Weierstraß–Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Reduction of the number of particles in the stochastic weighted particle method for the Boltzmann equation

Sergej Rjasanow[1] , Thomas Schreiber[2] , Wolfgang Wagner[3]

submitted: 10th March 1997

[1] University of Saarbrücken
Department of Mathematics
Postfach 15 11 50
D – 66041 Saarbrücken
Germany
email: rjasanow@num.uni-sb.de

[2] University of Kaiserslautern
Department of Informatics
D – 67653 Kaiserslautern
Germany
email: tschreib@informatik.uni-kl.de

[3] Weierstrass Institute
for Applied Analysis
and Stochastics
Mohrenstraße 39
D – 10117 Berlin
Germany
email: wagner@wias-berlin.de

Preprint No. 320
Berlin 1997

1991 Mathematics Subject Classification. 65C05, 76P05, 82C80.

Key words and phrases. Boltzmann equation, stochastic particle method, random weight transfer, collision mechanism, numerical experiments.

Edited by
Weierstraß–Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax: + 49 30 2044975
e–mail (X.400): c=de;a=d400-gw;p=WIAS-BERLIN;s=preprint
e–mail (Internet): preprint@wias-berlin.de

Abstract. Different ideas for reducing the number of particles in the stochastic weighted particle method for the Boltzmann equation are described and discussed. The corresponding error bounds are obtained and numerical tests for the spatially homogeneous Boltzmann equation presented. It is shown that if an appropriate reduction procedure is used then any effect on the accuracy of the numerical scheme is negligible.

Contents

1. Introduction

In this paper we continue the development and analysis of the Stochastic Weighted Particle Method (SWPM) for kinetic equations. This method was introduced in [14], where we presented first numerical results for the one-dimensional heat exchange problem. The convergence of the method was investigated in [15], where we were also able to show a drastic reduction of the stochastic fluctuations using the SWPM for one model kinetic equation. In [16] we presented a detailed study of different effects of the numerical solution of this equation. The computation of the macroscopic quantities in the regions with low particle density was of special interest. We refer to [9], [8] to complete the list of references for the SWPM.

The main object of our interest is the spatially inhomogeneous nonlinear Boltzmann equation for dilute monoatomic gases [4]

$$\frac{\partial f}{\partial t}(t,x,v) + (v, \mathrm{grad}_x f(t,x,v)) = \tag{1.1}$$
$$= \int_{\mathbb{R}^3}\int_{S^2} B(v,w,e)\,[f(t,x,v')f(t,x,w') - f(t,x,v)f(t,x,w)]\,de\,dw\,,$$

$$f(0,x,v) = f_0(x,v)\,, \tag{1.2}$$

where $t \geq 0$ is the time variable, $x \in \Omega \subset \mathbb{R}^3$ is the space and $v \in \mathbb{R}^3$ is the velocity space variable. The vector e is from the unit sphere S^2. The function $B(v,w,e)$, the so-called collision kernel, has the following form for the "hard spheres" model

$$B(v,w,e) = \frac{1}{2\sqrt{2}\pi\kappa}|(v-w,e)|, \tag{1.3}$$

where κ denotes the Knudsen number. The post-collision velocities v' and w' are defined by

$$v' = v - (v-w,e)\,e\,, \qquad w' = w + (v-w,e)\,e\,. \tag{1.4}$$

The main difference between the SWPM and other particle schemes for the Boltzmann equation [2], [12], [11] is the idea of a random weight transfer between particles during collisions. The distribution function $f(t,x,v)$ in the low density regions of the flow can then be resolved more accurately by producing many particles of low weight. This procedure usually leads to an increase in the number of particles in the system. If this increase cannot be compensated in some natural way, for example, if the new small particles cannot leave the computational domain (as in the heat exchange problem), then it becomes imperative to reduce the number of particles. The problem of reducing the number of particles has already been discussed in [10], [19], [14].

In the present paper we give a systematic study of the theoretical and numerical aspects of reducing the number of particles including the theoretical estimates for the error in the

bounded Lipschitz metric as well as in the Sobolev space $\mathbb{H}^{-2}$. Furthermore, we discuss a possible choice of the reduction parameters in both a random and a deterministic way. In the numerical tests we concentrate on the influence of the reduction on statistical values such as empirical mean and confidence intervals. It is shown that an appropriate reduction procedure has little effect on the accuracy of the numerical scheme.

The paper is organized as follows. A brief description of the SWPM is given in Section 2. In Section 3, the main part of the paper, we discuss different approaches to the reduction of the number of particles. The results of our numerical tests are presented in Section 4. Finally, we draw some concluding remarks.

2. Description of the SWPM

The main idea of all particle methods for the Boltzmann equation (1.1), (1.2) is an approximation of the sequence of measures

$$f(t_k, x, v)dxdv\,, \quad t_k = k\,\Delta t\,, \quad k = 0, 1, \dots\,, \quad \Delta t > 0\,,$$

by a system of point measures

$$\mu(t_k, dx, dv) = \sum_{j=1}^{n(t_k)} g_j(t_k)\delta_{(x_j(t_k),\, v_j(t_k))}(dx, dv)\,, \quad k = 0, 1, \dots\,, \tag{2.1}$$

defined by the families of particles

$$\Big(g_j(t_k), x_j(t_k), v_j(t_k)\Big)_{j=1}^{n(t_k)}\,, \quad k = 0, 1, \dots\,. \tag{2.2}$$

The behaviour of the system (2.2) can be briefly described as follows. The first step ($k = 0$) is an approximation of the initial measure

$$f_0(x, v)dxdv$$

by a system of particles (2.2) for $t_0 = 0$. Usually, one uses constant weights

$$g_j(0) = g\,, \quad j = 1, ..., n(0).$$

Then the particles move according to their velocities, i.e.

$$x_j(t) = x_j(t_k) + (t - t_k)v_j(t_k)\,, \quad t \in [t_k, t_{k+1}]\,.$$

If a particle crosses the "outflow boundary" during this step then this particle will be removed from the further simulation. The velocity of a particle changes according to the boundary condition if this particle hits the "boundary of the body": the particle continues

the movement with a new velocity for the rest of the time interval. The weights of particles remain the same during this "free flow step". Through the "inflow boundary" new particles of standard weight come into the computational domain.

The "collision step" can be described as follows. First, all particles are sorted in the spatial cells Ω_ℓ, $\ell = 1, ..., \ell_c$. These cells define a non-overlapping decomposition of the computational domain

$$\Omega = \bigcup_{\ell=1}^{\ell_c} \Omega_\ell \,.$$

In each cell Ω_ℓ, $\ell = 1, ..., \ell_c$, collisions of $n_\ell(t_k)$ particles are simulated. This is the most crucial part of the whole procedure. Here we also have the main difference between the SWPM and other particle methods which use constant weights. The collision simulation step in one spatial cell Ω_ℓ, $\ell = 1, ..., \ell_c$, corresponds to the mollified equation [4]

$$\frac{\partial f}{\partial t}(t,x,v) = \tag{2.3}$$
$$\int_\Omega \int_{\mathbb{R}^3} \int_{S^2} h_\ell(x,y) B(v,w,e) \left[f(t,x,v') f(t,y,w') - f(t,x,v) f(t,y,w) \right] de\, dw\, dy \,,$$

where

$$h_\ell(x,y) \;=\; \frac{1}{|\Omega_\ell|} \P_{\Omega_\ell}(x) \P_{\Omega_\ell}(y) \,, \tag{2.4}$$

is a spatial mollifier, $|\Omega_\ell|$ denotes the volume of the cell Ω_ℓ and $\P_{\Omega_\ell}(x)$ is the indicator function of the set Ω_ℓ.

The stochastic process of the collisions is

$$Z(t) = \{ (g_j(t), x_j(t), v_j(t)) \,,\; j = 1, ..., n \} \,, \quad t \geq t_k \,. \tag{2.5}$$

Here we now use the local numbering of particles in the cell Ω_ℓ and notate $n = n_\ell(t_k)$. The infinitesimal generator of the process (2.5) is given by

$$\mathcal{A}(\Phi)(z) = \sum_{1 \leq i \neq j \leq n} \int_{S^2} \frac{1}{2} q(z,i,j,e) \Big(\Phi(J(z,i,j,e)) - \Phi(z) \Big) \, de \,, \tag{2.6}$$

where Φ is a measurable function of the argument

$$z = ((g_1, x_1, v_1), \ldots, (g_n, x_n, v_n)) \tag{2.7}$$

and

$$\Big(J(z,i,j,e) \Big)_k = \begin{cases} (g_k, x_k, v_k) & , \text{if} \quad k \leq n \,,\; k \neq i,j \,, \\ (g_i - G(z,i,j,e), x_i, v_i)) & , \text{if} \quad k = i \,, \\ (g_j - G(z,i,j,e), x_j, v_j) & , \text{if} \quad k = j \,, \\ (G(z,i,j,e), x_i, v_i') & , \text{if} \quad k = n+1 \,, \\ (G(z,i,j,e), x_j, v_j') & , \text{if} \quad k = n+2 \,, \end{cases} \tag{2.8}$$

where v_i', v_j' are defined as in (1.4). The function $G(z,i,j,e)$ is called "weight transfer function". This function, the intensity kernel $q(z,i,j,e)$ of the generator (2.6) and the collision kernel of the Boltzmann equation (2.3), (2.4) are connected via the basic relationship

$$q(z,i,j,e)\,G(z,i,j,e) = h_\ell(x_i,x_j)\,B(v_i,v_j,e)\,g_i\,g_j \tag{2.9}$$

which has been proved [16] to be sufficient for the convergence of the method.

