Springer INdAM Series 48

Francesco Salvarani *Editor*

Recent Advances in Kinetic Equations and Applications

Springer

Springer INdAM Series

Volume 48

This series will publish textbooks, multi-authors books, thesis and monographs in English language resulting from workshops, conferences, courses, schools, seminars, doctoral thesis, and research activities carried out at INDAM - Istituto Nazionale di Alta Matematica, http://www.altamatematica.it/en. The books in the series will discuss recent results and analyze new trends in mathematics and its applications.
THE SERIES IS INDEXED IN SCOPUS

More information about this series at http://www.springer.com/series/10283

Francesco Salvarani
Editor

Recent Advances in Kinetic Equations and Applications

Springer

Editor
Francesco Salvarani
De Vinci Pôle Universitaire
Research Center
Courbevoie, France

Department of Mathematics "F. Casorati"
University of Pavia
Pavia, Italy

ISSN 2281-518X ISSN 2281-5198 (electronic)
Springer INdAM Series
ISBN 978-3-030-82945-2 ISBN 978-3-030-82946-9 (eBook)
https://doi.org/10.1007/978-3-030-82946-9

Mathematics Subject Classification: Primary: 82C40; Secondary: 76P05, 35H10, 35K65, 35P15, 35Q84

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Kinetic theory is a rapidly developing field of research, not only because of its intrinsic theoretical interests, but also because of its applications to many problems in science and technology.

The main interest of kinetic models is the possibility of studying collective phenomena for systems composed of a large number of elementary particles (for example, gas molecules or dust particles) by going beyond the microscopic approach. The description of these systems is based on one or more distribution functions (or number densities) in phase space (or in extended phase spaces).

In the case of collisional models, the effect of binary interactions between particles is treated from a statistical point of view. However, each such interaction is viewed as a simple collision (or scattering) event, and its description at microscopic scale involves only a few fundamental laws (such as the conservation of momentum, or of kinetic energy in the case of elastic collisions). For this reason, kinetic models depend much less on phenomenological laws than most models of continuum mechanics. In particular, the main macroscopic collective features of the system can be rigorously deduced by means of an appropriate limiting procedure on an ab initio kinetic model, rather than heuristically introduced in the macroscopic model.

Another important class of kinetic equations does not model binary collisions, but rather describes the interaction between particles through the mean field generated by the whole system and is well adapted for handling long-range interactions.

Because of the quality of research in kinetic theory and its future developments, the National Italian Institute of Higher Mathematics (INdAM) funded the workshop "Recent Advances in Kinetic Equations and Applications", which was held in Rome (Italy), from 11 to 15 November 2019. This volume collects several original contributions written by invited speakers at the workshop.

Anton Arnold, Jean Dolbeault, Christian Schmeiser and Tobias Wöhrer discuss, in their chapter, two L^2 hypocoercivity methods based on Fourier decomposition and mode-by-mode estimates for dynamical systems. In particular, their theory allows to study situations involving both a degenerate dissipative operator and a conservative operator, provided that their combination implies the convergence to a unique equilibrium.

Luigi Barletti, Philipp Holzinger and Ansgar Jüngel derive, by means of non-standard applications of several mathematical tools, such as Wigner transform, Moyal product expansion and Chapman-Enskog expansion, quantum drift-diffusion equations for a two-dimensional electron gas with spin-orbit interaction of Rashba type. In particular, the obtained quantum drift-diffusion equations involve the full spin vector.

The chapter by Marzia Bisi and Romina Travaglini introduces a BGK kinetic model for a mixture of four polyatomic gases, which may undergo bi-molecular reversible chemical reactions. The authors prove that all disposable parameters that appear in the BGK operators may be obtained in terms of some macroscopic fields, such as the densities, the velocities and the temperatures of the species. They show that the correct collision equilibria and collision invariants of the reactive Boltzmann equations are preserved, and that the H-theorem holds.

The contribution by Guido Cavallaro is a review on the theory of the Vlasov equation, with emphasis on a specific and interesting problem: what happens when the initial mass of the plasma is infinite, and what is the effect of the decay in the space of velocities of the initial datum. Several types of interactions are considered, such as smooth or singular potentials.

Frédérique Charles has studied, both theoretically and numerically, a rarefied mixture of gas and dust. The gas is treated as a Knudsen gas, whereas the interactions between dust particles and gas molecules are modelled by considering a moving domain free transport equation. Her chapter introduces a new numerical strategy, based on a splitting between the transport of the gas molecules and the movement of the boundary.

The chapter by Amic Frouvelle investigates phase transitions (including equilibria, stability and convergence rates) for a model of collectives dynamics based on body-attitude alignment. After a review of previous results, the author presents new results for a non-linear Fokker–Planck model.

François Golse examines, in his chapter, how to describe condensation evaporation phenomena. The starting point is the steady Boltzmann equation with slab symmetry for a monatomic, hard sphere gas in a half space above its condensed phase. He proves the existence and uniqueness of a uniformly, exponentially decaying solution in the vicinity of the Maxwellian equilibrium with zero bulk velocity, with the same temperature as that of the condensed phase, and whose pressure is the saturating vapor pressure at the temperature of the interface.

Megan Griffin–Pickering and Mikaela Iacobelli review, in their chapter, some results on quasineutral limits for Vlasov equations, modelling non-magnetized non-collisional plasmas. The electron case is described by the classical Vlasov–Poisson system, while the ion case requires the introduction of an additional exponential term in the Poisson equation. The authors especially focus on the latter case.

Juhi Jang and Chanwoo Kim introduce, in their contribution, a new Hilbert-type expansion of the Boltzmann equation with the acoustic scaling. By using recent L^p-L^∞ theory of the Boltzmann equation, they show the validity of the acoustic limit in optimal scaling.

The chapter by Yunbai Cao and Chanwoo Kim reviews some results on the Vlasov–Poisson–Boltzmann system in a bounded domain of $\mathbb{R}^3$ with diffuse boundary conditions in the case of strong solutions. After giving an accurate description of the state-of-the-art on the subject, they moreover extend their regularity result when the particles are surrounded by a conductor boundary.

Tomasz Komorowski and Stefano Olla study the effect of a thermal boundary, modelled by a Langevin dynamics, on the macroscopic evolution of the energy at different space-time scales. The authors analyze how the presence of a thermostat at the origin influences the asymptotics and show that a boundary condition appears in the kinetic limit.

The chapter by Nastassia Pouradier Duteil and by Benedetto Piccoli studies a control problem for a consensus model. The problem is studied for two different sets of controls. Each constraint on the control leads to a different result. The theoretical results are presented together with several numerical simulations.

M. Piedade M. Ramos, Carolina Ribeiro and Ana Jacinta Soares provide a review on the mathematical modelling of autoimmune diseases when the kinetic theory approach is used for describing the microscopic interactions between cells. Before studying the mathematical problem, the authors give a brief introduction to the biology of autoimmune diseases and clarify the role of different types of cells involved in the process.

The contribution by Satoshi Taguchi and Tetsuro Tsuji deals with the motion of a slightly rarefied gas caused by a discontinuous wall temperature in a simple two-surface problem and illustrates how the existing theory can be extended to this case.

Shigeru Takata, Shigenori Akasobe and Masanari Hattori revisit, in their chapter, the Cercignani–Lampis model for the gas-surface interaction from the Langevin dynamics viewpoint. The velocity of the gaseous molecule after the interaction with the surface is obtained through the sum of two operators: a drift part, which drives the normal velocity to a value proportional to the wall temperature and decreases the tangential velocities, and a diffusion process.

Finally, the chapter by Edoardo Zoni and Stefan Possanner studies the accuracy of gyrokinetic equations in fusion applications. In particular, it shows the necessity of considering high-order expansion for electrons.

Pavia, Italy
March 2021

Francesco Salvarani

Contents

Sharpening of Decay Rates in Fourier Based Hypocoercivity Methods 1
Anton Arnold, Jean Dolbeault, Christian Schmeiser, and Tobias Wöhrer

Quantum Drift-Diffusion Equations for a Two-Dimensional Electron Gas with Spin-Orbit Interaction 51
Luigi Barletti, Philipp Holzinger, and Ansgar Jüngel

A Kinetic BGK Relaxation Model for a Reacting Mixture of Polyatomic Gases 69
Marzia Bisi and Romina Travaglini

On Some Recent Progress in the Vlasov–Poisson–Boltzmann System with Diffuse Reflection Boundary 93
Yunbai Cao and Chanwoo Kim

The Vlasov Equation with Infinite Mass 115
Guido Cavallaro

Mathematical and Numerical Study of a Dusty Knudsen Gas Mixture: Extension to Non-spherical Dust Particles 129
Frédérique Charles

Body-Attitude Alignment: First Order Phase Transition, Link with Rodlike Polymers Through Quaternions, and Stability 147
Amic Frouvelle

The Half-Space Problem for the Boltzmann Equation with Phase Transition at the Boundary 183
François Golse

Recent Developments on Quasineutral Limits for Vlasov-Type Equations 211
Megan Griffin-Pickering and Mikaela Iacobelli

A Note on Acoustic Limit for the Boltzmann Equation 233
Juhi Jang and Chanwoo Kim

Thermal Boundaries in Kinetic and Hydrodynamic Limits 253
Tomasz Komorowski and Stefano Olla

Control of Collective Dynamics with Time-Varying Weights 289
Benedetto Piccoli and Nastassia Pouradier Duteil

Kinetic Modelling of Autoimmune Diseases 309
M. Piedade M. Ramos, C. Ribeiro, and Ana Jacinta Soares

A Generalized Slip-Flow Theory for a Slightly Rarefied Gas Flow Induced by Discontinuous Wall Temperature 327
Satoshi Taguchi and Tetsuro Tsuji

A Revisit to the Cercignani–Lampis Model: Langevin Picture and Its Numerical Simulation 345
Shigeru Takata, Shigenori Akasobe, and Masanari Hattori

On the Accuracy of Gyrokinetic Equations in Fusion Applications 367
Edoardo Zoni and Stefan Possanner

About the Editor

Francesco Salvarani is an expert in the mathematical and numerical study of collective phenomena arising both in physics and in social sciences. His scientific activities are mainly focused on kinetic equations and systems.

Sharpening of Decay Rates in Fourier Based Hypocoercivity Methods

Anton Arnold, Jean Dolbeault, Christian Schmeiser, and Tobias Wöhrer

Abstract This paper is dealing with two L^2 hypocoercivity methods based on Fourier decomposition and mode-by-mode estimates, with applications to rates of convergence or decay in kinetic equations on the torus and on the whole Euclidean space. The main idea is to perturb the standard L^2 norm by a twist obtained either by a nonlocal perturbation build upon diffusive macroscopic dynamics, or by a change of the scalar product based on Lyapunov matrix inequalities. We explore various estimates for equations involving a Fokker–Planck and a linear relaxation operator. We review existing results in simple cases and focus on the accuracy of the estimates of the rates. The two methods are compared in the case of the Goldstein–Taylor model in one-dimension.

Keywords Hypocoercivity · Linear kinetic equations · Entropy–entropy production inequalities · Goldstein–Taylor model · Fokker–Planck operator · Linear relaxation operator · Linear BGK operator · Transport operator · Fourier modes decomposition · Pseudo-differential operators · Nash's inequality

A. Arnold · T. Wöhrer
Technische Universität Wien, Institut für Analysis und Scientific Computing, Wien, Austria
e-mail: anton.arnold@tuwien.ac.at; tobias.woehrer@asc.tuwien.ac.at
https://www.asc.tuwien.ac.at/~arnold/

J. Dolbeault (✉)
CEREMADE (CNRS UMR n°7534), PSL University, Université Paris-Dauphine, Place de Lattre de Tassigny, Paris, France
e-mail: dolbeaul@ceremade.dauphine.fr
https://www.ceremade.dauphine.fr/~dolbeaul/

C. Schmeiser
Fakultät für Mathematik, Universität Wien, Wien, Austria
e-mail: Christian.Schmeiser@univie.ac.at
https://homepage.univie.ac.at/christian.schmeiser/

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_1

1 Introduction

We consider dynamical systems involving a degenerate dissipative operator and a conservative operator, such that the combination of both operators implies the convergence to a uniquely determined equilibrium state. In the typical case encountered in kinetic theory, the dissipative part is not coercive and has a kernel which is unstable under the action of the conservative part. Such dynamical systems are called *hypocoercive* according to [34]. We are interested in the decay rate of a natural dissipated functional, the *entropy*, in spite of the indefiniteness of the entropy dissipation term. In a linear setting, the functional typically is quadratic and can be interpreted as the square of a Hilbert space norm. Classical examples are evolutions of probability densities for Markov processes with positive equilibria. Over the last 15 years, various hypocoercivity methods have been developed, which rely either on Fisher type functionals (the H^1 approach) or on entropies which are built upon weighted L^2 norms, or even weaker norms as in [6]. In the L^2 approach, it is very natural to introduce spectral decompositions and handle the free transport operator, for instance in Fourier variables, as a simple multiplicative operator. In the appropriate functional setting, the problem is then reduced to the study of a system of ODEs, which might be finite or infinite. This is the point of view that we adopt here, with the purpose of comparing several methods and benchmarking them on some simple examples.

Decay rates are usually obtained by adding a twist to the entropy or squared Hilbert space norm. In hyperbolic systems with dissipation, early attempts can be traced back to the work of Kawashima and Shizuta [27, 32], where the twist is defined in terms of a *compensating function*. The similarities between hypocoercivity and hypoellipticity are not only motivated the creation of the latter terminology, as explained in [34], but also serve as a guideline for proofs of hypocoercivity [25, 28, 34] and in particular for the construction of the twist. Here we shall focus on two approaches to L^2-hypocoercivity.

In [20, 21], an abstract method motivated by [25] and by the compensating function approach has been formulated, which provides constructive hypocoercivity estimates. The twist is built upon a non-local term associated with the spectral gap of the diffusion operator obtained in the *diffusion limit* and controls the relaxation of the macroscopic part in the limiting diffusion equation, that is, the projection of the distribution function on the orthogonal of the kernel of the dissipative part of the evolution operator. The motivating applications are kinetic transport models with diffusive macroscopic dynamics, see, e.g., [5, 18, 19, 22, 24, 30], where the results yield decay estimates in an L^2 setting.

The goal of the second approach is to find sharp decay estimates in special situations, where sufficient explicit information about the dynamics is available. Examples are ODE systems [1] as well as problems where a spectral decomposition into ODE problems exists [7–9, 11]. In these situations, sharp decay estimates can be derived by employing *Lyapunov matrix inequalities*.

In the standard definitions of hypocoercivity, a spectral gap and an exponential decay to equilibrium are required. This, however, can be expected only in sufficiently confined situations, i.e., in bounded domains or for sufficiently strong confining forces. Problems without or with too weak confinement have been treated either by regaining spectral gaps pointwise in frequency after Fourier transformation as in [15, 33] or by employing specially adapted functional inequalities in [14–16], with the Nash inequality [29] as the most prominent example.

The aim of this work is to present a review and a comparison of the two approaches mentioned above, executed for both confined and unconfined situations, where for the former a periodic setting is chosen, such that the Fourier decomposition method can be used in all cases. A special emphasis is put on optimizing the procedures with the ultimate goal of proving sharp decay rates. Attention is restricted to abstract linear hyperbolic systems with linear relaxation, where 'abstract' means that infinite systems such as kinetic transport equations are allowed. Note that in the finite dimensional case, the setting is as in [33].

In Sect. 2 of this work, both methods are presented in an abstract framework. Concerning the method of [15, 21], the setting is abstract linear ODEs, where the dynamics is driven by the sum of a dissipative and a conservative operator such that the dissipation rate is indefinite, but the conservative operator provides enough mixing to create hypocoercivity. Then the approach based on Lyapunov matrix inequalities is discussed at the hand of hyperbolic systems with relaxation. By Fourier decomposition the problem is reduced to ODE systems and Lyapunov functionals with optimal decay rates are built. These results can be seen as a sharpening of the abstract decay estimates in [33].

Section 3 is concerned with sharpening the approach of [15, 21] applied to linear kinetic equations with centered Maxwellian equilibria (for sake of simplicity). It contains results on the optimal choice of parameters in the abstract setting, on the mode-by-mode application of the method after Fourier transformation, on the convergence of an optimized rate estimate to the sharp rate in the macroscopic diffusion limit and, finally, on the derivation of global convergence or decay rates for the cases of small tori and of the Euclidean space without confinement.

Section 4 is devoted to a comparison of both approaches for a particular example, the Goldstein–Taylor model with constant exchange rate, a hyperbolic system of two equations with an exchange term in one space dimension, which can be interpreted as a discrete velocity model with two velocities. It has already been used as a model problem in [21], and the sharp decay rate on the one-dimensional torus has been derived by the Lyapunov matrix inequality approach in [8]. The challenging problem of finding the sharp decay rate for a position dependent exchange rate has been treated in [12, 13]. It is shown that the mode-by-mode Lyapunov functionals derived by both methods, the Lyapunov matrix inequality approach and the modal optimization of the abstract framework outlined in Sect. 3, coincide for the Goldstein–Taylor model. On the torus the mode-by-mode Lyapunov functionals can be combined into a global Lyapunov functional which provides the sharp decay rate (see Theorem 5). On the real line, the modal results combine into a global estimate with sharp algebraic decay rate. Due to the presence of a defective

eigenvalue in the modal equations, the standard approach requires modifications to obtain reasonable multiplicative constants.

2 Review of Two Hypocoercivity Methods

We consider the abstract evolution equation

$$\frac{dF}{dt} + \mathsf{T}F = \mathsf{L}F\,, \quad t > 0\,, \tag{1}$$

with initial datum $F(t = 0, \cdot) = F_0$. Applied to kinetic equations, T and L are respectively the *transport* and the *collision* operators, but the abstract result of this section is not restricted to such operators. We shall assume that T and L are respectively anti-Hermitian and Hermitian operators defined on a complex Hilbert space $\big(\mathcal{H}, \langle\cdot,\cdot\rangle\big)$ with corresponding norm denoted by $\|\cdot\|$.

2.1 An Abstract Hypocoercivity Result Based on a Twisted $\mathbf{L^2}$ Norm

Let us start by recalling the basic method of [21]. This technique is inspired by diffusion limits and we invite the reader to consider [21] for detailed motivations. We define

$$\mathsf{A} := \Big(\mathrm{Id} + (\mathsf{T}\Pi)^*\mathsf{T}\Pi\Big)^{-1}(\mathsf{T}\Pi)^* \tag{2}$$

where * denotes the adjoint with respect to $\langle\cdot,\cdot\rangle$ and Π is the orthogonal projection onto the null space of L. We assume that positive constants λ_m, λ_M, and C_M exist, such that, for any $F \in \mathcal{H}$, the following properties hold:

- *microscopic coercivity*

$$-\langle \mathsf{L}F, F\rangle \geq \lambda_m \|(\mathrm{Id} - \Pi)F\|^2\,, \tag{H1}$$

- *macroscopic coercivity*

$$\|\mathsf{T}\Pi F\|^2 \geq \lambda_M \|\Pi F\|^2\,, \tag{H2}$$

- *parabolic macroscopic dynamics*

$$\Pi\mathsf{T}\Pi\, F = 0\,, \tag{H3}$$

- *bounded auxiliary operators*

$$\|\mathsf{AT}(\mathrm{Id}-\Pi)F\| + \|\mathsf{AL}F\| \le C_M \,\|(\mathrm{Id}-\Pi)F\|\,. \tag{H4}$$

A simple computation shows that a solution F of (1) is such that

$$\frac{1}{2}\,\frac{d}{dt}\|F\|^2 = \langle \mathsf{L}F, F\rangle \le -\,\lambda_m\,\|(\mathrm{Id}-\Pi)F\|^2\,.$$

We assume that (1) has, up to normalization, a unique steady state F_∞. By linearity, we can replace F_0 by $F_0 - \langle F_0, F_\infty\rangle\, F_\infty$ or simply $F_0 - F_\infty$, with F_∞ appropriately normalized. With no loss of generality, we can therefore assume that $F_\infty = 0$. This is however not enough to conclude that $\|F(t,\cdot)\|^2$ decays exponentially with respect to $t \ge 0$. As in the *hypocoercivity* method introduced in [21] for real valued operators and extended in [15] to complex Hilbert spaces, we consider the Lyapunov functional

$$\mathsf{H}_1[F] := \frac{1}{2}\,\|F\|^2 + \delta\,\mathrm{Re}\langle \mathsf{A}F, F\rangle \tag{3}$$

for some $\delta > 0$ to be determined later. If F solves (1), then

$$\begin{aligned} -\,\tfrac{d}{dt}\mathsf{H}_1[F] = \mathsf{D}[F] := &-\langle \mathsf{L}F, F\rangle + \delta\,\langle \mathsf{AT}\Pi F, F\rangle \\ &-\delta\,\mathrm{Re}\langle \mathsf{TA}F, F\rangle + \delta\,\mathrm{Re}\langle \mathsf{AT}(\mathrm{Id}-\Pi)F, F\rangle - \delta\,\mathrm{Re}\langle \mathsf{AL}F, F\rangle\,. \end{aligned} \tag{4}$$

The following result has been established in [15, 21].

Theorem 1 *Let* L *and* T *be closed linear operators in the complex Hilbert space* $\big(\mathcal{H}, \langle\cdot,\cdot\rangle\big)$. *We assume that* L *is Hermitian and* T *is anti-Hermitian, and that* (H1)–(H4) *hold for some positive constants* λ_m, λ_M, *and* C_M. *Then for some* $\delta > 0$, *there exists* $\lambda > 0$ *and* $C > 1$ *such that, if* F *solves* (1) *with initial datum* $F_0 \in \mathcal{H}$, *then*

$$\mathsf{H}_1[F(t,\cdot)] \le \mathsf{H}_1[F_0]\,e^{-\lambda t} \quad \textit{and} \quad \|F(t,\cdot)\|^2 \le C\,e^{-\lambda t}\,\|F_0\|^2 \quad \forall\, t \ge 0\,. \tag{5}$$

Here we assume that the unique steady state is $F_\infty = 0$ otherwise we have to replace $F(t,\cdot)$ by $F(t,\cdot) - F_\infty$ and F_0 by $F_0 - F_\infty$ in (5). The strategy of [21], later extended in [15], is to prove that for any $\delta > 0$ small enough, we have

$$\lambda\,\mathsf{H}_1[F] \le \mathsf{D}[F] \tag{6}$$

for some $\lambda > 0$, and

$$c_-\,\|F\|^2 \le \mathsf{H}_1[F] \le c_+\,\|F\|^2 \tag{7}$$

for some constants c_- and c_+ such that $0 \le c_- \le 1/2 \le c_+$. As a consequence, if $c_- > 0$, we obtain the estimate $C \le c_+/c_-$. We learn from [15, Proposition 4] that Theorem 1 holds with $c_\pm = (1 \pm \delta)/2$,

$$\lambda = \frac{\lambda_M}{3\,(1+\lambda_M)} \min\left\{1, \lambda_m, \tfrac{\lambda_m\,\lambda_M}{(1+\lambda_M)\,C_M^2}\right\} \text{ and } \delta = \frac{1}{2} \min\left\{1, \lambda_m, \tfrac{\lambda_m\,\lambda_M}{(1+\lambda_M)\,C_M^2}\right\}. \tag{8}$$

Our primary goal of Sect. 3 is to obtain sharper estimates of λ, $c_\pm$ and C for an appropriate choice of δ in specific cases. Notice that it is convenient to work in an Hilbert space framework because this allows us to use Fourier transforms.

2.2 An Abstract Hypocoercivity Result Based on Lyapunov Matrix Inequalities

Here we review our second hypocoercivity method, as developed on various examples in [2, 3, 9], before comparing it with the method of Sect. 2.1.

As in Sect. 2.1, without loss of generality we assume that (1) has the unique steady state $F_\infty = 0$. We are interested in explicit decay rates for $\|F(t,\cdot)\|^2 \to 0$ as $t \to +\infty$. To fix the ideas we start with some prototypical examples:

1. Although almost trivial, the *stable ODEs* with constant-in-t coefficients

$$\frac{dF}{dt} = -\,C\,F \tag{9}$$

   is at the core of the method. Here $F(t) \in \mathbb{C}^n$, $\mathsf{T} := C_{AH} \in \mathbb{C}^{n\times n}$ is an anti-Hermitian matrix, and $\mathsf{L} := -\,C_H \in \mathbb{C}^{n\times n}$ is a Hermitian negative semi-definite matrix, where C_{AH} and C_H denote the anti-Hermitian and Hermitian parts of $C = C_{AH} + C_H$. Several other examples will be reduced to (9), mostly via Fourier transformation in x. We shall use the same index notation ('AH' and 'H') for matrix B in Example 4 and matrix C in Example 5. The hypocoercivity structure of (9) is discussed in [1].
2. *Discrete velocity BGK models*, i.e. transport-relaxation equations (see § 2.1 and § 4.1 in [2]) can be written in the form of (1) where $F(t,x) = \big(f_1(t,x), ..., f_n(t,x)\big)^\top$, $x \in \mathcal{X} \subset \mathbb{R}$, $\mathsf{T} := V\,\partial_x$ with the diagonal matrix $V \in \mathbb{R}^{n\times n}$ representing the velocities, and the collision operator $\mathsf{L} := \sigma\,B$ with $\sigma > 0$. Here, the matrix $B \in \mathbb{R}^{n\times n}$ is in *BGK form*

$$B = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \otimes (1, \ldots, 1) - \mathrm{Id}$$

with $b = (b_1, \dots, b_n)^\top \in (0,1)^n$ such that $\sum_{j=1}^n b_j = 1$, and Id denotes the identity matrix. The collision operator L is symmetric on the velocity-weighted L^2-space $\mathcal{H} = \mathrm{L}^2(\mathcal{X} \times \{1, \dots, n\}; \{b_j^{-1}\})$. Due to this structure, B has a simple eigenvalue 0 with corresponding left eigenvector $l_1 = (1, \dots, 1)$ associated with the mass conservation of the system. The corresponding right eigenvector b spans the local-in-x steady states, which are of the form $\rho(x)\, b$ for some arbitrary scalar function $\rho(x)$. The case with only two velocities, or *Goldstein–Taylor model*, is dealt with in Sect. 4.

3. A linear *kinetic BGK model* is analyzed in [2], where $F = f(t,x,v) \in \mathbb{R}$, $x \in \mathbb{T}$ (the 1-dimensional torus of length 2π), and $v \in \mathbb{R}$. The kinetic transport operator is $\mathsf{T} := v\, \partial_x$, and the BGK operator $\mathsf{L}f := \mathcal{M}_\vartheta(v) \int_{\mathbb{R}} f\, dv - f$ is symmetric in the weighted space $\mathcal{H} = \mathrm{L}^2\big(\mathbb{T} \times \mathbb{R}; dx\, dv/(2\pi\, \mathcal{M}_\vartheta(v))\big)$, where $\mathcal{M}_\vartheta(v)$ denotes the centered Maxwellian with variance (or temperature) ϑ. The kernel of L is spanned by $\mathcal{M}_\vartheta(v)$, which is also the global steady state $F_\infty(v)$, due to the setting on the torus.
4. The (degenerate) *reaction-diffusion systems* of [23] can also be written as in (1), with $F(t,x) = (f_1(t,x), ..., f_n(t,x))^\top$, $\mathsf{T} := -B_{AH}$, $\mathsf{L} := D\,\Delta + B_H$. Here, $0 \le D \in \mathbb{R}^{n\times n}$ is a diagonal matrix, $B \in \mathbb{R}^{n\times n}$ is an essentially non-negative matrix, i.e., $b_{ij} \ge 0$ for any $i \ne j$, and $b_{ii} = -\sum_{i\ne j} b_{ij}$, and B_{AH} and B_H are its anti-symmetric and symmetric parts.
5. As a final example, let us mention (possibly degenerate) *Fokker–Planck equations* with linear-in-x drift for $F = f(t,x)$, $x \in \mathbb{R}^d$. After normalization (in the sense of [11]), they can be identified with (1), where $\mathsf{T}f := -\operatorname{div}\big(f_\infty\, C_{AH}\, \nabla(f/f_\infty)\big)$, $\mathsf{L}f := \operatorname{div}\big(f_\infty\, C_H\, \nabla(f/f_\infty)\big)$, with a positive stable drift matrix $C \in \mathbb{R}^{d\times d}$ such that $C_H \ge 0$, and $f_\infty = (2\pi)^{-d/2} \exp\big(-|x|^2/2\big)$ is the unique normalized steady state. As shown in [11], these Fokker–Planck equations are equivalent to (9) and tensorized versions of it.

In the articles cited above for Examples 1–3, an L^2-based hypocoercive entropy method has been used to derive sharp decay estimates for the solution $F(t)$ towards its steady state F_∞, and the same strategy can also be applied to Example 4. In [9] an H^1-based hypocoercive entropy method was developed for the Fokker–Planck equations in Example 5. But in view of its subspace decomposition given in [11], an L^2-analysis is also feasible.

In our second hypocoercive entropy method, we construct a problem adapted Lypunov functional that is able to reveal the sharp decay behavior as $t \to +\infty$. We shall illustrate this strategy for Examples 2 and 3, where the anti-Hermitian operator T is either $V\, \partial_x$ (for discrete velocities) or $v\, \partial_x$ (for continuous velocities). In order to establish the *mode-by-mode hypocoercivity*, we Fourier transform (1) w.r.t. $x \in \mathcal{X}$, with either $\mathcal{X} = \mathbb{T}^1 =: \mathbb{T}$ or $\mathcal{X} = \mathbb{R}^d$. In the torus case, we assume $d = 1$ for simplicity, but the method extends to higher dimensions (see [3]). With the

abuse of notations of keeping F for the distribution function written in the variables (t, ξ, v), this yields

$$\frac{dF}{dt} = -i\,\xi\,VF + \mathsf{L}F =: -C(\xi)\,F\,, \tag{10}$$

with a discrete modal variable $\xi \in \mathbb{Z}$ for the torus and $\xi \in \mathbb{R}^d$ in the whole space case. In (10), V is a diagonal matrix for Example 2, and for Example 3 it either represents the multiplication operator by v or, when using a basis in the v-variable, a symmetric, real-valued "infinite matrix" (*cf.* [2, § 4]).

For each fixed mode ξ, (10) is now an ODE with constant coefficients (of dimension $n < \infty$ for Example 2, and infinite dimensional for Example 3). For finite n, we define the *modal spectral gap* of $C(\xi)$ as

$$\mu(\xi) := \min_{0 \neq \lambda_j \in \sigma(C(\xi))} \mathrm{Re}(\lambda_j)\,. \tag{11}$$

If no eigenvalue of $C(\xi)$ with $\mathrm{Re}(\lambda_j) = \mu(\xi)$ is defective (i.e., all eigenvalues have matching algebraic and geometric multiplicities), see e.g. [26], then the exponential decay of $\|F(t,\xi)\|^2$ with the sharp rate $2\,\mu(\xi)$ is shown using a Lyapunov functional obtained as a twisted Euclidean norm on $\mathbb{C}^n$. To this end we use the following algebraic result.

Lemma 1 ([2, Lemma 2]) *For a given matrix $C \in \mathbb{C}^{n\times n}$, let μ be defined as in (11). Assume that $0 \notin \sigma(C)$ and that C has no defective eigenvalues with $\mathrm{Re}(\lambda_j) = \mu$. Then there exists a positive definite Hermitian matrix $P \in \mathbb{C}^{n\times n}$ such that*

$$C^*P + P\,C \geq 2\,\mu\,P\,. \tag{12}$$

Moreover, if all eigenvalues of C are non-defective, any matrix

$$P := \sum_{j=1}^{n} c_j\, w_j \otimes w_j^* \tag{13}$$

satisfies (12), where $w_j \in \mathbb{C}^n$ denote the normalized (right) eigenvectors of C^* and, for all $j = 1, \ldots, n$, the coefficient $c_j \in (0, +\infty)$ is an arbitrary weight.

For the extension of this lemma to the case $0 \in \sigma(C)$ we refer to [2, Lemma 3], but, anyhow, this is typically relevant only for $\xi = 0$. The more technical case when C has defective eigenvalues was analyzed in [9, Lemma 4.3(i)]. In the case $n = \infty$ (occuring in the kinetic BGK models of Example 3), the eigenfunction construction of the operator (or "infinite matrix") P via (13) is, in general, not feasible. A systematic construction of *approximate* matrices P with a suboptimal value compared with μ in (12) was presented in [2, § 4.3–4.4] and [3, § 2.3].

Using the deformation matrix P, we define the "twisted Euclidean norm" in $\mathbb{C}^n$ as

$$\|F\|_P^2 := \langle F, P\, F\rangle\,,$$

which is equivalent to the Euclidean norm $\|\cdot\|$ through the estimate

$$\lambda_1^P\,\|F\|^2 \le \|F\|_P^2 \le \lambda_n^P\,\|F\|^2\,, \tag{14}$$

where λ_1^P and λ_n^P are the smallest and largest eigenvalues of P, respectively. From (9) and (12) follows that

$$\frac{d}{dt}\|F\|_P^2 = -\,\langle F, (C^*P + PC)F\rangle \le -\,2\,\mu\,\|F\|_P^2\,.$$

This shows that solutions to (9) satisfy

$$\|F(t)\|_P^2 \le e^{-2\,\mu\,t}\,\|F_0\|_P^2 \quad \forall\, t \ge 0\,,$$

and hence in the Euclidean norm:

$$\|F(t)\|^2 \le \mathrm{cond}(P)\,e^{-2\,\mu\,t}\,\|F_0\|^2 \quad \forall\, t \ge 0\,, \tag{15}$$

where $\mathrm{cond}(P) := \lambda_n^P/\lambda_1^P$ denotes the *condition number* of P. We recall from [4] that $\mathrm{cond}(P)$ is in general not the minimal multiplicative constant for (15). In fact, in general it is impossible to obtain that optimal constant from a Lyapunov functional, even for $n = 2$, see [4, Theorem 4.1]. We also remark that the matrix P from (13) is not uniquely determined (even beyond trivial multiples). As a consequence, $\mathrm{cond}(P)$ may be different for different admissible choices of P. For an example with $n = 3$, we refer to [4, § 3].

Analogous decay estimates hold for solutions $F(t,\xi)$ to the modal ODEs (10), and they involve the deformation matrices $P(\xi)$ and the modal spectral gaps $\mu(\xi)$:

$$\|F(t,\xi)\|_{P(\xi)}^2 \le e^{-2\,\mu(\xi)\,t}\,\|F_0(\xi)\|_{P(\xi)}^2 \quad \forall\, t \ge 0\,. \tag{16}$$

This motivates the definition of a *modal-based Lyapunov functional* by assembling the modal functionals. We present two variants of this approach.

Strategy 1 We consider the global Lyapunov functional

$$\mathsf{H}_2[F] := \sum_{\xi\in\mathbb{Z}} \|F(\xi)\|_{P(\xi)}^2\,, \tag{17}$$

which is written here for the case of discrete modes, i.e., $\mathcal{X} = \mathbb{T}$.

We recall that the matrix $P(\xi)$ is not unique. In the kinetic BGK examples studied so far (*cf.* [2, 3]) it was convenient to choose P depending continuously on ξ (for $\xi \in \mathbb{R}^d$) and such that $P(\xi) \to \mathrm{Id}$ as $|\xi| \to +\infty$. For kinetic equations with a local-in-x dissipative operator L, the matrix $C(\xi)$ has the form given in (10). Under the assumption of a uniform spectral gap $\overline{\mu} := \inf_\xi \mu(\xi) > 0$, the form $P(\xi) = \mathrm{Id} + O(1/|\xi|)$ is very natural (see $P^{(1)}(\xi)$ in (54) for an example) in view of the matrix inequality (12).

The modal decay (16) implies the following decay estimate for the solution to (1):

$$\mathsf{H}_2[F(t)] \le e^{-2\,\bar{\mu}\,t}\,\mathsf{H}_2[F_0] \quad \forall\, t \ge 0\,, \quad \text{for any } F_0 \perp F_\infty\,.$$

Using Parseval's identity and the norm equivalence from (14), this yields

$$\|F(t)\|^2 \le \bar{c}_P\, e^{-2\,\bar{\mu}\,t}\,\|F_0\|^2 \quad \forall\, t \ge 0\,, \quad \text{for any } F_0 \perp F_\infty\,, \tag{18}$$

where $\bar{c}_P := \sup_\xi \mathrm{cond}(P(\xi))$.

Strategy 2 If all modes ξ have the same spectral gap $\mu(\xi)$, then the estimate (18) clearly yields the minimal multiplicative constant $\bar{c}_P$ (obtainable by Lyapunov methods). This is the case when the relaxation rate $\sigma < 2$ in the Goldstein–Taylor model, which is studied in [7] and in Sect. 4 below. But faster decaying modes may have a "too large" condition number $\mathrm{cond}(P(\xi))$, as it is the case for $\sigma > 2$ in [7]. Then, the matrices $P(\xi)$ from (12) have to be modified in order to reduce $\mathrm{cond}(P(\xi))$ by lowering $\mu = \mu(\xi)$ in (12). For simplicity we detail this strategy only for the case that the infimum $\bar{\mu}$ is actually attained. Main steps are:

- Let $\Xi := \{\xi : \mu(\xi) = \bar{\mu}\}$ be the set of the modes with slowest decay. Set $c_\Xi := \sup_{\xi\in\Xi} \mathrm{cond}(P(\xi))$, i.e. the worst common multiplicative constant for these slow modes.
- For all modes $\xi \notin \Xi$, we distinguish several cases:
  - If $\mathrm{cond}(P(\xi)) \le c_\Xi$, set $\tilde{P}(\xi) := P(\xi)$.
  - If $\mathrm{cond}(P(\xi)) > c_\Xi$, then replace $P(\xi)$ by $\tilde{P}(\xi) \in \mathbb{C}^{n\times n}$, which is a positive definite Hermitian solution to the matrix inequality

$$C(\xi)^* P + P\,C(\xi) \ge 2\,\bar{\mu}\,P\,.$$

    In particular, P should be either chosen as any such solution that satisfies $\mathrm{cond}(P(\xi)) \le c_\Xi$ or, if this is impossible, then by a solution P having the least condition number.
- Let $\tilde{c}_\Xi := \sup_{\xi\notin\Xi} \mathrm{cond}\left(\tilde{P}(\xi)\right)$ be the best multiplicative constant for the faster modes.

- Set $\tilde{c}_P := \max\{c_\Xi, \tilde{c}_\Xi\}$. With this construction we define a second, refined Lyapunov functional (again written for the case $\mathcal{X} = \mathbb{T}$) by

$$\widetilde{\mathsf{H}}_2[F] := \sum_{\xi\in\Xi} \|F(\xi)\|^2_{P(\xi)} + \sum_{\xi\in\Xi^c} \|F(\xi)\|^2_{\tilde{P}(\xi)}\,, \tag{19}$$

  where $\Xi^c := \mathbb{Z} \setminus \Xi$.

This yields the improved decay estimate (*w.r.t.* the multiplicative constant):

$$\|F(t)\|^2 \le \tilde{c}_P\, e^{-2\,\bar{\mu}\,t}\, \|F_0\|^2 \quad \forall\, t \ge 0\,, \quad \text{for any } F_0 \perp F_\infty\,. \tag{20}$$

Note that, by construction, $\tilde{c}_P \le \bar{c}_P$. Altogether, our estimates on a solution to the evolution equation (1) rewritten as (10) in Fourier variables can be summarized into the following result.

Proposition 1 *On $\mathbb{T}$, let us consider an operator C such that, in Fourier variables, $C(\xi)$ takes values in $\mathbb{C}^{n\times n}$ for any $\xi \in \mathbb{Z}$. Assume the existence of a uniform spectral gap $\overline{\mu} := \inf_{\xi\in\mathbb{Z}} \mu(\xi) > 0$ where $\mu(\xi)$ is defined by (11).*

(a) If the corresponding modal deformation matrices $P(\xi)$ satisfy $\bar{c}_P < \infty$, then the solutions of (1) satisfy the decay estimate (18).
(b) If the modified deformation matrices $\tilde{P}(\xi)$ satisfy $\tilde{c}_P < \infty$, then the solutions of (1) satisfy the decay estimate (20).

The above procedure was applied in [7] to the Goldstein–Taylor model, and in [2] to Examples 2–3, considered on $\mathbb{T}$.

The hypocoercivity results based on the Lyapunov matrix inequalities (12) and mode-by-mode estimates as in (16) have the advantage that, in simple cases, it is possible to identify the optimal decay rates. They are less flexible than the hypocoercivity results based on the twisted L^2 norm inspired by diffusion limits of Sect. 2.1. Our purpose of Sect. 4 is to detail several variants of these methods in simple cases, draw a few consequences and compare the estimates of the two methods.

3 Optimization of Twisted L^2 Norms

This part is devoted to accurate hypocoercivity estimates in Fourier variables based on our *first abstract method*, for two simple kinetic equations with Gaussian local equilibria. It is a refined version of the paper [15] devoted to a larger class of equilibria, but to the price of weaker bounds. Here we underline some key ideas of mode-by-mode hypocoercivity and perform more accurate and explicit computations. New estimates are obtained, which numerically improve upon known ones. Rates and constants are discussed and numerically illustrated, with the purpose of establishing benchmarks for the L^2-*hypocoercivity theory based upon a twist*

inspired by diffusion limits. Exponential rates are obtained on the torus, with a discussion on high frequency estimates. On the whole space case, low frequencies are involved in the computation of the asymptotic decay rates. We also detail how spectral estimates of the mode-by-mode L^2 hypocoercivity method can be systematically turned into rates of decay using the ideas of the original proof of Nash's inequality.

3.1 A Detailed Mode-by-Mode Approach

3.1.1 Introduction

We consider the Cauchy problem

$$\partial_t f + v \cdot \nabla_x f = \mathsf{L} f \,, \quad f(0,x,v) = f_0(x,v) \,, \tag{21}$$

for a distribution function $f(t,x,v)$, where $x \in \mathbb{R}^d$ denotes the position variable, $v \in \mathbb{R}^d$ is the velocity variable, and $t \geq 0$ is the time. Concerning the collision operator, L denotes the *Fokker–Planck operator* L_1 or, as in [20], the *linear BGK operator* L_2, which are defined respectively by

$$\mathsf{L}_1 f := \Delta_v f + \nabla_v \cdot (v\, f) \quad \text{and} \quad \mathsf{L}_2 f := \rho_f\, \mathcal{M} - f \,.$$

Here $\mathcal{M}$ is the normalized Gaussian function

$$\mathcal{M}(v) = \frac{e^{-\frac{1}{2}|v|^2}}{(2\,\pi)^{d/2}} \quad \forall\, v \in \mathbb{R}^d$$

and $\rho_f := \int_{\mathbb{R}^d} f\, dv$ is the spatial density. Notice that $\mathcal{M}$ spans the kernel of L. We introduce the *weight*

$$d\gamma := \gamma(v)\, dv \quad \text{where} \quad \gamma := \frac{1}{\mathcal{M}}$$

and the weighted norm

$$\|f\|^2_{\mathrm{L}^2(dx\,d\gamma)} := \iint_{\mathcal{X}\times\mathbb{R}^d} |f(x,v)|^2\, dx\, d\gamma \,,$$

where $\mathcal{X}$ denotes either the cube $[0,L)^d$ with periodic boundary conditions or $\mathcal{X} = \mathbb{R}^d$, that is, the whole Euclidean space.

Let us consider the Fourier transform of f in x defined by

$$\hat f(t,\xi,v)=\int_{\mathcal{X}} e^{-i\,x\cdot\xi}\,f(t,x,v)\,dx\,, \tag{22}$$

where either $\mathcal{X}=[0,L)^d$ (with periodic boundary conditions), or $\mathcal{X}=\mathbb{R}^d$. We denote by $\xi\in(2\pi/L)^d\,\mathbb{Z}^d\subset\mathbb{R}^d$ or $\xi\in\mathbb{R}^d$ the Fourier variable. Details will be given in Sect. 3.1.3. Next, we rewrite Eq. (21) for $F=\hat f$ as

$$\partial_t F+\mathsf{T}F=\mathsf{L}F\,,\quad F(0,\xi,v)=\hat f_0(\xi,v)\,,\quad \mathsf{T}F=i\,(v\cdot\xi)F\,. \tag{23}$$

Here we abusively use the same notation T for the transport operator in the original variables and after the Fourier transform, where it is a simple multiplication operator. We shall also consider ξ as a given, fixed parameter and omit it whenever possible, so that we shall write that F is a function of (t,v), for sake of simplicity. Let us define

$$\mathcal{H}=\mathrm{L}^2(d\gamma)\,,\quad \|F\|^2=\int_{\mathbb{R}^d}|F|^2\,d\gamma\,,\quad \Pi F=\mathcal{M}\int_{\mathbb{R}^d}F\,dv=\mathcal{M}\,\rho_F\,. \tag{24}$$

Our goal is to obtain decay estimates of $\|F\|$ parameterized by ξ and this is why such an approach can be qualified as a *mode-by-mode hypocoercivity method.*

3.1.2 A First Optimization in the General Setting

The estimates of [15, Proposition 4] are rough and it is possible to improve upon the choice for δ and λ. On the triangle

$$\mathcal{T}_m:=\Big\{(\delta,\lambda)\in(0,\lambda_m)\times(0,2\,\lambda_m)\,:\,\lambda<2\,(\lambda_m-\delta)\Big\}\,,$$

let us define

$$h_\star(\delta,\lambda):=\delta^2\left(C_M+\frac{\lambda}{2}\right)^2-4\left(\lambda_m-\delta-\frac{\lambda}{2}\right)\left(\frac{\delta\,\lambda_M}{1+\lambda_M}-\frac{\lambda}{2}\right)\,,$$

$$\lambda_\star(\delta):=\sup\Big\{\lambda\in(0,2\,\lambda_m)\,:\,h_\star(\delta,\lambda)\le0\Big\}\quad\text{and}\quad C_\star(\delta):=\frac{2+\delta}{2-\delta}\,.$$

We will also need later

$$K_M:=\frac{\lambda_M}{1+\lambda_M}<1\quad\text{and}\quad\delta_\star:=\frac{4\,K_M\,\lambda_m}{4\,K_M+C_M^2}<\lambda_m\,.$$

Our first result provides us with the following refinement of (5).

Proposition 2 *Under the assumptions of Theorem 1, we have*

$$\mathsf{H}_1[F(t,\cdot)] \le \mathsf{H}_1[F_0]\, e^{-\lambda\, t} \quad \forall\, t \ge 0$$

with $\lambda = \max\Big\{\lambda_\star(\delta) \,:\, \delta \in (0,\delta_\star)\Big\}$. *Moreover, for any* $\delta < \min\{2,\delta_\star\}$, *if* F *solves* (1) *with initial datum* $F_0 \in \mathcal{H}$, *then*

$$\|F(t)\|^2 \le C_\star(\delta)\, e^{-\,\lambda_\star(\delta)\, t}\, \|F_0\|^2 \quad \forall\, t \ge 0\,.$$

On the boundary of the triangle $\mathcal{T}_m$, we notice that

$$h_\star(0,\lambda) = \lambda\,(2\,\lambda_m - \lambda) > 0 \quad \forall\, \lambda \in (0, 2\,\lambda_m)\,,$$

$$h_\star\Big(\delta, 2\,(\lambda_m - \delta)\Big) = (C_M + \lambda_m - \delta)^2\,\delta^2 > 0 \quad \forall\, \delta \in (0,\lambda_m)\,,$$

and $h_\star(\delta,0)/\delta = \big(C_M^2 + 4\,K_M\big)\,\delta - 4\,K_M\,\lambda_m$ is negative if $0 < \delta < \delta_\star$. As a consequence, the set $\{(\delta,\lambda) \in \mathcal{T}_m \,:\, h_\star(\delta,\lambda) \le 0\}$ is non-empty. The functions $\lambda \mapsto h_\star(\delta,\lambda)$ for a fixed $\delta \in (0,\lambda_m)$ and $\delta \mapsto h_\star(\delta,\lambda)$ for a fixed $\lambda \in (0, 2\,\lambda_m)$ are both polynomials of second degree. The expression of $\lambda_\star(\delta)$ is explicitly computed as the smallest root of $\lambda \mapsto h_\star(\delta,\lambda)$ but has no interest by itself. It is also elementary to check that $h_\star$ is positive if $(\delta,\lambda) \in \mathcal{T}_m$ with $\delta > \delta_\star$.

Proof The method is the same as in [21] and [15, Proposition 4], except that we use sharper estimates.

Since $\mathsf{AT}\Pi$ can be interpreted as $z \mapsto (1+z)^{-1}\,z$ applied to $(\mathsf{T}\Pi)^*\mathsf{T}\Pi$, the spectral theorem and conditions (H1) and (H2) imply that

$$-\,\langle \mathsf{L}F, F\rangle + \delta\,\langle \mathsf{AT}\Pi F, F\rangle \ge \lambda_m\,\|(\mathrm{Id}-\Pi)F\|^2 + \frac{\delta\,\lambda_M}{1+\lambda_M}\,\|\Pi F\|^2\,. \tag{25}$$

From that point, one has to prove that $-\,\langle \mathsf{L}F, F\rangle + \delta\,\langle \mathsf{AT}\Pi F, F\rangle$ controls the other terms in the expression of $\mathsf{D}[F]$. By (H4), we know that

$$|\mathrm{Re}\langle \mathsf{AT}(\mathrm{Id}-\Pi)F, F\rangle + \mathrm{Re}\langle \mathsf{AL}F, F\rangle| \le C_M\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\|\,. \tag{26}$$

As in [21, Lemma 1], if $G = \mathsf{A}F$, i.e., if $(\mathsf{T}\Pi)^*F = G + (\mathsf{T}\Pi)^*\,\mathsf{T}\Pi\,G$, then

$$\langle \mathsf{TA}F, F\rangle = \langle G, (\mathsf{T}\Pi)^*\,F\rangle = \|G\|^2 + \|\mathsf{T}\Pi G\|^2 = \|\mathsf{A}F\|^2 + \|\mathsf{TA}F\|^2\,.$$

By the Cauchy-Schwarz inequality, we know that

$$\begin{aligned}\langle G, (\mathsf{T}\Pi)^*\,F\rangle &= \langle \mathsf{TA}F, (\mathrm{Id}-\Pi)F\rangle\\ &\le \|\mathsf{TA}F\|\,\|(\mathrm{Id}-\Pi)F\| \le \frac{1}{2\,\mu}\,\|\mathsf{TA}F\|^2 + \frac{\mu}{2}\,\|(\mathrm{Id}-\Pi)F\|^2\end{aligned}$$

for any $\mu > 0$. Hence

$$2\,\|\mathsf{A}F\|^2 + \left(2 - \frac{1}{\mu}\right)\|\mathsf{TA}F\|^2 \le \mu\,\|(\mathrm{Id} - \Pi)F\|^2\,,$$

which, by taking either $\mu = 1/2$ or $\mu = 1$, proves that

$$\|\mathsf{A}F\| \le \frac{1}{2}\,\|(\mathrm{Id} - \Pi)F\|\,, \quad \|\mathsf{TA}F\| \le \|(\mathrm{Id} - \Pi)F\|$$

and establishes (7). Incidentally, this proves that

$$|\langle \mathsf{TA}F, F\rangle| = |\langle \mathsf{TA}F, (\mathrm{Id} - \Pi)F\rangle| \le \|(\mathrm{Id} - \Pi)F\|^2\,, \tag{27}$$

and also that

$$|\langle \mathsf{A}F, F\rangle| \le \frac{1}{2}\,\|\Pi F\|\,\|(\mathrm{Id} - \Pi)F\| \le \frac{1}{4}\,\|F\|^2\,. \tag{28}$$

As a consequence of this last identity, we obtain

$$\left|\mathsf{H}_1[F] - \tfrac{1}{2}\,\|F\|^2\right| = \delta\,\left|\langle \mathsf{A}F, F\rangle\right| \le \frac{\delta}{4}\,\|F\|^2\,,$$

which, under the condition $\delta < 2$, is a proof of (7) with the improved constant

$$c_\pm = \frac{2 \pm \delta}{4}\,. \tag{29}$$

Now let us come back to the proof of (6). Collecting (25), (26), and (27) with the definition of $\mathsf{D}[F]$, we find that

$$\mathsf{D}[F] \ge (\lambda_m - \delta)\,X^2 + \frac{\delta\,\lambda_M}{1 + \lambda_M}\,Y^2 - \delta\,C_M\,X\,Y$$

with $X := \|(\mathrm{Id} - \Pi)F\|$ and $Y := \|\Pi F\|$. Using (28), we observe that

$$\mathsf{H}_1[F] \le \frac{1}{2}\left(X^2 + Y^2\right) + \frac{\delta}{2}\,X\,Y\,.$$

Hence the largest value of λ for which

$$\mathsf{D}[F] \ge \lambda\,\mathsf{H}_1[F]$$

can be estimated by the largest value of λ for which

$$Q(X,Y) := (\lambda_m - \delta)\, X^2 + \frac{\delta\,\lambda_M}{1+\lambda_M}\, Y^2 - \delta\, C_M\, X\, Y - \frac{\lambda}{2}\left(X^2+Y^2\right) - \frac{\lambda}{2}\,\delta\, X\, Y$$
$$= \left(\lambda_m - \delta - \frac{\lambda}{2}\right) X^2 - \delta\left(C_M + \frac{\lambda}{2}\right) X\, Y + \left(\frac{\delta\,\lambda_M}{1+\lambda_M} - \frac{\lambda}{2}\right) Y^2$$

is a nonnegative quadratic form. It is characterized by the discriminant condition $h_\star(\delta,\lambda) \le 0$, and the condition $\lambda_m - \delta - \lambda/2 > 0$ which determines $\mathcal{T}_m$ with the two other conditions: $\delta > 0$ and $\lambda > 0$. From (6), we deduce the decay of $\mathsf{H}_1[F(t,\cdot)]$ and the decay of $\|F(t)\|^2$ by (7) using (29). □

Remark 1 The estimate (8) of [15, Proposition 4] is easily recovered as follows. Using

$$\mathsf{D}[F] \ge (\lambda_m - \delta)\, X^2 + \frac{\delta\,\lambda_M}{1+\lambda_M}\, Y^2 - \delta\, C_M\, X\, Y$$
$$\ge (\lambda_m - \delta)\, X^2 + \frac{\delta\,\lambda_M}{1+\lambda_M}\, Y^2 - \frac{\delta}{2}\left(C_M^2\, X^2 + Y^2\right)$$

and

$$\mathsf{H}_1[F] \le \frac{2+\delta}{4}\left(X^2+Y^2\right),$$

with δ defined as in (8), we obtain

$$\mathsf{D}[F] \ge \frac{\lambda_m}{4}\, X^2 + \frac{\delta\,\lambda_M}{2\,(1+\lambda_M)}\, Y^2$$
$$\ge \frac{1}{4}\,\min\left\{\lambda_m, \frac{2\,\delta\,\lambda_M}{1+\lambda_M}\right\}\|F\|^2 \ge \frac{2\,\delta\,\lambda_M}{3\,(1+\lambda_M)}\,\mathsf{H}[F]\,.$$

Hence we have that $\frac{1}{4}\,\|F\|^2 \ge \frac{1}{3}\,\mathsf{H}[F]$ because $4/(2+\delta) \ge 8/5 > 4/3$ if $\delta < 1/2$. This estimate is non-optimal and it is improved in the proof of Proposition 2.

Remark 2 In the discussion of the positivity of Q, we can observe that (X,Y) is restricted to the upper right quadrant corresponding to $X > 0$ and $Y > 0$. The discriminant condition $h_\star(\delta,\lambda) \le 0$ and the condition $\lambda_m - \delta - \lambda/2 > 0$ guarantee that $Q(X,Y) \ge 0$ *for any* $X,\, Y \in \mathbb{R}$, which is of course a sufficient condition. It is also necessary because the coefficient of Y^2 is positive (otherwise one can find some $X > 0$ and $Y > 0$ such that $Q(X,Y) < 0$) and then by solving a second degree equation, one could again find a region in the upper right quadrant such that Q takes negative values.

Hence we produce a necessary and sufficient condition for Q to be a nonnegative quadratic form. This does not mean that the condition of Proposition 2 is necessary

because we have made various estimates, which are not generically optimal, in order to reduce the problem to the discussion of the sign of Q. In special cases, we can indeed improve upon Proposition 2. We will discuss such improvements in the next section.

3.1.3 Mode-by-Mode Hypocoercivity

Fourier Representation and Mode-by-Mode Estimates Let us consider the Fourier transform in x, take the Fourier variable $\xi \in \mathbb{R}^d$ as a parameter, and study, *for a given* ξ, Eq. (23). For a given $\xi \in \mathbb{R}^d$, let us implement the strategy of Theorem 1 and Proposition 2 applied to $(t, v) \mapsto F(t, \xi, v)$, with the choices (24). The operator A is defined by

$$(\mathsf{A}F)(v) = -\frac{i\,\xi}{1+|\xi|^2} \cdot \int_{\mathbb{R}^d} w\, F(w)\, dw\, \mathcal{M}(v)\,.$$

Taking advantage of the explicit form of A, we can reapply the method of Sect. 3.1.2 with explicit numerical values, and actually improve upon the previous results. Let us give some details, which will be useful for benchmarks and numerical computations. Again we aim at relating the Lyapunov functional

$$\mathsf{H}_1[F] := \frac{1}{2}\,\|F\|^2 + \delta\,\mathrm{Re}\langle \mathsf{A}F,\, F\rangle$$

defined as in (3) with $\mathsf{D}[F]$ defined by (4), i.e.,

$$\begin{aligned}\mathsf{D}[F] := &-\langle \mathsf{L}F,\, F\rangle + \delta\,\langle \mathsf{A}\mathsf{T}\Pi F,\, F\rangle \\ &-\delta\,\mathrm{Re}\langle \mathsf{T}\mathsf{A}F,\, F\rangle + \delta\,\mathrm{Re}\langle \mathsf{A}\mathsf{T}(\mathrm{Id}-\Pi)F,\, F\rangle - \delta\,\mathrm{Re}\langle \mathsf{A}\mathsf{L}F,\, F\rangle\,.\end{aligned}$$

In other words, we want to estimate the optimal constant $\lambda(\xi)$ in the *entropy–entropy production inequality*

$$\mathsf{D}[F] \ge \lambda(\xi)\,\mathsf{H}_1[F] \tag{30}$$

corresponding to the best possible choice of δ, for a given $\xi \in \mathbb{R}^d$.

If $\mathsf{L} = \mathsf{L}_1$, $\lambda_m = 1$ is given by the Gaussian Poincaré inequality. If $\mathsf{L} = \mathsf{L}_2$, it is straightforward to check that $\lambda_m = 1$. In both cases, it follows from the definition of T that $\lambda_M = |\xi|^2$. With $X := \|(\mathrm{Id}-\Pi)F\|$ and $Y := \|\Pi F\|$, using (25) we have

$$-\langle \mathsf{L}F,\, F\rangle + \delta\,\langle \mathsf{A}\mathsf{T}\Pi F,\, F\rangle \ge X^2 + \frac{\delta\,|\xi|^2}{1+|\xi|^2}\,Y^2\,. \tag{31}$$

By a Cauchy-Schwarz estimate, we know that

$$\left|\xi\cdot\int_{\mathbb{R}^d} w\,F(w)\,dw\right| = |\xi|\,\left|\int_{\mathbb{R}^d}\frac{\xi}{|\xi|}\cdot w\,\sqrt{\mathcal{M}}\,\frac{(\mathrm{Id}-\Pi)F}{\sqrt{\mathcal{M}}}\,dw\right| \le |\xi|\,\|(\mathrm{Id}-\Pi)F\|$$

and therefore obtain that

$$\|\mathsf{A}F\| \le \frac{|\xi|}{1+|\xi|^2}\,\|(\mathrm{Id}-\Pi)F\| \quad\text{and}\quad \|\mathsf{AL}F\| \le \frac{|\xi|}{1+|\xi|^2}\,\|(\mathrm{Id}-\Pi)F\|\,, \tag{32}$$

where the second estimate is a consequence of $\mathsf{AL}F = -\,\mathsf{A}F$ when $\mathsf{L}=\mathsf{L}_1$ or $\mathsf{L}=\mathsf{L}_2$. Notice that the estimate of $\|\mathsf{A}F\|$ is sharper than the one used in the introduction.

Using (32), we have that

$$|\mathrm{Re}\langle \mathsf{A}F,\,F\rangle| \le \frac{|\xi|}{1+|\xi|^2}\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\| \le \frac12\,\frac{|\xi|}{1+|\xi|^2}\,\|F\|^2 \tag{33}$$

and obtain an improved version of (7) given by

$$\frac12\left(1-\frac{\delta\,|\xi|}{1+|\xi|^2}\right)\|F\|^2 \le \mathsf{H}_1[F] \le \frac12\left(1+\frac{\delta\,|\xi|}{1+|\xi|^2}\right)\|F\|^2\,. \tag{34}$$

We also deduce from (32) that

$$\mathsf{H}_1[F] \le \frac12\left(X^2+Y^2\right)+\frac{\delta\,|\xi|}{1+|\xi|^2}\,X\,Y \tag{35}$$

and, using $\mathsf{AL}F = -\,\mathsf{A}F$ and (33),

$$|\mathrm{Re}\langle \mathsf{AL}F,\,F\rangle| \le \frac{|\xi|}{1+|\xi|^2}\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\|\,. \tag{36}$$

As for estimating $\|\mathsf{A}F\|$, by a Cauchy-Schwarz estimate we obtain

$$\|\mathsf{TA}F\| \le \frac{|\xi|^2}{1+|\xi|^2}\,\|(\mathrm{Id}-\Pi)F\|\,,$$

so that

$$\delta\,|\mathrm{Re}\langle \mathsf{TA}F,\,F\rangle| \le \frac{\delta\,|\xi|^2}{1+|\xi|^2}\,X^2\,. \tag{37}$$

As in [15], we can also estimate

$$\|\mathsf{AT}(\mathrm{Id}-\Pi)F\| = \frac{\left|\int_{\mathbb{R}^d}(v'\cdot\xi)^2\,(\mathrm{Id}-\Pi)F(v')\,dv'\right|}{1+|\xi|^2}$$
$$\le \frac{\left(\int_{\mathbb{R}^d}(v'\cdot\xi)^4\mathcal{M}(v')\,dv'\right)^{1/2}}{1+|\xi|^2}\,\|(\mathrm{Id}-\Pi)F\| = \frac{\sqrt{3}\,|\xi|^2}{1+|\xi|^2}\,\|(\mathrm{Id}-\Pi)F\|\,.$$

This inequality and (32) establish that (H4) holds with $C_M = \frac{|\xi|\,(1+\sqrt{3}\,|\xi|)}{1+|\xi|^2}$. Let us finally notice that

$$\delta\,|\mathrm{Re}\langle\mathsf{AT}(\mathrm{Id}-\Pi)F,\,F\rangle| + \delta\,|\mathrm{Re}\langle\mathsf{AL}F,\,F\rangle| \le \delta\,\frac{|\xi|\left(1+\sqrt{3}\,|\xi|\right)}{1+|\xi|^2}\,X\,Y\,. \tag{38}$$

Improved Estimates with Some Plots In this section, our purpose is to provide constructive estimates of the rate λ in Theorem 1 and get improved estimates using various refinements in the mode-by-mode approach. Let us start with the one given in (8).

With $s := |\xi|$, we read from section "Mode-by-Mode Hypocoercivity" that

$$\lambda_m = 1\,,\quad \lambda_M = s^2 \quad\text{and}\quad C_M = \frac{s\left(1+\sqrt{3}\,s\right)}{1+s^2}\,. \tag{39}$$

In that case, the estimate (8) becomes $\lambda \ge \lambda_0(s)$ for $\delta = \delta_0(s)$ with

$$\lambda_0(s) := \frac{1}{3}\,\frac{s^2}{\left(1+\sqrt{3}\,s\right)^2} \quad\text{and}\quad \delta_0(s) := \frac{1}{2}\,\frac{1+s^2}{\left(1+\sqrt{3}\,s\right)^2}\,.$$

With (39) in hand, we can also apply the result of Proposition 2. In order to take into account the dependence on s, the function $h_\star$ has to be replaced by a function h_1 defined by

$$h_1(\delta,\lambda,s) := \delta^2\left(\frac{s\left(1+\sqrt{3}\,s\right)}{1+s^2}+\frac{\lambda}{2}\right)^2 - 4\left(1-\delta-\frac{\lambda}{2}\right)\left(\frac{\delta\,s^2}{1+s^2}-\frac{\lambda}{2}\right),$$

so that the whole game is now reduced, for a given value of $s > 0$, to study the conditions on $(\delta,\lambda) \in \mathcal{T}_m$ such that $h_1(\delta,\lambda,s) \le 0$. In particular, we are interested in computing the largest value $\lambda_1(s)$ of λ for which there exists $\delta > 0$ for which $h_1(\delta,\lambda,s) \le 0$ with $(\delta,\lambda) \in \mathcal{T}_m$, and denote it by $\delta_1(s)$. The triangle $\mathcal{T}_m$ is shown in Fig. 1 and the curves $s \mapsto \lambda_1(s)$ and $s \mapsto \delta_1(s)$ in Figs. 2 and 3. Solutions are numerically contained in $\mathcal{T}_m$ in the sense that $s \mapsto \big(\delta_1(s),\lambda_1(s)\big) \in \mathcal{T}_m$ for any $s >$

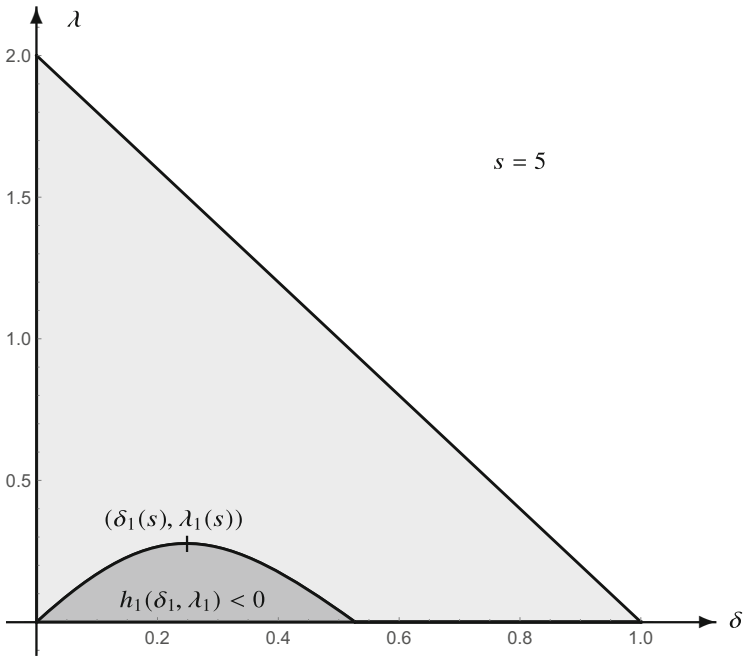

Fig. 1 With λ_m, λ_M and C_M given by (39), the admissible range $\mathcal{T}_m$ of the parameters (δ, λ) is shown in grey for $s = 5$. The darker area is the region in which $h_1(\delta, \lambda, s)$ takes negative values, and $(\delta_1(s), \lambda_1(s))$ are the coordinates of the maximum point of the curve which separates the two regions in the triangle $\mathcal{T}_m$

0. As already noted, some estimates in section "Mode-by-Mode Hypocoercivity" (namely (32), (34), (35), (36) and (37)) are slightly more accurate then the estimates of the proof of Proposition 2. By collecting (31), (35), (37) and (38), we obtain

$$\begin{aligned} &\mathsf{D}[F] - \lambda\,\mathsf{H}_1[F] \\ &\geq \left(1 - \frac{\delta\,s^2}{1+s^2} - \frac{\lambda}{2}\right) X^2 - \frac{\delta\,s}{1+s^2}\left(1 + \sqrt{3}\,s + \lambda\right) X\,Y + \left(\frac{\delta\,s^2}{1+s^2} - \frac{\lambda}{2}\right) Y^2 \end{aligned} \tag{40}$$

is nonnegative for any X and Y under the discriminant condition which amounts to the nonpositivity of

$$h_2(\delta, \lambda, s) := \delta^2\,s^2 \left(\frac{1 + \sqrt{3}\,s + \lambda}{1+s^2}\right)^2 - 4\left(1 - \frac{\delta\,s^2}{1+s^2} - \frac{\lambda}{2}\right)\left(\frac{\delta\,s^2}{1+s^2} - \frac{\lambda}{2}\right),$$

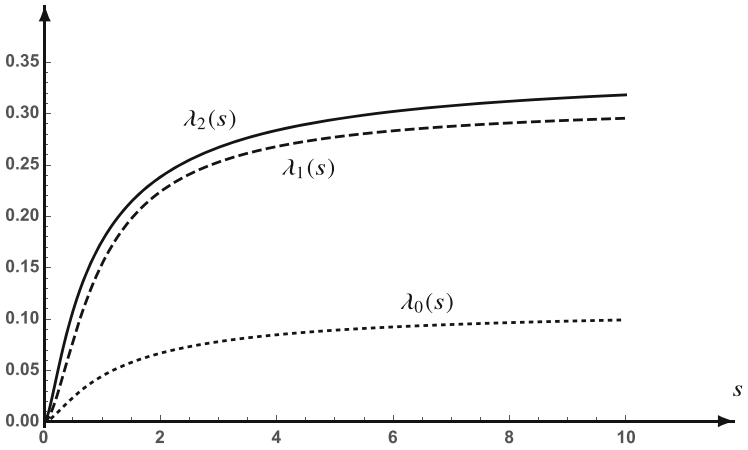

Fig. 2 With λ_m, λ_M and C_M given by (39), curves $s \mapsto \lambda_i(s)$ with $i = 0$, 1 and 2 are shown. The improvement of λ_2 upon λ_0 is of the order of a factor 5

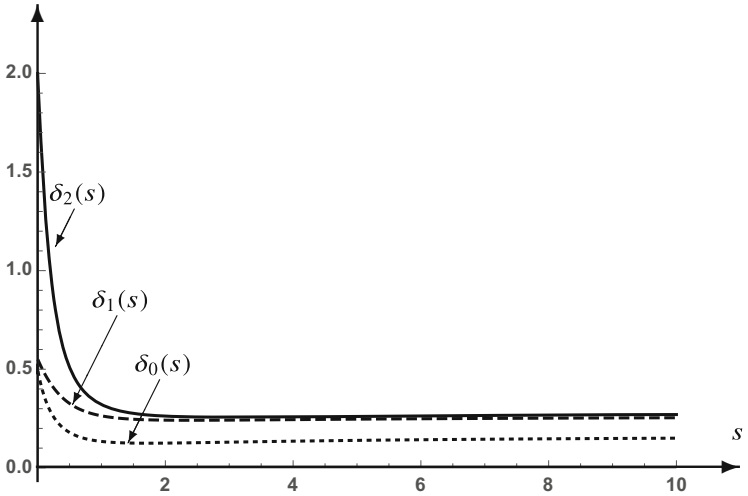

Fig. 3 With λ_m, λ_M and C_M given by (39), curves $s \mapsto \delta_i(s)$ with $i = 0$, 1 and 2 are shown. The dotted curve $s \mapsto \delta_0(s)$ shows the estimate (8) of [15, Proposition 4]. It can be checked numerically that the numerical curves $s \mapsto \big(\delta_i(s), \lambda_i(s)\big)$ with $i = 1, 2$ satisfy the constraints, i.e., stay in their respective triangles for all $s > 0$, as shown in Fig. 1

in the triangle

$$\mathcal{T}_m(s) := \left\{(\delta, \lambda) \in \left(0, \lambda_m \tfrac{1+s^2}{s^2}\right) \times (0, 2\,\lambda_m) \,:\, \lambda < 2\left(\lambda_m - \tfrac{\delta\, s^2}{1+s^2}\right)\right\}$$

with $\lambda_m = 1$. Exactly the same discussion as for $s \mapsto \lambda_1(s)$ and $s \mapsto \delta_1(s)$ determines the curves $s \mapsto \lambda_2(s)$ and $s \mapsto \delta_2(s)$ shown in Figs. 2 and 3. Solutions satisfy $s \mapsto \big(\delta_2(s), \lambda_2(s)\big) \in \mathcal{T}_m(s)$ for any $s > 0$.

3.1.4 Further Observations

In this section, we collect various observations, which are of practical interest, and rely all on the same computations as the ones of Sects. 3.1.2 and 3.1.3.

Explicit Estimates The explicit computation of δ_2 and λ_2 is delicate as it involves finding the roots of high degree polynomials, but it is possible to obtain a very good approximation as follows. After estimating $\lambda\, X\, Y$ by $\lambda\left(X^2 + Y^2\right)/2$, we obtain that

$$\begin{aligned}
&\mathsf{D}[F] - \lambda\, \mathsf{H}_1[F] \\
&\geq \left(1 - \frac{\delta\, s^2}{1+s^2} - \frac{\lambda}{2}\right) X^2 - \frac{\delta\, s}{1+s^2}\left(1 + \sqrt{3}\, s + \lambda\right) X\, Y + \left(\frac{\delta\, s^2}{1+s^2} - \frac{\lambda}{2}\right) Y^2 \\
&\qquad \geq \left(1 - \frac{\delta\, s^2}{1+s^2} - \frac{\lambda}{2}\left(1 + \frac{\delta\, s}{1+s^2}\right)\right) X^2 - \frac{\delta\, s}{1+s^2}\left(1 + \sqrt{3}\, s\right) X\, Y \\
&\qquad\qquad + \left(\frac{\delta\, s^2}{1+s^2} - \frac{\lambda}{2}\left(1 + \frac{\delta\, s}{1+s^2}\right)\right) Y^2 =: \widetilde{Q}(X, Y)
\end{aligned}$$

is nonnegative for any X and Y, under the discriminant condition which amounts to the nonpositivity of

$$\begin{aligned}
&\tilde{h}_2(\delta, \lambda, s) := \delta^2\, s^2 \left(\frac{1 + \sqrt{3}\, s}{1+s^2}\right)^2 \\
&- 4\left(1 - \frac{\delta\, s^2}{1+s^2} - \frac{\lambda}{2}\left(1 + \frac{\delta\, s}{1+s^2}\right)\right)\left(\frac{\delta\, s^2}{1+s^2} - \frac{\lambda}{2}\left(1 + \frac{\delta\, s}{1+s^2}\right)\right). \qquad (41)
\end{aligned}$$

By doing a computation as in section "Improved estimates with some plots", we can find an explicit result, which goes as follows.

Proposition 3 *Assume* (39)*. The largest value of* $\lambda > 0$ *for which there is some* $\delta > 0$ *such that the quadratic form* $\widetilde{Q}$ *is nonnegative is*

$$\tilde{\lambda}_2(s) := \frac{7\, s^2 - \sqrt{21\, s^4 + 4\,(3+5\sqrt{3})\, s^3 + (22+8\sqrt{3})\, s^2 + 4\,(1+\sqrt{3})\, s + 1} + 2\,(1+\sqrt{3})\, s + 1}{7\, s^2 + 2\,(2+\sqrt{3})\, s + 2}$$

with corresponding δ *given by*

$$\tilde{\delta}_2(s) := \frac{s^2+1}{s}\, \frac{\tilde{\lambda}_2(s)^2 - \tilde{\lambda}_2(s) + 2\, s}{7\, s^2 + 2\sqrt{3}\, s + 1 - \tilde{\lambda}_2(s)^2}\,.$$

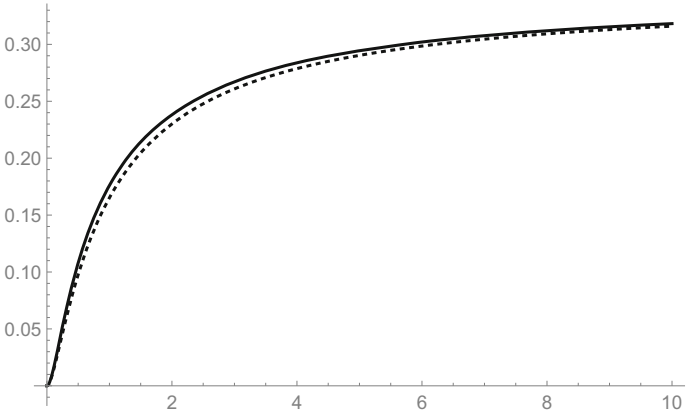

Fig. 4 Plot of $s \mapsto \lambda_2(s)$ and of $s \mapsto \tilde{\lambda}_2(s)$, represented, respectively, by the plain and by the dotted curves

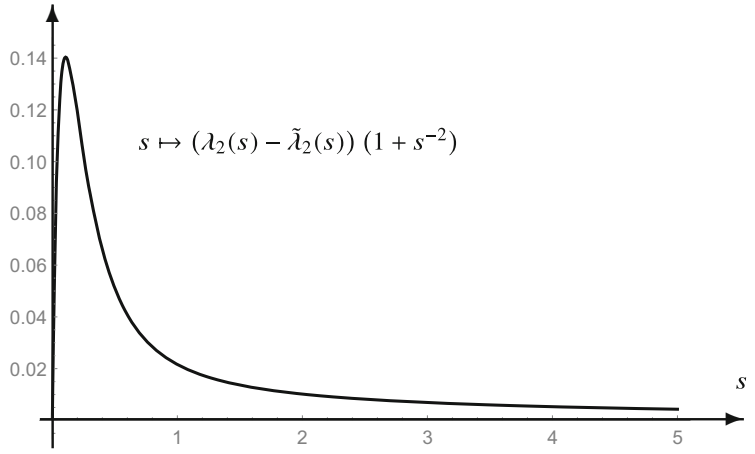

Fig. 5 The curves $s \mapsto \lambda_2(s)$ and $s \mapsto \tilde{\lambda}_2(s)$ have the same asymptotic behaviour as $s \to 0_+$ and as $s \to +\infty$

The proof is tedious but elementary and we shall skip it. By construction, we know that

$$\lambda_2(s) \geq \tilde{\lambda}_2(s) \quad \forall s > 0$$

and the approximation of $\lambda_2(s)$ by $\tilde{\lambda}_2(s)$ is numerically quite good (with a relative error of the order of about 10 %), with exact asymptotics in the limits as $s \to 0_+$, in the sense that $\lambda_2(s)/s^2 \sim \tilde{\lambda}_2(s)/s^2$, and $s \to +\infty$. See Figs. 4 and 5. The approximation of $\delta_2(s)$ by $\tilde{\delta}_2(s)$ is also very good.

Let us summarize some properties which, as a special case, are of interest for Sects. 3.2.2 and "Mode-by-mode diffusion limit".

Lemma 2 *With the notation of Proposition 3, the function $\tilde\lambda_2$ is monotone increasing, the function $\tilde\delta_2$ is monotone increasing for $s > 0$ large enough, and*

$$\lim_{s\to 0_+} \frac{\tilde\lambda_2(s)}{s^2} = 2\,, \quad \lim_{s\to+\infty} \tilde\lambda_2(s) = 1 - \sqrt{3/7} \approx 0.345346 \quad \textit{and} \quad \lim_{s\to+\infty} \tilde\delta_2(s) = 2/7\,.$$

The proof of this result is purely computational and will be omitted here.

Mode-by-Mode Diffusion Limit We consider the diffusion limit which corresponds to the parabolic scaling applied to the abstract equation (1), that is, the limit as $\varepsilon \to 0_+$ of

$$\varepsilon\,\frac{dF}{dt} + \mathsf{T}F = \frac{1}{\varepsilon}\,\mathsf{L}F\,.$$

We will not go to the details and should simply mention that this amounts to replace λ by $\lambda\,\varepsilon$ when we look for a rate λ which is asymptotically independent of ε. We also have to replace (H4) by the assumption

$$\|\mathsf{AT}(\mathrm{Id}-\Pi)F\| + \frac{1}{\varepsilon}\,\|\mathsf{AL}F\| \le C_M^\varepsilon\,\|(\mathrm{Id}-\Pi)F\| \tag{H4$_\varepsilon$}$$

in order to clarify the dependence on ε. Since A, T, Π and L do not depend on ε, this simply means that we can write $C_M^\varepsilon = C_M^{(1)} + \frac{1}{\varepsilon}\,C_M^{(2)}$ where $C_M^{(1)}$ and $C_M^{(2)}$ are the bounds corresponding to

$$\|\mathsf{AT}(\mathrm{Id}-\Pi)F\| \le C_M^{(1)}\,\|(\mathrm{Id}-\Pi)F\| \quad \text{and} \quad \|\mathsf{AL}F\| \le C_M^{(2)}\,\|(\mathrm{Id}-\Pi)F\|\,.$$

With these considerations taken into account, proving an entropy–entropy production inequality is equivalent to proving the nonnegativity of

$$\mathsf{D}[F] - \lambda\,\varepsilon\,\mathsf{H}_1[F]$$
$$\ge \left(\frac{1}{\varepsilon} - \frac{\delta\,s^2}{1+s^2} - \frac{\lambda\,\varepsilon}{2}\right) X^2 - \frac{\delta\,s}{1+s^2}\left(\frac{1}{\varepsilon} + \sqrt{3}\,s + \lambda\,\varepsilon\right) X\,Y + \left(\frac{\delta\,s^2}{1+s^2} - \frac{\lambda\,\varepsilon}{2}\right) Y^2$$

for any X and Y, and the discriminant condition amounts to the nonpositivity of

$$\frac{\delta^2\,s^2}{\varepsilon^2}\left(\frac{1+\sqrt{3}\,\varepsilon\,s+\lambda\,\varepsilon^2}{1+s^2}\right)^2 - 4\left(\frac{1}{\varepsilon} - \frac{\delta\,s^2}{1+s^2} - \frac{\lambda\,\varepsilon}{2}\right)\left(\frac{\delta\,s^2}{1+s^2} - \frac{\lambda\,\varepsilon}{2}\right).$$

In the limit as $\varepsilon \to 0_+$, we find that the optimal choice for λ is given by $(\lambda_\varepsilon(s), \delta_\varepsilon(s))$ with

$$\lim_{\varepsilon \to 0_+} \lambda_\varepsilon(s) = 2\,s^2 \quad \text{and} \quad \delta_\varepsilon(s) = 2\,(1+s^2)\,\varepsilon\,\big(1+o(1)\big)\,.$$

Notice that $\lambda(s) = 2\,s^2$ corresponds to the expected value of the spectrum associated with the heat equation obtained in the diffusion limit. This also corresponds to the limiting behaviour as $s \to 0_+$ of $\tilde{\lambda}_2$ obtained in Lemma 2.

Towards an Optimized Mode-by-Mode Hypocoercivity Approach? In our method, the essential property of the operator $\mathsf{A} := \Big(\mathrm{Id} + (\mathsf{T}\Pi)^*\mathsf{T}\Pi\Big)^{-1}(\mathsf{T}\Pi)^*$ is the equivalence of $\langle \mathsf{A}\mathsf{T}\Pi F, F\rangle$ with $\|\Pi F\|^2$ given by the estimate

$$\frac{\lambda_M}{1+\lambda_M}\,\|\Pi F\|^2 \le \langle \mathsf{A}\mathsf{T}\Pi F, F\rangle \le \|\Pi F\|^2\,.$$

These inequalities arise from the *macroscopic coercivity* condition (H2) and, using the spectral theorem, from the elementary estimate $z/(1+z) \le 1$ for any $z \ge 0$. On the one hand the Lyapunov functional $\mathsf{H}_1[F] := \frac{1}{2}\,\|F\|^2 + \delta\,\mathrm{Re}\langle \mathsf{A}F, F\rangle$ is equivalent to $\|F\|^2$ for $\delta > 0$ small enough because A is a bounded operator. On the other hand, $\mathsf{D}[F] = -\frac{d}{dt}\mathsf{H}_1[F]$ can be compared directly with $\|F\|^2$ because, up to terms that can be controlled, as in the proof of Proposition 2, in the limit as $\delta \to 0_+$, $\mathsf{D}[F]$ is bounded from below by

$$-\langle \mathsf{L}F, F\rangle + \delta\,\langle \mathsf{A}\mathsf{T}\Pi F, F\rangle \ge \lambda_m\,\|(\mathrm{Id}-\Pi)F\|^2 + \frac{\delta\,\lambda_M}{1+\lambda_M}\,\|\Pi F\|^2$$

by Assumptions (H1) and (H2). Notice that this estimate holds for any $\delta > 0$. The choice of $z/(1+z)$ picks a specific scale and one may wonder if $z/(\varepsilon+z)$ would not be a better choice for some value of $\varepsilon > 0$ to be determined. By "better", we simply have in mind to get a larger decay rate as $t \to +\infty$, without trying to optimize on the constant C in (5). It turns out that *the answer is negative*, as ε can be scaled out. Let us give some details.

