Numerical methods for kinetic equations

Giacomo Dimarco, Lorenzo Pareschi

Numerical methods for kinetic equations

G. Dimarco*
Institut de Mathématiques de Toulouse,
Université de Toulouse,
UPS, INSA, UT1, UTM, 31062 Toulouse,
France
E-mail: giacomo.dimarco@math.univ-toulouse.fr

L. Pareschi†
Department of Mathematics and Computer Science,
University of Ferrara,
Via Machiavelli 35, 44121 Ferrara,
Italy
E-mail: lorenzo.pareschi@unife.it

In this survey we consider the development and the mathematical analysis of numerical methods for kinetic partial differential equations. Kinetic equations represent a way of describing the time evolution of a system consisting of a large number of particles. Due to the high number of dimensions and their intrinsic physical properties, the construction of numerical methods represents a challenge and requires a careful balance between accuracy and computational complexity. Here we review the basic numerical techniques for dealing with such equations, including the case of semi-Lagrangian methods, discrete velocity models and spectral methods. In addition we give an overview of the current state of the art of numerical methods for kinetic equations. This covers the derivation of fast algorithms, the notion of asymptotic preserving methods and the construction of hybrid schemes.

* Partially supported by ANR grant BOOST: *"Building the future of numerical methods for Iter"*
† Partially supported by PRIN-MIUR grant *"Advanced numerical methods for kinetic equations and balance laws with source terms"*.

CONTENTS

1. Introduction

Kinetic equations are used to describe a variety of phenomena in different fields, ranging from rarefied gas dynamics and plasma physics to biology and socio-economy, and appear naturally when one considers a statistical description of a large particle system evolving in time.

At the microscopic level the particles motion is described by systems of ordinary differential equations. Such systems, however, are extremely costly to solve numerically and bring little insight on the behavior of a large set of particles. Therefore, one seeks for reduced models of the particle dynamics which are still able to describe the physical reality with sufficient accuracy.

In the classical kinetic theory of rarefied gases, the variation of a non-negative function $f = f(x, v, t)$, characterizing the particle densities having velocity $v \in \mathbb{R}^3$ in position $x \in \mathbb{R}^3$ at time t, is obtained through the equation

$$\frac{df}{dt} = Q(f), \tag{1.1}$$

where, by the chain rule,

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + v \cdot \nabla_x f. \tag{1.2}$$

In (1.2) we used the fact that all particles issued from the same point (x, v) of the phase-space follow the same trajectory

$$\frac{dx}{dt} = v, \qquad \frac{dv}{dt} = 0. \tag{1.3}$$

The operator $Q(f)$, on the right hand side in equation (1.1), describes the effects of internal forces due to particle interactions and its form depends on the details of the microscopic dynamic. The most well-known example is represented by the nonlinear Boltzmann collision integral of rarefied gas dynamics (Cercignani 1988, Cercignani, Illner and Pulvirenti 1994).

Typically this operator characterizes the conservation properties of the physical system, i.e.

$$\int_{\mathbb{R}^3} Q(f)\varphi(v)\,dv = 0, \tag{1.4}$$

where $U(x,t) = \int f(x,v,t)\varphi(v)\,dv \in \mathbb{R}^m$ defines a certain set of moments of the distribution function f. Classically $m = 5$ and $\varphi(v) = 1, v, |v|^2$ correspond to conservation of mass, momentum and energy respectively. Therefore integrating (1.1) against $\varphi(v)$ yields a system of macroscopic conservation laws

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^3} f\varphi(v)\,dv + \int_{\mathbb{R}^3} v\cdot\nabla_x f\,\varphi(v)\,dv = 0. \tag{1.5}$$

The above moment system, however, is not closed since the second term involves higher order moments of the distribution function f. Therefore, to obtain a closed set of equations, one is led to make assumptions about the form of the distribution function. A simple way to find and approximate closure is based on an additional property of the operator $Q(f)$. In fact, the distribution functions belonging to the kernel of the operator satisfy

$$Q(f) = 0 \quad \text{iff} \quad f = M[f], \tag{1.6}$$

where $M[f] = M[f](x,v,t)$ can be expressed univocally in terms of the set of moments $U(x,t)$. Using a terminology borrowed from the rarefied gas dynamic case, such functions are referred to as Maxwellian equilibria and the closed macroscopic system obtained by approximating f with $M[f]$ in (1.5) corresponds to the set of compressible Euler equations. Under standard moments boundedness assumptions, the Euler system can be written as

$$\frac{\partial U}{\partial t} + \nabla_x \cdot F(U) = 0, \tag{1.7}$$

with $F(U) = \int M[f]v\varphi(v)\,dv$. Note that the simplest operator satisfying (1.4) and (1.6) is the linear relaxation operator (Bhatnagar, Gross and Krook 1954)

$$Q(f) = \nu(M[f] - f), \tag{1.8}$$

where $\nu = \nu(x,t) > 0$. Other closure strategies, like the Navier-Stokes approach (Cercignani 1988), lead to more accurate macroscopic approximations of the moment system (1.5). In general, however, finding the right closure is a very difficult problem which is far from being solved.

Besides rarefied gas dynamics, kinetic equations play an important role in modeling plasmas (Landau 1981), granular gases (Pöschel and Brilliantov 2003), semiconductors (Markowich, Ringhofer and Schmeiser 1989), neutron transport (Lewis and Miller 1993) and quantum gases (Uehling and Uhlenbeck 1933). For a recent introduction to the Boltzmann equation and

related kinetic models we refer the reader to Villani (2002) and Degond, Pareschi and Russo (2004).

From the above picture it is clear that the numerical solution of a kinetic equation involves several problems of different nature. Aside from the high dimensionality of the problem, in general $(x, v, t) \in \mathbb{R}^7$, let us shortly summarize some of the additional numerical difficulties and requirements specific to kinetic equations:

(i) *Conservation properties.* Physical conservation properties (1.4) are very important since they characterize the steady states. Methods that do not maintain such properties at the discrete level need special care in practical applications.

(ii) *Computational cost.* The operator $Q(f)$ may be described by a high dimensional integral in velocity space at each point x in physical space. In such cases fast solvers are essential to avoid excessive computational cost.

(iii) *Velocity range.* The significant velocity range may vary strongly with space position (steady states are not compactly supported in velocity space and in some applications may present power law tails). Methods that use a finite velocity range may be inadequate in some circumstances.

(iv) *Presence of multiple scales.* In presence of multiple space-time scales and/or large velocities the kinetic equation becomes stiff. Classical stiff solvers may be hard to use when we have to invert a very large nonlinear system.

In this paper we review some of the main results in this field for deterministic numerical methods. Another class of methods, that we will not cover in the present survey, is based on stochastic Monte-Carlo techniques. The most famous examples are the Direct Simulation Monte-Carlo (DSMC) methods by Bird (1994) and by Nanbu (1980). These methods guarantee efficiency (their computational cost is linear with respect to the number of particles) and preservation of the main physical properties. However, avoiding statistical fluctuations in the results becomes extremely expensive in presence of non-stationary flows or close to continuum regimes. We refer the interested reader to the review by Pareschi and Russo (1999) and the book by Rjasanow and Wagner (2006). Some related topics based on the use of hybrid stochastic-deterministic methods are described in Section 9. This survey and the selected bibliography are obviously biased by the personal taste and knowledge of the authors. Numerical methods for kinetic equations are such a broad and active field of research that it is impossible to give credit to all relevant contributions.

After this introduction, the plan of the manuscript is organized as follows. Section 2 is devoted to a short introduction of some mathematical

and physical properties of the kinetic equations we will consider for the development of the different numerical approaches. Although the scope of our insights is wider, here we will focus mainly on the mean field Vlasov equation of plasma physics and the classical Boltzmann equation of rarefied gas dynamics. This is motivated by their relevance for applications and by the fact that these equations can be considered as prototype models containing most major difficulties present in other kinetic models. Other models, including linear Boltzmann equations and the Landau equation of collisional plasmas, are briefly recalled at the end of the section.

We start our presentation of numerical methods in Section 3 dealing with the case of semi-Lagrangian schemes designed for an accurate and efficient approximation of the Vlasov equation and particles transport (Cheng and Knorr 1976, Crouseilles, Mehrenberger and Sonnendrücker 2010, Sonnendrücker, Roche, Bertrand and Ghizzo 1999). The idea exploited by this kind of schemes is to use the theory of characteristics for computing the distribution function for the successive times either through forward or backward reconstructions. Conservative semi-Lagrangian schemes are also useful to ensure positivity of the solutions as well as conservations (Crouseilles, Respaud and Sonnendrücker 2009, Filbet, Sonnendrücker and Bertrand 2001). The methods are presented in the case of the Vlasov model and applied also to simple relaxation systems. Time splitting techniques or Runge-Kutta methods are used to link the discretization of the transport term with the discretization of the force term or the collision operator (Cheng and Knorr 1976, Filbet and Russo 2009, Dimarco and Loubère 2013*a*).

In Section 4 we deal with the discretization of the velocity space and its relevance in the deterministic approximations of the Boltzmann collision integral. Historically, one of the most popular method is represented by the so called Discrete Velocity Models (DVM) of the Boltzmann equation. These methods (Goldstein, Sturtevant and Broadwell 1989, Martin, Rogier and Schneider 1992, Rogier and Schneider 1994, Panferov and Heintz 2002) are based on a regular grid in velocity and on a discrete collision mechanism on the points of the grid that preserves the main physical properties. As we will see, the main drawback of this approach applied to the full Boltzmann integral is its high computational cost (larger that $O(n^2)$ for a quadrature formula based on n grid points) and relatively low accuracy (Palczewski, Schneider and Bobylev 1997, Buet 1996, Panferov and Heintz 2002, Fainsilber, Kurlberg and Wennberg 2006, Mouhot, Pareschi and Rey 2013). On the other hand the methods provide a robust framework for the derivation of conservative schemes in the case of simplified models, like the case of relaxation operators (Mieussens 2000, Mieussens 2001).

Next in Section 5 we focus on another relevant class of numerical techniques for the Boltzmann integral based on the use of spectral methods. The

foundations of the method were first proposed by Pareschi and Perthame (1996), inspired by previous works on the use of Fourier transform techniques (Bobylev 1988). The numerical scheme is based on Fourier-Galerkin spectral approximation of the distribution function represented by its Fourier series in velocity space and can be evaluated in exactly $O(n^2)$ operations. Related approaches, based on a direct discretization of the Fourier transformed Boltzmann equation, has been derived by Bobylev and Rjasanow (1999) and, more recently, by Gamba and Tharkabhushanam (2009). The method was further developed by Pareschi and Russo (2000*b*) where evolution equations for the Fourier modes were explicitly derived and spectral accuracy of the method was proved. Extensions of spectral methods to other fields, like plasma physics (Pareschi, Russo and Toscani 2000) and granular gases (Filbet, Pareschi and Toscani 2005) are also discussed. In the plasma physics case, algorithms that brings the overall cost to $O(n \log_2 n)$ are readily derived. A velocity rescaling technique is also presented for the granular gases case.

Section 6 is devoted to the issue of computational complexity in the numerical approximation of the Boltzmann equation by deterministic schemes. Recently Mouhot and Pareschi (2006), using a suitable representation of the collision operator, derived fast spectral solvers (for certain classes of particle interactions, including hard spheres, the cost of the method is reduced to approximately $O(n \log_2 n)$) without loosing the spectral accuracy. Previous attempts in this direction were based on discretizing directly the Fourier transformed Boltzmann equation for Maxwell molecules (Grigoriev and Mikhalitsyn 1983, Gabetta and Pareschi 1994, Bobylev and Rjasanow 2000). See also Bobylev and Rjasanow (1999) for a similar approach for the hard spheres case. The method has been subsequently extended to the case of quantum gases (Filbet, Hu and Jin 2012, Hu and Ying 2012) where fast solvers are mandatory due to the cubic nonlinearity of the operator. Using a pseudo-spectral formulation, this kind of approach has been applied successfully to the construction of fast algorithms for DVM discretizations of the Boltzmann operator (Mouhot et al. 2013).

Section 7 is concerned with the challenging problem of the severe time step restrictions in regions close to continuum regimes. Two situations are considered, the classical fluid limit problem and the diffusion regime. Several authors have tackled these problems in the past, and there is a large literature on the subject (see the recent surveys by Jin (2012), Degond (2014) and Pareschi and Russo (2011)). The feature shared by these techniques is that the resulting schemes avoid the solution of large systems of nonlinear equations, are unconditionally stable and capture the asymptotic limit automatically without resolving the small time scales. They are commonly referred to as asymptotic-preserving (AP) methods. Here we focus on two classes of AP methods of particular relevance for Boltzmann-type

equations in the fluid dynamic scaling. The first one is based on the use of exponential techniques (Gabetta, Pareschi and Toscani 1997, Dimarco and Pareschi 2011, Li and Pareschi 2014) and the second one is based on the use Implicit-Explicit (IMEX) methods (Filbet and Jin 2010, Dimarco and Pareschi 2013). We conclude the section with a discussion on the diffusion limit for linear transport equations. Extensions of the IMEX methods and a short review of other approaches are reported (Jin, Pareschi and Toscani 2000, Klar 1998*a*, Lemou and Mieussens 2008, Boscarino, Pareschi and Russo 2013).

In Section 8 we face the same kind of multiscale problems, but from a different perspective. The idea is to combine in a unique solver different numerical methods and models, each one specifically designed to deal with a particular regime. In such a vast research field, we confine ourselves to review some recent contributions where a dynamic fluid-kinetic coupling is realized through a time evolving transition zone (Degond, Jin and Mieussens 2005, Degond, Dimarco and Mieussens 2007, Degond, Dimarco and Mieussens 2010, Degond and Dimarco 2012). A closely related idea is based on the construction of hybrid methods. In these methods the numerical solution in each computational cell is obtained from the hybridization of two numerical solvers coupled together. Typically, one solver aims at constructing the equilibrium (continuum) part of the solution with a deterministic method and the other solver yields the non equilibrium (kinetic) part of the solution using a Monte Carlo strategy (Pareschi and Caflisch 2004, Dimarco and Pareschi 2007, Degond, Dimarco and Pareschi 2011, Burt and Boyd 2009, Homolle and Hadjiconstantinou 2007*a*, Radtke, Hadjiconstantinou and Wagner 2011, Alaia and Puppo 2012). As we will see, the construction of such multiscale schemes poses several new difficulties for numerical methods since it requires the definition of new concepts and methodologies (see also Abdulle, E, Engquist and Vanden-Eijnden (2012) for a general framework for designing multiscale algorithms).

At the end of this document, in Section 9, we include some final considerations and a non exhaustive list of related topics which we feel are important but are not included in this survey.

2. Preliminaries on kinetic equations

In this Section we give a short description of the kinetic models we will consider in the rest of the review. Due to the variety of fields where kinetic models have been applied it is impossible to give a fair description of all of them. Therefore we concentrate on two models that can be considered as prototypes for the development of the numerical methods, the Vlasov mean-field equation and the Boltzmann equation. Other relevant models are briefly mentioned at the end of the Section. We refer the reader to

the books by Cercignani (1988), Cercignani et al. (1994), the recent survey by Villani (2002), and the volume edited by Degond et al. (2004) for further insights.

2.1. The Vlasov equation

In the case of particles interacting through a smooth potential the right hand side in (1.1) is of the form $Q(f) = -F_m \cdot \nabla_v f$ where F_m is the mean-field force given by

$$F_m(t) = \int_{\mathbb{R}^3\times\mathbb{R}^3} F_i(x-y) f(y,v,t)\, dv\, dy = \int_{\mathbb{R}^3} F_i(x-y)\rho(y,t)\, dy, \quad (2.1)$$

where

$$\rho(x,t) = \int_{\mathbb{R}^3} f(x,v,t)\, dv, \quad (2.2)$$

is the density and F_i the internal force acting among particles. Typically we shall restrict to internal forces which derive from an interaction potential Φ_i so that $F_i = -\nabla_x \Phi_i$, where Φ_i is a scalar potential function.

This leads to the so-called Vlasov mean-field equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + F_m \cdot \nabla_v f = 0, \quad (2.3)$$

where F_m can also be written as

$$F_m = -\nabla_x \Phi_m, \quad \Phi_m = \int_{\mathbb{R}^3} \Phi_i(x-y)\rho(y)\, dy. \quad (2.4)$$

One of the most important examples for applications is the Coulomb potential

$$\Phi_i(x) = \frac{q}{4\pi r}, \quad r = |x|, \quad (2.5)$$

where $q = 1$ corresponds to the repulsive case (like e.g. the electrostatic interaction) and $q = -1$ to the attractive case (like e.g. gravitation). Then $\Delta\Phi_i(x) = -q\delta(x)$, where $\delta(x)$ is the delta distribution at 0. This gives

$$\Delta\Phi_m(x,t) = \int_{\mathbb{R}^3} \Delta\Phi_i(x-y)\rho(y,t)\, dy = -q\rho(x,t), \quad (2.6)$$

and we obtain the Vlasov-Poisson system

$$\begin{aligned} &\frac{\partial f}{\partial t} + v \cdot \nabla_x f - \nabla_x \Phi_m \cdot \nabla_v f = 0, \\ &\Delta\Phi_m(x,t) = -q\rho(x,t). \end{aligned} \quad (2.7)$$

In the case of negative charged particles in a uniform neutralizing background the Poisson equation reads

$$\Delta\Phi_m(x,t) = 1 - \rho(x,t). \quad (2.8)$$

The Vlasov-Poisson model is the most relevant model in plasma physics, we introduce some of the main numerical approaches in Section 3. For a rigorous derivation of the Vlasov mean-field system (2.3) we refer to the book by Spohn (1991).

2.2. The Boltzmann equation

The Boltzmann equation is the fundamental model for the description of the dynamics of particles in a dilute gas. The classical hard sphere case considers particles as solid spheres of diameter d which do not interact as long as they do not enter in contact. Note that, in contrast with the Vlasov description, here the interaction potential is non smooth since we have

$$F_i(x-y) = 0, \qquad \forall\, x, y \ \text{ s.t. } |x-y| > d. \tag{2.9}$$

The Boltzmann equation for colliding hard spheres takes the form

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q(f,f). \tag{2.10}$$

In this case the interaction operator has a bilinear structure, $Q = Q(f,f)$ and is obtained in the Boltzmann-Grad limit where the number of particles $N \to \infty$, $d \to 0$ in such a way that Nd^2 is kept constant

$$Q(f,f)(v) = \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} |v - v_*| [f(v')f(v'_*) - f(v)f(v_*)]\, d\omega\, dv_*. \tag{2.11}$$

The operator acts only on the velocity variable and is local in space, moreover the memory of the positions of the spheres before and after the collision has been lost. In the above expression, ω is a unit vector of the sphere $\mathbb{S}^2$ and (v', v'_*) represent the collisional velocities associated with (v, v_*). The collisional velocities satisfy microscopic momentum and energy conservation

$$v' + v'_* = v + v_*, \quad |v'|^2 + |v'_*|^2 = |v|^2 + |v_*|^2. \tag{2.12}$$

The above system of algebraic equations has the following parametrized solution

$$v' = \frac{1}{2}(v + v_* + |v - v_*|\omega), \quad v'_* = \frac{1}{2}(v + v_* - |v - v_*|\omega) \tag{2.13}$$

where $v - v_*$ is the relative velocity. For details on the rigorous derivation of the Boltzmann equation from the hard sphere dynamics we refer to Lanford III (1975) and to the book by Cercignani (1988).

Although a mathematical theory is still lacking, the Boltzmann equation is often used in connection with smooth potentials. Formally, for interactions forces described by an inverse power law we have the Boltzmann collision operator

$$Q(f,f)(v) = \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} B(v, v_*, \omega)[f(v')f(v'_*) - f(v)f(v_*)]\, d\omega\, dv_*, \tag{2.14}$$

where the collision kernel B is a nonnegative function depending only on $|v - v_*|$ and the scattering angle θ between relative velocities $v - v_*$ and $v' - v'_* = |v - v_*|\omega$

$$\cos\theta = \frac{(v' - v'_*) \cdot (v - v_*)}{|v - v_*|^2} = \frac{(v - v_*) \cdot \omega}{|v - v_*|}.$$

The kernel has the from

$$B(v, v_*, \omega) = |v - v_*|\sigma(|v - v_*|, \cos\theta), \tag{2.15}$$

where the scattering cross-section σ, in the case of inverse k-th power forces between particles, can be written as

$$\sigma(|v - v_*|, \cos\theta) = b_\alpha(\cos\theta)|v - v_*|^{\alpha-1}, \tag{2.16}$$

with $\alpha = (k-5)/(k-1)$. The potential with $k > 5$ are called hard potentials and for $k < 5$ we have soft potentials. The special situation $k = 5$ gives the so-called Maxwell pseudo-molecules model with

$$B(v, v_*, \omega) = b_0(\cos\theta). \tag{2.17}$$

For the Maxwell case the collision kernel is independent of the relative velocity. This case has been widely studied theoretically, in particular exact analytic solutions can be found in the space homogeneous case where $f = f(v, t)$ (Bobylev 1975). For an overview of existence results for the Boltzmann equation we refer the interested reader to the book by Cercignani et al. (1994).

Remark 2.1.

- For numerical purposes, a widely used model is the variable hard sphere (VHS) model introduced by Bird (1994) in order to correct the non-realistic scattering law of the hard spheres model in rarefied gas simulations. The model corresponds to $b_\alpha(\cos\theta) = C_\alpha$, where C_α is a positive constant, and hence

$$\sigma(|v - v_*|, \cos\theta) = C_\alpha|v - v_*|^{\alpha-1}. \tag{2.18}$$

- Along the survey we made the conventional assumptions $x \in \mathbb{R}^{d_x}$ and $v \in \mathbb{R}^{d_v}$ with $d_x = d_v = 3$. However, lower dimensional models can be constructed in order to simplify the mathematical analysis, or due to the different physical meaning of the independent variables. This lower dimensional model are particularly useful when testing numerical methods. When simplified models are considered we will emphasize the relevant differences.

2.3. Other parametrizations of the Boltzmann operator

The collision integral $Q(f, f)$ can be written in different equivalent forms, according to the parametrization used for the collisional velocities. Using

the identity

$$\int_{\mathbb{S}^2} F\left(\frac{u-|u|\omega}{2}\right)\,d\omega = \frac{2}{|u|}\int_{\mathbb{S}^2}|u\cdot n|F(n(u\cdot n))\,dn, \tag{2.19}$$

obtained by the transformation

$$\omega = e - 2(e\cdot n)n, \quad e = \frac{u}{|u|},$$

we get the frequently used form

$$Q(f,f)(v) = \int_{\mathbb{R}^3}\int_{\mathbb{S}^2}\bar{B}(v,v_*,n)[f(v')f(v'_*) - f(v)f(v_*)]\,dn\,dv_* \tag{2.20}$$

with

$$v' = v - ((v-v_*)\cdot n)n, \quad v'_* = v_* + ((v-v_*)\cdot n)n, \tag{2.21}$$

and

$$\bar{B}(v,v_*,n) = 2|v-v_*|\sigma(|v-v_*|, 1-2|\cos\varphi|^2)|\cos\varphi|, \tag{2.22}$$

where

$$|\cos\varphi| = \frac{|(v-v_*)\cdot n|}{|v-v_*|},$$

and the angle φ is related to the scattering angle θ by $\varphi = (\pi-\theta)/2$. The hard sphere case now corresponds to

$$\bar{B}(v,v_*,n) = 2|v-v_*||\cos\varphi|. \tag{2.23}$$

Another well-known parametrization of the collisional velocities is due to Carleman (1932). Here we report a closely related representation that we will use in the development of fast numerical solvers. From (2.19) we obtain the identity

$$\int_{\mathbb{S}^2} F\left(\frac{u-|u|\omega}{2}\right)\,d\omega = \frac{16}{|u|}\int_{\mathbb{R}^3}\delta(4x\cdot u + 4|x|^2)\,F(x)\,dx. \tag{2.24}$$

Equation (2.24) yields

$$Q(f,f)(v) = 16\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\sigma(|v-v_*|,\cos\theta)\,\delta(4x\cdot(v-v_*)+4|x|^2) \\ [f(v_*-x)\,f(v+x) - f(v_*)\,f(v)]\,dx\,dv_*$$

and then setting $y = v_* - v - x$ in v_* we obtain

$$Q(f,f)(v) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\tilde{B}(x,y)\delta(x\cdot y) \\ [f(v+y)\,f(v+x) - f(v+x+y)\,f(v)]\,dx\,dy, \tag{2.25}$$

with

$$\tilde{B}(x,y) = 4\sigma\left(|x+y|, -\frac{x\cdot(x+y)}{|x||x+y|}\right). \tag{2.26}$$

We refer to Section 6 for further details on the use of parametrization (2.25) of the Boltzmann operator for the derivation of fast summation methods.

2.4. Physical properties of the Boltzmann operator

During the evolution process, the collision operator preserves mass, momentum and energy, i.e.,

$$\int_{\mathbb{R}^3} Q(f,f)\varphi(v)\,dv = 0, \quad \varphi(v) = 1, v, |v|^2, \tag{2.27}$$

and in addition it satisfies Boltzmann's well-known H-theorem

$$\int_{\mathbb{R}^3} Q(f,f)\ln(f(v))dv \le 0. \tag{2.28}$$

Since the collision operator is local in space, the dependence from the variable x in this paragraph is omitted. The above properties are a consequence of the following identity that can be easily proved for any test function $\varphi(v)$

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q(f,f)\varphi(v)\,dv \\
&\quad = \int_{\mathbb{R}^6}\int_{\mathbb{S}^2} B(v,v_*,\omega)[ff_*][\varphi' - \varphi]\,d\omega\,dv_*\,dv \\
&\quad = -\frac{1}{4}\int_{\mathbb{R}^6}\int_{\mathbb{S}^2} B(v,v_*,\omega)[f'f'_* - ff_*][\varphi' + \varphi'_* - \varphi - \varphi_*]\,d\omega\,dv_*\,dv,
\end{aligned} \tag{2.29}$$

where we omitted the explicit dependence from v, v_*, v', v'_* to simplify the notation. In order to prove (2.29) one uses the micro-reversibility property $B(v,v_*,\omega) = B(v_*,v,\omega)$ and the fact that the Jacobian of the transformation of $(v,v_*) \to (v',v'_*)$ is equal to one.

It is useful to introduce the following definition.

Definition 2.1. A function $\varphi(v) : \mathbb{R}^3 \to \mathbb{R}$ such that

$$\varphi(v') + \varphi(v'_*) - \varphi(v) - \varphi(v_*) = 0, \quad \forall\, v, v_* \in \mathbb{R}^3,\ \omega \in \mathbb{S}^2$$

with v', v'_* defined by (2.13) is called a *collision invariant.*

It can be shown that

Theorem 2.1. A continuous function φ is a collision invariant if and only if $\varphi \in \text{span}\{1, v, |v|^2\}$ or equivalently

$$\varphi(v) = a + b\cdot v + c|v|^2, \quad a, c \in \mathbb{R}, \quad b \in \mathbb{R}^3.$$

Assuming f strictly positive, for $\varphi(v) = \ln(f(v))$ form (2.29) we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^3} Q(f,f)\ln(f)dv \\
&\quad = -\frac{1}{4}\int_{\mathbb{R}^6}\int_{\mathbb{S}^2} B(v,v_*,\omega)[f'f'_* - ff_*] \\
&\qquad\qquad [\ln(f') + \ln(f'_*) - \ln(f) - \ln(f_*)]\, d\omega\, dv_*\, dv \\
&\quad = -\frac{1}{4}\int_{\mathbb{R}^6}\int_{\mathbb{S}^2} B(v,v_*,\omega)[f'f'_* - ff_*]\ln\left(\frac{f'f'_*}{ff_*}\right) d\omega\, dv_*\, dv \le 0,
\end{aligned} \tag{2.30}$$

since the function

$$z(x,y) = (x-y)\ln(x/y) \ge 0 \tag{2.31}$$

and $z(x,y) = 0$ only if $x = y$.

In (2.30) the equality sign holds only if $\ln(f)$ is a collision invariant, which implies

$$f = \exp(a + b\cdot v + c|v|^2), \quad c < 0.$$

If we define the macroscopic density, mean velocity and temperature

$$\rho = \int_{\mathbb{R}^3} f\, dv, \qquad u = \frac{1}{\rho}\int_{\mathbb{R}^3} vf\, dv, \qquad T = \frac{1}{3R\rho}\int_{\mathbb{R}^3}(v-u)^2 f\, dv, \tag{2.32}$$

we obtain that the distribution function has the form

$$M[f](v,t) = M[\rho,u,T](v,t) = \frac{\rho}{(2\pi RT)^{3/2}}\exp\left(-\frac{|u-v|^2}{2RT}\right). \tag{2.33}$$

The constant $R = K_B/m$ is called the gas constant, K_B is the Boltzmann constant and m the mass of a particle.

Definition 2.2. A function of the form (2.33) is called a *Maxwellian* distribution.

Boltzmann's H-theorem (2.30) 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.

Let $f(v,t)$ be a solution of the homogeneous Boltzmann equation

$$\frac{\partial f}{\partial t} = Q(f,f), \tag{2.34}$$

where now the density, mean velocity and the temperature are constants defined by the initial distribution.

Definition 2.3. The functional

$$H(f) = \int_{\mathbb{R}^3} f\ln(f)\, dv, \tag{2.35}$$

is called the *H-functional*.

From (2.30) and (2.34), using mass conservation, we have

$$\frac{dH(f)}{dt} = \int_{\mathbb{R}^3} Q(f,f)\ln(f)\,dv \leq 0. \tag{2.36}$$

Thus the H-functional is monotonically decreasing until f reaches the equilibrium Maxwellian state. In a non homogeneous setting if we multiply the Boltzmann equation by $\ln(f)$ and integrate with respect to v we get the entropy dissipation equation

$$\frac{\partial H(f)}{\partial t} + \nabla_x \cdot \left(\int_{\mathbb{R}^3} f\ln(f) v\,dv \right) = \int_{\mathbb{R}^3} Q(f,f)\log f\,dv \leq 0. \tag{2.37}$$

If we further integrate with respect to x and ignore possible boundary terms we obtain

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^3} H(f)\,dx = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} Q(f,f)\ln(f)\,dv\,dx \leq 0. \tag{2.38}$$

By denoting the kinetic entropy with $-H(f)$, the above inequality express the fact that the total entropy is nondecreasing as time increases and therefore the Boltzmann dynamics is irreversible.

Finally, another possible way to characterize the Maxwellian state is through a minimization problem. Suppose we fix the moments $\rho, T \in \mathbb{R}_+$ and $u \in \mathbb{R}^3$. It can be shown that $M[f]$ is the unique solution of the entropy minimization problem

$$\min\left\{ H(f) = \int_{\mathbb{R}^3} f\log f\,dv, f \geq 0, \right. \\ \left. \int_{\mathbb{R}^3} f \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv = \begin{pmatrix} \rho \\ \rho u \\ \rho(u^2 + 3RT) \end{pmatrix} \right\}. \tag{2.39}$$

This means that $M[f]$ minimizes the entropy of all the possible states leading to the same macroscopic properties. This minimization problem is called the Gibbs principle and can be solved by a Lagrange multiplier method. We refer to Levermore (1996) for recent applications of the above principle to moment methods.

2.5. Moment equations and fluid limit

Fluid equations deal with averaged quantities over small volumes in position space. In order to express that we wish to look at the system at large scales, we are led to introduce the rescaling

$$x' = \varepsilon x, \qquad t' = \varepsilon t,$$

in the Boltzmann equation. Dropping the primes for notation simplicity we obtain the perturbation problem

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f, f), \tag{2.40}$$

where the scaling parameter $\varepsilon > 0$ is referred to as Knudsen number. If we multiply the scaled Boltzmann equation (2.40) by its collision invariants and integrate the result in velocity space we obtain the moment equations (1.5) that can be rewritten as

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^3} f\varphi(v)\, dv + \nabla_x \cdot \left(\int_{\mathbb{R}^3} v f \varphi(v)\, dv \right) = 0, \quad \varphi(v) = 1, v, |v|^2. \tag{2.41}$$

These equations describe the balance of mass, momentum and energy. As already observed, the above system is not closed since it involves higher order moments of the distribution function f.

As $\varepsilon \to 0$, from (2.40) we have formally $Q(f, f) \to 0$, and thus f approaches the local Maxwellian $M[f]$. In this case, substituting $f = M[f]$ into (2.41), the higher order moments of the distribution function can be computed as function of ρ, u, and T and we formally recover the closed system of compressible Euler equations

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) = 0 \\ &\frac{\partial \rho u}{\partial t} + \nabla_x \cdot (\rho u \otimes u + p) = 0 \\ &\frac{\partial E}{\partial t} + \nabla_x \cdot (E + p) u = 0 \end{aligned} \tag{2.42}$$

$$p = \rho R T, \quad E = \frac{1}{2} \rho (u^2 + 3RT) \tag{2.43}$$

where p is the gas pressure and $\otimes$ denotes the tensor product. We recall that the relation $p = \rho RT$ is called the perfect gas equation of state.

The rigorous passage from the Boltzmann equation to the compressible Euler equations has been investigated by several authors (Caflisch 1980, Nishida 1978). Higher order fluid models, such as the compressible Navier-Stokes model, can be derived using the expansions due to Chapmann-Enskog and to Grad (Müller and Ruggeri 1993, Struchtrup 2005). We refer to Levermore (1996) for a mathematical setting of the problem and to Golse and Saint-Raymond (2004) for recent theoretical results. Let us remark that, above the Navier-Stokes level, these classical expansions yield unsatisfactory equations, which are unstable in cases of the Chapman-Enskog expansion (Burnett and super-Burnett equations), and describe unphysical discontinuous shocks in case of the Grad method. We refer to Struchtrup (2005) for alternative approaches that avoid some of the short-comings of the classical high order closures.

Remark 2.2. In velocity dimension d_v, $d_v \geq 1$, the definition of the moments becomes

$$\rho = \int_{\mathbb{R}^{d_v}} f\,dv, \quad u = \frac{1}{\rho}\int_{\mathbb{R}^{d_v}} vf\,dv, \quad T = \frac{1}{d_v\,R\rho}\int_{\mathbb{R}^{d_v}} (v-u)^2 f\,dv, \quad (2.44)$$

and the corresponding Maxwellian reads

$$M[f](v,t) = \frac{\rho}{(2\pi RT)^{d_v/2}} \exp\left(-\frac{|u-v|^2}{2RT}\right). \quad (2.45)$$

In particular we have

$$E = \frac{1}{2}\int_{\mathbb{R}^{d_v}} f|v|^2\,dv = \frac{1}{2}\rho(u^2 + d_v\,RT).$$

2.6. Boundary conditions

The Boltzmann equation is complemented with the boundary conditions in space for $v \cdot n \geq 0$ and $x \in \partial\Omega$, where n denotes the unit normal, pointing inside the domain Ω. Mathematically, such boundary conditions are modelled by an expression of the form (Cercignani 1988)

$$|v\cdot n| f(x,v,t) = \int_{v_*\cdot n<0} |v_*\cdot n(x)| K(v_* \to v, x, t) f(x, v_*, t)\,dv_*. \quad (2.46)$$

The ingoing flux is defined in terms of the outgoing flux modified by a given boundary kernel K. This boundary kernel is such that positivity and mass conservation at the boundaries are guaranteed,

$$K(v_* \to v, x, t) \geq 0, \qquad \int_{v\cdot n(x)\geq 0} K(v_* \to v, x, t)\,dv = 1.$$

Commonly used reflecting boundary conditions are the so-called Maxwell's conditions. This is equivalent to impose for the ingoing velocities

$$f(x,v,t) = (1-\alpha)Rf(x,v,t) + \alpha Mf(x,v,t), \quad (2.47)$$

in which $x \in \partial\Omega$, $v \cdot n(x) \geq 0$. The coefficient α, with $0 \leq \alpha \leq 1$, is called the accommodation coefficient and

$$Rf(x,v,t) = f(x, v - 2n(n\cdot v), t), \quad Mf(x,v,t) = \mu(x,t)M_w(v). \quad (2.48)$$

If we denote by T_w the temperature of the solid boundary, M_w is given by

$$M_w(v) = \exp\left(-\frac{v^2}{2RT_w}\right),$$

and the value of μ is determined by mass conservation at the wall

$$\mu(x,t)\int_{v\cdot n\geq 0} M_w(v)|v\cdot n|dv = \int_{v\cdot n<0} f(x,v,t)|v\cdot n|dv. \quad (2.49)$$

The case $\alpha = 0$ corresponds to specular reflection and the re-emitted molecules have the same flow of mass, temperature and tangential momentum of the incoming molecules, while $\alpha = 1$ corresponds to full accommodation and the re-emitted molecules have completely lost memory of the incoming molecules, except for conservation of the number of molecules.

A different type of boundary condition are the inflow boundary conditions where one assumes that the distribution function of the particles entering the domain is known, i.e.

$$f(x, v, t) = g(v, t), \quad x \in \partial\Omega, \quad v \cdot n > 0.$$

A typical example of such condition is used in shock wave calculations, where one assumes that the distribution function at the boundary of the computational domain is a Maxwellian $M(v)$ and that the incoming flux is distributed according to the Maxwellian flux $(v \cdot n)M(v)$, $v \cdot n > 0$.

2.7. Other collision operators

BGK models

As mentioned in the introduction, a simplified model Boltzmann equation is represented by the relaxation operator (1.8). This model, which we rewrite below for the sake of completeness, is usually referred to as BGK model since its introduction by Bhatnagar et al. (1954)

$$Q(f)(v) = \nu(M[f] - f). \tag{2.50}$$

In (2.50) the function $M[f]$ is the local Maxwellian computed by the moments of the distribution function f and ν, in general, is proportional to the density and depends on the temperature $\nu(\rho, T) = C\rho T^{1-\mu}$, where $C > 0$ is a constant and μ is the exponent of the viscosity law of the gas (Mieussens 2000).

Conservation of mass, momentum and energy as well as Boltzmann's H-theorem are readily satisfied and the equilibrium solutions are Maxwellians. Furthermore, the model has the correct fluid dynamic limit, since under the scaling (2.40) as $\varepsilon \to 0$ formally the moments ρ, ρu, and E satisfy the compressible Euler equations (2.42).

The mathematical theory of the BGK equation is simpler then for the full Boltzmann equation (Perthame 1989). Numerical simulations are also easier, especially by deterministic methods (see Sections 3 and 4). However, this model exhibits some unphysical features, such as an unrealistic Prandtl number. The Prandtl number is a normalized ratio of the heat conductivity to the viscosity. In the case of the BGK operator, this ratio is one where it is smaller than one in the case of the Boltzmann operator (for example the hard-sphere model leads to a Prandtl number very close to 2/3). This causes the BGK model to have a different Navier-Stokes limit with respect to the

original Boltzmann equation. The correct Prandtl number, as well as the correct Navier-Stokes limit, can be recovered using more sophisticated BGK models, such as the velocity dependent collision frequency BGK models and the Ellipsoidal Statistical BGK (ES-BGK) models (Bouchut and Perthame 1993, Holway 1966).

Landau-Fokker-Planck models

The Landau-Fokker-Planck model is a common kinetic model in plasma physics and is obtained in the so-called grazing collision limit of the Boltzmann operator. In such limit the Boltzmann collision operator converges towards a nonlinear integro-differential diffusion operator (Landau 1981)

$$Q_L(f,f)(v) = \nabla_v \cdot \int_{\mathbb{R}^3} A(v-v_*)[f(v_*)\nabla_v f(v) - f(v)\nabla_{v_*} f(v_*)]\, dv_* \quad (2.51)$$

where $A(v-v_*) = \Psi(|v-v_*|)\Pi(v-v_*)$ is a 3×3 nonnegative symmetric matrix and

$$\Pi(v-v_*) = I - \frac{(v-v_*)(v-v_*)}{|v-v_*|^2},$$

with I the identity matrix, is the orthogonal projection upon the space orthogonal to $v-v_*$. We have $\Psi(|v-v_*|) = \Lambda|v-v_*|^{\alpha+2}$ for inverse-power laws, with $\alpha \geq -3$ and $\Lambda > 0$.

Since conservation of mass, momentum, and energy, as well as H-theorem for the entropy are satisfied, equilibrium states are Maxwellians. The case $\alpha = -3$ is the so-called Coulombian case, of primary importance for applications. In such case the Boltzmann collision operator has no meaning, due to the divergence of the integral, even for smooth functions, a cut-off angular approximation is then used and the Landau equation can be derived in the so called grazing collision limit (Villani 2002).

Linear Boltzmann models

Linear Boltzmann models occur when one considers a particle system moved by an external force which describes short-range interactions with some scatters (like fixed obstacles or a known distribution of target particles). In this situation, the interaction between the particles under consideration and the obstacles is described by a linear Boltzmann collision operator

$$Q(f) = \int_{\mathbb{R}^3} [W(v_* \to v)f(v_*) - W(v \to v_*)f(v)]\, dv_* \quad (2.52)$$

where $W(v_* \to v)$ is the scattering rate.

For example this operator is used in kinetic semiconductor models where $W(v_* \to v) = \sigma(v, v_*)M(v)$ with M the normalized equilibrium distribution (Maxwellian, Fermi-Dirac) at the temperature θ of the lattice. The function $\sigma(v, v_*)$ describes the interaction of carriers with phonons (Markowich et

al. 1989). A classical reference for linear transport is the book by Case and Zweifel (1967). More recently these models have found applications in biology (Perthame 2007).

Further models

The Enskog equation takes into account the nonlocality of the interactions induced by the diameter of the interacting spheres and describes the behavior of dense gases. The collision operator is delocalized in space since the nonlinear local integrand $f'f'_* - ff_*$ in (2.10) takes the form

$$f'(x)f'_*(x - \delta\omega) - f(x)f_*(x + \delta\omega),$$

where δ is the particle diameter, and this feature regularizes the singularity of the Boltzmann operator. As a result, the mathematical theory of the Enskog equation is more complete than that of the Boltzmann operator (Bellomo, Lachowicz, Polewczak and Toscani 1991). Modern applications of such models in one dimension are found in traffic flows (Klar and Wegener 1997).

Other generalizations of the Boltzmann operator deal with quantum or relativistic extensions. In the quantum-Boltzmann operator the nonlinear interactions $f'f'_* - ff_*$ in (2.14) are replaced by

$$f'f'(1 \pm \theta_0 f)(1 \pm \theta_0 f_*) - ff_*(1 \pm \theta_0 f')(1 \pm \theta_0 f'_*),$$

where $\theta_0 = \hbar^3$, $\hbar$ is the rescaled Planck constant. The minus sign corresponds corresponds to fermions (such as electrons), and the plus sign to bosons (such as photons). The collision operator are called Fermi-Dirac operator and Bose-Einstein operator respectively. References about these quantum collision operators and the associated challenging phoenomena of the Bose-Einstein condensation can be found in (Escobedo, Mischler and Valle 2003, Degond et al. 2004).

Finally, let us mention that Boltzmann operators have been also used to describe granular gases. In such models particles undergo binary inelastic collisions according to the rules

$$v' = \frac{1}{2}(v + v_*) + \frac{1+e}{4}(v - v_*) + \frac{1+e}{4}|v - v_*|\omega,$$
$$v'_* = \frac{1}{2}(v + v_*) - \frac{1+e}{4}(v - v_*) - \frac{1+e}{4}|v - v_*|\omega,$$

where $0 < e \leq 1$ is called the restitution coefficient. As a result, when $e < 1$ energy is dissipated by the model and the steady states (in absence of external sources of energy) are Dirac delta function centered in the mean velocity. We refer to Bobylev, Carrillo and Gamba (2000) for recent mathematical results. Similar models in a one-dimensional setting have been introduced recently to describe wealth distributions (Cordier, Pareschi and Toscani 2005).

2.8. Other scalings and diffusion limits

Different kind of scalings of the Boltzmann equation are frequently considered in applications. They consist in taking

$$x' = \varepsilon x, \quad t' = \varepsilon^{1+\alpha} t,$$

where $\alpha \geq 0$ yields different macroscopic limits when $\varepsilon \to 0$. In this case, reverting to the original notation, the Boltzmann equation reads

$$\varepsilon^{\alpha} \frac{\partial f}{\partial t} + v \cdot \nabla_{\mathbf{x}} f = \frac{1}{\varepsilon} Q(f, f)\,. \tag{2.53}$$

When $\alpha = 0$ we have the fluid limit discussed before, when $\alpha > 0$ a longer time scale is considered and the formal derivation of the asymptotic behavior for the Boltzmann equations becomes more delicate and some special assumptions on the structure of the initial data are required. In particular, for $0 < \alpha < 1$, formally one obtains the incompressible Euler equations, whereas for $\alpha = 1$ the incompressible Navier-Stokes equations are derived (Cercignani et al. 1994). We refer to Bardos, Golse and Levermore (1991) and Bardos, Golse and Levermore (1993) for a theoretical background.

Here we limit ourselves to show the formal derivation for the diffusion limit $\alpha = 1$ in the case of a simple linear collision term of the form

$$Q(f)(v) = \int_{\mathbb{R}^3} \sigma(v, v_*)[M(v) f(v_*) - M(v_*) f(v)]\, dv_*, \tag{2.54}$$

where M is the constant in time normalized Maxwellian

$$M(v) = \frac{1}{(2\pi)^{3/2}} \exp\left(-\frac{|v|^2}{2}\right)$$

corresponding to a spatially homogeneous fluid state with density and temperature equal to 1 and bulk velocity equal to 0. In (2.54), the anisotropic scattering kernel σ is rotationally invariant and satisfies

$$\sigma(v, v_*) = \sigma(v_*, v) > 0.$$

We assume that the collision frequency λ satisfies the following bound for some positive constant λ_M

$$0 < \lambda(v) = \int_{\mathbb{R}^3} \sigma(v, v_*)\, M(v_*)\, dv_* \leq \lambda_M. \tag{2.55}$$

When $\varepsilon \to 0$, formally we have $Q(f) = 0$ which implies $f(x, v, t) = \rho(x, t) M(v)$ where the mass ρ satisfies the diffusion equation (Markowich et al. 1989)

$$\partial_t \rho = \nabla_x \cdot (D \nabla_x \rho)\,. \tag{2.56}$$

In the above equation, D is the diffusion coefficient matrix defined implicitly

in terms of the cross section

$$D = \int_{\mathbb{R}^3} \frac{M(v)}{\lambda(v)} (v \otimes v)\, dv.$$

2.9. Splitting of the time scales

Operator splitting methods are a classical approach to the numerical approximation of partial differential equations. In the case of kinetic equations we introduce such methods ad the end of this section due to their transversal nature to many different numerical techniques.

The solution in the time interval $[0, \Delta t]$ of (1.1) is obtained as the sequence of two steps. First integrate the space homogeneous kinetic equation

$$\begin{aligned} &\frac{\partial f^*}{\partial t} = Q(f^*), \\ &f^*(x, v, 0) = f_0(x, v), \end{aligned} \tag{2.57}$$

on the time interval $[0, \Delta t]$ to obtain $f^* = \mathcal{C}_{\Delta t}(f_0)$, and then the transport equation using the output of the previous step as initial condition,

$$\begin{aligned} &\frac{\partial f}{\partial t} + v \cdot \nabla_x f = 0, \\ &f(x, v, 0) = f^*(x, v, \Delta t). \end{aligned} \tag{2.58}$$

in the same time interval to get $f = \mathcal{T}_{\Delta t}(f^*) = \mathcal{T}_{\Delta t}(\mathcal{C}_{\Delta t}(f_0))$.

After computing an approximation of the solution at time Δt, the process may be iterated to obtain the numerical solution at later times. For convergence results of this first order splitting in the case of Boltzmann and BGK models we refer to Desvillettes and Mischler (1996).

Splitting schemes are very popular since they share several nice properties.

- The homogeneous step acts only on v whereas the transport step acts on x. This makes the implementation of the resulting scheme simpler (it allows the use of any existing code designed to solve the free transport equation) and highly parallelizable.
- It is simpler to design schemes which preserves the physical properties (conservations, H-theorem), since these properties essentially depend on the treatment of the homogeneous step.

Higher order splitting formulas can be derived in different ways (Hairer, Lubich and Wanner 2002). For example the well-known second order Strang splitting (Strang 1968) can be written as

$$\mathcal{C}_{\Delta t/2}(\mathcal{T}_{\Delta t}(\mathcal{C}_{\Delta t/2}(f_0))). \tag{2.59}$$

Unfortunately for splitting methods of order higher then two it can be shown that it is impossible to avoid negative time steps both in the transport as well

as in the collision (Hairer et al. 2002). Higher order formulas which avoid negative time stepping can be obtained as suitable combination of splitting steps (Dia and Schatzman 1996). For example a fourth order scheme reads

$$\frac{4}{3}\mathcal{C}_{\Delta t/4}(\mathcal{T}_{\Delta t/2}(\mathcal{C}_{\Delta t/2}(\mathcal{T}_{\Delta t/2}(\mathcal{C}_{\Delta t/4}(f_0))))) - \frac{1}{3}\mathcal{C}_{\Delta t/2}(\mathcal{T}_{\Delta t}(\mathcal{C}_{\Delta t/2}(f_0))). \quad (2.60)$$

Clearly all the above splitting methods admit the symmetric formulation obtained by switching the transport and the collision operators. Beside the problem of the appearance of negative coefficients or negative time steps in high order formulas (which may originate lack of positivity), splitting methods suffer also of order reduction in the fluid-limit (Jin 1995).

3. Semi-Lagrangian schemes

In this section we give a short overview of semi-Lagrangian method for kinetic transport equation and the Vlasov equation. The methods are based on a fixed computational grid but take into account the Lagrangian nature of the transport process. For their structure semi-Lagrangian methods apply naturally to the linear transport part of kinetic equations, the full equation being often solved by splitting techniques. These methods can be designed in order to possess many desired properties for a numerical scheme for kinetic equations, namely positivity, physical conservations and robustness when dealing with large velocities. These restrictions often prevent a straightforward application of the usual schemes for hyperbolic conservation laws (Cockburn, Johnson, Shu and Tadmor 1998).

