Moment preserving Fourier-Galerkin spectral methods and application to the Boltzmann equation

Lorenzo Pareschi, Thomas Rey

MOMENT PRESERVING FOURIER-GALERKIN SPECTRAL METHODS AND APPLICATION TO THE BOLTZMANN EQUATION

LORENZO PARESCHI* AND THOMAS REY†

Abstract. Spectral methods, thanks to the high accuracy and the possibility of using fast algorithms, represent an effective way to approximate collisional kinetic equations in kinetic theory. On the other hand, the loss of some local invariants can lead to the wrong long time behavior of the numerical solution. We introduce in this paper a novel Fourier-Galerkin spectral method that improves the classical spectral method by making it conservative on the moments of the approximated distribution, without sacrificing its spectral accuracy or the possibility of using fast algorithms. The method is derived directly using a constrained best approximation in the space of trigonometric polynomials and can be applied to a wide class of problems where preservation of moments is essential. We then apply the new spectral method to the evaluation of the Boltzmann collision term, and prove spectral consistency and stability of the resulting Fourier-Galerkin approximation scheme. Various numerical experiments illustrate the theoretical findings.

KEYWORDS: Boltzmann equation, Fourier-Galerkin spectral method, conservative methods, spectral accuracy, stability, Maxwellian equilibrium.

2010 MATHEMATICS SUBJECT CLASSIFICATION: 76P05, 65N35, 82C40.

1. Introduction.

Spectral methods for collisional kinetic equations have a long history. They have been originally inspired by the pioneering works on the Fourier transformed Boltzmann equation for Maxwell molecules by A. Bobylev [5] which later inspired the first methods based on finite difference discretizations of the Fourier transform [6–8]. In such approaches the main purpose was to exploit the simplified form of the equation in Fourier space rather than the construction of a method providing spectral accuracy. The first Fourier-Galerkin type spectral method was introduced in the same years in [39, 42, 43]. Thanks to the new formalism, it was possible to prove spectral accuracy and consistency of the method. In particular, the method lent itself to be generalized to other collisional kinetic equations such as the Landau equation, the inelastic Boltzmann equation for granular gases and the quantum Boltzmann equation [14, 17, 21, 44, 45].

While in the case of the Landau equation the development of fast algorithms with spectral accuracy was achieved immediately due to the convolutional structure of the collision operator, in the case of the Boltzmann equation it represented a major breakthrough achieved later in [19, 35]. Subsequent developments that have extended the construction of fast algorithms to inelastic collisions, quantum interactions and general collisional kernels have been obtained in [23, 29, 31, 49]. Due to these advances, spectral methods have gained quite a bit of popularity in numerical simulations of the space non homogeneous Boltzmann equation [22, 27, 34] and today are successfully used in realistic multidimensional applications [15, 32, 49, 50]. For a more comprehensive introduction to this class of methods and further references we refer the interested reader to the survey in [16].

The main advantages of a Fourier-Galerkin type approach are the spectral accuracy for smooth solutions and the possibility to use fast algorithms that mitigate the curse of dimensionality. On the other hand, they typically lead to the loss of most physical properties of the Boltzmann equation, namely positivity, conservations, en-

*Department of Mathematics & Computer Science, University of Ferrara, Via Machiavelli 30, Ferrara, 44121, Italy (lorenzo.pareschi@unife.it).

†Univ. Lille, CNRS, UMR 8524, Inria – Laboratoire Paul Painlevé F-59000 Lille, France (thomas.rey@univ-lille.fr)

tropy dissipation and, as a consequence, long time behavior. The construction of deterministic numerical methods that can preserve the collisional invariants has always been a challenge in the approximation of the Boltzmann equation and related kinetic problems [2, 13, 16]. In fact, a major problem associated with deterministic methods is that the velocity space is approximated by a finite region. On the other hand, even starting from a compactly supported function in velocity, by the action of the collision term the solution becomes immediately positive in the whole velocity space. In particular, the local Maxwellian equilibrium states are characterized by exponential functions defined on the whole velocity space. Another line of research is based on the use of different orthogonal polynomials that do not require truncation of the velocity space, we refer the reader to [9, 25, 28, 48] and the references therein for more details.

Thus, at the numerical level some non physical conditions have to be imposed to keep the support of the function in velocity uniformly bounded. This can be done by neglecting collisions which will spread the support of the solution outside the finite region, as in discrete velocity models [38, 46], or by periodizing the function and the collision term, as in spectral methods [39, 42] or in finite difference schemes based on the Fourier transform [6, 7, 26]. In the former case, however, the symmetries of the Boltzmann collision operator that underlie the development of fast solvers are destroyed and a periodized formulation is therefore a necessary condition to derive computationally efficient algorithms also for discrete velocity models [36].

Spectral methods preserving some physical properties have been introduced by various authors using smoothing or renormalization techniques at different levels [7–9, 26, 39, 43]. However, these approaches typically may lead to the loss of spectral accuracy of the resulting approximation scheme. A way to overcome some of these drawbacks by keeping spectral accuracy has been proposed recently [20, 41], where Fourier-Galerkin steady state preserving spectral methods have been constructed. However, this class of methods does not exactly preserve moments and the approach can only be generalized to kinetic equations where the equilibrium state is known. On the other hand, it has been observed in [49], through numerical experiments, that the Lagrangian multiplier approach introduced in [26] when applied to the Fourier-Galerkin approximation is capable of maintaining spectral accuracy. We refer also to [1] for recent results on the convergence properties of the method in [26].

Motivated by the previous discussion, we consider in this paper a novel Fourier-Galerkin fast spectral methods that improves the classical fast spectral scheme by making it conservative on the moments of the approximated distribution, while maintaining the theoretical properties of spectral accuracy, consistency and stability. The method is derived directly starting from a constrained best approximation in the space of trigonometric polynomials and can be applied in a standard Fourier-Galerkin setting to a wide class of kinetic equations where preservation of moments is essential [47]. Due to its relevance in applications, even if the method is introduced in its generality, we discuss in details the application to the challenging case of the Boltzmann equation and show that, thanks to the new formalism, previous results on spectral consistency and stability can be extended to the present case.

The rest of the manuscript is organized as follows. In the next Section we recall some essential facts about the Boltzmann equation and the corresponding periodized space homogeneous problem. Section 3 is then devoted to the introduction of the conservative finite Fourier series as the constrained best approximation in the space of trigonometric polynomials. This permits to extend to the conservative case the classical result of spectral convergence for smooth solutions. In Section 4 we apply the new

method to the Boltzmann equation in a Fourier-Galerkin setting and prove spectral consistency and stability of the resulting scheme under a smallness assumption on the loss of conservation of the periodized collision term. To simplify the presentation and have a more self-contained treatment, the fast spectral method is summarized in Appendix A together with the details of other spectral schemes used in the numerical section. The subsequent Section illustrates through several numerical examples the performance of the new method in comparison with the classical fast spectral method [35] and the equilibrium preserving fast spectral method [20]. Finally we end the manuscript with some concluding remarks and future developments.

2. The Boltzmann equation.

2.1. Basic properties of the equation. The Boltzmann equation describes the behavior of a dilute gas of particles when the only interactions taken into account are binary elastic collisions [12, 47]. It reads for $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, $d \leq 3$

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = Q(f, f) \tag{2.1}$$

where $f(t, x, v)$ is the time-dependent particle distribution function in the phase space. The Boltzmann collision operator Q is a quadratic operator local in (t, x). The time and position acts only as parameters in Q and therefore will be omitted in its description

$$Q(f, f)(v) = \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} B(|v - v_*|, \cos\theta)\ (f'_* f' - f_* f)\ dv_*\, d\sigma. \tag{2.2}$$

In (2.2) we used the shorthand $f = f(v)$, $f_* = f(v_*)$, $f' = f(v')$, $f'_* = f(v'_*)$. The velocities of the colliding pairs (v, v_*) and (v', v'_*) can be parametrized as

$$v' = \frac{v + v_*}{2} + \frac{|v - v_*|}{2}\sigma, \qquad v'_* = \frac{v + v^*}{2} - \frac{|v - v_*|}{2}\sigma.$$

The collision kernel B is a non-negative function which by physical arguments of invariance only depends on $|v - v_*|$ and $\cos\theta = \hat{g} \cdot \sigma$ (where $\hat{g} = (v - v_*)/|v - v_*|$). It characterizes the details of the binary interactions, and has the form

$$B(|v - v_*|, \cos\theta) = |v - v_*|\sigma(|v - v_*|, \cos\theta) \tag{2.3}$$

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

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

with $\alpha = (k - 5)/(k - 1)$. The special situation $k = 5$ gives the so-called Maxwell pseudo-molecules model with

$$B(|v - v_*|, \cos\theta) = b_0(\cos\theta).$$

For the Maxwell case the collision kernel is independent of the relative velocity. For numerical purposes, a widely used model is the variable hard sphere (VHS) model introduced by Bird [4]. The model corresponds to $b_\alpha(\cos\theta) = C_\alpha$, where C_α is a positive constant, and hence

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

In the numerical test Section we will consider the Maxwell molecules case when dealing with a velocity space of dimension $d = 2$.

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

$$\int_{\mathbb{R}^d} Q(f,f)\,\phi(v)\,dv = 0, \quad \phi(v) = 1, v_1, \dots, v_d, |v|^2, \tag{2.4}$$

and satisfies the well-known Boltzmann's H theorem

$$\frac{d}{dt}\int_{\mathbb{R}^d} f\log f\,dv = \int_{\mathbb{R}^d} Q(f,f)\log(f)\,dv \le 0.$$

Boltzmann's H theorem implies that any equilibrium distribution function has the form of a locally Maxwellian distribution

$$M(\rho,u,T)(v) = \frac{\rho}{(2\pi T)^{d/2}}\exp\left\{-\frac{|u-v|^2}{2T}\right\}, \tag{2.5}$$

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

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

For further details on the physical background and derivation of the Boltzmann equation we refer to [12, 47].

In the sequel we will restrict out attention to the space homogeneous setting where $f = f(v,t)$ satisfies

$$\begin{cases} \dfrac{\partial f}{\partial t} = Q(f,f) \\ f(v,0) = f_0(v), \quad v \in \mathbb{R}^d. \end{cases} \tag{2.6}$$

In fact, the numerical solution of (2.6) contains all the major difficulties related to the Boltzmann equation (2.1), and can be easily extended to the full inhomogeneous case by splitting algorithms or IMEX methods (see [16]).