The behaviour of the process (2.5) is as follows. The waiting time $\hat{\tau}(z)$ between process jumps can be defined either as a random variable with the distribution

$$\text{Prob}\,\{\hat{\tau}(z) \geq t\} \;=\; \exp(-\hat{\pi}(z)\,t)\,,$$

where

$$\hat{\pi}(z) \;=\; \frac{1}{2}\sum_{1\leq i\neq j\leq n} \hat{q}_{max}(z,i,j) \tag{2.10}$$

and

$$\int_{S^2} q(z,i,j,e)\,de \;\leq\; \hat{q}_{max}(z,i,j)\,, \tag{2.11}$$

or as a deterministic object by

$$\hat{\tau}(z) = \hat{\pi}(z)^{-1}. \tag{2.12}$$

Then the collision partners (i.e. the indices i and j) must be chosen. The distribution of the parameters i and j is determined by the probabilities

$$\frac{\hat{q}_{max}(z,i,j)}{\sum_{1\leq i\neq j\leq n}\hat{q}_{max}(z,i,j)}\,. \tag{2.13}$$

For given i and j, the jump is fictitious with probability (cf. (2.11))

$$1 - \frac{\int_{S^2} q(z,i,j,e)\,de}{\hat{q}_{max}(z,i,j)}\,. \tag{2.14}$$

Otherwise the process (2.5) jumps to a new state $\tilde{z} = J(z,i,j,e)$ as described in (2.8). The distribution of the parameter e is

$$\frac{q(z,i,j,e)}{\int_{S^2} q(z,i,j,e)\,de}\,. \tag{2.15}$$

There is a degree of freedom in our method, namely an appropriate choice of the weight transfer function G. This function should always fulfil the condition

$$G(z,i,j,e) \leq \min(g_i,g_j)$$

in order to avoid negative weights in (2.8). We consider the function G in the form

$$G(z,i,j,e) = \frac{\min(g_i, g_j)}{1+\gamma(z,i,j,e)}\,, \tag{2.16}$$

where $\gamma(z,i,j,e) \geq 0$ is a parameter of our method which can be chosen arbitrarily, depending on our interest. The parameter γ can vary in different regions of the flow (cell Ω_ℓ), for different collision partners i and j or even as a function of the unit vector e. The jump intensity function q is then defined from the basic relationship (2.9) as

$$q(z,i,j,e) = (1+\gamma(z,i,j,e))\,\max(g_i,g_j)\,h_\ell(x_i,x_j)\,B(v_i,v_j,e)\,. \tag{2.17}$$

According to (2.11), we need a majorant for the function (2.17). Note that the function (2.4) is now just a constant, i.e.

$$h_\ell(x_i,x_j) = \frac{1}{|\Omega_\ell|}\,,$$

because we have assumed that all particles are sorted in cells. Furthermore, we use the majorants

$$\begin{aligned} 1+\gamma(z,i,j,e) &\leq 1+C_{\gamma,max}\,, && (2.18)\\ \int_{S^2} B(v_i,v_j,e)\,de &\leq C_{B,max}\,, && (2.19)\\ \max(g_i,g_j) &\leq g_i+g_j-g_{min}(z)\,, \end{aligned}$$

where (cf. (2.7))

$$g_{min}(z) = \min_{1\leq i\leq n} g_i\,, \tag{2.20}$$

to obtain

$$\hat{q}_{max}(z,i,j) = (1+C_{\gamma,max})\,C_{B,max}\,\frac{1}{|\Omega_\ell|}\,[g_i+g_j-g_{min}(z)]\,. \tag{2.21}$$

Now we are able to compute the waiting time parameter via (2.10)

$$\hat{\pi}(z) = \frac{1}{2}\,(1+C_{\gamma,max})\,C_{B,max}\,\frac{1}{|\Omega_\ell|}\,(n-1)\,[2\,g_{sum}(z)-n\,g_{min}(z)]\,, \tag{2.22}$$

where (cf. (2.7))

$$g_{sum}(z) = \sum_{i=1}^{n} g_i\,, \tag{2.23}$$

as well as all other parameters of our process. The probability of the parameters i and j is determined via (2.13) (cf. (2.21), (2.10), (2.22))

$$\frac{g_i + g_j - g_{min}(z)}{(n-1)\,[2\,g_{sum}(z) - n\,g_{min}(z)]}\,. \tag{2.24}$$

The parameter i is then to be chosen according to the probability

$$\frac{(n-2)\,g_i + g_{sum}(z) - (n-1)\,g_{min}(z)}{(n-1)\,[2\,g_{sum}(z) - n\,g_{min}(z)]}\,. \tag{2.25}$$

Given i, the parameter j is chosen according to the probability

$$\frac{g_i + g_j - g_{min}(z)}{(n-2)\,g_i + g_{sum}(z) - (n-1)\,g_{min}(z)}\,. \tag{2.26}$$

Given i and j, the jump is fictitious with probability (2.14) (cf. (2.17), (2.21))

$$1 - \frac{\int_{S^2}(1+\gamma(z,i,j,e))\,B(v_i,v_j,e)\,de}{(1+C_{\gamma,max})\,C_{B,max}}\,\frac{\max(g_i,g_j)}{g_i+g_j-g_{min}(z)}\,, \tag{2.27}$$

otherwise the distribution of the parameter e is (2.15) (cf. (2.17))

$$\frac{(1+\gamma(z,i,j,e))\,B(v_i,v_j,e)}{\int_{S^2}(1+\gamma(z,i,j,e))\,B(v_i,v_j,e)\,de}\,, \tag{2.28}$$

and the new state is $\tilde{z} = J(z,i,j,e)$ as defined in (2.8).

Now we shall consider some special cases. For the Boltzmann equation (1.1) with the collision kernel (1.3) we obtain for the constant $C_{B,max}$ (cf. (2.19))

$$\begin{aligned}\int_{S^2} B(v_i,v_j,e)\,de &= \frac{1}{2\sqrt{2}\pi\kappa}\int_{S^2}|(v_i-v_j,e)|\,de = \\ &= \frac{|v_i-v_j|}{2\sqrt{2}\pi\kappa}\int_0^{2\pi}\int_0^{\pi}|\cos\theta|\sin\theta\,d\theta\,d\varphi = \frac{|v_i-v_j|}{\sqrt{2}\kappa} \le \frac{U_\ell}{\sqrt{2}\kappa} = C_{B,max}\,,\end{aligned} \tag{2.29}$$

where U_ℓ denotes the maximum relative velocity in the cell Ω_ℓ.

Consider the **special case**

$$g_i = \text{const} = g \qquad \text{and} \qquad \gamma = 0\,.$$

From (2.16) we obtain

$$G(z,i,j,e) = g\,. \tag{2.30}$$

We then have (cf. (2.20), (2.23))

$$g_{min}(z) = g\,, \qquad g_{sum}(z) = n\,g\,, \tag{2.31}$$

and the waiting time parameter can be computed according to (2.22), (2.18), (2.29), (2.31) as

$$\hat{\pi}(z) = \frac{1}{2\sqrt{2}\kappa} \frac{U_\ell}{|\Omega_\ell|} g\, n\, (n-1)\,.$$

The deterministic time counter (2.12) is then nothing other than Bird's well-known "no time counter"

$$\hat{\tau}(z) = \hat{\pi}(z)^{-1} = \frac{2\sqrt{2}\kappa\, |\Omega_\ell|}{g\, n\, (n-1)\, U_\ell}\,.$$

The parameters i and j are distributed uniformly (cf. (2.24)). The jump is fictitious with probability (cf. (2.27), (2.29))

$$1 - \frac{|v_i - v_j|}{U_\ell}\,.$$

The vector e is distributed on the surface of the unit sphere S^2 according to (2.28), i.e.

$$\frac{B(v_i, v_j, e)}{\int_{S^2} B(v_i, v_j, e)\, de}\,. \tag{2.32}$$

There is no increase in the number of particles in the system. The particles for $k = i$ and $k = j$ in (2.8) have zero weights according to (2.30) and should therefore be removed from the system.

Consider the **second special case** where the weights of particles are different but the parameter γ is still considered to be zero,

$$g_i - \text{arbitrary} \qquad \text{and} \qquad \gamma = 0\,.$$

From (2.16) we obtain

$$G(z, i, j, e) = \min(g_i, g_j) \tag{2.33}$$

and from (2.22), (2.29)

$$\hat{\tau}(z) = \hat{\pi}(z)^{-1} = \frac{2\sqrt{2}\kappa\, |\Omega_\ell|}{(n-1)[2\, g_{sum}(z) - n\, g_{min}(z)]\, U_\ell}$$

for the deterministic time counter (2.12). The parameters i and j are distributed according to (2.24). The jump is fictitious with probability (cf. (2.27), (2.29))

$$1 - \frac{|v_i - v_j|}{U_\ell} \frac{\max(g_i, g_j)}{g_i + g_j - g_{min}(z)}\,.$$

The vector e is distributed according to (2.32).

The number of particles increases by one in each collision with unequal weights, according to (2.8) and (2.33). If all initial particles and all inflow particles have the same weight then this case is identical to the previous one. Here we would like to point out that our SWPM is a generalization of Bird's DSMC method.