Let us replace A by

$$\mathsf{A}_\varepsilon := \Big(\varepsilon^2\,\mathrm{Id} + (\mathsf{T}\Pi)^*\mathsf{T}\Pi\Big)^{-1}(\mathsf{T}\Pi)^*$$

for some $\varepsilon > 0$ that can be adjusted, without changing the general strategy, and consider the Lyapunov functional

$$\mathsf{H}_{1,\varepsilon}[F] := \tfrac{1}{2}\,\|F\|^2 + \delta\,\mathrm{Re}\langle \mathsf{A}_\varepsilon F, F\rangle$$

for some $\delta > 0$, so that

$$\begin{aligned}\mathsf{D}_\varepsilon[F] &:= -\frac{d}{dt}\mathsf{H}_{1,\varepsilon}[F]\\ &= -\langle \mathsf{L}F, F\rangle + \delta\,\langle \mathsf{A}_\varepsilon \mathsf{T}\Pi F, F\rangle\\ &\quad -\delta\,\mathrm{Re}\langle \mathsf{T}\mathsf{A}_\varepsilon F, F\rangle + \delta\,\mathrm{Re}\langle \mathsf{A}_\varepsilon \mathsf{T}(\mathrm{Id}-\Pi)F, F\rangle - \delta\,\mathrm{Re}\langle \mathsf{A}_\varepsilon \mathsf{L}F, F\rangle\,.\end{aligned} \tag{42}$$

For a given $\xi \in \mathbb{R}^d$ considered as a parameter, if f solves (21) and if $F = \hat{f}$, then we are back to the framework of Sect. 3.1.3. In this framework, the operator A_ε is given by

$$(\mathsf{A}_\varepsilon F)(v) = -\frac{i\,\xi}{\varepsilon^2 + |\xi|^2}\cdot \int_{\mathbb{R}^d} w\, F(w)\, dw\, \mathcal{M}(v)\,.$$

We have to adapt the computations of section "Mode-by-Mode Hypocoercivity" to $\varepsilon \neq 1$.

As a first remark, we notice that we do not need any estimate of $\|\mathsf{A}_\varepsilon F\|$: all quantities in (42) involving A_ε are directly computed except of $\mathrm{Re}\langle \mathsf{T}\mathsf{A}_\varepsilon F, F\rangle$. Estimating $\mathrm{Re}\langle \mathsf{T}\mathsf{A}_\varepsilon F, F\rangle$ provides a bound which is independent of ε for the following reason. When we solve $G = \mathsf{A}_\varepsilon F$, i.e., if $(\mathsf{T}\Pi)^* F = \varepsilon^2\, G + (\mathsf{T}\Pi)^*\,\mathsf{T}\Pi\, G$, then

$$\langle \mathsf{T}\mathsf{A}_\varepsilon F, F\rangle = \langle G, (\mathsf{T}\Pi)^*\, F\rangle = \varepsilon^2\,\|G\|^2 + \|\mathsf{T}\Pi G\|^2 = \varepsilon^2\,\|\mathsf{A}_\varepsilon F\|^2 + \|\mathsf{T}\mathsf{A}_\varepsilon F\|^2\,.$$

By the Cauchy-Schwarz inequality, we know that

$$\langle G, (\mathsf{T}\Pi)^* F\rangle = \langle \mathsf{T}\mathsf{A}_\varepsilon F, (\mathrm{Id}-\Pi)F\rangle \le \|\mathsf{T}\mathsf{A}_\varepsilon F\|\,\|(\mathrm{Id}-\Pi)F\|\,,$$

which proves that $\|\mathsf{T}\mathsf{A}_\varepsilon F\| \le \|(\mathrm{Id}-\Pi)F\|$ and, as a consequence,

$$|\mathrm{Re}\langle \mathsf{T}\mathsf{A}_\varepsilon F, F\rangle| \le \|(\mathrm{Id}-\Pi)F\|^2\,.$$

It is clear that the right-hand side is independent of $\varepsilon > 0$. A better estimate is obtained by computing as in (37). By doing so, we obtain

$$|\mathrm{Re}\langle \mathsf{T}\mathsf{A}_\varepsilon F, F\rangle| \le \frac{|\xi|^2}{\varepsilon^2 + |\xi|^2}\,\|(\mathrm{Id}-\Pi)F\|^2\,.$$

As in section "Fourier Representation and Mode-by-Mode Estimates", we have $\lambda_m = 1$, $\lambda_M = |\xi|^2$ and the same computations show that

$$|\mathrm{Re}\langle \mathsf{A}_\varepsilon F, F\rangle| \le \frac{|\xi|}{\varepsilon^2+|\xi|^2}\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\|\,,$$

$$|\mathrm{Re}\langle \mathsf{A}_\varepsilon \mathsf{L} F, F\rangle| \le \frac{|\xi|}{\varepsilon^2+|\xi|^2}\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\|\,,$$

$$|\mathrm{Re}\langle \mathsf{A}_\varepsilon \mathsf{T}(\mathrm{Id}-\Pi)F, F\rangle| \le \frac{\sqrt{3}\,|\xi|^2}{\varepsilon^2+|\xi|^2}\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\|\,.$$

This establishes that (H4) holds with

$$C_M = \frac{|\xi|\left(1+\sqrt{3}\,|\xi|\right)}{\varepsilon^2+|\xi|^2}\,,$$

$$\left|\mathsf{H}_{1,\varepsilon}[F] - \frac12\,\|F\|^2\right| \le \delta\,|\mathrm{Re}\langle \mathsf{A}_\varepsilon F, F\rangle| \le \frac{\delta\,|\xi|}{\varepsilon^2+|\xi|^2}\,\|\Pi F\|\,\|(\mathrm{Id}-\Pi)F\|\,,$$

and as a consequence it yields an improved version of (7) which reads

$$\frac12\left(1-\frac{\delta\,|\xi|}{\varepsilon^2+|\xi|^2}\right)\|F\|^2 \le \mathsf{H}_{1,\varepsilon}[F] \le \frac12\left(1+\frac{\delta\,|\xi|}{\varepsilon^2+|\xi|^2}\right)\|F\|^2\,. \tag{43}$$

Notice that the lower bound holds with a positive left-hand side for any ξ only under the additional condition that

$$\delta < 2\,\varepsilon\,.$$

Anyway, if we allow δ to depend on ξ, the whole method still applies, including for proving the hypocoercive estimate on $\|F\|^2$, if the condition

$$\varepsilon^2 - \delta\,|\xi| + |\xi|^2 \ge 0$$

is satisfied for every ξ.

With $X := \|(\mathrm{Id}-\Pi)F\|$, $Y := \|\Pi F\|$, and $s = |\xi|$, we look for the largest value of λ for which the right-hand side in

$$\begin{aligned}&\mathsf{D}_\varepsilon[F] - \lambda\,\mathsf{H}_{1,\varepsilon}[F]\\ &\ge \left(1-\frac{\delta\,s^2}{\varepsilon^2+s^2}-\frac{\lambda}{2}\right)X^2 - \frac{\delta\,s}{\varepsilon^2+s^2}\left(1+\sqrt{3}\,s+\lambda\right)X\,Y + \left(\frac{\delta\,s^2}{\varepsilon^2+s^2}-\frac{\lambda}{2}\right)Y^2\end{aligned} \tag{44}$$

is nonnegative for any X and Y. Recall that s is fixed and δ is a parameter to be adjusted. If we change the parameter δ into δ_* such that

$$\frac{\delta\, s}{\varepsilon^2 + s^2} = \frac{\delta_*\, s}{1 + s^2}\,, \tag{45}$$

then the nonnegativity problem of the r.h.s. in (44) is reduced to the same problem with $\varepsilon = 1$, provided that no additional constraint is added. Let us define

$$h_3(\delta, \lambda, \varepsilon, s) := \delta^2\, s^2 \left(\frac{1 + \sqrt{3}\, s + \lambda}{\varepsilon^2 + s^2}\right)^2 - 4\left(1 - \frac{\delta\, s^2}{\varepsilon^2 + s^2} - \frac{\lambda}{2}\right)\left(\frac{\delta\, s^2}{\varepsilon^2 + s^2} - \frac{\lambda}{2}\right)$$

with (δ, λ) in the triangle

$$\mathcal{T}_m^\varepsilon(s) := \left\{(\delta, \lambda) \in \left(0, \left(1 + \varepsilon^2\, s^{-2}\right)\lambda_m\right) \times (0, 2\,\lambda_m) \,:\, \lambda < 2\left(\lambda_m - \tfrac{\delta\, s^2}{\varepsilon^2 + s^2}\right)\right\}$$

and $\lambda_m = 1$. Exactly the same method as in section "Mode-by-Mode Hypocoercivity" determines the curves $s \mapsto \lambda_3(s)$ and $s \mapsto \delta_3(s, \varepsilon)$, but we have $\lambda_3(s) = \lambda_2(s)$ for any $s > 0$ while $\delta_2(s)$ and $\delta_3(s, \varepsilon)$ can be deduced from each other using (45). Solutions have to satisfy the constraint $s \mapsto \big(\delta_3(s, \varepsilon), \lambda_3(s)\big) \in \mathcal{T}_m^\varepsilon(s)$ for any $s > 0$. It is straightforward to check that $(\delta, \lambda) \in \mathcal{T}_m^\varepsilon(s)$ if and only if $(\delta_*, \lambda) \in \mathcal{T}_m^1(s) = \mathcal{T}_m(s)$, where δ_* is determined by (45). Altogether, our observations can be reformulated as follows.

Lemma 3 *Assume* (39). *Then for any $s > 0$, we have*

$$\max\left\{\lambda > 0 \,:\, (\delta, \lambda) \in \mathcal{T}_m^\varepsilon(s)\,,\ h_3(\delta, \lambda, \varepsilon, s) \le 0\right\}$$

is independent of $\varepsilon > 0$.

To conclude this subsection, we note that, while the Lyapunov functionals $\mathsf{H}_{1,\varepsilon}$ are clearly different for different values of $\varepsilon > 0$, mode-by-mode, i.e., for a given value of $s = |\xi|$, they all yield the same exponential decay rate $\lambda = \lambda_2(s)$, when choosing the best parameter $\delta = \delta_3(s, \varepsilon)$. Similarly, no improvement on the constant C as in (5) is achieved by adjusting $\varepsilon > 0$ when ε is taken into account in (43), for proving the equivalence of $\mathsf{H}_{1,\varepsilon}[F]$ and $\|F\|^2$. This reflects a deep scaling invariance of the method.

3.2 Convergence Rates and Decay Rates

In this section, we come back to the study of (21) and consider two situations. A periodic solution on a *small torus* has a behaviour driven by high frequencies corresponding to $|\xi|$ large, while the decay rate of a solution on the whole Euclidean

space is asymptotically determined by the low frequency regime with $\xi \to 0$. In the latter case, we use estimates as in Nash type inequalities and relate the time decay with the behaviour of $\lambda(\xi)$ in a neighbourhood of $\xi = 0$.

3.2.1 Exponential Convergence Rate on a Small Torus

Let us assume that $\mathcal{X} = [0, L)^d$ (with periodic boundary conditions) and consider the limit as $L \to 0_+$. With the notation of Sect. 3.1.1 and the Fourier transform (22), the periodicity implies that $\xi \in (2\pi/L)\,\mathbb{Z}^d$ and in particular, for any fixed $j \in \mathbb{Z}^d$ and $\xi = 2\pi\, j/L$, we have $|\xi| \to +\infty$ as $L \to 0_+$, unless $j = 0$. Let us denote by $\lambda_L(\xi)$ the optimal constant in (30) when ξ is limited to $(2\pi/L)\,\mathbb{Z}^d \setminus \{0\}$. We recall that

$$\lambda_\star := \liminf_{L\to 0_+} \inf_{\xi\in(2\pi/L)\,\mathbb{Z}^d\setminus\{0\}} \lambda_L(\xi) \ge 1 - \sqrt{3/7}$$

according to Lemma 2. As a consequence, we have the following result.

Proposition 4 *For any $\varepsilon > 0$, small, there exists some $L_\varepsilon > 0$ such that, if $\mathcal{X} = [0, L)^d$ for an arbitrary $L \le L_\varepsilon$, if f solves (21) with $f_0 \in \mathrm{L}^2(\mathcal{X}\times\mathbb{R}^d, dx\,d\gamma)$ and $\mathsf{L} = \mathsf{L}_1$ or $\mathsf{L} = \mathsf{L}_2$, then we have*

$$\|f(t,\cdot,\cdot) - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(dx\,d\gamma)} \le (1+\varepsilon)\,\|f_0 - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(dx\,d\gamma)}\, e^{-\min\{2,\lambda_\star-\varepsilon\}\,t} \quad \forall\, t \ge 0\,,$$

with $\bar f := \frac{1}{L^d}\iint_{\mathcal{X}\times\mathbb{R}^d} f_0(x,v)\,dx\,dv$.

Proof Let us notice that $g(t,v) := \hat f(t,0,v) = \int_{\mathcal{X}} f(t,x,v)\,dx$ solves

$$\partial_t g = \mathsf{L} g\,.$$

As a consequence either of the definition of $\mathsf{L} = \mathsf{L}_2$, or of the Gaussian Poincaré inequality

$$\|g - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(d\gamma)} \le \|\nabla g\|^2_{\mathrm{L}^2(d\gamma)}$$

if $\mathsf{L} = \mathsf{L}_1$, we know that

$$\|g(t,\cdot) - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(d\gamma)} \le \|g(0,\cdot) - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(d\gamma)}\, e^{-2t} \quad \forall\, t \ge 0\,.$$

By the Plancherel formula, we have

$$\|f(t,\cdot,\cdot) - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(dx\,d\gamma)} = \|g(t,\cdot) - \bar f\,\mathcal{M}\|^2_{\mathrm{L}^2(d\gamma)} + (2\pi)^{-d} \sum_{\xi\in(2\pi/L)\,\mathbb{Z}^d\setminus\{0\}} \|\hat f(t,\xi,\cdot)\|^2_{\mathrm{L}^2(d\gamma)}\,.$$

The conclusion follows from $C(s) = \left(1 + s^2 + \tilde{\delta}_2(s)\, s\right) / \left(1 + s^2 - \tilde{\delta}_2(s)\, s\right)$,

$$\|\hat{f}(t,\xi,\cdot)\|^2_{\mathrm{L}^2(d\gamma)} \le C(|\xi|)\, \|\hat{f}_0(\xi,\cdot)\|^2_{\mathrm{L}^2(d\gamma)}\, e^{-\lambda_L(\xi)\, t}$$

for any $(t,\xi) \in \mathbb{R}^+ \times (2\,\pi/L)\,\mathbb{Z}^d \setminus \{0\}$, and the estimates of Lemma 2. □

3.2.2 Algebraic Decay Rate in the Whole Euclidean Space

As a refinement of [15], we investigate the decay estimates for the solution to (21) on $\mathcal{X} = \mathbb{R}^d$. Here we rely on *Nash type estimates*.

To start with, let us consider a model problem. Assume that $s \mapsto \lambda(s)$ is a positive non-decreasing bounded function on $(0, +\infty)$ and, for any $s > 0$, let

$$h_\lambda(M, R, s) := \lambda(R) \left(\omega_d\, R^d\, M^2 - s\right), \quad \lambda^*(M, s) := -\min_{R>0} h_\lambda(M, R, s),$$

where M is a positive parameter and $\omega_d = |\mathbb{S}^{d-1}|/d$. Since $h_\lambda(M, R, s) \sim -\lambda(R)\, s$ as $R \to 0_+$ and $h_\lambda(M, R, s) \ge c\, R^d$ for some $c > 0$ as $R \to +\infty$, there is indeed some $R > 0$ such that $\lambda^*(M, s) = -\,h_\lambda(M, R, s)$ and $\lambda^*(M, s)$ is positive for any $(M, s) \in (0, +\infty)^2$. We also define the monotone decreasing function

$$\psi_{\lambda,M}(s) := -\int_1^s \frac{dz}{\lambda^*(M, z)} \quad \forall\, s \ge 0\,.$$

Our first result is a decay rate on $\mathbb{R}^d$ for a solution of $\partial_t u = \mathcal{L} u$ where the operator $\mathcal{L}$ acts on the Fourier space as the multiplication of $\xi \mapsto \hat{u}(\xi)$ with some scalar function $-\lambda(\xi)/2$, for any $\xi \in \mathbb{R}^d$. With just a spectral inequality, we obtain the following estimate.

Lemma 4 *Assume that $s \mapsto \lambda(s)$ is a positive non-decreasing bounded function on $(0, +\infty)$ such that, with the above notation, $\lim_{s\to 0_+} \psi_{\lambda,\mu}(s) = +\infty$ for all $\mu > 0$. If $u \in C(\mathbb{R}^+, \mathrm{L}^1 \cap \mathrm{L}^2(dx))$ is such that $M := \|u(t,\cdot)\|_{\mathrm{L}^1(dx)}$ does not depend on t and*

$$\frac{d}{dt}|\hat{u}(t,\xi)|^2 \le -\,\lambda(|\xi|)\, |\hat{u}(t,\xi)|^2 \quad \forall\, (t,\xi) \in \mathbb{R}^+ \times \mathbb{R}^d\,,$$

then

$$\|u(t,\cdot)\|^2_{\mathrm{L}^2(dx)} \le \psi^{-1}_{\lambda,M}\left(t + \psi_{\lambda,M}\left(\|u(0,\cdot)\|^2_{\mathrm{L}^2(dx)}\right)\right) \quad \forall\, t \in \mathbb{R}^+\,.$$

Here $\hat{u}$ denotes the Fourier transform of u in x.

Proof The inspiration for the proof comes from [29, page 935]. Let

$$y(t) := \int_{\mathbb{R}^d} |\hat{u}(t,\xi)|^2\, d\xi \le \int_{|\xi|\le R} \|\hat{u}(t,.)\|^2_{\mathrm{L}^\infty(d\xi)}\, d\xi + \frac{1}{\lambda(R)} \int_{\mathbb{R}^d} \lambda(|\xi|)\, |\hat{u}(t,\xi)|^2 d\xi$$
$$\le \omega_d\, R^d\, \|u(t,.)\|^2_{\mathrm{L}^1(dx)} - \frac{1}{\lambda(R)}\, \frac{d}{dt} \int_{\mathbb{R}^d} |\hat{u}(t,\xi)|^2\, d\xi\,.$$

Hence,

$$y' \le h_\lambda(M, R, y)\,,$$

for any $R > 0$. Taking the minimum of the r.h.s. over $R > 0$, we obtain

$$y' \le -\lambda^*(M, y)\,,$$

and the conclusion follows after elementary computations. □

Example 1 Let us consider a case that we have already encountered in Sects. 3.2.2 and "Mode-by-mode diffusion limit". If $\lambda(s) = 2\, s^2$ for $s \in (0,1)$, we find that

$$\lambda^*(M,s) = 2\, d \left(\frac{2}{\omega_d\, M^2}\right)^{\frac{2}{d}} \left(\frac{s}{d+2}\right)^{1+\frac{2}{d}} \qquad \forall\, s \in (0,1)\,.$$

With $c_d := \frac{1}{2}\left((d+2)/2\right)^{1+2/d} \omega_d^{2/d}$, we find

$$\psi_{\lambda,M}(s) = c_d\, M^{\frac{4}{d}} \left(s^{-\frac{2}{d}} - 1\right)$$

and deduce from Lemma 4 that

$$\|u(t,\cdot)\|^2_{\mathrm{L}^2(dx)} \le \left(\|u(0,\cdot)\|^{-\frac{4}{d}}_{\mathrm{L}^2(dx)} + \frac{t}{c_d}\, \|u(0,\cdot)\|^{-\frac{4}{d}}_{\mathrm{L}^1(dx)}\right)^{-\frac{d}{2}} \qquad \forall\, t \in \mathbb{R}^+\,.$$

This estimate is similar to the estimate that one would deduce from Nash's inequality in the case of the heat equation on $\mathbb{R}^d$. Notice that differentiating this estimate at $t = 0$ gives a proof of Nash's inequality, with the same constant as in Nash's proof in [29] (see [17] for a discussion of the optimal constant).

The assumption on u in Lemma 4 is a coercivity estimate for the operator $\mathcal{L}$ with Fourier representation $-\lambda(|\xi|)/2$, which allows us to use a Bihari-LaSalle estimate, i.e., a nonlinear version of Grönwall's lemma. For the main application in this paper, we have to rely on a hypocoercivity estimate, which is slightly more complicated.

Lemma 5 *Assume that $s \mapsto \lambda(s)$ is a positive non-decreasing bounded function on $(0,+\infty)$ such that, with the above notation, $\lim_{s\to 0_+} \psi_{\lambda,\mu}(s) = +\infty$ for all $\mu > 0$. Let $u \in C(\mathbb{R}^+, \mathrm{L}^1 \cap \mathrm{L}^2(dx))$ be such that, for some bounded continuous function $s \mapsto C(s)$ such that $C(s) \geq 1$ for any $s > 0$,*

$$|\hat{u}(t,\xi)|^2 \leq C(|\xi|)\,|\hat{u}(0,\xi)|^2\,e^{-\lambda(|\xi|)\,t} \quad \forall\,(t,\xi) \in \mathbb{R}^+ \times \mathbb{R}^d$$

and $\|u(t,\cdot)\|_{\mathrm{L}^1(dx)} \leq M$ for some M which does not depend on t. Then, for any $t \geq 0$, we have

$$\|u(t,\cdot)\|^2_{\mathrm{L}^2(dx)} \leq \Psi_{M,Q}(t)\,, \tag{46}$$

where $Q := \|u(0,\cdot)\|_{\mathrm{L}^2(dx)}$ and

$$\Psi_{M,Q}(t) := \inf_{R>0}\left(\int_0^R C(s)\,e^{-\lambda(s)\,t}\,s^{d-1}\,ds\,\omega_d\,d\,M^2 + \sup_{s\geq R} C(s)\,e^{-\lambda(R)\,t}\,Q^2\right). \tag{47}$$

Proof For any $R > 0$, we have

$$\int_{|\xi|\leq R} |\hat{u}(t,\xi)|^2\,d\xi \leq \int_{|\xi|\leq R} C(|\xi|)\,e^{-\lambda(|\xi|)\,t}\,d\xi\,\|\hat{u}(0,\cdot)\|^2_{\mathrm{L}^\infty(\mathbb{R}^d,d\xi)}$$

with $\|\hat{u}(0,\cdot)\|_{\mathrm{L}^\infty(\mathbb{R}^d,d\xi)} \leq \|u(0,\cdot)\|_{\mathrm{L}^1(\mathbb{R}^d,dx)}$ on the one hand, and

$$\int_{|\xi|>R} |\hat{u}(t,\xi)|^2\,d\xi \leq \sup_{s>R} C(s)\,e^{-\lambda(R)\,t}\|\hat{u}(0,\cdot)\|^2_{\mathrm{L}^2(\mathbb{R}^d,d\xi)}$$

on the other hand. The result follows by optimizing on $R > 0$. □

The result of Lemma 5 is not as explicit as the result of Lemma 4, but it is useful to investigate, for instance, the limit as $t \to +\infty$: if $\lim_{s\to 0_+} C(s) = C(0) > 0$ and $\lambda(s) = 2\,s^2$ for any $s \in (0,1)$, then one can prove that

$$\|u(t,\cdot)\|^2_{\mathrm{L}^2(dx)} \leq O\left(t^{-d/2}\right) \quad \text{as} \quad t \to +\infty\,.$$

In the spirit of [15], let us draw some consequences for the solution of (21).

Theorem 2 *If f solves (21) for some nonnegative initial datum $f_0 \in \mathrm{L}^2(\mathbb{R}^d \times \mathbb{R}^d, dx\,d\gamma) \cap \mathrm{L}^2\big(\mathbb{R}^d, d\gamma; \mathrm{L}^1(\mathbb{R}^d, dx)\big)$ and $\mathsf{L} = \mathsf{L}_1$ or $\mathsf{L} = \mathsf{L}_2$, then we have the estimate*

$$\|f(t,\cdot,\cdot)\|^2_{\mathrm{L}^2(\mathbb{R}^d\times\mathbb{R}^d, dx\,d\gamma)} \le (2\pi)^{-d}\,\Psi_{M,Q}(t)$$

with $M = \|f_0\|_{\mathrm{L}^2(\mathbb{R}^d, d\gamma; \mathrm{L}^1(\mathbb{R}^d, dx))}$, $Q = \|f_0\|_{\mathrm{L}^2(\mathbb{R}^d\times\mathbb{R}^d, dx\,d\gamma)}$, and $\Psi_{M,Q}(t)$ defined by (47) using $C(s) = (2+\delta(s))/(2-\delta(s))$ and $\lambda(s)$, for any pair (δ,λ) of continuous functions on $(0,+\infty)$ taking values in $(0,2)\times(0,+\infty)$, with $s \mapsto \lambda(s)$ monotone non-decreasing, such that the entropy–entropy production inequality (30) *and the equivalence (34) hold.*

Here we abusively write $\lambda(\xi) = \lambda(s)$ and $\delta(\xi) = \lambda(s)$ with $s = |\xi|$.

Proof We estimate $\|f(t,\cdot,\cdot)\|^2_{\mathrm{L}^2(\mathbb{R}^d\times\mathbb{R}^d, dx\,d\gamma)}$ using (22) and Plancherel's theorem

$$\iint_{\mathbb{R}^d\times\mathbb{R}^d} |f(t,x,v)|^2\,dx\,d\gamma = \frac{1}{(2\pi)^d}\iint_{\mathbb{R}^d\times\mathbb{R}^d} |\hat f(t,\xi,v)|^2\,d\xi\,d\gamma\,.$$

Applying the results of Theorem 1 with

$$C(\xi) = \frac{1+|\xi|^2+\delta(\xi)\,|\xi|}{1+|\xi|^2-\delta(\xi)\,|\xi|}\,,$$

we learn that

$$\int_{\mathbb{R}^d} |\hat f(t,\xi,v)|^2\,d\gamma \le C(\xi)\int_{\mathbb{R}^d} |\hat f_0(\xi,v)|^2\,d\gamma\,e^{-\lambda(\xi)\,t}\,.$$

We can apply the same strategy as for Lemma 5, with

$$\iint_{B_R\times\mathbb{R}^d} C(\xi)\,|\hat f_0(\xi,v)|^2\,e^{-\lambda(\xi)\,t}\,d\xi\,d\gamma \le \int_{|\xi|\le R} C(\xi)\,e^{-\lambda(\xi)\,t}\,d\xi\,M^2\,,$$

$$\iint_{B_R^c\times\mathbb{R}^d} C(\xi)\,|\hat f_0(\xi,v)|^2\,e^{-\lambda(\xi)\,t}\,d\xi\,d\gamma \le \sup_{\xi\in B_R^c} C(\xi)\,e^{-\lambda(R)\,t}\,Q^2\,,$$

using $\sup_{\xi\in\mathbb{R}^d}|\hat f_0(\xi,v)| \le \int_{\mathbb{R}^d} f_0(x,v)\,dx$ for the first inequality, and the monotonicity of λ. □

In practice, any good estimate, for instance the estimate based on the functions $(\tilde\delta_2,\tilde\lambda_2)$ of Proposition 3, provides us with explicit and constructive decay rates of the solution to (21) on $\mathbb{R}^d$. As a concluding remark, it has to be made clear that the method is not limited to the operators L_1 and L_2.

4 The Goldstein–Taylor Model

4.1 General Setting and Fourier Decomposition

Consider the *two velocities Goldstein–Taylor (GT) model* (*cf.* [21, § 1.4]) with constant relaxation coefficient $\sigma > 0$, position variable $x \in \mathcal{X} \subseteq \mathbb{R}$, and $t > 0$:

$$\begin{aligned} \partial_t f_+(t,x) + \partial_x f_+(t,x) &= \frac{\sigma}{2}\big(f_-(t,x) - f_+(t,x)\big)\,, \\ \partial_t f_-(t,x) - \partial_x f_-(t,x) &= -\frac{\sigma}{2}\big(f_-(t,x) - f_+(t,x)\big)\,, \\ f_\pm(x,0) &= f_{\pm,0}(x)\,, \end{aligned} \tag{48}$$

where $f_\pm(t,x)$ are the density functions of finding a particle with a velocity ± 1 in a position x at time $t > 0$ and $f_{\pm,0} \in \mathrm{L}^1_+(\mathcal{X})$ is the initial configuration. This model is the prime example of discrete velocity BGK equations, as described in Sect. 2.2, Example 2, with $b = (1/2, 1/2)^T$ and $V = \mathrm{diag}(1,-1)$. We consider two situations for $\mathcal{X}$, the one-dimensional torus and the real line, i.e., $\mathcal{X} \in \{\mathbb{T}, \mathbb{R}\}$.

Rewriting (48) in the macroscopic variables of (mass and flux densities)

$$u(t,x) := f_+(t,x) + f_-(t,x) \geq 0\,, \quad v(t,x) := f_+(t,x) - f_-(t,x)\,,$$

leads to the transformed equations

$$\begin{aligned} \partial_t u(t,x) &= -\,\partial_x v(t,x)\,, \\ \partial_t v(t,x) &= -\,\partial_x u(t,x) - \sigma\, v(u,x)\,, \end{aligned} \tag{49}$$

for $x \in \mathcal{X}$, $t \geq 0$. Integrating these equations along $\mathcal{X}$ directly shows that the total mass is conserved for all times, i.e. $\int_{\mathcal{X}} u(t,x)\,dx \equiv \int_{\mathcal{X}} u(x,0)\,dx$, and that the total flux is decaying exponentially, i.e. $\int_{\mathcal{X}} v(t,x)\,dx = e^{-\sigma t}\int_{\mathcal{X}} v(x,0)\,dx$ for $t \geq 0$.

A Fourier transformation in the space variable $x \in \mathcal{X}$ leads to ODEs of form (10), given explicitly as

$$\partial_t \hat{y}(t,\xi) = -C(\xi,\sigma)\,\hat{y}(t,\xi) \tag{50}$$

with

$$\hat{y}(t,\xi) := \begin{pmatrix} \hat{u}(t,\xi) \\ \hat{v}(t,\xi) \end{pmatrix} \quad \text{and} \quad C(\xi,\sigma) := \begin{pmatrix} 0 & i\,\xi \\ i\,\xi & \sigma \end{pmatrix}\,,$$

for the Fourier modes $\xi \in \mathbb{Z}$ in the case of $\mathcal{X} = \mathbb{T}$, and $\xi \in \mathbb{R}$ for $\mathcal{X} = \mathbb{R}$.

The matrix $C(\xi,\sigma)$ from (50) has the eigenvalues

$$\lambda_{\pm}(\xi,\sigma) := \frac{\sigma}{2} \pm \sqrt{\frac{\sigma^2}{4} - \xi^2}$$

and hence its *modal spectral gap* is given by

$$\mu(\xi,\sigma) := \mathrm{Re}\left(\frac{\sigma}{2} - \sqrt{\frac{\sigma^2}{4} - \xi^2}\right), \quad \xi \neq 0\,. \tag{51}$$

For $\mathcal{X} = \mathbb{R}$, the modal spectral gap takes all values in the interval $(0, \sigma/2]$ with $\lim_{\xi\to 0}\mu(\xi,\sigma) = 0$. To obtain decay estimates with the sharp decay rate of solutions $y(t,\xi)$ to (49) it is therefore important to achieve precise estimates of the decay behavior as $\xi \to 0$. For $\mathcal{X} = \mathbb{T}$, the spectral gap for solutions to (49) corresponds to the *uniform-in-$\mathbb{Z}$ spectral gap*, i.e.

$$\overline{\mu}(\sigma) := \min_{\xi\in\mathbb{Z}\setminus\{0\}} \mu(\xi,\sigma)\,.$$

The set of modal spectral gaps which coincide with the uniform spectral gap is denoted by $\Xi(\sigma)$ and depends on the values of $\sigma > 0$:

- For $\sigma \in (0, 2]$ it follows that

$$\overline{\mu}(\sigma) = \frac{\sigma}{2}, \quad \Xi(\sigma) = \mathbb{Z}\setminus\{0\}\,.$$

- For $\sigma > 2$ the lowest modes determine the uniform-in-$\mathbb{Z}$ spectral gap,

$$\overline{\mu}(\sigma) = \mu(\pm 1,\sigma) = \frac{\sigma}{2} - \sqrt{\frac{\sigma^2}{4} - 1}\,, \quad \Xi = \{-1, 1\}\,. \tag{52}$$

Now, we consider the two hypocoercivity methods from Sects. 2.1 and 2.2 for solutions $\hat{y}(t,\xi)$ of (50) for fixed but arbitrary modes ξ.

Approach of Sect. 2.2 For equations of form (10), we consider the modal Lyapunov functionals $\|\hat{y}(t,\xi)\|^2_{P(\xi,\sigma)}$ with deformation matrices $P(\xi,\sigma)$. These functionals satisfy the explicit estimates of form (16), which go as follows:

- For fixed $|\xi| \neq \sigma/2$, $|\xi| > 0$ the matrix $C(\xi,\sigma)$ is not defective and it follows from Lemma 1 that

$$\|\hat{y}(t,\xi)\|^2_{P(\xi,\sigma)} \leq e^{-2\,\mu(\xi,\sigma)\,t}\,\|\hat{y}(\xi,0)\|^2_{P(\xi,\sigma)}\,, \tag{53}$$

with $P(\xi,\sigma) = P^{(1)}(\xi,\sigma)$ for $|\xi| > \sigma/2$ and $P(\xi,\sigma) = P^{(2)}(\xi,\sigma)$ for $|\xi| < \sigma/2$, where

$$P^{(1)}(\xi,\sigma) := \begin{pmatrix} 1 & -\frac{i\sigma}{2\xi} \\ \frac{i\sigma}{2\xi} & 1 \end{pmatrix}, \quad P^{(2)}(\xi,\sigma) := \begin{pmatrix} 1 & -\frac{2i\xi}{\sigma} \\ \frac{2i\xi}{\sigma} & 1 \end{pmatrix}. \tag{54}$$

- For $|\xi| = \sigma/2$ the matrix $C(\xi,\sigma)$ is defective. Then, due to [9, Lemma 4.3], for any $\varepsilon > 0$ there exists an ε-dependent matrix that yields the purely exponential decay $\mu(\sigma/2) - \varepsilon$. For later purposes it will be sufficient to investigate the case $\sigma = 2$ with $\xi = 1$, see Sect. 4.2.2. Hence, we will not state the general form here.

Approach of Sect. 2.1 With notation from Theorem 1, the Goldstein–Taylor equation in Fourier modes (50) can be written as

$$\partial_t \hat{y}(t,\xi) = \big(\mathsf{L}(\sigma) - \mathsf{T}(\xi)\big)\, \hat{y}(t,\xi)\,.$$

The Hermitian collision matrix and the anti-Hermitian transport matrix are, respectively, given as

$$\mathsf{L}(\sigma) := \begin{pmatrix} 0 & 0 \\ 0 & -\sigma \end{pmatrix}, \quad \mathsf{T}(\xi) := \begin{pmatrix} 0 & i\,\xi \\ i\,\xi & 0 \end{pmatrix}.$$

The projection on the space of local-in-x equilibria (satisfying $\mathsf{L}(\sigma)\Pi = 0$) is given by the matrix

$$\Pi := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We introduce the operator $\mathsf{A}(\xi)$ as in (2) for each mode ξ:

$$\mathsf{A}(\xi) := \Big(\mathrm{Id} + \big(\mathsf{T}(\xi)\,\Pi\big)^*\,\mathsf{T}(\xi)\,\Pi\Big)^{-1} \big(\mathsf{T}(\xi)\,\Pi\big)^* = \begin{pmatrix} 0 & -\frac{i\,\xi}{1+\xi^2} \\ 0 & 0 \end{pmatrix}.$$

The modal Lyapunov functional (3) is given as

$$\begin{aligned} \mathsf{H}_1(\xi,\delta)[\hat{y}(\xi)] &:= \frac{1}{2}\,\|\hat{y}(\xi)\|^2 + \delta\,\mathrm{Re}\big(\hat{y}(\xi)^*\,\mathsf{A}(\xi)\,\hat{y}(\xi)\big) \\ &= \frac{1}{2}\,\|\hat{y}(\xi)\|^2 + \delta\hat{y}(\xi)^*\mathsf{A}_H(\xi)\,\hat{y}(\xi) \\ &= \frac{1}{2}\,\hat{y}(\xi)^* \begin{pmatrix} 1 & -\frac{i\,\xi\,\delta}{1+\xi^2} \\ \frac{i\,\xi\,\delta}{1+\xi^2} & 1 \end{pmatrix} \hat{y}(\xi)\,, \end{aligned} \tag{55}$$

where we denote the Hermitian part of the matrix A by $\mathsf{A}_H := \frac{1}{2}(\mathsf{A} + \mathsf{A}^*)$.

4.2 Comparison of the Two Hypocoercivity Methods for $\mathcal{X} = \mathbb{T}$

In the next step we shall assemble, for both hypocoercivity methods, the modal Lyapunov functionals to form a global one. When appropriately optimizing both of these functionals, we shall see that they actually coincide and achieve optimal decay estimates in the class of all quadratic forms.

4.2.1 The Optimal Global Lyapunov Functional

We start by applying the strategies outlined in Sect. 2.2 to assemble a global $\mathrm{L}^2(\mathbb{T})$ functional. For simplicity, let us first assume that the matrix $C(\xi, \sigma)$, $\xi \in \mathbb{Z}$ of (50) is diagonalizable for all modes, i.e. $\sigma \notin 2\,\mathbb{Z}$. A brief discussion of the defective cases is deferred to the end of this section.

We first consider Strategy 1 of Sect. 2.2 that leads to the functional

$$\mathsf{H}_2[y] := \sum_{|\xi|>\sigma/2} \|\hat{y}(\xi)\|^2_{P^{(1)}(\xi)} + \sum_{|\xi|<\sigma/2} \|\hat{y}(\xi)\|^2_{P^{(2)}(\xi)}\,, \quad y \in (\mathrm{L}^2(\mathbb{T}))^2\,, \tag{56}$$

according to definition (17). Assuming that the system has total mass 0, i.e. $\int_{\mathbb{T}} u(x,0)\,dx = 0$, we obtain that solutions $y(t)$ of (49), (50) satisfy the estimate:

$$\|y(t)\|^2 \le \overline{c}_P\, e^{-2\,\overline{\mu}(\sigma)\,t}\, \|y(0)\|^2\,, \tag{57}$$

where

$$\bar{c}_P := \max\left\{ \sup_{|\xi|>\sigma/2} \left[\mathrm{cond}\left(P^{(1)}(\xi)\right)\right], \sup_{|\xi|<\sigma/2} \left[\mathrm{cond}\left(P^{(2)}(\xi)\right)\right] \right\}. \tag{58}$$

To improve upon the multiplicative constant $\overline{c}_P$ in (57), we continue with Strategy 2 of Sect. 2.2.

- For the case $\sigma < 2$ the functional H_2 and (58) directly yield the optimal multiplicative constant. With notation from Sect. 2.2 this follows from $\Xi = \mathbb{Z} \setminus \{0\}$ and $\overline{c}_P = c_\Xi = \mathrm{cond}\big(P^{(1)}(\pm 1)\big) = (2+\sigma)/(2-\sigma)$. In this case the two eigenvalues of the ODE system matrix $C(\xi)$ are distinct and form a complex conjugate pair. Hence, the multiplicative constant $\overline{c}_P$ is the optimal constant within the family of form (57), as has been shown in [4, Theorem 3.7].
- For the case $\sigma > 2$, $\sigma \notin 2\,\mathbb{Z}$, the lowest modes have the slowest decay: $\Xi = \{-1, 1\}$ with $c_\Xi = \mathrm{cond}\big(P^{(2)}(\pm 1)\big) = (\sigma+2)/(\sigma-2)$. The multiplicative constant c_Ξ is not the smallest possible multiplicative constant in (57). However, according to [4, Theorem 4.1] it is the best possible multiplicative constant achievable by Lyapunov functionals that are quadratic forms. As $\sigma > 2$ it follows that $\bar{c}_P > c_\Xi$, and hence we replace the functionals $\|\cdot\|^2_{P^{(1)}(\xi)}$ and $\|\cdot\|^2_{P^{(2)}(\xi)}$

for the faster decaying modes $\xi \notin \Xi, \xi \neq 0$. Let us define

$$\overline{P}^{(1)}(\xi) := \begin{pmatrix} 1 & -\frac{2i}{\sigma\xi} \\ \frac{2i}{\sigma\xi} & 1 \end{pmatrix}, \quad \xi \notin \Xi, \xi \neq 0,$$

and notice that $\overline{P}^{(1)}(\xi)$ satisfies the matrix inequality

$$C^*(\xi)\,\overline{P}^{(1)}(\xi) + \overline{P}^{(1)}(\xi)\,C(\xi) \geq 2\,\overline{\mu}\,\overline{P}^{(1)}(\xi)\,, \quad \xi \notin \Xi\,,\ \xi \neq 0\,, \tag{59}$$

where $\overline{\mu}$ is the explicitly given uniform-in-$\mathbb{Z}$ spectral gap (52). Furthermore, as $\mathrm{cond}\big(\overline{P}^{(1)}(\xi)\big) \leq \mathrm{cond}\big(\overline{P}^{(1)}(\pm 1)\big) = (\sigma+2)/(\sigma-2)$, it satisfies the estimate $\mathrm{cond}\big(\overline{P}^{(1)}(\xi)\big) \leq c_\Xi(\sigma)$. Thus, the choice $\|\cdot\|_{\overline{P}^{(1)}(\xi)}$ for $\xi \notin \Xi(\sigma)$ and $\xi \neq 0$ leads (via (19)) to the global functional for $y \in (\mathrm{L}^2(\mathbb{T}))^2$, given as

$$\widetilde{\mathsf{H}}_2[y] := \sum_{\xi\in\Xi} \|\hat{y}(\xi)\|^2_{P^{(2)}(\xi)} + \sum_{\xi\notin\Xi,\xi\neq 0} \|\hat{y}(\xi)\|^2_{\overline{P}^{(1)}(\xi)} = \sum_{\xi\in\mathbb{Z}\setminus\{0\}} \|\hat{y}(\xi)\|^2_{\overline{P}^{(1)}(\xi)}\,,$$

where the equality follows as $P^{(2)}(\pm 1) = \overline{P}^{(1)}(\pm 1)$. $\widetilde{\mathsf{H}}_2$ yields decay with sharp rate $2\,\overline{\mu}(\sigma)$ given by (52) and, within the family of quadratic forms, the optimal multiplicative constant $c_\Xi(\sigma)$ in (57).

In summary, for arbitrary $\sigma > 0$, $\sigma \notin 2\,\mathbb{Z}$, Strategy 2 of Sect. 2.2 yields the global Lyapunov functional

$$\widetilde{\mathsf{H}}_2(\sigma)[y] := \sum_{\xi\in\mathbb{Z}\setminus\{0\}} \|\hat{y}(\xi)\|^2_{\overline{P}(\xi,\theta(\sigma))}\,, \quad y \in (\mathrm{L}^2(\mathbb{T}))^2\,, \tag{60}$$

where

$$\overline{P}(\xi,\theta) := \begin{pmatrix} 1 & -\frac{i\,\theta}{2\,\xi} \\ \frac{i\,\theta}{2\,\xi} & 1 \end{pmatrix}, \quad \theta\,(\sigma) := \begin{cases} \sigma\,, & 0 < \sigma < 2\,, \\ \frac{4}{\sigma}\,, & \sigma > 2\,. \end{cases} \tag{61}$$

Next, we turn to the method of Sect. 2.1 and derive another global $\mathrm{L}^2(\mathbb{T})$ functional that is based on the modal functionals (55):

$$\mathsf{H}_1(\delta)[y] := \sum_{\xi\in\mathbb{Z}\setminus\{0\}} \mathsf{H}_1(\xi,\delta)[\hat{y}(\xi)]\,, \quad y \in (\mathrm{L}^2(\mathbb{T}))^2\,.$$

In [21, § 1.4] the parameter $\delta \in (0,2)$ was chosen independent of ξ. But optimizing the resulting decay rate of $\mathsf{H}_1(\delta)$ w.r.t. the parameter $\delta \in (0,2)$ yields non-sharp decay rates (as derived in [21, § 1.4] for $\lambda_m = \sigma = 1$). Hence, we shall optimize here each modal functional $\mathsf{H}_1(\xi,\delta(\xi))$ w.r.t. the parameter δ.

For $y \in (\mathrm{L}^2(\mathbb{T}))^2$ and arbitrary $\sigma > 0$, $\sigma \notin 2\,\mathbb{Z}$, the resulting functional is given as

$$\widetilde{\mathsf{H}}_1(\sigma)[y] := 2 \sum_{\xi \in \mathbb{Z}\setminus\{0\}} \mathsf{H}_1\left(\xi, \overline{\delta}(\xi,\sigma)\right)[\hat{y}(\xi)]\,, \tag{62}$$

with the optimal parameter $\overline{\delta}(\xi,\sigma) := \frac{\theta(\sigma)\,(1+\xi^2)}{2\,\xi^2} \in (0,2)$ and $\theta(\sigma)$ defined in (61). The following theorem relates $\widetilde{\mathsf{H}}_1$ to the previously defined functional $\widetilde{\mathsf{H}}_2$, given respectively by (62) and (60).

Theorem 3 *For $y \in (\mathrm{L}^2(\mathbb{T}))^2$ and arbitrary $\sigma > 0$, $\sigma \notin 2\,\mathbb{Z}$ it follows that*

$$\widetilde{\mathsf{H}}_1(\sigma)[y] = \widetilde{\mathsf{H}}_2(\sigma)[y]\,.$$

Proof Thanks to previous considerations, the proof is now straightfoward. For each mode $\xi \in \mathbb{Z}\setminus\{0\}$, the identity

$$2\,\mathsf{H}_1\left(\xi, \overline{\delta}(\xi,\sigma)\right)[\hat{y}] = \|\hat{y}\|^2_{\overline{P}(\xi,\theta(\sigma))}\,, \quad \hat{y} \in \mathbb{C}^2 \tag{63}$$

follows by setting $\delta = \overline{\delta}(\xi,\theta)$ in (55). □

4.2.2 The Defective Cases

For $\sigma \neq 2$ the defective modes $|\xi| = \sigma/2$ do not exhibit the slowest decay of all modes, i.e. $\xi \notin \Xi$ as defined in Sect. 2.2. The functional $\|\cdot\|^2_{\overline{P}^{(1)}(\xi)}$ yields the sufficient decay rate $2\,\bar{\mu}(\sigma)$, along with multiplicative constants that are small enough, i.e., $\mathrm{cond}\big(\overline{P}^{(1)}(\xi)\big) \leq c_\Xi(\sigma)$. It follows that Strategy 2 of Sect. 2.2 again yields the functional $\widetilde{\mathsf{H}}_2$ as defined in (60).

The case $\sigma = 2$ is the only case where the defective modes correspond to the slowest modal decay, i.e. $\xi = \pm 1 \in \Xi$. Then, for arbitrarily small $\varepsilon > 0$ the modified norm $\|\cdot\|^2_{\overline{P}(\xi,\theta_\varepsilon)}$, defined in (61), with

$$\theta_\varepsilon := 2\,\frac{2-\varepsilon^2}{2+\varepsilon^2}\,, \tag{64}$$

yields the exponential decay rate $2\left(\overline{\mu}(2)-\varepsilon\right)$ for all modes $|\xi| \neq 0$. Due to the lack of an eigenvector basis in the defective case, constructing the matrix $\overline{P}(\xi,\theta_\varepsilon)$ results in a decay estimate of form (57) with multiplicative constant $c_\varepsilon = \sqrt{2}/\varepsilon$. The blow-up $\lim_{\varepsilon\to 0_+} c_\varepsilon = +\infty$ reflects the fact that the true decay behaviour of solutions in this defective setting is not purely exponential with rate $2\,\overline{\mu}(2)$, but rather exponential times a polynomial in time t. An approach based on more involved time-dependent Lyapunov functionals yields estimates with the sharp defective decay

behaviour. As the time-dependent construction is besides our focus, we simply refer to [10] for further details.

4.2.3 Decay Results for the Case $\mathcal{X} = \mathbb{T}$

In this subsection we start by refining the general strategy of Sect. 3.1.3 to extract the sharp decay rate for the GT model from the functional $\widetilde{\mathsf{H}}_1$. Subsequently, we conclude the torus case by expressing the global Lyapunov functional in the spatial variable.

In Theorem 3 above, we establish that both functional constructions (as described in Sect. 2.1) coincide for the GT equation if one chooses the appropriate parameter $\delta(\xi)$ for $\widetilde{\mathsf{H}}_1$. Now, we compare both approaches of Sect. 2.1 to extract explicit decay rates from the functional.

The Strategy 2 of Sect. 2.2 for $\widetilde{\mathsf{H}}_2$ is based on modal matrix inequalities, (59), which prove the sharp global decay rate $2\,\overline{\mu}$ as already discussed in Sect. 4.2.1.

The general method of Sect. 2.1 for $\widetilde{\mathsf{H}}_1$ (and its improvements of Sect. 3.1) is to estimate the entropy–entropy production inequality $\mathrm{D}[y] - \lambda\,\widetilde{\mathsf{H}}_1[y] \geq 0$ in terms of $\|(\mathrm{Id} - \Pi)\,y\|$ and $\|\Pi y\|$ that are then optimized for λ. As assumed in Sect. 3.1, we restrict our discussion to $\lambda_m = 1$ which requires the relaxation rate $\sigma = 1$. Applying Proposition 3 to the modal equation (50) for $\xi = \pm 1$ yields the decay estimates

$$\mathsf{H}_1\left(\pm 1, \tilde{\delta}_2(1)\right)[\hat{y}(\pm 1, t)] \leq e^{-\tilde{\lambda}(1)\,t}\,\mathsf{H}_1\left(\pm 1, \tilde{\delta}_2(1)\right)[\hat{y}(\pm 1, 0)]$$

with non-optimal modal decay rate $\tilde{\lambda}(1) \approx 0.165$ and parameter $\tilde{\delta}_2(1) \approx 0.325$. Higher modes, $|\xi| > 1$, yield higher decay rates (*cf.* Lemma 2), but as

$$\inf_{\xi \in \mathbb{Z}\setminus\{0\}} \tilde{\lambda}(|\xi|) = \tilde{\lambda}(1) < 2\,\overline{\mu}(1) = 1\,,$$

the optimal global rate cannot be recovered. One cause for not reaching the sharp rate is that Proposition 3 approximates the entropy–entropy production inequality condition to obtain readable formulas (via the discriminant $\tilde{h}_2$ as defined in (41)). But even omitting approximations when optimizing δ does not yield sharp decay rates $\lambda(|\xi|) = 1$ for our example.

This is not surprising as Sect. 3.1.4 provides explicit estimates with a general hypocoercive setting in mind. In order to obtain sharp decay rates for the GT model, we sacrifice this generality and refine the strategy for the simple structure at hand. The reduction of the continuous velocity space velocity space $v \in \mathbb{R}$ (as defined in Sect. 3.1.1) to two discrete velocities in the GT setting allows the following modifications: With the notation of Sect. 4.1 it

$$\mathsf{A}(\xi)\,\mathsf{T}(\xi)(\mathrm{Id} - \Pi)\hat{y} = 0\,, \quad \hat{y} \in \mathbb{C}^2\,.$$

Thus, the constant C_M as defined in (39) improves to $C_M = \frac{|\xi|}{1+|\xi|^2}$. Additionally, as

$$\mathsf{A}(\xi)\,(\mathsf{L}+\lambda\,\mathrm{Id}) = \begin{pmatrix} 0 & \frac{i\,\xi\,(1-\lambda)}{1+\xi^2} \\ 0 & 0 \end{pmatrix},$$

it follows that

$$\left|\mathrm{Re}\langle \mathsf{A}(\xi)\,(\mathsf{L}+\lambda\,\mathrm{Id})\hat{y}(\xi)\,,\,\hat{y}(\xi)\rangle\right| \leq \frac{|\xi|}{1+|\xi|^2}\,(1-\lambda)\,X\,Y\,,$$

for $0 \leq \lambda \leq 1$, where $X := \|(\mathrm{Id}-\Pi)\hat{y}\| = \|v\|$ and $Y := \|\Pi\hat{y}\| = \|u\|$. Then, as a refinement of (40) for the GT equation with $\mathsf{H}_1(\xi,\delta)$ from (55) it follows that

$$\begin{aligned}
&\mathsf{D}(\xi,\delta)[\hat{y}] - \lambda\,\mathsf{H}_1(\xi,\delta)[\hat{y}] \\
&\geq \left(1 - \frac{\delta\,\xi^2}{1+\xi^2} - \frac{\lambda}{2}\right) X^2 - \delta\,\mathrm{Re}\langle \mathsf{A}\,(\mathsf{L}-\lambda\,\mathrm{Id})\,F,\,F\rangle + \left(\frac{\lambda\,\xi^2}{1+\xi^2} - \frac{\lambda}{2}\right) Y^2 \\
&\geq \left(1 - \frac{\delta\,\xi^2}{1+\xi^2} - \frac{\lambda}{2}\right) X^2 - \frac{\delta\,|\xi|\,(1-\lambda)}{1+\xi^2}\,X\,Y + \left(\frac{\lambda\,\xi^2}{1+\xi^2} - \frac{\lambda}{2}\right) Y^2\,.
\end{aligned}$$

The *refined discriminant condition* is then given by the non-positivity of

$$h_{\mathrm{GT}}(\delta,\lambda) := \frac{\delta^2\,\xi^2}{(1+\xi^2)^2}\,(1-\lambda)^2 - 4\left(1 - \frac{\delta\,\xi^2}{1+\xi^2} - \frac{\lambda}{2}\right)\left(\frac{\delta\,\xi^2}{1+\xi^2} - \frac{\lambda}{2}\right).$$

It can be verified directly that $\overline{\delta}(\xi) := \frac{1+\xi^2}{2\,\xi^2}$ for $\xi \neq 0$ yields $h_{\mathrm{GT}}\big(\overline{\delta}(\xi),1\big) = 0$. Hence we recover the sharp exponential decay rate $\lambda(|\xi|) = 2\,\overline{\mu}(1) = 1$ for the modal equations (50) for all $\xi \in \mathbb{Z}\setminus\{0\}$ with $\sigma = 1$.

With this we have shown that refining the method of Sect. 3.1.3 for the GT model (with $\sigma = 1$) allows us to recover the sharp global decay rate from the global functional $\widetilde{\mathsf{H}}_1$, as defined in (62).

In Sect. 4.2 above we show that both hypocoercive methods from § I lead to the same global Lyapunov functional for arbitrary $\sigma > 0$. We conclude this subsection by leaving the modal formulation behind and expressing this global functional in the spatial variable.

In [7] the authors define an explicit spatial Lyapunov functional that yields the sharp, purely exponential decay rates and best possible multiplicative constant (reachable via quadratic forms) for each $\sigma > 0$:

Definition 1 Let $u, v \in \mathrm{L}^2(\mathbb{T})$ be real-valued and let $\theta \in (0,2)$ be given. We then define the functional $\mathsf{E}_\theta[u,v]$ as

$$\mathsf{E}_\theta[u,v] := \|u\|^2_{\mathrm{L}^2(\mathbb{T})} + \|v\|^2_{\mathrm{L}^2(\mathbb{T})} - \frac{\theta}{2\pi}\int_0^{2\pi} v\,\partial_x^{-1}u\,dx\,.$$

Here, the *anti-derivative* of u is defined as

$$\partial_x^{-1} u(x) := \int_0^x u\, dy - \left(\int_0^x u(y)\, dy \right)_{\text{avg}} , \tag{65}$$

where $u_{\text{avg}} := \frac{1}{2\pi} \int_0^{2\pi} u\, dx = \hat{u}(0)$.

Theorem 4 *For $u, v \in \mathrm{L}^2(\mathbb{T})$ and arbitrary $\sigma > 0$, $\sigma \notin 2\,\mathbb{Z}$ it follows that*

$$\widetilde{\mathsf{H}}_1(\sigma) \left[u - u_{\text{avg}}, v \right] = \mathsf{E}_{\theta(\sigma)} \left[u - u_{\text{avg}}, v \right] , \tag{66}$$

with $\widetilde{\mathsf{H}}_1$ defined in (62).

Proof As in [7, § 4.3], we use Parseval's identity and the fact that $(i\,k)^{-1}$ is the (discrete) Fourier symbol of ∂_x^{-1} as defined in (65). For the total entropy of arbitrary $y := (u, v)^T \in (\mathrm{L}^2(\mathbb{T}))^2$, with $u_{\text{avg}} = 0$, we deduce from (63) that

$$\begin{aligned} \widetilde{\mathsf{H}}_1(\sigma)[y] + \|\hat{v}(0)\|^2 &= \sum_{k \in \mathbb{Z} \setminus \{0\}} \|\hat{y}(\xi)\|^2_{\overline{P}(\xi, \theta(\sigma))} + \|\hat{v}(0)\|^2 \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left(|u|^2 + |v|^2 - \theta(\sigma)\, v\, \partial_x^{-1} u \right) dx \\ &= \mathsf{E}_{\theta(\sigma)}[u, v] . \end{aligned}$$

□

In the following result we recall from [7, Theorem 2.2.a] the optimal exponential decay for $y(t)$ to the steady state $y_\infty = (u_{\text{avg}}, 0)^T$, both in the functional E_θ and in the Euclidean norm. *Mild solution* refers to the terminology of semigroup theory [31].

Theorem 5 *Let $(u, v) \in C([0, \infty); (\mathrm{L}^2(\mathbb{T}))^2)$ be a mild real valued solution to (49) with initial datum u_0, $v_0 \in \mathrm{L}^2(\mathbb{T})$ and define $u_{\text{avg}} := \frac{1}{2\pi} \int_0^{2\pi} u_0(x)\, dx$.*

- *If $\sigma \neq 2$ then*

$$\mathsf{E}_{\theta(\sigma)}[u(t) - u_{\text{avg}}, v(t)] \leq \mathsf{E}_{\theta(\sigma)}[u_0 - u_{\text{avg}}, v_0]\, e^{-2\,\mu(\sigma)\, t} \quad \forall\, t \geq 0 ,$$

where

$$\theta(\sigma) := \begin{cases} \sigma , & 0 < \sigma < 2 \\ \frac{4}{\sigma} , & \sigma > 2 \end{cases} , \quad \mu(\sigma) := \begin{cases} \frac{\sigma}{2} , & 0 < \sigma < 2 \\ \frac{\sigma}{2} - \sqrt{\frac{\sigma^2}{4} - 1} , & \sigma > 2 \end{cases} .$$

Consequently we obtain the decay estimate

$$\left\| f(t) - \begin{pmatrix} f_\infty \\ f_\infty \end{pmatrix} \right\|_{\mathrm{L}^2(\mathbb{T})} \leq \mathscr{C}_\sigma \left\| f_0 - \begin{pmatrix} f_\infty \\ f_\infty \end{pmatrix} \right\|_{\mathrm{L}^2(\mathbb{T})} e^{-\mu(\sigma)\, t} \quad \forall\, t \geq 0\,,$$

where the decay rate $\mu(\sigma)$ is sharp and

$$\mathscr{C}_\sigma := \sqrt{\frac{2+\sigma}{|2-\sigma|}}\,, \quad f(t) := \begin{pmatrix} f_+(t) \\ f_-(t) \end{pmatrix}, \quad f_\infty = \frac{1}{2}\, u_{\mathrm{avg}}\,, \quad f_0 := \begin{pmatrix} f_{+,0} \\ f_{-,0} \end{pmatrix}.$$

- *If $\sigma = 2$ then for any $0 < \varepsilon < 1$*

$$\mathsf{E}_{\theta_\varepsilon}[u(t) - u_{\mathrm{avg}}, v(t)] \leq \mathsf{E}_{\theta_\varepsilon}[u_0 - u_{\mathrm{avg}}, v_0]\, e^{-2\,(1-\varepsilon)\, t} \quad \forall\, t \geq 0$$

with θ_ε defined as in (64) *and we have that*

$$\left\| f(t) - \begin{pmatrix} f_\infty \\ f_\infty \end{pmatrix} \right\|_{\mathrm{L}^2(\mathbb{T})} \leq \frac{\sqrt{2}}{\varepsilon} \left\| f_0 - \begin{pmatrix} f_\infty \\ f_\infty \end{pmatrix} \right\|_{\mathrm{L}^2(\mathbb{T})} e^{-(1-\varepsilon)\, t} \quad \forall\, t \geq 0\,.$$

4.3 Decay Results for the Case $\mathcal{X} = \mathbb{R}$

We consider the GT model with position x on the real line and prove two global decay estimates with sharp algebraic rate. Our first goal is to obtain modal decay estimates of general form (46) with modal constants $C(|\xi|)$ as small as possible. As we discuss below, a straightforward application of Lemma 5 with Strategy 1 of Sect. 2.2 is not possible due to the appearance of a defective eigenvalue in the modal equation (50). To avoid this difficulty we shall use a non-sharp decay estimate as input to apply Lemma 5. Our second goal is to construct a simple spatial functional that closely approximates our first result. To achieve this, we construct modal Lyapunov functionals that yield slightly less precise estimates but have the advantage of representing a more convenient pseudo-differential operator.

To simplify notation, we assume that $\sigma = 1$ in the present section. This is no restriction, as the general case $\sigma > 0$ in (48) can always be reduced to the normalized one thanks to the rescaling $\tilde{t} = \sigma\, t$, $\tilde{x} = \sigma\, x$.

A natural approach to obtain a decay estimate for $\mathcal{X} = \mathbb{R}$ is an application of Lemma 5 to the decay estimates (53) with the matrices $P^{(1)}$ and $P^{(2)}$ of (54) for $\sigma = 1$. The extension of Strategy 1 of Sect. 2.2 to $\xi \in \mathbb{R}$ leads to (57). But as $\operatorname{cond}\big(P^{(1)}(\xi)\big) \to \infty$ and $\operatorname{cond}\big(P^{(2)}(\xi)\big) \to \infty$ for $|\xi| \to 1/2$, it follows that the multiplicative constants in (57) become unbounded. This is due to the defective limit of the modal equation (50) at $|\xi| = 1/2$. The modal Lyapunov functionals with sharp rate depend on the eigenspace structure, which has a discontinuity at

$|\xi| = 1/2$ and the decay rates are not purely exponential there. Hence, we cannot directly use Lemma 5 with sharp rates.

Therefore, the natural and in fact sharper approach is to start with the exact modal decay function (instead of an exponential approximation): for 2×2 ODE systems, this decay function was given in [4, Proposition 4.2]:

$$\|\hat{y}(t,\xi)\|_2^2 \leq h_+(t,\xi)\, \|\hat{y}(0,\xi)\|_2^2 \quad \forall\, t \geq 0\,, \tag{67}$$

where $h_+(t,\xi)$, the squared propagator norm associated with (50), is explicitly given in [4]. Since this function is continuous at the defective point $\xi = 1/2$ for all $t \geq 0$ (see Fig. 6), one could easily extend Lemma 5 to this setting. But, since $h_+(t,\xi)$ is a quite involved function, the minimization w.r.t. R (as in (47)) could only be carried out numerically. In order to come up with an explicit decay estimate, we shall therefore rather approximate the modal (exponential) decay estimates that are used as a starting point for Lemma 5. We now approximate the decay estimate for large frequencies $|\xi|$, but keep the sharp estimates for $|\xi|$ small.

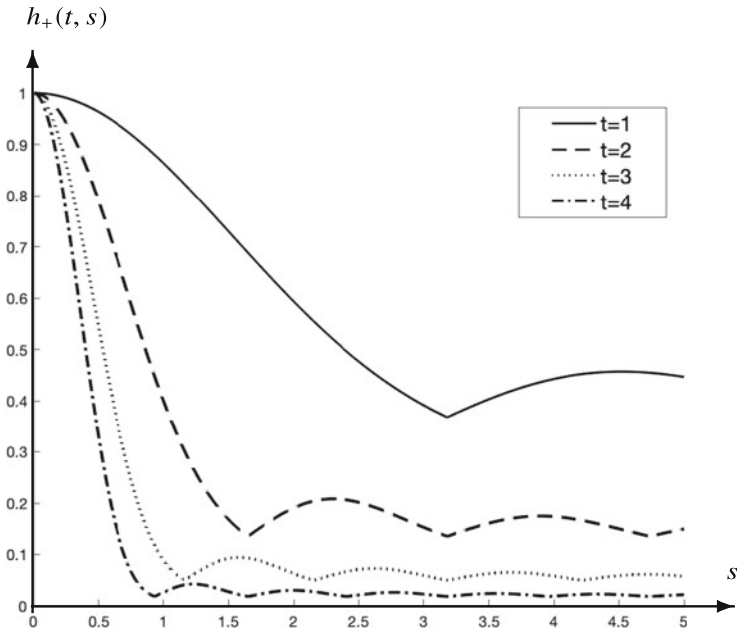

Fig. 6 The mapping $|\xi| = s \mapsto h_+(t,s)$ shows the continuous modal dependency of the squared propagator norm of $C(\xi, 1)$ for fixed times t. Note that the kinks are no numerical artefact

Lemma 6 *Assume that $R \in (0, 1/2)$ and let $\hat{y}(t, \xi)$ be a solution to (50). Then,*

$$\|\hat{y}(t,\xi)\|^2 \le c(\xi)\, e^{-\lambda(\xi)\, t}\, \|\hat{y}(0,\xi)\|^2\,, \quad \forall\, \xi \in \mathbb{R} \setminus \{0\}\,, \quad \forall\, t \ge 0\,, \tag{68}$$

with $c(\xi) = \begin{cases} \frac{1+2\,|\xi|}{1-2\,|\xi|}\,, & |\xi| < R\,, \\ \frac{|\xi|+2\,R^2}{|\xi|-2\,R^2}\,, & |\xi| \ge R\,, \end{cases} \quad \lambda(\xi) = \begin{cases} 2\,\mu(\xi)\,, & |\xi| < R\,, \\ 2\,\mu(R)\,, & |\xi| \ge R\,. \end{cases}$

Proof For every $|\xi| < R$, the modal functional $\|\cdot\|^2_{P^{(2)}(\xi)}$, as defined in (54), yields the sharp modal decay

$$2\,\mu(\xi) := 2\,\mu(\xi, 1) = 1 - \sqrt{1 - 4\,\xi^2}\,,$$

given by (51). The condition number of $P^{(2)}(\xi)$ is given as

$$c(\xi) := \mathrm{cond}\left(P^{(2)}(\xi)\right) = \frac{1+2\,|\xi|}{1-2\,|\xi|} \le \frac{1+2\,R}{1-2\,R}\,, \quad |\xi| < R\,. \tag{69}$$

For $|\xi| \ge R$ we use the rescaled version of $\|\cdot\|^2_{\overline{P}^{(1)}(\xi)}$ (from Sect. 4.2.1, but now for $\mathcal{X} = \mathbb{R}$), given as $\|\cdot\|^2_{\overline{P}(\xi)}$, with the matrix

$$\overline{P}(\xi) := \begin{pmatrix} 1 & -\frac{2\,i\,R^2}{\xi} \\ \frac{2\,i\,R^2}{\xi} & 1 \end{pmatrix}. \tag{70}$$

As this matrix satisfies the inequality

$$C^*(\xi)\overline{P}(\xi) + \overline{P}(\xi)C(\xi) \ge 2\mu(R)\overline{P}(\xi), \quad |\xi| \ge R, \tag{71}$$

the functional $\|\cdot\|^2_{\overline{P}(\xi)}$ yields an exponential decay $2\,\mu(R)$ for all modes $|\xi| \ge R$. The condition number of $\overline{P}(\xi)$ is given as:

$$c(\xi) := \mathrm{cond}\left(\overline{P}(\xi)\right) = \frac{|\xi| + 2\,R^2}{|\xi| - 2\,R^2} \le \mathrm{cond}\left(\overline{P}(R)\right) = \frac{1+2\,R}{1-2\,R}\,, \quad |\xi| \ge R\,, \tag{72}$$

from which the desired result follows. □

We can now apply Lemmas 5 and 6 to obtain following global decay estimate.

Proposition 5 *Let $y := (u, v)^T$ be a solution of the Goldstein–Taylor equation (49) on $\mathbb{R}$ with $\sigma = 1$ and initial datum $y_0 := (u_0, v_0)^T$, such that u_0, $v_0 \in \mathrm{L}^1(\mathbb{R}) \cap \mathrm{L}^2(\mathbb{R})$. Let the modal spectral gap, defined in (51), be denoted as $\mu(\xi) := \mu(\xi, 1)$.*

Then, for any $t \geq 0$ it follows that

$$\|y(t)\|^2_{\mathrm{L}^2(\mathbb{R})} \leq \inf_{0<R<\frac{1}{2}} \frac{1+2R}{1-2R} \left(2 \min\left\{ \frac{B(t,R)}{\sqrt{t}}, R \right\} \|y_0\|^2_{\mathrm{L}^1(\mathbb{R})} + e^{-2\mu(R)t} \|y_0\|^2_{\mathrm{L}^2(\mathbb{R})} \right)$$

with $B(t,R) := \sqrt{t} \int_0^R e^{-2\mu(s)t} ds \in \left[0, \sqrt{\pi/8}\right)$.

Proof Applying Lemma 5 to the modal decay estimates (68) and taking into account the estimates (69) and (72) leads to the decay result where, for $B(t,R)$, we use the estimate $\mu(|\xi|)/\xi^2 = \left(1/2 - \sqrt{1/4 - \xi^2}\right)/\xi^2 \geq 1$ for $0 \leq |\xi| \leq 1/2$. The bound on B follows from $B(t,R) \leq \sqrt{t} \int_0^\infty e^{-2\xi^2 t}\, d\xi$. □

Remark 3 The decay result of Proposition 5 is neither explicit in the optimization with respect to R (for fixed t), nor optimal, as this would require an approach starting from (67). It is however the best possible estimate of form (46) achievable with quadratic forms for each mode. This follows, as for one, the modal functionals for $|\xi| \leq R$, $\xi \neq 0$ are optimal for quadratic forms (*cf.* the discussion on $P^{(2)}(\xi)$ in Sect. 4.2.1). Additionally, the modal functionals for $|\xi| \geq R$ are sufficient (in light of Lemma 5) as they yield the sufficient decay $2\,\overline{\mu}(R)$ and the sufficient multiplicative constants $\sup_{|\xi|\geq R} c(\xi) = \sup_{|\xi|<R} c(\xi) = (1+2R)/(1-2R)$. In analogy to Sect. 2.2 the decay stated in Proposition 5 results from the global functional

$$\widehat{\mathsf{H}}_2(R)[y] := \int_{(-R,R)} \|\hat{y}(\xi)\|^2_{P^{(2)}(\xi)}\, d\xi + \int_{|\xi|\geq R} \|\hat{y}(\xi)\|^2_{\overline{P}(\xi)}\, d\xi\,.$$

As our final result, we shall consider an alternative modal functional for the GT equation on $\mathbb{R}$ that translates into a convenient representation in the spatial variable. The trade-off is a less accurate global decay estimate.

The result of Proposition 5 was based on the modal Lyapunov functional $\|\cdot\|^2_{\overline{P}(\xi)}$ for large modes and $\|\cdot\|^2_{P^{(2)}(\xi)}$ for small modes. Now, we replace both functionals by the single norm $\|\cdot\|^2_{\tilde{P}(\xi)}$ with the positive definite Hermitian matrix

$$\tilde{P}(\xi) := \begin{pmatrix} 1 & -\frac{2i\xi}{1+4\xi^2} \\ \frac{2i\xi}{1+4\xi^2} & 1 \end{pmatrix}, \quad \xi \neq 0\,, \tag{73}$$

which asymptotically approximates the matrices from (54) which yield sharp modal decay. For the off-diagonal matrix elements we have

$$\tilde{P}_{12}(\xi) - P^{(1)}_{12}(\xi) = o\left(P^{(1)}_{12}(\xi)\right) \quad \text{as} \quad |\xi| \to +\infty\,,$$
$$\tilde{P}_{12}(\xi) - P^{(2)}_{12}(\xi) = o\left(P^{(2)}_{12}(\xi)\right) \quad \text{as} \quad |\xi| \to 0\,.$$

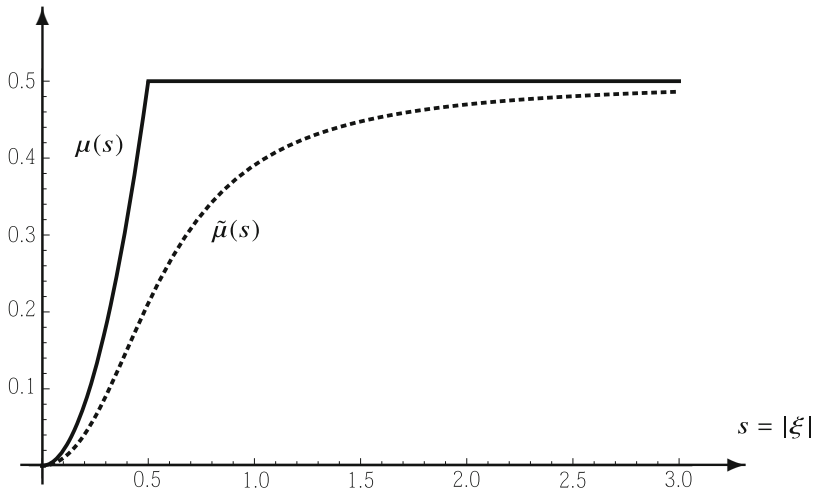

Fig. 7 Exponential decay rate $\tilde{\mu}$ in comparison to the sharp exponential rate μ, shown as functions of the spatial frequency $s = |\xi|$

It satisfies the matrix inequality (12) with $P = \tilde{P}(\xi)$ and the spectral gap μ replaced by

$$\tilde{\mu}(\xi) = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1 + 4\,\xi^2\,(1 + 4\,\xi^2)}}\right).$$

The rate $2\,\tilde{\mu}(\xi)$ is an approximation to the sharp decay rate $2\,\mu(\xi)$ of fifth order for modes ξ close to 0, see Fig. 7. The condition number of $\tilde{P}(\xi)$ is given by

$$\tilde{c}(\xi) := \mathrm{cond}\left(\tilde{P}(\xi)\right) := \frac{1 + 2\,|\xi| + 4\,\xi^2}{1 - 2\,|\xi| + 4\,\xi^2}, \tag{74}$$

and hence we arrive at the modal decay estimates for $\xi \neq 0$:

$$\|\hat{y}(t,\xi)\|^2 \leq \tilde{c}(\xi)\, e^{-2\,\tilde{\mu}(\xi)\,t}\, \|\hat{y}(0,\xi)\|^2, \quad t \geq 0. \tag{75}$$

We define the global Lyapunov functional

$$\mathsf{H}_3[y] := \int_{\mathbb{R}} \|\hat{y}(\xi)\|^2_{\tilde{P}(\xi)}\, d\xi\,.$$

As we shall see now, this can be rewritten in x-space (without resorting to the ξ-modes) in terms of a fairly simple pseudo-differential operator, similar to the functional $E_\theta[y]$ from Definition 1. Moreover, it is easily related to the functional

$\mathsf{H}_1[y]$ from Sect. 2.1: on the symbol level it holds that functional $\mathsf{H}_3(\xi) = 2\,\mathsf{H}_1(2\,\xi, \delta = 1)$, see (73), (55).

Proposition 6

(a) For $u, v \in \mathrm{L}^2(\mathbb{R})$, the functional H_3 can be expressed as

$$\mathsf{H}_3[u, v] = \|u\|^2_{\mathrm{L}^2(\mathbb{R})} + \|v\|^2_{\mathrm{L}^2(\mathbb{R})} - 4 \int_{\mathbb{R}} u(x)\, \partial_x \left(1 - 4\,\partial_x^2\right)^{-1} v(x)\, dx\,.$$

(b) Let $y := (u, v)^T \in C\big([0, \infty); (\mathrm{L}^2(\mathbb{R}))^2\big)$ be a mild real valued solution to (49) with $\sigma = 1$ and initial datum $u_0,\ v_0 \in \mathrm{L}^1(\mathbb{R}) \cap \mathrm{L}^2(\mathbb{R})$. Then, the functional H_3 yields the decay estimate

$$\begin{aligned} &\|y(t, x)\|^2_{\mathrm{L}^2(\mathbb{R})} \\ &\leq \inf_{0 < R \leq \frac{\sqrt{5}-1}{4}} \left(\tfrac{1+2\,R+4\,R^2}{1-2\,R+4\,R^2} \min\left\{2\,R, \sqrt{\tfrac{\pi}{2\,t}}\right\} \|y_0\|^2_{\mathrm{L}^1(\mathbb{R})} + 3\, e^{-2\,\tilde{\mu}(R)\,t}\, \|y_0\|^2_{\mathrm{L}^2(\mathbb{R})} \right). \end{aligned}$$

Proof With Plancherel's identity it follows that

$$\begin{aligned} \mathsf{H}_3[y] &= \int_{\mathbb{R}} \|\hat{y}(\xi)\|^2_{\hat{P}(\xi)}\, d\xi = \|u\|^2_{\mathrm{L}^2(\mathbb{R})} + \|v\|^2_{\mathrm{L}^2(\mathbb{R})} + 2\,\mathrm{Re}\left(\int_{\mathbb{R}} 2\,i\,\xi\,\hat{u}(\xi)\, \frac{\overline{\hat{v}(\xi)}}{1 + 4\,\xi^2}\, d\xi \right) \\ &= \|u\|^2_{\mathrm{L}^2(\mathbb{R})} + \|v\|^2_{\mathrm{L}^2(\mathbb{R})} - 4 \int_{\mathbb{R}} u(x)\, \partial_x \big(1 - 4\,\partial_x^2\big)^{-1} v(x)\, dx\,. \end{aligned}$$

To prove the decay estimate, we apply Lemma 5 to (75). The multiplicative constant $\tilde{c}(\xi)$ from (74) is monotonously increasing for $\xi \in [0, 1/2]$ to its global maximum $\mathrm{cond}\left(\tilde{P}(1/2)\right) = 3$. For the integral in (47) with $\tilde{c}(\xi)$, we estimate

$$\int_{|\xi| \leq R} \tilde{c}(\xi)\, e^{-2\,\tilde{\mu}(\xi)\,t}\, d\xi \leq \tilde{c}(R) \int_{|\xi| \leq R} e^{-\xi^2 \alpha(\xi)\,t}\, d\xi$$

with

$$\alpha(\xi) := \frac{1}{\xi^2} \left(1 - \frac{1}{\sqrt{1 + 4\,\xi^2\,(1 + 4\,\xi^2)}} \right).$$

One easily sees that α has a local minimum at $\xi = 0$ with $\alpha(0) = \alpha(\xi_1) = 1$, $\xi_1 = \left(\sqrt{5} - 1\right)/4 \approx 0.3$, i.e. for $0 < R < \left(\sqrt{5} - 1\right)/4$ it holds that $\alpha(\xi) \geq 1$ for all $|\xi| < R$. Thus, for $0 < R < \left(\sqrt{5} - 1\right)/4$ and $t > 0$, the desired result follows. □

Acknowledgments This work has been partially supported by the Project EFI (ANR-17-CE40-0030) of the French National Research Agency (ANR) and the Amadeus project *Hypocoercivity* no. 39453PH. J.D. and C.S. thank E. Bouin, S. Mischler and C. Mouhot for stimulating discussions that took place during the preparation of [15]: some questions raised at this occasion are the origin for this contribution. A.A., C.S., and T.W. were partially supported by the FWF (Austrian Science Fund) funded SFB F65.

References

1. Achleitner, F., Arnold, A., Carlen, E.A.: The hypocoercivity index for the short and large time behavior of ODEs. Preprint arXiv (2021). https://arxiv.org/abs/2109.10784
2. Achleitner, F., Arnold, A., Carlen, E.A.: On linear hypocoercive BGK models. In: From Particle Systems to Partial Differential Equations III, pp. 1–37. Springer, Berlin (2016). https://doi.org/10.1007/978-3-319-32144-8_1
3. Achleitner, F., Arnold, A., Carlen, E.A.: On multi-dimensional hypocoercive BGK models. Kinet. Relat. Models **11**(4), 953–1009 (2018). https://doi.org/10.3934/krm.2018038
4. Achleitner, F., Arnold, A., Signorello, B.: On optimal decay estimates for ODEs and PDEs with modal decomposition. In: Stochastic Dynamics Out of Equilibrium. Springer Proc. Math. Stat., vol. 282, pp. 241–264. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15096-9_6
5. Addala, L., Dolbeault, J., Li, X., Tayeb, M.L.: L^2-hypocoercivity and large time asymptotics of the linearized Vlasov–Poisson–Fokker-Planck system. J. Stat. Phys. **184**, 34 (2021). https://doi.org/10.1007/s10955-021-02784-4
6. Armstrong, S., Mourrat, J.C.: Variational methods for the kinetic Fokker-Planck equation. Preprint arXiv (2019). https://arxiv.org/abs/1409.5425
7. Arnold, A., Einav, A., Signorello, B., Wöhrer, T.: Large time convergence of the non-homogeneous Goldstein-Taylor equation. J. Stat. Phys. **182**, 35 (2021). https://doi.org/10.1007/s10955-021-02702-8
8. Arnold, A., Einav, A., Wöhrer, T.: On the rates of decay to equilibrium in degenerate and defective Fokker-Planck equations. J. Differ. Equ. **264**(11), 6843–6872 (2018). https://doi.org/10.1016/j.jde.2018.01.052
9. Arnold, A., Erb, J.: Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift. Preprint arXiv (2014). https://arxiv.org/abs/1409.5425
10. Arnold, A., Jin, S., Wöhrer, T.: Sharp decay estimates in local sensitivity analysis for evolution equations with uncertainties: from ODEs to linear kinetic equations. J. Differ. Equ. **268**(3), 1156–1204 (2020). https://doi.org/10.1016/j.jde.2019.08.047
11. Arnold, A., Schmeiser, C., Signorello, B.: Propagator norm and sharp decay estimates for Fokker-Planck equations with linear drift. Preprint arXiv (2020). https://arxiv.org/abs/2003.01405
12. Bernard, É., Salvarani, F.: Optimal estimate of the spectral gap for the degenerate Goldstein-Taylor model. J. Stat. Phys. **153**(2), 363–375 (2013). https://doi.org/10.1007/s10955-013-0825-6
13. Bernard, É., Salvarani, F.: Correction to: Optimal estimate of the spectral gap for the degenerate Goldstein-Taylor model. J. Stat. Phys. **181**(4), 1–2 (2020). https://doi.org/10.1007/s10955-020-02631-y
14. Bouin, E., Dolbeault, J., Lafleche, L., Schmeiser, C.: Hypocoercivity and sub-exponential local equilibria. Monatshefte für Mathematik (2020). https://doi.org/10.1007/s00605-020-01483-8
15. Bouin, E., Dolbeault, J., Mischler, S., Mouhot, C., Schmeiser, C.: Hypocoercivity without confinement. Pure Appl. Anal. **2**(2), 203–232 (2020). https://doi.org/10.2140/paa.2020.2.203

16. Bouin, E., Dolbeault, J., Schmeiser, C.: Diffusion and kinetic transport with very weak confinement. Kinet. Relat. Models **13**(2), 345–371 (2020). https://doi.org/10.3934/krm.2020012
17. Bouin, E., Dolbeault, J., Schmeiser, C.: A variational proof of Nash's inequality. Rend. Lincei Mate. Appl. **31**(1), 211–223 (2020). https://doi.org/10.4171/rlm/886
18. Calvez, V., Raoul, G.: Confinement by biased velocity jumps: aggregation of *escherichia coli*. Kinet. Relat. Models **8**, 651 (2015). https://doi.org/10.3934/krm.2015.8.651
19. Dolbeault, J., Klar, A., Mouhot, C., Schmeiser, C.: Exponential rate of convergence to equilibrium for a model describing fiber lay-down processes. Appl. Math. Res. eXpress (2012). https://doi.org/10.1093/amrx/abs015
20. Dolbeault, J., Mouhot, C., Schmeiser, C.: Hypocoercivity for kinetic equations with linear relaxation terms. C. R. Math. **347**(9–10), 511–516 (2009). https://doi.org/10.1016/j.crma.2009.02.025
21. Dolbeault, J., Mouhot, C., Schmeiser, C.: Hypocoercivity for linear kinetic equations conserving mass. Trans. Am. Math. Soc. **367**(6), 3807–3828 (2015). https://doi.org/10.1090/s0002-9947-2015-06012-7
22. Favre, G., Schmeiser, C.: Hypocoercivity and fast reaction limit for linear reaction networks with kinetic transport. J. Stat. Phys. **178**(6), 1319–1335 (2020). https://doi.org/10.1007/s10955-020-02503-5
23. Fellner, K., Prager, W., Tang, B.Q.: The entropy method for reaction-diffusion systems without detailed balance: First order chemical reaction networks. Kinet. Relat. Models **10**(4), 1055–1087 (2017). https://doi.org/10.3934/krm.2017042
24. Goudon, T., Alonso, R.J., Vavasseur, A.: Damping of particles interacting with a vibrating medium. Ann. Inst. Henri Poincaré (C) Non Linear Anal. (2016). https://doi.org/10.1016/j.anihpc.2016.12.005
25. Hérau, F.: Hypocoercivity and exponential time decay for the linear inhomogeneous relaxation Boltzmann equation. Asymptot. Anal. **46**(3–4), 349–359 (2006). https://content.iospress.com/articles/asymptotic-analysis/asy741
26. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2013). https://doi.org/10.1017/CBO9780511810817
27. Kawashima, S.: The Boltzmann equation and thirteen moments. Jpn. J. Appl. Math. **7**(2), 301–320 (1990). https://doi.org/10.1007/BF03167846
28. Mouhot, C., Neumann, L.: Quantitative perturbative study of convergence to equilibrium for collisional kinetic models in the torus. Nonlinearity **19**(4), 969–998 (2006). https://doi.org/10.1088/0951-7715/19/4/011
29. Nash, J.: Continuity of solutions of parabolic and elliptic equations. Am. J. Math. **80**, 931–954 (1958). https://doi.org/10.2307/2372841
30. Neumann, L., Schmeiser, C.: A kinetic reaction model: decay to equilibrium and macroscopic limit. Kinet. Relat. Models **9**, 571 (2016). https://doi.org/10.3934/krm.2016007
31. Pazy, A.: Semigroups of linear operators and applications to partial differential equations. Applied Mathematical Sciences, vol. 44. Springer, New York (1983). https://doi.org/10.1007/978-1-4612-5561-1
32. Shizuta, Y., Kawashima, S.: Systems of equations of hyperbolic-parabolic type with applications to the discrete Boltzmann equation. Hokkaido Math. J. **14**(2), 249–275 (1985). https://doi.org/10.14492/hokmj/1381757663
33. Ueda, Y., Duan, R., Kawashima, S.: Decay structure for symmetric hyperbolic systems with non-symmetric relaxation and its application. Arch. Ration. Mech. Anal. **205**(1), 239–266 (2012). https://doi.org/10.1007/s00205-012-0508-5
34. Villani, C.: Hypocoercivity. Mem. Am. Math. Soc. **202**(950), iv+141 (2009). https://doi.org/10.1090/S0065-9266-09-00567-5

Quantum Drift-Diffusion Equations for a Two-Dimensional Electron Gas with Spin-Orbit Interaction

Luigi Barletti, Philipp Holzinger, and Ansgar Jüngel

Abstract Quantum drift-diffusion equations are derived for a two-dimensional electron gas with spin-orbit interaction of Rashba type. The (formal) derivation turns out to be a non-standard application of the usual mathematical tools, such as Wigner transform, Moyal product expansion and Chapman–Enskog expansion. The main peculiarity consists in the fact that a non-vanishing current is already carried by the leading-order term in the Chapman–Enskog expansion. To our knowledge, this is the first example of quantum drift-diffusion equations involving the full spin vector. Indeed, previous models were either quantum bipolar (involving only the spin projection on a given axis) or full spin but semiclassical.

