Hyperbolic cross approximation for the spatially homogeneous Boltzmann equation

E. Fonn P. Grohs R. Hiptmair

March 22, 2013

Abstract

We consider the non-linear spatially homogeneous Boltzmann equation and its Fourier spectral discretization in velocity space involving periodic continuation of the density and a truncation of the collision operator. We allow discretization based on arbitrary sets of active Fourier modes with particular emphasis on the so-called hyperbolic cross approximation. We also discuss an offset method that takes advantage of the known equilibrium solutions.

Extending the analysis in [F. Filbet and C. Mouhot, *Analysis of spectral methods for the homogeneous Boltzmann equation*, Transactions of the American Mathematical Society 363(4):1947–1980, 2011] we establish consistency estimates for the discrete collision operators and stability of the semi-discrete evolution. Under an assumption of Gaussian-like decay of the discrete solution we give a detailed bound for H^s-Sobolev norms of the error due to Fourier spectral discretization.

1 Introduction

In this paper we are concerned with the discretization in velocity and time of the spatially homogeneous Boltzmann equation

$$\frac{\partial f}{\partial t}(t,\boldsymbol{v}) = Q(f,f)(t,\boldsymbol{v})\;, \tag{1}$$

where $f : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^+$ is an unknown density function, for which initial values $f_0 = f_0(\boldsymbol{v})$ at $t = 0$ are prescribed. The bilinear collision operator Q takes the form (dropping the variable t for the sake of readability)

$$Q(f,h)(\boldsymbol{v}) = \int_{\mathbb{R}^d}\int_{S^{d-1}} B(\|\boldsymbol{v}-\boldsymbol{v}_*\|,\cos\theta)(h'_*f' - h_*f)\,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{v}_*, \tag{2}$$

where the notation $Q(f,h)(\boldsymbol{v})$ means $Q(f,h)$ evaluated at $\boldsymbol{v}$, and we have used the common shorthand notation

$$f = f(\boldsymbol{v}), \qquad h_* = h(\boldsymbol{v}_*), \qquad f' = f(\boldsymbol{v}'), \qquad h'_* = h(\boldsymbol{v}'_*)\;.$$

The pre- and post-collision velocities $(\boldsymbol{v},\boldsymbol{v}_*)$ and $(\boldsymbol{v}',\boldsymbol{v}'_*)$ are related through the transformation

$$\boldsymbol{v}' = \frac{1}{2}\left(\boldsymbol{v}+\boldsymbol{v}_* + \|\boldsymbol{v}-\boldsymbol{v}_*\|\boldsymbol{\sigma}\right), \qquad \boldsymbol{v}'_* = \frac{1}{2}\left(\boldsymbol{v}+\boldsymbol{v}_* - \|\boldsymbol{v}-\boldsymbol{v}_*\|\boldsymbol{\sigma}\right)\;.$$

The two terms $h'_* f'$ and $h_* f$ are called *gain* and *loss* parts respectively, and it is often useful (and we will do so when necessary) to split the collision operator and write

$$Q(f,h) = Q^+(f,h) - Q^-(f,h) \; .$$

For the physical background and mathematical analysis of (1) we refer to [4, 23]. Throughout this paper we make the following customary assumption on the collision kernel B in (2).

Assumption 1. *Assume that B is separable, with a power law dependence on the relative velocity, that is,*

$$B(\|\boldsymbol{u}\|, \cos\theta) = \|\boldsymbol{u}\|^\lambda b(\cos\theta), \qquad \textit{with} \quad \lambda \geq -\frac{d}{2} \; , \tag{3}$$

and some function $b : [-1,1] \to \mathbb{R}$ that satisfies Grad's cutoff assumption

$$A(\boldsymbol{u}) := \int_{S^{d-1}} b(\cos\theta)\, \mathrm{d}\boldsymbol{\sigma} = \int_{S^{d-1}} b\left(\left\langle \frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \boldsymbol{\sigma} \right\rangle \right) \mathrm{d}\boldsymbol{\sigma} < \infty \; . \tag{4}$$

Owing to the translation invariance of the collision operator, it lends itself to a Fourier spectral discretization, which can exploit the convolution structure of Q, see [1, 14, 17] and [9, 10, 8, 19] for numerical simulations based on that method. The approach entails two fundamental modifications of the collision operator:

(i) the restriction of the integration over $\mathbb{R}^d$ in (2) to a ball of finite radius $R > 0$, see Section 2.1 for details,

(ii) the truncation of the velocity space $\mathbb{R}^d$ to a cube $\mathcal{D}_L := [-L, L[^d$ plus periodic continuation.

These result in a perturbed evolution

$$\frac{\partial}{\partial t} f_R = Q^R(f_R, f_R), \qquad f(0) = f_0 \; . \tag{5}$$

Clearly, its solution depends on both R and L. Subsequently, (5) is projected onto a subspace of $L^2(\mathcal{D}_L)$ spanned by a finite set of Fourier modes. An in-depth analysis of the discretization error due to this last step has recently been developed by Filbet and Mouhot in [7], and in parts of our work we are going to rely on their techniques.

Our contribution. One of our main goals was to provide a priori estimates of the discretization error in H^s-Sobolev norms, $s \geq 0$, for rather general sets of Fourier modes, which includes as a special cases the so-called hyperbolic crosses. For certain types of solutions, this scheme offers much more efficient approximation compared to the full tensor product space of Fourier modes. This extends the results of [7].

In another respect we go beyond [7]: we also quantify the impact of switching from (1) to (5), conditional on the following assumption.

Assumption 2. *For given Sobolev index $s \geq 0$, assume that there exist $C_0 > 0$ and $a > 0$ so that the solution f of (1) fulfills*

$$|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{v})| \leq C \exp\left(-a\|\boldsymbol{v}\|^2\right) \; ,$$

and that there exists a polynomial $\Pi(L)$ so that

$$\| \exp\left(a\|\boldsymbol{v}\|^2\right) \partial^{\boldsymbol{\alpha}} f_R \|_{L^2(\mathcal{D}_L)} \leq \Pi(L) \; ,$$

for all $|\boldsymbol{\alpha}|_1 \leq s$ and f_R solving (5).

Remark 3. The decay condition on f_R is formulated in this form since it is $2L$-periodic, and thus it cannot satisfy any classical decay condition on the whole of $\mathbb{R}^d$. The polynomial dependence on L is allowed so that equilibrium solutions fit the assumption.

The decay of f is well known, see for example [2]. Yet, so far, there is no rigorous proof that the Assumption 2 holds for f_R. We have only numerical evidence that this is true, see Section 6.4.

Outline. Our paper is structured as follows. In Section 2 we develop the details of the truncation of Q and the spectral discretization. The following Section 3 introduces the concrete Fourier spectral Galerkin discretization of (5) for families of hyperbolic cross Fourier modes. In Section 4 we develop the main a priori convergence estimates, which are summarized in Theorem 13 (consistency estimates), Theorem 23 (Fourier spectral discretization error for (5)), and Proposition 27 (estimate for $f - f_R$). We emphasize that Assumption 2 is essential only for this last result. Numerical experiments are reported in Section 6. They highlight the conditions that have to met for satisfactory performance of the hyperbolic cross approximation. They also convey the importance of choosing a sufficiently small ratio $R : L$ in order to avoid aliasing.

2 Fourier Spectral Discretization of the Boltzmann collision operator

In this section we aim to discretize the solution f by a finite Fourier series. Of course, integrals such as (6) and (7) will diverge for periodic f. Thus, a truncation in velocity is required.

2.1 Truncation in velocity space

In this section, we develop equivalent *truncated* forms for the collision operator as defined via

$$Q(f,h)(\boldsymbol{v}) = \int_{\mathbb{R}^d} \int_{S^{d-1}} B(\|\boldsymbol{g}\|, \cos\theta)(h'_* f' - h(\boldsymbol{v}-\boldsymbol{g})f(\boldsymbol{v}))\,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{g}, \tag{6}$$

which is achieved through a change of variables $\boldsymbol{g} = \boldsymbol{v} - \boldsymbol{v}_*$, and the *Carleman representation*

$$Q(f,h)(\boldsymbol{v}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{B}(\boldsymbol{x},\boldsymbol{y})\delta(\boldsymbol{x}\cdot\boldsymbol{y})\left[h(\boldsymbol{v}+\boldsymbol{y})f(\boldsymbol{v}+\boldsymbol{x}) - h(\boldsymbol{v}+\boldsymbol{x}+\boldsymbol{y})f(\boldsymbol{v})\right]\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}, \tag{7}$$

whose development is detailed in [16].

The following proposition is given in [17], with $\mathcal{B}_R$ denoting the ball of radius R centered at 0.

Proposition 4. *Let* $\operatorname{supp} f, h \subset \mathcal{B}_R$. *Then,* $\operatorname{supp} Q(f,h) \subset \mathcal{B}_{\sqrt{2}R}$, *and*

$$Q(f,h)(\boldsymbol{v}) = \int_{\mathcal{B}_{2R}} \int_{S^{d-1}} B(\|\boldsymbol{g}\|,\theta)(h'_* f' - h_* f)\,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{g}$$

for $\boldsymbol{v} \in \mathcal{B}_{\sqrt{2}R}$. *Under these assumptions,* $\boldsymbol{v}', \boldsymbol{v}'_*, \boldsymbol{v}-\boldsymbol{g} \in \mathcal{B}_{(2+\sqrt{2})R}$ *for all* $\boldsymbol{g} \in \mathcal{B}_{2R}$.

Thus, by considering f restricted on the cube $\mathcal{D}_L = [-L, L]^d$, with $f(\boldsymbol{v}) = 0$ on $\mathcal{D}_L \setminus \mathcal{B}_R$, extended periodically to all of $\mathbb{R}^d$, we can evaluate $Q(f, f)$ without aliasing if R/L is small enough. In practice, L should be chosen large enough to accommodate the necessary number of timesteps while minimizing the aliasing errors, as the support of f will grow.

Common practice [18, 16] is to choose R, L so that aliasing is avoided for one single application of the collision operator. This holds if

$$\kappa = \frac{R}{L} \leq \frac{2}{3 + \sqrt{2}}.$$

In Section 6.5 we present some numerical experiments showing how the choice of κ can significantly impact the performance of a numerical scheme.

The task is now to bring the Carleman representation (7) into truncated form also, so that the two truncated representations are equivalent. This is the content of the following proposition.

Proposition 5. *Consider f and h restricted to the cube $\mathcal{D}_L$ with $f(\boldsymbol{v}) = 0$ on $\mathcal{D}_L \setminus \mathcal{B}_R$, extended periodically to all of $\mathbb{R}^d$, with $\kappa = R/L \leq 2/(3+\sqrt{2})$. Then, for $v \in \mathcal{B}_{\sqrt{2}R}$, the following representations of $Q(f, h)(\boldsymbol{v})$ agree.*

$$Q(f,h)(\boldsymbol{v}) = \int_{\mathcal{B}_{2R}} \int_{S^{d-1}} B(\|\boldsymbol{g}\|, \theta)(h'_* f' - h_* f) \, \mathrm{d}\boldsymbol{\sigma} \, \mathrm{d}\boldsymbol{g} =: Q_P^R(f,g)(\boldsymbol{v}) \tag{8}$$

$$= \int_{\mathcal{B}_{\sqrt{2}R}} \int_{\mathcal{B}_{\sqrt{2}R}} \tilde{B}(\boldsymbol{x}, \boldsymbol{y}) \delta(\boldsymbol{x} \cdot \boldsymbol{y}) \left[h(\boldsymbol{v} + \boldsymbol{y}) f(\boldsymbol{v} + \boldsymbol{x}) - h(\boldsymbol{v} + \boldsymbol{x} + \boldsymbol{y}) f(\boldsymbol{v}) \right] \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \tag{9}$$

$$=: Q_M^R(f,g)(\boldsymbol{v}).$$

Proof. Representation (8) follows from Proposition 4. Following the policy from [16], it can be shown that

$$Q(f,h)(\boldsymbol{v}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{B}(\boldsymbol{x}, \boldsymbol{y}) \delta(\boldsymbol{x} \cdot \boldsymbol{y}) \chi_{\mathcal{B}_{2R}}(\boldsymbol{x} + \boldsymbol{y}) \\ \left[h(\boldsymbol{v} + \boldsymbol{y}) f(\boldsymbol{v} + \boldsymbol{x}) - h(\boldsymbol{v} + \boldsymbol{x} + \boldsymbol{y}) f(\boldsymbol{v}) \right] \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}.$$

It remains to show that when $\|\boldsymbol{x} + \boldsymbol{y}\| > 2R$, i.e. when $\chi_{\mathcal{B}_{2R}}(\boldsymbol{x} + \boldsymbol{y}) = 0$, we have $h(\boldsymbol{v} + \boldsymbol{y}) f(\boldsymbol{v} + \boldsymbol{x}) = h(\boldsymbol{v} + \boldsymbol{x} + \boldsymbol{y}) f(\boldsymbol{v}) = 0$.

Under these assumptions, it is clear that either $\|\boldsymbol{v}\| > R$ or $\|\boldsymbol{v} + \boldsymbol{x} + \boldsymbol{y}\| > R$. Furthermore, since $\boldsymbol{x} \perp \boldsymbol{y}$, we also have $\|\boldsymbol{x} - \boldsymbol{y}\| > 2R$, so either $\|\boldsymbol{v} + \boldsymbol{y}\| > R$ or $\|\boldsymbol{v} + \boldsymbol{x}\| = \|(\boldsymbol{v} + \boldsymbol{y}) + (\boldsymbol{x} - \boldsymbol{y})\| > R$.

Last, for $\|\boldsymbol{x}\|, \|\boldsymbol{y}\| \leq \sqrt{2}R$, we have $\max\{\|\boldsymbol{v}\|, \|\boldsymbol{v} + \boldsymbol{x}\|, \|\boldsymbol{v} + \boldsymbol{y}\|, \|\boldsymbol{v} + \boldsymbol{x} + \boldsymbol{y}\|\} \leq 2L - R$, so aliasing is avoided. This concludes the proof. □

Henceforth, we will denote by Q^R the truncated version of Q as defined in (8) and (9). We will also denote by $Q^{R,+}$ and $Q^{R,-}$ the corresponding truncated versions of Q^+ and Q^-. We will always assume that the ratio of R and L is fixed and denote it by the constant truncation parameter $\kappa = R/L$.