2.2. Reduction to a bounded domain and periodization. As shown in [39], we have the following

PROPOSITION 2.1. *Let the distribution function f be compactly supported on the ball $\mathcal{B}_0(R)$ of radius R centered in the origin, then*

$$\mathrm{Supp}(Q(f,f)(v)) \subset \mathcal{B}_0(\sqrt{2}R).$$

In order to write a spectral approximation which avoid superposition of periods, it is then sufficient that the distribution function $f(v)$ is restricted on the cube $[-T,T]^d$ with $T \ge (2+\sqrt{2})R$. Successively, one should assume $f(v) = 0$ on $[-T,T]^d \setminus \mathcal{B}_0(R)$ and extend $f(v)$ to a periodic function on the set $[-T,T]^d$. Let observe that the lower bound for T can be improved. For instance, the choice $T = (3+\sqrt{2})R/2$ guarantees the absence of intersection between periods where f is different from zero [42]. However, since in practice the support of f increases with time, we can just minimize the errors due to aliasing [10] with spectral accuracy. To further simplify the notation, let us take $T = \pi$ and hence $R = \lambda\pi$ with $\lambda = 2/(3+\sqrt{2})$ in the sequel.

We shall consider in the rest of the paper the following periodized, space homogeneous problem [18, 42]

$$\begin{cases} \dfrac{\partial f}{\partial t} = Q^R(f,f) \\ f(v,0) = f_0(v), \quad v \in [-\pi,\pi]^d, \end{cases} \tag{2.7}$$

where Q^R is the truncated and periodized collision term

$$Q^R(f,f)(v) = \int_{\mathcal{B}_0(2R)\times\mathbb{S}^{d-1}} B(|v-v_*|,\cos\theta)\ (f'_*f' - f_*f)\,dv_*\,d\sigma. \tag{2.8}$$

Note that, because of the reduction to a bounded domain by periodization, the collisional invariants of the original problem are lost, except for mass conservation. As a consequence the local equilibria m_∞ of problem (2.7)-(2.8) are the (piecewise) constant functions

$$m_\infty(v) := \frac{\rho}{2\pi^d}, \quad \forall\ v \in [-\pi,\pi]^d. \tag{2.9}$$

Concerning the loss of conservations of the periodized collision term in $[-\pi,\pi]^d$ we have the following result:

Proposition 2.2. *Assuming the solution to problem (2.7) satisfies for $\delta \ll 1$*

$$f(v,t) \le \delta, \quad v \in [-\pi,\pi]^d \setminus \mathcal{B}_0(R)$$

with $R = \lambda\pi$. Then, we have the bound for $\Phi(v) = (1, v_1, \dots, v_d, |v|^2)^T$

$$\left\| \int_{[-\pi,\pi]^d} Q^R(f,f)\,\Phi(v)\,dv \right\|_2 \le C\delta, \tag{2.10}$$

where $C = C(f,R)$ and $\|\cdot\|_2$ denotes the euclidean norm of the vector.

Proof. From the assumption on $f(v,t)$ we can write

$$f(v,t) = f^c(v,t) + f^\delta(v,t)$$

where $f^c(v,t)$ is compactly supported in $\mathcal{B}_0(R)$ and $f^\delta(v,t) = 0$ in $\mathcal{B}_0(R)$ with $f^\delta(v,t) \le \delta$ in $[-\pi,\pi]^d \setminus \mathcal{B}_0(R)$.

Since

$$Q^R(f,f) = Q^R(f^c,f^c) + Q^R(f^\delta,f^\delta) + Q^R(f^c,f^\delta) + Q^R(f^\delta,f^c),$$

estimate (2.10) follows from the weak form

$$\int_{[-\pi,\pi]^d} Q^R(f,f)\phi(v)\,dv = \int_{[-\pi,\pi]^d}\int_{\mathcal{B}_0(2R)\times\mathbb{S}^{d-1}} B(|v-v_*|,\cos\theta) f f_*(\phi(v') - \phi(v))\,dv_*\,d\sigma\,dv,$$

and the conservation property

$$\int_{[-\pi,\pi]^d} Q^R(f^c,f^c)\,\Phi(v)\,dv = 0.$$

∎

For the Boltzmann equation (2.6), it is well known (and precise estimates are available in [24]) that solutions f have uniform in time maxwellian upper bounds: There exists constants $C \geq 1$ and $0 < \mu \leq 1$, depending on the collision kernel B and the initial datum f_0 such that

$$0 \leq f(t,v) \leq C \frac{\rho}{(2\pi T)^{d/2}} e^{-\mu \frac{|v-u|^2}{2T}} = \frac{C}{\mu^{d/2}} M(\rho,u,T)(v) \quad \forall\, v \in \mathbb{R}^d, t > 0.$$

This, combined with the previous proposition would provide a bound on the loss of moments due to truncation. Note, however, that a similar estimate for the periodized problem (2.7) is not available since asymptotically the solution converges to the steady state (2.9).

Due to the physical relevance of the collisional invariants, which are the basis of the long time behavior (2.5), usually the truncated collision operator is modified to maintain the original conservation properties (2.4). For example, by modifying the collision kernel in order to neglect collisions originating velocities outside the bounded domain as in discrete velocity models [38]. This, however, destroying the convolutional structure of (2.8), leads to computationally inefficient quadratures that cannot take advantage of fast solvers [35,36]. In our setting, as we will illustrate, conservation are recovered as a constrained best approximation in the Fourier space which permits a standard implementation of fast algorithms.

3. Conservative approximations by trigonometric polynomials. In order to get conservative spectral methods, we construct a conservative projection of the periodized solution on the space of trigonometric polynomials following the constrained formulation approach introduced in [7] and [26] for finite difference discretizations of the Fourier transformed Boltzmann equation. In particular, we will give an explicit formulation of the trigonometric polynomial of best approximation in the least square sense, constrained by preservation of moments, and show that it preserves spectral accuracy for smooth solutions like the classical trigonometric approximation by finite Fourier series.

3.1. The finite Fourier series. Let us first set up the mathematical framework of our analysis. For the sake of the reader convenience, we shall restrain ourselves to the domain $[-\pi,\pi]^d$. Given a function $f(v) \in L^2_p([-\pi,\pi]^d)$, $d \geq 1$, we set as before its Fourier series representation as

$$f(v) = \sum_{k=-\infty}^{\infty} \hat{f}_k e^{ik\cdot v}, \qquad \hat{f}_k = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} f(v) e^{-ik\cdot v}\, dv, \tag{3.1}$$

where we use just one vector index $k = (k_1, \ldots, k_d)$ to denote the d-dimensional sums over the indexes k_j, $j = 1, \ldots, d$.

We define the space $\mathbb{P}^N$ of trigonometric polynomials of degree N in v as

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

Let $\mathcal{P}_N : L^2_p([-\pi,\pi]^d) \to \mathbb{P}^N$ be the orthogonal projection upon $\mathbb{P}^N$ in the inner product of $L^2_p([-\pi,\pi]^d)$

$$\langle f - \mathcal{P}_N f, \phi \rangle = 0, \qquad \forall\, \phi \in \mathbb{P}^N.$$

With these definitions $\mathcal{P}_N f = f_N$, where f_N is the finite Fourier series of f given by

$$f_N(v) = \sum_{k=-N}^{N} \hat{f}_k e^{ik\cdot v}, \tag{3.2}$$

Since the operator $\mathcal{P}_N$ is self-adjoint [10] the following property hold

$$\langle \mathcal{P}_N f, \varphi \rangle = \langle f, \mathcal{P}_N \varphi \rangle = \langle \mathcal{P}_N f, \mathcal{P}_N \varphi \rangle \quad \forall\ f,\ \varphi \in L^2_p([-\pi,\pi]^d).$$

We define the L^2_p-norm by

$$\|f\|^2_{L^2_p} = \langle f, f \rangle.$$

By the Parseval's identity we have

$$\|f\|^2_{L^2_p} = (2\pi)^d \sum_{k=-\infty}^{\infty} |\hat{f}_k|^2, \qquad \|f_N\|^2_{L^2_p} = (2\pi)^d \sum_{k=-N}^{N} |\hat{f}_k|^2.$$

An important feature of the orthogonal projection on $\mathbb{P}^N$ represented by the truncated Fourier series (3.2) is related to its spectral convergence properties for smooth solutions which are summarized in the following theorem.

THEOREM 3.1. *If $f \in H^r_p([-\pi,\pi]^d)$, where $r \geq 0$ is an integer and $H^r_p([-\pi,\pi]^d)$ is the subspace of the Sobolev space $H^r([-\pi,\pi]^d)$ which consists of periodic functions, we have*

$$\|f - f_N\|_{H^r_p} \leq \frac{C}{N^r} \|f\|_{H^r_p}. \tag{3.3}$$

Finally, we recall some approximation properties of the projection operator $\mathcal{P}_N$, in particular those concerning approximation of the macroscopic quantities. Let us remark that, in general, when we approximate a function by a partial sum of its Fourier series, except for the moment of order zero all other moments are not preserved by the projection.