1 Introduction

Spintronics is an alternative to electronics, where the bit of information is carried by the spin polarization and not by the current [25]. Spintronics must not be confused with quantum computing: in the latter, both the information and its processing are based on a relatively small number of spins and are completely subject to the laws of quantum mechanics; in the former, the spin carriers are a large population and only the polarization is the result of an average of many single spins. Also in the case of spintronics, each spin carrier is subject to the laws of quantum mechanics and, for an accurate simulation of the behaviour of a spintronic device, it is very important to include quantum mechanical effects in the mathematical models. A systematic way to construct mathematical models of quantum fluids (diffusive or hydrodynamic) has been introduced by Degond, Ringhofer, and Méhats in Refs. [8, 9] (see also the exposition in [16]). Their strategy is based on the quantum mechanical version of the

L. Barletti (✉)
Dip. di Matematica e Informatica "U. Dini", Università di Firenze, Firenze, Italy
e-mail: luigi.barletti@unifi.ir

P. Holzinger · A. Jüngel
Institute of Analysis and Scientific Computing, TU Wien, Vienna, Austria
e-mail: philipp.holzinger@tuwien.ac.at; juengel@tuwien.ac.at

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_2

Maximum Entropy Principle (MEP), which basically says that the fluid-dynamical (macroscopic) equations, derived from an underlying kinetic (microscopic) model, can be closed by assuming that the microscopic state is the most probable one compatible with the observed macroscopic quantities (densities, currents, etc.). In turn, the most probable state is the one that maximises a suitable entropy functional, dictated by the laws of statistical mechanics. The quantum MEP (Q-MEP) can be formulated in the standard (operator-based) formalism of statistical quantum mechanics or in the phase-space formalism due to Wigner [22]. The operator form is more general, to the extent that it can also be applied to Hamiltonians defined in bounded domains (while the Wigner formalism is only suited to the whole-space case). However, the Wigner framework, being a quasi-classical description, is more suited to the semiclassical expansion of the quantum model, resulting in "classical equations" with "quantum corrections".

Diffusive models of particles with spin, subject to spin-orbit interactions, have been previously derived in Refs. [4, 10, 20]. In Ref. [10], two kinds of models are considered: the bipolar one, where only the projection of the spin on a given axis is considered, and the spin-vector one, where all the components of the spin vector are present. Such models are "semiclassical", which means that the drift-diffusion equations are not the standard ones because (of course) they contain the spin components, but the models do not incorporate non-local effects, such as the Bohm potential [16]. This is because the postulated equilibrium state is a classical Maxwellian for each spin component, while non-local effects only arise from a quantum equilibrium state. Reference [20] is a generalisation of [10], where a more detailed collision operator is considered, with spin-dependent scattering rates.

The first application of the Q-MEP to the case of particles with spin-orbit interaction is given in Ref. [4]. There, a two-dimensional electron gas (2DEG) with spin-orbit interaction of Rashba type [25] is considered and the Q-MEP is used to derive bipolar quantum drift-diffusion equations (QDDE) for the spin polarisation in the direction perpendicular to the 2DEG plane. The obtained model is then expanded semiclassically in order to obtain classical drift-diffusion equations for the density and polarisation with quantum corrections.

Few results are available related to the existence analysis of spin drift- diffusion models. The bipolar model was investigated in [13, 14]. An existence result for a diffusion model for the spin accumulation with fixed electron current but non-constant magnetization was proved in [12, 21]. Matrix spin drift-diffusion models were analyzed in [15, 17] with constant precession axis and in [23] with non-constant precession vector. Numerical simulations for this model can be found in [7]. Assuming a mass- and spin-conserving relaxation mechanism, two full-spin drift-diffusion models were derived and analyzed in [24], including spin-orbit interactions. These model, however, do not contain "quantum correction" terms.

In the present paper, we derive spin-vector QDDE for the same spin-orbit system as in [4]. As remarked before, this means that the QDDE that we derive here involve all the components of the spin vector. The paper is organised as follows. In Sect. 2, we introduce the Rashba Hamiltonian, describing the spin-orbit interaction of each electron in the 2DEG. Moreover, some basic concepts of the spinorial Wigner-

Moyal formalism are recalled. In Sect. 3, we set up the model at the kinetic level, consisting of an evolution equation for the matrix-valued Wigner function, endowed with a collisional term that describes the relaxation of the system to an equilibrium Wigner function obtained by the Q-MEP. The formal diffusive limit of the kinetic model is analysed in Sect. 3, which leads to the spin-vector QDDE (Eqs. (17), (21), and (24)). In order to test the consistency of the obtained equations, we consider the semiclassical limit of the QDDE and show that it is in accordance with the semiclassical equations derived in [10].

2 Physical and Mathematical Background

Let us consider a population of electrons confined in a two-dimensional potential well, described by the coordinates (x_1, x_2) and subject to a spin-orbit interaction of Rashba type [25]. The Hamiltonian of each electron has therefore the form

$$H = \begin{pmatrix} -\frac{\hbar^2}{2m}\Delta & -i\hbar\alpha_R\left(\partial_{x_2} + i\,\partial_{x_1}\right) \\ -i\hbar\alpha_R\left(\partial_{x_2} - i\,\partial_{x_1}\right) & -\frac{\hbar^2}{2m}\Delta \end{pmatrix},$$

where α_R is the Rashba constant and m is the (effective) electron mass. In terms of the Pauli matrices, we can write

$$H = -\frac{\hbar^2}{2m}\Delta\,\sigma_0 - i\hbar\alpha_R\left(\partial_{x_2}\sigma_1 - \partial_{x_1}\sigma_2\right), \tag{1}$$

where

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

In the following, we will extensively make use of the algebra of the Pauli matrices. Each 2×2 matrix-valued quantity $a \in \mathbb{C}^{2\times 2}$ can be decomposed in Pauli components according to

$$a = \sum_{j=0}^{3} a_j\sigma_j = a_0\sigma_0 + \boldsymbol{a}\cdot\boldsymbol{\sigma},$$

where $\boldsymbol{a} = (a_1, a_2, a_3)$, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$, and the components a_k $(k = 0, 1, 2, 3)$ are real if and only if a is Hermitian. By using the well-known identity

$$\sigma_i\sigma_j = i\epsilon_{ijk}\sigma_k + \delta_{ij}\sigma_0, \qquad 1 \le i, j, k, \le 3,$$

(where ϵ_{ijk} and δ_{ij} are, respectively, the Levi-Civita and Kronecker symbols), it is straightforward to prove the following relations, mapping the matrix algebra on the Pauli components:

$$\mathrm{tr}(a) = 2a_0, \tag{2}$$

$$ab = (a_0 b_0 + \boldsymbol{a}\cdot\boldsymbol{b})\sigma_0 + (a_0\boldsymbol{b} + b_0\boldsymbol{a} + i\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{\sigma}, \tag{3}$$

$$ab - ba = i\boldsymbol{a}\times\boldsymbol{b}\cdot\boldsymbol{\sigma}. \tag{4}$$

The Hamiltonian (1) can be written more concisely as

$$H = -\frac{\hbar^2}{2m}\Delta\,\sigma_0 - i\hbar\alpha_R\nabla^\perp\cdot\boldsymbol{\sigma} \tag{5}$$

with the notation

$$\nabla = (\partial_{x_1}, \partial_{x_2}, 0), \quad \nabla^\perp = \nabla\times\boldsymbol{e}_3 = (\partial_{x_2}, -\partial_{x_1}, 0), \quad \boldsymbol{e}_3 = (0,0,1).$$

We now combine the matrix algebra with the Wigner-Moyal calculus. The following definitions and properties hold for suitably smooth functions. Let us recall the definition of the Wigner transform, $\varrho \mapsto a$, of a function $\varrho = \varrho(x, y)$, $x \in \mathbb{R}^d$, $y \in \mathbb{R}^d$, into a phase-space function $a = a(x, p)$, $x \in \mathbb{R}^d$, $p \in \mathbb{R}^d$:

$$a(x,p) = \mathcal{W}(\varrho)(x,p) = \int_{\mathbb{R}^d} \varrho\left(x + \frac{\xi}{2}, x - \frac{\xi}{2}\right) \mathrm{e}^{-ip\cdot\xi/\hbar} d\xi$$

(see also Ref. [22]). We remark that, in our framework, we have $d = 2$, and the Wigner transform acts on the matrix-valued functions ϱ and a componentwise. The Wigner transformation is closely related to the Weyl quantization, $a \mapsto A$, that maps the phase-space function a to an operator A, according to

$$\begin{aligned}(A\psi)(x) &= \left[\mathrm{Op}(a)\psi\right](x)\\ &= \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^{2d}} a\left(\frac{x+y}{2}, p\right)\psi(y)\,\mathrm{e}^{i(x-y)\cdot p/\hbar}\,dy\,dp.\end{aligned}$$

In the correspondence $A = \mathrm{Op}(a)$, the phase space function a is often called the "symbol" of A.

The Wigner transform is the inverse of the Weyl quantization if one identifies the operator A with its integral kernel ϱ_A. In fact,

$$(A\psi)(x) = \int_{\mathbb{R}^d} \varrho_A(x,y)\,\psi(y)\,dy = \int_{\mathbb{R}^d} \mathcal{W}^{-1}(a)(x,y)\,\psi(y)\,dy.$$

The Wigner–Weyl correspondence is summarized in Fig. 1.

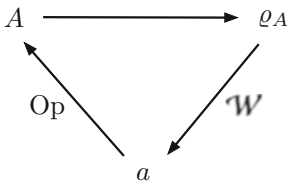

Fig. 1 The Wigner–Weyl correspondence: $A = \mathrm{Op}(a)$ is the operator associated to the phase-space function a, ϱ_A is the integral kernel of A, and $a = \mathcal{W}(\varrho_A)$ is the Wigner transform of ϱ_A

The operator algebra is transferred to phase-space functions by the Wigner–Weyl correspondence. In particular, the operator product gives rise to the definition of the Moyal product $a\#b = \mathrm{Op}^{-1}(AB)$, where $A = \mathrm{Op}(a)$ and $B = \mathrm{Op}(b)$. The Moyal product has an explicit expansion in powers of $\hbar$,

$$a\#b = \sum_{k=0}^{\infty} \hbar^k a\#_k b, \tag{6}$$

where

$$a\#_k b = \frac{1}{(2i)^k} \sum_{|\alpha|+|\beta|=k} \frac{(-1)^{|\alpha|}}{\alpha!\,\beta!} \left(\nabla_x^\alpha \nabla_p^\beta a\right) \left(\nabla_p^\alpha \nabla_x^\beta b\right).$$

At the leading order of the expansion, we find the ordinary product $a\#_0 b = ab$, while at the first order, it is related to the Poisson bracket,

$$a\#_1 b = \frac{i}{2} \sum_{j=1}^{2} \left(\partial_{x_j} a\, \partial_{p_j} b - \partial_{p_j} a\, \partial_{x_j} b\right).$$

The operator trace Tr is equivalent to the integral on the phase-space of the matrix trace tr of its symbol, i.e.

$$\mathrm{Tr}(A) = \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^{2d}} \mathrm{tr}(a)(x,\, p)\, dx\, dp.$$

In particular, if A represents some physical observable and S represents the state of the system, and if $a = \mathrm{Op}^{-1}(A)$ and $w = \mathrm{Op}^{-1}(S)$ are the corresponding phase-space functions (w is called the *Wigner function* of the system), then the expected value of the observable A in the state $S = \mathrm{Op}(w)$ is

$$\mathrm{Tr}(AS) = \frac{1}{(2\pi\hbar)^d} \int_{\mathbb{R}^{2d}} \mathrm{tr}(aw)(x,\, p)\, dx\, dp.$$

By expressing this identity in terms of Pauli components (by using (2) and (3)), we obtain the fundamental formula for the expected values:

$$\frac{1}{2}\operatorname{Tr}(AS) = \frac{1}{(2\pi\hbar)^d}\int_{\mathbb{R}^{2d}} (a_0 w_0 + \boldsymbol{a}\cdot\boldsymbol{w})(x,\,p)\,dx\,dp.$$

This relation suggests to define the *local density* n_A of the observable A as

$$n_A(x) = \int_{\mathbb{R}^d} (a_0 w_0 + \boldsymbol{a}\cdot\boldsymbol{w})(x,\,p)\,dp = \frac{1}{2}\langle \operatorname{tr}(aw)\rangle(x),$$

where we introduced the notation $\langle f\rangle = \int_{\mathbb{R}^d} f\,dp$. Note that we have omitted the constant factor $1/(2\pi\hbar)^d$, which is irrelevant to what follows. Since our goal is to derive a spinorial diffusive model, the local densities we are interested in are the position density n_0 (observable $\frac{1}{2}\sigma_0$) and the spin density $\boldsymbol{n}$ (observable $\frac{1}{2}\boldsymbol{\sigma}$), given by

$$n_0(x) = \int_{\mathbb{R}^{2d}} w_0(x,\,p)\,dp, \qquad \boldsymbol{n}(x) = \int_{\mathbb{R}^{2d}} \boldsymbol{w}(x,\,p)\,dp.$$

We remark that an operator S representing the state of a quantum system must be a positive operator with unit trace. In particular, $(S\psi)(x)$ is a positive definite matrix for all two-component wave functions ψ and for a.e. x. This fact leads to constraints on the functions n_k, $k = 0, 1, 2, 3$, namely $n_0 \geq 0$ and $n_1, n_2, n_3 \in \mathbb{R}$ with

$$n_1^2 + n_2^2 + n_3^2 \leq n_0^2$$

(for a.e. x). If $n_1^2 + n_2^2 + n_3^2 = n_0^2$, then S and $w = \operatorname{Op}^{-1}(S)$ represent a pure state while if $n_1^2 + n_2^2 + n_3^2 < n_0^2$, then S and w represent a "mixed" (statistical) state.

3 Transport Picture

We shall now derive a mesoscopic-level (kinetic) transport model for our two-dimensional electron gas.

3.1 Transport Equation

Let $S(t)$ be the time-dependent density operator, representing the statistical quantum mechanical state at time t, let $\varrho(x, y, t)$ be the associated density matrix (i.e. the integral kernel of $S(t)$) and $w(x, p, t) = \mathcal{W}(\varrho)$ the corresponding Wigner function. The evolution equation for $S(t)$ is the statistical version of the Schrödinger equation,

that is the Von Neumann equation

$$i\hbar\partial_t S = (H+V)S - S(H+V),$$

where H is the Rashba Hamiltonian (5) and $V = V(x)\sigma_0$ represents an external electrostatic potential (e.g. a gate potential). In terms of the density matrix, this equation reads as follows:

$$i\hbar\partial_t \varrho = \left(-\frac{\hbar^2}{2m}\left(\Delta_x - \Delta_y\right) + V(x) - V(y)\right)\varrho - i\hbar\alpha_R\left(\nabla_x^\perp \cdot \boldsymbol{\sigma}\varrho - \nabla_y^\perp \varrho \cdot \boldsymbol{\sigma}\right)$$

The evolution equation for the Wigner function w is obtained by applying the Wigner transformation to both sides of the last equation. This results in

$$i\hbar\partial_t w = \{h+V, w\}_\#,$$

where

$$h(x,p) = \frac{|p|^2}{2m}\sigma_0 + \alpha_R p^\perp \cdot \boldsymbol{\sigma}$$

is the symbol of the Rashba Hamiltonian (as usual $p^\perp = p \times \boldsymbol{e}_3 = (p_2, -p_1, 0)$) and $\{\cdot,\cdot\}_\#$ is the Moyal bracket

$$\{a,b\}_\# = a\#b - b\#a.$$

By explicitly computing this bracket and decomposing the matrix equation in the Pauli components, we obtain the following system for the trace and spin parts of w:

$$\begin{aligned}
\partial_t w_0 &= -\frac{1}{m}p\cdot\nabla_x w_0 - \alpha_R\nabla_x^\perp\cdot\boldsymbol{w} + \Theta_\hbar[V]w_0,\\
\partial_t \boldsymbol{w} &= -\frac{1}{m}p\cdot\nabla_x \boldsymbol{w} - \alpha_R\nabla_x^\perp w_0 + \Theta_\hbar[V]\boldsymbol{w} + \frac{2\alpha_R}{\hbar}p^\perp\times\boldsymbol{w},
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
\Theta_\hbar[f] &= \frac{1}{i\hbar}\left[f\left(x + \frac{i\hbar}{2}\nabla_p\right) - f\left(x - \frac{i\hbar}{2}\nabla_p\right)\right]\\
&= \sum_{j=0}^{\infty}(-1)^j\left(\frac{\hbar}{2}\right)^{2j}\sum_{|\alpha|=2j+1}\frac{1}{\alpha!}\nabla_x^\alpha f\nabla_p^\alpha
\end{aligned} \tag{8}$$

is the usual force term of the Wigner equation [3, 16, 22]. Note that the leading order term of the last expansion corresponds to the force term in the classical transport

equation, namely

$$\Theta_\hbar[V] \xrightarrow{\hbar\to 0} \nabla_x V \cdot \nabla_p .$$

In order to study the diffusion asymptotics of our system, the purely Hamiltonian dynamics described by Eq. (7) must be supplemented with a collisional mechanism. If we want to remain in a rigorous quantum-mechanical setting, we cannot expect to adopt a detailed description of collisions. However, since our goal is to obtain the diffusive limit of our model, only very general properties of the collision dynamics are needed, such as conservation properties. Therefore, the optimal strategy to insert a relatively simple collisional mechanism, and to respect at the same time the rules of quantum mechanics, is to adopt a relaxation-time term making the system relax to a suitable quantum equilibrium state [1, 8, 9, 16]. We therefore re-write Eq. (7) with suitable relaxation-time terms:

$$\begin{aligned}
\partial_t w_0 &= -\frac{1}{m} p \cdot \nabla_x w_0 - \alpha_R \nabla_x^\perp \cdot \boldsymbol{w} + \Theta_\hbar[V] w_0 + \frac{1}{\tau_p}(g_0 - w_0) \\
\partial_t \boldsymbol{w} &= -\frac{1}{m} p \cdot \nabla_x \boldsymbol{w} - \alpha_R \nabla_x^\perp w_0 + \Theta_\hbar[V] \boldsymbol{w} + \frac{2\alpha_R}{\hbar} p^\perp \times \boldsymbol{w} + \frac{1}{\tau_p}(\boldsymbol{g} - \boldsymbol{w})
\end{aligned} \tag{9}$$

where $g = g_0\sigma_0 + \boldsymbol{g} \cdot \boldsymbol{\sigma}$ is the equilibrium Wigner function that will be specified later on.

Before that, and in view of the diffusion asymptotics, let us rewrite Eq. (9) in a non-dimensional form. Let T_0 be the (given) temperature of the thermal bath with which our electron population is assumed to be in equilibrium. The reference energy E_0 is taken as the thermal energy, given by

$$k_B T_0 = E_0,$$

where k_B denotes the Boltzmann constant. The associated thermal momentum is

$$p_0 = \sqrt{m k_B T_0}.$$

Let us also fix a reference length x_0 (e.g., the device size) and take the reference time t_0 as

$$t_0 = \frac{m x_0}{p_0},$$

which is the time it takes an electron, traveling at the reference thermal velocity, to cross the reference length. Then, in Eq. (9) we switch to non-dimensional variables with the substitutions

$$x \to x_0 x, \qquad t \to t_0 t, \qquad p \to p_0 p, \qquad V \to E_0 V$$

(for the sake of simplicity, the new non-dimensional variables are denoted by the same symbols as the dimensional ones). We obtain in this way

$$\begin{aligned} \partial_t w_0 &= -p \cdot \nabla_x w_0 - \epsilon\alpha \nabla_x^\perp \cdot \boldsymbol{w} + \Theta_\epsilon[V] w_0 + \frac{1}{\tau}(g_0 - w_0)\,, \\ \partial_t \boldsymbol{w} &= -p \cdot \nabla_x \boldsymbol{w} - \epsilon\alpha \nabla_x^\perp w_0 + \Theta_\epsilon[V]\boldsymbol{w} + 2\alpha p^\perp \times \boldsymbol{w} + \frac{1}{\tau}(\boldsymbol{g} - \boldsymbol{w})\,. \end{aligned} \tag{10}$$

Here, two important non-dimensional parameters have been introduced,

$$\epsilon = \frac{\hbar}{x_0 p_0}, \qquad \tau = \frac{\tau_p}{t_0}.$$

The "semi-classical" parameter ϵ is the scaled Planck constant: roughly speaking, the smaller it is, the further we zoom out from the quantum scale and approach the classical scale. The diffusive parameter τ is the scaled collision time: the smaller it is, the more collisions occur in the reference time, making the diffusive regime predominate on the "ballistic" one. Moreover,

$$\alpha = \frac{m x_0 \alpha_R}{\hbar}$$

is the non-dimensional Rashba constant. Since $\epsilon\alpha = m\alpha_R/p_0$, we see that α is the coefficient of proportionality between ϵ and the ratio of α_R (which has the physical dimension of a velocity) and the thermal velocity p_0/m. This choice makes the Rashba constant scale with ϵ and gives the correct result in the semiclassical limit $\epsilon \to 0$ (see Sect. 4.3 and Ref. [4]).

3.2 Maximum Entropy Principle

We now come to the description of the quantum equilibrium function appearing in the transport equation (10). According to the theory developed in Refs. [8, 9] (see also [3, 16]), we choose the equilibrium Wigner function $g = g_0\sigma_0 + \boldsymbol{g} \cdot \boldsymbol{\sigma}$ as the minimiser of a suitable quantum entropy-like functional, with the constraint of positivity and fixed densities, which is the quantum version of the well-known Maximum Entropy Principle. Physically speaking, this means that g is assumed to be the most probable microscopic state compatible with the observed macroscopic density. This is rigorously expressed in our case as follows.

Quantum Maximum Entropy Principle (Q-MEP) *Let $n = n_0\sigma_0 + \boldsymbol{n} \cdot \boldsymbol{\sigma}$ be a given matrix-valued function of x and t, with*

$$n_0 > 0, \qquad n_1, n_2, n_3 \in \mathbb{R}, \qquad n_1^2 + n_2^2 + n_3^2 < n_0^2,$$

for a.e. $x \in \mathbb{R}^2$ *and* $t > 0$. *The equilibrium Wigner function* g *is given by*

$$g = \operatorname{argmin} \{\mathcal{H}(w) \mid \mathrm{Op}(w) > 0,\ \langle w \rangle = n\},$$

where $\mathcal{H}$ *is the quantum free-energy functional given (in non-dimensional variables) by*

$$\mathcal{H}(w) = \frac{1}{2} \operatorname{tr} \left(\int_{\mathbb{R}^6} (w \mathcal{L}og(w) - w + h\#w)(x, p)\, dx dp \right) \tag{11}$$

and $\mathcal{L}og$ *is the "quantum logarithm" defined as*

$$\mathcal{L}og(w) = \mathcal{W}(\log(\mathrm{Op}(w)))$$

*(*log *being the operator logarithm).*

Note that the constraints on n are consistent with the requirement that w represents a quantum mixed state, according to the remark at the end of Sec. 2 (see also [18, 19]).

Then, g is defined as the Wigner function that minimises the quantum entropy (or, more precisely, the free energy, which is the energy minus the entropy) under the constraint of the given density. Note that the condition $\mathrm{Op}(g) > 0$ means that g must be a genuine Wigner function (i.e. the Wigner transform of a density operator). The entropy functional (11) is the phase-space equivalent of the Von Neumann entropy (free energy, more precisely): if $S = \mathrm{Op}(w)$ is the density operator, then

$$\mathcal{H}(w) = \operatorname{Tr}(S \log(S) - S + HS).$$

A formal proof of the following theorem makes use of the mathematical techniques adopted in similar contexts (see, e.g., Ref. [2]); however the application of these techniques to the full-spin case is far from being straightforward and a detailed proof is deferred to a forthcoming paper. Rigorous proofs also exist, but only for the simpler case of a one-dimensional system of scalar (non-spinorial) particles in an interval with periodic boundary conditions, see Refs. [18, 19].

Theorem *The matrix-valued Wigner function* g, *satisfying the above constrained minimisation problem, exists and is given by*

$$g = \mathcal{E}xp(-h + a), \qquad \langle g \rangle = n, \tag{12}$$

where $a = a_0\sigma_0 + \boldsymbol{a} \cdot \boldsymbol{\sigma}$ *is a matrix of Lagrange multipliers and*

$$\mathcal{E}xp(w) = \mathcal{W}(\exp(\mathrm{Op}(w)))$$

(with exp *the operator exponential).*

Our model is now completed by using g given by (12) as the equilibrium function in the Wigner equation (9). We remark that the quantum equilibrium function g is quite a complicated object, it is a non-local function of the Lagrange multipliers, which are implicitly related to the densities n_0 and $\boldsymbol{n}$ by the four integral constraints $\langle g \rangle = n$, i.e. $\langle g_0 \rangle = n_0$ and $\langle \boldsymbol{g} \rangle = \boldsymbol{n}$. However, it is possible to make the model more explicit by performing a semiclassical expansion of g, made possible by the semiclassical expansion (6) of the Moyal product.

4 Diffusion Picture

Let us now formally derive the diffusion asymptotics of the kinetic model introduced in the previous section.

4.1 Chapman–Enskog Expansion

To shorten the notation, we denote by $\mathcal{T}$ the transport operator

$$\mathcal{T} w := \frac{1}{i\epsilon} \{h + V, w\}_{\#} = \left(-p \cdot \nabla_x w_0 - \epsilon \alpha \nabla_x^{\perp} \cdot \boldsymbol{w} + \Theta_\epsilon[V] w_0 \right) \sigma_0 + \left(-p \cdot \nabla_x \boldsymbol{w} - \epsilon \alpha \nabla_x^{\perp} w_0 + \Theta_\epsilon[V] \boldsymbol{w} + 2\alpha p^{\perp} \times \boldsymbol{w} \right) \cdot \boldsymbol{\sigma},$$

so that the scaled Wigner equation (10) is concisely written as

$$\tau \partial_t w = \tau \mathcal{T} w + g - w. \tag{13}$$

The diffusion asymptotics is obtained by means of the Chapman–Enskog expansion [6, 16], by expanding the equation for the macroscopic density $n = \langle w \rangle$,

$$\partial_t n = \partial_t^{(0)} n + \tau \partial_t^{(1)} n + \tau^2 \partial_t^{(2)} n + \cdots,$$

and the microscopic state,

$$w = w^{(0)} + \tau w^{(1)} + \tau^2 w^{(2)} + \cdots. \tag{14}$$

We remark that it is only the equation for n that is expanded, and not n itself, which is an $O(1)$ quantity with respect to τ.

Integrating (13) with respect to p and recalling that $\langle g - w \rangle = 0$ (which follows from (12) and reflects the conservation of the number of particles and the spin in the

collisions), we can identify the k-th order time derivative of n by

$$\partial_t^{(k)} n = \langle \mathcal{T} w^{(k)} \rangle.$$

To compute $w^{(k)}$, we substitute (14) in (13). This yields, at leading and at first order in τ,

$$w^{(0)} = g, \quad w^{(1)} = \mathcal{T} g - \partial_t g,$$

respectively. Therefore,

$$\partial_t^{(0)} n = \langle \mathcal{T} g \rangle, \quad \partial_t^{(1)} n = \langle \mathcal{T}\mathcal{T} g \rangle - \langle \mathcal{T} \partial_t g \rangle. \tag{15}$$

The function g depends on time only through its (functional) dependence on n, according to (12). Then, at the same order of approximation, we can also write

$$\partial_t g = \frac{\delta g}{\delta n} \circ \partial_t n \approx \frac{\delta g}{\delta n} \circ \partial_t^{(0)} n = \frac{\delta g}{\delta n} \circ \langle \mathcal{T} g \rangle, \tag{16}$$

where $\circ$ denotes the componentwise product, resulting from the chain rule

$$\frac{\delta g}{\delta n} \circ \partial_t n \equiv \sum_{k=0}^{3} \frac{\delta g}{\delta n_k} \partial_t n_k.$$

Using (15) and (16) and neglecting higher-order terms, we obtain the quantum drift-diffusion (QDDE) equation for n:

$$\partial_t n = \langle \mathcal{T} g \rangle + \tau \langle \mathcal{T}\mathcal{T} g \rangle - \tau \left\langle \mathcal{T} \frac{\delta g}{\delta n} \right\rangle \circ \langle \mathcal{T} g \rangle. \tag{17}$$

We remark the following:

1. The QDDE (17) is, formally, a closed equation for n, since g depends on n through (12).
2. The term $\tau \langle \mathcal{T}\mathcal{T} g \rangle$ is the truly diffusive term in the equation, to the extent that it is the only term that appears in the standard cases (i.e. classical or quantum scalar particles [8, 9, 16]).
3. The term $\langle \mathcal{T} g \rangle$, which is equal to zero for standard particles, does not vanish for spin-orbit electrons (see below). This is the reason why we were forced to use a hydrodynamic scaling instead of the usual diffusive one. As a consequence, the Chapman–Enskog procedure produces the additional terms $\langle \mathcal{T} g \rangle$ and $\tau \langle \mathcal{T} \frac{\delta g}{\delta n} \rangle \circ \langle \mathcal{T} g \rangle$ in the diffusive equations.

The last point deserves some additional comments. In the usual situation, the diffusion asymptotics is derived from the transport, or kinetic, equation in the so-

called diffusive scaling, i.e. obtained by a further rescaling of time, $t \mapsto t/\tau$. This means that the system is observed on a very long time scale, in which the collision time is τ^2 (the hydrodynamic scaling being instead the one in which the collision time is τ). This is because in the standard case, if collisions do not conserve the momentum, one has $\langle \mathcal{T} g \rangle = 0$, which reflects the fact that the equilibrium state carries no current. Therefore, a purely diffusive current manifests in the longer, diffusive, time scale. In the present situation, even though the collisions do not conserve the momentum, g still carries a current, that is due to the peculiar form of the spin-orbit interaction. This implies that a current, $\langle \mathcal{T} g \rangle$, already appears at the hydrodynamic scale. Moreover, at order τ the additional term $\tau \langle \mathcal{T} \frac{\delta g}{\delta n} \rangle \circ \langle \mathcal{T} g \rangle$ appears. A formally analogous term appears also in the derivation of the classical hydrodynamic equation: in that case it contains the viscosity [6]. In the present context, its interpretation is not so clear. We point out that the two non-standard terms $\langle \mathcal{T} g \rangle$ and $\tau \langle \mathcal{T} \frac{\delta g}{\delta n} \rangle \circ \langle \mathcal{T} g \rangle$ are "small" in a semiclassical perspective, because, as we shall see later, they vanish at leading order in ϵ.

4.2 Quantum Drift-Diffusion Equation

In order to recast (17) in a more explicit form, note that we can write

$$\mathcal{T} g = \frac{1}{i\epsilon} \{h + V, g\}_\# = \frac{1}{i\epsilon} \{h - a, g\}_\# + \frac{1}{i\epsilon} \{V + a, g\}_\# = \frac{1}{i\epsilon} \{V + a, g\}_\#,$$

where a is the matrix of Lagrange multipliers; see (12). In fact,

$$\{h - a, g\}_\# = 0, \tag{18}$$

because $g = \mathcal{E}xp(-h + a)$ and therefore, (18) is just the expression in the Wigner-Moyal formalism of the commutativity of the operator $H - A$ with its exponential $\exp(-H + A)$. Recalling that V and a do not depend on p, we find that

$$\begin{aligned} \mathcal{T} g = \frac{1}{i\epsilon} \{V + a, g\}_\# &= (\Theta_\epsilon[V + a_0] g_0 + \Theta_\epsilon[\boldsymbol{a}] \cdot \boldsymbol{g})\, \sigma_0 \\ &+ \left(\Theta_\epsilon[V + a_0] \boldsymbol{g} + \Theta_\epsilon[\boldsymbol{a}] g_0 + \epsilon^{-1} \Theta_\epsilon^+[\boldsymbol{a}] \times \boldsymbol{g} \right) \cdot \boldsymbol{\sigma}, \end{aligned} \tag{19}$$

where Θ_ϵ is given by (8) and Θ_ϵ^+ is defined as follows:

$$\begin{aligned} \Theta_\epsilon^+[f] &= \frac{1}{i\epsilon} \left[f\left(x + \frac{i\epsilon}{2} \nabla_p \right) + f\left(x - \frac{i\hbar}{2} \nabla_p \right) \right] \\ &= \sum_{j=0}^{\infty} (-1)^j \left(\frac{\epsilon}{2} \right)^{2j} \sum_{|\alpha|=2j} \frac{1}{\alpha!} \nabla_x^\alpha f \nabla_p^\alpha. \end{aligned} \tag{20}$$

We infer from (8) (with ϵ instead of $\hbar$) and (20) the following properties:

$$\langle \Theta_\epsilon[f]w\rangle = 0, \quad \langle p_j \Theta_\epsilon[f]w\rangle = -\partial_{x_j} f\,\langle w\rangle, \quad \langle \Theta_\epsilon^+[f]w\rangle = 2f\langle w\rangle.$$

Then, recalling that $\langle g\rangle = n$,

$$\langle \mathcal{T} g\rangle = 2\epsilon^{-1}\boldsymbol{a}\times\boldsymbol{n}\cdot\boldsymbol{\sigma}. \tag{21}$$

This represents explicitly the above-mentioned residual spin-orbit current at equilibrium. We see that a condition for this current to vanish is

$$\boldsymbol{a}\times\boldsymbol{n} = 0, \tag{22}$$

which is equivalent to the commutativity of the matrices n and a (see Eq. (4)). This explains why in Ref. [4], concerning the bipolar case, only the standard diffusion term $\langle \mathcal{T}\mathcal{T} g\rangle$ has been found: in that case the matrices n and a are both diagonal.

Now, for a generic w, we have

$$\begin{aligned}\langle \mathcal{T} w\rangle = &\left(-\partial_j\langle p_j w_0\rangle - \epsilon\alpha\nabla^\perp\cdot\langle\boldsymbol{w}\rangle\right)\sigma_0 \\ &+\left(-\partial_j\langle p_j\boldsymbol{w}\rangle - \epsilon\alpha\nabla^\perp\langle w_0\rangle + 2\alpha\langle p^\perp\times\boldsymbol{w}\rangle\right)\cdot\boldsymbol{\sigma}\end{aligned} \tag{23}$$

(where $\partial_j \equiv \partial_{x_j}$ and summation over $j = 1, 2$ is assumed). Substituting $w = \mathcal{T} g$ in (23), where $\mathcal{T}$ is defined in (19), yields

$$\begin{aligned}\langle \mathcal{T}\mathcal{T} g\rangle = &\left\{\partial_j\left[n_0\,\partial_j(V+a_0) + \boldsymbol{n}\cdot\partial_j\boldsymbol{a}\right] - 2\alpha\nabla^\perp\cdot(\boldsymbol{a}\times\boldsymbol{n})\right\}\sigma_0 \\ &+\Big\{\partial_j\left[\boldsymbol{n}\,\partial_j(V+a_0) + n_0\partial_j\boldsymbol{a} - 2\epsilon^{-1}\boldsymbol{a}\times\langle p_j\boldsymbol{g}\rangle\right] \\ &-2\alpha\left[\nabla^\perp(V+a_0)\times\boldsymbol{n} + (\nabla^\perp\times\boldsymbol{a})n_0 - 2\epsilon^{-1}\alpha\langle p^\perp\times(\boldsymbol{a}\times\boldsymbol{g})\rangle\right]\Big\}\cdot\boldsymbol{\sigma}.\end{aligned} \tag{24}$$

Equations (21) and (24) express the first and the second terms in the quantum drift-diffusion equations (17) in terms of the Lagrange multipliers (no such explicit expression has been found for the third term).

We remark that the Lagrange multipliers depend on the densities n via the constraint $\langle g\rangle = n$. Even though this fact makes (17) a closed equation for n, nevertheless the dependence of a on n is very implicit and non-local, since it comes from integral constraints on a quantum exponential, involving back and forth Wigner transforms. Numerical methods to solve QDDE of this kind exist [5, 11]. However, the optimal use of a QDDE is expanding it semiclassically (i.e. in powers of ϵ), in order to obtain "quantum corrections" to classical QDD [2, 4, 8, 9, 16]. This will be the subject of a future work. For the moment, we shall limit ourselves to consider the semiclassical limit $\epsilon \to 0$ of (17), just to check if our model allows us to recover

the semiclassical drift-diffusion equations for spin-orbit electrons already known in the literature [10].

4.3 Semiclassical Limit

The semiclassical limit is obtained from the fully quantum model (17), (21), and (24) by expanding g and a in powers of ϵ and retaining only the terms of order $O(\epsilon^0)$. This would require the expansions of g and a up to $O(\epsilon^1)$, because of the terms of order ϵ^{-1} appearing in (21) and (24). So it is easier to compute directly the right-hand side of (17), neglecting all terms of order ϵ and using the leading-order approximation of g, that is

$$g(x, p, t) \approx \mathrm{e}^{-p^2/2}\mathrm{e}^{a(x,t)} = \frac{1}{2\pi}\mathrm{e}^{-p^2/2}n(x, t).$$

We remark that this is indeed the semiclassical equilibrium distribution (see, e.g., Ref. [10]). Within this approximation, we have $\langle \mathcal{T} g\rangle \approx 0$ (and then, of course, also $\langle \mathcal{T}\frac{\delta g}{\delta n}\rangle \circ \langle \mathcal{T} g\rangle \approx 0$) as well as

$$\begin{aligned}\langle \mathcal{T}\mathcal{T} g\rangle \approx{}& \partial_j\left(\partial_j n_0 + n_0\partial_j V\right)\sigma_0\\ &+ \left\{\partial_j\left[\partial_j \boldsymbol{n} + \boldsymbol{n}\partial_j V + 4\alpha A_j(\boldsymbol{n})\right] - 2\alpha\nabla^{\perp}V\times\boldsymbol{n} - 4\alpha^2 B(\boldsymbol{n})\right\}\cdot\boldsymbol{\sigma},\end{aligned}$$

where

$$A_1(\boldsymbol{n}) = \begin{pmatrix} n_3\\ 0\\ -n_1\end{pmatrix}, \qquad A_2(\boldsymbol{n}) = \begin{pmatrix} 0\\ n_3\\ -n_2\end{pmatrix}, \qquad B(\boldsymbol{n}) = \begin{pmatrix} n_1\\ n_2\\ 2n_3\end{pmatrix}.$$

Then, as a leading-order approximation of the quantum drift-diffusion equations (17), we arrive to

$$\begin{aligned}\partial_t n_0 &= \partial_j\left(\partial_j n_0 + n_0\partial_j V\right),\\ \partial_t \boldsymbol{n} &= \partial_j\left[\partial_j \boldsymbol{n} + \boldsymbol{n}\partial_j V + 4\alpha A_j(\boldsymbol{n})\right] - 2\alpha\nabla^{\perp}V\times\boldsymbol{n} - 4\alpha^2 B(\boldsymbol{n}).\end{aligned}$$

The semiclassical drift-diffusion equations derived in Ref. [10] coincide with our equations in the case of constant relaxation time and purely spin-orbit interaction field. (In Ref. [10] an additional term, even in p, is introduced in the spinorial part of the Hamiltonian, $\boldsymbol{h}$, which can be used to model, e.g., an external magnetic field: this term could also be considered in our framework but we preferred to neglect it for the sake of simplicity.) We remark that each of the Pauli components diffuses according to a classical drift-diffusion equation and, moreover, the spin has the

additional current term $4\alpha A_j(\boldsymbol{n})$, coming from spin-orbit interactions, a relaxation term $-4\alpha^2 B(\boldsymbol{n})$, and an interaction with the external potential, $-2\alpha\nabla^\perp V \times \boldsymbol{n}$, which shows the capability of controlling the spin by means of an applied voltage.

5 Conclusions

In this paper, we have derived quantum drift diffusion equations (QDDE) for a 2DEG with spin-orbit interaction of Rashba type. The derivation is based on the quantum version of the maximum entropy principle (Q-MEP), as proposed in Refs. [8, 9]. To our knowledge, this is the first application of the Q-MEP to the full spin structure and not only to the spin polarization (i.e. the projection of the spin vector on a given axis).

Our derivation starts with the formulation of a kinetic model which has an Hamiltonian part (basically, the mixed-state Schrödinger equation in the phase-space formulation) and a non-conservative, collisional term in the relaxation time approximation. Here, the quantum equilibrium state given by the Q-MEP appears.

Assuming that the relaxation time is a small parameter in the problem, we apply the Chapman–Enskog technique to derive the quantum drift-diffusion model (17), (21), and (24). It forms a system of four equations: one for the charge density n_0 and three for the spin-vector components $\boldsymbol{n} = (n_1, n_2, n_3)$. Such equations are non-local in the components n_k, since they are expressed in terms of Lagrange multipliers that are connected with the densities by the (integral) constraint that the equilibrium state possesses such densities. This aspect of the model is not different from the analogous QDDE obtained in the scalar [8, 16] or bipolar [4] cases.

A new feature of the present model is that the application of the Chapman–Enskog technique is not the standard one for the diffusive case and resembles more to the Chapman–Enskog expansion of the hydrodynamic case. This is due to the fact that, due to the peculiar form of the spin-orbit interaction, the equilibrium state has no zero current. In the derivation, we have obtained a general condition, Eq. (22), for such current to vanish.

Typically, the QDDE are expanded semiclassically, i.e. in powers of the scaled Planck constant ϵ, which allows for an approximation of the QDDE by a local model consisting in classical diffusive equations with "quantum corrections". Here, we just computed the approximation at the leading order, in order to compare the semiclassical limit of our model with the semiclassical models already existing in the literature. The semiclassical expansion of our QDDE, which is not an easy task, goes beyond the aim of the present paper and is deferred to a work in preparation.

Acknowledgments The last two authors acknowledge partial support from the Austrian Science Fund (FWF), grants F65, P30000, P33010, and W1245.

References

1. Arnold, A.: Self-consistent relaxation-time models in quantum mechanics. Commun. Partial Differ. Equ. **21**, 473–506 (1996)
2. Barletti, G., Frosali, G.: Diffusive limit of the two-band K.P model for semiconductors. J. Stat. Phys. **139**, 280–306 (2010)
3. Barletti, L., Frosali, G., Morandi, O.: Kinetic and hydrodynamic models for multi-band quantum transport in crystals. In: Ehrhardt, M., Koprucki, T. (eds.) Multi-Band Effective Mass Approximations: Advanced Mathematical Models and Numerical Techniques, pp. 3–56. Springer, Berlin (2014)
4. Barletti, L., Méhats, F.: Quantum drift-diffusion modeling of spin transport in nanostructures. J. Math. Phys. **51**, 053304 (2010)
5. Barletti, L., Méhats, F., Negulescu, C., Possanner, S.: Numerical study of a quantum-diffusive spin model for two-dimensional electron gases. Commun. Math. Sci. **13**, 1347–1378 (2015)
6. Cercignani, C.: The Boltzmann Equation and Its Applications. Springer, New York (1988)
7. Chainais-Hillairet, C., Jüngel, A., Shpartko, P.: A finite-volume scheme for a spinorial matrix drift-diffusion model for semiconductors. Numer. Methods Partial Differ. Equ. **32**, 819–846 (2016)
8. Degond, P., Méhats, F., Ringhofer, C.: Quantum energy-transport and drift-diffusion models. J. Stat. Phys. **118**, 625–667 (2005)
9. Degond, P., Ringhofer, C.: Quantum moment hydrodynamics and the entropy principle. J. Stat. Phys. **112**, 587–628 (2003)
10. El Hajj, R.: Diffusion models for spin transport derived from the spinor Boltzmann equation. Commun. Math. Sci. **12**, 565–592 (2014)
11. Gallego, S., Méhats, F.: Entropic discretization of a quantum drift-diffusion model. SIAM J. Numer. Anal. **43**, 1828–1849 (2006)
12. García-Cervera, C., Wang, X.-P.: Spin-polarized transport: existence of weak solutions. Discrete Contin. Dyn. Sys. Ser. B **7**, 87–100 (2007)
13. Glitzky, A.: Analysis of a spin-polarized drift-diffusion model. Adv. Math. Sci. Appl. **18**, 401–427 (2008)
14. Glitzky A., Gärtner, K.: Existence of bounded steady state solutions to spin-polarized drift-diffusion systems. SIAM J. Math. Anal. **41**, 2489–2513 (2010)
15. Holzinger, P., Jüngel, A.: Large-time asymptotics for a matrix spin drift-diffusion model. J. Math. Anal. Appl. **486**, 123887 (2020)
16. Jüngel, A.: Transport Equations for Semiconductors. Springer, Berlin (2009)
17. Jüngel, A., Negulescu, C., Shpartko, P.: Bounded weak solutions to a matrix drift-diffusion model for spin-coherent electron transport in semiconductors. Math. Models Methods Appl. Sci. **25**, 929–958 (2015)
18. Méhats, F., Pinaud, O.: An inverse problem in quantum statistical physics. J. Stat. Phys. **140**, 565–602 (2010)
19. Méhats, F., Pinaud, O.: A problem of moment realizability in quantum statistical physics. Kinetic Relat. Models **4**, 1143–1158 (2011)
20. Possanner, S., Negulescu, C.: Diffusion limit of a generalized matrix Boltzmann equation for spin-polarized transport. Kinetic Relat. Models **4**, 1159–1191 (2011)
21. Pu, X., Gu, B.: Global smooth solutions for the one-dimensional spin-polarized transport equation. Nonlin. Anal. **72**, 1481–1487 (2010)
22. Zachos, C.K., Fairlie, D.B., Curtright, T.L. (eds.): Quantum Mechanics in Phase Space. An Overview with Selected Papers. World Scientific, Hackensack (2005)
23. Zamponi, N.: Analysis of a drift-diffusion model with velocity saturation for spin-polarized transport in semiconductors. J. Math. Anal. Appl. **420**, 1167–1181 (2014)
24. Zamponi, N., Jüngel, A.: Two spinorial drift-diffusion models for quantum electron transport in graphene. Commun. Math. Sci. **11**, 927–950 (2013)
25. Žutić, I., Fabian, J., Das Sarma, S.: Spintronics: fundamentals and applications. Rev. Mod. Phys. **76**, 323–410 (2002)

A Kinetic BGK Relaxation Model for a Reacting Mixture of Polyatomic Gases

Marzia Bisi and Romina Travaglini

Abstract We present a kinetic model of BGK-type for a mixture of four polyatomic gases, each one having its own number of internal energy levels, subject also to a bimolecular and reversible chemical reaction. A single relaxation operator is constructed for each gas component, with auxiliary parameters depending on main macroscopic fields and able to take into account all mechanical and reactive interactions affecting the considered component. Preservation of correct collision equilibria, conservation laws, and H-theorem is proved, and some numerical simulations in space homogeneous conditions are shown.

1 Introduction

Boltzmann kinetic equations for gas mixtures are very complicated to deal with, since they are integro-differential equations with a collision term provided by a sum of binary Boltzmann operators, each one describing elastic collisions between particles of the considered species and particles of only another constituent [10]. The kinetic system becomes even more cumbersome in the presence of polyatomic molecules or of chemical reactions that change the nature of the colliding particles. For this reason, simpler kinetic models have been presented in the literature, mainly in the spirit of the BGK relaxation model proposed by Bhatnagar, Gross, Krook in 1954 for a single gas [3].

The generalization of the BGK model to a gas mixture is not obvious, and several consistent ways of modelling have been investigated. Some pioneering works for inert mixtures, that have given rise to several generalizations and applications, are Ref. [21] by McCormack and Ref. [1] by Andries, Aoki, Perthame. The former is still used to face fluid-dynamic problems as flows in microchannels [23], and the latter has been extended also to mixtures of monoatomic gases undergoing

M. Bisi (✉) · R. Travaglini
Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università di Parma, Parma, Italy
e-mail: marzia.bisi@unipr.it; romina.travaglini@unipr.it

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_3

simple bimolecular and reversible chemical reactions [5, 15]. The BGK model in [21] has a linearized form, and shows a sum of binary relaxation operators in each kinetic equation; this idea has been generalized in different ways, constructing also non-linear BGK models for inert mixtures, see for instance [16, 18] and the more recent paper [9] where also a comparison among the existing models has been done. On the other hand, the BGK model proposed in [1] shows a unique relaxation operator for each species, being thus much simpler to be managed from the mathematical point of view, and for this reason it turns out to be well suited to take into account also chemical reactions for monoatomic particles, of course at the price of more complicated expressions for auxiliary parameters affecting Maxwellian attractors [5], or of suitable simplifying assumptions in the reactive contributions [15].

In view of physical applications, for instance the investigation of gas flows in the atmosphere, even polyatomic gases should be included in the kinetic description. Boltzmann-type models for polyatomic particles have been built up, modelling the non-translational degrees of freedom by means of an additional internal energy variable, that could be discrete [14] or continuous [12]. Macroscopic equations at different levels of accuracy have been consistently derived owing to an asymptotic Chapman–Enskog procedure in both cases of discrete [13] or continuous internal energy [2]. BGK approximations of these Boltzmann models for polyatomic gases have been recently proposed, with various assumptions. At first, mixtures of only polyatomic gases having the same number of internal energy levels have been considered [4], and then a consistent BGK model has been developed also in the case of continuous internal energy [7]. However, for physical applications, kinetic descriptions allowing the simultaneous presence of monoatomic and polyatomic particles are highly desirable. This has been the main motivation of the recent paper [8], where we have extended the BGK model proposed in [4] to an inert mixture constituted by both monoatomic and polyatomic species, with each polyatomic one characterized by its own number of discrete internal energy levels. In this paper, we aim at generalizing this model to a reacting mixture; specifically, we consider a mixture of four gas species, G^i, $i = 1, \ldots, 4$, whose particles, besides elastic collisions and inelastic transitions from one internal energy level to another, are subject also to the reversible chemical reaction $G^1 + G^2 \leftrightarrows G^3 + G^4$, where a pair of reacting particles of species (G^1, G^2) provides, as products, a pair belonging to (G^3, G^4), or vice versa.

The paper will be organized as follows. In Sect. 2, we briefly introduce the physical reacting frame that we are considering, and we present the construction of our BGK model; we prove that all disposable parameters appearing in the BGK operators may be determined in terms of the actual species densities, velocities and temperatures in such a way that correct collision equilibria of the reactive Boltzmann equations are preserved, as well as conservation laws and the validity of the H-theorem. Then, in Sect. 3 some numerical simulations of trends to equilibrium of main macroscopic fields in space homogeneous conditions are shown and commented on. Finally, Sect. 4 contains some concluding remarks.

2 Reacting BGK Model for Polyatomic Gases

We take into account a mixture of four polyatomic gas species, G^i, $i = 1, \ldots, 4$. The reversible chemical reaction in which the four gas species are involved is

$$G^1 + G^2 \leftrightarrows G^3 + G^4. \tag{1}$$

As done in [8], each gas species G^i will be characterized by a mass m^i and a certain number L^i of discrete energy levels. Thus, it will be seen as a mixture of components C^i_j, $j = 1, \ldots, L^i$, each one corresponding to a different energy level, denoted by E^i_j. In the frame of the same gas G^i the energy levels are assumed (without loss of generality) to be increasing with respect to the subindex j, namely $E^i_j < E^i_k$ for any $j, k = 1, \ldots, L^i$ with $j < k$. As concerns masses, according to the conservation law, they have to satisfy the relation $m^1 + m^2 = m^3 + m^4$.

The distribution function of the component C^i_j is denoted by

$$f^i_j(t, \mathbf{x}, \mathbf{v}), \quad i = 1, \ldots, 4, \quad j = 1, \ldots, L^i.$$

We now consider possible interactions between particles, that will be, as usual, only binary instantaneous collisions. We may have, besides classical elastic collisions, also inelastic encounters in which particles may change their internal energy (as those described in [8]), but also their nature, according to chemical reaction (1).

A generic encounter is given by

$$C^i_j + C^h_k \to C^l_m + C^n_p, \quad 1 \le i, h, l, n \le 4 \quad \begin{array}{l} 1 \le j \le L^i \\ 1 \le k \le L^h \\ 1 \le m \le L^l \\ 1 \le p \le L^n. \end{array} \tag{2}$$

Elastic encounters, where there is no change in the internal energy levels during the collision, correspond to

$$C^i_j + C^h_k \to C^i_j + C^h_k, \quad 1 \le i, h \le 4 \quad \begin{array}{l} 1 \le j \le L^i \\ 1 \le k \le L^h. \end{array} \tag{3}$$

Inelastic encounters, where there is a change of internal energy levels of components, are described by

$$C^i_j + C^h_k \to C^i_m + C^h_p, \quad 1 \le i, h \le 4 \quad \begin{array}{l} 1 \le j, m \le L^i \\ 1 \le k, p \le L^h, \end{array} \tag{4}$$

and they can be endothermic if $E^i_m + E^h_p - E^i_j - E^h_k \ge 0$ or exothermic if $E^i_m + E^h_p - E^i_j - E^h_k < 0$. Chemical encounters, where there is also a change in the gas

species, are given by

$$C_j^i + C_k^h \to C_m^l + C_p^n, \quad \begin{array}{l}(i,h) \neq (l,n),\\ (i,h),(l,n) \in \{(1,2),(3,4)\},\end{array} \quad \begin{array}{l}1 \le j \le L^i\\ 1 \le k \le L^h\\ 1 \le m \le L^l\\ 1 \le p \le L^n,\end{array} \tag{5}$$

that are endothermic if $E_m^l + E_p^n - E_j^i - E_k^h \ge 0$ or exothermic if $E_m^l + E_p^n - E_j^i - E_k^h < 0$.

Denoting with $(\mathbf{v}, \mathbf{w})$ the molecular velocities of the ingoing particles and with $(\mathbf{v}', \mathbf{w}')$ the corresponding post-collision velocities, we have preservation of mass, global momentum and total energy:

$$\begin{aligned} &m^i + m^h = m^l + m^n\,,\\ &m^i\,\mathbf{v} + m^h\,\mathbf{w} = m^l\,\mathbf{v}' + m^n\,\mathbf{w}'\,,\\ &\frac{1}{2}\,m^i\,|\mathbf{v}|^2 + E_j^i + \frac{1}{2}\,m^h\,|\mathbf{w}|^2 + E_k^h = \frac{1}{2}\,m^l\,|\mathbf{v}'|^2 + E_m^l + \frac{1}{2}\,m^n\,|\mathbf{w}'|^2 + E_p^n\,. \end{aligned} \tag{6}$$

We take into account the major moments of each component C_j^i, that are number densities n_j^i, drift velocity $\mathbf{u}_j^i$ and kinetic temperature T_j^i. Clearly, the total density of each gas species, given by

$$n^i = \sum_{j=1}^{L^i} n_j^i, \quad i = 1, \dots, 4,$$

is not constant in time, but thanks to conservation of total mass we have that three suitable combinations of them, for instance $n^1 + n^3$, $n^1 + n^4$, $n^2 + n^4$, are conserved, as well as global momentum and total energy of the mixture.

Collision equilibria in gas mixtures are provided by Maxwellian distributions in which all species share the same mean velocity $\mathbf{u}$ and the same temperature T [10, 11]. In particular, for a mixture of polyatomic gases with discrete energies, denoting by $M^i(\mathbf{v}; \mathbf{u}, T/m^i)$ the Maxwellian

$$M^i\left(\mathbf{v}; \mathbf{u}, \frac{T}{m^i}\right) = \left(\frac{m^i}{2\,\pi\,T}\right)^{3/2} \exp\left(-\frac{m^i}{2\,T}\,|\mathbf{v} - \mathbf{u}|^2\right), \tag{7}$$

the equilibrium state for gas components reads as

$$f_{jM}^i(\mathbf{v}) = n_j^i\,M^i\left(\mathbf{v}; \mathbf{u}, \frac{T}{m^i}\right), \qquad i = 1, \dots, 4, \quad j = 1, \dots, L^i\,, \tag{8}$$

where, as proven in [14], number densities of single components n_j^i are related to the total number density n^i of the gas G^i by the following relation depending on the internal energy levels:

$$n_j^i = n^i \frac{\exp\left(-\frac{E_j^i - E_1^i}{T}\right)}{\sum_{k=1}^{L^i} \exp\left(-\frac{E_k^i - E_1^i}{T}\right)} = n^i \frac{\exp\left(-\frac{E_j^i - E_1^i}{T}\right)}{Z^i(T)}. \tag{9}$$

In addition, in the present reactive frame, number densities of the four interacting gases must fulfill at equilibrium the mass action law of chemistry

$$\frac{n^1 n^2}{n^3 n^4} = \left(\frac{m^1 m^2}{m^3 m^4}\right)^{\frac{3}{2}} \frac{Z^1(T) Z^2(T)}{Z^3(T) Z^4(T)} \exp\left(\frac{\Delta E}{T}\right), \tag{10}$$

with $\Delta E = E_1^3 + E_1^4 - E_1^2 - E_1^1$.

2.1 BGK Model

We propose a BGK model analogous to [8] by writing a kinetic equation for each component's distribution function f_j^i ($i = 1, \ldots, 4$, $j = 1, \ldots, L^i$) with a collision operator constituted by a unique relaxation term. In this way, we get a set of $L^1 + \ldots + L^4$ BGK equations

$$\frac{\partial f_j^i}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f_j^i = \nu_j^i (\mathcal{M}_j^i - f_j^i), \qquad i = 1, \ldots, 4, \qquad j = 1, \ldots, L^i, \tag{11}$$

where ν_j^i are macroscopic collision frequencies (independent of molecular velocity $\mathbf{v}$, but possibly dependent on macroscopic fields). The distributions $\mathcal{M}_j^i$ are Maxwellian attractors:

$$\mathcal{M}_j^i(\mathbf{v}) = \tilde{n}_j^i \left(\frac{m^i}{2\pi \tilde{T}}\right)^{3/2} \exp\left[-\frac{m^i}{2\tilde{T}} |\mathbf{v} - \tilde{\mathbf{u}}|^2\right], \quad \begin{matrix} i = 1, \ldots, 4, \\ j = 1, \ldots, L^i, \end{matrix} \tag{12}$$

depending on auxiliary parameters $\tilde{n}_j^i$ ($i = 1, \ldots, 4$, $j = 1, \ldots, L^i$), $\tilde{\mathbf{u}}$, $\tilde{T}$, to be suitably determined in terms of the actual macroscopic fields.

For any gas species G^i, $i = 1, \dots, 4$, fictitious densities $\tilde{n}^i_j$ are taken bound together as

$$\tilde{n}^i_j = \tilde{n}^i \frac{\exp\left(-\frac{E^i_j - E^i_1}{\tilde{T}}\right)}{Z^i(\tilde{T})}, \tag{13}$$

and in addition fictitious total densities $\tilde{n}^i$ satisfy the constraint

$$\frac{\tilde{n}^1\tilde{n}^2}{\tilde{n}^3\tilde{n}^4} = \left(\frac{m^1 m^2}{m^3 m^4}\right)^{\frac{3}{2}} \frac{Z^1(\tilde{T})Z^2(\tilde{T})}{Z^3(\tilde{T})Z^4(\tilde{T})} \exp\left(\frac{\Delta E}{\tilde{T}}\right). \tag{14}$$

In this way, collision equilibria of the BGK model (11) are correctly provided by Maxwellian distributions sharing a common velocity and a common temperature, with number densities related to the total density of the gas by (9) and total densities bounded together by (10). Our aim is to find auxiliary parameters in terms of the actual ones imposing the preservation of the same (seven) collision invariants of the Boltzmann equations in the BGK model. These correspond to three suitable combinations of gas densities, for instance $n^1 + n^3$, $n^1 + n^4$, $n^2 + n^4$,

$$\sum_{j=1}^{L^1} \nu^1_j \int_{\mathbb{R}^3} (\mathcal{M}^1_j - f^1_j) d\mathbf{v} + \sum_{j=1}^{L^3} \nu^3_j \int_{\mathbb{R}^3} (\mathcal{M}^3_j - f^3_j) d\mathbf{v} = 0 \tag{15}$$

$$\sum_{j=1}^{L^1} \nu^1_j \int_{\mathbb{R}^3} (\mathcal{M}^1_j - f^1_j) d\mathbf{v} + \sum_{j=1}^{L^4} \nu^4_j \int_{\mathbb{R}^3} (\mathcal{M}^4_j - f^4_j) d\mathbf{v} = 0 \tag{16}$$

$$\sum_{j=1}^{L^2} \nu^2_j \int_{\mathbb{R}^3} (\mathcal{M}^2_j - f^2_j) d\mathbf{v} + \sum_{j=1}^{L^4} \nu^4_j \int_{\mathbb{R}^3} (\mathcal{M}^4_j - f^4_j) d\mathbf{v} = 0, \tag{17}$$

global momentum

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu^i_j m^i \int_{\mathbb{R}^3} \mathbf{v} (\mathcal{M}^i_j - f^i_j) d\mathbf{v} = \mathbf{0}, \tag{18}$$

and total energy

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu^i_j \int_{\mathbb{R}^3} \left(\frac{1}{2} m^i |\mathbf{v}|^2 + E^i_j\right) (\mathcal{M}^i_j - f^i_j) d\mathbf{v} = 0. \tag{19}$$

Relations (15)-(17) lead to

$$\sum_{j=1}^{L^i} \nu_j^i (\tilde{n}_j^i - n_j^i) = \lambda^i \sum_{j=1}^{L^1} \nu_j^1 (\tilde{n}_j^1 - n_j^1), \qquad i = 1, \ldots, 4, \tag{20}$$

with $\lambda^1 = \lambda^2 = -\lambda^3 = -\lambda^4 = 1$. A linear combination of previous equations, together with conservation of mass, gives as results

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \tilde{n}_j^i = \sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i n_j^i \tag{21}$$

and

$$\sum_{i=1}^{4} m^i \sum_{j=1}^{L^i} \nu_j^i \tilde{n}_j^i = \sum_{i=1}^{4} m^i \sum_{j=1}^{L^i} \nu_j^i n_j^i. \tag{22}$$

Expression (20) can be written as

$$\sum_{j=1}^{L^i} \nu_j^i \tilde{n}_j^i = \sum_{j=1}^{L^i} \nu_j^i n_j^i + \lambda^i \sum_{j=1}^{L^1} \nu_j^1 (\tilde{n}_j^1 - n_j^1), \qquad i = 1, \ldots, 4, \tag{23}$$

and, from relation (13), we have

$$\begin{aligned} &\sum_{j=1}^{L^i} \nu_j^i \frac{\tilde{n}^i}{Z^i(\tilde{T})} \left[\exp\left(- \frac{E_j^i - E_1^i}{\tilde{T}} \right) \right] \\ &= \sum_{j=1}^{L^i} \nu_j^i n_j^i + \lambda^i \sum_{j=1}^{L^1} \nu_j^1 \left(\frac{\tilde{n}^1}{Z^1(\tilde{T})} \left[\exp\left(- \frac{E_j^1 - E_1^1}{\tilde{T}} \right) \right] - n_j^1 \right), \qquad i = 1, \ldots, 4. \end{aligned} \tag{24}$$

This allows us to write three of the auxiliary total densities ($\tilde{n}^2$, $\tilde{n}^3$, $\tilde{n}^4$) as function of the remaining one ($\tilde{n}^1$):

$$\begin{aligned} \frac{\tilde{n}^i}{Z^i(\tilde{T})} = &\left[\sum_{j=1}^{L^i} \nu_j^i \exp\left(- \frac{E_j^i - E_1^i}{\tilde{T}} \right) \right]^{-1} \left\{ \sum_{j=1}^{L^i} \nu_j^i n_j^i - \lambda^i \sum_{j=1}^{L^1} \nu_j^1 n_j^1 \right. \\ &\left. + \lambda^i \left[\sum_{j=1}^{L^1} \nu_j^1 \exp\left(- \frac{E_j^1 - E_1^1}{\tilde{T}} \right) \right] \frac{\tilde{n}^1}{Z^1(\tilde{T})} \right\}, \end{aligned} \tag{25}$$

that holds for $i = 1, \ldots, 4$, since for $i = 1$ we get a trivial identity.

From momentum conservation (18) we get the equation involving auxiliary mean velocity

$$\sum_{i=1}^{4}\left(\sum_{h=1}^{L^i}\nu_h^i\, m^i\, \tilde{n}_h^i \tilde{\mathbf{u}} - \sum_{j=1}^{L^i}\nu_j^i\, m^i\, n_j^i \mathbf{u}_j^i\right) = \mathbf{0}\,, \tag{26}$$

that owing to (22) provides

$$\tilde{\mathbf{u}} = \frac{\displaystyle\sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, m^i\, n_j^i\, \mathbf{u}_j^i}{\displaystyle\sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, m^i\, n_j^i}, \tag{27}$$

hence $\tilde{\mathbf{u}}$ is an explicit combination of actual number densities and mean velocities of single gas components.

Total energy conservation (19) gives the equation

$$\frac{3}{2}\sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, n_j^i \tilde{T} + \sum_{i=1}^{4}\sum_{j=1}^{L^i}\tilde{n}_j^i \nu_j^i\, E_j^i = \Lambda \tag{28}$$

with Λ being a term explicitly depending on actual densities, velocities and energies

$$\Lambda = \frac{1}{2}\left(\sum_{i=1}^{4} m^i \sum_{j=1}^{L^i}\nu_j^i\, n_j^i\, \left(|\mathbf{u}_j^i|^2 - |\tilde{\mathbf{u}}|^2\right)\right) + \frac{3}{2}\left(\sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, n_j^i\, T_j^i\right) + \sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, E_j^i\, n_j^i\,. \tag{29}$$

By applying expression (13) to the left-hand side of (28) we get an equation of the form

$$\frac{3}{2}\sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, n_j^i \tilde{T} + \sum_{i=1}^{4}\frac{\tilde{n}^i}{Z^i(\tilde{T})}\sum_{j=1}^{L^i}\nu_j^i\, E_j^i \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right) = \Lambda. \tag{30}$$

At this point, we face the main difference with respect to the model without chemical reaction. Instead of a transcendental equation for $\tilde{T}$ having one positive solution and depending only on the actual parameters of the mixture, we have Eq. (30) containing both auxiliary parameters $\tilde{T}$ and $\tilde{n}^1$. Following the procedure applied in [4], it is possible to show that those two parameters are uniquely

determined, bearing in mind also the fictitious mass action law (14). At first, we find it convenient setting

$$Y^i = \frac{\tilde{n}^i}{Z^i(\tilde{T})} \sum_{j=1}^{L^i} \nu_j^i \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right), \tag{31}$$

and Eq. (30) thus becomes

$$\frac{3}{2} \sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \, n_j^i \tilde{T} + \sum_{i=1}^{4} Y^i \frac{\displaystyle\sum_{j=1}^{L^i} \nu_j^i \, E_j^i \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right)}{\displaystyle\sum_{k=1}^{L^i} \nu_k^i \exp\left(-\frac{E_k^i - E_1^i}{\tilde{T}}\right)} = \Lambda. \tag{32}$$

Since from (25) we have

$$Y^i = \sum_{j=1}^{L^i} \nu_j^i n_j^i - \lambda^i \sum_{j=1}^{L^1} \nu_j^1 n_j^1 + \lambda^i Y^1 \qquad i = 1, \ldots, 4, \tag{33}$$

we end up with an equation of the form

$$Y^1 = \sum_{j=1}^{L^1} \nu_j^1 n_j^1 + \mathcal{S}(\tilde{T}), \tag{34}$$

where $\mathcal{S}(\tilde{T})$ is written as

$$\mathcal{S}(\tilde{T}) = \frac{\mathcal{N}(\tilde{T})}{\mathcal{D}(\tilde{T})}, \tag{35}$$

with the numerator $\mathcal{N}$

$$\mathcal{N}(\tilde{T}) = \Lambda - \sum_{i=1}^{4} \left(\sum_{m=1}^{L^i} \nu_m^i \, n_m^i \right) \left[\frac{3}{2} \tilde{T} + \frac{\displaystyle\sum_{j=1}^{L^i} \nu_j^i \, E_j^i \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right)}{\displaystyle\sum_{k=1}^{L^i} \nu_k^i \exp\left(-\frac{E_k^i - E_1^i}{\tilde{T}}\right)} \right] \tag{36}$$

and the denominator $\mathcal{D}$

$$\mathcal{D}(\tilde{T}) = \sum_{i=1}^{4} \lambda^i \, \frac{\displaystyle\sum_{j=1}^{L^i} \nu_j^i \, E_j^i \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right)}{\displaystyle\sum_{k=1}^{L^i} \nu_k^i \exp\left(-\frac{E_k^i - E_1^i}{\tilde{T}}\right)}. \tag{37}$$

We observe that, if we had repeated the previous calculations choosing a different gas species to express the other three ones, we would have obtained

$$Y^i = \sum_{j=1}^{L^i} \nu_j^i n_j^i + \lambda^i \mathcal{S}(\tilde{T}) \qquad i = 1, \ldots, 4. \tag{38}$$

Putting these expressions in the constraint (14), we obtain a transcendental equation depending on $\tilde{T}$

$$\mathcal{G}(\tilde{T}) = \left(\frac{m^1 m^2}{m^3 m^4}\right)^{\frac{3}{2}}, \tag{39}$$

with

$$\mathcal{G}(\tilde{T}) = \mathcal{G}_1(\tilde{T}) \cdot \mathcal{G}_2(\tilde{T}) \cdot \mathcal{G}_3(\tilde{T}), \tag{40}$$

being

$$\mathcal{G}_1(\tilde{T}) = \frac{\left[\displaystyle\sum_{j=1}^{L^1} \nu_j^1 n_j^1 + \mathcal{S}(\tilde{T})\right] \left[\displaystyle\sum_{j=1}^{L^2} \nu_j^2 n_j^2 + \mathcal{S}(\tilde{T})\right]}{\left[\displaystyle\sum_{j=1}^{L^3} \nu_j^3 n_j^3 - \mathcal{S}(\tilde{T})\right] \left[\displaystyle\sum_{j=1}^{L^4} \nu_j^4 n_j^4 - \mathcal{S}(\tilde{T})\right]}, \tag{41}$$

$$\mathcal{G}_2(\tilde{T}) = \frac{\displaystyle\sum_{k=1}^{L^3} \nu_k^3 \exp\left(-\frac{E_k^3 - E_1^3}{\tilde{T}}\right) \sum_{k=1}^{L^4} \nu_k^4 \exp\left(-\frac{E_k^4 - E_1^4}{\tilde{T}}\right)}{\displaystyle\sum_{k=1}^{L^1} \nu_k^1 \exp\left(-\frac{E_k^1 - E_1^1}{\tilde{T}}\right) \sum_{k=1}^{L^2} \nu_k^2 \exp\left(-\frac{E_k^2 - E_1^2}{\tilde{T}}\right)}, \tag{42}$$

$$\mathcal{G}_3(\tilde{T}) = \exp\left(-\frac{\Delta E}{\tilde{T}}\right). \tag{43}$$

Our aim is now to show that Eq. (39) admits one positive solution in the range in which all the densities $\tilde{n}^i$ are positive. More precisely, referring to the quantities Y^i, we are looking for a solution in the set

$$A = \left\{ \tilde{T} > 0 \,:\, \max\left(-\sum_{j=1}^{L^1} \nu_j^1 n_j^1, -\sum_{j=1}^{L^2} \nu_j^2 n_j^2 \right) < \mathcal{S}(\tilde{T}) < \min\left(\sum_{j=1}^{L^3} \nu_j^3 n_j^3, \sum_{j=1}^{L^4} \nu_j^4 n_j^4 \right) \right\}. \tag{44}$$

We will go through the same proof performed in [4], adjusting it to the present frame of polyatomic gases with a different number of internal energy levels. The first result that we point out is the following.

Lemma 1 *Let $I = (\tilde{T}_1, \tilde{T}_2) \subseteq A$ be any interval in which the function $\mathcal{D}(\tilde{T})$ given in (37) is strictly negative (positive), then the function $\mathcal{S}(\tilde{T})$ given in (35) is strictly monotonically increasing (decreasing) in I.*

Proof From the expression of $\mathcal{S}(\tilde{T})$ we easily get

$$\mathcal{S}'(\tilde{T}) = \frac{\mathcal{N}'(\tilde{T})}{\mathcal{D}(\tilde{T})} - \mathcal{S}(\tilde{T}) \frac{\mathcal{D}'(\tilde{T})}{\mathcal{D}(\tilde{T})}. \tag{45}$$

Then we have that

$$\mathcal{D}'(\tilde{T}) = \sum_{i=1}^{4} \lambda^i \frac{\displaystyle\sum_{j=1}^{L^i} \sum_{k=1}^{L^i} \frac{\nu_j^i \nu_k^i}{\tilde{T}^2} \left[\left(E_j^i\right)^2 - E_j^i E_k^i \right] \exp\left(- \frac{E_j^i + E_k^i - 2E_1^i}{\tilde{T}} \right)}{\left[\displaystyle\sum_{k=1}^{L^i} \nu_k^i \exp\left(- \frac{E_k^i - E_1^i}{\tilde{T}} \right) \right]^2}; \tag{46}$$

performing the exchange of indices $j \leftrightarrow k$, Eq. (46) can be written as

$$\mathcal{D}'(\tilde{T}) = \sum_{i=1}^{4} \lambda^i \mathcal{F}^i(\tilde{T}) \tag{47}$$

with

$$\mathcal{F}^i(\tilde{T}) := \frac{\displaystyle\sum_{j=1}^{L^i} \sum_{k=1}^{L^i} \frac{\nu_j^i \nu_k^i}{2\tilde{T}^2} \left[E_j^i - E_k^i \right]^2 \exp\left(- \frac{E_j^i + E_k^i - 2E_1^i}{\tilde{T}} \right)}{\left[\displaystyle\sum_{k=1}^{L^i} \nu_k^i \exp\left(- \frac{E_k^i - E_1^i}{\tilde{T}} \right) \right]^2} \geq 0 \quad i = 1, \ldots, 4. \tag{48}$$

Analogously we get

$$\mathcal{N}'(\tilde{T}) = -\sum_{i=1}^{4}\left(\sum_{j=1}^{L^i} \nu_j^i\, n_j^i\right)\left[\frac{3}{2} + \mathcal{F}^i(\tilde{T})\right] < 0\,. \tag{49}$$

Eventually, from (45), the expression for $\mathcal{S}'(\tilde{T})$ is

$$\mathcal{S}'(\tilde{T}) = -\frac{1}{\mathcal{D}(\tilde{T})}\left\{\sum_{i=1}^{4}\frac{3}{2}\left(\sum_{j=1}^{L^i}\nu_j^i\, n_j^i\right) + \sum_{i=1}^{4}\left[\sum_{j=1}^{L^i}\nu_j^i\, n_j^i + \lambda^i\,\mathcal{S}(\tilde{T})\right]\mathcal{F}^i(\tilde{T})\right\}. \tag{50}$$

We notice that the content of the square brackets in (50) is strictly positive for $\tilde{T} \in A$ and so the content of the whole curly brackets is positive too, this means that $\mathcal{S}'(\tilde{T})$ and $\mathcal{D}(\tilde{T})$ have opposite sign. □

We focus now on the behavior of the function $\mathcal{S}(\tilde{T})$. For the numerator $\mathcal{N}(\tilde{T})$ we have

Lemma 2 *The function $\mathcal{N}(\tilde{T})$ has a unique positive root $\tilde{T}^*$.*

Proof First of all we recall that $\mathcal{N}'(\tilde{T}) < 0$. Moreover we have

$$\begin{aligned}
\lim_{\tilde{T}\to 0^+}\mathcal{N}(\tilde{T}) &= \Lambda - \lim_{\tilde{T}\to 0^+}\sum_{i=1}^{4}\left(\sum_{m=1}^{L^i}\nu_m^i\, n_m^i\right)\left[\frac{3}{2}\,\tilde{T} + \frac{E_1^i\nu_1^i + \displaystyle\sum_{j=2}^{L^i}\nu_j^i\, E_j^i \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right)}{\nu_1^i + \displaystyle\sum_{k=2}^{L^i}\nu_k^i \exp\left(-\frac{E_k^i - E_1^i}{\tilde{T}}\right)}\right] \\
&= \Lambda - \sum_{i=1}^{4}\left(\sum_{m=1}^{L^i}\nu_m^i\, n_m^i\right) E_1^i \\
&= \frac{1}{2}\left(\sum_{i=1}^{4} m^i \sum_{j=1}^{L^i}\nu_j^i\, n_j^i\,\left(|\mathbf{u}_j^i|^2 - |\tilde{\mathbf{u}}|^2\right)\right) + \frac{3}{2}\left(\sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, n_j^i\, T_j^i\right) \\
&+ \sum_{i=1}^{4}\sum_{j=1}^{L^i}\nu_j^i\, n_j^i\,\left(E_j^i - E_1^i\right).
\end{aligned} \tag{51}$$

As it is shown in [8], the sum involving mean velocities in (51) is non-negative, so the whole limit is strictly positive. In addition it holds

$$\lim_{\tilde{T}\to+\infty}\mathcal{N}(\tilde{T}) = -\infty. \tag{52}$$

Thus, $\mathcal{N}(\tilde{T})$ has a unique positive root $\tilde{T}^*$. □

Following the same argument used in [4], we shall omit the situation in which the choice of initial data, internal energies and collision frequencies is such that also $\mathcal{D}(\tilde{T}^*) = 0$. In this case, $\tilde{T} = \tilde{T}^*$ simplifies the function $\mathcal{S}(\tilde{T})$ and we have to deal with a simple algebraic equation for our unknown Y[1].

Now we make some considerations about $\mathcal{D}(\tilde{T})$.

Lemma 3 *On every interval* $\left(\tilde{T}_1, \tilde{T}_2\right) \subseteq A$ *the sign of* $\mathcal{D}(\tilde{T})$ *does not change.*

Proof Since

$$\lim_{\tilde{T}\to 0^+} \mathcal{D}(\tilde{T}) = -\Delta E < 0, \qquad \lim_{\tilde{T}\to +\infty} \mathcal{D}(\tilde{T}) = \sum_{i=1}^{4} \lambda^i \frac{\sum_{j=1}^{L^i} \nu_j^i E_j^i}{\sum_{j=1}^{L^i} \nu_j^i} \tag{53}$$

and the sign of $\mathcal{D}'(\tilde{T})$ given in (47) changes in relation to internal energy levels and collision frequencies, $\mathcal{D}(\tilde{T})$ may have a positive root, call it $\tilde{T}^{\#}$, with $\tilde{T}^{\#} \neq \tilde{T}^*$. But in this case we would have

$$\lim_{\tilde{T}\to \tilde{T}^{\#}} \mathcal{S}(\tilde{T}) = \pm\infty \tag{54}$$

getting a neighborhood of $\tilde{T}^{\#}$ not contained in A. Thus we can conclude that on every interval $\left(\tilde{T}_1, \tilde{T}_2\right) \subseteq A$ the sign of $\mathcal{D}(\tilde{T})$ does not change. □

Consequently, from Lemma 1, neither the sign of $\mathcal{S}'(\tilde{T})$ changes. This allows us to prove the following result.

Lemma 4 *The set A given in* (44) *is a connected set of* $\mathbb{R}^+$.

Proof Let $\left(\tilde{T}_a, \tilde{T}_b\right)$ be a connected component of A. If $\tilde{T}_a \neq 0$ the function $\mathcal{S}(\tilde{T})$ is continuous, strictly monotonically increasing or decreasing on it, hence it assumes all the values between its upper bound that is $\min\left(\sum_{j=1}^{L^3} \nu_j^3 n_j^3, \sum_{j=1}^{L^4} \nu_j^4 n_j^4\right)$ and its lower bound that is $\max\left(-\sum_{j=1}^{L^1} \nu_j^1 n_j^1, -\sum_{j=1}^{L^2} \nu_j^2 n_j^2\right)$. If $\tilde{T}_a = 0$, since $\lim_{\tilde{T}\to 0^+} \mathcal{D}(\tilde{T}) < 0$, $S(\tilde{T})$ has to be strictly monotonically increasing on $\left(\tilde{T}_a, \tilde{T}_b\right)$, going from $\lim_{\tilde{T}\to 0^+} \mathcal{S}(\tilde{T}) < 0$ to $\min\left(\sum_{j=1}^{L^3} \nu_j^3 n_j^3, \sum_{j=1}^{L^4} \nu_j^4 n_j^4\right)$. Thus $\mathcal{S}(\tilde{T})$ has a root in $\left(\tilde{T}_a, \tilde{T}_b\right)$, but we know that $\mathcal{S}(\tilde{T})$ has only one positive root, $\tilde{T}^*$. It follows that $\left(\tilde{T}_a, \tilde{T}_b\right)$ is the only connected component of A, i.e. A is a connected set. □

Now we are able to give the final result.