In the **third special case** we choose the constant γ unequal zero in one cell Ω_ℓ during the fixed time interval $[t_k, t_{k+1}]$, i.e. γ is independent of i, j and e,

$$g_i - \text{arbitrary} \qquad \text{and} \qquad \gamma = \text{const} > 0\,.$$

From (2.16) we obtain

$$G(z,i,j,e) = \frac{\min(g_i, g_j)}{1+\gamma}\,,$$

and from (2.22), (2.29)

$$\hat{\tau}(z) = \hat{\pi}(z)^{-1} = \frac{1}{1+\gamma}\,\frac{2\sqrt{2}\kappa\,|\Omega_\ell|}{(n-1)[2\,g_{sum}(z) - n\,g_{min}(z)]\,U_\ell}$$

for the deterministic time counter. All other parameters of the process remain the same.

In this case the number of particles increases by two in each collision. This procedure can be used efficiently for reducing stochastic fluctuations arising in computation of the macroscopic quantities in low particle density regions, as we showed in [15].

But the new small particles move and will probably reach the region where the particle density is normal. There it is necessary to use the second special case (2.33) for the collisions, which means the number of particles will increase further without any advantage being gained. The best situation is, of course, if the particles disappear through the "outflow boundary" of the computational domain at a rate corresponding to the "production rate" there. In such a situation we will still be dealing with an asymptotically constant number of particles, but with more small particles in the low density regions (this is our improvement) which are on the way to the "outflow boundary" (this is the price).

There are certainly many situations when the number of particles should be reduced during the calculations. For example, if we solve a problem in a closed computational domain then we have no chance for outflow. How should reduction be organized? How large is the additional error due to the reduction procedure? How much additional work will be required? We will try to answer these questions in the next section.

3. Reduction of the number of particles

Suppose the following system of particles is given

$$(g_i, x_i, v_i)\,, \qquad i = 1, \ldots, n\,, \tag{3.1}$$

where the number of particles n is too large and should be reduced. Thus the objective is to construct a new system

$$(\tilde{g}_i, \tilde{x}_i, \tilde{v}_i)\,, \qquad i = 1, \ldots, \tilde{n}\,, \qquad \tilde{n} < n\,, \tag{3.2}$$

having far fewer particles but such that the corresponding empirical measures still approximate the solution of the Boltzmann equation.

In fact, there are two problems. The first one is dividing the system (3.1) into a number $\hat{n}$ of groups or clusters

$$(g_{i,j}, x_{i,j}, v_{i,j})\,, \qquad i = 1, \ldots, \hat{n}\,, \qquad j = 1, \ldots, n_i\,, \tag{3.3}$$

with

$$\sum_{i=1}^{\hat{n}} n_i = n\,.$$

We will deal with this problem in subsection 3.4.

The second problem is replacing each cluster having $n_i \geq 3$ by few particles and in the simplest case by two particles

$$(\tilde{g}_{i,1}, \tilde{x}_{i,1}, \tilde{v}_{i,1})\,, \qquad (\tilde{g}_{i,2}, \tilde{x}_{i,2}, \tilde{v}_{i,2})\,, \qquad i:\ n_i \geq 3\,. \tag{3.4}$$

The new number of particles after reduction becomes

$$\tilde{n} \leq 2\hat{n}\,.$$

There are two things we have to consider in reduction: the conservation of the macroscopic quantities and control over the additional error.

Let us introduce the following notations for a cluster i having more than three particles:

$$g^{(i)} = \sum_{j=1}^{n_i} g_{i,j}\,, \tag{3.5}$$

for the mass of the cluster,

$$g^{(i)}V^{(i)} = \sum_{j=1}^{n_i} g_{i,j}v_{i,j}\,, \tag{3.6}$$

for the momentum of the cluster,

$$g^{(i)}M^{(i)} = \sum_{j=1}^{n_i} g_{i,j}v_{i,j}v_{i,j}^T\,, \tag{3.7}$$

for the flow of the momentum of the cluster,

$$g^{(i)}E^{(i)} = g^{(i)}\mathrm{tr}M^{(i)} = \sum_{j=1}^{n_i} g_{i,j}\|v_{i,j}\|^2\,, \tag{3.8}$$

$$\varepsilon^{(i)} = \sqrt{E^{(i)} - \|V^{(i)}\|^2}, \tag{3.9}$$

$$q^{(i)} = \frac{1}{2}\sum_{j=1}^{n_i} g_{i,j}(v_{i,j} - V^{(i)})\|v_{i,j} - V^{(i)}\|^2, \tag{3.10}$$

for the heat flux vector of the cluster.

We can easily see that if we conserve only those quantities $g^{(i)}$, $g^{(i)}V^{(i)}$ and $E^{(i)}$ which correspond to the conservation laws of the Boltzmann equation, then the simplest choice of the pair (3.4) is the following:

$$\tilde{g}_{i,1} = \tilde{g}_{i,2} = g^{(i)}/2\,, \tag{3.11}$$

$$\tilde{v}_{i,1} = V^{(i)} + \varepsilon^{(i)}e\,, \qquad \tilde{v}_{i,2} = V^{(i)} - \varepsilon^{(i)}e\,, \qquad e \in S^2\,. \tag{3.12}$$

The positions of new particles (3.4) $\tilde{x}_{i,1}$, $\tilde{x}_{i,2}$ can be randomly chosen from the old set of positions

$$X_i = \{x_{i,j},\ j = 1, \ldots, n_i\}\,. \tag{3.13}$$

Note that we do not use all degrees of freedom now, i.e. we choose two new particles of equal weights and randomly choose a vector e on the unit sphere. Here we have three additional degrees of freedom which can be used in different ways. In [19] the author requires the conservation of all main diagonal components of the flow of momentum (3.7) instead of the trace. By doing so the vector e can be defined (except for the sign of the single components) as follows:

$$e_k = \pm\frac{1}{\varepsilon^{(i)}}\sqrt{M^{(i)}_{kk} - \left[V^{(i)}_k\right]^2}, \quad k = 1, 2, 3\,. \tag{3.14}$$

The weights of the particles remain equal.

3.1. Conservation of the heat flux

In the following we show how to choose the pair of particles (3.4) using all possible degrees of freedom in order to conserve not only invariants of the collision integral but also the heat flux vector as defined in (3.10).

Let us choose the velocities of the particles (3.4) in the form

$$\tilde{v}_{i,1} = V^{(i)} + \alpha e\,, \qquad \tilde{v}_{i,2} = V^{(i)} - \beta e\,, \qquad e \in S^2\,, \tag{3.15}$$

where α and β are positive numbers. From (3.5)-(3.10) we obtain

$$\tilde{g}_{i,1} + \tilde{g}_{i,2} = g^{(i)}, \tag{3.16}$$

$$\tilde{g}_{i,1}\alpha - \tilde{g}_{i,2}\beta = 0\,, \tag{3.17}$$

$$\tilde{g}_{i,1}\alpha^2 + \tilde{g}_{i,2}\beta^2 = g^{(i)}\left(\varepsilon^{(i)}\right)^2, \tag{3.18}$$

$$\left(\tilde{g}_{i,1}\alpha^3 - \tilde{g}_{i,2}\beta^3\right)e = 2\,q^{(i)}. \tag{3.19}$$

From (3.19) it is clear that if $q^{(i)} \neq 0$ then vector e should be chosen as

$$e = \frac{q^{(i)}}{\|q^{(i)}\|} .$$

If $q^{(i)} = 0$ then vector e can be chosen randomly on the surface of the unit sphere S^2 or corresponding to (3.14).

We now solve the system (3.16)-(3.19) using the notation

$$\alpha = \theta \varepsilon^{(i)} , \qquad \theta \geq 1 . \tag{3.20}$$

From (3.17) we first obtain

$$\beta = \frac{\tilde{g}_{i,1}}{\tilde{g}_{i,2}} \alpha . \tag{3.21}$$

Then using (3.18), (3.21) and (3.20) we get

$$\begin{aligned} \tilde{g}_{i,1}\alpha^2 + \tilde{g}_{i,2}\beta^2 &= \tilde{g}_{i,1}\alpha^2 + \tilde{g}_{i,2}\frac{\tilde{g}_{i,1}^2}{\tilde{g}_{i,2}^2}\alpha^2 = \tilde{g}_{i,1}\theta^2 \left(\varepsilon^{(i)}\right)^2 + \frac{\tilde{g}_{i,1}^2}{\tilde{g}_{i,2}}\theta^2 \left(\varepsilon^{(i)}\right)^2 \\ &= \frac{\tilde{g}_{i,1}}{\tilde{g}_{i,2}}\theta^2 \left(\varepsilon^{(i)}\right)^2 (\tilde{g}_{i,1} + \tilde{g}_{i,2}) = g^{(i)} \left(\varepsilon^{(i)}\right)^2 \frac{\tilde{g}_{i,1}}{\tilde{g}_{i,2}}\theta^2 = g^{(i)} \left(\varepsilon^{(i)}\right)^2 . \end{aligned} \tag{3.22}$$