There are several approaches that can be used to solve efficiently the transport process in kinetic equations, ranging from particle in cell methods (Birdsall and Langdon 1991) and flux-balance methods (Boris and Book 1973) to WENO schemes (Carrillo and Vecil 2007) and Discontinuous-Galerkin methods (Qiu and Shu 2011, Ayuso, Carrillo and Shu 2011). Other schemes developed specifically for the Vlasov equation include Heath, Gamba, Morrison and Michler (2012) and Cheng, Gamba and Proft (2012). In this section we do not aspire to present a survey of such a general and broad topic and refer to the above references and the recent introductory notes by Sonnendrücker (2013) for a more complete overview of the methods.

We first present the basic concept of the semi-Lagrangian method in the simple context of the linear, one dimensional advection equation. We will then consider the applications of such technique to kinetic equations in particular we consider applications to the Vlasov-Poisson system (Cheng and Knorr 1976, Crouseilles et al. 2010, Sonnendrücker et al. 1999, Filbet et al. 2001) and to the BGK equation of rarefied gas dynamics (Filbet and Russo 2009, Santagati, Russo and Yun 2012, Dimarco and Loubère 2013*a*, Dimarco and Loubère 2013*b*).

3.1. Transport equations

Let us consider the one dimensional linear advection equation

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = 0, \qquad x \in \mathbb{R} \tag{3.1}$$

here $f = f(x,t)$, $v \in \mathbb{R}$, with initial datum $f(x,0) = f_0(x)$. It is then well known that the exact solution reads

$$f(x,t) = f_0(x - vt). \tag{3.2}$$

The Semi-Lagrangian methods use the knowledge of the exact solution which is explicitly represented in terms of the initial datum to construct a numerical approximation of the transport equation. In particular, the following expression holds

$$f(x_j, t^{n+1}) = f_0(x_j - vt^{n+1}) = f_0(x_j - v\Delta t - vt^n) = f(x_j - v\Delta t, t^n) \tag{3.3}$$

where we introduced a Cartesian uniform grid $x_j = j\Delta x$, $j \in \mathbb{Z}$ and a time discretization $t^n = n\Delta t$. The above equation provides the basis for semi-Lagrangian methods. The points in space that are used to compute the solution are the points that within a single time step are transported by the flow onto the computational mesh. These points does not lie in the general case on the grid. The backward semi-Lagrangian method can then be obtained as

$$f_j^{n+1} = f^n_{j - v\frac{\Delta t}{\Delta x}} = f^n_{j-k-\alpha}, \qquad k + \alpha = v\frac{\Delta t}{\Delta x}, \quad k = \left[v\frac{\Delta t}{\Delta x} \right], \tag{3.4}$$

where $[\,\cdot\,]$ denotes the integer part and $\alpha \in (0,1)$ is a non integer index unless the time and space grid satisfy $v\Delta t = k\Delta x$ in which case $\alpha = 0$. The expression $f^n_{j-k-\alpha}$ represents the value at the point $x_j - v\Delta t$ obtained by some interpolation procedure (see Figure 3.1). The type and the degree of interpolation defines then the type of semi-Lagrangian scheme. As an example we consider a simple linear interpolation, then the scheme reads

$$f_j^{n+1} = \alpha f^n_{j-k-1} + (1-\alpha) f^n_{j-k}. \tag{3.5}$$

Observe now that if $v\Delta t/\Delta x < 1$ one gets $k = 0$, $\alpha = v\Delta t/\Delta x$ and the resulting method is nothing else but the well-known upwind method. However, in contrast with standard upwind, scheme (3.5) holds for any value of $v\Delta t/\Delta x$ and, since the values of the solution at the new time level $n+1$ are obtained by a linear interpolation of the values at time level n with nonnegative coefficients, the discrete maximum principle holds. This means that no stability conditions are needed for such scheme and therefore it is well-suited to deal with arbitrary large values of v. Note also that (3.3) admits the formulation

$$f(x_j + v\Delta t, t^{n+1}) = f(x_j, t^n), \tag{3.6}$$

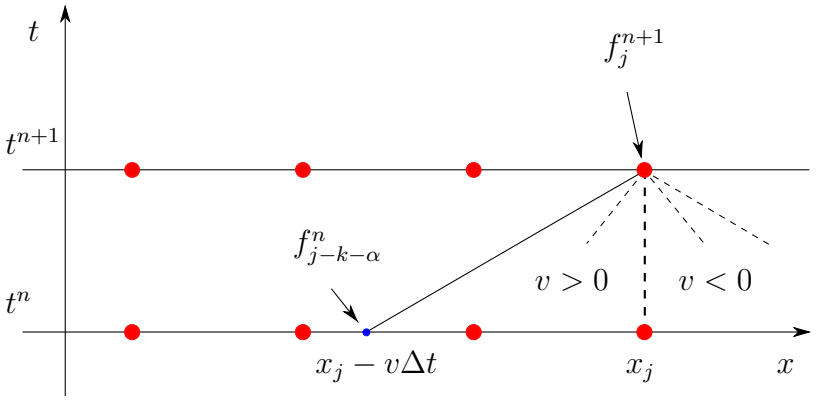

Figure 3.1. Sketch of the semi-Lagrangian approach for $v > 0$. The foot of the characteristics does not lie on the grid, and some interpolation is needed.

which gives the equivalent forward semi-Lagrangian scheme

$$f_{j+k+\alpha}^{n+1} = f_j^n, \qquad k + \alpha = v\frac{\Delta t}{\Delta x}, \quad k = \left[v\frac{\Delta t}{\Delta x}\right]. \tag{3.7}$$

The semi-Lagrangian method can be easily generalized to the multidimensional case by replacing one dimensional interpolation with multidimensional interpolation techniques.

In the more general case of a space and time dependent velocity field $V(x,t) \in \mathbb{R}^d$, one considers the equation

$$\frac{\partial f}{\partial t} + V(x,t) \cdot \nabla f = 0. \tag{3.8}$$

Under Lipschitz continuity assumptions on the velocity field, it can be proved that the characteristic curves exist. These are defined as the solutions $X(\,\cdot\,;t,x)$ of the ordinary differential equations

$$\frac{d}{ds}X(s;t,x) = V(X(s;t,x),s) \tag{3.9}$$

with initial data $X(t;t,x) = x$. It is then possible to prove that

$$f(x,t) = f(X(s;t,x),s) = f_0(X(0;t,x)). \tag{3.10}$$

This means that the solution at point x and at time t is nothing else but the initial datum at the foot of the characteristic indicated by $X(0;x,t)$ which passes in x at time t. Then a semi-Lagrangian approach can be derived provided that a numerical solution to equation (3.9) is computed. Using the formula (3.10) for the exact solution then a semi-Lagrangian method

for the approximation of the advection equation (3.8) can be derived. It can be described by the following two steps:

- At a given time level n compute for each mesh point x an approximate solution of (3.9) to determine an estimate of the characteristic $X^*(t^n; t^{n+1}, x)$ which passes at time t^{n+1} at position x.
- Compute an approximation of (3.10) by interpolating the mesh point values at time level n at the points $X^*(t^n; t^{n+1}, x)$.

This implies that the solution of the PDE (3.8) is reduced to solution of a large set ODEs combined with multidimensional interpolation. The most common reconstruction techniques found in literature are cubic splines, Hermite or Lagrange polynomials.

For example, the cubic spline interpolation $f_{\Delta x}$ is defined by $f_{\Delta x}(x_j) = f(x_j)$, $f_{\Delta x} \in \mathbb{P}_3([x_j, x_{j+1}])$ with $f_{\Delta x} \in \mathcal{C}^2(I)$ on the interval I. Writing $f_{\Delta x}$ using the cubic B-spline basis we get

$$f_{\Delta x}(x) = \sum_{h=0}^{M-1} a_h S^3(x - x_h) \tag{3.11}$$

where M are the number of points of the mesh and the coefficients a_h are given by the interpolation conditions

$$f(x_j) = f_{\Delta x}(x_j) = \sum_{h=0}^{M-1} a_h S^3(x_j - x_h) \tag{3.12}$$

while the cubic B-spline basis is given by

$$S^3(x) = \frac{1}{6} \begin{cases} (2 - |x|/\Delta x^3), & \text{if } \Delta x < |x| < 2\Delta x \\ 4 - 6(|x|/\Delta x)^2 + 3(|x|/\Delta x)^3, & \text{if } 0 < |x| < \Delta x \\ 0, & \text{otherwise.} \end{cases} \tag{3.13}$$

If the Hermite interpolation has to be used in the semi-Lagrangian approach one needs, in addition to the pointwise values of the solution f, the pointwise values of its derivative.

3.2. Semi-Lagrangian scheme for Vlasov type equations

To simplify notations we illustrate the methods for the one-dimensional Vlasov-Poisson system

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} - E\frac{\partial f}{\partial v} = 0, \tag{3.14}$$

$$\frac{\partial^2 \Phi_m}{\partial x^2}(x,t) = 1 - \rho(x,t) = 1 - \int_{\mathbb{R}} f(x,v,t)dv, \quad E = -\frac{\partial \Phi_m}{\partial x}. \tag{3.15}$$

Observe that the Vlasov equation can be rewritten in equivalent form as

$$\frac{\partial f}{\partial t} + V \cdot \nabla_{(x,v)} f = 0, \; V(x,v,t) = (v, E)^T \tag{3.16}$$

which is a linear transport equation in the phase space. Moreover since

$$\nabla_{(x,v)} \cdot V = \frac{\partial v}{\partial x} + \frac{\partial E}{\partial v} = 0, \tag{3.17}$$

the Vlasov equation can also be written in conservative form as

$$\frac{\partial f}{\partial t} + \nabla_{(x,v)} \cdot (Vf) = 0. \tag{3.18}$$

The semi-Lagrangian method by Cheng and Knorr (1976)
We start describing one of the firsts semi-Lagrangian schemes designed for the Vlasov-Poisson system (Cheng and Knorr 1976). The method is based on a Strang splitting between the transport and force term and on cubic spline interpolation. Starting from f^n we have the following algorithm based on the classical Strang splitting method to compute f^{n+1}.

1 Compute the electric filed E^n through the solution of the Poisson equation (3.15) with density ρ^n.
2 Compute f^* solving
$$\frac{\partial f}{\partial t} + E^n \frac{\partial f}{\partial v} = 0,$$
with initial data f^n, for a half time step $\Delta t/2$, through reconstruction of the distribution function on the characteristic curve in the velocity space.
3 Compute f^{**} solving
$$\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = 0,$$
with f^* as initial data, for a time step Δt, through reconstruction of the distribution function on the characteristic curve in the physical space.
4 Compute $\rho^{n+1} = \int_{\mathbb{R}} f^{**}(x,v)dv$ and evaluate the relative electric field E^{n+1} through the solution of the Poisson equation (3.15).
5 Compute f^{n+1} solving for a half time step $\Delta t/2$
$$\frac{\partial f}{\partial t} + E^{n+1} \frac{\partial f}{\partial v} = 0,$$
with initial data f^{**}.

To solve the Poisson equation any classical or more sophisticated numerical methods for elliptic equations can be employed, these techniques will not be discussed here.

The semi-Lagrangian method by Sonnendrücker et al. (1999)
The semi-Lagrangian methods with splitting for the resolution of the Vlasov-Poisson system has the big advantage that the characteristic equation can be solved explicitly at each step of the splitting procedure. However, the splitting introduce errors privileging the directions. It is then interesting to consider the construction of semi-Lagrangian methods directly without splitting. These methods, however, need a suitable numerical approximation of the characteristic equation (Sonnendrücker et al. 1999). This characteristic curve is solution of

$$\frac{dV}{dt} = E(X(t), t), \quad \frac{dX}{dt} = V. \tag{3.19}$$

Unfortunately the above equations cannot be solved exactly since the electric field E is computed through the Poisson equation which depends on the evolution of the distribution of particles f. An algorithm which permits to pass from time t^n to t^{n+1} can be written as following. Suppose to know at time t^n the distribution f^n and the electric potential E^n on the mesh points, then a second order in time approach is summarized below.

1 Compute a first or tentative value of the electric potential $\tilde{E}^{n+1}$ at time t^{n+1}.
2 Compute for all points in the phase space (x_j, v_k) the characteristics

$$\begin{aligned} V^{n+1/2} &= V^{n+1} - \frac{\Delta t}{2}\tilde{E}^{n+1}(X^{n+1}), \\ X^n &= X^{n+1} - \Delta t V^{n+1/2}, \\ V^n &= V^{n+1/2} - \frac{\Delta t}{2}\tilde{E}^n(X^n). \end{aligned}$$

3 Compute the interpolation of f^n at points (X^n, V^n).
4 We have then a first approximation of the distribution function on the mesh points $f^{n+1}(x_j, v_k) = f^n(X^n, V^n)$ which we can use for correct the value of E^{n+1}.
5 Perform successively step up to a prescribed error is reached.

The initial value of the electric field can be computed by solving the transport term of the Vlasov equation by using the classical semi-Lagrangian approach and then by solving the Poisson equation to get $\tilde{E}^{n+1}$.

Positive flux-conservative schemes
These schemes are based on a conservative reconstruction strategy along the characteristics curves. Using time splitting, we can restrict ourselves, to the discretization of the following one dimensional transport equation

$$\partial_t f + \partial_x (v\, f) = 0, \tag{3.20}$$

where we assume $v > 0$ is a constant velocity. By symmetry one constructs the method for $v < 0$. Now, let us introduce the mesh points $x_{j+1/2} = j\Delta x + \Delta x/2$, $j \in \mathbb{Z}$. Assume the values of the distribution function are known at time $t^n = n\,\Delta t$, we compute the new values at time t^{n+1} by integration of the distribution function on each cell. Thus, using the explicit expression of the solution, we have

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

then, setting

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

we obtain the conservative form

$$\int_{x_{j-1/2}}^{x_{j+1/2}} f(t^{n+1},x)dx = \int_{x_{j-1/2}}^{x_{j+1/2}} f(t^n,x)dx \,+\, G_{j-1/2}(t^n) \,-\, G_{j+1/2}(t^n).$$

The main step is now to choose an efficient method to reconstruct the distribution function from the values on each cell $[x_{j-1/2}, x_{j+1/2}]$. If we denote by

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

the simplest choice proposed by Fijalkow (1999) is based on a linear interpolation procedure

$$f_{\Delta x}(x) = f_j + (x - x_j)\frac{f_{j+1} - f_{j-1}}{2\Delta x}, \tag{3.21}$$

and permits an explicit computation of the fluxes. Unfortunately the resulting method does not preserve positivity.

Another approach is based on a reconstruction via primitive function (Filbet et al. 2001). A reconstruction method allowing to preserve positivity and maximum principle can be obtained using a third-order reconstruction with slope correctors

$$\begin{aligned} f_{\Delta x}(x) = f_j + \\ +\quad & \frac{\theta_j^+}{6\,\Delta x^2}\Big[2\,(x - x_j)(x - x_{j-3/2}) + (x - x_{j-1/2})(x - x_{j+1/2})\Big](f_{j+1} - f_j) \\ +\quad & \frac{\theta_j^-}{6\,\Delta x^2}\Big[2\,(x - x_j)(x - x_{j+3/2}) + (x - x_{j-1/2})(x - x_{j+1/2})\Big](f_j - f_{j-1}), \end{aligned} \tag{3.22}$$

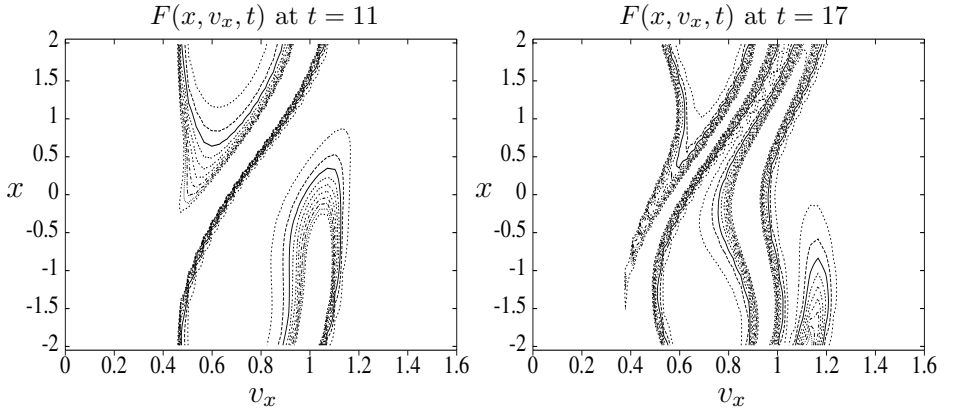

Figure 3.2. Vlasov-Poisson. Evolution of $F(x, v_x, t) = \int_{\mathbb{R}} f(x, v_x, v_y, t)\, dv_y$ in phase space computed with 32 cells in x and 64×64 in (v_x, v_y).

with

$$\theta_j^{\pm} = \begin{cases} \min\Big\{1; \dfrac{2\, f_j}{f_{j\pm1} - f_j}\Big\}, & \text{if } f_{j\pm1} - f_j > 0, \\ \\ \min\Big\{1; -\dfrac{2\,(f_{\max} - f_j)}{f_{j\pm1} - f_j}\Big\}, & \text{if } f_{j\pm1} - f_j < 0, \end{cases} \tag{3.23}$$

where $f_{\max} = \max_j\{f_j\}$. The theoretical properties of this reconstruction can be summarized by the following (Filbet et al. 2001)

Proposition 3.1. The approximation of the distribution function $f_{\Delta x}(x)$, defined by (3.22)-(3.23), satisfies:

(i) Conservation of the average

$$\int_{x_{j-1/2}}^{x_{j+1/2}} f_{\Delta x}(x)dx = \Delta x\, f_j, \quad \forall\, j.$$

(ii) Maximum principle

$$0 \le f_{\Delta x}(x) \le f_\infty, \quad \forall\, x.$$

As a numerical example we consider the above scheme applied to the Vlasov-Poisson equation, 1D in space and 2D in velocity, with initial data

$$f(0, x, v) = \frac{1}{2\pi\,\sigma^2}\, \mathrm{e}^{-|v|^2/2\sigma^2}\, \left(1 + \alpha \cos(2\pi\, x/L)\right), \quad \forall\, x \in (0, L), \quad v \in \mathbb{R}^2,$$

where $\sigma = 0.24$, $\alpha = 0.5$ and $L = 4$. The boundary conditions are periodic in space. The Vlasov equation develops thin filaments in phase space and the steep gradients in v are well-described by the method (see Figure 3.2).

Finally, concerning mathematical results on the convergence analysis of semi-Lagrangian schemes applied to the Vlasov-Poisson problem we refer to Filbet (2001), Besse (2004), Besse and Sonnendrücker (2003), Campos Pinto and Mehrenberger (2008), Respaud and Sonnendrücker (2011), Charles, Després and Mehrenberger (2013) and the references therein.

3.3. Semi-Lagrangian schemes for BGK type equations

Coupling the previous semi-Lagrangian schemes with a collision term can be done in a straightforward way through the splitting method (2.57)-(2.58). Here we present two approaches designed specifically to take advantage of the semi-Lagrangian formulation. We shall restrict to present the schemes for the BGK equation in one space dimension

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \nu(M[f] - f), \tag{3.24}$$

where $\nu > 0$ is assumed constant.

A semi-Lagrangian methods for the BGK equation

The numerical scheme for the solution of (3.24) is based on the characteristic formulation of the problem

$$\frac{df}{dt} = \nu(M\,[f] - f), \qquad \frac{dx}{dt} = v. \tag{3.25}$$

Let $f^n_{j,k}$ denote the approximate solution of the problem (3.25) at time t^n in each spatial and velocity node $x_j = j\Delta x$, $v_k = k\Delta v$, $j, k \in \mathbb{Z}$. An explicit first order semi-Lagrangian scheme could be constructed by first computing the distribution f at successive times at positions $x_j + v_k\Delta t$

$$f(x_j + v_k\Delta t, v_k, t^{n+1}) = f^n_{j,k}(1 - \Delta t\nu) + \Delta t\nu M^n_{j,k}, \tag{3.26}$$

which does not lie on a grid and then compute the values of $f^{n+1}_{j,k}$ by reconstruction from the computed values $f(x_j + v_k\Delta t, v_k, t^{n+1})$ by a suitable interpolation back on the grid points. Linear reconstruction will be sufficient for first order scheme, Hermite cubic splines have been used in Filbet and Russo (2009). Higher order reconstruction, such as ENO or WENO, could be used to provide higher order non oscillatory reconstruction (Shu 2009).

In order to advance in time we must define the approximated Maxwellian distribution $M^n_{j,k}$. The simplest method to do that is given by the following relation

$$M^n_{j,k} = \frac{\rho^n_j}{(2\pi RT^n_j)^{1/2}} \exp\left(-\frac{|v_k - u^n_j|^2}{2RT^n_j}\right), \tag{3.27}$$

where ρ^n_j, T^n_j and u^n_j are approximations of the moments at the grid point x_j

at time t^n. This formula requires the computation of the discrete moments of $f^n_{j,k}$ by some kind of quadrature. For example by simple summations

$$\rho^n_j = \Delta v \sum_h f^n_{j,h},\ u^n_j = \frac{\Delta v}{\rho^n_j} \sum_h v_h f^n_{j,h},\ T^n_j = \frac{\Delta v}{R\rho^n_j} \sum_h (v_h - u^n_j) f^n_{j,h}. \quad (3.28)$$

Note that (3.27) is not compactly supported in the velocity space and therefore this poses the question of the truncation of the velocity domain and the loss of conservation properties. We postpone a detailed discussion of this problem at the end of the Section.

Because of the semi-Lagrangian nature of the method, there is no CFL-type stability restriction on the time step due to convection. However, such scheme suffer from stability restriction on the time step when the collision rate ν is large. To circumvent the problem, it is possible to resort to an implicit formulation. By applying simple implicit Euler on the characteristic equation in order to compute $f^{n+1}_{j,k}$ one obtains

$$\begin{aligned} f^{n+1}_{j,k} &= f(t^n, x_j - v_k\Delta t, v_k) + \Delta t\nu(M^{n+1}_{j,k} - f^{n+1}_{j,k}) \\ &= \frac{1}{1+\Delta t\nu} f(t^n, x_j - v_k\Delta t, v_k) + \frac{\Delta t\nu}{1+\Delta t\nu} M^{n+1}_{j,k}. \end{aligned} \quad (3.29)$$

The quantity $f(t^n, x_j - v_k\Delta t, v_k)$ can be computed by suitable reconstruction from $f^n_{j,k}$ as in the explicit case.

The equation (3.29) cannot be immediately solved for $f^{n+1}_{j,k}$, because now the Maxwellian depends from $f^{n+1}_{j,k}$ itself. Note, however, that if the discrete Maxwellian at time t^{n+1} has exactly the same first three moments as $f^{n+1}_{j,k}$

$$\sum_h M^{n+1}_{j,h} \varphi_h = \sum_h f^{n+1}_{j,h} \varphi_h, \qquad \varphi_h = 1, v_h, |v_h|^2, \quad (3.30)$$

then from (3.29) we have

$$\sum_h f^{n+1}_{j,h} \varphi_h = \sum_h f(t^n, x_j - v_h\Delta t, v_h)\varphi_h, \qquad \varphi_h = 1, v_h, |v_h|^2. \quad (3.31)$$

As a consequence, the moments at time t^{n+1} can be computed from the solution at time t^n and then from those moments the equilibrium distribution $M^{n+1}_{j,k}$ can be obtained. Note that, for consistency, in this case we must construct the approximated Maxwellian values $M^{n+1}_{j,k}$ in such a way that (3.30) are exactly satisfied. Since this is a transversal problem to most schemes which use a finite grid over a bounded velocity domain we will discuss this at the end of this Section.

Higher order implicit semi-Lagrangian methods for relaxation operators can be constructed using L-stable diagonally implicit Runge Kutta (DIRK) schemes (Santagati et al. 2012, Pareschi and Russo 2011). See also Pareschi

(1998) for a related approach for more general collision terms based on time relaxed discretizations (Gabetta et al. 1997).

Remark 3.1.

- If the time step is such that $\Delta t = \Delta x/\Delta v$ then the foot of the characteristic is a grid point and no interpolation is required. In such case the schemes (3.26) and (3.29) becomes particular cases of Lattice Boltzmann Methods (LBM), although usually such methods are constructed from a BGK model with a limited number of velocities and mainly as a computational tool for the incompressible Navier-Stokes equations (Succi 2001).
- The resulting schemes are unconditionally stable, since no stability condition on Δt is required. However, taking large time steps will cause large numerical diffusion in the solution. In particular semi-Lagrangian schemes may suffer of accuracy degradation in the fluid limit, or equivalently for very large values of ν. The latter aspect can be understood by observing that the characteristic speeds of the system change in such a limit. Therefore, if one is interested in high order schemes close to fluid regimes usually other approaches are preferable (see Section 7).

Fast semi-Lagrangian methods for the BGK model

The method described in this paragraph combines the advantages of semi-Lagrangian methods with the structural simplicity of particle in cell methods with the aim to achieve maximum computational efficiency.

Let $f^n_{j,k}$ and $M^n_{j,k}$ be the pointwise distribution and equilibrium distribution at time n. After a splitting of the equation into transport and collisions

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = 0, \qquad \frac{\partial f}{\partial t} = \nu(M[f] - f),$$

a fast semi-Lagrangian method is constructed in the following way.

Transport stage. For each velocity v_k solve the transport equations

$$\frac{\partial f_k}{\partial t} + v_k\frac{\partial f_k}{\partial x} = 0, . \tag{3.32}$$

in the whole computational domain instead of solving them only on the mesh points. To this aim, let us define for each of the N equations a piecewise constant function in space as

$$\bar{f}_k(x, t^n) = f^n_{j,k} \quad \forall\, x \in [x_{j-1/2}, x_{j+1/2}]. \tag{3.33}$$

Now, the exact solution of the above equations is given by

$$\bar{f}^*_k(x) = \bar{f}_k(x - v_k\Delta t, t^n). \tag{3.34}$$

Relaxation stage. This step is local to the grid, this means that we solve the following ordinary differential equation

$$\frac{\partial f_{j,k}}{\partial t} = \nu(M_{j,k} - f_{j,k}),, \tag{3.35}$$

where the initial datum is the result of the transport step at points x_j

$$\bar{f}_k^*(x_j) = \bar{f}_k(x_j - v_k\Delta t, t^n). \tag{3.36}$$

To solve equation (3.35) we need the value of the equilibrium distribution at the center of the cell after the transport stage. The macroscopic quantities in the center of the cells are given by

$$\Delta v \sum_k f_{j,k}^* \varphi_k = \Delta v \sum_k \bar{f}_k^*(x_j)\varphi_k.$$

The discrete equilibrium distribution $M_{j,k}^* = M_{j,k}^{n+1}$ is then defined as the equilibrium distribution with the moments of $f_{j,k}^*$. This can be done efficiently using the conservative method described in the next paragraph. We can now solve the relaxation stage exactly

$$f_{j,k}^{n+1} = \exp(-\nu\Delta t) f_{j,k}^* + (1 - \exp(-\nu\Delta t)) M_{j,k}^*. \tag{3.37}$$

The above equation furnishes the new values of the distribution f in the center of each spatial cell for each velocity v_k. However, in order to continue the computation, we need the value of the distribution f in all points of the space. To this aim one defines

$$\bar{M}_k^*(x) = M_{j,k}^{n+1}, \ \ \forall\, x \quad \text{such that} \quad \bar{f}_k^*(x) = \bar{f}_k^*(x_j), \tag{3.38}$$

which consists in the assumption that the equilibrium distribution $M[f]$ has the same piecewise constant structure of the distribution f. We can now rewrite the relaxation term directly in term of spatial continuous function and define f at time $n+1$

$$\bar{f}_k^{n+1}(x) = \exp(-\nu\Delta t)\bar{f}_k^*(x) + (1 - \exp(-\nu\Delta t))\bar{M}_k^*(x). \tag{3.39}$$

The method essentially behaves like a particle method on piecewise constant functions and achieve high efficiency by avoiding interpolation. The scheme has been subsequently extended to achieve high order in the fluid limit (Dimarco and Loubère 2013*b*).

Conservative methods

Finally, we discuss the problem of loss of conservations due to the truncation of the velocity space and the introduction of a velocity grid (see also Section 4.1 for a different approach). Of course, a remedy to the problem consists in choosing the velocity support such that during all the simulation the source of error due to the velocity truncation is very small. In fact,

under the assumption that the distribution function is smooth and that the energy outside the domain is negligible the moments are approximated with spectral accuracy by replacing integrals with summations as in (3.28). In some circumstances, however, it may be desirable to construct a solver where conservations are exactly maintained during the simulation.

For the sake of simplicity in the sequel we omit the space and time dependence, since we deal with an approximation problem in the velocity space only. So let us consider $f = f(v)$, $v \in \mathbb{R}^d$, $d \geq 1$, and denote by $f_k \approx f(v_k)$, $k = 1, \dots, N$ the finite grid approximations. We want to define the grid values f_k in such a way that the macroscopic moments of f are preserved at a discrete level. We denote by $U \in \mathbb{R}^{2+d}$ the given set of moments

$$U = \int_{\mathbb{R}^d} f \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv.$$

We use notations $\mathrm{f} = (f_1, \dots, f_N)^T$ to denote the unknown set of values and $\tilde{\mathrm{f}} = (\tilde{f}_1, \dots, \tilde{f}_N)^T$ the point values $\tilde{f}_k = f(v_k)$. We also denote by $C \in \mathbb{R}^{(d+2)\times N}$ the matrix containing the parameters of the quadrature formula used to evaluate the discrete moments. Therefore we have

$$C\tilde{\mathrm{f}} \neq U,$$

and search for a vector f that it is "close" to $\tilde{\mathrm{f}}$ and such that

$$C\mathrm{f} = U. \tag{3.40}$$

In order to find a solution to the problem one can consider the constrained optimization problem: find $\mathrm{f} \in \mathbb{R}^N$ such that

$$\min \left\{ \|\tilde{\mathrm{f}} - \mathrm{f}\|_2^2 \ : \ C\mathrm{f} = U; C \in \mathbb{R}^{(d+2)\times N}, \tilde{\mathrm{f}} \in \mathbb{R}^N, U \in \mathbb{R}^{(d+2)} \right\}. \tag{3.41}$$

Problem (3.41) can be solved easily by a Lagrange multiplier method. Let $\lambda \in \mathbb{R}^{d+2}$ be the Lagrange multiplier vector, the objective function to be minimized is given by

$$L(\mathrm{f}, \lambda) = \sum_{k=1}^{N} |\tilde{f}_k - f_k|^2 + \lambda^T (C\mathrm{f} - U). \tag{3.42}$$

Next we impose

$$\begin{aligned} \frac{\partial L(\mathrm{f}, \lambda)}{\partial f_k} &= 0, \quad k = 1, \dots, N \\ \frac{\partial L(\mathrm{f}, \lambda)}{\partial \lambda_i} &= 0, \quad i = 1, \dots, d+2. \end{aligned}$$

The first condition implies $2\mathrm{f} = 2\tilde{\mathrm{f}} + C^T\lambda$ and the second $C\mathrm{f} = U$. Combining

the two results and using the fact that CC^T is symmetric and positive definite one gets

$$\lambda = 2(CC^T)^{-1}(U - C\tilde{\mathrm{f}}). \tag{3.43}$$

Therefore the problem can be solved explicitly and gives

$$\mathrm{f} = \tilde{\mathrm{f}} + C^T(CC^T)^{-1}(U - C\tilde{\mathrm{f}}). \tag{3.44}$$

In the same way, the approximated equilibrium distribution at the grid points can be defined in order to preserve some prescribed moments. Reverting now to the full space and time dependent notation used to describe the semi-Lagrangian schemes, we get

$$\mathrm{M}_j^n = \tilde{\mathrm{M}}_j^n + C^T(CC^T)^{-1}(U_j^n - C\tilde{\mathrm{M}}_j^n). \tag{3.45}$$

with U_j^n a prescribed set of moments, $\mathrm{M}_j^n = (M_{j,1}^n, \dots, M_{j,N}^n)^T$, $\tilde{\mathrm{M}}_j^n = (\tilde{M}_{j,1}^n, \dots, \tilde{M}_{j,N}^n)^T$ with $\tilde{M}_{j,k}^n = M[f](x_j, v_k, t^n)$.

The minimization problem just described was proposed in Bobylev and Rjasanow (1999) and generalized in Gamba and Tharkabhushanam (2010). We refer to the above references for examples of applications (see also Figure 4.1).

Remark 3.2. A remarkable feature of the method is that it only involves a matrix-vector multiplication. Moreover, since the matrix C depends only on the parameter of the discretization, the matrix $C^T(CC^T)^{-1}$ can be precomputed and stored in memory. This makes the technique extremely efficient for multi-dimensional computations. The method can be easily extended to preserve more given moments of the distribution. On the other hand positivity of the solution is lost in general, as well as the monotonicity property induced by the entropy inequality. For Maxwellian densities, these properties can be recovered considering a constrained minimization problem with respect to the entropy of the solution. As we will see in Section 4.1, however, solving such a minimization problem implies the solution of a system of $d + 2$ nonlinear equations at each time step.

4. Discrete velocity methods

Discrete-velocity methods represent a popular way for the approximation of the Boltzmann equation in the velocity space. Historically they originated as simplified models of the Boltzmann equation with the aim to provide a qualitative setting for the mathematical study of a rarefied gas (Carleman 1957, Broadwell 1964, Gatignol 1975, Cabannes, Gatignol and Luo 2003). Only recently they have been related to consistent velocity discretizations of the Boltzmann equation (Goldstein et al. 1989, Rogier and Schneider 1994, Schneider 1993, Palczewski et al. 1997, Mischler 1997, Panferov and Heintz 2002, Fainsilber et al. 2006). As we will see, however, their accuracy

is limited and their computational cost is considerably high if compared to stochastic techniques for the evaluation of the Boltzmann integral (see Section 6 for fast summation strategies). On the contrary they define a very robust setting when applied to simplified interaction operators, like the BGK model.

4.1. Discrete-velocity methods for the BGK model

Here we introduce the basic notions in discrete velocity methods by following the presentation in Mieussens (2000) for the BGK model defined by (2.50) that we rewrite here for completeness

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \nu(M[f] - f), \tag{4.1}$$

with initial data $f(x, v, 0) = f^0(x, v)$. The first step is based on the introduction of the discrete velocity grid which physically represents the set of admissible velocities in the discrete model.

Let $\mathcal{K} \subseteq \mathbb{Z}^3$ be a set of N_v integer vectors, and let

$$\mathcal{V} = \left\{ v_k \in \mathbb{R}^3, k \in \mathcal{K} \right\} \tag{4.2}$$

be a discrete-velocity grid of N_v points indexed by $k = (k_1, k_2, k_3) \in \mathcal{K}$, and defined by

$$v_k = (v_{k_1}, v_{k_2}, v_{k_3}) = (k_1 \Delta v_1, k_2 \Delta v_2, k_3 \Delta v_3), \tag{4.3}$$

where Δv_1, Δv_2, Δv_3 are three positive numbers characterizing the size of the mesh. To simplify notations we denote by $\Delta v_{\mathcal{K}} = \Delta v_1 \Delta v_2 \Delta v_3$. The velocity distribution f is then replaced by a vector $\mathbf{f} = (f_k(x,t))_{k \in \mathcal{K}} \in \mathbb{R}^{N_v}$ where each component $f_k(x,t)$ is assumed to be an approximation of $f(x, v_k, t)$. The fluid quantities are thus given as in the continuous case, except that integrals on $\mathbb{R}^3$ are replaced by discrete sums on $\mathcal{V}$. We can therefore define the discrete macroscopic quantities

$$\begin{aligned} \rho_{\mathcal{K}} &= \Delta v_{\mathcal{K}} \sum_{h \in \mathcal{K}} f_h, \\ \rho_{\mathcal{K}} u_{\mathcal{K}} &= \Delta v_{\mathcal{K}} \sum_{h \in \mathcal{K}} v_h f_h, \\ T_{\mathcal{K}} &= \frac{\Delta v_{\mathcal{K}}}{3R\rho_{\mathcal{K}}} \sum_{h \in \mathcal{K}} (v_h - u_{\mathcal{K}})^2 f_h, \end{aligned} \tag{4.4}$$

and discrete entropy functional

$$H_{\mathcal{K}}(\mathbf{f}) = \Delta v_{\mathcal{K}} \sum_{h \in \mathcal{K}} f_h \log(f_h). \tag{4.5}$$

This define the discrete velocity BGK model through a set of N_v differential

equations

$$\frac{\partial f_k}{\partial t} + v_k \cdot \nabla_x f_k = \nu_{\mathcal{K}}(M_k[\mathbf{f}] - f_k), \qquad \forall\, k \in \mathcal{K}, \tag{4.6}$$

where $\nu_{\mathcal{K}} = \nu(\rho_{\mathcal{K}}, T_{\mathcal{K}})$. Now the main difficulty is to define an approximation $M_k[\mathbf{f}]$ of the Maxwellian equilibrium such that conservation properties and entropy property still hold.

Discrete Maxwellian states
As already discussed in Section 3, the natural choice

$$M_k[\mathbf{f}] = \frac{\rho_{\mathcal{K}}}{(2\pi R T_{\mathcal{K}})^{3/2}} \exp\left(-\frac{|u_{\mathcal{K}} - v_k|^2}{2RT_{\mathcal{K}}}\right), \tag{4.7}$$

does not fulfill these requirements since it was derived starting from the continuous analogous of (4.4) and (4.5). The determination of a discrete Maxwellian state suggested in Gabetta et al. (1997) was based on the observation that the discrete equilibrium state $M_k[\mathbf{f}]$ should be such that $\log(M_k[\mathbf{f}]) \in \text{span}\{1, v_k, |v_k|^2\}$, i.e. it belongs to the space of collision invariants, which implies

$$M_k[\mathbf{f}] = \exp(a + b \cdot v_k + c|v_k|^2), \quad c < 0, \tag{4.8}$$

where $a, c \in \mathbb{R}$, $b \in \mathbb{R}^3$ are related to the macroscopic quantities. Note, however, that given a set of discrete moments $\rho_{\mathcal{K}}, T_{\mathcal{K}} \in \mathbb{R}_+$ and $u_{\mathcal{K}} \in \mathbb{R}^3$ we cannot compute explicitly, as for the continuum case, the solution of the nonlinear set of equations

$$\rho_{\mathcal{K}} = \Delta v_{\mathcal{K}} \sum_{h \in \mathcal{K}} M_h[\mathbf{f}], \tag{4.9}$$

$$\rho_{\mathcal{K}} u_{\mathcal{K}} = \Delta v_{\mathcal{K}} \sum_{h \in \mathcal{K}} v_h M_h[\mathbf{f}], \tag{4.10}$$

$$T_{\mathcal{K}} = \Delta v_{\mathcal{K}} \frac{1}{3R\rho_{\mathcal{K}}} \sum_{h \in \mathcal{K}} (v_h - u_{\mathcal{K}})^2 M_h[\mathbf{f}]. \tag{4.11}$$

Obviously, it must be checked that this problem admits a unique solution since, due to the particular choice of the grid, not all set of moments may be realizable by the discrete velocity model. In practice, in a numerical method, the above nonlinear system should be solved at each time step in order to compute a, b and c in functions of $\rho_{\mathcal{K}}$, $u_{\mathcal{K}}$ and $T_{\mathcal{K}}$. This can be done, for example, by a Newton-type method starting from the initial guess (4.7).

Since $\log(M_k[\mathbf{f}])$ is a linear combination of collision invariants we also have the discrete version of Boltzmann H-theorem for the BGK model

$$\sum_{h \in \mathcal{K}} (M_h[\mathbf{f}] - f_h) \log(f_h)$$

$$= \sum_{h\in\mathcal{K}} (M_h[\mathbf{f}] - f_h) \log\left(\frac{f_h}{M_h[\mathbf{f}]}\right) + \sum_{h\in\mathcal{K}} (M_h[\mathbf{f}] - f_h)\log(M_h[\mathbf{f}]) \le 0, \tag{4.12}$$

where we used (2.31) and the fact that the second summation on the right hand side of (4.12) is equal to zero.

In the space homogeneous case this corresponds to the monotonicity of the discrete H-functional

$$\frac{\partial}{\partial t} H_{\mathcal{K}}(\mathbf{f}) \le 0, \tag{4.13}$$

and

$$H_{\mathcal{K}}(\mathbf{f}) = 0 \quad \text{iff} \quad f_k = M_k[\mathbf{f}],\ \forall\, k \in \mathcal{K}. \tag{4.14}$$

The method derived in Mieussens (2000) propose to use the discrete version of entropy minimization problem (2.39). Let now $M_k[\mathbf{f}]$ be defined by the minimum of the discrete entropy, with the constraints that it must have a certain set of discrete moments , i.e. $M_k[\mathbf{f}]$ is the solution of the following problem

$$\min \left\{ H_{\mathcal{K}}(\mathbf{f}) = \Delta v_{\mathcal{K}} \sum_{h\in\mathcal{K}} f_h \log(f_h),\ f_k \ge 0, \forall\, k \in \mathcal{K} \right.$$
$$\left. \Delta v_{\mathcal{K}} \sum_{h\in\mathcal{K}} f_h \begin{pmatrix} 1 \\ v_h \\ |v_h|^2 \end{pmatrix} = \begin{pmatrix} \rho_{\mathcal{K}} \\ \rho_{\mathcal{K}} u_{\mathcal{K}} \\ \rho_{\mathcal{K}}(|u_{\mathcal{K}}|^2 + 3RT_{\mathcal{K}}) \end{pmatrix} \right\}. \tag{4.15}$$

From a numerical point of view, it is in general quite difficult to tackle the above formulation directly. In Mieussens (2000) it is proved that under a natural assumption on $\mathcal{V}$, the discrete equilibrium $M_k[\mathbf{f}]$ has an exponential form if, and only if, a strict realizability condition is fulfilled by $\mathcal{V}$.

Theorem 4.1. Let us consider a set of moments $\rho_{\mathcal{K}}, T_{\mathcal{K}} \in \mathbb{R}_+$ and $u_{\mathcal{K}} \in \mathbb{R}^3$ such that the set of nonnegative discrete distributions solutions to the corresponding minimization problem (4.15) is not empty. Then, the problem (4.15) has a unique solution $M_k[\mathbf{f}]$ called discrete equilibrium. Moreover, we assume that $\mathcal{V}$ has at least three points in each direction. Then there exists a unique vector $\alpha \in \mathbb{R}^5$, $\alpha = (a, b, c)$ such that $M_k[\mathbf{f}]$ is given by (4.8).

This shows the equivalence between the formulations of discrete equilibrium state used in Gabetta et al. (1997) and Mieussens (2000). For the sake of completeness we must point out that this notion of discrete Maxwellian was already present in the general theory of discrete velocity models (Gatignol 1975, Cabannes et al. 2003). We refer to Mieussens (2000) for further details on the computational aspects of the discrete equilibrium state $M_k[\mathbf{f}]$ from (4.9)-(4.11).

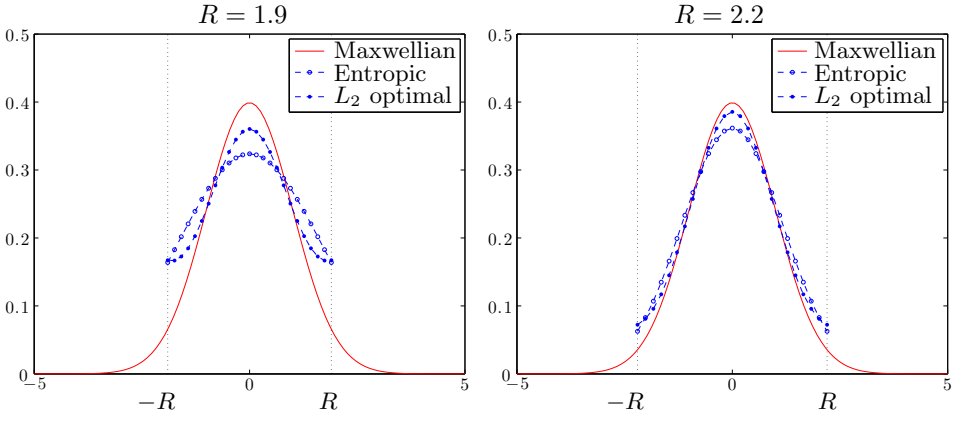

Figure 4.1. Discrete Maxwellian states in $[-R, R]$. The entropic Maxwellian is defined from the solution of (4.9)-(4.11), the L_2 optimal through (3.45). All functions share the same set of macroscopic moments $\rho_{\mathcal{K}} = 1$, $u_{\mathcal{K}} = 0$, $T_{\mathcal{K}} = 1$.

Remark 4.1.

- In contrast to the continuous case, where the minimization problem can be solved under the only assumption of nonnegative mass density and temperature, at the discrete level a moment realizability assumption is necessary since the introduction of a finite velocity grid imply also a bounded velocity domain. This clearly translates into a limited range of admissible temperatures and momentum.
- As a consequence, independently of the collision model, once $\mathcal{V}$ is chosen, a discrete velocity model cannot describe any rarefied gas flow. At variance, for a given rarefied gas flow, the discrete velocity set must be properly chosen to give a correct representation.
- For this same reason, the determination of a discrete Maxwellian state with all the relevant physical properties does not necessarily implies that this Maxwellian state is a good approximation of the continuous Maxwellian unless the support of the velocity grid is large enough (see Figure 4.1).

If we assume the grid $\mathcal{V}$ is such that the moment realizability condition is satisfied, then the following theorem summarizes the main properties of the discrete velocity BGK model.

Theorem 4.2. Let $\mathbf{f_0} = (f^0(x, v_k))_{k\in\mathcal{K}}$ be a strictly positive vector of $\mathbb{R}^{N_v}$. Consider the initial value problem associated with model (4.6), where $M_k[\mathbf{f}]$ has the form (4.8) and is implicitly defined either by (4.9)-(4.11), or by (4.15). If this problem has a solution $\mathbf{f}$, then the solution $\mathbf{f}$ remains strictly positive and the model satisfy the discrete moment equations and

entropy dissipation

$$\frac{\partial}{\partial t}\left(\Delta v_{\mathcal{K}} \sum_{h\in\mathcal{K}} f_h \varphi_h\right) + \nabla_x \cdot \left(\Delta v_{\mathcal{K}} \sum_{h\in\mathcal{K}} f_h \varphi_h v_h\right) = 0, \tag{4.16}$$

$$\frac{\partial}{\partial t}\left(\Delta v_{\mathcal{K}} \sum_{h\in\mathcal{K}} f_h \log f_h\right) + \nabla_x \cdot \left(\Delta v_{\mathcal{K}} \sum_{h\in\mathcal{K}} f_h \log f_h v_h\right) \leq 0, \tag{4.17}$$

for $\varphi_k = 1, v_k, |v_k|^2$, $k \in \mathcal{K}$.

Convergence to the BGK model

From a mathematical point of view it is interesting to consider the convergence of such approximation to the continuous BGK equation. Here, following Mieussens (2001) we sketch the main ideas of the proof which is based on three distinct steps

- convergence of the discrete velocity approximation of the collision operator (which is local in x and t);
- existence and uniqueness of solutions for the discrete velocity equation (4.6);
- convergence of the discrete velocity equation (4.6) to the continuous one (4.1).

The possibility to tackle the different steps of the above program is closely related to the simplified collision model used in the BGK formulation. In fact, in the case of discrete velocity methods for the full Boltzmann operator the first point strongly depends on the particular discrete velocity model considered and the validity of point 2 is not known in general. As we will see, for the BGK model, at least in the case of constant relaxation frequency $\nu = 1$, the whole convergence can be carried on successfully. We refer to Mieussens (2001) for the details of the proofs.

We introduce two sequences of real numbers Δv_m and B_m such that

$$\lim_{m\to\infty} \Delta v_m = 0, \quad \lim_{m\to\infty} \Delta v_m B_m = +\infty, \tag{4.18}$$

and a grid $\mathcal{V}_m$ of N_m velocities defined by

$$\mathcal{V}_m = \{v_k^m = i\Delta v_m,\, k \in \mathcal{K}_m\},$$

where $\mathcal{K}_m$ is the set of multi-indexes $\mathcal{K}_m = \{i \in \mathbb{Z}^3,\, |i| \leq B_m\}$. Moreover, we denote by $\mathbf{f}^m = (f_k^m)_{k\in\mathcal{K}_m}$, by ρ_m, u_m and T_m the corresponding discrete macroscopic quantities and by $(M_k^m[\mathbf{f}^m])_{k\in\mathcal{K}_m}$ the associated discrete equilibrium state in the entropic sense specified in the previous paragraph.

The first result shows that the discrete Maxwellian is consistent if the discrete moments are consistent.

Theorem 4.3. Let $\rho, T \in \mathbb{R}_+$ and $u \in \mathbb{R}^3$. If ρ_m, u_m and T_m are sequences of macroscopic moments strictly realizable on $\mathcal{V}_m$ such that

$$\lim_{m\to\infty} \rho_m = \rho, \quad \lim_{m\to\infty} u_m = u, \quad \lim_{m\to\infty} T_m = T, \tag{4.19}$$

then the discrete Maxwellian $(M_k^m[\mathbf{f}^m])_{k\in\mathcal{K}_m}$ converges to the continuous Maxwellian $M[f]$ given by (2.33).

The second result is a consequence of Theorem 4.2 and the theory of Perthame (1989), and shows existence and uniqueness for the discrete BGK model (4.6). However, it is important to note that the existence theory in Perthame (1989) is based on a constant relaxation time $\nu = 1$, therefore the same assumption will be made in the rest of this paragraph.

