Numerical Solution of the Boltzmann Equation I: Spectrally Accurate Approximation of the Collision Operator*

Lorenzo PARESCHI† and Giovanni RUSSO‡

Abstract

In this paper we show that the use of spectral-Galerkin methods for the approximation of the Boltzmann equation in the velocity space permits to obtain spectrally accurate numerical solutions at a reduced computational cost. We prove that the spectral algorithm preserves the total mass and approximates with infinite-order accuracy momentum and energy. Consistency of the method is also proved and a stability result for a smoothed positive scheme is given. We demonstrate that the Fourier coefficients associated to the collision kernel of the equation have a very simple structure and in some cases can be computed explicitly. Numerical examples for homogeneous test problems in two and three dimensions confirm the advantages of the method.

Key words: Boltzmann equation, Spectral Galerkin methods, Splitting algorithms.

AMS (MOS) subject classifications: 65L60, 65R20, 76P05, 82C40.

1 Introduction

This paper is the first of two papers devoted to the development of numerical schemes for the accurate computation of the Boltzmann equation and related kinetic models. Here we will mainly concentrate on the approximation of the collision operator and hence of the velocity space. In a companion paper we will consider the problem of space and time discretizations showing how to develop under resolved numerical schemes that work with uniform accuracy with respect to the *Knudsen number* [31].

*This work was partially supported by TMR project "Asymptotic Methods in Kinetic Theory", Contract Number ERB FMRX CT97 0157, and partially by CNR, grant # 98.01030.CT01.

†Universitá di Ferrara, Dipartimento di Matematica, Via Machiavelli 35, 44100 Ferrara, Italy. E-mail: `prl@dns.unife.it`

‡Universitá dell'Aquila, Dipartimento di Matematica, Via Vetoio, Loc. Coppito, 67010 L'Aquila, Italy. E-mail: `russo@univaq.it`

The numerical solution of the Boltzmann equation represents a real challenge for numerical methods. This is essentially due to the nonlinearity, to the large number of variables (t for the time, two three-dimensional vectors x for space and v for velocity) and to the five-fold integral that defines the collision operator. Furthermore, this integration has to be handled carefully since it is at the basis of the macroscopic properties of the Boltzmann equation.

For example, if we denote by n the number of parameters which characterize the density with respect to the velocity variables, the computational cost of a conventional deterministic method for the evaluation of the collisional integral is much larger than n^2. Moreover, we have to multiply this cost times the number of points in the physical space.

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 [3] and the modified Monte Carlo method by Nanbu [28]. For a detailed description of such methods we refer to [3, 11, 22, 34].

Probabilistic methods present different advantages. First, the computational cost is strongly reduced and approximatively can be considered of the order of the number of points n. Second, these methods do not need any artificial boundary in the velocity space. In fact, particles can have any velocity and thus the discretization points are always well defined independently of the physical problem.

At variance, in addition to the computational complexity, a major problem associated with deterministic methods that use a fixed discretization in the velocity domain is that the velocity space is approximated by a finite region. Hence, a large region with a huge number of discretization points is required in problems with very high Mach numbers, which greatly increase the computing effort.

On the other hand, particle methods yield low accurate and fluctuating results with respect to finite differences or finite volume methods and the convergence in general is quite low ($O(1/\sqrt{n})$). In particular, the convergence rate is very slow when the Knudsen number is small. For continuum flows such methods require so many particles that are not competitive with the corresponding finite volumes schemes for the fluid-dynamic equations [35, 13]. For this reason matching techniques with a Euler or Navier-Stokes solver are used for flows with large spatial variations in Knudsen number [26, 7]. In these approaches, the domain is subdivided into a *Boltzmann domain*, and an *aerodynamic domain*. In the first domain the Boltzmann equation is solved, and in the second domain Euler or Navier-Stokes equations are solved. Suitable matching condition are used to couple the two regions. One difficulty of this approach is the detection of the two regions. It is not trivial, in fact, to establish where the aerodynamic description is sufficient, and where it is necessary to use the full Boltzmann equation.

Similar problems are faced when explicit time discretizations are used with deterministic schemes. Some important partial progress in this direction have been recently presented in [16]. The possibility to extend the ideas developed in [16] by using Monte-Carlo methods has been recently investigated in [8, 32].

In order to avoid fluctuations several authors have proposed different deterministic computational approaches especially designed for the numerical solution of problems where particle methods become expensive in order to get sufficiently accurate results. Unfortunately, the largest part of them work well only for the linearized Boltzmann equation [37] or for some particular models [5, 13, 15, 21]. Other methods have been developed only for the homogeneous equation [27, 25] and have not been extensively tested. We also quote methods based on hybrid discretizations that combine finite difference techniques in space with some Monte-Carlo evaluation of the collision operator [38, 12].

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 [6, 17, 23, 36] 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. On the other hand, even if conservation properties are not imposed from the beginning, an accurate scheme would provide an accurate approximation of the conserved quantities.

From the above discussion it would be clearly desirable to possess a deterministic scheme for the Boltzmann equation with the following characteristics

- Greater accuracy in velocity (and therefore in the conserved quantities).
- Lower computational cost.
- Flexibility in the velocity range.
- Efficient time discretization even for small Knudsen numbers.

In this paper we mainly deal with the first two problems, with a particular emphasis on accuracy. The last two points will be the subject of a future paper [31].

Our starting point is a new method for the numerical solution of the homogeneous Boltzmann equation based on a Fourier spectral approximation of the collision operator recently presented by Pareschi and Perthame [29]. They 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.

Here we go further, by showing that in most interesting cases this approach leads to an explicit formula that can be computed directly without further approximations. We demonstrate that the method preserves the total mass and approximates with spectral accuracy momentum and energy. A proof of consistency and spectral accuracy for the numerical discretization is presented. A positive scheme is derived from the previous scheme, by suitable smoothing. A stability result for such smoothed positive scheme is given.

The rest of the paper is organized as follows. In Section 2 we recall some of the basic properties of the Boltzmann equation. After the usual splitting in time of the relaxation and the convection parts, in Section 3 we present the spectral projection of the homogeneous equation. Section 4 is devoted to the analysis of the resulting kernel modes. We show that for the variable hard sphere model (VHS) the computation of such modes reduces to a single integration that in some cases (hard spheres, Maxwell molecules) can be computed explicitly. The main theoretical properties of the methods, such as conservations, consistence, spectral accuracy and stability are studied in Section 5. Finally in Section 6 we present some numerical results for two-dimensional and three-dimensional homogeneous test problems that confirm our theoretical analysis. We conclude the paper with some remarks in Section 7.

2 The governing equation

In absence of external forces the time evolution of a single component mono atomic gas is governed by the Boltzmann equation (Cf.[10])

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\epsilon} Q(f,f), \quad x, v \in \mathbb{R}^3 \tag{1}$$

where $f = f(x,v,t)$ is a non negative function describing the time evolution of the distribution of particles which move with velocity v in the position x at time $t > 0$. The number $\epsilon > 0$ 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

$$Q(f,f)(v) = \int_{\mathbb{R}^3} \int_{S^2} B(|v - v_1|, \theta)[f(v')f(v_1') - f(v)f(v_1)]\, d\omega\, dv_1. \tag{2}$$

In the above integral v' and v_1' are the velocities after the collision of two particles which had the velocities v and v_1 before the encounter. The deflection angle θ is the angle between $v - v_1$ and $v' - v_1'$.

Here the post-collision velocities are parametrized by

$$v' = \frac{1}{2}(v + v_1 + |v - v_1|\omega), \quad v_1' = \frac{1}{2}(v + v_1 - |v - v_1|\omega), \tag{3}$$

where ω is a unit vector of the sphere S^2

$$S^2 = \left\{ \omega \in I\!R^3, |\omega|^2 = 1 \right\}. \tag{4}$$

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

$$B(|v - v_1|, \theta) = b_\alpha(\theta)|v - v_1|^\alpha, \tag{5}$$

where $\alpha = (k-5)/(k-1)$. In particular, we will consider the variable hard sphere(VHS) model [3] 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.

Boltzmann's collision operator has the fundamental properties of conserving mass, momentum and energy

$$\int_{I\!R^3} Q(f,f) \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv = 0, \tag{6}$$

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

$$\int_{I\!R^3} Q(f,f) \log(f) dv \le 0. \tag{7}$$

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

$$M(\rho, u, T)(v) = \frac{\rho}{(2\pi T)^{3/2}} \exp\left(-\frac{|u-v|^2}{2T}\right), \tag{8}$$

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

$$\rho = \int_{I\!R^3} f(v) dv, \quad u = \frac{1}{\rho} \int_{I\!R^3} v f(v) dv, \quad T = \frac{1}{3\rho} \int_{I\!R^3} |u - v|^2 f(v) dv. \tag{9}$$

Most numerical schemes for equation (1) are derived after a splitting in time of free molecular flow and relaxation due to collisions. In practice, in order to compute numerically the solution to (1), we solve in a time interval Δt, first a homogeneous Boltzmann equation

$$\frac{\partial f}{\partial t} = \frac{1}{\epsilon} Q(f,f), \tag{10}$$

and then a transport equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = 0. \tag{11}$$

The overall accuracy of this simple splitting is first order in time. Second order generalization of this method that work uniformly with respect to ϵ and that can be applied to the Boltzmann equation will be presented in a companion paper by the authors [31].

Since the main problems in the approximation to (1) are represented by the discretization of the collision operator and by the stiffness of this term for small values of ϵ, it is clear that after the splitting the most difficult part relies in the approximation of the collision step (10).

This is the reason why in this paper we will mainly concentrate on the approximation to the homogeneous equation (10).

3 Spectral projection of the Boltzmann equation

We now search for a discretization by spectral methods of the homogeneous Boltzmann equation

$$\frac{\partial f}{\partial t} + f\, L(f) = Q^+(f,f) \tag{12}$$

supplemented with the initial condition

$$f(x,v,t=0) = f_0(x,v). \tag{13}$$

In (12), to keep notation simple, we have fixed $\epsilon = 1$ and we have used the fact that the collision operator can be splitted into the gain and loss parts $Q(f,f) = Q^+(f,f) - f\,L(f)$.

A simple change of variables permits to write

$$Q^+(f,f) = \int_{\mathbb{R}^3}\int_{S^2} B(|g|,\theta) f(v')f(v_1')\,d\omega\,dg, \tag{14}$$

$$L(f) = \int_{\mathbb{R}^3}\int_{S^2} B(|g|,\theta) f(v-g)\,d\omega\,dg, \tag{15}$$

where $g = v - v_1$ and then

$$v' = v - \frac{1}{2}(g - |g|\omega), \quad v_1' = v - \frac{1}{2}(g + |g|\omega). \tag{16}$$

In (12) the definition domain of the distribution function f is the velocity space $\mathbb{R}^3$ whereas the variable x acts only as a parameter and it will be omitted in the sequel. During the evolution process, since the main physical properties depend only on the collision operator $Q(f,f)$, the solution to (12) preserves local mass, momentum and energy and satisfies Boltzmann's H-theorem.

