Eidgenössische
Technische Hochschule
Zürich

*Ecole polytechnique fédérale de Zurich*
*Politecnico federale di Zurigo*
*Swiss Federal Institute of Technology Zurich*

# Hyperbolic cross approximation for the spatially homogeneous Boltzmann equation

E. Fonn, Ph. Grohs and R. Hiptmair

Research Report No. 2012-28
October 2012

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

# Hyperbolic cross approximation for the spatially homogeneous Boltzmann equation

E. Fonn P. Grohs R. Hiptmair

October 2, 2012

**Abstract**

The nonlinear spatially homogeneous integro-differential Boltzmann equation is a uniquely challenging task for numerical solvers due to the difficulty of efficiently computing the collision operator. A popular method is to expand the solution in Fourier modes and to truncate the collision operator. We present an approach based on the hyperbolic cross, whereby the performance can be greatly enhanced in some situations, as well as an offset method, which takes advantage of the known equilibrium solutions. Some numerical experiments are presented in two dimensions with Maxwellian kernels.

Some error estimates are also given, where it is shown that under reasonable assumptions, the numerical solution converges to the analytical.

## 1 The Boltzmann equation

The Boltzmann equation reads

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f = Q(f, f), \tag{1}$$

for $f : \mathbb{R}^+ \times \Omega \times \mathbb{R}^d \to \mathbb{R}^+$, with initial conditon $f(0, \boldsymbol{x}, \boldsymbol{v}) = f_0(\boldsymbol{x}, \boldsymbol{v})$, and suitable boundary conditions. Here, $f(t, \boldsymbol{x}, \boldsymbol{v})$ is to be interpreted as the density of molecules located at $[\boldsymbol{x}, \boldsymbol{x} + \mathrm{d}\boldsymbol{x}]$ with velocity in $[\boldsymbol{v}, \boldsymbol{v} + \mathrm{d}\boldsymbol{v}]$ at time $t$

The bilinear collision operator $Q$ is what distinguishes the Boltzmann equation from the other kinetic transport equations, and it takes the form (dropping the variables $t$ and $x$ for 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 $Q(f, h)(\boldsymbol{v})$ is to mean $Q(f, h)$ evaluated at $\boldsymbol{v}$. We have shorthand notation

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

where the pre- and post-collision velocities are related by

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

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 later) to separate the kernel and write

$$Q(f,h) = Q_{\text{gain}}(f,h) - Q_{\text{loss}}(f,h).$$

Were it not for the collision term, (1) would have an analytical solution given by

$$f(t,\boldsymbol{x},\boldsymbol{v}) = f_0(\boldsymbol{x} - \boldsymbol{v}t, \boldsymbol{v}).$$

It is the collision term that makes (1) so numerically challenging, and so most attempts to tacke it will focus primarily on a discretization of $Q$. This paper is no different. To that end, we consider the related spatially homogeneous Boltzmann equation

$$\frac{\partial f}{\partial t}(t,\boldsymbol{v}) = Q(f,f)(t,\boldsymbol{v}) \tag{3}$$

for $f : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^+$ instead.

## 1.1 Equilibrium solutions

The conserved quantities for (3) are the observables mass, momentum and energy,

$$\rho(t) = \int_{\mathbb{R}^d} f(t,v)\,\mathrm{d}\boldsymbol{v}, \quad \boldsymbol{u}(t) = \frac{1}{\rho(t)}\int_{\mathbb{R}^d} f(t,\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}, \quad E(t) = \frac{1}{\rho(t)}\int_{\mathbb{R}^d} f(t,\boldsymbol{v})\|\boldsymbol{v}\|^2\,\mathrm{d}\boldsymbol{v}.$$

For all reasonable kernels $B$ and initial values $f_0$ the long-term nonzero equilibrium solutions to the Boltzmann equation are the Maxwellians (or Gaussians)

$$M(\boldsymbol{v}) = M(\rho,\boldsymbol{u},T)(\boldsymbol{v}) = \frac{\rho}{(2\pi T)^{d/2}}\exp\left(-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2T}\right), \tag{4}$$

which are fully characterized by the three observables. Here, $T$ is the temperature

$$T(t) = \frac{1}{d\rho(t)}\int_{\mathbb{R}^d}\|\boldsymbol{u}-\boldsymbol{v}\|^2 f(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = \frac{1}{d}\left(E(t) - \|\boldsymbol{u}(t)\|^2\right).$$

Thus, the equilibrium solution is available *a priori* through the observables $\rho$, $\boldsymbol{u}$ and $E$, and as such, the primary niche for numerical solvers ought to be situations where $f$ is far removed from equilibrium.

What is more, a recent proof by Gressman and Strain [5] has shown that $f$ converges to equilibrium exponentially fast. These are important features of the Boltzmann equation that must be kept in mind when designing numerical schemes.

## 1.2 Scaling

In the following, we use $\hat{g}$ to denote the Fourier transform of a function $g$. This lemma is due to Bobylev [1]:

**Lemma 1.** *If $B$ depends only on $\cos\theta$ we have*

$$\widehat{Q(f,f)}(\boldsymbol{\xi}) = \int_{S^{d-1}} B(\boldsymbol{e}_{\boldsymbol{\xi}}\cdot\boldsymbol{\sigma})\left[\hat{f}\left(\frac{1}{2}\|\boldsymbol{\xi}\|(\boldsymbol{e}_{\boldsymbol{\xi}}+\boldsymbol{\sigma})\right)\hat{f}\left(\frac{1}{2}\|\boldsymbol{\xi}\|(\boldsymbol{e}_{\boldsymbol{\xi}}-\boldsymbol{\sigma})\right) - \hat{f}(\boldsymbol{\xi})\hat{f}(\boldsymbol{0})\right]\mathrm{d}\boldsymbol{\sigma},$$

*where $e_x = x/\|x\|$.*

**Proposition 2.** *Assume $f(t, \boldsymbol{v})$ solves* (3) *and $\alpha > 0$. Then*

$$g_\alpha(t, \boldsymbol{v}) = \alpha f(\alpha t, \boldsymbol{v})$$

*is another solution. Moreover, if the kernel $B$ depends only on $\cos\theta$, then*

$$h_\alpha(t, \boldsymbol{v}) = \alpha^d f(t, \alpha\boldsymbol{v})$$

*is yet another solution.*

*Proof.* For the first claim, we have

$$\begin{aligned} \frac{\partial g_\alpha}{\partial t}(t, \boldsymbol{v}) &= \alpha^2 \frac{\partial f}{\partial t}(\alpha t, \boldsymbol{v}) = \alpha^2 Q(f, f)(\alpha t, \boldsymbol{v}) \\ &= Q(\alpha f, \alpha f)(\alpha t, \boldsymbol{v}) = Q(g_\alpha, g_\alpha)(t, \boldsymbol{v}). \end{aligned}$$

For the second, note first that

$$\hat{h}_\alpha(t, \boldsymbol{\xi}) = \hat{f}\left(t, \frac{\boldsymbol{\xi}}{\alpha}\right). \tag{5}$$

By assumption and lemma 1 we have that

$$\partial_t \hat{f}(t, \boldsymbol{\xi}) = \int_{S^{d-1}} B(\boldsymbol{e_\xi} \cdot \boldsymbol{\sigma}) \left[\hat{f}\left(\frac{1}{2}\|\boldsymbol{\xi}\|(\boldsymbol{e_\xi} + \boldsymbol{\sigma})\right) \hat{f}\left(\frac{1}{2}\|\boldsymbol{\xi}\|(\boldsymbol{e_\xi} - \boldsymbol{\sigma})\right) - \hat{f}(\boldsymbol{\xi})\hat{f}(\boldsymbol{0})\right] \mathrm{d}\boldsymbol{\sigma},$$

so, using $\boldsymbol{e_\xi} = e_{\alpha\boldsymbol{\xi}}$ and $\alpha > 0$, we get

$$\partial_t \hat{f}(t, \alpha\boldsymbol{\xi}) = \int_{S^{d-1}} B(\boldsymbol{e_\xi} \cdot \boldsymbol{\sigma}) \left[\hat{f}\left(\frac{1}{2\alpha}\|\boldsymbol{\xi}\|(\boldsymbol{e_\xi} + \boldsymbol{\sigma})\right) \hat{f}\left(\frac{1}{2\alpha}\|\boldsymbol{\xi}\|(\boldsymbol{e_\xi} - \boldsymbol{\sigma})\right) - \hat{f}\left(\frac{\boldsymbol{\xi}}{\alpha}\right)\hat{f}(\boldsymbol{0})\right] \mathrm{d}\boldsymbol{\sigma},$$

which by (5) is equivalent to

$$\partial_t \hat{h}_\alpha(t, \boldsymbol{\xi}) = \int_{S^{d-1}} B(\boldsymbol{e_\xi} \cdot \boldsymbol{\sigma}) \left[\hat{h}_\alpha\left(\frac{1}{2}\|\boldsymbol{\xi}\|(\boldsymbol{e_\xi} + \boldsymbol{\sigma})\right) \hat{h}_\alpha\left(\frac{1}{2}\|\boldsymbol{\xi}\|(\boldsymbol{e_\xi} - \boldsymbol{\sigma})\right) - \hat{h}_\alpha(\boldsymbol{\xi})\hat{h}_\alpha(\boldsymbol{0})\right] \mathrm{d}\boldsymbol{\sigma},$$

which by lemma 1 is $\widehat{Q(h_\alpha, h_\alpha)}(\boldsymbol{\xi})$. The claim follows by inverse Fourier transform. □

## 1.3 Integral representations of the Boltzmann collision operator

A useful form of the collision operator (2) is

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

which is achieved by a change of variables $\boldsymbol{g} = \boldsymbol{v} - \boldsymbol{v}_*$, and is used in [8] to develop a numerical scheme.

On the other hand, in [7], it is shown that an alternate representation of (2) is

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

This integral, due to the delta function, is also five-dimensional. The condition $\boldsymbol{x} \perp \boldsymbol{y}$ allows us to express the transformed collision kernel as

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

Of course, as $x \perp y$, we have $\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2$, and so it is clear that $\tilde{B}$ depends only on $\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}\|$.

While (6) and (7) are equivalent in the continuous formulation, this is not necessarily true post-discretization.

# 2 Discretization of the Boltzmann collision operator

## 2.1 Truncation in velocity space

The following proposition is given in [8], with $\mathcal{B}_R$ the ball of radius $R$ centered at 0.

**Proposition 3.** *Let* supp $f, h \subset \mathcal{B}_R$. *Then,* 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 $L \geq (2+\sqrt{2})R$. 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.

