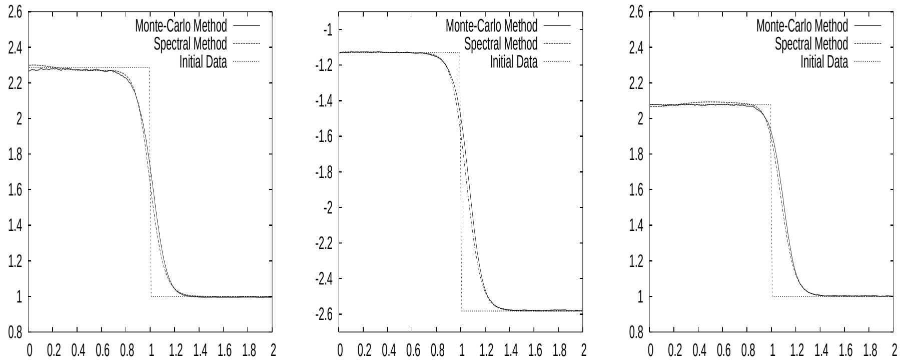


(1) (2) (3)

Fig. 6. Riemann problem ($k_n = 10^{-4}$): *(1) the density ρ, (2) mean velocity u and (3) temperature T at $t = 0.20$ obtained by the central scheme for Euler equations (up) and by Spectral-PFC and TRMC methods for Boltzmann equations.*

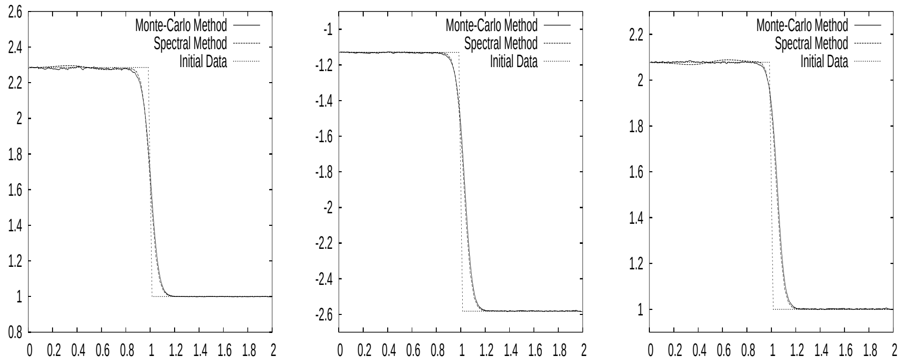
with $\gamma = 5/3$ since we have considered a 3D monoatomic gas in velocity space.

The results of the computation are shown in Fig. 5 and 6. On the one hand, we compute a solution using the Spectral-PFC scheme (128 cells in space and only $32 \times 32 \times 32$ modes in velocity) up to $t = 1.0$, so that the profile is practically stationary. On the other hand, Monte Carlo calculations (TRMC) are performed by time-averaging the numerical solution after time large enough ($t = 2.0$). We observe that there is a good agreement between the TRMC and Spectral-PFC method. However, the TRMC scheme is used with $n_x = 250$ cells in space in order to avoid a too large numerical diffusion. Indeed, with $n_x = 128$, the shock is not well described with this method. In any case, the TRMC is much cheaper in term of computational cost since we are only interested in the stationary solution.



(1) (2) (3)

Fig. 7. Shock profiles ($\epsilon = 1.\ 10^{-1}$): *(1) the density ρ, (2) mean velocity u and (3) temperature T obtained by the Spectral-PFC method and by the TRMC method.*



(1) (2) (3)

Fig. 8. Shock profiles ($\epsilon = 5.\ 10^{-2}$): *(1) the density ρ, (2) mean velocity u and (3) temperature T obtained by the Spectral-PFC method and by the TRMC method.*

6 Conclusion.

In this chapter we present an accurate deterministic method for the numerical approximation of the space non homogeneous, time dependent Boltzmann equation. The method, based on a fractional step approach, couples a Positive and Flux Conservative scheme for the treatment of the transport step with a Fourier spectral method for the collision step.

It possesses a high order of accuracy for this kind of problems. In fact it is second order accurate in time, third order accurate in space, and spectrally accurate in velocity. The high accuracy is evident from the quality of the numerical results that can be obtained with a relatively small number of grid points in velocity domain.

An effective time discretization allows the treatment of problems with a considerable range of mean free path, and the decoupling between the transport and the collision step makes it possible the use of parallel algorithms, which become competitive with state-of-the-art numerical methods for the Boltzmann equation.

The numerical results, and the comparison with other techniques, show the effectiveness of the present method for a wide class of problems.