The results are summarized in the following proposition [42]

PROPOSITION 3.2. *Let $f \in L^2_p([-\pi,\pi]^d)$ and let us define*

$$U = (\rho, \rho u, \rho e)^T = \langle f, \Phi \rangle,$$

where $\Phi = (1, v_1, \ldots, v_d, |v|^2)^T \in \mathbb{R}^{d+2}$.

i) The moments of f_N are given by

$$U_N = (\rho_N, \rho u_N, \rho e_N)^T = \langle f_N, \Phi \rangle = (2\pi)^d \sum_{k=-N}^{N} \hat{f}_k \hat{\Phi}_k, \tag{3.4}$$

where $\hat{\Phi}_k = (\delta_{k0}, \hat{v}_k, (\hat{v^2})_k)^T \in \mathbb{R}^{d+2}$, δ_{k0} is the Kronecker delta, and $\hat{v}_k$ and $(\hat{v^2})_k$ are the Fourier coefficients of v and v^2 characterized by $(\hat{v_j})_0 = 0$, $j = 1, \ldots, d$ and $(\hat{v^2})_0 = \pi^2$ whereas for $k \neq 0$ we get

(3.5)

$$(\hat{v_j})_k = -i \prod_{\substack{l=1 \\ l \neq j}}^{d} \delta_{k_l 0} \frac{(-1)^{k_j}}{k_j}, \quad j = 1, \ldots, d, \quad (\hat{v^2})_k = 2 \sum_{j=1}^{d} \prod_{\substack{l=1 \\ l \neq j}}^{d} \delta_{k_l 0} \frac{(-1)^{k_j}}{k_j^2}.$$

ii) The following relations hold

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

iii) If $f \in H^r_p([-\pi,\pi]^d)$, where $r \geq 0$ is an integer, for each $\phi \in L^2_p([-\pi,\pi]^d)$ we have

$$|\langle f, \phi\rangle - \langle f_N, \phi\rangle| \leq \|\phi\|_{L^2_p} \|f - f_N\|_{H^r_p} \leq \frac{C}{N^r} \|\phi\|_{L^2_p} \|f\|_{H^r_p}.$$

The last inequality shows that the projection error on the moments decay faster than algebraically when the solution is infinitely smooth.

3.2. Constrained best approximations. We want to define a different projection operator on the space of trigonometric polynomials, $\mathcal{P}^c_N : L^2_p([-\pi,\pi]^d) \to \mathbb{P}^N$ such that it satisfies

$$\langle \mathcal{P}^c_N f, \Phi\rangle = \langle f, \Phi\rangle,$$

but preserving the convergence properties of the finite Fourier series.

To this aim, we recall that, given a function $f \in L^2_p([-\pi,\pi]^d)$, the truncated Fourier series (3.2) represents the trigonometric polynomial of best approximation in the least squares sense, more precisely $f_N = \mathcal{P}_N f$ is the solution of the following minimization problem

$$f_N = \operatorname{argmin} \left\{ \|g_N - f\|^2_{L^2_p} \ : \ g_N \in \mathbb{P}^N \right\}.$$

Thus, it is natural to consider the following constrained best approximation problem in the space of trigonometric polynomials

$$f^c_N = \operatorname{argmin} \left\{ \|g_N - f\|^2_{L^2_p} \ : \ g_N \in \mathbb{P}^N, \ \langle g_N, \Phi\rangle = \langle f, \Phi\rangle \right\}. \tag{3.6}$$

Now, since $g_N \in \mathbb{P}^N$ we can represent it in the form

$$g_N = \sum_{k=-N}^{N} \hat{g}_k e^{ik\cdot v}$$

and then by Parseval's identity

$$\|g_N - f\|^2_{L^2_p} = (2\pi)^d \sum_{k=-\infty}^{\infty} |\hat{g}_k - \hat{f}_k|^2,$$

where we assumed $\hat{g}_k = 0$, $|k_j| > N$, $j = 1, \ldots, d$.

Note that, since conservation of moments is built in g_N, one necessarily needs that

$$\langle g_N, \Phi\rangle = (2\pi)^d \sum_{k=-N}^{N} \hat{g}_k \hat{\Phi}_k = \langle f, \Phi\rangle = U. \tag{3.7}$$

Let us now solve the minimization problem (3.6) using the Lagrange multiplier method. Let $\lambda \in \mathbb{R}^{d+2}$ be the vector of Lagrange multipliers, we consider the objective function

$$\mathcal{L}(\hat{g}, \lambda) = (2\pi)^d \sum_{k=-\infty}^{\infty} |\hat{g}_k - \hat{f}_k|^2 + \lambda^T \left((2\pi)^d \sum_{k=-N}^{N} \hat{g}_k \hat{\Phi}_k - U \right),$$

with $\hat{g} \in \mathbb{R}^{2N+1}$ the vector of coefficients $\hat{g}_k$, $k = -N, \dots, N$. Stationary points are found by imposing

$$\frac{\partial \mathcal{L}(\hat{g}, \lambda)}{\partial \hat{g}_k} = 0, \quad k = -N, \dots, N; \qquad \frac{\partial \mathcal{L}(\hat{g}, \lambda)}{\partial \lambda_j} = 0, \quad j = 1, \dots, d+2.$$

From the first condition, one gets

$$2(\hat{g}_k - \hat{f}_k) + \lambda^T \hat{\Phi}_k = 0, \tag{3.8}$$

whereas the second condition corresponds to (3.7).

Multiplying the above equation by $\hat{\Phi}_k$ and summing up over k we can write

$$2 \sum_{k=-N}^{N} (\hat{g}_k - \hat{f}_k)\hat{\Phi}_k + \sum_{k=-N}^{N} \hat{\Phi}_k \hat{\Phi}_k^T \lambda = 0$$

and, using (3.4), (3.7) and the fact that $\hat{\Phi}_k \hat{\Phi}_k^T$ are symmetric and positive definite matrices of size $d+2$ one obtains

$$\lambda = -\frac{2}{(2\pi)^d} \left(\sum_{k=-N}^{N} \hat{\Phi}_k \hat{\Phi}_k^T \right)^{-1} (U - U_N). \tag{3.9}$$

Now, rewriting the first condition (3.8) as

$$\hat{g}_k = \hat{f}_k - \frac{1}{2} \hat{\Phi}_k^T \lambda,$$

and plugging the expression (3.9) of λ in it, one obtains that the minimum is achieved for $\hat{g}_k = \hat{f}_k^c$, given by the following definition.

DEFINITION 3.3. *One can define a* conservative projection $\mathcal{P}_N^c f = f_N^c$ *in* $\mathbb{P}^N$, *where* f_N^c *is given by*

$$f_N^c = \sum_{k=-N}^{N} \hat{f}_k^c \, e^{ik \cdot v}. \tag{3.10}$$

with

$$\hat{f}_k^c = \hat{f}_k + \hat{C}_k^T (U - U_N), \qquad \hat{C}_k^T = \frac{1}{(2\pi)^d} \hat{\Phi}_k^T \left(\sum_{h=-N}^{N} \hat{\Phi}_h \hat{\Phi}_h^T \right)^{-1}. \tag{3.11}$$

The following result states the spectral accuracy of the conservative best approximation in least square sense (3.10), and generalizes Theorem 3.1.

THEOREM 3.4. *If* $f \in H_p^r([-\pi, \pi]^d)$, *where* $r \geq 0$ *is an integer, we have*

$$\|f - f_N^c\|_{H_p^r} \leq \frac{C_\Phi}{N^r} \|f\|_{H_p^r} \tag{3.12}$$

where the constant C_Φ *depends on the spectral radius of the matrix* $\langle \Phi, \Phi^T \rangle$, *and on* $\|\Phi\|_{2,L_p^2}^2 = \sum_{j=1}^{d+2} \|\Phi_j\|_{L_p^2}^2$ *where* Φ_j, $j = 1, \dots, d+2$ *are the components of the vector* Φ.

Proof. We can split

$$\|f - f_N^c\|_{L_p^2}^2 \leq \|f - f_N\|_{L_p^2}^2 + \|f_N - f_N^c\|_{L_p^2}^2.$$

The first term is bounded by the spectral estimate of truncated Fourier series (3.3), whereas for the second term by Parseval's identity we have

$$\|f_N^c - f_N\|_{L_p^2}^2 = (2\pi)^d \sum_{k=-N}^{N} |\hat{f}_k^c - \hat{f}_k|^2.$$

Now, using the definition (3.11) we get

$$\sum_{k=-N}^{N} |\hat{f}_k^c - \hat{f}_k|^2 = \sum_{k=-N}^{N} |\hat{C}_k^T (U - U_N)|^2 \leq \|U - U_N\|_2^2 \sum_{k=-N}^{N} \|\hat{C}_k^T\|_2^2.$$

From (3.2) spectral accuracy of moments implies

$$\|U - U_N\|_2 \leq \frac{C}{N^r} \|f\|_{H_p^r} \left(\sum_{j=1}^{d+2} \|\Phi_j\|_{L_p^2}^2 \right)^{1/2} = \frac{C}{N^r} \|f\|_{H_p^r} \|\Phi\|_{2,L_p^2}$$

where for $\Phi = (\Phi_1, \ldots, \Phi_{2+d})^T$ we defined $\|\Phi\|_{2,L_p^2}^2 = \sum_{j=1}^{d+2} \|\Phi_j\|_{L_p^2}^2$.

Finally, since $\hat{C}_k^T = (2\pi)^{-d} \hat{\Phi}_k^T \left(\sum_{h=-N}^{N} \hat{\Phi}_h \hat{\Phi}_h^T \right)^{-1}$, one has

$$\begin{aligned}(2\pi)^d \sum_{k=-N}^{N} \|\hat{C}_k^T\|_2^2 &\leq \sum_{k=-N}^{N} \|\hat{\Phi}_k\|_2^2 \left\| \left(\sum_{h=-N}^{N} \hat{\Phi}_h \hat{\Phi}_h^T \right)^{-1} \right\|_2^2 \\ &= \frac{\|\Phi\|_{2,L_p^2}^2}{\left\| \sum_{h=-N}^{N} \hat{\Phi}_h \hat{\Phi}_h^T \right\|_2^2} = \frac{\|\Phi\|_{2,L_p^2}^2}{\|\langle \Phi, \Phi^T \rangle\|_2^2} = \frac{\|\Phi\|_{2,L_p^2}^2}{\rho^2(\Phi)}\end{aligned}$$

where $\rho(\Phi)$ is the spectral radius of the matrix $\langle \Phi, \Phi^T \rangle$. One finally obtains

$$\|f_N^c - f_N\|_{L_p^2} \leq \frac{C}{N^r} \|f\|_{H_p^r} \frac{\|\Phi\|_{2,L_p^2}^2}{\rho(\Phi)}$$

which proves (3.12). ∎

Remark 3.5. The conservative best approximation in least square (3.11) can be represented in term of the standard projection as

$$\mathcal{P}_N^c f = \mathcal{P}_N f + \sum_{k=-N}^{N} \hat{C}_k^T \langle f - \mathcal{P}_N f, \Phi \rangle.$$

The above representation emphasizes the analogies with the L^2 minimization problem solved in [26]. The main difference is represented by the continuous representation of the solution in the space of trigonometric polynomials which permits to demonstrate spectral accuracy of the resulting approximation. Note also that the same conservative projection remains valid in the case we are interested in performing mesh changes for example by reducing or increasing the number of modes by keeping moment conservation and spectral accuracy.