2.2 Fourier discretization

Following ideas from [11, 3], let us now discretize f by representing it as a truncated d-dimensional Fourier series in $\boldsymbol{v}$,

$$f_{\mathcal{A}}(t,\boldsymbol{v}) = \sum_{\boldsymbol{k}\in\mathcal{A}} \hat{f}_{\boldsymbol{k}}(t) e^{i\boldsymbol{k}\cdot\boldsymbol{v}}, \tag{10}$$

where $\mathcal{A} \subset (\pi/L)\mathbb{Z}^d$ is some discrete and finite but so far unspecified set of Fourier modes. Then, (1) yields

$$\sum_{\boldsymbol{k}\in\mathcal{A}} \frac{d\hat{f}_{\boldsymbol{k}}}{dt} e^{i\boldsymbol{k}\cdot\boldsymbol{v}} = \sum_{\boldsymbol{l},\boldsymbol{m}\in\mathcal{A}} \hat{f}_{\boldsymbol{l}}\hat{f}_{\boldsymbol{m}} Q^R\left(e^{i\boldsymbol{l}\cdot\boldsymbol{v}}, e^{i\boldsymbol{m}\cdot\boldsymbol{v}}\right). \tag{11}$$

Substituting the Fourier modes into (6) or (7), we find that there exist coefficients $\hat{\beta}(\boldsymbol{l},\boldsymbol{m})$ so that

$$Q^R\left(e^{i\boldsymbol{l}\cdot\boldsymbol{v}}, e^{i\boldsymbol{m}\cdot\boldsymbol{v}}\right) = \hat{\beta}(\boldsymbol{l},\boldsymbol{m}) e^{i(\boldsymbol{m}+\boldsymbol{l})\cdot\boldsymbol{v}} \tag{12}$$

for all $\boldsymbol{l},\boldsymbol{m} \in (\pi/L)\mathbb{Z}^d$. Plugging this into (11), discarding coefficients outside $\mathcal{A}$ and comparing the remaining coefficients, gives us the following quadratic ordinary differential equation (ODE) for the coefficients $\hat{f}_{\boldsymbol{k}}$:

$$\frac{d\hat{f}_{\boldsymbol{k}}}{dt} = \sum_{\substack{\boldsymbol{l},\boldsymbol{m}\in\mathcal{A}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \hat{f}_{\boldsymbol{l}}\hat{f}_{\boldsymbol{m}}\hat{\beta}(\boldsymbol{l},\boldsymbol{m}), \quad \boldsymbol{k}\in\mathcal{A}\,. \tag{13}$$

This is the Fourier space equivalent to

$$\frac{\partial}{\partial t} f_{\mathcal{A}} = P_{\mathcal{A}} Q^R(f_{\mathcal{A}}, f_{\mathcal{A}}) \tag{14}$$

where $P_{\mathcal{A}}$ is the L^2-orthogonal projection on the the space spanned by the Fourier modes in $\mathcal{A}$.

The coefficients $\hat{\beta}(\boldsymbol{l},\boldsymbol{m})$ are called the *kernel modes*. These can be evaluated through (12). Note that since the Fourier modes $e^{i\boldsymbol{k}\cdot\boldsymbol{v}}$ do not satisfy the conditions of Proposition 5, representations (8) and (9) will yield *different* values for $\hat{\beta}$. In the following, we will denote by $\hat{\beta}_d^P$ those values arising from (8), and by $\hat{\beta}_d^M$ those arising from (9).

Separating gain and loss terms as described in Section 1, we find that we have

$$\hat{\beta}_d^*(\boldsymbol{l},\boldsymbol{m}) = \beta_d^*(\boldsymbol{l},\boldsymbol{m}) - \beta_d^*(\boldsymbol{m},\boldsymbol{m}), \qquad * = P, M$$

where the coefficients β_d^* are given, for general kernels B and $\tilde{B}$, as

$$\beta_d^P(\boldsymbol{l},\boldsymbol{m}) = \int_{\mathcal{B}_{2R}} \int_{S^{d-1}} B(\|\boldsymbol{g}\|, \cos\theta) \exp\left[-i\boldsymbol{g}\cdot\frac{\boldsymbol{l}+\boldsymbol{m}}{2} - i\|\boldsymbol{g}\|\sigma\cdot\frac{\boldsymbol{m}-\boldsymbol{l}}{2}\right] \mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{g} \tag{15}$$

$$\beta_d^M(\boldsymbol{l},\boldsymbol{m}) = \int_{\mathcal{B}_{\sqrt{2}R}} \int_{\mathcal{B}_{\sqrt{2}R}} \tilde{B}(\boldsymbol{x},\boldsymbol{y}) \delta(\boldsymbol{x}\cdot\boldsymbol{y}) e^{i\boldsymbol{l}\cdot\boldsymbol{x}} e^{i\boldsymbol{m}\cdot\boldsymbol{y}}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}, \tag{16}$$

for $\boldsymbol{l},\boldsymbol{m}\in\mathcal{A}$ [18, 16].

3 Choice of $\mathcal{A}$

So far, we have left the choice of Fourier space $\mathcal{A}$ unaccounted for. Of course, it is required that $\mathcal{A}$ is a subset of the scaled lattice grid

$$\mathcal{A} \subset \mathcal{A}_{\text{full}} = \frac{\pi}{L}\mathbb{Z}^d.$$

The obvious choice, and the one adopted in [18] and [16] is the full d-dimensional discrete Fourier representation with N degrees of freedom in each direction:

$$\mathcal{A}_{\text{FF}}(N) = \frac{\pi}{L}\left\{-\frac{N}{2}, -\frac{N}{2}+1, \ldots, \frac{N}{2}-1\right\}^d .$$

An alternative choice is the *hyperbolic cross* Fourier space, which can be considered the frequency-space equivalent of sparse grids (see for example [12] for more details),

$$\mathcal{A}_T(N) := \left\{ \boldsymbol{k} \in \frac{\pi}{L}\mathbb{Z}^d : \prod_{j=1}^{d}(1+|\tilde{k}_j|)\cdot(1+|\tilde{\boldsymbol{k}}|_\infty)^{-T} \leq (1+N)^{1-T} \right\} ,$$

where $T \leq 1$, with the limiting set $A_{-\infty}(N) = \mathcal{A}_{\text{FF}}(N)$, where the vectors $\boldsymbol{k}$ and $\tilde{\boldsymbol{k}}$ are, as in the previous section, related by

$$\boldsymbol{k} = \frac{\pi}{L}\tilde{\boldsymbol{k}} ,$$

rendering $\tilde{\boldsymbol{k}}$ integral.

Thus, we deal with two discretization parameters: (i) the parameter T, which controls the "fatness" of the hyperbolic cross, with the classical hyperbolic cross being given by $T = 0$, and (ii) the "resolution parameter" N. For $T = 0$, we have $\sharp\mathcal{A}_0(N) = O(N \log^d N)$ compared to $\sharp\mathcal{A}_{\text{FF}}(N) = N^d$.

Importantly, the hyperbolic cross does *not* provide good approximations of near-Maxwellians. Indeed, the spectrum of a Maxwellian is a Maxwellian centered at the origin, and its rotational symmetry makes it well suited for an approximation by $\mathcal{A}_{\text{FF}}(N)$, see Figure 1.

This indicates that the hyperbolic cross would be a poor choice for approximating near-equilibrium solutions. Of course, it could still provide a useful tool for certain situations with f far removed from equilibrium. Moreover, the results in this paper are not confined to the classical hyperbolic cross ($T = 0$) but to the whole family of sampling sets $\mathcal{A}_T$, as well as $\mathcal{A}_{\text{FF}}$, which allows fine-tuning according to the degree of isotropy in the solution.

3.1 Observables

For the linear functionals mass (ρ), momentum ($\boldsymbol{u}$) and energy (E), which define the equilibrium solution, we have representations in terms of the coefficients $\hat{f}_{\boldsymbol{k}}$ from (10):

$$\rho(f) = \frac{1}{(2L)^d}\sum_{\boldsymbol{k}\in\mathcal{A}} \hat{\rho}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{k}}, \quad \rho\boldsymbol{u}(f) = \frac{1}{(2L)^d}\sum_{\boldsymbol{k}\in\mathcal{A}} \hat{\boldsymbol{u}}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{k}}, \quad \rho E(f) = \frac{1}{(2L)^d}\sum_{\boldsymbol{k}\in\mathcal{A}} \hat{E}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{k}}.$$

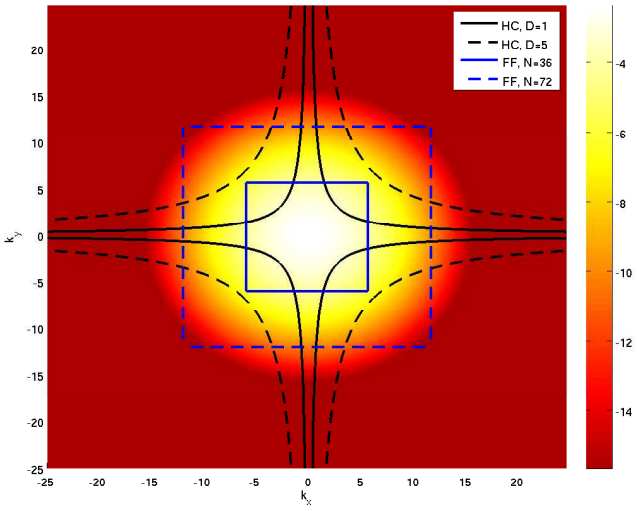


Figure 1: Amplitude of the spectrum of a Gaussian, with the Fourier modes belonging to hyperbolic crosses and full grids of comparable size delineated for comparison. It seems that full grids offer much better approximation of Gaussians.

The quantities $\hat{\rho}_{\boldsymbol{k}}$, $\hat{\boldsymbol{u}}_{\boldsymbol{k}}$ and $\hat{E}_{\boldsymbol{k}}$ are given as (rescaled) Fourier coefficients of the functions 1, $\boldsymbol{v}$ and $\|\boldsymbol{v}\|^2$:

$$\frac{1}{(2L)^d}\begin{pmatrix}\hat{\rho}_{\boldsymbol{k}}\\ \hat{\boldsymbol{u}}_{\boldsymbol{k}}\\ \hat{E}_{\boldsymbol{k}}\end{pmatrix} = \int_{\mathcal{D}_L}\begin{pmatrix}1\\ \boldsymbol{v}\\ \|\boldsymbol{v}\|^2\end{pmatrix} e^{i\boldsymbol{k}\cdot\boldsymbol{v}}\,\mathrm{d}\boldsymbol{v}$$

Clearly, for mass, we have $\hat{\rho}_{\boldsymbol{k}} = (2L)^d\delta_{\boldsymbol{k},\boldsymbol{0}}$.

For $\boldsymbol{u}$ and E, we find that $\hat{\boldsymbol{u}}_{\boldsymbol{k}} = \hat{E}_{\boldsymbol{k}} = 0$ whenever $\boldsymbol{k}$ is off the axes, i.e. there is more than one nonzero element of $\boldsymbol{k}$. Thus, let $\boldsymbol{k} = \frac{\pi}{L}\tilde{k}_j\boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the j'th Cartesian basis vector, and $\tilde{k}_j$ some integer.

Then we obtain

$$(\hat{\boldsymbol{u}}_{\boldsymbol{k}})_l = \begin{cases} -i\frac{(-1)^{\tilde{k}_j}}{k_j}, & l = j \quad \text{and} \quad \boldsymbol{k} \neq \boldsymbol{0} \\ 0, & l \neq j \quad \text{or} \quad \boldsymbol{k} = \boldsymbol{0}. \end{cases}$$

$$\hat{E}_{\boldsymbol{k}} = \begin{cases} 2\frac{(-1)^{\tilde{k}_j}}{k_j^2}, & \boldsymbol{k} \neq \boldsymbol{0}, \\ \frac{d}{3}L^2, & \boldsymbol{k} = \boldsymbol{0}. \end{cases}$$

Thus we see that even if $\mathcal{A}$ is relatively large, say a box of N^d degrees of freedom, the accurate evaluation of these functionals require only a subset of $\mathcal{A}$ containing the axes, which are of size dN.

It's also worth noting that $\mathcal{A}_{\mathrm{FF}}(N) \supseteq \mathcal{A}_T(N)$, yet for any functional ℓ that depends only on Fourier coefficients on the axes (such as ρ, $\boldsymbol{u}$ and E), we have

$$\ell\left(P_{\mathcal{A}_T(N)}f\right) = \ell\left(P_{\mathcal{A}_{\mathrm{FF}}(N)}f\right).$$

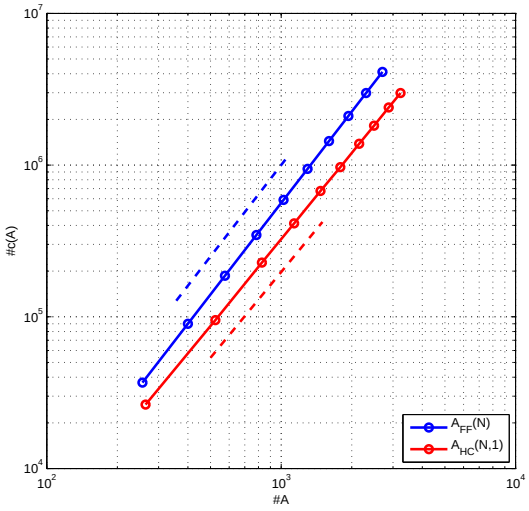


Figure 2: Logarithmic plot of $\sharp c(\mathcal{A})$ versus $\sharp\mathcal{A}$ for $\mathcal{A}_{\mathrm{FF}}(N)$ (blue) and $\mathcal{A}_0(N)$ (red), showing a minor improvement in complexity for the hyperbolic cross. The dashed lines represent $y = c_1 x^2$ and $y = c_2 x^{15/8}$.

The full grid offers no advantage over $\mathcal{A}_T$ in terms of such functionals.

3.2 Cost of evaluating $P_{\mathcal{A}} Q^R(f_{\mathcal{A}}, f_{\mathcal{A}})$

The evaluation of $Q(f_{\mathcal{A}}, f_{\mathcal{A}})$ requires the formation of the sum (13). There are no fast algorithms to compute this, unless some kind of separability of $\hat{\beta}$ is available, as in [16], which seems to be the case only for certain specific kernels B. In these cases the sum has a convolution structure.