Lemma 5 *The function $\mathcal{G}(\tilde{T})$ defined in (40) is strictly monotone in the set A, ranging from 0 to $+\infty$. More precisely, it is increasing if $\mathcal{D}(\tilde{T}) < 0$ and decreasing if $\mathcal{D}(\tilde{T}) > 0$.*

Proof We compute the derivative of the function $\mathcal{G}(\tilde{T})$. We have

$$\begin{aligned}\mathcal{G}_1'(\tilde{T}) &= \mathcal{S}'(\tilde{T}) \left(\sum_{j=1}^{L^3} v_j^3 n_j^3 - \mathcal{S}(\tilde{T})\right)^{-2} \left(\sum_{j=1}^{L^4} v_j^4 n_j^4 - \mathcal{S}(\tilde{T})\right)^{-2} \\ &\times \left[\left(\sum_{j=1}^{L^1} v_j^1 n_j^1 + \mathcal{S}(\tilde{T})\right) + \left(\sum_{j=1}^{L^2} v_j^2 n_j^2 + \mathcal{S}(\tilde{T})\right)\right] \left(\sum_{j=1}^{L^3} v_j^3 n_j^3 - \mathcal{S}(\tilde{T})\right) \left(\sum_{j=1}^{L^4} v_j^4 n_j^4 - \mathcal{S}(\tilde{T})\right) \\ &\times \left[\left(\sum_{j=1}^{L^3} v_j^3 n_j^3 - \mathcal{S}(\tilde{T})\right) + \left(\sum_{j=1}^{L^4} v_j^4 n_j^4 - \mathcal{S}(\tilde{T})\right)\right] \left(\sum_{j=1}^{L^1} v_j^1 n_j^1 + \mathcal{S}(\tilde{T})\right) \left(\sum_{j=1}^{L^2} v_j^2 n_j^2 + \mathcal{S}(\tilde{T})\right)\end{aligned} \tag{55}$$

that can be cast as

$$\mathcal{G}_1'(\tilde{T}) = \mathcal{G}_1(\tilde{T}) \mathcal{S}'(\tilde{T}) \sum_{i=1}^{4} \frac{1}{\sum_{j=1}^{L^i} v_j^i n_j^i + \lambda^i \mathcal{S}(\tilde{T})}. \tag{56}$$

Proceeding in a similar way we have

$$\mathcal{G}_2'(\tilde{T}) = -\mathcal{G}_2(\tilde{T}) \frac{1}{\tilde{T}^2} \sum_{i=1}^{4} \lambda^i \frac{\sum_{j=1}^{L^i} v_j^i \left(E_j^i - E_1^i\right) \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right)}{\sum_{k=1}^{L^i} v_k^i \exp\left(-\frac{E_k^i - E_1^i}{\tilde{T}}\right)}, \tag{57}$$

and, finally,

$$\mathcal{G}_3'(\tilde{T}) = \mathcal{G}_3(\tilde{T}) \frac{\Delta E}{\tilde{T}^2}. \tag{58}$$

In this way we can conclude that

$$\mathcal{G}'(\tilde{T}) = \mathcal{G}(\tilde{T}) \left\{ \mathcal{S}'(\tilde{T}) \sum_{i=1}^{4} \frac{1}{\sum_{j=1}^{L^i} \nu_j^i n_j^i + \lambda^i \mathcal{S}(\tilde{T})} - \frac{1}{\tilde{T}^2} \sum_{i=1}^{4} \lambda^i \frac{\sum_{j=1}^{L^i} \nu_j^i \left(E_j^i - E_1^i\right) \exp\left(-\frac{E_j^i - E_1^i}{\tilde{T}}\right)}{\sum_{k=1}^{L^i} \nu_k^i \exp\left(-\frac{E_k^i - E_1^i}{\tilde{T}}\right)} + \frac{\Delta E}{\tilde{T}^2} \right\}, \tag{59}$$

that can be written, using function $\mathcal{D}(\tilde{T})$ defined in (37), as

$$\mathcal{G}'(\tilde{T}) = \mathcal{G}(\tilde{T}) \left\{ \mathcal{S}'(\tilde{T}) \left[\sum_{i=1}^{4} \frac{1}{\sum_{j=1}^{L^i} \nu_j^i n_j^i + \lambda^i \mathcal{S}(\tilde{T})} \right] - \frac{1}{\tilde{T}^2} \mathcal{D}(\tilde{T}) \right\}. \tag{60}$$

We have that both $\mathcal{G}(\tilde{T})$ and the term in square brackets in (60) are positive in the set A that, as we proved in Lemma 4, is an interval $A = \left(\tilde{T}_{min}, \tilde{T}_{max}\right)$; moreover, thanks to Lemma 1, $\mathcal{S}'(\tilde{T})$ and $-\mathcal{D}(\tilde{T})$ have the same sign, that does not change in A, and this means that $\mathcal{G}(\tilde{T})$ is strictly monotone. Moreover, recalling the definition of $\mathcal{G}(\tilde{T})$ given in (40), when $\mathcal{S}(\tilde{T}) \to \min\left(\sum_{j=1}^{L^3} \nu_j^3 n_j^3, \sum_{j=1}^{L^4} \nu_j^4 n_j^4\right)$ we have $\mathcal{G}(\tilde{T}) \to +\infty$ and when $\mathcal{S}(\tilde{T}) \to \max\left(-\sum_{j=1}^{L^1} \nu_j^1 n_j^1, -\sum_{j=1}^{L^2} \nu_j^2 n_j^2\right)$ we have $\mathcal{G}(\tilde{T}) \to 0$; also in the case in which $\tilde{T}_{min} = 0$ it holds $\lim_{\tilde{T} \to 0^+} \mathcal{G}(\tilde{T}) = 0$. □

This final result allows us to assert that Eq. (39) has a unique solution, providing thus the auxiliary temperature $\tilde{T}$, and this completes the construction of Maxwellian attractors $\mathcal{M}_j^i$ of our BGK model.

The equilibrium states correspond to $f_j^i = \mathcal{M}_j^i$ for $i = 1, \ldots, 4$ and $j = 1, \ldots, L^i$, therefore

$$\mathbf{u}_j^i = \tilde{\mathbf{u}} = \mathbf{u}\,, \qquad T_j^i = \tilde{T} = T, \qquad i = 1, \ldots, 4, \qquad j = 1, \ldots, L^i,$$

and number densities of components n_j^i are related to global density of the corresponding gas n^i by the constraint (9), while the densities n^i are bound together by (10).

2.2 *H-Theorem for the Homogeneous Case*

We can also prove the asymptotic stability of collision equilibria. Indeed, in space homogeneous conditions, setting $\underline{\mathbf{f}} = (f_1^1, \dots, f_{L^4}^4)$, the physical entropy

$$H[\underline{\mathbf{f}}] = \sum_{i=1}^{4} \sum_{j=1}^{L^i} \int_{\mathbb{R}^3} f_j^i \, \log(f_j^i) \, d\mathbf{v} \tag{61}$$

is a Lyapunov functional for the present BGK model. Specifically, if $\underline{\mathbf{f}}_M$ denotes the stationary state corresponding to the initial state $\underline{\mathbf{f}}_0$, we have $H[\underline{\mathbf{f}}] > H[\underline{\mathbf{f}}_M]$ for any $\underline{\mathbf{f}} \neq \underline{\mathbf{f}}_M$ (this is a classical result, already shown for instance in [14]), and we can prove the entropy inequality $H'[\underline{\mathbf{f}}] < 0$ for any $\underline{\mathbf{f}} \neq \underline{\mathbf{f}}_M$, while $H'[\underline{\mathbf{f}}_M] = 0$.

The derivative of the H-functional (61) reads as

$$H'[\underline{\mathbf{f}}] = \sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \int_{\mathbb{R}^3} \left(\mathcal{M}_j^i - f_j^i \right) \log(f_j^i) \, d\mathbf{v}. \tag{62}$$

At first we can check that

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \int_{\mathbb{R}^3} \left(\mathcal{M}_j^i - f_j^i \right) \log(\mathcal{M}_j^i) \, d\mathbf{v} = 0 \,. \tag{63}$$

Indeed, we explicitly compute the logarithm of Maxwellian attractors, leading to

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \int_{\mathbb{R}^3} \left(\mathcal{M}_j^i - f_j^i \right) \left[\log \tilde{n}_j^i + \frac{3}{2} \log(m^i) - \frac{3}{2} \log(2\pi \tilde{T}) \right] d\mathbf{v}$$

$$+ \sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \int_{\mathbb{R}^3} \left(\mathcal{M}_j^i - f_j^i \right) \left[-\frac{m^i}{2\tilde{T}} \left(|\mathbf{v}|^2 - 2\tilde{\mathbf{u}} \cdot \mathbf{v} + |\tilde{\mathbf{u}}|^2 \right) \right] d\mathbf{v} \,. \tag{64}$$

Then, owing to conservation laws of momentum and total energy, it simplifies to

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \left(\tilde{n}_j^i - n_j^i \right) \left[\log \tilde{n}_j^i + \frac{E_j^i}{\tilde{T}} + \frac{3}{2} \log m^i \right]; \tag{65}$$

bearing in mind (13), the previous equation becomes

$$\sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \left(\tilde{n}_j^i - n_j^i \right) \left[\log \tilde{n}^i + \frac{E_1^i}{\tilde{T}} - \log \left(Z^i(\tilde{T}) \right) + \frac{3}{2} \log m^i \right]. \tag{66}$$

Using the relations (20), the quantity above can be written as

$$\sum_{j=1}^{L^1} \nu_j^1 (\tilde{n}_j^1 - n_j^1) \sum_{i=1}^{4} \lambda^i \left[\log \tilde{n}^i + \frac{E_1^i}{\tilde{T}} - \log \left(Z^i(\tilde{T}) \right) + \frac{3}{2} \log m^i \right]$$

$$= \sum_{j=1}^{L^1} \nu_j^1 (\tilde{n}_j^1 - n_j^1) \left\{ \log \left[\frac{\tilde{n}^1 \tilde{n}^2}{\tilde{n}^3 \tilde{n}^4} \left(\frac{m^3 m^4}{m^1 m^2} \right)^{\frac{3}{2}} \right] - \log \left[\frac{Z^1(\tilde{T}) Z^2(\tilde{T})}{Z^3(\tilde{T}) Z^4(\tilde{T})} \exp \left(\frac{\Delta E}{\tilde{T}} \right) \right] \right\} = 0 \tag{67}$$

due to the mass action law (14) for auxiliary parameters. Then, by subtracting (63) from (62) we easily get that for any $\underline{\mathbf{f}} \neq \underline{\mathbf{f}}_M$

$$H'[\underline{\mathbf{f}}] = - \sum_{i=1}^{4} \sum_{j=1}^{L^i} \nu_j^i \int_{\mathbb{R}^3} \left(f_j^i - \mathcal{M}_j^i \right) \log \left(\frac{f_j^i}{\mathcal{M}_j^i} \right) d\mathbf{v}$$

and the inequality $H'[\underline{\mathbf{f}}] < 0$ holds owing to usual convexity arguments.

3 Trend to Equilibrium in Space Homogeneous Conditions

Performing the same calculations done in [8], we are able to derive from BGK model (11) the evolution equations for the main macroscopic fields, i.e. number densities, mean velocities and temperatures of all components of the four reacting gases (n_j^i, $\mathbf{u}_j^i$, T_j^i, for $i = 1, \ldots, 4$ and $j = 1, \ldots, L^i$). We obtain the following system

$$\begin{cases} \dfrac{\partial n_j^i}{\partial t} + \nabla_{\mathbf{x}} \cdot (n_j^i \, \mathbf{u}_j^i) = \nu_j^i (\tilde{n}_j^i - n_j^i) \,, \\ n_j^i \left(\dfrac{\partial \mathbf{u}_j^i}{\partial t} + \mathbf{u}_j^i \cdot \nabla_{\mathbf{x}} \mathbf{u}_j^i \right) + \dfrac{1}{m^i} \nabla_{\mathbf{x}} \cdot \mathbf{P}_j^i = \nu_j^i \, \tilde{n}_j^i (\tilde{\mathbf{u}} - \mathbf{u}_j^i) \,, \\ \dfrac{3}{2} n_j^i \left(\dfrac{\partial T_j^i}{\partial t} + \mathbf{u}_j^i \cdot \nabla_{\mathbf{x}} T_j^i \right) + \mathbf{P}_j^i : \nabla_{\mathbf{x}} \mathbf{u}_j^i + \nabla_{\mathbf{x}} \cdot \mathbf{q}_j^i \\ \qquad\qquad = \nu_j^i \, \tilde{n}_j^i \left[\dfrac{3}{2} (\tilde{T} - T_j^i) + \dfrac{1}{2} m^i \, |\tilde{\mathbf{u}} - \mathbf{u}_j^i|^2 \right], \end{cases} \tag{68}$$

where $\mathbf{P}^i_j$ are pressure tensors and $\mathbf{q}^i_j$ heat fluxes for each component defined as

$$\mathbf{P}^i_j = m^i \int_{\mathbb{R}^3} (\mathbf{v}-\mathbf{u}^i_j)\otimes(\mathbf{v}-\mathbf{u}^i_j)\, f^i_j(\mathbf{v})\, d\mathbf{v}\,, \quad \mathbf{q}^i_j = \frac{m^i}{2} \int_{\mathbb{R}^3} (\mathbf{v}-\mathbf{u}^i_j)\, |\mathbf{v}-\mathbf{u}^i_j|^2\, f^i_j(\mathbf{v})\, d\mathbf{v}\,.$$

For illustrative purposes we will show some numerical results for two reacting mixtures taking into account the space homogeneous and one-dimensional version of evolution equations (68) that reads as

$$\begin{cases} \dfrac{\partial n^i_j}{\partial t} = \nu^i_j(\tilde{n}^i_j - n^i_j)\,, & \begin{matrix} i = 1,\dots,4, \\ j = 1,\dots,L^i, \end{matrix} \\ \dfrac{\partial u^i_j}{\partial t} = \nu^i_j \dfrac{\tilde{n}^i_j}{n^i_j}(\tilde{u} - u^i_j)\,, & \begin{matrix} i = 1,\dots,4, \\ j = 1,\dots,L^i, \end{matrix} \\ \dfrac{\partial T^i_j}{\partial t} = \nu^i_j \dfrac{\tilde{n}^i_j}{n^i_j}\left(\tilde{T} - T^i_j + \dfrac{1}{3} m^i\,(\tilde{u} - u^i_j)^2\right), & \begin{matrix} i = 1,\dots,4, \\ j = 1,\dots,L^i, \end{matrix} \end{cases} \tag{69}$$

where $\tilde{n}^i_j$ is provided by (13) and (25), $\tilde{u}$ is explicitly given in (27), and $\tilde{T}$ may be obtained as the unique solution of the transcendental equation (39). Equations (69) constitute thus a closed system of $3(L^1 + \ldots + L^4)$ equations, having as unknowns densities, velocities and temperatures of all components of polyatomic species. Once the initial state $(n^i_j)_0$, $(u^i_j)_0$, $(T^i_j)_0$ is assigned, the corresponding equilibrium configuration is unique and may be determined bearing in mind the conservations of three suitable combinations of total densities, global velocity and total energy.

The first mixture we model is inspired by the reversible reaction involving hydrogen H_2 (with mass 2.02 g/mol), iodine I_2 (253.8 g/mol) and hydrogen iodide HI (127.91 g/mol)

$$H_2 + I_2 \leftrightarrows HI + HI.$$

So we take into account four gases G^i, $i = 1, ..., 4$, with mass ratios reproducing the ones of the gases involved in the reaction: $m^1 = 0.1$, $m^2 = 12.8$, $m^3 = m^4 = 6.45$. Notice that in this bimolecular reaction the third and the fourth species coincide, therefore they are characterized by the same internal structure and by the same initial data. We suppose that gas species G^1, $G^3 \equiv G^4$ are endowed with two and G^2 is endowed with three discrete energy levels, respectively. Specifically, we assume the following configuration of internal energy levels

$$E^1_1 = 6.5, \quad E^1_2 = 7.5, \quad E^2_1 = 7, \quad E^2_2 = 8, \quad E^2_3 = 8.5,$$

$$E^3_1 = 6, \quad E^3_2 = 7, \quad E^4_1 = 6, \quad E^4_2 = 7. \tag{70}$$

From now on we will consider the components concentrations

$$c_j^i = \frac{n_j^i}{\sum_{h=1}^{4} \sum_{k=1}^{L^h} n_k^h},$$

and also velocities and temperatures will be suitably normalized with respect to the corresponding equilibrium values. Initial data for number concentrations, velocities and temperatures are given as follows

	C_1^1	C_2^1	C_1^2	C_2^2	C_3^2	C_1^3	C_2^3	C_1^4	C_2^4
c_0	0.13	0.07	0.08	0.06	0.15	0.14	0.11	0.14	0.11
u_0	0.3	0	0.1	0.4	0.2	0.6	0.1	0.6	0.1
T_0	2	4	1	2.5	2	6	1.5	6	1.5

(71)

The choice of collision frequencies is done as in [4], setting the sets of indices for $i = 1, \ldots, 4$ and $j = 1, \ldots, L^i$

$$\mathcal{D}^1{}_j^i = \left\{ \begin{array}{l} h = 1, \ldots, 4, \\ m = 1, \ldots, L^i, \\ k, p = 1, \ldots, L^h \end{array} : E_m^i + E_p^h - E_j^i - E_k^h \leq 0 \right\},$$

$$\mathcal{D}^2{}_j^i = \left\{ \begin{array}{l} h = 1, \ldots, 4, \\ m = 1, \ldots, L^i, \\ k, p = 1, \ldots, L^h \end{array} : E_m^i + E_p^h - E_j^i - E_k^h > 0 \right\},$$

$$\mathcal{D}^3{}_j^i = \left\{ h, l, n : \begin{array}{cl} (i,h) \neq (l,n), & k = 1, \ldots, L^h, \\ (i,h), (l,n) & m = 1, \ldots, L^l, \\ \in \{(1,2),(3,4)\}, & p = 1, \ldots, L^n \end{array} : E_m^l + E_p^n - E_j^i - E_k^h \leq 0 \right\},$$

$$\mathcal{D}^4{}_j^i = \left\{ h, l, n : \begin{array}{cl} (i,h) \neq (l,n), & k = 1, \ldots, L^h, \\ (i,h), (l,n) & m = 1, \ldots, L^l, \\ \in \{(1,2),(3,4)\}, & p = 1, \ldots, L^n \end{array} : E_m^l + E_p^n - E_j^i - E_k^h > 0 \right\},$$

and taking

$$\begin{aligned} \nu_j^i = & \sum_{h,k,m,p \in \mathcal{D}^2{}_j^i} \nu_{j,k}^{m,p} n_k^h \exp\left(- \frac{E_m^i + E_p^h - E_j^i - E_k^h}{T} \right) + \sum_{h,k,m,p \in \mathcal{D}^1{}_j^i} \nu_{j,k}^{m,p} n_k^h \\ & + \sum_{h,k,l,m,n,p \in \mathcal{D}^4{}_j^i} \nu_{j,k}^{m,p} n_k^h \left(\frac{m^i m^j}{m^l m^n} \right)^{\frac{3}{2}} \exp\left(- \frac{E_m^l + E_p^n - E_j^i - E_k^h}{T} \right) \\ & + \sum_{h,k,l,m,n,p \in \mathcal{D}^3{}_j^i} \nu_{j,k}^{m,p} n_k^h \end{aligned}$$

with $\nu_{j,k}^{m,p} = \dfrac{k+j}{20(m+p)}$. Equilibrium values for concentrations c_{jM}^i obtained in this setting are reported in the following table

	C_1^1	C_2^1	C_1^2	C_2^2	C_3^2	C_1^3	C_2^3	C_1^4	C_2^4
c_M	0.001	0.001	0.042	0.031	0.027	0.256	0.192	0.256	0.192

while equilibrium global mean velocity is $u_M = 0.29$ and temperature is $T_M = 3.47$. The evolution in time of species concentrations computed numerically is depicted in Fig. 1. It is possible to observe that, according to constraint (9), for each gas species the component corresponding to a higher energy level will have a lower concentration and vice-versa. Moreover, due to relation (10), the concentrations of species G^1 and G^2 are lower, in particular the one of G^1 (that has the lowest mass in the mixture) is the lowest, while concentrations of species G^3 and G^4 are higher. In other words, chemical equilibrium is achieved when species G^1 is almost completely disappeared, so that almost no reactive collision can occur. We also note that trend to equilibrium for concentrations may be non monotone, see for instance c_1^2 and c_2^2. In Fig. 2 we report the behavior of normalized velocities $\bar{u}_j^i = u_j^i/u_M$ and normalized temperatures $\bar{T}_j^i = T_j^i/T_M$. We can observe that the species G^1 takes a longer time to reach the equilibrium value for velocity and temperature, its components keep nearly constant values in the first stage of the evolution.

The second reacting mixture we take into account for our simulations is the following

$$ClNO_2 + NO \leftrightarrows NO_2 + ClNO,$$

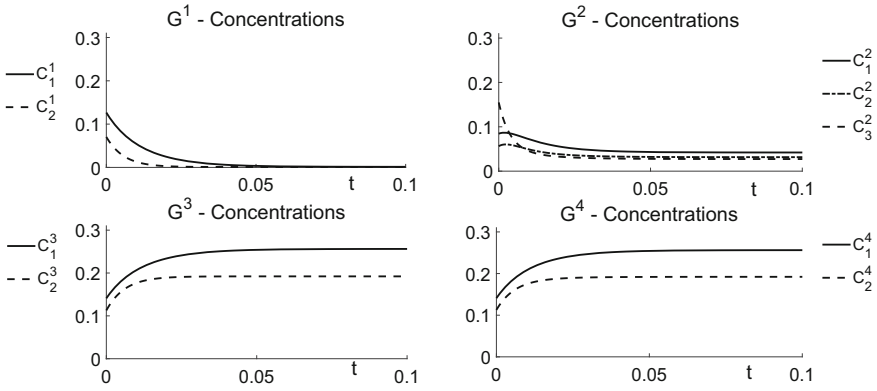

Fig. 1 Concentrations for a mixture of four reacting gases with energy levels and initial values as in (70), (71), considering masses $(m^1, m^2, m^3, m^4) = (0.1, 12.8, 6.45, 6.45)$

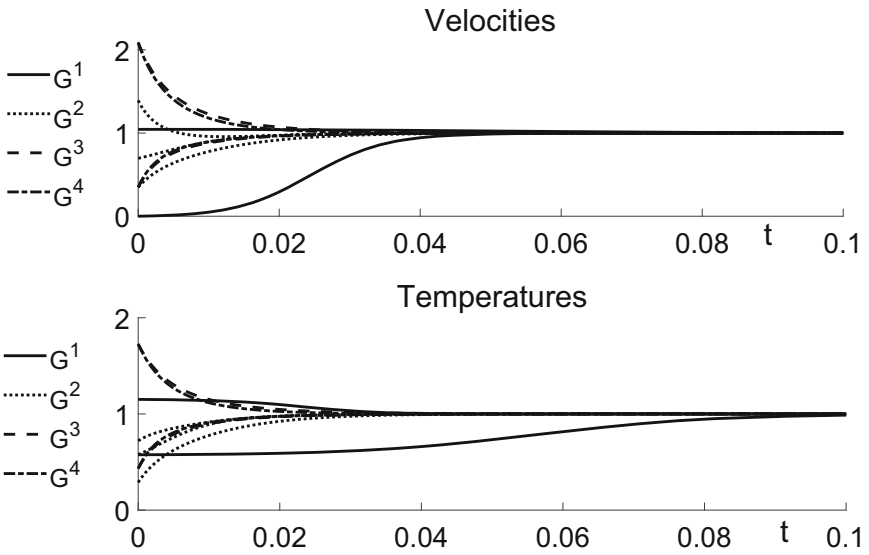

Fig. 2 Normalized velocities and temperatures for a mixture of four reacting gases with energy levels and initial values as in (70), (71), considering masses $(m^1, m^2, m^3, m^4) = (0.1, 12.8, 6.45, 6.45)$

where the chloro nitride $ClNO_2$ (81.46 g/mol) reacts with nitric oxide NO (30.01 g/mol) forming nitrosyl chloride $ClNO$ (65.46 g/mol) and nitrogen dioxide NO_2 (46.01 g/mol), and vice-versa in the reverse reaction. As before, we take in our model four gases having mass ratios similar to the ones involved in the real reaction: $m^1 = 1, m^2 = 2.72, m^3 = 2.18, m^4 = 1.53$. We make the assumption that the first gas G^1 is composed by two, the second gas G^2 is composed by four, and the other two gases G^3 and G^4 are composed by three components, respectively. Each component corresponds to a different internal energy level as follows

$$E_1^1 = 6, \quad E_2^1 = 7, \quad E_1^2 = 7, \quad E_2^2 = 8, \quad E_3^2 = 10, \quad E_4^2 = 12,$$

$$E_1^3 = 5.5, \quad E_2^3 = 6, \quad E_3^3 = 7.5, \quad E_1^4 = 4, \quad E_2^4 = 9, \quad E_3^4 = 10. \tag{72}$$

We set initial number concentrations, velocities and temperatures as reported in the following table

	C_1^1	C_2^1	C_1^2	C_2^2	C_3^2	C_4^2	C_1^3	C_2^3	C_3^3	C_1^4	C_2^4	C_3^4
c_0	0.12	0.09	0.07	0.08	0.11	0.06	0.07	0.05	0.13	0.03	0.08	0.09
u_0	0.3	0	0.1	0.4	0.2	0.6	0.1	0.4	0.5	0.3	0	0.2
T_0	2	4	1	2.5	2	6	1.5	2.5	3	4.5	5	1

(73)

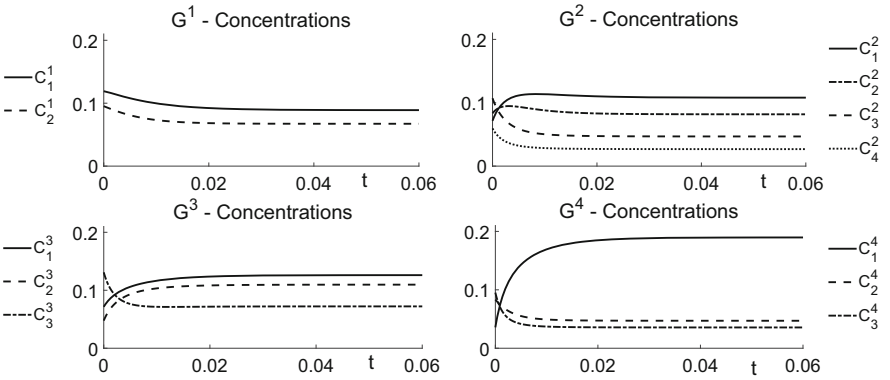

Fig. 3 Concentrations for a mixture of four reacting gases with energy levels and initial values as in (72), (73), considering masses $(m^1, m^2, m^3, m^4) = (1, 2.72, 2.18, 1.53)$

In this case, equilibrium values for equilibrium concentrations c^i_{jM} are

	C_1^1	C_2^1	C_1^2	C_2^2	C_3^2	C_4^2	C_1^3	C_2^3	C_3^3	C_1^4	C_2^4	C_3^4
c_M	0.09	0.07	0.12	0.08	0.04	0.02	0.13	0.11	0.07	0.2	0.05	0.04

while global mean velocity is $u_M = 0.28$ and temperature is $T_M = 3.59$. Numerical results for behavior in time of concentrations for all the components are showed in Fig. 3. In this case, since there is less difference among masses than in the previous case, final values of number densities are more similar; we only have a significantly higher concentration of component C_1^4 that corresponds to the lowest energy level. Trends for normalized mean velocities and temperatures of the species are reported in Fig. 4 and also in this case we can observe that the lighter gas G^1 takes a longer time to reach the equilibrium value.

4 Conclusions

We have generalized the BGK model proposed in [8] to a mixture of four polyatomic gases undergoing a bimolecular and reversible chemical reaction. The additional difficulties with respect to the inert frame are essentially due to two reasons. At first, single number densities are no more preserved during the evolution, since particles involved in a reactive collision change their nature; consequently, proper auxiliary number densities affect the Maxwellian attractors of the BGK collision operators, and they are related in a non-trivial way to species masses and concentrations, to global (auxiliary) temperature and to the chemical energy gap. Then, the mass action law of chemistry that characterizes chemical collision equilibrium, and that

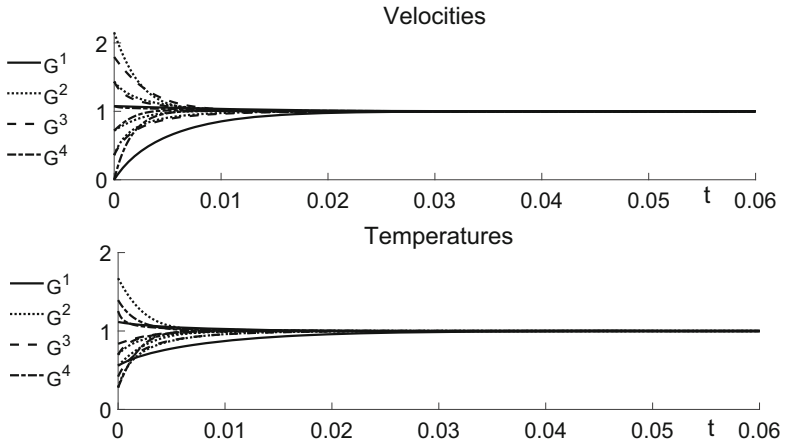

Fig. 4 Normalized velocities and temperatures for a mixture of four reacting gases with energy levels and initial values as in (72), (73), considering masses $(m^1, m^2, m^3, m^4) = (1, 2.72, 2.18, 1.53)$

is assumed to be valid also for auxiliary parameters, constitutes an additional transcendental equation to be combined to the energy conservation requirement (that is a transcendental law by itself) in order to prove well-posedness of auxiliary number densities and temperature. We have also performed some numerical simulations of evolution equations for species concentrations, velocities and temperatures corresponding to our BGK model. Specifically, we have considered two different mixtures, with particle mass ratios corresponding to real cases, namely to the ones of two real bimolecular and reversible chemical reactions. The trend to equilibrium turns out to be much slower for the species much lighter than the others, and this is in agreement with the separation of species with disparate masses observed in several physical problems [17]. Of course it will be interesting to test our BGK model and corresponding macroscopic equations also in space dependent problems, as for instance the shock wave structure, comparing our results with the ones obtained for inert mixtures in the frame of extended thermodynamics [20], or from kinetic systems for reactive monoatomic gases [6] or for a single polyatomic gas [19, 22].

Acknowledgments The author M.B. thanks INdAM for having supported the organization of the workshop *Recent Advances in Kinetic Equations and Applications* (Roma, November 11–15, 2019), where part of this work has been presented. The authors thank also the support by the University of Parma, by the Italian National Group of Mathematical Physics (GNFM-INdAM), and by the Italian National Research Project *Multiscale phenomena in Continuum Mechanics: singular limits, off-equilibrium and transitions* (Prin 2017YBKNCE).

References

1. Andries, P., Aoki, K., Perthame, B.: A consistent BGK-type model for gas mixtures. J. Stat. Phys. **106**, 993–1018 (2002)
2. Baranger, C., Bisi, M., Brull, S., Desvillettes, L.: On the Chapman–Enskog asymptotics for a mixture of monoatomic and polyatomic rarefied gases. Kinet. Relat. Models **11**, 821–858 (2018)
3. Bhatnagar, P.L., Gross, E.P., Krook, K.: A model for collision processes in gases. Phys. Rev. **94**, 511–524 (1954)
4. Bisi, M., Cáceres, M.J.: A BGK relaxation model for polyatomic gas mixtures. Commun. Math. Sci. **14**, 297–325 (2016)
5. Bisi, M., Groppi, M., Spiga, G.: Kinetic Bhatnagar–Gross–Krook model for fast reactive mixtures and its hydrodynamic limit. Phys. Rev. E **81**, 036327 (2010)
6. Bisi, M., Martalò, G., Spiga, G.: Multi–temperature fluid–dynamic model equations from kinetic theory in a reactive gas: the steady shock problem. Comput. Math. Appl. **66**, 1403–1417 (2013)
7. Bisi, M., Monaco, R., Soares, A.J.: A BGK model for reactive mixtures of polyatomic gases with continuous internal energy. J. Phys. A - Math. Theor. **51**, 125501 (2018)
8. Bisi, M., Travaglini, R.: A BGK model for mixtures of monoatomic and polyatomic gases with discrete internal energy. Phys. A: Stat. Mech. Appl. **547**, 124441 (2020)
9. Bobylev, A.V., Bisi, M., Groppi, M., Spiga, G., Potapenko, I.F.: A general consistent BGK model for gas mixtures. Kinet. Relat. Models **11**, 1377–1393 (2018).
10. Cercignani, C.: The Boltzmann Equation and Its Applications. Springer, New York (1988)
11. Chapman, S., Cowling, T.G.: The Mathematical Theory of Non-Uniform Gases. Cambridge University Press, Cambridge (1970)
12. Desvillettes, L., Monaco, R., Salvarani, F.: A kinetic model allowing to obtain the energy law of polytropic gases in the presence of chemical reactions. Eur. J. Mech. B/ Fluids **24**, 219–236 (2005)
13. Giovangigli, V.: Multicomponent Flow Modeling. Series on Modeling and Simulation in Science, Engineering and Technology. Birkhäuser, Boston (1999)
14. Groppi, M., Spiga, G.: Kinetic approach to chemical reactions and inelastic transitions in a rarefied gas. J. Math. Chem. **26**, 197–219 (1999)
15. Groppi, M., Spiga, G.: A Bhatnagar–Gross–Krook-type approach for chemically reacting gas mixtures. Phys. Fluids **16**, 4273–4284 (2004)
16. Haack, J.R., Hauck, C.D., Murillo, M.S.: A conservative, entropic multispecies BGK model. J. Stat. Phys. **168**, 826–856 (2017)
17. Huck, R.J., Johnson, E.A.: Possibility of double sound propagation in disparate-mass gas mixtures. Phys. Rev. Lett. **44**, 142–145 (1980)
18. Klingenberg, C., Pirner, M., Puppo, G.: A consistent kinetic model for a two-component mixture with an application to plasma. Kinet. Relat. Models **10**, 445–465 (2017)
19. Kosuge, S., Aoki, K.: Shock-wave structure for a polyatomic gas with large bulk viscosity. Phys. Rev. Fluids **3**, 023401 (2018)
20. Madjarevic, D., Ruggeri, T., Simic, S.: Shock structure and temperature overshoot in macroscopic multi-temperature model of mixtures. Phys. Fluids **26**, 106102 (2014)
21. McCormack, F.J.: Construction of linearized kinetic models for gaseous mixtures and molecular gases. Phys. Fluids **16**, 2095–2105 (1973)
22. Pavic-Colic, M., Madjarevic, D., Simic, S.: Polyatomic gases with dynamic pressure: kinetic non-linear closure and the shock structure. Int. J. Non Linear Mech. **92**, 160–175 (2017)
23. Szalmas, L., Pitakarnnop, J., Geoffroy, S., Colin, S., Valougeorgis, D.: Comparative study between computational and experimental results for binary rarefied gas flows through long microchannels. Microfluid. Nanofluidics **9**, 1103–1114 (2010)

On Some Recent Progress in the Vlasov–Poisson–Boltzmann System with Diffuse Reflection Boundary

Yunbai Cao and Chanwoo Kim

Abstract We discuss some recent development on the Vlasov–Poisson–Boltzmann system in bounded domains with diffuse reflection boundary condition. In addition we present a new regularity result when the particles are surrounded by conductor boundary.

1 Background

The object of kinetic theory is the modeling of particles by a distribution function in the phase space, which is denoted by $F(t, x, v)$ for $(t, x, v) \in [0, \infty) \times \Omega \times \mathbb{R}^3$ where Ω is an open bounded subset of $\mathbb{R}^3$. Dynamics and collision processes of dilute charged particles with an electric field E can be modeled by the (two-species) Vlasov–Poisson–Boltzmann equation

$$\begin{aligned} \partial_t F_+ + v \cdot \nabla_x F_+ + E \cdot \nabla_v F_+ &= Q(F_+, F_+) + Q(F_+, F_-), \\ \partial_t F_- + v \cdot \nabla_x F_- - E \cdot \nabla_v F_- &= Q(F_-, F_+) + Q(F_-, F_-). \end{aligned} \tag{1}$$

Here $F_\pm(t, x, v) \geq 0$ are the density functions for the ions $(+)$ and electrons $(-)$ respectively. The collision operator measures "the change rate" in binary hard sphere collisions and takes the form of ([14])

$$\begin{aligned} Q(F_1, F_2)(v) &:= Q_{\text{gain}}(F_1, F_2) - Q_{\text{loss}}(F_1, F_2) \\ &:= \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} |(v - u) \cdot \omega| [F_1(u') F_2(v') - F_1(u) F_2(v)] \mathrm{d}\omega \mathrm{d}u, \end{aligned} \tag{2}$$

Y. Cao · C. Kim (✉)
Department of Mathematics, University of Wisconsin-Madison, Madison, WI, USA
e-mail: ycao35@wisc.edu; chanwoo.kim@wisc.edu

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_4

where $u' = u - [(u - v) \cdot \omega]\omega$ and $v' = v + [(u - v) \cdot \omega]\omega$. The collision operator enjoys a collision invariance: for any measurable G, $\int_{\mathbb{R}^3} \left[1 \; v \; \frac{|v|^2-3}{2}\right] Q(G, G) \mathrm{d}v = \left[0 \; 0 \; 0\right]$. It is well-known that a global Maxwellian μ satisfies $Q(\cdot, \cdot) = 0$, where

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

The electric field E is given by

$$E(t, x) := -\nabla_x \phi(t, x), \tag{4}$$

where an electrostatic potential is determined by the Poisson equation:

$$- \Delta_x \phi(t, x) = \int_{\mathbb{R}^3} (F_+(t, x, v) - F_-(t, x, v))\, dv \;\text{ in } \Omega. \tag{5}$$

A simplified one-species Vlasov–Poisson–Boltzmann equation is often considered to reduce the complexity. Where we let $F(t, x, v)$ takes the role of $F_+(t, x, v)$, and assume $F_- = \rho_0 \mu$ where the constant $\rho_0 = \int_{\Omega \times \mathbb{R}^3} F_+(0, x, v)\, dvdx$. Then we get the system

$$\partial_t F + v \cdot \nabla_x F + E \cdot \nabla_v F = Q(F, F), \tag{6}$$

$$- \Delta_x \phi(t, x) = \int_{\mathbb{R}^3} F(t, x, v)\, dv - \rho_0 \;\text{ in } \Omega. \tag{7}$$

Here the background charge density ρ_0 is assumed to be a constant.

Throughout this paper, we use the notation

$$\iota = + \text{ or } -, \text{ and } -\iota = \begin{cases} - & , \text{ if } \iota = + \\ + & , \text{ if } \iota = -. \end{cases} \tag{8}$$

And for the one-species case, $F_\iota = F$.

In many physical applications, e.g. semiconductor and tokamak, the charged dilute gas is confined within a container, and its interaction with the boundary, which can be described by suitable boundary conditions, often plays a crucial role in global dynamics. In this paper we consider one of the physical conditions, a so-called diffuse boundary condition:

$$F_\iota(t, x, v) = \sqrt{2\pi}\mu(v) \int_{n(x) \cdot u > 0} F_\iota(t, x, u) \{n(x) \cdot u\} \mathrm{d}u \;\text{ for } (x, v) \in \gamma_-. \tag{9}$$

Here $\gamma_- := \{(x, v) \in \partial\Omega \times \mathbb{R}^3 : n(x) \cdot v < 0\}$, and $n(x)$ is the outward unit normal at a boundary point x.

Due to its importance, there have been many research activities in mathematical study of the Boltzmann equation. In [11], global strong solution of Boltzmann equation coupled with the Poisson equation has been established through the nonlinear energy method, when the initial data are close to the Maxwellian μ. In the large-amplitude regime, an almost exponential decay for Boltzmann solutions is established in [8], provided certain a priori strong Sobolev estimates can be verified. Such high regularity insures an L^∞-control of solutions which is crucial to handle the quadratic nonlinearity. Even though these estimates can be verified in periodic domains, their validity in general bounded domains have been doubted.

Despite its importance, mathematical theory on boundary problems of VPB, especially for strong solutions, hasn't been developed up to satisfactory (cf. renormalized solutions of VPB were constructed in [18]). One of the fundamental difficulties for the system in bounded domains is the lack of higher regularity, which originates from the characteristic nature of boundary conditions in the kinetic theory, and the nonlocal property of the collision term Q. This nonlocal term indicates that the local behavior of the solution could be affected globally by x and v, and thus prevents the localization of the solution. From that a seemingly inevitable singularity of the spatial normal derivative at the boundary $x \in \partial\Omega$ arises $\partial_n F_\iota(t,x,v) \sim \frac{1}{n(x)\cdot v} \notin L^1_{loc}$. Such singularity towards the grazing set $\gamma_0 := \{(x,v) \in \partial\Omega \times \mathbb{R}^3 : n(x)\cdot v = 0\}$ has been studied thoroughly in [13] for the Boltzmann equation in convex domain. For recent development of the boundary theory of the Boltzmann equation, we refer to [9, 12, 15–17] and the references therein. Here we clarify that a C^α domain means that for any $p \in \partial\Omega$, there exists sufficiently small $\delta_1 > 0, \delta_2 > 0$, and an one-to-one and onto C^α-map, $\eta_p : \{(x_{\|,1}, x_{\|,2}, x_n) \in \mathbb{R}^3 : x_n > 0\} \cap B(0;\delta_1) \to \Omega \cap B(p;\delta_2)$ with $\eta_p(x_{\|,1}, x_{\|,2}, x_n) = \eta_p(x_{\|,1}, x_{\|,2}, 0) + x_n[-n(\eta_p(x_{\|,1}, x_{\|,2}, 0))]$, such that $\eta_p(\cdot,\cdot,0) \in \partial\Omega$ ([10]). A *convex* domain means that there exists $C_\Omega > 0$ such that for all $p \in \partial\Omega$ and η_p and for all $x_\|$,

$$\sum_{i,j=1}^{2} \zeta_i \zeta_j \partial_i \partial_j \eta_p(x_\|) \cdot n(x_\|) \le -C_\Omega |\zeta|^2 \text{ for all } \zeta \in \mathbb{R}^2. \tag{10}$$

Construction of a unique global solution and proving its asymptotic stability of VPB in general domains has been a challenging open problem for any boundary condition. In [4] the authors give the *first* construction of a unique global *strong* solution of the one-species VPB system with the diffuse boundary condition when the domain is C^3 and *convex*. Moreover an asymptotic stability of the global Maxwellian μ is studied. The result was then extended to the two-species case in [6].

2 Global Strong Solution of VPB

In [4, 6], the authors take the first step toward comprehensive understanding of VPB in bounded domains. They consider the zero Neumann boundary condition for the potential ϕ: $n \cdot E|_{\partial\Omega} = \frac{\partial\phi}{\partial n}|_{\partial\Omega} = 0$, which corresponds to a so-called insulator boundary condition. In such setting $(F_\iota, E) = (\mu, 0)$ is a stationary solution.

The characteristics (trajectory) is determined by the Hamilton ODEs for f_+ and f_- separately

$$\frac{d}{ds}\begin{bmatrix} X_\iota^f(s;t,x,v) \\ V_\iota^f(s;t,x,v) \end{bmatrix} = \begin{bmatrix} V_\iota^f(s;t,x,v) \\ -\iota\nabla_x\phi_f(s, X_\iota^f(s;t,x,v)) \end{bmatrix} \quad \text{for} -\infty < s,t < \infty, \tag{11}$$

with $(X_\iota^f(t;t,x,v), V_\iota^f(t;t,x,v)) = (x,v)$. Where the potential is extended to negative time as $\phi_f(t,x) = e^{-|t|}\phi_{f_0}(x)$ for $t \le 0$. For $(t,x,v) \in \mathbb{R} \times \Omega \times \mathbb{R}^3$, define *the backward exit time* $t_{\mathbf{b},\iota}^f(t,x,v)$ as

$$t_{\mathbf{b},\iota}^f(t,x,v) := \sup\{s \ge 0 : X_\iota^f(\tau;t,x,v) \in \Omega \ \text{ for all } \tau \in (t-s,t)\}. \tag{12}$$

Furthermore, define $x_{\mathbf{b},\iota}^f(t,x,v) := X_\iota^f(t - t_{\mathbf{b},\iota}(t,x,v);t,x,v)$ and $v_{\mathbf{b},\iota}^f(t,x,v) := V_\iota^f(t - t_{\mathbf{b},\iota}(t,x,v);t,x,v)$. In order to handle the boundary singularity, they introduce the following notion

Definition 1 (Kinetic Weight) For $\varepsilon > 0$

$$\begin{aligned}\alpha_{f,\varepsilon,\iota}(t,x,v) := \chi\Big(\frac{t - t_{\mathbf{b},\iota}^f(t,x,v) + \varepsilon}{\varepsilon}\Big)|n(x_{\mathbf{b},\iota}^f(t,x,v)) \cdot v_{\mathbf{b},\iota}^f(t,x,v)| \\ + \Big[1 - \chi\Big(\frac{t - t_{\mathbf{b},\iota}^f(t,x,v) + \varepsilon}{\varepsilon}\Big)\Big].\end{aligned} \tag{13}$$

Here they use a smooth function $\chi : \mathbb{R} \to [0,1]$ satisfying

$$\chi(\tau) = 0, \ \tau \le 0, \ \text{and} \ \chi(\tau) = 1, \ \tau \ge 1. \ \frac{d}{d\tau}\chi(\tau) \in [0,4] \ \text{ for all } \tau \in \mathbb{R}. \tag{14}$$

Also, denote

$$\alpha_{f,\varepsilon}(t,x,v) := \begin{bmatrix} \alpha_{f,\varepsilon,+}(t,x,v) & 0 \\ 0 & \alpha_{f,\varepsilon,-}(t,x,v) \end{bmatrix}. \tag{15}$$

Note that $\alpha_{f,\varepsilon,\iota}(0,x,v) \equiv \alpha_{f_0,\varepsilon,\iota}(0,x,v)$ is determined by f_0. For the sake of simplicity, the superscription ^{f} in X_ι^f, V_ι^f, $t_{\mathbf{b},\iota}^f$, $x_{\mathbf{b},\iota}^f$, $v_{\mathbf{b},\iota}^f$ is dropped unless they could cause any confusion. One of the crucial properties of the kinetic weight in (13) is an invariance under the Vlasov operator: $\left[\partial_t + v\cdot\nabla_x - \nabla_x\phi_f\cdot\nabla_v\right]\alpha_{f,\varepsilon,\iota}(t,x,v) = 0$. This is due to the fact that the characteristics solves a deterministic system (11). This crucial invariant property under the Vlasov operator is one of the key points in their approach in [4, 6].

Theorem 1 ([4, 6]) *Let $w_\vartheta(v) = e^{\vartheta|v|^2}$. Assume a bounded open C^3 domain $\Omega \subset \mathbb{R}^3$ is convex (10). Let $0 < \tilde{\vartheta} < \vartheta \ll 1$. Assume the compatibility condition: (9) holds at $t = 0$. There exists a small constant $0 < \varepsilon_0 \ll 1$ such that for all $0 < \varepsilon \le \varepsilon_0$ if an initial datum $F_{0,\iota} = \mu + \sqrt{\mu} f_{0,\iota}$ satisfies*

$$\|w_\vartheta f_{0,\iota}\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} < \varepsilon, \ \|w_{\tilde{\vartheta}}\nabla_v f_{0,\iota}\|_{L^3(\Omega\times\mathbb{R}^3)} < \infty, \tag{16}$$

$$\|w_{\tilde{\vartheta}}\alpha_{f_{0,\iota},\varepsilon}^\beta \nabla_{x,v} f_{0,\iota}\|_{L^p(\Omega\times\mathbb{R}^3)} < \varepsilon \ \text{ for } \ 3 < p < 6, \ \ 1 - \frac{2}{p} < \beta < \frac{2}{3}, \tag{17}$$

then there exists a unique global-in-time solution $F_\iota(t) = \mu + \sqrt{\mu} f_\iota(t) \ge 0$ to (1), (4), (5), (9). Moreover there exists $\lambda_\infty > 0$ such that

$$\sup_{t\ge 0} e^{\lambda_\infty t}\|w_\vartheta f_\iota(t)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} + \sup_{t\ge 0} e^{\lambda_\infty t}\|\phi_f(t)\|_{C^2(\Omega)} \lesssim 1, \tag{18}$$

and, for some $C > 0$, and, for $0 < \delta = \delta(p,\beta)$,

$$\|w_{\tilde{\vartheta}}\alpha_{f,\varepsilon,\iota}^\beta \nabla_{x,v} f_\iota(t)\|_{L^p(\Omega\times\mathbb{R}^3)} \lesssim e^{Ct} \ \text{ for all } t \ge 0, \tag{19}$$

$$\|\nabla_v f_\iota(t)\|_{L_x^3(\Omega)L_v^{1+\delta}(\mathbb{R}^3)} \lesssim_t 1 \ \text{ for all } \ t \ge 0. \tag{20}$$

Furthermore, if F_ι nad G_ι are both solutions to (1), (4), (5), (9), then

$$\|f_\iota(t) - g_\iota(t)\|_{L^{1+\delta}(\Omega\times\mathbb{R}^3)} \lesssim_t \|f_\iota(0) - g_\iota(0)\|_{L^{1+\delta}(\Omega\times\mathbb{R}^3)} \ \text{ for all } \ t \ge 0. \tag{21}$$

Remark 1 The second author and his collaborators constructs a local-in-time solution for given general large datum in [7] for the generalized diffuse reflection boundary condition. By introducing a scattering kernel $R(u \to v; x, t)$, representing the probability of a molecule striking in the boundary at $x \in \partial\Omega$ with velocity u to be bounced back to the domain with velocity v, they consider

$$F(t,x,v)|n(x)\cdot v| = \int_{\gamma_+(x)} R(u \to v; x, t) F(t,x,u)\{n(x)\cdot u\}du, \qquad \text{on } \gamma_-. \tag{22}$$

In [7] they study a model proposed by Cercignani and Lampis in [2, 3]. With two accommodation coefficients $0 < r_\perp \le 1, \quad 0 < r_\| < 2$, the Cercignani-Lampis boundary condition (C-L boundary condition) can be written as

$$\begin{aligned} & R(u \to v; x, t) \\ :=& \frac{1}{r_\perp r_\| (2-r_\|)\pi/2} \frac{|n(x)\cdot v|}{(2T_w(x))^2} I_0\left(\frac{1}{2T_w(x)} \frac{2(1-r_\perp)^{1/2} v_\perp u_\perp}{r_\perp}\right) \\ & \times \exp\left(-\frac{1}{2T_w(x)}\left[\frac{|v_\perp|^2 + (1-r_\perp)|u_\perp|^2}{r_\perp} + \frac{|v_\| - (1-r_\|)u_\||^2}{r_\|(2-r_\|)}\right]\right). \end{aligned} \tag{23}$$

Here $T_w(x)$ is a wall temperature on the boundary and $I_0(y) := \pi^{-1}\int_0^\pi e^{y\cos\phi} d\phi$. In this formula, $v_\perp$ and $v_\|$ denote the normal and tangential components of the velocity respectively: $v_\perp = v \cdot n(x)$, $v_\| = v - v_\perp n(x)$.

In [4, 6] a global L^∞-bound is proven by $L^2 - L^\infty$ framework. The idea is to use Duhamel's principle to estimate the solution f along the characteristics (11) to reach

$$\|f_\iota(t)\|_{L^\infty(\bar\Omega\times\mathbb{R}^3)} \sim \|e^{-t} f_{0,\iota}\|_{L^\infty(\bar\Omega\times\mathbb{R}^3)} + \int_0^t e^{-(t-s)} \|f_\iota(s)\|_{L^2(\Omega\times\mathbb{R}^3)} ds.$$

And then use the decay of f_ι in L^2 norm to conclude the decay in L^∞. The key of this process is to verify

$$\begin{aligned} \frac{\partial X_\iota(s;t,x,v)}{\partial v} &\sim -(t-s)\mathrm{Id}_{3\times 3} + \int_s^t \int_{s'}^t \nabla_x^2 \phi(s'') \frac{\partial X_\iota(s'';t,x,v)}{\partial v} ds' ds'' \\ &\sim O(|t-s|)\mathrm{Id}_{3\times 3}. \end{aligned} \tag{24}$$

For which the C^2-bound of ϕ seems necessary. Unfortunately such C^2 estimate for ϕ falls short of the boarder line case of the Schauder elliptic regularity theory when the source term of the Poisson equation $\int_{\mathbb{R}^3}(F_+ - F_-)dv$ in (5) is merely continuous or bounded. They overcome such difficulty by interpolating the C^2 norm into a sum of a $C^{2,0+}$ norm and a $C^{1,1-}$ norm:

Lemma 1 *Assume $\Omega \subset \mathbb{R}^3$ with a C^2 boundary $\partial\Omega$. For $0 < D_1 < 1$, $0 < D_2 < 1$, and $\Lambda_0 > 0$,*

$$\|\nabla_x^2 \phi(t)\|_{L^\infty(\Omega)} \lesssim_{\Omega, D_1, D_2} e^{D_1\Lambda_0 t} \|\phi(t)\|_{C^{1,1-D_1}(\Omega)} + e^{-D_2\Lambda_0 t} \|\phi(t)\|_{C^{2,D_2}(\Omega)}. \tag{25}$$

While an exponential decay of the weaker $C^{1,1-}$ norm can be derived from the exponential decay of f_ι in L^∞, the $C^{2,0+}$ norm is controlled by Morrey's inequality

$$\|\phi\|_{C_x^{2,0+}} \lesssim \sum_{\iota=\pm} \| \int_{\mathbb{R}^3} f_\iota \sqrt{\mu} dv \|_{C_x^{0,0+}} \lesssim \sum_{\iota=\pm} \| \int_{\mathbb{R}^3} \nabla_x f_\iota \sqrt{\mu} dv \|_{L_x^p}, \text{ for } p > 3. \tag{26}$$

Now the spatial derivative of f_ι needs to be controlled. They develop an α_ι-weighted $W^{1,p}$ estimate by energy-type estimate of $\alpha_\iota \nabla_{x,v} f_\iota$, where the α_ι-multiplication compensates the boundary singularity. This allows us to bound (26) for $\frac{p-2}{p} < \beta < \frac{p-1}{p}$,

$$\| \int_{\mathbb{R}^3} \nabla_x f_\iota \sqrt{\mu} dv \|_{L_x^p} \lesssim \|\alpha_\iota^{-\beta}\|_{L^{\frac{p}{p-1}}} \|\alpha_\iota^\beta \nabla_x f_\iota \sqrt{\mu}\|_{L_{x,v}^p} \lesssim \|\alpha_\iota^\beta \nabla_x f_\iota \sqrt{\mu}\|_{L_{x,v}^p},$$

as long as

$$\alpha_\iota^{-\frac{\beta p}{p-1}} \sim \frac{1}{\alpha_\iota(t,x,v)^{1-}} \in L_v^1 \text{ uniformly for all } x. \tag{27}$$

A difficulty of the proof of (27) arises form lack of local representation of $\alpha_\iota(t,x,v)$. α_ι is only defined at some boundary point along (possibly very complicated) characteristics. They employ a geometric change of variables $v \mapsto (x_{\mathbf{b},\iota}(t,x,v), t_{\mathbf{b},\iota}(t,x,v))$ to exam (27). By computing the Jacobian there is an extra α-factor from $dv \sim \frac{\alpha_\iota}{|t_{\mathbf{b},\iota}|^3} dt_{\mathbf{b},\iota} dx_{\mathbf{b},\iota}$, which cancels the singularity of (27). Then they use a lower bound of $t_{\mathbf{b},\iota} \gtrsim \frac{|x_{\mathbf{b},\iota}^f - x|}{\max|V|}$ and a bound $\alpha \lesssim \frac{|(x-x_{\mathbf{b},\iota}^f)\cdot n(x_{\mathbf{b},\iota}^f)|}{t_{\mathbf{b},\iota}^f}$ to have

$$\int_{|v|\lesssim 1} \alpha_\iota^{-\frac{\beta p}{p-1}} \mathrm{d}v \lesssim \int_{\text{boundary}} \frac{|(x-x_{\mathbf{b},\iota})\cdot n(x_{\mathbf{b},\iota})|^{1-\frac{\beta p}{p-1}}}{|x-x_{\mathbf{b},\iota}|^{3-\frac{\beta p}{p-1}}} \mathrm{d}x_{\mathbf{b},\iota} + \text{good terms} < \infty, \tag{28}$$

which turns to be bounded as long as $\frac{\beta p}{p-1} < 1$.

From the above estimates and the interpolation, they derive *an exponential decay* of $\phi(t)$ in C_x^2 as long as $\|\alpha_\iota^\beta \nabla_x f(t)\|_{L_{x,v}^p}$ grows at most exponentially. With the C_x^2-bound of ϕ in hand, they control $\|\alpha_\iota^\beta \nabla_x f(t)\|_{L_{x,v}^p}$ via Gronwall's inequality and close the estimate by proving its (at most) exponential growth.

For the uniqueness and stability of approximating sequence they prove L^1-stability. The key observation is that v-derivatives of the diffuse BC (9) has no boundary singularity, thus is bounded. The equation of $\nabla_v f_\iota$ has a singular forcing term $\nabla_x f_\iota$. For which they control $\|\nabla_x f_\iota\|_{L_x^3 L_v^1}$ as $\|\alpha_\iota^{-\beta}\|_{L_v^{\frac{p}{p-1}}} \|\alpha_\iota^\beta \nabla_x f_\iota\|_{L_{x,v}^p}$, and this term is bounded from (27).

3 Improved Regularity Under the Sign Condition

One interesting question is to improve the regularity estimate beyond a weighted $W^{1,p}$ for $p < 6$ of f_t in [4, 6]. Some work in this direction has been done in [5].

In [5] the author consider the one-species VPB system (4), (6), where the potential consists of a self-generated electrostatic potential and an external potential. That is $E = \nabla\phi$, where

$$\phi(t,x) = \phi_F(t,x) + \phi_E(t,x), \text{ with } \frac{\partial \phi_E}{\partial n} > C_E > 0 \text{ on } \partial\Omega, \tag{29}$$

and ϕ_F satisfies (7) and the zero Neumann boundary condition $\frac{\partial \phi_F}{\partial n} = 0$ on $\partial\Omega$. Under such setting, the field E satisfies a crucial sign condition on the boundary

$$E(t,x)\cdot n(x) > C_E > 0 \text{ for all } t \text{ and all } x \in \partial\Omega. \tag{30}$$

With the help of the external potential ϕ_E with the crucial sign condition (29), they construct a short time weighted $W^{1,\infty}$ solution to the VPB system, which improves the regularity estimate of such system in Theorem 1. The key idea of the result is to incorporate a different distance function $\tilde{\alpha}$:

$$\tilde{\alpha} \sim \left[|v\cdot\nabla\xi(x)|^2 + \xi(x)^2 - 2(v\cdot\nabla^2\xi(x)\cdot v)\xi(x) - 2(E(t,\overline{x})\cdot\nabla\xi(\overline{x}))\xi(x) \right]^{1/2}, \tag{31}$$

where $\xi : \mathbb{R}^3 \to \mathbb{R}$ is a smooth function such that $\Omega = \{x \in \mathbb{R}^3 : \xi(x) < 0\}$, and the closest boundary point $\overline{x} := \{\bar{x} \in \partial\Omega : d(x,\bar{x}) = d(x,\partial\Omega)\}$ is uniquely defined for x closed to the boundary. Note that $\tilde{\alpha}|_{\gamma_-} \sim |n(x)\cdot v|$. A version of a distance function without the potential was used in [13]. One of the key contribution in [5] is to incorporate this different distance function (31) in the presence of an external field.

Theorem 2 ([5]) *Let $\phi_E(t,x)$ be a given external potential with $\nabla_x\phi_E$ satisfying (30), and $\|\nabla_x\phi_E(t,x)\|_{C^1_{t,x}(\mathbb{R}_+\times\bar{\Omega})} < \infty$. Assume that, for some $0 < \vartheta < \frac{1}{4}$, $\|w_\vartheta\tilde{\alpha}\nabla_{x,v}f_0\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} + \|w_\vartheta f_0\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} < \infty$. Then there exists a unique solution $F(t,x,v) = \sqrt{\mu}f(t,x,v)$ to (6), (4), (9), (29) for $t \in [0,T]$ with $0 < T \ll 1$, such that for some $0 < \vartheta' < \vartheta$, $\varpi \gg 1$, $\sup_{0\le t\le T}\|w_{\vartheta'}f(t)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} < \infty$, and*

$$\sup_{0\le t\le T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_{x,v}f(t,x,v)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} < \infty. \tag{32}$$

One of the crucial property $\tilde{\alpha}$ enjoys, under the assumption of the sign condition (30), is the invariance along the characteristics:

Lemma 2 (Velocity Lemma Near Boundary) *Suppose $E(t,x)$ satisfies the sign condition (30). Then for any $0 \le s < t$ and trajectory $X(\tau), V(\tau)$ solving (11), if $X(\tau) \in \Omega$ for all $s \le \tau \le t$, then*

$$\begin{aligned} e^{-C\int_s^t (|V(\tau')|+1)d\tau'}\tilde{\alpha}(s,X(s),V(s)) &\le \tilde{\alpha}(t,X(t),V(t)) \\ &\le e^{C\int_s^t (|V(\tau')|+1)d\tau'}\tilde{\alpha}(s,X(s),V(s)), \end{aligned} \tag{33}$$

for any $C \gtrsim (\|\nabla_x \phi_E(t,x)\|_{C^1_{t,x}(\mathbb{R}_+\times\bar{\Omega})}+1)/C_E$.

The key ingredient in the $\tilde{\alpha}$-weighted regularity estimate is a dynamical non-local to local estimate which can be stated as

Lemma 3 *Let $(t,x,v) \in [0,T]\times\Omega\times\mathbb{R}^3$, $1<\beta<3$, $0<\kappa\le 1$. Suppose E satisfies the sign condition (30). Then for $\varpi \gg 1$ large enough, and for any $0<C_\vartheta<\frac{1}{4}$, $0<\delta\ll 1$,*

$$\begin{aligned} &\int_{\max\{0,t-t_{\mathbf{b}}\}}^t \int_{\mathbb{R}^3} e^{-\int_s^t \frac{\varpi}{2}\langle V(\tau;t,x,v)\rangle d\tau} \frac{e^{-\frac{C_\vartheta}{2}|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}} \frac{1}{(\tilde{\alpha}(s,X(s),u))^\beta} du ds \\ &\lesssim e^{2C_\Omega \frac{\|\nabla E\|_\infty + \|E\|^2_{L^\infty_{t,x}} + \|E\|_{L^\infty_{t,x}}}{C_E}} \frac{\delta^{\frac{3-\beta}{2}}}{\langle v\rangle^2 (C_E+1)^{\frac{\beta-1}{2}} (\tilde{\alpha}(t,x,v))^{\beta-2} (\|E\|^2_{L^\infty_{t,x}}+1)^{\frac{3-\beta}{2}}} \\ &\quad + \frac{(\|E\|^2_{L^\infty_{t,x}}+1)^{\beta-1}}{C_E^{\beta-1}\delta^{\beta-1}(\tilde{\alpha}(t,x,v))^{\beta-1}} \frac{2}{\varpi}, \end{aligned} \tag{34}$$

where $(X(s),V(s)) = (X(s;t,x,v),V(s;t,x,v))$ as in (11).

The same estimate without the external field had been established by the second author and his collaborators in [13]. The proof of (34) is obtained by first making use of a series of change of variables to get the precise estimate of the velocity integration, which is bounded by,

$$\int_{\mathbb{R}^3} \frac{e^{-\vartheta|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}[\tilde{\alpha}(s,X(s),u)]^\beta} du \lesssim \frac{1}{(|V(s)|^2\xi(X(s)) - C_E\xi(X(s)))^{\frac{\beta-1}{2}}}, \tag{35}$$

then followed by relating the time integration back to $\tilde{\alpha}^{-1}$. For the later part of the proof, the velocity lemma (33) and the boundedness of the external field to ensure the monotonicity of $|\xi(X(s))|$ near the boundary, where the change of variable $dt \simeq \frac{d\xi}{|v\cdot\nabla\xi|}$, can be performed and recovers a power of $\tilde{\alpha}$ in the ξ-integration. On the other hand, the sign condition (29) is crucially used to establish a lower bound for $|\xi(X(s))|$ when it's away from the boundary, which helps to recover a power of $\tilde{\alpha}$ as wanted.

4 On the Vlasov–Poisson–Boltzmann System Surrounded by Conductor Boundary

In the second part of the paper, we consider the one-species VPB system surrounded by conductor boundary. More specifically, we consider the system (6), (4), where the electrostatic potential ϕ is obtained by

$$-\Delta_x\phi(t,x) = \int_{\mathbb{R}^3} F(t,x,v)dv,\ x\in\Omega,\quad \phi = 0,\ x\in\partial\Omega. \tag{36}$$

An important benefit in the conductor boundary setting (36) is that $E = -\nabla_x\phi$ enjoys the sign condition (30) from a quantitative Hopf lemma, without the need of an external potential.

Lemma 4 (Lemma 3.2 in [1]) *Suppose $h \geq 0$, and $h \in L^\infty(\Omega)$. Let v be the solution of*

$$-\Delta v = h \text{ in } \Omega,\quad v = 0 \text{ on } \partial\Omega. \tag{37}$$

Then for any $x \in \partial\Omega$,

$$\frac{\partial v(x)}{\partial n} \geq c\int_\Omega h(x)d(x,\partial\Omega)dx, \tag{38}$$

for some $c > 0$ depending only on Ω. Here $d(x,\partial\Omega)$ is the distance from x to the boundary $\partial\Omega$.

Our goal is to prove a local existence and regularity theorem for the system (6), (4), (9), (36). Let's first define our distance function $\tilde{\alpha}$.

Let $d(x,\partial\Omega) := \inf_{y\in\partial\Omega}\|x-y\|$. For any $\delta > 0$, let $\Omega^\delta := \{x\in\Omega : d(x,\partial\Omega) < \delta\}$. For $\delta \ll 1$ is small enough, we have for any $x\in\Omega^\delta$ there exists a unique $\bar{x}\in\partial\Omega$ such that $d(x,\bar{x}) = d(x,\partial\Omega)$ (cf. (2.44) in [5]).

Definition 2 First we define for all $(x, v) \in \Omega^\delta \times \mathbb{R}^3$,

$$\beta(t,x,v) = \left[|v\cdot\nabla\xi(x)|^2 + \xi(x)^2 - 2(v\cdot\nabla^2\xi(x)\cdot v)\xi(x) + 2(\nabla\phi(t,\overline{x})\cdot\nabla\xi(\overline{x}))\xi(x) \right]^{1/2}.$$

For any $\epsilon > 0$, let $\chi_\epsilon : [0,\infty) \to [0,\infty)$ be a smooth function satisfying $\chi_\epsilon(x) = x$ for $0 \le x \le \frac{\epsilon}{4}$, $\chi_\epsilon(x) = C_\epsilon$ for $x \ge \frac{\epsilon}{2}$, $\chi_\epsilon(x)$ is increasing for $\frac{\epsilon}{4} < x < \frac{\epsilon}{2}$, and $\chi'_\epsilon(x) \le 1$. Let $\delta' := \min\{|\xi(x)| : x \in \Omega, d(x,\partial\Omega) = \delta\}$, then we define our weight function to be:

$$\tilde{\alpha}(t,x,v) := \begin{cases} (\chi_{\delta'}(\beta(t,x,v))) & x \in \Omega^\delta, \\ C_{\delta'} & x \in \Omega \setminus \Omega^\delta. \end{cases} \tag{39}$$

Theorem 3 (Weighted $W^{1,\infty}$ Estimate for the VPB Surrounded by Conductor) *Assume $F_0 = \sqrt{\mu} f_0$ satisfies*

$$\|w_\vartheta \tilde{\alpha}\nabla_{x,v} f_0\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} + \|w_\vartheta f_0\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} + \|w_\vartheta \nabla_v f_0\|_{L^3(\bar{\Omega}\times\mathbb{R}^3)} < \infty, \tag{40}$$

for some $0 < \vartheta < \frac{1}{4}$. Then there exists a unique solution $F(t,x,v) = \sqrt{\mu} f(t,x,v)$ to (6), (4), (9), (36) for $t \in [0,T]$ with $0 < T \ll 1$, such that for some $0 < \vartheta' < \vartheta$, $\varpi \gg 1$,

$$\sup_{0\le t\le T} \|w_{\vartheta'} f(t)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} < \infty, \tag{41}$$

$$\sup_{0\le t\le T} \|w_{\vartheta'} e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_{x,v} f(t,x,v)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} < \infty, \tag{42}$$

$$\sup_{0\le t\le T} \|e^{-\varpi\langle v\rangle t}\nabla_v f(t)\|_{L^3_x(\Omega)L^{1+\delta}_v(\mathbb{R}^3)} < \infty \text{ for } 0 < \delta \ll 1. \tag{43}$$

The corresponding equation for $f = \frac{F}{\sqrt{\mu}}$ is

$$(\partial_t + v\cdot\nabla_x - \nabla\phi\cdot\nabla_v + \frac{v}{2}\cdot\nabla\phi + \nu(\sqrt{\mu}f))f = \Gamma_{\text{gain}}(f,f), \tag{44}$$

$$-\Delta_x\phi(t,x) = \int_{\mathbb{R}^3} \sqrt{\mu} f\,dv, \ \ \phi = 0 \text{ on } \partial\Omega, \tag{45}$$

$$f(t,x,v) = c_\mu\sqrt{\mu(v)}\int_{n\cdot u>0} f(t,x,v)\sqrt{\mu(u)}(n(x)\cdot u)du. \tag{46}$$

Here $\nu(\sqrt{\mu}f)(v) := \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |v-u|^\kappa q_0(\frac{v-u}{|v-u|}\cdot w)\sqrt{\mu(u)} f(u)d\omega du$, and $\Gamma_{\text{gain}}(f_1,f_2)(v) := \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} |v-u|^\kappa q_0(\frac{v-u}{|v-u|}\cdot w)\sqrt{\mu(u)} f_1(u')f_2(v')d\omega du$.

Let $\partial \in \{\nabla_x, \nabla_v\}$. Let $E = -\nabla_x \phi$. Denote

$$\nu_\varpi = \nu(\sqrt{\mu} f) + \frac{v}{2} \cdot E + \varpi \langle v \rangle + t\varpi \frac{v}{\langle v \rangle} \cdot E - \tilde{\alpha}^{-1}(\partial_t \tilde{\alpha} + v \cdot \nabla_x \tilde{\alpha} + E \cdot \nabla_v \tilde{\alpha}). \tag{47}$$

Then by direct computation we get

$$\begin{aligned} &\left\{ \partial_t + v \cdot \nabla_x + E \cdot \nabla_v + \nu_\varpi \right\} (e^{-\varpi \langle v \rangle t} \tilde{\alpha} \partial f) \\ &= e^{-\varpi \langle v \rangle t} \tilde{\alpha} \left(\partial \Gamma_{gain}(f, f) - \partial v \cdot \nabla_x f - \partial E \cdot \nabla_v f - \partial (\frac{v}{2} \cdot E) f - \partial (\nu(\sqrt{\mu} f)) f \right) \\ &:= \mathcal{N}(t, x, v). \end{aligned} \tag{48}$$

In order to deal with the diffuse boundary condition (9), we define the stochastic (diffuse) cycles as $(t^0, x^0, v^0) = (t, x, v)$,

$$\begin{aligned} &t^1 = t - t_{\mathbf{b}}(t, x, v),\ x^1 = x_{\mathbf{b}}(t, x, v) = X(t - t_{\mathbf{b}}(t, x, v); t, x, v), \\ &v_b^0 = V(t - t_{\mathbf{b}}(t, x, v); t, x, v) = v_{\mathbf{b}}(t, x, v), \end{aligned} \tag{49}$$

and $v^1 \in \mathbb{R}^3$ with $n(x^1) \cdot v^1 > 0$. For $l \geq 1$, define

$$\begin{aligned} &t^{l+1} = t^l - t_{\mathbf{b}}(t^l, x^l, v^l),\ x^{l+1} = x_{\mathbf{b}}(t^l, x^l, v^l), \\ &v_b^l = v_{\mathbf{b}}(t^l, x^l, v^l), \end{aligned}$$

and $v^{l+1} \in \mathbb{R}^3$ with $n(x^{l+1}) \cdot v^{l+1} > 0$. Also, define

$$X^l(s) = X(s; t^l, x^l, v^l),\ V^l(s) = V(s; t^l, x^l, v^l),$$

so $X(s) = X^0(s)$, $V(s) = V^0(s)$. We have the following lemma.

Lemma 5 (Lemma 12 in [5])
If $t^1 < 0$, then

$$\begin{aligned} &e^{-\varpi \langle v \rangle t} \tilde{\alpha} |\partial f(t, x, v)| \\ &\lesssim \tilde{\alpha}(0, X^0(0), V^0(0)) \partial f(0, X^0(0), V^0(0)) + \int_0^t \mathcal{N}(s, X^0(s), V^0(s)) ds. \end{aligned} \tag{50}$$

If $t^1 > 0$, then

$$
\begin{aligned}
& e^{-\varpi\langle v\rangle t}\tilde{\alpha}|\partial f(t,x,v)| \\
\lesssim & e^{-\frac{\vartheta}{2}|v_{\mathbf{b}}^0|^2}P(\|w_\vartheta f_0\|_\infty) + \int_{t^1}^{t} \mathcal{N}(s, X^0(s), V^0(s))ds \\
& + \sqrt{\mu(v_{\mathbf{b}}^0)}\langle v_{\mathbf{b}}^0\rangle^2 \int_{\prod_{j=1}^{l-1}\mathcal{V}_j} \sum_{i=1}^{l-1} \mathbf{1}_{\{t^{i+1}<0<t^i\}} |\tilde{\alpha}\partial f(0, X^i(0), V^i(0))|\, d\Sigma_i^{l-1} \\
& + \sqrt{\mu(v_{\mathbf{b}}^0)}\langle v_{\mathbf{b}}^0\rangle^2 \int_{\prod_{j=1}^{l-1}\mathcal{V}_j} \sum_{i=1}^{l-1} \mathbf{1}_{\{t^{i+1}<0<t^i\}} \int_0^{t^i} \mathcal{N}(s, X^i(s), V^i(s))ds\, d\Sigma_i^{l-1} \\
& + \sqrt{\mu(v_{\mathbf{b}}^0)}\langle v_{\mathbf{b}}^0\rangle^2 \int_{\prod_{j=1}^{l-1}\mathcal{V}_j} \sum_{i=1}^{l-1} \mathbf{1}_{\{t^{i+1}>0\}} \int_{t^{i+1}}^{t^i} \mathcal{N}(s, X^i(s), V^i(s))ds\, d\Sigma_i^{l-1} \\
& + \sqrt{\mu(v_{\mathbf{b}}^0)}\langle v_{\mathbf{b}}^0\rangle^2 \int_{\prod_{j=1}^{l-1}\mathcal{V}_j} \sum_{i=2}^{l-1} \mathbf{1}_{\{t^{i}>0\}} e^{-\frac{\vartheta}{2}|v_{\mathbf{b}}^{i-1}|^2} P(\|w_\vartheta f_0\|_\infty)\, d\Sigma_{i-1}^{l-1} \\
& + \sqrt{\mu(v_{\mathbf{b}}^0)}\langle v_{\mathbf{b}}^0\rangle^2 \int_{\prod_{j=1}^{l-1}\mathcal{V}_j} \mathbf{1}_{\{t^{l}>0\}} e^{-\varpi\langle v_{\mathbf{b}}^{l-1}\rangle t^l} \tilde{\alpha}(t^l, x^l, v_{\mathbf{b}}^{l-1})|\partial f(t^l, x^l, v_{\mathbf{b}}^{l-1})| d\Sigma_{l-1}^{l-1},
\end{aligned}
\tag{51}
$$

where $\mathcal{V}_j = \{v^j \in \mathbb{R}^3 : n(x^j)\cdot v^j > 0\}$, and

$$
\begin{aligned}
d\Sigma_i^{l-1} = & \{\prod_{j=i+1}^{l-1} \mu(v^j)c_\mu |n(x^j)\cdot v^j| dv^j\}\{e^{\varpi\langle v^i\rangle t^i}\mu^{1/4}(v^i)\langle v^i\rangle dv^i\} \\
& \{\prod_{j=1}^{i-1} \sqrt{\mu(v_{\mathbf{b}}^j)}\langle v_{\mathbf{b}}^j\rangle \mu^{1/4}(v^j)\langle v^j\rangle e^{\varpi\langle v^j\rangle t^j} dv^j\},
\end{aligned}
$$

where c_μ is the constant that $\int_{\mathbb{R}^3}\mu(v^j)c_\mu|n(x^j)\cdot v^j|dv^j = 1$.

The following lemma is necessary for us to establish Theorem 3.

Lemma 6 *If (F, ϕ) solves (36), write $f = \frac{F}{\sqrt{\mu}}$, then*

$$
\|\phi_F(t)\|_{C^{1,1-\delta}(\Omega)} \lesssim_{\delta,\Omega} \|w_\vartheta f(t)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)}, \text{ for any } 0<\delta<1, \tag{52}
$$

and

$$
\|\nabla^2\phi_F(t)\|_{L^\infty(\Omega)} \lesssim \|w_\vartheta f(t)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)} + \|e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_x f(t)\|_{L^\infty(\bar{\Omega}\times\mathbb{R}^3)}. \tag{53}
$$

Proof It is obvious to have (52) from the Morrey inequality and elliptic estimate. Next we show (53). By Schauder estimate, we have, for $p > 3$ and $\Omega \subset \mathbb{R}^3$,

$$\|\nabla^2 \phi_F(t)\|_{L^\infty(\Omega)} \leq \|\phi_F\|_{C^{2,1-\frac{3}{p}}(\Omega)} \lesssim_{p,\Omega} \| \int_{\mathbb{R}^3} f(t)\sqrt{\mu} dv\|_{C^{0,1-\frac{3}{p}}(\Omega)}.$$

Then by Morrey inequality, $W^{1,p} \subset C^{0,1-\frac{3}{p}}$ with $p > 3$ for a domain $\Omega \subset \mathbb{R}^3$ with a smooth boundary $\partial\Omega$, we derive

$$\begin{aligned} \| \int_{\mathbb{R}^3} f(t)\sqrt{\mu} dv\|_{C^{0,1-\frac{3}{p}}} &\lesssim \| \int_{\mathbb{R}^3} f(t)\sqrt{\mu} dv\|_{W^{1,p}} \\ &\lesssim \|w_\vartheta f(t)\|_\infty + \|e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_x f(t)\|_\infty \| \int_{\mathbb{R}^3} e^{\varpi\langle v\rangle t}\sqrt{\mu}\frac{1}{\tilde{\alpha}} dv\|_{L^p(\Omega)}. \end{aligned}$$

It suffices to show that for some $\beta > 1$,

$$\| \int_{\mathbb{R}^3} e^{-\frac{1}{8}|v|^2} \frac{1}{\tilde{\alpha}^\beta} dv\|_{L^p(\Omega)} < \infty. \tag{54}$$

Since $\tilde{\alpha}$ is bounded from below when x is away from the boundary of Ω, it suffices to only consider the case when x is close enough to $\partial\Omega$. From direct computation (see [5]), we get

$$\int_{\mathbb{R}^3} e^{-\frac{1}{8}|v|^2} \frac{1}{\tilde{\alpha}^\beta} dv \lesssim \frac{1}{(\xi(x)^2 - 2E(t,\bar{x})\cdot\nabla\xi(\bar{x})\xi(x))^{\frac{\beta-1}{2}}} \lesssim \frac{1}{|\xi(x)|^{\frac{\beta-1}{2}}}. \tag{55}$$

And since ξ is C^2, we have

$$\int_{d(x,\partial\Omega)\ll 1} \frac{1}{|\xi(x)|^{\frac{(\beta-1)p}{2}}} dx \lesssim \int_{d(x,\partial\Omega)\ll 1} \frac{1}{|x-\bar{x}|^{\frac{(\beta-1)p}{2}}} dx.$$

Now from (10),

$$\int_{\Omega\cap B(p;\delta_2)} \frac{1}{|x-\bar{x}|^{\frac{(\beta-1)p}{2}}} dx \lesssim \int_{|x_n|<\delta_1} \frac{1}{|x_n|^{\frac{(\beta-1)p}{2}}} d_{x_n} < \infty,$$

if we pick $\beta < \frac{2}{p} + 1$. And since $\partial\Omega$ is compact, we can cover $\partial\Omega$ with finitely many such balls, and therefore we get (54). □

Proof of Theorem 3 For the sake of simplicity we only show the a priori estimate. See [7] for the construction of the sequences of solutions and passing a limit.

The proof of (41) for f satisfying (44), (45), and (46) is standard. We refer to Theorem 4 in [5].

First from (45) and the fact that $\int_{\mathbb{R}^3} \sqrt{\mu} f dv \geq 0$, we apply Lemma 4 to get

$$-\frac{\partial \phi(t,x)}{\partial n} \geq c \iint_{\Omega\times\mathbb{R}^3} \sqrt{\mu} f(t,x,v)\delta(x) dv dx, \tag{56}$$

for some c depending only on Ω.

Denote

$$\iint_{\Omega\times\mathbb{R}^3} F_0(x,v)\delta(x) dv dx = c_{E_0}.$$

Then $\int_0^T \iint_{\Omega\times\mathbb{R}^3} \delta(x) \times (6)\, dv dx dt$ gives

$$\begin{aligned}&\iint_{\Omega\times\mathbb{R}^3} F(T,x,v)\delta(x) dv dx \\ &= \iint_{\Omega\times\mathbb{R}^3} F_0(x,v)\delta(x) dv dx + \int_0^T \iint_{\Omega\times\mathbb{R}^3} Fv\cdot\nabla_x \delta(x) dv dx dt.\end{aligned}$$

Together with (41) and (56) we deduce

$$-\frac{\partial \phi(t,x)}{\partial n} \geq c \iint_{\Omega\times\mathbb{R}^3} F(t,x,v)\delta(x) dv dx > \frac{c c_{E_0}}{2}, \tag{57}$$

as long as $T \lesssim \frac{c_{E_0}}{2M}$.

Next, we investigate (48). Since

$$w_\vartheta \Gamma_{gain}(\partial f, f) \lesssim \|e^{2\vartheta' |v|^2} f\|_\infty \int_{\mathbb{R}^3} \frac{e^{-C_{\vartheta'}|u-v|^2}}{|u-v|^{2-\kappa}} |e^{\vartheta' |u|^2} \partial f(t,x,u)| du,$$

and

$$w_\vartheta \nu(\sqrt{\mu}\partial f) f \lesssim \|e^{2\vartheta' |v|^2} f\|_\infty \int_{\mathbb{R}^3} \frac{e^{-C_{\vartheta'}|u-v|^2}}{|u-v|^{2-\kappa}} |\partial f(t,x,u)| du.$$

Thus from (41) we have the following bound for $\mathcal{N}$:

$$\begin{aligned}|\mathcal{N}(t,x,v)| \lesssim &(1+\|\nabla^2\phi\|_\infty)[P(\|w_\vartheta f_0\|_\infty) + |w_{\vartheta'} e^{-\varpi\langle v\rangle t}\tilde{\alpha}\partial f(t,x,v)|] \\ &+ \|w_\vartheta f_0\|_\infty e^{-\varpi\langle v\rangle t}\tilde{\alpha}(t,x,v) \int_{\mathbb{R}^3} \frac{e^{-C_\vartheta |u-v|^2}}{|u-v|^{2-\kappa}} |e^{\vartheta' |u|^2} \partial f(t,x,u)| du.\end{aligned} \tag{58}$$

Recall the definition of ν_{ϖ} in (47), note that from the velocity lemma (33), and (57) we have

$$\begin{aligned}
&\tilde{\alpha}^{-1}(\partial_t\tilde{\alpha} + v\cdot\nabla_x\tilde{\alpha} - \nabla\phi\cdot\nabla_v\tilde{\alpha})\\
\lesssim&(\|\nabla\phi\|_\infty + \|\nabla^2\phi\|_\infty)\langle v\rangle\\
\lesssim&(\|w_{\vartheta'}f(t)\|_\infty + \|e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_x f(t)\|_\infty)\langle v\rangle\\
\lesssim&(P(\|w_\vartheta f_0\|_\infty) + \|\tilde{\alpha}\partial f_0\|_\infty)\langle v\rangle.
\end{aligned}$$

Therefore we have

$$\nu_{\varpi} \geq \frac{\varpi}{2}\langle v\rangle, \tag{59}$$

once we choose $\varpi \gg 1$ large enough.