4. Application to the Boltzmann equation. The moment-preserving Fourier approximation introduced in Section 3 allows a direct construction of a moment-preserving Fourier-Galerkin approximation of the Boltzmann equation. The algorithmic details of this construction, together with the fast implementation strategy, are reported in Appendix A. Here, we will focus our attention on the general mathematical formulation and on the study of its spectral accuracy and stability properties. Let us define the moment constrained Fourier approximation of the truncated Boltzmann operator (2.8) as the solution of the following problem

$$Q_N^{R,c}(f,f) = \operatorname{argmin}\left\{ \|g_N - Q^R(f,f)\|_{L^2_p}^2 \ : \ g_N \in \mathbb{P}^N, \ \langle g_N, \Phi\rangle = 0 \right\} \tag{4.1}$$

or equivalently

$$Q_N^{R,c}(f,f) = \mathcal{P}_N Q^R(f,f) - \sum_{k=-N}^{N} \hat{C}_k^T \langle \mathcal{P}_N Q^R(f,f), \Phi\rangle \tag{4.2}$$

and the constrained, Fourier projected, homogeneous Boltzman equation then reads

$$\begin{cases} \dfrac{\partial f_N^c}{\partial t} = Q_N^{R,c}(f_N^c, f_N^c) \\ f_N^c(v,0) = \mathcal{P}_N^c f_0(v), \quad v \in [-\pi,\pi]^d. \end{cases} \tag{4.3}$$

Let us underline that (4.2) differs from the conservative projection in Definition 3.3 in the sense that the constrained minimization problem (4.1) is solved with respect to the physical conservation laws of the collision term in the whole space and not in the periodic box (see the discussion at the end of Section 2). A direct application of Definition 3.3, for which we have proved the spectral accuracy property in Theorem 3.4, will lead to the projected operator

$$\tilde{Q}_N^{R,c}(f,f) = \mathcal{P}_N Q^R(f,f) - \sum_{k=-N}^{N} \hat{C}_k^T \langle Q^R(f,f) - \mathcal{P}_N Q^R(f,f), \Phi\rangle. \tag{4.4}$$

It is therefore clear, that some smallness on $\langle Q^R(f,f), \Phi\rangle$ is necessary in order to prove spectral consistency of (4.3). As discussed in Section 2.2, see Proposition 2.2, this is guaranteed if the solution satisfy a smallness assumption outside the ball $\mathcal{B}_0(R)$.

In the sequel, we discuss consistency and stability of the moment preserving spectral method.

4.1. Spectral consistency. First, we shall prove that the moment preserving method for the Boltzmann equation is spectrally accurate. Let us recall some of the regularity properties of the truncated collision operator [18, 42].

LEMMA 4.1 (Lemma 4.1 of [18], Lemma 5.2 of [42]). *For all $p \in (1, +\infty]$ there exists a positive constant $C = C(R,p)$ such that if $f \in L^1([-\pi,\pi]^d)$, $g \in L^p([-\pi,\pi]^d)$ one has*

$$\|Q^R(f,g)\|_{L^p} \leq C \|f\|_{L^1} \|g\|_{L^p}.$$

Moreover, if $f, g \in L^2([-\pi,\pi]^d)$, one also has

$$\|Q^R(f,g)\|_{L^2} \leq C \|f\|_{L^2} \|g\|_{L^2}.$$

Proof. The proof of these inequalities are classical since the collision kernel is bounded and the functions compactly supported, see e.g. [37]. ∎

Under a smallness assumption on the moments of the truncated collision term (see Proposition 2.2) one can then prove a consistency property for equation (4.3).

THEOREM 4.2. *Let $f \in L^2([-\pi,\pi]^d)$ be such that for $\delta > 0$ there exists $R = R(\delta)$ providing the following smallness estimate on the moments of the truncated collision term*

$$\|\langle Q^R(f,f), \Phi\rangle\|_2 \leq \tilde{C}\delta. \tag{4.5}$$

Then for any $r \geq 1$, there exists a constant $C = C\left(\|f\|_{L^2}, R, r\right)$ such that

$$\|Q^R(f,f) - Q_N^{R,c}(f_N^c, f_N^c)\|_{L^2} \leq C\left(\|f - f_N^c\|_{L^2} + \frac{\|Q^R(f_N^c, f_N^c)\|_{H_p^r}}{N^r} + \delta\right).$$

Proof. We first split

$$\begin{aligned} \|Q^R(f,f) - Q_N^{R,c}(f_N^c, f_N^c)\|_{L^2} &\leq \|Q^R(f,f) - Q^R(f_N^c, f_N^c)\|_{L^2} \\ &+ \|Q^R(f_N^c, f_N^c) - \tilde{Q}_N^{R,c}(f_N^c, f_N^c)\|_{L^2} \\ &+ \|\tilde{Q}_N^{R,c}(f_N^c, f_N^c) - Q_N^{R,c}(f_N^c, f_N^c)\|_{L^2}, \end{aligned} \tag{4.6}$$

where the projected collision term $\tilde{Q}_N^{R,c}(f_N^c, f_N^c)$ is defined by (4.4).

Since $Q^R(f_N^c, f_N^c) \in \mathbb{P}^N \subset H_p^r$, the second term in the RHS of (4.6) is spectrally small, according to Theorem 3.4 there exists $C_1 > 0$ such that

$$\|Q^R(f_N^c, f_N^c) - \tilde{Q}_N^{R,c}(f_N^c, f_N^c)\|_{L^2} \leq C_1 \frac{\|Q^R(f_N^c, f_N^c)\|_{H_p^r}}{N^r}. \tag{4.7}$$

Moreover, using Lemma 4.1 and the bilinearity of Q^R, there exists $C_2 > 0$ such that

$$\begin{aligned} \|Q^R(f,f) - Q^R(f_N^c, f_N^c)\|_{L^2} &= \|Q^R(f + f_N^c, f - f_N^c)\|_{L^2} \\ &\leq C_2 \|f + f_N^c\|_{L^2} \|f - f_N^c|_{L^2} \\ &\leq 2C_2 \|f\|_{L^2} \|f - f_N^c\|_{L^2} \end{aligned} \tag{4.8}$$

Finally using assumption (4.5) there exists $C_3 > 0$ such that

$$\begin{aligned} \|\tilde{Q}_N^{R,c}(f_N^c, f_N^c) - Q_N^{R,c}(f_N^c, f_N^c)\|_{L^2}^2 &= (2\pi)^d \sum_{k=-N}^{N} |\hat{C}_k^T \langle Q^R(f_N^c, f_N^c), \Phi\rangle|^2 \\ &\leq (2\pi)^d \left(\|\langle Q^R(f,f), \Phi\rangle\|_2^2 + \|\langle Q^R(f,f) - Q^R(f_N^c, f_N^c), \Phi\rangle\|_2^2\right) \sum_{k=-N}^{N} \|C_k^T\|_2^2 \\ &\leq C_3(\delta^2 + \|f - f_N^c\|_{L^2}^2). \end{aligned} \tag{4.9}$$

Collecting together (4.7), (4.8) and (4.9) in (4.6) completes the proof. ∎

Theorem 4.2 states that the rate of convergence of the moment constrained spectral approximation of the truncated Boltzmann collision operator depends on the regularity of the distribution f. However, in contrast to the classical spectral estimates in [35, 42], here the regularity has to be complemented with assumption (4.5)

on the smallness of the moments of the truncated collision term evaluated at f. From a practical viewpoint, this is equivalent to assume a suitable decay of the tails of the initial data and a computational domain large enough to guarantee minimal loss of the collision invariants. Gathering this result with the spectral accuracy (3.12) of the moment constrained projection, one then has the following spectral consistency result:

COROLLARY 4.3. *Let $f \in H^r_p([-\pi,\pi]^d)$ for a given $r \geq 1$ be such that there exist $R = R(r,N)$ providing estimate (4.5) with $\delta = \delta' N^{-r}$. There exists a constant $C = C\left(\|f\|_{H^r_p}, R\right)$ such that*

$$\|Q^R(f,f) - Q^{R,c}_N(f^c_N, f^c_N)\|_{L^2} \leq \frac{C}{N^r}\left(\|f\|_{H^p_r} + \|Q^R(f^c_N, f^c_N)\|_{H^r_p} + \delta'\right).$$

Note that, achieving consistency and spectral accuracy for increasing values of N requires a vanishing error in terms of moments of the collision operator and, as a consequence, a truncation domain which depends on the number of modes N. This agrees well with the intuition that a larger computational support has to be used when the number of modes is increased as already observed in practice in [42].

4.2. Stability of the moment constrained spectral methods. Let us recall the general stability result of Filbet and Mouhot [18], that can be used to prove the stability and large time behavior of spectral methods for the Boltzmann equation. Let us introduced the following perturbed, truncated Boltzmann equation

$$\begin{cases} \dfrac{\partial f}{\partial t} = Q^R(f,f) + P_\varepsilon(f) \\ f(v,0) = f_{0,\varepsilon}(v), \quad v \in [-\pi,\pi]^d, \end{cases} \tag{4.10}$$

where the perturbation $P_\varepsilon(f)$ is smooth and "balanced", in the following sense

DEFINITION 4.4. *A family of operators P_ε is said to be a* stable *perturbation of the Boltzmann equation if it verifies the following properties:*

i) $\int P_\varepsilon(f)\,dv = 0$;

ii) there exists $r \geq 1$, C_1, $C_r \geq 0$ such that

$$\begin{cases} \|P_\varepsilon(f)\|_{L^1} \leq C_1 \|f\|_{L^1}\|f\|_{L^1}, \\ \|P_\varepsilon(f)\|_{H^r_p} \leq C_r \|f\|_{L^1}\|f\|_{H^r_p}. \end{cases}$$

iii) there exists a function $\varphi(\varepsilon)$ which goes to 0 when ε goes to 0 and such that

$$\|P_\varepsilon(f)\|_{H^r_p} \leq \varphi(\varepsilon), \quad \forall r \geq 1.$$

One then has the following stability and trends to equilibrium result for the perturbed equation (4.10).