Theorem 4.4. The initial value problem

$$\frac{\partial f_k^m}{\partial t} + v_k^m \cdot \nabla_x f_k^m = M_k^m[\mathbf{f}^m] - f_k^m, \qquad \forall\, k \in \mathcal{K}_m, \tag{4.20}$$

$$f_k^m(x,0) = f^0(x, v_k^m), \tag{4.21}$$

has a unique solution $(f_k^m)_{k\in\mathcal{K}_m}$ in $L^\infty(]0, t_{max}[\times\mathbb{R}^3)^{N_m}$ for all $t_{max} > 0$.

For the last point, following Mischler (1997), one defines a continuous form of the discrete velocity BGK model, and use the stability proof of Perthame for the BGK equation (Perthame 1989). We define the constant per velocity cell functions

$$f^m(x,v,t) = \sum_{h\in\mathcal{K}_m} f_h^m(x,t)\chi_h^m(v), \tag{4.22}$$

$$M^m[f^m](x,v,t) = \sum_{h\in\mathcal{K}_m} M_h^m[\mathbf{f}^m](x,t)\chi_h^m(v), \tag{4.23}$$

where $\chi_k^m(\cdot)$ is the indicator function of the velocity cell C_k^m centered in v_k^m of side Δv_m

$$C_k^m = \left[v_{k_1}^m - \frac{\Delta v_m}{2}, v_{k_1}^m + \frac{\Delta v_m}{2}\right[\times \ldots \times \left[v_{k_3}^m - \frac{\Delta v_m}{2}, v_{k_3}^m + \frac{\Delta v_m}{2}\right[. \tag{4.24}$$

Then we can relate the discrete model (4.6) with the continuous model (4.1) by the equation

$$\frac{\partial f^m}{\partial t} + \nu^m \cdot \nabla_x f^m = M^m[f^m] - f^m, \tag{4.25}$$

with

$$\nu^m(v) = \sum_{h\in\mathcal{K}_m} v_h^m \chi_h^m(v),$$

and initial data

$$f^m(x,v,0) = \sum_{h\in\mathcal{K}_m} f_h^m(x,0)\chi_h^m(v). \tag{4.26}$$

We can finally state the convergence result:

Theorem 4.5. For all sequences Δv_m, B_m satisfying (4.18), the sequence $\{f^m\}_{m\geq 0}$ solution to the initial value problem (4.25)-(4.26) is weakly convergent in $L_1(]0,t_{max}[\times\mathbb{R}^3\times\mathbb{R}^3)$, $\forall\, t_{max} > 0$, up to the extraction of a subsequence, to a distribution solution of the BGK equation (4.1).

Fully discrete schemes

To obtain a fully discrete scheme suitable choices of space and time approximations have to be chosen. Concerning the space discretization there are several options since in principle one can apply the whole literature of numerical methods for hyperbolic conservations laws. We already discussed semi-lagrangian approaches in Section 3 and how they can be coupled with source terms. Here we report as an example the second order finite volume method used in Mieussens (2000) based on the use of flux limiters. We first discuss the properties of explicit time discretization and then consider the case of implicit methods by postponing a more complete discussion on the fluid-limit problem to Section 7.

For the sake of simplicity, the scheme is presented here in two spatial dimensions but all the results extends naturally to dimension three. We consider a spatial Cartesian grid defined by nodes $(x_i,y_i) = (i\Delta x, j\Delta y)$ and cells

$$I_{ij} = \Big]x_{i-\frac12}, x_{i+\frac12}\Big[\times \Big]y_{j-\frac12}, y_{j+\frac12}\Big[.$$

Introducing a time step Δt and the time levels $t^n = n\Delta t$, using the formalism of the first paragraph, we denote by $f_{ij,k}^n = f_k(x_i,y_j,t^n)$, $\forall\, k\in\mathcal{K}$, by $\rho_{ij}^n = \rho_{\mathcal{K}}(x_i,y_j,t^n)$, $u_{ij}^n = u_{\mathcal{K}}(x_i,y_j,t^n)$ and $T_{ij}^n = T_{\mathcal{K}}(i_h,y_j,t^n)$ the discrete moments and by $M_{ij,k}^n[\mathbf{f}] = M_k[\mathbf{f}](x_i,y_j,t^n)$ the discrete Maxwellian equilibrium. We further denote with $\nu_{ij}^n = \nu(\rho_{ij}^n, T_{ij}^n)$.

A finite volume explicit discretization in the time interval $[t, t+\Delta t]$ reads

$$\begin{aligned} f_{ij,k}^{n+1} &= f_{ij,k}^n - \frac{\Delta t}{\Delta x}\left(F_{i+\frac12,j,k}^n - F_{i-\frac12,j,k}^n\right) \\ &- \frac{\Delta t}{\Delta y}\left(F_{i,j+\frac12,k}^n - F_{i,j-\frac12,k}^n\right) + \Delta t\nu_{ij}^n\left(M_{ij,k}^n[\mathbf{f}] - f_{ij,k}^n\right), \end{aligned} \tag{4.27}$$

where

$$\begin{aligned} F_{i+\frac12,j,k}^n &= \frac12\left[v_{k_1}\left(f_{i+1,j,k}^n + f_{ij,k}^n\right) - |v_{k_1}|\left(\Delta f_{i+\frac12,j,k}^n - \Phi_{i+\frac12,j,k}^n\right)\right], \\ F_{i,j+\frac12,k}^n &= \frac12\left[v_{k_2}\left(f_{i,j+1,k}^n + f_{ij,k}^n\right) - |v_{k_2}|\left(\Delta f_{i,j+\frac12,k}^n - \Phi_{i,j+\frac12,k}^n\right)\right], \end{aligned} \tag{4.28}$$

are the numerical fluxes, $\Delta f^n_{i+\frac{1}{2},j,k} = f^n_{i+1,j,k} - f^n_{ij,k}$, and $\Phi^n_{i+1/2,j,k}$ is the slope limiter function. A first order method is obtained when $\Phi^n_{i+1/2,j,k} = \Delta f^n_{i+1/2,j,k}$, second order accuracy is achieved with a suitable limiter choice, for example

$$\Phi^n_{i+\frac{1}{2},j,k} = \text{minmod}\left(\Delta f^n_{i-\frac{1}{2},j,k}, \Delta f^n_{i+\frac{1}{2},j,k}, \Delta f^n_{i+\frac{3}{2},j,k}\right), \tag{4.29}$$

where the minmod limiter is defined as

$$\text{minmod}\,(z_1, z_2, z_3) = \begin{cases} \min\limits_i\{z_i\}, & \text{if} \quad z_i > 0 \quad \forall\, i, \\ \max\limits_i\{z_i\}, & \text{if} \quad z_i < 0 \quad \forall\, i, \\ 0 & \text{otherwise.} \end{cases}$$

For first order space discretizations it is possible to prove (Mieussens 2000)

Proposition 4.1. Let $f^n_{ij,k}$ be strictly positive. If the time steps satisfy

$$\Delta t \left(\max_{ij} \nu^n_{ij} + \max_{k\in\mathcal{K}} \left(\frac{|v_{k_1}|}{\Delta x} + \frac{|v_{k_2}|}{\Delta y}\right)\right) \leq 1, \tag{4.30}$$

then the numerical solution $f^{n+1}_{ij,k}$ defined by the first order scheme (4.27)-(4.28) remains strictly positive. Furthermore, the total mass, momentum, and energy are conserved, and the total entropy is decreasing.

In dense or rapidly varying regimes the value of ν can be very large and the above time step condition becomes very restrictive. This is the case, for example, of regions close to fluid regimes. A classical way to overcome this difficulty is to use an implicit scheme. Similarly to the semi-Lagrangian scheme (3.29) we can avoid the solution of large nonlinear system of equation at each time step thanks to the conservation properties of the resulting method. An implicit evaluation leads to the scheme

$$\begin{aligned} f^{n+1}_{ij,k} &= f^n_{ij,k} - \frac{\Delta t}{\Delta x}\left(F^n_{i+\frac{1}{2},j,k} - F^n_{i-\frac{1}{2},j,k}\right) \\ &- \frac{\Delta t}{\Delta y}\left(F^n_{i,j+\frac{1}{2},k} - F^n_{i,j-\frac{1}{2},k}\right) + \Delta t \nu^{n+1}_{ij}\left(M^{n+1}_{ij,k}[\mathbf{f}] - f^{n+1}_{ij,k}\right). \end{aligned} \tag{4.31}$$

If we now multiply the scheme by the collision invariants $\varphi_k = 1, v_k, |v_k|^2$ and sum over $k \in \mathcal{K}$ we obtain the explicit moment scheme

$$\begin{aligned} \sum_{h\in\mathcal{K}} f^{n+1}_{ij,h}\varphi_h &= \sum_{h\in\mathcal{K}} f^n_{ij,h}\varphi_h - \frac{\Delta t}{\Delta x}\sum_{h\in\mathcal{K}}\left(F^n_{i+\frac{1}{2},j,h} - F^n_{i-\frac{1}{2},j,h}\right)\varphi_h \\ &- \frac{\Delta t}{\Delta y}\sum_{h\in\mathcal{K}}\left(F^n_{i,j+\frac{1}{2},h} - F^n_{i,j-\frac{1}{2},h}\right)\varphi_h = 0. \end{aligned} \tag{4.32}$$

Note that the above result strictly depends on the exact conservation properties of the discrete Maxwellian equilibrium. This shows that the values of

ρ_{ij}^{n+1}, u_{ij}^{n+1} and T_{ij}^{n+1} can be computed explicitly and permits to define the conservative and entropic discrete Maxwellian at time t^{n+1} in the sense of (4.15). Therefore we obtain the explicit formulation

$$\begin{aligned} f_{ij,k}^{n+1} &= \frac{1}{1+\Delta t\nu_{ij}^{n+1}}\left[f_{ij,k}^{n} - \frac{\Delta t}{\Delta x}\left(F_{i+\frac{1}{2},j,k}^{n} - F_{i-\frac{1}{2},j,k}^{n}\right)\right. \\ &- \left.\frac{\Delta t}{\Delta y}\left(F_{i,j+\frac{1}{2},k}^{n} - F_{i,j-\frac{1}{2},k}^{n}\right)\right] + \frac{\Delta t\nu_{ij}^{n+1}}{1+\Delta t\nu_{ij}^{n+1}} M_{ij,k}^{n+1}[\mathbf{f}]. \end{aligned} \tag{4.33}$$

Remark 4.2.

- The possibility to evaluate explicitly the implicit BGK collision term through the moment system has been used by several authors, for example in Pieraccini and Puppo (2007), Filbet and Russo (2009), Filbet and Jin (2010), Dimarco and Pareschi (2013). More in general the property extends also to higher order schemes (see Remark 7.4 in Section 7) and to more refined BGK models, like the ES-BGK (Filbet and Jin 2011).
- The stability condition (4.30) may become very restrictive also in the case of large velocities. Implicit evaluations of both the source term and the fluxes have been considered in Mieussens (2000), where the large linear system originated by the implicit fluxes is solved using a suitable iterative method adapted to the different sparse structures of the matrices. Alternative implicit strategies for the same problem have been proposed in Pieraccini and Puppo (2012).

4.2. Discrete velocity methods for the Boltzmann equation

The discrete velocity models of the Boltzmann equation supply a clarifying example of the difficulties one encounters when the discretization of the full Boltzmann collision operator is tackled. To introduce the problem let us rewrite here the Boltzmann equation

$$\frac{\partial f}{\partial t} + v\cdot\nabla_x f = Q(f,f), \tag{4.34}$$

and the expression of the integral collision term

$$Q(f,f)(v) = \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} B(v,v_*,\omega)[f(v')f(v'_*) - f(v)f(v_*)]\,d\omega\,dv_*. \tag{4.35}$$

Once a discrete velocity grid of the type (4.2)-(4.3) has been introduced, the relatively simple problem in the BGK case of the definition of a discrete Maxwellian that preserves conservation and entropy property, here turns into the challenging problem of developing a quadrature formula for (4.35) preserving the same kind of properties.

To start with we consider a regular grid $\mathcal{V}$ in $\mathbb{R}^3$, $\mathcal{K} \subseteq \mathbb{Z}^3$ a set of integer vectors, and use the same notation introduced in (4.2)-(4.3) with $v_i = i\Delta v = (i_1\Delta v, i_2\Delta v, i_3\Delta v)$, $i \in \mathcal{K}$, Δv the mesh size (note that the reference velocity index in the sequel is 'i' instead of 'k'). The general discrete velocity Boltzmann model is written in the form

$$\frac{\partial f_i}{\partial t} + v_i \cdot \nabla f_i = Q_i^{\mathcal{K}}(\mathbf{f}, \mathbf{f}), \quad i \in \mathcal{K}, \tag{4.36}$$

and the crucial point is now the definition of the quadrature method $Q_i^{\mathcal{K}}(\mathbf{f}, \mathbf{f})$ for $Q(f, f)(v_i)$.

Discretization of the collision operator
A general quadrature formula for (4.35) can be written in the form

$$Q_i^{\mathcal{K}}(\mathbf{f}, \mathbf{f}) = \sum_{j \in \mathcal{K}} \sum_{\alpha \in \Theta} W_{ij}^{\alpha} B(v_i, v_j, \omega_\alpha) \left(f_i^{\alpha} f_j^{\alpha} - f_i f_j \right),$$

where W_{ij}^{α} are the weight of the quadrature formula, Θ is a set of indices characterizing the discrete set of angles such that $\alpha = (\alpha_1, \alpha_2) \in \Theta$ and $\omega_\alpha = (\omega_{\alpha_1}, \omega_{\alpha_2}) \in \Omega \subset \mathbb{S}^2$. The main problem is that $f_i^{\alpha} = f(v_i^{\alpha})$ and $f_j^{\alpha} = f(v_j^{\alpha})$ requires some kind of interpolation, since the collisional velocities v_i^{α} and v_j^{α} defined through

$$v_i^{\alpha} = \frac{1}{2}(v_i + v_j + |v_i - v_j|\omega_\alpha), \quad v_j^{\alpha} = \frac{1}{2}(v_i + v_j - |v_i - v_j|\omega_\alpha), \quad i, j \in \mathcal{K} \tag{4.37}$$

are not in $\mathcal{V}$. Unfortunately the construction of interpolation formulas such that the conservation properties and entropy dissipation are preserved at a discrete level results into a very difficult problem, see Tcheremissine (2006) for interpolation strategies and Buet, Cordier and Degond (1998) for an approach based on a regularized collision term.

Therefore discrete velocity methods focus on quadrature formulas of the form

$$Q_i^{\mathcal{K}}(\mathbf{f}, \mathbf{f}) = \sum_{j,k,l \in \mathcal{K}} \Gamma_{ij}^{kl} (f_k f_l - f_i f_j), \tag{4.38}$$

where the sum acts only over the grid points that satisfy local conservation of momentum and energy

$$v_k + v_l = v_i + v_j, \quad |v_k|^2 + |v_l|^2 = |v_i|^2 + |v_j|^2. \tag{4.39}$$

Note that this choice intrinsically implies a selection of discrete angular vectors $\omega_{ij}^{kl} \in \mathbb{S}^2$ as functions of the colliding pairs (v_i, v_j) and (v_h, v_k). We will discuss this aspect later when considering the consistency of such methods. The quantities $\Gamma_{ij}^{kl} \geq 0$ are related to the weights W_{ij}^{kl} through the equations

$$\Gamma_{ij}^{kl} = \mathbf{1}(i + j - k - l)\, \mathbf{1}(|i|^2 + |j|^2 - |k|^2 - |l|^2)\, W_{ij}^{kl} B_{ij}^{kl}, \tag{4.40}$$

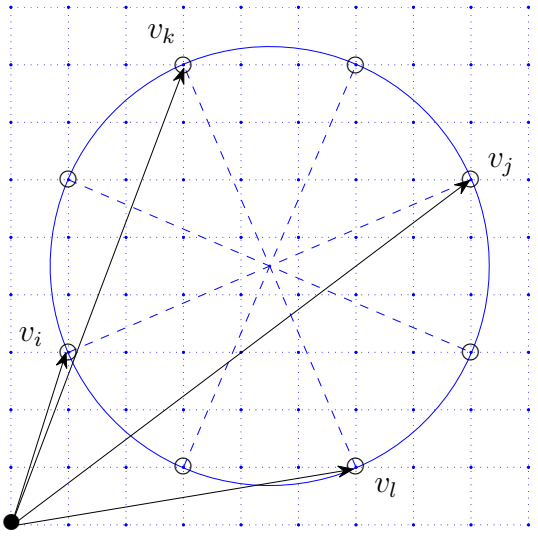

Figure 4.2. Sketch of the collision circle in a planar discrete velocity regular grid.

where $\mathbf{1}$ denotes the function on $\mathbb{Z}$ defined by $\mathbf{1}(z) = 1$ if $z = 0$ and 0 elsewhere, and $B_{ij}^{kl} = B(v_i, v_j, \omega_{ij}^{kl})$. From a physical viewpoint these quantities are linked to the probability that two particles with velocities v_i and v_j collide and come out of the collision with velocities v_k and v_l. We have the following

Proposition 4.2. If Γ_{ij}^{kl} in (4.38) satisfy the following symmetry properties

$$\Gamma_{ij}^{kl} = \Gamma_{kl}^{ij}, \qquad \Gamma_{ij}^{kl} = \Gamma_{ji}^{kl} = \Gamma_{ij}^{lk} = \Gamma_{ji}^{lk}, \quad \forall\, i,j,k,l \in \mathcal{K} \tag{4.41}$$

then the discrete Boltzmann equation (4.36)-(4.38) inherit the properties of the solution of the Boltzmann equation, namely, conservation of mass, momentum and energy and entropy inequality.

In fact, thanks to the symmetries (4.41) we have

$$\begin{aligned} &\sum_{i,j,k,l\in\mathcal{K}} \Gamma_{ij}^{kl}(f_k f_l - f_i f_j)\varphi_i \\ &\quad = -\frac{1}{4} \sum_{i,j,k,l\in\mathcal{K}} \Gamma_{ij}^{kl}(f_k f_l - f_i f_j)(\varphi_k + \varphi_l - \varphi_i - \varphi_j) \end{aligned} \tag{4.42}$$

and this shows that the model admits the collision invariants $\varphi_i = 1, v_i, |v_i|^2$. Moreover, by choosing $\varphi_i = \ln(f_i)$ we obtain the discrete analogue of Boltz-

mann H-theorem

$$\begin{aligned}&\sum_{i,j,k,l\in\mathcal{K}} \Gamma_{ij}^{kl}(f_k f_l - f_i f_j)\ln(f_i)\\ &\quad = -\frac{1}{4}\sum_{i,j,k,l\in\mathcal{K}} \Gamma_{ij}^{kl}(f_k f_l - f_i f_j)\ln\left(\frac{f_k f_l}{f_i f_j}\right) \le 0.\end{aligned} \tag{4.43}$$

Because of symmetry properties, due to the particular choice of the allowed velocities, some discrete velocity grids may originate models that possess a number of collision invariants greater than the usual one. In these cases, there is no possibility to define a unique equilibrium state with the same moments of the initial density. In what follows we ignore this possibility and refer to models that have only the usual conserved quantities. A characterization of such models is due to Cercignani (1985). For such models the discrete Maxwellian equilibrium states have the form

$$M_i[\mathbf{f}] = \exp(a + b\cdot v_i + c|v_i|^2), \quad c < 0, \tag{4.44}$$

where $a, c \in \mathbb{R}$, $b \in \mathbb{R}^3$ are related to the macroscopic quantities as in the BGK case.

Remark 4.3. From the numerical point of view, one of the main difficulty with discrete velocity quadrature formulas is the small number of pairs of discrete post collisional velocities for a given pair of pre-collisional velocities (see Figure 4.2). Indeed, the number of intersection points between the collision sphere and the discrete velocity grid may be very small. In addition the evaluation of (4.38) at each time step has at least a quadratic cost. If N_v is the total number of velocity grid points the overall cost is more then $O(N_v^2)$ whereas for probabilistic methods it is only $O(N_v)$. In Section 6 we will see how using the convolution-like structure of certain models, this cost can be reduced significantly.

Consistency of discrete velocity methods

In the general setting described in the previous paragraph one has to specify the choices of the weight accordingly to the procedure used for the derivation of the quadrature formula. Here we discuss shortly the methods proposed by Goldstein et al. (1989), by Martin et al. (1992) and Panferov and Heintz (2002), and their properties studied in Palczewski et al. (1997), Fainsilber et al. (2006), Rogier and Schneider (1994), Panferov and Heintz (2002) and Mouhot et al. (2013).

The method by Goldstein et al. (1989) . The first family of methods (Goldstein et al. 1989) can be derived after the change of variable $q = (v_* - v)/2$, and

then $v_* = v + 2q$ to get

$$Q(f,f)(v) = 2^3 \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} g(v,q,\omega)\, d\omega\, dq, \tag{4.45}$$

where we have set

$$g(v,q,\omega) = B(v, v+2q, \omega)[f(v')f(v'_*) - f(v)f(v+2q)], \tag{4.46}$$

and now $v' = v + q + |q|\omega$, $v'_* = v + q - |q|\omega$.

Using a simple rectangular formula on the $\mathbb{R}^3$ integral we can write

$$Q(f,f)(v_i) \approx (2\Delta v)^3 \sum_{j \in \mathbb{Z}^3} \int_{\mathbb{S}^2} g(v_i, q_{ji}, \omega)\, d\omega. \tag{4.47}$$

Here v_i is a given point of the grid and the sum is taken over all such points v_j so that $q_{ji} = (v_j - v_i)/2$ belongs to the grid.

The second step is the most delicate, namely the evaluation of the inner integral in terms of the values of g on the grid points. This is achieved approximating

$$\int_{\mathbb{S}^2} g(v_i, q_{ji}, \omega)\, d\omega \approx \frac{1}{N_{ij}} \sum_{(v_k, v_l) \in S_{ij}} g(v_i, q_{ij}, \omega_{ij}^{kl}), \tag{4.48}$$

where $\omega_{ij}^{kl} = (v_k - v_l)/|q_{ji}| = (k-l)/|i-j|$ and the sum is taken over all pairs of antipodal integer points that are on the sphere S_{ij} spanned by (v_i, v_j), i.e., the sphere of diameter $|q_{ij}| = |v_i - v_j|$ on which are v_i and v_j, and N_{ij} is the number of such points. For example, in Figure 4.2 we report a planar case with $N_{ij} = 3$. The above model can be cast in the general setting (4.38) by taking $\Gamma_{ij}^{kl} = (2\Delta v)^3 B_{ij}^{kl}/N_{ij}$, with the convention just described on the points involved in the sum. It is easy to verify that Proposition 4.2 is satisfied thanks to the symmetries of N_{ij} and B_{ij}^{kl} (it is nevertheless not automatic to prove that the space of summational invariants is reduced to mass, momentum, and energy).

The natural question now concerns the convergence of the resulting quadrature formula to the Boltzmann integral as $\Delta v \to 0$, $\forall\, v_i$. If g is sufficiently regular (continuous), and decays sufficiently rapidly for large q, then the quadrature formula for the outer integral (4.47) converges. As a consequence, the consistency of the whole quadrature method is closely related to the repartition of integer roots of the equation $x^2+y^2+z^2 = n$, $n \in \mathbb{N}$. Such consistency results have been obtained via number theory in Palczewski et al. (1997) in dimension $d \geq 3$ and Fainsilber et al. (2006) for the case $d = 2$. The mathematical proofs, although elegant, are rather technical and go beyond the scopes of the present survey. We limit ourselves to observe that for the convergence over the unit sphere one has to show that:

1 the number of integer points on the sphere S_{ij} is increasing to infinity as $\Delta v \to 0$;
2 the asymptotic distribution of these points on the sphere is uniform;
3 the answers to problems 1 and 2 are uniform with respect to vectors v_i and v_j.

Finally, under the following assumptions on B

$$0 \le B(v, v + 2q, \omega) \le \alpha + \beta|q|, \quad \forall\, q \in \mathbb{R}^3,\, \omega \in \mathbb{S}^2, \tag{4.49}$$

where α and β are positive constants, and considering positive and continuous solutions f of the Boltzmann equation with given polynomial decay

$$\|f\| = \sup_{v \in \mathbb{R}^3} f(v)(1 + v^2)^3 < +\infty, \tag{4.50}$$

it is possible to prove:

Theorem 4.6. Let B and f satisfy (4.49)-(4.50), then

$$\left| \sum_{j \in \mathbb{Z}^3} \frac{(\Delta v)^3}{N_{ij}} \sum_{(v_k, v_l) \in S_{ij}} g(v_i, q_{ij}, \omega_{ij}^{kl}) - \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} g(v, q, \omega)\, d\omega\, dq \right| \to 0,$$

as $\Delta v \to 0$, uniformly with respect to v_i.

Remark 4.4. Under additional smoothness assumptions on B it is also possible to derive error estimates (Palczewski et al. 1997, Fainsilber et al. 2006). These error estimates, although far from being obvious, however, are of theoretical rather than practical interest since the estimated rate of convergence is so slow that a numerical method based on such estimates would hardly ever become useful (only in particular circumstances the rate of convergence is $O((\Delta v)^{1/2})$). On the other hand, these consistency results are interesting, because they provide the necessary basis for the convergence results of Mischler (1997), who proved that solutions to families of discrete velocity models can converge to DiPerna-Lions solutions of the full Boltzmann equation if certain conditions are satisfied.

The method by Martin et al. (1992). A similar approach was introduced in Martin et al. (1992) and Rogier and Schneider (1994) using the theory of Farey series to discretize the angular variable in the collision integral. The main idea is to inverse the order of integration and integrate first over $\mathbb{S}^2$. The starting point now is the Boltzmann collision operator parametrized accordingly to (2.20), which after the change of variables $q = v_* - v$ reads

$$Q(f, f)(v) = \int_{\mathbb{S}^1} \int_{\mathbb{R}^2} q\bar{\sigma}(q, n)[f(v')f(v'_*) - f(v)f(v + q)]\, dq\, dn \tag{4.51}$$

with

$$v' = v + (q \cdot n)n, \quad v'_* = v_* - (q \cdot n)n. \tag{4.52}$$

For simplicity, we recall here the method in the 2D case. We can express $v_* = v + q$ in the cartesian frame $(n, n^\perp)$ centered in v

$$v_* = v + |v' - v|n + |v'_* - v|n^\perp = v + rn + sn^\perp, \tag{4.53}$$

and write

$$Q(f,f)(v) = \int_{\mathbb{S}^1} \int_{\mathbb{R}^2} F(v,r,s,n)\, dr\, ds\, dn \tag{4.54}$$

where we have set

$$F(v,r,s,n) = q\bar{\sigma}(q,n)[f(v+rn)f(v+sn^\perp) - f(v)f(v+rn+sn^\perp)]. \tag{4.55}$$

The quadrature formula at a grid point v_i is then based on choosing a particular set of angles $n_h \in \Theta$ such that $v_i + rn_h$ and $v_i + sn_h^\perp$ cross the grid points

$$Q(f,f)(v_i) \approx \sum_{n_h \in \Theta} W_h \int_{\mathbb{R}^2} F(v_i, r, s, n_h)\, dr\, ds. \tag{4.56}$$

Finally, since each angle has necessarily the form $n_h = (p_h, q_h)/\sqrt{p_h^2 + q_h^2}$, $n_h^\perp = (-q_h, p_h)/\sqrt{p_h^2 + q_h^2}$, $(p_h, q_h) \in \mathbb{Z}^2$ one considers the subgrid of $\mathcal{V}$ spanned by $\Delta v(p_h, q_h)$ and $\Delta v(-q_h, p_h)$ and use a rectangular formula

$$\int_{\mathbb{R}^2} F(v_i, r, s, n_h)\, dr\, ds \approx (\Delta v)^2 \sum_{(k,l)\in\mathbb{Z}^2} F(v_i, r_k^h, s_l^h, n_h), \tag{4.57}$$

where $\Delta v_h = \Delta v\sqrt{p_h^2 + q_h^2}$, $r_h^h = k\Delta v_h$ and $s_l^h = l\Delta v_h$. The set of possible angles n_h can be chosen such that either $|p_h/q_h|$ or $|q_h/p_h|$ belongs to the so-called Farey series of ordered rational numbers (Hardy and Wright 1979). Concerning the precise expressions of the weights W_h and the construction of the set Θ we refer to Schneider (1993), (see also Section 6.4 on the Farey series). It is possible to prove the following consistency result

Theorem 4.7. Let F defined in (4.55) be in $C_0^2(\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{S}^1)$ then

$$\left| \int_{\mathbb{S}^1} \int_{\mathbb{R}^2} F(v_i, r, s, n)\, dr\, ds\, dn - (\Delta v)^2 \sum_{n_h \in \Theta} W_h \sum_{(k,l)\in\mathbb{Z}^2} F(v_i, r_k^h, s_l^h, n_h) \right| \to 0,$$

as $\Delta v \to 0$ uniformly with respect to v_i.

A similar result holds true also in the three dimensional case (Martin et al. 1992, Rogier and Schneider 1994). Note that, similarly to the formula by Goldstein et al. (1989), the consistency result is non trivial and requires a careful analysis of the angular approximation based on the Farey series. Error estimates are also possible but the estimated rate of convergence is

very slow (the maximum estimated order is $O((\Delta v)^{3/7})$) and their interest is mainly theoretical.

The method by Panferov and Heintz (2002). Finally, let us now consider a different approach in the derivation of a discrete velocity quadrature method (Panferov and Heintz 2002, Mouhot et al. 2013). The method is based on the Carleman representation of the Boltzmann integral. In fact, there are some advantages in using the alternative form (4.58) of the Boltzmann integral that we rewrite here

$$Q(f,f)(v) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \tilde{B}(x,y)\delta(x\cdot y) \\ [f(v+y)\,f(v+x) - f(v+x+y)\,f(v)]\,dx\,dy, \quad (4.58)$$

with $\tilde{B}(x,y)$ given by (2.26). Thanks to equations (4.38) and (4.40), we can write at the discrete level the same representation as in the continuous case

$$Q_i(\mathbf{f},\mathbf{f}) = \sum_{k,l\in\mathbb{Z}^3} \widetilde{\Gamma}_{kl}\big[f_{i+k}f_{i+l} - f_i f_{i+k+l}\big] \quad (4.59)$$

with

$$\widetilde{\Gamma}_{kl} = \tilde{B}(k,l)\mathbf{1}(k\cdot l)\,W_{kl}, \quad (4.60)$$

where W_{kl} are the weights of the quadrature method (for the precise quadrature weights see Panferov and Heintz (2002)). One can derive the following consistency result from Panferov and Heintz (2002) in the case of hard spheres kernels

Theorem 4.8. Assume that $f \in C^k(\mathbb{R}^3)$ $(k \geq 1)$ with compact support. Then for $h > 0$ sufficiently small

$$\|Q(f,f)(v_i) - Q_i(\mathbf{f},\mathbf{f})\|_{L^\infty(\mathbb{Z}_h)} \leq C\,(\Delta v)^r$$

where Q_i is the discrete operator defined in (4.59). Here $r = k/(k+3)$ and the constant C is independent on Δv.

The proof, although far from being trivial, is somehow easier if compared to method derived from the standard representation, because the integration over a sphere in the collision term is replaced by the integration over planes in $\mathbb{R}^3$ and this allows to use simpler and more direct techniques. As can be seen from the error estimate in Theorem 4.8 and confirmed in the numerical experiments the method is expected to have slightly less then first order accuracy. A remarkable feature of this approach is that it can be evaluated with the aid of fast summation methods, as discussed in Section 6.4.

5. Spectral methods

Spectral methods for solving the Boltzmann equation originated in the works of Pareschi and Perthame (1996) and Pareschi and Russo (2000*b*), and their properties were further studied in Pareschi and Russo (2000*c*) and in Filbet and Mouhot (2011). Related approaches, based on the use of the Fourier transform have been introduced by Bobylev and Rjasanow (1999), Ibragimov and Rjasanow (2002) and Gamba and Tharkabhushanam (2009). Historically the very first attempts of this type for the Boltzmann equation were presented in Grigoriev and Mikhalitsyn (1983) and Gabetta and Pareschi (1994). All these methods share the fact that they were inspired to the Fourier transform theory for the Boltzmann equation for Maxwell molecules by Bobylev (1988). Extensions to other kinetic problems have been considered in Pareschi et al. (2000) and Filbet and Pareschi (2003) for the Landau equation, in Filbet et al. (2005) for granular gases, in Filbet et al. (2012) for the quantum case, and in Pareschi, Toscani and Villani (2003) and Gamba and Haack (2014) for grazing limits. Here we use the word spectral method in a standard way commonly used in numerical analysis and scientific computing, namely to denote a class of method involving the use of the Fast Fourier Transform (FFT), which exhibit excellent error properties, with the so-called spectral accuracy (the error tends to zero faster than any fixed power of N, where N is the number of grid points), when the solution is smooth (Canuto, Hussaini, Quarteroni and Zang 1988). Beside the spectral accuracy, as we will discuss in Section 6, the fundamental property of spectral methods is the possibility to speed up their evaluation through fast summation algorithms (Mouhot and Pareschi 2006), making them competitive with direct simulation Monte Carlo methods in the case of non stationary flows (Filbet and Russo 2003, Filbet, Mouhot and Pareschi 2006, Gamba and Tharkabhushanam 2010, Filbet 2012, Wu, White, Scanlon, Reese and Zhang 2013).

5.1. Spectral methods for the Boltzmann equation

Following Pareschi and Russo (2000*b*) let us start with the change of variables $q = v_* - v$ in the collision operator (2.14) to get

$$\begin{aligned} Q(f,f)(v) &= \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos\theta) \\ &\quad [f(v+q^+)f(v+q^-) - f(v)f(v+q)]d\omega\, dq, \end{aligned} \tag{5.1}$$

where we used the fact that

$$B(v, v_*, \omega) = |q|\sigma(|q|, \cos\theta), \tag{5.2}$$

and the vectors q^+ and q^- that parameterize the post-collisional velocities are given by

$$q^+ = \frac{1}{2}(q + |q|\omega), \quad q^- = \frac{1}{2}(q - |q|\omega). \tag{5.3}$$

We recall the following identity for the weak form of the collision operator

$$\begin{aligned} &\int_{\mathbb{R}^3} Q(f,f)\varphi(v)\,dv \\ &\quad = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos\theta) f(v) f(v+q)[\varphi(v+q^+) - \varphi(v)]\,d\omega\,dq\,dv, \end{aligned} \tag{5.4}$$

for all test functions φ.

Periodization and choice of the integration domain

The first step in the construction of the method is the reduction of the integral over a finite integration domain. To this aim, we observe that if a distribution function f has compact support, 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

$$(v')^2 \le v^2 + v_*^2 \le 2R^2,$$

and similarly we have $(v'_*)^2 \le 2R^2$. Moreover $|q| \le 2R$ and thus the collision operator satisfies the following (Pareschi and Perthame 1996)

Proposition 5.1. Let supp $(f(v)) \subset \mathcal{B}_0(R)$ then

i) supp $(Q(f,f)(v)) \subset \mathcal{B}_0(\sqrt{2}R)$,
ii)

$$\begin{aligned} \int_{\mathbb{R}^3} Q(f,f)\varphi(v)\,dv \;=\; &\int_{\mathcal{B}_0(\sqrt{2}R)}\int_{\mathcal{B}_0(2R)}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos\theta) \\ &f(v) f(v+q)[\varphi(v+q^+) - \varphi(v)]\,d\omega\,dq\,dv, \end{aligned} \tag{5.5}$$

with $v + q^+, v + q \in \mathcal{B}_0((2+\sqrt{2})R)$.

In order to write a spectral approximation to (5.1) we can consider the distribution function $f(v)$ restricted on the cube $[-T,T]^3$ with $T \ge (2+\sqrt{2})R$, 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 lower bound for T can be improved using the periodicity of the function and allowing intersections of periods where the function f is zero, to get $T \ge (3+\sqrt{2})R/2$ (see Figure 5.1).

Remark 5.1. The choice $T = (3+\sqrt{2})R/2$ guarantees the absence of intersection between periods of the distribution function where f is different from zero and thus permits to develop spectral methods on the cube without aliasing errors (Canuto et al. 1988). However, since in practice the support of f increases with time, we can just minimize the errors due to aliasing.

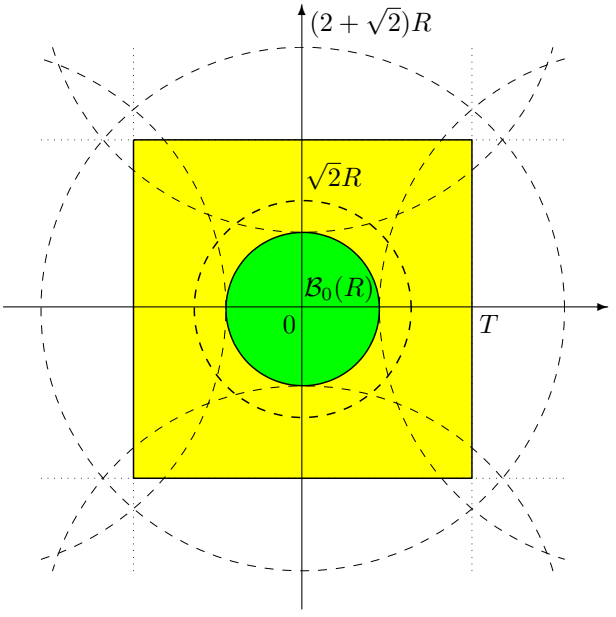

Figure 5.1. Restriction of the distribution function on the periodic box $[-T,T]\times[-T,T]$, with $T=(3+\sqrt{2})R/2$.

Spectral projection of the collision operator

To simplify the notation let us take $T=\pi$ and hence $R=\lambda\pi$ with $\lambda = 2/(3+\sqrt{2})$. We denote by $Q^R(f,f)$ the Boltzmann operator with cut-off on the relative velocity on $\mathcal{B}_0(2R)$. Hereafter, we use just one index to denote the three-dimensional sums with respect to the vector $k=(k_1,k_2,k_3)\in\mathbb{Z}^3$. The approximate function f_N is represented as the truncated Fourier series

$$f_N(v)=\sum_{k=-N}^{N}\hat{f}_k e^{ik\cdot v}, \tag{5.6}$$

$$\hat{f}_k=\frac{1}{(2\pi)^3}\int_{[-\pi,\pi]^3} f(v)e^{-ik\cdot v}\,dv. \tag{5.7}$$

We obtain a spectral quadrature based on the $\hat{f}_k$ coefficients by projecting (5.1) on the space of trigonometric polynomials of degree $\le N$ (Canuto et al. 1988, Gottlieb and Orszag 1977). Hence, we have

$$\hat{Q}_k=\int_{[-\pi,\pi]^3} Q^R(f_N,f_N)e^{-ik\cdot v}\,dv,\quad k=-N,\dots,N. \tag{5.8}$$

By substituting expression (5.6) in (5.8) and using the identity (5.4) for $\varphi = e^{-ik\cdot v}$ we get

$$\hat{Q}_k = \sum_{\substack{l,m=-N \\ l+m=k}}^{N} \hat{f}_l \, \hat{f}_m \hat{\beta}(l,m), \quad k = -N, \dots, N, \tag{5.9}$$

where the Boltzmann kernel modes $\hat{\beta}(l,m) = \hat{B}(l,m) - \hat{B}(m,m)$ are now given by

$$\hat{B}(l,m) = \int_{\mathcal{B}_0(2\lambda\pi)} \int_{\mathbb{S}^2} |q|\sigma(|q|, \cos\theta) e^{-i(l\cdot q^+ + m\cdot q^-)} \, d\omega \, dq. \tag{5.10}$$

It is remarkable that (5.10) is a scalar quantity completely independent on the function f_N and on the argument v, depending just on the particular kernel structure. In practice these quantities can be computed in advance and then stored in a multidimensional matrix. It is immediate to verify that the kernel modes satisfy

$$\hat{B}(l,m) = \hat{B}(-l,m) = \overline{\hat{B}(l,-m)} = \hat{B}(l,-m), \tag{5.11}$$

where the last equality states that the coefficients are real. Moreover they depend only of $|l-m|$, $|l+m|$ and of the angle between $\eta = (l+m)$ and $\mu = (l-m)$ (Pareschi and Perthame 1996). Obviously, these properties are useful to reduce the storage requirements of the method. We emphasize that the straightforward evaluation of (5.9) requires $O(n^2)$ operations, where $n = N^3$, and hence it is less expensive than a usual discrete-velocity based algorithm. We refer to Section 6 for a more detailed discussion on this topic and the construction of fast summation methods.

Finally we can rewrite scheme (5.9) in the form, $k = -N, \dots, N$

$$\hat{Q}_k = \sum_{m=-N}^{N} \hat{f}_{k-m} \, \hat{f}_m \hat{\beta}(k-m, m). \tag{5.12}$$

In the previous expression we assume that the Fourier coefficients are extended to zero for $|k_j| > N$, $j = 1, 2, 3$.

In the VHS case, $|q|\sigma(|q|, \cos\theta) = C_\alpha |q|^\alpha$, the dependence on the scattering angle disappears and it is easy to obtain the bound

$$|\hat{B}(l,m)| \le \frac{1}{3+\alpha}, \quad \alpha > -3, \tag{5.13}$$

where we have chosen $C_\alpha = ((4\pi)^2 (2\lambda\pi)^{3+\alpha})^{-1}$. Moreover (5.10) reduces to a one-dimensional integral (Pareschi and Russo 2000*b*)

$$\hat{B}(l,m) = \int_0^1 r^{2+\alpha} \, \mathrm{Sinc}(\xi r) \, \mathrm{Sinc}(\eta r) \, dr = F_\alpha(\xi, \eta), \tag{5.14}$$

where $\xi = |l+m|\lambda\pi$, $\eta = |l-m|\lambda\pi$. In addition for integer values of $\alpha < -3$ the previous integral can be computed explicitly. We report the expressions for the the case of Maxwell molecules $\alpha = 0$ and hard spheres $\alpha = 1$

$$F_0(\xi,\eta) = \frac{p\sin(q) - q\sin(p)}{2\xi\eta pq} \tag{5.15}$$

$$F_1(\xi,\eta) = \frac{q\sin(q)+\cos(q)}{2\xi\eta q^2} - \frac{p\sin(p)+\cos(p)}{2\xi\eta p^2} - \frac{2}{p^2q^2} \tag{5.16}$$

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

Remark 5.2.

- Formula (5.9) or equivalently (5.12) can be derived also from the weak form (5.4) in a bounded domain

$$\int_{[-\pi,\pi]^3} Q^R(f,f)\varphi(v)\,dv = \int_{[-\pi,\pi]^3}\int_{\mathcal{B}_0(2\lambda\pi)}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos\theta) f(v)f(v+q)[\varphi(v+q^+)-\varphi(v)]\,d\omega\,dq\,dv, \tag{5.17}$$

  taking $\varphi(v) = e^{-ikv}$.
- The spectral method can be applied directly to the Boltzmann equation by requiring that the residual is L^2-orthogonal to all Fourier modes. A natural application is given by the space homogeneous case where we require

$$\int_{[-\pi,\pi]^3}\left(\frac{\partial f_N}{\partial t} - Q^R(f_N,f_N)\right)e^{-ikv}\,dv = 0, \qquad k = -N,\dots,N, \tag{5.18}$$

  to obtain a set of ordinary differential equations satisfied by the Fourier coefficients (Pareschi and Russo 2000*b*)

$$\frac{\partial \hat{f}_k}{\partial t} = \hat{Q}_k, \quad k = -N,\dots,N. \tag{5.19}$$

Methods based on the Fourier transform

In the approaches developed in Bobylev and Rjasanow (1997), Bobylev and Rjasanow (2000), Bobylev and Rjasanow (1999), Ibragimov and Rjasanow (2002), Gamba and Tharkabhushanam (2009) and Gamba and Tharkabhushanam (2010) the concept is dual from the spectral method just described, since it consists in first using Fourier transform of the collision operator and then approximating the truncated Fourier transformed operator by a quadrature formula. The same ideas motivated the early attempts in Grigoriev and Mikhalitsyn (1983) and Gabetta and Pareschi (1994).

Here we don't review all the different strategies used to tackle the Fourier

discretization problem but we limit ourselves to present the general idea following Gamba and Tharkabhushanam (2009). Let us introduce the Fourier transform

$$\hat{f}(\xi) = \mathcal{F}_v[f](\xi) = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} f(v)e^{-i\xi v}\,dv, \quad \mathcal{F}_\xi^{-1}[\hat{f}](v) = \int_{\mathbb{R}^3}\hat{f}(\xi)e^{i\xi v}\,d\xi.$$

We can then consider the weak form (5.4) of the Boltzmann equation with the choice $\varphi(v) = e^{-i\xi v}/(2\pi)^3$ to get the Fourier transformed operator

$$\begin{aligned}\hat{Q}(\xi) &= \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} Q(f,f)e^{-i\xi v}\,dv \\ &= \frac{1}{(2\pi)^3}\int_{\mathbb{R}^6}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos(\theta))f(v)f(v+q)[e^{-i\xi(v+q^+)} - e^{-i\xi v}]\,d\omega\,dq\,dv.\end{aligned}$$

Now denoting by $f_q(v) = f(v+q)$ and using the convolution theorem $\mathcal{F}_v[fg](\xi) = \int \hat{f}(\xi-\mu)\hat{g}(\mu)\,d\mu$ we get

$$\begin{aligned}\hat{Q}(\xi) &= \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos(\theta))\mathcal{F}_v[ff_q](\xi)\left[e^{-i\xi q^+} - 1\right]\,d\omega\,dq \\ &= \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos(\theta))\int_{\mathbb{R}^3}\hat{f}(\xi-\mu)\hat{f}(\mu)e^{-i\mu q}\,d\mu\left[e^{-i\xi q^+} - 1\right]\,d\omega\,dq \\ &= \int_{\mathbb{R}^3}\hat{\beta}(\xi-\mu,\mu)\hat{f}(\xi-\mu)\hat{f}(\mu)\,d\mu \qquad (5.20)\end{aligned}$$

where we used the property $\hat{f}_q(\mu) = \hat{f}(\mu)e^{-i\mu q}$. In (5.20), setting $\hat{\beta}(\lambda,\mu) = \hat{B}(\lambda,\mu) - \hat{B}(\mu,\mu)$ with $\xi = \lambda + \mu$, we have

$$\hat{B}(\lambda,\mu) = \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |q|\sigma(|q|,\cos(\theta))e^{-i\lambda q^+ - i\mu q^-}\,d\omega\,dq. \qquad (5.21)$$

The above representation satisfies the same symmetries as in (5.11). Note that (5.20) and (5.21) are the Fourier transform representation of the collision term analogous of the Fourier series in (5.12) and (5.10). In particular the same computations as in Pareschi and Russo (2000*b*) for VHS are possible and we get the analogous of (5.14)

$$\hat{B}(\lambda,\mu) = \int_{\mathbb{R}} r^{2+\alpha}\,\mathrm{Sinc}\left(\frac{|\lambda+\mu|r}{2}\right)\mathrm{Sinc}\left(\frac{|\lambda-\mu|r}{2}\right)\,dr, \qquad (5.22)$$

where we have set $C_\alpha = (4\pi)^{-2}$.

The method is then applied by truncating the above formulation in a bounded domain and then approximating the Fourier transform using the FFT. Of course, since this is done for the transformed equation, it should be realized properly to avoid aliasing as described in Proposition 5.1. In particular, when approximating the Fourier transform by the FFT, for an appropriate choice of the computational parameters, one goes back to the

spectral method (5.12)-(5.10) or to a pseudo-spectral formulation of the same. A word of caution is in order when one first applies the Fourier transform over the whole space $\mathbb{R}^3$ and then truncates the computational domain in Fourier variables. To illustrate this, let us observe that, in the case of Maxwell molecules $\sigma(|q|, \cos(\theta)) = C_0|q|^{-1}$, the terms in (5.21) originates a Dirac δ distribution in (5.20). It is immediate to verify that $\hat{B}(\mu, \mu)$ concentrates at $\mu = 0$, whereas $\hat{B}(\lambda, \mu)$, after a change of variables, concentrates at $\mu = (\xi + |\xi|\omega)/2$. This yields the well-known simplification (Bobylev 1988)

$$\hat{Q}(\xi) = \int_{\mathbb{S}^2} (\hat{f}(\xi^+)\hat{f}(\xi^-) - \hat{f}(\xi)\hat{f}(0))\, d\omega, \tag{5.23}$$

with $\xi^{\pm} = (\xi \pm |\xi|\omega)/2$ and where we have set $C_0 = (2\pi)^{-3}$. A direct discretization of (5.23) originates in a straightforward way an $O(n \log_2 n)$ fast solver (the cost of the FFT) and this was the motivation behind the approximations in Grigoriev and Mikhalitsyn (1983), Gabetta and Pareschi (1994), Bobylev and Rjasanow (1997) and Bobylev and Rjasanow (2000). Note, however, that the above simplification is no more valid starting from the Fourier series representation of f in a bounded domain (Pareschi and Perthame 1996).

Finally, in these approaches conservation of momentum and kinetic energy are enforced, either by a renormalization procedure or by solving a constrained optimization problem, analogous to the one discussed in Section 3 for the definition of the numerical Maxwellian states. We refer also to Narayan and Klöckner (2009) for a recent review of the different approaches.