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

**Proposition 4.** *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 $L \geq (2+\sqrt{2})R$. Then, for $v \in \mathcal{B}_{\sqrt{2}R}$, the following representations of $Q(f,h)(\boldsymbol{v})$ are equivalent.*

$$\begin{aligned} Q(f,h)(\boldsymbol{v}) &= \int_{\mathcal{B}_{2R}} \int_{S^{d-1}} B(\|\boldsymbol{g}\|, \theta)(h'_* f' - h_* f) \, \mathrm{d}\boldsymbol{\sigma} \, \mathrm{d}\boldsymbol{g} &(8)\\ &= \int_{\mathcal{B}_{\sqrt{2}R}} \int_{\mathcal{B}_{\sqrt{2}R}} \tilde{B}(\boldsymbol{x}, \boldsymbol{y}) \delta(\boldsymbol{x} \cdot \boldsymbol{y}) &(9)\\ &\quad \cdot [h(\boldsymbol{v}+\boldsymbol{y}) f(\boldsymbol{v}+\boldsymbol{x}) - h(\boldsymbol{v}+\boldsymbol{x}+\boldsymbol{y}) f(\boldsymbol{v})] \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}. &(10) \end{aligned}$$

*Proof.* Representation (8) follows from proposition 3. Following the outline from [7], it can be shown that

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

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 $\max\{\|\boldsymbol{v}\|, \|\boldsymbol{v}+\boldsymbol{x}\|, \|\boldsymbol{v}+\boldsymbol{y}\|, \|\boldsymbol{v}+\boldsymbol{x}+\boldsymbol{y}\|\} \leq L$. This concludes the proof. □

Henceforth, we will denote by $Q^R$ the bounded version of $Q$ as defined in (8) and (9).

### 2.2 Fourier discretization

Let us now discretize $f$ by representing it as a truncated $d$-dimensional Fourier series in $v$,

$$f_N(t,\boldsymbol{v}) = \sum_{\boldsymbol{k}\in\mathcal{A}} \hat{f}_{\boldsymbol{k}}(t) e^{i\boldsymbol{k}\cdot\boldsymbol{v}}, \tag{11}$$

where $\mathcal{A} \subset \frac{\pi}{L}\mathbb{Z}^d$ is some discrete and finite but so far unspecified set of Fourier modes. Then, (3) yields

$$\sum_{\boldsymbol{k}\in\mathcal{A}} \hat{f}'_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{v}} = \sum_{\boldsymbol{l},\boldsymbol{m}\in\mathcal{A}} \hat{f}_{\boldsymbol{l}} \hat{f}_{\boldsymbol{m}} Q^R\left(e^{i\boldsymbol{l}\cdot\boldsymbol{v}}, e^{i\boldsymbol{m}\cdot\boldsymbol{v}}\right). \tag{12}$$

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

$$Q^R\left(e^{i\boldsymbol{l}\cdot\boldsymbol{v}}, e^{i\boldsymbol{m}\cdot\boldsymbol{v}}\right) = \hat{\beta}(\boldsymbol{l},\boldsymbol{m}) e^{i(\boldsymbol{m}+\boldsymbol{l})\cdot\boldsymbol{v}}. \tag{13}$$

Plugging this into (12), discarding coefficients outside $\mathcal{A}$ and comparing the remaining coefficients, gives us the following quadratic ODE for the coefficients $\hat{f}_{\boldsymbol{k}}$:

$$\hat{f}'_{\boldsymbol{k}} = \sum_{\substack{\boldsymbol{l},\boldsymbol{m}\in\mathcal{A} \\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \hat{f}_{\boldsymbol{l}} \hat{f}_{\boldsymbol{m}} \hat{\beta}(\boldsymbol{l},\boldsymbol{m}), \tag{14}$$

where $'$ stands for differentiation with respect to time.

The coefficients $\hat{\beta}(\boldsymbol{l},\boldsymbol{m})$ are called the *kernel modes*. These can be evaluated through (13). Note that since the Fourier modes $e^{i\boldsymbol{k}\cdot\boldsymbol{v}}$ do not satisfy the conditions of proposition 4, representations (6) and (7) will yield *different* values for $\hat{\beta}$. In the following, we will denote by $\hat{\beta}_d^P$ those values arising from (6), and by $\hat{\beta}_d^M$ those arising from (7).

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 $\beta_d^\cdot$ 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 $\boldsymbol{l},\boldsymbol{m} \in \mathcal{A}$.

### 2.3 Observables

For the linear functionals $\rho$, $\boldsymbol{u}$ and $E$, we necessarily have representations in terms of the coefficients $\hat{f}_k$ from (11):

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

The values $\hat{\rho}^{\boldsymbol{k}}$, $\hat{\boldsymbol{u}}^{\boldsymbol{k}}$ 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 $u$ 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. 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, for momentum, we have

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

And finally the energy:

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

Thus we see that even if $\mathcal{A}$ is relatively full, 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 $dN$.

## 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 pursued in [8] and [7] is

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

i.e. the full $d$-dimensional discrete Fourier representation with $N$ degrees of freedom in each direction.

An alternative choice is the *hyperbolic cross* Fourier transform, which can be considered the frequency-space equivalent of sparse grids (see [4] for more details),

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

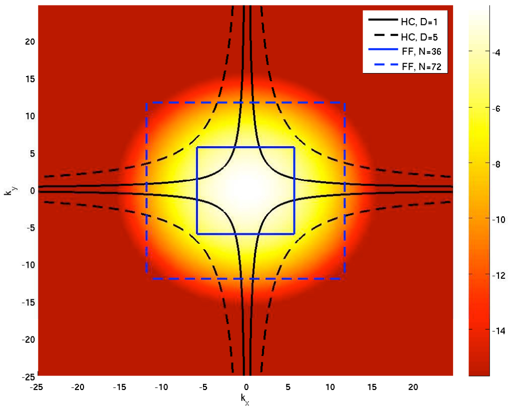


Figure 1: Shows the amplitude of the spectrum of a Gaussian, with the Fourier modes belonging to hyperbolic crosses and full grids of comparable size delineated for comparison. It seems that full grids offer much better approximation of Gaussians.

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

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

making $\tilde{\boldsymbol{k}}$ integral. The parameter $T \leq 0$ controls the "fatness" of the hyperbolic cross, with the classical hyperbolic cross being given by $T = 0$.

For $T = 0$, we have $|\mathcal{A}_0(N)| = O(N \log^d N)$ compared to $|\mathcal{A}_0(N)| = N^d$.

It's also worth noting that $\mathcal{A}_{\mathrm{FF}}(N) \supseteq \mathcal{A}_T(N)$, yet for any functional $\ell$ that depends only on Fourier coefficients on the axes (such as $\rho$, $\boldsymbol{u}$ and $E$), we have for $L^2$-projections $f_{\mathcal{A}_0}$, $f_{\mathcal{A}_{\mathrm{FF}}}$ of $f$ onto span $\left\{e^{i\boldsymbol{k}\cdot\boldsymbol{v}} \mid \boldsymbol{k} \in \mathcal{A}_*\right\}$ that

$$\ell\left(f_{\mathcal{A}_0(N,D)}\right) = \ell\left(f_{\mathcal{A}_{\mathrm{FF}}(N)}\right).$$

In spite of this property, it is not necessarily the case that the hyperbolic cross provides a good approximation for $f$ itself, and by extension, $Q(f,f)$ and $\partial_t f$. In particular, it 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 $\mathcal{A}_{\mathrm{FF}}(N)$, see figure 1.

This indicates that the hyperbolic cross would be a poor choice for approximating near-equilibrium solutions. It could still provide a useful tool for certain situations with $f$ far removed from equilibrium.

## 3.1 Complexity

The evaluation of $Q(f,f)$ requires the formation of the sum (14). There are no generally fast algorithms to compute this, unless some kind of separability of $\hat{\beta}$ is available, as in [7], which seems to be the case only for certain specific kernels $B$. A straightforwardly naive implementation has cost on the order of the number of

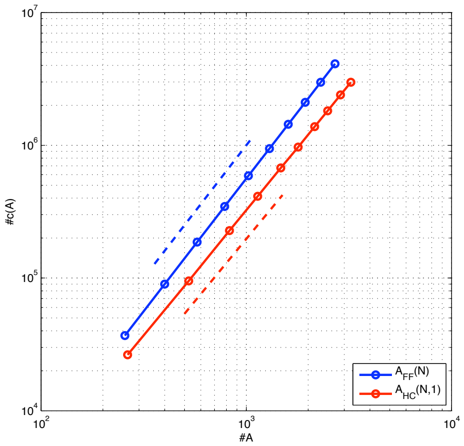


Figure 2: Logarithmic plot of $\sharp c(\mathcal{A})$ versus $\sharp\mathcal{A}$ for $\mathcal{A}_{\mathrm{FF}}(N)$ (blue) and $\mathcal{A}_0(N)$ (red), showing a minor improvement in complexity for the hyperbolic cross. The dashed lines are the power laws $y = c_1 x^2$ and $y = c_2 x^{15/8}$.

pairs of vectors $\boldsymbol{l}, \boldsymbol{m} \in \mathcal{A}$ such that $\boldsymbol{l} + \boldsymbol{m} \in \mathcal{A}$. We are asking for the cardinality of the combination set

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

**Proposition 5.** *For $d$ fixed,*

$$\sharp c(\mathcal{A}_{FF}(N)) = O(\sharp\mathcal{A}_{FF}(N)^2) = O(N^{2d}).$$

*Proof.* For simplicity, let $T = \pi$, so that $\mathcal{A}_{\mathrm{FF}}(N)$ has integral elements. Let $M = N/2$. Then, $\boldsymbol{l} + \boldsymbol{m} \in \mathcal{A}$ is equivalent with

$$-M \leq l_i + m_i \leq M - 1 \qquad \forall i$$

Whatever the value of $l_i$, there are at least $M$ values of $m_i$ that satisfies this inequality (say, $0 \leq -\operatorname{sgn}(l_i) m_i < M$). Thus,

$$\sharp c(\mathcal{A}_{\mathrm{FF}}(N)) = O((NM)^d) = O(N^{2d}).$$

□

This is clearly the worst possible case for any $\mathcal{A}$, counting by degrees of freedom. However, the hyperbolic cross can do better. Experimentally we have, for instance,

$$\sharp c(\mathcal{A}_0(N)) = O(\sharp\mathcal{A}_0(N)^{\frac{15}{8}}).$$

See figure 2.

# 4 Approximation Error

## 4.1 Basic Assumptions and Estimates

In our analysis we assume the VHS model

$$B(\|\boldsymbol{u}\|, \cos\theta) = \|\boldsymbol{u}\|^{\lambda} b(\cos\theta), \qquad \lambda \in \mathbb{R} \tag{17}$$

with $b$ satisfying *Grad's cutoff assumption*

$$A(\boldsymbol{u}) := \int_{S^{d-1}} b(\cos\theta)\,\mathrm{d}\boldsymbol{\sigma} = \int_{S^{d-1}} b\left(\left\langle \frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \boldsymbol{\sigma} \right\rangle\right) \mathrm{d}\boldsymbol{\sigma} < \infty. \tag{18}$$

During all of this section we fix $R$ and $L$ as introduced in Section 2.1. Our first result is a boundedness result of $Q^R$ in $L^2$ for periodic functions.

**Theorem 6.** *Assume that $\lambda \geq -\frac{d}{2}$. Then we have for functions $f, g$ which are $L$-periodic with fundamental domain $\mathcal{D}_L = [-L, L]^d$ the estimate*

$$\|Q^R(f,g)\|_{L^2(\mathcal{D}_L)} \leq C_{\mathrm{pc}} L^{\frac{d}{2}+\max(2\lambda,0)} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)},$$

*where $C_{\mathrm{pc}}$ is a constant that may depend on $d$ and $\lambda$. Moreover, the function $Q^R(f,g)$ is $L$-periodic.*

The proof of this theorem treats the gain- and loss term seperatly. In order to handle the gain term we need the following result from [9].