Thus, (3.22), (3.16) and (3.21) yield

$$\tilde{g}_{i,1} = g^{(i)} \frac{1}{1+\theta^2} , \qquad \tilde{g}_{i,2} = g^{(i)} \frac{\theta^2}{1+\theta^2} , \tag{3.23}$$

$$\beta = \frac{\varepsilon^{(i)}}{\theta} . \tag{3.24}$$

All unknowns are now represented by θ. If we put (3.20), (3.23) and (3.24) in (3.19) then we obtain the final equation for θ

$$\begin{aligned} \tilde{g}_{i,1}\alpha^3 - \tilde{g}_{i,2}\beta^3 &= g^{(i)} \frac{1}{1+\theta^2} \theta^3 \left(\varepsilon^{(i)}\right)^3 - g^{(i)} \frac{\theta^2}{1+\theta^2} \frac{\left(\varepsilon^{(i)}\right)^3}{\theta^3} \\ &= g^{(i)} \frac{\left(\varepsilon^{(i)}\right)^3}{1+\theta^2} \left(\theta^3 - \frac{1}{\theta}\right) = g^{(i)} \frac{\left(\varepsilon^{(i)}\right)^3}{\theta} \left(\theta^2 - 1\right) = 2\,\|q^{(i)}\| , \end{aligned}$$

or

$$\theta^2 - 2\frac{\|q^{(i)}\|}{g^{(i)} \left(\varepsilon^{(i)}\right)^3} \theta - 1 = 0 . \tag{3.25}$$

The equation (3.25) is always solvable and only one of its solutions, namely

$$\theta^{(i)} = \frac{\|q^{(i)}\|}{g^{(i)} \left(\varepsilon^{(i)}\right)^3} + \sqrt{1 + \frac{\|q^{(i)}\|^2}{\left(g^{(i)}\right)^2 \left(\varepsilon^{(i)}\right)^6}} \tag{3.26}$$

fulfils the condition (3.20).

Note that if $q^{(i)} = 0$ we will automatically obtain the simplest solution (3.11), (3.12).

3.2. Lipschitz metric

In this subsection we give a brief summary of the results published in [14]. We consider the bounded Lipschitz metric as a distance between two measures $\nu_1(dx, dv)$ and $\nu_2(dx, dv)$ defined as

$$\varrho(\nu_1, \nu_2) = \sup_{\|\varphi\|_L \leq 1} \left| \int_{\Omega \times \mathbb{R}^3} \varphi(x, v)\, \nu_1(dx, dv) - \int_{\Omega \times \mathbb{R}^3} \varphi(x, v)\, \nu_2(dx, dv) \right|, \tag{3.27}$$

where

$$\|\varphi\|_L = \max \left(\sup_{(x,v)} |\varphi(x, v)|, \sup_{(x,v) \neq (y,w)} \frac{|\varphi(x, v) - \varphi(y, w)|}{\|x - y\| + \|v - w\|} \right).$$

The main result is the following lemma.

Lemma 1 *Let (3.3) be a given system of particles in a cluster and the particles*

$$(\tilde{g}_{i,1}, \tilde{x}_{i,1}, \tilde{v}_{i,1}), \qquad (\tilde{g}_{i,2}, \tilde{x}_{i,2}, \tilde{v}_{i,2})$$

be chosen according to (3.11), (3.12). Then for the bounded Lipschitz metric (3.27) between the measures

$$\mu^{(i)} = \sum_{j=1}^{n_i} g_{i,j} \delta_{(x_{i,j}, v_{i,j})} \tag{3.28}$$

and

$$\tilde{\mu}^{(i)} = \frac{g^{(i)}}{2} \left(\delta_{(\tilde{x}_{i,1}, \tilde{v}_{i,1})} + \delta_{(\tilde{x}_{i,2}, \tilde{v}_{i,2})} \right) \tag{3.29}$$

the following estimate is valid

$$\varrho(\mu^{(i)}, \tilde{\mu}^{(i)}) \leq 2\, g^{(i)} \left(\varepsilon^{(i)} + \operatorname{diam}(X_i) \right),$$

where $\varepsilon^{(i)}$ and X_i are defined in (3.9) and (3.13), respectively.

Using the triangle inequality we obtain the corresponding result for the whole systems (3.1) and (3.2):

$$\varrho(\mu, \tilde{\mu}) \leq 2 \sum_{i=1}^{\hat{n}} g^{(i)} \left(\varepsilon^{(i)} + \operatorname{diam}(X_i) \right). \tag{3.30}$$

On the other hand, with a similar technique for the reduction procedure (3.15), (3.20), (3.23), (3.24), (3.26) we obtain an estimate

$$\varrho(\mu, \tilde{\mu}) \leq \sum_{i=1}^{\hat{n}} g^{(i)} \left(\left[1 + \frac{\theta^{(i)}}{1 + (\theta^{(i)})^2} \right] \varepsilon^{(i)} + 2 \operatorname{diam}(X_i) \right), \tag{3.31}$$

which is slightly better than the previous one.

Note that the dependence of the particles (3.4) on the choice of the unit vector e is lost in both estimates (3.30) and (3.31). This means that for the reduction technique corresponding to (3.14) the estimate (3.30) holds too. We would like to neglect the error relating to the influence of the space distribution, because we assume that the value $\mathrm{diam}(X_i)$ is small enough already. On the other hand the estimates (3.30), (3.31) show us the possibility of clustering the particles. The clusters have to be chosen so that the product of the mass of the cluster $g^{(i)}$ with its "temperature" $\varepsilon^{(i)}$ is small. The corresponding discussion can be found in subsection 3.4.

3.3. Sobolev norm

In this subsection we use a different distance between the measures (3.28), (3.29) which is the norm in the Sobolev space $\mathbb{H}^{-2}$. The equivalence of the weak* convergence of the measures and of the convergence in the Sobolev norms $\mathbb{H}^s$, $s < -d/2$, where d denotes the space dimension ($d = 3$ in our case), was proved in [20].

Let us first introduce some notations which are needed. If $\mu(dv)$ is a measure then the complex-valued function

$$\hat{\mu}(\xi) = \int_{\mathbb{R}^3} \exp(\imath(\xi, v))\mu(dv)$$

is called the Fourier transformation of the measure $\mu(dv)$. The Sobolev norm of this measure is then defined by

$$\|\mu\|_s^2 = \int_{\mathbb{R}^3} \left(1 + |\xi|^2\right)^s |\hat{\mu}(\xi)|^2 \, d\xi \,.$$

In this subsection we neglect the error due to the spatial distribution of the particles and compute only the Sobolev norm of the difference between the measures μ and $\tilde{\mu}$ (cf. (3.28), (3.29)) defined by the systems of the particles

$$\Big((g_1, v_1), \ldots, (g_n, v_n)\Big) \quad \text{and} \quad \Big((g/2, V + \varepsilon e)\,, (g/2, V - \varepsilon e)\Big)\,,$$

where g, V and ε are defined corresponding to (3.5)-(3.8). We do not use the cluster index i in this subsection in order not to overload the formulae, bearing in mind that all the things we consider here will have to be summed up later for all clusters.

Lemma 2 *The Sobolev norm of the difference of the measures μ and $\tilde{\mu}$ in $\mathbb{H}^{-2}$ is given by*

$$\begin{aligned} \|\mu - \tilde{\mu}\|_{-2}^2 \;=\; & \frac{1}{8}\Bigg((1 + \exp(-2\varepsilon))g^2 + 2\sum_{k,l=1}^{n} g_k g_l \exp(-|v_k - v_l|) - \\ & -2\,g \sum_{k=1}^{n} g_k \left[\exp(-|v_k - V - \varepsilon e|) + \exp(-|v_k - V + \varepsilon e|)\right] \Bigg)\,. \end{aligned} \tag{3.32}$$

Proof : We begin the proof by computing the Fourier transformation of the measures μ and $\tilde{\mu}$:

$$\begin{aligned}
\hat{\mu}(\xi) &= \int_{\mathbb{R}^3} \exp(\imath(\xi, v))\, \mu(dv) = \sum_{j=1}^{n} g_j \exp(\imath(\xi, v_j))\,, \\
\hat{\tilde{\mu}}(\xi) &= \int_{\mathbb{R}^3} \exp(\imath(\xi, v))\, \tilde{\mu}(dv) = \frac{g}{2} \exp(\imath(\xi, V - \varepsilon e)) + \frac{g}{2} \exp(\imath(\xi, V + \varepsilon e)) \\
&= \sum_{j=1}^{n} g_j \left(\frac{1}{2} \exp(\imath(\xi, V - \varepsilon e)) + \frac{1}{2} \exp(\imath(\xi, V + \varepsilon e)) \right).
\end{aligned}$$

Thus we obtain

$$\begin{aligned}
\left| \hat{\mu} - \hat{\tilde{\mu}} \right| &= \sum_{k,l=1}^{n} g_k g_l \left(\cos(\xi, v_k - v_l) - \frac{1}{2} \cos(\xi, v_k - V - \varepsilon e) - \right. \\
&\quad - \frac{1}{2} \cos(\xi, v_k - V + \varepsilon e) - \frac{1}{2} \cos(\xi, v_l - V - \varepsilon e) - \\
&\quad \left. - \frac{1}{2} \cos(\xi, v_l - V + \varepsilon e) + \frac{1}{2} + \frac{1}{2} \cos(\xi, 2\varepsilon e) \right).
\end{aligned} \tag{3.33}$$