For $t^1 < 0$, using the Duhamel's formulation we have from (48)

$$\begin{aligned}
&w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde{\alpha}|\partial f(t,x,v)|\\
\leq&e^{-\int_s^t \nu_\varpi(\tau,X(\tau),V(\tau)d\tau}e^{\vartheta'|V(0)|^2}\tilde{\alpha}\partial f(0,X(0),V(0))\\
&+\int_0^t e^{-\int_s^t \nu_\varpi(\tau,X(\tau),V(\tau)d\tau}\mathcal{N}(s,X(s),V(s))ds.
\end{aligned} \tag{60}$$

Thus by (58) we have

$$\begin{aligned}
&\sup_{0\leq t\leq T}\|\mathbf{1}_{\{t^1<0\}}e^{-\varpi\langle v\rangle t}w_{\vartheta'}\tilde{\alpha}\partial f(t,x,v)\|_\infty\\
&\leq \sup_{0\leq t\leq T}\|e^{-\int_0^t \nu_\varpi(\tau,X(\tau),V(\tau)d\tau}e^{\vartheta'|V(0)|^2}\tilde{\alpha}\partial f(0,X(0),V(0))\\
&\quad+\int_0^t e^{-\int_s^t \nu_\varpi(\tau,X(\tau),V(\tau)d\tau}\mathcal{N}(s,X(s),V(s))ds\|_\infty\\
&\leq\|w_{\vartheta'}\tilde{\alpha}\partial f_0\|_\infty + P(\|w_\vartheta f_0\|_\infty)\sup_{0\leq t\leq T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde{\alpha}\partial f(t,x,v)\|_\infty\\
&\quad+T(1+\|\nabla^2\phi\|_\infty)[P(\|w_\vartheta f_0\|_\infty)+\sup_{0\leq t\leq T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde{\alpha}\partial f(t,x,v)\|_\infty]\\
&\quad\times\int_0^t\int_{\mathbb{R}^3}e^{-\int_s^t\frac{\varpi}{2}\langle V(\tau;t,x,v)\rangle d\tau}\frac{e^{-\varpi\langle(s;t,x,v)\rangle s}}{e^{-\varpi\langle u\rangle s}}\frac{e^{-C_\vartheta|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}}\frac{\tilde{\alpha}(s,X(s),V(s))}{\tilde{\alpha}(s,X(s),u)}duds.
\end{aligned}$$

Now since $\langle u\rangle - \langle V(s;t,x,v)\rangle \le 2\langle u - V(s;t,x,v)\rangle$, we have $\frac{e^{-\varpi\langle (s;t,x,v)\rangle s}}{e^{-\varpi\langle u\rangle s}}$ $e^{-C_\vartheta|V(s)-u|^2} \lesssim e^{-\frac{C_\vartheta|V(s)-u|^2}{2}}$. Thus

$$\begin{aligned}
&\int_0^t\int_{\mathbb{R}^3} e^{-\int_s^t \frac{\varpi}{2}\langle V(\tau;t,x,v)\rangle d\tau}\frac{e^{-\varpi\langle (s;t,x,v)\rangle s}}{e^{-\varpi\langle u\rangle s}}\frac{e^{-C_\vartheta|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}}\frac{\tilde\alpha(s,X(s),V(s))}{\tilde\alpha(s,X(s),u)}duds\\
&\lesssim\int_0^t\int_{\mathbb{R}^3} e^{-\int_s^t \frac{\varpi}{2}\langle V(\tau;t,x,v)\rangle d\tau}\frac{e^{-\frac{C_\vartheta}{2}|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}}\frac{\tilde\alpha(s,X(s),V(s))}{\tilde\alpha(s,X(s),u)}duds.
\end{aligned}\tag{61}$$

Note that, for any $\beta > 1$, $\frac{1}{\tilde\alpha(x,X(s),u)} \lesssim \frac{1}{(\tilde\alpha(x,X(s),u))^\beta} + 1$. So from (57) we can let $1 < \beta \le 2$, and apply the nonlocal-to-local estimate (34)–(61) to have

$$\begin{aligned}
&\int_0^t\int_{\mathbb{R}^3} e^{-\int_s^t \frac{\varpi}{2}\langle V(\tau;t,x,v)\rangle d\tau}\frac{e^{-\varpi\langle (s;t,x,v)\rangle s}}{e^{-\varpi\langle u\rangle s}}\frac{e^{-C_\vartheta|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}}\frac{\tilde\alpha(s,X(s),V(s))}{\tilde\alpha(s,X(s),u)}duds\\
&\lesssim e^{C(\|\nabla\phi\|_\infty^2+\|\nabla^2\phi\|_\infty)}\left(\frac{\delta^{\frac{3-\beta}{2}}(\tilde\alpha(t,x,v))^{3-\beta}}{(|v|^2+1)^{\frac{3-\beta}{2}}}+\frac{(|v|+1)^{\beta-1}(\tilde\alpha(t,x,v))^{2-\beta}}{\delta^{\beta-1}\varpi\langle v\rangle}\right)\\
&\lesssim e^{C(\|\nabla\phi\|_\infty^2+\|\nabla^2\phi\|_\infty)}\left(\delta^{\frac{3-\beta}{2}}+\frac{1}{\delta^{\beta-1}\varpi}\right),
\end{aligned}\tag{62}$$

where we used $\tilde\alpha(s,X(s),V(s)) \lesssim e^{C(\|\nabla\phi\|_\infty^2+\|\nabla^2\phi\|_\infty)}\tilde\alpha(t,x,v)$.

Similarly, for $t^1(t,x,v) \ge 0$, we again apply the nonlocal-to-local estimate (34) to get

$$\begin{aligned}
&|\mathbf{1}_{\{t^1>0\}}w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde\alpha\partial f(t,x,v)|\\
&\lesssim C_l e^{Clt^2}\left(\delta^{\frac{3-\beta}{2}}+\frac{1}{\delta^{\beta-1}\varpi}\right)P(\|w_\vartheta f_0\|_\infty)\max_{0\le i\le l-1}e^{C(\|\nabla\phi\|_\infty^2+\|\nabla^2\phi\|_\infty)}\\
&\quad\times\sup_{0\le t\le T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde\alpha\partial f(t,x,v)\|_\infty\\
&+T(1+\|\nabla^2\phi\|_\infty)\sup_{0\le t\le T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde\alpha\partial f(t,x,v)\|_\infty\\
&+Tl(Ce^{Ct^2})^l(1+\|\nabla^2\phi\|_\infty)\sup_{0\le t\le T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde\alpha\partial f(t,x,v)\|_\infty\\
&+Tl(Ce^{Ct^2})^l(1+\|\nabla^2\phi\|_\infty)P(\|w_\vartheta f_0\|_\infty)+l(Ce^{Ct^2})^l\|\tilde\alpha\partial f_0\|_\infty+P(\|w_\vartheta f_0\|_\infty)\\
&\quad+C\left(\frac{1}{2}\right)^l\sup_{0\le t\le T}\|w_{\vartheta'}e^{-\varpi\langle v\rangle t}\tilde\alpha\partial f(t,x,v)\|_\infty.
\end{aligned}$$

Finally from (53), we can choose a large l then large C then small δ then large ϖ and finally small T to conclude

$$\sup_{0\le t\le T}\|e^{-\varpi\langle v\rangle t}\tilde{\alpha}\partial f(t,x,v)\|_\infty \le \frac{C_1}{2}\left(\|w_\vartheta\tilde{\alpha}\partial f_0\|_\infty + P(\|w_\vartheta f_0\|_\infty)\right)$$

This proves (42).

Next we prove (43). Consider taking ∇_v derivative of (44) and adding the weight function $e^{-\varpi\langle v\rangle t}$, we get

$$\begin{aligned}&[\partial_t + v\cdot\nabla_x - \nabla_x\phi\cdot\nabla_v + \frac{v}{2}\cdot\nabla_x\phi + \varpi\langle v\rangle - \frac{v}{\langle v\rangle}\varpi t\cdot\nabla_x\phi + \nu(\sqrt{\mu}f)](e^{-\varpi\langle v\rangle t}\nabla_v f)\\ =&e^{-\varpi\langle v\rangle t}\left(-\nabla_v\nu(\sqrt{\mu}f)f - \nabla_x f - \frac{1}{2}\nabla_x\phi f + \nabla_v\Gamma_{\text{gain}}(f,f)\right),\end{aligned} \tag{63}$$

with the boundary bound

$$\left|\nabla_v f\right| \lesssim |v|\sqrt{\mu}\int_{n\cdot u>0}|f|\sqrt{\mu}\{n\cdot u\}\mathrm{d}u \quad\text{on}\quad \gamma_-. \tag{64}$$

And

$$\frac{v}{2}\cdot\nabla_x\phi + \varpi\langle v\rangle - \frac{v}{\langle v\rangle}\varpi t\cdot\nabla_x\phi + \nu(\sqrt{\mu}f) > \frac{\varpi}{2}\langle v\rangle,$$

for $\varpi \gg 1$.

Using the Duhamel's formulation, from (63) we obtain the following bound along the characteristics

$$|e^{-\varpi\langle v\rangle t}\nabla_v f(t,x,v)|$$

$$\le \mathbf{1}_{\{t_{\mathbf{b}}(t,x,v)>t\}}e^{-\int_0^t -\frac{C}{2}\langle V(\tau)\rangle d\tau}|\nabla_v f(0,X(0;t,x,v),V(0;t,x,v))| \tag{65}$$

$$+\ \mathbf{1}_{\{t_{\mathbf{b}}(t,x,v)<t\}}e^{-\varpi\langle v_{\mathbf{b}}\rangle t_{\mathbf{b}}}\mu(v_{\mathbf{b}})^{\frac{1}{4}}\int_{n(x_{\mathbf{b}})\cdot u>0}|f(t-t_{\mathbf{b}},x_{\mathbf{b}},u)|\sqrt{\mu}\{n(x_{\mathbf{b}})\cdot u\}\mathrm{d}u \tag{66}$$

$$+\int_{\max\{t-t_{\mathbf{b}},0\}}^t e^{-\int_s^t -\frac{\varpi}{2}\langle V(\tau)\rangle d\tau}e^{-\varpi\langle V(s)\rangle s}|\nabla_x f(s,X(s),V(s))|\mathrm{d}s \tag{67}$$

$$+\int_{\max\{t-t_{\mathbf{b}},0\}}^{t}(1+\|w_{\vartheta'}f\|_\infty)e^{-\int_s^t-\frac{\varpi}{2}\langle V(\tau)\rangle d\tau}e^{-\varpi\langle V(s)\rangle s} \tag{68}$$

$$\times\int_{\mathbb{R}^3}\frac{e^{-C_{\vartheta'}|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}}\nabla_v f(s,X(s),u)|\mathrm{d}u\mathrm{d}s$$

$$+\|w_{\vartheta'}f\|_\infty\int_{\max\{t-t_{\mathbf{b}},0\}}^{t}e^{-\int_s^t-\frac{\varpi}{2}\langle V(\tau)\rangle d\tau}e^{-\varpi\langle V(s)\rangle s}e^{-\vartheta'|V(s)|^2} \tag{69}$$

$$\times|\nabla_x\phi(s,X(s;t,x,v))|\mathrm{d}s.$$

We first have

$$\begin{aligned}
&\|(65)\|_{L^3_xL^{1+\delta}_v}\\
&\lesssim\left(\int_\Omega\left(\int_{\mathbb{R}^3}|e^{\vartheta'|V(0)|^2}\nabla_v f(0,X(0),V(0))|^3\right)\left(\int_{\mathbb{R}^3}e^{-(1+\delta)\frac{3}{2-\delta}\vartheta'|V(0)|^2}\mathrm{d}v\right)^{\frac{2-\delta}{1+\delta}}\right)^{1/3}\\
&\lesssim\left(\iint_{\Omega\times\mathbb{R}^3}|e^{\vartheta'|V(0)|^2}\nabla_v f(0,X(0;t,x,v),V(0;t,x,v))|^3\mathrm{d}v\mathrm{d}x\right)^{1/3}\\
&\lesssim\|w_{\vartheta'}\nabla_v f(0)\|_{L^3_{x,v}},
\end{aligned} \tag{70}$$

where we have used a change of variables $(x,v)\mapsto(X(0;t,x,v),V(0;t,x,v))$.

Clearly

$$\|(66)\|_{L^3_xL^{1+\delta}_v}\lesssim\sup_{0\le s\le t}\|w_{\vartheta'}f(s)\|_\infty. \tag{71}$$

From $\|\nabla_x\phi\|_{L^3}\lesssim\|\phi\|_{W^{2,2}_x}$ for a bounded $\Omega\subset\mathbb{R}^3$, and the change of variables $(x,v)\mapsto(X(s;t,x,v),V(s;t,x,v))$ for fixed $s\in(\max\{t-t_{\mathbf{b}},0\},t)$,

$$\begin{aligned}
\|(69)\|_{L^3_xL^{1+\delta}_v}&\lesssim\|w_{\vartheta'}f\|_\infty\int_{\max\{t-t_{\mathbf{b}},0\}}^{t}\|\phi(s)\|_{W^{2,2}_x}\\
&\lesssim\|w_{\vartheta'}f\|_\infty\int_{\max\{t-t_{\mathbf{b}},0\}}^{t}\|\int_{\mathbb{R}^3}\sqrt{\mu}f(s)dv\|_2.\lesssim t\|w_{\vartheta'}f\|_\infty\|w_{\vartheta'}f\|_\infty.
\end{aligned} \tag{72}$$

Next we have from (55), for $\frac{3\delta}{2(1+\delta)} < 1$, equivalently $0 < \delta < 2$,

$$
\begin{aligned}
\|(67)\|_{L^3_x L^{1+\delta}_v} &\le \left\| \left\| \int_{\max\{t-t_{\mathbf{b}},0\}}^{t} \nabla_x f(s, X(s), V(s)) \mathrm{d}s \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x} \\
&= \left\| \left\| \int_{\max\{t-t_{\mathbf{b}},0\}}^{t} \frac{e^{\vartheta' |V(s)|^2} e^{-\varpi \langle V(s)\rangle s} \tilde{\alpha} \nabla_x f(s, X(s), V(s))}{e^{\vartheta' |V(s)|^2} e^{-\varpi \langle V(s)\rangle s} \tilde{\alpha}} \mathrm{d}s \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x} \\
&\le \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi \langle v\rangle t} \tilde{\alpha} \nabla_x f \right\|_{\infty} \\
&\quad \times \left\| \left\| \int_{\max\{t-t_{\mathbf{b}},0\}}^{t} \frac{e^{-\vartheta' |V(s)|^2} e^{\varpi \langle V(s)\rangle s}}{\tilde{\alpha}(s, X(s), V(s))} \mathrm{d}s \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x} \\
&\lesssim e^{C(\|\nabla\phi\|_\infty + \|\nabla\phi\|_\infty^2 + \|\nabla^2\phi\|_\infty)} \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi \langle v\rangle t} \tilde{\alpha} \nabla_x f \right\|_{\infty} \\
&\quad \times t \int_{\Omega} \left(\int_{\mathbb{R}^3} \frac{e^{-\frac{\vartheta'}{2}|v|^2}}{(\tilde{\alpha}(t, x, v))^{1+\delta}} \mathrm{d}v \right)^{\frac{3}{1+\delta}} \mathrm{d}x \\
&\lesssim t e^{C(\|\nabla\phi\|_\infty^2 + \|\nabla^2\phi\|_\infty)} \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi \langle v\rangle t} \tilde{\alpha} \nabla_x f \right\|_{\infty}.
\end{aligned}
\tag{73}
$$

Next, we consider (68). From the computations in (55), and using the fact that $\frac{1}{\tilde{\alpha}} \lesssim \frac{1}{\tilde{\alpha}^\beta}$, we have

$$
\begin{aligned}
&\|(68)\|_{L^3_x L^{1+\delta}_v} \\
&\le \left\| \left\| \int_{\max\{t-t_{\mathbf{b}},0\}}^{t} e^{-\int_s^t -\frac{\varpi}{2}\langle V(\tau)\rangle d\tau} e^{-\varpi \langle V(s)\rangle s} \right. \right. \\
&\qquad \left. \left. \times \int_{\mathbb{R}^3} \frac{e^{-C_{\vartheta'}|V(s)-u|^2}}{|V(s)-u|^{2-\kappa}} \nabla_v f(s, X(s), u) |\mathrm{d}u \mathrm{d}s \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x} \\
&\lesssim e^{C\|\nabla\phi\|_\infty} \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi \langle v\rangle t} \tilde{\alpha} \nabla_x f \right\|_{\infty} \\
&\quad \times \left\| \left\| \int_{\max\{t-t_{\mathbf{b}},0\}}^{t} e^{-\int_s^t -\frac{\varpi}{2}\langle V(\tau)\rangle d\tau} \int_{\mathbb{R}^3} \frac{e^{-C_{\vartheta'}|V(s)-u|^2}}{|(s)-u|^{2-\kappa}} \frac{e^{-\frac{\vartheta'}{2}|u|^2}}{(\tilde{\alpha}(s, X(s), u))^\beta} \mathrm{d}u \mathrm{d}s \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x}.
\end{aligned}
\tag{74}
$$

And then applying the nonlocal-to-local estimate (34) to (74) , we conclude

$$
\begin{aligned}
\|(68)\|_{L^3_x L^{1+\delta}_v} \lesssim & e^{C(\|\nabla\phi\|_\infty+\|\nabla^2\phi\|_\infty)} \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_x f \right\|_\infty \\
& \times \left\| \left\| \frac{\delta^{\frac{3-\beta}{2}}}{(\tilde{\alpha}(t,x,v))^{\beta-2}(|v|^2+1)^{\frac{3-\beta}{2}}} + \frac{(|v|+1)^{\beta-1}}{\delta^{\beta-1}\varpi\langle v\rangle(\tilde{\alpha}(t,x,v))^{\beta-1}} \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x} \\
\lesssim & e^{C(\|\nabla\phi\|_\infty+\|\nabla^2\phi\|_\infty)} \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_x f \right\|_\infty \\
& \times \left(O(\delta^{\frac{3-\beta}{2}}) + \frac{1}{\delta^{\beta-1}\varpi} \left\| \left\| \frac{1}{\langle v\rangle^{2-\beta}(\tilde{\alpha}(t,x,v))^{\beta-1}} \right\|_{L^{1+\delta}_v(\mathbb{R}^3)} \right\|_{L^3_x} \right) \\
\lesssim & C(\delta^{\frac{3-\beta}{2}} + \frac{1}{\delta^{\beta-1}\varpi}) e^{C(\|\nabla\phi\|_\infty+\|\nabla^2\phi\|_\infty)} \sup_{0\le t\le T} \left\| w_{\vartheta'} e^{-\varpi\langle v\rangle t}\tilde{\alpha}\nabla_x f \right\|_\infty ,
\end{aligned}
\tag{75}
$$

for β satisfies $\frac{(\beta-1)(1+\delta)-1}{2}\frac{3}{1+\delta} < 1$, which is equivalent to $\beta < \frac{5}{3} + \frac{1}{1+\delta}$. Therefore any $1 < \beta < \frac{5}{3}$ would work.

Collecting terms from (65)–(69), and (70), (71), (72), (73), (75), we derive

$$
\begin{aligned}
& \sup_{0\le s\le t} \|e^{-\varpi\langle v\rangle t}\nabla_v f(s)\|_{L^3_x L^{1+\delta}_v} \\
\lesssim & \|w_{\vartheta'}\nabla_v f(0)\|_{L^3_{x,v}} + \|w_{\vartheta'} f\|_\infty)^2 + \|w_{\vartheta'} f\|_\infty \\
< & \infty.
\end{aligned}
\tag{76}
$$

This proves (43) and conclude Theorem 3. □

Acknowledgments This work was supported in part by National Science Foundation under Grant No. 1501031, Grant No. 1900923, NRF-2021H1D3A2A01039047 (Brain Pool Program of the Ministry of Science and ICT of Korea), and the Wisconsin Alumni Research Foundation.

References

1. Brezis, H., Cabré, X.: Some simple nonlinear PDE's without solutions. Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8) **1**(2), 223–262 (1998)
2. Cercignani, C., Illner, R., Pulvirenti, M.: The mathematical theory of dilute gases. In: Applied Mathematical Sciences, vol. 106, viii+347 pp. Springer, New York (1994)
3. Cercignani, C., Lampis, M.: Kinetic models for gas-surface interactions. Transp. Theory Stat. Phys. **1**(2), 101–114 (1971)
4. Cao, Y., Kim, C., Lee, D.: Global strong solutions of the Vlasov–Poisson–Boltzmann system in bounded domains. Arch. Ration. Mech. Anal. **233**(3), 1027–1130 (2019)

5. Cao, Y.: Regularity of Boltzmann equation with external fields in convex domains of diffuse reflection. SIAM J. Math. Anal. **51**(4), 3195–3275 (2019)
6. Cao, Y.: A note on two species collisional plasma in bounded domains. Kinet. Relat. Models **12**(6), 1359–1429 (2019)
7. Chen, H., Kim, C., Li, Q.: Local well-posedness of Vlasov–Poisson–Boltzmann equation with generalized diffuse boundary condition. J. Stat. Phys. **179**(2), 535–631 (2020)
8. Devillettes, L., Villani, C.: On the trend to global equilibrium for spatially inhomogeneous kinetic systems: the Boltzmann equation. Invent. Math. **159**(2), 245–316 (2005)
9. Esposito, R., Guo, Y., Kim, C., Marra, R.: Non-Isothermal Boundary in the Boltzmann Theory and Fourier Law. Commun. Math. Phys. **323**(1), 177–239 (2003)
10. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence (1998)
11. Guo, Y.: The Vlasov–Poisson–Boltzmann system near Maxwellians. Commun. Pure Appl. Math. **55**(9), 1104–1135 (2002)
12. Guo, Y.: Decay and continuity of Boltzmann equation in bounded domains. Arch. Ration. Mech. Anal. **197**(3), 713–809 (2010)
13. Guo, Y., Kim, C., Tonon, D., Trescases, A.: Regularity of the Boltzmann equation in convex domains. Invent. Math. **207**(1), 115–290 (2017)
14. Glassey, R.: The Cauchy Problems in Kinetic Theory. SIAM, Philadelphia (1996)
15. Kim, C.: Boltzmann equation with a large potential in a periodic box. Commun. PDE. **39**(8), 1393–1423 (2014)
16. Kim, C.: Formation and propagation of discontinuity for Boltzmann equation in non-convex domains. Commun. Math. Phys. **308**(3), 641–701 (2011)
17. Kim, C., Lee, D.: The Boltzmann equation with specular boundary condition in convex domains. Commun. Pure Appl. Math. **71**(3), 411–504 (2018)
18. Mischler, S.: On the initial boundary value problem for the Vlasov–Poisson–Boltzmann system. Commun. Math. Phys. **210**, 447–466 (2000)

The Vlasov Equation with Infinite Mass

Guido Cavallaro

Abstract We discuss about the initial value problem for the Vlasov equation in case of unbounded total mass. The problem strongly depends on the dimension d of the physical space and on the singularity of the interaction. In particular, for $d = 3$, the more singular the interaction, the faster must be the spatial decay at infinity of the initial distribution. We describe also an application which gives rise to a viscous friction model.

1 Introduction

We want to review some known results, and to address some open problems, related to the initial value problem for the Vlasov equation when it describes the time evolution of a plasma in an unbounded domain with an unbounded mass distribution. The problem is not trivial, since it is not easy in this case to exclude a priori a blow-up of the mass distribution in a finite time. We summarize some results on the Vlasov equation, starting from finite mass systems.

A *Vlasov system* is described by a function $f(\mathbf{x}, \mathbf{v}, t)$ which gives the density of mass at time t in the point $(\mathbf{x}, \mathbf{v})$ of the one-particle phase space. A general way to write the equation governing its time evolution is based on the characteristics. More precisely, we look for a pairs of functions,

$$(\mathbf{x}, \mathbf{v}) \to (\mathbf{X}(\mathbf{x}, \mathbf{v}; t), \mathbf{V}(\mathbf{x}, \mathbf{v}; t)) \ , \qquad f_0(\mathbf{x}, \mathbf{v}) \to f(\mathbf{x}, \mathbf{v}; t) \ , \tag{1}$$

G. Cavallaro (✉)
Dipartimento di Matematica, Sapienza Università di Roma, Roma, Italy
e-mail: cavallar@mat.uniroma1.it

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_5

where $(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^d \times \mathbb{R}^d$, $t \in \mathbb{R}$, are solution to

$$\begin{cases} \dot{\mathbf{X}}(\mathbf{x}, \mathbf{v}; t) = \mathbf{V}(\mathbf{x}, \mathbf{v}; t) \ , \\ \dot{\mathbf{V}}(\mathbf{x}, \mathbf{v}; t) = \mathbf{F}(\mathbf{X}(\mathbf{x}, \mathbf{v}; t), t) \ , \\ (\mathbf{X}(\mathbf{x}, \mathbf{v}; 0), \mathbf{V}(\mathbf{x}, \mathbf{v}; 0)) = (\mathbf{x}, \mathbf{v}) \ , \\ f\ (\mathbf{X}(\mathbf{x}, \mathbf{v}; t), \mathbf{V}(\mathbf{x}, \mathbf{v}; t); t) = f_0(\mathbf{x}, \mathbf{v}) \ , \end{cases} \tag{2}$$

where

$$\mathbf{F}(\mathbf{x}, t) = -\int_{\mathbb{R}^d} \mathrm{d}\mathbf{y}\ \nabla \Phi\ (\mathbf{x} - \mathbf{y})\ \rho(\mathbf{y}, t) \ , \tag{3}$$

$$\rho(\mathbf{x}, t) = \int_{\mathbb{R}^d} \mathrm{d}\mathbf{v}\ f(\mathbf{x}, \mathbf{v}; t) \tag{4}$$

is the density of mass, and Φ is the interaction potential between two particles.

In the case of smooth initial data, the condition $(2)_4$ implies that $f(\mathbf{x}, \mathbf{v}; t)$ satisfies the (transport) differential equation,

$$(\partial_t + \mathbf{v} \cdot \nabla_{\mathbf{x}} + F(\mathbf{x}, t) \cdot \nabla_{\mathbf{v}})\ f(\mathbf{x}, \mathbf{v}; t) = 0 \ . \tag{5}$$

As the phase space velocity $(\mathbf{V}, \mathbf{F}(\mathbf{X}, t))$ has zero divergence, by Liouville's Theorem the Jacobian $|J(\mathbf{X}, \mathbf{V}|\mathbf{x}, \mathbf{v})|$ is equal one and the Lebesgue measure $\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{v}$ is thus conserved along the motion. This implies that $(2)_4$ (i.e., (5) for smooth initial data) corresponds to the conservation of mass.

The total mass of the gas is defined as the integral in the whole space of the density of mass,

$$M_{\text{total}} = \int_{\mathbb{R}^d} \mathrm{d}\mathbf{x}\ \rho(\mathbf{x}, t) \ .$$

For finite total mass and smooth mutual interaction Φ, the existence and uniqueness of the solution is simple and follows from standard methods. Singular interactions are more delicate to deal, especially the physically relevant case of Coulomb interaction, where (5) is known as the Vlasov-Poisson equation. The problem has been solved both for the attractive (gravitational) case and for the repulsive (plasma) case [22, 23, 26, 27, 32–34, 38–40]. (See also [17, 28–30]).

However we deal with unbounded masses, and on this topic there are fewer results. The case of smooth interaction in three dimensions is discussed in [4], where the authors assume positive interaction, but with some technical effort it can be extended to superstable interactions, as done in the case of the point particles dynamics [5]. In dimension $d = 2$, the Coulomb interaction gives rise to the so called Vlasov-Helmholtz equation, which is discussed in [19]. More recently, the case in which in dimension $d = 2$ a point body interacts via a Coulomb potential

with a plasma with a charge of the same sign has been treated in [17]. Perhaps, it is possible to study also the three dimensional case, but with a cylindrical symmetry and the body moving along the symmetry axis. The general problem in dimension $d = 3$ with singular interaction is open (and we believe very hard).

All these results, at least in dimension $d > 1$, concern the well-posedness of the problems but give no information on the long time behavior of the solutions. For the case of unbounded plasma see also [24, 31, 35–37], and when an external magnetic field acting on the plasma is present see [6–8, 10, 12, 13, 15].

We remark that in [4] the authors assume an initial distribution of the form

$$0 \leq f_0(\mathbf{x}, \mathbf{v}) \leq C_0 \mathrm{e}^{-\lambda|\mathbf{v}|^2}$$

in the whole space $\mathbb{R}^3$. This means that it is sufficient to take an exponential decay in the velocities, and a spatial distribution simply bounded by a constant, to obtain existence and uniqueness of the solution. Whenever the mutual interaction potential is singular at the origin, it is necessary to introduce a suitable spatial decay in the initial datum to obtain an analogous result [9, 14, 16]. This holds in the whole space $\mathbb{R}^3$. In special unbounded domains contained in $\mathbb{R}^3$ it is possible to consider an initial spatial density simply bounded by a constant: in [7] it is considered the evolution in an infinite cylinder when the potential has a coulomb singularity. A singularity of the form $1/r^\alpha$, $\alpha > 0$, and a spatial density $\rho(\mathbf{x}, 0) \leq const$ in the whole space $\mathbb{R}^3$ is an open problem. It would be also interesting (but not explicitly done) to investigate, in $\mathbb{R}^3$, the relation between the exponent α of the singularity of the potential, and the exponent β of the decay of the spatial density at infinity

$$\rho(\mathbf{x}, 0) \leq \frac{const}{(1 + |\mathbf{x}|)^\beta}$$

which is necessary for the well posedness of the problem.

An interesting application which takes inspiration from the previous results is the motion of a heavy body immersed in a Vlasov gas and interacting with it. The resulting coupled dynamics of the gas and the body is very hard to be studied in general, as mentioned before, and some assumptions and approximations are necessary in order to tackle the problem (see for instance [1]). In Sect. 3 we deal with this model. In the next one we give some hints on the initial value problem for the Vlasov equation with infinite mass.

2 Initial Value Problem

We give here an overview of the initial value problem for the Vlasov equation when it describes the time evolution of a plasma distributed in the whole space $\mathbb{R}^d$ and with infinite total mass. As for point particle systems [5], this problem is not trivial since it is not easy to exclude *a priori* the blow-up of the mass distribution

in a finite time. There are many studies on the Vlasov equations, here we focus our attention on the difficulty related to the assumption of infinite total mass. In order to separate the difficulties, we assume the interaction is positive, smooth, and short-range. In analogy to the case of point particle systems, we believe that the positiveness and short-range assumptions can be relaxed by assuming that the interaction is superstable and satisfies some decaying property at large distance. But this task needs a not trivial effort and it has not been done. The case of singular interaction (the Coulomb interaction being the most interesting one) is discussed later on.

The difficulty of the problem grows with the dimension of the physical space. We start with an heuristic consideration, which shows the importance of the physical space dimension in this framework.

Consider the Vlasov equation (2) and assume $\Phi = \Phi(|\mathbf{x}|)$ to be a non-negative function such that

$$\Phi \in C^2(\mathbb{R}) , \quad \Phi(0) > 0 , \quad \Phi(|\mathbf{x}|) = 0 \quad \text{if} \quad |\mathbf{x}| > r \quad (r > 0) . \tag{6}$$

Moreover, we assume that the initial distribution f_0 satisfies

$$0 \le f_0(\mathbf{x}, \mathbf{v}) \le C_0 \mathrm{e}^{-\lambda |\mathbf{v}|^2} \quad (C_0, \lambda > 0) . \tag{7}$$

We remark that we really need to postulate some decay in the velocity variable as shown by the following example. Consider the free evolution in one dimension of an initial datum $f_0(x, v)$ which is the characteristic function of the set $\{(x, v) \colon x > 0, \ -(x+1) < v < -x\}$. Therefore, the initial density of mass is equal to zero for $x \le 0$ and to one for $x > 0$. It is clear that for $t = 1$ we have a blow-up of the density.

The main issue in proving the existence of solutions is to show that the force $\mathbf{F}(\mathbf{x}, t)$ acting on the element of fluid located in $\mathbf{x} \in \mathbb{R}^d$ is bounded. By (3) and (4) we have,

$$|\mathbf{F}(\mathbf{x}, t)| \le \|\nabla \Phi\|_\infty \int_{B(\mathbf{x}, r)} \mathrm{d}\mathbf{y} \ \rho(\mathbf{y}, t) = \|\nabla \Phi\|_\infty \ m(B(\mathbf{x}, r), t) , \tag{8}$$

where $B(\mathbf{x}, r)$ in an open ball around $\mathbf{x}$ of radius r, $m(B(\mathbf{x}, r), t)$ is the mass contained in such a ball at time t and r is defined in (6). To simplify the situation we first assume that

$$f_0(\mathbf{x}, \mathbf{v}) \le C_0 \ \chi(|\mathbf{v}| < \widehat{V}_0) \tag{9}$$

where $\chi(\cdot)$ is the characteristic function. Letting

$$\widehat{V}(t) = \sup_{0 \le s \le t} \sup_{\mathbf{x}, \mathbf{v}} |\mathbf{V}(\mathbf{x}, \mathbf{v}, s)| ,$$

we have, for any $\mathbf{a} \in \mathbb{R}^d$,

$$m(B(\mathbf{a}, r), t) = \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{v}\; f_0(\mathbf{x}, \mathbf{v})\; \chi(\mathbf{X}(\mathbf{x}, \mathbf{v}; t) \in B(\mathbf{a}, r))$$
$$\leq \|f_0\|_{L^\infty}\; \widehat{V}_0^d\; [r + \widehat{V}(t)t]^d\ .$$

The last inequality follows from the fact that $\chi(\mathbf{X}(\mathbf{x}, \mathbf{v}; t) \in B(\mathbf{a}, r)) = 0$ if $|\mathbf{x} - \mathbf{a}|$ is larger than $r + \widehat{V}(t)t$. On the other hand,

$$\mathbf{V}(\mathbf{x}, \mathbf{v}; t) = \mathbf{v} + \int_0^t \mathrm{d}s\; \mathbf{F}(\mathbf{X}(\mathbf{x}, \mathbf{v}; s), s)\ ,$$

which gives

$$\widehat{V}(t) \leq \widehat{V}_0 + \|\nabla \Phi\|_\infty\; \|f_0\|_\infty\; \widehat{V}_0^d \int_0^t \mathrm{d}s\; [r + \widehat{V}(s)s]^d\ . \tag{10}$$

The above inequality is solvable globally in time only if $d = 1$. We remark that, as for the particle systems, a rigorous proof where the assumption (9) is relaxed requires some care. In dimension $d > 1$, other tools are needed besides the naive use of mass conservation. More precisely, to control the maximal velocity $\widehat{V}$ a deep use of energy conservation is needed in $d = 2$, while for $d = 3$ suitable time averages have to been used. We discuss directly the more difficult case, i.e., the three dimensional one.

For a given function $f(\mathbf{x}, \mathbf{v})$ and any couple $(\mu, R) \in (\mathbb{R}^3 \times \mathbb{R}^+)$ we introduce a sort of "smoothed energy" of a ball of center μ and radius R,

$$W(f; \mu, R) = \frac{1}{2} \int \mathrm{d}\mathbf{x}\; g^{\mu,R}(\mathbf{x}) \left[\int \mathrm{d}\mathbf{v}\; |\mathbf{v}|^2 f(\mathbf{x}, \mathbf{v}) + \rho(\mathbf{x}) \int \mathrm{d}\mathbf{y}\; \rho(\mathbf{y})\; \Phi(|\mathbf{x} - \mathbf{y}|) \right]\ ,$$

where $g^{\mu,R}$ is a smoothing function defined as

$$g^{\mu,R}(\mathbf{x}) = g\left(\frac{|\mathbf{x} - \mu|}{R} \right)\ ,$$

with $g \in C^\infty(\mathbb{R}^+)$ such that

$$g(\eta) = 1 \quad \text{if} \quad \eta \in [0, 1]\ , \quad g(\eta) = 0 \quad \text{if} \quad \eta \in [2, \infty)\ , \quad -2 \leq g'(\eta) \leq 0\ .$$

For the positivity of the potential Φ, W is a well-defined positive functional for any f satisfying (7). Moreover, it is straightforward to see that there exists a positive

constant C_1 such that

$$\sup_{(\mu,R)\in\mathbb{R}^3\times\mathbb{R}^+} \frac{W(f;\mu,R)}{R^3} \leq C_1 \ .$$

The following theorem is proved in [4].

Theorem 1 *Let f_0 satisfy (7). Then, there exists a pair of functions*

$$(\mathbf{x},\mathbf{v}) \to (\mathbf{X}(\mathbf{x},\mathbf{v};t),\mathbf{V}(\mathbf{x},\mathbf{v};t)) \ , \quad f_0(\mathbf{x},\mathbf{v}) \to f(\mathbf{x},\mathbf{v};t) \ ,$$
$$(\mathbf{x},\mathbf{v},t) \in \mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}^+ \ ,$$

satisfying the Vlasov equations (2). This is the unique solution in the class of functions $f(t) = f(\cdot,\cdot;t)$ such that

$$\sup_{t\in[0,T]} \sup_{(\mu,R)\in\mathbb{R}^3\times\mathbb{R}^+} \frac{W(f(t);\mu,R)}{R^3} < \infty \quad \forall\, T > 0 \ .$$

Moreover, for each $\lambda_1 < \lambda$ and $T > 0$ there exists $C_2 > 0$ such that

$$f(\mathbf{x},\mathbf{v};t) \leq C_2 \mathrm{e}^{-\lambda_1|\mathbf{v}|^2} \quad \forall\, t \in [0,T] \ .$$

The proof is obtained in analogy with the case of point particle systems in three dimensions [5]. First, we introduce a partial dynamics with a cut-of on the positions and the velocities, i.e., we introduce the sequence of problems,

$$\dot{\mathbf{X}}^{M,N}(\mathbf{x},\mathbf{v};t) = \mathbf{V}^{M,N}(\mathbf{x},\mathbf{v};t) \ , \quad \dot{\mathbf{V}}^{M,N}(\mathbf{x},\mathbf{v};t) = \mathbf{F}^{M,N}(\mathbf{X}(\mathbf{x},\mathbf{v},t),t) \ ,$$
$$\mathbf{X}^{M,N}(\mathbf{x},\mathbf{v},0) = \mathbf{x} \ , \quad \mathbf{V}^{M,N}(\mathbf{x},\mathbf{v},0) = \mathbf{v} \ , \quad |\mathbf{x}| \leq M \ , \quad |\mathbf{v}| \leq N \ ,$$

where M, N are positive integers,

$$\mathbf{F}^{M,N}(\mathbf{x},t) = \int \mathrm{d}\mathbf{y}\ \nabla\Phi(|\mathbf{x}-\mathbf{y}|) \int \mathrm{d}\mathbf{v}\ f^{M,N}(\mathbf{x},\mathbf{v};t) \ ,$$

$$f^{M,N}(\mathbf{X}^{M,N}(\mathbf{x},\mathbf{v};t),\mathbf{V}^{M,N}(\mathbf{x},\mathbf{v},t),t) = f_0^{M,N}(\mathbf{x},\mathbf{v}) \ ,$$

and

$$f_0^{M,N}(\mathbf{x},\mathbf{v}) = f_0(\mathbf{x},\mathbf{v})\ \chi(|\mathbf{x}| \leq M)\ \chi(|\mathbf{v}| \leq N) \ .$$

The above problem is a well posed Vlasov evolution with finite mass, which admits an unique positive solution $f^{M,N}(\mathbf{x},\mathbf{v};t)$ (see for instance [20] and the references quoted in).

We next investigate the limit $M, N \to \infty$. We introduce the quantity,

$$\widehat{V}^{M,N}(t) = \sup_{0 \le s \le t} \sup_{\mathbf{x},\mathbf{v}} |\mathbf{V}^{M,N}(\mathbf{x}, \mathbf{v}, s)| \, .$$

We can prove, after many efforts, that for each $T > 0$ there exists a positive constant C such that

$$V^{M,N}(T) \le CN.$$

The last step is to remove the cut-off by means of an iterative procedure, under the hypothesis that $M = N^{\alpha}$, with $\alpha > 1$ to be fixed conveniently. This is the idea of the proof, that develops through complicated steps. We forward to the original paper [4].

In the proof the smoothness of the interaction plays an essential role. In fact, in this case the only way to obtain a large growth of the velocity in a point is to crowd a lot of mass in a point. In this case, the superstability condition imposes a large energy, that in its turn controls the maximal velocity. This proof fails in three dimensions. In two dimensions indeed it is possible to do it if the interaction is not too singular [19], while in a three dimensional domain which is unbounded in one direction only the Vlasov equation can be studied for interactions with a singularity almost Coulomb-like [11] (and Coulomb-like when the initial velocities are bounded by a constant [7]). Another direction of (not trivial) generalization is to consider also long range interactions.

Some results have been obtained in this direction in the physically relevant case of the so-called Vlasov-Helmholtz equation, where the interaction at short distance behaves as the Coulomb one and decays exponentially at large distances by a screening effect [19].

It is interesting to consider also situations where point particles coexists with a Vlasov fluid. Of course, the Coulomb interaction plays a privileged role because of its physical importance. Some results have been obtained for localized Vlasov fluid that we do not quote here, but only one: a two-dimensional system composed by a point charge particle that interacts with an unbounded Vlasov fluid with charges of the same sign. The interaction behaves at short distance as the Coulomb one and it is exponentially decreasing at large distances [18].

We mention that in this direction it would be interesting the study of the following case: a Vlasov gas in three dimensions with a cylindrical symmetry and a point particle moving along the symmetry axis. Of course, it would be a model of viscous friction. We will talk about this model in the next section. A study of the long time behavior is too hard, but at least the existence of the infinite dynamics, or the existence of stationary states, seem to be approachable issues.

3 A Viscous Friction Model

An interesting application of the previous ideas and results consists in considering the motion of a heavy point body in a Vlasov fluid, experiencing a drag force which slows down its motion. Such force should derive by the interaction between the body and the fluid particles. The stronger is this interaction, the stronger we expect the drag force exerted by the fluid over the body. We analyse a simple schematic model in which such conjecture can be supported rigorously, and which can give some hints on more realistic models. The more drastic assumption is to consider a *non* self-interacting Vlasov fluid.

We consider a point body of mass $M = 1$ under the action of a constant force $\mathbf{E}$ of intensity E and directed along the x_1-axis, i.e., $\mathbf{E} = (E, 0, 0)$. The body is immersed in a gas of free particles (see below), which interacts with the body via a force of pair potential $\Psi(|\mathbf{r}|)$, $\mathbf{r} \in \mathbb{R}^3$. We assume that $\Psi(r)$ is a twice differentiable function for $r > 0$, and that there exist two positive constants $r_1 < r_0 < \infty$ such that $\Psi(r) = g\, r^{-\alpha}$ for $r < r_1$, with $g, \alpha > 0$, $\Psi(r)$ is a decreasing function for $r_1 \le r \le r_0$, and $\Psi(r) = 0$ for $r > r_0$. Here, we assume Ψ singular at the origin, but of course we can also consider the case in which Ψ is bounded everywhere. The analysis of this case is simpler and it is essentially contained in the present one.

We assume the medium to be a three-dimensional Vlasov system of free particles in the mean field approximation (Knudsen gas), namely the pairs of functions

$$(\mathbf{x}, \mathbf{v}) \to (\mathbf{X}(\mathbf{x}, \mathbf{v}; t), \mathbf{V}(\mathbf{x}, \mathbf{v}; t)) , \qquad f_0(\mathbf{x}, \mathbf{v}) \to f(\mathbf{x}, \mathbf{v}; t) ,$$

where $(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^3 \times \mathbb{R}^3$, $t \in \mathbb{R}$, solution to

$$\begin{cases} \dot{\mathbf{X}}(\mathbf{x}, \mathbf{v}; t) = \mathbf{V}(\mathbf{x}, \mathbf{v}; t) , \\ \dot{\mathbf{V}}(\mathbf{x}, \mathbf{v}; t) = -\nabla_{\mathbf{X}} \Psi\, (|\mathbf{X}(\mathbf{x}, \mathbf{v}; t) - \mathbf{r}(t)|) , \\ (\mathbf{X}(\mathbf{x}, \mathbf{v}; 0), \mathbf{V}(\mathbf{x}, \mathbf{v}; 0)) = (\mathbf{x}, \mathbf{v}) , \\ f\, (\mathbf{X}(\mathbf{x}, \mathbf{v}; t), \mathbf{V}(\mathbf{x}, \mathbf{v}; t); t) = f_0(\mathbf{x}, \mathbf{v}) , \end{cases} \tag{11}$$

where $\mathbf{r}(t) \in \mathbb{R}^3$ is the trajectory of the point body.

In the case of smooth initial data $f(\mathbf{x}, \mathbf{v}; t)$ satisfies the differential equation

$$\big(\partial_t + \mathbf{v} \cdot \nabla_{\mathbf{x}} - \nabla_{\mathbf{x}} \Psi\, (|\mathbf{x} - \mathbf{r}(t)|) \cdot \nabla_{\mathbf{v}}\big) f(\mathbf{x}, \mathbf{v}; t) = 0 , \tag{12}$$

which is coupled to the equation of motion of the body,

$$\ddot{\mathbf{r}}(t) = \mathbf{E} - \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{v}\, \nabla_{\mathbf{r}} \Psi\, (|\mathbf{r}(t) - \mathbf{x}|)\; f(\mathbf{x}, \mathbf{v}; t) , \tag{13}$$

with initial conditions

$$\mathbf{r}(0) = \mathbf{r}_0 , \qquad \dot{\mathbf{r}}(0) = \dot{\mathbf{r}}_0 . \tag{14}$$

This model, in which the medium is coupled with a massive body, has been introduced in connection with the so-called piston problem, see [21, 25] and references therein. In our problem, the medium is unbounded and so in principle the existence of the solution is not obvious. In fact, it is easy to exhibit initial conditions for which the time evolution produces singularity in the motion after a finite time. This happens because very far away particles could arrive very fastly close to the body. These situations are pathological from a physical point of view, and they are removed by the assumption that, initially,

$$f_0(\mathbf{x}, \mathbf{v}) \leq \rho \left(\frac{\beta}{\pi}\right)^{3/2} \mathrm{e}^{-\beta|\mathbf{v}|^2} , \qquad \beta = (kT)^{-1} , \quad \rho > 0 , \tag{15}$$

where T is the temperature and k the Boltzmann constant. Hence, f is bounded by a homogeneous Maxwellian distribution.

We now look for a particular steady state. We assume that the body has a constant velocity $V > 0$, and by a Galilean transformation we consider a reference system in which the body is at rest. Then, in this reference system we assume $f_0(\mathbf{x}, \mathbf{v}) \equiv \hat{f}$, given by a scattering state with incoming particles having velocity $(-V, 0, 0)$ and constant density, which produce a friction force on the body. By construction this state (if it exists) is stationary. In this set up it can be proved the following theorem.

Theorem 2 *Fix $E > 0$. In the limit $V \to \infty$, the friction force tends to $-\infty$ for $\alpha > 2$, to a constant for $\alpha = 2$, to zero for $0 < \alpha < 2$ and for any bounded interaction.*

As a consequence, whatever intensity $E > 0$ is considered, in the case $\alpha > 2$ large enough values of the velocity V produce a friction force opposite to E with an intensity larger than E, and so for some value of V there is a stationary state $\hat{f}$.

On the contrary, if $\alpha \leq 2$ or if the interaction is bounded, for sufficiently large value of E and any value of V, the friction force has an intensity smaller than E, and hence a stationary state $\hat{f}$ cannot exist.

Proof The proof is elementary and we only sketch it. We assume spherical coordinates (r, θ, ϕ) with the axis corresponding to $\theta = 0$ in the x_1-direction. The problem has an axial symmetry and, for the moment, we put $\phi = 0$. We study the motion of a particle starting at time $-\infty$ from the point $(r, \theta) = (\infty, 0)$, with velocity $-V$ directed along the x_1-axis and an impact parameter s. After a scattering the particle escapes at $r = \infty$ with an angle $\theta = \theta_f$ and the same absolute value of the velocity. The energy and angular momentum conservations determine the motion of the particle,

$$\dot{\theta} r^2 = sV = \text{const} , \tag{16}$$

$$\frac{1}{2}\left(\dot{r}^2 + \dot{\theta}^2 r^2\right) + \Psi(r) = \frac{1}{2} V^2 . \tag{17}$$

Performing the explicit calculation (for $\alpha > 0$) we have,

$$\theta_f = \int_{r_{\min}}^{r_1} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1-2V^{-2}g\,r^{-\alpha}-(s/r)^2}} + 2\arcsin\left(\frac{s}{r_0}\right) + \int_{r_1}^{r_0} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1-2V^{-2}\Psi(r)-(s/r)^2}}\,, \tag{18}$$

where $r_{\min}$ is the value of r for which the square root in the first integral vanishes.

The momentum transferred from the particle to the body is equal to the difference of the x_1-component of the initial and final velocities,

$$\Delta p = [\text{lost momentum by particle}] = V(1+\cos\theta_f)\,. \tag{19}$$

The flux is proportional to the incoming velocity, $[\text{Flux}] = CV$, $C < 0$. Hence, by integration over all particles with impact parameter $s \le r_0$,

$$F = [\text{total friction force}] = 2\pi\, C \int_0^{Vr_0} \mathrm{d}z\; z\,(1+\cos\theta_f)\,, \tag{20}$$

where $z = Vs$.

Our aim is to show that for $\alpha > 2$, $|F| \to \infty$ for $V \to \infty$, whereas for $\alpha \in [0, 2]$ $|F|$ converges to zero or remains bounded in the same limit. Since the interaction is repulsive, $\theta_f \in [0, \pi]$, range in which the cosine is decreasing with respect to its argument, so that for $\alpha > 2$ we must increase θ_f to decrease $|F|$, and, viceversa, for $\alpha \in [0, 2]$ we must decrease θ_f to increase $|F|$.

We start with the case $\alpha = 2$, and by it we study the other cases.

(i) $\alpha = 2$

We can do an explicit calculation and we obtain:

$$\theta_f = \frac{2z}{\sqrt{z^2+2g}}\left[\frac{\pi}{2} - \arcsin\left(\frac{s}{r_1}\sqrt{1+\frac{2g}{z^2}}\right)\right] + 2\arcsin\left(\frac{s}{r_0}\right) + 2\int_{r_1}^{r_0} \frac{s\,\mathrm{d}r}{r^2\sqrt{1-2V^{-2}\Psi(r)-(s/r)^2}}\,, \tag{21}$$

and hence

$$\lim_{V\to\infty} |F| = 2\pi|C| \int_0^{\infty} \mathrm{d}z\; z\,\left(1+\cos\theta_f\right) < \infty\,. \tag{22}$$

(We remark that the integral in the previous formula is bounded since the integrand for large values of z behaves like z^{-3}).

Since the friction force vanishes for $V = 0$, is bounded at infinity and it is continuous, it is always bounded. It is enough to take E larger than the maximum of $|F|$ to forbid a stationary state $\hat{f}$.

(ii) $0 < \alpha < 2$

We want to obtain integrals similar to those of the previous case. For this purpose we change the term $2gV^{-2}r^{-\alpha}$ in the first integral of Eq. (18) into as^2r^{-2}, where a is determined by the relation:

$$a = 2gV^{-2}s^{-2}(r_{\min})^{2-\alpha} \ . \tag{23}$$

By this operation we decrease θ_f and we have

$$\begin{aligned} \theta_f > & \int_{r_{\min}}^{r_1} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1 - as^2r^{-2} - (s/r)^2}} + 2\arcsin\left(\frac{s}{r_0}\right) \\ & + \int_{r_1}^{r_0} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1 - 2V^{-2}\Psi(r) - (s/r)^2}} \ . \end{aligned} \tag{24}$$

We neglect a positive term, perform the integral and obtain:

$$\theta_f > \frac{2}{\sqrt{1+a}}\left[\frac{\pi}{2} - \arcsin\left(s\,r_1^{-1}\sqrt{1+a}\right)\right] + 2\arcsin\left(\frac{s}{r_0}\right) \ . \tag{25}$$

We fix z; in the limit $V \to \infty$, we have $s \to 0$, $r_{\min} \to 0$, $a \to 0$, and hence $\theta_f \to \pi$. This fact implies that $|F| \to 0$, by using the Dominated Convergence Theorem and observing that in Eq. (20) the integrand is bounded by an L_1 function. (A rough but simple choice of this function can be made considering the scattering process with interaction at short distance $\Psi = g_1\,r^{-2}$, which produces a larger friction force and, as proved in i), gives a bounded force).

(iii) Bounded interaction.

We can prove that the friction force vanishes for $V \to \infty$ in a similar (but simpler) way to that of point (ii).

(iv) $\alpha > 2$

We must show that $F \to -\infty$ for $V \to \infty$. It is enough to investigate a quantity smaller than $|F|$.

First we restrict the set of integration in Eq. (20) to $z = sV \le b$, where b is a fixed constant. Second we decrease $|F|$ by increasing θ_f. We proceed like (ii). We change the term $2gV^{-2}r^{-\alpha}$ into as^2r^{-2}, where $a = 2gV^{-2}s^{-2}(r_{\min})^{2-\alpha}$. By this

operation we increase θ_f and we have

$$\theta_f < \int_{r_{\min}}^{r_1} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1-as^2r^{-2}-(s/r)^2}} + 2\arcsin\left(\frac{s}{r_0}\right) + \int_{r_1}^{r_0} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1-2V^{-2}\Psi(r)-(s/r)^2}}\,. \tag{26}$$

Performing the integral and neglecting a negative term, we obtain

$$\theta_f < \frac{\pi}{\sqrt{1+a}} + \int_{r_1}^{r_0} \frac{2s\,\mathrm{d}r}{r^2\sqrt{1-2V^{-2}\Psi(r)-(s/r)^2}} + 2\arcsin\left(\frac{s}{r_0}\right)\,. \tag{27}$$

For $V \to \infty$, for z fixed, $s \to 0$, $r_{\min} \approx (2g/V^2)^{1/\alpha}$, and so $a \to \infty$. Hence in this limit $\theta_f \to 0$ and $\cos\theta_f \to 1$. Now we allow $b \to \infty$ and we have proved that $|F| \to \infty$.

We have proved that for large V the friction force is larger than E. On the contrary for $V = 0$ the friction force vanishes. Since it is a continuous function of V, there exists at least a velocity of the body for which the friction force balances exactly the external force E, and this value gives a stationary state $\hat{f}$. □

The stationary state $\hat{f}$ corresponds to a scattering state with particles of the medium at rest (in the original reference system in which the body moves at speed V), and hence at zero temperature. A similar study can be made for a scattering state with temperature $T > 0$, corresponding to a distribution which is of the form $f = \rho(\beta/\pi)^{3/2}\mathrm{e}^{-\beta|\mathbf{v}|^2}$ for incoming particles at a distance from the body larger than the range of the interaction. The analysis could be performed in strict analogy to the one previously discussed, with some geometrical complications. We also remark that, although we have assumed the potential to be a decreasing function on $[0, r_0]$, this assumption can be relaxed, by choosing $\Psi(r)$ a bounded but possibly attractive potential in the interval $[r_1, r_0]$. The analysis of the scattering process leads to similar conclusions.

To summarize, we have shown that, in the framework of a fully Hamiltonian system, the motion of a body in a medium can represent a good model of viscous friction if the interaction between the body and the medium particles is sufficiently strong at short distances, i.e., if the potential has a singularity at the origin stronger than $r^{-\alpha}$, with $\alpha > 2$. In fact, this condition assures that the force exerted by the medium on the body during its motion increases to infinity when the velocity of the body diverges, relation that we expect to occur in any reasonable model of viscous friction. On the contrary, if the potential at short distances is of the form $r^{-\alpha}$, with $0 \le \alpha \le 2$, the force exerted by the medium remains finite, then in this case the model cannot be considered as a model of viscous friction.

This analysis is preliminary to the much more complicated case of a nonstationary motion of the body, that is the study of the Cauchy problem with arbitrary initial data, as expressed in (11)–(13), proving the reaching of an asymptotic velocity

(for $\alpha > 2$), or an unbounded growth of it (for $0 \leq \alpha \leq 2$), finding the rate of convergence of the velocity of the body towards the asymptotic velocity, or towards a "uniformly accelerated motion". The latter case corresponds to the *runaway particle effect*, as discussed in [2, 3]. Another level of difficulty would consist to assume a *real*, self-interacting, Vlasov fluid, which is an open problem.

References

1. Buttà, P., Cavallaro, G., Marchioro, C.: Mathematical Models of Viscous Friction. Lecture Notes in Mathematics, vol. 2135. Springer, Berlin (2015)
2. Buttà, P., Ferrari, G., Marchioro, C.: Speedy motions of a body immersed in an infinitely extended medium. J. Stat. Phys. **140**, 1182–1194 (2010)
3. Buttà, P., Manzo, F., Marchioro, C.: A simple Hamiltonian model of runaway particle with singular interaction. Math. Models Methods Appl. Sci. **15**, 753–766 (2005)
4. Caglioti, E., Caprino, S., Marchioro, C., Pulvirenti, M.: The Vlasov equation with infinite Mass. Arch. Ration. Mech. Anal. **159**, 85–108 (2001)
5. Caglioti, E., Marchioro, C., Pulvirenti, M.: Non-equilibrium dynamics of three-dimensional infinite particle systems. Commun. Math. Phys. **215**, 25–43 (2000)
6. Caprino, S., Cavallaro, G., Marchioro, C.: Time evolution of a Vlasov-Poisson plasma with magnetic confinement. Kinet. Relat. Models **5**, 729–742 (2012)
7. Caprino, S., Cavallaro, G., Marchioro, C.: On a magnetically confined plasma with infinite charge. SIAM J. Math. Anal. **46**, 133–164 (2014)
8. Caprino, S., Cavallaro, G., Marchioro, C.: Remark on a magnetically confined plasma with infinite charge. Rend. Mat. Appl. **35**, 69–98 (2014)
9. Caprino, S., Cavallaro, G., Marchioro, C.: Time evolution of a Vlasov-Poisson plasma with infinite charge in $\mathbb{R}^3$. Commun. Partial Differ. Equ. **40**, 357–385 (2015)
10. Caprino, S., Cavallaro, G., Marchioro, C.: On a Vlasov-Poisson plasma confined in a torus by a magnetic mirror. J. Math. Anal. Appl. **427**, 31–46 (2015)
11. Caprino, S., Cavallaro, G., Marchioro, C.: Time evolution of an infinitely extended Vlasov system with singular mutual interaction. J. Stat. Phys. **162**, 426–456 (2016)
12. Caprino, S., Cavallaro, G., Marchioro, C.: A Vlasov-Poisson plasma with unbounded mass and velocities confined in a cylinder by a magnetic mirror. Kinet. Relat. Models **9**, 657–686 (2016)
13. Caprino, S., Cavallaro, G., Marchioro, C.: On the magnetic shield for a Vlasov-Poisson plasma. J. Stat. Phys. **169**, 1066–1097 (2017)
14. Caprino, S., Cavallaro, G., Marchioro, C.: The Vlasov-Poisson equation in $\mathbb{R}^3$ with infinite charge and velocities. J. Hyperbolic Differ. Equ. **15**, 407–442 (2018)
15. Caprino, S., Cavallaro, G., Marchioro, C.: Efficacy of a magnetic shield against a Vlasov-Poisson plasma. Rep. Math. Phys. **84**, 85–116 (2019)
16. Caprino, S., Cavallaro, G., Marchioro, C.: Time evolution of a Vlasov-Poisson plasma with different species and infinite mass in $\mathbb{R}^3$. Z. Angew. Math. Phys. **71**, 1–9 (2020)
17. Caprino, S. Marchioro, C.: On the plasma-charge model. Kinet. Relat. Models **3**, 241–254 (2010)
18. Caprino, S., Marchioro, C.: On a charge interacting with a plasma of unbounded mass. Kinet. Relat. Models **4**, 215–226 (2011)
19. Caprino, S., Marchioro, C., Pulvirenti, M.: On the Vlasov-Helmholtz equations with infinite mass. Commun. Partial Differ. Equ. **27**, 791–808 (2002)
20. Glassey, R.: The Cauchy Problem in Kinetic Theory. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1996)
21. Gruber, Ch., Piasecki, Jb.: Stationary motion of the adiabatic piston. Phys. A **268**, 412–423 (1999)

22. Horst, E.: On the classical solutions of the initial value problem for the unmodified non-linear Vlasov equation I. Math. Methods Appl. Sci. **3**, 229–248 (1981)
23. Horst, E.: On the classical solutions of the initial value problem for the unmodified non-linear Vlasov equation II. Math. Methods Appl. Sci. **4**, 19–32 (1982)
24. Jabin, P.E.: The Vlasov-Poisson system with infinite mass and energy. J. Stat. Phys. **103**, 1107–1123 (2001)
25. Lebowitz, J. L., Piasecki, J., Sinai, Ya.: Scaling dynamics of a massive piston in an ideal gas. Hard ball systems and the Lorentz gas, 217–227, Encyclopaedia Math. Sci., vol. 101. Springer, Berlin (2000)
26. Lions, P.-L., Perthame, B.: Propagation of moments and regularity for the 3-dimensional Vlasov-Poisson system. Invent. Math. **105**, 415–430 (1991)
27. Loeper, G: Uniqueness of the solution to the Vlasov-Poisson system with bounded density. J. Math. Pure Appl. **86**, 68–79 (2006)
28. Majda, A., Majda, G., Zheng, Y.: Concentrations in the one-dimensional Vlasov-Poisson equations. I: temporal development and non-unique weak solutions in the single component case. Phys. D **74**, 268–300 (1994)
29. Majda A., Majda G., Zheng, Y.: Concentrations in the one-dimensional Vlasov-Poisson equations. II: screening and the necessity for measure-valued solutions in the two component case. Phys. D **79**, 41–76 (1994)
30. Marchioro, C., Miot, E., Pulvirenti, M.: The Cauchy problem for the 3-D Vlasov-Poisson system with point charges. Arch. Ration. Mech. Anal. **201**, 1–26 (2011)
31. Pankavich, S.: Global existence for a three dimensional Vlasov-Poisson system with steady spatial asymptotics. Commun. Partial Differ. Equ. **31**, 349–370 (2006)
32. Pfaffelmoser, K.: Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data. J. Differ. Equ. **95** 281–303 (1992)
33. Salort, D.: Transport equations with unbounded force fields and application to the Vlasov-Poisson equation. Math. Models Methods Appl. Sci. **19**, 199–228 (2009)
34. Schaeffer, J.: Global existence of smooth solutions to the Vlasov-Poisson system in three dimensions. Commun. Partial Differ. Equ. **16**, 1313–1335 (1991)
35. Schaeffer, J: Steady spatial asymptotics for the Vlasov-Poisson system. Math. Methods Appl. Sci. **26**, 273–296 (2003)
36. Schaeffer, J: The Vlasov Poisson system with steady spatial asymptotics. Commun. Partial Differ. Equ. **28**, 1057–108 (2003)
37. Schaeffer, J.: Global existence for the Vlasov-Poisson system with steady spatial asymptotic behavior. Kinet. Relat. Models **5**, 129–153 (2012)
38. Ukai, S., Okabe, T.: On classical solutions in the large in time of two-dimensional Vlasov's equation. Osaka J. Math. **15**, 245–261 (1978)
39. Wollman, S.: Global in time solutions to the two-dimensional Vlasov-Poisson system. Commun. Pure Appl. Math. **33**, 173–197 (1980)
40. Wollman S: Global in time solution to the three-dimensional Vlasov-Poisson system. J. Math. Anal. Appl. **176**, 76–91 (1993)

Mathematical and Numerical Study of a Dusty Knudsen Gas Mixture: Extension to Non-spherical Dust Particles

Frédérique Charles

Abstract In this work, we consider the model introduced in Charles and Salvarani (Acta Appl Math 168:17–31, 2020) describing the movement of dust particles in a very rarefied atmosphere. The gas is treated as a Knudsen gas, whereas the interaction between dust particles and gas molecules is modeled by considering a moving domain free transport equation (including the boundary with the particles and the boundary of the domain). We here precise the proof of existence of solutions to the initial-boundary value problem announced in Charles and Salvarani (Acta Appl Math 168:17–31, 2020). Moreover, we introduce a new numerical strategy, based on a splitting between the transport of the gas molecules and the movement of the boundary. This strategy allow to perform 2d-numerical simulations with elliptical-shaped particles.

1 Introduction

We consider here a mixture of a rarefied gas and macroscopic particles (such as dust particles). A typical example of such a situation is the study of the dynamics of gases inside a microelectromechanical system (MEMS). More precisely, we place ourselves in the physical situation described by the order of magnitude of the physical constants in Table 1. Under these assumptions, the mean free path of the gas is equal to $\lambda_g = 2 \cdot 10^{-3}$ m, and the Knudsen number of the gas (that is, the ratio between the mean free path and the characteristic length of the domain) inside the container is $K_n = 10$. In this context, a kinetic description of the gas is more suitable than a description with fluid models. Moreover, one of the advantages of kinetic models is that they depend much less on phenomenological laws than most models of continuum mechanics. We therefore consider a mesoscopic scale and describe the gas thanks to a density function defined in the phase space (no

F. Charles (✉)
Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Université, CNRS, Université de Paris, Paris, France
e-mail: frederique.charles@sorbonne-universite.fr

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_6

Table 1 Order of magnitude of physical quantities in the situation under study

Temperature of the gas	T_g	293 K
Mach number	Ma	0.1
Gas Pressure	P	5 Pa
Size of the Container	L	$2 \cdot 10^{-4}$ m
Radius of particles	r	10^{-5} m

distinction is made here between the different types of molecules constituting the gas). Without any particles, a rarefied gas inside a vessel could typically be described by the Boltzmann equation (see [4]) with suitable boundary conditions. A kinetic description of a gas-particle mixture was introduced in [5], where the flow of particles is described thanks to another density function, and interactions between particles and molecules are modeled by integral collisions operators. We can also mention [11], where the movement of spherical particles is described through equations on their momentum and velocity, and where the gas is described by a Boltzmann equation with an integral operator describing gas-particles interactions. In [13], the motion of a rigid body immersed in a gas is governed by the Newton-Euler equations, where the force and the torque on this body are computed from the momentum transfer of the gas molecules colliding with the body; the gas is described by a Boltzmann equation without any effect of the body on the gas.

The point of view adopted in [7] is rather different. The interaction between the gas and the particles (in finite number) is modeled by considering the evolution of the gas in a moving domain, where the boundary of the domain include the surface of the particles. This approach has already been introduced in [9] and [12]. However, in the later works, authors use an Eulerian numerical method (Finite-Difference and Semi-Lagrangien method respectively) which makes the treatment of boundary conditions rather complicated; the numerical study is therefore only carried out in dimension 1.

Moreover, for large Knudsen number (typically larger to 10), it is generally admitted [10] that the gas can be considered as a Knudsen gas (or molecular flow), and we therefore neglect here collisions between molecules. Theoritical studies of the convergence to equilibrium of a particle in a Knudsen gas have been carried out in [2] and [3], but no numerical simulation has been performed. The study of a Knudsen system in a moving domain, both at the theoretical and at the numerical level, has been the subject of [8], but in the context of a gas in a vessel with absorbing boundary conditions.

The paper is organised as follows. We first recall the model introduced in [7] for spherical particles, that we extend to any shape of particle. We then precise the proof of the existence of solutions announced in [7]. Finally we present a new numerical strategy that allows to perform numerical simulations with non-spherical particles, and some scenarios of numerical simulations.

2 Description of the Model

We briefly recall the model introduced in [7] and generalize it to non-spherical particles. We consider a free transport equation in a open bounded and regular spatial domain $D \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$, which describes the evolution of the molecules density $f := f(t, x, v)$, with $(t, x, v) \in \mathbb{R}^+ \times D \times \mathbb{R}^d$. The motion of particles is supposed to be known, and we denote $B_i(t)$ the closed set corresponding to the region occupied by the particle indexed by i at time t. We assume that the domain D and the particles are smooth enough (C^1 for example) to define a normal vector on the boundaries. We introduce the time $\mathcal{T}_1$ which guarantees the non-overlapping of particles

$$\mathcal{T}_1 = \sup\{t \geq 0 : \forall s \in [0, t[,\ B_j(s) \cap B_i(s) = \emptyset \quad \text{for all } j,\ i = 1, \ldots, N_d,\ j \neq i\} \tag{1}$$

and the time $\mathcal{T}_2$ which guarantee the non-exit of particles out of the domain

$$\mathcal{T}_2 = \sup\{t \geq 0 : \forall s \in [0, t[, B_i(s) \cap \partial D = \emptyset \quad \text{for all } i = 1, \ldots, N_d\}. \tag{2}$$

We do not consider here collisions of a particle with another particle or with the boundary of the domain, and therefore consider the problem for $t \in [0, \mathcal{T})$, with $\mathcal{T} \leq \min(\mathcal{T}_1, \mathcal{T}_2)$. For $t \in [0, \mathcal{T})$ we denote Ω^t the domain occupied by the gas at time t

$$\Omega^t := D \setminus \bigcup_{i=1}^{N_d} B_i(t),$$

and $\partial\Omega^t = \partial D \cup \Gamma^t$ its boundary, with

$$\Gamma^t = \bigcup_{i=1}^{N_d} \partial B_i(t)$$

(see Fig. 1). The motion of the domain is described through the velocity law of each point of the boundary at a given time t. We define a field $c : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$ which gives the local velocity of each point $x \in \partial\Omega^t$, for any $t \in [0, \mathcal{T}[$. We note that for any $x \in \partial D$, we have $c(t, x) = 0$. We assume that the interaction between molecules and particles is described by a diffuse reflection on the surface of the particle, and that all particles have the same temperature $T_p > 0$, uniform on the surface. Following this assumption, the boundary condition on the surface of

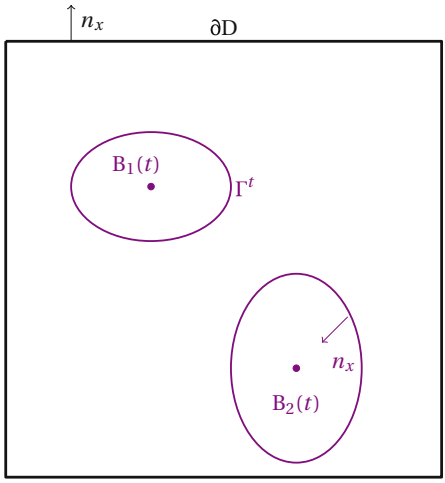

Fig. 1 Graphical description of the problem

particles, that is for $x \in \Gamma^t$, writes

$$f(t,x,v) = \begin{cases} \displaystyle\int_{\{(w-c(t,x))\cdot n_x \geq 0\}} k_{d,T_p}(x, v-c(t,x), w-c(t,x)) f(t,x,w) dw \\ \qquad \text{for } x \in \Gamma^t,\ (v-c(t,x))\cdot n_x < 0 \\ 0 \quad \text{for } x \in \Gamma^t,\ (v-c(t,x))\cdot n_x \geq 0 \end{cases} \tag{3}$$

where $n_x \in \mathbb{S}^{d-1}$ is the outward normal originated in x, and k_{d,T_p} a kernel modeling a diffuse reflection at temperature T_p, defined by (see [14])

$$k_{d,T_p}(x,v,w) = \sqrt{\frac{2\pi}{T_p}} \mathcal{M}_{T_p}(v)(w \cdot n_x), \tag{4}$$

where $\mathcal{M}_{T_p}$ is the centered Maxwellian at temperature T_p:

$$\mathcal{M}_{T_p}(v) = \frac{1}{(2\pi T_p)^{d/2}} e^{-\frac{|v|^2}{2T_p}}.$$

This kernel verifies the following property

$$\int_{\{w\cdot n_x \geq 0\}} k_{d,T_p}(x,v,w)\, M_{T_p}(w) dw = M_{T_p}(v). \tag{5}$$

For $x \in \partial D$, that is on the boundary of the external domain, which is assumed to be still ($c(t,x) = 0$ for $x \in \partial D$), the boundary condition writes

$$f(t,x,v) = \begin{cases} \displaystyle\int_{\{w\cdot n_x \geq 0\}} k(x,v,w) f(t,x,w) dw & \text{for } x \in \partial D,\ v \cdot n_x < 0 \\ 0 & \text{for } x \in \partial D,\ v \cdot n_x \geq 0 \end{cases} \tag{6}$$

In (6), k is a kernel modeling an accommodation reflection on ∂D at temperature T_p

$$k(x,v,w) = \gamma k_s(x,v,w) + (1-\gamma) k_{d,T_p}(x,v,w)$$

where $\gamma \in [0,1]$ is the accommodation coefficient, and k_s a kernel modeling a specular reflection

$$k_s(x,v,w) = \delta(w - v + 2(v \cdot n_x) n_x).$$

The kernel k_s verifies, for all function φ defined on $\mathbb{R}^+$:

$$\int_{\{w\cdot n_x \geq 0\}} k_s(x,v,w)\, \varphi(|w|) dw = \varphi(|v|) \mathbb{1}_{\{v\cdot n_x \leq 0\}}. \tag{7}$$

One can summarize the boundary conditions by

$$f(t,x,v) = \int_{\mathbb{R}^3} K(t,x,v,w) f(t,x,w) dw\, \mathbb{1}_{\{(v-c(t,x))\cdot n_x < 0\}} \quad \text{for } x \in \partial\Omega^t, \tag{8}$$

with

$$K(t,x,v,w) = \begin{cases} k_{d,T_p}(x, v - c(t,x), w - c(t,x))\, \mathbb{1}_{\{w - c(t,x)\cdot n_x \geq 0\}} & \text{if } x \in \Gamma^t \\ \left(\gamma k_s(x,v,w) + (1-\gamma) k_{d,T_p}(x,v,w)\right) \mathbb{1}_{\{w\cdot n_x \geq 0\}} & \text{if } x \in \partial D. \end{cases} \tag{9}$$

We end-up with the following model

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = 0 \qquad (t,x,v) \in \mathbb{R}^+ \times \Omega^t \times \mathbb{R}^d, \tag{10}$$

with the initial condition

$$f(0,x,v) = f^{\text{in}}(x,v) \mathbb{1}_{\{\Omega^0 \times \mathbb{R}^d\}}(x,v) \tag{11}$$

and the boundary conditions (8)–(9).

3 Existence

We slightly modify and precise the Theorem 3.3 in [7].

Theorem 1 *Let $\mathcal{T} \in (0, \min(T_1, T2))$, where T_1 and T_2 are defined by (1) and (2). We assume that $c \in L^\infty((0,T)\times D)$. Let $f^{\text{in}} \in L^\infty(\Omega^0 \times \mathbb{R}^d, e^{|v|^2/T_p}\mathrm{d}v\,\mathrm{d}x)$, $f^{\text{in}} \geq 0$ for a.e. $(x,v) \in \Omega^0 \times \mathbb{R}^d$. Then there exists at least one non-negative weak solution $f \in L^\infty((0,T); L^\infty(\bar{\Omega}^t, \mathbb{R}^d))$ of the initial-boundary value problem (10)–(11)–(8)–(9). Moreover $(t,x,v) \mapsto f(t,x,v)e^{\frac{|v-c(t,x)|^2}{2T_p}} \in L^\infty((0,T); L^\infty(\bar{\Omega}^t, \mathbb{R}^d))$.*

Proof We follow and adapt the proof for a fixed domain made in [1]. We first consider the auxiliary problem for the function $g : \mathbb{R}^+ \times \Omega^t \times \mathbb{R}^d \to \mathbb{R}$

$$\frac{\partial g}{\partial t} + v \cdot \nabla_x g = 0, \qquad (t,x,v) \in \mathbb{R}^+ \times \Omega^t \times \mathbb{R}^d, \tag{12}$$

with initial data

$$g(0,x,v) = f^{\text{in}}(x,v)\mathbb{1}_{\{\Omega^0 \times \mathbb{R}^d\}}(x,v) \tag{13}$$

and boundary conditions

$$g(t,x,v) = \Phi(t,x,v)\mathbb{1}_{\{(v-c(t,x))\cdot n_x < 0\}} \tag{14}$$

for a.e. $(x,v) \in \partial\Omega^t \times \mathbb{R}^d$, where $\Phi \in L^\infty((0,T)\times\partial\Omega^t \times \mathbb{R}^d)$ is a given function. The problem (12)–(13)–(14) has a unique weak solution, given by the method of characteristics

$$g(t,x,v) = f^{\text{in}}(x - vt, v)\mathbb{1}_{\{\tau_{\Omega^t}(x,v) > t\}} + \Phi(t, x - \tau_{\Omega^t}(x,v)v, v)\mathbb{1}_{\{\tau_{\Omega^t}(x,v) < t\}},$$

where

$$\tau_{\Omega^t}(x,v) = \begin{cases} +\infty & \text{if } \{\theta > 0 \,:\, x - \theta v \in \Gamma^{t-\theta} \cup \partial D\} = \emptyset \\ \inf\{\theta > 0 : \, x - \theta v \in \Gamma^{t-\theta} \cup \partial D\} & \text{otherwise.} \end{cases}$$

$\tau_{\Omega^t}(x,v)$ correspond to the arrival time on the boundary when we follow backward the characteristic starting from $x \in \Omega^t$ at velocity $v \in \mathbb{R}^d$. We deduce that

$$\|g\|_{L^\infty((0,T)\times\Omega^t\times\mathbb{R}^d)} \leq \max\{\|f^{\text{in}}\|_{L^\infty(\Omega^0\times\mathbb{R}^d)}, \, \|\Phi\|_{L^\infty((0,T)\times\partial\Omega^t\times\mathbb{R}^d}\}. \tag{15}$$

We now consider the sequence $(f_n)_{n\in\mathbb{N}}$ of functions, such that

$$f_0(t,x,v) = 0 \quad \text{for a.e. } (t,x,v) \in [0,T) \times \bar{\Omega}^t \times \mathbb{R}^d$$

and, for all $n \in \mathbb{N}$, $n \geq 1$, f_n is the solution of the following initial-boundary value problems:

$$\frac{\partial f_n}{\partial t} + v \cdot \nabla_x f_n = 0, \qquad (t, x, v) \in \mathbb{R}^+ \times \Omega^t \times \mathbb{R}^d, \tag{16}$$

with initial data

$$f_n(0, x, v) = f^{\text{in}}(x, v) \mathbb{1}_{\{\Omega^0 \times \mathbb{R}^d\}}(x, v) \tag{17}$$

and boundary conditions

$$f_n(t, x, v) = \int_{\mathbb{R}^3} K(t, x, v, w) f_{n-1}(t, x, w) dw \mathbb{1}_{\{(v - c(t,x)) \cdot n_x < 0\}} \tag{18}$$

for $(x, v) \in \partial\Omega^t \times \mathbb{R}^d$, where K is defined in (9). Thanks to properties (5) and (7), the boundary condition (18) leads to the estimate on the boundary

$$\left\| \frac{f_n}{\mathcal{M}_{T_p}(v - c(t,x))} \right\|_{L^\infty((0,T) \times \partial\Omega^t \times \mathbb{R}^d)} \leq \left\| \frac{f_{n-1}}{\mathcal{M}_{T_p}(v - c(t,x))} \right\|_{L^\infty((0,T) \times \partial\Omega^t \times \mathbb{R}^d)} \tag{19}$$

Then the estimate (15) allow to prove by induction that

$$\left\| \frac{f_n}{\mathcal{M}_{T_p}(v - c(t,x))} \right\|_{L^\infty((0,T) \times \bar{\Omega}^t \times \mathbb{R}^d)} \leq \left\| \frac{f^{\text{in}}}{\mathcal{M}_{T_p}(v - c(t,x))} \right\|_{L^\infty((0,T) \times \Omega^0 \times \mathbb{R}^d)} \tag{20}$$

Moreover, an immediate induction argument prove that $f_n \geq 0$ for all $n \geq 0$. We obtain

$$0 \leq f_n(t, x, v) \leq f_n(t, x, v) e^{\frac{|v - c(t,x)|^2}{2T_p}} \leq \left\| f^{\text{in}} e^{\frac{|v|^2}{T_p}} \right\|_{L^\infty(\Omega^0 \times \mathbb{R}^d)} e^{\frac{\|c\|_\infty^2}{T_p}} \tag{21}$$

for a.e. $(t, x, v) \in [0, T) \times \bar{\Omega}^t \times \mathbb{R}^d$. We then can prove that the sequence is non decreasing by considering the sequence $h_n := f_{n+1} - f_n$, for all $n \geq 0$. By linearity, for all $n \geq 0$, h_n satisfy the free transport equation (12) with the initial condition

$$\forall (x, v) \in \Omega^0 \times \mathbb{R}^d, \qquad \begin{cases} h_0(0, x, v) = f^{\text{in}}(x, v) \mathbb{1}_{\{\Omega^0 \times \mathbb{R}^d\}}(x, v) \\ h_n(0, x, v) = 0 \quad \text{for } n \geq 1; \end{cases}$$

and the boundary condition

$$h_n(t,x,v) = \int_{\mathbb{R}^3} K(t,x,v,w) h_{n-1}(t,x,w) dw \mathbb{1}_{\{(v-c(t,x))\cdot n_x<0\}} \tag{22}$$

for $(x,v) \in \partial\Omega^t \times \mathbb{R}^d$. We deduce that $h_n(t,x,v) \geq 0$ for a.e. $(t,x,v) \in [0,T) \times \bar{\Omega}^t \times \mathbb{R}^d$. We have hence built a monotone non-decreasing sequence $(f_n)_{n\in\mathbb{N}}$ composed by non-negative and uniformly bounded functions a.e. in the domain of definition of the problem. By consequence, the sequence $(f_n)_{n\in\mathbb{N}}$ pointwise converges to a limit f, which is by construction a non-negative solution of the initial-boundary value problem (10)–(11)–(8), and we can pass to the limit in estimate (21). □

4 Numerical Simulations

4.1 Numerical Method

We describe here a new strategy for the numerical study of the model (10)–(11)–(8), which is a modification of the particle method proposed in [7]. The initial density f^{in} of the gas is discretized by mean of a collection of weighted smooth shape functions centered on the particle positions, that is

$$f^{\text{in}}_{\varepsilon,N_m}(x,v) = \sum_{k=1}^{N_m} \omega_k \, \varphi_\varepsilon(x - x_k^0) \, \varphi_\varepsilon(v - v_k^0), \tag{23}$$

where N_m represents the number of numerical particles, ω_k is the weight of the k-th numerical particle (which represents ω_k molecules). In (23), the shape function $\varphi_\varepsilon(x) = \varphi(\varepsilon^{-1}x)/\varepsilon^d$ is a smooth function with compact support. The term "numerical particles" is here used for avoiding any confusion with the (real) number of dust particles. Once the number N_m of numerical particles has been chosen, the initial positions $(x_k^0)_{1\leq k\leq N_m}$ and velocities $(v_k^0)_{1\leq k\leq N_m}$ are sampled according to the initial density f^{in} (either in a deterministic way, either thanks to a Monte-Carlo procedure). Then, the positions and velocities of the numerical particles evolve in time by taking into account the different phenomena listed below:

(i) the free flow of the numerical particles in the absence of any interaction, mathematically represented by the transport operator $v \cdot \nabla$;
(ii) the boundary condition on ∂D; we can consider here specular reflection or accommodation reflection, but also a periodic condition in order to hide the effects of the boundary.

(iii) the diffuse reflection between gas molecules and dust particles;
(iv) the time evolution of the set of dust particles.

We introduce a time discretization of step Δt and we set $t^n = n\Delta t$. The density of gaseous molecules at time t^n—i.e. $f(t^n, \cdot, \cdot)$ where f is the solution of (10)–(11)–(8) is then approximated by

$$f^n_{\varepsilon,N_m}(x, v) = \sum_{k=1}^{N_m} \omega_k \, \varphi_\varepsilon(x - x^n_k) \, \varphi_\varepsilon(v - v^n_k), \tag{24}$$

where $(x^n_k)_{1\le k\le N}$ and $(v^n_k)_{1\le k\le N}$ are the positions and the velocities of the numerical particles at time t^n.

In [7], our strategy was to compute simultaneously the steps *(i)*, *(iii)*, *(iv)* previously described. For that purpose, we compute for each numerical particle the position of the possible intersection of its trajectory with the dust particle during the time Δt. To do that, we compute if the condition

$$\min_{1\le i\le N_p} \min_{t\in[t^n, t^n+\Delta t]} \|\xi_i(t) - x^n_k(t)\| \le r,$$

is verified or not, where $\xi_i(t)$ is the position of the center of the i-th spherical particle, r its radius, and $x^n_k(t) = x^n_k + (t - t^n)v^n_k$ the trajectory of the k-th numerical particle between time t^n and t^{n+1}. However, this strategy is hardy adaptable to non-spherical particles.

We consider here a splitting between the advection stage of the dust particles *(iv)* and the evolution of gas molecules, corresponding to stages *(i)–(ii)–(iii)*. In other word, we first transport dust particles independently of molecules during the time Δt, and we then transport numerical particles and perform the treatment of the boundary conditions. We thus come back to dealing with conditions at the boundaries of a fixed domain instead of a mobile domain. We first test on every numerical particle if $x^{n+1} \in \Omega^t$, and otherwise we compute the boundary condition. To do so, we only need a Cartesian equation of the surface of dust particles, in order to calculate the intersection of this surface with a straight line as well as the normal vector at each point of the surface.

The latter strategy, which gives graphically similar results to the first one for spherical particles, allows to consider easily some ellipse-shaped particles. For such particles, our objective is in particular to observe the effect of the rotational velocity of the particle on the gas. This effect was not visible for spherical particles because the gas has no viscosity.