**Theorem 7.** *Assume that $\lambda \geq 0$ and $f, g$ general functions on $\mathbb{R}^d$. Then we have the estimate*

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \leq C_{\mathrm{pos}} \|f\|_{L^2_\lambda(\mathcal{D}_{3L})} \|g\|_{L^1_\lambda(\mathcal{D}_{3L})},$$

*where we define for $\nu > 0$, $p \geq 1$ and a domain $D \subset \mathbb{R}^d$*

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

*For $-d < \lambda < 0$ and*

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

*we have*

$$\|Q^{R,+}(f,g)\|_{L^r(\mathcal{D}_L)} \leq C_{\mathrm{neg}} \|f\|_{L^p(\mathcal{D}_{3L})} \|g\|_{L^q(\mathcal{D}_{3L})}.$$

*Proof.* This result has been shown in [9] for the non-truncated gain operator $Q^+$ and with the norms for the terms $Q^{R,+}(f,g), f, g$ taken over all of $\mathbb{R}^d$.

As to the effect of the truncation we remark that exactly the same arguments as in [9] 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 $[-3L, 3L]^d$ we remark that by the definition of $Q^{R,+}$, the values of $Q^{R,+}(f,g)(v)$ for $v \in [-L, L]^d$ only depend on $f$ and $g$ restricted to $[-3L, 3L]^d$. □

*Proof of Theorem 6.* We first show the desired statement for the loss term $Q^{R,-}$. We have

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

where

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

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

$$\|A_\lambda\|_{L^\infty(\mathcal{D}_L)} \le C_A L^{\max(\lambda,0)}$$

which entails

$$\begin{aligned}\|Q^{R,-}(f,g)\|_{L^2(\mathcal{D}_L)} &\le \|f\|_{L^2(\mathcal{D}_L)} \|A_\lambda * g\|_{L^\infty(\mathcal{D}_L)} \\ &\le \|f\|_{L^2(\mathcal{D}_L)} \|A_\lambda\|_{L^\infty(\mathcal{D}_L)} \|g\|_{L^1(\mathcal{D}_{3L})}. \quad (19)\end{aligned}$$

The last inequality holds since only the values of $g$ restricted to $[-3L, 3L]^d$ are used for the evaluation of $A_\lambda * g(\boldsymbol{v})$, $\boldsymbol{v} \in [-L, L]^d$.

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

We can further estimate

$$\|g\|_{L^1(\mathcal{D}_{3L})} \le (6L)^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_{3L})}.$$

Now we observe that, due to periodicity of $g$, the above quantity can be bounded by

$$3(6L)^{\frac{d}{2}} \|g\|_{L^2(\mathcal{D}_L)}$$

Plugging this estimate into (19) yields the desired estimate for the loss term.

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

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \le C_{\text{pos}} \|f\|_{L^2_\lambda(\mathcal{D}_{3L})} \|g\|_{L^1_\lambda(\mathcal{D}_{3L})}. \quad (20)$$

Since $f$ and $g$ are periodic, we have

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \le 27 C_{\text{pos}} \|f\|_{L^2_\lambda(\mathcal{D}_L)} \|g\|_{L^1_\lambda(\mathcal{D}_L)}. \quad (21)$$

Since

$$\|f\|_{L^p_\lambda(\mathcal{D}_L)} \le 2(\sqrt{d}L)^\lambda \|f\|_{L^p(\mathcal{D}_L)}$$

for all $p \ge 1$ and

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

we arrive at

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \le 108 C_{\text{pos}} d^\lambda 2^{\frac{d}{2}} L^{\frac{d}{2}+2\lambda} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)}$$

whenever $\lambda \ge 0$. Now we turn to the case $-d/2 \le \lambda < 0$. By our assumptions on $d$ and $\lambda$, 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

$$\|Q^{R,+}(f,g)\|_{L^2(\mathcal{D}_L)} \leq 9C_{\mathrm{neg}}(6L)^{\frac{d}{2}}\|f\|_{L^2(\mathcal{D}_L)}\|g\|_{L^2(\mathcal{D}_L)}.$$

using the estimate

$$\|f\|_{L^p(\mathcal{D}_{3L})} \leq 3(6L)^{d\left(\frac{1}{p}-\frac{1}{2}\right)}\|f\|_{L^2(\mathcal{D}_L)}$$

and choosing

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

This gives us the desired estimate with, say

$$C_{\mathrm{pc}} = (3C_A + 108C_{\mathrm{pos}} + 9C_{\mathrm{neg}}) \cdot 6^{\frac{d}{2}} d^{\max(\lambda,0)}.$$

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 [11].

**Proposition 8.** *We have*

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

*where $\partial_j$ denotes the derivative in the $j$-th coordinate direction.*

*Proof.* This result has been proven in [11] but we give a simpler proof which applies to the case when $f$ and $g$ are $L$-periodic, which is of interest to us. In this case we can write

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

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. □

In the following we will develop estimates for the approximation error

$$\|Q(f,f) - P_{\mathcal{A}} Q^R(P_{\mathcal{A}} f, P_{\mathcal{A}} f)\|_{L_2(\mathbb{R}^d)},$$

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

## 4.2 Consistency

In this section we develop error bounds for $L^2$-error between the approximate application of the discretized collision operator $Q^R(f,g)$ and the application of the true operator $Q(f,g)$ in terms of $f$ and $g$. Our main result Theorem 13 may be viewed as a generalization of the approximation results in [8]. Our results are more general in several aspects:

(i) The main approximation result in [8] estimates the approximation error in terms of a Sobolev norm of $f, g$ and $Q(f, g)$. In contrast we notice that, due to the product rule shown in Proposition 8, actually any function norm of $Q(f, g)$ based on derivatives can be estimated by the corresponding norms for $f$ and $g$. Therefore, our Theorem 13 only requires finiteness of the norms of the functions $f$ and $g$.

(ii) Our results are not confined to approximation on full Fourier grids. In fact, in the next section we develop error estimates for a whole family of Fourier grids, including hyperbolic cross approximation and approximation on full grids.

(iii) While the results of [8] only hold for the case of VHS kernels, we also treat more general kernels satisfying Grad's cutoff assumption (17), (18).

We will develop approximation errors 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 no more difficult but it would require a heavier notation.

The corresponding smoothness spaces are the following *mixed Sobolev spaces* as defined in [6].

**Definition 9.** We define the smoothness spaces

$$\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L) := \left\{ f \in L^2(\mathcal{D}_L) : \sum_{\alpha \leq (t,\dots,t),\ |\beta|_\infty \leq l} \left\| \partial^\alpha \partial^\beta f \right\|_{L^2(\mathcal{D}_L)} < \infty \right\}. \tag{22}$$

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

In [6] the following approximation result is shown.

**Theorem 11.** *We have*

$$\|f - \mathcal{P}_{\mathcal{A}_T(N)} f\|_{L^2(\mathcal{D}_L)} \lesssim \begin{cases} (1+N)^{-l-t+Tt\frac{d-1}{d-T}} \|f\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)} & \textit{for} \quad T \geq -\frac{l}{t} \\ (1+N)^{-l-t} \|f\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)} & \textit{for} \quad T \leq -\frac{l}{t} \end{cases}$$

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*

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

*Proof.* Note that by Proposition 8, every derivative

$$\partial^\alpha Q^R(f, g)$$

can be expressed as a linear combination of terms

$$Q^R(\partial^{\alpha_1} f, \partial^{\alpha_2} g), \quad \alpha_1 + \alpha_2 = \alpha.$$

It follows that we can estimate

$$\|\partial^\alpha Q^R(f, g)\|_{L^2(\mathcal{D}_L)} \lesssim \sum_{\alpha_1 + \alpha_2 = \alpha} \|Q^R(\partial^{\alpha_1} f, \partial^{\alpha_2} g)\|_{L^2(\mathcal{D}_L)}.$$

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*

$$\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{L^2(\mathcal{D}_L)} \lesssim \begin{cases} (1+N)^{-l-t+Tt\frac{d-1}{d-T}}\left(1+\|f\|^2_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)}\right) & \textit{for} \quad T \geq -\frac{l}{t}, \\ (1+N)^{-l-t}\left(1+\|f\|^2_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)}\right) & \textit{for} \quad T \leq -\frac{l}{t}. \end{cases}$$

*Proof.* We write

$$\begin{aligned}\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{L^2(\mathcal{D}_L)} \\ \leq \left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(f,f)\right\|_{L^2(\mathcal{D}_L)} \\ + \left\|P_{\mathcal{A}_T(N)}Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{L^2(\mathcal{D}_L)}.\end{aligned}$$

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

$$\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(f,f)\right\|_{L^2(\mathcal{D}_L)} + \left\|Q^R(f,f) - Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{L^2(\mathcal{D}_L)}.$$

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

$$\left\|Q^R(f,f) - P_{\mathcal{A}_T(N)}Q^R(f,f)\right\|_{L^2(\mathcal{D}_L)} \lesssim \begin{cases} (1+N)^{-l-t+Tt\frac{d-1}{d-T}}\|Q^R(f,f)\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)} & \text{for} \quad T \geq -\frac{l}{t}, \\ (1+N)^{-l-t}\|Q^R(f,f)\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)} & \text{for} \quad T \leq -\frac{l}{t}. \end{cases}$$

Now, all we need to do is estimate the quantity $\|Q^R(f,f)\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)}$ in terms of the mixed Sobelev 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

$$\left\|Q^R(f,f) - Q^R(P_{\mathcal{A}_T(N)}f, P_{\mathcal{A}_T(N)}f)\right\|_{L^2(\mathcal{D}_L)},$$

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

$$\left\|Q^R\left(f - P_{\mathcal{A}_T(N)}f, f\right) + Q^R\left(P_{\mathcal{A}_T(N)}f, f - P_{\mathcal{A}_T(N)}f\right)\right\|_{L^2(\mathcal{D}_L)}.$$

Now we can invoke the bound of Theorem 6 to bound this quantity by

$$C_{10}(R,L)\left\|f - P_{\mathcal{A}_T(N)}f\right\|_{L_2(\mathcal{D}_L)}\left(\|f\|_{L^2(\mathcal{D}_L)} + \left\|P_{\mathcal{A}_T(N)}f\right\|_{L^2(\mathcal{D}_L)}\right).$$

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

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

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

*Remark* 14. The previous result opens up the door 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 this section is to establish estimates for the time-dependent error between the solution to the actual Boltzmann equation

$$\dot{f} = Q(f, f), \qquad f(0) = f_0$$

and the truncated and projected equation

$$\dot{f}_{\mathcal{A}} = P_{\mathcal{A}} Q^R(f_{\mathcal{A}}, f_{\mathcal{A}}), \qquad f_{\mathcal{A}}(0) = P_{\mathcal{A}} f_0.$$

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

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

The estimates cannot be completed without a number of realistic assumptions, which are:

- that $f$ and $f_R$ satisfy exponential decay for all $t$. They are bounded by some constant times a function

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

  The constant in question will be denoted by $C_{\mathrm{exp}}$ and $C_{\mathrm{exp},R}$, respectively.
- that $f_{\mathcal{A}}$ is bounded in the $L^2(\mathcal{D}_L)$-norm, for all $t$:

  $$\|f_{\mathcal{A}}(t)\|_{L^2(\mathcal{D}_L)} \leq C_{\mathrm{b}}.$$

- that $f_{\mathcal{A}}$ also satisfies an exponential decay in the sense

  $$\|f_{\mathcal{A}(L)} \mathcal{M}_a^{-1}\|_{L^2(\mathcal{D}_L)} \leq C_{\mathrm{dec}} L^{\frac{d}{2}} \tag{24}$$

  for all $t$ and $L$, where we have assumed some dependency between $\mathcal{A}$ and $L$ to which we will return in section 6.4.