At variance to DSMC methods, where no artificial boundary in velocity space is needed, when constructing a numerical solution for (12) by means of a velocity discretization of $f(v)$ a classical problem that one has to deal with is the fact that any numerical approximation to (12) requires the given initial density f_0 in a compact support whereas the steady state solution is a local Maxwellian like (8), that clearly does not have this property. As a consequence of this, relations (6) are no longer verified in time within a bounded domain. A general way to avoid this problem consists in modifying the collision mechanics in order to satisfy the original conservation laws in a bounded domain. However, it is clear that even with this modification the accuracy of the solution with respect to the original equation is guaranteed only if the size of the velocity domain is large enough.

In [29] the authors considered a different approach to the problem. They observe that if a distribution function f has compact support, $\mathrm{Supp}(f(v)) \subset \mathcal{B}(0,R)$, where $\mathcal{B}(0,R)$ is the ball of radius R centered in the origin, then by conservation of energy and momentum the collision operator satisfies the following

Proposition 3.1 *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)

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

with $v', v'_1, v-g \in \mathcal{B}(0,(2+\sqrt{2})R)$.

Hence in order to write a spectral approximation to (12) we can consider the distribution function $f(v)$ restricted on the cube $[-T,T]^3$ with $T \geq (2+\sqrt{2})R$, assuming $f(v)=0$ on $[-T,T]^3 \setminus \mathcal{B}(0,R)$, and extend it by periodicity to a periodic function on $[-T,T]^3$.

The previous proposition when $\mathrm{Supp}(f) \in \mathcal{B}(0,R)$ guarantees the absence of intersection between periods of the distribution function and thus permits to develop spectral methods on the cube without aliasing errors. However, since the support of f increases with time in practice we should choose T large enough in order to minimize the errors due to aliasing.

Note that by virtue of the periodicity property of the function it is enough to take $T \geq (3+\sqrt{2})R/2$ to prevent intersections of the regions where f is different from zero (see figure 1).

Remark 3.1 The choice of the integration domain $\mathcal{B}(0,2R)$ for g guarantees the equality ii). However (see [11]) it is important to note that for any $M > 0$ the collision operator with cut-off $Q_M(f,f)$, where g is integrated on the ball $\mathcal{B}(0,M)$, satisfies the conservation laws and the Boltzmann H-theorem we saw in Section 2. Clearly, the choice of $M < 2R$ will not satisfy ii), and gives rise to different approximations of $Q(f,f)$. Aliasing errors in this general case are avoided if $T \geq (M + (1+\sqrt{2})R)/2$.

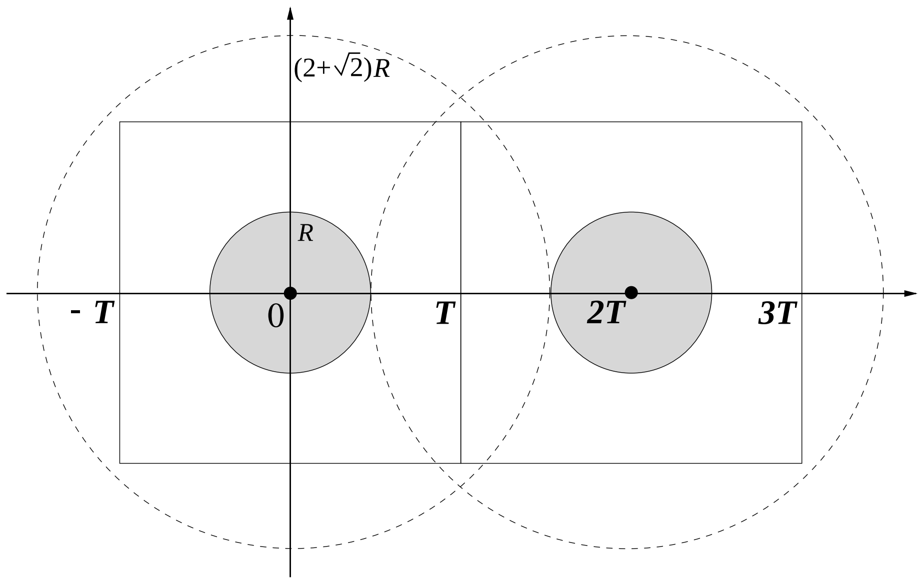

Figure 1: Restriction of the distribution function on the periodic box $[-T,T]\times[-T,T]$ in two dimensions.

To simplify the notation let us take $T=\pi$ and hence $R=\lambda\pi$ with $\lambda = 2/(3+\sqrt{2})$. Hereafter, we use just one index to denote the three-dimensional sums with respect to the vector $k=(k_1,k_2,k_3)\in \mathbb{Z}^3$, hence we set

$$\sum_{k=-N}^{N} = \sum_{k_1,k_2,k_3=-N}^{N}.$$

The approximate function f_N is represented as the truncated Fourier series

$$f_N(v) = \sum_{k=-N}^{N} \hat{f}_k e^{ik\cdot v}, \tag{17}$$

$$\hat{f}_k = \frac{1}{(2\pi)^3}\int_{[-\pi,\pi]^3} f(v)e^{-ik\cdot v}\,dv. \tag{18}$$

In a Fourier-Galerkin method the fundamental unknowns are the coefficients $\hat{f}_k$, $k = -N,\dots,N$. We obtain a set of Ode's for the $\hat{f}_k$ coefficients by requiring that the residual of (12) be orthogonal to all trigonometric polynomials of degree $\leq N$ [9, 18].

Hence for $k=-N,\dots,N$

$$\int_{[-\pi,\pi]^3}\left(\frac{\partial f_N}{\partial t} + f_N\,L(f_N) - Q^+(f_N,f_N)\right)e^{-ik\cdot v}\,dv = 0. \tag{19}$$

By substituting expression (17) in (15) and (14) 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

$$\hat{B}(l,m) = \int_{\mathcal{B}(0,2\lambda\pi)} \int_{S^2} B(|g|,\theta) e^{-ig\cdot\frac{(l+m)}{2} - i|g|\omega\cdot\frac{(m-l)}{2}}\, d\omega\, dg. \tag{20}$$

Using the orthogonality property we get from (19)

$$\frac{\partial \hat{f}_k}{\partial t} + \sum_{\substack{l+m=k \\ l,m=-N}}^{N} \hat{f}_l\, \hat{f}_m \hat{B}(m,m) = \sum_{\substack{l+m=k \\ l,m=-N}}^{N} \hat{f}_l\, \hat{f}_m \hat{B}(l,m), \quad k = -N, \dots, N, \tag{21}$$

with the initial condition

$$\hat{f}_k(0) = \frac{1}{(2\pi)^3} \int_{[-\pi,\pi]^3} f_0(v) e^{-ik\cdot v}\, dv. \tag{22}$$

Now it is very remarkable that (20) is a scalar quantity completely independent on the function f_N and on the argument v, depending just on the particular kernel structure. In addition it can be proved that [29],

Proposition 3.2 *Let $\hat{B}(l,m)$ be defined by (20) then*

i) $\hat{B}(l,m)$ is a function of $|l-m|$, $|l+m|$ and $(l-m)\cdot(l+m)$.

ii) If the kernel $B(|g|,\theta) = C_\alpha |g|^\alpha$ then $\hat{B}(l,m)$ is a function of $|l-m|$, $|l+m|$.

The previous proposition characterizes the number of kernel modes $\hat{B}(l,m)$ that need to be computed and stored for the implementation of the method. Note that in the situation *ii*), the most interesting for applications, the kernel modes depend just on two parameters and can be stored in a two-dimensional matrix. The dimension of the matrix will be $\mathcal{O}(N^6)$, however, thanks to symmetry considerations, this number is strongly reduced in practice.

Finally we can rewrite scheme (21) in the form, $k = -N, \dots, N$

$$\frac{\partial \hat{f}_k}{\partial t} + \sum_{m=-N}^{N} \hat{f}_{k-m}\, \hat{f}_m \hat{B}(m,m) = \sum_{m=-N}^{N} \hat{f}_{k-m}\, \hat{f}_m \hat{B}(k-m,m). \tag{23}$$

In the previous expression we assume that the Fourier coefficients are extended to zero for $|k_j| > N$, $j = 1, 2, 3$. The evaluation of (23) requires exactly $\mathcal{O}(N^6)$ operations. We emphasize that the usual cost for a method based on N^3 parameters for f in the velocity space is $\mathcal{O}(n_a N^6)$ where n_a is the numbers of angle discretizations. Thus the straightforward evaluation of (23) it is less expensive than a usual discrete-velocity algorithm.

Remark 3.2 *The first sum on the left hand side is a convolution sum and thus transform methods [9, 18] allow this term to be evaluated only in $\mathcal{O}(N^3 \log N)$ operations. Hence the most expensive part of the computation is represented by the sum on the right hand side which in general cannot be evaluated with fast algorithms.*

4 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 Eq.(20). In the VHS model, the kernel does not depend on the angle θ: $B = C_\alpha |g|^\alpha$. One has

$$\hat{B}(l,m) = C_\alpha \int_{\mathcal{B}(0,2\lambda\pi)} |g|^\alpha \exp\left(-ig \cdot \frac{l+m}{2}\right) I_2(|g|, l-m)\, dg \tag{24}$$

where

$$I_2(|g|, l-m) = \int_{S^2} \exp\left(i|g|\omega \cdot \frac{l-m}{2}\right) d\omega. \tag{25}$$

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.

4.1 3D case

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

$$\begin{aligned} I_2(|g|, l-m) &= \int_{S^2} e^{iq\cdot\omega}\, d\omega = 2\pi \int_0^\pi e^{i|q|\cos\theta} \sin\theta\, d\theta \\ &= 2\pi \int_{-1}^{1} e^{i|q|\mu}\, d\mu = 4\pi\, \mathrm{Sinc}(|q|) \\ &= 4\pi\, \mathrm{Sinc}\left(\frac{|g||l-m|}{2}\right) \end{aligned} \tag{26}$$

where

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

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

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

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

$$\begin{aligned}
\hat{B}(l,m) &= C_\alpha 8\pi^2 \int_0^{2\pi\lambda} \rho^{2+\alpha} \operatorname{Sinc}(|l-m|\rho/2)\, d\rho \int_0^\pi \exp(-i|p|\rho\cos\theta)\sin\theta\, d\theta \\
&= C_\alpha 16\pi^2 \int_0^{2\pi\lambda} \rho^{2+\alpha} \operatorname{Sinc}(|l-m|\rho/2) \operatorname{Sinc}(|l+m|\rho/2)\, d\rho. \qquad (27)
\end{aligned}$$

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

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

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

$$F_\alpha(\xi,\eta) = \int_0^1 r^{2+\alpha} \operatorname{Sinc}(\xi r) \operatorname{Sinc}(\eta r)\, dr. \qquad (28)$$

From (28) it is easy to prove the following

Proposition 4.1 *Let $F_\alpha(\xi,\eta)$ be defined by (28) 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)

$$F_0(\xi,\eta) = \frac{p\sin(q) - q\sin(p)}{2\xi\eta pq} \qquad (29)$$

$$F_1(\xi,\eta) = \frac{p^2(q\sin(q)+\cos(q)) - q^2(p\sin(p)+\cos(p)) - 4\xi\eta}{2\xi\eta p^2 q^2} \qquad (30)$$

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

Note that, since the five-fold integral (20) which defines the $\hat{B}(l,m)$ has been reduced to a one-dimensional integral (28), 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.