4.2 Numerical Results

We describe here a series of numerical experiments in dimension $d = 2$. We suppose that the initial density is uniform in space and that it is described by a Maxwellian function in velocity, that is

$$f^{\text{in}}(x, v) = f^{\text{in}}(v) = \frac{n_0 m}{2\pi k_B T^{\text{in}}} e^{-\frac{m|v-u_g|^2}{2k_B T^{\text{in}}}}, \tag{25}$$

where m is the mass of a gas molecule, u^{in} and T^{in} are respectively the initial macroscopic velocity and the temperature of the gas (in K), and n_0 correspond to $\|f^{\text{in}}\|_{L^1(\Omega^0 \times R^3)}/|\Omega^0|$. In this case, each component $(v_k^0)_i$, for $1 \leq k \leq N_p$ of the initial velocities of the gas particles is sampled according a Gaussian law of mean $(u^{\text{in}})_i$ and variance $k_B T^{\text{in}}/m$ (see [6] for details). The weights of the particles are identical, and are tuned in order to reproduce the mass of the initial condition:

$$\omega_k = \frac{\|f^{\text{in}}\|_{L^1(\Omega^0 \times R^3)}}{N_p} = \frac{n_0 |\Omega^0|}{N_p}, \qquad \text{for all } 1 \leq k \leq N_p.$$

The initial positions of the numerical particles have been fixed on a regular grid, except inside the dust particles. In some scenarios, the gas has a macroscopic velocity along the first axis equal $u^{\text{in}} = V_s M_a$, where M_a is the Mach number and V_s is the sound velocity in air at temperature T^{in}. We take here $M_a = 0.1$ and $T^{\text{in}} = 293\,\text{K}$; then $u^{\text{in}} = 34.41\,\text{m/s}$. The temperature of the surface of particles is 500 K. The value n_0 has been normalized to 1. Indeed, the values of n_0 have no impact, neither on the transport of molecules and of dust particles (since these ones are no influenced by the surrounding gas) nor on the collisions between molecules and dust particles (the number of collisions is not computed as in DSMC methods). The domain D is the square $[-1, 1] \times [-1, 1]$ (in 10^{-5} m), with specular reflection at the top and bottom boundary. We use a periodic boundary condition at the left and at the right sides of ∂D, in order to mimic an infinite domain in the x direction. We use B_3-splines (see [6]) as shape functions φ, with a shape size $\varepsilon = h^{0.5}$, where h is the initial distance between two numerical gas particles in each direction (and which is obviously linked to N_m).

4.2.1 Scenarios 1 and 2

The first simulations presents the rotation of a particle with no translational velocity. The particle is an ellipse, with axes equal to $a = 2.5 \cdot 10^{-5}$ and $b = 1 \cdot 10^{-5}$. In the first scenario, the macroscopic velocity of the gas is $u_g = (0, 0)$, whereas in the second one the macroscopic velocity of the gas is $u_g = (-u^{\text{in}}, 0)$. The rotational

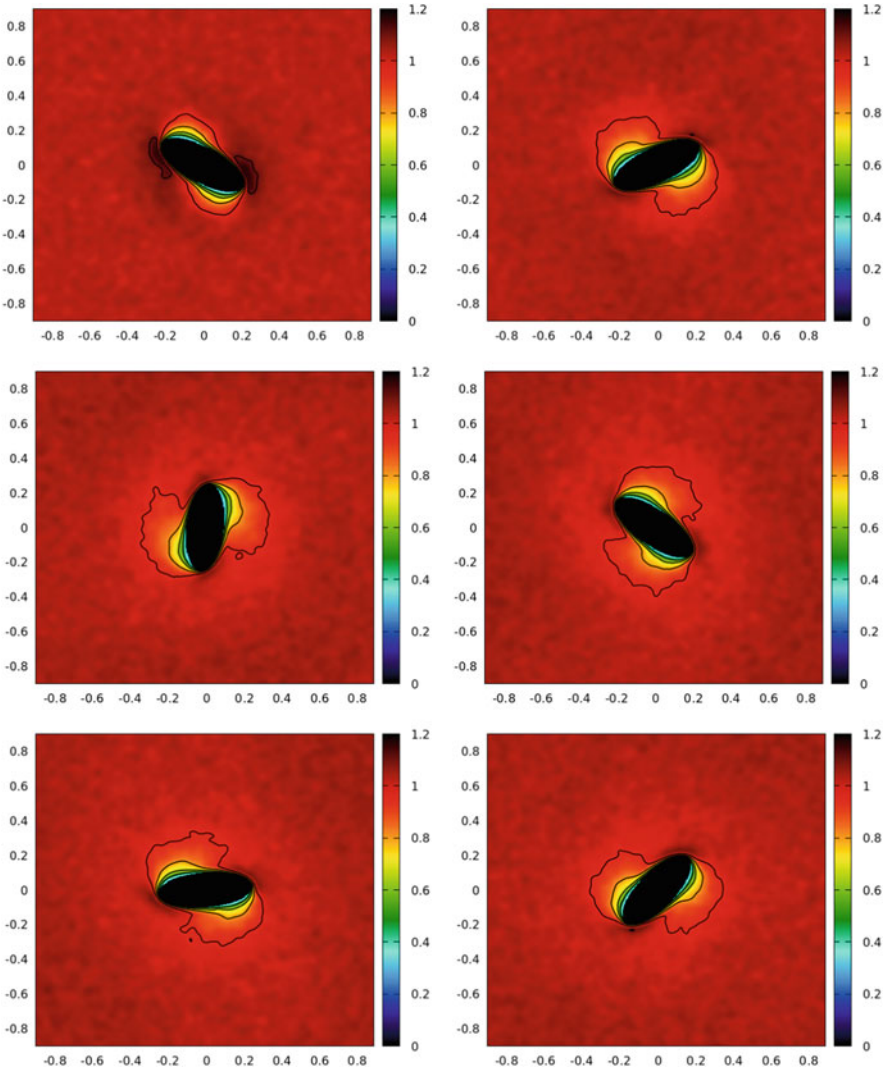

Fig. 2 Time history of the gas density in Scenario 1, from left to right and from top to bottom, at times $8 \cdot 10^{-8}$ s, $2.4 \cdot 10^{-7}$ s, $4 \cdot 10^{-7}$ s, $5.6 \cdot 10^{-7}$ s, $7.2 \cdot 10^{-7}$ s, $8 \cdot 10^{-7}$ s. The axis are scaled according to the length scale $L^\circ = 10^{-4}$ m

velocity of the particle is equal to $\Omega = 2\pi \cdot 10^6$ rad/s in both scenarios. Figures 2 and 3 show the time evolution of the number density $\rho(t, x) = \int_{\mathbb{R}^3} f(t, x, v)\mathrm{d}v$ of scenario 1 and 2 respectively. In particular, we can observe the effect of the macroscopic velocity of the gas, which acts as a side wind. Figure 4 show the

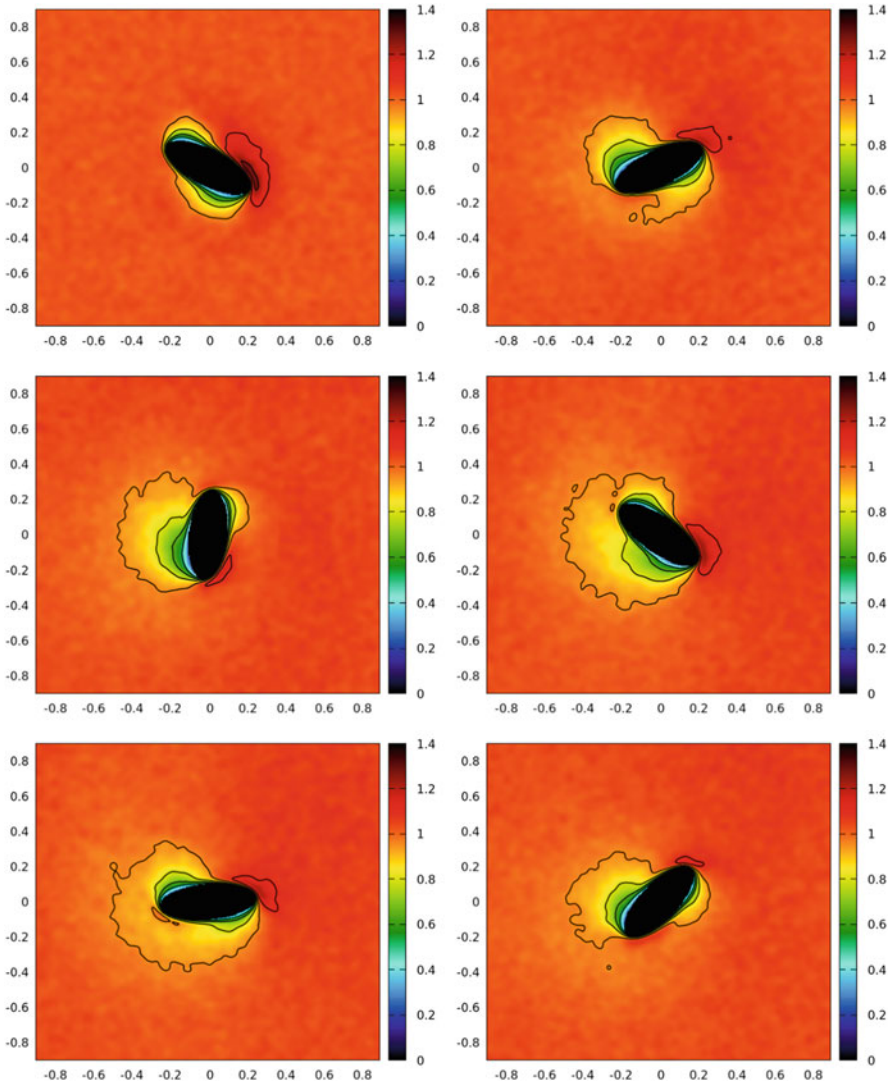

Fig. 3 Time history of the gas density in Scenario 2, from left to right and from top to bottom, at times $8 \cdot 10^{-8}$ s, $2.4 \cdot 10^{-7}$ s, $4 \cdot 10^{-7}$ s, $5.6 \cdot 10^{-7}$ s, $7.2 \cdot 10^{-7}$ s, $8 \cdot 10^{-7}$ s. The axis are scaled according to the length scale $L^\circ = 10^{-4}$ m

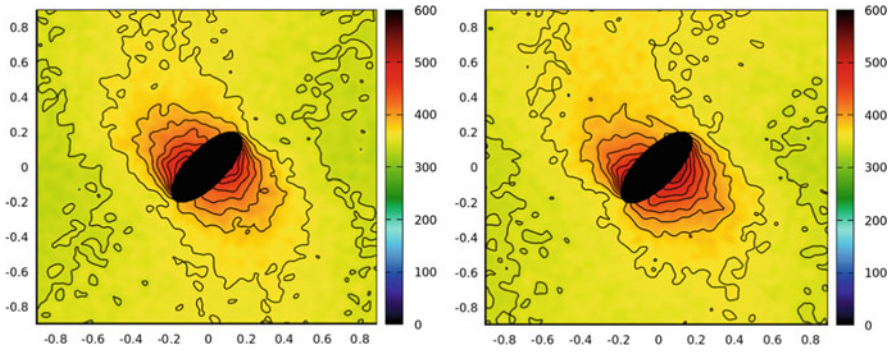

Fig. 4 Temperature (in K) at times $8 \cdot 10^{-7}$ s for scenario 1 (left) and for scenario 2 (right). The axis are scaled according to the length scale $L^\circ = 10^{-4}$ m

comparison at a given time between the kinetic temperature of the gas

$$T(t,x) = \frac{m}{2k_B\rho(t,x)}\left(\int_{\mathbb{R}^3} f(t,x,v)v^2\mathrm{d}v - \left|\frac{1}{\rho(t,x)}\int_{\mathbb{R}^3} f(t,x,v)v\mathrm{d}v\right|^2\right)$$

in scenarios 1 and 2. Here the macroscopic speed of the gas (which is much smaller than the kinetic velocity of molecules) does not have much influence on the temperature.

4.2.2 Scenarios 3 and 4

In scenarios 3 and 4, two particles are crossing each other with opposite velocities: $u_p^1 = (0, 2u^{\text{in}})$, and $u_p^2 = (0, -2u^{\text{in}})$. The gas has a macroscopic velocity equal to $(-u^{\text{in}}, 0)$. In scenario 3 the dust particles have no rotational velocity, whereas in scenario 4 they have rotational velocities equal to $\Omega^1 = 2\pi \cdot 10^6$ rad/s and $\Omega^2 = -\pi \cdot 10^6$ rad/s. Figures 5 and 6 show the evolution of the number density $\rho(t,x) = \int_{\mathbb{R}^3} f(t,x,v)\mathrm{d}v$ of scenario 3 and 4 respectively, and Figure 7 shows the time evolution of the kinetic temperature of the gas in scenario 4.

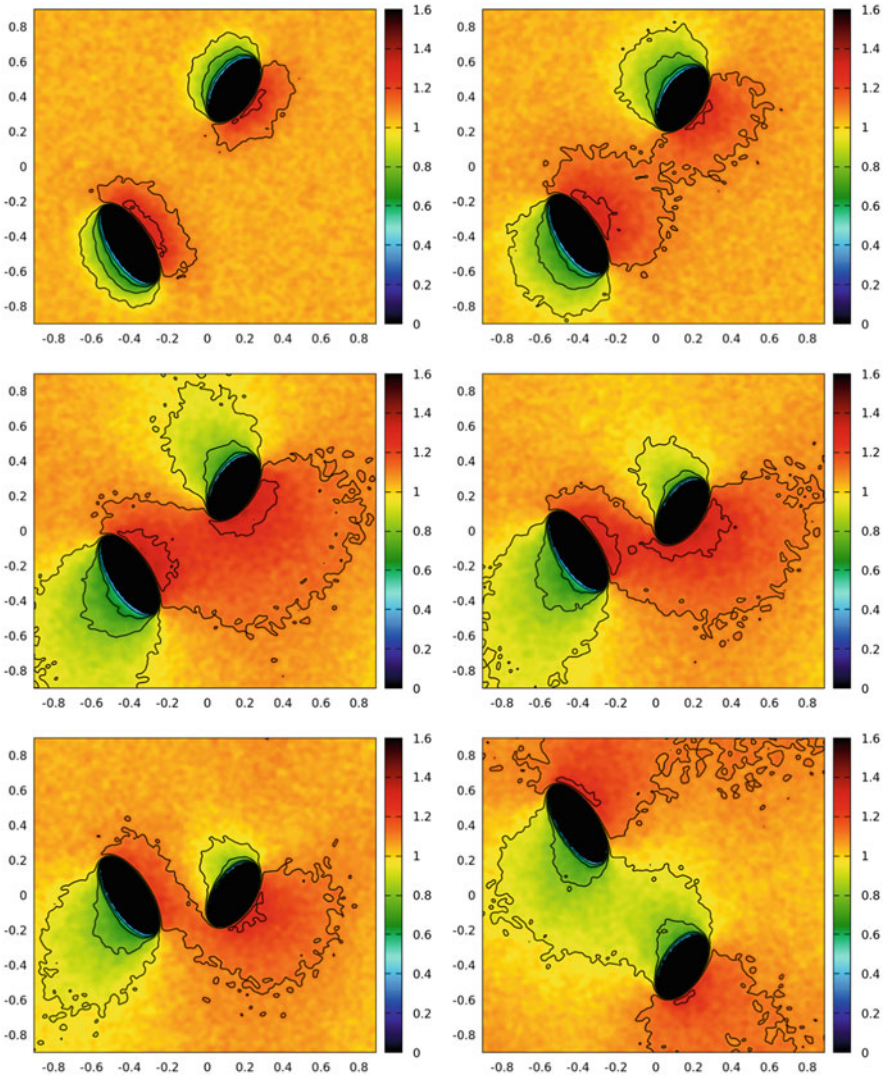

Fig. 5 Time history of the gas density, from left to right and from top to bottom, in Scenario 3 at times $8 \cdot 10^{-8}$ s, $1.6 \cdot 10^{-7}$ s, $3.6 \cdot 10^{-7}$ s, $5.6 \cdot 10^{-7}$ s, $7.2 \cdot 10^{-7}$ s, $1.32 \cdot 10^{-6}$ s. The axis are scaled according to the length scale $L^{\circ} = 10^{-4}$ m

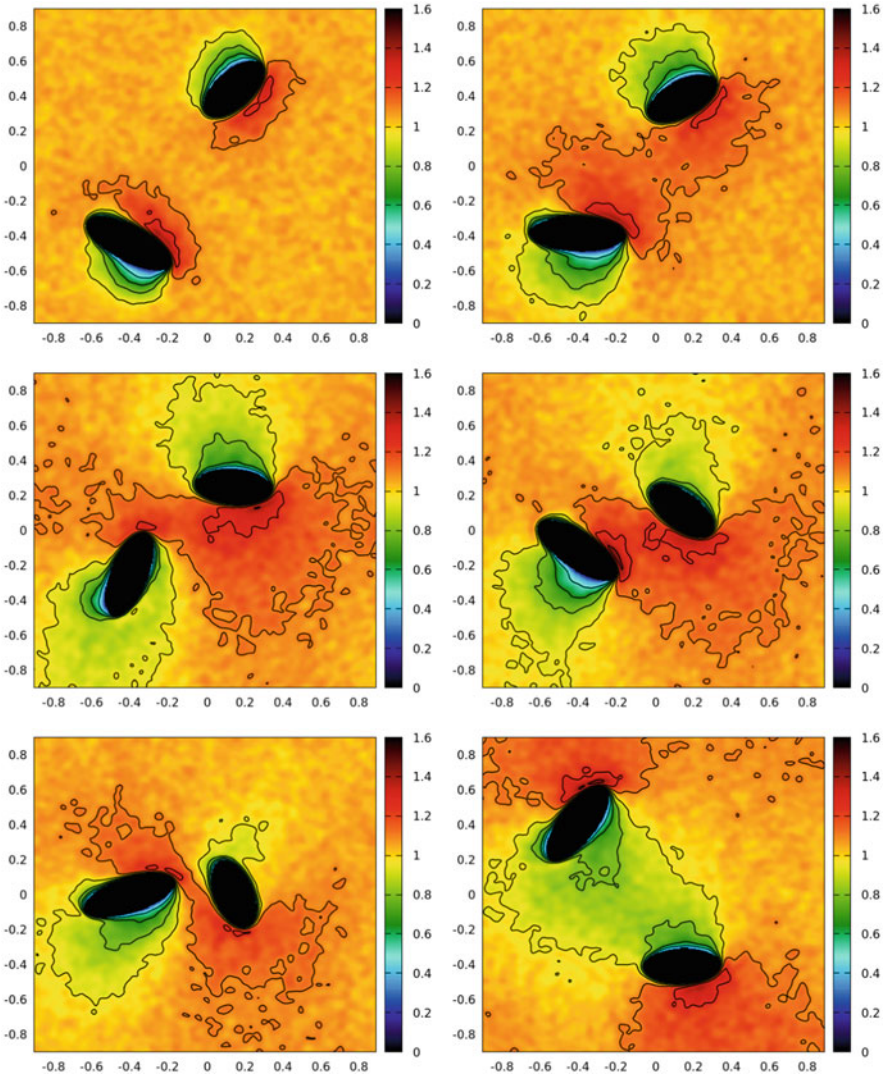

Fig. 6 Time history of the gas density, from left to right and from top to bottom, in Scenario 4 at times $8 \cdot 10^{-8}$ s, $1.6 \cdot 10^{-7}$ s, $3.6 \cdot 10^{-7}$ s, $5.6 \cdot 10^{-7}$ s, $7.2 \cdot 10^{-7}$ s, $1.32 \cdot 10^{-6}$ s. The axis are scaled according to the length scale $L^{\circ} = 10^{-4}$ m

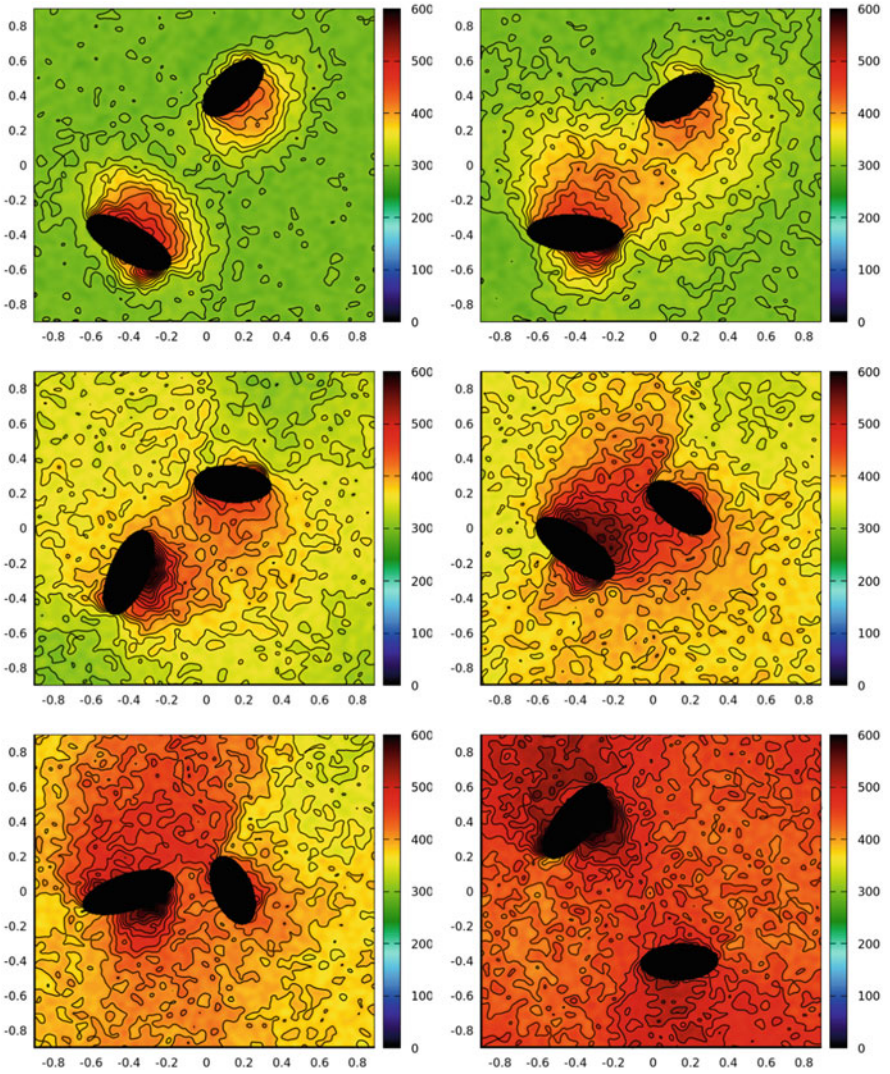

Fig. 7 Time history of the temperature (in K) in Scenario 4, from left to right and from top to bottom, at times $8 \cdot 10^{-8}$ s, $1.6 \cdot 10^{-7}$ s, $3.6 \cdot 10^{-7}$ s, $5.6 \cdot 10^{-7}$ s, $7.2 \cdot 10^{-7}$ s, $1.32 \cdot 10^{-6}$ s. The axis are scaled according to the length scale $L^\circ = 10^{-4}$ m

References

1. Allaire, G., Blanc, X., Despres, B., Golse, F.: Transport et Diffusion. Éditions de l'École de Polytechnique (2018)
2. Aoki, K., Golse, F.: On the speed of approach to equilibrium for a collisionless gas. Kinet. Relat. Models **4**(1), 87–107 (2011)

3. Cavallaro, G., Marchioro, C.: On the motion of an elastic body in a free gas. Rep. Math. Phys. **69**(2), 251–264 (2012)
4. Cercignani, C.: Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations, vol. 21. Cambridge University Press, Cambridge (2000)
5. F. Charles, Kinetic Modelling and Numerical Simulations using Particle Methods for the Transport of Dust in a Rarefied Gas. In: Proceeding of the 26th Symposium Rarefied Gas Dynamics, vol. 1084, pp. 409–414. American Institute of Physics, College Park (2009)
6. Charles, F., Copol, C., Dellacherie, S., Mounsamy, J.-M.: Numerical simulation by a random particle method of deuterium-tritium reactions in a plasma. ESAIM Proc. **38**, 220–240 (2012)
7. Charles, F., Salvarani, F.: Mathematical and numerical study of a dusty Knudsen gas mixture. Acta Appl. Math. **168**, 17–31 (2020)
8. De Vuyst, F., Salvarani, F.: GPU-accelerated numerical simulations of the Knudsen gas on time-dependent domains. Comput. Phys. Commun. **184**(3), 532–536 (2013)
9. Dechristé, G., Mieussens, L.: Numerical simulation of micro flows with moving obstacles. In: Journal of Physics: Conference Series, vol. 362, p. 012030. IOP Publishing, Bristol (2012)
10. Laurendeau, N.M.: Statistical Thermodynamics: Fundamentals and Applications. Cambridge University Press, Cambridge (2005)
11. Ostmo, S., Frezzotti, A., Ytrehus, T.: Kinetic theory study of steady evaporation from a spherical condensed phase containing inert solid particles. Phys. Fluids **9**(1), 211–225 (1997)
12. Russo, G., Filbet, F.: Semi-Lagrangian schemes applied to moving boundary problems for the BGK model of rarefied gas dynamics. Kinet. Relat. Models **2**(1), 231–250 (2009).
13. Shrestha, S., Tiwari, S., Klar, A., Hardt, S.: Numerical simulation of moving rigid bodies in rarefied gases. J. Comp. Phys **292**, 239–252 (2015)
14. Sone, Y.: Molecular Gas Dynamics. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser, Boston (2007). Theory, techniques, and applications

Body-Attitude Alignment: First Order Phase Transition, Link with Rodlike Polymers Through Quaternions, and Stability

Amic Frouvelle

Abstract We present a simple model of alignment of a large number of rigid bodies (modeled by rotation matrices) subject to internal rotational noise. The numerical simulations exhibit a phenomenon of first order phase transition with respect the alignment intensity, with abrupt transition at two thresholds. Below the first threshold, the system is disordered in large time: the rotation matrices are uniformly distributed. Above the second threshold, the long time behaviour of the system is to concentrate around a given rotation matrix. When the intensity is between the two thresholds, both situations may occur. We then study the mean-field limit of this model, as the number of particles tends to infinity, which takes the form of a nonlinear Fokker–Planck equation. We describe the complete classification of the steady states of this equation, which fits with numerical experiments. This classification was obtained in a previous work by Degond, Diez, Merino-Aceituno and the author, thanks to the link between this model and a four-dimensional generalization of the Doi–Onsager equation for suspensions of rodlike polymers interacting through Maier–Saupe potential. This previous study concerned a similar equation of BGK type for which the steady-states were the same. We take advantage of the stability results obtained in this framework, and are able to prove the exponential stability of two families of steady-states: the disordered uniform distribution when the intensity of alignment is less than the second threshold, and a family of non-isotropic steady states (one for each possible rotation matrix, concentrated around it), when the intensity is greater than the first threshold. We also show that the other families of steady-states are unstable, in agreement with the numerical observations.

A. Frouvelle (✉)
CEREMADE, CNRS, Université Paris-Dauphine, Université PSL, Paris, France

Laboratoire de Mathématiques et Applications (LMA), CNRS, Université de Poitiers, UMR 7348, Poitiers, France
e-mail: frouvelle@ceremade.dauphine.fr; amic.frouvelle@math.univ-poitiers.fr

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_7

1 Introduction

The mathematical study of active matter, such as aligning self-propelled particles, is now a well established field of research, inspired for instance by phase transition phenomena that appear in the Vicsek model [4, 25]. Following the kinetic approach introduced in [13], a simple model of alignment of unit vectors subject to internal rotational noise gives rise to a continuous phase transition at the kinetic level [17]. When the alignment intensity (that we call ρ, since it is related to the local density ρ of particles in the inhomogeneous version [8], where the unit vectors represent the velocities of self-propelled particles) is below a threshold ρ_c, the only stable steady-state is the uniform distribution on the unit sphere. On the other hand, when $\rho > \rho_c$, this isotropic equilibria becomes unstable and a family of stable equilibria arises: von Mises distributions with concentration parameter depending on ρ, around a given unit vector. When setting the intensity of alignment as a nonlinear function of the order parameter of the system [9], this continuous phase transition may become a discontinuous one (or first order), with hysteresis phenomenon: a second threshold $\rho^* < \rho_c$ appears, the uniform equilibrium distribution being stable for $\rho < \rho_c$ and the concentrated distributions being stable for $\rho > \rho_*$. Around those thresholds, the order parameter cannot vary continuously from a family of equilibria to the other.

Recently, in a work with Degond and Merino-Aceituno [10] we extended the model of self-propelled particles of Degond and Motsch [13] to the case where the orientation of particles are not only given by their velocity (a unit vector) but by their whole body attitude (an orthonormal frame, given by a rotation matrix). Then, still with Degond and Merino-Aceituno, together with Trescases [11] we proposed a similar model based on quaternion representation for rotation matrices, and the models appeared to be equivalent. In these models, the interaction was normalized and no phenomenon of phase transition could occur, but we remarked that the non-normalized version may lead to such a phenomena. Finally, with Degond, Diez and Merino-Aceituno [7] we managed to treat this phenomenon of phase transition in a homogeneous Bhatnagar–Gross–Krook (BGK) model, thanks to this link with unit quaternions and an analogy with a four-dimensional generalization of the Doi–Onsager equation for suspensions of rodlike polymers interacting through Maier–Saupe potential. Indeed, the compatibility equation we need to solve to determine the possible steady-states can be reformulated in this quaternionic formulation, and leads to a compatibility equation for the Maier–Saupe potential in dimension 4, which was solved in [27]. We obtain a discontinuous phase transition with two thresholds $\rho^* < \rho_c$, still with the same two types of stable equilibria: the uniform distribution for $\rho < \rho_c$, and a family of generalized von Mises distributions, concentrated around a given rotation matrix when $\rho > \rho^*$.

The aim of this paper is twofold. We first want to introduce the model of alignment of rigid bodies through numerical simulations of the particle system, in order to present the first order phase transition that we observe numerically. And then we want to provide a rigorous mathematical description of this phase transition

phenomenon, at the kinetic level: the mean-field limit of the particle system when the number of particles is large is given by a nonlinear Fokker–Planck equation, for which the steady states are the same as those characterized in [7] for the BGK equation. The main result of this article is that we have a fine description of the long-time behaviour of the solution to the Fokker–Planck equation: we classify all the families of equilibria regarding their stability, and prove the exponential stability of the uniform equilibrium when $\rho < \rho_c$ and of the concentrated von Mises distributions when $\rho > \rho^*$.

In Sect. 2, we present the framework of our model: a system of coupled stochastic differential equations for N matrices in $SO_3(\mathbb{R})$. We present a time discretization scheme of Euler–Naruyama type, and provide numerical simulations which illustrate the phenomenon of first order phase transition. In Sect. 3, we describe the mean-field limit of this system, which takes the form of a nonlinear Fokker–Planck equation. We give general results on the behaviour of the solution of this evolution equation, and we show that the determination of its steady states amounts to solve a matrix compatibility equation. Thanks to the free energy associated to the Fokker–Planck equation, the uniform equilibria is shown to be unstable for $\rho > \rho_c = 6$, proving that in that case there are others solutions than 0 for the compatibility equation. Section 4 is a summary of the results of [7] to solve this compatibility equation: we present the link between rotation matrices and unit quaternions, and the fact that the compatibility equation can be transformed to a compatibility equation for Q-tensors which was solved in [27]. We therefore get a precise description of all the steady-states of the equation, and a way to obtain the second threshold ρ^* (as the minimum of a one-dimensional function) such that for $\rho \geqslant \rho^*$ there exists non-trivial steady-states. In Sect. 5, we summarize the results of [7] regarding the stability of these equilibria in the framework a BGK equation (which shares the same steady-states), and we are able to use these results to obtain the classification of the steady-states, as critical points of the free energy. In particular we show that three families of equilibria are unstable, and the remaining two other types are local minimizers of the free energy: the uniform distribution when $\rho < \rho_c$ and the concentrated von Mises distributions for $\rho > \rho^*$. Finally, Sect. 6 is devoted to the main new result of this paper: the exponential stability of these two types of steady-states. In Theorem 3, we prove that if a function f_0 is sufficiently close to the set of equilibria (in relative entropy), then there exist such an equilibrium f_∞ such that the solution of the Fokker–Planck equation converges exponentially fast towards f_∞ (still in relative entropy). We finish this last section by some comments and perspectives.

2 Numerical Evidence of a First-Order Phase Transition in a System of Interacting Particles

2.1 A Simple SDE on $SO_3(\mathbb{R})$ and Its Time-Discretization

First of all let us recall some basic facts about $SO_3(\mathbb{R})$.

Definition 1 For any $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \in \mathbb{R}^3$ we denote by $[\mathbf{u}]_\times = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}$ the (antisymmetric) matrix associated to the linear map $\mathbf{v} \in \mathbb{R}^3 \mapsto \mathbf{u} \times \mathbf{v}$ in the canonical basis.

Proposition 1 (Rodrigues' Formula) *Any special orthogonal matrix $A \in SO_3(\mathbb{R})$ can be written as a rotation around an axis in $\mathbb{R}^3$. More precisely, there exists a unique angle $\theta \in [0, \pi]$ and a unit vector $\mathbf{n} \in \mathbb{S}_2$ such that A is the rotation $R(\theta, \mathbf{n})$ of angle θ around the axis directed by $\mathbf{n}$, given by the following formula:*

$$R(\theta, \mathbf{n}) = \exp(\theta[\mathbf{n}]_\times) = \cos\theta\, I_3 + \sin\theta[\mathbf{n}]_\times + (1 - \cos\theta)\mathbf{n}\mathbf{n}^\top. \tag{1}$$

where I_3 is the identity matrix. When $\theta \in (0, \pi)$, the unit vector $\mathbf{n}$ is unique. When $\theta = \pi$ there are two such vectors $\mathbf{n}$, opposite one to the other. And when $\theta = 0$, any unit vector $\mathbf{n}$ can be used.

To introduce the model and some important notations, we first start with a simple stochastic differential equation (SDE) modeling a rotation matrix $A(t) \in SO_3(\mathbb{R})$ trying to align with another fixed rotation matrix $A_0 \in SO_3(\mathbb{R})$, with strength of alignment $\nu > 0$, and subject to angular noise of intensity $\tau > 0$:

$$\mathrm{d}A = -\nu\nabla_A(\tfrac{1}{2}\|A - A_0\|^2)\mathrm{d}t + 2\sqrt{\tau}P_{T_A} \circ \mathrm{d}B_t. \tag{2}$$

To give a meaning to the previous equation, let us describe the terms one by one, from left to right. We need to define a metric on $SO_3(\mathbb{R})$ in order to define the gradient ∇_A. As it is usually the case in $SO_3(\mathbb{R})$, we will take the metric induced by the scalar product in $M_3(\mathbb{R})$ given by

$$A \cdot B = \frac{1}{2}\mathrm{Tr}(AB^\top). \tag{3}$$

One of the reasons to take this metric is that the geodesic distance between a matrix $A \in SO_3(\mathbb{R})$ and its composition by a rotation of angle $\theta \in [0, \pi]$ is exactly θ. Said differently, if $\mathbf{n} \in \mathbb{S}_2$, then the curve $\theta \in \mathbb{R} \mapsto R(\theta, \mathbf{n})A$ given by the formula (1) is a geodesic travelled at unit speed. The other reason is that the map $\mathbf{u} \in \mathbb{R}^3 \mapsto [\mathbf{u}]_\times$ given by Definition 1 is an isometry from $\mathbb{R}^3$ to the antisymmetric matrices (which is the Lie algebra of $SO_3(\mathbb{R})$). The norm $\|A - A_0\|$

in the SDE (2) is the one associated to this scalar product. The operator P_{T_A} is the orthogonal projection on the tangent space of $SO_3(\mathbb{R})$ at A, given by $P_{T_A}H = \frac{1}{2}(H - AH^\top A)$. The notation $\circ$ in the SDE (2) means that it must be understood in the Stratonovich sense, and the Brownian motion B_t is a 3×3 matrix whose entries are independent real standard Brownian motions.[1] This ensures that the matrix A stays on $SO_3(\mathbb{R})$ for all time, and this is the usual way of defining SDEs on manifolds (we refer to [20] for a reference on this topic). Therefore the first term in the right-hand side of (2) may be written $\nu\nabla_A(A \cdot A_0)$ since $\|A\|^2 = \frac{3}{2}$ whenever $A \in SO_3(\mathbb{R})$. Finally, the law $t \mapsto \mu(t, \cdot)$ (with values in $\mathcal{P}(SO_3(\mathbb{R}))$, the set of probability measures on $SO_3(\mathbb{R})$) of such a process satisfies the following Fokker–Planck equation:

$$\partial_t \mu + \nu\nabla_A \cdot (\nabla_A(A \cdot A_0)\mu) = \tau \Delta_A \mu, \tag{4}$$

where $\nabla_A \cdot$ and Δ_A are the divergence and Laplace-Beltrami operators on $SO_3(\mathbb{R})$. Up to a time rescaling, we see that the important parameter is $\kappa = \frac{\nu}{\tau}$, and we can then without loss of generality study the following PDE, obtained by replacing τ by 1 and ν by κ in (4):

$$\begin{aligned} \partial_t \mu &= -\kappa\nabla_A \cdot (\nabla_A(A \cdot A_0)\mu) + \Delta_A \mu \\ &= \nabla_A \cdot \Big[\exp(\kappa\, A \cdot A_0)\nabla_A\Big(\frac{\mu}{\exp(\kappa\, A \cdot A_0)}\Big)\Big]. \end{aligned} \tag{5}$$

In view of the above formulation, we now define the generalized von Mises distribution (a probability measure) on $SO_3(\mathbb{R})$ of parameter $J \in M_3(\mathbb{R})$ by

$$M_J(A) = \frac{1}{\mathcal{Z}(J)} \exp(J \cdot A), \text{ where } \mathcal{Z}(J) = \int_{SO_3(\mathbb{R})} \exp(J \cdot A)\mathrm{d}A, \tag{6}$$

the normalized volume form on $SO_3(\mathbb{R})$ being its Haar probability measure (this comes from invariance of the metric with respect to left or right multiplication by a given rotation matrix). Therefore it is for instance easy to see that $\mathcal{Z}(\kappa A_0)$ only depends on κ when $A_0 \in SO_3(\mathbb{R})$. With this notation, we can multiply the PDE (5) by $\frac{\mu}{M_{\kappa A_0}}$, integrate by parts and take advantage of the fact that the integral of μ on $SO_3(\mathbb{R})$ remains constant in time, to obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_{SO_3(\mathbb{R})} |\tfrac{\mu}{M_{\kappa A_0}} - 1|^2 M_{\kappa A_0}\mathrm{d}A = -\int_{SO_3(\mathbb{R})} \|\nabla_A(\tfrac{\mu}{M_{\kappa A_0}} - 1)\|^2 M_{\kappa A_0}\mathrm{d}A. \tag{7}$$

[1] Note that this does not give a standard Brownian motion on the Euclidean space $M_3(\mathbb{R})$, equipped with this scalar product, but $\widetilde{B}_t = \sqrt{2}B_t$ is such a standard Brownian motion. The SDE for a standard Brownian motion on the manifold, with generator $\frac{1}{2}\Delta_A$, would be $\mathrm{d}A = P_{T_A} \circ \mathrm{d}\widetilde{B}_t$, which explain the choice of $2\sqrt{\tau}$ instead of the usual $\sqrt{2\tau}$ in the SDE (2) so that the Fokker–Planck equation (4) has the simplest coefficients.

Together with a weighted Poincaré inequality on $SO_3(\mathbb{R})$, this shows that the solution to the PDE (5) converges exponentially fast to the von Mises distribution $M_{\kappa A_0}$. Let us remark that when κ is small (strong noise, or weak alignment), this distribution tends to be uniform on $SO_3(\mathbb{R})$, and when κ is large (strong alignment or low level of noise), it is concentrated around the maximizer of $A \mapsto A \cdot A_0$, which is exactly A_0, as expected.

Let us finish this subsection by describing a numerical discretization of the SDE (2). By using the fact that $\nabla_A(A \cdot A_0) = P_{T_A} A_0$, and denoting by Π the orthogonal projection on $SO_3(\mathbb{R})$ (well-defined in a neighborhood of the manifold), a naive projected Euler-Naruyama scheme would read as follows:

$$A(t + \Delta t) \approx \Pi(A(t) + \nu \Delta t \, P_{T_{A(t)}} A_0 + \sqrt{\Delta t}\, 2\sqrt{\tau} P_{T_{A(t)}} \mathcal{N}_9), \tag{8}$$

where $\mathcal{N}_9$ is a three by three matrix whose 9 entries are independent samples of standard Gaussian distribution. One could even remove the projections on the tangent plane and use this model, easy to describe as a starting point: "Start from $A \in SO_3(\mathbb{R})$, move with step $\nu \Delta t$ in the direction of the target A_0, add some noise of intensity $2\sqrt{\tau \Delta t}$ and project the result back on $SO_3(\mathbb{R})$". However, there is a way to avoid sampling 9 entries per step and to take advantage of the Lie group structure of $SO_3(\mathbb{R})$ instead of computing the projection on $SO_3(\mathbb{R})$ (which is the polar decomposition of matrices and may have some cost). Indeed, the right-hand side of the scheme (8) can be written

$$\Pi(I_3 + \tfrac{1}{2}\nu \Delta t \, [A_0 A(t)^\top - A(t) A_0^\top] + \sqrt{\tau \Delta t} [\mathcal{N}_9 A(t)^\top - A(t) \mathcal{N}_9^\top]) A(t).$$

Since a rotation of a standard Gaussian vector is still a standard Gaussian vector, one can see that the matrix $\mathcal{N}_9 A(t)^\top$ is also a matrix whose 9 entries are independent samples of standard Gaussian distribution. Therefore $\mathcal{N}_9 A(t)^\top - A(t)\mathcal{N}_9^\top$ is an antisymmetric matrix whose independent entries are samples of centered Gaussian distribution of variance 2. It is then a matrix of the form $\sqrt{2}[\boldsymbol{\eta}]_\times$ (see Definition 1), where $\boldsymbol{\eta}$ is a standard Gaussian vector in $\mathbb{R}^3$. When H is a small antisymmetric matrix, a consistent approximation to $\Pi(I_3 + H)$ is given by $\exp(H)$ and can be computed thanks to Rodrigues' formula (1). Therefore a numerical scheme consistent with the naive scheme (8) is given by

$$A(t + \Delta t) \approx \exp(\tfrac{1}{2}\nu \Delta t \, [A_0 A(t)^\top - A(t) A_0^\top] + \sqrt{2\tau \Delta t} [\boldsymbol{\eta}]_\times) A(t), \tag{9}$$

where $\boldsymbol{\eta}$ is a standard Gaussian vector in $\mathbb{R}^3$.

2.2 A System of SDEs and Its Numerical Simulations

We are now ready to introduce our model. In the article [10], we considered N individuals located at positions $X_i \in \mathbb{R}^3$ for $1 \leqslant i \leqslant N$ and with body orientations $A_i \in SO_3(\mathbb{R})$, moving at unit speed in the direction of their first

vector $A_i \mathbf{e}_1$ and aligning their orientations with their neighbours, as in the simple SDE (2). This could take the following form:[2]

$$\begin{cases} \mathrm{d}X_k = A_k \mathbf{e}_1 \mathrm{d}t \\ \mathrm{d}A_k = -\sum_{j=1}^{N} \nu_{j,k} \nabla_{A_k} (\frac{1}{2} \|A_k - A_j\|^2) \mathrm{d}t + 2\sqrt{\tau} P_{T_{A_k}} \circ \mathrm{d}B_{t,k}, \end{cases} \tag{10}$$

where $\nu_{j,k}$ is the intensity at which particle k aligns with particle j, and which may depend for instance on the distance $\|X_j - X_k\|$ between the particles. We consider here a much simpler model, homogeneous in space, so we only look at N rotation matrices $(A_i)_{1 \leqslant i \leqslant n} \in SO_3(\mathbb{R})$, with the same intensity $\frac{\rho}{N}$ of alignment between any pair of particles. We are therefore interested in the following system of SDEs:

$$\forall k \in 1 \dots N, \quad \mathrm{d}A_k = \frac{\rho}{N} \sum_{j=1}^{N} P_{T_{A_k}} A_j \mathrm{d}t + 2\sqrt{\tau} P_{T_{A_k}} \circ \mathrm{d}B_{t,k},$$

where we used the fact that $\nabla_A (\frac{1}{2} \|A - A_0\|^2) = -\nabla_A (A \cdot A_0) = -P_{T_A} A_0$.

In this model, when all the rotation matrices are close to a given one A_0, the behaviour of the system can be expected to be similar to the one of the simple SDE (2), and we may expect the matrices to concentrate if the alignment intensity ρ is high (or τ is low). Conversely, if they are not concentrated around some target, the average of the alignment forces is small and the noise level may prevent the matrices to align if ρ is low (or τ is high). From now on, up to rescaling time (and dividing ρ by τ), we consider the case $\tau = 1$ and we denote by J the average "flux", so our system has the following form:

$$\begin{cases} \mathrm{d}A_k = P_{T_{A_k}} J \mathrm{d}t + 2 P_{T_{A_k}} \circ \mathrm{d}B_{t,k}, \quad (1 \leqslant k \leqslant N) \\ J(t) = \frac{\rho}{N} \sum_{j=1}^{N} A_j(t). \end{cases} \tag{11}$$

We are then interested in the different behaviours of the system (11) for different values of ρ. One way to measure how much matrices are concentrated is to compute the variance $\langle \|A - \langle A \rangle\|^2 \rangle$ (where for any function h, we write $\langle h(A) \rangle = \frac{1}{N} \sum_{j=1}^{N} h(A_j)$). This nonnegative quantity is equal to $\langle \|A\|^2 \rangle - \|\langle A \rangle\|^2 = \frac{3}{2} - \|\frac{J}{\rho}\|^2$, which implies that if we define the order parameter $c(t)$ by

$$c(t) = \frac{\sqrt{2}}{\sqrt{3}\rho} \|J(t)\|, \tag{12}$$

[2] Actually, the model studied in [10] (which does not present the phenomenon of phase transition we are studying here) is a little bit more involved: each particle first chose an average target and aligns with it, instead of averaging the "forces of alignment" as it is the case in the system of SDEs (10).

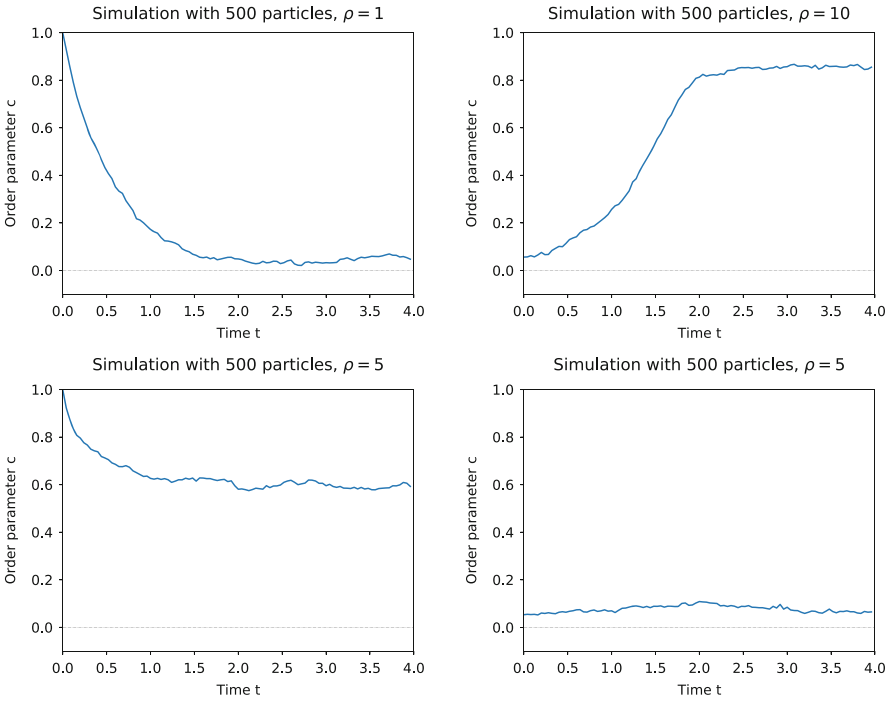

Fig. 1 Time evolution of the order parameter in four situations

we obtain a quantity between 0 (when the variance is maximal) and 1 (the variance is 0, all matrices are the same). To give a numerical illustration of the phenomenon we are interested in, we use a scheme similar to the scheme (9) of the previous subsection: we take N matrices $A_k \in SO_3(\mathbb{R})$ for $1 \leqslant k \leqslant N$), a time step Δt, and at each time iteration, we compute $J = \frac{\rho}{N} \sum_{j=1}^{N} A_j$ and we update each A_k for $1 \leqslant k \leqslant N$ with the matrix

$$\exp(\tfrac{1}{2}\Delta t\,[J A_k^\top - A_k J^\top] + \sqrt{2\Delta t}[\boldsymbol{\eta}_k]_\times) A_k,$$

where $(\boldsymbol{\eta}_k)_{1 \leqslant k \leqslant N}$ are independent samples of a standard Gaussian vector in $\mathbb{R}^3$.

Figure 1 depicts the time evolution of the order parameter $c(t)$ given by the formula (12) for two realisations of this numerical scheme. In both cases the number of particles is $N = 500$, the time step is $\Delta t = 0.04$ and we run the simulation for 100 time iterations. In the top-left part of Fig. 1 where we took $\rho = 1$, even if we started with all the particles in the same position (order parameter equal to 1), as time evolves, the order parameter becomes very small. In the top-right part, with $\rho = 10$, even if the particles were uniformly sampled on $SO_3(\mathbb{R})$ (order parameter close to 0), as time evolves, the order parameter stabilizes around a quite high value, indicating that the matrices are concentrated around a given rotation

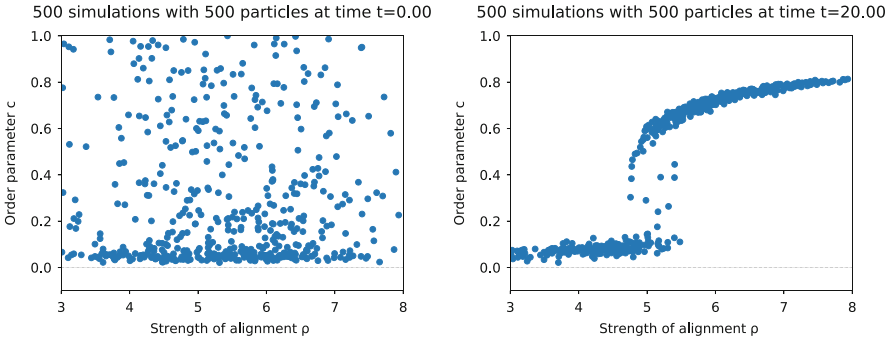

Fig. 2 Numerical illustration of a first-order phase transition

matrix. This indicates that a phase transition phenomenon is occurring with respect to the parameter ρ. However, for some intermediate values of ρ, as in the bottom part of Fig. 1 where $\rho = 5$, two different behaviours may happen: starting with concentrated particles lead to an order parameter stabilizing around a non-zero value, while the configuration starting with particles uniformly sampled on $SO_3(\mathbb{R})$ stays with an order parameter close to 0 as time evolves.

In order to obtain a more precise illustration of this phenomenon, we ran 500 such simulations with various values of the parameter ρ and different initial conditions,[3] still with $N = 500$ and $\Delta t = 0.04$, for 500 time iterations. Figure 2 depicts the initial order parameters c and strengths ρ, and their value after 500 iterations ($t = 20$). We clearly see two thresholds for ρ. The first threshold that we will denote ρ^*, is such that for all simulations with $\rho < \rho^*$, the order parameter seems to be close to 0 for large times. The second threshold, that we will denote ρ_c (with $\rho^* < \rho_c$), is such that for all simulations with $\rho > \rho_c$, the order parameter does not stay close to 0 for large times, and stabilizes around a quite high value. In the intermediate regime $\rho^* < \rho < \rho_c$, both behaviours occur. This is what is called first-order (or discontinuous) phase transition: the order parameter does not vary continuously when going from one behaviour to the other.

The aim of the next sections is to present a rigorous mathematical description of this phenomenon in the framework of a kinetic equation corresponding to the limiting behaviour of the system of SDEs (11) when $N \to \infty$.

[3] For a better illustration, the parameter ρ and the initial order parameter c are not uniformly sampled, in order to see more points in the region of interest.

3 Mean-Field Limit and Compatibility Equation

Let us first consider the first part of the system (11), as if $t \mapsto J(t)$ was a prescribed regular function from $\mathbb{R}$ to $M_3(\mathbb{R})$:

$$\mathrm{d}A = P_{T_A} J \,\mathrm{d}t + 2 P_{T_A} \circ \mathrm{d}B_t. \tag{13}$$

As before for the simple SDE (2), the law $t \mapsto \mu(t,\cdot)$ of such a stochastic process would satisfy the following (linear) Fokker–Planck equation:

$$\partial_t \mu = -\nabla_A \cdot (\mu\, P_{T_A} J) + \Delta_A \mu = \nabla_A \cdot \left[M_J(A) \nabla_A \left(\frac{\mu}{M_J(A)} \right) \right], \tag{14}$$

where the definition of the generalized von Mises distribution M_J is given by the formula (6). Let us now suppose that several such processes A_k satisfying the SDE (13) were independently drawn, with different independent Brownian motions $B_{t,k}$, and independent initial conditions following a probability measure μ_0 on SO_3. Their law at time t would be given by $\mu(t,\cdot)$, solution of the Fokker–Planck equation (14) with initial condition μ_0 by the law of large numbers the average $\frac{1}{N}\sum_{k=1}^N A_k(t)$ would converge to the expectation of one of this process, that we call $\mathcal{J}[\mu(t,\cdot)]$. More generally, we define $\mathcal{J}[f]$ for any finite measure f on $SO_3(\mathbb{R})$ (not necessarily a probability measure, it may also be a signed measure):

$$\mathcal{J}[f] = \int_{SO_3(\mathbb{R})} A\, f(A) \mathrm{d}A. \tag{15}$$

To deal with the system (11), where $J(t) = \frac{\rho}{N}\sum_{k=1}^N A_k(t)$ is not prescribed but depends on all the particles, we cannot expect the particles A_k to behave independently. However one can show that in the limit $N \to \infty$, their behaviour is close to independent particles. This is called the propagation of chaos property, and we refer to [24] for an introduction on this subject. One of the typical results in this theory is that the empirical measure of the particle system converges to a solution to the (now nonlinear) Fokker–Planck equation corresponding to (14) with $J(t) = \rho \mathcal{J}[\mu(t,\cdot)]$:

Proposition 2 *If $A_{k,0}$ are independent random rotation matrices distributed according to the probability measure μ_0, then the empirical measure associated to the solution of the system of SDEs (11), given by $\mu^N(t) = \frac{1}{N}\sum_{k=1}^N \delta_{A_k(t)}$, converges (in Wasserstein distance) to the solution μ of the following nonlinear Fokker–Planck equation, with initial condition μ_0:*

$$\partial_t \mu = -\rho\, \nabla_A \cdot (\mu\, P_{T_A} \mathcal{J}[\mu]) + \Delta_A \mu. \tag{16}$$

The convergence is uniform on $[0,T]$ for all $T > 0$.

Proof We will not provide the proof in detail here, as it follows the classical theory of propagation of chaos for coupled drift-diffusion processes, but we will recall some important steps. It has to be adapted to the framework of SDEs on a manifold, but this is not a real problem in this compact case (see for instance [3] in the case of the Vicsek model on the sphere). Let us recall the coupling argument such as the one in [24]. We start by proving the well-posedness of this following SDE (the coupling process):

$$\begin{cases} \mathrm{d}A = \rho\, P_{T_A}\mathcal{J}[\mu]\mathrm{d}t + 2P_{T_A}\circ \mathrm{d}B_t, \\ \mu(t,\cdot) \text{ is the law of } A(t). \end{cases} \tag{17}$$

The proof of this well-posedness, seen as a fixed point problem (either for the function $J(t) = \rho\mathcal{J}[\mu]$ or directly on the law μ) is done thanks to a Picard iteration which leads to a contraction in the appropriate Wasserstein metric.

We then construct independent solutions to this coupling process $\overline{A}_k$ with independent Brownian motions $B_{t,k}$ and initial conditions $A_{k,0}$: the same as the Brownian motions and initial conditions used for the original system of SDEs (11). All these processes $\overline{A}_k$ have the same law, which is the solution μ of the Fokker–Planck equation (16) starting with μ_0. By the law of large numbers, the empirical distribution $\overline{\mu}^N$ of the coupling processes converges to μ, and therefore it is enough to estimate the distance between $\overline{\mu}^N$ and μ^N. This can be done by obtaining estimates of the form

$$\mathbb{E}[\|A_k - \overline{A}_k\|^2] \leqslant \frac{\exp(CT)}{N}, \tag{18}$$

for all $1 \leqslant k \leqslant N$, which gives control on the 2-Wasserstein distance between $\overline{\mu}^N$ and μ^N on the time interval $[0, T]$. □

We now want to study the long time behaviour of the nonlinear Fokker–Planck equation (16), that we will rewrite in function of $f = \rho\,\mu$ (in that case, ρ represents the total "mass" of f). Since $\rho\mathcal{J}[\mu] = \mathcal{J}[f]$, it therefore has the following form, without any parameter on the equation:

$$\partial_t f = -\nabla_A \cdot (f\, P_{T_A}\mathcal{J}[f]) + \Delta_A f. \tag{19}$$

This is an equation of the form $\partial_t f = \mathcal{C}[f]$ where $\mathcal{C}[f]$ can also be written, using the definition (6) of the von Mises distribution M_J, under the following factorized form:

$$\mathcal{C}[f] = \nabla_A \cdot \Big[M_{\mathcal{J}[f]}(A)\nabla_A\Big(\frac{f}{M_{\mathcal{J}[f]}(A)}\Big)\Big].$$

In order to understand the long time behaviour of the solution, let us first look at stationary solutions.

Proposition 3 *A measure f on $SO_3(\mathbb{R})$ is a stationary solution of the Fokker–Planck equation (19) if and only if it is of the form $f = \rho M_J$, where J satisfies the following compatibility equation*

$$J = \rho \mathcal{J}[M_J]. \tag{20}$$

Proof Since we have, by integration by parts,

$$\int_{SO_3(\mathbb{R})} \frac{f}{M_{\mathcal{J}[f]}(A)} \mathcal{C}[f] \mathrm{d}A = -\int_{SO_3(\mathbb{R})} \left\| \nabla_A \left(\frac{f}{M_{\mathcal{J}[f]}(A)} \right) \right\|^2 M_{\mathcal{J}[f]}(A) \mathrm{d}A,$$

we immediately get that if $\mathcal{C}[f] = 0$ then f has to be proportional to $M_{\mathcal{J}[f]}$, and the total mass of f, denoted by ρ, gives the coefficient of proportionality. Then, taking the average on $SO_3(\mathbb{R})$ against A, thanks to the definition (15) of $\mathcal{J}$, we obtain, denoting $J = \mathcal{J}[f]$:

$$J = \mathcal{J}[f] = \mathcal{J}[\rho M_{\mathcal{J}[f]}] = \rho \mathcal{J}[M_J],$$

which is the compatibility equation for J. Conversely, if J is a fixed point of this map $J \mapsto \rho \mathcal{J}[M_J]$, then setting $f = \rho M_J$, we get $\mathcal{J}[f] = J$, and then $\mathcal{C}[f] = 0$. □

Before obtaining a simple characterization of the solutions of the compatibility equation (20), which is the object of the next section, let us give some more results on the solutions to the Fokker–Planck equation (19).

Proposition 4 *For all nonnegative measure f_0 on $SO_3(\mathbb{R})$, with a total mass $\rho > 0$, there exists a unique weak solution f to the nonlinear Fokker–Planck equation (19) such that $f(t, \cdot)$ converges to f_0 (in Wasserstein distance) as $t \to 0$. This solution belongs to $C^\infty((0, +\infty), SO_3(\mathbb{R}))$ and is positive for any positive time. Furthermore, we have the following uniform estimates in time: for all $t_0 > 0$, and $s \in \mathbb{R}$, the solution f is uniformly bounded on $[t_0, +\infty)$ in the Sobolev space $H^s(SO_3(\mathbb{R}))$.*

The proof of this proposition can be obtained through simple energy estimates in $H^s(SO_3(\mathbb{R}))$, using Poincaré inequalities for high modes and the fact that the low modes are uniformly bounded in time. Indeed, the nonlinearity in the Fokker–Planck equation (19) is only through $\mathcal{J}[f]$, which is uniformly bounded thanks to its definition (15) and the fact that $SO_3(\mathbb{R})$ is compact, together with the fact that the total mass ρ is preserved. The positivity comes from the maximum principle. We refer to [17] to a detailed proof of such results on the unit sphere instead of $SO_3(\mathbb{R})$, for which all the arguments may be used similarly.

Let us now describe the free energy associated to this Fokker–Planck equation, which may be rewritten

$$\partial_t f = \nabla_A \cdot \big(f \, \nabla_A (\ln f - A \cdot \mathcal{J}[f]) \big).$$

Multiplying by $\ln f - A \cdot \mathcal{J}[f]$ and integrating over $SO_3(\mathbb{R})$, the left-hand side of the equality can be seen as a time derivative, and the right-hand side can be integrated by parts, to obtain the following dissipation relation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[f] + \mathcal{D}[f] = 0, \tag{21}$$

where

$$\mathcal{F}[f] = \int_{SO_3(\mathbb{R})} f(A) \ln f(A) \mathrm{d}A - \frac{1}{2}\|\mathcal{J}[f]\|^2, \tag{22}$$

$$\mathcal{D}[f] = \int_{SO_3(\mathbb{R})} f(A) \|\nabla_A (\ln f - A \cdot \mathcal{J}[f])\|^2 \mathrm{d}A. \tag{23}$$

We can then prove, as in [17] that being a stationary state of the Fokker–Planck equation (see Proposition 3) is equivalent to be a critical point of $\mathcal{F}$ under the constraint of mass ρ, and that is also equivalent to be a function with no dissipation ($\mathcal{D}[f] = 0$).

We then have a decreasing free energy $\mathcal{F}[f]$, and thanks to a kind of LaSalle's principle, we obtain that the solution converges to a set of equilibria:

Proposition 5 *Let f_0 be a nonnegative measure on $SO_3(\mathbb{R})$ with mass $\rho > 0$. We denote by $\mathcal{F}_\infty$ the limit of $\mathcal{F}[f(t,\cdot)]$ as $t \to +\infty$, where f is the solution to the Fokker–Planck equation (19) with initial condition f_0. Then the set of equilibria $\mathcal{E}_\infty$, given by*

$$\mathcal{E}_\infty = \{\rho M_J \text{ such that } J = \rho \mathcal{J}[M_J] \text{ and } \mathcal{F}[\rho M_J] = \mathcal{F}_\infty\},$$

is not empty. Furthermore, the solution f converges in any Sobolev space H^s to this set of equilibria in the following sense:

$$\lim_{t\to\infty} \inf_{g\in\mathcal{E}_\infty} \|f(t,\cdot) - g\|_{H^s} = 0.$$

Once more, the proof of this proposition follows exactly the one given in [17]. The important point of this proposition is that once the structure of the solutions of the compatibility equation (20) is known (which is the aim of the next section), it gives a lot of information on the large time behaviour of the solutions to the Fokker–Planck equation.

Before giving a precise description of these solutions, let us remark that $J = 0$ is always a solution to the compatibility equation, since $\mathcal{J}[\rho] = 0$, therefore the uniform distribution with mass ρ is a steady-state. We want to expand the free

energy $\mathcal{F}$ around this steady-state. We will need the following lemma (Lemma 3.3 of [7]):

Lemma 1 *For all $J \in M_3(\mathbb{R})$,*

$$\int_{SO_3(\mathbb{R})} (J \cdot A) A \, \mathrm{d}A = \frac{1}{6} J. \tag{24}$$

Consequently, if f is a finite measure, the orthogonal projection of f on the space of functions of the form $A \mapsto J \cdot A$ for $J \in M_3(\mathbb{R})$ is $A \mapsto 6\mathcal{J}[f] \cdot A$. Now, let us take a nonnegative measure f with mass ρ, we write $J = \mathcal{J}[f]$ and $g(A) = 6\, J \cdot A$. We suppose that $\|J\|$ is sufficiently small, so that $\rho + g > 0$ on $SO_3(\mathbb{R})$. We write $h = f - \rho - g$, so h is a finite measure with zero average and $\mathcal{J}[h] = 0$. Then we obtain, by convexity of $x \mapsto x \ln x$ on $\mathbb{R}_+$:

$$\begin{aligned}
\mathcal{F}[f] &\geqslant \int_{SO_3(\mathbb{R})} [(\rho + g(A)) \ln(\rho + g(A)) + h(A)(\ln(\rho + g(A)) + 1)] \, \mathrm{d}A - \tfrac{1}{2}\|J\|^2 \\
&\geqslant \mathcal{F}[\rho + g] + \int_{SO_3(\mathbb{R})} h(A)\Big(1 + \ln \rho + \frac{g(A)}{\rho}\Big) \mathrm{d}A \\
&\qquad\qquad - O(\|g\|_\infty^2) \int_{SO_3(\mathbb{R})} |h(A)| \mathrm{d}A. \\
&\geqslant \mathcal{F}[\rho + g] - O(\|J\|^2)\Big(\int_{SO_3(\mathbb{R})} |f(A) - \rho| \, \mathrm{d}A + O(\|J\|)\Big).
\end{aligned} \tag{25}$$

Next we compute

$$\begin{aligned}
\mathcal{F}[\rho + g] &= \rho \ln \rho + \frac{1}{2\rho} \int_{SO_3(\mathbb{R})} (6A \cdot J)^2 \mathrm{d}A + O(\|g\|_\infty^3) - \tfrac{1}{2}\|J\|^2 \\
&= \mathcal{F}[\rho] + \frac{6 - \rho}{2\rho} \|J\|^2 + O(\|J\|^3),
\end{aligned} \tag{26}$$

thanks to Lemma 24. We therefore see that the sign of $6 - \rho$ plays a role to study the nature, as a critical point of $\mathcal{F}$, of the uniform distribution of mass ρ:

Proposition 6 *We set $\rho_c = 6$.*

- *If $\rho < \rho_c$, then the uniform distribution with mass ρ is a local strict minimizer of the free energy $\mathcal{F}$ under the constraint of total mass ρ.*
- *If $\rho > \rho_c$, the uniform distribution with mass ρ is not a local minimizer of the free energy $\mathcal{F}$ under the constraint of total mass ρ.*

Proof When $\rho > 6$, it is clear thanks to (25) and (26) that when $J \neq 0$, if $\|J\|$ and $\int_{SO_3(\mathbb{R})} |f - \rho|$ are sufficiently small, then $\mathcal{F}[f] > \mathcal{F}[\rho]$. If $J = 0$ but $f \neq \rho$, then by strict convexity of $x \mapsto x \ln x$ on $\mathbb{R}_+$, we get

$$\mathcal{F}[f] = \int_{SO_3(\mathbb{R})} f(A) \ln f(A) \, \mathrm{d}A > \int_{SO_3(\mathbb{R})} (\rho \ln \rho + (f(A) - \rho) \ln \rho) \, \mathrm{d}A = \mathcal{F}[\rho].$$

The second point follows directly from (26). □

This last proposition gives an insight on the stability of the uniform steady-state (we will indeed see later that this uniform steady-state is isolated). In summary, we have shown that there is a phenomenon of phase transition at the threshold $\rho = \rho_c$, and we know thanks to Proposition 5 that there must exist other types of steady-states, at least when $\rho > \rho_c$. We are now ready to give a precise description of those non-isotropic equilibria.

4 Link with Higher Dimensional Polymers, Solutions to the Compatibility Equation

This section is the summary of the results we obtained in [7] to solve the compatibility equation (20) (in a slightly different context, see Sect. 5), therefore we will omit the proofs.

Let us first recall some definitions. We denote by $\mathbb{H}$ the set of quaternions: objects of the form $q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, where $(a, b, c, d) \in \mathbb{R}^4$ and the imaginary quaternions satisfy $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$. For such a quaternion q, we denote by $q^* = a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$ its conjugate. If we identify the Euclidean space $\mathbb{R}^4$ with $\mathbb{H}$, it satisfies $|q|^2 = a^2 + b^2 + c^2 + d^2 = qq^* = q^*q$. We denote then by $\mathbb{H}_1$ the set of units quaternions: those for which $|q|^2 = 1$.

We say that a quaternion q of the previous form is purely imaginary if its real part a is zero. It allows now to identify $\mathbb{R}^3$ with the set of purely imaginary quaternions. We will use boldface letters when using this identification.

The first proposition is a link between $SO_3(\mathbb{R})$ and $\mathbb{H}_1/\{\pm 1\}$.

Proposition 7 *For any $q \in \mathbb{H}_1$, the linear map $\mathbf{u} \mapsto q\mathbf{u}q^*$ sends purely imaginary quaternions on purely imaginary quaternions of the same norm. It is therefore identified as a rotation of $\mathbb{R}^3$, and the corresponding rotation matrix is denoted $\Phi(q)$. Conversely for any rotation matrix $A \in SO_3(\mathbb{R})$, there exists a unit quaternion q such that $A = \Phi(q)$ (this quaternion is not unique, the only other possibility being $-q$). The map Φ can then be seen as a group isomorphism between $SO_3(\mathbb{R})$ and $\mathbb{H}_1$ (this is actually a local isometry between the manifolds). In practice, the matrix $R(\theta, \mathbf{n})$ given by Rodrigues' formula (1) corresponds to the quaternion $q = \cos(\frac{\theta}{2}) + \sin(\frac{\theta}{2})\mathbf{n}$ (remember that vectors in $\mathbb{R}^3$ are seen as purely imaginary quaternions, and remark that if we replace θ by $\theta + 2\pi$, we get the same rotation matrix, but the opposite quaternion).*

This allows to represent a rotation matrix by a unit quaternion up to multiplication by ± 1. This is reminiscent of describing rodlike polymers as unit vectors up to multiplication by ± 1, but generalized in dimension 4. This analogy was the starting point of our work [11], where we used those unit quaternions for the modeling of alignment of rigid bodies. In the present case, we will see that this analogy will actually be very helpful, by transforming the compatibility equation (20) into another one which has already been solved in [27], in the context of suspensions of diluted polymers.

We denote by $\mathcal{S}_4^0(\mathbb{R})$ the space of symmetric and trace-free matrices of dimension 4, which are called Q-tensors. To a unit quaternion q, we can associate the Q-tensor given by $q \otimes q - \frac{1}{4} I_4$. Remark that two unit quaternions q and $\widetilde{q}$ are associated to the same Q-tensor if and only if $q = \pm \widetilde{q}$ (this is a unit vector in the eigenspace of this Q-tensor associated to the eigenvalue $\frac{3}{4}$, which is one-dimensional). So we have another way to represent unit quaternions up to multiplication by ± 1 in this space. The important fact to notice is that those two embeddings are actually the same, up to a linear isomorphism between the spaces $M_3(\mathbb{R})$ and $\mathcal{S}_4^0(\mathbb{R})$, which has nice properties.

Proposition 8 *There exists a linear isomorphism between the spaces $M_3(\mathbb{R})$ and $\mathcal{S}_4^0(\mathbb{R})$ (both of dimension 9), denoted ϕ, with the following properties:*

$$\forall q \in \mathbb{H}_1, \quad \phi(\Phi(q)) = q \otimes q - \tfrac{1}{4} I_4, \tag{27}$$

$$\forall J \in M_3(\mathbb{R}), \forall q \in \mathbb{H}_1, \quad \frac{1}{2} J \cdot \Phi(q) = q \cdot \phi(J) q, \tag{28}$$

where the map Φ is given by Proposition 7. The dot product in the left-hand side of (28) is the metric in the space $M_3(\mathbb{R})$ given in (3), while the one in the right-hand side is the canonical scalar product of $\mathbb{R}^4$. Furthermore, the isomorphism ϕ preserves the diagonal structure: $J \in M_3(\mathbb{R})$ is diagonal if and only if $\phi(J)$ is diagonal in $\mathcal{S}_4^0(\mathbb{R})$.

The proof of this proposition is done in [7]. The expression (28) is actually the definition of ϕ: the left-hand side is a quadratic form in q (seen as an element of $\mathbb{R}^4$), defined for any unit quaternion, which defines a symmetric bilinear form on all quaternions, the matrix of which is $\phi(J)$. The expression of $\phi(J)$ is given in the appendix of [7], which gives the fact that it is bijective and with values in trace-free matrices, and the provides the property (27). With this isomorphism, we can rewrite the compatibility equation in the framework of Q-tensors. For a finite measure f on $\mathbb{H}_1$, we define its averaged Q-tensor by

$$\mathcal{Q}[f] = \int_{\mathbb{H}_1} f(q)(q \otimes q - \tfrac{1}{4} I_4) \mathrm{d}q.$$

Therefore, thanks to the definition (15) of $\mathcal{J}$ and the fact that Φ is a local isometry, we obtain, for a finite measure f on $SO_3(\mathbb{R})$

$$\phi(\mathcal{J}[f]) = \int_{SO_3(\mathbb{R})} \phi(A) f(A) \mathrm{d}A = \int_{\mathbb{H}_1} \phi(\Phi(q)) f(\Phi(q)) \mathrm{d}q = \mathcal{Q}[f \circ \Phi].$$

Finally, we also define the generalized von Mises associated to $Q \in \mathcal{S}_4^0(\mathbb{R})$ by

$$M_Q(q) = \frac{1}{\mathcal{Z}(Q)} \exp(q \cdot Qq), \text{ where } \mathcal{Z}(Q) = \int_{\mathbb{H}_1} \exp(q \cdot Qq) \mathrm{d}q,$$

where we use the same notation as in (6) for the generalized von Mises on $SO_3(\mathbb{R})$, but it will always be clear following the context which definition is concerned. Using the property (28), it gives $M_J(\Phi(q)) = M_{2\phi(J)}(q)$. Therefore, the compatibility equation (20) becomes, writing $Q = 2\phi(J)$:

$$Q = 2\phi(J) = 2\rho\, \phi(\mathcal{J}[M_J]) = 2\rho\, \mathcal{Q}[M_Q].$$

It happens that this equation is exactly the compatibility equation that we obtain when we try to obtain the steady states of the following Fokker–Planck equation, for a probability measure μ on $\mathbb{H}_1$:

$$\partial_t \mu = -2\rho\, \nabla_q \cdot (\mu \nabla_q (q \cdot \mathcal{Q}[\mu] q)) - \Delta_q \mu.$$

This corresponds to the Smoluchowski (or Doi–Onsager) equation for suspensions of dilute rodlike polymers with Maier–Saupe potential of strength 2ρ, and is nothing else than our Fokker–Planck equation (14), up to a change of variable thanks to the map Φ. It happens that this compatibility equation has been studied a lot in dimension 3 (instead of 4 here), with the independent works [5, 15, 21]. And in the work [27], a unified approach has been proposed, which allows to treat the case of higher dimensional space. The main result is that a solution $Q \in \mathcal{S}_n^0(\mathbb{R})$ of the compatibility equation $Q = \alpha\, \mathcal{Q}[M_Q]$ can have at most two different eigenvalues. In dimension 4, it means that if Q is different from zero, there are only two cases: either one eigenvalue is simple and the other one is triple, or both are double. In the first case, if we take q a unit quaternion in the eigenspace of dimension one, we get that Q is proportional to $q \otimes q - \frac{1}{4} I_4$, which means that $J = \phi^{-1}(Q)$ is proportional to the rotation matrix $\Phi(q)$. And indeed it is possible to see that if A_0 is a rotation matrix and $\alpha \in \mathbb{R}$, then $\mathcal{J}[M_{\alpha A_0}]$ is proportional to A_0, with a coefficient $c_1(\alpha)$ (that can be expressed using an appropriate volume form on $SO_3(\mathbb{R})$ and will be given later on). Therefore the compatibility equation (20) becomes the one-dimensional equation $\alpha = \rho c_1(\alpha)$. For the second case, it is a little bit more subtle, but it

still leads to a one-dimensional equation of the form $\alpha = \rho c_2(\alpha)$. The results are summarized in the following proposition (corresponding to Theorem 5 of [7]):

Proposition 9 *The solutions to the compatibility equation (20) are:*

- *The matrix $J = 0$,*
- *the matrices of the form $J = \alpha A_0$ with $A_0 \in SO_3(\mathbb{R})$ and where $\alpha \in \mathbb{R} \setminus \{0\}$ satisfies the scalar compatibility equation*

$$\alpha = \rho c_1(\alpha), \tag{29}$$

- *the matrices of the form $J = \alpha\sqrt{3}\,\mathbf{a}_0 \otimes \mathbf{b}_0$ where $\mathbf{a}_0$ and $\mathbf{b}_0$ are two unit vectors of $\mathbb{R}^3$ and $\alpha > 0$ satisfies the scalar compatibility equation*

$$\alpha = \rho c_2(\alpha), \tag{30}$$

with the functions c_1 and c_2 given by

$$c_1(\alpha) = \frac{\int_0^\pi \frac{1}{3}(2\cos\theta + 1)\sin^2(\frac{\theta}{2})\exp(\alpha\cos\theta)\mathrm{d}\theta}{\int_0^\pi \sin^2(\frac{\theta}{2})\exp(\alpha\cos\theta)\mathrm{d}\theta},$$

$$c_2(\alpha) = \frac{1}{\sqrt{3}}\frac{\int_0^\pi \cos\varphi\sin\varphi\exp(\frac{\sqrt{3}}{2}\alpha\cos\varphi)\mathrm{d}\varphi}{\int_0^\pi \sin\varphi\exp(\frac{\sqrt{3}}{2}\alpha\cos\varphi)\mathrm{d}\varphi}.$$

Compared to the convention taken in [7], we chose to add the constant $\sqrt{3}$ in the last type of solutions (changing accordingly the expression of $c_2(\alpha)$). The reason is that if J is a solution to the compatibility equation (20), where α satisfies (29) or (30), then $\|J\|^2 = \frac{3}{2}\alpha^2$. The order parameter c associated to the steady state ρM_J by the formula (12) is then equal to $\frac{|\alpha|}{\rho}$ which is $|c_1(\alpha)|$ or $|c_2(\alpha)|$. These functions c_1 and c_2 then provide the values of the order parameter of the considered steady-state. The study of these functions (and more precisely the behaviour of $\frac{\alpha}{c_1(\alpha)}$ and $\frac{\alpha}{c_2(\alpha)}$) is the key to provide a complete description of the possible steady-states. Once more, the following proposition is taken from [7].

Proposition 10 *The functions c_1 and c_2 are both strictly increasing on $\mathbb{R}$ having value 0 at 0. Therefore 0 is always a solution to the scalar compatibility Eqs. (29) and (30). If we set $\rho_c = 6$, then the functions $\rho_1 : \alpha \mapsto \frac{\alpha}{c_1(\alpha)}$ and $\rho_2 : \alpha \mapsto \frac{\alpha}{c_2(\alpha)}$ both have a limit equal to ρ_c when $\alpha \to 0$,. Furthermore:*

- *There exists $\alpha^* > 0$ such that ρ_1 is decreasing on $(-\infty, \alpha^*]$ and increasing on $[\alpha^*, +\infty)$, converging to $+\infty$ at $\pm\infty$. We set $\rho^* = \rho_1(\alpha^*)$ (which is less than ρ_c). For all $\rho \geqslant \rho^*$, we define $\alpha_1^{\uparrow}(\rho)$ (resp. $\alpha_1^{\downarrow}(\rho)$) to be the unique value of $\alpha \geqslant \alpha^*$ (resp $\alpha \leqslant \alpha^*$) such that $\rho_1(\alpha) = \rho$. Finally, we define $\widetilde{c}_1^{\uparrow}(\rho) = c_1(\alpha_1^{\uparrow}(\rho))$ and $\widetilde{c}_1^{\downarrow}(\rho) = c_1(\alpha_1^{\downarrow}(\rho))$. Setting $c^* = c_1(\alpha^*)$, the function $\widetilde{c}_1^{\uparrow}$*

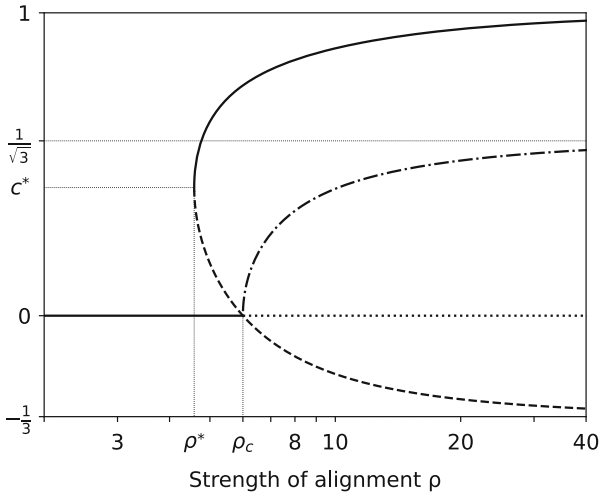

Fig. 3 Behaviors of the functions $\widetilde{c}_1^{\uparrow}$ (solid line), $\widetilde{c}_1^{\downarrow}$ (dashed line) and $\widetilde{c}_2$ (dashed-dot line)

(resp. $\widetilde{c}_1^{\downarrow}$) is increasing (resp. decreasing) on $[\rho^*, +\infty)$, with value c^* at ρ^*, and converging to 1 (resp. $-\frac{1}{3}$) at $+\infty$.

Numerically, we obtain $\alpha^* \approx 1.9395$, $\rho^* \approx 4.5832$, and $c^* \approx 0.4232$.

- *The function ρ_2 is (even and) increasing on $[0, +\infty)$, converging to $+\infty$ at $+\infty$. For all $\rho \geqslant \rho_c$, we define $\alpha_2(\rho)$ to be the unique value of $\alpha \geqslant \alpha^*$ such that $\rho_2(\alpha) = \rho$. Finally, we define $\widetilde{c}_2(\rho) = c_2(\alpha_2(\rho))$. The function $\widetilde{c}_2$ is increasing on $[\rho_c, +\infty)$, with value 0 at ρ_c and converging to $\frac{1}{\sqrt{3}}$ at $+\infty$.*

Figure 3 depicts a plot of these functions $\widetilde{c}_1^{\uparrow}$ (solid), $\widetilde{c}_1^{\downarrow}$ (dashed), and $\widetilde{c}_2$ (dashed-dot line), in log-scale for $\rho \in [2, 40]$. They represent the order parameters (up to sign) of the different families of steady-states. We also drew a solid line at level 0 for $\rho < \rho_c$ and a dotted line at level 0 for $\rho > \rho_c$, corresponding to the order parameter of the uniform steady-state (and illustrating the result of Proposition 6 regarding its stability).

We can therefore describe more precisely the long time behaviour of the solution to the Fokker–Planck equation according to the value of ρ, thanks to Proposition 5.