A straightforward naive implementation has cost proportional to the cardinality of the combination set

$$c(\mathcal{A}) = \left\{(\boldsymbol{l}, \boldsymbol{m}) \in \mathcal{A}^2 \mid \boldsymbol{l} + \boldsymbol{m} \in \mathcal{A}\right\}.$$

The worst possible case for any $\mathcal{A}$ is $\sharp c(\mathcal{A}(N)) = O((\sharp\mathcal{A}(N))^2)$, and for the the full grid this bound is sharp. However, the hyperbolic cross can do better. For $T = 0$, we have experimentally that

$$\sharp c(\mathcal{A}_0(N)) = O(\sharp\mathcal{A}_0(N)^{15/8});,$$

see Figure 2.

4 Approximation Error

When necessary, we will assume hyperbolic cross approximations to give specific rates, however the theory is equally valid for full spaces.

We are primarily concerned with the dependence of the errors on the discretization parameters N and L. In line with Section 2.1 we will assume the ratio

$$\frac{R}{L} = \kappa$$

for some fixed $\kappa > 0$.

Throughout the section, we will take Assumption 1 for granted. We also make use of the notation

$$\lambda^{(+)} = \max(0, \lambda) \quad \text{for } \lambda \text{ from (3).}$$

Constants that do not depend on the truncation parameters L, R, the discretization parameters T, N, or f are suppressed. Thus, a statement such as $A \lesssim B$ means that there exists some quantity C that does not depend on L, R, N or f, such that $A \leq CB$.

4.1 Basic estimates

Our first result is a boundedness result for Q^R in L^2 for periodic functions.

Theorem 6. *Under Assumption 1 we have for functions f, g, which are $2L$-periodic with fundamental domain $\mathcal{D}_L = [-L, L]^d$, the estimate*

$$\|Q^R(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim L^{2\lambda^{(+)}+\frac{d}{2}} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)}, \qquad \forall f, g \in L^2(\mathcal{D}_L)\ .$$

Moreover, the function $Q^R(f,g)$ is L-periodic.

The proof of this theorem treats the gain- and loss term separately. In order to handle the gain term we need the following result from [20, Theorems 1, 2].

Theorem 7. *If $\lambda \geq 0$, with constants also depending on λ,*

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim \|f\|_{L^2_\lambda(\mathcal{D}_{\tilde{\kappa}L})} \|g\|_{L^1_\lambda(\mathcal{D}_{\tilde{\kappa}L})}, \qquad \forall f, g \in L^2_\lambda(\mathcal{D}_{\tilde{\kappa}L})\ ,$$

where $\tilde{\kappa} = \sqrt{2} + 2\kappa$, and we define for $\nu > 0$, $p \geq 1$, and a domain $D \subset \mathbb{R}^d$

$$\|f\|^p_{L^p_\nu(D)} := \int_D |f(\boldsymbol{v})|^p (1 + |\boldsymbol{v}|^{p\nu})\, \mathrm{d}\boldsymbol{v}\ .$$

For $-d < \lambda < 0$ and

$$\frac{1}{p} + \frac{1}{q} = 1 + \frac{\lambda}{d} + \frac{1}{r}$$

we have

$$\|Q^{R,+}(f,g)\|_{L^r(\mathcal{D}_L)} \lesssim \|f\|_{L^p(\mathcal{D}_{\tilde{\kappa}L})} \|g\|_{L^q(\mathcal{D}_{\tilde{\kappa}L})} \qquad \forall f \in L^p(\mathcal{D}_{\tilde{\kappa}L}), g \in L^q(\mathcal{D}_{\tilde{\kappa}L})$$

Proof. This result has been shown in [20, Theorems 1, 2] for the non-truncated gain operator Q^+ and with the norms for the terms $Q^{R,+}(f,g), f, g$ taken over all of $\mathbb{R}^d$ instead of bounded subsets.

As to the effect of the truncation we remark that exactly the same arguments as in [20] apply to the truncated case by replacing

$$B(\|\boldsymbol{g}\|, \cos\theta) \leftrightarrow B(\|\boldsymbol{g}\|, \cos\theta)\chi_{\mathcal{B}_{2R}},$$

with $\chi_{\mathcal{B}_{2R}}$ denoting the indicator function of $\mathcal{B}_{2R}$.

To justify the fact that in our estimates the norms on the right-hand sides are just taken over $[-\tilde{\kappa}L, \tilde{\kappa}L]^d$ we remark that by the definition of $Q^{R,+}$, the values of $Q^{R,+}(f,g)(\boldsymbol{v})$ for $\boldsymbol{v} \in [-L, L]^d$ depend only on f and g restricted to $[-\tilde{\kappa}L, \tilde{\kappa}L]^d$. $\square$

Proof of Theorem 6. We first show the desired statement for the loss term

$$Q^{R,-}(f,g)(\boldsymbol{v}) = \int_{\mathcal{B}_{2R}} \int_{S^{d-1}} B(\|\boldsymbol{u}\|, \cos\theta) g(\boldsymbol{v}-\boldsymbol{u}) \,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{u} \cdot f(\boldsymbol{v}) = (A_\lambda * g)(\boldsymbol{v}) \cdot f(\boldsymbol{v}) \,,$$

where, with A from (4),

$$A_\lambda(\boldsymbol{u}) := \|\boldsymbol{u}\|^\lambda A(\boldsymbol{u}) \,.$$

Since the integral runs over a bounded domain and, by Grad's cutoff assumption (4), A_λ is uniformly bounded if $\lambda > 0$, and we can assume

$$\|A_\lambda\|_{L^\infty(\mathcal{D}_L)} \lesssim L^{\lambda^{(+)}} \,,$$

which entails

$$\|Q^{R,-}(f,g)\|_{L^2(\mathcal{D}_L)} \le \|f\|_{L^2(\mathcal{D}_L)} \|A_\lambda * g\|_{L^\infty(\mathcal{D}_L)} \tag{17}$$

$$\le \|f\|_{L^2(\mathcal{D}_L)} \|A_\lambda\|_{L^\infty(\mathcal{D}_L)} \|g\|_{L^1(\mathcal{D}_{\tilde\kappa L})}. \tag{18}$$

The last inequality holds since only the values of g restricted to $[-\tilde\kappa L, \tilde\kappa L]^d$ are used for the evaluation of $A_\lambda * g(\boldsymbol{v})$, $\boldsymbol{v} \in [-L, L]^d$.

If $\lambda < 0$, A_λ is still integrable by the assumption $\lambda \ge -d/2$. Thus, A_λ has bounded Fourier transform, and the convolution operator is bounded in L^2. We can further estimate

$$\|g\|_{L^1(\mathcal{D}_{\tilde\kappa L})} \lesssim L^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_{\tilde\kappa L})} \,.$$

Now we observe that, due to periodicity of g, the previous quantity can be bounded by a constant times

$$L^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)}$$

Plugging this estimate into (18) yields the desired estimate for the loss term.

For the gain term we first focus on the case $\lambda \ge 0$, in which case we appeal to the first part of Theorem 7, which states that

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim \|f\|_{L^2_\lambda(\mathcal{D}_{\tilde\kappa L})} \|g\|_{L^1_\lambda(\mathcal{D}_{\tilde\kappa L})} \,. \tag{19}$$

Since f and g are periodic, we have

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim \|f\|_{L^2_\lambda(\mathcal{D}_L)} \|g\|_{L^1_\lambda(\mathcal{D}_L)} \,. \tag{20}$$

Since

$$\|f\|_{L^p_\lambda(\mathcal{D}_L)} \lesssim L^\lambda \|f\|_{L^p(\mathcal{D}_L)}$$

for all $p \ge 1$ and

$$\|g\|_{L^1(\mathcal{D}_L)} \lesssim L^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)} \,,$$

we arrive at

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim L^{\frac{d}{2}+2\lambda} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)}$$

whenever $\lambda \ge 0$.

Now we turn to the case $-d/2 \le \lambda < 0$. By our assumptions on d and λ, we can find $p, q \le 2$ such that with $r = 2$ we have

$$\frac{1}{p} + \frac{1}{q} = 1 + \frac{\lambda}{d} + \frac{1}{r}$$

By the second part of Theorem 7 and arguing as above, we get

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim L^{\frac{d}{2}} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)} \ ,$$

using the estimate

$$\|f\|_{L^p(\mathcal{D}_{\tilde{\kappa}L})} \lesssim L^{d\left(\frac{1}{p}-\frac{1}{2}\right)} \|f\|_{L^2(\mathcal{D}_L)}$$

and choosing

$$\frac{2}{p} = \frac{2}{q} = \frac{3}{2} + \frac{\lambda}{d} \leq \frac{3}{2}.$$

This yields the desired estimate.

To see that also $Q^R(f,g)$ is L-periodic, we simply write both f and g as a Fourier series which directly yields the Fourier series representation of Section 2.2 for $Q^R(f,g)$. This proves the theorem. □

The second result we will require is a product rule for derivatives of the collision operator which can be found in [22].

Proposition 8. *We have*

$$\partial_j Q^R(f,g) = Q^R(\partial_j f, g) + Q^R(f, \partial_j g),$$

where ∂_j denotes the derivative in the j-th coordinate direction.

Proof. This result has been proven in [22], but we give a simpler proof which applies to the case when f and g are $2L$-periodic, which is of interest to us. In this case we can write

$$\begin{aligned}\mathcal{F}(\partial_j Q^R(f,g))(\boldsymbol{k}) &= \sum_{\boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}} \hat{\beta}(\boldsymbol{l},\boldsymbol{m}) i k_j \hat{f}_{\boldsymbol{l}} \hat{g}_{\boldsymbol{m}} \\ &= \sum_{\boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}} \hat{\beta}(\boldsymbol{l},\boldsymbol{m}) i l_j \hat{f}_{\boldsymbol{l}} \hat{g}_{\boldsymbol{m}} + \sum_{\boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}} \hat{\beta}(\boldsymbol{l},\boldsymbol{m}) \hat{f}_{\boldsymbol{l}} i m_j \hat{g}_{\boldsymbol{m}}.\end{aligned}$$

The latter sum is equal to

$$\mathcal{F}(Q^R(\partial_j f, g))(\boldsymbol{k}) + \mathcal{F}(Q^R(f, \partial_j g))(\boldsymbol{k})$$

which proves the statement. □

4.2 Consistency

In the following we will develop estimates for the consistency error

$$\|Q^R(f,f) - P_{\mathcal{A}} Q^R(P_{\mathcal{A}} f, P_{\mathcal{A}} f)\|_{H^s(\mathcal{D}_L)},$$

where $P_{\mathcal{A}}$ denotes the projection operators onto the Fourier modes contained in $\mathcal{A}$.

If the support of f lies in a ball $\mathcal{B}_R$, by Proposition 5 such a bound yields error bounds for the H^s-norm of the error between the application of the discretized collision operator $P_{\mathcal{A}} Q^R(P_{\mathcal{A}} f, P_{\mathcal{A}} f)$ and the application of the true operator $Q(f,f)$ in terms of f. These results generalize the results of [18] to more general sets of Fourier modes.

We will examine this consistency error for a family of different Fourier discretizations. Only for simplicity we will assume $L = 1$ and, therefore, all function spaces

to follow are defined on $[-1,1]^d$. The case of general L is hardly more difficult, but it would require heavier notation.

The corresponding smoothness spaces are the following *mixed Sobolev spaces* as defined in [15].

Definition 9. We define the smoothness spaces

$$\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L) := \left\{ f \in L^2(\mathcal{D}_L) : \sum_{|\boldsymbol{\alpha}|_\infty \le t,\ |\boldsymbol{\beta}|_1 \le l} \left\| \partial^{\boldsymbol{\alpha}} \partial^{\boldsymbol{\beta}} f \right\|_{L^2(\mathcal{D}_L)} < \infty \right\}. \tag{21}$$

Remark 10. For $t = 0$ we get the usual Sobolev spaces, for $l = 0$ we get the Sobolev spaces with dominating mixed smoothness.

In [15] the following approximation result is shown.

Theorem 11. *We have*

$$\|f - \mathcal{P}_{\mathcal{A}_T(N)} f\|_{H^s(\mathcal{D}_L)} \lesssim (1+N)^\rho \|f\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)}$$

Where

$$\rho = \rho(s,l,t,T,d) = \begin{cases} s - l - t + (Tt - s + l)\frac{d-1}{d-T}, & T \ge \frac{s-l}{t}, \\ s - l - t, & T \le \frac{s-l}{t}. \end{cases} \tag{22}$$

In the remainder of the present section we establish the important fact that the previous optimal approximation order can be retained for the application of the truncated collision operator.

A crucial tool will be the following boundedness result for the collision operator.

Theorem 12. *Under the assumptions of Theorem 6 we have that*

$$\left\| Q^R(f,g) \right\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)} \lesssim \|f\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)} \|g\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)}. \tag{23}$$

Proof. Note that by Proposition 8, every derivative $\partial^{\boldsymbol{\alpha}} Q^R(f,g)$ can be expressed as a linear combination of terms

$$Q^R(\partial^{\boldsymbol{\alpha}_1} f, \partial^{\boldsymbol{\alpha}_2} g), \quad \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}.$$

It follows that we can estimate

$$\|\partial^{\boldsymbol{\alpha}} Q^R(f,g)\|_{L^2(\mathcal{D}_L)} \lesssim \sum_{\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}} \|Q^R(\partial^{\boldsymbol{\alpha}_1} f, \partial^{\boldsymbol{\alpha}_2} g)\|_{L^2(\mathcal{D}_L)}.$$

Now we can apply Theorem 6 to the summands in the above expression and arrive at the desired result. □

The following theorem is our main result concerning the approximation error in the Fourier discretization of the collision operator.

Theorem 13. *We have the estimate*

$$\left\| Q^R(f,f) - P_{\mathcal{A}_T(N)} Q^R(P_{\mathcal{A}_T(N)} f, P_{\mathcal{A}_T(N)} f) \right\|_{H^s(\mathcal{D}_L)} \lesssim (1+N)^\rho \left(1 + \|f\|^2_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)} \right),$$

where $\rho = \rho(s,l,t,T,d)$ *is given by* (22).

Proof. We write

$$\begin{aligned}\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{H^s(\mathcal{D}_L)} \\ \leq \left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(f,f)\right\|_{H^s(\mathcal{D}_L)} \\ + \left\|P_{\mathcal{A}_T(N)}Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{H^s(\mathcal{D}_L)}.\end{aligned}$$

Since the operator $P_{\mathcal{A}_T(N)}$ is a projection, this can be further bounded from above by

$$\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(f,f)\right\|_{H^s(\mathcal{D}_L)} + \left\|Q^R(f,f) - Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{H^s(\mathcal{D}_L)}.$$

To handle the first term we first invoke Theorem 11 to obtain

$$\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(f,f)\right\|_{H^s(\mathcal{D}_L)} \lesssim (1+N)^\rho \|Q^R(f,f)\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)}$$

Now, all we need to do is estimate the quantity $\|Q^R(f,f)\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)}$ in terms of the mixed Sobolev norm of f, which has been done in Theorem 12.