We will also take for granted that $L$ depends on $R$ in such a fashion that only values in $\mathcal{D}_L$ are used for the evaluation of $Q^R$ in $\mathcal{B}_R$. For this it suffices that $L(R) = 3L$. We also assume that the collision kernel is of VHS type with $\lambda \geq -\frac{d}{2}$ and that $t$ is restricted to some interval $[0, T]$.

We are primarily interested in error estimates for fixed initial data, asymptotically as $\mathcal{A}$ and $L$ grow. It is clear, and we will see so later, that there must be some dependence $\mathcal{A} = \mathcal{A}(L)$ if this is to work. Quantities that depend only on $d$, $\lambda$

and the assumed decay properties ($a$, $C_{\text{exp}}$ and $C_{\text{b}}$, for example) will be treated as constants.

Our results are valid for *sufficiently large* $L$. It is worth noting that this condition is in reality quite modest.

Throughout this section, we will simplify notation by using

$$Q(f) \doteq Q(f,f), \qquad Q^R(f) \doteq Q^R(f,f)$$

whenever appropriate.

We begin with a Lipschitz continuity type result for the collision operator $Q^R$.

**Lemma 15.** *Assume that $f$ satisfies the exponential decay*

$$|f(\boldsymbol{v})| \le C_{\text{exp},f} \exp(-a\|\boldsymbol{v}\|^2), \qquad \boldsymbol{v} \in \mathbb{R}^d,$$

*and that $g$ is $2L$-periodic, satisfying*

$$\|g\mathcal{M}_a^{-1}\|_{L^2(\mathcal{D}_L)} \le C_{\text{dec}} L^{\frac{d}{2}},$$

*as well as $\|g\|_{L^2(\mathcal{D}_L)} \le C_{\text{b}}$, where $C_{\text{exp},f}$, $C_{\text{dec}}$ and $C_{\text{b}}$ are constants that do not depend on $L$.*

*Then there exists constants $C_{\text{lip}}$ and $C_{\text{d}}$ that may depend on $\lambda$ and $d$, such that for sufficiently large $L$,*

$$\|Q^R(f,f) - Q^R(g,g)\|_{L^2(\mathcal{B}_R)} \le C_{\text{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} \left( \|f-g\|_{L^2(\mathcal{B}_R)} + C_{\text{d}} L^{\frac{d}{2}} e^{-aR^2} \right).$$

*Proof.* From the proof of theorem 6 we know that

$$\begin{aligned} \|Q^{R,-}(f,g)\|_{L^2(\mathcal{B}_R)} &\le C_A R^{\max(\lambda,0)} \|f\|_{L^2(\mathcal{B}_R)} \|g\|_{L^1(\mathcal{D}_L)} \\ &\le C_A R^{\max(\lambda,0)} (2L)^{\frac{d}{2}} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)}. \end{aligned}$$

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

$$\begin{aligned} \|Q^{R,+}(f,g)\|_{L^2(\mathcal{B}_R)} &\le C_{\text{pos}} \|f\|_{L^2_\lambda(\mathcal{D}_L)} \|g\|_{L^1_\lambda(\mathcal{D}_L)} \\ &\le 4C_{\text{pos}} (\sqrt{d}L)^{2\lambda} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^1(\mathcal{D}_L)} \\ &\le 4C_{\text{pos}} (\sqrt{d}L)^{2\lambda} (2L)^{\frac{d}{2}} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)} \end{aligned}$$

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

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

that

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

whenever $2L \ge 1$.

In summary, under the given assumptions,

$$\|Q^R(f,g)\|_{L^2(\mathcal{B}_R)} \leq C_{\mathrm{c}} L^{\max(2\lambda,0)+\frac{d}{2}} \|f\|_{L^2(\mathcal{D}_L)} \|g\|_{L^2(\mathcal{D}_L)}$$

with, say,

$$C_{\mathrm{c}} = (C_A + 4C_{\mathrm{pos}} + C_{\mathrm{neg}}) \cdot 2^{\frac{d}{2}} d^{\max(\lambda,0)}.$$

Also, the same bound will hold for the symmetrized operator

$$Q^R_{\mathrm{sym}}(f,g) = \frac{1}{2}\left(Q^R(f,g) + Q^R(g,f)\right).$$

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

$$\begin{aligned}\|Q^R(f,f) - Q^R(g,g)\|_{L^2(\mathcal{B}_R)} &= \|Q^R_{\mathrm{sym}}(f,f) - Q^R_{\mathrm{sym}}(g,g)\|_{L^2(\mathcal{B}_R)} \\ &= \|Q^R_{\mathrm{sym}}(f+g, f-g)\|_{L^2(\mathcal{B}_R)} \\ &\leq C_{\mathrm{c}} L^{\max(2\lambda,0)+\frac{d}{2}} \|f+g\|_{L^2(\mathcal{D}_L)} \|f-g\|_{L^2(\mathcal{D}_L)}.\end{aligned}$$

Since $f$ and $g$ are bounded, we can reduce this to

$$\begin{aligned}&\|Q^R(f,f) - Q^R(g,g)\|_{L^2(\mathcal{B}_R)} \leq \\ &\qquad C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} \left(\|f-g\|_{L^2(\mathcal{B}_R)} + \|f\|_{L^2(\mathcal{D}_L\setminus\mathcal{B}_R)} + \|g\|_{L^2(\mathcal{D}_L\setminus\mathcal{B}_R)}\right)\end{aligned}$$

Where

$$C_{\mathrm{lip}} = C_{\mathrm{c}}\left(C_{\exp,f}\left(\frac{\pi}{2a}\right)^{\frac{d}{4}} + C_{\mathrm{b}}\right).$$

From the exponential decay of $f$ we can conclude

$$\|f\|_{L^2(\mathcal{D}_L\setminus\mathcal{B}_R)} \leq C_{\exp,f} e^{-aR^2} (2L)^{\frac{d}{2}},$$

and likewise for $g$ we have

$$\begin{aligned}\|g\|_{L^2(\mathcal{D}_L\setminus\mathcal{B}_R)} &= \|g\mathcal{M}_a^{-1}\mathcal{M}_a\|_{L^2(\mathcal{D}_L\setminus\mathcal{B}_R)} \\ &\leq \|\mathcal{M}_a\|_{L^\infty(\mathcal{D}_L\setminus\mathcal{B}_R)} \|g\mathcal{M}_a^{-1}\|_{L^2(\mathcal{D}_L\setminus\mathcal{B}_R)} \leq C_{\mathrm{dec}} e^{-aR^2} L^{\frac{d}{2}}\end{aligned}$$

Thus,

$$\|Q^R(f,f) - Q^R(g,g)\|_{L^2(\mathcal{B}_R)} \leq C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} \left(\|f-g\|_{L^2(\mathcal{B}_R)} + C_{\mathrm{d}} L^{\frac{d}{2}} e^{-aR^2}\right).$$

with $C_{\mathrm{d}} = 2^{\frac{d}{2}} C_{\exp,f} + C_{\mathrm{dec}}$. □

We will also require the following generalization of the Grönwall inequality from [2], here stated in a more particular form.

**Theorem 16** (Theorem 21 in [2])**.** *Let $u(t)$ be a nonnegative function satisfying*

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

*where $a, b, c$ are nonnegative. Then*

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

### 4.3.1 Error due to projection

Denote by $e_{\mathcal{A}}(t) = f_R(t) - f_{\mathcal{A}}(t)$ the discretization error due to $\mathcal{A}$. Then

$$\dot{e}_{\mathcal{A}} = Q^R(f_R) - P_{\mathcal{A}} Q^R(f_{\mathcal{A}})$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|e_{\mathcal{A}}\|^2_{L^2(\mathcal{B}_R)} = (\dot{e}_{\mathcal{A}}, e_{\mathcal{A}})_{L^2(\mathcal{B}_R)}.$$

**Proposition 17.** *Assume that $f_R$ satisfies the exponential decay*

$$|f_R(\boldsymbol{v})| \leq C_{\exp,R} \exp(-a\|\boldsymbol{v}\|^2)$$

*and that $f_{\mathcal{A}}$ satisfies*

$$\|f_{\mathcal{A}}\mathcal{M}_{\mathrm{a}}^{-1}\|_{L^2(\mathcal{D}_L)} \leq C_{\mathrm{dec}} L^{\frac{d}{2}}$$

*as well as $\|f_{\mathcal{A}}\|_{L^2(\mathcal{D}_L)} \leq C_{\mathrm{b}}$, where $C_{\exp,f}$, $C_{\mathrm{dec}}$ and $C_{\mathrm{b}}$ are constants that do not depend on $L$ or $t$.*

*Then there exists constants $C_{\mathrm{lip}}$ and $C_{\mathrm{d}}$, that may depend on $\lambda$ and $d$ such that for sufficiently large $L$, for the error $e_{\mathcal{A}}$ induced by discretizing the velocity space, we have the bound*

$$\begin{aligned}\|e_{\mathcal{A}}(t)\|_{L^2(\mathcal{B}_R)} \leq{}& \left(\|e_{\mathcal{A}}(0)\|^2_{L^2(\mathcal{B}_R)} + \rho_{\mathcal{A}}(t)\right)^{\frac{1}{2}} \exp\left(C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) \\ &+ C_{\mathrm{d}} L^{\frac{d}{2}} \left(\exp\left(C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) - 1\right) \exp(-aR^2)\end{aligned}$$

*with*

$$\rho_{\mathcal{A}} = 2\int_0^T \|(\mathrm{Id} - P_{\mathcal{A}})Q^R(f_R(\tau))\|_{L^2(\mathcal{B}_R)} \|(\mathrm{Id} - P_{\mathcal{A}})f_R(\tau)\|_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau. \tag{25}$$

*Proof.* Integrating, we get

$$\begin{aligned}\frac{1}{2}\|e_{\mathcal{A}}(t)\|^2_{L^2(\mathcal{B}_R)} ={}& \frac{1}{2}\|e_{\mathcal{A}}(0)\|^2_{L^2(\mathcal{B}_R)} + \int_0^t \langle Q^R(f_R(\tau)) - P_{\mathcal{A}}Q^R(f_{\mathcal{A}}(\tau)), e_{\mathcal{A}}(\tau)\rangle_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau \\ ={}& \frac{1}{2}\|e_{\mathcal{A}}(0)\|^2_{L^2(\mathcal{B}_R)} + \int_0^t \langle (\mathrm{Id} - P_{\mathcal{A}})Q^R(f(\tau)), e_{\mathcal{A}}(\tau)\rangle_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau \\ &+ \int_0^t \langle P_{\mathcal{A}}\left[Q^R(f_R(\tau)) - Q^R(f_{\mathcal{A}}(\tau))\right], e_{\mathcal{A}}(\tau)\rangle_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau.\end{aligned}$$