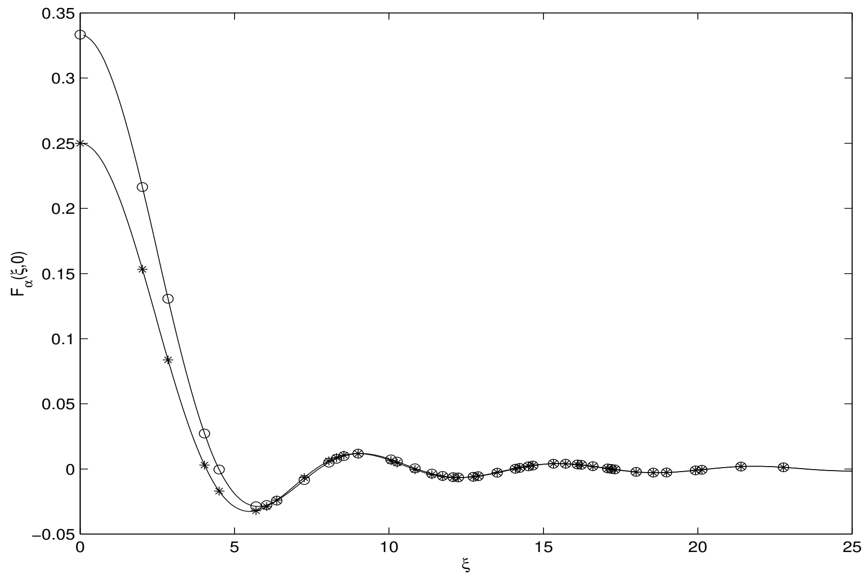

Figure 2: Distribution of the kernel modes $\hat{B}(l,m) = F_\alpha(\xi,\eta)$ on the line defined by $l = m$ ($\eta = 0$) for $N = 8$ and $\alpha = 0\,(\circ)$ and $\alpha = 1\,(*)$.

4.2 2D case

For the computation in 2D we start from Eqs.(24,25). We will consider only the VHS model.

In this case it is

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

where $r = |q| = |g||l-m|/2$, and J_0 is the Bessel function of order 0. By inserting the result in the expression (24) for $\hat{B}(l,m)$, one has

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

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

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

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 4.1 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.

5 Properties of the spectral method

In this section we will analyze in detail the main properties of the spectral approach presented in Section 3, with a particular attention to the physical properties and to the accuracy and stability properties of the spectral method.

Let us first set up the mathematical framework of our analysis. For any $t \geq 0$, $f_N(v,t)$ is a trigonometric polynomial of degree N in v, i.e. $f_N(t) \in I\!P^N$ where

$$I\!P^N = span\left\{ e^{ik\cdot v} \mid -N \leq k_j \leq N,\, j = 1,2,3 \right\}.$$

Moreover, let $\mathcal{P}_N : L^2([-\pi,\pi]^3) \to I\!P^N$ be the orthogonal projection upon $I\!P^N$ in the inner product of $L^2([-\pi,\pi]^3)$ (see (19)):

$$< f - \mathcal{P}_N f, \phi >= 0, \qquad \forall\, \phi \in I\!P^N.$$

We denote the L^2-norm by

$$||f||_2 = (< f, f >)^{1/2}.$$

With this definition $\mathcal{P}_N f = f_N$, where f_N is the truncated Fourier series of f (17). Since the operator $\mathcal{P}_N$ is self-adjoint the following property hold

$$< \mathcal{P}_N f, \varphi >=< f, \mathcal{P}_N \varphi >=< \mathcal{P}_N f, \mathcal{P}_N \varphi > \quad \forall\, f,\, \varphi \in L^2([-\pi,\pi]^3). \tag{34}$$

We will also define the smoothing operator $\mathcal{S} : I\!P^N \to I\!P^N$ for the truncated Fourier series which is defined by a multiplication of each Fourier coefficient by a factor σ_k

$$\mathcal{S}\, f_N(v) = f_N^s = \sum_{k=-N}^{N} \sigma_k \hat{f}_k e^{ik\cdot v}, \tag{35}$$

where $\sigma_k = \sigma_{k_1}\sigma_{k_2}\sigma_{k_3}$ are required to be real non-negative numbers such that $\sigma_{k_j} = \sigma_{-k_j}$, $\sigma_0 = 1$ and $\sigma_{|k_j|} \leq \sigma_{|k_j-1|}, j = 1,2,3$.

5.1 Approximation by truncated Fourier series

First we prove some approximation properties of the projection operator $\mathcal{P}_N$, in particular those concerning positivity of the density function and approximation of the macroscopic quantities. Let us remark that, in general, when we approximate a non negative function by a partial sum of its Fourier series, that partial sum may be negative. The results are summarized in the following proposition

Proposition 5.1 *Let $f \in L^2([-\pi,\pi]^3)$, $f \geq 0, \forall\, v$, and let us define*

$$\begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix} := \int_{[-\pi,\pi]^3} f \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv. \tag{36}$$

Then we have:

i) If the factors σ_k are such that

$$1 + 2\sum_{k=1}^{N} \sigma_k \cos(k \cdot v) \geq 0, \tag{37}$$

then

$$f_N^s(v) \geq 0, \quad \forall\, v.$$

ii) The moments of f_N can be defined equivalently as

$$\begin{pmatrix} \rho_N \\ \rho u_N \\ \rho e_N \end{pmatrix} := \int_{[-\pi,\pi]^3} f_N \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv = \int_{[-\pi,\pi]^3} f \begin{pmatrix} 1 \\ v_N \\ (v^2)_N \end{pmatrix} dv =$$

$$\int_{[-\pi,\pi]^3} f_N \begin{pmatrix} 1 \\ v_N \\ (v^2)_N \end{pmatrix} dv = (2\pi)^3 \sum_{k=-N}^{N} \hat{f}_k \begin{pmatrix} \delta_{k0} \\ \hat{v}_k \\ \hat{(v^2)}_k \end{pmatrix},$$

where $v_N = \mathcal{P}_N v$, $(v^2)_N = \mathcal{P}_N v^2$, δ_{k0} is the Kronecker delta, and $\hat{v}_k$ and $\hat{(v^2)}_k$ are the Fourier coefficients of v and v^2.

iii) The following relations hold

$$\rho = \rho_N,$$

$$|\rho u - \rho u_N| \leq \frac{C_1}{N^{1/2}} ||f||_2,$$

$$|\rho e - \rho e_N| \leq \frac{C_2}{N^{3/2}} ||f||_2.$$

Proof
Note that the smoothed series (35) can be represented in integral form, as [9]

$$f_N^s(v) = \frac{1}{(2\pi)^3}\int_{[-\pi,\pi]^3} K_N(v-w)f(w)\,dw,$$

where the kernel K_N is given by

$$K_N(v) = 1 + 2\sum_{k=1}^{N}\sigma_k\cos(k\cdot v).$$

From the positivity of K_N follows the positivity of f_N^s and hence property i).

Property ii) is a simple consequence of the fact that the projection operator is self-adjoint and follows choosing $\varphi(v) = 1, v, |v|^2$ in (34). The last identity can be obtained by direct substitution of the truncated Fourier series f_N and using the properties that $\hat{f}_k = \bar{\hat{f}}_{-k}$, where $\bar{\hat{f}}_{-k}$ denotes the complex conjugate of $\hat{f}_k$, and the following relations:

$$\delta_{k0} = \frac{1}{(2\pi)^3}\int_{[-\pi,\pi]^3} e^{ik\cdot v}\,dv,$$

$$(\hat{v^2})_k = \frac{1}{(2\pi)^3}\int_{[-\pi,\pi]^3} v^2 e^{-ik\cdot v}\,dv,$$

$$\hat{v}_k = \frac{1}{(2\pi)^3}\int_{[-\pi,\pi]^3} v e^{-ik\cdot v}\,dv.$$

The first equality in iii) is a consequence of

$$\rho_N = (2\pi)^3\hat{f}_0 = \int_{[-\pi,\pi]^3} f(v)\,dv = \rho.$$

The estimates for ρu_N and ρe_N are derived observing that for each $\varphi \in L^2([-\pi,\pi]^3)$, using Schwartz inequality, we have

$$| < f,\varphi > - < f,\varphi_N > | \leq < f, |\varphi - \varphi_N| > \leq ||f||_2\,||\varphi-\varphi_N||_2.$$

Now, by direct computation we obtain $(\hat{v_j})_0 = 0$, $j = 1,2,3$ and $(\hat{v^2})_0 = \pi^2$ whereas for $k \neq 0$ we have

$$(\hat{v_j})_k = -\frac{i}{(2\pi)}\prod_{\substack{l=1\\l\neq j}}^{3}\delta_{k_l 0}\int_{[-\pi,\pi]} v_j\sin(k_j v_j)\,dv_j = -i\prod_{\substack{l=1\\l\neq j}}^{3}\delta_{k_l 0}\frac{(-1)^{k_j}}{k_j}, \quad j = 1,2,3, \tag{38}$$

and

$$(\hat{v^2})_k = \frac{1}{(2\pi)}\sum_{j=1}^{3}\prod_{\substack{l=1\\l\neq j}}^{3}\delta_{k_l 0}\int_{[-\pi,\pi]}(v_j)^2\cos(k_j v_j)\,dv_j = 2\sum_{j=1}^{3}\prod_{\substack{l=1\\l\neq j}}^{3}\delta_{k_l 0}\frac{(-1)^{k_j}}{k_j^2}. \tag{39}$$

Using Parseval's identity and (38)-(39) we obtain the following estimates

$$||v - v_N||_2 \le \frac{C_1}{N^{1/2}},$$

$$||v^2 - (v^2)_N||_2 \le \frac{C_2}{N^{3/2}},$$

the conclusion follows taking $\varphi(v) = v,\, v^2$.

■

Remark 5.1

- *The positivity requirement i) can be satisfied using the factors [9]*

$$\sigma_{k_j} = (1 - \frac{|k_j|}{N}), \quad j = 1, 2, 3. \tag{40}$$

  In fact, these smoothing factors correspond to a nonnegative kernel, the so called Fejer's kernel, K_N given by

$$K_N(v) = 1 + 2\sum_{k=1}^{N} \sigma_k \cos(k \cdot v) = \frac{1}{N^3}\prod_{i=1}^{3}\left(\frac{\sin(Nv_i/2)}{\sin(v_i/2)}\right)^2 \ge 0.$$

  This is also equivalent to replace the truncated Fourier series by the arithmetic means, or Cesaro sums, of the truncated series. However, Fejer's kernel produce a heavy smearing of the function near a singularity point. In most applications it is desirable to have a sharper representation of the function by using different smoothing at the expense of retaining some oscillations or small negative values.