This takes care of the first term. In order to estimate the second term, given by

$$\left\|Q^R(f,f) - Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{H^s(\mathcal{D}_L)},$$

we invoke the bilinearity of Q^R which allows us to rewrite this expression as

$$\left\|Q^R\left(f - P_{\mathcal{A}_T(N)}f, f\right) + Q^R\left(P_{\mathcal{A}_T(N)}f, f - P_{\mathcal{A}_T(N)}f\right)\right\|_{H^s(\mathcal{D}_L)}.$$

Now we can invoke the bound of Theorem 12 with $t = 0$ and $l = s$ to bound this quantity by

$$\left\|f - P_{\mathcal{A}_T(N)}f\right\|_{H^s(\mathcal{D}_L)} \left(\|f\|_{H^s(\mathcal{D}_L)} + \left\|P_{\mathcal{A}_T(N)}f\right\|_{H^s(\mathcal{D}_L)}\right).$$

The first factor in this product can be estimated using Theorem 11, the second one is bounded by

$$2\|f\|_{H^s(\mathcal{D}_L)}.$$

Summing up these estimates we arrive at the desired result. □

Remark 14. The previous result paves the way for adaptively enlarging or shrinking the set of active Fourier modes in each timestep. To this end, we envision to solve the homogenous Boltzmann equation over three Fourier grids, corresponding to different values of T and decide to switch to a larger/smaller grid based on the relative errors between these three different solutions. We consider this approach to be especially promising in cases where the solution is well-approximable by a sparse (HC-type) grid initially. As the solution approaches the Maxwellian distribution, the approximation grid can be modified to yield a full Fourier grid more suitable for the approximation of radially symmetric functions. We leave the further exploration of this idea to future work.

4.3 Error for the projected equation

The aim of the remainder of Section 4 will be to establish estimates for the error $\|f_{\mathcal{A}} - f\|$, where f is the solution to the actual Boltzmann equation

$$\frac{\partial}{\partial t} f = Q(f, f), \qquad f(0) = f_0,$$

and $f_{\mathcal{A}}$ is the solution to the truncated and projected equation

$$\frac{\partial}{\partial t} f_{\mathcal{A}} = P_{\mathcal{A}} Q^R(f_{\mathcal{A}}, f_{\mathcal{A}}), \qquad f_{\mathcal{A}}(0) = P_{\mathcal{A}} f_0 \ .$$

We will also require the intermediate solution of the truncated, but not projected, equation

$$\dot{f}_R = Q^R(f_R, f_R), \qquad f(0) = f_0.$$

Due to the periodicity of $f_{\mathcal{A}}$, the error will necessarily have to be estimated on a bounded domain.

In Section 4.3.1 we will estimate the error $\|f_{\mathcal{A}} - f_R\|$ by relying on work by Filbet and Mouhot [7]. In Section 4.3.2 we will estimate the error $\|f_R - f\|$ due to truncation.

Throughout, we will make Assumption 1 and in Section 4.3.2 we make Assumption 2.

To ease notation we will also write $Q(f) := Q(f, f)$ when necessary, as well as $\mathcal{A} = \mathcal{A}_T(N)$.

4.3.1 Error due to discretization

First, we will aim to generalize the non-global results of [7], by which me mean everything up to, and including, [7, Proposition 4.6][1]. The program involves studying a perturbation from the *truncated* equation, of the form

$$\frac{\partial}{\partial t} f_{\mathcal{A}} = Q^R(f_{\mathcal{A}}) + r_{\mathcal{A}} \tag{24}$$

where the residual term $r_{\mathcal{A}}$ has the form

$$r_{\mathcal{A}} = (P_{\mathcal{A}} - \mathrm{Id}) Q^R(f_{\mathcal{A}}).$$

First, let us recall

Lemma 15. *Assume g and h are $2L$-periodic functions on $\mathbb{R}^d$, and that B is separable with a power law dependence on the relative velocity with exponent $\lambda \geq -\frac{d}{2}$, satisfying Grad's cutoff assumption* (4)*. Then, for all $p \in [1, \infty]$*

$$\|Q^R(g, h)\|_{L^p(\mathcal{D}_L)}, \ \|Q^R(h, g)\|_{L^p(\mathcal{D}_L)} \lesssim L^{2\lambda^{(+)}} \|g\|_{L^1(\mathcal{D}_L)} \|h\|_{L^p(\mathcal{D}_L)}.$$

Proof. This is a straightforward generalization of the proofs of Theorems 6 and 7. It is also the content of [7, Lemma 4.1]. □

[1] The remainder of [7] involves global estimates under an assumption of $\kappa \geq \sqrt{2}$. This assumptions appears unphysical, however, and causes heavy aliasing, see section 6.5

Furthermore we have the product rule

$$\partial^{\boldsymbol{\alpha}} Q^R(g,h) = Q^R(\partial^{\boldsymbol{\alpha}} g, h) + Q^R(g, \partial^{\boldsymbol{\alpha}} h), \tag{25}$$

for all $\boldsymbol{\alpha} \in \mathbb{N}^d$, $|\boldsymbol{\alpha}| = 1$. It is straightforward to obtain a similar Leibnitz-type product formula for general $\boldsymbol{\alpha} \in \mathbb{N}^d$.

Definition 16. Let $\boldsymbol{\alpha} \in \mathbb{N}^d$. Then define the smoothness spaces $H^{\boldsymbol{\alpha}}(\mathcal{D}_L)$, $H^{<\boldsymbol{\alpha}}(\mathcal{D}_L)$ with norms

$$\|g\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)} := \sum_{\boldsymbol{\beta} \leq \boldsymbol{\alpha}} \|\partial^{\boldsymbol{\beta}} g\|_{L^2(\mathcal{D}_L)}, \quad \|g\|_{H^{<\boldsymbol{\alpha}}(\mathcal{D}_L)} := \sum_{\boldsymbol{\beta} < \boldsymbol{\alpha}} \|\partial^{\boldsymbol{\beta}} g\|_{L^2(\mathcal{D}_L)} \,.$$

Remark 17. Note that we have

$$\|g\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)} = \sum_{|\boldsymbol{\alpha}_1|_\infty = t,\ |\boldsymbol{\alpha}_2|_1 = l} \|g\|_{H^{\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2}(\mathcal{D}_L)} \,,$$

so any smoothness result which applies to all norms $\|g\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)}$ also applies to the mixed smoothness Sobolev norms.

Note that for any $\mathcal{A}$ and $\boldsymbol{\alpha} \in \mathbb{N}^d$, it is obvious that

$$\|\partial^{\boldsymbol{\alpha}} P_{\mathcal{A}} f\|_{L^2(\mathcal{D}_L)} \leq \|\partial^{\boldsymbol{\alpha}} f\|_{L^2(\mathcal{D}_L)} \,. \tag{26}$$

Lemma 18. *Let $\boldsymbol{\alpha} \in \mathbb{N}^d$. Then we have*

$$\|Q^R(f)\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)},\ \|P_{\mathcal{A}} Q^R(f)\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)} \lesssim L^{2\lambda^{(+)} + \frac{d}{2}} \|f\|_{H^{<\boldsymbol{\alpha}}(\mathcal{D}_L)} \|f\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)}.$$

Proof. By (26) it suffices to show the estimate for Q^R. Hence we get

$$\begin{aligned}
\|\partial^{\boldsymbol{\alpha}} Q^R(f,f)\|_{L^2(\mathcal{D}_L)} &\lesssim L^{2\lambda^{(+)}} \sum_{0 \leq \boldsymbol{\beta} \leq \boldsymbol{\alpha}} \|\partial^{\boldsymbol{\beta}} f\|_{L^1(\mathcal{D}_L)} \|\partial^{\boldsymbol{\alpha} - \boldsymbol{\beta}} f\|_{L^2(\mathcal{D}_L)} \\
&\lesssim L^{2\lambda^{(+)} + \frac{d}{2}} \sum_{0 \leq \boldsymbol{\beta} \leq \boldsymbol{\alpha}} \|\partial^{\boldsymbol{\beta}} f\|_{L^2(\mathcal{D}_L)} \|\partial^{\boldsymbol{\alpha} - \boldsymbol{\beta}} f\|_{L^2(\mathcal{D}_L)} \\
&\lesssim L^{2\lambda^{(+)} + \frac{d}{2}} \sum_{0 \leq \boldsymbol{\beta} < \boldsymbol{\alpha}} \|\partial^{\boldsymbol{\beta}} f\|_{L^2(\mathcal{D}_L)} \|f\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)}.
\end{aligned}$$

We have used the product rule (25), the bound of Lemma 15 and the fact that the 1-norm can be bounded by the 2-norm on compact sets.

The implicit constant depends on the implicit constant from Lemma 15 and a combinatorial expression depending on $\boldsymbol{\alpha}$. □

These results allow us to state the following analogue of [7, Lemma 4.2]:

Lemma 19. *Consider the evolution problem* (24) *and assume that we have a uniform L^1-bound*

$$\sup_{t \in [0, T_{max}]} \|f_{\mathcal{A}}(t, \cdot)\|_{L^1(\mathcal{D}_L)} \leq M. \tag{27}$$

Then for any $\boldsymbol{\alpha} \in \mathbb{N}^d$, and $f_0 \in H^{\boldsymbol{\alpha}}(\mathcal{D}_L)$, there exists a quantity $C_{\boldsymbol{\alpha}}$, depending only on M, L, T and $\|f_0\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)}$ such that

$$\sup_{t \in [0, T_{max}]} \|f_{\mathcal{A}}(t, \cdot)\|_{H^{\boldsymbol{\alpha}}(\mathcal{D}_L)} \leq C_{\boldsymbol{\alpha}}(L). \tag{28}$$

Proof. For $0 \leq k \leq |\boldsymbol{\alpha}|$ we denote the sets

$$A_{\boldsymbol{\alpha}}^k := \left\{ \boldsymbol{\beta} \in \mathbb{N}^d : \ \boldsymbol{\beta} \leq \boldsymbol{\alpha} \text{ and } |\boldsymbol{\beta}| = k \right\}.$$

We will show inductively that for all such $k \in \mathbb{N}$, there exists a constant $C_k(L) < \infty$ such that

$$\sup_{t\in[0,T]} \sup_{\boldsymbol{\beta}\in A_{\boldsymbol{\alpha}}^k} \|f_{\mathcal{A}}(t,\cdot)\|_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)} \leq C_k(L). \tag{29}$$

Clearly, (29) with $k = |\boldsymbol{\alpha}|_\infty$ is simply the desired statement (28).

We start by showing (29) for $k = 0$ which amounts to deriving an L^2-bound for $f_{\mathcal{A}}$. To this end consider the ODE

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|f_{\mathcal{A}}\|^2_{L^2(\mathcal{D}_L)} \leq \|Q^R(f_{\mathcal{A}}) + r_{\mathcal{A}})\|_{L^2(\mathcal{D}_L)}\|f_{\mathcal{A}}\|_{L^2(\mathcal{D}_L)}$$

From Lemma 15 with $p = 2$ we deduce that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|f_{\mathcal{A}}\|^2_{L^2(\mathcal{D}_L)} \lesssim L^{2\lambda^{(+)}}\|f_{\mathcal{A}}\|_{L^1(\mathcal{D}_L)}\|f_{\mathcal{A}}\|^2_{L^2(\mathcal{D}_L)} \leq L^{2\lambda^{(+)}} M\|f_{\mathcal{A}}\|^2_{L^2(\mathcal{D}_L)}.$$

Grönwall's lemma yields (29) for $k = 0$.

Now the induction step. Assume we have (29) for $k-1$. Then we have for any $\boldsymbol{\beta} \in A_{\boldsymbol{\alpha}}^k$ that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|f_{\mathcal{A}}\|^2_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)} \leq \|Q^R(f) + r_{\mathcal{A}}\|_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)}\|f_{\mathcal{A}}\|_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)}.$$

Now, Lemma 18 yields that

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|f_{\mathcal{A}}\|^2_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)} &\lesssim L^{2\lambda^{(+)}+\frac{d}{2}}\|f_{\mathcal{A}}\|_{H^{<\boldsymbol{\beta}}(\mathcal{D}_L)}\|f_{\mathcal{A}}\|_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)}\|f_{\mathcal{A}}\|_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)} \\ &\leq L^{2\lambda^{(+)}+\frac{d}{2}} C_{k-1}(L)\|f_{\mathcal{A}}\|^2_{H^{\boldsymbol{\beta}}(\mathcal{D}_L)}.\end{aligned}$$

We have used the induction hypothesis in the last inequality. Using Grönwall's lemma again yields the desired claim. □

The following proposition establishes existence, uniqueness and control of L^1- and $H^{\boldsymbol{\alpha}}$-norms.

Proposition 20. *Consider the evolution problem* (24)*, make Assumption 1 and set* $W = \|P_{\mathcal{A}} f_0\|_{L^2(\mathcal{D}_L)}$.

Then there exists $\overline{\tau} = \overline{\tau}(W, L) > 0$ *so that* (24) *admits a unique solution on* $[0, \overline{\tau}]$*, and on this interval* $\|f_{\mathcal{A}}(t)\|_{L^1(\mathcal{D}_L)}$ *is bounded.*

Proof. We closely follow [7, Proposition 4.3]. Using Lemma 15 with $p = 1$, we find that for some $K > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|f_{\mathcal{A}}(t)\|_{L^2(\mathcal{D}_L)} \leq KL^{2\lambda^{(+)}+\frac{d}{2}}\|f_{\mathcal{A}}(t)\|^2_{L^2(\mathcal{D}_L)},$$

yielding

$$\|f_{\mathcal{A}}(t)\|_{L^2(\mathcal{D}_L)} \leq \frac{\|P_{\mathcal{A}} f_0\|_{L^2(\mathcal{D}_L)}}{1 - KL^{2\lambda^{(+)}+\frac{d}{2}}\|P_{\mathcal{A}} f_0\|_{L^2(\mathcal{D}_L)} t}.$$

Since $\mathcal{D}_L$ is bounded, this can be used to show that the L^1-norm is bounded on $[0, \overline{\tau}]$ provided the interval is small enough. The bound is explicit from the above inequality.