THEOREM 4.5 (Thm 3.1 of [18]). *Let us consider the perturbed truncated Boltzmann equation (4.10), with a stable family of perturbations (P_ε) satisfying the hypotheses of Definition 4.4. Let $f_0 \in H^r_p$ for $r > d/2$ be a nonzero, nonnegative function and $(f_{0,\varepsilon})$ be a family of smooth perturbations of f_0:*

$$\int_{[-\pi,\pi]^d} f_{0,\varepsilon}\,dv = \int_{[-\pi,\pi]^d} f_0\,dv, \qquad \|f_0 - f_{0,\varepsilon}\| \leq \psi(\varepsilon),$$

where $\psi(\varepsilon)$ goes to 0 when ε goes to 0. Then there exists ε_0 depending only on the truncation parameter R, the collision kernel B, the constant in Definition 4.4 and $\|f_0\|_{H^r_p}$ such that, for any $\varepsilon \in (0, \varepsilon_0)$,

1. *there exists a unique global smooth solution f_ε to (4.10);*
2. *for any $k < r$, $f_\varepsilon(t, \cdot) \in H^k_p$, uniformly in time;*
3. *the quantity of negative values of f_ε vanishes when ε goes to 0;*
4. *for any $T > 0$, the solution f_ε of (4.10) converges in $L^\infty([0, T]; H^r_p)$ towards a solution f to the unperturbed equation (2.7) when ε goes to 0;*
5. *as time goes to infinity, the solution f_ε converges in H^r_p towards the piecewise constant equilibrium m_∞ defined in (2.9).*

Note that there are other more recent stability results, see [1, 30], but they do not cover the very critical point of the large-time behavior of the methods.

After investigating the consistency of the moment preserving spectral methods, it is natural to wonder whether the constrained projection impacts the stability and long time behavior properties of the new spectral approach. We shall prove the following theorem.

THEOREM 4.6. *Let us consider a nonnegative initial condition $f_0 \in H^k_p([-\pi, \pi]^d)$ for a given $k \geq d/2$ be such that there exist $R = R(k, N)$ providing an uniform in time estimate (4.5) for the solution $f(t, v)$ to problem (2.7) with $\delta = \delta' N^{-k+r}$ for any $r < k$. There exists $N_0 \in \mathbb{N}$ depending on the H^k_p norm of f_0 and on the spectral radius of the matrix $\langle \Phi, \Phi^T \rangle$ such that for all $N \geq N_0$:*

1. *There is a unique global smooth solution f^c_N to the problem (4.3);*
2. *For any $r < k$, there exists $C_r > 0$ such that*

$$\|f^c_N(t, \cdot)\|_{H^r_p} \leq C_r;$$

3. *this solution in everywhere positive for time large enough;*
4. *this solution $f^c_N(t, \cdot)$ converges to a solution $f(t)$ of the truncated Boltzmann equation (2.7) with spectral accuracy, uniformly in time;*

In order to prove this result, we follow the perturbative framework developed in [18] and summarized in Theorem 4.5. The main difference here is that the constrained method preserves not only mass, but also momentum and kinetic energy, on the finite hypercube $[-\pi, \pi]^d$, without an H-theorem-like decay of the Boltzmann entropy a priori. As such, the equilibria of this new operator are not necessarily Gaussian (not even explicit), and one won't be able to perform the same spectral analysis of the linearized collision operator as was done in [18] to study its long time behavior. The same difficulties were faced in [41] for the equilibrium preserving spectral method. In addition, to recover spectral accuracy we need a smallness assumption on the error in the moments approximation of the collision term.

To mimic the proof of the first points of the general stability Theorem 4.5, one needs to rewrite the moment constrained spectral method as a stable perturbation of the truncated Boltzmann equation that fits the hypotheses of Definition 4.4. Let us introduce the perturbation operator $P^{R,c}_N$ as

$$P^{R,c}_N(f^c_N) := Q^{R,c}_N(f^c_N, f^c_N) - Q^R(f^c_N, f^c_N).$$

Plugging this expression into the constrained, Fourier projected, homogeneous Boltz-

mann equation (4.3) yields the following perturbed equation

$$\begin{cases} \dfrac{\partial f_N^c}{\partial t} = Q^R(f_N^c, f_N^c) + P_N^{R,c}(f_N^c) \\ f_N^c(v,0) = \mathcal{P}_N^c f_0(v), \quad v \in [-\pi,\pi]^d. \end{cases}$$

Proof of Theorem 4.6. In order to prove the result, one need to check that the perturbation operator $P_N^{R,c}$ is a stable perturbation as in 4.4:

i) the mass conservation is clear, since the constrained projection is conservative

$$\int P_N^{R,c}(f)(v)\,dv = \int \left[Q_N^{R,c}(f,f) - Q^R(f,f)\right] dv = 0.$$

ii) Let $r \geq 1$, one has using iteratively Leibniz formula for the derivatives of a product of smooth function on the results of Lemma 4.1, along with the smoothness of the spectral projection $\mathcal{P}_N^c$ that

$$\begin{aligned} \left\|P_N^{R,c}(f)\right\|_{H_p^r} &\leq \left\|Q^R(f,f)\right\|_{H_p^r} + \left\|Q_N^{R,c}(f,f)\right\|_{H_p^r} \\ &\leq C_r(R)\,\|f\|_{L^1}\,\|f\|_{H_p^r}. \end{aligned}$$

Note that this estimate is similar to the one obtained in [18] for the nonconservative spectral method.

iii) Using the same regularity estimate of the truncated collision operator, along with the spectral convergence property (3.12) of the constrained spectral projection, one has the spectral smallness of the perturbation

$$\begin{aligned} \|P_N^{R,c}(f)\|_{H_p^r} &= \left\|Q^R(f,f) - Q_N^{R,c}(f,f)\right\|_{H_p^r} \\ &\leq \frac{C_\Phi}{N^{k-r}}\left(\left\|Q^R(f,f)\right\|_{H_p^r} + \delta'\right) \\ &\leq C_r(R)\frac{C_\Phi}{N^{k-r}}\left(\|f\|_{L^1}\,\|f\|_{H_p^r} + \delta'\right). \end{aligned}$$

These three properties allow us to obtain the global existence, positivity for large time and spectral convergence (points 1, 3, and 4 of Theorem 4.6) by noticing that for any $f_0 \in H_p^k([\pi,\pi]^d)$ with $k > d/2$ we have

$$\|f_N^c(0,\cdot)\|_{H_p^k} \leq \|f_0\|_{H_p^k}, \qquad \|f_N^c(0,\cdot) - f_0\|_{H_p^k} \leq \frac{C_\Phi}{N^k}.$$

Finally, the global boundedness of the H^r norm is just a consequence of a classical Grönwall-type argument, that can be reproduced as in [41]. This concludes the proof of Theorem 4.6. ∎

5. Numerical examples. In this section we present several numerical examples to validate our theoretical findings. First we consider the moment preserving approximation in the Fourier space and analyze its spectral convergence properties for smooth solutions. Next we compare the results obtained with the new method with those computed using the fast spectral method [35], the equilibrium preserving spectral method recently introduced in [20] and a novel moment preserving and equilibrium preserving spectral method obtained by combining the two previous approaches.

function (5.1)

N	$\lvert\rho - \rho_N^c\rvert$	$\lvert\rho u - \rho u_N^c\rvert$	$\lvert\rho e - \rho e_N^c\rvert$	$\lVert f - f_N^c\rVert_2$
8	0	8.462e-16	2.776e-16	1.654e-09
16	0	6.540e-17	3.053e-16	1.443e-13
32	0	1.496e-16	5.551e-16	3.342e-16

function (5.2)

N	$\lvert\rho - \rho_N^c\rvert$	$\lvert\rho u - \rho u_N^c\rvert$	$\lvert\rho e - \rho e_N^c\rvert$	$\lVert f - f_N^c\rVert_2$
8	0	1.756e-16	8.770e-15	8.402e-05
16	0	2.165e-16	6.808e-16	1.545e-11
32	0	3.849e-16	1.314e-16	6.007e-15

TABLE 1

Test 1. *Approximation errors of the moment constrained Fourier truncation f_N^c in (3.10)-(3.11) for the reduced centered Gaussian (5.1) and for the two bumps function (5.2).*

5.1. Conservative approximations in the Fourier space. We are interested in validating the theoretical results on the moment constrained spectral representation (3.10)-(3.11) for the first three moments, namely mass, momentum and kinetic energy.

Test 1. Conservation of moments and spectral accuracy. We consider the following simple problem to verify conservations and spectral accuracy: Compute the unidimensional truncated spectral projection (3.2) f_N of a given function f with moments U, then compute its moment constrained Fourier transform f_N^c using formula (3.10)-(3.11) by prescribing the moments of the original function f and compute both the error on the moments $U_N^c := (\rho_N^c, \rho u_N^c, \rho e_N^c)$ of f_N^c with respect to U, and the L^2-error between the original f and f_N^c.

We shall perform this numerical test for the reduced centered Gaussian

$$f(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-v^2/2\right), \quad v \in [-6, 6], \tag{5.1}$$

and for the following asymmetric sum of Gaussians

$$f(v) = \frac{1}{2\sqrt{2\pi}} \left(\exp\left(-(v-4)^2/2\right) + \exp\left(-(v+2)^2/2\right)\right), \quad v \in [-12, 12]. \tag{5.2}$$

We present in Table 1 the errors on the mass, momentum, energy and on the distribution for the moment constrained spectral projection. As expected, the constrained projection preserves the moments up to machine precision even in the case of an asymmetric distribution. We also observe the spectral convergence of the constrained distribution f_N^c towards the exact solution f as predicted in Theorem 3.4.

We emphasize that in Table 1 we used the exact Fourier expansion of the moment vector given by (3.5). One other possibility would be to use a Discrete Fourier Transform on the original moment vector, by sampling uniformly its value (and using the discrete version of Parseval identity, see identity (6.1) in [3]). This, however, will introduce an additional spectrally small error on the moment of the constrained function f_N^c which may reduce accuracy when using a small number of nodes.