5.2. Properties of the spectral method

In this section we will analyze in detail the main properties of the spectral method (5.9)-(5.10), with a particular attention to the conservation properties, the accuracy and the stability of the 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. $f_N(t) \in \mathbb{P}^N$ where

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

Moreover, let $\mathcal{P}_N : L^2([-\pi, \pi]^3) \to \mathbb{P}^N$ be the orthogonal projection upon $\mathbb{P}^N$ in the inner product of $L^2([-\pi, \pi]^3)$ (see (5.8)):

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

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 given by (5.6). Since the operator $\mathcal{P}_N$ is self-adjoint the following property hold

$$< \mathcal{P}_N f, \varphi >=< f, \mathcal{P}_N \varphi >=< \mathcal{P}_N f, \mathcal{P}_N \varphi > \quad \forall\ f,\ \varphi \in L^2([-\pi,\pi]^3). \tag{5.24}$$

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

$$\mathcal{S}\, f_N(v) = f_N^s = \sum_{k=-N}^{N} s_k \hat{f}_k e^{ik\cdot v}, \tag{5.25}$$

where $s_k = s_{k_1} s_{k_2} s_{k_3}$ are required to be real non-negative numbers such that $s_{k_j} = s_{-k_j},\ s_0 = 1$ and $s_{|k_j|} \le s_{|k_j-1|}, j = 1, 2, 3.$

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 5.2. Let $f \in L^2([-\pi,\pi]^3)$ and let us define

$$\begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix} := \int_{[-\pi,\pi]^3} f \begin{pmatrix} 1 \\ v \\ |v|^2 \end{pmatrix} dv. \tag{5.26}$$

Then we have:

i) If $f \ge 0$, $\forall\ v$ and the factors s_k are such that

$$1 + 2\sum_{k=1}^{N} s_k \cos(k \cdot v) \ge 0, \tag{5.27}$$

then $f_N^s(v) \ge 0$, $\forall\ v$.

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

$$\rho = \rho_N, \quad |\rho u - \rho u_N| \leq \frac{C_1}{N^{1/2}} ||f||_2, \quad |\rho e - \rho e_N| \leq \frac{C_2}{N^{3/2}} ||f||_2. \tag{5.28}$$

Proof. The smoothed series (5.25) can be represented in integral form

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

where the kernel K_N is given by

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

From the positivity of K_N follows the positivity of f_N^s and hence 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 (5.24). 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

$$(\hat{v_j})_k = -\frac{i}{(2\pi)} \prod_{\substack{l=1 \\ l\neq j}}^{3} \delta_{k_l 0} \int_{[-\pi,\pi]} v_j \sin(k_j v_j)\, dv_j = -i \prod_{\substack{l=1 \\ l\neq j}}^{3} \delta_{k_l 0} \frac{(-1)^{k_j}}{k_j}, \quad j = 1,2,3, \tag{5.29}$$

and

$$(\hat{v^2})_k = \frac{1}{(2\pi)} \sum_{j=1}^{3} \prod_{\substack{l=1 \\ l\neq j}}^{3} \delta_{k_l 0} \int_{[-\pi,\pi]} (v_j)^2 \cos(k_j v_j)\, dv_j = 2 \sum_{j=1}^{3} \prod_{\substack{l=1 \\ l\neq j}}^{3} \delta_{k_l 0} \frac{(-1)^{k_j}}{k_j^2}. \tag{5.30}$$

Using Parseval's identity and (5.29)-(5.30) we obtain the following estimates

$$||v - v_N||_2 \leq \frac{C_1}{N^{1/2}}, \quad ||v^2 - (v^2)_N||_2 \leq \frac{C_2}{N^{3/2}},$$

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

Remark 5.3.

- The positivity requirement i) can be satisfied using the factors

$$s_{k_j} = \left(1 - \frac{|k_j|}{N}\right), \quad j = 1,2,3. \tag{5.31}$$

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

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

  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 $\mathcal{S}\mathcal{P}_N = \mathcal{P}_N^s$ is also self-adjoint by using the properties of s_k, the analogue of points *ii*) and *iii*) can be proved for f_N^s.
- From *ii*) and (5.29) 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 (5.30) 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 $r \geq 0$ 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, for each $\varphi \in L^2([-\pi,\pi]^3)$ we have

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

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

Consistency and spectral accuracy

Now we can discuss the consistency properties of the spectral quadrature method defined by (5.9). For simplicity we will restrict our discussion to the VHS model introduced previously. As we will see the consistency proof follows in a direct way from the construction of the method. We will denote by $Q^R_N(f_N) = \mathcal{P}_N Q^R(f_N, f_N)$. 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 $Q^R(f_N, g_N) \in \mathbb{P}_{2N}$ and

$$||Q^R(f,g)||_2 \le C||f||_2||g||_2. \tag{5.32}$$

Proof. The first statement follows immediately from the representation formula

$$Q^R(f_N, g_N) = \sum_{l=-N}^{N} \sum_{m=-N}^{N} \hat{f}_l \, \hat{g}_m \hat{\beta}(l,m) e^{i(l+m)\cdot v} = \sum_{k=-2N}^{2N} \hat{Q}^N_k e^{ik\cdot v},$$

where

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

The estimate (5.32) for the gain part of the collision operator is a consequence of the estimates in (Gustaffson 1986, Lions 1994)

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

The corresponding result for the loss part can be computed directly observing that $L^R(f) = f *_q B^\lambda$ 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 $\mathcal{B}_0(2R)$, $R = \lambda\pi$. Hence

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

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

□

Then we observe that the method defined by equation (5.9) implies the following spectral approximation of the collision integral

$$Q_N^R(f_N) = \sum_{k=-N}^{N} \hat{Q}_k e^{ikv}, \tag{5.33}$$

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

From Proposition 5.2 it is also clear that the projected collision operator $Q_N^R(f_N)$ will preserve the mass in time. This can be also derived directly from the properties of the kernel modes, in fact

$$\int_{[-\pi,\pi]^3} Q_N(f_N)\, dv = \hat{Q}_0 = \sum_{m=-N}^{N} \hat{f}_{-m}\hat{f}_m\hat{\beta}(-m,m) = 0,$$

since $\hat{\beta}(-m,m) = 0$.

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

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

$$||Q^R(f) - Q_N^R(f_N)||_2 \le C\left(||f - f_N||_2 + \frac{||Q^R(f_N)||_{H_p^r}}{N^r}\right), \quad \forall\, r \ge 0, \tag{5.34}$$

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

Proof. First, we can split the error in two parts

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

Now from Lemma 5.1, $Q^R(f_N) \in \mathbb{P}_{2N}$ and hence $Q^R(f_N)$ is periodic and infinitely smooth together with all its derivatives thus

$$||Q^R(f_N) - Q_N^R(f_N)||_2 \le \frac{C}{N^r}||Q^R(f_N)||_{H_p^r}, \quad \forall\, r \ge 0. \tag{5.35}$$

By application of Lemma 5.1 and from the identity

$$Q^R(f) - Q^R(g) = Q^R(f+g, f-g),$$

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

$$||Q^R(f) - Q^R(f_N)||_2 = ||Q^R(f+f_N, f-f_N)||_2 \le C_1||f+f_N||_2||f-f_N||_2$$
$$\le 2C_1||f||_2||f-f_N||_2.$$

This concludes the proof. □

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

Theorem 5.2. Let $f \in H_p^r([-\pi,\pi]^3)$, $r \geq 0$ then

$$||Q^R(f) - Q_N^R(f_N)||_2 \leq \frac{C}{N^r}\left(||f||_{H_p^r} + ||Q^R(f_N)||_{H_p^r}\right), \tag{5.36}$$

Proof. It is enough to observe that

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

□

From the previous corollary it follows

$$| < Q^R(f), \varphi > - < Q_N^R(f_N), \varphi > | \leq \frac{C}{N^r}||\varphi||_2\left(||f||_{H_p^r} + ||Q^R(f_N)||_{H_p^r}\right), \tag{5.37}$$

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

Remark 5.4.

- From (5.35) it follows that, except for the projection errors on the initial data, the variations of momentum and energy introduced by the spectral 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 (5.35), from Schwarz inequality we have

  $$| < Q^R(f_N), \varphi > - < Q_N^R(f_N), \varphi > | \leq \frac{C}{N^r}||\varphi||_2||Q^R(f_N)||_{H_p^r}. \tag{5.38}$$

  The estimates on the conservation laws can be derived by considering $\varphi = v, v^2$.
- Of course, the method can be made exactly conservative by enforcing conservations through the L_2 projection (3.44) applied to the approximated collision term $Q_N^R(f_N)$ over the grid. Since the correction is proportional to the moments deviation, from the above estimates it follows that spectral accuracy may be maintained provided that the moments are approximated with spectral accuracy and aliasing errors are controlled by the choice of a sufficiently large computational domain.

Numerical validation of spectral accuracy

Finally we report some numerical results which show that the method defined by (5.9)-(5.10) is capable to achieve spectral accuracy of the numerical

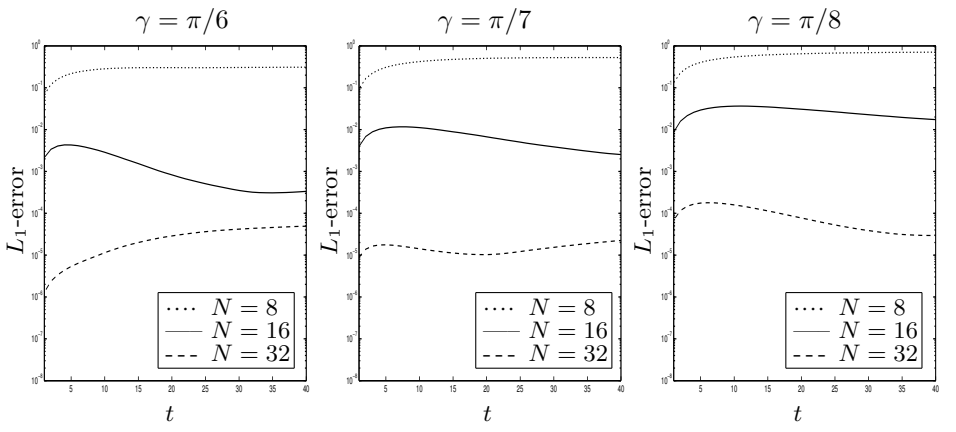

Figure 5.2. L_1 relative norm of the error vs time.

Table 5.1. Convergence test for the homogenous Boltzmann equation.

# modes	Error at time $t = 5$			Convergence rate		
n	$\pi/6$	$\pi/7$	$\pi/8$	$\pi/6$	$\pi/7$	$\pi/8$
8×8	1.51E-01	2.09E-01	2.71E-01	5.64	4.41	3.91
16×16	3.01E-03	9.78E-03	1.78E-03	9.64	10.23	8.19
32×32	3.79E-06	0.81E-05	0.61E-05			

solution. We consider the two-dimensional Maxwellian molecules case (i.e. $\alpha = 0$, $v \in \mathbb{R}^2$), with $C_0 = 1/(2\pi)$. In the space homogeneous case, this problem has an exact solution given by (Bobylev 1975, Ernst 1983)

$$f(v,t) = \frac{1}{4\pi S^2\gamma^2}\left(2S - 1 + \frac{1-S}{2S}\frac{v^2}{\gamma^2}\right)\exp\left(-\frac{v^2}{2S\gamma^2}\right), \quad t \geq 0, \tag{5.39}$$

where $S = 1 - \exp(-\gamma^2 t/8)/2$. In the computation the scaling parameter $\gamma = \pi/6$ is chosen in such a way that the numerical support of the initial condition is well approximated by $\mathcal{B}_0(R)$ and and the integration time: $t_{\max} = 40$. Figures 5.2 show the relative L_1 norm of the error in the density function for $N = 8, 16$, and 32 modes per direction. Note that the relative 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. If

we fix the attention to Figure 5.2 at time $t = 5$ we observe the results for the relative L^1 norms of the error reported in table 5.1 for different choices of the scaling parameter $\gamma = \pi/6, \pi/7, \pi/8$. In the last three columns the order of accuracy is reported. The decay of the error with the increase of the number of modes is an indication of spectral accuracy.

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 Fig. (5.2) (right). 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.

Stability results for the homogeneous Boltzmann equation

Next we consider the problem of stability of the numerical solution in the space homogeneous setting for VHS kernels with cut-off over the relative velocity. We consider the initial value problem

$$\begin{aligned} &\frac{\partial f}{\partial t} = Q^R(f,f), \qquad v \in [-\pi,\pi]^3,\ t > 0, \\ &f(v, t=0) = f_0(v), \end{aligned} \tag{5.40}$$

First, following Pareschi and Russo (2000*c*), we rewrite the equation in the equivalent form

$$\frac{\partial f}{\partial t} + \mu f = P^R(f,f), \tag{5.41}$$

with $P^R(f,f) = Q^R(f,f) + \mu f$.

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

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

Thus for

$$\mu \ge C_\alpha 4\pi(2\lambda\pi)^\alpha \rho, \tag{5.42}$$

$P^R(f,f)$ is a positive monotone operator in the sense that

$$P^R(f,f) \ge P^R(g,g) \ge 0 \quad \text{if} \quad f \ge g \ge 0.$$

Then we have the following result regarding the stability of a smoothed spectral scheme (Pareschi and Russo 2000*c*)

Theorem 5.3. There exists a unique solution $f_N(t,v) \in C^1([0,T], \mathbb{P}^N)$, $f_N \ge 0$, with $||f_N||_1 = \rho$, for arbitrary time $T > 0$ to the initial value

problem

$$\begin{aligned} &\frac{\partial f_N}{\partial t} + \mu f_N = \mathcal{S} P_N^R(f_N), \\ &f_N(v, t=0) = f_{0,N}^s(v), \end{aligned} \tag{5.43}$$

provided that $f_0 \in L^2([-\pi,\pi]^3)$ is non negative, $||f_0||_1 = \rho$, μ satisfies (5.42) and the smoothing operator $\mathcal{S}$ satisfies (5.27).

From Theorem 5.3 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 (5.43) introduces too much smoothing, and spectral accuracy is lost. However, it is interesting to remark that, as pointed out in Pareschi and Russo (2000*c*), the main reason in the lack of accuracy of (5.43) is represented by the smoothed projection of the initial data.

A more general stability result in absence of smoothing has been proved recently by Filbet and Mouhot (2011). Their approach is based on considering homogeneous Boltzmann equations perturbed by smoothed balanced operators which do not preserve positivity of the distribution. They proved the following result in dimension $d \geq 2$.

Theorem 5.4. Consider any nonnegative initial datum $f_0 \in H_p^r([-\pi,\pi]^3)$, with $r > d/2$, which is not zero everywhere. Then there exists $N_0 \in N$ (depending on the mass and $\|f\|_{H_p^r}$) such that for all $N \geq N_0$:

(i) there is a unique global solution $f_N = f_N(\cdot, t)$ to the following problem

$$\begin{aligned} &\frac{\partial f_N}{\partial t} = Q_N^R(f_N), \\ &f_N(v, t=0) = f_{0,N}^s(v); \end{aligned} \tag{5.44}$$

(ii) for any $k < r$, there exists $C > 0$ such that

$$\forall\, t \geq 0, \quad \|f_N(\cdot, t)\|_{H_p^k} \leq C; \tag{5.45}$$

(iii) this solution is everywhere positive for time large enough, and the mass of its negative values can be made uniformly (in times) L^∞ small as $N \to \infty$;

(iv) this solution f_N converges to $f(t)$, the periodized solution in $[-\pi,\pi]^3$ to (5.40), with spectral accuracy, uniformly in time;

(v) this solution converges exponentially in time to a constant solution prescribed by the mass conservation law.

Remark 5.6. As a consequence of the results just described the steady

states of the spectral method (5.44) are the same of the periodized homogeneous Boltzmann equation and are characterized by constant functions with given mass density. Of course, if the velocity support is large enough this behavior is never observed in practical calculations. Nevertheless one may be interested in developing a spectral method that preserves exactly the Maxwellian equilibrium state (which is not guaranteed by enforcing moments conservations). A possible approach is based on the decomposition (Liu and Yu 2004)

$$f = M[f] + g, \tag{5.46}$$

with g such that $\int_{\mathbb{R}^3} g\,\varphi\,dv = 0$, $\varphi = 1, v, |v|^2$, which inserted into the collision operator gives

$$Q(f,f) = Q(g, M[f]) + Q(M[f], g) + Q(g,g), \tag{5.47}$$

where we used the fact that $Q(M[f], M[f]) = 0$.

Now instead of periodizing f, one periodizes g and applies the spectral method to the decomposition (5.47). The major difference is that the steady state of (5.47) is given by $g = 0$, $\forall\ v$ which belongs to the space of trigonometric polynomials and therefore it can be described correctly by the spectral scheme.

5.3. Spectral methods for other collision terms

The general methodology described in Section 5.1 has been successfully applied also to other collision terms, like the Landau equation (Pareschi et al. 2000, Filbet and Pareschi 2003), the inelastic Boltzmann equation (Filbet et al. 2005) and the quantum Boltzmann equation (Filbet et al. 2012). In the sequel we give a short description of the first two cases and refer to Section 6 for some details on the quantum case.

Landau equation

One case of particular relevance is that of the Landau equation of plasma physics (Pareschi et al. 2000). We rewrite here the expression of the Landau integral after the change of variables $q = v - v_*$

$$Q_L(f,f)(v) = \nabla_v \cdot \int_{\mathbb{R}^3} A(q)[f(v-q)\nabla_v f(v) - f(v)\nabla_q f(v-q)]\,dq \tag{5.48}$$

where $A(q) = \Psi(|q|)\Pi(q)$ is a 3×3 nonnegative symmetric matrix and $\Pi(q) = I - qq/|q|^2$, with I the identity matrix, is the orthogonal projection upon the space orthogonal to q. We have $\Psi(|q|) = \Lambda|q|^{\alpha+2}$ for inverse-power laws, with $\alpha \geq -3$ and $\Lambda > 0$.

In the Landau case, similarly to the previous situation, it can be proved that if supp $(f(v)) \subset \mathcal{B}_0(R)$ then supp $(Q_B(f,f)(v)) \subset \mathcal{B}_0(3R)$. By denoting $R = \lambda\pi$, aliasing errors for compactly supported functions can be

avoided with $\lambda = 1/2$ and the integration over $\mathbb{R}^3$ in (5.48) can be replaced by an integration over $\mathcal{B}_0(\pi)$ (Pareschi et al. 2000).

By substituting expression (5.6) in (5.8) with $Q^R = Q_L^R$, where Q_L^R is the Landau operator with cut-off over the relative velocity in the ball $\mathcal{B}_0(2R)$, using the orthogonality property of trigonometric polynomials, we get

$$\hat{Q}_{L,k} = \sum_{\substack{l+m=k \\ l,m=-N}}^{N} \hat{f}_l \, \hat{f}_m \hat{\psi}(l,m), \quad k = -N, \dots, N \tag{5.49}$$

where $\hat{\psi}(l,m) = \hat{\Psi}(l,m) - \hat{\Psi}(m,m)$, and the Landau kernel modes $\hat{\Psi}(l,m)$ are given by

$$\hat{\Psi}(l,m) = \int_{\mathcal{B}_0(\pi)} \Psi(|g|) \left[l^2 - \left(l \cdot \frac{g}{|g|} \right)^2 \right] e^{ig \cdot m} dg. \tag{5.50}$$

As in the Boltzmann case, $\hat{\Psi}(l,m)$ are scalar quantities independent on the function f that satisfy the same symmetry property, so that they are functions of $|l+m|$ and $|l-m|$ only. Moreover, for inverse power laws, taking $\Lambda = ((4\pi)(\pi)^{5+\alpha})^{-1}$ we have the bound

$$|\hat{\Psi}(l,m)| \leq \frac{3N^2}{5+\alpha}. \tag{5.51}$$

Note that the estimates on the Fourier coefficients $\hat{B}(l,m)$ and $\hat{\Psi}(l,m)$ are quite different. In particular, $\hat{\Psi}(l,m)$ grow with N, and this is the cause of the stiffness observed in the time integration of the equation (Filbet and Pareschi 2003, Lemou and Mieussens 2005). This reflects the fact that the Landau equation suffers of the stiffness typical of diffusion equations. Stability condition of grid based methods requires that the time step scales with the square of the velocity mesh size. Although this is a very important issue, it is beyond our scope here. Some considerations on possible methods to overcome the problem are found in Remark 7.5 in Section 7.

By similar arguments as in the Boltzmann case it is possible to prove consistency and spectral accuracy of the method (Pareschi et al. 2000).

A fast summation method

Let us remark that schemes (5.9) and (5.49) have exactly the same structure. The only difference is given by the expressions of the Boltzmann kernel modes (5.10) and the Landau kernel modes (5.50). As a consequence the straightforward computation of (5.49) has the same $O(n^2)$, $n = N^3$, cost of a conventional discretization applied to the Landau equation.

On the other hand we can rewrite (5.49) as

$$\sum_{m=-N}^{N} \hat{f}_{k-m}\,\hat{f}_m \hat{\Psi}(k-m,m) - \sum_{m=-N}^{N} \hat{f}_{k-m}\,\hat{f}_m \hat{\Psi}(m,m), \quad k=-N,\dots,N.$$

Clearly the second sum is a convolution sum and thus transform methods allow this term to be evaluated in $O(n\log_2 n)$ operations. For the details of the implementation of this standard technique for spectral methods we refer the reader to (Canuto et al. 1988). Hence the most expensive part of the computation is represented by the first sum which in general cannot be evaluated with fast algorithms.

In the case of the Landau equation, however, $\hat{\Psi}(l,m)$ splits as

$$\hat{\Psi}(l,m) := l^2\tilde{F}(m) - \sum_{p,q=1}^{3} l_p\, l_q I_{pq}(m) = l^2\tilde{F}(m) - l\,\mathcal{I}(m)\,l^T,$$

where l^T denotes the transpose of the vector l, $\mathcal{I} = (I_{pq})$ is a 3×3 symmetric matrix, and taking $\Lambda = 1$

$$\tilde{F}(m) = \int_{\mathcal{B}_0(\pi)} |g|^{2+\gamma} e^{ig\cdot m} dg, \tag{5.52}$$

$$I_{pq}(m) = \int_{\mathcal{B}_0(\pi)} |g|^{\gamma} g_p\, g_q e^{ig\cdot m} dg, \quad p,q=1,\dots,3. \tag{5.53}$$

Thus we can write

$$\hat{\psi}(l,m) = l^2\tilde{F}(m) - l\,\mathcal{I}(m)\,l^T - \hat{\Psi}_L(m,m). \tag{5.54}$$

The resulting scheme requires the evaluation of 8 convolution sums (the number of distinct elements of $\mathcal{I}$ plus two single convolution sums for $\tilde{F}(m)$ and $\hat{\Psi}_L(m,m)$). Hence, the overall cost of the scheme is only $O(n\log_2 n)$.

For the implementation of the algorithm we need to evaluate the quantities (5.52)-(5.53). For simplicity, we will treat here only the two-dimensional case $v\in\mathbb{R}^2$. We have

$$\begin{aligned} I_{11}(m) &= \frac{1}{2}\left[F(|m|) + \frac{m_1^2-m_2^2}{|m|^2}G(|m|)\right], \\ I_{22}(m) &= \frac{1}{2}\left[F(|m|) - \frac{m_1^2-m_2^2}{|m|^2}G(|m|)\right], \\ I_{12}(m) &= I_{21}(m) = \frac{m_1 m_2}{|m|^2}G(|m|), \end{aligned}$$

where

$$\tilde{F}(m) = F(|m|) = 2\pi\int_0^{\pi} r^{\gamma+3} J_0(|m|r)\,dr, \tag{5.55}$$

with J_0 the Bessel function of order 0 and

$$G(|m|) = \int_0^{\pi} r^{\gamma+3} \int_0^{2\pi} \cos(|m| r \cos\varphi) \cos(2\varphi)\, d\varphi\, dr. \tag{5.56}$$

Thus the computation reduces simply to the computation of two one-dimensional integrals $F(|m|)$ and $G(|m|)$. These quantities can be computed very accurately once and then stored in two bidimensional arrays. A similar reduction can be performed in the full three dimensional case.

Remark 5.7. Let us recall that the Landau equation is obtained in the so-called grazing collision limit of the Boltzmann operator for Coulombian collisions. In such case, corresponding to $\alpha = -3$, the Boltzmann integral diverges and the Landau equation can be derived in the grazing collision limit (Villani 2002). It is remarkable that the spectral method, thanks to its weak formulation, can be successfully applied also to the study of the Boltzmann equation in this challenging asymptotic behavior where $\hat{B}(l,m) \to \hat{B}_L(l,m)$ (Pareschi et al. 2003). We refer also to Gamba and Haack (2014) for recent results based on the above ideas.

In particular, during this asymptotic process it is possible to obtain intermediate kinetic approximations that can be evaluated with fast algorithms at a cost of $O(n \log_2 n)$. This idea has been used in Pareschi (2003) to construct fast approximated algorithms for the Boltzmann equation (but with loss of spectral accuracy).

Inelastic Boltzmann equation

The case of the Boltzmann equation for granular gases has been studied in Filbet et al. (2005). In such case the method is applied to the Boltzmann integral for inelastic hard spheres which for any smooth test function ψ can be written as

$$\begin{aligned} &\int Q_e(f,f)(v)\psi(v)dv \\ &\quad = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |q| f(v)\, f(v-g)\left(\psi(v') - \psi(v)\right) d\omega\, dq dv, \end{aligned} \tag{5.57}$$

where

$$v' = v - \frac{1+e}{4}(q - |q|\omega) = v - \frac{1+e}{2} q^-,$$

and the inelastic collisions are characterized by a restitution coefficient e ($0 < e < 1$). The derivation of the spectral method is pretty straightforward and originates a scheme which has the same structure as (5.9) and (5.49). In this case it reads

$$\hat{Q}_{e,k} = \sum_{\substack{l,m=-N \\ l+m=k}}^{N} \hat{f}_l\, \hat{f}_m \hat{\beta}_e(l,m), \quad k = -N, \dots, N, \tag{5.58}$$

where the inelastic Boltzmann kernel modes $\hat{\beta}_e(l,m) = \hat{B}_e(l,m) - \hat{B}_e(m,m)$ are now given by

$$\hat{B}_e(l,m) = \int_{\mathcal{B}_0(2\lambda\pi)} \int_{\mathbb{S}^2} |q| e^{-im\cdot q + i(l+m)\cdot(1+e)q^-/2}\, d\omega\, dq. \tag{5.59}$$

As for the case of elastic hard spheres, it can be shown that the kernel modes can be reduced to one-dimensional integrals and computed explicitly. In this case the coefficient can be written as

$$\hat{B}_e(l,m) = C_\lambda F(\xi,\eta) = C_\lambda \int_0^1 r^3 \,\mathrm{Sinc}(\xi r)\,\mathrm{Sinc}(\eta r)\, dr, \tag{5.60}$$

where now $\xi = |l+m|(1+e)\lambda\pi/2$, $\eta = |l(1+e) - m(3-e)|\lambda\pi/2$ and $C_\lambda = (8\pi^2(2\lambda\pi)^4)$. It is easy to prove that for constant coefficient of restitution (5.60) has the same explicit analytical expression given by (5.16) except for the different definitions of ξ and η.

Remark 5.8. For non constant coefficient of restitution the spectral method has been analyzed in Naldi, Pareschi and Toscani (2003) for a one-dimensional inelastic model. In particular, it was shown that the method is well-defined in the numerical passage of the Boltzmann equation with singular kernel to nonlinear friction equations in the so-called quasi elastic limit.

A velocity-rescaling method

A major difficulty with the inelastic Boltzmann equation, is that the solution formally converges to a Dirac delta equilibrium state. A velocity rescaling technique in this case is necessary to approximate accurately the asymptotic behavior of the equation with the spectral method and avoid Gibbs phenomenon (see Figure 5.3). Here we recall the basic ideas of the method. We refer to Filbet et al. (2005), Filbet and Russo (2006) and Filbet and Rey (2013) for more details on the construction of schemes based on this strategy.

Given the distribution function $f(t,x,v)$, $x \in \mathbb{R}^{d_x}$, $v \in \mathbb{R}^{d_v}$ we introduce a new distribution $g(t,x,\xi)$ by setting

$$f(t,x,v) \;=\; \frac{1}{\omega^{d_v}}\, g\,(t,x,\xi)\,, \quad \xi = \frac{v}{\omega}, \tag{5.61}$$

where the function ω is assumed to be an accurate measure of the "support" or scale of the distribution f in velocity variables. Then according to this scaling, the distribution g should naturally "follow" either the concentration or the spreading in velocity of the distribution f.

Let us now derive the kinetic equation verified by the distribution g. Differentiating relation (5.61) with respect to time yields

$$\frac{\partial f}{\partial t} = \frac{1}{\omega^{d_v}} \left[\frac{\partial g}{\partial t} - \frac{1}{\omega} \frac{\partial \omega}{\partial t} \nabla_\xi \cdot (\xi\, g) \right].$$

One also has

$$v \cdot \nabla_x f = \frac{\xi}{\omega^{d_v-1}} \cdot \left[\nabla_x g - \frac{1}{\omega} \nabla_x \omega \nabla_\xi \cdot (\xi\, g) \right].$$

Then, if f is solution to the Boltzmann equation (2.10), with the inelastic collision term given by (5.57), the distribution g given by (5.61) is solution to the following equation

$$\frac{\partial g}{\partial t} + \omega\, \xi \cdot \nabla_x g - \frac{1}{\omega} \left[\left(\frac{\partial \omega}{\partial t} + \omega\, \xi \cdot \nabla_x \omega \right) \nabla_\xi \cdot (\xi\, g) \right] = \omega\, \tilde{Q}_e(g, g),$$

where $\tilde{Q}$ is such that $\tilde{Q}_e(g, g) = \omega^{d_v-1} Q_e(f, f)$. This equation can actually be written in the more convenient conservative form

$$\frac{\partial g}{\partial t} + \nabla_x \cdot (\omega\, \xi\, g) - \nabla_\xi \cdot \left[\left(\frac{1}{\omega} \frac{\partial \omega}{\partial t}\, \xi + \xi \otimes \xi\, \nabla_x \omega \right) g \right] = \omega\, \tilde{Q}_e(g, g). \quad (5.62)$$

In order to make this rescaling efficient, the main difficulty is now to choose an appropriate scaling function ω to define completely the distribution g. The natural idea which follows from Filbet et al. (2005) and Filbet and Russo (2006) consists in computing the function ω directly from the distribution function f, by setting

$$\omega \;:=\; \sqrt{\frac{2\,E}{d_v\,\rho}}, \quad (5.63)$$

where $E = (d_v T + u^2)\rho/2$ and the macroscopic quantities ρ, u and T have been defined in (2.32).

Assuming that f is nonnegative, the quantity ω will then provide correct information on its support: if f is concentrated, ω will be small, whereas it will be large for scattered distributions. This approach has been shown to be very accurate for the space-homogeneous setting (Filbet et al. 2005, Filbet and Russo 2006), but in the space non homogeneous case it poses some limitations. First, the definition of ω yields very restrictive constraints on the moments of g, namely

$$\int_{\mathbb{R}^{d_v}} g(t, x, \xi)\, d\xi \;=\; \frac{1}{d_v} \int_{\mathbb{R}^{d_v}} g(t, x, \xi)\, |\xi|^2\, d\xi.$$

Second, the strong coupling between ω and g leads to nonlinear terms in the evolution equation. A simple way to overcome these aspects has been proposed in Filbet and Rey (2013) and is based on evaluating ω from macroscopic quantities which are assumed to be close enough to the one computed from f. A good candidate will be a solution to the system of macroscopic equations obtained from a suitable closure of the kinetic model (see for example Pareschi and Toscani (2004)). Once a good macroscopic moments

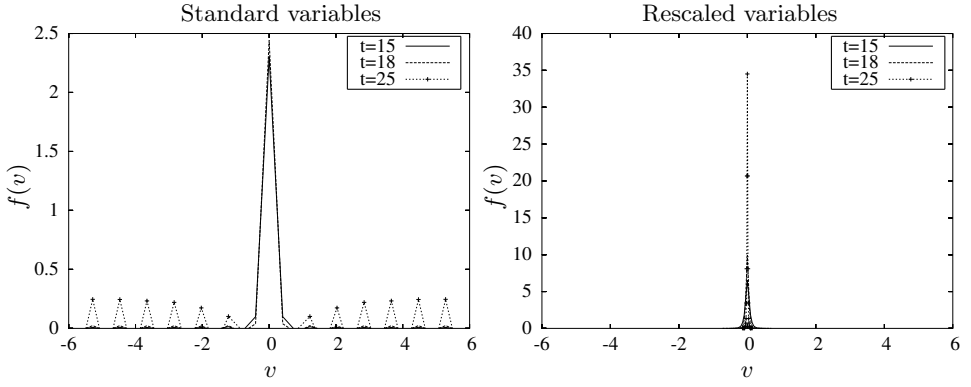

Figure 5.3. Large time behavior of the spectral method for the inelastic homogeneous Boltzmann equation in one-dimension.

estimator as been selected we then set

$$\omega := \sqrt{\frac{2\,E_m}{d_v\,\rho_m}}, \tag{5.64}$$

where E_m and ρ_m have been computed with the macroscopic model.

Remark 5.9. The velocity-rescaling method described above is not restricted to spectral methods and to inelastic collision operators, but applies also to any other kind of discretizations of a kinetic equation with the aim to avoid the drawbacks caused by a fixed grid in velocity space. See Filbet and Rey (2013) for several examples.

6. Fast summation methods

In this section we shall focus on the main question which plague the use of deterministic approximation of the Boltzmann equation, that is their computational complexity. As already observed in the previous sections, quadrature approximations of the five fold Boltzmann integral lead to a computational cost which is at least quadratic $O(n^{2+\delta})$, $\delta \geq 0$ with respect to the total number of velocity grid points n used to approximate the distribution function. In contrast stochastic methods, although less accurate, can be evaluated at a cost which is linear with respect to the number of particles (Bird 1994, Nanbu 1980). In the case of spectral methods a reduction from $O(n^2)$ to $O(n \log_2 n)$ was readily deducible for the Landau equation (see Pareschi et al. (2000) and Section 5.3). The convolution structure of the Landau equation, in fact, allows naturally the development of fast solvers (Buet and Cordier 1999, Lemou and Mieussens 2005). In the Boltzmann case such a straightforward reduction is not possible and only recently, in Mouhot and Pareschi (2004), Mouhot and Pareschi (2006) for

spectral methods and Mouhot et al. (2013) for discrete velocity methods, such result has been achieved for a particular class of interaction kernels. We refer also to Bobylev and Rjasanow (1999) for a similar approach. Let us mention that several acceleration techniques have been proposed in the past literature (Bokanowski and Lemou 2005, Buet 1996, Kowalczyk, Palczewski, Russo and Walenta 2008, Platkowski and Walús 2000, Valougeorgis and Naris 2003). We do not seek to review all of them here, and refer the reader to the above articles and the references therein.

6.1. The Boltzmann operator in bounded domains

As observed 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. In general the collision process "spreads" the support and at the numerical level some non physical condition has to be imposed to keep the support of the function in velocity uniformly bounded. In order to do this there are two main strategies.

1 One can remove the physical binary collisions that will lead outside the bounded velocity domain. If this is done properly, the scheme remains conservative (without spurious invariants). However this truncation breaks down the convolution-like structure of the collision operator, which requires the invariance in velocity. This truncation is at the basis of most discrete velocity methods in a bounded domain.
2 One can add some non physical binary collisions by periodizing the function and the collision operator. This implies the loss of some local invariants (some non physical collisions are added). Thus the scheme is not conservative anymore, except for the mass. In this way the structural properties of the collision operator are maintained and thus they can be exploited to derive fast algorithms. This approach is at the basis of spectral methods.

Of course, a third possibility, which can be used in conjunction with the periodization in point 2, is based on projecting the collision operator back to a compact support in such a way that conservations are enforced. This is the case, for example, of the L_2 projection (3.44) discussed in Section 3. In all cases, however, by enlarging enough the computational domain the number of removed or added collisions (or the projection) can be made negligible as well as the error in the local invariants.

Here, in order to derive fast summation methods, we shall focus on the second approach. The starting point in the development of fast summation methods is to approximate the collision operator starting from representation (2.25) which somehow conserves more symmetries of the collision operator when one truncates it in a bounded domain. This representation was used before by several authors to construct quadrature formu-

las (Bobylev and Rjasanow 1999, Ibragimov and Rjasanow 2002, Panferov and Heintz 2002). We consider the collision operator in the form (2.25) that we rewrite in dimension $d \geq 2$ as

$$Q(f,f)(v) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{B}(x,y)\delta(x\cdot y) \\ [f(v+y)\, f(v+x) - f(v+x+y)\, f(v)]\, dx\, dy, \quad (6.1)$$

with

$$\tilde{B}(x,y) = 2^{d-1}\, \sigma\left(|x+y|, -\frac{x\cdot(x+y)}{|x||x+y|}\right) |x+y|^{-(d-3)}.$$

One can easily see that on the manifold defined by $x \cdot y = 0$, we have $\tilde{B}(x,y) = \tilde{B}(|x|,|y|)$ (using the parities of the collision kernel) with

$$\tilde{B}(|x|,|y|) = 2^{d-1}\, \sigma\left(\sqrt{|x|^2+|y|^2}, \frac{|x|}{\sqrt{|x|^2+|y|^2}}\right) (|x|^2+|y|^2)^{-\frac{d-3}{2}}. \quad (6.2)$$

Now let us consider the function f periodized on the bounded domain $\mathcal{D}_T = [-T,T]^d$ and truncate the integration in x and y by setting them to vary in $\mathcal{B}_0(R)$, the ball of center 0 and radius R. As seen in Section 5, a geometrical argument shows that using the periodicity 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.

The operator now reads

$$Q^R(f,f)(v) = \int_{\mathcal{B}_0(R)} \int_{\mathcal{B}_0(R)} \tilde{B}(x,y)\, \delta(x\cdot y) \\ [f(v+y)f(v+x) - f(v+x+y)f(v)]\, dx\, dy \quad (6.3)$$

for $v \in \mathcal{D}_T$.

By making some translation changes of variable on v (by x, y and $x+y$), using the changes $x \to -x$ and $y \to -y$ and the fact that

$$\tilde{B}(-x,y)\, \delta(-x\cdot y) = \tilde{B}(x,y)\, \delta(x\cdot y) = \tilde{B}(x,-y)\, \delta(x\cdot -y)$$

one can easily prove that for any function φ *periodic* on $\mathcal{D}_T$ the following weak form is satisfied

$$\int_{\mathcal{D}_T} Q^R(f,f)\, \varphi(v)\, dv = \frac{1}{4} \int_{\mathcal{D}_T} \int_{\mathcal{B}_0(R)} \int_{\mathcal{B}_0(R)} \tilde{B}(x,y)\, \delta(x\cdot y) f(v+x+y) f(v) \\ [\varphi(v+y) + \varphi(v+x) - \varphi(v+x+y) - \varphi(v)]\, dx\, dy\, dv. \quad (6.4)$$

It can be shown that if f has compact support included in $\mathcal{B}_0(R)$ with $T \geq (3+\sqrt{2})R/2$, then no unphysical collisions occur and thus mass, momentum and energy are preserved. Obviously this compactness is not preserved with

time since the collision operator spreads the support of f by a factor $\sqrt{2}$. In the rest of the paper we will focus on the periodized truncation Q^R.

6.2. Fast spectral methods

Now we use the above formulatuon to derive spectral methods that can be evaluated through fas algorithms. The presentation here follows Mouhot and Pareschi (2006) and Filbet et al. (2006).

To simplify notations let us take $T = \pi$. Using the same notations of Section 5 a straightforward computation leads to the following spectral quadrature for the collision operator

$$\hat{Q}_k = \sum_{\substack{l,m=-N \\ l+m=k}}^{N} \hat{\beta}_F(l,m)\, \hat{f}_l\, \hat{f}_m, \quad k = -N, ..., N \tag{6.5}$$

where $\hat{\beta}_F(l,m) = \hat{B}_F(l,m) - \hat{B}_F(m,m)$ are now given by

$$\hat{B}_F(l,m) = \int_{\mathcal{B}_0(R)} \int_{\mathcal{B}_0(R)} \tilde{B}(x,y)\, \delta(x \cdot y)\, e^{i(l\cdot x + m\cdot y)}\, dx\, dy. \tag{6.6}$$

Finally let us compare the new kernel modes with the ones in (5.10). The usual kernel modes written in the x and y variables reads

$$\hat{B}(l,m) = \int_{\mathcal{B}_0(R)} \int_{\mathcal{B}_0(R)} \tilde{B}(x,y)\, \delta(x \cdot y)\, \chi_{\{|x+y|\le R\}}\, e^{i(l\cdot x + m\cdot y)}\, dx\, dy. \tag{6.7}$$

Thus the usual representation contains a strong coupling between x and y which makes it very hard the construction of fast algorithms.

Now we search for a convolution structure in the equations (6.5). The aim is to approximate each $\hat{\beta}_F(l,m)$ by a sum

$$\hat{\beta}_F(l,m) \simeq \sum_{p=1}^{A} \alpha_p(l)\alpha'_p(m).$$

This gives a sum of A discrete convolutions and so the algorithm can be computed in $O(A\, N^d \log_2 N)$ operations by means of standard FFT techniques (Canuto et al. 1988, Cooley and Tukey 1965). Obviously this is equivalent to obtain such a decomposition on $\hat{B}_F$. To this purpose we shall use a further approximated collision operator where the number of possible directions of collision is reduced to a finite set.

A semi-discrete collision operator

We write x and y in spherical coordinates as follows

$$Q^R(f,f)(v) = \frac{1}{4} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} \delta(e \cdot e') \Bigg\{ \int_{-R}^{R} \int_{-R}^{R} \rho^{d-2}\, (\rho')^{d-2} \tilde{B}(\rho, \rho')$$

$$\left[f(v+\rho' e')f(v+\rho e) - f(v+\rho e+\rho' e')f(v)\right]\, d\rho\, d\rho' \Bigg\}\, de\, de'. \quad (6.8)$$

Let us take $\mathcal{A}$ a discrete set of orthogonal couples of unit vectors (e, e'), which is even: $(e, e') \in \mathcal{A}$ implies that $(-e, e')$, $(e, -e')$ and $(-e, -e')$ belong to $\mathcal{A}$ (this property on the set $\mathcal{A}$ is required to preserve the conservation properties of the operator). Now we define $Q_R^{\mathcal{A}}$ to be

$$Q^{R,\mathcal{A}}(f,f)(v) = \frac{1}{4}\int_{(e,e')\in\mathcal{A}} \Bigg\{ \int_{-R}^{R}\int_{-R}^{R} \rho^{d-2}\,(\rho')^{d-2}\,\tilde{B}(\rho,\rho') \left[f(v+\rho' e')f(v+\rho e) - f(v+\rho e+\rho' e')f(v)\right] d\rho\, d\rho' \Bigg\}\, d\mathcal{A}$$

where $d\mathcal{A}$ denotes a discrete measure on $\mathcal{A}$ which is also even in the sense that $d\mathcal{A}(e,e') = d\mathcal{A}(-e,e') = d\mathcal{A}(e,-e') = d\mathcal{A}(-e,-e')$. Using again translation change of variable on v by ρe, $\rho' e'$ and $\rho e + \rho' e'$ and the symmetries of the set $\mathcal{A}$ one can easily derive the following weak form on $Q_R^{\mathcal{A}}$. For any function φ periodic on $\mathcal{D}_T$,

$$\begin{aligned}\int_{\mathcal{D}_T} & Q^{R,\mathcal{A}}(f,f)\,\varphi(v)\,dv \\ &= \frac{1}{16}\int_{v\in\mathcal{D}_T}\int_{(e,e')\in\mathcal{A}}\int_{-R}^{R}\int_{-R}^{R} \rho^{d-2}\,(\rho')^{d-2}\,\tilde{B}(\rho,\rho') f(v+\rho e+\rho' e')f(v) \\ &\qquad \left[\varphi(v+\rho' e') + \varphi(v+\rho e) - \varphi(v+\rho e+\rho' e') - \varphi(v)\right]\, d\rho\, d\rho'\, d\mathcal{A}\, dv.\end{aligned}$$

This immediately gives the same conservations properties as Q_R.

Expansion of the kernel modes
We make the decoupling assumption that

$$\tilde{B}(x,y) = a(|x|)\, b(|y|). \quad (6.9)$$

This assumption is obviously satisfied if $\tilde{B}$ is constant. This is the case of Maxwellian molecules in dimension two, and hard spheres in dimension three (the most relevant kernel for applications). Extensions to more general interactions are discussed in Mouhot and Pareschi (2006).

First let us deal with dimension 2 with $\tilde{B} = 1$ to explain the method. Here we write x and y in spherical coordinates $x = \rho e$ and $y = \rho' e'$ to get

$$\hat{B}_F(l,m) = \frac{1}{4}\int_{\mathbb{S}^1}\int_{\mathbb{S}^1} \delta(e\cdot e') \left[\int_{-R}^{R} e^{i\rho(l\cdot e)}\, d\rho\right]\left[\int_{-R}^{R} e^{i\rho'(m\cdot e')}\, d\rho'\right] de\, de'.$$

Let us denote by

$$\varphi_R^2(s) = \int_{-R}^{R} e^{i\rho s}\, d\rho,$$

for $s \in \mathbb{R}$. It is easy to see that φ_R^2 is even and we can give the explicit formula

$$\varphi_R^2(s) = 2\,R\,\mathrm{Sinc}(Rs).$$

Thus we have

$$\hat{B}_F(l,m) = \frac{1}{4}\int_{\mathbb{S}^1}\int_{\mathbb{S}^1} \delta(e\cdot e')\,\varphi_R^2(l\cdot e)\,\varphi_R^2(m\cdot e')\,de\,de'$$

and thanks to the parity property of φ_R^2 we can adopt the following periodic parametrization

$$\hat{B}_F(l,m) = \int_0^{\pi} \varphi_R^2(l\cdot e_\theta)\,\varphi_R^2(m\cdot e_{\theta+\pi/2})\,d\theta.$$

The function $\theta \to \varphi_R^2(l\cdot e_\theta)\,\varphi_R^2(m\cdot e_{\theta+\pi/2})$ is periodic on $[0,\pi]$ and thus the rectangular quadrature rule is of infinite order and optimal. A regular discretization of M equally spaced points thus gives

$$\hat{B}_F(l,m) = \frac{\pi}{M}\sum_{p=0}^{M-1} \alpha_p(l)\alpha_p'(m) \tag{6.10}$$

with

$$\alpha_p(l) = \varphi_R^2(l\cdot e_{\theta_p}), \qquad \alpha_p'(m) = \varphi_R^2(m\cdot e_{\theta_p+\pi/2}) \tag{6.11}$$

where $\theta_p = \pi p/M$.

More generally under the decoupling assumption (6.9) on $\tilde{B}$, we get the following decomposition formula

$$\hat{B}_F(l,m) = \frac{\pi}{M}\sum_{p=0}^{M-1} \alpha_p(l)\alpha_p'(m) \tag{6.12}$$

where

$$\alpha_p(l) = \varphi_{R,a}^2(l\cdot e_{\theta_p}), \qquad \alpha_p'(m) = \varphi_{R,b}^2(m\cdot e_{\theta_p+\pi/2}) \tag{6.13}$$

and

$$\varphi_{R,a}^2(s) = \int_{-R}^{R} a(\rho)\,e^{i\rho s}\,d\rho, \qquad \varphi_{R,b}^2(s) = \int_{-R}^{R} b(\rho')\,e^{i\rho' s}\,d\rho' \tag{6.14}$$

with $\theta_p = \pi p/M$.

Remark 6.1. In the symmetric case $a = b$ (for instance for hard spheres) it is possible to parametrize $\hat{B}_F(l,m)$ as

$$\hat{B}_F(l,m) = 2\int_0^{\pi/2} \varphi_{R,a}^2(l\cdot e_\theta)\,\varphi_{R,a}^2(m\cdot e_{\theta+\pi/2})\,d\theta$$

and the function $\theta \to \varphi_{R,a}^2(l\cdot e_\theta)\,\varphi_{R,a}^2(m\cdot e_{\theta+\pi/2})$ is periodic on $[0,\pi/2]$.

Thus the decomposition can be obtained by applying the rectangular rule on this interval.

Now let us deal with dimension $d = 3$ with $\tilde{B}$ satisfying the decoupling assumption (6.9). First we change to the spherical coordinates

$$\hat{B}_F(l,m) = \frac{1}{4}\int_{\mathbb{S}^2}\int_{\mathbb{S}^2} \delta(e\cdot e') \left[\int_{-R}^{R} \rho a(\rho) e^{i\rho(l\cdot e)} d\rho\right] \left[\int_{-R}^{R} \rho' b(\rho') e^{i\rho'(m\cdot e')} d\rho'\right] dede'$$

and then we integrate first e' on the intersection of the unit sphere with the plane $e^{\perp}$,

$$\hat{B}_F(l,m) = \frac{1}{4}\int_{e\in\mathbb{S}^2} \varphi^3_{R,a}(l\cdot e) \left[\int_{e'\in\mathbb{S}^2\cap e^{\perp}} \varphi^3_{R,b}(m\cdot e')\, de'\right] de$$

where

$$\varphi^3_{R,a}(s) = \int_{-R}^{R} \rho\, a(\rho)\, e^{i\rho s}\, d\rho.$$

Thus we get the following decoupling formula with two degrees of freedom

$$\hat{B}_F(l,m) = \int_{e\in\mathbb{S}^2_+} \varphi^3_{R,a}(l\cdot e)\, \psi^3_{R,b}\big(\Pi_{e^{\perp}}(m)\big)\, de$$

where $\mathbb{S}^2_+$ denotes the half-sphere and

$$\psi^3_{R,b}\big(\Pi_{e^{\perp}}(m)\big) = \int_0^{\pi} \sin\theta\, \varphi_{R,b}\big(|\Pi_{e^{\perp}}(m)|\,\cos\theta\big)\, d\theta,$$

(this formula can be derived performing the change of variable $de' = \sin\theta\, d\theta\, d\varphi$ with the basis $(e, u = \Pi_{e^{\perp}}(m)/|\Pi_{e^{\perp}}(m)|, e\times u)$).