Then, existence and uniqueness follows from the theorem of Picard-Lindelöf, since Q^R is a bounded bilinear operator from $L^2(\mathcal{D}_L) \times L^2(\mathcal{D}_L)$ to $L^2(\mathcal{D}_L)$. □

Uniform control of the L^2-norm on $[0, T_{\max}]$ will, by repeatedly applying Proposition 20, yield existence and uniqueness on this interval. Then, Lemma 19 will provide smoothness, assuming smoothness of the initial condition.

Such control of the L^1-norm is clear for the classical (non-perturbed) Boltzmann equation due to preservation of positivity and mass. In [7], Lemmas 4.4 and 4.5 provide the following positivity result.

Lemma 21. *Consider the evolution perturbed problem* (24)*, make Assumption 1, and also assume that $f_0 \in L^\infty(\mathcal{D}_L)$ is non-negative, and that it satisfies*

$$\|f_0 - P_{\mathcal{A}} f_0\|_{L^\infty(\mathcal{D}_L)} \to 0$$

as $N \to \infty$.

Set $M = 2\|f_0\|_{L^1(\mathcal{D}_L)}$. Then there exists $\hat{\tau} \in (0, \overline{\tau})$ depending only on M, L and B, and there exists $\hat{N} < \infty$ depending only on $\hat{\tau}$ and f_0, such that for all $N > \hat{N}$, and for any smooth solution $f_{\mathcal{A}}$ to (24)*, we have*

$$\forall \boldsymbol{v} \in \mathcal{D}_L, \qquad f_{\mathcal{A}}(\hat{\tau}, \boldsymbol{v}) > 0 \, .$$

Furthermore, there exists $\eta(N)$, tending to 0 as $N \to \infty$, such that for all $t \in [0, \hat{\tau}]$, the negative part of $f_{\mathcal{A}}$ satisfies the bound

$$\|f_{\mathcal{A}}^-(t)\|_{L^\infty(\mathcal{D}_L)} \leq \eta(N) \, .$$

Proof. This is exactly [7, Lemma 4.5]. The proof is identical. □

The following theorem summarizes the argument, and corresponds to [7, Proposition 4.6]. It shows propagation of smoothness for the perturbed Boltzmann equation, and provides an estimate for the error $f_R - f_{\mathcal{A}}$.

Its proof employs the following nonlinear generalization of Grönwall's lemma used in the proof of the error estimate above.

Lemma 22 (Theorem 21 in [5])**.** *Let $u(t)$ be a nonnegative function satisfying*

$$u(t) \leq c + \int_0^t \left(a u(s) + b\sqrt{u(s)} \right) \mathrm{d}s$$

where a, b, c are nonnegative. Then

$$u(t) \leq \left[\sqrt{c} e^{at/2} + \frac{b}{a} \left(e^{at/2} - 1 \right) \right]^2 .$$

Theorem 23. *Consider the evolution problem* (24) *for a fixed time interval $[0, T_{max}]$. In addition, make Assumption 1, and assume that $f_0 \in L^\infty(\mathcal{D}_L) \cap \mathcal{H}^{t,l}_{mix}(\mathcal{D}_L)$ is non-negative, and satisfies*

$$\|f_0 - P_{\mathcal{A}} f_0\|_{L^\infty(\mathcal{D}_L)} \to 0 \quad \text{for } N \to \infty \, .$$

Then there exists $\hat{N} < \infty$ such that for all $N > \hat{N}$,

(i) there is a unique solution $f_{\mathcal{A}}(t,\cdot) \in \mathcal{H}^{t,l}_{mix}(\mathcal{D}_L)$ *on* $[0, T_{max}]$ *to the perturbed equation* (24) *with initial data* $P_{\mathcal{A}} f_0$*, uniformly bounded in* L^1*.*

(ii) There is $\eta(N) \to 0$ *as* $N \to \infty$ *such that for all* $t \in [0, T_{max}]$*,*

$$\|f^-_{\mathcal{A}}(t,\cdot)\|_{L^\infty(\mathcal{D}_L)} \leq \eta(N).$$

(iii) The solution satisfies the error bound

$$\|(f_R - f_{\mathcal{A}})(t)\|_{H^s(\mathcal{D}_L)} \lesssim \left(\frac{1 + C_{t+l}(L)^2}{L^{2\lambda^{(+)}+\frac{d}{2}} C_s(L)} + \|f_0\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}} \right) (1+N)^\rho \exp\left[\gamma L^{2\lambda^{(+)}+\frac{d}{2}} C_s(L) t \right]$$

for some $\gamma > 0$ *that does not depend on* L*,* s*,* t*,* l *or* N*, and where* $\rho = \rho(s,l,t,T,d)$ *is given in* (22)*.*

Proof. Set $M = 2\|f_0\|_{L^1(\mathcal{D}_L)}$. By Proposition 20, there exists a unique smooth solution to the perturbed problem on some interval $[0, \overline{\tau}]$. By Lemma 21, there exists $\hat{\tau} \leq \overline{\tau}$ and $\hat{N}$, such that for all $N > \hat{N}$, $f_{\mathcal{A}}(\hat{\tau}, \cdot)$ is positive.

By preservation of mass (see e.g., [7, Equation (3.4)]) and positivity, we thus have

$$\|f_{\mathcal{A}}(\hat{\tau})\|_{L^1(\mathcal{D}_L)} = \|f_0\|_{L^1(\mathcal{D}_L)} = \frac{M}{2}.$$

By Lemma 19, $f_{\mathcal{A}}(\hat{\tau})$ also retains the smoothness, and by Lemma 21, it is positive. Thus we can use induction to prove the existence for all of $[0, T_{\max}]$. The L^1-bound is M, following from Proposition 20.

The error bound can be deduced from the observation

$$\frac{\partial}{\partial t}(f_R - f_{\mathcal{A}}) = Q^R_{\mathrm{sym}}(f_R - f_{\mathcal{A}}, f_R + f_{\mathcal{A}}) + r_{\mathcal{A}}$$

where Q^R_{sym} is the symmetrized collision operator. From Theorem 12 we have

$$\begin{aligned} \|Q^R_{\mathrm{sym}}(f_R - f_{\mathcal{A}}, f_R + f_{\mathcal{A}})\|_{H^s(\mathcal{D}_L)} &\lesssim L^{2\lambda^{(+)}+\frac{d}{2}} \|f_R + f_{\mathcal{A}}\|_{H^s(\mathcal{D}_L)} \|f_R - f_{\mathcal{A}}\|_{H^s(\mathcal{D}_L)} \\ &\lesssim L^{2\lambda^{(+)}+\frac{d}{2}} C_s(L) \|f_R - f_{\mathcal{A}}\|_{H^s(\mathcal{D}_L)}. \end{aligned}$$

The bound for $r_{\mathcal{A}} = (\mathrm{Id} - P_{\mathcal{A}_T(N)}) Q^R(f_{\mathcal{A}})$ comes from Theorem 13:

$$\begin{aligned} \|r_{\mathcal{A}}\|_{H^s(\mathcal{D}_L)} &\lesssim (1+N)^\rho \left(1 + \|f_{\mathcal{A}}\|^2_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)} \right) \\ &= (1+N)^\rho \left(1 + C_{t+l}(L)^2 \right) . \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\partial}{\partial t} \|f_R - f_{\mathcal{A}}\|^2_{H^s(\mathcal{D}_L)} \lesssim L^{2\lambda^{(+)}+\frac{d}{2}} C_s(L) \|f_R - f_{\mathcal{A}}\|^2_{H^s(\mathcal{D}_L)} + \\ (1+N)^\rho \left(1 + C_{t+l}(L)^2 \right) \|f_R - f_{\mathcal{A}}\|_{H^s(\mathcal{D}_L)} . \end{aligned}$$

Integrating this inequality, and applying Lemma 22, as well as the initial approximation of f_0 (using Theorem 11), we arrive at the desired result. □

Note that all mixed smoothness Sobolev spaces can be written as intersections of spaces of the type $H^{\boldsymbol{\alpha}}$ from Definition 16, which connects the estimates made in the proof of Theorem 23 to those from Lemmas 18 and 19.

4.3.2 Error due to truncation

We will now turn our attention to the error induced by truncating the collision operator, namely $e_R(t) = f(t) - f_R(t)$. In the following, we will assume that κ is sufficiently small so that the evaluation of $Q^R(f)$ in the ball $\mathcal{B}_R$ requires only values within $\mathcal{D}_L$. This condition is slightly stricter than that which was given earlier.

Throughout this section we will make Assumption 2 concerning the decay of f and f_R.

Definition 24. Given a Sobolev index $s \geq 0$, we say that a function $g : \mathbb{R}^d \to \mathbb{R}$ has a Gauss-like decay, if there exists $C > 0$ and $a > 0$ such that

$$|\partial^{\boldsymbol{\alpha}} g(\boldsymbol{v})| \leq C \exp\left(-a\|\boldsymbol{v}\|^2\right)$$

for all $|\boldsymbol{\alpha}|_1 \leq s$.

Lemma 25. *Fix $s > 0$ and assume that g, h satisfy Gauss-like decay in the sense of Definition 24. Then, for sufficiently large L ($2L \geq 1$)*

$$\|Q^R(g,g) - Q^R(h,h)\|_{H^s(\mathcal{B}_R)} \lesssim L^{2\lambda^{(+)}+\frac{d}{2}} \left(\|g-h\|_{H^s(\mathcal{B}_R)} + L^{\frac{d}{2}} e^{-aR^2}\right) .$$

Proof. From the proof of Theorem 6 we know that

$$\begin{aligned}\|Q^{R,-}(g,h)\|_{L^2(\mathcal{B}_R)} &\lesssim R^{\lambda^{(+)}} \|g\|_{L^2(\mathcal{B}_R)} \|h\|_{L^1(\mathcal{D}_L)} \\ &\lesssim R^{\lambda^{(+)}} (2L)^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)} \|h\|_{L^2(\mathcal{D}_L)}.\end{aligned}$$

Moreover, by Theorem 7 and the proof of Theorem 6 we have that if $\lambda \geq 0$,

$$\begin{aligned}\|Q^{R,+}(g,h)\|_{L^2(\mathcal{B}_R)} &\lesssim \|g\|_{L^2_\lambda(\mathcal{D}_L)} \|h\|_{L^1_\lambda(\mathcal{D}_L)} \\ &\lesssim Lk2\lambda \|g\|_{L^2(\mathcal{D}_L)} \|h\|_{L^1(\mathcal{D}_L)} \\ &\lesssim L^{2\lambda+\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)} \|h\|_{L^2(\mathcal{D}_L)}\end{aligned}$$

and if $-\frac{d}{2} \leq \lambda < 0$, just as in the proof for Theorem 6, we find, with

$$\frac{1}{p} = \frac{3}{4} + \frac{\lambda}{2d} \leq \frac{3}{4}$$

that

$$\begin{aligned}\|Q^{R,+}(g,h)\|_{L^2(\mathcal{B}_R)} &\lesssim \|g\|_{L^p(\mathcal{D}_L)} \|h\|_{L^p(\mathcal{D}_L)} \\ &\lesssim L^{d\left(\frac{2}{p}-1\right)} \|g\|_{L^2(\mathcal{D}_L)} \|h\|_{L^2(\mathcal{D}_L)} \\ &\lesssim L^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)} \|h\|_{L^2(\mathcal{D}_L)}\end{aligned}$$

whenever $2L \geq 1$.

In summary, under the given assumptions,

$$\|Q^R(g,h)\|_{L^2(\mathcal{B}_R)} \lesssim L^{2\lambda^{(+)}+\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)} \|h\|_{L^2(\mathcal{D}_L)}.$$

Also, the same bound will hold for the symmetrized operator Q^R_{sym}.

Now, given two functions g and h satisfying the given assumptions, we have

$$\begin{aligned}\|Q^R(g,g)-Q^R(h,h)\|_{H^s(\mathcal{B}_R)} &= \|Q^R_{\mathrm{sym}}(g,g)-Q^R_{\mathrm{sym}}(h,h)\|_{H^s(\mathcal{B}_R)}\\ &= \|Q^R_{\mathrm{sym}}(g+h,g-h)\|_{H^s(\mathcal{B}_R)}\\ &\lesssim \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|Q^R_{\mathrm{sym}}\left(\partial^{\boldsymbol{\alpha}}(g+h),\partial^{\boldsymbol{\alpha}-\boldsymbol{\beta}}(g-h)\right)\|_{L^2(\mathcal{B}_R)}\\ &\lesssim L^{2\lambda^{(+)}+\frac{d}{2}}\|g+h\|_{H^s(\mathcal{D}_L)}\|g-h\|_{H^s(\mathcal{D}_L)}.\end{aligned}$$

Since g and h are bounded, we can reduce this to

$$\begin{aligned}\|Q^R(g,g)-Q^R(h,h)\|_{H^s(\mathcal{B}_R)} \lesssim\\ L^{2\lambda^{(+)}+\frac{d}{2}}\left(\|g-h\|_{H^s(\mathcal{B}_R)}+\|g\|_{H^s(\mathcal{D}_L\setminus\mathcal{B}_R)}+\|h\|_{H^s(\mathcal{D}_L\setminus\mathcal{B}_R)}\right)\end{aligned}$$

From the exponential decay of g and h we can conclude

$$\|g\|_{H^s(\mathcal{D}_L\setminus\mathcal{B}_R)},\|h\|_{H^s(\mathcal{D}_L\setminus\mathcal{B}_R)} \lesssim e^{-aR^2}L^{\frac{d}{2}}.$$

The result follows. □

The following lemma is purely geometric in nature and is need for the proof of the next Proposition 28.

Lemma 26. *For $\boldsymbol{v},\boldsymbol{w}\in\mathbb{R}^d$, it holds that*

$$\|\boldsymbol{v}+\boldsymbol{w}\|^2+\|\boldsymbol{v}\|^2 \geq (2-\phi)\left(\|\boldsymbol{v}\|^2+\|\boldsymbol{w}\|^2\right),$$

where ϕ is the golden ratio.