- *Since the smoothed projection $\mathcal{SP}_N = \mathcal{P}_N^s$ satisfies*

$$< \mathcal{P}_N^s f, \varphi > = < f, \mathcal{P}_N^s \varphi > = < \mathcal{P}_N f, \mathcal{P}_N^s \varphi > \quad \forall\ f,\ \varphi \in L^2([-\pi, \pi]^3), \tag{41}$$

  by using the properties of σ_k, the analogue of points ii) and iii) can be proved for f_N^s.

- *From ii) and (38) the j-component of the momentum of f_N depends only on the imaginary part of $\hat{f}_k$ and only on the N Fourier coefficients on the axis $k_l = 0, l \neq j$. Similarly from ii) and (39) the energy of f_N depends only on the real part of f_k and only on the $3N$ Fourier coefficients on the three orthogonal axes.*

- *The estimates given in iii) can be strongly improved if f is smooth. If $f \in H_p^r([-\pi, \pi]^3)$, where $r \ge 0$ is an integer and $H_p^r([-\pi, \pi]^3)$ is the subspace of the Sobolev space*

$H^r([-\pi,\pi]^3)$, which consists of periodic functions [9], for each $\varphi \in L^2([-\pi,\pi]^3)$ we have

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

where $||\cdot||_{H_p^r}$ denotes the norm in $H_p^r([-\pi,\pi]^3)$. This inequality shows that the projection error on the moments decay faster than algebraically when the solution is infinitely smooth.

5.2 Consistency and spectral accuracy

Now we can discuss the main properties of the semi-discrete scheme defined by (21). For simplicity we will restrict our discussion to the VHS model introduced previously. We will denote by $Q^\lambda(f,f)$ the Boltzmann operator with cut-off on the relative velocity on $\mathcal{B}(0,2\lambda\pi)$ and, as usual, we denote by $Q_N^\lambda(f_N) = \mathcal{P}_N Q^\lambda(f_N, f_N)$. In order to prove a consistency result for the method we need the following

Lemma 5.1 *Let $f, g \in L^2([-\pi,\pi]^3)$, then*

i)

$$Q^\lambda(f_N, g_N) \in I\!P_{2N},$$

ii)

$$||Q^\lambda(f,g)||_2 \leq C||f||_2||g||_2.$$

Proof

The first statement follows immediately from the representation formula

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

$$= \sum_{k=-2N}^{2N} \hat{Q}_k^N e^{ik\cdot v},$$

where

$$\hat{Q}_k^N = \sum_{\substack{l+m=k\\ l,m=-N}}^{N} \hat{f}_l \hat{g}_m \hat{D}(l,m),$$

and $\hat{D}(l,m) = \hat{B}(l,m) - \hat{B}(m,m)$.

The point *ii*) for the gain part of the collision operator is a consequence of the estimates in [20, 24]

$$||Q^{\lambda,+}(f,g)||_2 \leq C_1||f||_2||g||_1 \leq (2\pi)^{3/2}C_1||f||_2||g||_2.$$

The corresponding result for the loss part can be computed directely observing that $L^\lambda(f) = f *_q B^\lambda$ where $*_q$ denotes the convolution operation with respect to q and B^λ is the VHS kernel with cut-off over the relative velocity q on the ball $\mathcal{B}(0, 2\lambda\pi)$.

Hence

$$||f(g *_q B^\lambda)||_2 \leq ||f||_2 ||g *_q B^\lambda||_\infty \leq ||f||_2 ||B^\lambda||_\infty ||g||_1 \leq C_2 ||f||_2 ||g||_2,$$

with $C_2 = (2\pi)^{3/2} ||B^\lambda||_\infty = (2\pi)^{3/2} C_\alpha (2\lambda\pi)^\alpha$.

■

Remark 5.2 *In the previous proposition the constant C depends on the size of the domain of integration in v. Thus the estimate becomes meaningless when the integration domain is the whole $I\!R^3$ space as in the original equation. This drawback does not affect us, since here we are interested in proving the spectral accuracy of the method when the integration domain is fixed.*

Then we observe that the method defined by equation (21) can be written in equivalent form as

$$\frac{\partial f_N}{\partial t} = Q_N^\lambda(f_N) \tag{42}$$

with the initial condition

$$f_N(v, t = 0) = f_{0,N}(v), \tag{43}$$

where, to simplify the notation, we use $Q^\lambda(f)$ instead of $Q^\lambda(f, f)$. We point out that because of the periodicity assumption on f, and hence on $Q^\lambda(f)$, the collision operator $Q^\lambda(f)$ preserves in time the mass contained in the period. On the contrary, momentum and energy are not preserved in time.

From Proposition 5.1 it is also clear that the projected collision operator $Q_N^\lambda(f_N)$ will preserve the mass in time. This can be also derived directly from the properties of the kernel modes, in fact, from $\rho = (2\pi)^3 \hat{f}_0$ we have

$$\frac{d\hat{f}_0}{dt} = \sum_{m=-N}^{N} \hat{f}_{-m} \hat{f}_m (\hat{B}(-m, m) - \hat{B}(m, m)) = 0,$$

since $\hat{B}(-m, m) = \hat{B}(m, m)$.

Next we state the consistency in the L^2-norm for the approximation of the collision operator $Q^\lambda(f)$ with $Q_N^\lambda(f_N)$,

Theorem 5.1 *Let $f \in L^2([-\pi, \pi]^3)$, then*

$$||Q^\lambda(f) - Q_N^\lambda(f_N)||_2 \leq C \left(||f - f_N||_2 + \frac{||Q^\lambda(f_N)||_{H_p^r}}{N^r} \right), \quad \forall r \geq 0, \tag{44}$$

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

Proof
First, we can split the error in two parts

$$||Q^\lambda(f) - Q_N^\lambda(f_N)||_2 \le ||Q^\lambda(f) - Q^\lambda(f_N)||_2 + ||Q^\lambda(f_N) - Q_N^\lambda(f_N)||_2.$$

Now from Lemma 5.1, $Q^\lambda(f_N) \in I\!P_{2N}$ and hence $Q^\lambda(f_N)$ is periodic and infinitely smooth together with all its derivatives thus [9]

$$||Q^\lambda(f_N) - Q_N^\lambda(f_N)||_2 \le \frac{C}{N^r}||Q^\lambda(f_N)||_{H_p^r}, \quad \forall r \ge 0. \tag{45}$$

By application of Lemma 5.1 and from the identity [11]

$$Q^\lambda(f) - Q^\lambda(g) = Q^\lambda(f+g, f-g),$$

(which is a direct consequence of the bilinearity of Q^λ), we have

$$\begin{aligned}||Q^\lambda(f) - Q^\lambda(f_N)||_2 = ||Q^\lambda(f+f_N, f-f_N)||_2 &\le C_1||f+f_N||_2||f-f_N||_2 \\ &\le 2C_1||f||_2||f-f_N||_2.\end{aligned}$$

This concludes the proof.

■

Remark 5.3 *From (45) it follows that, except for the projection errors on the initial data, the variations of momentum and energy introduced by the semi-discrete scheme are spectrally small and hence the observed variations with respect to the projected moments are mainly due to the aliasing of periods. In fact, using (45), from Schwarz inequality we have*

$$| < Q^\lambda(f_N), \varphi > - < Q_N^\lambda(f_N), \varphi > | \le \frac{C}{N^r}||\varphi||_2||Q^\lambda(f_N)||_{H_p^r}. \tag{46}$$

The estimates on the conservation laws can be derived by considering $\varphi = v, v^2$.

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

Corollary 5.1 *Let $f \in H_p^r([-\pi,\pi]^3)$, $r \ge 0$ then*

$$||Q^\lambda(f) - Q_N^\lambda(f_N)||_2 \le \frac{C}{N^r}\left(||f||_{H_p^r} + ||Q^\lambda(f_N)||_{H_p^r}\right), \tag{47}$$

Proof
It is enough to observe that

$$||f - f_N||_2 \le \frac{C}{N^r}||f||_{H_p^r}.$$

■

Remark 5.4 *From the previous corollary it follows*

$$|< Q^\lambda(f),\varphi> - < Q^\lambda_N(f_N),\varphi>| \leq \frac{C}{N^r}||\varphi||_2\left(||f||_{H^r_p} + ||Q^\lambda(f_N)||_{H^r_p}\right), \tag{48}$$

and hence, by taking $\varphi = v, v^2$, the spectral accuracy of the moments.

5.3 Stability of a smoothed positive scheme

Let us rewrite the homogeneous Boltzmann equation for VHS kernels with cut-off over the relative velocity in the equivalent form

$$\frac{\partial f}{\partial t} + \mu f = J^\lambda(f,f), \tag{49}$$

with $J^\lambda(f,f) = Q^\lambda(f,f) + \mu f$.

It is easy to check that for the loss part of the collision operator $fL^\lambda(f)$ we have the inequality

$$L^\lambda(f) \leq C_\alpha 4\pi(2\lambda\pi)^\alpha \int_{[-\pi,\pi]^3} f(v)\,dv = C_\alpha 4\pi(2\lambda\pi)^\alpha\rho.$$

Thus for

$$\mu \geq C_\alpha 4\pi(2\lambda\pi)^\alpha\rho, \tag{50}$$

$J^\lambda(f,f)$ is a positive monotone operator in the sense that

$$J^\lambda(f,f) \geq J^\lambda(g,g) \geq 0 \quad \text{if} \quad f \geq g \geq 0.$$

Then we have the following result regarding the stability of a smoothed spectral scheme [30]

Theorem 5.2 *There exist a unique positive solution $f_N(t,v) \in C^1([0,T], I\!P^N)$, with $||f_N||_1 = \rho$, for arbitrary time $T > 0$ to the initial value problem*

$$\begin{aligned} &\frac{\partial f_N}{\partial t} + \mu f_N = \mathcal{S}J^\lambda_N(f_N), \qquad v \in I\!R^3,\ t > 0, \\ &f_N(v,t=0) = f^s_{0,N}(v), \end{aligned} \tag{51}$$

provided that $f_0 \in L^2([-\pi,\pi]^3)$ is non negative, $||f_0||_1 = \rho$, μ satisfies (50) and the smoothing operator $\mathcal{S}$ satisfies (37).

The proof is similar to the proof of existence and uniqueness of a solution for the homogeneous Boltzmann equation for hard sphere gases with cut-off on the relative velocity. It makes use of the conservation of mass of $\mathcal{S}Q^\lambda_N(f_N)$ and on the positivity preserving properties of $\mathcal{P}^s_N$ [30]. From Theorem 5.2 it follows that the L^1 norm of the spectral solution is constant in time and hence the smoothed positive scheme is stable in the L^1-norm.