Therefore it is necessary to compute the integral

$$\int_{\mathbb{R}^3} \frac{\cos(\xi, u)}{\left(1 + |\xi|^2\right)^2}\, d\xi$$

for various u involved in (3.33). We use the spherical coordinates $(\varrho, \varphi, \theta)$ whereby the $z-$axis has the same direction as u. Using $(\xi, u) = |\xi||u| \cos\theta = \varrho\alpha(\theta)$ we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^3} \frac{\cos(\xi, u)}{\left(1 + |\xi|^2\right)^2}\, d\xi = 2 \int_0^{2\pi} d\varphi \int_0^{\pi/2} \sin\theta d\theta \int_0^{\infty} \frac{\varrho^2 \cos(\varrho\alpha(\theta))}{\left(1 + \varrho^2\right)^2}\, d\varrho = \\
&= 4\pi \int_0^{\pi/2} \sin\theta d\theta \frac{1}{2} \operatorname{Re} \int_{\mathcal{D}} \frac{z^2 \exp(\imath\alpha(\theta) z)}{\left(1 + z^2\right)^2} dz = \\
&= 4\pi \int_0^{\pi/2} \pi\imath \operatorname{Res} \left[\frac{z^2 \exp(\imath\alpha(\theta) z)}{\left(1 + z^2\right)^2}, \imath \right] \sin\theta d\theta = \\
&= 4\pi \int_0^{\pi/2} \frac{\pi}{4} (1 - \alpha(\theta)) \exp(-\alpha(\theta)) \sin\theta d\theta = \\
&= \pi^2 \int_0^{\pi/2} (1 - |u| \cos\theta) \exp(-|u| \cos\theta) \sin\theta d\theta = \pi^2 \exp(-|u|)\,.
\end{aligned}$$

If we use this result for the values $u = v_k - v_l$, $u = v_k - V - \varepsilon e$, $u = v_k - V + \varepsilon e$, $u = v_l - V - \varepsilon e$, $u = v_l - V + \varepsilon e$, $u = 0$ and $u = 2\varepsilon e$ we obtain from (3.33) the formula (3.32). □

The main advantage of the distance (3.32) is, of course, that this formula is exact, and the dependence of the distance on the vector e is shown clearly in the third term. On the other

hand, this formula includes as the second term a double sum, which requires a numerical work of the order n_i^2 in the cluster i having n_i elements. Note that our aim is not to produce only few clusters of many elements but rather to produce many clusters with 4-5 particles in each which should be replaced by two. In such a situation the whole work required for computing all distances corresponding to (3.32) remains of the capital order $O(n)$. The next observation is that the unit vector e is only involved in the third term of the formula (3.32) which requires $O(n_i)$ numerical work. Our idea now is to try to maximize the third term in (3.32) in order to minimize the distance between both measures. Unfortunately, the dependence of the Sobolev distance on the vector e is very complicated, having a lot of local minima and maxima. We would like to illustrate this behaviour using the following example. We randomly generate 128 particles corresponding to the distribution $f_0(v)$ (see Section 4) and compute the vector e via (3.14). This vector is defined by

$$e = \Big(\cos(\varphi)\sin(\theta), \sin(\varphi)\sin(\theta), \cos(\theta)\Big)^T, \quad 0 \le \varphi < 2\pi, \quad 0 \le \theta \le \pi.$$

In our example we obtain $\varphi \approx 0.479$ and $\theta \approx 1.094$. We now fix the value of θ and plot the Sobolev distance as a function of φ. The result is shown in **Figure 1**. It is to be concluded that we will not have a chance to determine the optimal value of φ numerically because of the presence of many local extrema. On the other hand, the dependence of the Sobolev distance on e is rather weak. We will see in Section 4 that clustering particles correctly is much more important than the choice of the vector e.

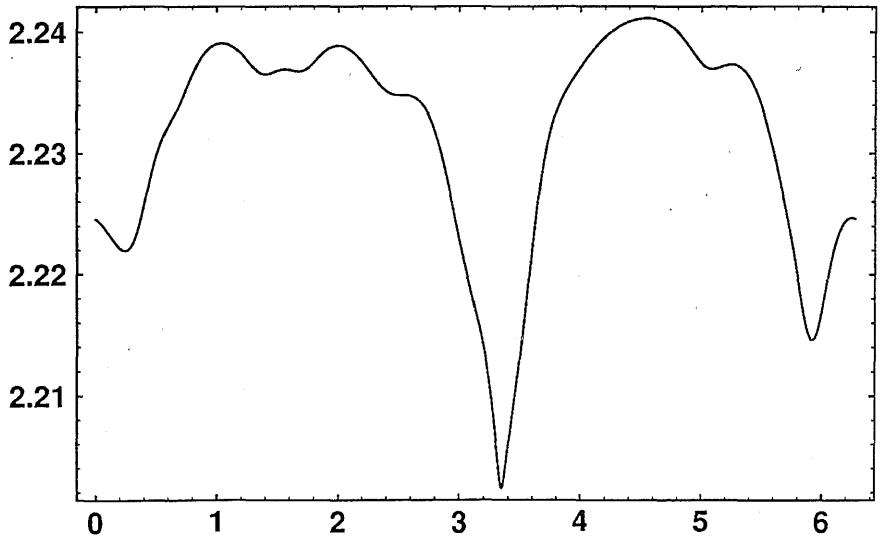

Figure 1: Sobolev distance via polar angle φ

3.4. Clustering the particles

Clustering means grouping similar objects by minimizing a certain criterion function or other object-dependent properties. Clustering techniques are very common and useful in many applications such as data analysis, data reduction, digital image processing, pattern recognition and computer graphics. In the past many algorithms have been developed (see, e.g., [1], [5], [6], [7], [13], [17], [18], [21]).

In this section we propose a solution to the problem stated earlier: finding a way to partition the system of particles (3.1). As mentioned before, however, we are not interested in position. Therefore the problem can be reduced to finding a set of clusters such that for each cluster $i = 1, \ldots, \hat{n}$ the quantity

$$\varrho_i = g^{(i)}\, \varepsilon^{(i)}\,, \tag{3.34}$$

i.e. the product of the cluster mass and the cluster temperature, is minimized (cf. (3.30), (3.5), (3.9)). In addition, all ϱ_i should be nearly equal and lower than a given ϱ. Thus the resulting number of clusters $\hat{n}$ will depend on ϱ.

However, clustering is known to be np-complete. Our intention here is not to find a method which is as close as possible to the global optimum but to find an appropriate method which is both acceptable for the problem we are faced with and efficient enough to run several times on large data bases.

In the following we propose a solution which is related to the method introduced by Orchard and Bouman [13]. It is based on a hierarchical binary space subdivision and constrains the partitioning to have the structure of a binary tree. Each node of the tree represents a subset, and the children of any node partition the members of the parent node. The method of generating the binary tree is specified by ϱ and by the method of splitting a node into its two children. The algorithm starts with the whole data set in the root of the tree and partitions each node until the quantity (3.34) is lower than ϱ.

In order to limit the complexity of the splitting algorithm, a splitting plane is used. In the algorithm proposed we determine the direction in which the cluster variation is greatest, and then split the cluster with a plane perpendicular to that direction through the cluster mean. More specifically, we determine the cluster covariance

$$R^{(i)} = M^{(i)} - V^{(i)}\,[V^{(i)}]^T$$

(cf. (3.6), (3.7)), where $V^{(i)}$ is the cluster mean. The normal direction of the splitting plane is parallel to the eigenvector corresponding to the largest eigenvalue of $R^{(i)}$. Note that (cf. (3.8), (3.9))

$$\mathrm{tr}R^{(i)} = \mathrm{tr}M^{(i)} - \|V^{(i)}\|^2 = E^{(i)} - \|V^{(i)}\|^2 = \left[\varepsilon^{(i)}\right]^2 .$$

4. Numerical experiments

It is clear that reducing the number of particles produces an additional error in the computational process. From the theoretical point of view this error can be held in check by the estimates (3.30), (3.31). From the practical point of view it is extremely important to investigate this additional error very carefully in order to be sure that the error due to reduction algorithms does not become dominant in the computations. Since the numerical solution of the spatially inhomogeneous Boltzmann equation is always faced with different kinds of discretization errors, i.e. discretization of the computational domain, splitting free flow and collision phases, sorting the particles in spatial cells, finite (and usually small) number of particles per cell, etc., it is difficult to check the additional effect of reduction, especially if we would like to compare different reduction strategies.

In our opinion is better for our purpose to solve the spatially homogeneous Boltzmann equation, i.e. to model the situation in one spatial cell. It is also useful to choose the collision kernel which corresponds to pseudo-Maxwell molecules because in this case exact formulae for the time development of the moments are known even for non-trivial initial distributions (cf. [3]).

We consider the problem of calculating the second moments

$$m_{i,j}(t) = \int_{\mathbb{R}^3} v_i v_j f(t,v)\,dv\,, \quad i,j = 1,2,3\,, \tag{4.1}$$

and the third moments

$$r_i(t) = \int_{\mathbb{R}^3} v_i \|v\|^2 f(t,v)\,dv\,, \quad i = 1,2,3\,. \tag{4.2}$$

The stochastic weighted particle method described in Section 2 is used with the parameter $\gamma = 1$ (cf. (2.16)). This means that during each collision two additional particles are created. The initial distribution is a mixture of two Maxwellians, namely

$$f_0(v) = \frac{1}{2}\frac{1}{(2\pi T_1)^{3/2}} \exp\left(-\frac{\|v-V_1\|^2}{2T_1}\right) + \frac{1}{2}\frac{1}{(2\pi T_2)^{3/2}} \exp\left(-\frac{\|v-V_2\|^2}{2T_2}\right),$$

with

$$V_1 = (2,0,0)\,, \quad V_2 = (-2,0,0)\,, \quad T_1 = 2\,, \quad T_2 = 1\,.$$

4.1. Statistical notions

First we introduce some definitions and notations that are helpful for the understanding of stochastic numerical procedures.