Proof. First, we have

$$\begin{aligned}-2(\boldsymbol{v},\boldsymbol{w}) &\leq 2\|\sqrt{\phi}\boldsymbol{v}\|\|\frac{1}{\sqrt{\phi}}\boldsymbol{w}\|\\ &\leq \phi\|\boldsymbol{v}\|^2+\frac{1}{\phi}\|\boldsymbol{w}\|^2.\end{aligned}$$

Thus

$$2(\boldsymbol{v},\boldsymbol{w})+\phi\|\boldsymbol{v}\|^2+\frac{1}{\phi}\|\boldsymbol{w}\|^2\geq 0,$$

and

$$\begin{aligned}\|\boldsymbol{v}\|^2+\|\boldsymbol{v}+\boldsymbol{w}\|^2 &= 2\|\boldsymbol{v}\|^2+\|\boldsymbol{w}\|^2+2(\boldsymbol{v},\boldsymbol{w})\\ &\geq (2-\phi)\|\boldsymbol{v}\|^2+\left(1-\frac{1}{\phi}\right)\|\boldsymbol{w}\|^2.\end{aligned}$$

The result follows since $\phi^{-1}=\phi-1$. □

Proposition 27. *Fix $s>0$ and make Assumption 2 about the solutions f and f_R. Then, on a bounded time interval $[0,T_{max}]$, and for sufficiently large L, we have the following bound for the error $e_R=f-f_R$ due to truncating the collision operator:*

$$\|e_R(t)\|_{H^s(\mathcal{B}_R)} \lesssim \left(\sqrt{\mu_R}+L^{\frac{d}{2}}e^{-aR^2}\right)\exp\left(\delta t L^{2\lambda^{(+)}+\frac{d}{2}}\right) \tag{30}$$

with

$$\mu_R = \int_0^{T_{max}} \|(Q - Q^R)(f(\tau))\|_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau,$$

and where δ is a constant that does not depend on L.

Proof. We have

$$\begin{aligned}
\frac{1}{2}\frac{\partial}{\partial t}\|e_R(t)\|^2_{H^s(\mathcal{B}_R)} &= \int_0^t \langle Q(f(\tau)) - Q^R(f_R(\tau)), e_R(\tau)\rangle_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau \\
&= \int_0^t \langle (Q - Q^R)(f(\tau)), e_R(\tau)\rangle_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau \\
&\quad + \int_0^t \langle Q^R(f(\tau)) - Q^R(f_R(\tau)), e_R(\tau)\rangle_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau.
\end{aligned}$$

noting that $e_R(0) = 0$.

Now note that

$$\int_0^t \langle (Q - Q^R)(f(\tau)), e_R(\tau)\rangle_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau \lesssim \int_0^{T_{\max}} \|(Q - Q^R)(f(\tau))\|_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau,$$

where the implicit constant is

$$\sup_t \|f\|_{H^s(\mathcal{B}_R)} + \sup_t \|f_R\|_{H^s(B_R)},$$

which exists by assumption. Thus define

$$\mu_R = \int_0^{T_{\max}} \|(Q - Q^R)(f(\tau))\|_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau.$$

Then,

$$\begin{aligned}
\|e_R(t)\|^2_{H^s(\mathcal{B}_R)} &\lesssim \mu_R + \int_0^t \langle Q^R(f(\tau)) - Q^R(f_R(\tau)), e_R(\tau)\rangle_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau \\
&\leq \mu_R + L^{2\lambda^{(+)} + \frac{d}{2}} \int_0^t \left(\|e_R(\tau)\|_{H^s(\mathcal{B}_R)} + L^{\frac{d}{2}} e^{-aR^2} \right) \|e_R(\tau)\|_{H^s(\mathcal{B}_R)} \, \mathrm{d}\tau
\end{aligned}$$

by Lemma 25.

As in Theorem 23, we use Lemma 22 to complete the proof. □

The quantity μ_R will decrease exponentially as $e^{-(2-\phi)aR^2}$, where ϕ is the golden ratio. This is shown in the following proposition.

Proposition 28. *Fix $s > 0$ and make Assumption 2 on f. Then*

$$\|(Q - Q^R)(f)\|_{H^s(\Omega)} \lesssim e^{-2a(2-\phi)R^2}$$

where ϕ is the golden ratio.

The implicit constant may depend on the domain $\Omega \subseteq \mathbb{R}^d$, the decay rates of f, as well as λ and d.

Proof. First, note that, by Lemma 26, we have

$$\begin{aligned}\|\boldsymbol{v}+\boldsymbol{x}+\boldsymbol{y}\|^2+\|\boldsymbol{v}\|^2 &\geq (2-\phi)(\|\boldsymbol{v}\|^2+\|\boldsymbol{x}+\boldsymbol{y}\|^2)\\ \|\boldsymbol{v}+\boldsymbol{x}\|^2+\|\boldsymbol{v}+\boldsymbol{y}\|^2 &\geq (2-\phi)(\|\boldsymbol{v}+\boldsymbol{x}\|^2+\|\boldsymbol{x}+\boldsymbol{y}\|^2)\\ &\geq (2-\phi)^2\|\boldsymbol{v}\|^2+(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2.\end{aligned}$$

Thus,

$$\begin{aligned}|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{v}+\boldsymbol{x})\partial^{\boldsymbol{\beta}} f(\boldsymbol{v}+\boldsymbol{y})| &\lesssim \exp(-a\|\boldsymbol{v}+\boldsymbol{x}\|^2-a\|\boldsymbol{v}+\boldsymbol{y}\|^2)\\ &\lesssim \exp\left(-a(2-\phi)^2\|\boldsymbol{v}\|^2-a(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2\right)\end{aligned}$$

and

$$\begin{aligned}|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{v}+\boldsymbol{x}+\boldsymbol{y})\partial^{\boldsymbol{\beta}} f(\boldsymbol{v})| &\lesssim \exp(-a\|\boldsymbol{v}+\boldsymbol{x}+\boldsymbol{y}\|^2-a\|\boldsymbol{v}\|^2)\\ &\lesssim \exp\left(-a(2-\phi)^2\|\boldsymbol{v}\|^2-a(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2\right),\end{aligned}$$

where we added another power of $(2-\phi)$ in the last step to ease notation.

Then, using the Carleman representation for Q^R, and the summation formula for derivatives, we get

$$|\partial^{\alpha}(Q-Q^R)(f,f)(\boldsymbol{v})| \lesssim e^{-a(2-\phi)^2\|\boldsymbol{v}\|^2}\int_S \|\boldsymbol{x}+\boldsymbol{y}\|^{\lambda+2-d}\delta(\boldsymbol{x}\cdot\boldsymbol{y})e^{-a(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y},$$

where the power $\lambda+2-d$ comes from the transformed collision kernel $\tilde{B}$, and the integral runs over the set

$$S=\left(\mathcal{B}^2_{\sqrt{2}R}\right)^c,$$

i.e. the complement of the squared $\sqrt{2}R$-ball.

The proof is completed by writing

$$e^{-a(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2} \leq e^{-2a(2-\phi)R^2}e^{-\frac{a}{2}(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2},$$

since $\|\boldsymbol{x}+\boldsymbol{y}\|^2=\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2\geq 4R^2$. The remaining integrals over $\boldsymbol{v}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ clearly converge, and they are bounded from above as $L,R\to\infty$. □

4.3.3 Summary

The following theorem summarizes the results from the previous two sections.

Theorem 29. *Fix $s>0$, and make Assumptions 1 and 2. Then, for the error induced by truncation and projection, for sufficiently large L we have the estimate*

$$\begin{aligned}\|(f-f_{\mathcal{A}})(t)\|_{H^s(\mathcal{D}_L)} \lesssim \left(\frac{1+C_{t+l}(L)^2}{L^{2\lambda^{(+)}+\frac{d}{2}}C_s(L)}+\|f_0\|_{\mathcal{H}^{t,l}_{\mathrm{mix}}}\right)(1+N)^{\rho}\exp\left[\gamma L^{2\lambda^{(+)}+\frac{d}{2}}C_s(L)t\right]\\ +\exp\left[\delta L^{2\lambda^{(+)}+\frac{d}{2}}t-a(2-\phi)R^2\right],\end{aligned}$$

for some $\gamma,\delta>0$ that do not depend on L, s, t, l or N, and where $\rho=\rho(s,l,t,T,d)$ is given in (22).

Proof. This is a combination of the results from Propositions 28, 27 and Theorem 23. □

As $\rho < 0$, we see that the first term can be controlled, for a given L, by an appropriate choice $N = N(L)$.

The second term depends only on L, and it can be controlled only under the condition

$$\lambda^{(+)} \leq 1 - \frac{d}{4} \,. \tag{31}$$

Remark 30. The condition (31) is certainly restrictive. It allows Maxwellian and soft potentials in all dimensions. For hard potentials, it requires $\lambda \leq 1/2$ in two dimensions and $\lambda \leq 1/4$ in three dimensions. Specifically, hard spheres are always excluded.

A decay of the solution of type $e^{-a\|\boldsymbol{v}\|^\omega}$ with $\omega > 2$ would imposed a weaker constraint on λ, but this would rule out equilibrium solutions, which decay with rate $\omega = 2$.

It should be noted that one should not expect the estimate of Theorem 29 to be sharp for large t. Indeed, the exact and numerical solutions will both tend towards equilibria, as can be seen in Section 6.

5 An offset method

On the one hand the hyperbolic cross $\mathcal{A}_0(N)$ offers a great benefit in terms of efficiency as compared to the full Fourier grid $\mathcal{A}_{\mathrm{FF}}(N)$. However, as $\mathcal{A}_0(N)$ is not very well adapted to isotropic near-equilibrium solutions, a more flexible idea for the near-equilibrium regime may be to consider f as a perturbation from equilibrium

$$f(\boldsymbol{v}) = f^p(\boldsymbol{v}) + \mathcal{M}(\rho, \boldsymbol{u}, T)(\boldsymbol{v}) \,. \tag{32}$$

The spectrum of $\mathcal{M}$ is known *a priori* and f^p can be approximated using any coefficient set $\mathcal{A}$. Since we would keep $\mathcal{M}$ constant in time, we then get

$$\partial_t f = \partial_t f^p = Q(f,f) = Q(f^p, f^p) + Q(\mathcal{M}, f^p) + Q(f^p, \mathcal{M}) \tag{33}$$

as $Q(\mathcal{M}, \mathcal{M}) = 0$ by necessity. Moreover, the terms $Q(\mathcal{M}, f^p)$ and $Q(f^p, \mathcal{M})$ represent a linear function of f^p which can be assembled *a priori* using the spectrum of $\mathcal{M}$ to any desired accuracy. The only quadratic part is $Q(f^p, f^p)$. As we would expect $f_p \to 0$, the collision operator becomes near linear over time.

The spectrum of the periodically continued Maxwellian in terms of the k from (10) is

$$\hat{\mathcal{M}}(\rho, \boldsymbol{u}, T)(\boldsymbol{k}) = \frac{\rho}{(2L)^d} \exp\left(-\frac{T}{2}\|\boldsymbol{k}\|^2 - i\boldsymbol{u}\cdot\boldsymbol{k}\right). \tag{34}$$

Knowing this, we can formulate the linear part of the right hand side of (33) as

$$\begin{aligned}
\left[Q(f^p, \mathcal{M}) + Q(\mathcal{M}, f^p)\right](\boldsymbol{k}) &= \sum_{\substack{\boldsymbol{l}\in\mathcal{A}\\ \boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \hat{\beta}(\boldsymbol{l},\boldsymbol{m})\hat{\mathcal{M}}(\boldsymbol{m})f_{\boldsymbol{l}}^p + \sum_{\substack{\boldsymbol{l}\in\mathcal{A}\\ \boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \hat{\beta}(\boldsymbol{m},\boldsymbol{l})\hat{\mathcal{M}}(\boldsymbol{m})f_{\boldsymbol{l}}^p \\
&= \sum_{\boldsymbol{l}\in\mathcal{A}} \sum_{\substack{\boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \left(\hat{\beta}(\boldsymbol{l},\boldsymbol{m}) + \hat{\beta}(\boldsymbol{m},\boldsymbol{l})\right)\hat{\mathcal{M}}(\boldsymbol{m})f_{\boldsymbol{l}}^p \\
&= \sum_{\boldsymbol{l}\in\mathcal{A}} Q_{\mathrm{lin}}(\boldsymbol{k},\boldsymbol{l})f_{\boldsymbol{l}}^p \qquad (35)
\end{aligned}$$

which can be recognized as simple matrix multiplication with

$$Q_{\mathrm{lin}}(\boldsymbol{k},\boldsymbol{l}) := \sum_{\substack{\boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \left(\hat{\beta}(\boldsymbol{l},\boldsymbol{m}) + \hat{\beta}(\boldsymbol{m},\boldsymbol{l})\right)\hat{\mathcal{M}}(\boldsymbol{m})$$

defined for all $\boldsymbol{k}, \boldsymbol{l} \in \mathcal{A}$.

The set $\mathcal{B}$ can be any suitable set of Fourier coefficients for approximating the Maxwellian. As can be seen from (34), $|\hat{\mathcal{M}}|$ is a Gaussian centered at zero, so a suitable choice might be $\mathcal{B} = \mathcal{A}_{\mathrm{FF}}(N)$ with a sufficiently large N. It is worth noting that $\mathcal{B}$ can be very large, since as soon as Q_{lin} is assembled, the size of $\mathcal{B}$ does not affect the cost of applying (35).

To ensure

$$\sup_{\boldsymbol{m}\notin\mathcal{B}} |\hat{\mathcal{M}}(\boldsymbol{m})| \leq \epsilon$$

the condition on N is

$$N^2 \geq \frac{8L^2}{T\pi^2} \log\left(\frac{\rho}{(2L)^d\epsilon}\right) .$$

6 Numerical results

We study the accuracy of Fourier spectral discretizations numerically for various situations, some of which are "friendly" to the hyperbolic cross, whereas others are not. To summarize, the three methods we have used are

- FF: The full grid Fourier approximation, see Section 3.
- HC: The "raw" hyperbolic cross method, see Section 3.
- OM: The offset method with hyperbolic cross, see Section 5.

In Section 6.1 we apply the FF and HC methods to the BKW solution to verify correctness, and to show that the hyperbolic cross does indeed to a poor job of resolving Maxwellians. We also show numerical evidence of the monotonic increase of entropy and its convergence to the theoretical maximum, given by the entropy of the equilibrium distribution.

In Section 6.2 we apply the method to a highly anisotropic solution where the hyperbolic cross performs much better.

In Section 6.3 we give numerical evidence that the relaxation to equilibrium is indeed exponential in time, as demonstrated by Gressman and Strain [13].

In Section 6.4 we provide some numerical evidence for the exponential decay of f_R stated in Assumption 2, which allowed us to prove the error estimates in Section 4.3.2.

Finally, in Section 6.5 we investigate the effect the choice of the ratio $\kappa = \frac{R}{L}$ can have on the numerical solution.