Theorem 1 *Let f_0 be a nonnegative measure with mass $\rho > 0$, and f the solution to the Fokker–Planck equation (19) with initial condition f_0. For the following statements, the notion of convergence is with respect to any H^s norm on $SO_3(\mathbb{R})$.*

- *If $\rho < \rho^*$, the only steady-state is the uniform distribution on $SO_3(\mathbb{R})$, and the solution $f(t, \cdot)$ converges to this steady state as $t \to +\infty$.*
- *If $\rho^* \leqslant \rho \leqslant \rho_c$, there are three families of steady-states (two of which are equal when $\rho = \rho^*$ or $\rho = \rho_c$), and $f(t, \cdot)$ converges to one of these families:*
  - *either there exists $A_0(t) \in SO_3(\mathbb{R})$ such that $f(t, \cdot) - \rho M_{\alpha_1^{\uparrow}(\rho) A_0(t)}$ converges to zero,*

  - *either $f(t,\cdot)$ converges to the uniform distribution on $SO_3(\mathbb{R})$,*
  - *or there exists $A_0(t) \in SO_3(\mathbb{R})$ such that $f(t,\cdot) - \rho M_{\alpha_1^\downarrow(\rho)A_0(t)}$ converges to zero, as $t \to +\infty$.*

- *If $\rho > \rho_c$, there is an additional family of steady-states, and $f(t,\cdot)$ converges to one of these four families:*
  - *either there exists $A_0(t) \in SO_3(\mathbb{R})$ such that $f(t,\cdot) - \rho M_{\alpha_1^\uparrow(\rho)A_0(t)}$ converges to zero,*
  - *either $f(t,\cdot)$ converges to the uniform distribution on $SO_3(\mathbb{R})$,*
  - *either there exists $A_0(t) \in SO_3(\mathbb{R})$ such that $f(t,\cdot) - \rho M_{\alpha_1^\downarrow(\rho)A_0(t)}$ converges to zero,*
  - *or there exist unit vectors $\mathbf{a}_0(t), \mathbf{b}_0(t)$, with $f(t,\cdot) - \rho M_{\alpha_2(\rho)\sqrt{3}\mathbf{a}_0(t)\otimes\mathbf{b}_0(t)}$ converging to zero, as $t \to +\infty$.*

Proof This result is a summary of the possible steady-states according to Proposition 3 and 9. The convergence of f to one of this families comes from Proposition 5 and from the fact that, even if the limit set $\mathcal{E}_\infty$ of equilibria may consist of several distinct such families, they would belong to different connected components of $\mathcal{E}_\infty$. □

Let us now try to understand the stability of each of these families of equilibria. Figure 4 is a zoom on the region $\rho \in [3, 8]$ of the plots of the functions $\widetilde{c}_1^\uparrow$, $|\widetilde{c}_1^\downarrow|$ and $\widetilde{c}_2$ (remember that these functions are the order parameters of the corresponding

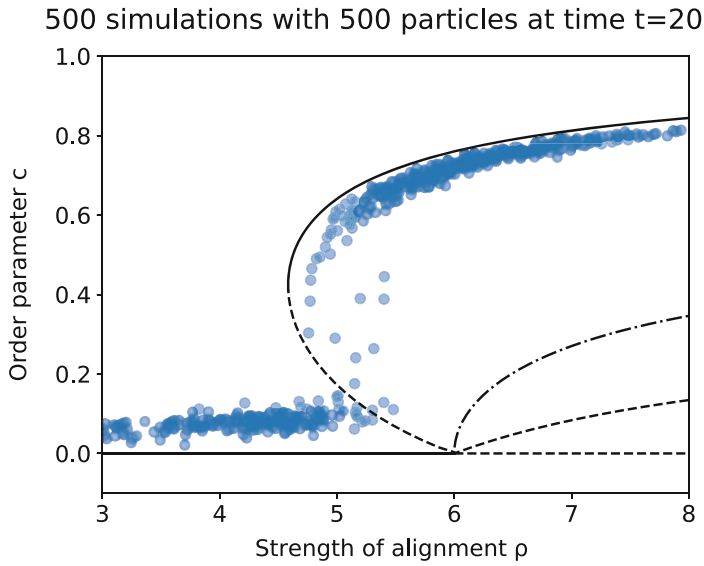

Fig. 4 Behaviors of the functions $\widetilde{c}_1^\uparrow$, $|\widetilde{c}_1^\downarrow|$ and $\widetilde{c}_2$ and final order parameters of the numerical simulations

steady-states), on top of the final values of the order parameters of the numerical simulations which were given in the right part of Fig. 2. It suggests the only stable equilibria, apart from the uniform one when $\rho < \rho_c$, are those corresponding to the curve $\widetilde{c}_1^{\uparrow}$. This is indeed what we will show in the next section.

5 Stability Results Thanks to a BGK Model

Instead of the Fokker–Planck equation (19), let us consider the following BGK equation:

$$\partial_t f = \rho M_{\mathcal{J}[f]} - f. \tag{31}$$

This is still an equation where the total mass is preserved and for which the steady states satisfy the same compatibility equation: if f is a steady-state, it has to be of the form ρM_J where $J = \mathcal{J}[f] = \rho \mathcal{J}[M_J]$. Therefore these two evolution equations share the same steady-states, which were determined in [7] and summarized in the previous section. Let us now give a summary of the results of stability of these equilibria which were obtained in [7]. It happens that these two evolution equations (BGK and Fokker–Planck) also share the same property of dissipation of the free energy $\mathcal{F}$: if f is a positive solution to (31), then by multiplying both sides by $\ln f(A) - A \cdot \mathcal{J}[f]$ and integrating on $SO_3(\mathbb{R})$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[f] + \widetilde{\mathcal{D}}[f] = 0,$$

where $\mathcal{F}[f]$ is given by (22) and

$$\widetilde{\mathcal{D}}[f] = \int_{SO_3(\mathbb{R})} (f - \rho M_{\mathcal{J}[f]})\big(\ln f - \ln(\rho M_{\mathcal{J}[f]})\big)\mathrm{d}A \geqslant 0.$$

Then, by writing $J(t) = \mathcal{J}[f(t,\cdot)]$ where f is a solution of the BGK equation (31), we obtain that J satisfies an ordinary differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}J = \rho \mathcal{J}[M_J] - J. \tag{32}$$

The long-time behaviour of the solution of the BGK equation is much simpler to study, since it can be reduced to the study of a finite dimensional ODE.

A further reduction can be done through the special singular value decomposition, for which we state a result which will be useful in the following.

Proposition 11 *If $J \in M_3(\mathbb{R})$, we call Special Singular Value Decomposition (SSVD) of J a decomposition of the form $J = PDQ$ where $P, Q \in SO_3(\mathbb{R})$ and $D = \mathrm{diag}(d_1, d_2, d_3)$ is a diagonal matrix satisfying $d_1 \geqslant d_2 \geqslant |d_3|$.*

Such a SSVD always exists, and the matrix D is unique (the rotations P and Q may not be unique). Furthermore, we have

$$\min_{A \in SO_3(\mathbb{R})} \|J - A\| = \|J - PQ\| = \|D - I_3\|. \tag{33}$$

Proof The existence and uniqueness can be obtained through the singular value decomposition, and modifying the orthogonal matrices if necessary to change the sign of the last entry of the diagonal part and get special orthogonal matrices, see [7]. We now compute

$$\|J - A\|^2 = \|D - P^\top A Q^\top\|^2 = \|D\|^2 - 2\, B \cdot D + \frac{3}{2},$$

where $B = P^\top A Q^\top$. Therefore minimizing $\|J - A\|$ for $A \in SO_3(\mathbb{R})$ amounts to maximizing $B \cdot D$, for $B \in SO_3(\mathbb{R})$. The set of diagonal parts of rotation matrices (seen as vectors of $\mathbb{R}^3$) is given by Horn's tetrahedron [19]: this is the convex hull $\mathcal{T}$ of the points $(\pm 1, \pm 1, \pm 1)$ with an even number of minus signs. Therefore we want to maximize $\mathbf{x} \cdot \mathbf{d}$ for $\mathbf{x} \in \mathcal{T}$ and $\mathbf{d} = (d_1, d_2, d_3)$. This convex function reaches it maximum on extremal points of $\mathcal{T}$, that is to say on one of the vertices of $\mathcal{T}$. Since we have

$$d_1 + d_2 + d_3 \geqslant d_1 - d_2 - d_3 \geqslant -d_1 + d_2 - d_3 \geqslant -d_1 - d_2 - d_3,$$

we see that the maximum is reached for $\mathbf{x} = (1, 1, 1)$. Therefore the maximum of $B \cdot D$ for $B \in SO_3(\mathbb{R})$ is reached for $B = I_3$, which ends the proof.[4] □

With this definition of the SSVD, the reduction that can be done is that the flow of the ODE (32) preserves the SSVD: if a SSVD of the initial condition is given by $J(0) = P D_0 Q$, then for all time t, we have the following SSVD: $J(t) = P D(t) Q$, with the same rotation matrices P and Q, and where $D(t) = (d_1(t), d_2(t), d_3(t))$ is a diagonal matrix satisfying the same ODE (32) as J, with initial condition D_0 (the fact that the matrix is diagonal and the inequalities $d_1(t) \geqslant d_2(t) \geqslant |d_3(t)|$ are preserved by the flow of this ODE). We therefore only have to study a three-dimensional ODE. Finally, the last observation

[4] Let us remark that if $d_2 > -d_3$, the maximum of $\mathbf{x} \cdot \mathbf{d}$ is unique on $\mathcal{T}$ and since the only rotation matrix for which the diagonal part is $(1, 1, 1)$ is the identity matrix I_3, we get that the minimizer PQ of (33) is unique. So even if P and Q may not be unique, in that case the matrix PQ is unique, and could be seen as a Special Polar Decomposition of J (with the analogy with the fact that if $\det J > 0$, then $J = PDQ$ is the singular value decomposition of J and PQ is the polar decomposition of J [10]).

we can do is that the flow of the ODE (32) is actually a gradient flow of a potential: if we write

$$V(J) = \frac{1}{2}\|J\|^2 - \rho \ln \mathcal{Z}(J), \text{ where } \mathcal{Z}(J) = \int_{SO_3(\mathbb{R})} \exp(J \cdot A)\mathrm{d}A, \tag{34}$$

as in the definition (6) of the generalized von Mises distribution, we obtain

$$\nabla V(J) = J - \rho \mathcal{J}[M_J], \tag{35}$$

where the gradient is taken with respect to the inner product of $M_3(\mathbb{R})$ given by (3).

Therefore the ODE (32) is simply $\frac{\mathrm{d}}{\mathrm{d}t} J = -\nabla V(J)$, and one can prove that any solution will converge to a critical point of V, which corresponds to a solution of the compatibility equation (20). We then obtain the same type of convergence as in Theorem 1. The main difference is that we have convergence to a unique steady-state (and not to a set of steady-states), that can be determined by knowing a special singular value decomposition of $\mathcal{J}[f_0]$. The other difference is that the convergence does not takes place in any Sobolev space H^s: the BGK equation is not regularizing in time. The following proposition is a summary of results in [7]:

Proposition 12 *Let f_0 be a finite nonnegative measure with mass $\rho > 0$, and f the solution to the BGK equation (31) with initial condition f_0. We write the decomposition $\mathcal{J}[f_0] = P_0 D_0 Q_0$, where $P_0, Q_0 \in SO_3(\mathbb{R})$ and $D_0 = \mathrm{diag}(d_{1,0}, d_{2,0}, d_{3,0})$, with $d_{1,0} \geqslant d_{2,0} \geqslant |d_{3,0}|$ (special singular value decomposition). Then for all $t \in \mathbb{R}$, we have $\mathcal{J}[f(t,\cdot)] = P_0 D(t) Q_0$, where $D(t) = \mathrm{diag}(d_1(t), d_2(t), d_3(t))$ is the solution to the ODE (32) with initial condition D_0, satisfying $d_1(t) \geqslant d_2(t) \geqslant |d_3(t)|$. In the following statements, the notion of convergence of $f(t,\cdot)$ is in the space of measures (or any normed space for which f_0 is an element and for which the map $f \mapsto \mathcal{J}[f]$ is continuous).*

- *If $\rho < \rho^*$, then $D(t) \to 0$ and $f(t,\cdot)$ converges to the uniform distribution as $t \to +\infty$.*
- *If $\rho^* \leqslant \rho \leqslant \rho_c$, there are three families of steady-states (two of which are equal when $\rho = \rho^*$ or $\rho = \rho_c$), and $f(t,\cdot)$ converges to one of these steady-states, as $t \to +\infty$:*
  - *either $D(t) \to 0$, and $f(t,\cdot)$ converges to the uniform distribution,*
  - *either $D(t) \to \alpha_1^{\uparrow}(\rho) I_3$, and $f(t,\cdot) \to \rho M_{\alpha_1^{\uparrow}(\rho) A_0}$ where $A_0 = P_0 Q_0$,*
  - *or $D(t) \to \alpha_1^{\downarrow}(\rho) I_3$, and $f(t,\cdot) \to \rho M_{\alpha_1^{\downarrow}(\rho) A_0}$ where $A_0 = P_0 Q_0$.*
- *If $\rho > \rho_c$, there is an additional family of steady-state, and $f(t,\cdot)$ converges to one of these steady-states, as $t \to +\infty$:*
  - *either $D(t) \to 0$, and $f(t,\cdot)$ converges to the uniform distribution,*
  - *either $D(t) \to \alpha_1^{\uparrow}(\rho) I_3$, and $f(t,\cdot) \to \rho M_{\alpha_1^{\uparrow}(\rho) A_0}$ where $A_0 = P_0 Q_0$,*

- *either* $D(t) \to \alpha_1^{\downarrow}(\rho)\,\mathrm{diag}(-1,-1,1)$, *and* $f(t,\cdot)$ *converges to* $\rho M_{\alpha_1^{\downarrow}(\rho)A_0}$, *where* $A_0 = P_0\,\mathrm{diag}(-1,-1,1)\,Q_0$
- *or* $D(t) \to \alpha_2(\rho)\mathrm{diag}(\sqrt{3},0,0)$, *and* $f(t,\cdot)$ *converges to* $\rho M_{\alpha_2(\rho)\sqrt{3}\mathbf{a}_0\otimes\mathbf{b}_0}$, *with* $\mathbf{a}_0 = P_0\mathbf{e}_1$ *and* $\mathbf{b}_0 = Q_0^{\top}\mathbf{e}_1$ *(where* $\mathbf{e}_1$ *is the first element of the canonical basis of* $\mathbb{R}^3$*).*

We now turn to stability results. For convenience, we will denote $\overline{V}$ the restriction of V to the space of diagonal matrices. Its Hessian $\mathrm{Hess}\,\overline{V}$ is then a symmetric bilinear form on a space of dimension 3. Thanks to the study of the signature of this Hessian, we obtained in [7] the characterization of the stability of all steady-states. The next proposition is a summary of these results (without details on the domains of convergence):

Proposition 13 *The uniform steady-state for the BGK equation* (31) *corresponds to the critical point* 0 *of the potential* $\overline{V}$ *(and* V*).*

- *If* $0 < \rho < \rho_c$, *the Hessian* $\mathrm{Hess}\,\overline{V}(0)$ *has signature* $(+++)$ *(and so* 0 *is a local minimizer of* V*). Therefore the uniform steady-state is locally asymptotically stable (with exponential rate of convergence).*
- *If* $\rho > \rho_c$, *the signature is* $(---)$ *(therefore* 0 *is not a local minimizer of* V*), and the uniform steady-state is unstable.*

When $\rho \geqslant \rho^*$, *the steady-states of the form* $\rho M_{\alpha_1^{\uparrow}(\rho)A_0}$ *(resp.* $\rho M_{\alpha_1^{\downarrow}(\rho)A_0}$*) with* A_0 *in* $SO_3(\mathbb{R})$ *(see Theorem 1) correspond to the critical points of the form* $\alpha_1^{\uparrow}(\rho)A_0$ *(resp.* $\alpha_1^{\downarrow}(\rho)A_0$*) of* V. *Their nature can be reduced to the study of the critical point* $D_\infty^{\uparrow} = \alpha_1^{\uparrow}(\rho)I_3$ *(resp.* $D_\infty^{\downarrow} = \alpha_1^{\downarrow}(\rho)I_3$*) of* $\overline{V}$.

- *If* $\rho > \rho^*$, *the Hessian* $\mathrm{Hess}\,\overline{V}(D_\infty^{\uparrow})$ *has signature* $(+++)$ *(and so* $\alpha_1^{\uparrow}(\rho)A_0$ *is a local minimizer of* V*). Thus the steady-states of the form* $\rho M_{\alpha_1^{\uparrow}(\rho)A_0}$ *are locally asymptotically stable (with exponential rate of convergence).*
- *If* $\rho^* < \rho < \rho_c$ *(resp.* $\rho > \rho_c$*), then* $\mathrm{Hess}\,\overline{V}(D_\infty^{\downarrow})$ *has signature* $(-++)$ *(resp.* $(+--)$*) (therefore* $\alpha_1^{\downarrow}(\rho)A_0$ *is not a local minimizer of* V*), and the steady-states of the form* $\rho M_{\alpha_1^{\uparrow}(\rho)A_0}$ *are unstable.*

When $\rho > \rho_c$, *the steady-states of the form* $\rho M_{\alpha_2(\rho)\sqrt{3}\mathbf{a}_0\otimes\mathbf{b}_0}$ *with* $\mathbf{a}_0, \mathbf{b}_0 \in \mathbb{S}^2$ *(see Theorem 1) correspond to the critical points of the form* $\alpha_2(\rho)\sqrt{3}\mathbf{a}_0 \otimes \mathbf{b}_0$ *of* V, *which reduces to the study of the critical point* $D_\infty = \alpha_2(\rho)\mathrm{diag}(1,0,0)$ *of* $\overline{V}$.

- *The Hessian* $\mathrm{Hess}\,\overline{V}(D_\infty)$ *has signature* $(++-)$ *(so* $\alpha_2(\rho)\sqrt{3}\mathbf{a}_0 \otimes \mathbf{b}_0$ *is not a local minimizer of* V*), and the steady-states of the form* $\rho M_{\alpha_2(\rho)\sqrt{3}\mathbf{a}_0\otimes\mathbf{b}_0}$ *are unstable.*

Furthermore, the critical cases are unstable: the uniform steady-state is unstable for $\rho = \rho_c$, *and the steady-states of the form* $\rho M_{\alpha^* A_0}$ *are unstable when* $\rho = \rho^*$ *(the corresponding matrices* $J = 0$ *or* $J = \alpha^* A_0$ *are not local minimizers of* V*).*

The main object of this section is to show, as it was claimed in Remark 5.5 of [7], that we can directly use these results of (in)stability for the BGK equation (and more precisely for the potential V) to obtain (in)stability results for the Fokker–Planck equation, in order to complete the results around the uniform distribution given by Proposition 6. We provide a proposition and a theorem which give details on this statement. The first proposition allows to compare the behaviours of V and of $J \mapsto \mathcal{F}[\rho M_J]$.

Proposition 14 *Let us define for* $J \in M_3(\mathbb{R})$

$$W(J) = \mathcal{F}[\rho M_J],$$

Then, we have that $\nabla W(J) = 0$ *if and only if* $\nabla V(J) = 0$, *that is to say* J *is a solution to the compatibility equation* (20). *Furthermore, if* J *is such a critical point, the Hessian* Hess W *has the same signature as* Hess V *(and more precisely, if* $\overline{W}$ *is the restriction of* W *to the diagonal matrices, then* Hess $\overline{W}$ *and* Hess $\overline{V}$ *have the same signature).*

Proof We first compute

$$\begin{aligned} W(J) &= \int \rho(\ln\rho + A\cdot J - \ln \mathcal{Z}(J))M_J(A)\mathrm{d}A - \frac{\rho^2}{2}\|\mathcal{J}[M_J]\|^2 \\ &= \rho\ln\rho - \ln\mathcal{Z}(J) + \frac{1}{2}\|J\|^2 - \frac{1}{2}\|J - \rho\mathcal{J}[M_J]\|^2 \\ &= V(J) - \frac{1}{2}\|\nabla V(J)\|^2 + \rho\ln\rho, \end{aligned}$$

thanks to (35). Therefore we obtain

$$\nabla W(J) = \nabla V(J) - \mathrm{Hess}\, V(J)(\nabla V(J)). \tag{36}$$

We want to compute the Hessian of V, seen as a linear mapping from $M_3(\mathbb{R})$ to $M_3(\mathbb{R})$, symmetric with respect to the inner product of $M_3(\mathbb{R})$. For a small $H \in M_3(\mathbb{R})$, we first have $\mathcal{Z}(J+H) = (1 + \mathcal{J}[M_J]\cdot H)\mathcal{Z}(J) + O(\|H\|^2)$. Thus we get $M_{J+H}(A) = (1 + A\cdot H - \mathcal{J}[M_J]\cdot H)M_J(A) + O(\|H\|^2)$. Finally we obtain

$$\begin{aligned} \mathcal{J}[M_{J+H}] = \mathcal{J}[M_J] &- (\mathcal{J}[M_J]\cdot H)\mathcal{J}[M_J] \\ &+ \int_{SO_3(\mathbb{R})} A(A\cdot H)M_J(A)\mathrm{d}A + O(\|H\|^2). \end{aligned}$$

Now, using the expression (35) of ∇V, we get

$$\mathrm{Hess}\, V(J)(H) = H - \rho\Big[(\mathcal{J}[M_J]\cdot H)\mathcal{J}[M_J] - \int_{SO_3(\mathbb{R})} A(A\cdot H)M_J(A)\mathrm{d}A\Big].$$

Said differently, seeing now Hess V as a symmetric bilinear form on $M_3(\mathbb{R})$:

$$\begin{aligned}\text{Hess } V(J)(H,H) &= \|H\|^2 - \rho\Big[(\mathcal{J}[M_J]\cdot H)^2 - \int_{SO_3(\mathbb{R})}(A\cdot H)^2 M_J(A)\mathrm{d}A\Big]\\ &= \|H\|^2 - \rho\int[(A-\mathcal{J}[M_J])\cdot H]^2 M_J(A)\mathrm{d}A, \end{aligned} \tag{37}$$

and we see that all the eigenvalues of Hess V are strictly less than 1. Therefore the (symmetric) linear mapping $\mathrm{Id} - \text{Hess } V$ from $M_3(\mathbb{R})$ to $M_3(\mathbb{R})$ has only strictly positive eigenvalues, and is therefore an isomorphism. The expression (36) of ∇W then provides the equivalence between critical points for V and for W.

Finally, at a point J for which $\nabla V(J) = 0$, we obtain

$$\text{Hess } W(J) = \text{Hess } V(J) - [\text{Hess } V(J)]^2.$$

Therefore, the eigenvalues of Hess $W(J)$ are given by $\lambda(1-\lambda)$, where λ are the eigenvalues of Hess $W(J)$, which all satisfy $\lambda < 1$. Therefore their signs are the same. And this is also true when restricted to the space of diagonal matrices. □

We can now state the final theorem of this section.

Theorem 2 *The nature of all the critical points of the free energy $\mathcal{F}$ is given by the following statements.*

- *For $\rho < \rho_c$, the uniform equilibrium of mass ρ is a local strict minimizer of the free energy $\mathcal{F}$.*
- *For $\rho > \rho^*$, the set $\mathcal{E} = \{\rho M_{\alpha_1^\uparrow(\rho)A_0},\ A_0 \in SO_3(\mathbb{R})\}$ is a local strict minimizer of the free energy $\mathcal{F}$, in the sense that there exists a neighborhood $\mathcal{V}$ of $\mathcal{E}$ (in the space of nonnegative measures of mass ρ) such that if $f \in \mathcal{V}\setminus\mathcal{E}$, then $\mathcal{F}[f] > \mathcal{F}_\infty$, where $\mathcal{F}_\infty$ is the common value of $\mathcal{F}$ on $\mathcal{E}$.*
- *For $\rho \geqslant \rho_c$, the uniform equilibrium of mass ρ is not a local minimizer of the free energy $\mathcal{F}$.*
- *For $\rho \geqslant \rho^*$ (and $\rho \neq \rho_c$), any steady-state of the form $\rho M_{\alpha_1^\downarrow(\rho)A_0}$ for A_0 in $SO_3(\mathbb{R})$ is not a local minimizer of the free energy $\mathcal{F}$.*
- *For $\rho > \rho_c$, any steady-state of the form $\rho M_{\alpha_2(\rho)\sqrt{3}\mathbf{a}_0\otimes\mathbf{b}_0}$ for A_0 in $SO_3(\mathbb{R})$ is not a local minimizer of the free energy $\mathcal{F}$.*

Therefore, the last three families of steady-states are unstable for the Fokker–Planck equation (19): there exist initial conditions arbitrarily close to these families (in any H^s norm), such that the solution to the Fokker–Planck equation converges in long time towards another family of equilibria (see Theorem 1).

Proof The first point has been proven in Proposition 6. For the second one, if it was not true, there would exist f_0 as close as we want from $\mathcal{E}$ such that $\mathcal{F}(f_0) \leqslant \mathcal{F}_\infty$, and $f_0 \notin \mathcal{E}$. Since the different families of steady-states are isolated, f_0 cannot be a steady-state. By letting f be the solution of the BGK equation with initial

condition f_0, we would have $\widetilde{Q}[f_0] > 0$ and therefore $\mathcal{F}[f(t,\cdot)] < \mathcal{F}_\infty$ for all $t > 0$. Combined with the fact that $\mathcal{F}[f(t,\cdot)]$ is nonincreasing in time, this would be in contradiction with the fact that $f(t,\cdot)$ converges towards the set $\mathcal{E}$, thanks to the asymptotic stability of those steady-states for the BGK equation given by Proposition 13. Let us remark that the first point of the theorem could be proven in the same way, without having to expand the free energy, but only using the known results for the BGK equation and the fact that $\mathcal{F}$ is nonincreasing.

To prove the last three points, let us take such a steady state, of the form ρM_{J_0}. We want to prove that J_0 is not a local minimizer of W, therefore ρM_{J_0} is not a local minimizer of $\mathcal{F}$. We write a SSVD of the form $J_0 = PD_0Q$ where D_0 is a diagonal matrix and $P, Q \in SO_3(\mathbb{R})$. If $J = PDQ$ where D is a diagonal matrix close to D_0, then $W(J) = \overline{W}(D)$. Therefore we only need to prove that D_0 is not a local minimizer of $\overline{W}$. In the case where $\rho \neq \rho_c$ and $\rho \neq \rho^*$, since the signature of Hess $\overline{W}(D_0)$ has negative components (thanks to Propositions 13 and 14), we directly get the results. In the critical cases we will use a mountain-pass lemma argument. In the case where $\rho = \rho_c$, suppose that 0 is a local minimizer of $\overline{W}$. Then it is a local strict minimizer, since this critical point is isolated. Therefore by looking at the other local strict minimizer $\alpha_1^{\uparrow}(\rho_c)I_3$ of $\overline{W}$ (for which the signature of the Hessian is $(+\,+\,+)$, thanks again to Propositions 13 and 14), we would obtain, by the mountain-pass lemma, a third critical point D of $\overline{W}$, which would satisfy $\overline{W}(D) > \max(\overline{W}(0), \overline{W}(\alpha_1^{\uparrow}(\rho_c)I_3))$. This is in contradiction with the fact that we only have two families of equilibria for this value of ρ. The same argument can be used to show that when $\rho = \rho^*$, the point $\alpha^* I_3$ is not a local minimizer of $\overline{W}$, using as other local strict minimizer the point 0.

The conclusion of the statement of the theorem comes from the fact that we actually proved that the critical points were not local minimizers of W, which is the evaluation of $\mathcal{F}$ on smooth functions of the form ρM_J, so the H^s norm of $\rho M_J - \rho M_{J_0}$ is small when J is close to J_0. □

For the first two points of Theorem 2, we did not provide the corresponding stability results. Indeed, in the next section, a more detailed study will show that they are exponentially stable.

6 Exponential Convergence for the Stable Steady-States

We will now show that the two families of steady-states that correspond to what we observe in the numerical simulations are locally exponentially attracting. In particular, when f is a solution to the Fokker–Planck equation in the neighborhood of those steady-states, we will show that $\mathcal{J}[f(t,\cdot)]$ will converge to a solution J_∞ of the compatibility equation (20). However, since this J_∞ (if it is non-zero) is not known from the initial condition (contrary to the case of the BGK equation), it is not easy to control directly the distance between f and ρM_{J_∞}, but we will see that controlling the distance from f and $\rho M_{\mathcal{J}[f]}$, even if this last one is not a steady-

state, will be the key to our analysis. A convenient framework is to use the relative entropy, for which we will need the following results.

Proposition 15 *Let $\rho > 0$. If f, g are two measurable nonnegative functions on $SO_3(\mathbb{R})$ with total mass ρ and with $g > 0$, we define the relative entropy $\mathcal{H}(f|g)$ and Fisher information $\mathcal{I}(f|g)$:*

$$\mathcal{H}(f|g) = \int_{SO_3(\mathbb{R})} f(A) \ln\left(\frac{f(A)}{g(A)}\right) dA,$$
$$\mathcal{I}(f|g) = \int_{SO_3(\mathbb{R})} f(A) \left\| \nabla \ln\left(\frac{f(A)}{g(A)}\right)\right\|^2 dA.$$

Then, for two such functions, we have the Csiszár–Kullback–Pinsker inequality:

$$\int_{SO_3(\mathbb{R})} |f(A) - g(A)| \, \mathrm{d}A \leqslant \sqrt{2\rho \, \mathcal{H}(f|g)}. \tag{38}$$

Finally, we have the following families of (weighted) logarithmic Sobolev inequalities: there exists a constant $\lambda > 0$ such that for all $J \in M_3(\mathbb{R})$ with $\|J\| \leqslant \frac{\sqrt{3}}{\sqrt{2}}\rho$, and all measurable nonnegative function f with total mass ρ, we have

$$\mathcal{H}(f|\rho M_J) \leqslant \frac{1}{2\lambda} \mathcal{I}(f|\rho M_J). \tag{39}$$

Proof The Csiszár–Kullback–Pinsker inequality is well-known [6, 23], we just notice the factor ρ since we do not work with probability measures here. The logarithmic Sobolev inequality (39) in the case $J = 0$ (uniform measure on $SO_3(\mathbb{R})$) comes for instance from the Bakry–Émery criterion [1] since $SO_3(\mathbb{R})$ has positive Ricci curvature (this is the same as the curvature of $\mathbb{S}^3$, thanks to the local isometry Φ given in Proposition 7).[5] Then, we use the fact that the logarithmic Sobolev inequality is stable by bounded perturbation [18, 26]. Since $\|J\| \leqslant \frac{\sqrt{3}}{\sqrt{2}}\rho$, then M_J is bounded above and below, uniformly in J, which ends the proof. □

Let us now compute the relative entropy of f with respect to ρM_J for J in $M_3(\mathbb{R})$. Using the definition (6), we obtain

$$\begin{aligned}\mathcal{H}(f|\rho M_J) &= \int_{SO_3(\mathbb{R})} \big(f(A) \ln f(A) - f(A)\, A \cdot J\big) \mathrm{d}A + \rho \ln \mathcal{Z}(J) - \rho \ln \rho \\ &= \mathcal{F}[f] + \frac{1}{2}\|J - \mathcal{J}[f]\|^2 - V(J) - \rho \ln \rho, \end{aligned} \tag{40}$$

[5] Actually, as already stated by Bakry and Émery [1], this criterion does not give the optimal constant in $\mathbb{S}^3$, which was given by Mueller and Weissler in [22], but here even the optimal constant in $\mathbb{S}^3$ would not be necessarily optimal in $SO_3(\mathbb{R})$, since we only want the logarithmic Sobolev inequality for even functions on $\mathbb{S}^3$.

thanks to the definitions (22) and (34) of $\mathcal{F}$ and V. Therefore, if $J_{\rm eq}$ is a solution to the compatibility equation and $f_{\rm eq} = \rho M_{J_{\rm eq}}$, we apply (40) with $f = f_{\rm eq}$ and $J = J_{\rm eq}$ to obtain $\rho \ln \rho = \mathcal{F}[f_{\rm eq}] - V(J_{\rm eq})$. Now applying (40) with $J = J_{\rm eq}$ or with $J = \mathcal{J}[f]$, we obtain

$$\mathcal{F}[f] - \mathcal{F}[f_{\rm eq}] = \mathcal{H}(f|\rho M_{\mathcal{J}[f]}) + V(\mathcal{J}[f]) - V(J_{\rm eq}), \tag{41}$$

$$\mathcal{H}(f|f_{\rm eq}) = \mathcal{F}[f] - \mathcal{F}[f_{\rm eq}] + \frac{1}{2}\|J_{\rm eq} - \mathcal{J}[f]\|^2. \tag{42}$$

Furthermore, it is straightforward to see, thanks to the definition (23) of $\mathcal{D}[f]$, that

$$\mathcal{D}[f] = \mathcal{I}(f|\rho M_{\mathcal{J}[f]}). \tag{43}$$

These links between the free energy, its dissipation, the relative entropy, the Fisher information, and the potential V associated to the BGK equation are the key points to prove the stability of the steady-states associated to solutions of the compatibility equation corresponding to local minimizers of V.

Theorem 3 *Let $\rho > \rho^*$ (resp. $\rho < \rho_c$).*

We define the set of equilibria $\mathcal{E}_\infty = \{\rho M_{\alpha_1^\uparrow(\rho)A_0},\, A_0 \in SO_3(\mathbb{R})\}$ (resp. $\mathcal{E}_\infty$ reduced to the uniform distribution on $SO_3(\mathbb{R})$ of mass ρ).

Then there exists $\delta > 0$, $\widetilde{\lambda} > 0$ and $C > 0$ such that for all nonnegative measurable function f_0 with mass ρ, if there exists $f_{{\rm eq},0} \in \mathcal{E}_\infty$ such that $\mathcal{H}(f_0|f_{{\rm eq},0}) < \delta$, then there exists $f_\infty \in \mathcal{E}_\infty$ such that for all time $t \geqslant 0$, we have

$$\mathcal{H}(f(t,\cdot)|f_\infty) \leqslant C\, e^{-2\widetilde{\lambda}t}\, \mathcal{H}(f_0|f_{{\rm eq},0}).$$

Proof For convenience, we write $\alpha = \alpha_1^\uparrow(\rho)$ (resp. $\alpha = 0$ for the study of stability of the uniform equilibrium) and $V_\infty = V(\alpha I_3)$. We also denote by E_∞ the set of matrices $J_{\rm eq}$ solutions to the compatibility equation (20) corresponding to the family of equilibria we are interested in, that is to say $E_\infty = \{\alpha A_0,\, A_0 \in SO_3(\mathbb{R})\}$.

Since the signature of Hess$\overline{V}(\alpha I_3)$ is $(+\,+\,+)$ (thanks to Proposition 13), by continuity of Hess $\overline{V}$ (and of its smallest eigenvalue), there exists $\delta_0 > 0$ and $\eta > 0$ such that for all diagonal matrix D with $\|D - \alpha I_3\| < \delta_0$, Hess $\overline{V}(D)$ is positive definite with lowest eigenvalue being greater than or equal to η (we recall that thanks to (37), its highest eigenvalue is always less than 1). By the following Taylor formulas, for all such D, we have

$$\|\nabla\overline{V}(D)\|^2 = (D - \alpha I_3) \cdot \Big(\int_0^1 \mathrm{Hess}\,\overline{V}(\alpha I_3 + t(D - \alpha I_3))\mathrm{d}t \Big)^2 (D - \alpha I_3),$$

$$V(D) - V_\infty = \int_0^1 (1-t)(D - \alpha I_3) \cdot \mathrm{Hess}\,\overline{V}(\alpha I_3 + t(D - \alpha I_3))(D - \alpha I_3)\mathrm{d}t$$

and therefore

$$\|\nabla \overline{V}(D)\|^2 \geqslant \eta \, \|D - \alpha I_3\|^2,$$

$$\frac{\eta}{2} \|D - \alpha I_3\|^2 \leqslant V(D) - V_\infty \leqslant \frac{1}{2} \|D - \alpha I_3\|^2 \leqslant \frac{1}{2\eta} \|\nabla \overline{V}(D)\|^2.$$

Therefore, we write $U = \{J \in M_3(\mathbb{R}), \min_{J_{\text{eq}} \in E_\infty} \|J - J_{\text{eq}}\| < \delta_0\}$, which is a neighborhood of E_∞. If $J \in U$ and we write the SSVD $J = PDQ$, we obtain by Proposition 11 that $\min_{J_{\text{eq}} \in E_\infty} \|J - J_{\text{eq}}\| = \|D - \alpha I_3\| \leqslant \delta_0$ (when $\alpha > 0$, and the result is still true if $\alpha = 0$ since $E_\infty = \{0\}$ in that case). Therefore, since $V(J) = V(D)$ we obtain that there exists $J_{\text{eq}} \in E_\infty$ (which is equal to αPQ) such that

$$\frac{\eta}{2} \|J - J_{\text{eq}}\|^2 \leqslant V(J) - V_\infty \leqslant \frac{1}{2\eta} \|\nabla V(J)\|^2 = \frac{1}{2\eta} \|J - \rho \mathcal{J}[M_J]\|^2. \tag{44}$$

By the Csiszár–Kullback–Pinsker inequality (38), we have that if g is a nonnegative measure with mass ρ:

$$\|\mathcal{J}[f] - \mathcal{J}[g]\| \leqslant \int_{SO_3(\mathbb{R})} \|A\| |f(A) - g(A)| \mathrm{d}A \leqslant \frac{\sqrt{3}}{\sqrt{2}} \sqrt{2\rho \mathcal{H}(f|g)}, \tag{45}$$

and therefore for $g = \rho M_{\mathcal{J}[f]}$, we obtain

$$\|\mathcal{J}[f] - \mathcal{J}[\rho M_{\mathcal{J}[f]}]\| \leqslant \sqrt{3\rho \mathcal{H}(f|\rho M_{\mathcal{J}[f]})}.$$

Combining this with (41) and (44) with $J = \mathcal{J}[f]$, we get that if $\mathcal{J}[f] \in U$, then

$$\mathcal{F}[f] - \mathcal{F}_\infty \leqslant (1 + \tfrac{3\rho}{2\eta}) \, \mathcal{H}(f|\rho M_{\mathcal{J}[f]}).$$

Therefore, as soon as $\mathcal{J}[f] \in U$, we have by (43) and the logarithmic Sobolev inequality (39) (we recall that $\|\mathcal{J}[f]\| \leqslant \frac{\sqrt{3}}{\sqrt{2}} \rho$ if the total mass of f is ρ):

$$\mathcal{D}[f] \geqslant \frac{2\lambda}{1 + \frac{3\rho}{2\eta}} (\mathcal{F}[f] - \mathcal{F}_\infty).$$

By the dissipation of the free energy (21), writing $\widetilde{\lambda} = \frac{\lambda}{1 + \frac{3\rho}{2\eta}}$ we obtain that as long as $\mathcal{J}[f] \in U$,

$$0 \leqslant \mathcal{F}[f] - \mathcal{F}_\infty \leqslant e^{-2\widetilde{\lambda} t} (\mathcal{F}[f_0] - \mathcal{F}_\infty) \leqslant e^{-2\widetilde{\lambda} t} \mathcal{H}(f_0|f_{\text{eq},0}), \tag{46}$$

the first inequality coming from (41) and the fact that $V(\mathcal{J}[f]) - V_\infty \geqslant 0$ thanks to (44), and the last inequality coming from (42). Finally, thanks to (44), (41)

and (46), we obtain that still as long as $\mathcal{J}[f(t,\cdot)] \in U$, there exists $J_{\mathrm{eq}}(t)$ such that

$$\|\mathcal{J}[f(t,\cdot)] - J_{\mathrm{eq}}(t)\| \leqslant \sqrt{\frac{2}{\eta}(V(\mathcal{J}[f(t)]) - V_\infty)} \leqslant \frac{\sqrt{2}}{\sqrt{\eta}}\, e^{-\widetilde{\lambda}t} \sqrt{\mathcal{H}(f_0|f_{\mathrm{eq},0})}. \tag{47}$$

Therefore, by taking $\delta = \min(\frac{\eta}{2}\delta_0^2, \frac{1}{3\rho}\delta_0^2)$, and using (45) with $g = f_{\mathrm{eq},0}$, we obtain that if $\mathcal{H}(f_0|f_{\mathrm{eq},0}) < \delta$, then $\|\mathcal{J}[f_0] - \mathcal{J}[f_{\mathrm{eq},0}]\| < \delta_0$, so $\mathcal{J}[f_0] \in U$, and for all positive time $\|\mathcal{J}[f(t,\cdot)] - J_{\mathrm{eq}}(t)\| < \delta_0$ (and therefore $\mathcal{J}[f(t,\cdot)]$ stays in U) thanks to (47). Indeed, if it was not the case, for the first exit time $t_0 > 0$ of U, we would have $\|\mathcal{J}[f(t_0,\cdot)] - J_{\mathrm{eq}}(t_0)\| \leqslant \frac{\sqrt{2}}{\sqrt{\eta}}\sqrt{\mathcal{H}(f_0|f_{\mathrm{eq},0})} < \delta_0$ which is a contradiction. From now on we suppose that $\mathcal{H}(f_0|f_{\mathrm{eq},0}) < \delta$, so that (47) and (46) are valid for all time $t \geqslant 0$.

Let us now find a way to control the displacement of $\mathcal{J}[f]$. For $J \in M_3(\mathbb{R})$, using the Fokker–Planck equation (19) and integrating by parts, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} J \cdot \mathcal{J}[f] = \int_{SO_3(\mathbb{R})} [\nabla_A(A\cdot J)\cdot\nabla_A(A\cdot\mathcal{J}[f]) - \Delta_A(A\cdot J)] f(A)\,\mathrm{d}A,$$

which can be written

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{J}[f] = \mathcal{M}[f](\mathcal{J}[f]) - \mathcal{L}[f], \tag{48}$$

where, when g is an integrable function on $SO_3(\mathbb{R})$, we define $\mathcal{M}[g]$ as the linear operator from $M_3(\mathbb{R})$ to $M_3(\mathbb{R})$ given by the fact that for any J, J' in $M_3(\mathbb{R})$,

$$J\cdot\mathcal{M}[g](J') = \int_{SO_3(\mathbb{R})} \nabla_A(A\cdot J)\cdot\nabla_A(A\cdot J')g(A)\,\mathrm{d}A,$$

and $\mathcal{L}[g]$ as the matrix[6] such that for all $J \in M_3(\mathbb{R})$,

$$J\cdot\mathcal{L}[g] = \int_{SO_3(\mathbb{R})} \Delta_A(A\cdot J)g(A)\,\mathrm{d}A.$$

[6] We can actually show (but we do not need it here) that $\mathcal{L}[f]$ is proportional to $\mathcal{J}[f]$. Indeed, since $q \mapsto q\cdot Qq$ is an eigenfunction of the Laplacian on the unit sphere of $\mathbb{R}^4$ (more precisely a spherical harmonic of degree 2) when Q is a symmetric trace-free matrix, we get, thanks to the local isometry Φ and Proposition (8), that $A \mapsto A\cdot J$ is also an eigenfunction of the Laplacian on $SO_3(\mathbb{R})$.

We therefore see that since the functions under the integral are smooth and bounded, there exists $C_0 > 0$ such that for all $J \in M_3(\mathbb{R})$ and for any integrable function g on $SO_3(\mathbb{R})$,

$$\|\mathcal{L}[g]\| \leqslant C_0 \int_{SO_3(\mathbb{R})} |g(A)| \, \mathrm{d}A, \tag{49}$$

and

$$\|\mathcal{M}[g](J)\| \leqslant C_0 \|J\| \int_{SO_3(\mathbb{R})} |g(A)| \, \mathrm{d}A. \tag{50}$$

Therefore, defining $f_{\mathrm{eq}}(t,\cdot) = \rho M_{J_{\mathrm{eq}}(t)}$, and using the fact that it is a stationary solution, thus giving by (48) that $\mathcal{M}[f_{\mathrm{eq}}](\mathcal{J}[f_{\mathrm{eq}}]) - \mathcal{L}[f_{\mathrm{eq}}] = 0$, we obtain

$$\begin{aligned}\left\| \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{J}[f] \right\| &= \|\mathcal{M}[f](\mathcal{J}[f]) - \mathcal{M}[f_{\mathrm{eq}}](\mathcal{J}[f_{\mathrm{eq}}]) - \mathcal{L}[f] + \mathcal{L}[f_{\mathrm{eq}}]\| \\ &\leqslant \|\mathcal{M}[f](\mathcal{J}[f - f_{\mathrm{eq}}])\| + \|\mathcal{M}[f - f_{\mathrm{eq}}](\mathcal{J}[f_{\mathrm{eq}}])\| + \|\mathcal{L}[f - f_{\mathrm{eq}}]\|.\end{aligned}$$

Therefore, by using (49)–(50) and the Csiszár–Kullback–Pinsker inequalities (38) and (45), we get that there exists a constant $C_1 > 0$ (only depending on ρ) such that

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{J}[f] \right\| \leqslant \sqrt{C_1 \mathcal{H}(f|f_{\mathrm{eq}})}.$$

Combining this with (42), (47), and (46), we then get that there exists a constant C_2 (not depending on f_0) such that for all $t \geqslant 0$

$$\left\| \frac{\mathrm{d}}{\mathrm{d}\,t} \mathcal{J}[f] \right\| \leqslant e^{-\widetilde{\lambda} t} \sqrt{C_2 \mathcal{H}(f_0|f_{\mathrm{eq},0})}.$$

Finally, this gives that $\mathcal{J}[f]$ converges exponentially fast with rate $\widetilde{\lambda}$ towards a given matrix $J_\infty \in M_3(\mathbb{R})$ and since the distance between $\mathcal{J}[f]$ and E_∞ converges to 0 thanks to (47), we obtain that $J_\infty \in E_\infty$. More precisely, we have

$$\|\mathcal{J}[f(t,\cdot)] - J_\infty\| \leqslant \int_t^{+\infty} \left\| \frac{\mathrm{d}}{\mathrm{d}s} \mathcal{J}[f(s,\cdot)] \right\| \mathrm{d}t \leqslant \frac{e^{-\widetilde{\lambda} t}}{\widetilde{\lambda}} \sqrt{C_2 \mathcal{H}(f_0|f_{\mathrm{eq},0})}. \tag{51}$$

Defining $f_\infty = \rho M_{J_\infty}$ and using (42) with $f_{\mathrm{eq}} = f_\infty$, (46), and (51), we then get that there exists a constant $C_3 > 0$ (not depending on f_0) such that

$$\mathcal{H}(f(t,\cdot)|f_\infty) \leqslant C_3 e^{-2\widetilde{\lambda} t} \, \mathcal{H}(f_0|f_{\mathrm{eq},0}),$$

which ends the proof. Let us remark that this proof covers the case $\alpha = 0$, but if we only want to do this case, it can be simplified a lot since $E_\infty = \{0\}$. □

Let us finish this section by some comments. The proof of Theorem 3 has been done here in relative entropy. It may look similar in some points to [16], but the main idea is above all based on the fact that we measure the relative entropy with respect to a target measure $\rho M_{\mathcal{J}[f]}$ which is not itself a steady-state. The fine control of the potential V around the solutions of the compatibility equation is the key to link all these different quantities. The proof would have worked the same in L^2, by using the regularizing effect of the equation (and L^∞ bounds), as was done in [17] for the Vicsek model, but the main difference is again that we would compare $\mathcal{D}[f]$ and $\mathcal{F}[f] - \mathcal{F}_\infty$ with $\|f - \rho M_{\mathcal{J}[f]}\|_2^2$. This proof seems to be adaptable to a lot of different models of Fokker–Planck type, such as the Doi–Onsager theory for suspensions of rodlike polymers, for which, as far as we know, no proof of exponential convergence is available (but the analog to the potential V has been studied, therefore the nature of the critical points is well-known). This is left for future work.

Finally, now that we have a good understanding of the long time behaviour of the Fokker–Planck equation (19), we could try to further understand the limit of the particle system as $N \to \infty$. Since the mean-field limit is essentially a law of large numbers, we expect fluctuations of order $\frac{1}{\sqrt{N}}$, which explains why the order parameters of the numerical simulations in Fig. 4 are not so close to 0 for what is expected to be the uniform distribution. More precisely, as indicated by the estimate (18), the distance between the empirical measure and the solution to the Fokker–Planck equation can be bounded by $\frac{e^{\tilde{C}T}}{\sqrt{N}}$, for all t in $[0, T]$. Therefore if we want such an estimate for a large time T, we cannot do better than T of order $\ln N$. However, since the equilibria are exponentially stable, the fluctuations that would push the empirical distribution away from the family of stable equilibria, are compensated by the deterministic dynamics of the Fokker–Planck equation. Therefore the only remaining fluctuations would cause the solution to fluctuate mainly in the tangential component of the family of equilibria. This approach has been made rigorous in the case of identical Kuramoto oscillators in [2] (which corresponds to the Vicsek model studied in [17] in dimension two), where it is proved that the solution stays close to the set of equilibria up to times of order N, but with the center of synchronization of the distribution performing a Brownian motion on the circle at these time scales. In analogy with this result, we could expect in our case that, close to the family of von Mises distributions $\rho M_{\alpha A}$ with $\alpha > 0$ and $A \in SO_3(\mathbb{R})$, the long time behaviour at time $t = sN$ of the empirical measure of particle system would be close to $\rho M_{\alpha A(s)}$, where $A(s)$ performs a Brownian motion on $SO_3(\mathbb{R})$. This is also left for future work.

Acknowledgments The author wants to thank his collaborators Pierre Degond, Sara Merino-Aceituno, Ariane Trescases and Antoine Diez for all the work done together on body-attitude coordination models [7, 10–12, 14], which inspired the talk given at [iℕδA℣] in November 2019, and finally led to the present paper. This work has been supported by the Project EFI ANR-17-CE40-0030 of the French National Research Agency.

References

1. Bakry, D., Émery, M.: Diffusions hypercontractives. In: Séminaire de probabilités, XIX, 1983/1984. Lecture Notes in Math., vol. 1123, pp. 177–206. Springer, Berlin (1985). https://doi.org/10.1007/BFb0075847
2. Bertini, L., Giacomin, G., Poquet, C.: Synchronization and random long time dynamics for mean-field plane rotators. Probab. Theory Related Fields **160**(3–4), 593–653 (2014). https://doi.org/10.1007/s00440-013-0536-6
3. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Mean-field limit for the stochastic Vicsek model. Appl. Math. Lett. **3**(25), 339–343 (2012)
4. Chaté, H., Ginelli, F., Grégoire, G., Raynaud, F.: Collective motion of self-propelled particles interacting without cohesion. Phys. Rev. E **77**(4), 046113 (2008)
5. Constantin, P., Kevrekidis, I.G., Titi, E.S.: Asymptotic states of a Smoluchowski equation. Arch. Ration. Mech. Anal. **174**(3), 365–384 (2004)
6. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Stud. Sci. Math. Hungar. **2**, 299–318 (1967)
7. Degond, P., Diez, A., Frouvelle, A., Merino-Aceituno, S.: Phase transitions and macroscopic limits in a BGK model of body-attitude coordination. J. Nonlinear Sci. (2020). https://doi.org/10.1007/s00332-020-09632-x
8. Degond, P., Frouvelle, A., Liu, J.G.: Macroscopic limits and phase transition in a system of self-propelled particles. J. Nonlinear Sci. **23**(3), 427–456 (2013). https://doi.org/10.1007/s00332-012-9157-y
9. Degond, P., Frouvelle, A., Liu, J.G.: Phase transitions, hysteresis, and hyperbolicity for self-organized alignment dynamics. Arch. Ration. Mech. Anal. **216**(1), 63–115 (2015). https://doi.org/10.1007/s00205-014-0800-7
10. Degond, P., Frouvelle, A., Merino-Aceituno, S.: A new flocking model through body attitude coordination. Math. Models Methods Appl. Sci. **27**(06), 1005–1049 (2017). https://doi.org/10.1142/S0218202517400085
11. Degond, P., Frouvelle, A., Merino-Aceituno, S., Trescases, A.: Quaternions in collective dynamics. Multiscale Mod. Simul. **16**(1), 28–77 (2018). https://doi.org/10.1137/17M1135207
12. Degond, P., Frouvelle, A., Merino-Aceituno, S., Trescases, A.: Alignment of self-propelled rigid bodies : from particle systems to macroscopic equations. In: Giacomin, G., Olla, S., Saada, E., Spohn, H., Stoltz, G., Stoltz, G. (eds.) Stochastic Dynamics Out of Equilibrium. IHPStochDyn 2017, Springer Proceedings in Mathematics and Statistics, vol. 282, pp. 28–66. Springer (2019)
13. Degond, P., Motsch, S.: Continuum limit of self-driven particles with orientation interaction. Math. Models Methods Appl. Sci. **18**, 1193–1215 (2008)
14. Diez, A.: Propagation of chaos and moderate interaction for a piecewise deterministic system of geometrically enriched particles. Electron. J. Probab. **25**, Paper No. 90, 38 (2020). https://doi.org/10.1214/20-ejp496
15. Fatkullin, I., Slastikov, V.: Critical points of the Onsager functional on a sphere. Nonlinearity **18**, 2565–2580 (2005)
16. Figalli, A., Kang, M.J., Morales, J.: Global well-posedness of the spatially homogeneous Kolmogorov-Vicsek model as a gradient flow. Arch. Ration. Mech. Anal. **227**(3), 869–896 (2018). https://doi.org/10.1007/s00205-017-1176-2
17. Frouvelle, A., Liu, J.G.: Dynamics in a kinetic model of oriented particles with phase transition. SIAM J. Math. Anal. **44**(2), 791–826 (2012)
18. Holley, R., Stroock, D.: Logarithmic Sobolev inequalities and stochastic Ising models. J. Stat. Phys. **46**(5–6), 1159–1194 (1987). https://doi.org/10.1007/BF01011161
19. Horn, A.: Doubly stochastic matrices and the diagonal of a rotation matrix. Am. J. Math. **76**, 620–630 (1954). https://doi.org/10.2307/2372705
20. Hsu, E.P.: Stochastic Analysis on Manifolds. Graduate Series in Mathematics, vol. 38. American Mathematical Society, Providence (2002)

21. Liu, H., Zhang, H., Zhang, P.: Axial symmetry and classification of stationary solutions of Doi-Onsager equation on the sphere with Maier-Saupe potential. Commun. Math. Sci. **3**(2), 201–218 (2005)
22. Mueller, C.E., Weissler, F.B.: Hypercontractivity for the heat semigroup for ultraspherical polynomials and on the n-sphere. J. Funct. Anal. **48**(2), 252–283 (1982). https://doi.org/10.1016/0022-1236(82)90069-6
23. Pinsker, M.S.: Information and information stability of random variables and processes. Translated and edited by Amiel Feinstein. Holden-Day, San Francisco, Calif.-London-Amsterdam (1964)
24. Sznitman, A.S.: Topics in propagation of chaos. In: École d'Été de Probabilités de Saint-Flour XIX—1989. Lecture Notes in Mathematics, vol. 1464, pp. 165–251. Springer, Berlin (1991)
25. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., Shochet, O.: Novel type of phase transition in a system of self-driven particles. Phys. Rev. Lett. **75**(6), 1226–1229 (1995)
26. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003). https://doi.org/10.1090/gsm/058
27. Wang, H., Hoffman, P.J.: A unified view on the rotational symmetry of equilibria of nematic polymers, dipolar nematic polymers and polymers in higher dimensional space. Commun. Math. Sci. **6**(4), 949–974 (2008)

The Half-Space Problem for the Boltzmann Equation with Phase Transition at the Boundary

François Golse

In memory of Basil Nicolaenko (1942–2007)

Abstract Consider the steady Boltzmann equation with slab symmetry for a monatomic, hard sphere gas in a half space above its condensed phase. The present paper studies the existence and uniqueness of a uniformly, exponentially decaying solution in the vicinity of the Maxwellian equilibrium with zero bulk velocity, with the same temperature as that of the condensed phase, and whose pressure is the saturating vapor pressure at the temperature of the interface. This problem has been studied numerically by Y. Sone, K. Aoki and their collaborators—see section 2 of (Bardos et al., J Stat Phys 124:275–300, 2006) for a detailed presentation of these works. More recently Liu and Yu (Arch Ration Mech Anal 209:869–997, 2013) have proposed a mathematical strategy to handle problems of this type. In this paper, we describe an alternative approach to one of their results obtained in collaboration with Bernhoff (Arch Ration Mech Anal 240:51–98, 2021).

1 The Sone Half-Space Problem with Condensation/Evaporation

Consider a monatomic, hard sphere gas filling the half space

$$\mathbf{R}^3_+ := \{(x, y, z) \in \mathbf{R}^3 \text{ s.t. } z > 0\},$$

on top of its condensed phase, filling the domain $\{(x, y, z) \in \mathbf{R}^3 \text{ s.t. } z \leq 0\}$. We are concerned with the existence and uniqueness of a steady solution with slab symmetry to the Boltzmann equation for the gas, assuming that the state of the gas at

F. Golse (✉)
Ecole Polytechnique, CMLS, Palaiseau Cedex, France
e-mail: francois.golse@polytechnique.edu

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_8

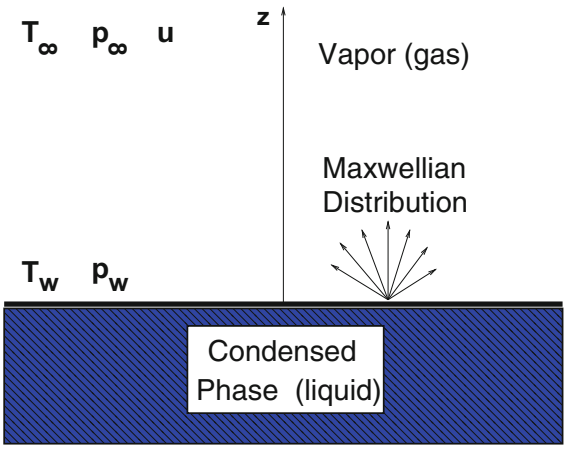

Fig. 1 Interface temperature of the liquid T_w, saturating vapor pressure p_w. As $z \to +\infty$, the gas distribution function converges to a Maxwellian with temperature T_∞, pressure p_∞ and bulk velocity $(0, 0, u)$

infinity is a Maxwellian equilibrium with prescribed pressure p_∞, temperature T_∞ and bulk velocity $(0, 0, u)$, while the velocity distribution function of gas molecules emitted towards the gas at the interface with the condensed phase is the centered Maxwellian with the temperature T_w of the condensed phase at the interface, and with pressure p_w equal to the saturating vapor pressure at the temperature T_w: see Fig. 1.

The Boltzmann equation describing this situation takes the form

$$v_z \partial_z F(z, v) = C(F)(z, v)\,, \qquad z > 0\,,\ v \in \mathbf{R}^3\,. \tag{1}$$

The unknown is the velocity distribution function $F \equiv F(z, v) \geq 0$, which depends on the three components of the molecular velocity $v = (v_x, v_y, v_z) \in \mathbf{R}^3$, and on the only height variable $z > 0$—this being precisely the assumption of "slab symmetry" often used in the context of half-space problems. The Boltzmann collision integral is given by the formula

$$C(F)(z, v) := \iint_{\mathbf{R}^3 \times \mathbf{S}^2} (F(z, v')F(z, v'_*) - F(z, v)F(z, v_*))|(v - v_*) \cdot \omega| dv_* d\omega \tag{2}$$

(up to some unessential scaling factor involving the molecular radius), with the notation

$$v' := v - (v - v_*) \cdot \omega\omega\,, \qquad v'_* := v_* + (v - v_*) \cdot \omega\omega\,. \tag{3}$$

The Boltzmann equation (1) is supplemented with "boundary" conditions as $z \to +\infty$ and as $z \to 0^+$. Henceforth, we shall use systematically the following notation

for Maxwellians:

$$\mathcal{M}_{p,u,T}(v) := \frac{p}{(2\pi)^{3/2}T^{5/2}} \exp\left(-\frac{v_x^2 + v_y^2 + (v_z - u)^2}{2T}\right). \tag{4}$$

Notice that we consider only Maxwellians which are centered in the tangential velocity variable v_x and v_y. In other words, the bulk velocity of all the Maxwellians considered here is of the form $(0, 0, u)$. Besides, we have chosen to use the pressure p and temperature T, instead of the density and temperature as the thermodynamic parameters in the definition of the Maxwellian.

At infinity, it is assumed that the solution of the Boltzmann equation (1) converges to a Maxwellian equilibrium of the form

$$F(z, v) \to \mathcal{M}_{p_\infty,u,T_\infty}(v), \qquad v \in \mathbf{R}^3, \ z \to +\infty. \tag{5}$$

On the other hand, it is assumed that, at the gas-liquid interface $z = 0$, the velocity distribution function of gas molecules emitted in the direction of the gas is

$$F(0, v) = \mathcal{M}_{p_w,0,T_w}(v), \qquad v \in \mathbf{R}^3, \ v_z > 0. \tag{6}$$

Typically, one chooses p_w to be the saturating vapor pressure at the temperature T_w. Evaporation corresponds to solutions of (1)–(5)–(6) for which $u > 0$, while condensation corresponds to solutions of the same equations with $u < 0$.

A natural question is to understand how many of the parameters p_∞, T_∞, p_w, T_w and u can be chosen freely. Obvious scaling considerations show that the relevant dimensionless parameters are the pressure ratio p_∞/p_w, the temperature ratio T_∞/T_w and the (signed) Mach number at infinity $-u/c_\infty$, where $c_\infty = \sqrt{5T_\infty/3}$ is the speed of sound at infinity. The numerical simulations conducted by Y. Sone, K. Aoki and their collaborators from the Kyoto school suggest that there is a dramatic change in the number of free parameters in this problem as u crosses the values 0 and $\pm c_\infty$. More precisely, the number of solvability conditions on the parameters p_∞/p_w, T_∞/T_w, $-u/c_\infty$ for the existence (and uniqueness) of a solution of the steady Boltzmann equation (1) satisfying the interface condition (6) and the condition at infinity (5) are summarized in Table 1. The interested reader is referred to section 2.1 in the survey paper [5] for a more complete description of these compatibility conditions, taking into account the possibility of a tangential bulk velocity—which is set to 0 throughout the present paper for the sake of simplicity. The relevant original references to the numerous contributions of the Kyoto school to this important problem are [1, 2, 23, 24, 26–30] and can be found in the bibliography section of [5]. Otherwise, a complete description is provided in chapters 6 and 7 of [25]—see especially section 6.1 in chapter 6, which deals with the case of a plane interface as in the present paper. Other geometries (spherical or cylindrical interfaces) are also treated in great detail in chapter 6 of [25].

Table 1 Sone's solvability conditions for the problem (1)–(6)–(5)

Normal velocity	Phase transition	Solvability condition(s)
$u > c_\infty$	Supersonic evaporation	No solution
$0 \leq u \leq c_\infty$	Subsonic evaporation	$p_\infty/p_w = h_1(u/c_\infty)$ and $T_\infty/T_w = h_2(u/c_\infty)$
$0 > u > -c_\infty$	Subsonic condensation	$p_\infty/p_w = F_s(u/c_\infty, T_\infty/T_w)$
$u = -c_\infty$	Sonic condensation	$p_\infty/p_w \geq F_b(-1, T_\infty/T_w)$
$u < -c_\infty$	Supersonic condensation	$p_\infty/p_w > F_b(u/c_\infty, T_\infty/T_w)$

Some parts of this table (in particular the fact that solutions corresponding to supersonic condensation cannot exist for arbitrarily small p_∞/p_w) can be confirmed by elementary computations based on conservation laws and the Boltzmann H Theorem: see [9, 31].

From a mathematical point of view, this table indicates a change of dimension in the set of solutions to the problem (1)–(6)–(5) as the Mach number at infinity u/c_∞ varies over the real line—assuming of course that the functions h_1, h_2, F_s and F_b are of class C^1 at least, so that the equations in the right column of the table above define bona fide differential manifolds. (Needless to say, the functions h_1, h_2, F_s and F_b are not known explicitly, but tabulated. One can therefore not hope to check that these functions are of class C^1 by inspection.) It is natural to surmise that there is some deep topological interpretation for this picture, yet to be fully understood.

2 Transition from Evaporation to Condensation

One of the difficulties in arriving at a complete mathematical justification of the results reported in Table 1 is due to the global nature of the problem. In other words, the solutions described in this table are in general not perturbations around some well-known, explicit solution of the problem. There is however an obvious, but important exception:

$$p_w = p_\infty\,, \quad T_w = T_\infty\,, \quad \text{and} \quad u = 0 \implies \mathcal{M}_{p_\infty,u,T_\infty} \text{ is a solution of (1)–(6)–(5).}$$

According to Theorem 5.1 in [5], the only nonnegative, classical solution of (1) satisfying the conservation laws of mass, momentum and energy, and Boltzmann's H theorem, together with the condition (5) at infinity with $u = 0$ is the uniform Maxwellian $F(z, v) = \mathcal{M}_{p_\infty,0,T_\infty}(v)$, so that, with the notations used in Table 1,

$$h_1(0) = h_2(0) = 1\,.$$

In the present paper, we seek to understand the situation described in Table 1 in the vicinity of $p_\infty/p_w = T_\infty/T_w = 1$ and $u = 0$. This is a very interesting regime, corresponding to the transition from evaporation to condensation. Table 1 suggests

that the dimensionality of the set of solutions of (1)–(6)–(5) jumps from 1 to 2 as u crosses the value 0. This transition has been studied in detail by means of formal asymptotic analysis in chapter 7 of [25], especially in sections 7.1–7.2. Besides, Sone's discussion of the problem reported in chapter 6 of [25] makes it clear that this is the only explicit exact solution of (1)–(6)–(5) around which one can hope to study the problem by perturbation arguments.

Sone's asymptotic analysis of the transition from evaporation to condensation in chapter 7 suggests that the sudden change of dimension of the set of solutions of (1)–(6)–(5) comes from the existence, for $u < 0$ and $|u| \ll c_\infty$ of a *slowly varying* solution, specifically of a solution which is a function of the slow variable $|u|z/c_\infty$. This solution is a local Maxwellian up to the first order in the small parameter $|u|/c_\infty$, with constant pressure and normal velocity fields, and with a temperature field of the form

$$T_\infty - ae^{buz},$$

where $b > 0$ is a constant and $a \in \mathbf{R}$ a free parameter. Clearly $e^{buz} \to +\infty$ as $z \to \infty$ if $u > 0$, so that the only admissible solution in this case is $a = 0$. However, if $u < 0$, the term $ae^{buz} \to 0$ as $z \to +\infty$, so that the first order temperature field above remains bounded (and even converges to the constant T_∞ as $z \to +\infty$). Then these asymptotic solutions are corrected by a rapidly varying boundary layer term, which decays exponentially fast as $z \to +\infty$. The analysis with the temperature correction presented here explains the jump in dimensionality across $u = 0$ in the set of solutions of (1)–(6)–(5): indeed, the free parameter a in the case $u < 0$, i.e. in the case of condensation, accounts for the extra degree of freedom in the set of solutions of (1)–(6)–(5).

In other words, if one eliminates the slowly varying component in solutions of (1)–(6)–(5) for $u < 0$, i.e. in the case of condensation, and for $|u| \ll 1$, one can hope that the evaporation curve given by the parametric representation

$$p_\infty/p_w = h_1(u/c_\infty) \quad \text{and} \quad T_\infty/T_w = h_2(u/c_\infty)$$

for $0 < u < c_\infty$ extends in a curve drawn on the condensation surface of equation

$$p_\infty/p_w = F_s(u/c_\infty, T_\infty/T_w)$$

for $0 < -u \ll c_\infty$. This curve corresponds to solutions of (1)–(6)–(5) for $u < 0$ which decay exponentially fast as $z \to \infty$, uniformly as $u \to 0^-$, i.e. near the edge of the condensation surface.

3 Perturbation Setting and Main Result

Our purpose is to investigate the diagram represented on Fig. 2 in the vicinity of the point where the evaporation curve C meets the condensation surface S, which corresponds to

$$p_w = p_\infty\,, \quad T_w = T_\infty\,, \quad \text{and} \quad u = 0\,.$$

Throughout the present paper, we shall operate under the following smallness assumption:

$$\left|\frac{p_\infty}{p_w} - 1\right| + \left|\frac{u}{c_\infty}\right| + \left|\frac{T_\infty}{T_w} - 1\right| \ll 1\,. \tag{7}$$

Without loss of generality, we henceforth set

$$p_\infty = T_\infty = 1\,,$$

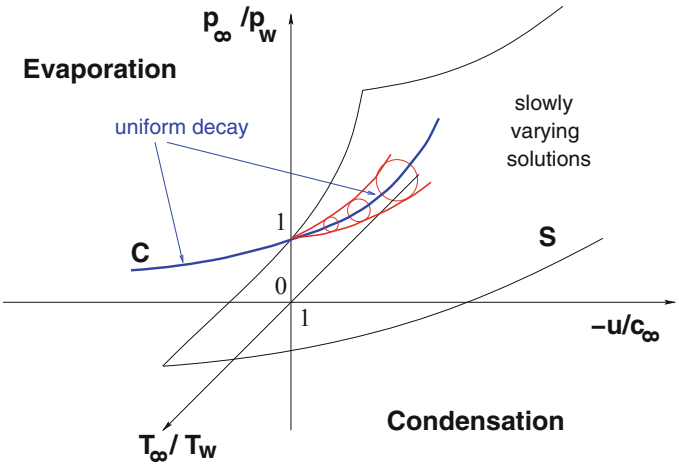

Fig. 2 The blue, evaporation curve C of parametric equation $p_\infty/p_w = h_1(u/c_\infty)$ and $T_\infty/T_w = h_2(u/c_\infty)$ extends in a curve drawn on the condensation surface S of equation $p_\infty/p_w = F_s(u/c_\infty, T_\infty/T_w)$. This blue curve drawn on the surface S corresponds to parameters for which the problem (1)–(6)–(5) has a solution which decays exponentially fast as $z \to +\infty$, uniformly in u as $u \to 0^-$. One possible line of investigation for future work could be to perturb the solution corresponding to the parameters on the blue curve by a slowly varying mode on the condensation surface S. It is expected that one could obtain in this way a piece of the surface S represented by the red cusp. Most likely, the size of the domain in the set of parameters for which a solution of (1)–(6)–(5) can be constructed by such a perturbation argument from a point on the extended blue curve will shrink as this point approaches the edge of the condensation surface S. This is represented on this diagram by the red disks of diminishing radius whose envelope is precisely the red cusp. This part of the picture, however, remains to be confirmed

and introduce the notation

$$M := \mathcal{M}_{1,0,1}$$

for the centered, reduced Gaussian distribution. Henceforth, it will be especially convenient to shift the velocity variable by u (the bulk velocity at infinity) and to use the velocity variable

$$\xi := v - (0, 0, u)$$

instead of the original variable $v \in \mathbf{R}^3$. In this way, the normal distribution M is the equilibrium state at infinity, and the distribution function will be sought in the form of a perturbation of the state at infinity, i.e.

$$F(z, v) = M(\xi)(1 + f(z, \xi)) .$$

In these new variables, and with this new unknown function, the half-space problem (1)–(6)–(5) becomes

$$\begin{cases} (\xi_z + u)\partial_z f(z, v) + \mathcal{L}f(z, \xi) = Q(f)(z, \xi) , & z > 0, \ \xi \in \mathbf{R}^3, \\ f(0, \xi) = f_b(\xi) & \text{for } \xi_z > -u , \\ f(z, \xi) \to 0 & \text{as } z \to +\infty , \end{cases} \tag{8}$$

where

$$\mathcal{L}f := -M^{-1} DC(M) \cdot (Mf) , \qquad Q(f) := M^{-1} C(Mf) .$$

Translations in the velocity variables in all directions parallel to the gas-liquid interface are much less important for this problem than in the direction orthogonal to that interface. For this reason, we shall discuss in the present paper only the case where the boundary data f_b is an even function of (ξ_x, ξ_y), and, accordingly, seek the unknown distribution function fluctuation f in the form

$$f(z, \xi_x, \xi_y, \xi_z) = f(z, -\xi_x, -\xi_y, \xi_z) .$$

Our main result on this problem is summarized in the following theorem. This result—and all the work reported in the present paper—has been obtained in collaboration with N. Bernhoff, and is discussed in detail in our joint paper [6].

Theorem 1 *There exists constants $\epsilon, \gamma^*, E, R > 0$ such that, for each boundary data $f_b \equiv f(\xi)$ which is even in ξ_x, ξ_y and satisfies*

$$\|(1 + |\xi|)^3 \sqrt{M} f_b\|_{L^\infty_\xi} \leq \epsilon ,$$

and for each u s.t. $0 < |u| < R$, the half-space problem (8) has a solution $f_u \equiv f(z, \xi)$ which is even in ξ_x, ξ_y and satisfies the uniform decay bound

$$\|(1 + |\xi|)^3 \sqrt{M} f_u(z, \cdot)\|_{L^\infty_\xi} \leq E e^{-\gamma z}$$

for all $0 < \gamma < \gamma^*$ if and only if

$$\begin{cases} \displaystyle\int_{\mathbf{R}^3} (\xi_z + u) Y_1[u](\xi) \mathfrak{R}_u[f_b](\xi) M(\xi) d\xi = 0\,, \\ \displaystyle\int_{\mathbf{R}^3} (\xi_z + u) Y_2[u](\xi) \mathfrak{R}_u[f_b](\xi) M(\xi) d\xi = 0\,. \end{cases} \tag{9}$$

In these compatibility conditions, the functions $Y_j[u]$ and $\mathfrak{R}_u[f_b]$ are defined below, in (16) and (18) respectively.

It remains to check that Sone's original problem is solved by the theorem above. Indeed, Sone's boundary data

$$F(0, v) = \mathcal{M}_{p_w, 0, T_w}(v)$$

or, equivalently

$$M(\xi)(1 + f_b(\xi)) = \mathcal{M}_{p_w, -u, T_w}(\xi)\,,$$

satisfy the assumptions of our main theorem above. Observe indeed that

$$0 < T_w < 2 \implies \sup_{0 \leq p_w + |u| \leq C} \frac{\mathcal{M}_{p_w, -u, T_w}(\xi)}{\sqrt{M}(\xi)} \to 0 \text{ as } |\xi| \to \infty\,.$$

Therefore, provided that the smallness condition (7) is satisfied,

$$f_b(\xi) := \frac{\mathcal{M}_{p_w, -u, T_w}}{M} - 1 \text{ satisfies } \|(1 + |\xi|)^3 \sqrt{M} f_b\|_{L^\infty_\xi} < \epsilon\,,$$

so that such an f_b is an example of boundary data to which the main theorem above applies.

Theorem 1 shows that the three-parameter family of Sone's boundary data must satisfy the two compatibility conditions (9) in order for the solution f_u of the half-space problem (8) to vanish at infinity exponentially fast uniformly in $0 < |u| < \epsilon$.

Notice however that Theorem 1 does not guarantee that these two compatibility conditions are C^1 functions of $(p_w, -u, T_w)$, and that their differentials at $(1, 0, 1)$ are linearly independent. Therefore, we cannot apply the implicit function theorem to deduce that applying the two compatibility conditions (9) to the Sone boundary data

$$f_b(\xi) := \frac{\mathcal{M}_{p_w, -u, T_w}}{M} - 1$$

results in a bona fide C^1 curve in the three-dimensional set of parameters

$$\{(p_w/p_\infty, -u/c_\infty, T_w/T_\infty) \quad \text{s.t. } p_w, p_\infty, T_w, T_\infty > 0\,,\ u \in \mathbf{R}\}\,.$$

There is however one suggestive remark on the tangent line to this "curve" at the point $(p_w/p_\infty, -u/c_\infty, T_w/T_\infty) = (1, 0, 1)$. Assume that the solution (8) whose existence and uniqueness is predicted by Theorem 1 is a C^1 function of u near $u = 0$. Differentiating formally in u at $u = 0$ in the problem (8), and observing that $f\Big|_{u=0} = 0$ according to Theorem 5.1 in [5] suggests that

$$\begin{cases} \xi_z \partial_z \dfrac{df}{du}\Big|_{u=0}(z,\xi) + \mathcal{L}\dfrac{df}{du}\Big|_{u=0}(z,\xi) = 0\,, & z > 0,\ \xi \in \mathbf{R}^3, \\ \dfrac{df}{du}\Big|_{u=0}(0,\xi) = \dfrac{dp_w}{du}\Big|_{u=0} - \xi_z + \dfrac{dT_w}{du}\Big|_{u=0}\frac{1}{2}(|\xi|^2 - 5) & \text{for } \xi_z > 0\,, \\ \dfrac{df}{du}\Big|_{u=0}(z,\xi) \to 0 & \text{as } z \to +\infty\,. \end{cases}$$

Bardos et al. [4] have proved that for each boundary data

$$g_b \equiv g_b(\xi) = g_b(\xi_x, \xi_y, \xi_z)$$

in the space $L^2(\mathbf{R}^3; (1 + |\xi|)Md\xi)$ that is even in the variables ξ_x, ξ_y, the linear half-space problem

$$\begin{cases} \xi_z \partial_z g(z,\xi) + \mathcal{L}g(z,\xi) = 0\,, \quad z > 0\,,\ \xi \in \mathbf{R}^3\,, \\ \displaystyle\int_{\mathbf{R}^3} \xi_z g M d\xi = 0\,, \quad g(0,\xi) = g_b(\xi) \text{ for } \xi_z > 0\,, \end{cases}$$

has a unique solution $g \equiv g(z,\xi) \in L^\infty((0,+\infty); L^2(\mathbf{R}^3; (1 + |\xi|)Md\xi))$. This solution satisfies

$$g(z,\xi) \to \Lambda_1[g_b] + \Lambda_2[g_b]\frac{1}{2}(|\xi|^2 - 3) \quad \text{as } z \to \infty\,,$$

where Λ_1, Λ_2 are continuous linear functionals on $L^2(\mathbf{R}^3; (1 + |\xi|)Md\xi)$. Assuming, as in Sone's Table 1, that our compatibility conditions are of the form

$$p_w/p_\infty = h_1(u/c_\infty)\,, \quad \text{and} \quad T_w/T_\infty = h_2(u/c_\infty)\,,$$

and deriving formally in u near 0 suggests that

$$\dot{h}_1(0) = -\sqrt{\tfrac{5}{3}}\,(\Lambda_1[\xi_z] + \Lambda_2[\xi_z])\,, \quad \dot{h}_2(0) = -\sqrt{\tfrac{5}{3}}\Lambda_2[\xi_z]\,.$$

If one could prove somehow that the solution f whose existence and uniqueness is granted by Theorem 1 is a C^1 function of u near $u = 0$ in some appropriate functional setting, one could hope to conclude that the two compatibility conditions obtained in Theorem 1 define indeed a C^1 curve near the point $(1, 0, 1)$ in the set of parameters

$$\{(p_w/p_\infty, -u/c_\infty, T_w/T_\infty) \quad \text{s.t. } p_w, p_\infty, T_w, T_\infty > 0\,,\ u \in \mathbf{R}\}\,.$$

by the implicit function theorem. To the best of our knowledge, this remains at present an open problem.

4 Comparison with Previous Results

We shall in this section discuss the differences and similarities between our Theorem 1, and earlier, related results in the literature.

First, the result obtained in [4] is a special case of the following, more general one. For each $p, T > 0$ and $u \in \mathbf{R}$, denote

$$\mathcal{L}_{p,u,T} f := -\mathcal{M}_{p,u,T}^{-1} DC(\mathcal{M}_{p,u,T}) \cdot (\mathcal{M}_{p,u,T} f)\,,$$

where

$$DC(F) \cdot G := \frac{d}{d\theta} C(F + \theta G)\Big|_{\theta=0}$$

designates the Gateaux derivative of the Boltzmann collision integral C at F in the direction G. In accordance with the notation used earlier in this paper, $\mathcal{L} = \mathcal{L}_{1,0,1}$. Consider the half-space problem for the Boltzmann equation linearized at $\mathcal{M}_{p,u,T}$:

$$\begin{cases} v_z \partial_z h(z, v) + \mathcal{L}_{p,u,T} h(z, v) = 0\,, & z > 0\,,\ v \in \mathbf{R}^3\,, \\ h(0, v) = h_b(v) & \text{for } v_z > 0\,, \\ h(z, v) \to 0 & \text{as } z \to +\infty\,. \end{cases} \tag{10}$$

It is assumed that $h_b \in L^2(\mathbf{R}^3; (1 + |v|)\mathcal{M}_{p,u,T} dv)$, and is even in v_x, v_y. Cercignagni had conjectured in [12] the existence and uniqueness of an even in v_x, v_y solution h of the problem above in $L^2((0, +\infty); L^2(\mathbf{R}^3; (1+|v|)\mathcal{M}_{p,u,T} dv))$ if and only if h_b satisfies N linear compatibility conditions, where N is given in Table 2.

Table 2 Cercignani's solvability conditions for the problem (10). Here $c := \sqrt{5T/3}$ is the speed of sound for the Maxwellian state $\mathcal{M}_{p,u,T}$

Normal velocity	Number N of solvability conditions
$c \leq u$	3
$0 \leq u < c$	2
$-c \leq u < 0$	1
$u < -c$	0

Equivalently, N is the maximum dimension of a subspace of $\mathrm{Ker}\mathcal{L}_{p,u,T}$ on which the quadratic form

$$\mathrm{Ker}\mathcal{L}_{p,u,T} \ni g \mapsto \int_{\mathbf{R}^3} v_z g(v)^2 \mathcal{M}_{p,u,T}(v) dv$$

is nonnegative, which is easily seen to be

$$\#\left(\{u-c, u, u+c\} \cap \mathbf{R}_+\right) \qquad \text{where } c := \sqrt{\tfrac{5}{3}T}\,.$$

Cercignagni's conjecture has been completely proved in [14] (see [3] for a partial result on a much simpler relaxation model of the Boltzmann equation).

Obviously, N increases from 1 to 2 as u increases across 0 (in other words, at the transition between evaporation and condensation), exactly as in Sone's Table 1.

Later, S. Ukai, T. Yang and S.-H. Yu studied a weakly nonlinear variant of the Cercignani's conjecture proved in [14]. They consider the half-space problem (8), and study $\mathbf{S}[u]$, the set of boundary data f_b which are even in ξ_x, ξ_y and such that $f_b/\sqrt{M}$ is rapidly decaying in $|\xi|$, and such that the problem (8) has a solution. One can think of $\mathbf{S}[u]$ as the stable manifold of $f = 0$ for the half-space equation (8), viewed as an evolution problem in the variable z.

Of course $f_b \equiv 0 \in \mathbf{S}[u]$ for all u. For $u \neq 0, \pm\sqrt{5/3}$, S. Ukai, T. Yang and S.-H. Yu prove in [32] that $\mathbf{S}[u]$ is, locally near 0, a C^1-manifold of codimension N. How to treat the degenerate cases $u \in \{0, \pm\sqrt{5/3}\}$ is explained in [15].

Yet, however interesting, this result does not solve Sone's original problem, except in the obvious case $u = 0$, dealt with more completely, without any smallness assumption in Theorem 5.1 of [5]. Indeed, as $u \to 0$, the local stable manifold $\mathbf{S}[u]$ constructed in [32] may shrink near 0 to the point that it may fail to contain Sone's boundary data

$$f_b(\xi) := \frac{\mathcal{M}_{p_W, -uT_W}}{M} - 1\,.$$

More recently, T.-P. Liu and S.-H. Yu [18] have studied Sone's problem from a stability point of view, obtaining solutions to the steady Boltzmann equation as long time limits of solutions to the evolution Boltzmann equation. Their paper is based on rather involved central manifold arguments, together with their previous work on the structure of the Green function for the linearized Boltzmann equation [17]. They obtain in this way a complete picture of the half-space problem with phase transition

at the boundary, which corroborates the very precise numerical exploration of the set of parameters conducted by Sone, Aoki and their collaborators in Kyoto in the 1980s-1990s.

As mentioned above (in the abstract and before the statement of Theorem 1), the present paper is a survey of the results obtained in the joint article [6] with N. Bernhoff. Our main results in this work, reported here in Theorem 1, correspond to cases 2 and 4 in Theorem 28 of [18]. Of course, the proofs of cases 2 and 4 in Theorem 28 of [18] are only sketched, but should follow from the general strategy presented in that paper. At variance with the argument presented in [18], our proof is self-contained and based on rather standard energy estimates, instead of the much more involved theory of Green functions.

Perhaps the novel element in our work lies in the combination of two earlier techniques: (a) the Ukai-Yang-Yu penalization technique, and (b) the much older Nicolaenko-Thurber theory of the generalized eigenvalue problem for the linearized Boltzmann collision integral, which we believe had not been used in the context of half-space problems until now.

5 The Nicolaenko-Thurber Generalized Eigenvalue Problem

We henceforth consider the Hilbert space

$$\mathfrak{H} := \{f \in L^2(\mathbf{R}^3; Md\xi) \text{ such that } f(\xi_x, \xi_y, \xi_z) = f(-\xi_x, -\xi_y, \xi_z) \text{ for a.e. } \xi \in \mathbf{R}^3\}.$$

For each $\phi \in L^1(\mathbf{R}^3; Md\xi)$, we set

$$\langle \phi \rangle := \int_{\mathbf{R}^3} \phi(\xi) M(\xi) d\xi \, .$$

The Generalized Eigenvalue Problem
For each real u near 0, find $\phi_u \in \mathfrak{H} \cap \mathrm{Dom}(\mathcal{L})$ such that

$$\text{(GEP)} \qquad \mathcal{L}\phi_u(\xi) = \tau_u(\xi_z + u)\phi_u \, , \qquad \langle (\xi_z + u)\phi_u^2 \rangle = -u \, .$$

Our main result on this problem is summarized in the next proposition.