Again in the particular case where $\tilde{B} = 1$ (hard spheres model), we can compute explicitly the functions φ^3_R (in this case $a = b = 1$),

$$\varphi^3_R(s) = R^2\,\left[2\mathrm{Sinc}(Rs) - \mathrm{Sinc}^2(Rs/2)\right], \qquad \psi^3_R(s) = 2\,R^2\,\mathrm{Sinc}^2(Rs/2).$$

Now the function $e \to \varphi^3_{R,a}(l\cdot e)\,\psi^3_{R,b}\big(\Pi_{e^{\perp}}(m)\big)$ is periodic on $\mathbb{S}^2_+$ and so the rectangular rule is of infinite order and optimal. Taking a spherical parametrization (θ,φ) of $e \in \mathbb{S}^2_+$ and uniform grids of respective size M_1 and M_2 for θ and φ we get

$$\hat{B}_F(l,m) = \frac{\pi^2}{M_1 M_2} \sum_{p,q=0}^{M_1,M_2} \alpha_{p,q}(l)\alpha'_{p,q}(m)$$

where

$$\alpha_{p,q}(l) = \varphi^3_{R,a}(l\cdot e_{(\theta_p,\varphi_q)}), \qquad \alpha'_{p,q}(m) = \psi^3_{R,b}(\Pi_{e^{\perp}_{(\theta_p,\varphi_q)}}(m))$$

and

$$\varphi^3_{R,a}(s) = \int_{-R}^{R} \rho\, a(\rho)\, e^{i\rho s}\, d\rho, \qquad \psi^3_{R,b}(s) = \int_0^{\pi} \sin\theta\, \varphi^3_{R,b}(s\cos\theta)\, d\theta$$

and

$$(\theta_p, \varphi_q) = \Big(\frac{p\,\pi}{M_1}, \frac{q\,\pi}{M_2}\Big).$$

From now on we shall consider this expansion with $M = M_1 = M_2$ to avoid anisotropy in the computational grid.

Remark 6.2. For any dimension, we can construct as above an approximated collision operator $Q^{R,\mathcal{A}_M}$ with

$$\mathcal{A}_M = \Big\{(e, e') \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mid e \in \mathbb{S}^{d-1}_{M,+}, \quad e' \in e^{\perp} \cap \mathbb{S}^{d-1}\Big\}$$

where $\mathbb{S}^{d-1}_{M,+}$ denotes a uniform angular discretization of the half sphere with M points in each angular coordinate (the other half sphere is obtained by parity). Let us remark that this discretization contains exactly M^{d-1} points. From now on we shall denote

$$Q^{R,M} = Q^{R,\mathcal{A}_M} = \sum_{p=1}^{M^{d-1}} Q^{R,M}_p.$$

Spectral accuracy

We are interested in computing the accuracy of the scheme according to the three parameters N (the number of modes), R (the truncation parameter), and M (the number of angular directions for each angular coordinate). Instead of looking at the error on each kernel mode it is more convenient to look at the error on the global operator. Here the Lebesgue spaces L^p, $p = 1 \ldots + \infty$, and the periodic Sobolev spaces H^r_p, $r = 0 \ldots + \infty$ refer to $\mathcal{D}_\pi$.

So in order to give a consistency result, the first step will be to prove a consistency result for the approximation of Q^R by $Q^{R,M}$ (see Mouhot and Pareschi (2006) for details).

Lemma 6.1. The error on the approximation of the collision operator is spectrally small, i.e for all $r > d-1$ such that $f \in H^r_p$

$$\|Q^R(g,f) - Q^{R,M}(g,f)\|_{L^2} \le C_1\, \frac{R^r}{M^r} \|g\|_{H^r_p} \|f\|_{H^r_p}.$$

For the second step we shall use the consistency result of Theorem 5.2 on the operator Q^R. Combining these two results, one gets the following consistency result

Table 6.1. Relative L_1 norm of the error for different values of N and M for the fast spectral method.

N	M=2	M=4	M=8	M=16
32	2.129E-04	1.993E-05	2.153E-05	2.262E-05
64	2.109E-04	7.122E-10	6.830E-10	6.843E-10
128	2.112E-04	3.116E-12	3.117E-12	3.117E-12

Theorem 6.1. For all $r > d-1$ such that $f \in H^r_p(\mathcal{D}_\pi)$

$$\|Q^R(f) - Q^{R,M}_N(f_N)\|_{L^2} \le C_1 \frac{R^r}{M^r}\|f_N\|^2_{H^r_p} + \frac{C_2}{N^r}\left(\|f\|_{H^r_p} + \|Q^R(f_N)\|_{H^r_p}\right).$$

In the above theorem, to simplify notations, we used $Q^R(f)$ instead of $Q^R(f,f)$.

Now let us focus briefly on the macroscopic quantities. First with Lemma 6.1 at hand one can establish the estimate

$$\|Q^{R,M}(g,f)\|_{L^2} \le C\,\|g\|_{H^d_p}\|f\|_{H^d_p},$$

for a constant uniform in M. Then following the method of (Pareschi and Russo 2000*b*, Remark 5.4) and using this estimate we obtain the following spectral accuracy result

$$\begin{aligned}&\left|\langle Q^{R,M}(f,f),\varphi\rangle - \langle Q^{R,M}_N(f_N,f_N),\varphi\rangle\right|_{L^2}\\ &\quad\le \frac{C_3}{N^r}\,\|\varphi\|_{L^2}\left(\|f\|_{H^{k+d}_p} + \|Q^{R,M}(f_N,f_N)\|_{H^r_p}\right)\end{aligned}$$

where φ can be replaced by $v, |v|^2$. Indeed there is no need to compare the momenta of $Q^{R,M}_N(f_N,f_N)$ with those of $Q^R(f,f)$ since $Q^{R,M}$ is also conservative, and so they can be compared directly to those of $Q^{R,M}$. Thus the error on momentum and energy is independent on M and is spectrally small according to N even for very small value of the parameter M. The same considerations of Remark 5.4 remain valid concerning the derivation of exactly conservative methods based on the L_2 projection (3.44).

Implementation aspects

The method of the previous subsections yields a decomposition of the collision operator, which after projection on $\mathbb{P}^N$ gives the following decomposi-

tion

$$Q_N^{R,M} = \sum_{p=1}^{M^{d-1}} \mathcal{P}_N Q_p^{R,M}. \tag{6.15}$$

Each $\mathcal{P}_N Q_p^{R,M}$ can be computed with a cost $O(N^d \log_2 N)$. Thus for a general choice of M and N we obtain the cost $O(M^{d-1} N^d \log_2 N)$. The decomposition (6.15) is completely parallelizable and thus the cost can be strongly reduced on a parallel machine (formally up to $O(N^d \log_2 N)$). One just has to make independent computations for the M^{d-1} terms of the decomposition. The decomposition can be also interesting from the storage viewpoint, as the classical spectral method requires the storage of a $N^d \times N^d$ matrix whereas the fast method requires the storage of $2M^{d-1}$ vectors of size N^d. As a numerical example we report the results obtained in the case of space homogeneous two-dimensional Maxwellian molecules using as a comparison the exact analytic solution (5.39). The results for the relative L_1 norm of the error at time $t = 0.01$ are reported in Table 6.1 and indicate a very low influence of the number of directions over the accuracy of the scheme.

6.3. Fast spectral methods for the quantum-Boltzmann equation

The fast method studied in the previous section is closely related to the possibility of representing the collision operator in the form (6.1). A relevant example is given by the quantum-Boltzmann equation. Here we follow the derivation of the fast method by Filbet et al. (2012) and the improvements obtained in Hu and Ying (2012). Related fast solvers in a discrete velocity setting for a one-dimensional model have been constructed in Markowich and Pareschi (2005).

In this case the collision term contains a cubic nonlinearity and in dimension d reads

$$\begin{aligned} Q_q(f)(v) \;=\; & \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} B(v, v_*, \omega) \left[f' f'_* (1 \pm \theta_0 f)(1 \pm \theta_0 f_*) \right. \\ & \left. - f f_* (1 \pm \theta_0 f')(1 \pm \theta_0 f'_*) \right] d\omega dv_* \end{aligned} \tag{6.16}$$

where $\theta_0 = \hbar^d$, $\hbar$ is the rescaled Planck constant. Here, the upper sign ('+') corresponds to the Bose gas while the lower sign ('-') to the Fermi gas. For the Fermi gas, we also need $f \le {\theta_0}^{-1}$ by the Pauli exclusion principle.

We first write (6.16) as

$$Q_q = Q \pm \theta_0 (Q_1 + Q_2 - Q_3 - Q_4), \tag{6.17}$$

where Q is the classical collision operator of rarefied gas dynamics. The

cubic terms Q_1 – Q_4 are

$$\begin{aligned} Q_1(f)(v) &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} B(v-v_*,\omega) f' f'_* f_* d\omega dv, \\ Q_2(f)(v) &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} B(v-v_*,\omega) f' f'_* f d\omega dv, \\ Q_3(f)(v) &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} B(v-v_*,\omega) f f_* f' d\omega dv, \\ Q_4(f)(v) &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} B(v-v_*,\omega) f f_* f'_* d\omega dv. \end{aligned} \tag{6.18}$$

In order to derive a spectral method, we periodize the function f in the conventional way over the domain $\mathcal{D}_T = [-T,T]^d$ where T is chosen such that $T \geq (3+\sqrt{2})R/2$, R is the truncation of the collision integral (see Section 5.1).

For the sake of simplicity we report the details in the case of $d = 2$ for Maxwell molecules. We can apply the identity (2.24) used to represent Q in the form (6.1) to the cubic terms to get

$$\begin{aligned} Q_1(f)(v) &= \int_{\mathcal{B}_0(R)}\int_{\mathcal{B}_0(R)} \delta(x\cdot y) f(v+x) f(v+y) f(v+x+y) dxdy, \\ Q_2(f)(v) &= \int_{\mathcal{B}_0(R)}\int_{\mathcal{B}_0(R)} \delta(x\cdot y) f(v+x) f(v+y) f(v) dxdy, \\ Q_3(f)(v) &= \int_{\mathcal{B}_0(R)}\int_{\mathcal{B}_0(R)} \delta(x\cdot y) f(v+x) f(v+x+y) f(v) dxdy, \\ Q_4(f)(v) &= \int_{\mathcal{B}_0(R)}\int_{\mathcal{B}_0(R)} \delta(x\cdot y) f(v+y) f(v+x+y) f(v) dxdy. \end{aligned} \tag{6.19}$$

Now the spectral method applies in a standard way and we focus on the cubic terms, since the classical part can be treated through the fast spectral scheme described in the last section. Starting from the kernel modes defined in (6.6) we assume that they can be decomposed accordingly to (6.10). We have:

- The k-th coefficient of $\hat{Q}_1$ is

$$\sum_{\substack{l,m,n=-N \\ l+m+n=k}}^{N} \hat{B}_F(l+n,m+n)\hat{f}_l\hat{f}_m\hat{f}_n$$

$$= \frac{\pi}{M} \sum_{p=0}^{M-1} \sum_{n=-N}^{N} \left[\sum_{\substack{l,m=-N \\ l+m=k-n}}^{N} \alpha_p(l+n)\alpha'_p(m+n)\hat{f}_l\hat{f}_m \right] \hat{f}_n$$

$$= \frac{\pi}{M} \sum_{p=0}^{M-1} \sum_{n=-N}^{N} \hat{g}_{k-n}(n)\hat{f}_n. \tag{6.20}$$

Terms inside the bracket is a convolution (defined as $\hat{g}_{k-n}(n)$), which can be computed by the Fast Fourier Transform (FFT). However, the outside structure is not a convolution, since $\hat{g}_{k-n}(n)$ itself depends on n.

- The k-th coefficient of $\hat{Q}_2$ is

$$\sum_{\substack{l,m,n=-N \\ l+m+n=k}}^{N} \hat{B}_F(l,m)\hat{f}_l\hat{f}_m\hat{f}_n$$

$$= \frac{\pi}{M} \sum_{p=0}^{M-1} \sum_{n=-N}^{N} \left[\sum_{\substack{l,m=-N \\ l+m=k-n}}^{N} \alpha_p(l)\alpha'_p(m)\hat{f}_l\hat{f}_m \right] \hat{f}_n. \tag{6.21}$$

In this case, both inside and outside are convolutions. The FFT can be implemented easily.

- The k-th coefficient of $\hat{Q}_3$ is

$$\sum_{\substack{l,m,n=-N \\ l+m+n=k}}^{N} \hat{B}_F(l+m,m)\hat{f}_l\hat{f}_m\hat{f}_n$$

$$= \frac{\pi}{M} \sum_{p=0}^{M-1} \sum_{n=-N}^{N} \alpha_p(l+m) \left[\sum_{\substack{l,m=-N \\ l+m=k-n}}^{N} \alpha'_p(m)\hat{f}_l\hat{f}_m \right] \hat{f}_n. \tag{6.22}$$

Factoring out $\alpha_p(l+m)$, both inside and outside are convolutions again.

- The k-th coefficient of $\hat{Q}_4$ is

$$\sum_{\substack{l,m,n=-N \\ l+m+n=k}}^{N} \hat{B}_F(m,l+m)\hat{f}_l\hat{f}_m\hat{f}_n$$

$$= \frac{\pi}{M} \sum_{p=0}^{M-1} \sum_{n=-N}^{N} \alpha'_p(l+m) \left[\sum_{\substack{l,m=-N \\ l+m=k-n}}^{N} \alpha_p(m)\hat{f}_l\hat{f}_m \right] \hat{f}_n. \tag{6.23}$$

Table 6.2. Comparison of the fast quantum solver on different Maxwellians.

model	θ_0	16×16	32×32	64×64	convergence rate
classical gas	0	2.1746E-04	3.8063E-12	1.9095E-16	20.03
Bose gas	0.01	2.1084E-04	2.5512E-10	1.9080E-16	20.00
	9	4.891E-01	3.10E-02	1.3496E-04	5.91
Fermi gas	0.01	2.2397E-04	1.6485E-10	1.9152E-16	20.05
	9	8.9338E-04	2.0192E-06	1.5962E-10	11.21

This term can be evaluated similarly as $\hat{Q}_3$.

The computational cost of the method is $O(M^{d-1}N^{2d}\log N)$, which mainly comes from computing Q_1. The cost is slightly higher than $O(M^{d-1}N^{2d})$ typical of a discrete velocity model based on a product quadrature rule. But taking into account the high accuracy, the method may be considered still more attractive than the quadrature method. A deeper analysis on the structure of Q_1 based on an exponential decomposition shows that the cost can be further reduced to $O(M^{d-1}N^{d+1}\log N)$ which in dimension $d = 3$ implies a gain of a factor N^2 (Hu and Ying 2012).

To illustrate the spectral accuracy of the above method, we consider a steady state problem, namely, we compute the max norm of $Q_q(M_q[f])$ where $M_q[f]$ is the quantum Maxwellian given by

$$M_q[f] = \frac{1}{\theta_0}\frac{1}{z^{-1}e^{\frac{(v-u)^2}{2T}} \mp 1}, \tag{6.24}$$

with z the fugacity and T the temperature (see Escobedo et al. (2003) for more details). This corresponds to the well-known Bose-Einstein ('-') and Fermi-Dirac ('+') distributions. We consider also the classical case corresponding to $\theta_0 = 0$. In all the numerical simulations, the particles are assumed to be the 2-D Maxwellian molecules. In Table 6.2, we list the values of $\| Q_q(M_q[f]) \|_{L^\infty}$ computed on different meshes $N = 16$, 32, 64, $M = 4$. The computational domain is $[-8, 8] \times [-8, 8]$.

The results confirm the spectral accuracy of the method, although the accuracy becomes worse when θ_0 is increasing since the domain has been kept fixed. To remedy this problem, one can add more grid points or more effectively, shorten the computational domain.

6.4. Fast discrete velocity methods

In this section we will see how the fast algorithms developed for the spectral method can be extended to periodized discrete velocity methods. The method that originates is in some sense related to the direct FFT approach proposed in Bobylev and Rjasanow (1999).

In Section 4.2 starting from (4.58) we have seen how to derive discrete velocity methods (Panferov and Heintz 2002, Mouhot et al. 2013). Similarly to the spectral method the representation of Q^R in (6.3) can also be used to derive fast discrete velocity methods in the form (4.59). Here, however, instead of the usual truncation based on neglecting collisions violating the velocity bounds we periodize the function f over the box and truncate the sum in k and l. It yields for a given truncation parameter $\tilde{N} \in \mathbb{N}^*$

$$Q_i^{\tilde{N}}(f,f) = \sum_{-\tilde{N} \le k,l \le \tilde{N}} \widetilde{\Gamma}_{k,l} \big[f_{i+k} f_{i+l} - f_i f_{i+k+l} \big], \tag{6.25}$$

for any $i \in [\![-N, N]\!]^d$. In the above sum we have

$$\widetilde{\Gamma}_{kl} = \tilde{B}(k,l) \mathbf{1}(k \cdot l)\, W_{k,l}.$$

It can be shown that $Q^{\tilde{N}}$ preserves exactly the mass and it preserves momentum and energy up to aliasing issues. Thus, for a sufficiently large computational domain, it is exact up to machine precision. This is slightly different from spectral methods where the truncation of Fourier modes introduces a spectrally small error in the conservation laws. Moreover in a space homogenous setting it preserves also non negativity of the solution and therefore we have also the stability of the method (Mouhot et al. 2013). Finally a consistency results analogous of Theorem 4.8 holds true (Mouhot et al. 2013).

Principle of the method: a pseudo-spectral viewpoint

We start from the periodized DVM in $[\![-N, N]\!]^d$ with representation (6.25) and as in the continuous case we set, for $k, l \in -\tilde{N} \le k, l \le \tilde{N}$,

$$\tilde{B}(|k|,|l|) = 2^{d-1}\, B\left(\frac{|k|}{\sqrt{|k|^2+|l|^2}}, \sqrt{|k|^2+|l|^2} \right) (|k|^2+|l|^2)^{-\frac{d-2}{2}}.$$

With this notation

$$\widetilde{\Gamma}_{k,l} = \mathbf{1}(k \cdot l)\, \tilde{B}(|k|,|l|)\, W_{k,l},$$

and thus discrete collision operator becomes

$$Q_i^{\tilde{N}} = \sum_{-\tilde{N} \le k,l \le \tilde{N}} \mathbf{1}(k \cdot l)\, \tilde{B}(|k|,|l|)\, W_{k,l} \left[f_{i+k} f_{i+l} - f_i f_{i+k+l} \right].$$

Now we transform this discrete operator into a new one using the involution transformation of the discrete Fourier transform on the vector $(f_i)_{-N\le i\le N}$. This involution reads for $I \in [\![-N, N]\!]^d$

$$\tilde{f}_I = \frac{1}{2N+1} \sum_{i=-N}^{N} f_i \, \mathbf{e}_{-I}(i), \qquad f_i = \sum_{I=-N}^{N} \tilde{f}_I \, e_I(i)$$

where $\mathbf{e}_K(k)$ denotes $e^{\frac{2i\pi K\cdot k}{2N+1}}$, and thus we have

$$Q_i^{\tilde{N}} = \sum_{I=-N}^{N} \tilde{Q}_I \, e_I(i) \tag{6.26}$$

with

$$\tilde{Q}_I = \sum_{K,L=-N}^{N} \left(\frac{1}{2N+1} \sum_{i=-N}^{N} \mathbf{e}_{K+L-I}(i) \right) \left[\sum_{-\tilde{N}\le k,l\le \tilde{N}} \mathbf{1}(k\cdot l)\, \tilde{B}(|k|,|l|)\, W_{k,l}\, \left(\mathbf{e}_K(k)\mathbf{e}_L(l) - \mathbf{e}_L(k+l)\right) \right] \tilde{f}_K \, \tilde{f}_L$$

for $-N \le I \le N$. We have the following identity

$$\frac{1}{2N+1} \sum_{i=-N}^{N} \mathbf{e}_{K+L-I}(i) = \mathbf{1}(K+L-I)$$

and so

$$\tilde{Q}_I = \sum_{\substack{K,L=-N \\ K+L=I}}^{N} \tilde{\beta}(K,L)\, \tilde{f}_K \, \tilde{f}_L \tag{6.27}$$

with $\tilde{\beta}(K,L) = \beta(K,L) - \beta(L,L)$ where

$$\beta(K,L) = \sum_{-\tilde{N}\le k,l\le \tilde{N}} \mathbf{1}(k\cdot l)\, \tilde{B}(|k|,|l|)\, W_{k,l}\, \mathbf{e}_K(k)\mathbf{e}_L(l). \tag{6.28}$$

Let us first remark that this new formulation allows to reduce the usual cost of computation of a discrete velocity model exactly to $O(N^{2d})$ (as with the usual spectral method) instead of $O(N^{2d+\delta})$ for $\delta \sim 1$ (Buet 1996, Panferov and Heintz 2002). Note however that the $(2N+1)^d \times (2N+1)^d$ matrix of coefficients $(\beta(K,L))_{K,L}$ has to be computed and stored first, thus the storage requirements are larger. Nevertheless symmetries in the matrix can substantially reduce this cost.

Now, as for the spectral method, the aim is to give an expansion of $\beta(K,L)$

of the form

$$\beta_{K,L} \simeq \sum_{p=1}^{M} \alpha_p(K)\, \alpha'_p(L),$$

for a parameter $M \in \mathbb{N}^*$ to be defined later.

Expansion of the discrete kernel modes
We make a decoupling assumption on the collision kernel as in the spectral case

$$\tilde{B}(|k|, |l|)\, W_{k,l} = a(k)\, b(l). \tag{6.29}$$

Note that the DVM constructed by quadrature in dimension 3 for hard spheres in (Panferov and Heintz 2002) on the cartesian velocity grid $h\,\mathbb{Z}^3$ (for $h > 0$) satisfies this decoupling assumption with $a(k) = h^5\,|k|/\gcd(k_1, k_2, k_3)$ and $b(l) = 1$, and $\gcd(k_1, k_2, k_3)$ denotes the greater common divisor of the three integers. For Maxwell molecules in dimension 2 on the grid $h\,\mathbb{Z}^2$, these coefficients are $a(k) = h^3\,|k|/\gcd(k_1, k_2)$ and $b(l) = 1$.

The difference here with the spectral method, which is a continuous numerical method, is that we have to enumerate the set of $\{-\tilde{N} \le k, l \le \tilde{N} \mid k \perp l\}$. This motivates for a detailed study of the number of lines passing through 0 and another point in the grid (this is equivalent to the study of this set), in order to compute the complexity of the method in term of N.

To this purpose let us introduce the Farey series and a new parameter $0 \le \bar{N} \le \tilde{N}$ for the size of the grid used to compute the number of directions. The usual Farey series is

$$\mathcal{F}^1_{\bar{N}} = \left\{(p, q) \in [\![0, \bar{N}]\!]^2 \mid 0 \le p \le q \le \bar{N},\ q \ge 1,\ \text{and } \gcd(p, q) = 1\right\}$$

where $\gcd(p, q)$ denotes again the greater common divisor of the two integers (more details can be found in (Hardy and Wright 1979)). It is straightforward to see that the number of lines $A^1_{\bar{N}}$ passing through 0 in the grid $[\![-\bar{N}, \bar{N}]\!]^2$ is

$$A^1_{\bar{N}} = 4\left(\left|\mathcal{F}^1_{\bar{N}}\right| - 1\right),$$

where the factor 4 allows to take into account the permutations when counting the couples (p, q) as well as the ordering, minus the line which is repeated during the symmetry process. We gave a schematic representation of the two dimensional Farey series in Figure 6.1.

Similarly one can define the set

$$\mathcal{F}^2_{\bar{N}} = \left\{(p, q, r) \in [\![0, \bar{N}]\!]^3 \mid 0 \le p \le q \le r \le \bar{N},\ r \ge 1,\ \text{and } \gcd(p, q, r) = 1\right\}$$

and the number of lines $A^2_{\bar{N}}$ passing through 0 in the grid $[\![-\bar{N}, \bar{N}]\!]^3$ is

$$A^2_{\bar{N}} = 24\left(\left|\mathcal{F}^2_{\bar{N}}\right| - \left|\mathcal{F}^1_{\bar{N}}\right|\right) - 2\,A^1_{\bar{N}}$$

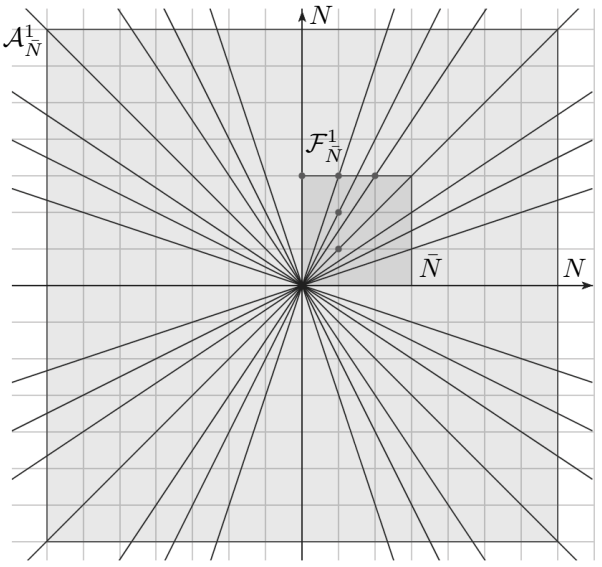

Figure 6.1. Representation of the Farey series $\mathcal{F}^1_{\bar N}$ and $\mathcal{A}^1_{\bar N}$, the associated primal representant of lines in $[-N, N]$, for $N = 7$ and $\bar N = 3$.

all possible permutations of the three numbers times 4 and minus the interfaces $2A^1_{\bar N}$ accounting for the possible negative values by symmetry, minus $24\left|\mathcal{F}^1_{\bar N}\right|$ for the spurious terms when two equal numbers are swapped. The exponents of the Farey series refer to the dimension of the space of lines (which is $d-1$). Now we can estimate the cardinals of $\mathcal{F}^1_{\bar N}$ and $\mathcal{F}^2_{\bar N}$ (Mouhot et al. 2013).

Lemma 6.2. The Farey series in dimension $d = 2$ and $d = 3$ satisfy the following asymptotic behavior

$$\begin{aligned}\left|\mathcal{F}^1_{\bar N}\right| &= \frac{\bar N^2}{2\,\zeta(2)} + O(\bar N\,\log \bar N) = \frac{3\bar N^2}{\pi^2} + O(\bar N\,\log \bar N),\\ \left|\mathcal{F}^2_{\bar N}\right| &= \frac{\bar N^3}{12\,\zeta(3)} + O(\bar N^2),\end{aligned}$$

where $\zeta(s) = \sum_{n\geq 0} n^{-s}$ denotes the usual Riemann zeta function.

Next one can deduce the following decomposition of the kernel modes using their definition (6.28) and the decoupling assumption (6.29) on the discrete kernel

$$\beta(K, L) = \sum_{-\tilde N\leq k,l\leq \tilde N} \mathbf{1}(k\cdot l)\, a(|k|)\, b(|l|)\, e_K(k) e_L(l)$$

$$\simeq \quad \beta^{\bar{N}}(K,L) = \sum_{e\in\mathcal{A}_{\bar{N}}^{d-1}} \Big[\sum_{\substack{k\in e\mathbb{Z} \\ -\tilde{N}\le k\le\tilde{N}}} a(|k|)\, e_K(k) \Big] \Big[\sum_{\substack{l\in e^\perp \\ -\tilde{N}\le l\le\tilde{N}}} b(|l|)\, e_L(l) \Big]$$

with equality if $\bar{N} = \tilde{N}$. Here $\mathcal{A}_{\bar{N}}^{d-1}$ denotes the set of primal representants of directions of lines in $[\![-\bar{N}, \bar{N}]\!]$ passing through 0. After indexing this set, which has cardinal $A_{\bar{N}}^{d-1}$, one gets

$$\beta^{\bar{N}}(K,L) = \sum_{p=1}^{A_{\bar{N}}^{d-1}} \alpha_p(K)\, \alpha'_p(L) \tag{6.30}$$

with

$$\alpha_p(K) = \sum_{\substack{k\in e_p\,\mathbb{Z} \\ -\tilde{N}\le k\le\tilde{N}}} a(|k|)\, e_K(k), \qquad \alpha'_p(L) = \sum_{\substack{l\in e_p^\perp \\ -\tilde{N}\le l\le\tilde{N}}} b(|l|)\, e_L(l).$$

After inversion of the discrete Fourier transform, this method yields a decomposition of the discrete collision operator

$$Q_i^{\tilde{N}} \simeq Q_i^{\tilde{N},\bar{N}} = \sum_{p=1}^{A_{\bar{N}}^{d-1}} Q_i^{\tilde{N},\bar{N},p}, \qquad i \in [\![-N, N]\!]^d, \tag{6.31}$$

with equality with (6.25) if $\bar{N} = \tilde{N}$. Each $Q_i^{\tilde{N},\bar{N},p}(f,f)$ is defined by the p-th term of the decomposition of the kernel modes (6.30). Each term $Q^{\tilde{N},\bar{N},p}$ of the sum is a discrete convolution operator when it is written in Fourier space. Moreover, each α_p (resp. α'_p) is defined as the discrete Fourier transform of some non-negative coefficients $a(|k|)$ times the characteristic function of $k \in e_p\mathbb{Z}$ (resp. $b(|l|)$ times the characteristic function of $l \in e_p^\perp$). Hence, we get after inversion of the transform that $Q^{\tilde{N},\bar{N},p}$ is a discrete convolution with non-negative coefficients.

By using the approximate kernel modes $\beta^{\bar{N}}(K,L)$, we obtain a new discrete collision operator, which inherits the same nice stability properties as the usual DVM schemes (Mouhot et al. 2013).

Computational considerations

The fast DVM method described in the last subsection depends on the three parameters N (the size of the gridbox), R (the truncation parameter) and $\bar{N}$ (the size of the box in the space of lines). Thus one can see thanks to Lemma 6.2 that even if we take $\bar{N} = \tilde{N} = N$, i.e. we take all possible directions in the grid $[\![-N, N]\!]^d$, we get the computational cost $O(N^{2d} \log_2 N)$ which is better than the usual cost of the DVM, $O(N^{2d+1})$ (but slightly worse than the cost $O(N^{2d})$ obtained by solving directly the pseudo-spectral scheme, thanks to a bigger storage requirement).

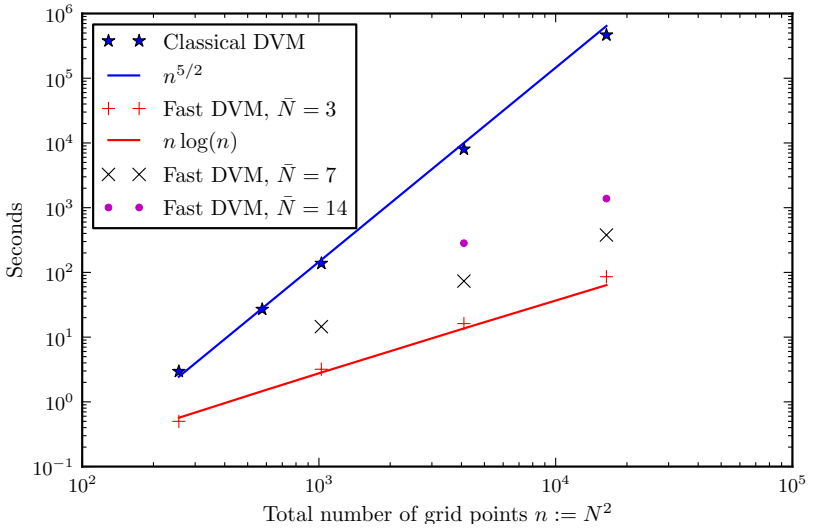

Figure 6.2. Computational time with respect to the total number of points for the classical and fast DVM methods for various values of $\bar{N}$ in two dimensions.

More generally for a choice of $\bar{N} < N$ we obtain the cost $O(\bar{N}^d N^d \log_2 N)$, which is slightly worse than the cost of the fast spectral algorithm (namely $O(M^{d-1} N^d \log_2 N)$ where M is the number of discrete angle, but interesting given that the algorithm is accurate for small values of $\bar{N}$, and more stable. The justification for this is the low accuracy of the method (the reduction of the number of direction has a small effect on the overall accuracy of the scheme).

Remark 6.3.

1 Concerning the construction of the set of directions $\mathcal{A}^d_{\bar{N}}$, it can be done with systematic algorithms of iterated subdivisions of a simplex, thanks to the properties of the Farey series. In dimension $d = 2$ this construction is quite simple (see Hardy and Wright (1979)). In dimension 3 we refer to Nogueira and Sevennec (2006).

2 Let us remark that in order to get a *regular* scheme (i.e with no other conservation laws than the usual ones) in spite of the reduction of directions, it is enough that the schemes contains the directions 0 and $\pi/2$ (Cercignani 1985). This is satisfied if we take the directions contained in $\mathcal{F}_1^{d-1}$, i.e. as soon as $\bar{N} \geq 1$.

3 In the practical implementation of the algorithm one has to take advantage of the symmetry of the decomposition (6.30) in order to reduce the

Table 6.3. Comparison of the L^1 error between the classical DVM method and the fast DVM method with different values of $\bar{N}$ at time $T = 0.01$, after one iteration.

Number of points N	Classical DVM	Fast DVM with $\bar{N} = 1$	Fast DVM with $\bar{N} = 3$	Fast DVM with $\bar{N} = 7$	Fast DVM with $\bar{N} = 14$
8	1.445E-03	1.4511E-03	x	x	x
16	8.912E-04	9.887E-04	8.9646E-04	x	x
32	6.1054E-04	6.5209E-04	5.8397E-04	6.1328E-04	x
64	2.6351E-04	4.094E-04	2.906E-04	3.667E-04	2.7341E-04
128	x	2.6669E-04	1.8245E-04	2.0371E-04	1.6341E-04

number of terms in the sum: for instance in dimension 2, if $a = b = 1$, one can write a decomposition with $A_{\bar{N}}^{d-1}/2$ terms.

Finally we report the results of an accuracy test for the exact solution (5.39) of the homogeneous Boltzmann equation in dimension 2, with Maxwell molecules. We will compare the fast DVM method with the method introduced in Panferov and Heintz (2002), referred to as classical DVM. We compare the error at a given time T_{end} when using $N = 8$ to $N = 128$ grid points for each coordinate (the case $N = 128$ for the classical DVM has been omitted due to its large computational cost). In Table 6.3 we give the results obtained by the classical DVM method and the fast one, with different numbers of $\bar{N}$. The corresponding computational times are plotted in Figure 6.2. We choose the value $\tilde{N}$ such that the classical method is convergent according to Theorem 4.8, namely

$$\tilde{N} = \left[\frac{2N}{3 + \sqrt{2}} \right].$$

Then, one has $\tilde{N} = 1$ when $N = 8$, $\tilde{N} = 3$ when $N = 16$, $\tilde{N} = 7$ when $N = 32$ and $\tilde{N} = 14$ when $N = 64$. These values give a result corresponding to the kernel mode (6.28), namely that no truncation of the number of lines has been done: the solution obtained is essentially the same obtained with the classical DVM method. Note that $\bar{N}$ must be chosen less of equal than $\tilde{N}$ and this is why we do not present the results with, e.g., $N = 16$ and $\bar{N} = 7$. The size of the domain has to be chosen carefully in order to minimize the aliasing error. In this test, we used $T = 5$ for $N = 8$, $T = 5.5$ for $N = 16$, $T = 7$ for $N = 32$ and $T = 8$ for $N = 64,\ 128$. We can see that, even with very few directions, there is a small loss of accuracy for the fast

DVM method compared to the classical one, and that taking all possible directions we recover the original DVM solution. The observed order of convergence in N is close to 1, as predicted by Theorem 4.8

From the computational cost point of view, taking *e.g.* $N = 64$ points in each direction, the fast method is more than 28 time faster than the classical one when no truncation is done (i.e. when we take $\bar{N} = \tilde{N} = 14$), and even 109 times faster with a small loss of accuracy when taking $\bar{N} = 7$.

7. Asymptotic-preserving schemes

The numerical solution of kinetic equations in stiff regimes represents a challenge in the construction of computational methods. In such regimes particles interactions typically drive the underlying kinetic densities toward local equilibria. This fact allows solutions of the kinetic equation to be approximated by solutions of a reduced system, typically a fluid-dynamical system or diffusion equations (Cercignani 1988, Cercignani et al. 1994, Degond et al. 2004), that can be solved efficiently by classical numerical methods. However there are regimes, where collisions are plentiful enough to make the kinetic equation stiff but not enough to drive the kinetic system close to local equilibria. These transition regions are typically the most difficult to solve numerically. In these cases, the multiscale nature of the physical problem leads both to time or space step limitations which may become extremely restrictive for numerical simulations, either at deterministic or at stochastic level. On the other hand, the use of implicit schemes would allow larger time steps but presents considerable limitations in the Boltzmann case since the collisional operator is highly nonlinear and nonlocal and therefore its inversion prohibitive.

Asymptotic-preserving (AP) schemes have been particularly successful in the construction of unconditionally stable numerical methods which are capable to capture the correct asymptotic behavior of the system by avoiding the resolution of small scales (Coron and Perthame 1991, Jin 1995, Caflisch, Jin and Russo 1997, Jin, Pareschi and Toscani 1998, Jin 1999, Jin and Pareschi 2001, Klar 1998*a*, Gosse and Toscani 2002, Pareschi and Russo 2005). The main common idea of asymptotic preserving techniques is to allow the use of the same numerical scheme to discretize a perturbation problem and its limit problem, with fixed discretization parameters. This permits to match regions where the perturbation parameter has very different orders of magnitude. In this case, the AP scheme realizes an automatic transition between the perturbation problem and its limit problem, therefore avoiding most drawbacks of techniques based on model coupling. It is such an active field of research that it is essentially impossible to do a comprehensive review of the different techniques developed. For recent surveys on asymptotic-preserving schemes for various kinds of

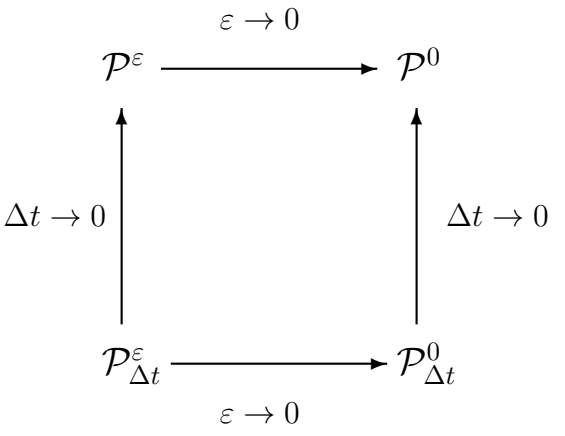

Figure 7.1. $\mathcal{P}^\varepsilon$ is a singular perturbation problem (the kinetic equation) and $\mathcal{P}^\varepsilon_{\Delta t}$ its numerical approximation characterized by the discretization parameter Δt. The method is asymptotic-preserving (AP) if $\mathcal{P}^\varepsilon_{\Delta t}$ is a consistent and stable discretization of $\mathcal{P}^0$ (the macroscopic limit) as $\varepsilon \to 0$ for a fixed Δt.

systems we refer to the review papers by Jin (2012), Pareschi and Russo (2011) and Degond (2014). We also mention the book by Gosse (2013) for related problems for balance laws. For its relevance in applications, in this section we focus our attention on asymptotic preserving schemes specifically designed for the full Boltzmann equation in the classical fluid-limit (Gabetta et al. 1997, Filbet and Jin 2010, Lemou 2010, Dimarco and Pareschi 2011, Dimarco and Pareschi 2013) and report some recent advancements on the diffusion limit (Jin et al. 2000, Bennoune, Lemou and Mieussens 2008, Lemou and Mieussens 2008, Boscarino et al. 2013, Dimarco, Pareschi and Rispoli 2014). The general problem we will consider can be written in the form (2.53) that we rewrite here

$$\varepsilon^\alpha \frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f, f), \tag{7.1}$$

where $\varepsilon > 0$ is a small scaling parameter. The fluid-limit corresponds to $\alpha = 0$ and the diffusion limit to $\alpha = 1$. For small values of ε we have the presence of two scales, the scale $\varepsilon^{1+\alpha}$ that forces f towards its local Maxwellian and the scale ε^α which may originate some diffusive behavior in the asymptotic process. In order to clarify the meaning of AP that will be used in the sequel it is useful to introduce the following general definition (see Figure 7.1).

Definition 7.1. A consistent and stable time discretization method for (7.1) of stepsize Δt is *asymptotic preserving (AP)* if, for a fixed time step

Δt, in the limit $\varepsilon \to 0$ becomes a consistent and stable time discretization method for the reduced equilibrium system.

The method discussed in this section are therefore based on solving directly problem (7.1) in the whole computational domain independently on the order of magnitude of the scaling parameter ε. Other approaches, based on dynamic domain decomposition strategies and hybrid methods will be reviewed in Section 8.

7.1. Splitting based exponential methods in the fluid regime

In this section we present the derivation of exponential schemes combined with the operator splitting (2.57)-(2.58). A remarkable feature of the schemes is that they can be designed to preserve all relevant physical properties of the system including non negativity of the solution and entropy inequality (Gabetta et al. 1997, Dimarco and Pareschi 2011).

We mention here that exponential methods for parabolic partial differential equations and highly oscillatory problems have a long tradition and have been extensively studied by several authors (see the recent review by Hochbruck and Ostermann (2010)).

AP splitting and problem reformulation

In the case of operator splitting methods (2.57)-(2.58) applied to (7.1) in the fluid limit $\alpha = 1$, it easy to see that if the scheme used in the collision step is AP then the whole scheme is AP. In a time interval $[0, \Delta t]$ the collision step $\mathcal{C}^{\varepsilon}_{\Delta t}$ now reads

$$\begin{aligned} &\frac{\partial f}{\partial t} = \frac{1}{\varepsilon} Q(f, f), \\ &f(x, v, 0) = f_0(x, v), \end{aligned} \tag{7.2}$$

where $Q(f, f)$ is given by (2.14). As $\varepsilon \to 0$ (7.2) formally yields the algebraic equation $Q(f, f) = 0$ which, thanks to the conservation properties (2.27) of Q, can be solved as a function of the initial data to get $\mathcal{C}^{0}_{\Delta t}(f_0) = M[f_0]$. Coupling this projection with the transport step (2.58) originates a so-called kinetic scheme (Coron and Perthame 1991, Godlewski and Raviart 1996) for the Euler equation (2.42) given by $\mathcal{T}_{\Delta t}(M[f_0])$. Analogous results hold true for the higher order splitting methods (2.59) and (2.60).

The starting point of exponential schemes is to rewrite the homogeneous equation (7.2) in the form

$$\frac{\partial f}{\partial t} = \frac{1}{\varepsilon}(P(f, f) - \mu f), \tag{7.3}$$

where $P(f, f) = Q(f, f) + \mu f$ and $\mu > 0$ is such that $P(f, f) \geq 0$. Typically μ is proportional to the density and is an estimate of the largest value taken

by the loss part of the collision operator (2.14)

$$\mu \geq \int_{\mathbb{R}^3 \times S^2} B(v, v_*, \omega) f(v_*)\, dv_*\, d\omega. \tag{7.4}$$

In the sequel we will assume that μ depends linearly on the density and therefore that $P(f, f)$ is a bilinear operator. By construction we have the following

$$\frac{1}{\mu} \int_{\mathbb{R}^3} P(f, f)(v)\varphi(v)\, dv = \int_{\mathbb{R}^3} f(v)\varphi(v)\, dv, \quad \varphi(v) = 1, v, |v|^2. \tag{7.5}$$

Thus $P(f, f)/\mu$ is a density function. Let us denote $f_1 = P(f, f)/\mu$, we can consider the following decomposition

$$f_1 = M + g, \tag{7.6}$$

where M is the Maxwellian with the same moments of f (and hence of f_1). Here we omit the explicit dependence of M on the solution f since, thanks to the conservation properties (2.27), it remains constant during the collision step. Moreover since f_1 and M share the same moments we have

$$\int_{\mathbb{R}^3} g(v)\varphi(v)\, dv = 0, \quad \varphi(v) = 1, v, |v|^2. \tag{7.7}$$

The homogeneous Boltzmann equation can be then written in the form

$$\frac{\partial f}{\partial t} = \frac{\mu}{\varepsilon} g + \frac{\mu}{\varepsilon}(M - f) = \frac{\mu}{\varepsilon}\left(\frac{P(f, f)}{\mu} - M\right) + \frac{\mu}{\varepsilon}(M - f). \tag{7.8}$$

The above system is equivalent to the penalization method introduced in Filbet and Jin (2010).

Exponential Runge-Kutta methods

This class of methods has been proposed in Dimarco and Pareschi (2011). Using the fact that M does not depend on time we can rewrite (7.8) as

$$\frac{\partial (f - M) e^{\mu t/\varepsilon}}{\partial t} = \frac{1}{\varepsilon}(P(f, f) - \mu M) e^{\mu t/\varepsilon}. \tag{7.9}$$

Starting from the above reformulation we consider the family of methods characterized by the stages

$$\begin{aligned} F^{(i)} &= e^{-c_i \lambda} f^n + \lambda \sum_{j=1}^{i-1} A_{ij}(\lambda) \left(\frac{P(F^{(j)}, F^{(j)})}{\mu} - M^n \right) \\ &+ \left(1 - e^{-c_i \lambda}\right) M^n, \qquad i = 1, \ldots, \nu \end{aligned} \tag{7.10}$$

and by the numerical solution

$$\begin{aligned} f^{n+1} &= e^{-\lambda} f^n + \lambda \sum_{i=1}^{\nu} W_i(\lambda) \left(\frac{P(F^{(i)}, F^{(i)})}{\mu} - M^n \right) \\ &+ \left(1 - e^{-\lambda}\right) M^n, \end{aligned} \tag{7.11}$$

where $\lambda = \mu \Delta t/\varepsilon$, Δt is the time step, $f^n = f(t^n)$, $M^n = M(t^n)$, $c_i \geq 0$, and the coefficients A_{ij} and the weights W_i are such that

$$A_{ij}(0) = a_{ij}, \quad W_i(0) = w_i, \quad i, j = 1, \ldots, \nu$$

with coefficients a_{ij} and weights w_i given by a standard explicit Runge-Kutta method called the underlying method. Different methods originate from the different choices of the underlying method. The most popular approaches are the Integrating Factor (IF) and the Exponential Time Differencing (ETD) methods (Maset and Zennaro 2009). Since M^n does not depend on time during the collision process to simplify notations in the sequel we will omit the index n.

For the so-called IF methods we have

$$\begin{aligned} A_{ij}(\lambda) &= a_{ij} e^{-(c_i - c_j)\lambda}, \quad i, j = 1, \ldots, \nu, \quad i > j \\ W_i(\lambda) &= w_i e^{-(1-c_i)\lambda}, \quad i = 1, \ldots, \nu. \end{aligned} \tag{7.12}$$

For such methods the order of accuracy is the same as the order of the underlying method.

We recall the main properties for an IF exponential scheme in the form (7.10)-(7.11). We refer to Dimarco and Pareschi (2011) for further details. Let us define

$$R(\lambda) = e^{-\lambda} + \sum_{k=0}^{\nu-1} \lambda^{k+1} \bar{w}(\lambda)^T \bar{A}(\lambda)^k \bar{E}(\lambda) \bar{e}, \tag{7.13}$$

where $\bar{A}(\lambda)$ is the $\nu \times \nu$ matrix of elements $|A_{ij}(\lambda)|$, $\bar{w}(\lambda)$ the $\nu \times 1$ vector of elements $|W_i(\lambda)|$, $\bar{e}$ the $\nu \times 1$ unit vector and $\bar{E}(\lambda) = \mathrm{diag}(e^{-c_1\lambda}, \ldots, e^{-c_\nu \lambda})$. We have

Theorem 7.1. If an explicit exponential Runge-Kutta method in the form (7.10)-(7.11) satisfies

$$\lim_{\lambda \to \infty} R(\lambda) = 0, \tag{7.14}$$

with $R(\lambda)$ given by (7.13) then it is asymptotic preserving.

Note that for an IF method we have

$$|A_{ij}(\lambda)| \leq |a_{ij}| e^{-(c_i - c_j)\lambda}, \quad |W_i(\lambda)| \leq |w_i| e^{-(1-c_i)\lambda},$$

thus we require

$$0 = c_1 \leq c_2 \ldots \leq c_\nu \leq 1, \tag{7.15}$$

in order for the above quantities to be bounded independently of λ.

It can also be proved that if the underlying Runge-Kutta method is a ν-stage explicit Runge-Kutta method of order ν, with nonnegative coefficients and weights satisfying (7.15), then the scheme is unconditionally stable and contractive (Dimarco and Pareschi 2010). As pointed out in Maset and Zennaro (2009) examples of such methods are well-known up to $\nu = 4$ and the classical RK method of order four is the sole method with four stages.

For practical applications it may be convenient to require that as $\lambda \to \infty$ the numerical solution f^{n+1} and each level $F^{(i)}$ of the IF method are projected towards the local Maxwellian. It is straightforward to verify that this stronger AP property is satisfied if we replace condition (7.15) by

$$0 = c_1 < c_2 < \ldots < c_\nu < 1. \tag{7.16}$$

An important result concerns the convexity property of IF schemes (Dimarco and Pareschi 2010).

Proposition 7.1. An explicit IF method is unconditionally positive and convex if the underlying Runge-Kutta method has nonnegative coefficients and weights satisfying

$$\sum_{j=1}^{i-1} a_{ij} {c_j}^k \leq \frac{c_i^k}{k+1}, \quad k = 0, 1, 2, \ldots, \quad i = 1, \ldots, \nu \tag{7.17}$$

$$\sum_{i=1}^{\nu} w_i c_i^k \leq \frac{1}{k+1}, \quad k = 0, 1, 2, \ldots, \tag{7.18}$$

In the above conditions we did not use the bilinearity of $P(f, f)$ which would lead to weaker constraints on a_{ij} and w_i. Examples of methods that satisfy convexity are the second order modified Euler method and the third order Heun method but not the classical fourth order Runge-Kutta scheme. Let us finally remark that the convexity property implies that the scheme preserves at a discrete level the entropy inequality (2.36).