In each case, we worked with $d = 2$, and $\lambda = 0$ (which is the Maxwellian collision kernel). All timestepping was done using an explicit 4^{th} order Runge-Kutta method. The kernel modes were computed with sufficiently high order Gaussian quadrature.

6.1 Verification (BKW)

As a verification of correctness, one can use the only known analytical non-equilibrium solution to the Boltzmann equation—the rotationally symmetric BKW solution [21], [6]. It takes the form

$$f(t, v) = (2\pi s)^{-d/2} \exp\left(-\frac{\|\boldsymbol{v}\|^2}{2s}\right)\left(1 - \frac{1-s}{2s}\left(d - \frac{\|\boldsymbol{v}\|^2}{s}\right)\right) \tag{36}$$

where

$$s = s(t) = 1 - e^{-\lambda(t+t_0)},$$

and $B = \text{const.}$ (also called the *Maxwellian* kernel). Here, λ is a parameter given in terms of B, and for $d = 2$ we find $B = \frac{1}{2\pi}$ and $\lambda = \frac{1}{8}$. Finally, t_0 is any reasonable starting time so that $f(t, \boldsymbol{v}) \geq 0$ everywhere. We will use a t_0 determined by $s(0) = \frac{1}{2}$, which gives the initial distribution

$$f_0(\boldsymbol{v}) = \frac{1}{\pi}\|\boldsymbol{v}\|^2 e^{-\|\boldsymbol{v}\|^2}.$$

This can be scaled to ensure that it meets the conditions of Proposition 4 to a sufficient degree. We use $L = 3\pi$ and a scaling in $\boldsymbol{v}$ with factor 5 (i.e. $f_0(5\boldsymbol{v})$). This is a relatively large value, which eliminates aliasing to machine precision level, but which also makes the Fourier series approximation quite poor. Cold gases have narrow support in $\boldsymbol{v}$-space and wide support in $\boldsymbol{k}$-space.

Figure 3 shows the results for this experiment. Note the poor approximation properties due to the very narrow support of f_0, as well as the poor performance of the hyperbolic cross, due to the rotational symmetry. Note also how inaccurate initial data can still yield accurate long-term solutions. The solutions themselves are shown in Figure 4. Finally, Figure 5 shows, for $N = 56$, how the entropy of the solution,

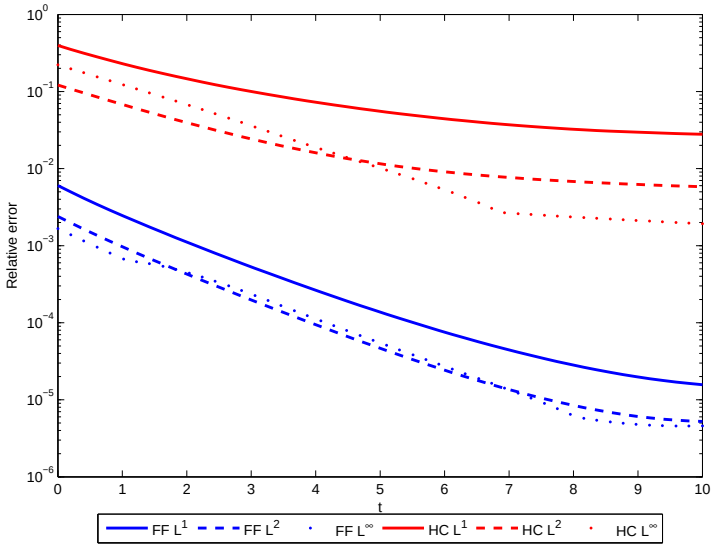
$$-\int_{\mathcal{D}_L} f(\boldsymbol{v}) \log f(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}$$

converges to the theoretical maximum as given by the equilibrium distribution.

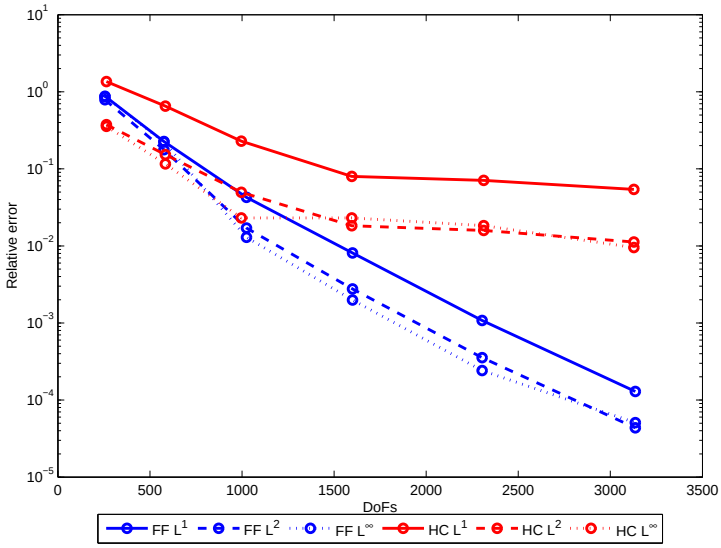
6.2 Crossed beams

This is an example of a case where the hyperbolic cross approximation works very well. The initial condition is

$$f_0(\boldsymbol{v}) = \left[(1 + \sin(sv_x))e^{-sv_x^2} + (1 + \sin(sv_y))e^{-sv_y^2}\right] e^{-2\|\boldsymbol{v}\|^2}.$$

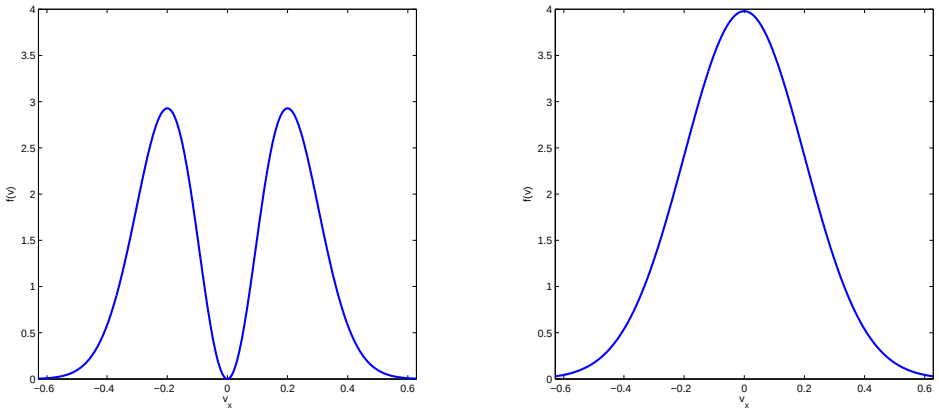


(a) Relative error w.r.t. time for about 3100 degrees of freedom.



(b) Relative error w.r.t. degrees of freedom at time $t = 5.1$.

Figure 3: Relative errors for the BKW solution for the full grid and hyperbolic cross methods, in L^1-, L^2- and L^∞-norms. For timestepping, we used a fixed-timestep explicit 4^{th} order Runge-Kutta method with timestep 10^{-2}.



(a) Initial distribution. (b) Equilibrium.

Figure 4: Cross-section plots of the BKW solution. Rotational symmetry applies.

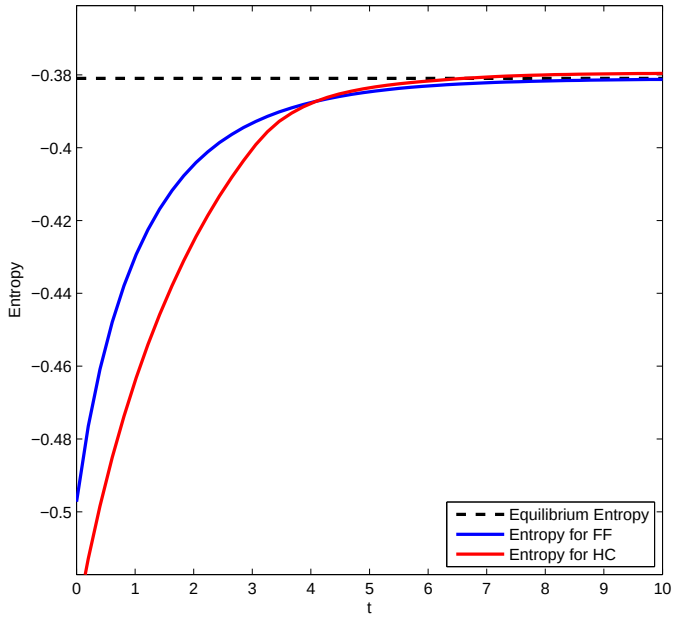


Figure 5: Convergence to entropy for $\mathcal{A}_{\mathrm{FF}}(56)$ and $\mathcal{A}_0(234)$.

There parameter $s > 0$ can be tweaked to make f_0 more or less isotropic. For our experiment we have used $s = 10$, and a hyperbolic cross with standard fatness $D = 1$. We computed the solution over four time units, with an explicit 4^{th} order Runge-Kutta method with timestep $5 \cdot 10^{-3}$.

Since this is beyond the scope of known analytic solutions, we used a reference solution computed on a full grid with $N = 80$. The results for global L^2-errors are shown in Figure 6, and for observables in Figure 7.

In these cases, the hyperbolic cross can be seen to outperform the full grid method, although the latter catches up over time and eventually wins out as the solution approaches equilibrium, see Figure 6(d).

6.3 Relaxation to equilibrium

This is an example of the offset method, using the initial distribution

$$f_0(\boldsymbol{v}) = e^{-2\|\boldsymbol{v}\|^2} + \epsilon \left(e^{-sv_x^2 - v_y^2} + e^{-v_x^2 - sv_y^2} \right),$$

where again the parameter $s > 1$ controls the anisotropy of f_0, and $\epsilon > 0$ represents the fact that f_0 is a minor perturbation from equilibrium.[2]

Using $s = 7$ and $\epsilon = 10^{-2}$, we have plotted $\|f^p\|$ from (32) versus time for three different norms in Figure 9. The convergence halts at $\|f^p\| \approx 10^{-5}$ due to truncation error in v for all norms, but prior to this, exhibits behavior in accordance with [13].

6.4 Exponential decay of the numerical solution

We will here provide some numerical evidence that a certain exponential decay can be expected to hold for the numerical solution using both $\mathcal{A}_{\text{FF}}(N)$ and $\mathcal{A}_0(N)$ with some unspecified dependence $N = N(L)$. The decay we are after is the existence of some C independent of L so that

$$\|f_{\mathcal{A}(L)} \mathcal{M}_a^{-1}\|_{L^2(\mathcal{D}_L)} \leq C L^{\frac{d}{2}}$$

where $\mathcal{M}_a$ is a Maxwellian with some decay rate a,

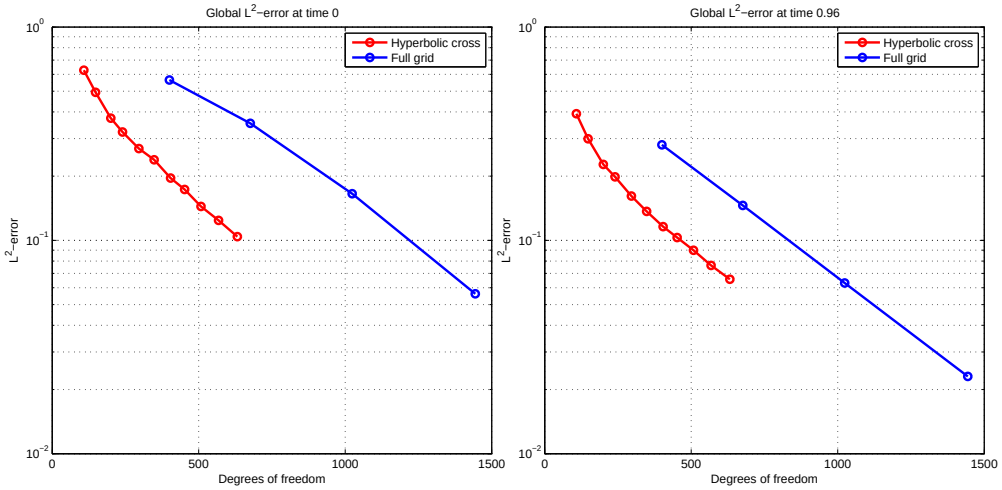
$$\mathcal{M}_a(\boldsymbol{v}) = e^{-a\|\boldsymbol{v}\|^2}.$$

Note that the dependence on L on the right hand side is necessary in the limit $f_{\mathcal{A}(L)} \to \mathcal{M}_a$.

Table 1 shows these norms for various N and L. For the L chosen, a stable result of $\|f_{\mathcal{A}(L)} \mathcal{M}_a^{-1}\|_{L^2(\mathcal{D}_L)} \approx 1.3$ was achieved with relatively modest numbers of degrees of freedom.

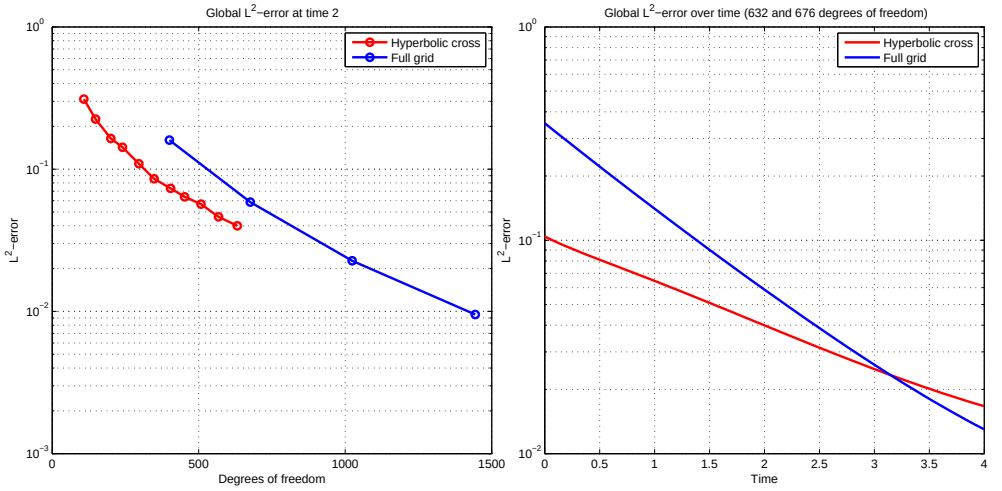
A test with value L would be difficult to perform with the current implementation, as the exponential weight at the boundaries of $\mathcal{D}_L$ is much larger than machine precision, artificially inflating the norms.