5.2. Moment constrained and equilibrium preserving spectral methods. Let us now apply the conservative approximations to the fast spectral method

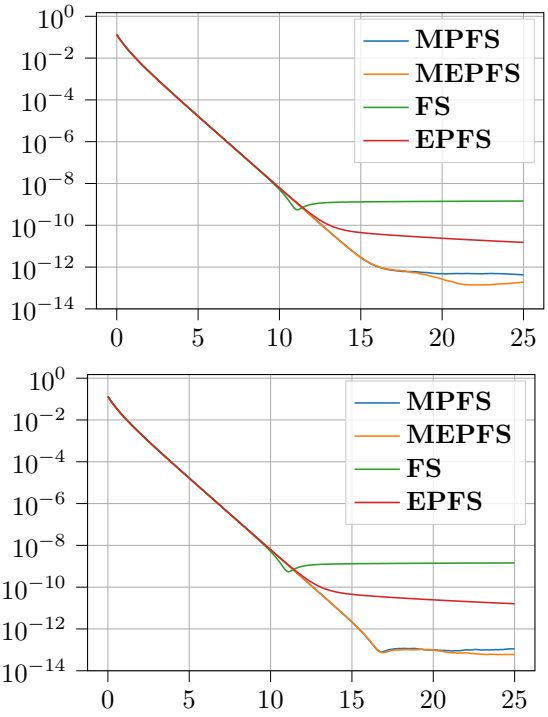

FIGURE 1. ***Test 2.*** *Time evolution of $\|f_N - M\|_{L^2}$ for the BKW solution (5.3), where M is the local Maxwellian, using $N = 32^2$ (top) and $N = 64^2$ (bottom).*

for the Boltzmann equation (2.7). For the velocity discretization, we choose $\Omega = [-12, 12]^2$, $N = 64^2$ then 128^2 points and $M = 8$ angular discretizations. Because the problem is not stiff, we use an explicit Runge-Kutta of order 4 in time, with $\Delta t = 0.01$. We shall compare the fast spectral (**FS**) method [36] with the moment preserving fast spectral (**MPFS**) method (A.4). We shall also compare our numerical experiments with the equilibrium preserving fast spectral (**EPFS**) method introduced in [20], and by taking a combination of the two approaches where we apply the equilibrium preserving method together with the moment preserving technique. Note that, this latter approach (referred to as **MEPFS**) preserves not only the moments but also the local Maxwellian equilibrium state. For the reader convenience, its details are summarized in Appendix A.2 (see equation (A.7)).

In the following, for the sake of brevity, we omit to present the error for short times since all three methods give similar results due to their spectral convergence properties, instead we will focus on the behavior of the methods for long times.

Test 2. Trends to equilibrium. We shall consider an exact solution of the homogeneous Boltzmann equation, the so called Bobylev-Krook-Wu solution [5, 33]. It is given in two dimensions of velocity by

$$f_{BKW}(t, v) = \frac{\exp(-v^2/2S)}{2\pi S^2} \left[2\,S - 1 + \frac{1-S}{2\,S}\,v^2\right] \tag{5.3}$$

with $S = S(t) = 1 - \exp(-t/8)/2$.

Figure 1 presents the time evolution of the L^2 error of the solution f_N with respect to the equilibrium distribution M, where we observe an exponential convergence

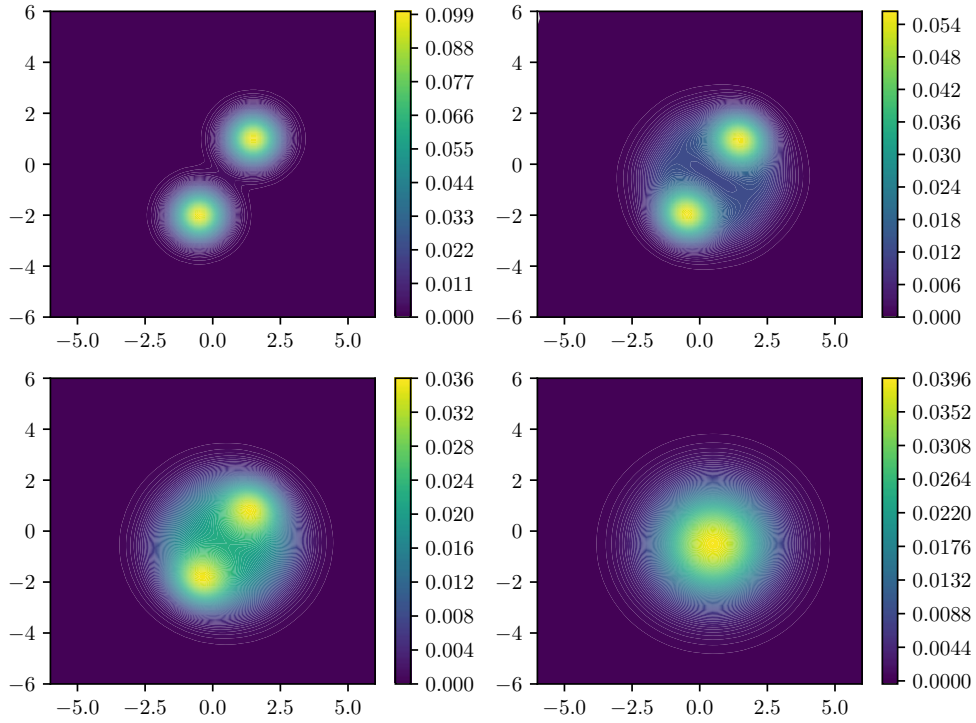

FIGURE 2. ***Test 3.*** *Contour plot of the solution f_N^c of the **MPFS** method, for the two bumps initial data (5.4), with $N = 64^2$ points, at time $t = 0,\ 0.5,\ 1,\ 10$ (bottom).*

towards M. As observed in [20, 40], the behavior of the **EPFS** method is better than the classical **FS** which saturates around 10^{-9}, but both are outperformed by almost two order of magnitude by the new moment constrained **MPFS** method. Adding the equilibrium preserving feature to this latter method slightly improve its accuracy, and the **MEPFS** scheme outperforms all the others in very large time.

Test 3. Error on the temperature. We now consider the following asymmetric two bumps initial data:

$$f(v) = \frac{1}{4\pi}\left(e^{-(|v-1|^2+|v-2|^2)/2} + e^{-(|v+2|^2+|v+1|^2)/2}\right), \quad \forall\, v \in [-12, 12]^2. \tag{5.4}$$

Figure 2 presents the isovalues of the solution computed with the **MPFS** method, for $N = 64^2$. We can observe the correct convergence towards the Maxwellian equilibrium state.

We then compute numerically the time evolution of the temperature

$$T = \frac{1}{2\rho}\int f|v-u|^2\,dv$$

associated to this test case. Figure 3 shows the temperature error $|T_N(t) - T(0)|$ of the three numerical methods. As expected, for the classical spectral method **FS** one can observe a linear growth of the error. This growth is almost grid independent, and is actually due to the Fourier truncation. As was observed in [20], using the equilibrium preserving correction **EPFS** improves this conservation error by one order of magnitude, and again independently on the grid. Nevertheless, one needs to perform

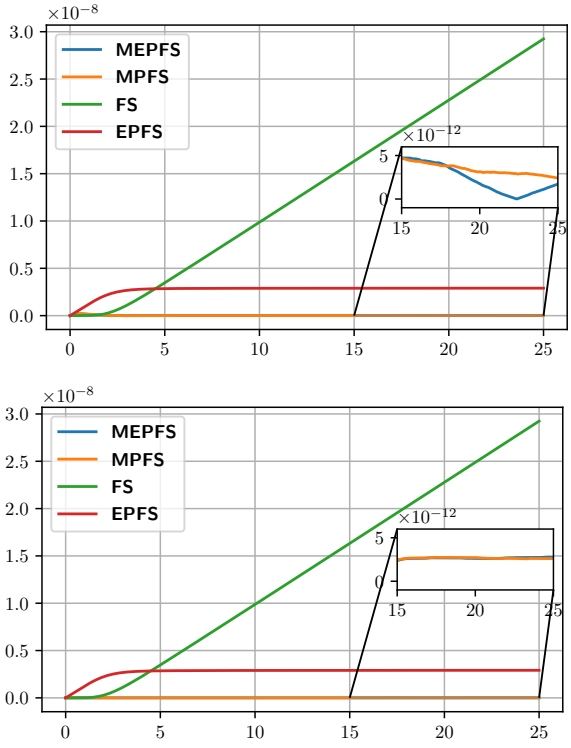

FIGURE 3. ***Test 3.*** *Error on the temperature for the two bumps initial data (5.4), for* $N = 32^2$ *(top) and* $N = 64^2$ *(bottom).*

the moment constrained correction **MPFS** in order to (almost) attain the machine precision in the evolution of the temperature. Note that including the equilibrium preserving approach **MEPFS** improves the results only marginally.

6. Conclusions. We introduced and analyzed a new class of Fourier-Galerkin spectral methods for kinetic equations that can preserve the moments of the distribution function. The method was introduced using a best constrained approximation formalism in the space of trigonometric polynomials. Due to the general formulation, the method allows the spectral accuracy property of the solution to be maintained for the conservative finite Fourier series. The new approximation was then used to derive fast Fourier-Galerkin methods for the Boltzmann equation that preserve the collisional invariants, and their theoretical properties have been analyzed. Compared to other conservative schemes for the Boltzmann equation based on a constrained minimization approach [26], the analysis of the theoretical properties of the new method, such as spectral consistency, is greatly simplified thanks to the Fourier-Galerkin setting. Due to the enforcement of conservations, the corresponding estimates contain an additional error term that depends on the smallness of the moments of the collision operator. We also introduced a modification of the method which is capable to preserve exactly the equilibrium state represented by the conservative projection of the local Maxwellian.

Given its generality, the method admits numerous extensions. Among the most interesting are certainly the construction of spectrally accurate and conservative meth-

ods for the Landau equation. Another interesting direction is the application to kinetic models in the socio-economic domain where equilibrium states are often unknown and thanks to the present approach can be computed with spectral accuracy.