References

1. Babovsky, H.: On a simulation scheme for the Boltzmann equation. Mathematical Methods in the Applied Sciences **8**, 223–233 (1986)
2. Bird, G.A.: Molecular gas dynamics. Clarendon Press, Oxford (1994)
3. Bobylev, A.V. and Rjasanow, S.: Difference scheme for the Boltzmann equation based on the Fast Fourier Transform. Eur. J. Mech. B/Fluids **16**, 293-306 (1997)
4. Buet, C.: A discrete velocity scheme for the Boltzmann operator of rarefied gas dynamics. Trans. Theo. Stat. Phys. **25**, 33-60 (1996)
5. Canuto, C., Hussaini, M.Y., Quarteroni, A. and Zang, T.A.: Spectral methods in fluid dynamics. Springer Verlag, New York (1988)
6. Cercignani, C.: The Boltzmann equation and its applications. Springer-Verlag, Berlin (1988)
7. Filbet, F. and Pareschi, L.: A numerical method for the accurate solution of the Landau-Fokker-Planck equation in the non homogeneous case. J. Comput. Phys. **179**, 1–26 (2002)
8. Filbet, F., Sonnendrücker, E. and Bertrand, P.: Conservative Numerical schemes for the Vlasov equation. J. Comput. Phys. **172**, 166–187 (2001)
9. Filbet, F. and Sonnendrücker, E.: Comparison of Eulerian Vlasov solvers. Comput. Phys. Communications **151**, 247–266 (2003)
10. Filbet, F. and Russo, G.: High order numerical methods for the space non homogeneous Boltzmann equation. J. Comput. Phys. **186**, 457–480 (2003)
11. Goldstein, D., Sturtevant, B. and Broadwell, J.E.: Investigation of the motion of discrete velocity gases. Rar. Gas. Dynam., Progress in Astronautics e Aeronautics **118** AIAA, Washington (1989)
12. Inamuro, T. and Sturtevant B.: Numerical study of discrete velocity gases. Phys. Fluids A **12**, 2196–2203 (1990)
13. Jin, S.: Runge-Kutta methods for hyperbolic systems with stiff relaxation terms. J.Comput.Phys., **122**, 51–67 (1995)
14. Nanbu, K.: Direct simulation scheme derived from the Boltzmann equation. I. Monocomponent Gases. J. Phys. Soc. Japan **52**, 2042–2049 (1983)
15. Nessyahu, H. and Tadmor, E.: Nonoscillatory central differencing for hyperbolic conservation laws. J. Comput. Phys. **87**, 408–463 (1990)
16. Ohwada, T.: Higher Order Approximation Methods for the Boltzmann Equation. J. Comput. Phys. **139**, 1–14 (1998)
17. Pareschi, L: Computational methods and fast algorithms for Boltzmann equations, Chapter 7, Lecture Notes on the discretization of the Boltzmann equation, ed. N.Bellomo, World Scientific, 46 pagg. (to appear).

18. Pareschi, L. and Perthame, B.: A Fourier spectral method for homogeneous Boltzmann equations. Transp. Theo. Stat. Phys. **25**, 369–383 (1996)
19. Pareschi, L. and Russo, G.: On the stability of spectral methods for the homogeneous Boltzmann equation. Proceedings of the Fifth International Workshop on Mathematical Aspects of Fluid and Plasma Dynamics (Maui, HI, 1998). Transport Theory Statist. Phys. **29**, 431–447 (2000)
20. Pareschi, L. and Russo, G.: Numerical solution of the Boltzmann equation. I. Spectrally accurate approximation of the collision operator. SIAM J. Numer. Anal. **37**, 1217–1245 (2000)
21. Pareschi, L. and Russo, G.: Time Relaxed Monte Carlo methods for the Boltzmann equation. SIAM J. Sci. Comp. **23**, 1253–1273 (2001)
22. Pareschi, L.; Toscani, G. and Villani, C.: Spectral methods for the non cut-off Boltzmann equation and numerical grazing collision limit. Numer. Math. **93**, no. 3, 527–548 (2003)
23. Rogier, F. and Schneider, J.: A direct method for solving the Boltzmann equation. Trans. Theo. Stat. Phys. **23**, 313–338 (1994)
24. Shu, C.-W.: Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu and E. Tadmor (Editor: A. Quarteroni), Lecture Notes in Mathematics, Springer **1697**, 325–432 (1998)
25. Sone, Y., Aoki, K., Takata, S., Sugimoto, H. and Bobylev, A.V: Inappropriateness of the heat-conduction equation for description of a temperature field of a stationary gas in the continuum limit: examination by asymptotic analysis and numerical computation of the Boltzmann equation. Phys. Fluids **8**, 628–638 (1996)
26. Strang, G.: On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**, 506–517 (1968)
27. Villani, C.: A review of mathematical topics in collisional kinetic theory. Handbook of mathematical fluid dynamics **I**, 71–305, North-Holland, Amsterdam (2002)
28. Whitham, G. B.: Linear and nonlinear waves. Wiley Interscience (1974).
29. Wild, E: On Boltzmann's equation in the kinetic theory of gases. Proc. Cambridge Philos. Soc. **47**, 602–609, (1951).