The functionals to be calculated (4.1), (4.2) are of the form

$$F(t) = \int_{\mathbb{R}^3} \varphi(v) f(t,v)\,dv\,. \tag{4.3}$$

According to (2.1), a functional (4.3) is approximated by the random variable

$$\xi(t) = \int_{\mathbb{R}^3} \varphi(v)\,\mu(t,dv) = \sum_{i=1}^{n(t)} g_i(t)\,\varphi(v_i(t))\,. \tag{4.4}$$

Note that this random variable depends on the value $n = n(0)$, which determines the quality of approximation of the initial distribution by means of a point measure.

In order to estimate and to reduce the random fluctuations of the estimator (4.4), a number N of independent ensembles of particles is generated. The corresponding values of the random variable are denoted by

$$\xi_1^{(n)}(t), \ldots, \xi_N^{(n)}(t)\,.$$

The **empirical mean value** of the random variable (4.4)

$$\eta_1^{(n,N)}(t) = \frac{1}{N}\sum_{j=1}^{N} \xi_j^{(n)}(t) \tag{4.5}$$

is then used as an approximation to the functional (4.3). The error of this approximation is

$$e^{(n,N)}(t) = |\eta_1^{(n,N)}(t) - F(t)| \tag{4.6}$$

and consists of the following two components.

The **systematic error** is the difference between the mathematical expectation of the random variable (4.4) and the exact value of the functional, i.e.

$$e_{sys}^{(n)}(t) = E\xi^{(n)}(t) - F(t)\,.$$

The **statistical error** is the difference between the empirical mean value and the expected value of the random variable, i.e.

$$e_{stat}^{(n,N)}(t) = \eta_1^{(n,N)}(t) - E\xi^{(n)}(t)\,.$$

A **confidence interval** for the expectation of the random variable $\xi^{(n)}(t)$ is obtained as

$$I_p = \left[\eta_1^{(n,N)}(t) - \lambda_p\sqrt{\frac{\operatorname{Var}\xi^{(n)}(t)}{N}}\,,\ \eta_1^{(n,N)}(t) + \lambda_p\sqrt{\frac{\operatorname{Var}\xi^{(n)}(t)}{N}}\right],$$

where

$$\operatorname{Var}\xi^{(n)}(t) := E\left[\xi^{(n)}(t) - E\xi^{(n)}(t)\right]^2 = E\left[\xi^{(n)}(t)\right]^2 - \left[E\xi^{(n)}(t)\right]^2 \tag{4.7}$$

is the **variance** of the random variable (4.4), and $p \in (0,1)$ is the **confidence level**. This means that

$$\operatorname{Prob}\left\{E\xi^{(n)}(t) \notin I_p\right\} = \operatorname{Prob}\left\{|e_{stat}^{(n,N)}(t)| \geq \lambda_p\sqrt{\frac{\operatorname{Var}\xi^{(n)}(t)}{N}}\right\} \sim p\,.$$

Thus, the value

$$c^{(n,N)}(t) = \lambda_p \sqrt{\frac{\operatorname{Var} \xi^{(n)}(t)}{N}} \tag{4.8}$$

is a probabilistic upper bound for the statistical error.

In the calculations we use a confidence level of $p = 0.999$ and $\lambda_p = 3.2$. The variance is approximated by the corresponding empirical value (cf. (4.7)), i.e.

$$\operatorname{Var} \xi^{(n)}(t) \sim \eta_2^{(n,N)}(t) - \left[\eta_1^{(n,N)}(t)\right]^2,$$

where

$$\eta_2^{(n,N)}(t) = \frac{1}{N} \sum_{j=1}^{N} \left[\xi_j^{(n)}(t)\right]^2$$

is the **empirical second moment** of the random variable (4.4).

4.2. Systematic error – long time behaviour

First we study the long time behaviour of the approximations (4.5) to the functionals (4.1), (4.2). We consider the time interval $[0., 30.]$.

The typical behaviour can best be observed from **Figure 2**. The exact curves are displayed by dashed lines and the confidence bands by solid lines. The stationary state is reached at about $t = 10$. A systematic error can be detected clearly up to $n = 64$.

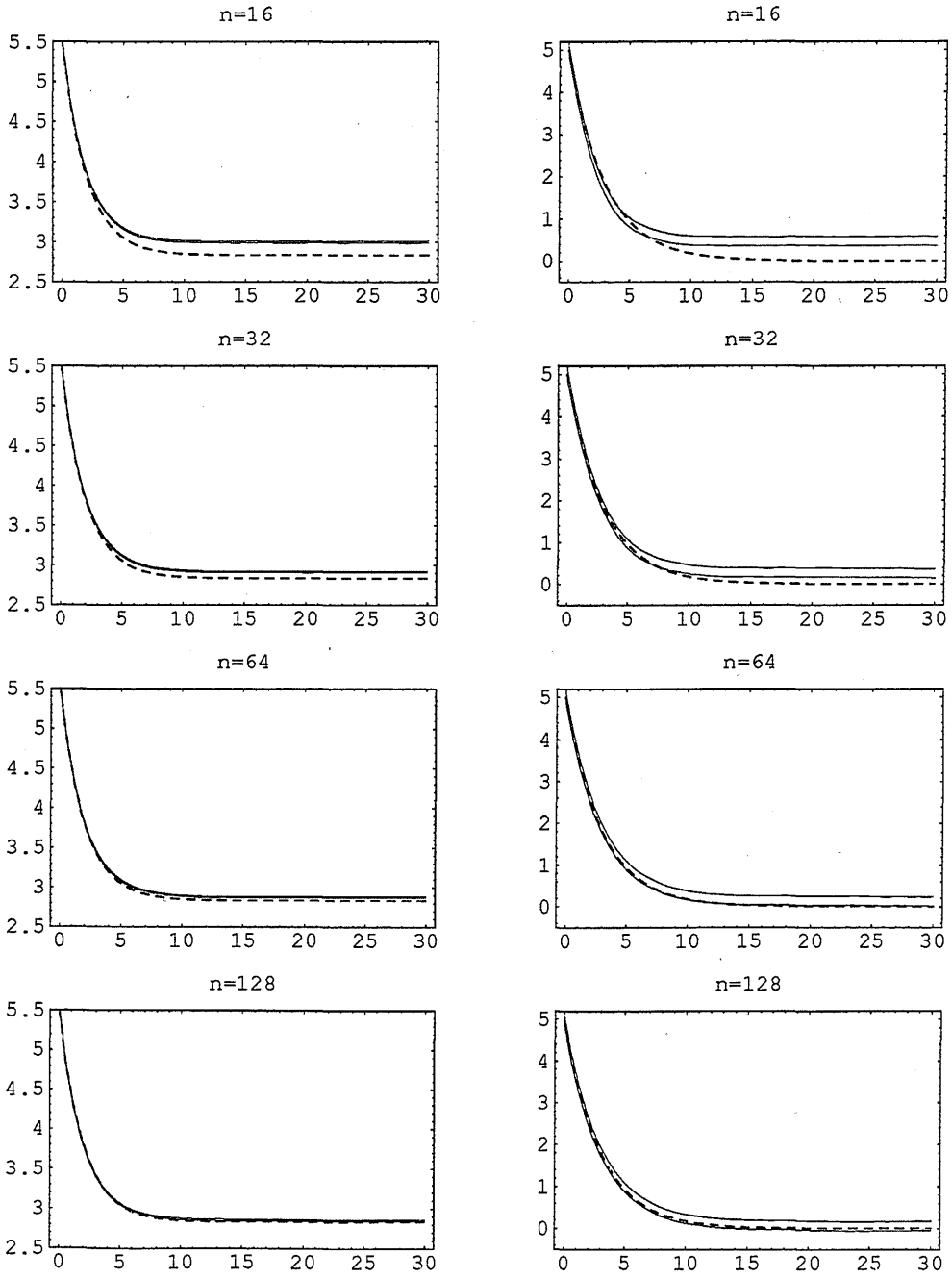

Figure 2: Moments $m_{1,1}(t)$ (left) and $r_1(t)$ (right) for different n

More complete data is contained in **Table 1**. The supremum over the time interval of the absolute error (4.6) is denoted by $err\text{-}m_{1,1}$ and $err\text{-}r_1$ for the functionals $m_{1,1}(t)$ and $r_1(t)$, respectively. The corresponding statistical error bounds (4.8) are denoted by $conf\text{-}m_{1,1}$ and $conf\text{-}r_1$. Several other quantities relevant to the stochastic particle method are also displayed. Here, *ired* denotes the number of reductions on the time interval, while *ipart* denotes the number of particles in the system averaged over 50 observation points. Finally, *gmin* and *gmax* denote the averaged minimal and maximal weights in the system.