Acknowledgment. This work has been written within the activities of GNCS groups of INdAM (National Institute of High Mathematics). L.P. acknowledge the partial support of MIUR-PRIN Project 2017, No. 2017KKJP4X "Innovative numerical methods for evolutionary partial differential equations and applications". T.R. would like to thanks Maxime Herda for fruitful discussions on the implementation of the method. T.R. was partially funded by Labex CEMPI (ANR-11-LABX-0007-01) and ANR Project MoHyCon (ANR-17-CE40-0027-01).

Appendix A. The conservative fast spectral method. The fast spectral method developed in [35] can be applied directly in the conservative formulation. In the sequel we will summarize briefly the details of the method. Since the collision operator is local in space and time, only the dependency on the velocity variable v is considered for the distribution function f, i.e. $f = f(v)$. We suppose the distribution function f to have compact support on the ball $\mathcal{B}_0(R)$ of radius R centered in the origin. Then, since one can prove [42] that

$$\mathrm{Supp}(Q(f)(v)) \subset \mathcal{B}_0(\sqrt{2}R).$$

In order to write a spectral approximation which avoid aliasing, it is sufficient that the distribution function $f(v)$ is restricted on the cube $[-T,T]^d$ with $T \geq (2+\sqrt{2})R$. Successively, one should assume $f(v) = 0$ on $[-T,T]^d \setminus \mathcal{B}_0(R)$ and extend $f(v)$ to a periodic function on the set $[-T,T]^d$. Let observe that the lower bound for T can be improved. For instance, the choice $T = (3+\sqrt{2})R/2$ guarantees the absence of intersection between periods where f is different from zero. However, since in practice the support of f increases with time, we can just minimize the errors due to aliasing [10] with spectral accuracy.

To further simplify the notation, let us take $T = \pi$ and hence $R = \lambda\pi$ with $\lambda = 2/(3+\sqrt{2})$ in the following. Hereafter, using one index to denote the d-dimensional sums, we have that the conservative approximate Fourier truncation f_N^c can be represented as

$$f_N^c(v) = \sum_{k=-N}^{N} \hat{f}_k^c e^{ik\cdot v}, \qquad \hat{f}_k^c = \hat{f}_k + \hat{C}_k^T \left(\langle f, \Phi\rangle - \langle f_N, \Phi\rangle\right)$$

where $\hat{f}_k$ are the standard Fourier coefficients in (3.1), Φ is the moment vector defined in Proposition 3.2, and $\hat{C}_k^T$ is defined in (3.11).

A.1. A conservative spectral quadrature. We then obtain a conservative spectral quadrature of our collision operator by a constrained projection of Q^R defined in (4.1) on the space of trigonometric polynomials of degree less or equal to N, i.e.

$$Q_N^{R,c}(f_N^c, f_N^c)(v) = \sum_{k=-N}^{N} \hat{Q}_k^c e^{ik\cdot v}$$

where using Definition 3.3 we have

$$\hat{Q}_k^c = \sum_{\substack{l,m=-N \\ l+m=k}}^{N} \hat{f}_l^c\, \hat{f}_m^c\, \hat{\beta}(l,m) - \hat{C}_k^T \langle Q_N^R(f_N^c, f_N^c), \Phi\rangle, \quad k = -N, \dots, N. \tag{A.1}$$

In the above equation $\hat{\beta}(l,m) = \mathcal{B}(l,m) - \mathcal{B}(m,m)$ are given by

$$\mathcal{B}(l,m) = \int_{\mathcal{B}_0(2\lambda\pi)} \int_{\mathbb{S}^{d-1}} |q|\sigma(|q|,\cos\theta) e^{-i(l\cdot q^+ + m\cdot q^-)}\, d\omega\, dq$$

with

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

Let us notice that the naive evaluation of (A.1) requires $O(n^2)$ operations, where $n = N^d$. This causes the spectral method to be computationally very expensive, especially in dimension three. In order to reduce the number of operations needed to evaluate the collision integral, the main idea is to use another representation of the collision operator, the so-called Carleman representation [11] which is obtained by using the following identity

$$\frac{1}{2}\int_{\mathbb{S}^{d-1}} F(|u|\sigma - u)\, d\sigma = \frac{1}{|u|^{d-2}} \int_{\mathbb{R}^d} \delta(2\, x\cdot u + |x|^2)\, F(x)\, dx.$$

This gives in our context the Boltzmann integral representation

$$Q(f,f) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \tilde{B}(x,y)\delta(x\cdot y)\,[f(v+y)\, f(v+x) - f(v+x+y)\, f(v)]\; dx\, dy, \tag{A.2}$$

with

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

Denoting by $Q^{F,R}(f,f)$ the truncation on $\mathcal{B}_0(R)$ of (A.2) we have the following conservative spectral quadrature formula

$$\hat{Q}_k^{F,c} = \sum_{\substack{l,m=-N\\ l+m=k}}^{N} \hat{\beta}_F(l,m)\, \hat{f}_l^c\, \hat{f}_m^c - \hat{C}_k^T \langle Q_N^{F,R}(f_N^c, f_N^c), \Phi\rangle, \quad k = -N, ..., N \tag{A.4}$$

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

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

A.2. Other conservative and equilibrium preserving methods. The constrained modes of the collision term in (A.1) (or the fast ones (A.4)) have the general form

$$\hat{Q}_k^c(f_N^c, f_N^c) = \hat{Q}_k(f_N^c, f_N^c) - \hat{C}_k^T \langle Q_N^R(f_N^c, f_N^c), \Phi\rangle. \tag{A.5}$$

The above representation is reminiscent of the conservative spectral method introduced in [26], with one major difference: the latter method performs the moment constrained projection using the grid points in the original velocity space and a finite difference discretization in the Fourier space. As a consequence its theoretical analysis is more difficult, in particular regarding its consistency properties [1]. Although for the

sake of brevity we have omitted the results in the numerical section, the constrained approach in [26] was also applied in combination with the classical Fourier-Galerkin approximation, and the accuracy obtained was comparable to that of the conservative spectral method, indicating that spectral consistency could also be demonstrated for this latter approach. The same thing was also suggested in [49] by combining the constrained projection in the original velocity space using the Fourier-Galerkin approximation in [36]. However, we have not explored these analogies further from a theoretical viewpoint, as they are beyond the scope of this article.
Let us also note that the equilibrium-preserving fast spectral method in [20, 41] has also a similar form and the corresponding Fourier modes can be written as

$$\hat{Q}_k^e(f_N, f_N) = \hat{Q}_k(f_N, f_N) - \hat{Q}_k(M_N, M_N), \tag{A.6}$$

where $M_N = \mathcal{P}_N M$ and M is the local Maxwellian equilibrium. The two projections (A.5) and (A.6) can be combined to originate a conservative and equilibrium preserving spectral method defined as

$$\begin{aligned} \hat{Q}_k^{c,e}(f_N^c, f_N^c) &= \hat{Q}_k^c(f_N^c, f_N^c) - \hat{Q}_k^c(M_N^c, M_N^c) \\ &= \hat{Q}_k(f_N^c, f_N^c) - \hat{Q}_k(M_N^c, M_N^c) \\ &\quad - \hat{C}_k^T \left(\langle Q_N^R(f_N^c, f_N^c) - Q_N^R(M_N^c, M_N^c), \Phi \rangle \right), \end{aligned} \tag{A.7}$$

where now $M_N^c = \mathcal{P}_N^c M$. The spectral consistency of the method in (A.7) is a direct consequence of the theoretical results in this article and we will omit the details. The formulations (A.5), (A.6) and (A.7) show a common analogy between all the different approaches, namely to use a spectrally small correction to recover the conservation properties and/or the correct equilibrium state.

A.3. The fast algorithm. To reduce the number of operation needed to evaluate (A.4), we look for a convolution structure. The aim is to approximate each $\hat{\beta}_F(l, m)$ by a sum

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

where A represents the number of finite possible directions of collisions. This finally gives a sum of A discrete convolutions and, consequently, the algorithm can be computed in $O(A\, N \log_2 N)$ operations by means of standard FFT technique [10].

In order to get this convolution form, we make the decoupling assumption

$$\tilde{B}(x, y) = a(|x|)\, b(|y|).$$

This assumption is satisfied if $\tilde{B}$ is constant. This is the case of Maxwellian molecules in dimension two, and hard spheres in dimension three. Indeed, using kernel (2.3) in (A.3), one has

$$\tilde{B}(x, y) = 2^{d-1} C_\alpha (|x|^2 + |y|^2)^{-\frac{d-\alpha-2}{2}},$$

so that $\tilde{B}$ is constant if $d = 2$, $\alpha = 0$ and $d = 3$, $\alpha = 1$.

Remark A.1. Let us detail the computations in dimension 2 for $\tilde{B} = 1$, i.e. Maxwellian molecules. Here we write x and y in spherical coordinates $x = \rho e$ and $y = \rho' e'$ to get

$$\mathcal{B}_F(l, m) = \frac{1}{4} \int_{\mathbb{S}^1} \int_{\mathbb{S}^1} \delta(e \cdot e') \left[\int_{-R}^{R} e^{i\rho(l \cdot e)}\, d\rho \right] \left[\int_{-R}^{R} e^{i\rho'(m \cdot e')}\, d\rho' \right] de\, de'.$$

Then, denoting $\phi_R^2(s) = \int_{-R}^{R} e^{i\rho s}\, d\rho$, for $s \in \mathbb{R}$, we have the explicit formula

$$\phi_R^2(s) = 2\, R\, \mathrm{Sinc}(Rs),$$

where $\mathrm{Sinc}(x) = \frac{\sin(x)}{x}$ for $x \neq 0$. This explicit formula is further plugged in the expression of $\mathcal{B}_F(l, m)$ and using its parity property, this yields

$$\mathcal{B}_F(l, m) = \int_0^{\pi} \phi_R^2(l \cdot e_\theta)\, \phi_R^2(m \cdot e_{\theta+\pi/2})\, d\theta.$$

Finally, a regular discretization of A equally spaced points, which is spectrally accurate because of the periodicity of the function, gives

$$\mathcal{B}_F(l, m) = \frac{\pi}{M} \sum_{p=1}^{A} \alpha_p(l)\alpha'_p(m),$$

with

$$\alpha_p(l) = \phi_R^2(l \cdot e_{\theta_p}), \qquad \alpha'_p(m) = \phi_R^2(m \cdot e_{\theta_p+\pi/2})$$

where $\theta_p = \pi p/A$.