Remark 5.5 *For practical purposes the positive scheme (51) introduces too much smoothing, and spectral accuracy is lost. Note that, without smoothing, scheme (51) and scheme (42) are the same. However, it is interesting to remark that, as pointed out in [30], the main reason in the lack of accuracy of (51) is represented by the smoothed projection of the initial data.*

6 Time discretization and numerical tests

The time integration of the Boltzmann equation represents in itself a challenging problem, since the nonlinear collision operator becomes highly *stiff* near the fluid regime ($\epsilon << 1$). A general method for the development of unconditionally stable schemes with the correct fluid limit has been recently presented in [16]. This method will be at the basis of our construction of uniformly accurate second order schemes for the Boltzmann equation [31]. However, here we are just interested in testing the accuracy of the velocity discretization we have seen in the previous sections and therefore we will always assume a fixed $\epsilon = 1$.

Clearly, for a low-order time discretization, one expects that the error of the fully discretized scheme will be dominated by the temporal errors. Thus if we want to check the spectral accuracy of the method either very small time-steps or else high-order time-differencing methods are required. In our tests we use the spectral method coupled with an explicit fourth-order Runge-Kutta method in time. This Runge-Kutta scheme provides the high temporal accuracy needed to demonstrate spectral accuracy in velocity.

6.1 Numerical tests

In this section we study the performance of our spectral methods by applying them to the following test problems:

Test #1 2D Maxwellian molecules (i.e. $\alpha = 0$), with $C_0 = 1/(2\pi)$.

Initial condition:

$$f(v,0) = \frac{v^2}{\pi\sigma^2}\exp(-v^2/\sigma^2).$$

Integration time: $t_{\max} = 40$. This problem has an exact solution given by [14]

$$f(v,t) = \frac{1}{4\pi S^2\sigma^2}\left(2S - 1 + \frac{1-S}{2S}\frac{v^2}{\sigma^2}\right)\exp\left(-\frac{v^2}{2S\sigma^2}\right),$$

where $S = 1 - \exp(-\sigma^2 t/8)/2$. In the computation the scaling parameter $\sigma = \pi/6$ is chosen in such a way that the numerical support of the initial condition is well approximated by $\mathcal{B}(0, R)$.

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, and to check the accuracy in the conservation of energy and of the fourth moment. Because of the symmetry of the problem, the moments of order 1 are conserved, within round-off error.

Test #2 3D Maxwellian molecules, with $C_0 = 1/(4\pi)$.

Initial condition: sum of two Gaussians

$$f(v,0) = \frac{1}{2(2\pi\sigma^2)^{3/2}}\left[\exp\left(-\frac{|v-2\sigma e_1|^2}{2\sigma^2}\right) + \exp\left(-\frac{|v+2\sigma e_1|^2}{2\sigma^2}\right)\right]$$

with $\sigma = \lambda\pi/6$. Integration time: $t_{\max} = 10$.

Here $e_1 \equiv (1,0,0)$ denotes the unit vector in the direction v_x.

This test is used to check the evolution laws for the moments of the distribution function (temperature and stress tensor). The derivation of the evolution law for the stress tensor is presented in [5]. Let $P_{i,j}$, $i,j = 1,\cdots,3$ denote the components of the stress tensor defined as

$$P_{ij} = \int_{R^3} f(v)(v_i - u_i)(v_j - u_j)\, dv$$

where u_i are the components of the mean velocity. Because of the axial symmetry of the initial condition, the stress tensor is diagonal, with $P_{22} = P_{33}$. The expression of the components is given by

$$P_{11} = p + \frac{2}{3}w(t), \quad P_{22} = p - \frac{1}{3}w(t)$$

and $w(t) = w_0 \exp(-2\pi\rho C_0 t)$. $w_0 = 2\sigma^2$, $p = \rho T$, $T = (7/3)\sigma^2$. In our case it is $\rho = 1$ and $C_0 = 1/(4\pi)$, and therefore $2\pi\rho C_0 = 1/2$.

Test #3 2D VHS molecules, with $C_0 = 1/(2\pi)$.

Initial condition: sum of two Gaussians

$$f(v,0) = \frac{1}{2(2\pi\sigma^2)}\left[\exp\left(-\frac{|v-2\sigma e_1|^2}{2\sigma^2}\right) + \exp\left(-\frac{|v+2\sigma e_1|^2}{2\sigma^2}\right)\right]$$

with $\sigma = \lambda\pi/6$. Integration time: $t_{\max} = 10$. This test is used to check the evolution of moments (temperature and stress tensor), and to compare the relaxation to equilibrium of Maxwellian molecules, with the relaxation of other VHS molecules. For Maxwellian molecules, an expression similar to the 3D case can be derived for the components of the stress tensor:

$$P_{11}(t) = p + \frac{1}{2}w(t), \quad P_{22}(t) = p - \frac{1}{2}w(t)$$

where $w(t) = w_0 \exp(-\pi\rho C_0 t)$. In our case $\rho = 1$, and $C_0 = 1/(2\pi)$, therefore $\pi\rho C_0 = 1/2$.

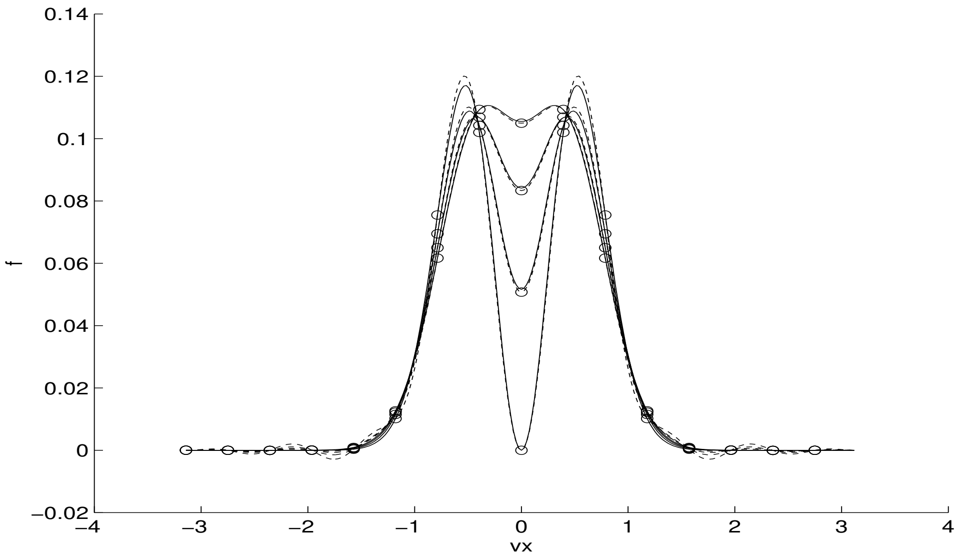

Figure 3: Test # 1. Density function $f(v_x, 0, t)$, at time $t = 0, 3, 6, 9$. $N = 16$ modes per direction. Line: exact solution. Circles: numerical solution. Dots: trigonometric interpolation.

6.2 Numerical results

The results of Test #1 are shown in Figs. 3–8. Figure 3 shows the cross section of the exact (lines) and numerical (circles) solution at different times, obtained with $N = 16$ modes per direction. The dotted lines represent the numerical solution reconstructed by trigonometric interpolation. It is possible to see that the reconstructed solution loses positivity near the tail. Positivity may be recovered by using Fejer's filter (see remark 5.1), but this would introduce some additional smearing in the solution. However it is remarkable that the solution remain positive at the nodes.

Figure 4 shows the relative L_∞, L_1, and L_2 norms of the difference between the numerical and the exact solution, as a function of time. They are computed according to the formula

$$E_p = \left(\frac{\sum_{i=-N}^{N} |f_i(t) - f(v_i, t)|^p}{\sum_{i=-N}^{N} |f(v_i, t)|^p} \right)^{1/p}$$

with $i = (i_1, i_2)$, $p = 1$ and $p = 2$. A similar expression is used for the L_∞ error.

Note that the error increases initially, and then it decreases almost monotonically in time. After a long time the error starts increasing again. This effect is due to aliasing. When the number of Fourier modes increases, the effect of aliasing becomes dominant over the error due to the spectral approximation.

This effect is more evident in Figure 5 where we show the relative L_1 norm of the error in the density function for $N = 8, 16$, and 32 modes per direction. A time step $\Delta t = 1$ has been used.

For short times the spectral accuracy is evident. If we fix the attention, for example, to time $t = 5$, then the error decays by a factor 5 when going from 8 to 16 modes, and by a factor 800 when going from 16 to 32 modes. The order of accuracy would therefore be 5.64 in the first case and 9.64 in the second. This is a clear indication that the convergence is more than algebraic.

For longer times the aliasing effect dominates in the case $N = 32$. This suggests that a compromise should be obtained between aliasing and spectral error, and that the period should be increased when increasing the number of Fourier modes. Roughly speaking, the period should be chosen in such a way that the two contribution of the error are of the same order of magnitude.

To show that the effect of aliasing becomes dominant when the number of Fourier modes is increased, we repeat the previous calculation using a more compact initial condition, which is equivalent to using a larger period. The results of the computation are shown in Figs. (6) and (7). With 32 modes per direction, when increasing the period with respect to the numerical support of the function, the error for short times increases, because of the loss of resolution (since the number of modes is the same), but the error for long time decreases, since the effect of aliasing is less pronounced.

In Figure 8 the error in the conservation of energy is shown. Energy is not exactly conserved, but if aliasing error is small, energy is conserved within spectral accuracy. The cusp in the figure is due to the fact that the exact and the numerical values of the energy cross at a certain time, and therefore at that time the error is zero (which is shown by a cusp in a semi-log plot). Mass (i.e. $\int f(v)\, dv$) is conserved exactly by the scheme, while conservation of momentum is maintained within round-off error, because of the particular symmetry of the problem. Because of spectral accuracy, a semi-log plot has been used for the norms in most figures.

The results of Test # 2 are shown in Figs. 9-12. A time step $\Delta t = 0.5$ has been used. Figure 9 shows the time evolution of the temperature for $N = 8$ and $N = 16$ modes per direction. The result with 8 modes is very inaccurate, because 8 modes are not enough to furnish an accurate representation of the initial condition with this period. It is possible to obtain a better result with 8 modes by simply reducing the period (see section 6.3). On the other hand, with 16 modes the maximum error remains less than 0.39 %, and is mainly due to aliasing.

Note that only the zeroth moment is conserved by the scheme. However, because of loss of positivity, the L_1 norm of f is not conserved. The loss of positivity can be measured by the difference between the L_1 norm of the function and its initial value. This difference (which we do not report) is however spectrally small.

The evolution of the stress tensor is shown in Figs. 10–11. The relative error in the computation of the stress tensor are shown in Fig. 12.

The agreement between the numerical solution obtained with only 16 modes and the analytical solution is remarkable.

The results of Test # 3 are shown in Figs. 13–15. The relative error in the computation of the stress is below 10^{-4}, and this allows an accurate computation of the relaxation of the moments for general VHS molecules. Note that, at variance with the case shown in the previous figure, in this case the error increases monotonically. The reason is that with this choice of the number of modes and size of the computational domain, the aliasing error is dominant with respect to the resolution error.