Remark 7.1.

- The estimation of μ plays an important role in practical computations. An overestimate may lead to over relaxation of the distribution f towards the equilibrium. For example, choosing μ as the upper bound in (7.4)

$$\mu_p = \sup_v \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} B(v, v_*, \omega) f(v_*) \, dv_* \, d\omega, \tag{7.19}$$

clearly guarantees positivity of the resulting exponential method but leads to overestimate the true spectrum of the collision operator. Better estimates may be obtained taking the average collision frequency or, as suggested in Filbet and Jin (2010), an estimate of the spectral radius of the linearized operator Q around the Maxwellian M. In fact

$$Q(f,f) \approx Q(M,M) + \nabla Q(M,M)(M-f) = \nabla Q(M)(M-f),$$

where $\nabla Q(M,M)$ is the Frechet derivative of Q evaluated at M. For example one can take

$$\mu_s = \sup_v \left| \frac{Q(f,f)}{f-M} \right|. \tag{7.20}$$

For simplicity, we assumed μ constant during the time stepping. In general one can take $\mu = \mu(t)$ and rewrite the exponential methods for a time dependent μ (Dimarco and Pareschi 2011).

- The exponential integrators can be applied in conjunction with the micro-macro decomposition (5.46). Inserting $f = M[f] + g$ in (7.2) gives the reformulated space homogeneous equation (Lemou 2010)

$$\frac{\partial g}{\partial t} = \frac{1}{\varepsilon}\left[\mathcal{L}_M(g) + Q(g,g)\right], \tag{7.21}$$

where $\mathcal{L}_M(g) = Q(M[f],g) + Q(g,M[f])$. Now the AP requirement corresponds to the fact that $g \to 0$ as $\varepsilon \to \infty$.
The exponential Runge-Kutta scheme (7.10)-(7.11) can be written in the equivalent form as

$$\begin{aligned} G^{(i)} &= e^{-c_i\lambda} g^n + \frac{\Delta t}{\varepsilon} \sum_{j=1}^{i-1} a_{ij} e^{-(c_i-c_j)\lambda} \Big[\mathcal{L}_M(G^{(j)}) \\ &+ Q(G^{(j)}, G^{(j)}) + G^{(j)} \Big], \end{aligned} \tag{7.22}$$

$$\begin{aligned} g^{n+1} &= e^{-\lambda} g^n + \frac{\Delta t}{\varepsilon} \sum_{i=1}^{\nu} b_i e^{-(1-c_i)\lambda} \Big[\mathcal{L}_M(G^{(i)}) \\ &+ Q(G^{(i)}, G^{(i)}) + G^{(i)} \Big]. \end{aligned} \tag{7.23}$$

Alternatively one can apply the exponential methods directly to the form (7.21), where the exact flow of the linear part $\mathcal{L}_M(g)$, i.e. an operator exponential, is used in the construction of the schemes.

Time Relaxed methods

Another important class of exponential methods for the numerical approximation of the space homogeneous Boltzmann equation, the so-called Time Relaxed (TR) methods, was introduced in Gabetta et al. (1997). These

schemes were originally derived from the Wild sum expansion of the homogeneous Boltzmann equation (Wild 1951) together with a suitable Maxwellian truncation. Historically they represent the first asymptotic preserving methods for the full Boltzmann equation and have found several application in numerous contests, including Monte Carlo techniques (Pareschi and Caflisch 1999, Pareschi and Russo 2000*a*, Filbet and Russo 2003, Pareschi, Trazzi and Wennberg 2008).

Here we show that the schemes can be derived starting from a suitable Taylor expansion of (7.9). To this aim, let us first introduce the change of variables

$$\tau = 1 - \exp(-\mu t/\varepsilon),$$

which, using the bilinearity of $P(f, f)$, gives the equation

$$\frac{\partial}{\partial \tau}\left[(f - M)\frac{1}{1-\tau}\right] = (P(f, f) - \mu M)\frac{1}{\mu(1-\tau)^2}. \tag{7.24}$$

By taking the Taylor expansion of $(f - M)/(1 - \tau)$ around $\tau = 0$ in (7.24) we get

$$(f - M)/(1 - \tau) = (f_0 - M) + \tau\left[\frac{P(f_0, f_0)}{\mu} - M\right] +$$
$$+\frac{\tau^2}{2}\left[\frac{P(P(f_0, f_0), f_0) + P(f_0, P(f_0, f_0))}{\mu^2} - 2M\right] + O(\tau^3)$$

where we have used the bilinearity of the operator $P(f, f)$.

If we compute all the terms in the expansion and use recursively the bilinearity of $P(f, f)$ we can state the following

Proposition 7.2. The solution to problem (7.24) can be represented as

$$f(v, t) = (1 - \tau)f_0(v) + (1 - \tau)\sum_{k=1}^{\infty} \tau^k (f_k^n(v) - M(v)) + \tau M(v), \tag{7.25}$$

where f_0 is the initial data and the functions f_k are given by the recurrence formula

$$f_{k+1}(v) = \frac{1}{k+1}\sum_{h=0}^{k}\frac{1}{\mu}P(f_h, f_{k-h})(v), \quad k = 0, 1, \ldots. \tag{7.26}$$

By truncating expansion (7.25) at the order m, and reverting to the old variables in a time interval Δt, we recover the TR schemes presented in (Gabetta et al. 1997)

$$f^{n+1} = e^{-\lambda} f^n + e^{-\lambda}\sum_{k=1}^{m}(1 - e^{-\lambda})^k (f_k^n - M) + (1 - e^{-\lambda})M,$$

which, using the fact that

$$1 - e^{-\lambda} \sum_{k=0}^{m} (1 - e^{-\lambda})^k = (1 - e^{-\lambda})^{m+1},$$

can be rewritten in the usual form emphasizing their convexity properties

$$f^{n+1} = e^{-\lambda} \sum_{k=0}^{m} (1 - e^{-\lambda})^k f_k^n + (1 - e^{-\lambda})^{m+1} M. \tag{7.27}$$

The above class of schemes satisfy the following (Gabetta et al. 1997)

Theorem 7.2. The time discretization defined by (7.27) is such that

i) If $\sup_{k>m}\{|f_k - M|\} \le C$ for a constant $C = C(v)$ then it is at least a m-order approximation in λ of (7.25) with $f_0 = f^n$ and

$$|f(v, \Delta t) - f^{n+1}(v)| \le C\,(1 - e^{-\lambda})^{m+1}.$$

ii) The coefficients f_k^n satisfy

$$\int_{\mathbb{R}^3} f_k^n \varphi \, dv = \int_{\mathbb{R}^3} f^n \varphi \, dv, \qquad \varphi = 1, v, |v|^2.$$

iii) The solution f^{n+1} is a convex combination of the coefficients f_k^n independently of λ. Moreover, if P is a nonnegative operator then all f_k^n and f^{n+1} are nonnegative densities independently of λ.

iv) For any $m, n \ge 0$ we have

$$\lim_{\lambda \to 0} f^{n+1}(v) = M(v).$$

Moreover, with the same assumptions of i), the following holds

$$|f^{n+1}(v) - M| \le C\,[1 - (1 - e^{-\lambda})^{m+1}].$$

Remark 7.2. Thanks to the nonnegativity and convexity properties of exponential methods in Pareschi and Russo (2000*a*), Pareschi and Russo (2001) a class of asymptotic preserving Monte Carlo methods has been constructed. To illustrate the idea let us consider a general first order exponential scheme in the form

$$f^{n+1} = A_0 f^n + A_1 f_1^n + A_2 M \tag{7.28}$$

where $f_1^n = P(f^n, f^n)/\mu$, $A_0 = e^{-\lambda}$, $A_2 = 1 - A_0 - A_1$, and $A_1 = e^{-\lambda}(1 - e^{\lambda})$ for the TR scheme or $A_1 = \lambda e^{\lambda}$ for the IF scheme.

The probabilistic interpretation of the above equation is given below. Given a set of samples from f^n to generate a sample from f^{n+1} do the following:

- with probability A_0 take a sample from f^n;

- with probability A_1 extract a sample from f_1^n (this is equivalent to apply the usual DSMC collision algorithm between samples (Bird 1994));
- with probability A_2 extract a sample from the Maxwellian M.

In this formulation the probabilistic interpretation holds uniformly in λ, at variance with standard DSMC, which requires $\lambda \leq 1$. Furthermore, as $\lambda \to \infty$, the distribution at time $n+1$ is sampled from a Maxwellian. In a space non homogeneous case, this would be equivalent to the particle method for the Euler equations proposed by Pullin (1980).

7.2. High order exponential Runge-Kutta methods

In this paragraph we present the basic framework for the derivation of AP exponential Runge-Kutta methods without time splitting (Li and Pareschi 2014). See also Lemou (2010) for a related approach based on a suitable exponential integration of the space non homogeneous Boltzmann equation. The major difficulty is related to the time dependent nature of the local Maxwellian equilibrium that does not allow a direct application of the methods developed for the space homogeneous equation. On the other hand, since splitting methods suffer from order reduction in the fluid limit (see Remark 7.3), it represents an important step for the derivation of uniformly accurate high order methods.

We reformulate, similarly to the previous paragraph the complete Boltzmann equation (7.1) as

$$\begin{aligned}\frac{\partial}{\partial t}\left[(f-\tilde{M})e^{\mu t/\varepsilon}\right] &= \frac{\partial}{\partial t}(f-\tilde{M})e^{\mu t/\varepsilon} + (f-\tilde{M})\frac{\mu}{\varepsilon}e^{\mu t/\varepsilon} \\ &= \left[\frac{1}{\varepsilon}(Q+\mu f-\mu\tilde{M}) - \frac{\partial\tilde{M}}{\partial t} - v\cdot\nabla_x f\right]e^{\mu t/\varepsilon} \\ &= \left[\frac{1}{\varepsilon}(P-\mu\tilde{M}) - \frac{\partial\tilde{M}}{\partial t} - v\cdot\nabla_x f\right]e^{\mu t/\varepsilon}. \qquad (7.29)\end{aligned}$$

Here, μ is independent of time and $\tilde{M}$ could be any arbitrary non-negative function. The main problem in constructing a numerical method based on (7.29) is to select the right $\tilde{M}$ and μ to meet stability and monotonicity requirement.

The first possibility is to assume $\tilde{M}$ as a time independent function given a-priori. Therefore, $\partial\tilde{M}/\partial t$ cancels and (7.29) becomes

$$\frac{\partial}{\partial t}\left[(f-\tilde{M})e^{\mu t/\varepsilon}\right] = \left[\frac{1}{\varepsilon}(P-\mu\tilde{M}) - v\cdot\nabla_x f\right]e^{\mu t/\varepsilon}. \qquad (7.30)$$

A direct application of a standard explicit Runge-Kutta method to (7.30)

yields then to the following scheme

$$
\begin{aligned}
(f^{(i)} - \tilde{M})e^{c_i\lambda} &= (f^n - \tilde{M}) + \sum_{j=1}^{i-1} a_{ij}\frac{\Delta t}{\varepsilon}\left[P^{(j)} - \mu\tilde{M} - \varepsilon v\cdot\nabla_x f^{(j)}\right]e^{c_j\lambda},\\
(f^{n+1} - \tilde{M})e^{\lambda} &= (f^n - \tilde{M}) + \sum_{i=1}^{\nu} w_i\frac{\Delta t}{\varepsilon}\left[P^{(i)} - \mu\tilde{M} - \varepsilon v\cdot\nabla_x f^{(i)}\right]e^{c_i\lambda},
\end{aligned}
\tag{7.31}
$$

where i indicates the stages, $\lambda = \mu\Delta t/\varepsilon$, and $P^{(j)} = P(f^{(j)}, f^{(j)})$. Since in the fluid regime the distribution function f^{n+1} should be projected to the Maxwellian function M^{n+1} whose macroscopic quantities satisfy the limiting Euler equation, the simplest choice to obtain an AP scheme is to take $\tilde{M} = M_E^{n+1}$, where M_E^{n+1} is the local Maxwellian computed from the macroscopic quantities satisfying the Euler system (2.42). In order to do that one can compute the limiting Euler equation using the same explicit Runge-Kutta scheme used for the kinetic equation and then, with these moments, defines $\tilde{M}$. Note that as $\varepsilon \to 0$, then f^{n+1} and $\tilde{M}$ share the same moments, so the scheme is AP. However, M_E^{n+1} does not share moments with f^{n+1} unless $\varepsilon = 0$ so that, with this choice, the scheme may be inaccurate in regimes when $\Delta t \sim \varepsilon$.

We discuss now the case $\tilde{M} = M[f]$, the time dependent local Maxwellian that shares mass, momentum and energy with f. We reformulate the Boltzmann equation as

$$
\frac{\partial}{\partial t}\left[(f - M[f])\,e^{\mu t/\varepsilon}\right] = \left(\frac{P - \mu M[f]}{\varepsilon} - v\cdot\nabla_x f - \frac{\partial M[f]}{\partial t}\right)e^{\mu t/\varepsilon}. \tag{7.32}
$$

What we need now is to define the Maxwellian distribution for each stage of the numerical scheme. This means we need the moments of the distribution function at each level of the Runge-Kutta procedure. We then numerically solve using an explicit Runge-Kutta method the following coupled system of equations

$$
\begin{aligned}
\frac{\partial}{\partial t}(f - M[f])e^{\mu t/\varepsilon} &= \frac{1}{\varepsilon}\left(P - \mu M[f] - \varepsilon v\cdot\nabla_x f - \varepsilon\frac{\partial M[f]}{\partial t}\right)e^{\mu t/\varepsilon},\\
\frac{\partial}{\partial t}\int_{\mathbb{R}^3}\varphi f dv &= -\int_{\mathbb{R}^3}\varphi v\cdot\nabla_x f dv, \quad \varphi(v) = 1, v, |v|^2
\end{aligned}
\tag{7.33}
$$

where the second equation corresponds to the time evolution of the moments

of the distribution f. Thus we have the following scheme for the stages

$$\begin{aligned}
&(f^{(i)} - M[f^{(i)}])e^{c_i\lambda} = (f^n - M[f^n]) + \\
&\quad \sum_{j=1}^{i-1} a_{ij}\frac{\Delta t}{\varepsilon}\left[P^{(j)} - \mu M[f^{(j)}] - \varepsilon v\cdot\nabla_x f^{(j)} - \varepsilon\frac{\partial M[f^{(j)}]}{\partial t}\right]e^{c_j\lambda}, \\
&\int_{\mathbb{R}^3}\varphi f^{(i)}dv = \int_{\mathbb{R}^3}\varphi f^n dv + \sum_{j=1}^{i-1} a_{ij}\left(-\Delta t\int_{\mathbb{R}^3}\varphi v\cdot\nabla_x f^{(j)}dv\right);
\end{aligned} \tag{7.34}$$

and for the numerical solution

$$\begin{aligned}
&(f^{n+1} - M[f^{n+1}])e^{\lambda} = (f^n - M[f^n]) + \\
&\quad \sum_{i=1}^{\nu} w_i\frac{\Delta t}{\varepsilon}\left[P^{(i)} - \mu M[f^{(i)}] - \varepsilon v\cdot\nabla_x f^{(i)} - \varepsilon\frac{\partial M[f^{(i)}]}{\partial t}\right]e^{c_i\lambda}, \\
&\int_{\mathbb{R}^3}\varphi f^{n+1}dv = \int_{\mathbb{R}^3}\varphi f^n dv + \sum_{i=1}^{\nu} w_i\left(-\Delta t\int_{\mathbb{R}^3}\varphi v\cdot\nabla_x f^{(i)}dv\right).
\end{aligned} \tag{7.35}$$

The two equations, the one for the kinetic equation and the one for the moments, are intrinsically coupled, as to evaluate $f^{(i)}$, one needs to compute $M[f^{(i)}]$, whose macroscopic quantities are obtained in the second equation, and $\partial M[f]^{(j)}/\partial t$ for all $j < i$. The time derivative of the Maxwellian is computed as following, as $M[f]$ only depends on ρ, u and T, one has

$$\frac{\partial M[f]}{\partial t} = \frac{\partial M[f]}{\partial\rho}\frac{\partial\rho}{\partial t} + \nabla_u M[f]\cdot\frac{\partial u}{\partial t} + \frac{\partial M[f]}{\partial T}\frac{\partial T}{\partial t}, \tag{7.36}$$

where

$$\begin{aligned}
&\frac{\partial M[f]}{\partial\rho} = \frac{M[f]}{\rho}, \quad \nabla_u M[f] = M[f]\frac{v-u}{RT}, \\
&\frac{\partial M[f]}{\partial T} = \frac{M[f]}{2}\left(\frac{|v-u|^2}{RT^2} - \frac{3}{2\pi RT}\right),
\end{aligned} \tag{7.37}$$

and $\partial\rho/\partial t$, $\partial u/\partial t$ and $\partial T/\partial t$ could be evaluated numerically taking the first three moments of the distribution function

$$\frac{\partial}{\partial t}\begin{pmatrix}\rho\\ \rho u\\ \dfrac{3\rho RT}{2} + \dfrac{1}{2}\rho u^2\end{pmatrix} = -\int_{\mathbb{R}^3}\begin{pmatrix}1\\ v\\ \dfrac{|v|^2}{2}\end{pmatrix} v\cdot\nabla_x f dv. \tag{7.38}$$

The asymptotic preservation property for this class of schemes is stated in the following

Theorem 7.3. The class of exponential Runge-Kutta methods defined by (7.31) or by (7.34)-(7.35) are AP for a general explicit Runge-Kutta methods with $0 \le c_1 \le c_2 \le \cdots \le c_\nu < 1$.

In addition, using the Shu-Osher representation of Runge-Kutta methods (Shu and Osher 1989) it is possible to prove that the schemes defined by (7.31), under a suitable CLF condition, preserve non negativity of the numerical solution. We refer to Li and Pareschi (2014) for further details.

Remark 7.3.

- The main feature of the AP exponential methods just described is the possibility to achieve very high order uniformly in time. In contrast, AP methods based on splitting suffer of order reduction for small values of ε. This can be understood by observing that the collision step (7.2) in such singular limit becomes independent of the time step since it reduces to a projection over the local equilibrium

$$\lim_{\varepsilon \to 0} \mathcal{C}^{\varepsilon}_{\Delta t}(f_0) = M[f_0], \quad \forall\ \Delta t > 0.$$

- As for the exponential splitting method, the high order scheme (7.34)-(7.35) admits a natural extension to other Boltzmann-type equations where the inversion of the collision operator is difficult. In the case of the quantum Boltzmann equation this has been done in Li, Hu and Pareschi (2014).

A convergence rate test

The following test is reported to illustrate the convergence rates for the non splitting exponential schemes applied to the full Boltzmann equation. The collision operator is solved with the fast spectral method of Section 6.1 and a third order WENO discretization is used for the space derivatives. The four schemes for which we report the rates of convergence are denoted as ExpRK2-F, ExpRK2-V, ExpRK3-F and ExpRK3-V. They correspond to the second order Runge-Kutta and the third order Heun method (Hairer and Wanner 1996). The letters F and V indicate, respectively, the scheme with a fixed $\tilde{M} = M_E$ given by the solution of the Euler equations and the scheme with a time variable $\tilde{M} = M[f]$.

The initial data is

$$\begin{aligned}
&\rho_0(x) = \frac{1}{2}\left(2 + \sin\left(2\pi x\right)\right), \\
&u_1(x) = [0.75, -0.75]^T, \qquad u_2(x) = [-0.75, 0.75]^T, \\
&T_0(x) = \frac{1}{20}\left(5 + 2\cos\left(2\pi x\right)\right).
\end{aligned}$$

The L_1 norm of the error for the density for different values of the Knudsen number, i.e. $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-6}$ is reported. Both equilibrium

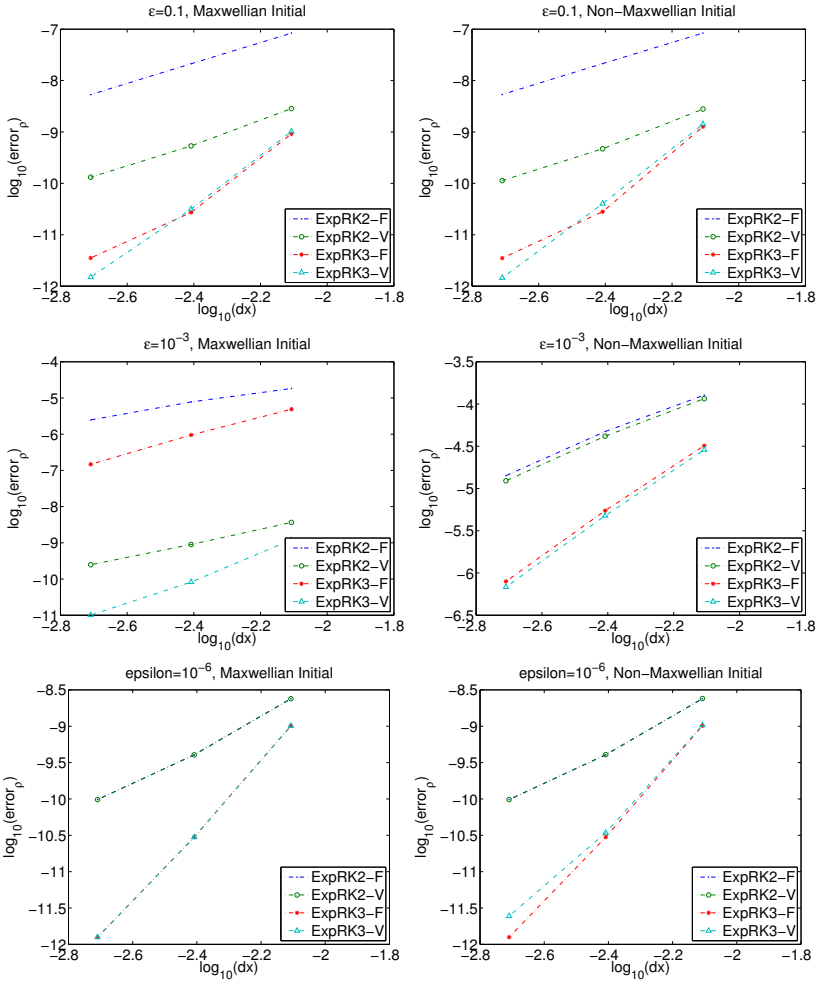

Figure 7.2. L_1 error for the density ρ for different exponential schemes. From the top to the bottom, results show $\varepsilon = 0.1$, 10^{-3}, 10^{-6} respectively. Left column equilibrium initial data, right column non equilibrium initial data.

initial data $f_0(x,v) = M[f_0]$ and non equilibrium initial data

$$f(t=0,x,v) = \frac{\rho_0(x)}{2}\left(e^{-\frac{|v-u_1(x)|^2}{T_0(x)}} + e^{\frac{|v-u_2(x)|^2}{T_0(x)}}\right) \tag{7.39}$$

are considered. Here $x \in [0,1]$ and periodic boundary conditions on x are applied. We use 32×32 grid points in velocity and $N_x = 128$, 256, 512, 1024 grid points in space. Time stepping Δt is chosen to satisfy the CFL transport condition with CFL number fixed to 0.5. One can observe in Figure 7.2 that, as expected, in the kinetic and intermediate regimes, $\varepsilon = 0.1$ and $\varepsilon = 10^{-3}$, ExpRK-V generally performs better than ExpRK-F.

On the other hand, when $\varepsilon = 10^{-6}$, i.e. in the hydrodynamic regime both methods achieve high order accuracy.

7.3. Implicit-Explicit Runge-Kutta schemes in the fluid regime

Another class of Asymptotic Preserving schemes is based on the use of Implici-Explicit (IMEX) Runge-Kutta methods. Such schemes were developed originally in Ascher, Ruuth and Spiteri (1997) for parabolic partial differential equations and later extended to hyperbolic system with relaxation in Pareschi and Russo (2005). Early examples of such schemes were developed in Jin (1995) and Caflisch et al. (1997). Only recently, they have been designed to achieve asymptotic preservation for the Bolztmann equation in the fluid limit without requiring the inversion of the collision operator (Filbet and Jin 2010, Dimarco and Pareschi 2012, Dimarco and Pareschi 2013).

IMEX Runge-Kutta schemes

A standard IMEX Runge-Kutta method applied to a kinetic equation of the type (7.1) for $\alpha = 1$, reads

$$F^{(i)} = f^n - \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} v \cdot \nabla_x F^{(j)} + \Delta t \sum_{j=1}^{\nu} a_{ij} \frac{1}{\varepsilon} Q(F^{(j)}, F^{(j)}) \quad (7.40)$$

$$f^{n+1} = f^n - \Delta t \sum_{i=1}^{\nu} \tilde{w}_i v \cdot \nabla_x F^{(i)} + \Delta t \sum_{i=1}^{\nu} w_i \frac{1}{\varepsilon} Q(F^{(i)}, F^{(i)}), \quad (7.41)$$

where f^{n+1} represents the numerical solution and $F^{(i)}$ the stage values. The matrices $\tilde{A} = (\tilde{a}_{ij})$, $\tilde{a}_{ij} = 0$ for $j \geq i$ and $A = (a_{ij})$ are $\nu \times \nu$ matrices such that the resulting scheme is explicit in $v \cdot \nabla_x f$, and implicit in $Q(f, f)$. We restrict to diagonally implicit Runge-Kutta (DIRK) schemes for the collision operator ($a_{ij} = 0$, for $j > i$) and observe that this ensures that the transport term $v \cdot \nabla_x f$ is always evaluated explicitly. Using the vector notations the schemes can be written in compact form

$$F = f^n e + \Delta t \tilde{A}\, L(F) + \frac{\Delta t}{\varepsilon} A Q(F) \quad (7.42)$$

$$f^{n+1} = f^n + \Delta t \tilde{w}^T L(F) + \frac{\Delta t}{\varepsilon} w^T Q(F), \quad (7.43)$$

where $e = (1, .., 1)^T \in \mathbb{R}^\nu$, $F = (F^{(1)}, \ldots, F^{(\nu)})^T$, $Q(F) = (Q(F^{(1)}, F^{(1)}), \ldots, Q(F^{(\nu)}, F^{(\nu)}))^T$ and $L(F) = (L(F^{(1)}), \ldots, L(F^{(\nu)}))$ with $L(F^{(i)}) = -v \cdot \nabla_x F^{(i)}$. In addition, suitable order conditions must be satisfied by the coefficients of the Runge-Kutta schemes. We refer to Ascher et al. (1997), Kennedy and Carpenter (2003) and Pareschi and Russo (2005) for details. Here we

just point out that, in general, additional coupling conditions for the two Runge-Kutta methods must be satisfied.

It is useful to characterize the different IMEX schemes developed in the literature accordingly to the structure of the DIRK method (Boscarino et al. 2013).

Definition 7.2. We call an IMEX-RK method of *type A* if the matrix $A \in \mathbb{R}^{\nu\times\nu}$ is invertible, or equivalently $a_{ii} \neq 0$, $i = 1, \dots, \nu$. We call an IMEX-RK method of *type CK* if the matrix A can be written as

$$A = \begin{pmatrix} 0 & 0 \\ a & \hat{A} \end{pmatrix}, \tag{7.44}$$

with $a = (a_{21}, \dots, a_{\nu 1})^T \in \mathbb{R}^{(\nu-1)}$ and the submatrix $\hat{A} \in \mathbb{R}^{(\nu-1)\times(\nu-1)}$ invertible, or equivalently $a_{ii} \neq 0$, $i = 2, \dots, \nu$. In the special case $a = 0$, $w_1 = 0$ the scheme is said to be of *type ARS*.

Finally, we recall the following definition that will be used in the sequel.

Definition 7.3. We call an IMEX-RK method *globally stiffly accurate (GSA)* if the corresponding DIRK method is stiffly accurate, namely

$$a_{\nu i} = w_i, \quad i = 1, \dots, \nu, \tag{7.45}$$

and in addition the explicit method satisfies

$$\tilde{a}_{\nu i} = \tilde{w}_i, \quad i = 1, \dots, \nu. \tag{7.46}$$

Note that for GSA methods the numerical solution coincides with the last stage value of the method.

Keeping these definitions in mind, we summarize the results in Dimarco and Pareschi (2013) concerning the AP property of the different IMEX schemes.

Theorem 7.4. (AP-type A) If the IMEX method is of type A then in the limit $\varepsilon \to 0$, scheme (7.42)-(7.43) becomes the explicit RK scheme characterized by $(\tilde{A}, \tilde{w}, \tilde{c})$ applied to the limit Euler system (2.43). Moreover if the scheme satisfies the GSA property we have

$$\lim_{\varepsilon\to 0} f^{n+1} = M[f^{n+1}]. \tag{7.47}$$

The first part of the result is an immediate consequence of the fact that, as $\varepsilon \to 0$ in (7.42), we get $AQ(F) = 0$ which, since A is invertible, implies $Q(F) = 0$ and hence $F = M$. Plugging this into the numerical solution (7.43) yields the desired result.

After a little algebra one observes that the additional property (7.47) is achieved if the following conditions are satisfied

$$w^T A^{-1} e = 1, \qquad \tilde{w}^T = w^T A^{-1} \tilde{A}, \qquad w^T A^{-1} M[F] = M[f^{n+1}]. \tag{7.48}$$

The first condition corresponds to the classical L-stability requirement of the DIRK method (Hairer and Wanner 1996). Since the third condition depends on the stage values vector F the only possibility to satisfy (7.48) is that the IMEX scheme is GSA. In this case, in fact, we have

$$w^T A^{-1} = (0, \dots, 0, 1)^T, \quad M[F^\nu] = M[f^{n+1}].$$

The request that the matrix A is invertible can be quite restrictive for high order methods. However, under additional hypothesis, one can obtain schemes which are asymptotic preserving for CK matrices. In order to do this, we first introduce the notion of initial data consistent with the limit problem.

Definition 7.4. The initial data for equation (7.1) are said *consistent* or *well prepared* if

$$f_0(x, v) = M[f_0(x, v)] + g^\varepsilon(x, v), \qquad \lim_{\varepsilon \to 0} g^\varepsilon(x, v) = 0. \tag{7.49}$$

It is possible to prove the following

Theorem 7.5. (AP-type CK) If the IMEX scheme is of type CK and GSA then for consistent initial data, in the limit $\varepsilon \to 0$, the scheme (7.42)-(7.43) becomes the explicit RK scheme characterized by $(\tilde{A}, \tilde{w}, \tilde{c})$ applied to the limit Euler system (2.43). Moreover if one of the following conditions is satisfied

(a) the initial data is consistent;

(b)

$$\hat{e}_\nu^T \hat{A}^{-1} a = 0, \tag{7.50}$$

where $\hat{e}_\nu = (0, \dots, 0, 1)^T \in \mathbb{R}^{\nu-1}$,

then

$$\lim_{\varepsilon \to 0} f^{n+1} = M[f^{n+1}], \tag{7.51}$$

The proof is similar to the one for type A schemes, except that one has to work with the invertible submatrix $\hat{A}$. One can then use the fact that, since the initial data is consistent, $F^{(1)} = M^{(1)}$ and that, thanks to the GSA property, this projection is maintained at subsequent time levels. We refer to Dimarco and Pareschi (2013) for the details.

Remark 7.4. If we restrict to the particular case where the collision term is given by a BGK relaxation operator $Q(f, f) = \nu(M[f] - f)$, a fundamental property of equations (7.42)-(7.43) is that they can be solved explicitly. In fact, since the implicit scheme is a DIRK method, the stage values take the

form

$$F^{(i)} = f^n + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} L(F^{(j)}) + \Delta t \sum_{j=1}^{i-1} a_{ij} \frac{\nu^{(j)}}{\varepsilon} (M[F^{(j)}] - F^{(j)}) + a_{ii} \Delta t \frac{\nu^{(i)}}{\varepsilon} (M[F^{(i)}] - F^{(i)}) \tag{7.52}$$

where the only implicit term is the diagonal factor $\nu^{(i)}(M[F^{(i)}] - F^{(i)})$ in which $M[F^{(i)}]$ and $\nu^{(i)}$ depend only on the moments of $F^{(i)}$. If we now integrate the above equation against the collision invariants $\varphi = 1, v, |v|^2$, we obtain the explicit moment scheme

$$\int_{\mathbb{R}^3} \varphi F^{(i)}\, dv = \int_{\mathbb{R}^3} \varphi f^n\, dv + \Delta t \sum_{j=1}^{i-1} \tilde{a}_{ij} \int_{\mathbb{R}^3} \varphi L(F^{(j)})\, dv. \tag{7.53}$$

Thus $M[F^{(i)}]$ and $\nu^{(i)}$, can be computed directly from the moments of $F^{(i)}$ and system (7.52) is explicitly solvable.

Penalized IMEX Runge-Kutta schemes

The IMEX methods just described provide high order AP schemes for collision operator that are easy to invert, so that the solution of the implicit term can be performed efficiently. This is the case, for example, of the BGK model where it can be solved explicitly. For more general collision operators, like in the Boltzmann case, one can use the penalty method proposed in Filbet and Jin (2010) to construct a class of AP IMEX methods. The idea has been already used in the derivation of the exponential schemes for the homogeneous equation, but let us illustrate it here in the general case.

We denote with $Q_P(f)$ an arbitrary collision operator, easy to invert, possessing the same physical properties of the Boltzmann integral so that $Q_P(f) = 0$ implies $f = M[f]$. Next we rewrite the collision operator in the form

$$Q(f, f) = (Q(f, f) - Q_P(f)) + Q_P(f) = G_P(f) + Q_P(f), \tag{7.54}$$

where by construction $\int_{\mathbb{R}^3} G_P(f)\varphi\, dv = 0$, $\varphi = 1, v, |v|^2$, and the corresponding kinetic equation reads

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\varepsilon} G_P(f) + \frac{1}{\varepsilon} Q_P(f). \tag{7.55}$$

The general class of penalized IMEX Runge-Kutta schemes for the Boltzmann equation now reads

$$F = f^n e + \Delta t \tilde{A} \left(\frac{1}{\varepsilon} G_P(F) + L(F) \right) + \Delta t A \frac{1}{\varepsilon} Q_P(F) \tag{7.56}$$

$$f^{n+1} = f^n + \Delta t \tilde{w}^T \left(\frac{1}{\varepsilon} G_P(F) + L(F) \right) + \Delta t w^T \frac{1}{\varepsilon} Q_P(F), \quad (7.57)$$

where $G_P(F) = (G_P(F^{(1)}), \dots, G_P(F^{(\nu)}))^T$.

We resume now the main results concerning the AP properties of penalized IMEX schemes (Dimarco and Pareschi 2013).

Theorem 7.6. (AP-type A penalized) If the penalized IMEX method is of type A and satisfies

$$\tilde{w}^T = w^T A^{-1} \tilde{A}, \quad (7.58)$$

then in the limit $\varepsilon \to 0$, scheme (7.56)-(7.57) becomes the explicit RK scheme characterized by $(\tilde{A}, \tilde{w}, \tilde{c})$ applied to the limit Euler system (2.43). Moreover if the penalized IMEX satisfied the GSA property then

$$\lim_{\varepsilon \to 0} f^{n+1} = M[f^{n+1}]. \quad (7.59)$$

Note that condition (7.58) is automatically satisfied if the IMEX scheme is GSA.

We consider now the case of penalized IMEX schemes of type CK. We can state an analogous result of Theorem 7.5 for standard IMEX schemes of type CK.

Theorem 7.7. (AP-type CK penalized) If the penalized IMEX scheme is of type CK and GSA then for consistent initial data in the limit $\varepsilon \to 0$ scheme (7.56)-(7.57) becomes the explicit RK scheme characterized by $(\tilde{A}, \tilde{w}, \tilde{c})$ applied to the limit Euler system (2.43). Moreover if one of the following conditions is satisfied

(a) the initial data is consistent;

(b)
$$\hat{e}_\nu^T \hat{A}^{-1} \hat{\tilde{A}} = 0, \qquad \hat{e}_\nu^T \hat{A}^{-1} \tilde{a} = 0, \qquad \hat{e}_\nu^T \hat{A}^{-1} a = 0, \quad (7.60)$$
where $\hat{e}_\nu = (0, \dots, 0, 1)^T \in \mathbb{R}^{\nu-1}$;

(c)
$$\hat{A}^{-1} \tilde{a} = 0, \qquad \hat{A}^{-1} a = 0, \quad (7.61)$$

then

$$\lim_{\varepsilon \to 0} f^{n+1} = M[f^{n+1}]. \quad (7.62)$$

Remark 7.5.

- The general penalization approach described in (7.54) can be applied to any collision term $Q(f)$ that can be efficiently penalized through a suitable, easy to invert, operator sharing the same asymptotic behavior and conservation properties. The simplest choice of penalizing operator satisfying the above requirements is the BGK-like operator

$Q_P(f) = \mu(M[f] - f)$, where $\mu > 0$ is a suitable constant, used in the development of exponential schemes. Other possibilities are given by the ES-BGK relaxation (Holway 1966), which has the advantage of matching the $O(\varepsilon)$ expansion, i.e. the compressible Navier-Stokes system, or the linearized Boltzmann operator $Q_P(f) = Q(f, M[f])$.

- A simple BGK-like penalization suffices to give the correct behavior also for the quantum Boltzmann equation (Filbet et al. 2012) and of the multi species Boltzmann equation (Jin and Li 2013). However, beside the question of the choice of the optimal operator for the penalization, there are cases in which the simple BGK operator is not suitable as a penalization any more. Most noticeably the Landau operator (2.51) and the inelastic Boltzmann operator (5.57).
  For the Landau equation one has the additional difficulty of the diffusive nature of the operator $Q_L(f, f)$, which introduces a parabolic stiffness relating the time step to the square of the velocity mesh. In Jin and Yan (2011) using the following Fokker-Planck operator as the penalty operator for the Landau term

$$Q_P(f) = \nabla_v \cdot \left(M[f] \nabla_v \left(\frac{f}{M[f]} \right) \right), \tag{7.63}$$

  an asymptotic preserving scheme has been derived. Note that, the use of a diffusive penalization term is essential in removing the parabolic stiffness.
  For the inelastic Boltzmann equation the steady states are Dirac delta distributions and it is not immediate to identify a simplified operator that can be used as a penalization term. Asymptotic preservation is achieved, for example, using the BGK-like model with friction proposed in Astillero and Santos (2004)

$$Q_P(f) = \mu(M[f] - f) + \beta_e \nabla_v \cdot [(v - u)f], \tag{7.64}$$

  where $\beta_e \geq 0$ is a suitable constant depending of the restitution coefficient e. Alternatively in Filbet and Rey (2013) a standard BGK operator has been used in the rescaled setting described in Section 5.2 by equations (5.61)-(5.62).

Convergence Rate for different IMEX schemes

Several examples of IMEX schemes satisfying the GSA property have been developed in the literature. We mention the following methods: ARS$(2, 2, 2)$ and ARS$(4, 4, 3)$ from Sections 2.6 and 2.8 in Ascher et al. (1997), JF-CK$(2, 3, 2)$ from (2.8) Section 2 in Filbet and Jin (2010), BPR-CK$(3, 5, 3)$ from the Appendix in Boscarino et al. (2013). Schemes DP1-A$(1, 2, 1)$ and DP2-A$(2, 4, 2)$ from Dimarco and Pareschi (2013). In the above list, we used

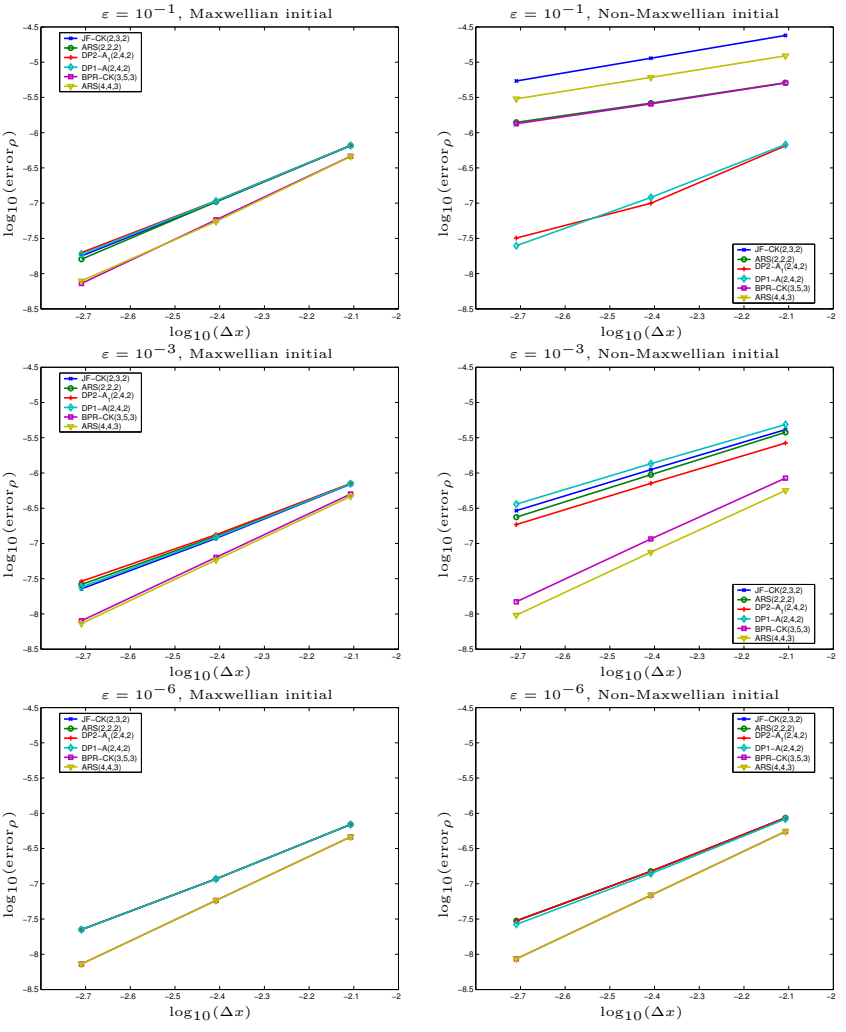

Figure 7.3. L_1 error for the density ρ for different second and third order IMEX schemes. Left column equilibrium initial data, right column non equilibrium initial data. Top $\varepsilon = 10^{-1}$, center $\varepsilon = 10^{-3}$, bottom $\varepsilon = 10^{-6}$.

the notation NAME(ν_E, ν_I, p) where ν_E, ν_I are, respectively, the number of function evaluations of the explicit and the implicit methods and p is the combined order of the IMEX scheme. The field NAME of the schemes is composed by the initials of the authors and the scheme type. We emphasize that the computational cost of penalized IMEX schemes is characterized by the number of stages of the explicit method since, by construction, the implicit part is applied to the easy invertible term used for penalization.

The test is performed on $(x, v) \in [0, 1] \times [-v_{\max}, v_{\max}]^2$, with $v_{\max} = 8$. A 3rd order WENO scheme for the space discretization and a fast spectral

method for solving the collision integral were employed. The number of grid points in each velocity direction is $N_v = 32$. The time step is $\Delta t = \Delta x/(2v_{\max})$. The initial data is

$$\rho_0(x) = \frac{2+\sin(2\pi x)}{3},\ u_0(x) = \frac{\cos(2\pi x)}{5},\ T_0(x) = \frac{3+\cos(2\pi x)}{4}. \quad (7.65)$$

The L_1 norm of the error for the density for different values of the Knudsen number, i.e. $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-6}$ is reported in Figure 7.3. Both equilibrium initial data $f_0(x,v) = M[f_0]$ and non equilibrium initial data

$$f_0(x,v) = \frac{\rho_0(x)}{(2\pi T_0(x))^{1/2}}\frac{1}{2}\left(\exp^{-\frac{|v-u_0(x)|^2}{2T_0(x)}} + \exp^{-\frac{|v+3u_0(x)|^2}{2T_0(x)}}\right), \quad (7.66)$$

are considered. As expected, all the schemes exhibit the prescribed order of convergence for equilibrium initial data while degradation of accuracy is observed for type CK schemes and initial values far from equilibrium.

7.4. Asymptotic preserving methods in the diffusion regime

Similar to the classical fluid limit the development of efficient numerical methods for kinetic equations in diffusion regimes has been studied by several authors (Jin et al. 1998, Jin and Pareschi 2000, Klar 1998*a*, Naldi and Pareschi 2000, Gosse and Toscani 2003, Buet and Cordier 2007, Lemou and Mieussens 2008, Carrillo, Goudon, Lafitte and Vecil 2008, Lafitte and Samaey 2012, Boscarino et al. 2013, Dimarco et al. 2014). In this section, without aiming at being exhaustive, we illustrate some of the main strategies which have been developed in the recent literature to tackle the problem and construct AP schemes. As a prototype model we consider here the linear Boltzmann equation, characterized by the collision term (2.54) introduced in Section 2.8, in the diffusive scaling

$$\varepsilon\frac{\partial f}{\partial t} + v\cdot\nabla_{\mathbf{x}} f = \frac{1}{\varepsilon}Q(f)\,, \quad (7.67)$$

with $Q(f)$ given by (2.54).

Parity decomposition and AP splitting method

The numerical approach is based on a reformulation of equation (7.67) through the even and odd parities formalism : we split equation (7.67) into two equations, one for v and one for $-v$

$$\begin{aligned} \varepsilon\,\partial_t f + v\cdot\nabla_x f &= \frac{1}{\varepsilon}Q(f)(v),\\ \varepsilon\,\partial_t f - v\cdot\nabla_x f &= \frac{1}{\varepsilon}Q(f)(-v). \end{aligned} \quad (7.68)$$

Next, we introduce the so called even parity r and odd parity j defined by

$$r(x,v,t) = \frac{1}{2}\Big(f(x,v,t) + f(x,-v,t)\Big), \tag{7.69}$$

$$j(x,v,t) = \frac{1}{2\varepsilon}\Big(f(x,v,t) - f(x,-v,t)\Big). \tag{7.70}$$

This is equivalent to the decomposition

$$f(x,v,t) = r(x,v,t) + \varepsilon j(x,v,t), \quad f(x,-v,t) = r(x,v,t) - \varepsilon j(x,v,t).$$

Adding and subtracting the two equations in (7.68) leads to

$$\begin{aligned} \partial_t r + v\cdot\nabla_x j &= \frac{1}{\varepsilon^2}Q(r), \\ \partial_t j + \frac{1}{\varepsilon^2}v\cdot\nabla_x r &= -\frac{1}{\varepsilon^2}\lambda j\,, \end{aligned} \tag{7.71}$$

where λ is the collision frequency defined in (2.55) and where the property

$$\int_{\mathbb{R}^3}\sigma(v,w)j(w)\,dw = 0$$

has been used. An important advantage of this formulation is that now only one time scale appears in our new system (7.71). Note that the standard splitting method, based on the separation of convection and collision processes, applied to (7.71) originates the wrong asymptotic behavior (Jin et al. 1998, Naldi and Pareschi 2000).

The AP splitting method presented in Jin et al. (2000) is based on rewriting the above system into the following form

$$\begin{aligned} \partial_t r + v\nabla_x j &= \frac{1}{\varepsilon^2}Q(r)(f), \\ \partial_t j + \psi v\nabla_x r &= -\frac{1}{\varepsilon^2}\left(\lambda j + (1-\varepsilon^2\psi)v\nabla_x r\right), \end{aligned} \tag{7.72}$$

where $\psi = \psi(\varepsilon)$ is such that $0 \le \psi \le 1/\varepsilon^2$. This restriction on ψ guarantees the positivity of $\psi(\varepsilon)$ and $(1-\varepsilon^2\psi(\varepsilon))$ so the problem remain well-posed uniformly in ε. The simplest choice of ψ is

$$\psi(\varepsilon) = \min\left\{1,\, \varepsilon^{-2}\right\}. \tag{7.73}$$

Clearly system (7.72) preserves the same asymptotic behavior of the corresponding kinetic equation (7.67) since it is mathematically equivalent. Thus it will preserve the correct diffusion limit if discretized in a suitable way. Note that the left hand side is a simple non stiff transport operator whereas the right hand side contains all the stiff terms. Thus a natural splitting of the previous system will be

$$\begin{aligned} \partial_t r + v\nabla_x j &= 0, \\ \partial_t j + \psi v\nabla_x r &= 0, \end{aligned} \tag{7.74}$$

together with

$$\begin{aligned} \partial_t r &= \frac{1}{\varepsilon^2} Q(r), \\ \partial_t j &= -\frac{1}{\varepsilon^2} \left(\lambda j + (1 - \varepsilon^2 \psi) v \nabla_x r \right). \end{aligned} \tag{7.75}$$

In the limit $\varepsilon \to 0$ the relaxation step (7.75) gives

$$Q(r) = 0, \qquad \lambda j = -v \nabla_x r,$$

or equivalently

$$r = \rho M, \qquad j = -\frac{M}{\lambda} [v \nabla_x \rho]. \tag{7.76}$$

Inserting (7.76) in the transport step (7.74) and integrating over v one gets the drift diffusion equation (2.56). From a numerical point of view, fully explicit schemes to treat (7.74) are then combined with fully implicit schemes for (7.75) (see Jin et al. (2000) for further details). Note, however, that as a consequence the limiting scheme will be an explicit scheme for the diffusion equation (2.56) which may suffer of the usual parabolic CFL condition which requires the time step Δt to be of the order of the square of the space grid Δx. Moreover, the resulting space discretization of the diffusive terms is not optimal, since usually it originates a non compact stencil (typically a five point rather than a three point discretization of a second order space derivative). Finally let us mention that the above AP splitting in the case $\psi = 0$ was used in Klar (1998*a*), Klar (1998*b*), Naldi and Pareschi (1998).