[2] Note that equilibrium is *not* $e^{-2\|\boldsymbol{v}\|^2}$ — the perturbation adds both mass and temperature.



(a) Relative error at time $t = 0$.

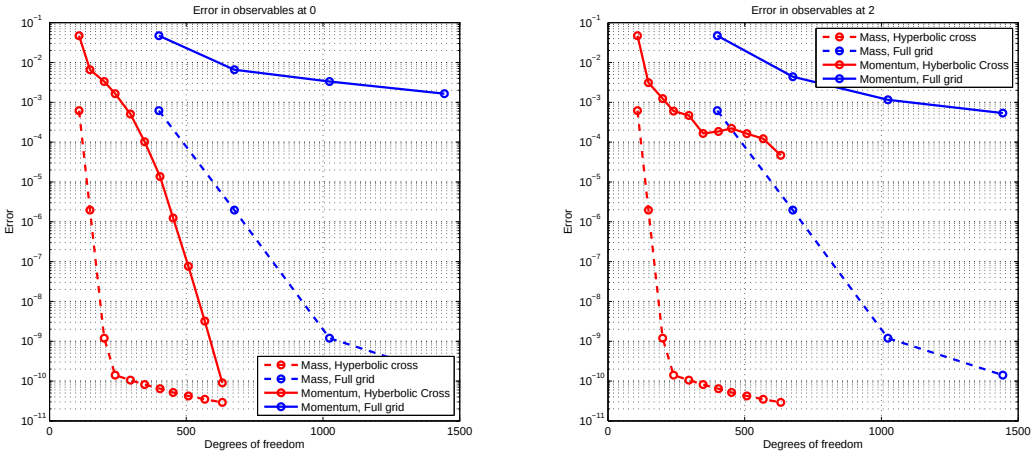
(b) Relative error at time $t = 0.96$.



(c) Relative error at time $t = 2$.

(d) Relative error over time for 676 and 632 degrees of freedom, respectively.

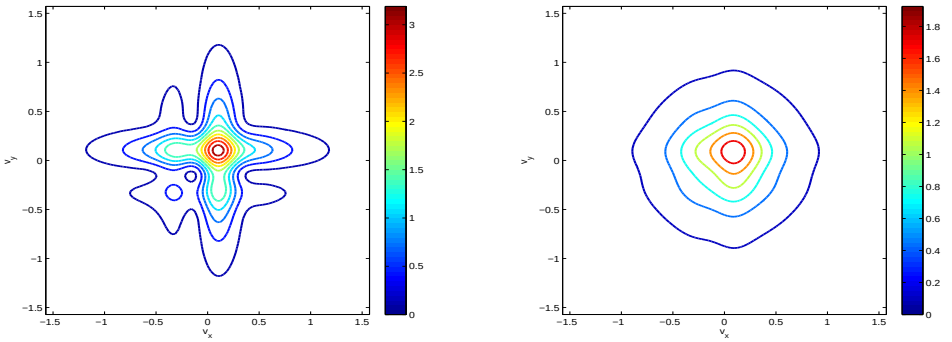
Figure 6: Relative L^2-errors for Section 6.2.



(a) Error in observables at time $t = 0$.

(b) Error in observables at time $t = 2$.

Figure 7: Errors in obserables for Section 6.2.



(a) Initial distribution.

(b) Solution at time $t = 2$.

Figure 8: The crossed beams solution.

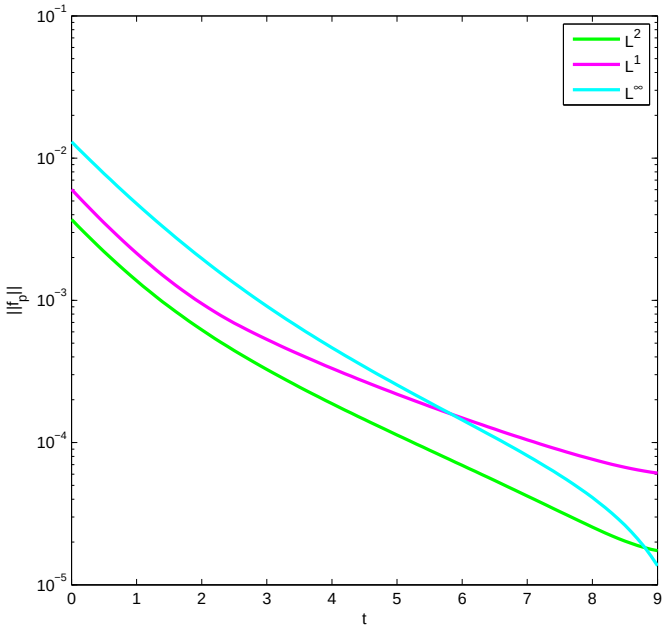
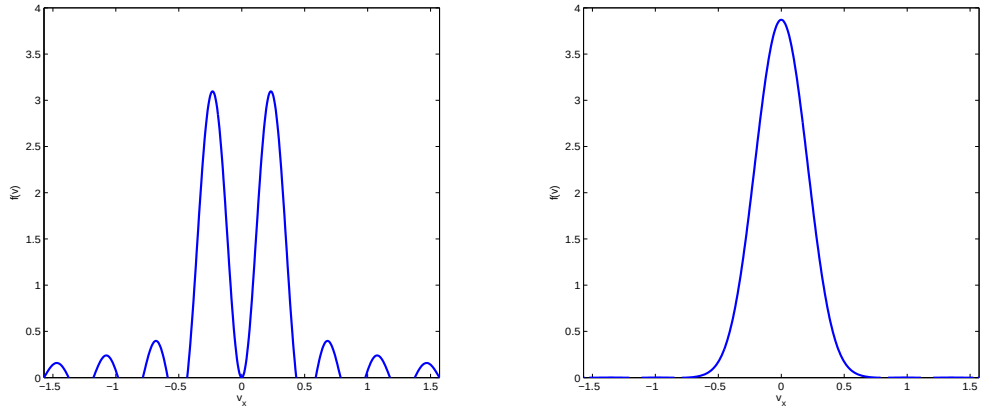


Figure 9: Relaxation to equilibrium (Section 6.3): norms of the perturbation versus time.

N	$L = 0.5\pi$	$L = 0.75\pi$	$L = \pi$
12	1.25	1.25	299
16	1.25	1.25	1.74
20	1.25	1.25	1.28
24	1.25	1.25	1.28
28	1.25	1.25	1.28

N	$L = 0.5\pi$	$L = 0.75\pi$	$L = \pi$
50	1.25	1.30	$5.25 \cdot 10^4$
70	1.25	1.25	$5.67 \cdot 10^3$
90	1.25	1.25	$1.15 \cdot 10^3$
110	1.25	1.25	193
130	1.25	1.25	10.5
150	1.25	1.25	1.99
170	1.25	1.25	1.29
190	1.25	1.25	1.28

Table 1: Numerical evidence for the exponential decay of $f_{\mathcal{A}}$. The table shows $\sup_t \|f_{\mathcal{A}} \mathcal{M}_a^{-1}\|_{L^2(\mathcal{D}_L)}$ for various N and L. The left table is for $\mathcal{A}_{\mathrm{FF}}$ and the right table for $\mathcal{A}_0$. The test case was the same as for Section 6.2, and integrated until $T = 10$.



(a) Initial distribution.

(b) Solution at time $t = 40$.

Figure 10: Initial and final solution for $\kappa \approx 0.45$. Compare to figure 11. Note the y-axis scales.

6.5 Aliasing and antialiasing

The ratio $\kappa = {}^{R}/_{L}$ can be adjusted to control aliasing between different periods of the numerical solution. A small choice of κ will avoid aliasing for several timesteps, while a large choice will allow all possible collisions to be treated in a single timestep.

The following Figures show the evolution for the BKW solution using $N = 32$ degrees of freedom in each direction using a full Fourier grid over the time interval $[0, 10]$.

The initial and final distribution for $\kappa = \frac{2}{3+\sqrt{2}} \approx 0.45$ is shown in Figure 10. It is clear that the solution maintains its equilibrium very well due to the low aliasing.

In [7], time-global consistency and convergence is shown under various assumptions, one of which is $\kappa \geq \sqrt{2}$. The distribution for this scheme at $t = 0.6$ is shown in Figure 11. This time is well before equilibrium is reached, and the effect of significant aliasing can be seen. The solution is clearly unphysical.

Figure 12 shows the time-evolution of the L^2-error for the BKW solution for various κ, ranging from 0.45 to $\sqrt{2}$. For long times, the error will always tend toward 0.9935, which happens to be the L^2-norm between the physical equilibrium solution (the Maxwellian), and the numerical equilibrium solution (a constant function). We observe that for large κ, the numerical solution collapses to a constant very rapidly, and for smaller κ, the solution remains qualitatively correct for much longer times.

References

[1] A. V. Bobylev. The theory of the nonlinear spatially uniform Boltzmann equation for Maxwell molecules. *Soviet Sci. Rev. Sect. C Math. Phys. Rev.*, 7:110–233, 1988.

[2] A. V. Bobylev. Moment inequalities for the Boltzmann equation and applications to spatially homogeneous problems. *J. Stat. Phys*, 88:1183–1214, 1997.

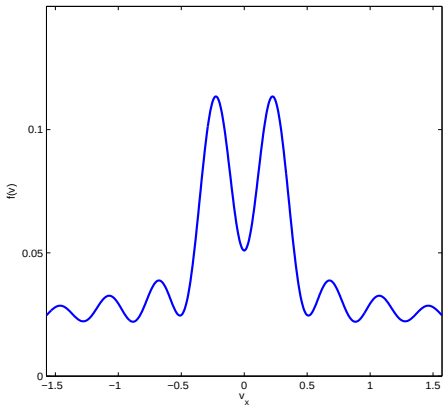


Figure 11: Solution at time $t = 0.6$ for $\kappa = \sqrt{2}$. Compare to Figure 10. Note the y-axis scales.

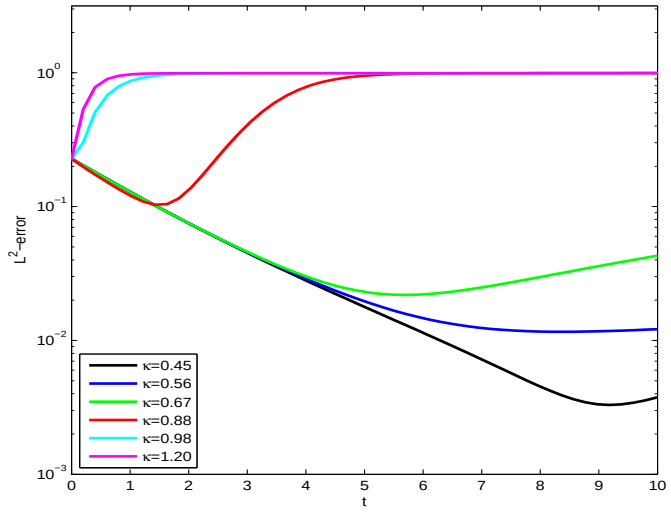


Figure 12: L^2-error versus time for the BKW solution at various κ.

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods in fluid dynamics*. Springer-Verlag, 1988.

[4] C. Cercignani. The boltzmann equation and fluid dynamics. In *Handbook of Mathematical Fluid Mechanics*, volume 1, chapter 1, pages 1–69. Elsevier, Amsterdam, 2002.

[5] S. S. Dragomir. *Some Gronwall type inequalities and applications*. Nova Science Pub Inc, 2003.

[6] M. H. Ernst. Exact solutions of the nonlinear Boltzmann equation. *Journal of Statistical Physics*, 34:1001–1017, 1984.

[7] F. Filbet and C. Mouhot. Analysis of spectral methods for the homogeneous Boltzmann equation. *Transactions of the American Mathematical Society*, 363(4):1947–1980, 2011.

[8] F. Filbet, C. Mouhot, and L. Pareschi. Solving the Boltzmann equation in $N \log_2 N$. *SIAM J. Sci. Comput.*, 28(3):1029–1053, 2006.

[9] Irene M. Gamba and Sri Harsha Tharkabhushanam. Spectral-lagrangian methods for collisional models of non-equilibrium statistical states. *Journal of Computational Physics*, 228(6):2012 – 2036, 2009.

[10] Irene M. Gamba and Sri Harsha Tharkabhushanam. Shock and boundary structure formation by spectral-Lagrangian methods for the inhomogeneous Boltzmann transport equation. *J. Comput. Math.*, 28(4):430–460, 2010.

[11] David Gottlieb and Steven A. Orszag. *Numerical analysis of spectral methods: theory and applications*. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977. CBMS-NSF Regional Conference Series in Applied Mathematics, No. 26.

[12] V. Gradinaru. Fourier transform on sparse grids: code design and the time dependent Schrödinger equation. *Computing*, 80(1):1–22, 2007.

[13] P. T. Gressman and R. M. Strain. Global classical solutions of the Boltzmann equation without angular cut-off. *J. Amer. Math. Soc*, 24:771–847, 2011.

[14] R. Kirsch and S. Rjasanow. A weak formulation of the Boltzmann equation based on the Fourier transform. *J. Stat. Phys.*, 129:483–492, 2007.

[15] S. Knapek. Hyperbolic cross approximation of integral operators with smooth kernel. submitted.

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

[17] L. Pareschi and B. Perthame. A Fourier spectral method for homogeneous boltzmann equations. *Transport Theory Stat. Phys.*, 25:369–383, 1996.

[18] L. Pareschi and G. Russo. Numerical solution of the Boltzmann equation I: spectrally accurate approximation of the collision operator. *SIAM J. Numer. Anal.*, 37:1217–1245, 2000.

[19] L. Pareschi, G. Toscani, and C. Villani. Spectral methods for the non cut-off boltzmann equation and numerical grazing collision limit. *Numerische Mathematik*, 93:527–548, 2003.

[20] E. Carneiro R. J. Alonso and I. M. Gamba. Convolution inequalities for the boltzmann collision operator. *Communications in Mathematical Physics*, 298(2):293–322, 2010.

[21] C. J. Tourenne. The entropy of the BKW solution. *Journal of Statistical Physics*, 32:71–80, 1983.

[22] C. Villani. Fisher information estimates for boltzmann's collision operator. *Journal de Mathematiques Pures et Appliquees*, 77:821–837, 1998.

[23] C. Villani. A review of mathematical topics in collisional kinetic theory. In *Handbook of Mathematical Fluid Mechanics*, volume 1, chapter 2, pages 71–305. Elsevier, Amsterdam, 2002.