Table 1

n	N	*ired*	*ipart*	*gmax/gmin*	$err\text{-}m_{1,1}$	$conf\text{-}m_{1,1}$	$err\text{-}r_1$	$conf\text{-}r_1$
4	256000	36.6	7.42	1.18/0.46	0.667	0.013	1.916	0.110
8	128000	47.5	16.2	1.33/0.28	0.344	0.012	1.030	0.110
16	64000	53.4	32.0	1.46/0.20	0.168	0.010	0.480	0.109
32	32000	56.6	64.9	1.59/0.14	0.084	0.009	0.276	0.107
64	16000	58.3	130.	1.71/0.10	0.044	0.009	0.142	0.108
128	8000	59.4	263.	1.93/0.07	0.023	0.008	0.073	0.108
256	4000	60.1	527.	2.13/0.05	0.016	0.008	0.053	0.107
512	2000	60.8	1054.	2.29/0.04	0.006	0.008	0.040	0.106
1024	1000	61.0	2102.	2.41/0.03	0.007	0.008	0.069	0.111

The systematic error is displayed in the logarithmic scale in **Figure 3**. Here the small points correspond to $err\text{-}m_{1,1}$ and the big points to $err\text{-}r_1$. As long as the error is larger than the statistical error bound there is clear linear behaviour (corresponding to the order n^{-1}). Inside the confidence interval the error fluctuates.

Note that the systematic error in the stochastic weighted particle method is comparable to that in the standard method (cf. (2.30)), as **Table 2** shows.

Table 2

n	N	$err\text{-}m_{1,1}$	$conf\text{-}m_{1,1}$	$err\text{-}r_1$	$conf\text{-}r_1$
8	128000	0.338	0.013	1.013	0.113
16	64000	0.166	0.012	0.497	0.113
32	32000	0.089	0.012	0.344	0.114
64	16000	0.051	0.012	0.167	0.116
128	8000	0.024	0.012	0.078	0.114
256	4000	0.015	0.012	0.100	0.114
512	2000	0.010	0.012	0.042	0.116
1024	1000	0.011	0.012	0.063	0.117

Thus, the method provides a correct approximation of the moments despite the permanent blow-up and the frequent reductions of the system. These properties are illustrated by **Figure 4**, where one single trajectory is displayed in the case $n = 128$ (cf. line 6 of Table 1).

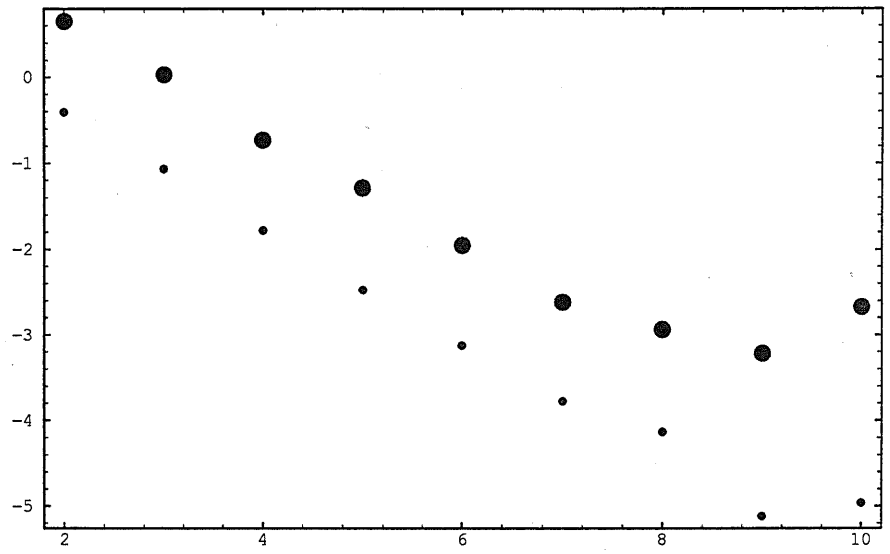

Figure 3: Systematic error dependent on n

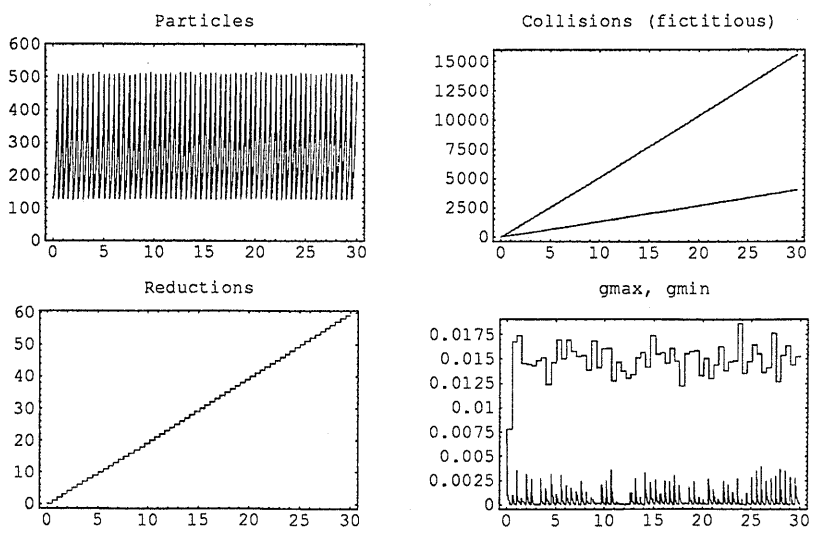

Figure 4: One trajectory for $n = 128$

4.3. Systematic error – short time behaviour

Figure 4 gives a long term picture of the behaviour of the number of particles and collisions in the system. A more precise description can be obtained by looking at the shorter time interval $[0.,3.]$. The functionals (4.1), (4.2) are calculated with the parameters $n = 10240$ and $N = 100$. If the number of particles reaches $4n$ then this number is reduced to $n/4$. **Figure 5** shows the behaviour of the number of particles, which grows exponentially up to the corresponding maximum. Thus, on a small scale, the number of collisions is not linear as Figure 4 shows on a large time scale. Despite the strong fluctuations of the number of particles in the system, the moments $m_{i,i}(t)$, $i = 1,2,3$, and $r_1(t)$ are calculated correctly. Here, as before, exact curves are displayed by dashed lines, and the confidence bands by solid lines.

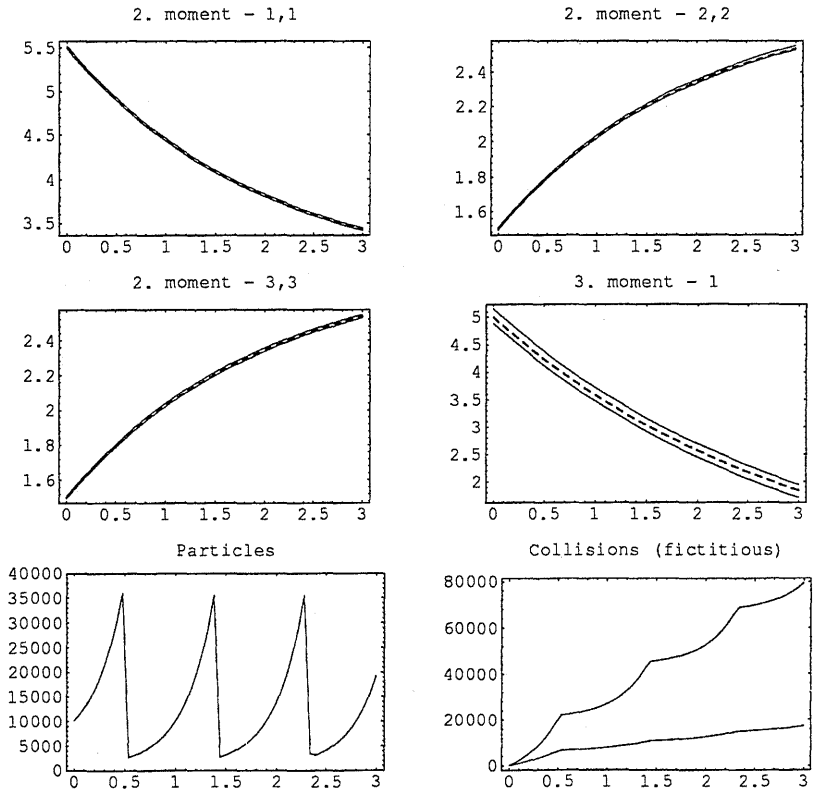

Figure 5: Short time interval ($n = 10240$)

4.4. Reduction error

Finally, we study the behaviour of the reduction error (cf. the right-hand side of (3.30)) dependent on n. During the calculation of the functionals on the time interval $[0., 3.]$ the error bounds were evaluated and averaged. We considered different reduction strategies, reducing the number of particles from $4n$ to $n/4$, $n/2$, and n, and from $2n$ to n. The corresponding values of the error are displayed in **Table 3**.

Table 3

n	4n:n/4	4n:n/2	4n:n	2n:n
16	2.234	1.815	1.436	1.322
32	1.900	1.527	1.195	1.086
64	1.572	1.274	0.994	0.895
128	1.324	1.063	0.815	0.736
1024	0.725	0.566	0.429	0.387
10240	0.350	0.269	0.202	0.183
102400	0.164	0.126	0.095	0.085

The reduction error is displayed in the logarithmic scale in **Figure 6**. Here the big points correspond to the first column of Table 3, while the small points correspond to the third column. The other columns would look similar. Figure 6 shows linear behaviour of the reduction error. The lines corresponding to different columns of Table 3 are roughly parallel. A comparison of the particular values suggests an order of convergence close to $n^{-\frac{1}{3}}$.