IMEX Runge-Kutta methods

The method recently derived in Boscarino et al. (2013) is based on a different reformulation of the system with the goal to construct IMEX Runge-Kutta methods that in the limit are capable to avoid the severe stability condition $\Delta t \leq (\Delta x)^2$. First let us note that a standard IMEX Runge-Kutta method can be applied to the original parity system in the form

$$\begin{aligned} \partial_t r &= -\underbrace{v \cdot \nabla_x j}_{explicit} + \frac{1}{\varepsilon^2} \underbrace{Q(r)}_{implicit}, \\ \partial_t j &= -\frac{1}{\varepsilon^2} \underbrace{(v \cdot \nabla_x r + \lambda j)}_{implicit}. \end{aligned} \tag{7.77}$$

Note that the inversion of the implicit term in the second equation is done explicitly if one is able to compute efficiently r from the first equation. With above partitioning it is easy to see that the results discussed in Section 7.3 applies. In particular the limiting scheme corresponds to the explicit Runge-Kutta scheme applied to the limiting system (2.56). Therefore we obtain

an AP scheme but with the same limitations on the time step induced by the parabolic nature of the reduced limiting model.

The idea in Boscarino et al. (2013) is to write the parity system in the equivalent form

$$\begin{aligned} &\partial_t r + v\cdot\nabla_x\left(j+\eta\frac{v}{\lambda}\cdot\nabla_x r\right) = \frac{1}{\varepsilon^2}Q(r) + v\cdot\nabla_x\left(\eta\frac{v}{\lambda}\cdot\nabla_x r\right),\\ &\partial_t j + \frac{1}{\varepsilon^2}v\cdot\nabla_x = -\frac{1}{\varepsilon^2}\,\lambda j\,, \end{aligned} \tag{7.78}$$

where $\eta = \eta(\varepsilon)$ is a numerical positive function such that $\eta(0) = 1$. Note that in (7.78) the fluxes have been penalized using the equilibrium fluxes of the limiting behavior and so higher order space derivatives appear. Different choices for η are possible, for example in Boscarino et al. (2013) it was used

$$\eta(\varepsilon) = \exp(-\varepsilon^2/\Delta x).$$

An IMEX method is then applied following the partitioning below

$$\begin{aligned} \partial_t r &= -\underbrace{v\cdot\nabla_x\left(j+\eta\frac{v}{\lambda}\cdot\nabla_x r\right)}_{explicit} + \frac{1}{\varepsilon^2}\underbrace{Q(r) + v\cdot\nabla_x\left(\eta\frac{v}{\lambda}\cdot\nabla_x r\right)}_{implicit},\\ \partial_t j &= -\frac{1}{\varepsilon^2}\underbrace{(v\cdot\nabla_x r + \lambda j)}_{implicit}. \end{aligned} \tag{7.79}$$

In particular for $\eta = 0$ we recover (7.77). Compared to the partitioning (7.77) one has the additional problem of the inversion of $v\cdot\nabla_x\left(\eta\frac{v}{\lambda}\cdot\nabla_x r\right)$. However this term it is exactly the equilibrium flux that originates the limit equation. Therefore it contains the same difficulties of an implicit integrator applied to (2.56), which we cannot skip if our goal is to achieve an implicit scheme for (2.56). In addition the second order space derivatives in this additional term can be discretized accordingly to the desired scheme for the limiting parabolic equation and therefore non compact stencils are avoided. In other words the value of η at $\varepsilon = 0$ realizes a transition between an explicit solver ($\eta = 0$, non compact stencil) and an implicit solver ($\eta = 1$, compact stencil) for the limiting diffusion system.

Remark 7.6.

- In the numerical methods just described the collision operator has to be implicitly computed. For simple linear terms, like the case of neutron transport $Q(r) = \rho - r$ the implicit step can be solved explicitly. Otherwise, similarly to the fluid limit one can use the penalization technique by Filbet and Jin (2010) to avoid the inversion of the operator. In the case of the semiconductor Boltzmann equation this has been done for IMEX schemes in Dimarco et al. (2014) and for the AP splitting method in Deng (2014).

- The diffusion limit needs a particular care in the treatment of the space derivatives, this is due to the fact that in the limit the equation changes of character passing from an hyperbolic to a parabolic type equation. Different strategy can be adopted at this stage, see for instance Jin et al. (2000), Jin and Levermore (1996), Naldi and Pareschi (2000) and the references therein.

Micro-macro decomposition methods

Another approach to the problem is based on the so-called micro-macro decomposition (Lemou and Mieussens 2008). The starting point is the decomposition

$$f(x,v,t) = \rho(x,t)M(v) + \varepsilon g(x,v,t), \tag{7.80}$$

where the non equilibrium part g clearly is such that $\int_{\mathbb{R}^3} g\,dv = 0$. By direct substitution into the original equation (7.67) we get

$$\varepsilon\, M\frac{\partial \rho}{\partial t} + \varepsilon^2 \frac{\partial g}{\partial t} + v\cdot M\nabla_x \rho + \varepsilon v\cdot \nabla_x g = Q(g). \tag{7.81}$$

Note that integrating over v yields the equation for ρ

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot \int_{\mathbb{R}^3} vg\,dv = 0. \tag{7.82}$$

An evolution equation for g is found by defining the operator Π such that $\Pi(f) = M\int_{\mathbb{R}^3} f\,dv$ and the identity operator I. If we now apply the operator $\Pi - I$ to the equation (7.81) we get

$$\varepsilon^2 \frac{\partial g}{\partial t} + v\cdot M\nabla_x \rho + \varepsilon(I-\Pi)(v\cdot\nabla_x g) = Q(g). \tag{7.83}$$

The micro-macro approach is then based on discretizing (7.82) and (7.83). The method proposed in Lemou and Mieussens (2008) use the following implicit-explicit partitioning

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= \underbrace{-\nabla_x \cdot \int_{\mathbb{R}^3} vg\,dv}_{implicit},\\
\frac{\partial g}{\partial t} &= \underbrace{-\frac{1}{\varepsilon^2} v\cdot M\nabla_x\rho + \frac{1}{\varepsilon}(I-\Pi)(v\cdot\nabla_x g)}_{explicit} + \frac{1}{\varepsilon^2}\underbrace{Q(g)}_{implicit}.
\end{aligned} \tag{7.84}$$

Note that, as for the schemes derived before, it is the possibility to invert the collision operator that makes the whole scheme explicitly solvable. The above approach is clearly AP, since as $\varepsilon \to 0$ the second equation gives

$$Q(g) = v\cdot M\nabla_x \rho,$$

which gives

$$g = Q^{-1}(vM) \cdot \nabla_x \rho = \frac{M}{\lambda} \left[\int_{\mathbb{R}^3} \sigma(v, v_*) g(v_*)\, dv_* - v \cdot \nabla_x \rho \right].$$

Substituting this into the first equation in (7.84) gives the desired diffusion limit (2.56). More precisely, at a time discrete level, one gets an explicit scheme for the diffusion limit. A modified approach that permits to recover an implicit scheme in the diffusion limit has been proposed in Lemou (2010).

8. Fluid-kinetic coupling and hybrid methods

We discuss in this section numerical methods which address specifically the multiscale nature of several physical problems described through kinetic equations. In contrast with asymptotic preserving schemes, discussed in Section 7, which aim at solving the kinetic equation in the whole computational domain for all the different regimes, here we consider other complementary approaches based on fluid-kinetic coupling strategies and hybrid schemes. The fundamental principles are extremely simple and intuitive but their practical realization poses several difficulties. Roughly speaking, one would like to avoid the expensive cost of solving the kinetic equation in regions well described by continuum fluid models (since the latter are easily solvable by classical numerical methods). On the other hand, far away from equilibrium, it is desirable to maintain the flexibility and efficiency of stochastic techniques, such as DSMC methods (Bird 1994, Nanbu 1980, Pareschi and Russo 1999, Rjasanow and Wagner 2006). The crucial difficulty is the identification, the modeling and the numerics of the transition zone between the fluid and the kinetic descriptions.

The amount of literature in this direction is enormous, since several different techniques are possible and often the implementation details of the schemes are of fundamental importance for the effective understanding of the simulation process (Burt and Boyd 2008, Burt and Boyd 2009, Degond et al. 2007, Dimarco and Pareschi 2007, Caflisch, Wang, Dimarco, Cohen and Dimits 2008, Degond et al. 2011, Wijesinghe and Hadjiconstantinou 2004, Homolle and Hadjiconstantinou 2007*a*, Crestetto, Crouseilles and Lemou 2012, Alaia and Puppo 2011, Alaia and Puppo 2012). Here we limit ourselves to illustrate some examples, that we consider to be representative of the most common approaches used in this context. We refer also to the recent review by Radtke, Péraud and Hadjiconstantinou (2013).

8.1. Dynamic fluid-kinetic coupling methods

Domain decomposition techniques represent the most natural way to tackle the problem through a subdivision of the computational domain into two

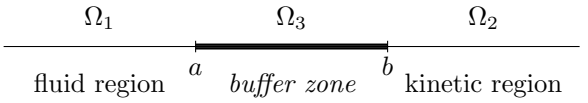

Figure 8.1. A schematic representation in one-dimension of the buffer zone between the kinetic and the fluid regions.

complementary domains (Bourgat, LeTallec, Perthame and Qiu 1992, Bourgat, LeTallec and Tidriri 1996, Schneider 1996, Tiwari and Klar 1998, Tiwari 1998*a*). In the continuum region the gas is well described by either Euler or Navier-Stokes equations, while in the kinetic region the gas needs a kinetic description. Unfortunately, in most cases the two regions are themselves unknown, and therefore they have to be computed and evolved as part of the solution. Along this direction we review some recent contributions which proposes a moving interface method to deal with the coupling of the different regions (Degond et al. 2007, Degond et al. 2010, Degond and Dimarco 2012). Related approaches realizing automatic domain decomposition methods were derived in Kolobov, Arslanbekov, Aristov, Frolova and Zabelok (2007) and in Tiwari (1998*b*).

For sake of simplicity we describe the methods in one space and velocity dimensions for the Boltzmann equation in the fluid dynamic scaling

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \frac{1}{\varepsilon}Q(f), \quad x, v \in \mathbb{R} \tag{8.1}$$

with initial data $f(x, v, 0) = f_0(x, v)$ and where the collision term is the BGK relaxation operator $Q(f) = \nu(M[f] - f)$. Extension of the approach to the full Boltzmann equation is discussed at the end of the section.

A moving interface method

The method here described has been proposed in Degond et al. (2007) and is based on a previous work of Degond et al. (2005), where a stationary smooth transition strategy was proposed for this coupling. The main idea is to derive the time evolution of the buffer zones between the kinetic and the fluid regions based on several microscopic and macroscopic criteria. As we will see, the construction of this buffer zone is based on the choice of a cut-off function updated in time by certain out-of-equilibrium indicators.

Let now Ω_1, Ω_2, and Ω_3 be three disjointed sets such that $\Omega_1 \cup \Omega_2 \cup \Omega_3 = \mathbb{R}^1$. The first set Ω_1 is supposed to be a domain in which the flow is far from the equilibrium (the "kinetic region"), while the flow is supposed to be close to the equilibrium in Ω_2 (the "fluid region") and also in Ω_3 (the

"buffer zone"), see Figure 8.1. We define a function $h(x,t)$ such that

$$h(x,t) = \begin{cases} 1, & \text{for} \quad x \in \Omega_1, \\ 0, & \text{for} \quad x \in \Omega_2, \\ 0 \le h(x,t) \le 1, & \text{for} \quad x \in \Omega_3, \end{cases} \tag{8.2}$$

and set $h(x,0) = h_0(x)$.

The time dependence of h means that we account for possibly dynamically changing fluid and kinetic zones. We will denote $\Omega_3 = [a,b]$ so that $\Omega_1 = (-\infty, a)$ and $\Omega_2 = (b, \infty)$. The topology and geometry of these zones is directly encoded in h and may change dynamically as well. For instance, h can be chosen piecewise linear in $[a,b]$

$$h(x,t) = \frac{x-b}{a-b} \quad \text{for} \quad x \in [a,b].$$

We define two distribution functions such that

$$f_K = hf, \qquad f_F = (1-h)f. \tag{8.3}$$

We look now for an evolution equation for f_K and for f_F. We write

$$\begin{aligned} \frac{\partial f_K}{\partial t} &= \frac{\partial}{\partial t}(hf) = f\,\frac{\partial h}{\partial t} + h\frac{\partial f}{\partial t}, \\ \frac{\partial f_F}{\partial t} &= \frac{\partial}{\partial t}((1-h)f) = -f\,\frac{\partial h}{\partial t} + (1-h)\frac{\partial f}{\partial t}. \end{aligned}$$

Thus multiplying the Boltzmann equation (8.1) by h and $1-h$ respectively, (8.1) can be rewritten in the following form

$$\begin{aligned} \frac{\partial f_K}{\partial t} &= f\,\frac{\partial h}{\partial t} + h\left(-v\frac{\partial f}{\partial x} + \frac{\nu}{\varepsilon}(M[f]-f)\right), \\ \frac{\partial f_F}{\partial t} &= -f\,\frac{\partial h}{\partial t} + h\left(-v\frac{\partial f}{\partial x} + \frac{\nu}{\varepsilon}(M[f]-f)\right), \end{aligned}$$

which finally leads to the following system for f_F and f_K

$$\frac{\partial f_K}{\partial t} + hv\left[\frac{\partial f_K}{\partial x} + \frac{\partial f_F}{\partial x}\right] = \frac{h\nu}{\varepsilon}(M[f]-f) + f\frac{\partial h}{\partial t}, \tag{8.4}$$

$$\frac{\partial f_F}{\partial t} + (1-h)v\left[\frac{\partial f_F}{\partial x} + \frac{\partial f_K}{\partial x}\right] = \frac{(1-h)\nu}{\varepsilon}(M[f]-f) - f\frac{\partial h}{\partial t}, \tag{8.5}$$

with initial data

$$f_K(x,v,0) = h_0(x)f_0(x,v)\,, \quad f_F(x,v,0) = (1-h_0(x))f_0(x,v). \tag{8.6}$$

Let us note that if $f = f_F + f_K$ is the solution of (8.1) with initial data $f_0(x,v)$, then (f_F, f_K) is the solution of (8.4)-(8.5) with initial data (8.6) and conversely.

Let us now assume that the domain can be subdivided in two regions: in

one of the regions, the distribution function is close to a local Maxwellian while in the other, it is far from it. We choose to set $h = 0$ in the region where f is close to the Maxwellian. Therefore, $f_F = f$ is close to its associated Maxwellian $M[f_F] = M[f]$ and we can replace the Boltzmann equation by the Euler equations without making any significant error. We also suppose that in the buffer zone, f_F remains close to the equilibrium and thus, it can be replaced by $M[f_F]$ in the whole interval $\Omega_1 \cup \Omega_3$.

We introduce the notations

$$\langle f \rangle = \int_{\mathbb{R}} f \, dv, \qquad \langle f \, m \rangle = \int_{\mathbb{R}} f \begin{pmatrix} 1 \\ v \\ \frac{|v|^2}{2} \end{pmatrix} dv = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}, \tag{8.7}$$

where $m = (1, v, |v|^2/2)^T$. Replacing f_F by $M[f_F]$ in (8.5) and taking the hydrodynamic moments, leads to the following modified Euler system defined in the interval $x \leq b$

$$\begin{aligned}
&\frac{\partial \rho_F}{\partial t} + (1-h)\frac{\partial}{\partial x}(\rho_F u_F) = -(1-h)\frac{\partial}{\partial x}\langle v f_K \rangle - \rho \frac{\partial h}{\partial t}, \\
&\frac{\partial \rho_F u_F}{\partial t} + (1-h)\frac{\partial}{\partial x}(\rho_F u_F^2 + p_F) = -(1-h)\frac{\partial}{\partial x}\langle v^2 f_K \rangle - \rho u \frac{\partial h}{\partial t}, \\
&\frac{\partial E_F}{\partial t} + (1-h)\frac{\partial}{\partial x}((E_F + p_F)u_F) = -(1-h)\frac{\partial}{\partial x}\langle v \frac{|v|^2}{2} f_K \rangle - E \frac{\partial h}{\partial t},
\end{aligned} \tag{8.8}$$

where $p_F = \rho_F T_F$, $E_F = \rho_F(T_F + u_F^2)/2$, and initial data

$$(\rho_F, u_F, T_F)|_{(x,0)} = (1 - h_0(x))(\rho, u, T)|_{(x,0)}.$$

Under these assumptions, we have $f = f_K + M[f_F]$, where f_K is a solution of

$$\frac{\partial f_K}{\partial t} + hv\frac{\partial f_K}{\partial x} + hv\frac{\partial}{\partial x}M[f_F] = \frac{h\nu}{\varepsilon}(M[f] - f) + f\frac{\partial h}{\partial t}, \tag{8.9}$$

in the interval $\Omega_3 \cup \Omega_2$. The coupling model consists of system (8.8) for the hydrodynamic moments in the region $\Omega_1 \cup \Omega_3$ and eq. (8.9) for the kinetic distribution function in the region $\Omega_3 \cup \Omega_2$.

When $h = 0$, system (8.8) coincides with the Euler system (2.43) because $f_K = 0$ and $f_F = M[f_F]$. Moreover no boundary condition is needed at the boundary $x = b$ because $h = 1$ at this point, and the factors in front of the spatial derivatives of (8.8) vanish. A similar remark is true for f_K. Indeed, when $h = 0$, $f_K = 0$ and no boundary condition is needed for the kinetic equation at $x = a$ because $h = 0$ at this point and the factor in front of the spatial derivatives in (8.9) vanishes. In the buffer zone Ω_3, the solution of the full kinetic problem f is computed as the sum of the Maxwellian $M[f_F]$ and of the function f_K. To summarize, the solution of the full kinetic problem is given by f_K in Ω_2, by $M[f_F]$ in Ω_1 and by $M[f_F] + f_K$ in Ω_3.

Once the coupled model is derived one can use classical finite volume techniques to discretize both the kinetic equation (8.9) and the modified Euler system (8.8). We refer to Degond et al. (2007) for more details.

Remark 8.1. An important feature of the method is that it is very easy to divide the domain in more than two zones. Thus we can define as many buffers and as many kinetic regions as necessary. Additionally, we can create new buffer zones and new kinetic zones during the simulation. For this purpose, one can update the cut-off function h according to convenient criteria to a new value and reset $f_K = hf$ and $f_F = (1-h)f$ at the time when h is changed.

A micro-macro moving interface method

The method just described can be improved (Degond et al. 2010) by considering the micro-macro decomposition of the distribution function

$$f = M[f] + g. \tag{8.10}$$

Because the equilibrium distribution has the same first three moments as f we have $\langle g\, m\rangle = 0$. Then it can be easily proved that the following coupled system

$$\frac{\partial}{\partial t}\langle f\, m\rangle + \frac{\partial}{\partial x}\langle v\, M[f]\, m\rangle + \frac{\partial}{\partial x}\langle v\, g\, m\rangle = 0 \tag{8.11}$$

$$\frac{\partial g}{\partial t} + v\frac{\partial g}{\partial x} = -\frac{\nu}{\varepsilon}g - \left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right) M[f] \tag{8.12}$$

is satisfied. The corresponding initial data are

$$\begin{aligned} &\langle f(x,v,0)\ m\rangle = \langle f_0(x,v)\ m\rangle, \\ &g(x,v,0) = g_0(x,v) = f_0(x,v) - M[f_0]. \end{aligned} \tag{8.13}$$

The converse statement is also true: if $\langle f\, m\rangle$ and g satisfy system (8.11) and (8.12) with initial data (8.13), then $f = M[f] + g$ satisfies the kinetic equation (8.1) (see Degond et al. (2005) for details).

Next, we introduce the buffer zone function $h(x,t)$ as in (8.2) and split the perturbation term in two distribution functions $g_K = hg$ and $g_F = (1-h)g$. It is therefore easy to derive the following coupled system of equations

$$\frac{\partial}{\partial t}\langle f\, m\rangle + \frac{\partial}{\partial x}\langle v\, M[f]\, m\rangle + \frac{\partial}{\partial x}\langle v g_K\, m\rangle + \frac{\partial}{\partial x}\langle v g_F\, m\rangle = 0, \tag{8.14}$$

$$\begin{aligned} &\frac{\partial g_K}{\partial t} + hv\frac{\partial g_K}{\partial x} + hv\frac{\partial g_F}{\partial x} \\ &\qquad = -\frac{\nu}{\varepsilon}g_K - h\left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right) M[f] + \frac{g_K}{h}\frac{\partial h}{\partial t}, \end{aligned} \tag{8.15}$$

$$\frac{\partial g_F}{\partial t} + (1-h)v\frac{\partial g_K}{\partial x} + (1-h)v\frac{\partial g_F}{\partial x}$$

$$= -\frac{\nu}{\varepsilon} g_F - (1-h)\left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right) M[f] - \frac{g_F}{1-h}\frac{\partial h}{\partial t}, \tag{8.16}$$

with initial data

$$\begin{aligned} g_K(x,v,0) &= h_0(v) g_0(x,v), \\ g_F(x,v,0) &= (1-h_0(x)) g_0(x,v), \\ \langle f(x,v,0)\rangle &= \langle f_0(x,v)\rangle. \end{aligned} \tag{8.17}$$

Again, system (8.14)-(8.16) with initial data (8.17) is equivalent to system (8.11)-(8.12) with initial data (8.13).

Now assume that the flow is very close to equilibrium in $\Omega_2 \cup \Omega_3$. This means that g is very small in these domains and can be set to zero. Since $g = g_F$ in Ω_2, we set $g_F = 0$ in this domain. In Ω_3, we also set $g_F = 0$, which means that we approximate g by g_K. Consequently, g_F can be eliminated from (8.14)-(8.16) to get

$$\frac{\partial}{\partial t}\langle f\, m\rangle + \frac{\partial}{\partial x}\langle v\, M[f]\, m\rangle + \frac{\partial}{\partial x}\langle v g_K\, m\rangle = 0 \tag{8.18}$$

$$\frac{\partial g_K}{\partial t} + hv\frac{\partial g_K}{\partial x} = -\frac{\nu}{\varepsilon} g_K - h\left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x}\right) M[f] + \frac{g_K}{h}\frac{\partial h}{\partial t}, \tag{8.19}$$

with initial data (8.17).

Note that since by definition g_K is zero in the fluid zone Ω_2, the kinetic equation equation (8.19) is solved in the kinetic and buffer zones Ω_1 and Ω_3 only. Indeed, in the fluid zone, we only solve (8.18) with $g_K = 0$, which is nothing but the Euler equations. In the kinetic zone, we have $g_K = g$ and hence system (8.18)-(8.19) is nothing but system (8.11)-(8.12), which is equivalent to the original BGK equation. With this system, the distribution function f is approximated by $M[f] + g_K$. Similarly to the previous paragraph finite volume techniques can be used to discretize the coupled system (8.18)-(8.19). We refer to (Degond et al. 2010) for details on the discretizations and on the breakdown criteria of the fluid model.

As a numerical example, we consider the classical Sod test in the computational domain $[-20, 20]$ for the micro-macro fluid-kinetic coupling. The simulations are initialized with a thermodynamic equilibrium with $h = 0$ and $g_K = 0$ everywhere. In Figure 8.2 the density on the left, the transition function, the heat flux and the local Knudsen number on the right are reported. For the density we plot the solution computed with the moving interface micro-macro method (mic-mac in the legend) and as a reference the solution computed with the full kinetic model. We also report the solution computed with a macroscopic fluid solver. We observe that due to the initial shock, a kinetic region appears immediately and starts to grow in time, but as soon as the different non equilibrium regions separate, the kinetic region itself splits into three: one around the rarefaction wave, one around the contact discontinuity, and one around the shock. In this test

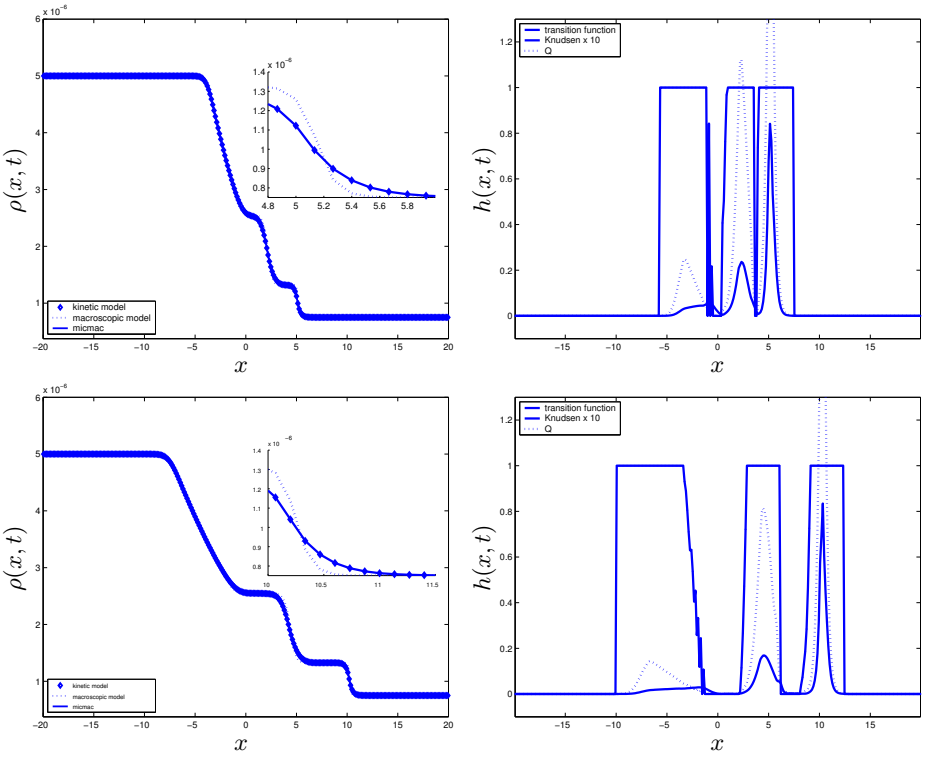

Figure 8.2. Moving interface method in Sod shock tube problem. Density (left) and transition function (right). Solution at $t = 12 \times 10^{-3}$ (top) and $t = 24 \times 10^{-3}$ (bottom). The small panels are a magnification of the solution close to the shock.

case the dynamic coupling allows for a reduction of approximately 60% of the computational time employed by the full kinetic scheme. We refer to Degond et al. (2010) for further numerical results.

Remark 8.2. The method can be extended to the full Boltzmann equation

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f, f), \quad x, v \in \mathbb{R}^3 \tag{8.20}$$

by introducing the same cut-off function $h(x,t)$ as in (8.2) except that now $\Omega_i \subset \mathbb{R}^3$ are disjoint sets such that $\cup_{i=1}^3 \Omega_i = \mathbb{R}^3$.

By repeating similar calculations as before we arrive to the system (Degond and Dimarco 2012)

$$\frac{\partial}{\partial t}\langle f\, m\rangle + \nabla_x \cdot \langle v\, M[f]\, m\rangle + \nabla_x \cdot \langle v g_K\, m\rangle = 0 \tag{8.21}$$

$$\frac{\partial f_K}{\partial t} + v \cdot \nabla_x f_K = \frac{1}{\varepsilon} Q(M[f_F] + f_K, M[f_F] + f_K) \tag{8.22}$$

$$+\left(\frac{\partial h}{\partial t}+v\cdot\nabla_x h\right)\frac{f_K}{h}.$$

In this case, to achieve maximum efficiency one can combine a Monte Carlo solver for (8.22) with a finite volume method for (8.21). This originates an hybrid scheme of the type of those we will discuss in the next subsections (Degond and Dimarco 2012).

8.2. Low-variance deviational Monte Carlo methods

Variance reduction methods are a popular way to improve the accuracy of Monte Carlo methods by reducing the amount of fluctuations in the results (Caflisch 1988). Control variates methods have been studied for DSMC simulations of low speed gas flows (Homolle and Hadjiconstantinou 2007*a*, Homolle and Hadjiconstantinou 2007*b*, Baker and Hadjiconstantinou 2005, Radtke and Hadjiconstantinou 2009, Radtke et al. 2011). Other methods are based on the use of weighted samples in each computational cell (Rjasanow and Wagner 1996, Rjasanow and Wagner 2001).

Numerical simulations of low speed gas flows typically involve small deviations from equilibrium, which translates to small hydrodynamic signals (e.g. flow velocity, heat flux, etc.). In this limit, DSMC methods become very expensive because resolution of the hydrodynamic signals requires very large numbers of samples. Here we describe the method proposed in Radtke and Hadjiconstantinou (2009) in the case of the BGK equation

$$\frac{\partial f}{\partial t}+v\cdot\nabla_x f=\nu(M[f]-f). \tag{8.23}$$

The starting point is to represent the kinetic solution using the decomposition

$$f(x,v,t)=M^B(x,v)+f^D(x,v,t), \tag{8.24}$$

where $M^B(x,v)$ is a suitable time-independent Maxwellian equilibrium function, characterized by the macroscopic quantities $\rho^B(x)$, $u^B(x)$ and $T^B(x)$, and f^D represents a deviation from the equilibrium distribution, characterized by signed particles (with sign $\mathrm{sgn}(M^B-f)$). The rationale behind this approach is that it leads naturally to a control variate integration

$$\int_{\mathbb{R}^3} f\varphi\,dv=\int_{\mathbb{R}^3}(f-M^B)\varphi\,dv+\int_{\mathbb{R}^3} M^B\,\varphi\,dv, \tag{8.25}$$

for a given function $\varphi(v)$ where $M^B(x,v)$ must be close to the solution $M^B\approx f$. Using a Monte Carlo method to evaluate only the second integral in (8.25) results in significantly reduced statistical uncertainty, because most of the statistical uncertainty is removed through the deterministic evaluation of the last integral.

The simulation proceeds by using the splitting method (2.58)-(2.57) of

the BGK model (8.23). The collision and advection substeps are described in details as follows.

1 The collision step reads

$$\frac{\partial f^D}{\partial t} = \nu(M[f] - M^B) - \nu f^D, \tag{8.26}$$

which can be exactly solved in a time step Δt to give

$$f^{D,n+1}(x,v) = e^{-\lambda} f^{D,n}(x,v) + (1 - e^{-\lambda})(M[f^n] - M^B(x,v)),$$

where $\lambda = \nu\Delta t$. The first term describe a particle deletion with probability $e^{-\lambda}$ whereas the second a signed particle generation (with sign $\mathrm{sgn}(M[f^n] - M^B(x,v))$) with probability $(1 - e^{-\lambda})$ from the distribution $|M[f^n] - M^B(x,v)|$.

2 The transport step becomes

$$\frac{\partial f^D}{\partial t} + v \cdot \nabla_x f^D = -v \cdot \nabla_x M^B. \tag{8.27}$$

This is solved by free advection of particles and generating additional particles from the cell interfaces at every time step to satisfy the inhomogeneous term on the right hand side. This can be done using the ratio-of-uniforms sampling method. We refer to Radtke and Hadjiconstantinou (2009) for the details.

In practice the value $M^B(x,v)$ is adjusted along the computations in order to achieve maximum efficiency and minimize the number of deviational particles at each time step. We refer to Homolle and Hadjiconstantinou (2007*a*), Homolle and Hadjiconstantinou (2007*b*), Radtke et al. (2011) for extensions to the full Boltzmann equation.

8.3. Moment guided Monte Carlo methods

The basic idea described here consists in obtaining reduced variance Monte Carlo methods by forcing the statistical samples to match prescribed sets of moments given by the solution of deterministic macroscopic fluid equations (Degond et al. 2011, Dimarco 2013). These macroscopic models, in order to represent the correct physics for all range of Knudsen numbers include a kinetic correction term, which takes into account departures from thermodynamical equilibrium and couples the kinetic and fluid models.

The starting point of the method is the micro-macro decomposition

$$f = M[f] + g, \tag{8.28}$$

where the function g, with $\langle g\, m\rangle = 0$, represents the non-equilibrium part of the distribution function. In the case of the Boltzmann equation in the

fluid-limit scaling it is easy to see that f and g satisfy the coupled system of equations

$$\frac{\partial U}{\partial t} + \nabla_x \cdot F(U) + \nabla_x \cdot \langle vmg \rangle = 0, \tag{8.29}$$

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\varepsilon} Q(f, f). \tag{8.30}$$

where $U = \langle f\, m \rangle$ and $F(U) = \langle vM[f]\, m \rangle$. The method aims at solving the *macro-scale* moment system (8.29) using as a closure the time evolution of g, which is given by the solution of the *micro-scale* kinetic equation (8.30). Of course, solving equation (8.30) also mean knowing the solution to (8.29). However, if the numerical solution of equation (8.30) is computed by DSMC methods and the solution of the moment system (8.29) by a finite volume scheme, we may expect this latter solution to be less affected by fluctuations and therefore to represent a better estimate of true solution. This is the statement on which the method is build.

The method can be summarized as follows. Given an initial data f^n at time $t^n = n\Delta t$ the new solution f^{n+1} is computed as a sequence of three steps.

1 Solve the kinetic equation (8.30) with a DSMC scheme and obtain a first set of moments $U^* = \langle mf^* \rangle$.
2 Solve the fluid equation (8.29) with a finite volume/difference scheme using the DSMC solution to close the system and obtain a second set of moments U^{n+1}.
3 Match the moments of the kinetic solution with the fluid solution through a transformation of the samples values $f^{n+1} = T(f^*)$ so that $\langle mf^{n+1} \rangle = U^{n+1}$.

Note that, if we want the method to be efficient in the fluid regime a robust Monte Carlo solver for small values of ε is required. In fact, both solvers, the deterministic and the DSMC methods are used in the whole computational domain. To this aim one can use the asymptotic-preserving Monte Carlo methods introduced in Remark 7.2 (Pareschi and Russo 2001).

An AP moment guided method

Here we describe the basic structure of the first order method based on splitting (2.58)-(2.57), higher order can be achieved by Strang splitting. We assume to apply first the transport and then the collision part. If we denote with $\tilde{f}$ the solution after the transport, the solution of the kinetic equation is computed approximating the collision part with the scheme

$$f^* = A_0 \tilde{f} + A_1 \tilde{f}_1 + A_2 M[\tilde{f}] \tag{8.31}$$

where $f_1 = P(f, f)/\mu$, $P(f, f) = Q(f, f) + \mu f$, $A_0 = e^{-\lambda}$, $A_1 = e^{-\lambda}(1 - e^{\lambda})$ for TR scheme or $A_1 = \lambda e^{\lambda}$ for IF scheme, and $A_2 = 1 - A_0 - A_1$. Here $\lambda =$

$\mu\Delta t/\varepsilon$ and $\mu > 0$ is such that P is a nonnegative operator (see Section 7.1). The AP property of the scheme correspond to the fact that the coefficients satisfy

$$\lim_{\lambda\to\infty} A_2 = 1, \quad \lim_{\lambda\to\infty} A_i = 0,\ i = 0, 1. \tag{8.32}$$

Next, since g is computed from the particle solution we can write

$$g^* = f^* - M[f^*] = A_0\tilde{f} + A_1\tilde{f}_1 + (A_2 - 1)M[\tilde{f}], \tag{8.33}$$

where we used the fact that $M[f^*] = M[\tilde{f}]$ for the conservation properties of the collision term. The above expression tells that the moments $\langle vmg^*\rangle$ can be obtained as a contribution of three terms

$$A_0\langle vm\tilde{f}\rangle + A_1\langle vm\tilde{f}_1\rangle + (A_2 - 1)\langle vmM[\tilde{f}]\rangle. \tag{8.34}$$

The first two terms are obtained by simple evaluation of the particles moments. The third term is obtained by integrating over the velocity space the analytic expression of the Maxwellian distribution. Finally, observe that in the limit $\varepsilon \to 0$ the contribution of the perturbation g goes to zero

$$\lim_{\varepsilon\to 0} g^* = 0, \tag{8.35}$$

because of (8.32) and the kinetic correction in (8.29) disappears. As a consequence, in such limit we obtain a purely deterministic solver for the Euler equations.

There are several possible types of space and time discretizations that can be used to approach the moments equations (8.29). We did not discuss it here. We remark that the most delicate point is the reconstruction of the value $\langle vmg^*\rangle$ which is used as a correction term (Degond et al. 2011, Dimarco 2013).

The moment matching procedure permits to the scheme to reduce the variance of DSMC even far from the fluid regime and to advance in time the method to the next time step. Note that, after solving (8.29) and (8.30) we obtain two different sets of moments U^* and U^{n+1}. Since we assume U^{n+1} to be a better estimate of the expected values of the statistical samples we force the particles to have the prescribed moments. For the mass density this can be achieved only introducing a particle weight in each cell, whereas for momentum and energy one can use the classical transformation described in Caflisch (1988).

In Figure 8.3 we gave a simple example of the variance reduction obtained with the present method. An unsteady shock for the BGK model is considered in a regime close to the fluid limit. The shock is produced pushing the particles against a wall which is located on the left boundary. For this model the collision step is solved exactly which corresponds to $A_0 = e^{-\lambda}$, $A_1 = 0$ and $A_2 = (1 - e^{-\lambda})$ in (8.31). We report also in each figure the

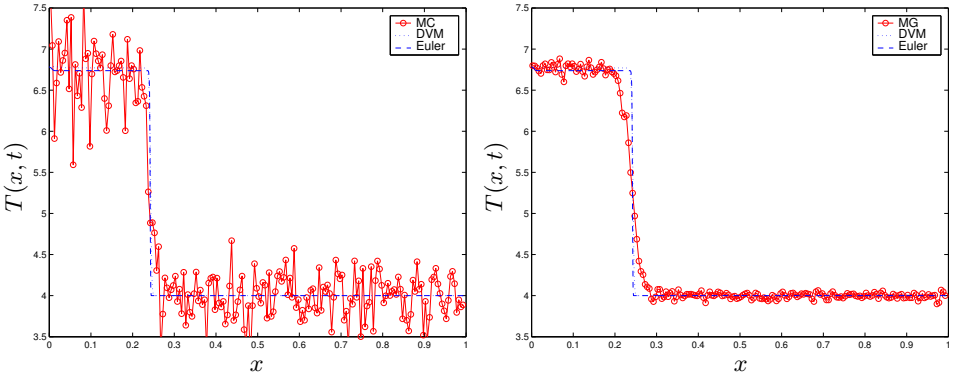

Figure 8.3. Temperature solution. Monte Carlo method (left) and Moment guided method (right). Knudsen number $\varepsilon = 10^{-4}$.

results for the compressible Euler equation. Finally, the reference solutions are obtained by solving the BGK scheme with a discrete velocity method.

8.4. Hybrid multiscale methods

In this paragraph we discuss another idea to blend together deterministic and stochastic solvers. The starting point of the method is a different kind of representation of the solution as equilibrium and non equilibrium part first introduced in Pareschi and Caflisch (1999). In these schemes, the solution in each cell is represented as a combination of two different parts, a stochastic particle representation of the non equilibrium fraction and a deterministic representation of the equilibrium part (Pareschi and Caflisch 2004, Dimarco and Pareschi 2006, Dimarco and Pareschi 2007, Dimarco and Pareschi 2010). To simplify the presentation we make use of the one-dimensional BGK model. Extensions of the methods to the multidimensional case are straightforward. The full Boltzmann case can be treated by using a suitable AP Monte Carlo method and will be addressed at the end of the paragraph.

The methods are based on the following definition of hybrid representation of the solution.

Definition 8.1. Given a density $f(v,t)$, and a Maxwellian density $M[f](v,t)$ we define $w(v,t) \in [0,1]$ and $\tilde{f}(v,t) \geq 0$ in the following way

$$w(v,t) = \begin{cases} \dfrac{f(v,t)}{M[f](v,t)}, & f(v,t) \leq M[f](v,t) \neq 0 \\ 1, & f(v,t) \geq M[f](v,t) \end{cases} \tag{8.36}$$

and

$$\tilde{f}(v,t) = f(v,t) - w(v,t)M[f](v,t). \tag{8.37}$$

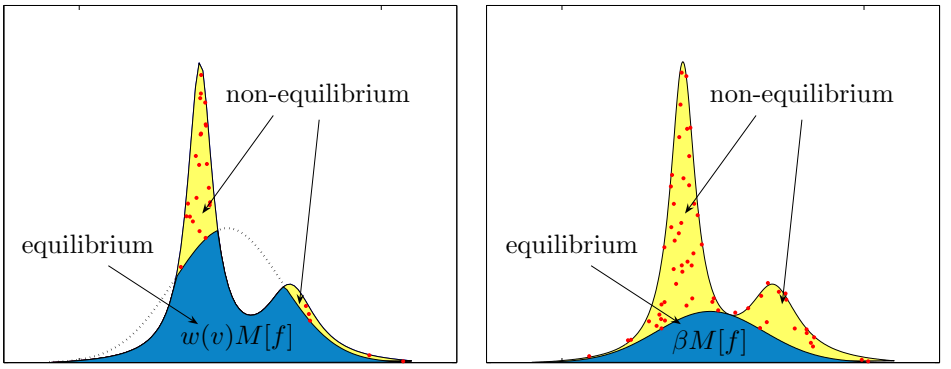

Figure 8.4. Distribution function as a combination of equilibrium and non-equilibrium part. Representation (8.38) (left) and (8.39) (right).

Thus $f(v,t)$ can be represented as (Figure 8.4)

$$f(v,t) = \tilde{f}(v,t) + w(v,t)M[f](v,t). \tag{8.38}$$

If one takes $\beta(t) = \min_v\{w(v,t)\}$ and $\tilde{f}(v,t) = f(v,t) - \beta(t)M[f](v,t)$, this leads to

$$\int_v \tilde{f}(v,t)dv = 1 - \beta(t).$$

Defining for $\beta(t) \neq 1$ the probability density

$$f^p(v,t) = \frac{\tilde{f}(v,t)}{1-\beta(t)},$$

then $= f(v,t)$, can be written as a convex combination of two probability densities in the form (Pareschi and Caflisch 1999, Pareschi and Caflisch 2004)

$$f(v,t) = (1-\beta(t))f^p(v,t) + \beta(t)M[f](v,t). \tag{8.39}$$

The case $\beta(t) = 1$ is trivial since it implies $f(v,t) = M[f](v,t)$.

We consider now the following general representation, including space dependence

$$f(x,v,t) = \tilde{f}(x,v,t) + w(x,v,t)M[f](x,v,t), \tag{8.40}$$

where $w(x,v,t) \geq 0$ is a function characterizing the equilibrium fraction and $\tilde{f}(x,v,t)$ the non equilibrium part of the distribution function.

A hybrid method for the BGK model

The core idea in the method is to represent $\tilde{f}(x,v,t)$ with particles and $M[f]$ by its analytical expression through the moments of f. Then we must find a way to compute the evolution of the hybrid representation (8.40) in time. The starting point of the method is again the classical operator splitting where the collision step for the BGK model $Q(f) = \nu(M[f] - f)$ is solved exactly.

In a single time step Δt a simple hybrid method can be summarized as follows.

1 Starting from a function $f^n = \tilde{f}^n + w^n M[f^n]$ in the form (8.40) solve the relaxation step

$$\frac{\partial f^*}{\partial t} = \frac{\nu}{\varepsilon}(M[f^*] - f^*). \tag{8.41}$$

For example, using the exact solution we have

$$\begin{aligned} f^* &= e^{-\lambda} f^n + (1 - e^{-\lambda}) M[f^n] \\ &= e^{-\lambda} \tilde{f}^n + (1 - e^{-\lambda} + e^{-\lambda} w^n) M[f^n], \end{aligned}$$

where $\lambda = \nu \Delta t/\varepsilon$. This decomposition can be recast in the form

$$f^* = \tilde{f}^* + w^* M[f^n],$$

taking $\tilde{f}^* = e^{-\lambda}\tilde{f}$ and $w^* = 1 - e^{-\lambda} + e^{-\lambda} w^n$. Note that $w^* > w^n$.

2 Discard a fraction $e^{-\lambda}$ of the sample particles since $\tilde{f}^* = e^{-\lambda}\tilde{f}$.

3 Starting from the function f^* solve the transport step.

(a) Transport the particle fraction $\tilde{f}^*$ by simple particles shifts which solve

$$\frac{\partial \tilde{f}^*}{\partial t} + v \cdot \nabla_x \tilde{f}^* = 0.$$

(b) Transport the deterministic fraction $w^* M[f]$ by a deterministic scheme for

$$\frac{\partial w^* M[f]}{\partial t} + v \cdot \nabla_x w^* M[f] = 0. \tag{8.42}$$

(c) Project the computed hybrid solution f^{n+1} to the form (8.40) using Definition 1. This gives w^{n+1} and $\tilde{f}^{n+1}$.

Note that as $\varepsilon \to 0$ we have $w^* \to 1$ and therefore the solution becomes fully deterministic.

Remark 8.3.

- An important aspect in the previous method is the possibility to use a cut-off weight function

$$w_R(x,v,t) = w(x,v,t), \quad |v| \le R, \qquad w_R(x,v,t) = 0, \quad |v| > R,$$

where the choice of R is done in such a way that the deterministic solver does not suffer of restrictive CFL conditions due to high velocities. In this way, high speed velocities $|v| > R$ are treated by a full particle scheme whereas for $|v| \le R$ the hybrid method is used.

- In the case of representation (8.39) the method can be modified to avoid the limitations induced by the use of a kinetic scheme in step 3(b) for the equilibrium part and to allow the coupling with an arbitrary fluid solver for the Euler equations (Dimarco and Pareschi 2010).

Extension to the Boltzmann equation

We conclude by showing how the scheme modifies in the case of the full Boltzmann equation. The only modifications concern Step 1 of the above algorithm. In this case the starting point is the AP scheme (8.31) for the collision step. Assuming now the convex representation (8.39)

$$f^n = \tilde{f}^n + \beta^n M[f^n]$$

with $\tilde{f}^n = (1-\beta^n) f_p^n$ in (8.31) we have

$$\tilde{f}^* + \beta^* M[f^n] = A_0(\tilde{f}^n + \beta^n M[f^n]) + A_1 f_1^n + A_2 M[f^n] \tag{8.43}$$

where

$$\begin{aligned} f_1^n = \frac{P(f^n, f^n)}{\mu} \quad &= \quad \frac{1}{\mu}\Big[P(\tilde{f}^n, \tilde{f}^n) + \beta^n P(\tilde{f}^n, M[f^n]) \\ &\qquad + \beta^n P(M[f^n], \tilde{f}^n) + (\beta^n)^2 \mu M[f^n]\Big]. \end{aligned}$$

Collecting the various terms we obtain the evolution equation

$$\begin{aligned} \tilde{f}^* = A_0 \tilde{f}^n + \frac{A_1}{\mu}\Big[&P(\tilde{f}^n, \tilde{f}^n) + \beta^n P(\tilde{f}^n, M[f^n]) \\ &+ \beta^n P(M[f^n], \tilde{f}^n)\Big] \end{aligned} \tag{8.44}$$

and the equilibrium fraction

$$\beta^* = A_0 \beta^n + A_1 (\beta^n)^2 + A_2. \tag{8.45}$$

Note that $\beta^* \to 1$ as $\varepsilon \to 0$ and therefore in the limit we have a pure deterministic method. Finally using the fact that $\tilde{f}^* = (1-\beta^*) f_p^*$ and (8.45) the non equilibrium particle density satisfies

$$f_p^* = p_1 f_p^n + p_2 \left[q_1 \frac{P(f_p^n, f_p^n)}{\mu} + q_2 \frac{P(f_p^n, M[f^n]) + P(M[f^n], f_p^n)}{2\mu} \right], \tag{8.46}$$

in which

$$p_1 = \frac{A_0}{A_0 + A_1(1+\beta^n)}, \quad p_2 = \frac{A_1(1+\beta^n)}{A_0 + A_1(1+\beta^n)}, \tag{8.47}$$

$$q_1 = \frac{1-\beta^n}{1+\beta^n}, \quad q_2 = \frac{2\beta^n}{1+\beta^n}. \tag{8.48}$$

Note that, $p_1 \geq 0$, $p_2 \geq 0$, $p_1 + p_2 = 1$, $q_1 \geq 0$, $q_2 \geq 0$, $q_1 + q_2 = 1$. Therefore (8.46) is a convex combination of probability density that is suitable for the construction of DSMC methods. See Caflisch, Chen, Luo and Pareschi (2006) for space non homogenous results based on the above hybrid scheme for the Boltzmann equation.

9. Concluding remarks

The development of approximation methods for solving the Boltzmann equation has a long history which goes back to Hilbert, Chapmann and Enskog (Cercignani 1988) at the beginning of the last century. Only later, starting in the 70s with the pioneering works by Chorin (1972) and Sod (1977), the problem has been tackled numerically with particular care to accuracy and computational cost. After those pioneering works there has been an enormous amount of literature on the subject, with a strong increase in recent years. This made it virtually impossible to give a comprehensive survey on such a vast research topic.

The focus of the survey was on deterministic methods, therefore we did not cover Monte-Carlo techniques (Bird 1994, Nanbu 1980) except for few remarks on their use coupled with deterministic solvers in the development of hybrid schemes. Compared to DSMC techniques, deterministic methods offer clear advantages for problems where high accuracy and low noise are required. In addition, the possibility to compute accurate solutions makes them an important source of validation for large-scale DSMC simulations. We expect that future progress, both in more powerful computers and improved numerical algorithms, will continue to act in favor of deterministic methods.

There are several important aspects concerning the numerical solution of kinetic equations that we skipped or quickly mentioned in the present survey, here is a non exhaustive list of them.

- Other deterministic methods for the the Boltzmann integral. Several other approaches have been pursued to discretize the collision operator, such as finite difference methods (Aoki 1989, Ohwada 1993), methods based on polar coordinates (Preziosi and Longo 1997), discontinuous Galerkin methods (Majorana 2011, Alekssenko and Josyula 2012), wavelet based methods (Antoine and Lemou 2003, Tran 2013) and pseudo-spectral discretization (Ghiroldi and Gibelli 2014). See also the books edited by Bellomo and Gatignol (2003) and Degond et al. (2004)
- AP schemes for other kind of asymptotic behaviors. For instance plasmas with strong magnetic field and drift limits (Crouseilles and

Lemou 2011, Hauck, Chacon and del Castillo-Negrete 2014), highly oscillatory Vlasov-Poisson models (Crouseilles, Lemou and Mehats 2013) and high-field regimes for kinetic semiconductors equations (Jin and Wang 2013). We refer to the recent review by Degond (2014) for further examples.