In practice, for the two dimensional case in velocity we choose a number of discretization angles $A = 8$. This is enough to guarantee a good accuracy of the results in many situations. However, when close to shock waves or a boundary layers, it may happen that spectral accuracy is lost [23]. In these situations, in order to keep a good accuracy in the resolution of the collision operator, one may need a larger set of angles.

REFERENCES

[1] Alonso, R. J., Gamba, I. M., and Tharkabhushanam, S. H. Convergence and error estimates for the Lagrangian-based conservative spectral method for Boltzmann equations. *SIAM J. Numer. Anal. 56*, 6 (2018), 3534–3579.

[2] Bellomo, N., and Gatignol, R., Eds. *Lecture Notes on the Discretization of the Boltzmann Equation*, vol. 63 of *Series on Advances in Mathematics for Applied Sciences*. World Scientific, 2003.

[3] Bessemoulin-Chatard, M., Herda, M., and Rey, T. Hypocoercivity and diffusion limit of a finite volume scheme for linear kinetic equations,. *Math. Comp.* (2020).

[4] Bird, G. *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, 2nd ed. Oxford University Press, 1994.

[5] Bobylev, A. V. Exact solutions of the Boltzmann equation. *Dokl. Akad. Nauk SSSR 225*, 6 (1975), 1296–1299.

[6] Bobylev, A. V., and Rjasanow, S. Difference scheme for the Boltzmann equation based on the fast Fourier transform. *Eur. J. Mech. B Fluids 16*, 2 (1997), 293–306.

[7] Bobylev, A. V., and Rjasanow, S. Fast deterministic method of solving the Boltzmann equation for hard spheres. *Eur. J. Mech. B Fluids 18*, 5 (1999), 869–887.

[8] Bobylev, A. V., and Rjasanow, S. 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 (2000), 289–310.

[9] Cai, Z., Fan, Y., and Ying, L. An entropic Fourier method for the Boltzmann equation. *SIAM J. Sci. Comput. 40*, 5 (2018), A2858–A2882.

[10] Canuto, C., Hussaini, M., Quarteroni, A., and Zang, T. *Spectral Methods in Fluid Dynamics*. Springer Series in Computational Physics. Springer-Verlag, New York, 1988.

[11] Carleman, T. Sur la théorie de l'équation intégrodifférentielle de Boltzmann. *Acta Math. 60*, 1 (1933), 91–146.

[12] Cercignani, C., Illner, R., and Pulvirenti, M. *The Mathematical Theory of Dilute Gases*, vol. 106 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994.

[13] Degond, P., Pareschi, L., and Russo, G., Eds. *Modeling and computational methods for kinetic equations*. Modeling and Simulation in Science, Engineering and Technology Series. Birkhäuser, 2004.

[14] Dimarco, G., Li, Q., Pareschi, L., and Yan, B. Numerical methods for plasma physics in collisional regimes. *J. Plasma Phys. 81*, 1 (2015), 305810106.

[15] Dimarco, G., Loubère, R., Narski, J., and Rey, T. An efficient numerical method for solving the Boltzmann equation in multidimensions. *J. Comput. Phys. 353* (2018), 46–81.

[16] Dimarco, G., and Pareschi, L. Numerical methods for kinetic equations. *Acta Num. 23* (2014), 369–520.

[17] Filbet, F., Hu, J., and Jin, S. A numerical scheme for the quantum Boltzmann equation with stiff collision terms. *ESAIM Math. Model. Numer. Anal. 46*, 2 (2012), 443–463.

[18] Filbet, F., and Mouhot, C. Analysis of Spectral Methods for the Homogeneous Boltzmann Equation. *Trans. Amer. Math. Soc. 363* (2011), 1947–1980.

[19] Filbet, F., Mouhot, C., and Pareschi, L. Solving the Boltzmann Equation in N log2 N. *SIAM J. Sci. Comput. 28*, 3 (2007), 1029–1053.

[20] Filbet, F., Pareschi, L., and Rey, T. On steady-state preserving spectral methods for homogeneous Boltzmann equations. *C. R. Acad. Sci. Paris, Ser. I 353*, 4 (2015), 309–314.

[21] Filbet, F., Pareschi, L., and Toscani, G. Accurate Numerical Methods for the Collisional Motion of (Heated) Granular Flows. *J. Comput. Phys. 202*, 1 (2005), 216–235.

[22] Filbet, F., and Russo, G. High order numerical methods for the space non-homogeneous Boltzmann equation. *J. Comput. Phys. 186*, 2 (2003), 457–480.

[23] Gamba, I. M., Haack, J. R., Hauck, C. D., and Hu, J. A fast spectral method for the Boltzmann collision operator with general collision kernels. *SIAM J. Sci. Comput. 39*, 4 (2017), B658–B674.

[24] Gamba, I. M., Panferov, V., and Villani, C. Upper Maxwellian bounds for the spatially homogeneous Boltzmann equation. *Arch. Rat. Mech. Anal 194*, 1 (2009), 253–282.

[25] Gamba, I. M., and Rjasanow, S. Galerkin–Petrov approach for the Boltzmann equation. *Journal of Computational Physics 366* (2018), 341–365.

[26] Gamba, I. M., and Tharkabhushanam, S. H. Spectral-Lagrangian methods for collisional models of non-equilibrium statistical states. *J. Comput. Phys. 228*, 6 (Apr. 2009), 2012–2036.

[27] Gamba, I. M., and Tharkabhushanam, S. H. Shock and Boundary Structure formation by Spectral-Lagrangian methods for the Inhomogeneous Boltzmann Transport Equation. *J. Comput. Math. 28*, 4 (2010), 430–460.

[28] Hu, J., Huang, X., Shen, J., and Yang, H. A fast Petrov-Galerkin spectral method for the multi-dimensional Boltzmann equation using mapped Chebyshev functions. *preprint arXiv:2105.08806* (2021).

[29] Hu, J., and Ma, Z. A fast spectral method for the inelastic Boltzmann collision operator and application to heated granular gases. *J. Comput. Phys. 385* (2019), 119–134.

[30] Hu, J., Qi, K., and Yang, T. A new stability and convergence proof of the Fourier-Galerkin spectral method for the spatially homogeneous Boltzmann equation. *arXiv preprint arXiv:2007.05184* (2020).

[31] Hu, J., and Ying, L. A fast spectral algorithm for the quantum Boltzmann collision operator. *Commun. Math. Sci. 10*, 3 (2012), 989–999.

[32] Jaiswal, S., Hu, J., and Alexeenko, A. A. Fast deterministic solution of the full Boltzmann equation on graphics processing units. In *AIP Conference Proceedings* (2019), vol. 2132, AIP Publishing LLC, p. 060001.

[33] Krook, M., and Wu, T. T. Exact solutions of the Boltzmann equation. *Phys. Fluids 20*, 10 (1977), 1589.

[34] Li, Q., and Pareschi, L. Exponential Runge–Kutta for the inhomogeneous Boltzmann equations with high order of accuracy. *J. Comput. Phys. 259* (2014), 402–420.

[35] Mouhot, C., and Pareschi, L. Fast algorithms for computing the Boltzmann collision operator. *Math. Comp. 75*, 256 (2006), 1833–1852 (electronic).

[36] Mouhot, C., Pareschi, L., and Rey, T. Convolutive Decomposition and Fast Summation Methods for Discrete-Velocity Approximations of the Boltzmann Equation. *ESAIM Math. Model. Numer. Anal. 47*, 5 (2013), 1515–1531.

[37] Mouhot, C., and Villani, C. Regularity theory for the spatially homogeneous Boltzmann equation with cut-off. *Arch. Rat. Mech. Anal. 173*, 2 (2004), 169–212.

[38] Panferov, V. A., and Heintz, A. G. A New Consistent Discrete Velocity Model for the Boltzmann Equation. *Math. Models Methods Appl. Sci. 25*, 7 (2002), 571–593.

[39] Pareschi, L., and Perthame, B. A Fourier Spectral Method for Homogeneous Boltzmann Equations. *Transport Theory Statist. Phys. 25*, 3 (1996), 369–382.

[40] Pareschi, L., and Rey, T. Residual equilibrium schemes for time dependent partial differential equations. *Comput. Fluids 156* (2017), 329–342.

[41] Pareschi, L., and Rey, T. On the stability of equilibrium preserving spectral methods for the homogeneous Boltzmann equation. *App. Math. Letters* (2021).

[42] Pareschi, L., and Russo, G. Numerical Solution of the Boltzmann Equation I : Spectrally Accurate Approximation of the Collision Operator. *SIAM J. Numer. Anal. 37*, 4 (2000), 1217–1245.

[43] Pareschi, L., and Russo, G. On the stability of spectral methods for the homogeneous Boltzmann equation. *Transport Theory Statist. Phys. 29*, 3-5 (2000), 431–447.

[44] Pareschi, L., Russo, G., and Toscani, G. Fast spectral methods for the Fokker-Planck-Landau collision operator. *J. Comput. Phys. 165*, 1 (2000), 216–236.

[45] Pareschi, L., Toscani, G., and Villani, C. Spectral methods for the non cut-off Boltzmann equation and numerical grazing collision limit. *Num. Math. 93*, 3 (Jan. 2003), 527–548.

[46] Rogier, F., and Schneider, J. A direct method for solving the Boltzmann equation. *Transport Theory Statist. Phys. 23*, 1-3 (1994), 313–338.

[47] Villani, C. *A Review of Mathematical Topics in Collisional Kinetic Theory*. Elsevier Science, 2002.

[48] Wang, Y., and Cai, Z. Approximation of the Boltzmann collision operator based on Hermite spectral method. *Journal of Computational Physics 397* (2019), 108815.

[49] Wu, L., White, C., Scanlon, T. J., Reese, J. M., and Zhang, Y. Deterministic numerical solutions of the Boltzmann equation using the fast spectral method. *J. Comput. Phys. 250* (2013), 27–52.

[50] Wu, L., Zhang, J., Reese, J. M., and Zhang, Y. A fast spectral method for the Boltzmann equation for monoatomic gas mixtures. *J. Comput. Phys. 298* (Oct. 2015), 602–621.