As an example, in Fig. 14 we show the relaxation of P_{11} for molecules with $\alpha = 0, 0.5, 1,$ and 2. A semi-log plot of the relaxation shows that the decay is essentially exponential (see Fig. 15).

6.3 Comparison with Monte Carlo

Finally we compare the results obtained with our spectral method with those obtained by a Monte Carlo scheme. We used a standard version of Monte Carlo, which may be refereed as the Nanbu-Babowsky scheme [28, 2]. We apply Monte Carlo to Maxwell molecules. More specifically we consider two tests. In the first, referred as Test A, we repeat the computation of Test #3 with Monte Carlo, and we show the relative error in the second moment P_{11}, during the time interval $[0, 10]$. In the second test, referred as Test B, we apply the Monte Carlo method to Test #1, and compute the relative error in the fourth moment M_4. The exact value of the fourth moment is $\int_{R^2} v^4 f(v,t)\, d^2v = 8S(t)(2 - S(t))$.

The Monte Carlo scheme is described as follows. The homogeneous Boltzmann equation (for Maxwell molecules) is discretized as

$$f^{n+1}(v) = (1 - \mu\Delta t) f^n(v) + \Delta t Q^+(f^n, f^n)(v),$$

where $f^n(v)$ denotes an approximation of the density function at time $n\Delta t$, and, because we treat only space homogeneous problems, we assume that the density function is a probability density. In our case (Test #1 and Test # 3) it is $\mu = 1$.

Then, the Monte Carlo scheme is obtained from the probabilistic interpretation of the previous equation. Given N particles at time $t_n = n\Delta t$, at each time step,

i) $(1 - \mu\Delta t)N$ particles will not collide,

ii) $N\mu\Delta t/2$ collisions are performed with collision pairs randomly selected uniformly among all possible pairs.

During each collision, the new velocities are obtained according to Eq. (3), with ω uniformly sampled on the unit sphere (in our case in the unit circle, since the computation is in 2D in velocity space).

The result of the two tests are reported in Tables 1 and 2. The first column of each table reports the CPU time of the calculation. All the computations were performed on an

Alpha Server 2000 Model 4/275. In Table 1, the second column indicates the number of modes per direction. The third column shows the time step used, and the fourth column indicates the Runge-Kutta scheme. The values of σ are $\sigma = \lambda\pi/6$ for Test A and $\sigma = 1$ for Test B.

In both tests the error denotes the relative error between the numerical and the exact solution. More specifically, for Test A one has:

$$E_1 = \frac{1}{N_o}\sum_{k=1}^{N_o}\left|1 - \frac{P_{11}(t_k)}{P_{11}^{\text{exa}}(t_k)}\right|, \quad E_\infty = \max_{k=1\ldots N_o}\left|1 - \frac{P_{11}(t_k)}{P_{11}^{\text{exa}}(t_k)}\right|$$

Similar expressions are computed for the fourth moment in Test B. Here N_o denotes the number of output times, which is 10 or 20 (depending on the time step used for the spectral method) in both spectral and Monte Carlo calculations. We tried to choose parameters which give good performance for both methods. For example, with the spectral method, different periods T have been used for different numbers of Fourier modes, as indicated in column 5 and 8 of Table 1, in order to balance resolution and aliasing.

In the case of Monte Carlo, the moments are computed by using unbiased estimators averaged over several runs. Number of runs N_{runs}, number of particles N_p, and time step Δt, have been chosen in such a way to balance time discretization error with statistical fluctuations. Because the relative error obtained by Monte Carlo is still a random variable with a considerable variance, each result reported in Table 2 is obtained by taking an average of 10 different sets of runs (but the CPU time refers, of course, to a single set of runs !).

We remark that the computations have been obtained by using Maxwellian molecules. This case is particularly favorable to the Monte Carlo scheme, since the cross section is constant and no rejection is needed. From the comparison, it is evident that, for two dimensional computations, the greater cost of the spectral scheme (with the same number of degrees of freedom) is compensated by a much greater accuracy, allowing better results with the same computational cost.

Remark 6.1 *Note that, in the 3D case for α integer, using the result of section 4.1 one can avoid to store the coefficients $\hat{B}(l,m)$ in a matrix. Thus, the storage requirement of the spectral method and of the Monte Carlo method are comparable, since the number of particles used in the latter is of the same order of the number of modes used by the former. In the general case, by storing the coefficients in a matrix, the storage requirement of the spectral scheme is much larger, since the number of coefficients is $\mathcal{O}(N^{2d})$, where d is the dimension of the velocity space. However, in a space non homogeneous problem the storage requirement would not represent a serious drawback, since the kernel modes $\hat{B}(l,m)$ would be the same for all space cells. Therefore if the number of space cells is the same (or larger) than the number of Fourier modes, then the storage requirements of the Monte Carlo method and the spectral method would be comparable.*

Spectral parameters				Test A (# 3)			Test B (# 1)		
Time (s)	N	Δt	RK	T	$E_1(\%)$	$E_\infty(\%)$	T	$E_1(\%)$	$E_\infty(\%)$
0.05	8	1.0	2	0.55 π	5.24	6.85	4.0	16.7	23.7
1.44	16	1.0	4	0.75 π	0.0826	0.140	5.5	0.350	0.48
55.1	32	0.5	4	0.90 π	0.00285	0.00646	6.0	0.0261	0.0375

Table 1: Relative error in P_{11} (Test A) and M_4 (Test B) obtained by the spectral method for different numbers of modes.

Monte Carlo parameters				Test A (# 3)		Test B (# 1)	
Time (s)	N_p	Δt	N_{runs}	$E_1(\%)$	$E_\infty(\%)$	$E_1(\%)$	$E_\infty(\%)$
0.05	100	0.5	10	3.5	7.2	3.91	6.12
1.5	500	0.1	60	0.54	1.3	0.58	1.18
56.6	2000	0.01	200	0.13	0.30	0.24	0.41

Table 2: Relative error in P_{11} (Test A) and M_4 (Test B) obtained by the Monte Carlo method.

7 Conclusions

In this paper we introduce a spectrally accurate scheme for the evaluation of the collisional integral of the Boltzmann equation. To our knowledge, this is the most accurate scheme present in the literature.

The computational domain in velocity space must be chosen large enough, so that aliasing error is not too large. The problem of finding a sharp estimate of the aliasing error, and therefore the optimal balance between resolution in velocity and aliasing error deserves further investigation.

The complexity of the scheme is $O(n^2)$, where n is the total number of discretization points in velocity space. The efficiency of the scheme is therefore lower than Monte Carlo methods, but it is higher than other deterministic schemes available in the literature. We hope that in a near future it will be possible to find a fast algorithm for the computation of the gain term (the loss term can be computed in $O(n \log n)$ operations). We believe that if this is the case, then this scheme will become more effective than present schemes in several applications. A further reduction in the computational cost is in fact almost necessary for realistic problems for the full space non-homogeneous Boltzmann equation with three dimensions in velocity.

In view of future applications to space non-homogeneous problems, a couple of remarks are in order. For non-stiff problems, the use of Strang splitting between the convection and collision step, together with a second order-shock capturing scheme for the convection step, will guarantee second order accuracy in space, and spectral accuracy in velocity. For stiff problems, the situation is more complicated. Schemes based on the Wild sum representation of the solution seem to be suitable to this purpose [16]. Uniformly accurate second order schemes based on this discretization will be presented in a companion paper [31].

Finally, we remark that a similar approach can be successfully applied to the Landau-Fokker-Planck collision operator [33].

Acknowledgements. L. Pareschi is grateful for the encouragement and support of Benôit Perthame and Giuseppe Toscani. Both authors would like to thank Bernt Wennberg and Cedric Villani for helpful discussions and suggestions.

References

[1] V.V. Aristov, F.G. Ceremisin *The conservative splitting method for solving Boltzmann's equation*, USSR Comput. Maths. Math. Phys., 20, 1, (1980), pp. 208-225.

[2] H. Babovsky *On a simulation scheme for the Boltzmann equation*, Math. Methods Appl. Sci., **8** (1986), pp. 223-233.

[3] G.A. Bird *Molecular gas dynamics*, Clarendon Press, Oxford, (1994).

[4] A.V. Bobylev *The method of Fourier transform in the theory of the Boltzmann equation for Maxwell molecules*, Dokl. Akad. Nauk SSSR, 225, (1975), pp. 1041-1044.

[5] A.V. Bobylev, S. Rjasanow *Difference scheme for the Boltzmann equation based on the Fast Fourier Transform*, Eur. J. Mech., B/Fluids, 16, 2, (1997), pp. 293-306.

[6] C. Buet *A discrete velocity scheme for the Boltzmann operator of rarefied gas dynamics*, Trans. Theo. Stat. Phys. 25 1996, pp. 33-60.

[7] J.F. Bourgat, P. Le Tallec, B. Perthame, Y. Qiu *Coupling Boltzmann and Euler Equations without overlapping*, Domain decomposition methods in science and engineering, Como 1992, Contemp. Math., 157, AMS (1994), pp. 377-398.

[8] R.E. Caflisch, L. Pareschi *Implicit Monte-Carlo methods for rarefied gas dynamics I: the space homogeneous case*, J. Comp. Phys. submitted.

[9] C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang Spectral methods in fluid dynamics, Springer Verlag, New York, (1988).

[10] C. Cercignani The Boltzmann equation and its applications, Springer-Verlag, Berlin, (1988).

[11] C. Cercignani, R. Illenr, M. Pulvirenti, The mathematical theory of dilute gases, Springer-Verlag, New York, (1995).

[12] F.G. Ceremisin *Fast solution of the Boltzmann Equation*, Proceedings of the 17th International Symposium on Rarefied Gas Dynamics, edited by A. Beylich, Weinheim, (1991), pp. 273-284.

[13] F. Coron, B. Perthame *Numerical passage from kinetic to fluid equations*, SIAM J. Numer. Anal., 28, (1991), pp. 26-42.

[14] M. H. Ernst *Exact solutions of the nonlinear Boltzmann equation and related kinetic models*, Nonequilibrium Phenomena I. The Boltzmann equation, North-Holland, (1983), pp. 52-119.

[15] E. Gabetta, L. Pareschi *The Maxwell gas and its Fourier transform towards a numerical approximation*, Waves and stability in continuous media (Bologna 1993), Ser. Adv. Math. for Appl. Sci., 23, (1994), pp. 197-201.

[16] E. Gabetta, L. Pareschi, G. Toscani *Relaxation schemes for nonlinear kinetic equations*, SIAM J. Numer. Anal., 34, 6, (1997), pp. 1967-2011.

[17] D. Goldstein, B. Sturtevant, J.E. Broadwell *Investigation of the motion of discrete velocity gases*, Rar. Gas. Dynam., Progress in Astronautics e Aeronautics, 118, AIAA, Washington, (1989).

[18] D. Gottlieb, S.A. Orszag Numerical Analysis of Spectral Methods: Theory and Applications, SIAM CBMS-NSF Series, (1977).