Now define $\rho_{\mathcal{A}}$ as in (25), which is clearly an upper bound for the first integral term. Then, by lemma 15,

$$\begin{aligned}\|e_{\mathcal{A}}(t)\|^2_{L^2(\mathcal{B}_R)} ={}& \|e_{\mathcal{A}}(0)\|^2_{L^2(\mathcal{B}_R)} + \rho_{\mathcal{A}} + 2\int_0^t \langle Q^R(f_R(\tau)) - Q^R(f_{\mathcal{A}}(\tau)), P_{\mathcal{A}}e_{\mathcal{A}}(\tau)\rangle_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau \\ \leq{}& \|e_{\mathcal{A}}(0)\|^2_{L^2(\mathcal{B}_R)} + \rho_{\mathcal{A}} + 2C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} \\ & \int_0^t \left(\|e_{\mathcal{A}}(\tau)\|_{L^2(\mathcal{B}_R)} + C_{\mathrm{d}} L^{\frac{d}{2}} e^{-aR^2}\right) \|e_{\mathcal{A}}(\tau)\|_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau.\end{aligned}$$

Now, we can apply theorem 16, giving

$$\begin{aligned}\|e_{\mathcal{A}}(t)\|_{L^2(\mathcal{B}_R)} \leq{}& \left(\|e_{\mathcal{A}}(0)\|^2_{L^2(\mathcal{B}_R)} + \rho_{\mathcal{A}}\right)^{\frac{1}{2}} \exp\left(C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) \\ &+ C_{\mathrm{d}} L^{\frac{d}{2}} \left(\exp\left(C_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) - 1\right) \exp(-aR^2).\end{aligned} \tag{26}$$

□

Given a fixed $L$, the first term in (26) can be controlled by choosing $\mathcal{A}$ large enough. However, the second term can only be controlled in certain specific cases. Since we have $L \propto R$, this term will only approach zero as $L \to \infty$ if

$$\max(2\lambda, 0) + \frac{d}{2} < 2,$$

which means that for Maxwellian and soft potentials ($\lambda \leq 0$), the error can be controlled for $d \leq 3$, and for hard potentials ($\lambda > 0$), it can be controlled if $d < 4(1-\lambda)$.

### 4.3.2 Error due to truncation

We will now turn our attention to the error induced by truncating the collision operator, namely $e_R(t) = f(t) - f_R(t)$.

**Proposition 18.** *Assume that $f$ and $f_R$ satisfy the exponential decay conditions*

$$|f(\boldsymbol{v})| \leq C_{\mathrm{exp}} \exp(-a\|\boldsymbol{v}\|^2), \qquad |f_R(\boldsymbol{v})| \leq C_{\mathrm{exp},R} \exp(-a\|\boldsymbol{v}\|^2)$$

*where $C_{\mathrm{exp}}$ and $C_{\mathrm{exp},R}$ do not depend on $L$ or $t$.*

*Then there exists constants $C'_{\mathrm{lip}}$, $C'_{\mathrm{d}}$ and $C_{\mathrm{n}}$, that may depend on $\lambda$ and $d$, such that for sufficiently large $L$, for the error $e_R$ induced by truncating the collision operator, we have the bound*

$$\begin{aligned}\|e_R(t)\|_{L^2(\mathcal{B}_R)} \leq& \sqrt{\rho_R} \exp\left(C'_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) \\ &+ C'_{\mathrm{d}} L^{\frac{d}{2}} \left(\exp\left(C'_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) - 1\right) \exp(-aR^2). \qquad (27)\end{aligned}$$

*with*

$$\rho_R = 2C_{\mathrm{n}} \int_0^T \|(Q - Q^R)(f(\tau))\|_{L^2(\mathcal{B}_R)}.$$

*Proof.* As before, we have

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

noting that $e_R(0) = 0$.

Now note that

$$\int_0^t \langle (Q - Q^R)(f(\tau)), e_R(\tau)\rangle_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau \leq C_{\mathrm{n}} \int_0^T \|(Q - Q^R)(f(\tau))\|_{L^2(\mathcal{B}_R)} \,\mathrm{d}\tau,$$

where

$$C_{\mathrm{n}} = \sup_t \|f\|_{L^2(\mathcal{B}_R)} + \sup_t \|f_R\|_{L^2(B_R)} \leq (C_{\mathrm{exp}} + C_{\mathrm{exp},R}) \left(\frac{\pi}{2a}\right)^{\frac{d}{4}}.$$

exists by assumption. Thus define

$$\rho_R = 2C_{\text{n}} \int_0^T \|(Q - Q^R)(f(\tau))\|_{L^2(\mathcal{B}_R)}.$$

Then,

$$\begin{aligned}\|e_R(t)\|_{L^2(\mathcal{B}_R)} &\leq \rho_R + 2\int_0^t \langle Q^R(f(\tau)) - Q^R(f_R(\tau)), e_R(\tau)\rangle_{L^2(\mathcal{B}_R)}\,\mathrm{d}\tau \\ &\leq \rho_R + 2C'_{\text{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} \int_0^t \left(\|e_R(\tau)\|_{L^2(\mathcal{B}_R)} + C'_{\text{d}} L^{\frac{d}{2}} e^{-aR^2}\right) \|e_R(\tau)\|_{L^2(\mathcal{B}_R)}\,\mathrm{d}\tau\end{aligned}$$

by the proof of lemma 15, which is easily tweaked to allow for the case where both $f$ and $g$ are exponentially decaying. We get $C'_{\text{d}} = 2^{\frac{d}{2}}(C_{\exp} + C_{\exp,R})$ and

$$C'_{\text{lip}} = C_{\text{c}} \left(\frac{\pi}{2a}\right)^{\frac{d}{4}} (C_{\exp} + C_{\exp,R}).$$

As in proposition 17, we use theorem 16 to complete the proof. □

The quantity $\rho_R$ will decrease exponentially as $e^{-(2-\phi)aR^2}$, where $\phi$ is the golden ratio, which means that both terms can be controlled, and will converge to zero as $L$ grows, under the same condition on $d$ and $\lambda$ as before. This is shown in proposition 19.

**Proposition 19.** *Assume that $f$ satisfies the exponential decay condition*

$$|f(\boldsymbol{v})| \leq C_{\exp} \exp(-a\|\boldsymbol{v}\|^2)$$

*where $C_{\exp}$ does not depend on $L$.*

*Then,*

$$\|(Q - Q^R)(f)\|_{L^2(\Omega)} \leq C_{\text{tr}} e^{-2a(2-\phi)R^2}$$

*where $\phi$ is the golden ratio, and the quantity $C_{\text{tr}}$ may depend on the domain $\Omega \subseteq \mathbb{R}^d$, the decay rates of $f$, as well as $\lambda$ and $d$.*

*Proof.* First, note that, by lemma 20, 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,

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

and

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

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

Then, using the Carleman representation for $Q^R$ we get

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

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 $\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2 \geq 4R^2$, and defining

$$C_{\mathrm{tr}} = 2C_{\exp}^2 \int_\Omega e^{-a(2-\phi)^2\|\boldsymbol{v}\|^2}\,\mathrm{d}\boldsymbol{v} \int_S \|\boldsymbol{x}+\boldsymbol{y}\|^{\lambda+2-d}\delta(\boldsymbol{x}\cdot\boldsymbol{y})e^{-\frac{a}{2}(2-\phi)\|\boldsymbol{x}+\boldsymbol{y}\|^2}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.$$

□

**Lemma 20.** *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 $\phi$ 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 $\phi^{-1}=\phi-1$. □

### 4.3.3 Summary

The following corollary summarizes everything so far.

**Corollary 21.** *Assume that $f$ and $f_R$ satisfy the exponential decay conditions*

$$|f(\boldsymbol{v})| \leq C_{\exp} \exp(-a\|\boldsymbol{v}\|^2), \qquad |f_R(\boldsymbol{v})| \leq C_{\exp,R} \exp(-a\|\boldsymbol{v}\|^2)$$

*and that $f_{\mathcal{A}}(L)$ satisfies*

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

*as well as $\|f_{\mathcal{A}(L)}\|_{L^2(\mathcal{D}_L)} \leq C_{\text{b}}$, where $C_{\exp,R}$, $C_{\exp}$, $C_{\text{dec}}$ and $C_{\text{b}}$ are constants that do not depend on $L$ or $t$.*

*Then, for the error $e$ induced by truncation and projection, there exists constants $C^*_{\text{tr}}$, $C^*_{\text{lip}}$ and $C^*_{\text{d}}$, that may depend on $\lambda$ and $d$, such that for sufficiently large $L$ we have the estimate*

$$\begin{aligned}\|e(t)\|_{L^2(\mathcal{B}_R)} \leq & \left[\left(\|e(0)\|^2_{L^2(\mathcal{B}_R)} + \rho_{\mathcal{A}(L)}\right)^{\frac{1}{2}} + C^*_{\text{tr}} e^{-a(2-\phi)R^2}\right] \exp\left(C^*_{\text{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) \\ & + C^*_{\text{d}} L^{\frac{d}{2}} \left(\exp\left(C^*_{\text{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) - 1\right) \exp(-aR^2). \qquad (28)\end{aligned}$$

*Proof.* We merely combine the error estimates from propositions 17, 18 and 19, and write

$$C^*_{\text{lip}} = \max(C_{\text{lip}}, C'_{\text{lip}}), \qquad C^*_{\text{d}} = C_{\text{d}} + C'_{\text{d}} \qquad C^*_{\text{tr}} = \sqrt{2C_{\text{n}}C_{\text{tr}}T}.$$

□

The only remaining estimates to complete are those of $\|e(0)\|_{L^2(\mathcal{B}_R)}$ and $\rho_{\mathcal{A}(L)}$, which depend on the approximation power of $\mathcal{A}$.

#### 4.3.4 Approximation using $\mathcal{A}_0(N)$

We present here the error estimates for the classical hyperbolic cross $\mathcal{A}_0(N)$. The case for general $T$ is not much different.

We note first that we have, for general $f \in \mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)$ we have the result from theorem 11 that

$$\|f - P_{\mathcal{A}_0(N)} f\|_{L^2(\mathcal{D}_L)} \leq C_{\text{appr}} (1+N)^{-l-t} \|f\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)}. \qquad (29)$$

Also, an easy generalisation of theorem 12 to the case where $f$ and $g$ are non-periodic (using the estimates derived in the proof of proposition 17), shows that

$$\|Q^R(f,g)\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{B}_R)} \leq C_{\text{mix}} L^{\max(2\lambda,0)+\frac{d}{2}} \|f\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)} \|g\|_{\mathcal{H}^{t,l}_{\text{mix}}(\mathcal{D}_L)}. \qquad (30)$$

This gives us the following theorem.