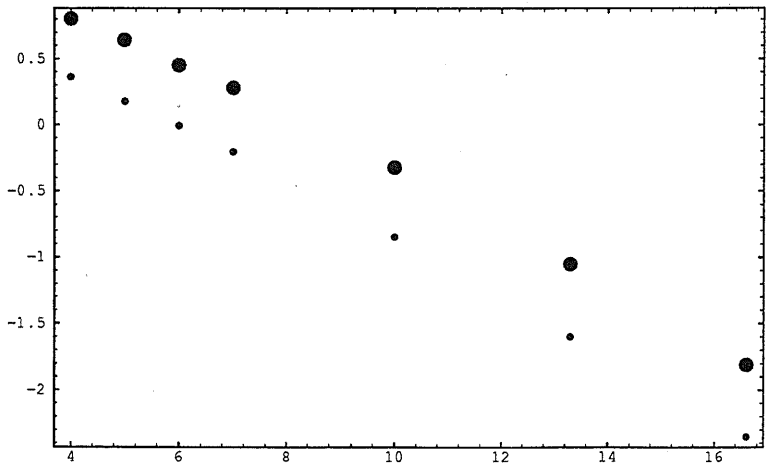

Figure 6: Reduction error dependent on n

5. Concluding remarks

In this paper we presented a detailed study of reduction procedures for the stochastic weighted particle method (SWPM). These procedures are based on appropriate clustering of the particle system in the velocity space. Different methods are provided which possess conservation properties for all physically relevant macroscopic moments. These results represent a significant, necessary improvement of the SWPM, which can now be used for calculations for long time intervals.

Theoretical error bounds have been obtained both in the bounded Lipschitz metric and in a particular Sobolev norm. These results were illustrated by detailed numerical tests for the spatially homogeneous Boltzmann equation. The convergence order with respect to the particle number n was found to be n^{-1} for the macroscopic moments. A comparison with the standard method (complete weight transfer, no reduction) shows that the SWPM not only has no additional error but also contains several useful degrees of freedom. Calculations for long time intervals (far beyond the relaxation time) show the stability of the SWPM.

Our main objective for future research is coupling the spatially inhomogeneous nonlinear Boltzmann equation with the system of Euler equations in regions of local equilibrium. In terms of numerical procedures we will face the problem of combining stochastic and deterministic algorithms. The robust determination of the coupling boundary, i.e. automatic domain decomposition, requires reliable computation of several first moments of the density function. The results obtained by stochastic particle methods are perturbed by stochastic fluctuations, especially in regions of low particle density. Here we expect a significant improvement of numerical results using the SWPM.

References

[1] N. R. ANDERBERG, *Cluster analysis for application*, Academic Press, 1973.

[2] G. A. BIRD, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Clarendon Press, Oxford, 1994.

[3] A. BOBYLEV AND S. RJASANOW, *Difference Scheme for the Boltzmann equation based on fast Fourier Transform*, European J. Mech. B Fluids, (1997). to appear.

[4] C. CERCIGNANI, R. ILLNER, AND M. PULVIRENTI, *The Mathematical Theory of Dilute Gases*, Springer, New York, 1994.

[5] R. O. DUDA AND P. E. HART, *Pattern classification and scene analysis*, Wiley, 1973.

[6] J. A. HARTIGAN, *Clustering algorithms*, Wiley, 1975.

[7] L. HYAFIL AND R. L. RIVERT, *Constructing optimal binary decision trees is np-complete*, Inf. Proc. Let., 5 (1976), pp. 15–17.

[8] R. ILLNER AND S. RJASANOW, *Numerical solution of the Boltzmann equation by random discrete velocity models*, European J. Mech. B Fluids, 13 (1994), pp. 197–210.

[9] R. ILLNER AND W. WAGNER, *A random discrete velocity model and approximation of the Boltzmann equation*, J. Statist. Phys., 70 (1993), pp. 773–792.

[10] ——, *Random discrete velocity models and approximation of the Boltzmann equation. Conservation of momentum and energy*, Transport Theory Statist. Phys., 23 (1994), pp. 27–38.

[11] M. S. IVANOV AND S. V. ROGAZINSKI, *Analysis of numerical techniques of the direct simulation Monte Carlo method in the rarefied gas dynamics*, Soviet J. Numer. Anal. Math. Modelling, 3 (1988), pp. 453–465.

[12] H. NEUNZERT, F. GROPENGIESSER, AND J. STRUCKMEIER, *Computational methods for the Boltzmann equation*, in Applied and Industrial Mathematics, R. Spigler, ed., Kluwer Acad. Publ., Dordrecht, 1991, pp. 111–140.

[13] M. T. ORCHARD AND C. A. BOUMAN, *Color optimization of images*, IEEE Trans. Sig. Proc., 39 (1991), pp. 2677–2690.

[14] S. RJASANOW AND W. WAGNER, *A stochastic weighted particle method for the Boltzmann equation*, J. Comput. Phys., 124 (1996), pp. 243–253.

[15] ——, *A generalized collision mechanism for stochastic particle schemes approximating Boltzmann type equations*, Comput. Math. Appl., (1997). to appear.

[16] ——, *Numerical study of a stochastic weighted particle method for a model kinetic equation*, J. Comput. Phys., 128 (1996), pp. 351–362.

[17] T. SCHREIBER, *A Voronoi-diagram based data reduction and approximation*, in LNCS 553, Springer, 1991, pp. 265–275.

[18] ——, *Clustering for data reduction and approximation*, in Proc. GraphiCon 93, St. Petersburg, Russia, 1993.

[19] M. SCHREINER, *Weighted particles in the finite pointset method*, Transport Theory Statist. Phys., 22 (1993), pp. 793–817.

[20] W. SCHREINER, *Partikelverfahren für kinetische Schemata zu den Euler Gleichungen*, PhD-Thesis, University of Kaiserslautern, 1994.

[21] H. SPÄTH, *Cluster analysis algorithms*, Wiley, 1980.

Recent publications of the Weierstraß–Institut für Angewandte Analysis und Stochastik

Preprints 1996

291. Vladimir G. Spokoiny: Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice.

292. Peter E. Kloeden, Eckhard Platen, Henri Schurz, Michael Sørensen: On effects of discretization on estimators of drift parameters for diffusion processes.

293. Erlend Arge, Angela Kunoth: An efficient ADI–solver for scattered data problems with global smoothing.

294. Alfred Liemant, Ludwig Brehmer: A mean field approximation for hopping transport in disordered materials.

295. Michael H. Neumann: Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations.

296. Lida V. Gilyova, Vladimir V. Shaidurov: A cascadic multigrid algorithm in the finite element method for the plane elasticity problem.

297. Oleg Lepski, Arkadi Nemirovski, Vladimir Spokoiny: On estimation of non–smooth functionals.

298. Luis Barreira, Yakov Pesin, Jörg Schmeling: On a general concept of multifractality: Multifractal spectra for dimensions, entropies, and Lyapunov exponents. Multifractal rigidity.

299. Luis Barreira, Jörg Schmeling: Any set of irregular points has full Hausdorff dimension and full topological entropy.

300. Jörg Schmeling: On the completeness of multifractal spectra.

301. Yury Kutoyants, Vladimir Spokoiny: Optimal choice of observation window for Poisson observations.

302. Sanjeeva Balasuriya, Christopher K.R.T. Jones, Björn Sandstede: Viscous perturbations of vorticity conserving flows and separatrix splitting.

303. Pascal Chossat, Frédéric Guyard, Reiner Lauterbach: Generalized heteroclinic cycles in spherically invariant systems and their perturbations.

304. Klaus R. Schneider: Decomposition and diagonalization in solving large systems.

305. Klaus Fleischmann, Achim Klenke: Convergence to a non–trivial equilibrium for two–dimensional catalytic super–Brownian motion.

306. Klaus Fleischmann, Vladimir A. Vatutin: Reduced subcritical Galton–Watson processes in a random environment.

Preprints 1997

307. Andreas Rathsfeld: On the stability of piecewise linear wavelet collocation and the solution of the double layer equation over polygonal curves.

308. Georg Hebermehl, Rainer Schlundt, Horst Zscheile, Wolfgang Heinrich: Eigen mode solver for microwave transmission lines.

309. Georg Hebermehl, Rainer Schlundt, Horst Zscheile, Wolfgang Heinrich: Improved numerical solutions for the simulation of microwave circuits.

310. Krzysztof Wilmański: The thermodynamical model of compressible porous materials with the balance equation of porosity.

311. Hans Günter Bothe: Strange attractors with topologically simple basins.

312. Krzysztof Wilmański: On the acoustic waves in two–component linear poroelastic materials.

313. Peter Mathé: Relaxation of product Markov chains on product spaces.

314. Vladimir Spokoiny: Testing a linear hypothesis using Haar transform.

315. Dietmar Hömberg, Jan Sokołowski: Optimal control of laser hardening.

316. Georg Hebermehl, Rainer Schlundt, Horst Zscheile, Wolfgang Heinrich: Numerical solutions for the simulation of monolithic microwave integrated circuits.

317. Donald A. Dawson, Klaus Fleischmann, Guillaume Leduc: Continuous dependence of a class of superprocesses on branching parameters, and applications.

318. Peter Mathé: Asymptotically optimal weighted numerical integration.

319. Guillaume Leduc: Martingale problem for (ξ, Φ, k)–superprocesses.