[19] Y.N. Grigoriev, A.N. Mikhalitsyn *A spectral method of solving Boltzmann's kinetic equation numerically*, U.S.S.R. Comput. Maths. Math. Phys., 23, 6, (1983), pp. 105-111.

[20] T. Gustaffson *L^p-estimates for the nonlinear spatially homogeneous Boltzmann equation*, Arch. Rat. Mech. Anal., 92, (1986) pp. 23-57.

[21] J.C. Hang, J.Y. Yang *Rarefied flow using nonlinear model Boltzmann equations*, J. Comp. Phys., 120, (1995), pp. 323-339.

[22] R. Illner, H. Neunzert *On simulation methods for the Boltzmann equation*, Trans, Theo. Stat. Phys., 16, (1987), pp. 141-154.

[23] T. Inamuro, B. Sturtevant *Numerical study of discrete velocity gases*, Phys. Fluids A, 2, 12, (1990), pp. 2196-2203.

[24] P.L.Lions *Compactness in Boltzmann's equation via Fourier integral operators and applications. I*, J. Math. Kyoto Univ. 34-2, (1994), pp. 391-427.

[25] E. Longo, L. Preziosi *On a conservative polar discretization of the Boltzmann equation*, Japan J. Indust. App. Math., 14, (1997), pp. 399-435.

[26] P. Le Tallec, F. Mallinger *Coupling Boltzmann and Navier-Stokes Equations by Half Fluxes*, J. Comp. Phys., 136, (1997), pp. 51-67.

[27] Y. Morchoisne *Une méthode de différences finies pour la résolution del l'equation de Boltzmann: Traitement du terme de collision* C. R. Acad. Sci. Paris, t.313. Série II, (1991), pp. 1513-1518.

[28] K. Nanbu *Direct simulation scheme derived from the Boltzmann equation. I. Monocomponent Gases*, J. Phys. Soc. Japan, 52, (1983), pp. 2042-2049.

[29] L. Pareschi, B. Perthame *A Fourier spectral method for homogeneous Boltzmann equations*, Trans. Theo. and Stat. Phys., 25, 5, (1996), pp. 369-383.

[30] L. Pareschi, G. Russo *On the stability of spectral methods for the Boltzmann equation*, Transp. Theo. Stat. Phys. submitted

[31] L. Pareschi, G. Russo *Numerical solution of the Boltzmann equation II: Uniformly accurate second order schemes*, work in preparation.

[32] L. Pareschi, G. Russo *Asymptotic preserving Monte Carlo methods for the Boltzmann equation*, Transp. Theo. Stat. Phys. to appear.

[33] L. Pareschi, G. Russo, G.Toscani *Fast spectral methods for the Landau-Fokker-Planck equation of plasma physics*, preprint 1999.

[34] B. Perthame *Introduction to the theory of random particle methods for Boltzmann equation*, in Advances in Kinetic Theory and Computing, B. Perthame Editor, World Scientific, (1994).

[35] D.I. Pullin *Direct simulation methods for compressible inviscid ideal gas-flow*, J. Comp. Phys., 34, (1980), pp. 231-244.

[36] F. Rogier, J. Schneider *A direct method for solving the Boltzmann equation*, Trans. Theo. Stat. Phys., 23, (1994), pp. 313-338.

[37] Y. Sone, T. Owada, K. Aoki *Temperature jump and Knudsen layer in a rarefied gas over a plane wall: Numerical analysis of the linearized Boltzmann equation for hard-sphere molecules*, Phys. Fluids A, 1, 2, (1989), pp. 363-370.

[38] Z.Tan, P. Varghese *The Δ-ϵ method for the Boltzmann equation*, J. Comp. Phys., 110, (1994), pp. 327-340.

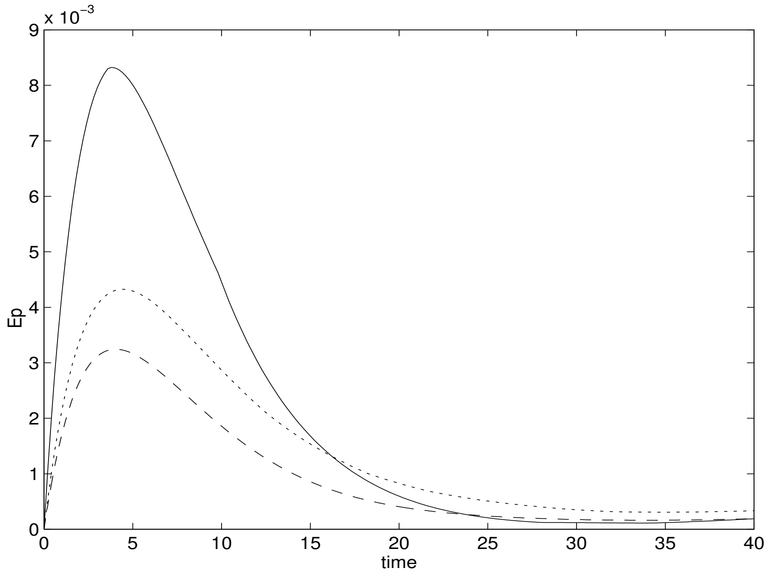

Figure 4: Test # 1. Relative norms of the error in the density function vs time, for $N = 16$ modes per direction. Dots: L_∞, line: L_1, dash: L_2.

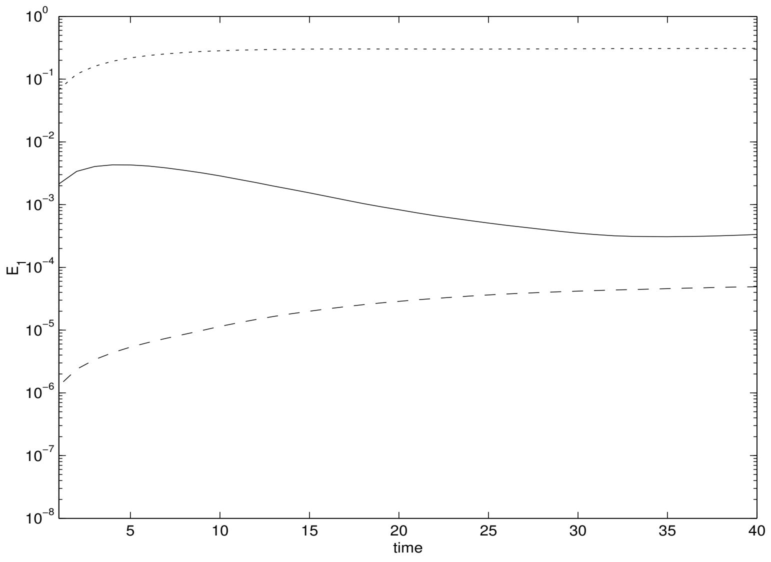

Figure 5: Test#1. L_1 relative norm of the error vs time for $N = 8, 16$, and 32 modes per direction.

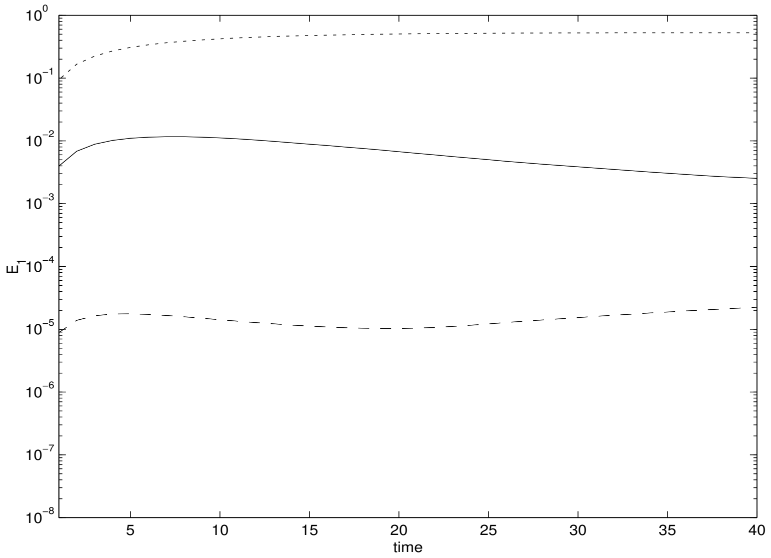

Figure 6: Same as figure 5, but with a more compact initial condition ($\sigma = \pi/7$).

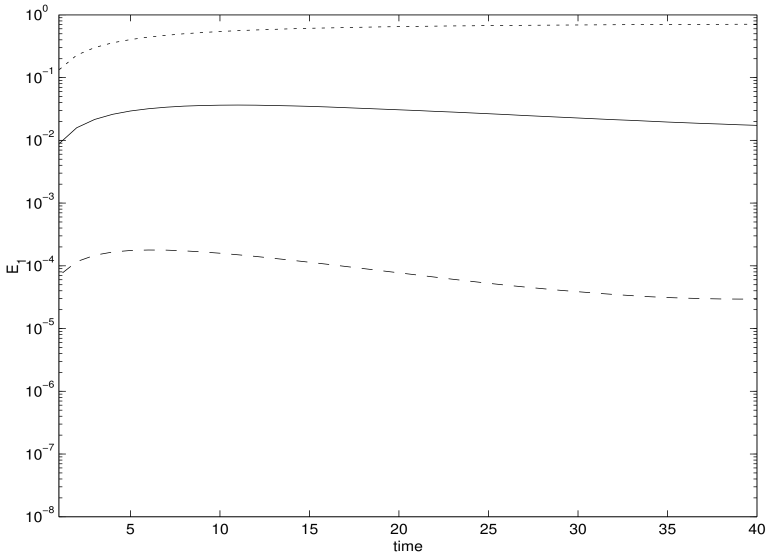

Figure 7: Same as figure 6, but with a more compact initial condition ($\sigma = \pi/8$).

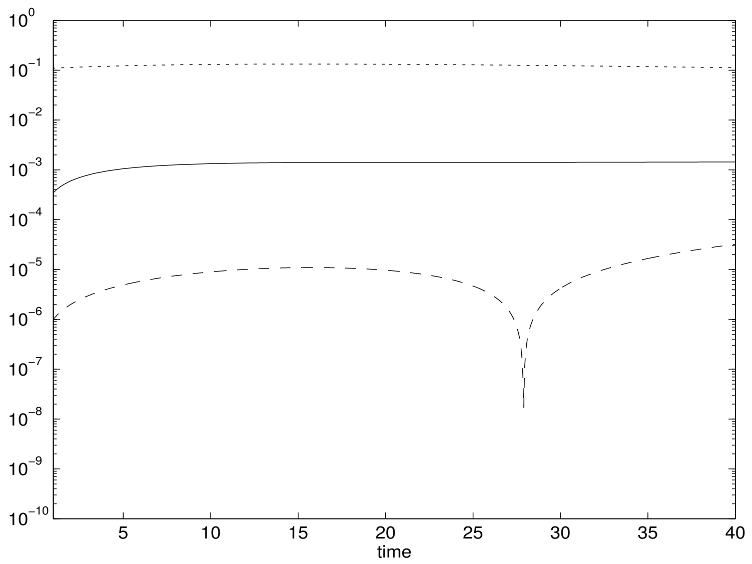

Figure 8: Test#1. Error in energy vs time for $n = 8, 16$, and 32 modes per direction.