**Theorem 22.** *Assume that $f$ and $f_R$ satisfy the exponential decay conditions*

$$|f(\boldsymbol{v})| \leq C_{\exp} \exp(-a\|\boldsymbol{v}\|^2), \qquad |f_R(\boldsymbol{v})| \leq C_{\exp,R} \exp(-a\|\boldsymbol{v}\|^2)$$

*and that $f_{\mathcal{A}}$ satisfies*

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

*as well as $\|f_{\mathcal{A}}\|_{L^2(\mathcal{D}_L)} \leq C_{\text{b}}$, where $C_{\exp,R}$, $C_{\exp}$, $C_{\mathcal{A}}$ and $C_{\text{b}}$ are constants that do not depend on $L$ or $t$.*

*Also assume that* $f_0$ *and* $f_R$ *is in* $\mathcal{H}^{t,l}_{\mathrm{mix}}$ *for some* $t, l$ *and that the mapping* $\tau \mapsto \|f_R(\tau)\|^3_{\mathcal{H}^{t,l}_{\mathrm{mix}}}$ *is integrable with respect to* $\tau$*. Then there is a function* $\Gamma(\lambda, d, L)$ *so that for sufficiently large* $L$ *it holds*

$$\begin{aligned}\|e(t)\|_{L^2(\mathcal{B}_R)} \leq \Big[C_{\mathrm{appr}}(1+N)^{-l-t}\Gamma(\lambda, d, L) + C^*_{\mathrm{tr}} e^{-a(2-\phi)R^2}\Big] \exp\left(C^*_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) \\ + C^*_{\mathrm{d}} L^{\frac{d}{2}} \left(\exp\left(C^*_{\mathrm{lip}} L^{\max(2\lambda,0)+\frac{d}{2}} t\right) - 1\right) \exp(-aR^2). \qquad (31)\end{aligned}$$

*Proof.* This is achieved by substituting equations (29) and (30) in the definition of $\rho_{\mathcal{A}}$ (25), extending the domain of norms from $\mathcal{B}_R$ to $\mathcal{D}_L$ where necessary.

The function $\Gamma$ will take the form

$$\Gamma(\lambda, d, L)^2 = 2C_{\mathrm{mix}} L^{\max(2\lambda,0)+\frac{d}{2}} \int_0^T \|f_R(\tau)\|^3_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)} \,\mathrm{d}\tau + \|f_0\|^2_{\mathcal{H}^{t,l}_{\mathrm{mix}}(\mathcal{D}_L)}.$$

□

We can see that given the aforementioned condition on $\lambda$ and $d$, all these terms can be controlled, first by choosing $L$ and then by choosing $N(L)$, through which the dependence $\mathcal{A} = \mathcal{A}(L)$ is realized.

It should be noted that one should not expect this estimate 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

Since the hyperbolic cross performs so poorly with near-equilibrium solutions, a more flexible idea may be to consider $f$ as a perturbation from equilibrium

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

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

$$\partial_t f = \partial_t f^p = Q(f, f) = Q(f^p, f^p) + Q(M, f^p) + Q(f^p, M) \qquad (32)$$

as $Q(M, M) = 0$ by necessity. Moreover, the terms $Q(M, f^p)$ and $Q(f^p, M)$ represent a linear function of $f^p$ which can be assembled *a priori* using the spectrum of $M$ 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 (11) is

$$\hat{M}(\rho, \boldsymbol{u}, T)(\boldsymbol{k}) = \frac{\rho}{(2L)^d} \exp\left(-\frac{T}{2}\|\boldsymbol{k}\|^2 - i\boldsymbol{u} \cdot \boldsymbol{k}\right). \qquad (33)$$

Knowing this, we can formulate the linear part of the right hand side of (32) as

$$\begin{aligned}
\left[Q(f^p, M) + Q(M, f^p)\right](\boldsymbol{k}) &= \sum_{\substack{\boldsymbol{l}\in\mathcal{A}\\ \boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \hat{\beta}(\boldsymbol{l},\boldsymbol{m})\hat{M}(\boldsymbol{m}) f^p_{\boldsymbol{l}} + \sum_{\substack{\boldsymbol{l}\in\mathcal{A}\\ \boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \hat{\beta}(\boldsymbol{m},\boldsymbol{l})\hat{M}(\boldsymbol{m}) f^p_{\boldsymbol{l}} \\
&= \sum_{\boldsymbol{l}\in\mathcal{A}} \sum_{\substack{\boldsymbol{m}\in\mathcal{B}\\ \boldsymbol{l}+\boldsymbol{m}=\boldsymbol{k}}} \left(\hat{\beta}(\boldsymbol{l},\boldsymbol{m}) + \hat{\beta}(\boldsymbol{m},\boldsymbol{l})\right)\hat{M}(\boldsymbol{m}) f^p_{\boldsymbol{l}} \\
&= \sum_{\boldsymbol{l}\in\mathcal{A}} Q_{\text{lin}}(\boldsymbol{k},\boldsymbol{l}) f^p_{\boldsymbol{l}} \qquad (34)
\end{aligned}$$

which can be recognized as simple matrix multiplication with

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

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 (33), $|\hat{M}|$ is a Gaussian centered at zero, so a suitable choice might be $\mathcal{B} = \mathcal{A}_{\text{FF}}(N)$ with a sufficiently large $N$. It is worth noting that $\mathcal{B}$ can be very large, since as soon as $Q_{\text{lin}}$ is assembled, the size of $\mathcal{B}$ does not affect the cost of applying (34).

To ensure

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

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 have numerical results for various solutions, some of which are "friendly" to the hyperbolic cross, others which are not. To summarize, the three methods we have used are

- FF: The full grid fourier approximation.
- HC: The "raw" hyperbolic cross method.
- OM: The offset method with hyperbolic cross.

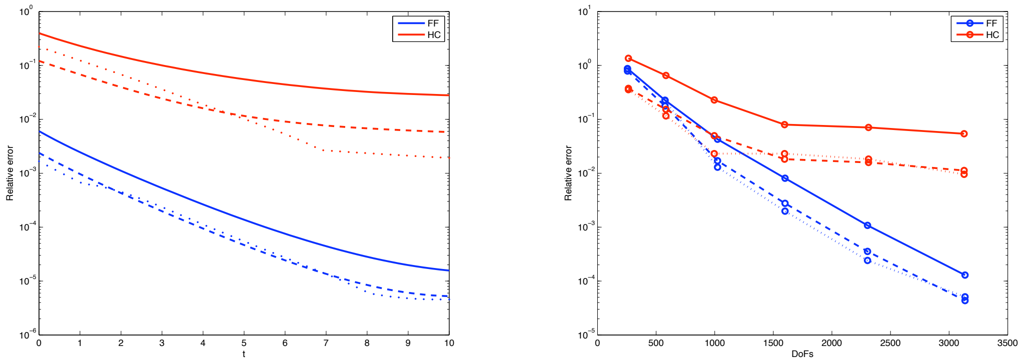
## 6.1 Verification (BKW)

As a verification of correctness, one can use the only known analytical non-equilibrium solution to the Boltzmann equation—the BKW solution [10], [3]. It takes the form

$$f(t,v) = (2\pi s)^{-d/2} \exp\left(-\frac{\|\boldsymbol{v}\|^2}{2s}\right)\left(1 - \frac{1-s}{2s}\left(d - \frac{\|\boldsymbol{v}\|^2}{s}\right)\right) \qquad (35)$$

where

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



(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. Full lines are $L^1$-norm, dashed lines are $L^2$-norm, and dotted lines are $L^\infty$-norm errors. Blue lines represent the full grid $\mathcal{A}_{\text{FF}}$ and red lines represent $\mathcal{A}_{\text{HC}}$. For timestepping, we used a fixed-timestep explicit 4th order Runge-Kutta method with timestep $10^{-2}$.

and $B = \text{const.}$ (also called the *Maxwellian* kernel). Here, $\lambda$ 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 $f(0, \boldsymbol{v}) \geq 0$ everywhere. We will use a $t_0$ determined by $s(0) = \frac{1}{2}$, which gives the initial distribution

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

It can further be scaled using proposition 2 to ensure that it meets the conditions of proposition 3 to a sufficient degree. We use $L = 3\pi$ and a scaling in $\boldsymbol{v}$ with $\sigma = 5$. 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 arising from the very narrow support of $f_0$, as well as the poor performance of the hyperbolic cross, arising from 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,

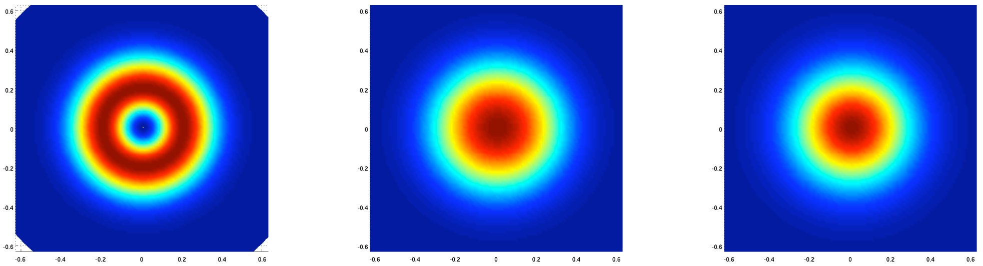
$$-\int_{\mathcal{D}_L} f(\boldsymbol{v}) \log f(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}$$

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

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



(a) Initial distribution.

(b) Solution at time $t = 5.1$.

(c) Equilibrium.

Figure 4: The BKW solution.

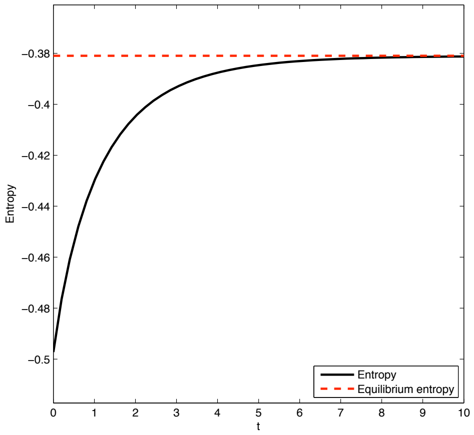


Figure 5: Convergence to entropy for $\mathcal{A}_{\mathrm{FF}}(56)$.

There parameter $s > 0$ can be tweaked to make $f_0$ more or less "hyperbolic". 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 4th 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 hyperbolic cross with $N = 200$, with 1928 degrees of freedom. 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 "hyperbolicity", and $\epsilon > 0$ represents the fact that $f_0$ is a minor perturbation from equilibrium.[1]

Using $s = 7$ and $\epsilon = 10^{-2}$, we have plotted $\|f_p\|$ 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 behaviour in accordance with [5].

### 6.4 Exponential decay of the numerical solution

A crucial assumption was made in section 4.3 that allowed us to produce theorem 22, namely that $f_{\mathcal{A}}$ satisfies the decay property

$$\|f_{\mathcal{A}(L)} M_a^{-1}\|_{L^2(\mathcal{D}_L)} \leq C_{\text{dec}} L^{\frac{d}{2}}.$$

We will here provide some numerical evidence that this can be expected to hold for both $\mathcal{A}_{\text{FF}}(N)$ and $\mathcal{A}_0(N)$ with some dependence $N = N(L)$.