- Numerical treatment of boundary conditions. It depends on the geometry of the domain and on the details of the space discretization. In particular, adequate space discretization are necessary in presence of boundary layers (Sone, Ohwada and Aoki 1989). We refer to Carrillo, Gamba, Majorana and Shu (2006), Gamba and Tharkabhushanam (2010) and Filbet (2012) for a numerical treatment of boundary conditions using deterministic schemes. Few studies have been addressed to AP schemes for boundary value problems (Lemou and Mehats 2012, Jin and Levermore 1993).
- Well-balanced techniques for stationary flows. Stationary solutions, thanks to averaging procedures, are usually computed efficiently with Monte Carlo methods (Bird 1994). However, when high accuracy is required one may be interested in schemes aimed to capture the stationary state (Greenberg and Leroux 1996, Botchorishvili, Perthame and Vasseur 2003). We refer to Gosse (2012) and Gosse (2013) for their application to kinetic equations.

We also mention the following topics which, while being connected to the content of this review, have not been discussed in the text.

- Moment based methods. The problem of finding high order closures to the moment system for small and moderate Knudsen numbers has been tackled by several authors with the goal to avoid the expensive solution of the kinetic equation (Müller and Ruggeri 1993, Struchtrup 2005). The numerical discretization of the resulting systems, however, may pose new difficulties (Jin, Pareschi and Slemrod 2002, Rana, Torrilhon and Struchtrup 2013). In particular, when a large number of moments is considered the methods are closely related to discrete velocity models for the BGK equation.
- Kinetic and relaxation schemes, Lattice Boltzmann methods. The asymptotic procedure that leads from a kinetic equation to its corresponding fluid or diffusion limit can be used as a framework for the derivation of new schemes for the limiting equations. This idea is at the basis of the kinetic schemes for the compressible Euler (Deshpande 1986, Perthame 1990) and Navier-Stokes equations (Xu 2001), the relaxation schemes for conservation laws (Jin and Xin 1995), and the Lattice Boltzmann schemes for the incompressible Navier-Stokes equations (Succi 2001, Banda, Klar, Pareschi and Seaïd 2008).

In recent years, kinetic equations have found applications in new emerging

areas like car traffic flows (Klar and Wegener 1997), tumor immune cells competition (Bellomo and Bellouquid 2004), bacterial movement (Perthame 2007), wealth distributions (Cordier et al. 2005), supply chains (Armbruster, Degond and Ringhofer 2007), flows on networks (Herty and Ringhofer 2011), flocking dynamics (Ha and Tadmor 2008) and many other. Surveys of applications to socio-economic and life sciences can be found in Naldi, Pareschi and Toscani (2010) and Pareschi and Toscani (2013). These represent new emerging fields where the construction of accurate numerical methods for kinetic equations will play a major rule in the future.

Acknowledgments

We thank the members of our research groups for many helpful comments on an earlier version of the paper. We are particularly grateful to Francis Filbet, Qin Li and Jacques Schneider for their careful reading and suggestions. Of fundamental importance for the development of our research in this field has been the collaboration with Russ Caflisch, Pierre Degond, Francis Filbet, Shi Jin, Giovanni Naldi, Luc Mieussens, Clément Mouhot, Benoît Perthame, Giovanni Russo and Giuseppe Toscani. To all of them goes our sincere gratitude.

REFERENCES

A. Abdulle, W. E, B. Engquist and E. Vanden-Eijnden (2012), 'The heterogeneous multiscale method', *Acta Numerica* **21**, 1–87.

A. Alaia and G. Puppo (2011), 'A hybrid method for hydrodynamic-kinetic flow Part I: a particle-grid method for reducing stochastic noise in kinetic regimes', *J. Comput. Phys.* **230**(14), 5660–5683.

A. Alaia and G. Puppo (2012), 'A hybrid method for hydrodynamic-kinetic flow—Part II—Coupling of hydrodynamic and kinetic models', *J. Comput. Phys.* **231**(16), 5217–5242.

A. Alekssenko and E. Josyula (2012), Deterministic solution of the Boltzmann equation using a discontinuous Galerkin velocity discretization, in *Proceedings of the 28th International Symposium on Rarefied Gas Dynamics* (A. C. P. A. I. of Physics, ed.), Vol. 1501, pp. 279–286.

X. Antoine and M. Lemou (2003), 'Wavelet approximations of a collision operator in kinetic theory', *C. R. Acad. Sci. Paris. Ser. I* **337**, 353–358.

K. Aoki (1989), Numerical analysis of rarefied gas flows by finite-difference method, in *AIAA, Rarefied Gas Dynamics: Theoretical and Computational Techniques* (E. Muntz, D. Weaver, and D. Campbell, eds), Washington, pp. 297–322.

D. Armbruster, P. Degond and C. Ringhofer (2007), 'Kinetic and fluid models for supply chains supporting policy attributes', *Bulletin of the Institute of Mathematics* **2**, 433–460.

U. Ascher, S. Ruuth and R. Spiteri (1997), 'Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations', *Appl. Numer. Math.* **25**(2-3), 151–167. Special issue on time integration (Amsterdam, 1996).

A. Astillero and A. Santos (2004), A granular fluid modeled as a driven system of elastic hard spheres, in *The Physics of Complex Systems: New Advances and Perspectives* (F. Mallamace and H. Stanley, eds), Vol. 155, IOS Press.

B. Ayuso, J. Carrillo and C.-W. Shu (2011), 'Discontinuous Galerkin methods for the one-dimensional Vlasov-Poisson system', *Kinet. Relat. Models* **4**(4), 955–989.

L. Baker and N. Hadjiconstantinou (2005), 'Variance reduction for Monte Carlo solutions of the Boltzmann equation', *Phys. Fluids* **17**, 051703.

M. Banda, A. Klar, L. Pareschi and M. Seaïd (2008), 'Lattice-Boltzmann type relaxation systems and high order relaxation schemes for the incompressible Navier-Stokes equations', *Math. Comp.* **77**(262), 943–965.

C. Bardos, F. Golse and D. Levermore (1991), 'Fluid dynamic limits of kinetic equations I: Formal derivations', *J. Stat. Phys.* **63**, 323–344.

C. Bardos, F. Golse and D. Levermore (1993), 'Fluid dynamic limits of kinetic equations II: Convergence proofs for the Boltzmann equation', *Commun. Pure Appl. Math.* **46**, 667–753.

N. Bellomo and A. Bellouquid (2004), 'From a class of kinetic models to the macroscopic equations for multicellular systems in biology', *Discrete Contin. Dyn. Syst. Ser. B* **4**, 59–80.

N. Bellomo and R. Gatignol, eds (2003), *Lecture notes on the discretization of the Boltzmann equation*, Vol. 63 of *Series on Advances in Mathematics for Applied Sciences*, World Scientific Publishing Co. Inc., River Edge, NJ.

N. Bellomo, M. Lachowicz, J. Polewczak and G. Toscani (1991), *Mathematical topics in nonlinear kinetic theory, II The Enskog equation.*, World Scientific, London.

M. Bennoune, M. Lemou and L. Mieussens (2008), 'Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier-Stokes asymptotics', *J. Comp. Phys.* **227**, 3781–3803.

N. Besse (2004), 'Convergence of a semi-lagrangian scheme for the one dimensional Vlasov-Poisson system', *SIAM J. Num. Anal.* **42**, 350–382.

N. Besse and E. Sonnendrücker (2003), 'Semi-Lagrangian schemes for the Vlasov equation on an unstructured mesh of phase space', *J. Comp. Phys.* **191**, 341–376.

P. Bhatnagar, E. Gross and M. Krook (1954), 'A model for collision processes in gases i. small amplitute processes in charged and neutral one component systems', *Phys. Rev.* **94**, 511–525.

G. Bird (1994), *Molecular gas dynamics and direct simulation of gas flows*, Clarendon Press, Oxford.

C. Birdsall and A. Langdon (1991), *Plasma Physics via Computer Simulation*, Inst. of Phys. Publishing, Bristol/Philadelphia.

A. Bobylev (1975), 'Exact solutions of the Boltzmann equation (russian)', *Dokl. Akad. Nauk. S.S.S.R.* **225**, 1296–1299.

A. Bobylev (1988), 'The theory of the nonlinear spatially uniform Boltzmann equation for maxwell molecules', *Math. Phys. Reviews* **7**, 111–233.

A. Bobylev and S. Rjasanow (1997), 'Difference scheme for the Boltzmann equation based on the fast fourier transform', *European J. Mech. B Fluids* **16**, 293–306.

A. Bobylev and S. Rjasanow (1999), 'Fast deterministic method of solving the Boltzmann equation for hard spheres', *Eur. J. Mech. B Fluids* **18**, 869–887.

A. Bobylev and S. Rjasanow (2000), 'Numerical solution of the Boltzmann equation using a fully conservative difference scheme based on the fast Fourier transform', *Transport Theory Statist. Phys.* **29**(3-5), 289–310.

A. Bobylev, J. Carrillo and I. Gamba (2000), 'On some properties of kinetic and hydrodynamics equations for inelastic interactions', *J. Statist. Phys.* **98**, 743–773.

O. Bokanowski and M. Lemou (2005), 'Fast multipole method for multivariable integrals', *SIAM J. Numer. Anal.* **42**, 2098–2117.

J. Boris and D. Book (1973), 'Flux-corrected transport i. SHASTA, a fluid transport algorithm that works', *J. Compu. Phys.* **11**, 38–69.

S. Boscarino, L. Pareschi and G. Russo (2013), 'Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit', *SIAM J. Sci. Comp.* **35**, A22–A51.

R. Botchorishvili, P. Perthame and A. Vasseur (2003), 'Equilibrium schemes for scalar conservation laws with stiff sources', *Math. Comp.* **72**(241), 131–157 (electronic).

F. Bouchut and B. Perthame (1993), 'A BGK model for small prandtl numbers in the navier-stokes approximation', *J. Stat. Phys.* **71**, 191–207.

F. Bourgat, P. LeTallec, B. Perthame and Y. Qiu (1992), *Coupling Boltzmann and Euler equations without overlapping*, AMS, Providence, RI.

J. Bourgat, P. LeTallec and M. Tidriri (1996), 'Coupling Boltzmann and navier-stokes equations by friction', *J. Comput. Phys.* **127**, 227–245.

J. Broadwell (1964), 'Shock structure in a simple discrete velocity gas', *Phys. Fluids* **7**, 1243–1247.

C. Buet (1996), 'A discrete velocity scheme for the Boltzmann operator of rarefied gas dynamics', *Trans. Theo. Stat. Phys.* **25**, 33–60.

C. Buet and S. Cordier (1999), 'Numerical analysis of conservative and entropy schemes for the Fokker-Planck-Landau equation', *SIAM J. Numer. Anal.* **36**(3), 953–973 (electronic).

C. Buet and S. Cordier (2007), 'An asymptotic preserving scheme for hydrodynamics radiative transfer models: numerics for radiative transfer', *Numer. Math.* **108**(2), 199–221.

C. Buet, S. Cordier and P. Degond (1998), 'Regularized Boltzmann operators. Simmolation methods in kinetic theory', *Comp. Math. App.* **35**, 55–74.

J. Burt and I. Boyd (2008), 'A low diffusion particle method for simulating compressible inviscid flows', *J. Comp. Phys.* **227**, 4653–4670.

J. Burt and I. Boyd (2009), 'A hybrid particle approach for continuum and rarefied flow simulation', *J. Comp. Phys.* **228**, 460–475.

H. Cabannes, R. Gatignol and L. Luo (2003), *The discrete Boltzmann equation (theory and applications)*, Henri Cabannes, Paris. Revised from the lecture notes given at the University of California, Berkeley, CA, 1980, `http://henri.cabannes.free.fr/Cours_de_Berkeley.pdf`.

R. Caflisch (1980), 'The fluid dynamical limit of the nonlinear Boltzmann equation', *Commun. Pure Appl. Math.* **33**, 651–666.

R. Caflisch (1988), 'Monte Carlo and quasi-Monte Carlo methods', *Acta Numerica* pp. 1–49.

R. Caflisch, H. Chen, E. Luo and L. Pareschi (2006), A hybrid method that interpolates between DSMC and CFD, in *44TH AIAA Aerospace Sciences Meeting and Exhibit*, Reno, pp. AIAA–2006–987.

R. Caflisch, S. Jin and G. Russo (1997), 'Uniformly accurate schemes for hyperbolic systems with relaxation', *SIAM J. Numer. Anal.* **34**, 246–281.

R. Caflisch, C. Wang, G. Dimarco, B. Cohen and A. Dimits (2008), 'A hybrid method for accelerated simulation of Coulomb collisions in a plasma', *Multiscale Model. Simul.* **7**(2), 865–887.

M. Campos Pinto and M. Mehrenberger (2008), 'Convergence of an adaptive scheme for the one dimensional Vlasov-Poisson system', *Num. Math.* **108**, 407–444.

C. Canuto, M. Hussaini, A. Quarteroni and T. Zang (1988), *Spectral methods in fluid dynamics*, Springer Series in Computational Physics, Springer-Verlag, New York.

T. Carleman (1932), 'Sur la théorie de l"équation intégrodifférentielle de Boltzmann', *Acta Math.*

T. Carleman (1957), *Problemes mathematiques dans la theorie cinetique des gas*, Publ. Sci. Inst. Mittag-Leffler, Almqvist-Wiksell, Upsala.

J. A. Carrillo and F. Vecil (2007), 'Nonoscillatory interpolation methods applied to Vlasov-based models', *SIAM J. Sci. Comp.* **29**, 11791206.

J. Carrillo, I. Gamba, A. Majorana and C.-W. Shu (2006), '2D semiconductor device simulations by WENO-Boltzmann schemes: efficiency, boundary conditions and comparison to Monte Carlo methods', *J. Comput. Phys.* **214**(1), 55–80.

J. Carrillo, T. Goudon, P. Lafitte and F. Vecil (2008), 'Numerical schemes of diffusion asymptotics and moment closures for kinetic equations', *J. Sci. Comput.* **36**(1), 113–149.

K. Case and P. Zweifel (1967), *Linear transport theory*, Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

C. Cercignani (1985), 'Sur des critères d'existence globale en théorie cinétique discrète', *C. R. Acad. Sc. Paris* **3**, 89–92.

C. Cercignani (1988), *The Boltzmann equation and its applications*, Springer Verlag, New York.

C. Cercignani, R. Illner and M. Pulvirenti (1994), *The mathematical theory of dilute gases*, Vol. 106, Applied Mathematical Sciences.

F. Charles, B. Després and M. Mehrenberger (2013), 'Enhanced convergence estimates for semi-lagrangian schemes. application to the Vlasov-Poisson equation', *SIAM J. Num. Anal.* **2**, 840–863.

C. Cheng and G. Knorr (1976), 'The integration of the Vlasov equation in configuration space', *Comput. Phys. Comm.* **22**, 330–335.

Y. Cheng, I. Gamba and J. Proft (2012), 'Positivity-preserving discontinuous Galerkin schemes for linear Vlasov-Boltzmann transport equations', *Math. Comp.* **81**(277), 153–190.

A. Chorin (1972), 'Numerical solution of Boltzmann's equation', *Comm. Pure Appl. Math.* pp. 171–186.

B. Cockburn, C. Johnson, C.-W. Shu and E. Tadmor (1998), *Advanced numerical approximation of nonlinear hyperbolic equations*, Vol. 1697 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin. Papers from the C.I.M.E. Summer School held in Cetraro, June 23–28, 1997, Edited by Alfio Quarteroni, Fondazione C.I.M.E.. [C.I.M.E. Foundation].

J. Cooley and J. Tukey (1965), 'An algorithm for the machine calculation of complex Fourier series', *Math. Comput.* **19**, 297–301.

S. Cordier, L. Pareschi and G. Toscani (2005), 'On a kinetic model for a simple market economy', *J. Stat. Phys.* **120**, 253–277.

F. Coron and B. Perthame (1991), 'Numerical passage from kinetic to fluid equations', *SIAM J. Numer. Anal.* **28**, 26–42.

A. Crestetto, N. Crouseilles and M. Lemou (2012), 'Kinetic/fluid micro-macro numerical schemes for Vlasov-Poisson-BGK equation using particles', *Kin. Rel. Mod.* **5**, 787– 816.

N. Crouseilles and M. Lemou (2011), 'An asymptotic preserving scheme based on a micro-macro decomposition for collisional Vlasov equations: diffusion and high-field scaling limits', *Kin. Rel. Mod.,* **4**, 441–477.

N. Crouseilles, M. Lemou and F. Mehats (2013), 'Asymptotic preserving schemes for highly oscillatory Vlasov-Poisson equations', *J. Comp. Phys.* **248**, 287308.

N. Crouseilles, M. Mehrenberger and E. Sonnendrücker (2010), 'Conservative semilagrangian schemes for Vlasov-type equations', *J. Comput. Phys.* **229**, 1927–1953.

N. Crouseilles, T. Respaud and E. Sonnendrücker (2009), 'A forward semilagrangian scheme for the numerical solution of the Vlasov equation', *J. Comput. Phys.* **180**, 1730–1745.

P. Degond (2014), 'Asymptotic-preserving schemes for fluid models of plasmas', *Panoramas et Syntheses*.

P. Degond and G. Dimarco (2012), 'Fluid simulations with localized Boltzmann upscaling by direct Monte Carlo.', *J. Comp. Phys.* **231**, 2414–2437.

P. Degond, G. Dimarco and L. Mieussens (2007), 'A moving interface method for dynamic kinetic-fluid coupling', *J. Comput. Phys.* **227**(2), 1176–1208.

P. Degond, G. Dimarco and L. Mieussens (2010), 'A multiscale kinetic-fluid solver with dynamic localization of kinetic effects.', *J. Comp. Phys.* **229**, 4097–4133.

P. Degond, G. Dimarco and L. Pareschi (2011), 'The moment-guided Monte Carlo method', *Internat. J. Numer. Methods Fluids* **67**(2), 189–213.

P. Degond, S. Jin and L. Mieussens (2005), 'A smooth transition between kinetic and hydrodynamic equations', *J. Comp. Phys.* **209**, 665–694.

P. Degond, L. Pareschi and G. Russo, eds (2004), *Modeling and computational methods for kinetic equations*, Modeling and Simulation in Science, Engineering and Technology, Birkhäuser Boston Inc., Boston, MA.

J. Deng (2014), 'Implicit asymptotic preserving schemes for semiconductor Boltzmann equation in the diffusive regime', *Int. J. Num. Anal. Model.* **11**, 1–23.

S. Deshpande (1986), Kinetic theory based new upwind methods for inviscid compressible flows, Technical report, AIAA Paper 86-0275.

L. Desvillettes and S. Mischler (1996), 'About the splitting algorithm for boltzmann and BGK equations', *Math. Mod. & Meth. in App. Sci.* **6**, 1079–1101.

B. Dia and M. Schatzman (1996), 'Commutateurs de certains semi-groupes holomorphes et applications aux directions alternèes', *M2AN Math. Model. Num. Anal.* **30**, 343–383.

G. Dimarco (2013), 'The hybrid moment guided Monte Carlo method for the Boltzmann equation', *Kin. Rel. Mod.* **6**, 291–315.

G. Dimarco and R. Loubère (2013*a*), 'Towards an ultra efficient kinetic scheme. Part I: Basics on the BGK equation.', *J. Comput. Phys.* **255**, 680–698.

G. Dimarco and R. Loubère (2013*b*), 'Towards an ultra efficient kinetic scheme. Part II: The high order case.', *J. Comput. Phys.* **255**, 699–719.

G. Dimarco and L. Pareschi (2006), 'Hybrid multiscale methods. I. Hyperbolic relaxation problems', *Commun. Math. Sci.* **4**(1), 155–177.

G. Dimarco and L. Pareschi (2007), 'Hybrid multiscale methods. II. Kinetic equations', *Multiscale Model. Simul.* **6**(4), 1169–1197.

G. Dimarco and L. Pareschi (2010), 'Fluid solver independent hybrid methods for multiscale kinetic equations', *SIAM J. Sci. Comput.* **32**(2), 603–634.

G. Dimarco and L. Pareschi (2011), 'Exponential Runge-Kutta methods for stiff kinetic equations', *SIAM J. Numer. Anal.* **49**(5), 2057–2077.

G. Dimarco and L. Pareschi (2012), 'High order asymptotic-preserving schemes for the Boltzmann equation', *C. R. Math. Acad. Sci. Paris* **350**(9-10), 481–486.

G. Dimarco and L. Pareschi (2013), 'Asymptotic preserving implicit-explicit Runge-Kutta methods for nonlinear kinetic equations', *SIAM J. Numer. Anal.* **51**(2), 1064–1087.

G. Dimarco, L. Pareschi and V. Rispoli (2014), 'Implicit-Explicit Runge-Kutta schemes for the Boltzmann-Poisson system for semiconductors', *Comm. Comput. Phys.* To appear.

M. Ernst (1983), *Exact solutions of the nonlinear Boltzmann equation and related kinetic models*, Nonequilibrium Phenomena I. The Boltzmann equation, North-Holland.

M. Escobedo, S. Mischler and M. Valle (2003), *Homogeneous Boltzmann equation in quantum relativistic kinetic theory*, Vol. 4 of *Electronic Journal of Differential Equations. Monograph*, Southwest Texas State University, San Marcos, TX.

L. Fainsilber, P. Kurlberg and B. Wennberg (2006), 'Lattice points on circles and discrete velocity models for the Boltzmann equation', *SIAM J. Math. Analysis* **37**, 1903–1922.

E. Fijalkow (1999), 'A numerical solution to the Vlasov equation', *Comput. Phys. Comm.* **116**(2-3), 319–328.

F. Filbet (2001), 'Convergence of a finite volume scheme for the Vlasov-Poisson system', *SIAM J. Numer. Anal.* **39**(4), 1146–1169 (electronic).

F. Filbet (2012), 'On deterministic approximation of the Boltzmann equation in a bounded domain', *Multiscale Model. and Sim.* **10**, 792–817.

F. Filbet and S. Jin (2010), 'A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources', *J. Comput. Phys.* **229**, 7625–7648.

F. Filbet and S. Jin (2011), 'An asymptotic preserving scheme for the ES-BGK model of the Boltzmann equation', *J. Sci. Comp.* **46**, 204–224.

F. Filbet and C. Mouhot (2011), 'Analysis of spectral methods for the homogeneous Boltzmann equation', *Trans. Amer. Math. Soc.* **363**, 1947–1980.

F. Filbet and L. Pareschi (2003), 'A numerical method for the accurate solution of the Fokker-Planck-Landau equation in the non homogeneous case', *J. Comput. Phys.* **186**, 457–480.

F. Filbet and T. Rey (2013), 'A rescaling velocity method for dissipative kinetic equations Applications to granular media', *J. Comput. Phys.* **248**, 177–199.

F. Filbet and G. Russo (2003), 'High order numerical methods for the space non-homogeneous Boltzmann equation', *J. Comput. Phys.* **186**, 457–480.

F. Filbet and G. Russo (2006), A rescaling velocity method for kinetic equations: the homogeneous case, in *Modelling and numerics of kinetic dissipative systems*, Nova Sci. Publ., Hauppauge, NY, pp. 191–202.

F. Filbet and G. Russo (2009), 'Semi-lagrangian schemes applied to moving boundary problems for the BGK model of rarefied gas dynamics', *Kinet. relat. models* **2**, 231–250.

F. Filbet, J. Hu and S. Jin (2012), 'A numerical scheme for the quantum Boltzmann equation efficient in the fluid regime', *M2AN Math. Model. Numer. Anal.* **46**, 443–463.

F. Filbet, C. Mouhot and L. Pareschi (2006), 'Solving the Boltzmann equation in $O(N \log N)$', *SIAM J. Sci. Comput.* **28**, 1029–1053.

F. Filbet, L. Pareschi and G. Toscani (2005), 'Accurate numerical methods for the collisional motion of (heated) granular flows', *J. Comput. Phys.* **202**, 216–235.

F. Filbet, E. Sonnendrücker and P. Bertrand (2001), 'Conservative numerical schemes for the Vlasov equation', *J. Comput. Phys.* **172**, 166–187.

E. Gabetta and L. Pareschi (1994), 'The maxwell gas and its fourier transform towards a numerical approximation', *Series on Advances in Math. for App. Scie.* **23**, 197–201.

E. Gabetta, L. Pareschi and G. Toscani (1997), 'Relaxation schemes for nonlinear kinetic equations', *SIAM J. Numer. Anal.* **34**, 2168–2194.

I. Gamba and J. Haack (2014), 'A conservative spectral method for the boltzmann equation with anisotropic scattering and the grazing collisions limit', *J. Comp. Phys.*

I. Gamba and S. Tharkabhushanam (2009), 'Spectral-Lagrangian methods for collisional models of non-equilibrium statistical states', *J. Comput. Phys.* **228**(6), 2012–2036.

I. Gamba and S. Tharkabhushanam (2010), 'Shock and boundary structure formation by spectral-Lagrangian methods for the inhomogeneous Boltzmann transport equation', *J. Comput. Math.* **28**(4), 430–460.

R. Gatignol (1975), *Théorie cinétique d'un gas à répartition discrète de vitesses*, Vol. 36 of *Lecture Notes in Phys.*, Springer Verlag, Heidelberg.

G. Ghiroldi and L. Gibelli (2014), 'A direct method for the Boltzmann equation based on a pseudo-spectral velocity space discretization', *J. Comput. Phys.* **258**, 568–584.

E. Godlewski and P. Raviart (1996), *Numerical Approximation of Hyperbolic System of Conservation Laws*, Springer-Verlag, New York.

D. Goldstein, B. Sturtevant and J. Broadwell (1989), Investigation of the motion of discrete velocity gases, Technical Report 118, Rar. Gas. Dynam., Progress in Astronautics e Aeronautics, AIAA, Washington.

F. Golse and L. Saint-Raymond (2004), 'The navier-stokes limit of the Boltzmann equation for bounded collision kernels', *Invent. Math.* **155**, 81–161.

L. Gosse (2012), 'Well-balanced schemes using elementary solutions for linear models of the Boltzmann equation in one space dimension', *Kinet. Relat. Models* **5**(2), 283–323.

L. Gosse (2013), *Computing Qualitatively Correct Approximations of Balance Laws. Exponential-Fit, Well-Balanced and Asymptotic-Preserving*, SEMA SIMAI Springer Series, Springer.

L. Gosse and G. Toscani (2002), 'An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations', *C. R. Math., Acad. Sci. Paris* **334**, 337–342.

L. Gosse and G. Toscani (2003), 'Space localization and well-balanced schemes for discrete kinetic models in diffusive regimes', *SIAM J. Numer. Anal.* **41**(2), 641–658 (electronic).

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

J. Greenberg and A. Leroux (1996), 'A well-balanced scheme for the numerical processing of source terms in hyperbolic equations', *SIAM J. Numer. Anal.* **33**(1), 1–16.

A. Grigoriev and A. Mikhalitsyn (1983), 'A spectral method of solving Boltzmann's kinetic equation numerically', *U.S.S.R. Comput. Maths. Math. Phys.* **23**, 105–111.

T. Gustaffson (1986), 'L^p-estimates for the nonlinear spatially homogeneous Boltzmann equation', *Arch. Rat. Mech. Anal.* **92**, 23–57.

S.-Y. Ha and E. Tadmor (2008), 'From particle to kinetic and hydrodynamic descriptions of flocking', *Kinet. Relat. Models* **1**(3), 415–435.

E. Hairer and G. Wanner (1996), *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Vol. 14 of *Springer Series in Computational Mathematics*, second revised edn, Springer-Verlag.

E. Hairer, C. Lubich and G. Wanner (2002), *Geometric Numerical Integration. Structure- Preserving Algorithms for Ordinary Differential Equations*, Springer, Berlin.

G. Hardy and E. Wright (1979), *An introduction to the theory of numbers*, fifth edn, The Clarendon Press Oxford University Press, New York.

C. Hauck, L. Chacon and D. del Castillo-Negrete (2014), 'Asymptotic-preserving lagrangian approach for modeling anisotropic transport in magnetized plasmas for arbitrary magnetic fields', *J. Comp. Phys.*

R. Heath, I. Gamba, P. Morrison and C. Michler (2012), 'A discontinuous Galerkin method for the Vlasov-Poisson system', *J. Comput. Phys.* **231**(4), 1140–1174.

M. Herty and C. Ringhofer (2011), 'Averaged kinetic models for flows on unstructured networks', *Kin. Rel. Models* **4**, 1081–1096.

M. Hochbruck and A. Ostermann (2010), 'Exponential integrators', *Acta Numerica* **19**, 209–286.

L. Holway (1966), 'New statistical models for kinetic theory: methods of construction', *Phys. Fluid.* **9**, 1658–1673.

T. Homolle and N. Hadjiconstantinou (2007*a*), 'Low-variance deviational simulation Monte Carlo', *J. Comp. Phys.* **19**, 041701.

T. Homolle and N. Hadjiconstantinou (2007*b*), 'A low-variance deviational simulation Monte Carlo for the Boltzmann equation', *J. Comp. Phys.* **226**, 2341–2358.

J. Hu and L. Ying (2012), 'A fast spectral algorithm for the quantum Boltzmann collision operator', *Commun. Math. Sci.* **10**, 989–999.

I. Ibragimov and S. Rjasanow (2002), 'Numerical solution of the Boltzmann equation on the uniform grid', *Computing* **69**, 163–186.

S. Jin (1995), 'Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation terms', *J. Comp. Phys.* **122**, 51–67.

S. Jin (1999), 'Efficient asymptotic-preserving (ap) schemes for some multiscale kinetic equations', *SIAM J. Sci. Comput.* **21**, 441–454.

S. Jin (2012), 'Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review.', *Riv. Mat. Univ. Parma* **3**, 177–216.

S. Jin and C. Levermore (1993), 'Fully-discrete numerical transfer in diffusive regimes', *Trans. Theo. Stat. Phys.* **22**, 739–791.

S. Jin and C. Levermore (1996), 'Numerical schemes for hyperbolic conservation laws with stiff relaxation terms', *J. Comput. Phys.* **126**(2), 449–467.

S. Jin and Q. Li (2013), 'A BGK-penalization-based asymptotic-preserving scheme for the multispecies Boltzmann equation', *Numer. Methods Partial Differential Equations* **29**(3), 1056–1080.

S. Jin and L. Pareschi (2000), 'Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes', *J. Comp. Phys.* **161**, 312–330.

S. Jin and L. Pareschi (2001), *Asymptotic-Preserving (AP) Schemes for Multiscale Kinetic Equations: a Unified Approach*, Vol. 141 of *ISNM International Series of Numerical Mathematics*, Hyperbolic Problems: Theory, Numerics, Applications, pp. 573–582.

S. Jin and L. Wang (2013), 'Asymptotic-preserving numerical schemes for the semiconductor Boltzmann equation efficient in the high field regime', *SIAM J. Sci. Comput.* **35**(3), B799–B819.

S. Jin and Z. Xin (1995), 'The relaxation schemes for systems of conservation laws in arbitrary space dimensions', *Comm. Pure Appl. Math.* **48**(3), 235–276.

S. Jin and B. Yan (2011), 'A class of asymmptotic-preserving schemes for the Fokker-Planck-Landau equation', *J. Comput. Phys.* **230**, 6420–6437.

S. Jin, L. Pareschi and M. Slemrod (2002), 'A relaxation scheme for solving the Boltzmann equation based on the Chapman-Enskog expansion', *Acta Math. Appl. Sin. Engl. Ser.* **18**(1), 37–62.

S. Jin, L. Pareschi and G. Toscani (1998), 'Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations', *SIAM J. Numer. Anal.* **35**, 2405–2439.

S. Jin, L. Pareschi and G. Toscani (2000), 'Uniformly accurate diffusive relaxation schemes for multiscale transport equations', *SIAM J. Numer. Anal.* **38**, 913–936.

C. Kennedy and M. Carpenter (2003), 'Additive Runge-Kutta schemes for convection-diffusion-reaction equations', *Appl. Numer. Math.* **44**(1-2), 139–181.

A. Klar (1998*a*), 'An asymptotic-induced scheme for non stationary transport equations in the diffusive limit', *SIAM J. Num. Anal.* **35**, 1073–1094.

A. Klar (1998*b*), 'A numerical method for kinetic semiconductor equations in the drift diffusion limit,', *J. Sci. Comp.,* **19**, 2032–2050.

A. Klar and R. Wegener (1997), 'Enskog-like kinetic models for vehicular traffic', *J. Stat. Phys.* **87**, 91–114.

V. Kolobov, R. Arslanbekov, V. Aristov, A. Frolova and S. Zabelok (2007), 'Unified solver for rarefied and continuum flows with adaptive mesh and algorithm refinement', *J. Comp. Phys.* **223**, 589–608.

P. Kowalczyk, A. Palczewski, G. Russo and Z. Walenta (2008), 'Numerical solutions of the Boltzmann equation: comparison of different algorithms', *Eur. J. Mech. B Fluids* **27**, 62–74.

P. Lafitte and G. Samaey (2012), 'Asymptotic-preserving projective integration schemes for kinetic equations in the diffusion limit', *SIAM J. Sci. Comp.* **34**, 579–602.

L. Landau (1981), *The transport equation in the case of the Coulomb interaction*, D.ter Haar Ed. Collected papers of L.D. Landau, Pergamon press, pp. 163–170.

O. Lanford III (1975), The evolution of large classical systems, in *Dynamical systems, theory and applications* (J. Moser, ed.), Vol. LNP 35, Springer, Berlin, p. 1111.

M. Lemou (2010), 'Relaxed micro-macro schemes for kinetic equations', *Compt. Rendus Math.* **348**, 455–460.

M. Lemou and F. Mehats (2012), 'Micro-macro schemes for kinetic equations including boundary layers', *SIAM J. Sci. Comp.* **34**, 134–160.

M. Lemou and L. Mieussens (2005), 'Implicit schemes for the Fokker-Planck-Landau equation', *SIAM J. Sci. Comput.* **27**, 809– 830.

M. Lemou and L. Mieussens (2008), 'A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit', *SIAM J. Sci. Comput.* **31**(1), 334–368.

C. Levermore (1996), 'Moment closure hierarchies for kinetic theories', *J. Stat. Phys.* **83**, 1021–1065.

E. Lewis and W. F. Miller (1993), *Computational methods of neutron transport*, American Nuclear Society, La Grange Park.

Q. Li and L. Pareschi (2014), 'Exponential Runge-Kutta schemes for inhomogeneous Boltzmann equations with high order of accuracy', *J. Comp. Phys.*

Q. Li, J. Hu and L. Pareschi (2014), 'Asymptotic preserving exponential methods for the Quantum Boltzmann equations with high order of accuracy', *preprint.*

P. Lions (1994), 'Compactness in Boltzmann's equation via Fourier integral operators and applications. I', *J. Math. Kyoto Univ.* **34**, 391–427.

T.-P. Liu and S.-H. Yu (2004), 'Boltzmann equation: micro-macro decompositions and positivity of shock profiles', *Comm. Math. Phys.* **246**(1), 133–179.

A. Majorana (2011), 'A numerical model of the Boltzmann equation related to the discontinuous Galerkin method', *Kinet. Relat. Models* **4**(1), 139–151.

P. Markowich and L. Pareschi (2005), 'Fast conservative and entropic numerical methods for the boson Boltzmann equation', *Num. Math.* **99**, 509–532.

P. Markowich, C. Ringhofer and C. Schmeiser (1989), *Semiconductor equations*, Springer-Verlag.

Y.-L. Martin, F. Rogier and J. Schneider (1992), 'Une méthode déterministe pour la résolution de l'équation de Boltzmann inhomogène', *C. R. Acad. Sci. Paris Sér. I Math.* **314**, 483–487.

S. Maset and M. Zennaro (2009), 'Unconditional stability of explicit exponential Runge-Kutta methods for semi-linear ordinary differential equations', *Math. Comp.* **78**, 957–967.

L. Mieussens (2000), 'Discrete Velocity Model and implicit scheme for the BGK equation of rarefied gas dynamics', *Math. Models and Meth. Appl. Sci.* **8**, 1121–1149.

L. Mieussens (2001), 'Convergence of a Discrete-Velocity Model for the Boltzmann-BGK equation', *Comp. Math. App.* **41**, 83–96.

S. Mischler (1997), 'Convergence of discrete velocity schemes for the Boltzmann equation', *Arch. Rat. Mech. Anal.* **140**, 53–77.

C. Mouhot and L. Pareschi (2004), 'Fast methods for the Boltzmann collision integral', *C. R. Math. Acad. Sci. Paris* **339**(1), 71–76.

C. Mouhot and L. Pareschi (2006), 'Fast algorithms for computing the Boltzmann collision operator', *Math. Comp.* **75**(256), 1833–1852 (electronic).

C. Mouhot, L. Pareschi and T. Rey (2013), 'Convolutive decomposition and fast summation methods for discrete-velocity approximations of the Boltzmann equation', *Math. Mod. Num. Anal.* **47**, 1515–1531.

I. Müller and T. Ruggeri (1993), *Extended thermodynamics*, Vol. 37 of *Springer Tracts in Natural Philosophy*, Springer-Verlag, New York.

G. Naldi and L. Pareschi (1998), 'Numerical schemes for kinetic equations in diffusive regimes', *Appl. Math. Lett.* **11**(2), 29–35.

G. Naldi and L. Pareschi (2000), 'Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation', *SIAM J. Numerical Analysis* **37**, 1246–1270.

G. Naldi, L. Pareschi and G. Toscani (2003), 'Spectral methods for one-dimensional kinetic models of granular flows and numerical quasi elastic limit', *ESAIM RAIRO Math. Model. Numer. Anal.* **37**, 73–90.

G. Naldi, L. Pareschi and G. Toscani, eds (2010), *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, Series: Modeling and Simulation in Science, Engineering and Technology, Birkhauser, Boston.

K. Nanbu (1980), 'Direct simulation scheme derived from the Boltzmann equation I. Monocomponent gases.', *J. Phys. Soc. Japan* **49**, 2042–2049.

A. Narayan and A. Klöckner (2009), Deterministic numerical schemes for the Boltzmann equation, Technical report, arXiv:0911.3589v1.

T. Nishida (1978), 'Fluid dynamical limit of the nonlinear Boltzmann equation at the level of the compressible euler equations', *Commun. Math. Phys.* **61**, 119–148.

A. Nogueira and B. Sevennec (2006), 'Multidimensional Farey partitions', *Indag. Math. (N.S.)* **17**(3), 437–456.

T. Ohwada (1993), 'Structure of normal shock waves: Direct numerical analysis of the Boltzmann equation for hard sphere molecules', *Phys. Fluids A* **5**, 217–234.

A. Palczewski, J. Schneider and A. Bobylev (1997), 'A consistency result for a

discrete-velocity model of the Boltzmann equation', *SIAM J. Numer. Anal.* **34**, 1865–1883.

V. Panferov and A. Heintz (2002), 'A new consistent discrete-velocity model for the Boltzmann equation', *Math. Methods Appl. Sci.* **25**, 571–593.

L. Pareschi (1998), Characteristic-based numerical schemes for hyperbolic systems with nonlinear relaxation, in *Proceedings of the IX International Conference on Waves and Stability in Continuous Media (Bari, 1997)*, number 57, pp. 375–380.

L. Pareschi (2003), Computational methods and fast algorithms for Boltzmann equations, in *Chapter 7 Lecture Notes on the discretization of the Boltzmann equation* (N. Bellomo and R. Gatignol, eds), pp. 527–548.

L. Pareschi and R. Caflisch (1999), 'Implicit Monte Carlo methods for rarefied gas dynamics I: The space homogeneous case', *J. Comput. Phys.* **154**, 90–116.

L. Pareschi and R. Caflisch (2004), 'Towards an hybrid method for rarefied gas dynamics', *IMA Vol. App. Math.* **135**, 57–73.

L. Pareschi and B. Perthame (1996), 'A spectral method for the homogeneous Boltzmann equation', *Trans. Theo. Stat. Phys.* **25**, 369–383.

L. Pareschi and G. Russo (1999), 'An introduction to Monte Carlo methods for the Boltzmann equation', *Esaim Proceedings, EDP Sciences, ESAIM* **10**, 1–38.

L. Pareschi and G. Russo (2000*a*), 'Asymptotic preserving Monte Carlo methods for the Boltzmann equation', *Transport Theory Statist. Phys.* **29**, 415–430.

L. Pareschi and G. Russo (2000*b*), 'Numerical solution of the Boltzmann equation I. Spectrally accurate approximation of the collision operator', *SIAM J. Numer. Anal.* **37**, 1217–1245.

L. Pareschi and G. Russo (2000*c*), 'On the stability of spectral methods for the homogeneous Boltzmann equation', *Trans. Theo. Stat. Phys.* **29**, 431–447.

L. Pareschi and G. Russo (2001), 'Time relaxed Monte Carlo methods for the Boltzmann equation', *SIAM J. Sci. Comput.* **23**, 1253–1273.

L. Pareschi and G. Russo (2005), 'Implicit-explicit Runge-Kutta methods and applications to hyperbolic systems with relaxation', *J. Sci. Comp.* **25**, 129–155.

L. Pareschi and G. Russo (2011), Efficient asymptotic preserving deterministic methods for the Boltzmann equation, in *Models and Computational Methods for Rarefied Flows*, AVT- 194 RTO AVT/VKI, Rhode St. Genese, Belgium.

L. Pareschi and G. Toscani (2004), Modelling and numerical methods for granular gases, in *Modeling and computational methods for kinetic equations*, Model. Simul. Sci. Eng. Technol., Birkhäuser Boston, Boston, MA, pp. 259–285.

L. Pareschi and G. Toscani (2013), *Interacting multi-agent systems. Kinetic equations and Monte Carlo methods*, Oxford University Press, USA.

L. Pareschi, G. Russo and G. Toscani (2000), 'Fast spectral methods for the Fokker-Planck-Landau collision operator', *J. Comput. Phys.* **165**, 216–236.

L. Pareschi, G. Toscani and C. Villani (2003), 'Spectral methods for the non cut-off Boltzmann equation and numerical grazing collision limit', *Numer. Math.* **93**(3), 527–548.

L. Pareschi, S. Trazzi and B. Wennberg (2008), 'Adaptive and recursive time relaxed Monte Carlo method for rarefied gas dynamics', *SIAM J. Sci. Comp.* **31**, 1379–1398.

B. Perthame (1989), 'Global existence to the BGK model of Boltzmann equation', *J. Diff. Eq.* **82**, 191–205.

B. Perthame (1990), 'Boltzmann type schemes for gas dynamics and the entropy property', *SIAM J. Numer. Anal.* **27**(6), 1405–1421.

B. Perthame (2007), *Transport Equations in Biology*, Frontiers in mathematics, Springer, Boston.

S. Pieraccini and G. Puppo (2007), 'Implicit-Explicit schemes for BGK kinetic equations', *Journal of Scientific Computing* **32**, 1–28.

S. Pieraccini and G. Puppo (2012), 'Microscopically implicitmacroscopically explicit schemes for the BGK equation', *Journal of Computational Physics* **231**, 299–327.

T. Platkowski and W. Walús (2000), 'An acceleration procedure for discrete velocity approximation of the Boltzmann collision operator', *Comp. Math. Appl.* **39**, 151–163.

T. Pöschel and N. Brilliantov (2003), *Granular Gas Dynamics*, Vol. 624 of *Lecture Notes in Physics*, Springer-Verlag, New York.

L. Preziosi and E. Longo (1997), 'On a conservative polar discretization of the Boltzmann equation', *Japan J. Indust. Appl. Math.* **14**(3), 399–435.

D. Pullin (1980), 'Direct simulation methods for compressible inviscid ideal gas flow', *J. Comput. Phys.* **34**, 231–244.

J.-M. Qiu and C.-W. Shu (2011), 'Positivity preserving semi-lagrangian discontinuous Galerkin formulation: Theoretical analysis and application to the VlasovPoisson system', *J. Comp. Phys.* **230**, 8386–8409.

G. Radtke and N. Hadjiconstantinou (2009), 'Variance-reduced particle simulation of the Boltzmann transport equation in the relaxation-time approximation.', *Phys. Rev. E* **79**, 056711.

G. Radtke, N. Hadjiconstantinou and W. Wagner (2011), 'Low-noise Monte Carlo simulation of the variable hard sphere gas.', *Phys. Fluids* **23**, 030606.

G. Radtke, J.-P. Péraud and N. Hadjiconstantinou (2013), 'On efficient simulations of multiscale kinetic transport', *Phil. Trans. R. Soc. A* **23**, 030606.

A. Rana, M. Torrilhon and H. Struchtrup (2013), 'A robust numerical method for the R13 equations of rarefied gas dynamics: application to lid driven cavity', *J. Comput. Phys.* **236**, 169–186.

T. Respaud and E. Sonnendrücker (2011), 'Analysis of a new class of forward semi-Lagrangian schemes for the 1d Vlasov Poisson equations', *Num. Math.* **118**, 329–366.

S. Rjasanow and W. Wagner (1996), 'A stochastic weighted particle method for the Boltzmann equation', *J. Comput. Phys.* **124**, 243–253.

S. Rjasanow and W. Wagner (2001), 'Simulation of rare events by the stochastic weighted particle method for the Boltzmann equation', *Math. Comput. Modelling* **33**, 907–926.

S. Rjasanow and W. Wagner (2006), *Stochastic Numerics for the Boltzmann Equation*, Vol. 37 of *Springer series in computational mathematics*, Springer.

F. Rogier and J. Schneider (1994), 'A direct method for solving the Boltzmann equation', *Trans. Theo. Stat. Phys.* **23**, 313–338.

P. Santagati, G. Russo and S.-B. Yun (2012), 'Convergence of a semi-Lagrangian

scheme for the BGK model of the Boltzmann equation', *SIAM J. Num. Anal.* **50**, 1111–1135.

J. Schneider (1993), *Une méthode déterministe pour la résolution de l'équation de Boltzmann*, PhD thesis, University Pierre et Marie Curie (Paris VI).

J. Schneider (1996), 'Direct coupling of fluid and kinetic equations', *Trans. Theo. Stat. Phys.* **25**, 681–698.

C.-W. Shu (2009), 'High order weighted essentially non-oscillatory schemes for convection dominated problems', *SIAM Review* **51**, 82–126.

C.-W. Shu and S. Osher (1989), 'Efficient implementation of essentially non-oscillatory shock-capturing schemes, II.', *J. Comp. Phys.* **83**, 32–78.

G. Sod (1977), 'A numerical solution of Boltzmann's equation', *Comm. Pure Appl. Math.* **30**, 391–419.

Y. Sone, T. Ohwada and K. Aoki (1989), '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**, 363–370.

E. Sonnendrücker (2013), Numerical methods for Vlasov equations, Technical report, MPI TU Munich. (`http://www-m16.ma.tum.de/foswiki/pub/M16/Allgemeines/NumMethVlasov/Num-Meth-Vlasov-Notes.pdf`).

E. Sonnendrücker, J. Roche, P. Bertrand and A. Ghizzo (1999), 'The semi-lagrangian method for the numerical resolution of the Vlasov equation', *J. Comput. Phys.* **149**, 201–220.

H. Spohn (1991), *Large scale dynamics of interacting particles*, Springer, Berlin.

G. Strang (1968), 'On the construction and the comparison of difference schemes', *SIAM J. Numer. Anal.* **5**, 506–517.

H. Struchtrup (2005), *Macroscopic transport equations for rarefied gas flows*, Interaction of Mechanics and Mathematics, Springer, Berlin. Approximation methods in kinetic theory.

S. Succi (2001), *The lattice Boltzmann equation for fluid dynamics and beyond*, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York. Oxford Science Publications.

F. Tcheremissine (2006), 'Solution to the Boltzmann kinetic equation for high-speed flows', *Comp. Math. and Math. Phys.* **46**, 315–329.

S. Tiwari (1998*a*), 'Coupling of the Boltzmann and euler equations with automatic domain decomposition', *J. Comput. Phys.* **144**, 710–726.

S. Tiwari (1998*b*), 'Coupling of the Boltzmann and Euler equations with automatic domain decomposition', *J. Comput. Phys.* **144**, 710–726.

S. Tiwari and A. Klar (1998), 'An adaptive domain decomposition procedure for Boltzmann and euler equations', *J. Comp. Appl. Math.* **90**, 223–237.

M.-B. Tran (2013), 'Nonlinear approximation theory for the homogeneous Boltzmann equation', *arXiv:1305.1667* p. 82.

E. Uehling and G. Uhlenbeck (1933), 'Transport phenomena in Einstein-Bose and Fermi-Dirac gases. i', *Phys. Rev.* **43**, 552–561.

D. Valougeorgis and S. Naris (2003), 'Acceleration schemes of the discrete velocity method: Gaseous flows in rectangular microchannels', *SIAM J. Sci. Comput.* **25**, 534–552.

C. Villani (2002), *A survey of mathematical topics in kinetic theory*, Handbook of fluid mechanics, S. Friedlander and D. Serre, Eds. Elsevier Publ.

H. Wijesinghe and N. Hadjiconstantinou (2004), 'Discussion of hybrid atomistic-continuum methods for multiscale hydrodynamics', *Int. J. Multi. Comp. Eng.* **2**, 189–202.

E. Wild (1951), 'On Boltzmann's equation in the kinetic theory of gases', *Proc. Camb. Phil. Soc.* **47**, 602–609.

L. Wu, C. White, T. Scanlon, J. Reese and Y. Zhang (2013), 'Deterministic numerical solutions of the Boltzmann equation using the fast spectral method', *J. of Comp. Phys.* **250**, 27–52.

K. Xu (2001), 'A gas-kinetic BGK scheme for the Navier-Stokes equations and its connection with artificial dissipation and Godunov method', *J. Comput. Phys.* **171**(1), 289–335.