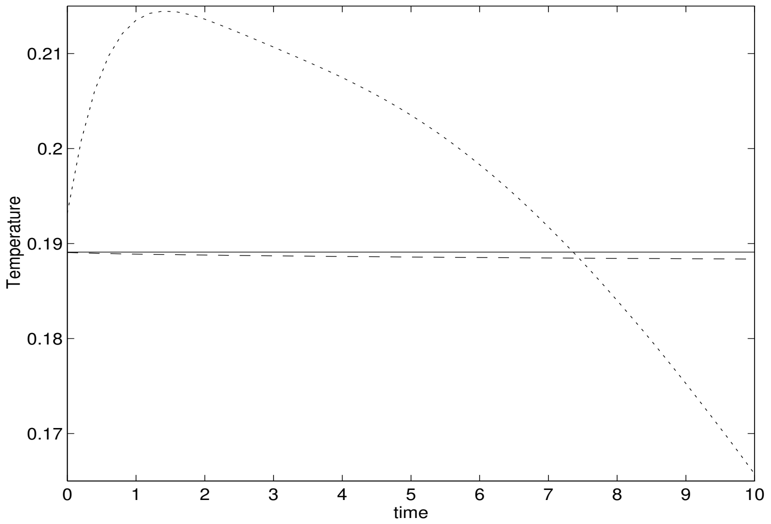

Figure 9: Test#2. Time evolution of the temperature. Exact (line), N=8 (dots), N=16 (dash).

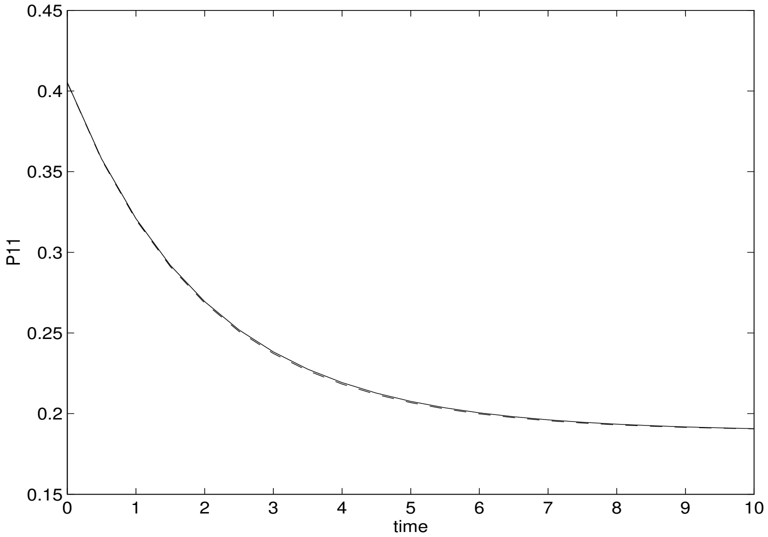

Figure 10: Test#2. Stress tensor component P_{11} vs time for N=16. Exact (line) and numerical (dash)

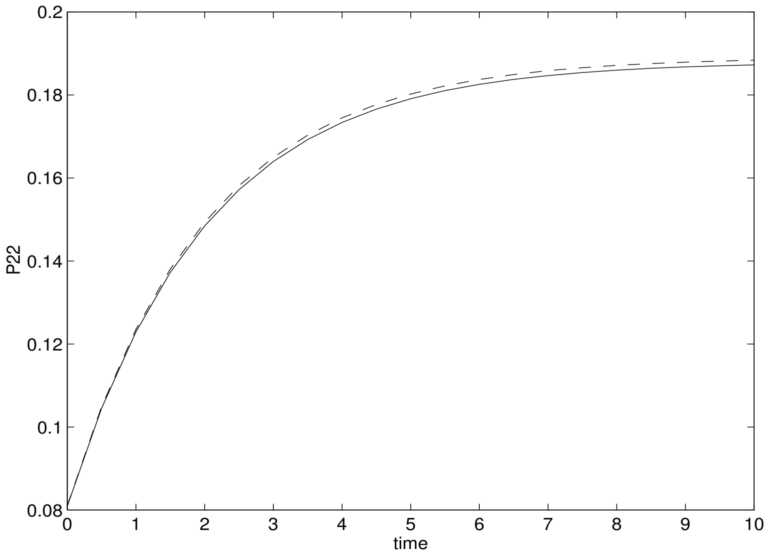

Figure 11: Test#2. Stress tensor component P_{22} vs time for N=16. Exact (line) and numerical (dash)

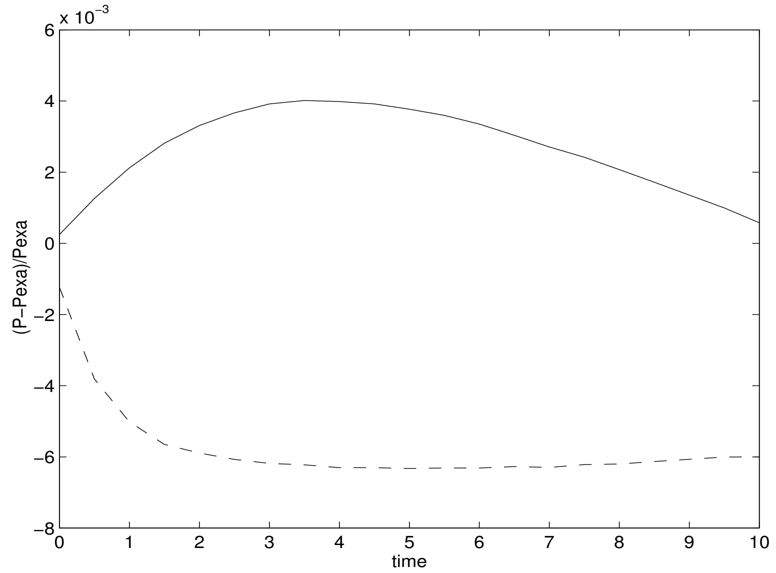

Figure 12: Test#2. Relative error in the computation of the stress tensor: P11 (line) and P22 (dash) for N=16.

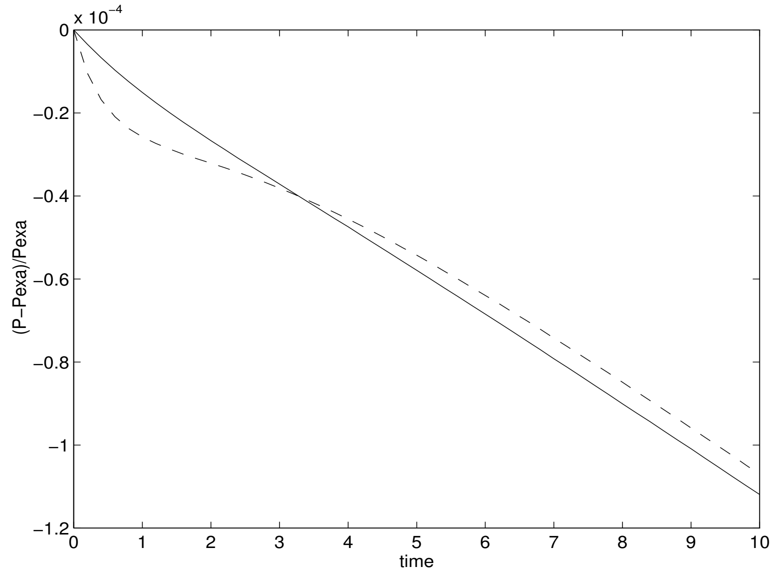

Figure 13: Test#3. Relative error in the computation of the stress tensor: P11 (line) and P22 (dash) for N=32.

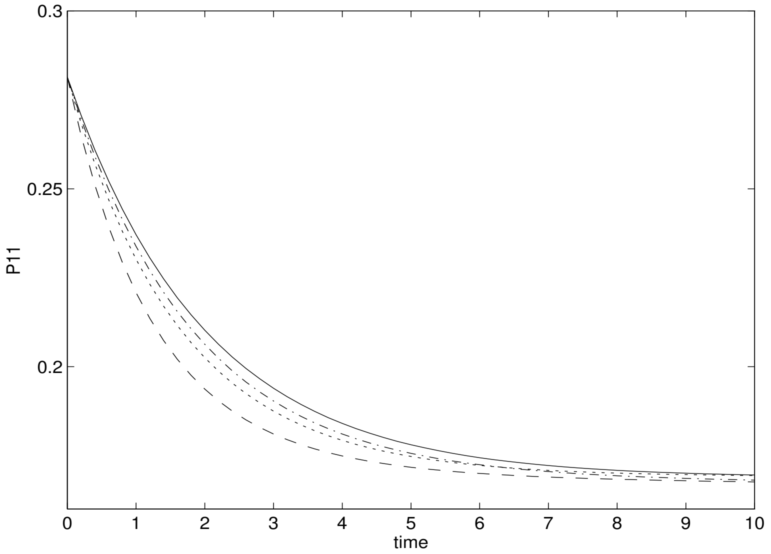

Figure 14: Test#3. Relaxation of the stress for different VHS molecules. $\alpha = 0$ (line), $\alpha = 0.5$ (dot-dash), $\alpha = 1$ (dots), and $\alpha = 2$ (dash).

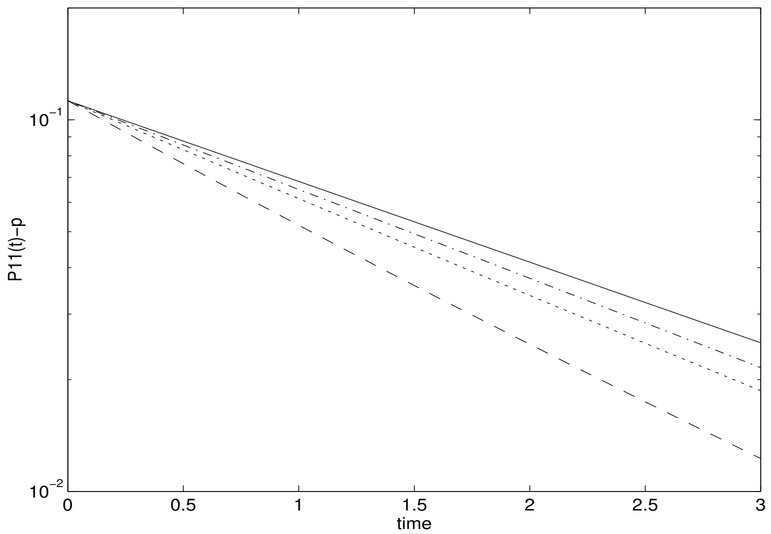

Figure 15: Test#3. Relaxation of the stress for different VHS molecules. $\alpha = 0$ (line), $\alpha = 0.5$ (dot-dash), $\alpha = 1$ (dots), and $\alpha = 2$ (dash).