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

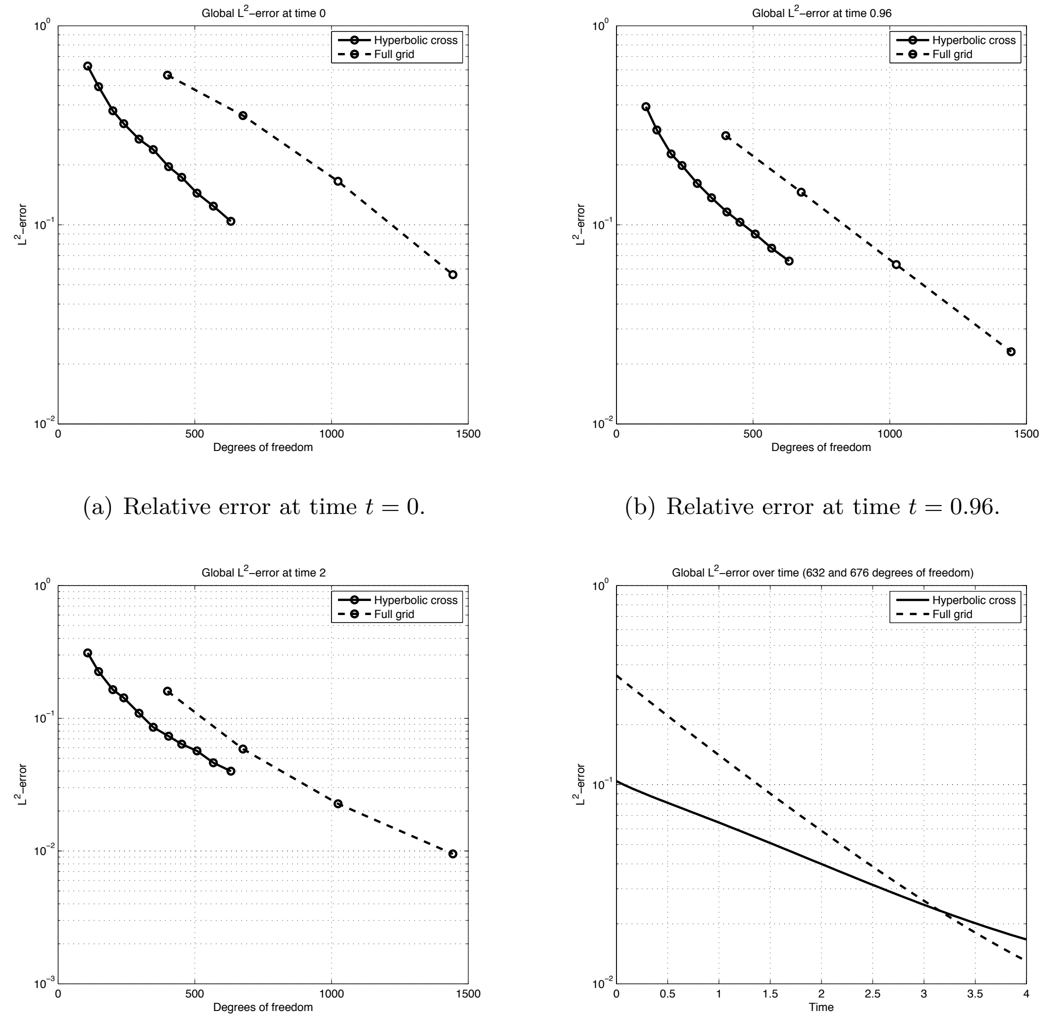
A test with larger $L$ would be difficult to perform with the current implementation, as the exponential weight at the boundaries of $\mathcal{D}_L$ is much larger than machine precision, artificially inflating the norms.

## A Kernel modes for the VHS model

We proceed to develop expressions for the kernel modes in the particular case of the VHS (Variable Hard Sphere) collision model

$$B(\|\boldsymbol{g}\|, \cos\theta) = C_\alpha \|\boldsymbol{g}\|^\alpha,$$

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



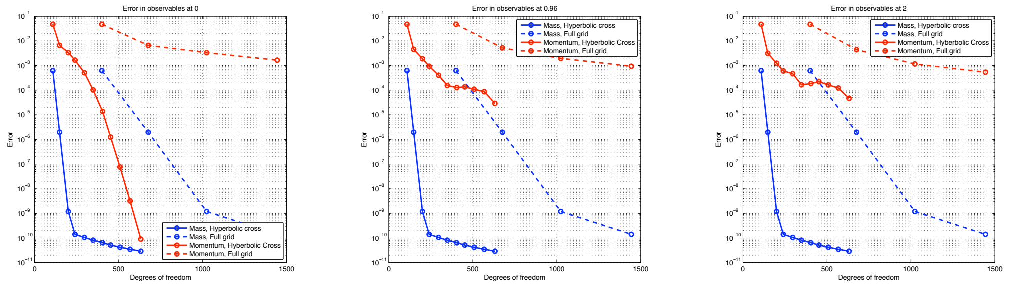
(a) Relative error at time $t = 0$.

(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.

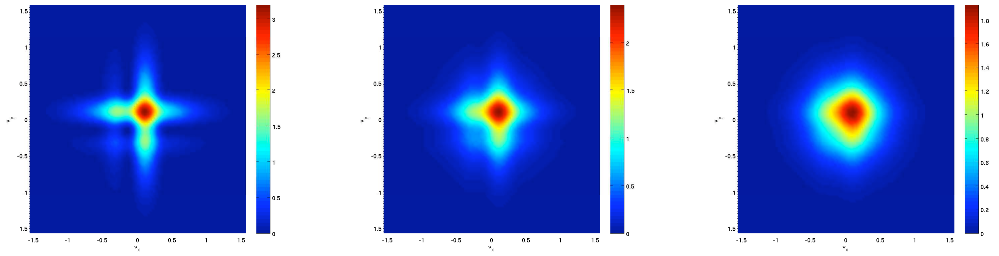


(a) Error in observables at time $t = 0$.

(b) Error in observables at time $t = 0.96$.

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

Figure 7: Errors in obserables for section 6.2.



(a) Initial distribution.

(b) Solution at time $t = 0.96$.

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

Figure 8: The crossed beams solution.

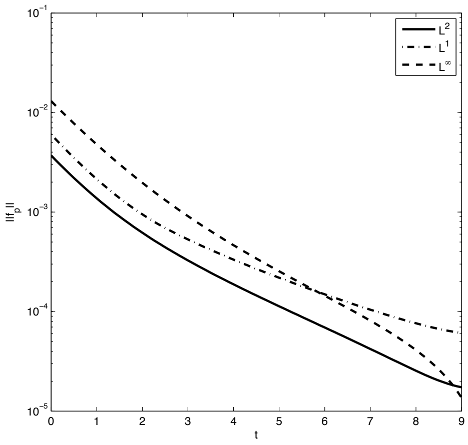


Figure 9: Relaxation to equilibrium: the norm of the perturbation versus time.

| $N$ | $L=0.5\pi$ | $L=0.75\pi$ | $L=\pi$ |
|---|---|---|---|
| 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 | $5.25\cdot10^4$ |
| 70 | 1.25 | 1.25 | $5.67\cdot10^3$ |
| 90 | 1.25 | 1.25 | $1.15\cdot10^3$ |
| 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 $\sup_t \|f_{\mathcal{A}} M_a^{-1}\|_{L^2(\mathcal{D}_L)}$ 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 $T = 10$.

for $\alpha \leq 1$. The case $\alpha = 1$ is the hard sphere case, and $\alpha = 0$ is the Maxwellian molecule case (where the collision kernel is constant).

As in [8], we have for $\beta_d^P$,

$$\beta_d^P(\boldsymbol{l},\boldsymbol{m}) = C_\alpha \int_{\mathcal{B}_{2R}} \|\boldsymbol{g}\|^\alpha \exp\left[-i\boldsymbol{g}\cdot\frac{\boldsymbol{l}+\boldsymbol{m}}{2}\right] I_d(\|\boldsymbol{g}\|,\boldsymbol{l}-\boldsymbol{m})\,\mathrm{d}\boldsymbol{g},$$

where $I_d^P(\|\boldsymbol{g}\|,\boldsymbol{l}-\boldsymbol{m})$ is the integral

$$I_d^P(\|\boldsymbol{g}\|,\boldsymbol{l}-\boldsymbol{m}) = \int_{S^{d-1}} \exp\left[i\|\boldsymbol{g}\|\boldsymbol{\sigma}\cdot\frac{\boldsymbol{l}-\boldsymbol{m}}{2}\right]\,\mathrm{d}\boldsymbol{\sigma}.$$

## A.1 Two dimensions

For $\beta_2^P$, we first find that

$$\begin{aligned} I_1^P(\|\boldsymbol{g}\|,\boldsymbol{l}-\boldsymbol{m}) &= \int_{S^1} \exp\left[i\|\boldsymbol{g}\|\boldsymbol{\sigma}\cdot\frac{\boldsymbol{l}-\boldsymbol{m}}{2}\right]\,\mathrm{d}\boldsymbol{\sigma} \\ &= \int_0^{2\pi} \exp\left[i\frac{\|\boldsymbol{g}\|\|\boldsymbol{l}-\boldsymbol{m}\|}{2}\cos\theta\right]\,\mathrm{d}\theta = 2\pi J_0(\|\boldsymbol{g}\|\|\boldsymbol{l}-\boldsymbol{m}\|/2), \end{aligned}$$

where $J_0$ is a Bessel function of the first kind. Continuing,

$$\begin{aligned} \beta_2^P(\boldsymbol{l},\boldsymbol{m}) &= 2\pi C_\alpha \int_{\mathcal{B}_{2R}} \|\boldsymbol{g}\|^\alpha \exp\left[-i\boldsymbol{g}\cdot\frac{\boldsymbol{l}+\boldsymbol{m}}{2}\right] J_0(\|\boldsymbol{g}\|\|\boldsymbol{l}-\boldsymbol{m}\|/2)\,\mathrm{d}\boldsymbol{g} \\ &= 2\pi C_\alpha \int_0^{2R} \rho^{1+\alpha} \int_0^{2\pi} \exp\left[-i\rho\hat{\boldsymbol{e}}_\theta\cdot\frac{\boldsymbol{l}+\boldsymbol{m}}{2}\right]\,\mathrm{d}\theta J_0(\|\boldsymbol{l}-\boldsymbol{m}\|\rho/2)\,\mathrm{d}\rho \\ &= 4\pi^2 C_\alpha \int_0^{2R} \rho^{1+\alpha} J_0(\|\boldsymbol{l}+\boldsymbol{m}\|\rho/2) J_0(\|\boldsymbol{l}-\boldsymbol{m}\|\rho/2)\,\mathrm{d}\rho \\ &= P_2(R,\alpha)\int_0^1 r^{1+\alpha} J_0(\|\boldsymbol{l}+\boldsymbol{m}\|Rr) J_0(\|\boldsymbol{l}-\boldsymbol{m}\|Rr)\,\mathrm{d}r, \qquad (36) \end{aligned}$$

where $\hat{\boldsymbol{e}}_\theta$ is the unit vector in the direction $\theta$, and $P_2(R,\alpha) = 4\pi^2(2R)^{2+\alpha}C_\alpha$.

For $\beta_2^M$, we first find the transformed collision kernel to be

$$\tilde{B}(\boldsymbol{x},\boldsymbol{y}) = 2^{d-1}C_\alpha \left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2\right)^{1+(\alpha-d)/2}.$$

And so

$$\begin{aligned}
\beta_2^M(\boldsymbol{l},\boldsymbol{m}) &= \int_0^{\sqrt{2}R}\int_0^{\sqrt{2}R} \tilde{B}(\rho,\rho')\rho\rho' \int_0^{2\pi} e^{i\rho\boldsymbol{l}\cdot\hat{\boldsymbol{e}}_\phi} \int_0^{2\pi} \delta(\hat{\boldsymbol{e}}_\phi\cdot\hat{\boldsymbol{e}}_\theta) e^{i\rho'\boldsymbol{m}\cdot\hat{\boldsymbol{e}}_\theta}\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}\rho\,\mathrm{d}\rho' \\
&= \int_0^{\sqrt{2}R}\int_0^{\sqrt{2}R} \tilde{B}(\rho,\rho')\rho\rho' \int_0^{2\pi} e^{i\rho\boldsymbol{l}\cdot\hat{\boldsymbol{e}}_\phi} \left(e^{i\rho'\boldsymbol{m}\cdot\hat{\boldsymbol{e}}_\phi^\perp} + e^{-i\rho'\boldsymbol{m}\cdot\hat{\boldsymbol{e}}_\phi^\perp}\right)\mathrm{d}\phi\,\mathrm{d}\rho\,\mathrm{d}\rho' \\
&= \int_0^{\sqrt{2}R}\int_0^{\sqrt{2}R} \tilde{B}(\rho,\rho')\rho\rho' \int_0^{2\pi} \left(e^{i(\rho\boldsymbol{l}+\rho'\boldsymbol{m}^\perp)\cdot\hat{\boldsymbol{e}}_\phi} + e^{i(\rho\boldsymbol{l}-\rho'\boldsymbol{m}^\perp)\cdot\hat{\boldsymbol{e}}_\phi}\right)\mathrm{d}\phi\,\mathrm{d}\rho\,\mathrm{d}\rho' \\
&= 2\pi C_\alpha \int_0^{\sqrt{2}R}\int_0^{\sqrt{2}R} \left(\rho^2+\rho'^2\right)^{\alpha/2}\rho\rho' \\
&\qquad \left(J_0(\|\rho\boldsymbol{l}+\rho'\boldsymbol{m}^\perp\|) + J_0(\|\rho\boldsymbol{l}-\rho'\boldsymbol{m}^\perp\|)\right)\mathrm{d}\rho\,\mathrm{d}\rho' \\
&= M_2(R,\alpha)\int_0^1\int_0^1 (r^2+r'^2)^{\alpha/2} rr' \\
&\qquad \left(J_0\left(\sqrt{2}R\|r\boldsymbol{l}+r'\boldsymbol{m}^\perp\|\right) + J_0\left(\sqrt{2}M\|r\boldsymbol{l}-r'\boldsymbol{m}^\perp\|\right)\right)\mathrm{d}r\,\mathrm{d}r',
\end{aligned}$$

where $\boldsymbol{m}^\perp$ is $\boldsymbol{m}$ rotated through $\pi/2$, and $M_2(R,\alpha) = 4\pi(2R^2)^{2+\alpha/2}C_\alpha$.

## A.2 Three dimensions

For $\beta_3^P$, we can use a know identity involving the sinc function, namely

$$\int_{S^2} e^{i\boldsymbol{q}\cdot\boldsymbol{\sigma}}\,\mathrm{d}\boldsymbol{\sigma} = 2\pi\int_0^\pi e^{i\|\boldsymbol{q}\|\cos\theta}\sin\theta\,\mathrm{d}\theta = 4\pi\,\mathrm{sinc}(\|\boldsymbol{q}\|),$$

which can be shown choosing a spherical coordinate system centered around $\boldsymbol{q}$, to get

$$I_3^P(\|\boldsymbol{g}\|,\boldsymbol{l}-\boldsymbol{m}) = 4\pi\,\mathrm{sinc}(\|\boldsymbol{g}\|\|\boldsymbol{l}-\boldsymbol{m}\|/2).$$

To finish,

$$\begin{aligned}
\beta_3^P(\boldsymbol{l},\boldsymbol{m}) &= 4\pi C_\alpha \int_{\mathcal{B}_{2R}} \|\boldsymbol{g}\|^\alpha \exp\left[-i\boldsymbol{g}\cdot\frac{\boldsymbol{l}+\boldsymbol{m}}{2}\right]\mathrm{sinc}(\|\boldsymbol{g}\|\|\boldsymbol{l}-\boldsymbol{m}\|/2)\,\mathrm{d}\boldsymbol{g} \\
&= 4\pi^2 C_\alpha \int_0^{2R} \rho^{2+\alpha}\,\mathrm{sinc}(\|\boldsymbol{l}-\boldsymbol{m}\|\rho/2)\int_0^\pi \exp\left[-i\|\boldsymbol{l}+\boldsymbol{m}\|\rho\cos\theta/2\right]\sin\theta \\
&= 16\pi^2 C_\alpha \int_0^{2R} \rho^{2+\alpha}\,\mathrm{sinc}(\|\boldsymbol{l}-\boldsymbol{m}\|\rho/2)\,\mathrm{sinc}(\|\boldsymbol{l}+\boldsymbol{m}\|\rho/2)\,\mathrm{d}\rho \\
&= P_3(R,\alpha)\int_0^1 r^{2+\alpha}\,\mathrm{sinc}(\|\boldsymbol{l}+\boldsymbol{m}\|Rr)\,\mathrm{sinc}(\|\boldsymbol{l}-\boldsymbol{m}\|Rr)\,\mathrm{d}r
\end{aligned}$$

where $P_3(R,\alpha) = 16\pi^2(2R)^{3+\alpha}C_\alpha$.

It is worth mentioning that the integral

$$F_\alpha(\xi,\eta) = \int_0^1 r^{2+\alpha}\,\mathrm{sinc}(\xi r)\,\mathrm{sinc}(\eta r)\,\mathrm{d}r$$

has closed form expressions for integral $\alpha$, and, as given in [8], for Maxwellian molecules and hard spheres, we have

$$\begin{aligned} F_0(\xi,\eta) &= \frac{1}{2\xi\eta pq}(p\sin q - q\sin p), \\ F_1(\xi,\eta) &= \frac{1}{2\xi\eta p^2q^2}(p^2(q\sin q+\cos q) - q^2(p\sin p+\cos p) - 4\xi\eta), \end{aligned}$$

for $p = \xi+\eta$, $q = \xi-\eta$.

For $\beta_3^M$, we get

$$\beta_3^M(\boldsymbol{l},\boldsymbol{m}) = \int_0^{\sqrt{2}R}\int_0^{\sqrt{2}R} B(\rho,\rho')(\rho\rho')^2 \int_{S^2} e^{i\rho\boldsymbol{l}\cdot\boldsymbol{\sigma}} \int_{S^1_\perp(\boldsymbol{\sigma})} e^{i\rho'\boldsymbol{m}\cdot\boldsymbol{\sigma}'}\,\mathrm{d}\boldsymbol{\sigma}'\,\mathrm{d}\boldsymbol{\sigma}\,\mathrm{d}\rho\,\mathrm{d}\rho',$$

where $S^1_\perp(\boldsymbol{\sigma})$ is the unit circle in three dimensions orthogonal to $\boldsymbol{\sigma}$. For $\boldsymbol{\sigma}' \in S^1_\perp(\boldsymbol{\sigma})$, we have that $\boldsymbol{\sigma}'\cdot\boldsymbol{m} = \boldsymbol{\sigma}'\cdot P_{\boldsymbol{\sigma}}\boldsymbol{m}$, where $P_{\boldsymbol{\sigma}}$ is the orthogonal projection onto the subspace $\{\boldsymbol{x}\in\mathbb{R}^3 \,|\, \boldsymbol{x}\perp\boldsymbol{\sigma}\}$. Thus

$$\int_{S^1_\perp(\boldsymbol{\sigma})} e^{i\rho'\boldsymbol{m}\cdot\boldsymbol{\sigma}'}\,\mathrm{d}\boldsymbol{\sigma}' = 2\pi J_0(\rho'\|P_{\boldsymbol{\sigma}}\boldsymbol{m}\|),$$

and the objective is to resolve

$$\int_{S^2} e^{i\rho\boldsymbol{l}\cdot\boldsymbol{\sigma}} J_0(\rho'\|\boldsymbol{m}\||\sin q(\boldsymbol{\sigma})|)\,\mathrm{d}\boldsymbol{\sigma}$$

where $q(\boldsymbol{\sigma})$ is the angle between $\boldsymbol{m}$ and $\boldsymbol{\sigma}$, which can be evaluated with quadrature.

### A.3 Evaluation

For our experiments, which run exclusively in two dimensions, we have been using the $\beta_2^P$ kernel modes, since these are given as an integral of lower dimension.

The evaluation of (36) has been done using "overkill" Gauss-Legendre quadrature on $[0,1]$. Since the integrals are, or can be, highly oscillatory, it would have been desirable to have developed a more efficient quadrature. Even so, we have experienced that quadrature is not the bottleneck of this method.

The method we have used is to successively increase the number of quadrature points by 25 until the absolute *and* relative errors between two successive approximations are smaller than $10^{-4}$. Usually this can be accomplished with less than 200 points for most $l$ and $m$.

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

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

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

[5] 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.

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

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

[8] 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.

[9] 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.

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

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

# Research Reports

| No. | Authors/Title |
|---|---|
| 12-28 | *E. Fonn, Ph. Grohs and R. Hiptmair*<br>Hyperbolic cross approximation for the spatially homogeneous Boltzmann equation |
| 12-27 | *P. Grohs*<br>Wolfowitz's theorem and consensus algorithms in Hadamard spaces |
| 12-26 | *H. Heumann and R. Hiptmair*<br>Stabilized Galerkin methods for magnetic advection |
| 12-25 | *F.Y. Kuo, Ch. Schwab and I.H. Sloan*<br>Multi-level quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients |
| 12-24 | *St. Pauli, P. Arbenz and Ch. Schwab*<br>Intrinsic fault tolerance of multi level Monte Carlo methods |
| 12-23 | *V.H. Hoang, Ch. Schwab and A.M. Stuart*<br>Sparse MCMC gpc Finite Element Methods for Bayesian Inverse Problems |
| 12-22 | *A. Chkifa, A. Cohen and Ch. Schwab*<br>High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs |
| 12-21 | *V. Nistor and Ch. Schwab*<br>High order Galerkin approximations for parametric second order elliptic partial differential equations |
| 12-20 | *X. Claeys, R. Hiptmair, and C. Jerez-Hanckes*<br>Multi-trace boundary integral equations |
| 12-19 | *Šukys, Ch. Schwab and S. Mishra*<br>Multi-level Monte Carlo finite difference and finite volume methods for stochastic linear hyperbolic systems |
| 12-18 | *Ch. Schwab*<br>QMC Galerkin discretization of parametric operator equations |
| 12-17 | *N.H. Risebro, Ch. Schwab and F. Weber*<br>Multilevel Monte-Carlo front tracking for random scalar conservation laws |
| 12-16 | *R. Andreev and Ch. Tobler*<br>Multilevel preconditioning and low rank tensor iteration for space-time simultaneous discretizations of parabolic PDEs |
| 12-15 | *V. Gradinaru and G.A. Hagedorn*<br>A timesplitting for the semiclassical Schrödinger equation |