Proposition 1 *There exists $r > 0$ and a real-analytic map*

$$(-r, r) \ni u \mapsto (\tau_u, \phi_u) \in \mathbf{R} \times (\mathfrak{H} \cap Dom(\mathcal{L}))$$

of solutions to the generalized eigenvalue problem (GEP) such that

$$u\tau_u < 0 \quad \textit{for all } u \in (-r, r) \, .$$

In particular

$$\tau_u = u\dot{\tau}_0 + O(u^2) \qquad \textit{with } \dot{\tau}_0 < 0\,.$$

Moreover, for each $s > 0$, one has

$$\sup_{|u|<r} \|(1+|\xi|)^s \sqrt{M}\phi_u\|_{L^\infty_\xi} \leq C_s < \infty\,.$$

Here is a good reason for studying the generalized eigenvalue problem (GEP) in connection with the transition between evaporation and condensation in Sone's half-space problem for the Boltzmann equation. Define

$$\Phi_u(z,\xi) := e^{-\tau_u z}\phi_u(\xi)$$

where ϕ_u is the solution of (GEP) provided by the proposition above. Observe that Φ_u satisfies

$$(\xi_z + u)\partial_z \Phi_u(z,\xi) + \mathcal{L}\Phi(z,\xi) = 0\,.$$

Besides

$$0 < -u \ll 1 \implies \Phi_u(z,\xi) = O(\exp(-\tfrac{1}{2}|u||\dot{\tau}_0|z)) \to 0 \quad \text{as } z \to +\infty\,,$$
$$0 < +u \ll 1 \implies \exp(\tfrac{1}{2}u|\dot{\tau}_0|z) = O(\Phi_u(z,\xi)) \to +\infty \quad \text{as } z \to +\infty\,.$$

This shows that the Nicolaenko-Thurber generalized eigenvalue problem (GEP) provides us with a smooth branch of slowly varying (i.e. depending on the slow variable $\zeta = |u|z$ for u near 0) solutions to the linearized Boltzmann equation, depending smoothly on u, and admissible only for $u < 0$ (i.e. in the condensation case). Indeed, it is only for $u < 0$ that these slowly varying solutions are bounded as $z \to +\infty$.

The proof of the Proposition 1 can be obtained by following the method sketched in [21], where the generalized eigenvalue problem is solved in the vicinity of the sonic speed.[1] Instead of following the careful description of the zeros of some appropriate Fredholm determinant as in [21], one can apply instead the Kato theory of holomophic families of unbounded self-adjoint operators to

$$L(z) := \mathcal{L} - z\xi_1\,.$$

[1] The possibility of extending the Nicolaenko-Thurber theory to the case $|u| \ll 1$ was mentioned to me by Prof. Nicolaenko in the late 1990s during one of my visits to his department at Arizona State University in Phoenix.

See the discussion in §3, section 1 of chapter VII in [16] (especially the penultimate paragraph on p. 386). Proceeding in this way, we obtain $u\tau_u$ as a usual eigenvalue $\lambda(\tau_u)$ of $L(\tau_u)$, so that one needs in the end to use the Open Mapping Theorem from complex analysis in order to invert the relation $u\tau_u = \lambda_0(\tau_u)$. See section 3 of [6] for a complete write-up of this argument.

6 Sketch of the Proof of Theorem 1

Before embarking on the proof of Theorem 1, we need some preparations.

6.1 The Linearized Collision Integral

First, we recall a few basic, but important facts about the linearized collision integral.

Lemma 1 (Hilbert, 1912) *The linearized collision integral $\mathcal{L}$ is an unbounded self-adjoint, nonnegative and Fredholm operator on $L^2(\mathbf{R}^3; Md\xi)$, with*

$$Dom\,\mathcal{L} = L^2(\mathbf{R}^3; (1+|\xi|)Md\xi) \quad \text{and} \quad (Ker\,\mathcal{L}) \cap \mathfrak{H} = span\,\{1, \xi_z, |\xi|^2\}\,.$$

This lemma is standard material in the theory of the Boltzmann equation: see for instance Theorem 7.2.1 in chapter 7, section 2 of [13].

One easily checks that the following functions :

$$X_\pm = \frac{|\xi|^2 \pm \sqrt{15}\xi_z}{\sqrt{30}}\,, \qquad X_0 \equiv \frac{|\xi|^2 - 5}{\sqrt{10}}$$

form an $\mathfrak{H}$-orthonormal basis of $(\mathrm{Ker}\,\mathcal{L}) \cap \mathfrak{H}$, which is orthogonal for the bilinear functional on $\mathrm{Dom}\,\mathcal{L}$:

$$(\phi, \psi) \mapsto \langle \xi_z \phi \psi \rangle\,.$$

Moreover

$$\langle \xi_z X_\pm^2 \rangle = \pm\sqrt{5/3}\,, \qquad \langle \xi_z X_0^2 \rangle = 0\,.$$

Since $\mathcal{L}$ is a nonnegative, self-adjoint Fredholm operator on $L^2(\mathbf{R}^3; Md\xi)$, there exists $\lambda_0 > 0$ such that $\mathcal{L}$ satisfies the following spectral gap inequality

$$g \in \mathrm{Dom}\,\mathcal{L} \cap (\mathrm{Ker}\,\mathcal{L})^\perp \implies \langle g\mathcal{L}g \rangle \geq \lambda_0 \langle g^2 \rangle\,.$$

This inequality is not sufficient for a priori estimates on (8). In their work, C. Bardos, R. Caflisch and B. Nicolaenko [4] have improved it into the following *weighted* spectral gap inequality.

Bardos-Caflisch-Nicolaenko Weighted Spectral Gap Inequality
There exists $\kappa_0 > 0$ such that

$$g \in \text{Dom}\,\mathcal{L} \cap (\text{Ker}\,\mathcal{L})^\perp \implies \langle g\mathcal{L}g\rangle \geq \kappa_0 \langle (1+|\xi|)g^2\rangle\,.$$

6.2 Lyapunov-Schmidt Reduction

In view of the role of slowly varying solutions in the half-space problem (8) for the Boltzmann equation, we must seek a way to filter out the slowly varying component of solutions to (8) in the condensation case $0 < -u \ll 1$. One way of doing this is by using a Lyapunov-Schmidt reduction—a tool often used in connection with bifurcation problems, and which appears for instance in the work of B. Nicolaenko and his collaborators on the shock profile problem for the Boltzmann equation: see [11, 19–21].

With the solution ϕ_u to the generalized eigenvalue problem obtained in Proposition 1, we construct the following pair of projections, in complete analogy with the procedure described in [11]:

$$\mathbf{p}_u g := -\langle (\xi_z + u)\psi_u g\rangle \phi_u\,, \qquad \mathbf{P}_u g := -\langle \psi_u g\rangle (\xi_z + u)\phi_u\,,$$

with the notation

$$\psi_u := \frac{\phi_u - \phi_0}{u}\,.$$

Lemma 2 *The linear maps* $\mathbf{p}_u$ *and* $\mathbf{P}_u$ *introduced above, are bounded operators on* $\mathfrak{H}$*, and satisfy the following properties:*

(a) both $\mathbf{p}_u$ *and* $\mathbf{P}_u$ *are projections on* $\mathfrak{H}$*, i.e.*

$$\mathbf{p}_u^2 = \mathbf{p}_u\,, \qquad \mathbf{P}_u^2 = \mathbf{P}_u\,, \qquad \textit{rank}\,\mathbf{p}_u = \textit{rank}\,\mathbf{P}_u = 1\,;$$

(b) one has

$$\textit{Ran}\,\mathbf{P}_u \subset (\textit{Ker}\mathcal{L})^\perp\,;$$

(c) for each $g \in \textit{Dom}\,\mathcal{L}$*, one has*

$$\mathbf{P}_u((\xi_z + u)g) = (\xi_z + u)\mathbf{p}_u g\,;$$

(d) for each $g \in \mathfrak{H} \cap Dom\,\mathcal{L}$*, one has*

$$(\xi_z + u)g \perp |\xi|^2 - 5 \implies \mathbf{P}_u(\mathcal{L}g) = \mathcal{L}(\mathbf{p}_u g)\,.$$

Here is how the Lyapunov-Schmidt reduction is applied to the half-space problem (8). Let $f \equiv f(z,\xi)$ solve the linear half-space problem with source

$$\begin{cases} (\xi_z + u)\partial_z f(z,v) + \mathcal{L}f(z,\xi) = Q(z,\xi)\,, & z > 0,\ \xi \in \mathbf{R}^3, \\ f(0,\xi) = f_b(\xi) & \text{for } \xi_z > -u\,, \\ f(z,\xi) \to 0 & \text{as } z \to +\infty\,. \end{cases} \tag{11}$$

The Lyapunov-Schmidt reduction consists in splitting $f(z,\cdot)$ into its images by $\mathbf{p}_u$ and $I - \mathbf{p}_u$. The result of this procedure is summarized in the following proposition; see [6] for a detailed proof.

Proposition 2 *Assume that* $0 < |u| < r$ *and that, for some* $\gamma > \max(\tau_u, 0)$*,*

$$e^{\gamma z} Q \in L^\infty((0,+\infty); \mathfrak{H} \cap (Ker\,\mathcal{L})^\perp)$$

while

$$e^{\gamma z} f \in L^\infty((0,+\infty); \mathfrak{H})\,.$$

Then f *is of the form* $f \equiv g(x,\xi) - h(x)\phi_u(\xi)$ *with*

$$g(z,\cdot) = (I - \mathbf{p}_u) f(z,\cdot)\,, \quad \text{and} \quad h(z)\phi_u = -\mathbf{p}_u f(z,\cdot)\,,$$

and

$$\begin{cases} (\xi_z + u)\partial_z g(z,\xi) + \mathcal{L}g(z,\xi) = (I - \mathbf{P}_u)Q(z,\xi)\,, & \xi \in \mathbf{R}^3\,,\ z > 0\,, \\ \langle (\xi_z + u)\psi_u g(z,\cdot)\rangle = 0\,, & z > 0\,, \\ \lim\limits_{z \to +\infty} g(z,\xi) = 0\,, & \xi \in \mathbf{R}^3\,, \\ h(z) = -\displaystyle\int_0^\infty e^{\tau_u y} \langle \psi_u Q\rangle(z+y)\,dy\,, & z > 0\,. \end{cases} \tag{12}$$

6.3 Adapting the Ukai-Yang-Yu Penalization Method

With the material prepared in the previous sections, we are ready to explain how the Ukai-Yang-Yu penalization method introduced in [32] can be used to handle the half-space problem (8).

Along with the non self-adjoint projections $\mathbf{p}_u$ and $\mathbf{P}_u$, it will be convenient to use the following rank-one, self-adjoint projections:

$$\Pi_\pm g := \langle g X_\pm \rangle X_\pm \,, \quad \Pi_0 g := \langle g X_0 \rangle X_0 \,, \quad \text{and} \quad \Pi := \Pi_+ + \Pi_0 + \Pi_- \,.$$

Observe that, if $g \in L^\infty((0,+\infty); \mathfrak{H} \cap \mathrm{Dom}\, \mathcal{L}))$ satisfies

$$(\xi_z + u)\partial_z g + \mathcal{L} g = (I - \mathbf{P}_u) Q \,, \qquad \text{and} \qquad \lim_{z \to \infty} g = 0 \,,$$

with $Q \in L^\infty((0,+\infty); (\mathrm{Ker}\, \mathcal{L})^\perp)$, then

$$\Pi((\xi_z + u) g) = 0 \,.$$

Hence, under the assumptions of, and with the notations used in Proposition 2, the function

$$g_\gamma(z,\xi) := e^{\gamma z} g(z,\xi)$$

is a solution to the penalized problem

$$(\xi_z + u)\partial_z g_\gamma(z,\xi) + \mathcal{L}^p g_\gamma(z,\xi) = (I - \mathbf{P}_u) e^{\gamma z} Q(z,\xi)$$

for all $\alpha, \beta, \gamma > 0$, where the penalized collision integral is defined by the formula[2]

$$\mathcal{L}^p g := \mathcal{L} g + \alpha \Pi_+((\xi_z + u) g) + \beta \mathbf{p}_u g - \gamma (\xi_z + u) g \,.$$

[2] During the meeting Prof. Schmeiser kindly reminded me that a somewhat reminiscent penalization of the linearized collision integral had been used in the paper [11], which predates the introduction of the penalization method in [32]. See the definition of the operator denoted M in formula (3.39) of [11], and Proposition 3.3 on p. 171 in the same reference. However, the idea of penalizing the collision integral is used quite differently in [11] and [32]. That the penalization method of [32] escaped the notice of the authors of the first fundamental contribution [4] to the theory of the half-space problem for the Boltzmann equation, who were obviously aware of its importance in the shock profile problem treated in [11], says a lot about the originality and depth of the ideas in [32].

Conversely, if g_γ solves the penalized problem for some $\alpha, \beta, \gamma > 0$, then

$$\frac{d}{dz}\begin{pmatrix}\langle(\xi_z+u)X_+ g_\gamma\rangle\\ \langle(\xi_z+u)X_0 g_\gamma\rangle\\ \langle(\xi_z+u)\psi_u g_\gamma\rangle\end{pmatrix} + (\mathcal{A}_u - \gamma I)\begin{pmatrix}\langle(\xi_z+u)X_+ g_\gamma\rangle\\ \langle(\xi_z+u)X_0 g_\gamma\rangle\\ \langle(\xi_z+u)\psi_u g_\gamma\rangle\end{pmatrix} = 0\,,$$

where we have denoted

$$\mathcal{A}_u := \begin{pmatrix}\alpha & 0 & -u\beta\langle\psi_u X_+\rangle\\ 0 & 0 & -\beta\langle\phi_u X_0\rangle\\ \alpha\langle\psi_u X_+\rangle & \tau_u/u & \tau_u - \beta\langle\psi_u\phi_u\rangle\end{pmatrix}.$$

One of the key ingredients in the proof of Theorem 1 is the following description of the spectrum of the matrix $\mathcal{A}_u$.

First, we observe that $u \mapsto \mathcal{A}_u$ is real-analytic for $|u| < r$, and that

$$\det(\mathcal{A}_0 - \lambda I) = (\alpha - \lambda)(\lambda^2 - \beta\langle\psi_0 X_0\rangle\lambda + \dot{\tau}_0\beta)\,.$$

Hence there exists $r' \in (0, r)$ so that, for $|u| < r'$, the matrix $\mathcal{A}_u$ has 3 simple real eigenvalues which are real-analytic functions of u and satisfy the ordering

$$\lambda_1(u) > \lambda_2(u) > 0 > \lambda_3(u)\,,$$

and more precisely, the uniform inequality:

$$\inf_{0<|u|<r'} \lambda_2(u) > 0 > \sup_{0<|u|<r'} \lambda_3(u)\,. \tag{13}$$

Henceforth, we denote by $u \mapsto (l_1(u), l_2(u), l_3(u))$ a real-analytic basis of eigenvectors of $\mathcal{A}_u^T$ for $|u| < r'$, such that

$$\mathcal{A}_u^T l_j(u) = \lambda_j(u) l_j(u)\,, \quad j = 1, 2, 3\,.$$

See [6] for the missing details.

6.4 A Strategy for Proving Theorem 1

With the preparations described above, we can now explain how the proof of Theorem 1 unfolds. It involves four main steps as indicated below.

6.4.1 Step 1: Defining a Penalized Collision Operator

Our first task is to choose the penalization parameters α, β, γ so that the penalized collision integral $\mathcal{L}^p$ satisfies the Bardos-Caflisch-Nicolaenko weighted spectral gap inequality *uniformly* in $|u| \ll 1$. Specifically, we prove the following lemma.

Lemma 3 *There exists $R, \Gamma, \kappa_1 > 0$ such that, whenever $0 < \alpha = \beta = 2\gamma < 2\Gamma$ and $|u| < R$, the penalized linearized collision integral*

$$\mathcal{L}^p g := \mathcal{L}g + \alpha \Pi_+((\xi_z + u)g) + \beta \mathbf{p}_u g - \gamma(\xi_z + u)g$$

satisfies

$$g \in (Dom\ \mathcal{L}) \cap \mathfrak{H} \implies \langle g\mathcal{L}^p g\rangle \geq \kappa_1 \langle (1 + |\xi|)g^2\rangle\,.$$

How to fit the parameters α, β, γ in order to obtain the weighted positivity property in the lemma above is done by inspection, and involves some tedious manipulations. However, these computations are rather elementary, and do not require knowing more than the Bardos-Caflisch-Nicolaenko spectral gap inequality recalled above. The argument follows [32] and [15]; see [6] for a complete proof.

The key point in connection with this lemma is that the uniform in $|u| < R$, weighted spectral gap constant κ_1 is related both to the exponential decay rate γ and to the "norm of the inverse" of $(\xi_z + u)\partial_z + \mathcal{L}^p$, the penalized linearized Boltzmann operator.

6.4.2 Step 2: Solving the Linearized, Penalized Half-Space Problem

This section and the next are based on the usual energy method for the penalized half-space problem: see for instance [15] for a detailed description of the method, which parallels the proof in [6]. The interested reader is referred to the latter reference for a complete write-up—which is rather lengthy, but without remarkable difference from earlier results, such as [32] or [15]. The only difference with these earlier references is the uniformity in $|u| \ll 1$ of the estimates so obtained, which must be checked carefully—and ultimately depends on the result of Step 1.

With $\alpha = \beta = 2\gamma > 0$ chosen as in Lemma 3, solve for $g_{u,\gamma}$ the problem

$$\begin{cases} (\xi_z + u)\partial_z g_{u,\gamma}(z,\xi) + \mathcal{L}^p g_{u,\gamma}(z,\xi) = e^{\gamma x}(I - \mathbf{P}_u)Q(z,\xi)\,, & z > 0\,,\ \xi \in \mathbf{R}^3, \\ g_{u,\gamma}(0,\xi) = g_b(\xi)\,, & \xi_z + u > 0\,. \end{cases} \tag{14}$$

More precisely, we solve this problem successively

(i) in $L^2((0,+\infty), \mathfrak{H} \cap \mathrm{Dom}\, \mathcal{L})$ by using some variant of the Riesz representation theorem, then
(ii) in $L^2((1+|\xi|)Md\xi; L^\infty(0,+\infty))$, by using the integral equation as in [14] to "improve" the bound on the z-dependence[3] from L^2 to L^∞, and finally
(iii) in $(1+|\xi|)^{-3}M^{-1/2}L^\infty((0,+\infty)\times\mathbf{R}^3)$ by using Grad's decay estimates for the gain part of the linearized collision integral, which can be found for instance in [10].

The key point in this step is that, by filtering out the slowly varying component of the solution, i.e. by looking at $g_{u,\gamma}$ instead of f, one manages to prove that the linear solution map

$$(g_b, Q) \mapsto g_{u,\gamma}$$

is bounded *uniformly* in u for $|u| < R$. This uniformity will be crucial in the next step.

6.4.3 Step 3: Solving the Nonlinear, Penalized Half-Space Problem

Apply the standard fixed point theorem, replacing the source term Q in (14) with

$$Q(e^{-\gamma z} g_{u,\gamma} - e^{-\gamma z} h_{u,\gamma}\phi_u)\,,$$

and keeping in mind that

$$h_{u,\gamma}(z) = -e^{-\gamma z}\int_0^\infty e^{(\tau_u - 2\gamma)y}\langle \psi_u Q(g_{u,\gamma} - h_{u,\gamma}\phi_u)\rangle(z+y)dy\,.$$

With the resulting fixed point $(g_{u,\gamma}, h_{u,\gamma})$, we construct the function

$$(z,\xi) \mapsto \tilde{f_u}(z,\xi) := e^{-\gamma z}(g_{u,\gamma}(z,\xi) - h_{u,\gamma}(z)\phi_u(\xi))\,,$$

[3] One should pay attention to the fact that the appropriate function space used in this argument is an anisotropic, or mixed Lebesgue space of the form $L^2_\xi(L^\infty_z)$, and not $L^\infty_z(L^2_\xi)$. That $L^2_\xi(L^\infty_z)$ is the function space of interest for this type of problem has been known for a long time—for instance it was already used in [14].

which is a solution of the problem

$$\begin{cases} (\xi_z + u)\partial_z \tilde{f}_u + \mathcal{L}\tilde{f}_u + \alpha\Pi_+((\xi_z + u)\tilde{f}_u) + \beta \mathbf{p}_u \tilde{f}_u = Q(\tilde{f}_u)\,, & \xi \in \mathbf{R}^3\,,\ z > 0\,, \\ \tilde{f}_u(0,\xi) = f_b(\xi)\,, & \xi_z + u > 0\,, \\ \lim\limits_{z\to+\infty} \tilde{f}_u(z,\xi) = 0\,, & \xi \in \mathbf{R}^3\,. \end{cases} \tag{15}$$

Specifically, we prove the existence of $\epsilon > 0$ such that, for each $f_b \equiv f_b(\xi)$, even in ξ_x, ξ_y and satisfying the bound

$$\|(1 + |\xi|)^3\sqrt{M}\, f_b\|_{L^\infty(\mathbf{R}^3)} \le \epsilon,$$

the problem (15) has a unique solution such that

$$(1 + |\xi|)^3\sqrt{M}|\tilde{f}_u(z,\xi)| \le O(\epsilon)e^{-\gamma z}$$

for all u such that $|u| < r''$, where $0 < r'' < \inf(r', R)$ is a small enough positive number.

The key point in this step is that the uniform in u bound on the linear solution operator obtained in Step 2 implies that the nonlinear solution operator is well defined on a small neighborhood of the origin *whose size is uniform in u* for $|u| < R$. All the constructions in the previous sections, especially the Lyapunov-Schmidt reduction in Sect. 6.2, based on the resolution of the generalized eigenvalue problem (GEP), and the resulting modification in the penalization method, i.e. introducing the projection $\mathbf{p}_u$ in the definition of $\mathcal{L}^p$, are aimed at obtaining this uniformity. In this way, we avoid the objection reported in Sect. 4 against using the result in [32] on Sone's half-space problem with evaporation or condensation at the gas-liquid interface.

6.4.4 Step 4: Removing the Penalization

At the end of Step 3, we have solved the nonlinear, penalized half-space problem (15) for *all* small enough boundary data f_b. The solution $\tilde{f}_u$ decays exponentially fast to 0 as z tends to infinity, and the exponential decay rate γ is uniform in u for $|u| \ll 1$. While the uniform in u exponential decay was one of our goals, we have not yet solved the original problem (8), which is the physically relevant one. In other words, we still have to remove the penalization in order to arrive at a proof of Theorem 1.

The origin of the compatibility conditions in Theorem 1 is to be found precisely in this part of the procedure. Since these compatibility conditions are at the core of the main result in this paper, we shall describe in full detail how to remove the penalization, and how this leads to the compatibility conditions in Theorem 1.

This is done as follows. Choose γ so that

$$0 < \gamma < \min\left(\Gamma, \inf_{0<|u|<r'} \lambda_2(u)\right),$$

where $\Gamma > 0$ appeared in Lemma 3, and where we have used one of the uniform inequalities in (13).

Set

$$Y_j[u](\xi) = l_j(u)_1 X_+(\xi) + l_j(u)_2 X_0(\xi) + l_j(u)_3 \psi_u(\xi)\,, \qquad 1 \le j \le 3\,, \tag{16}$$

where $l_j(u)_k$ is the k-th component of the eigenvector $l_j(u)$ of $\mathcal{A}(u)^T$.

Lemma 4 *If $g_\gamma \in L^\infty((0,+\infty); \mathfrak{H} \cap Dom\,\mathcal{L})$ solves the penalized problem (14), then*

$$\Pi_+ g_\gamma = \mathbf{p}_u g_\gamma = 0 \iff \langle(\xi_z + u) Y_j[u] g_\gamma\rangle\Big|_{z=0} = 0 \quad \textit{for } j = 1, 2\,.$$

Proof Observe that, for $j = 1, 2, 3$, one has

$$\begin{aligned}
\frac{d}{dz}\langle(\xi_z + u) Y_j[u] g_\gamma(z,\cdot)\rangle =& l_j(u)_1 \frac{d}{dz}\langle(\xi_z + u) X_+ g_\gamma(z,\cdot)\rangle \\
&+ l_j(u)_2 \frac{d}{dz}\langle(\xi_z + u) X_0 g_\gamma(z,\cdot)\rangle \\
&+ l_j(u)_3 \frac{d}{dz}\langle(\xi_z + u) \psi_u g_\gamma(z,\cdot)\rangle \\
=& - l_j(u)^T (\mathcal{A}_u - \gamma I) \begin{pmatrix} \langle(\xi_z + u) X_+ g_\gamma(z,\cdot)\rangle \\ \langle(\xi_z + u) X_0 g_\gamma(z,\cdot)\rangle \\ \langle(\xi_z + u) \psi_u g_\gamma(z,\cdot)\rangle \end{pmatrix}.
\end{aligned}$$

By definition of $l_j(u)$, one has

$$\begin{aligned}
\frac{d}{dz}\langle(\xi_z + u) Y_j[u] g_\gamma(z,\cdot)\rangle &= - (\lambda_j(u) - \gamma) l_j(u)^T \begin{pmatrix} \langle(\xi_z + u) X_+ g_\gamma(z,\cdot)\rangle \\ \langle(\xi_z + u) X_0 g_\gamma(z,\cdot)\rangle \\ \langle(\xi_z + u) \psi_u g_\gamma(z,\cdot)\rangle \end{pmatrix} \\
&= - (\lambda_j(u) - \gamma)\langle(\xi_z + u) Y_j[u] g_\gamma(z,\cdot)\rangle\,,
\end{aligned}$$

so that

$$\langle(\xi_z + u) Y_j[u] g_\gamma(z,\cdot)\rangle = e^{-(\lambda_j(u)-\gamma)z}\langle(\xi_z + u) Y_j[u] g_\gamma(0,\cdot)\rangle\,. \tag{17}$$

By the second uniform inequality in (13), one has $\lambda_3(u) < \gamma < 0$ for $0 < |u| < r''$. Since

$$g_\gamma \in L^\infty((0,+\infty);\mathfrak{H}\cap \mathrm{Dom}\,\mathcal{L}) \implies \langle(\xi_z+u)Y_j[u]g_\gamma(z,\cdot)\rangle \in L^\infty(0,+\infty)\,,$$

the equality (17) for $j=3$ implies that

$$\langle(\xi_z+u)Y_3[u]g_\gamma(z,\cdot)\rangle = 0 \quad \text{for all } z\geq 0 \quad \text{and} \quad 0<|u|<r''\,.$$

On the other hand, our choice of γ implies that

$$\lambda_j(u)-\gamma > 0 \quad \text{for all } j=1,2 \quad \text{and} \quad 0<|u|<r''\,,$$

so that (17) for $j=1,2$ implies that $\langle(\xi_z+u)Y_j[u]g_\gamma\rangle \in L^\infty(0,+\infty)$, without any restriction on the values of $\langle(\xi_z+u)Y_j[u]g_\gamma(0,\cdot)\rangle$.

If one assumes that $\langle(\xi_z+u)Y_j[u]g_\gamma(0,\cdot)\rangle = 0$ for $j=1,2$, then

$$\langle(\xi_z+u)Y_j[u]g_\gamma(z,\cdot)\rangle = 0 \quad \text{for all } j=1,2,3, \text{ all } z>0, \text{ and all } 0<|u|<r''\,.$$

Since the eigenvectors $l_1(u), l_2(u), l_3(u)$ are linearly independent, this implies that

$$\langle(\xi_z+u)X_+g_\gamma(z,\cdot)\rangle = \langle(\xi_z+u)X_0g_\gamma(z,\cdot)\rangle = \langle(\xi_z+u)\psi_u g_\gamma(z,\cdot)\rangle = 0$$

for all $z\geq 0$, which implies in turn that $\Pi_+ g_\gamma = \mathbf{p}_u g_\gamma = 0$.

Conversely, if $\Pi_+ g_\gamma = \mathbf{p}_u g_\gamma = 0$, then

$$(\xi_z+u)\partial_z g_\gamma + \mathcal{L}g_\gamma - \gamma(\xi_z+u)g_\gamma = (I-\mathbf{P}_u)Q,$$

so that

$$\frac{d}{dz}\langle(\xi_z+u)X_0g_\gamma\rangle = \gamma\langle(\xi_z+u)X_0g_\gamma\rangle,$$

and hence

$$\langle(\xi_z+u)X_0g_\gamma(z,\cdot)\rangle = e^{\gamma z}\langle(\xi_z+u)X_0g_\gamma(0,\cdot)\rangle\,.$$

Since

$$g_\gamma \in L^\infty((0,+\infty);\mathfrak{H}\cap \mathrm{Dom}\,\mathcal{L}) \implies \langle(\xi_z+u)X_0g_\gamma\rangle \in L^\infty((0,+\infty)),$$

we conclude from the equality above and the fact that $\gamma>0$ that

$$\langle(\xi_z+u)X_0g_\gamma(z,\cdot)\rangle = 0 \quad \text{for all } z\geq 0\,.$$

With $\Pi_+ g_\gamma = \mathbf{p}_u g_\gamma = 0$, this implies that

$$\langle (\xi_z + u) Y_j[u] g_\gamma(z, \cdot) \rangle = 0 \quad \text{for all } 1 \leq j \leq 3 \text{ and all } z \geq 0 \,.$$

In fact, as mentioned above, this equality is obvious for $j = 3$. In any case, it holds for $j = 1, 2$, and this completes the proof of the lemma. □

With Lemma 4, it is easy to conclude the proof of Theorem 1. Starting from the solution $\tilde{f}_u$ of (15) obtained in Step 3, we define

$$g_{u,\gamma}(z, \cdot) = e^{\gamma z}(I - \mathbf{p}_u)\tilde{f}_u(z, \cdot) \,, \qquad z > 0 \,,$$

and we set

$$\mathfrak{R}_u[f_b](\xi) = g_{u,\gamma}(0, \cdot) = (I - \mathbf{p}_u)\tilde{f}_u(0, \cdot) \,. \tag{18}$$

Since $\tilde{f}_u$ solves (15), the function $g_{u,\gamma}$ solves (14), with $Q(z, \xi) := Q(\tilde{f}_u)(z, \xi)$. According to Lemma 4, one can remove the penalization in (15) if and only if

$$0 = \langle (\xi_z + u) Y_j[u] g_\gamma(0, \cdot) \rangle = \langle (\xi_z + u) Y_j[u] \mathfrak{R}_u[f_b] \rangle$$

for $j = 1, 2$, which are precisely the compatibility conditions in Theorem 1.

The interested reader is referred to [6] for a complete proof.

7 Conclusion

We have proved that, near the stationary ($u_\infty = 0$) equilibrium Maxwellian state with $T_\infty = T_w$ and $p_\infty = p_w$, there exists a unique branch of solutions to Sone's half-space problem with uniform in u_∞ exponential decay far away from the liquid-gas interface

This branch of solutions extends the evaporation curve into the condensation surface in Sone's diagram, denoted S on Fig. 2. In other words, it is defined in the space of parameters

$$(p_\infty/p_w, -u/c_\infty, T_\infty/T_w)$$

by the same two compatibility conditions which define admissible parameters in the evaporation case, i.e.

$$p_\infty/p_w = h_1(u/c_\infty) \,, \qquad T_\infty/T_w = h_2(u/c_\infty) \,,$$

where the functions h_1, h_2 are extended to $u < 0$ with $|u| \ll 1$.

Our analysis is based on a perturbative argument for the steady Boltzmann equation and fails to establish positivity of the solution—exactly as in the treatment of the weak shock profile problem by B. Nicolaenko and his collaborators [11, 19–21]. However, it should be possible to remove this difficulty by using the Liu-Yu stability technique described in [18].

There are several open problems in connection with the result presented here in Theorem 1.

First, we have assumed everywhere in the paper that the bulk velocity at infinity is $(0, 0, u)$, in other words, that it is normal to the liquid-gas interface. One should consider the more general situation where the bulk velocity at infinity has a nonzero component tangential to the liquid-gas interface. In other words, one should consider the same half-space problem (8) without seeking the solution $f(z, \cdot)$ in the space $\mathfrak{H}$ of functions which are even in ξ_x, ξ_y. Since the tangential component of the bulk velocity at infinity does not appear in the streaming operator $(\xi_z + u)\partial_z$, including it in the discussion is not expected to lead to serious mathematical difficulties.

More serious mathematical difficulties are expected to be met if one seeks to recover Sone's condensation surface (denoted S on Fig. 2). Indeed, at this point, one must face the obviously challenging problem of handling a change in the topology (specifically, in the dimensionality, which is expected to jump from 1 to 2) of the set of solutions to the half-space problem as u decreases across the value 0.

One possibility for handling this problem could be to perturb about a solution of the half-space problem corresponding to parameters $(p_\infty/p_w, -u/c_\infty, T_\infty/T_w)$ lying on the extension of the evaporation curve C on the condensation surface S obtained in the present paper. One can expect that the maximal size of the perturbation for which the existence and uniqueness of a solution including a nontrivial slowly varying component can be proved by a standard fixed point method will vanish as one approaches the edge of the condensation surface S. The part of the condensation surface S which one could hope to obtain in this way is represented as a red cusp on Fig. 2. Since this cusp intersects the edge of S only at the only point corresponding to the temperature and pressure ratio $p_\infty/p_w = T_\infty/T_w = 1$, in other words to the trivial solution $F = \mathcal{M}_{p_w, 0, T_w}$, this problem might be tractable with the tools discussed in the present paper.

Finally, there obviously remains the issue of justifying completely the picture in Table 1 for all $u \in \mathbf{R}$, in other words, in nonperturbative regimes. While the work of T.-P. Liu and S.-H. Yu [18] provides us with a strategy to do so, it would certainly be interesting to investigate other approaches to this problem—or to the related shock profile problem for the Boltzmann equation without restriction on the shock strength. Topological methods in the style of those described in Part IV of [22] could perhaps be of some help in both problem. The work of A. Bobylev and N. Bernhoff [8] on shock profiles and half-space problems for discrete velocity models of the Boltzmann equation suggests that something along these lines could be attempted on the Boltzmann equation itself.

Acknowledgments I wish to thank Profs. Sone and Aoki who spent a lot of time teaching me their theory of the half-space problem with condensation/evaporation during several visits in Kyoto. I am also very much indebted to the late Prof. Nicolaenko for explaining to me his approach to the generalized eigenvalue problem described in Sect. 5—along with so many other things in mathematics. The question reported in the last paragraph of this paper had been a long standing program of ours, which we unfortunately did not have the time to complete. My collaboration on this problem with N. Bernhoff started from our joint interest in the topics discussed in [7, 8], to which I was introduced by Prof. Bobylev. Finally, I would like to thank Prof. Aoki for his friendly interest and support in our joint work [6].

References

1. Aoki, K., Nishino, K., Sone Y., Sugimoto H.: Numerical analysis of steady flows of a gas condensing on or evaporating from its plane condensed phase on the basis of kinetic theory: effect of gas motion along the condensed phase. Phys. Fluids A **3**, 2260–2275 (1991)
2. Aoki, K., Sone Y., Yamada T.: Numerical analysis of gas flows condensing on its plane condensed phase on the basis of kinetic theory. Phys. Fluids A **2**, 1867–1878 (1990)
3. Arthur, M.D., Cercignani, C.: Nonexistence of a steady rarefied supersonic flow in a half-space. Z. Angew. Math. Phys. **31**, 634–645 (1980)
4. Bardos, C., Caflisch, R.E., Nicolaenko, B.: The Milne and Kramers problems for the Boltzmann equation of a hard sphere gas. Commun. Pure and Appl. Math. **39**, 323–352 (1986)
5. Bardos, C., Golse, F., Sone, Y.: Half-space problems for the Boltzmann equation: a survey. J. Stat. Phys. **124**, 275–300 (2006)
6. Bernhoff, N., Golse, F.: On the boundary layer equations with phase transition in the kinetic theory of gases. Arch. Ration. Mech. Anal. **240**, 51–98 (2021)
7. Bernhoff, N., Bobylev, A.V.: Weak shock waves for the general discrete velocity model of the Boltzmann equation. Commun. Math. Sci. **5**, 815–832 (2007)
8. Bobylev, A.V., Bernhoff, N.: Discrete velocity models and dynamical systems. In: Bellomo, N., Gatignol, R. (eds.) Lecture Notes on the Discretization of the Boltzmann Equation, pp. 203–222, World Scientific, Singapore (2003)
9. Bobylev, A.V., Grzhibovskis, R., Heintz, A.: Entropy inequalities for evaporation/condensation problem in rarefied gas. J. Stat. Phys. **102**, 1156–1176 (2001)
10. Caflisch, R.E.: The Boltzmann equation with a soft potential I. Linear, space-homogeneous. Commun. Math. Phys. **74**, 71–95 (1980)
11. Caflisch, R.E., Nicolaenko, B.: Shock profile solutions of the Boltzmann equation. Commun. Math. Phys. **86**, 161–194 (1982)
12. Cercignani, C.: Half-space problems in the kinetic theory of gases. In: Kröner, E., Kirchgässner, K. (eds.) Trends in Applications of Pure Mathematics to Mechanics (Bad Honnef, 1985), pp. 35–50. Lecture Notes in Phys., vol. 249. Springer, Berlin (1986)
13. Cercignani, C., Illner, R., Pulvirenti, M.: The Mathematical Theory of Dilute Gases. Springer, New York (1994)
14. Coron, F., Golse, F., Sulem, C.: A Classification of Well-Posed Kinetic Layer Problems. Comm. Pure Appl. Math. **41**, 409–435 (1988)
15. Golse, F.: Analysis of the boundary layer equation in the kinetic theory of gases. Bull. Inst. Math. Acad. Sin. **3**, 211–242 (2008)
16. Kato, T.: Perturbation Theory for Linear Operators. Springer, Berlin, Heidelberg, New-York (1980)
17. Liu, T.-P., Yu S.-H.: The Green function and large-time behavior of solutions for one-dimensional Boltzmann equation. Commun. Pure Appl. Math. **57**, 1543–1608 (2004)
18. Liu, T.-P., Yu S.-H.: Invariant manifolds for steady Boltzmann flows and applications. Arch. Ration. Mech. Anal. **209**, 869–997 (2013)

19. Nicolaenko, B.: Shock wave solutions of the Boltzmann equation as a nonlinear bifurcation problem from the essential spectrum. In: Théories cinétiques classiques et relativistes. (Colloq. Internat. CNRS, No. 236, Paris, 1974), pp. 127–150, CNRS, Paris (1975)
20. Nicolaenko, B.: A general class of nonlinear bifurcation problems from a point in the essential spectrum. Application to shock wave solutions of kinetic equations. In: Applications of bifurcation theory (Proc. Advanced Sem., Univ. Wisconsin, Madison, Wis., 1976), pp. 333–357, Publ. Math. Res. Center Univ. Wisconsin, 38, Academic Press, New York-London (1977)
21. Nicolaenko, B., Thurber, J.K.: Weak shock and bifurcating solutions of the non-linear Boltzmann equation. J. de Mécanique **14**, 305–338 (1975)
22. Smoller, J.: Shock Waves and Reaction-Diffusion Equations. 2nd edn. Springer, New York (1994)
23. Sone, Y.: Kinetic theory of evaporation and condensation—Linear and nonlinear problems. J. Phys. Soc. Jpn. **45**, 315–320 (1978)
24. Sone, Y.: Kinetic theoretical studies of the half-space problem of evaporation and condensation. Transp. Theory Stat. Phys. **29**, 227–260 (2000)
25. Sone, Y.: Molecular Gas Dynamics. Theory, Techniques and Applications. Birkhäuser, Boston (2007)
26. Sone, Y., Aoki, K., Yamashita, I.: A study of unsteady strong condensation on a plane condensed phase with special interest in formation of steady profile. In: Boffi, V., Cercignani, C. (eds.) Rarefied Gas Dynamics, vol. 2, pp. 323–333. Teubner, Stuttgart (1986)
27. Sone, Y., Golse, F., Ohwada, T., Doi, T.: Analytical study of transonic flows of a gas condensing onto its plane condensed phase on the basis of kinetic theory. Eur. J. Mech. B/Fluids **17**, 277–306 (1998)
28. Sone, Y., Ohwada, T., Aoki, K.: Evaporation and condensation on a plane condensed phase: numerical analysis of the linearized Boltzmann equation for hard-sphere molecules. Phys. Fluids A **1**, 1398–1405 (1989)
29. Sone, Y., Onishi, Y.: Kinetic theory of evaporation and condensation. J. Phys. Soc. Jpn. **35**, 1773–1776 (1973)
30. Sone, Y., Sugimoto, H.: Strong evaporation from a plane condensed phase. In: Meier, G.E.A., Thompson, P.A. (eds.) Adiabatic Waves in Liquid-Vapor Systems, pp. 293–304, Springer, Berlin (1990)
31. Sone, Y., Takata, S., Golse, F: Notes on the boundary conditions for fluid-dynamic equations on the interface of a gas and its condensed phase. Phys. Fluids **13**, 324–334 (2001)
32. Ukai, S., Yang, T., Yu S.-H.: Nonlinear boundary layers of the Boltzmann equation I: existence. Commun. Math. Phys. **236**, 373–393 (2003)

Recent Developments on Quasineutral Limits for Vlasov-Type Equations

Megan Griffin-Pickering and Mikaela Iacobelli

Abstract Kinetic equations of Vlasov type are in widespread use as models in plasma physics. A well known example is the Vlasov-Poisson system for collisionless, unmagnetised plasma. In these notes, we discuss recent progress on the quasineutral limit in which the Debye length of the plasma tends to zero, an approximation widely assumed in applications. The models formally obtained from Vlasov-Poisson systems in this limit can be seen as kinetic formulations of the Euler equations. However, rigorous results on this limit typically require a structural or strong regularity condition. Here we present recent results for a variant of the Vlasov-Poisson system, modelling ions in a regime of massless electrons. We discuss the quasineutral limit from this system to the kinetic isothermal Euler system, in a setting with rough initial data. Then, we consider the connection between the quasineutral limit and the problem of deriving these models from particle systems. We begin by presenting a recent result on the derivation of the Vlasov-Poisson system with massless electrons from a system of extended charges. Finally, we discuss a combined limit in which the kinetic isothermal Euler system is derived.

Keywords Vlasov-Poisson · Plasma physics · Quasineutral limit · Mean-field limit

M. Griffin-Pickering
Department of Mathematical Sciences, Durham University, Durham, UK
e-mail: megan.k.griffin-pickering@durham.ac.uk

M. Iacobelli (✉)
ETH Zürich, Zürich, Switzerland
e-mail: mikaela.iacobelli@math.ethz.ch

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_9

1 Introduction

Plasma is a state of matter consisting of an ionised gas, formed by the dissociation of a neutral gas under the influence of, for example, high temperatures or a strong magnetic field. Various mathematical models are available to describe plasma, corresponding to different physical regimes (such as typical length and time scales). Here we will focus on systems of Vlasov-Poisson type, which are kinetic equations describing dilute, collisionless, weakly magnetised plasmas.

The charged particles in a plasma typically fall into two distinguished types: electrons and positively charged ions. The respective masses of these two species differ significantly—note that the proton-to-electron mass ratio is of order 10^3 [8]. The result is a separation between the relevant timescales of evolution for the two species. As a consequence, it is a reasonable approximation to model the two species to some extent separately, and moreover the two species require different models.

The best known version of the Vlasov-Poisson system is a kinetic model for the electrons in a plasma, evolving in a background of ions that are assumed to be stationary. This approximation is justified by the aforementioned separation of timescales. For simplicity we leave aside the issue of boundary conditions by discussing the system posed on the d-dimensional flat torus $\mathbb{T}^d$, which reads as follows:

$$(VP) := \begin{cases} \partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0 \,, \\ E = -\nabla_x U \,, \ -\Delta U = \rho_f - 1 \,, \\ f|_{t=0} = f_0 \,, \ \displaystyle\int_{\mathbb{T}^d \times \mathbb{R}^d} f_0(x, v) \, \mathrm{d}x \mathrm{d}v = 1. \end{cases} \tag{1}$$

In these notes, we instead focus on a related model for the ions in a plasma. On the ions' timescale, the electrons are comparatively fast moving. In particular, the electron-electron collision frequency ν_e is much higher than the ion-ion collision frequency ν_i. For example, Bellan [8, Section 1.9] gives a relation of the form $\nu_e \sim (m_e/m_i)^{-1/2}\nu_i$ for plasmas with similar ion and electron temperatures, where m_e and m_i denote the masses of, respectively, a single electron and a single ion. Thus, when the mass ratio m_e/m_i is small, the frequency of electron-electron collisions can be significant even when ion-ion collisions are negligible.

In the *massless electrons* limit, the mass ratio m_e/m_i is assumed to tend to zero, motivated by the fact that it is small in applications. As a consequence, the electron collision frequency tends to infinity. In the formal limiting regime, the electrons are thermalised, instantaneously assuming their equilibrium distribution, which is a Maxwell-Boltzmann law of the form

$$\rho_e \sim e^{q_e \beta_e \Phi} \,,$$

where q_e is the charge of a single electron, β_e is the inverse electron temperature, and Φ is the ambient potential.

Combining the Vlasov-Poisson system (1) with a Maxwell-Boltzmann law for the electron distribution leads to the *Vlasov-Poisson system with massless electrons*, or VPME system. After an appropriate rescaling of physical constants, this reads as follows:

$$(VPME) := \begin{cases} \partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0 \,, \\ E = -\nabla_x U, \ \Delta U = e^U - \rho_f \,, \\ f|_{t=0} = f_0, \ \displaystyle\int_{\mathbb{T}^d \times \mathbb{R}^d} f_0(x, v)\, \mathrm{d}x\, \mathrm{d}v = 1. \end{cases} \tag{2}$$

This model is used in the plasma physics literature to model ion plasma. For a more detailed introduction to the model in a physics context, see Gurevich and Pitaevsky [31]. The VPME system has been used to study the formation of ion-acoustic shocks [52, 58], the development of phase-space vortices behind these shocks [10], and the expansion of plasma into vacuum [53], among other applications.

From a mathematical perspective, the VPME system has been studied less than the electron Vlasov-Poisson system (1). The systems differ through the additional exponential nonlinearity in the elliptic equation for the electrostatic potential in the VPME system. The nonlinearity of this coupling leads to additional difficulties. For example, while the well-posedness theory of the Vlasov-Poisson system is well established (see for example [49, 50, 57, 60]), for the VPME system this theory was developed more recently. The existence of weak solutions was shown in $\mathbb{R}^3$ by Bouchut [12], while global well-posedness was proved recently by the authors in [28].

The massless electrons limit itself is not yet resolved in full generality. Bouchut and Dolbeault [13] considered the problem for a one species model described by the Vlasov-Poisson-Fokker-Planck system. Bardos, Golse, Nguyen and Sentis [7] studied a two-species model represented by a system of coupled kinetic equations. Under the assumption that this system has sufficiently regular solutions, in the massless electron limit they derive the Maxwell-Boltzmann law for the electron distribution, and a limiting system for the ions that is very similar to the VPME system (2), but with a time-dependent electron temperature. We also refer to Herda [44] for the massless electron limit in the case with an external magnetic field.

In these notes, we summarise some recent progress on two problems related to the VPME system. In Sect. 2, we consider the quasineutral limit, in which a characteristic parameter of the plasma known as the Debye length tends to zero. The limit of the VPME system in this regime is a singular Vlasov equation known as the kinetic isothermal Euler system. In Sect. 3 we consider the derivation of the VPME and kinetic isothermal Euler systems from a particle system. The underlying microscopic system consists of 'ions', here represented as extended charges, interacting with each other and a background of thermalised electrons.

2 Quasineutrality

2.1 The Debye Length

Plasmas have several important characteristic scales, one of which is the *Debye (screening) length*, λ_D. The Debye length has a key role in describing the physics of plasmas: broadly speaking, it governs the scale of electrostatic phenomena in the plasma. For example, it characterises charge separation within the plasma, describing the scale at which it can be observed that the plasma contains areas with a net positive or negative charge, and so is not microscopically neutral.

In terms of the physical constants of the plasma, the electron Debye length λ_D is defined by

$$\lambda_D := \left(\frac{\epsilon_0 k_B T_e}{n_e q_e^2} \right)^{1/2} . \tag{3}$$

In the above formula, ϵ_0 denotes the vacuum permittivity, k_B is the Boltzmann constant, T_e is the electron temperature and n_e is the electron density. The ions similarly have an associated Debye length, which may differ from the electron Debye length. It is defined by the formula (3), replacing the electron density, temperature and charge with the corresponding values for the ions.

Since the Debye length is related to observable quantities such as the density and temperature, it can be found for a real plasma. Typically, λ_D is much smaller than the typical length scale of observation L. The parameter $\varepsilon := \lambda_D / L$ is therefore expected to be small. In this case the plasma is called *quasineutral*: since the scale of charge separation is small, the plasma appears to be neutral at the scale of observation. Quasineutrality is a very common property of real plasmas—for example Chen [20, Section 1.2] includes quasineutrality as one of the key properties distinguishing plasmas from ionised gases more generally.

The significance for Vlasov-Poisson systems becomes apparent after a rescaling. When written in appropriate dimensionless variables, the Vlasov-Poisson systems acquire a scaling of ε^2 in front of the Laplacian in the Poisson equation for the electric field. For example, the VPME system (2) takes the form

$$(VPME)_\varepsilon := \begin{cases} \partial_t f_\varepsilon + v \cdot \nabla_x f_\varepsilon + E \cdot \nabla_v f_\varepsilon = 0 , \\ E = -\nabla_x U , \\ \varepsilon^2 \Delta U = e^U - \rho_{f_\varepsilon} , \\ f_\varepsilon|_{t=0} = f_\varepsilon(0) , \ \displaystyle\int_{\mathbb{T}^d \times \mathbb{R}^d} f_\varepsilon(0, x, v) \, \mathrm{d}x \, \mathrm{d}v = 1. \end{cases} \tag{4}$$

In plasma physics literature, the approximation that $\varepsilon \approx 0$ is widely used. For this reason, it is important to understand what happens to the Vlasov-Poisson system in the limit as ε tends to zero. This is known as the *quasineutral limit*. Taking this limit leads to other models for plasma known as kinetic Euler systems.

2.2 Kinetic Euler Systems

Formally setting $\varepsilon = 0$ in the system (4) results in the *kinetic isothermal Euler system* (KIsE):

$$(KIsE) := \begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x U \cdot \nabla_v f = 0 \,, \\ U = \log \rho_f \,, \\ f|_{t=0} = f_0 \,, \quad \displaystyle\int_{\mathbb{T}^d \times \mathbb{R}^d} f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v = 1. \end{cases} \tag{5}$$

This system was described and studied in a physics context in [31–33]. The name arises from the fact that, for monokinetic solutions f, of the form

$$f(t,x,v) = \rho(t,x)\delta_0(v - u(t,x))$$

for some density ρ and velocity field u, the KIsE system is equivalent to the following isothermal Euler system:

$$(IsE) := \begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0 \,, \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u) - \nabla_x \rho = 0. \end{cases} \tag{6}$$

The KIsE system (5) can be thought of as a kinetic formulation of the isothermal Euler system (6) To see this, consider a solution in the form of a superposition of monokinetic profiles: let

$$f(t,x,v) = \int_\Theta \rho_\theta(t,x)\delta_0(v - u_\theta(t,x))\pi(\mathrm{d}\theta) \,, \tag{7}$$

for a measure space (Θ, π) and a family of fluids $(\rho_\theta, u_\theta)_{\theta\in\Theta}$. The multi-fluid representation (7) can be used in the case where f has a density with respect to Lebesgue measure on $\mathbb{T}^d \times \mathbb{R}^d$. However, it can also accommodate more singular situations. For example, if π is a sum of N Dirac masses, then the distribution (7) can be used to describe a system of N phases.

With this multi-fluid representation in mind, consider the following system of PDEs for the unknowns $(\rho_\theta, u_\theta)_{\theta\in\Theta}$:

$$(KIsE)_{MF} := \begin{cases} \partial_t \rho_\theta + \nabla_x \cdot (\rho_\theta u_\theta) = 0 \,, \\ \partial_t (\rho_\theta u_\theta) + \nabla_x \cdot (\rho_\theta u_\theta \otimes u_\theta) = -\rho_\theta \nabla_x U \,, \\ U = \log \displaystyle\int_\Theta \rho_\theta(t,x)\pi(\mathrm{d}\theta). \end{cases} \tag{8}$$

Given a (distributional) solution of this multi-fluid system, the formula (7) then defines a distributional solution of the KIsE system (5). Thus (8) is a multi-fluid

formulation of KIsE (5) and KIsE is a kinetic formulation of the isothermal Euler system (6). The use of multi-fluid representations of this type for Vlasov-type equations is discussed, for example, in [16, 26, 61].

A system closely related to the KIsE system can be formally obtained by linearising the coupling $U = \log \rho_f$ between U and ρ_f around the constant density 1: since $\log t \approx t - 1$ for t close to one, one gets

$$(VDB) := \begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x U \cdot \nabla_v f = 0 \,, \\ U = \rho_f - 1 \,, \\ f|_{t=0} = f_0 \,, \ \int_{\mathbb{T}^d \times \mathbb{R}^d} f_0(x, v) \, \mathrm{d}x \, \mathrm{d}v = 1. \end{cases} \tag{9}$$

This system was named the *Vlasov-Dirac-Benney* (VDB) system by Bardos [2]. The name 'Benney' was chosen due to a connection with the Benney equations for water waves, in particular as formulated by Zakharov [61].

The VDB system formally has the structure of a general Vlasov equation, in which the potential U is of the form $U = \Phi *_x (\rho_f - 1)$ for some kernel Φ. In this case, the kernel would be a Dirac mass; this is the origin of the reference to Dirac. In particular, this demonstrates the additional singularity of the VDB system in comparison to the Vlasov-Poisson system: in the Vlasov-Poisson system the potential U gains two derivatives compared to the density ρ_f, while in the VDB system this regularisation does not occur.

For the Vlasov-Poisson system for electrons (1), the quasineutral limit leads to the following *kinetic incompressible Euler* system (KInE):

$$(KInE) := \begin{cases} \partial_t f + v \cdot \nabla_x f - \nabla_x U \cdot \nabla_v f = 0 \,, \\ \rho_f = 1 \,, \\ f|_{t=0} = f_0 \,, \ \int_{\mathbb{T}^d \times \mathbb{R}^d} f_0(x, v) \, \mathrm{d}x \, \mathrm{d}v = 1. \end{cases} \tag{10}$$

The force $-\nabla_x U$ is defined implicitly through the incompressibility constraint $\rho_f = 1$, and may be thought of as a Lagrange multiplier associated to this constraint. The system (10) was discussed by Brenier in [15] as a kinetic formulation of the incompressible Euler equations.

All three kinetic Euler systems described above (8), (9), and (10) as well as the two Vlasov-Poisson systems (1), (2), have a large family of stationary solutions: the spatially homogeneous profiles $f(t, x, v) = \mu(v)$. As is well-known for the Vlasov-Poisson system, some of these profiles may be unstable [56]. For the kinetic Euler systems, the corresponding linearised problems have unbounded unstable spectrum: see [3, 6, 39]. As a consequence, they are in general ill-posed. For example, ill-posedness in Sobolev spaces was shown for the VDB system by Bardos and Nouri [6]. Han-Kwan and Nguyen [39] further extended this by showing that the solution map cannot be Hölder continuous with respect to the initial datum in Sobolev spaces, for both the VDB system (9) and the KInE system (10). See also Baradat [1] for the generalisation when the unstable profile μ is only a measure.

Due to these instability properties, well-posedness results for the kinetic Euler systems typically involve either a strong regularity restriction or a structural condition. For instance, in the monokinetic case one may appeal to the results known for the corresponding Euler system.

Without imposing any structural condition, the most general results available are in analytic regularity. Local existence of analytic solutions for the VDB system was proven by Jabin and Nouri [46] in the one-dimensional case, and also follows from [54, Section 9]. Bossy, Fontbona, Jabin and Jabir [11] proved an analogous result for a class of kinetic equations involving an incompressibility constraint, generalising the KInE system (10) to include, for example, noise terms. Local existence of analytic solutions for the multi-fluid system corresponding to KInE (10) was shown by Grenier [26] as part of a study of the quasineutral limit; note that, due to the multi-fluid formulation, the required regularity is only imposed in the x variable.

In Sobolev regularity, local well-posedness is known for the VDB system for initial data satisfying a Penrose-style stability criterion, following the results of Bardos and Besse [3] and Han-Kwan and Rousset [40]. We do not know of any global-in-time existence results for any of the kinetic Euler systems (5), (9) or (10).

The VDB system also appears in the semiclassical limit of an infinite dimensional system of coupled nonlinear Schrödinger equations: for more details, see for example [3–5]. See also [19, 22] for discussion of semiclassical limits involving the KIsE model.

2.3 Failure of the Quasineutral Limit

The mathematical justification of the quasineutral limit is a non-trivial problem, since in general the limit can be false. The failure of the limit can be linked to known phenomena in plasma physics. We note for instance the example of Medvedev [53] regarding the expansion of ion plasma into vacuum. For a one-dimensional hydrodynamic model it is found that the quasineutral approximation $U = \log \rho$ is not valid everywhere, and this is corroborated by numerical simulations for a kinetic model.

Another important issue, well-known in plasma physics, is the 'two stream' instability. From a physics perspective, this instability is typically introduced through a model problem in which two jets of electrons are fired towards each other (whence the name). Configurations of this kind are known to be unstable (see for example [8, Section 5.1], [20, Section 6.6]), with the resulting dynamics producing a vortex-like behaviour in phase space. See [9] for simulations and experimental results on this phenomenon. The streaming instability is seen in kinetic models by considering profiles with a 'double bump' structure in the velocity variable. These profiles are unstable for the linearised problem in the Penrose sense discussed above.

The relevance of instability for the quasineutral limit can be indicated by looking at a time rescaling of the Vlasov-Poisson system. If f is a solution of the unscaled

Vlasov-Poisson system (1), then $f_\varepsilon(t, x, v) = f\left(\frac{t}{\varepsilon}, \frac{x}{\varepsilon}, v\right)$ is a solution of the system with quasineutral scaling. The limit as ε tends to zero is thus a form of long time limit. Grenier outlined this obstruction to the quasineutral limit in [26, 27], for a one-dimensional two-stream configuration. Subsequently, Han-Kwan and Hauray [35] constructed counterexamples to the quasineutral limit in the Sobolev spaces H^s for arbitrary large s, by considering initial data around unstable profiles.

2.4 Results on the Quasineutral Limit

Positive results on the quasineutral limit can be categorised along the lines of the well-posedness results known for the kinetic Euler systems; these problems are closely related. The mathematical study of the quasineutral limit can be traced back to the 1990s, with the works of Brenier and Grenier [18] and Grenier [25], using an approach based on defect measures, and the result of Grenier [27] for the one-dimensional case.

A particular case is the 'cold electrons' or 'cold ions' regime, in which the initial data for the Vlasov-Poisson system is assumed to converge to a monokinetic profile. The limiting kinetic Euler system is therefore reduced to its corresponding Euler system. Brenier [17] and Masmoudi [51] considered the electron case, from the Vlasov-Poisson system to the incompressible Euler equations. Han-Kwan [34] considered the ions case, from the VPME system to the isothermal Euler equations. See also the work of Golse and Saint-Raymond [24], obtaining a '2.5 dimensional' incompressible Euler system through a combined quasineutral and gyrokinetic limit (a limit of strong magnetic field).

In [26], Grenier proved the quasineutral limit from the electron Vlasov-Poisson system to KInE in analytic regularity. The result is framed in terms of the corresponding multi-fluid formulations. If the initial data for the multi-fluid Vlasov-Poisson system are uniformly analytic in x, then the quasineutral limit to the multi-fluid KInE system holds locally in time. By the same techniques, similar results can be shown for the ion quasineutral limits, obtaining the VDB and KIsE systems, as observed in [37], in the discussion after Proposition 4.1.

Under a Penrose-type stability criterion, Han-Kwan and Rousset [40] proved that the quasineutral limit holds in Sobolev regularity, for the passage from a variant of the VPME system, with linearised Poisson-Boltzmann coupling for the electric field, to the VDB system.

2.5 Quasineutral Limit with Rough Data

An alternative direction for relaxing the regularity constraint for the quasineutral limit was investigated in a series of works, by Han-Kwan and the second author [36, 37] and by the authors [30]. In this setting, one considers rough initial

data (measures in the one-dimensional case, L^∞ for $d = 2, 3$) that are small perturbations of the uniformly analytic case. The smallness of the perturbation is measured in a Wasserstein (Monge-Kantorovich) distance.

Definition 1 (Wasserstein Distances) Let $p \in [1, \infty)$. Let μ and ν be probability measures on $\mathbb{T}^d \times \mathbb{R}^d$ for which the moment of order p is finite. Then the pth order Wasserstein distance between μ and ν, $W_p(\mu, \nu)$, is defined by

$$W_p(\mu, \nu) = \left(\inf \int_{(z_1, z_2) \in (\mathbb{T}^d \times \mathbb{R}^d)^2} d(z_1, z_2)^p \, \mathrm{d}\pi(z_1, z_2) \right)^{1/p},$$

with the infimum taken over measures π on $(\mathbb{T}^d \times \mathbb{R}^d)^2$ such that for all Borel sets $A \subset \mathbb{T}^d \times \mathbb{R}^d$,

$$\pi(A \times \mathbb{T}^d \times \mathbb{R}^d) = \mu(A) \,, \qquad \pi(\mathbb{T}^d \times \mathbb{R}^d \times A) = \nu(A) \,,$$

and d denotes the standard metric on $\mathbb{T}^d \times \mathbb{R}^d$.

The article [37] deals with the one-dimensional case for both electron and ion models, while in higher dimensions $d = 2, 3$, the limit for the electron models is considered in [36]. Then, for the VPME system, we proved a rough data quasineutral limit in [30].

Below we give the statement of this result. We use the notation $\overline{\exp}_{\mathbf{n}}$ to denote the n-fold iteration of the exponential function, for example

$$\overline{\exp}_{\mathbf{3}}(x) = \exp \exp \exp(x).$$

We also use the analytic norms $\|\cdot\|_{B_\delta}$, defined for $\delta > 1$ by

$$\|g\|_{B_\delta} := \sum_{k \in \mathbb{Z}^d} |\hat{g}(k)| \delta^{|k|} \,,$$

where $\hat{g}(k)$ denotes the Fourier coefficient of g of index k.

Theorem 1 (Quasineutral Limit) *Let $d = 2, 3$. Consider initial data $f_\varepsilon(0)$ satisfying the following conditions:*

- *(Uniform bounds) $f_\varepsilon(0)$ is bounded and has bounded energy, uniformly with respect to ε: for some constant $C_0 > 0$,*

$$\|f_\varepsilon(0)\|_{L^\infty(\mathbb{T}^d \times \mathbb{R}^d)} \leq C_0 \,,$$
$$\frac{1}{2} \int_{\mathbb{T}^d \times \mathbb{R}^d} |v|^2 f \, \mathrm{d}x \, \mathrm{d}v + \frac{\varepsilon^2}{2} \int_{\mathbb{T}^d} |\nabla U|^2 \, \mathrm{d}x + \int_{\mathbb{T}^d} U e^U \, \mathrm{d}x \leq C_0.$$

- *(Control of support) There exists $C_1 > 0$ such that*

$$f_\varepsilon(0, x, v) = 0 \quad \text{for } |v| > \exp(C_1 \varepsilon^{-2}). \tag{11}$$

- *(Perturbation of an analytic function) There exist $g_\varepsilon(0)$ satisfying, for some $\delta > 1$, $\eta > 0$, and $C > 0$,*

$$\begin{aligned} &\sup_{\varepsilon>0} \sup_{v\in\mathbb{R}^d} (1 + |v|^{d+1}) \|g_\varepsilon(0, \cdot, v)\|_{B_\delta} \leq C \ , \\ &\sup_{\varepsilon>0} \left\| \int_{\mathbb{R}^d} g_\varepsilon(0, \cdot, v) \, \mathrm{d}v - 1 \right\|_{B_\delta} \leq \eta \ , \end{aligned} \tag{12}$$

  as well as the support condition (11), such that, for all $\varepsilon > 0$,

$$W_2(f_\varepsilon(0), g_\varepsilon(0)) \leq \left[\overline{\exp}_{\mathbf{4}}(C\varepsilon^{-2}) \right]^{-1} \tag{13}$$

  for C sufficiently large with respect to C_0, C_1.
- *(Convergence of data) $g_\varepsilon(0)$ has a limit $g(0)$ in the sense of distributions as $\varepsilon \to 0$.*

Let f_ε denote the unique solution of (4) with bounded density and initial datum $f_\varepsilon(0)$. Then there exists a time horizon $T_* > 0$, independent of ε but depending on the collection $\{g_{0,\varepsilon}\}_\varepsilon$, and a solution g of (5) on the time interval $[0, T_*]$ with initial datum $g(0)$, such that

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T_*]} W_1(f_\varepsilon(t), g(t)) = 0.$$

Remark 1 As an example of a choice of initial data satisfying these assumptions, consider any compactly supported, spatially homogeneous profile $\mu = \mu(v) \geq 0$ with unit mass. Then

$$f_\varepsilon(0) = \mu(v) \left(1 + \sin(2\pi N_\varepsilon x_1)\right), \quad N_\varepsilon \gtrsim \overline{\exp}_{\mathbf{4}}(C\varepsilon^{-2})$$

satisfies the assumptions of Theorem 1.

2.6 Remarks on the Strategy

The strategy of proof for the rough data quasineutral limits [30, 36, 37] is based on stability results for the Vlasov-Poisson systems in Wasserstein distances. Stability results of this type have been known for Vlasov-type equations since the work of Dobrushin [21] for the case of Lipschitz force kernels.

The Vlasov-Poisson case was considered by Loeper [50], for solutions whose mass density ρ_f is bounded in L^∞. This is an estimate of the form

$$W_2(f_1(t), f_2(t)) \leq \mathcal{F}\left[W_2(f_1(0), f_2(0)), \max_{i=1,2} \|\rho_{f_i}\|_{L^\infty([0,t]\times\mathbb{T}^d)}\right],$$

for some suitable $\mathcal{F}$. The corresponding estimate for the VPME system was proved recently in [28].

The proof of Theorem 1 relies on a quantification of the W_2 stability estimate in terms of ε. This has two steps: first, the stability estimate itself is quantified, in the sense that

$$W_2(f_\varepsilon^{(1)}(t), f_\varepsilon^{(2)}(t)) \leq \mathcal{F}_\varepsilon\left[W_2(f_\varepsilon^{(1)}(0), f_\varepsilon^{(2)}(0)), \max_{i=1,2} \|\rho_{f_\varepsilon^{(i)}}\|_{L^\infty([0,t]\times\mathbb{T}^d)}\right].$$

Then, a bound is proved for the mass density $\|\rho_{f_\varepsilon^{(i)}}\|_{L^\infty([0,t]\times\mathbb{T}^d)}$ in terms of the initial data. This is achieved by controlling the rate of growth of the support of a solution f_ε in terms of the initial data, via an analysis of the characteristic trajectories of the system. This is the reason for the compact support assumption in Theorem 1.

The quantified stability estimate is then used to make a perturbation around the analytic regime. More specifically, we consider the analytic functions $g_\varepsilon(0)$ defined in the statement as initial data for the VPME system (2). The assumptions (12) are chosen precisely so that the resulting solutions g_ε satisfy the quasineutral limit: on some time interval $[0, T_*]$, as ε tends to zero, g_ε converges to a solution g of the KIsE system (5). This follows from the techniques of Grenier [26], and implies convergence in a Wasserstein distance.

The proof is concluded by the triangle inequality:

$$W_1(f_\varepsilon(t), g(t)) \leq W_1(f_\varepsilon(t), g_\varepsilon(t)) + W_1(g_\varepsilon(t), g(t)) ,$$

choosing the envelope of initial data (13) so that the perturbation term $W_1(f_\varepsilon(t), g_\varepsilon(t))$ vanishes in the limit.

3 Derivations from Particle Systems

It is a fundamental problem to derive effective equations, such as Vlasov-Poisson systems, from the physical systems they are intended to describe. In a reasonably general setting, we may consider a system of N point particles with binary interactions. The dynamics of such a system are modelled in classical mechanics by a system of ODEs of the following form, describing the phase space positions

$(X_i, V_i)_{i=1}^N$ of the particles:

$$\begin{cases} \dot{X}_i = V_i \ , \\ \dot{V}_i = \alpha(N) \sum_{j \neq i} \nabla W(X_i - X_j) + \nabla V(X_i). \end{cases} \tag{14}$$

In this setting ∇W denotes the interaction force between pairs of particles, which here depends only on the spatial separation of the particles and is derived from an interaction potential W. We also include an external force ∇V. The parameter $\alpha(N)$ rescales the system with N and can be thought of as a rescaling of the physical constants of the system. The choice of $\alpha(N)$ determines the model that is obtained as N tends to infinity.

The case $\alpha(N) = 1/N$ is known as the *mean field limit*. The formal limiting system is the Vlasov-type equation

$$\partial_t f + v \cdot \nabla_x f + (\nabla W *_x \rho_f + \nabla V) \cdot \nabla_v f = 0 \ , \tag{15}$$

in the sense that the empirical measures μ^N defined by the formula

$$\mu^N := \frac{1}{N} \sum_{i=1}^N \delta_{(X_i, V_i)}$$

are expected to converge to a solution of the Vlasov equation (15) in the limit as N tends to infinity. The Vlasov-Poisson system fits into this framework by choosing $\nabla V = 0$ and ∇W to be the Coulomb kernel K on the torus $\mathbb{T}^d$. This is the function $K = -\nabla G$, where G satisfies

$$-\Delta G = \delta_0 - 1 \quad \text{on } \mathbb{T}^d.$$

The corresponding microscopic system (14) then describes a system of interacting electrons modelled as point charges, while (15) is the Vlasov-Poisson system (1).

To derive the VPME system, a natural choice for the underlying microscopic system is to consider the dynamics of N ions, modelled as point charges, in a background of thermalised electrons. On the torus, this is modelled by an ODE system of the form

$$\begin{cases} \dot{X}_i = V_i \ , \\ \dot{V}_i = \dfrac{1}{N} \displaystyle\sum_{j \neq i}^N K(X_i - X_j) - K * e^U \ , \end{cases}$$

where the electrostatic potential U satisfies

$$\Delta U = e^U - \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}.$$

We can think of this system as being of the form (14) by taking $\nabla W = K$ and an 'external' force $\nabla V = K * e^U$, even though ∇V is not truly external due to its nonlinear dependence on the particle configuration through U. In this way it can be seen that the VPME system formally describes the limit as N tends to infinity.

Other choices are possible for $\alpha(N)$, in which case the limit as N tends to infinity may produce models of other forms. This approach can be used to derive the kinetic Euler systems discussed above in Sect. 2.2. In the papers [29, 30], the scaling $\alpha(N) \approx \frac{1}{N \log\log N}$ is used to derive the kinetic Euler systems (10) and (5). The method is based on passing via the associated Vlasov-Poisson system, and this limit can thus be thought of as a simultaneous mean field and quasineutral limit. In the recent paper [38], a similar limit is proved in the monokinetic regime, to derive the incompressible Euler equations.

3.1 Mean Field Limits

For a detailed survey of mathematical results on the mean field limit, see [23, 45]. For our purposes we emphasise that the theory of mean field limits depends on the regularity of the interaction force ∇W chosen in the system (15).

Early contributions on the problem include the works of Braun–Hepp [14], Neunzert–Wick [55] and Dobrushin [21]. In particular, the limit holds in the case where the forces are Lipschitz: $\nabla W,\ \nabla V \in W^{1,\infty}$.

However, the Vlasov-Poisson system is not included in this setting, due to the singularity of the Coulomb kernel. Identifying the torus $\mathbb{T}^d$ with $\left[-\frac{1}{2}, \frac{1}{2}\right]^d$, with appropriate identifications of the boundary, we note the following properties of the Coulomb kernel K. $K \in C^\infty(\mathbb{T}^d \setminus \{0\})$ is smooth function apart from a point singularity at the origin. In a neighbourhood of the origin, K can be written in the form

$$K(x) = C_d \frac{x}{|x|^d} + K_0(x)\,, \qquad K_0 \in C^\infty.$$

The kernel therefore has a strong singularity of the form $K \sim |x|^{-(d-1)}$.

Forces with a point singularity are of interest in physical applications, since this class includes inverse power laws. From here on, we discuss forces satisfying

bounds of the following form: for some $\beta \in (0, d-1]$,

$$\frac{|\nabla W(x)|}{|x|^\beta} \leq C\,, \quad \frac{|\nabla^2 W(x)|}{|x|^{\beta+1}} \leq C \quad \text{for all } x \in \mathbb{R}^d \setminus \{0\}. \tag{16}$$

Note that the Vlasov-Poisson case corresponds to $\beta = d-1$.

Several works have studied the mean field limit problem for singular forces of the form (16) by considering a regularisation of the limit. The singular force ∇W is replaced by a smooth approximation ∇W_r such that $\lim_{r\to 0} \nabla W_r = \nabla W$. Then, the limits as N tends to infinity and as r tends to zero are taken simultaneously. In this way, one derives the Vlasov equation with singular force in the limit from a sequence of regularised particle systems. In this formulation, the goal is to optimise the regime $r = r(N)$ for which this limit is valid. That is, r should be as small as possible, so that the regularised particle systems are close to the original particle system with singular interaction.

Hauray and Jabin [43] considered the case $\beta < d-1$. The force is regularised by truncation at a certain distance from the singularity. In this case the regularisation parameter $r(N)$ represents the order of this truncation distance. If $r(N)$ tends to zero sufficiently slowly as N tends to infinity, they prove that the regularised mean field limit holds for a large set of initial configurations. For 'weakly singular' forces with $\beta < 1$, in [42, 43] they also prove the mean field limit without truncation.

For Coulomb interactions, the results available depend on the dimension of the problem. In one dimension, the interaction force is less singular. As a consequence, the mean field limit holds, as proved by Hauray [41]. The corresponding result for the VPME system was proved by Han-Kwan and the second author in [37].

In higher dimensions, the Coulomb force is of the form (16). It has a strong singularity corresponding to the endpoint case $\beta = d-1$ not covered by the results of Hauray and Jabin [43]. Regularised approaches were considered by Lazarovici [47] and Lazarovici and Pickl [48]. By a truncation method, Lazarovici and Pickl prove a regularised mean field limit for the Vlasov-Poisson system, for a truncation radius of order $r(N) \sim N^{-1/d+\eta}$ for any $\eta > 0$. To put this in context, note that $N^{-1/d}$ is the order of separation of particles in x if their spatial distribution is close to uniform.

In a recent breakthrough [59], Serfaty introduced a modulated energy method to prove the validity of the mean-field limit for systems of points evolving along the gradient flow of their interaction energy when the interaction is the Coulomb potential or a super-coulombic Riesz potential, in arbitrary dimension. In the appendix (in collaboration with Duerinckx), they adapt this method to prove the mean-field convergence of the solutions to Newton's law with Coulomb interaction in the monokinetic case to solutions of an Euler-Poisson type system.

For the VPME system, a regularised mean field limit was considered by the authors in [30]. The regularisation used is a regularisation by convolution, similar to the setting of Lazarovici [47] that we describe below in Sect. 3.1.1. With this regularisation, the resulting microscopic system represents a system of interacting extended charges, where the parameter r gives the order of the radius of the charges.

Lazarovici [47] derived the Vlasov-Poisson system from a system of extended electrons for $r(N) \geq CN^{-\frac{1}{d(d+2)}+\eta}$ for some $\eta > 0$. In [30], the authors proved a similar derivation for the VPME system from a system of extended ions, for the same range of r. We present this result below in Sect. 3.1.1. To our knowledge, this is the first derivation of the VPME system from a particle system in three dimensions.

3.1.1 Mean Field Limits for VPME

For the VPME system, the mean field limit was proved in the one-dimensional setting in [37]. In the article [30], we considered the problem in higher dimensions $d = 2, 3$, deriving the VPME system from a particle system. The microscopic system is regularised with the regularisation used by Lazarovici [47] for the Vlasov-Poisson system. It consists of a system of 'extended ions': instead of representing the ions as point charges, we consider charges of shape χ for some non-negative, radially symmetric function $\chi \in C_c^\infty(\mathbb{R}^d)$ with unit mass (see Fig. 1). The charges are rescaled as follows: for $r > 0$, let

$$\chi_r(x) := r^{-d}\chi\left(\frac{x}{r}\right).$$

The extended ions interact with a background of thermalised electrons, leading to the following system of ODEs:

$$\begin{cases} \dot{X}_i = V_i \ , \\ \dot{V}_i = -\chi_r * \nabla_x U_r(X_i) \ , \\ \Delta U_r = e^{U_r} - \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \chi_r(X_i). \end{cases} \tag{17}$$

We are able to derive the VPME system (2) from this regularised system, under a condition on the initial data that is satisfied with high probability for $r(N) \geq CN^{-\frac{1}{d(d+2)}+\eta}$. This matches the rate found in Lazarovici's result for the Vlasov-Poisson system.

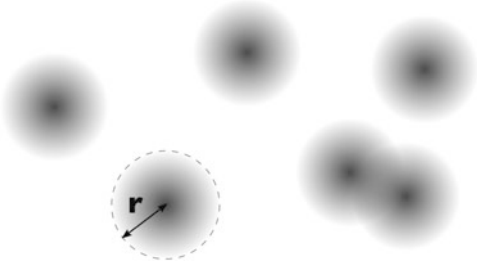

Fig. 1 A system of extended charges. Here χ is supported in the unit ball and thus r represents the radius of each charge

Theorem 2 (Regularised Mean Field Limit) *Let $d = 2, 3$, and let $f_0 \in L^1 \cap L^\infty(\mathbb{T}^d \times \mathbb{R}^d)$ be compactly supported. Let f denote the unique bounded density solution of the VPME system (2) with initial datum f_0. Fix $T_* > 0$.*

Assume that $r = r(N)$ and the initial configurations for (17) are chosen such that the corresponding empirical measures satisfy, for some sufficiently large constant $C > 0$, depending on T_* and the support of f_0,

$$\limsup_{N\to\infty} \frac{W_2^2(f_0, \mu_r^N(0))}{r^{d+2+C|\log r|^{-1/2}}} < 1.$$

Then the empirical measure μ_r^N associated to the particle system dynamics starting from this configuration converges to f:

$$\lim_{N\to\infty} \sup_{t\in[0,T_*]} W_2(f(t), \mu_r^N(t)) = 0. \tag{18}$$

In particular, choose $r(N) = N^{-\gamma}$ for some $\gamma < \frac{1}{d(d+2)}$. For each N, let the initial configurations for the regularised N-particle system (17) be chosen by taking N independent samples from f_0. Then (18) holds with probability one.

This theorem is proved by introducing a regularised version of the VPME system:

$$\begin{cases} \partial_t f_r + v \cdot \nabla_x f_r + E_r \cdot \nabla_v f_r = 0\,, \\ E = -\chi_r *_x \nabla_x U,\ \Delta U = e^U - \chi_r *_x \rho_f\,, \\ f_r|_{t=0} = f_0\,,\ \displaystyle\int_{\mathbb{T}^d\times\mathbb{R}^d} f_0(x,v)\,\mathrm{d}x\,\mathrm{d}v = 1. \end{cases} \tag{19}$$

The solution f_r of this system is used as an intermediate step between the particle system and the VPME system, as illustrated in Fig. 2.

The proof proceeds as follows:

- We estimate the discrepancy between μ_r^N and f_r, and that between f_r and f, in a Wasserstein distance. This uses similar techniques to the stability estimate discussed in Sect. 2.6.
- This estimate is carefully quantified and the regularisation parameter r is allowed to depend on N. This allows us to identify a relationship between r and N such that μ_r^N converges to f for almost all initial data drawn as N independent samples from f_0.

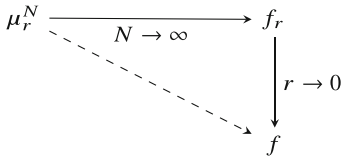

Fig. 2 Strategy for the proof of Theorem 2

3.2 Derivation of Kinetic Euler Systems

The kinetic Euler systems (5) and (10) can be derived from particle systems, by using a modified scaling instead of the mean field scaling. In the articles [29, 30] we consider an approach based on a combined mean field and quasineutral limit. In terms of the scaling $\alpha(N)$, this means that we write $\alpha = (N\varepsilon^2)^{-1}$, and then consider allowing ε to depend on N. We then seek a rate of decay of $\varepsilon(N)$ to zero as N tends to infinity for which it possible to take the mean field and quasineutral limits simultaneously.

Due to the challenges involved in the mean field limit for Vlasov-Poisson system, as discussed above, we again use the extended charges model. For the KIsE system we therefore work with the following microscopic system:

$$\begin{cases} \dot{X}_i = V_i \ , \\ \dot{V}_i = -\chi_r * \nabla_x U(X_i) \ , \\ \varepsilon^2 \Delta U = e^U - \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \chi_r(x - X_i). \end{cases} \tag{20}$$

In [30], we prove the following result.

Theorem 3 (From Extended Ions to Kinetic Isothermal Euler) *Let $d = 2$ or 3, and let $f_\varepsilon(0)$, $g_\varepsilon(0)$ and $g(0)$ satisfy the assumptions of Theorem 1. Let $T_* > 0$ be the maximal time of convergence from Theorem 1 and let g denote the solution of the KIsE system (5) with initial data $g(0)$ on the time interval $[0, T_*]$ appearing in the conclusion of Theorem 1.*

Let $r = r(N)$ be of the form

$$r(N) = cN^{-\frac{1}{d(d+2)}+\eta}, \quad \text{for some } \eta > 0\,,\ c > 0.$$

There exists a constant C, depending on d, η, c and $\{f_\varepsilon(0)\}_\varepsilon$, such that the following holds.

Let $\varepsilon = \varepsilon(N)$ satisfy

$$\varepsilon(N) \geq \frac{C}{\sqrt{\log\log\log N}}\,, \qquad \lim_{N\to\infty} \varepsilon(N) = 0.$$

For each N, let the initial conditions for the regularised and scaled N-particle ODE system (20) be chosen randomly with law $f_{\varepsilon(N)}(0)^{\otimes N}$. Let $\mu^N_{\varepsilon,r}(t)$ denote the empirical measure associated to the solution of (20).

Then, with probability one,

$$\lim_{N\to\infty} \sup_{t\in[0,T_*]} W_1\left(\mu^N_{\varepsilon,r}(t), g(t)\right) = 0.$$

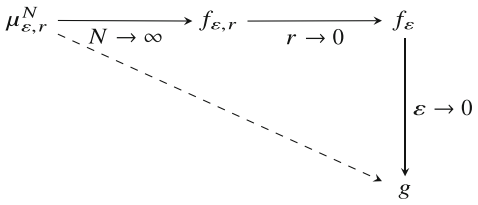

Fig. 3 Strategy for the proof of Theorem 3

This theorem is proved using the strategy illustrated in Fig. 3. Here $f_{\varepsilon,r}$ denotes the solution of a version of the regularised VPME system (19) with quasineutral scaling.

The proof proceeds as follows:

- As in the proof of Theorem 2, we estimate the Wasserstein distance between $\mu^N_{\varepsilon,r}$ and $f_{\varepsilon,r}$ and between $f_{\varepsilon,r}$ and f_ε.
- We carefully quantify these estimates in terms of all three parameters N, r and here also ε.
- For the convergence of f_ε to g, we appeal to Theorem 1.
- Using this, we are able to identify a dependence $r = r(N)$ and $\varepsilon = \varepsilon(N)$ of the parameters on the number of particles, and a relation between r and ε, so that the convergence from the particle system to the KIsE system holds for almost all initial data drawn as independent samples from $f_\varepsilon(0)$.

References

1. Baradat, A.: Nonlinear instability in Vlasov type equations around rough velocity profiles. Annales de l'Institut Henri Poincaré C, Analyse non linéaire **37**(3), 489–547 (2020)
2. Bardos, C.: About a Variant of the 1d Vlasov equation, dubbed "Vlasov-Dirac-Benney equation". In: Séminaire Laurent Schwartz—Équations aux dérivées partielles et applications. Année 2012–2013., Sémin. Équ. Dériv. Partielles, pp. 1–21. École Polytechnique, Centre de Mathématiques, Palaiseau (2014)
3. Bardos, C., Besse, N.: The Cauchy problem for the Vlasov-Dirac-Benney equation and related issues in fluid mechanics and semi-classical limits. Kinet. Relat. Models **6**(4), 893–917 (2013)
4. Bardos, C., Besse, N.: Hamiltonian structure, fluid representation and stability for the Vlasov-Dirac-Benney equation. In: Hamiltonian Partial Differential Equations and Applications. *Fields Institute Communications*, vol. 75, pp. 1–30. Fields Institute Research Mathematical Science, Toronto (2015)
5. Bardos, C., Besse, N.: Semi-classical limit of an infinite dimensional system of nonlinear Schrödinger equations. Bull. Inst. Math. Acad. Sin. (N.S.) **11**(1), 43–61 (2016)
6. Bardos, C., Nouri, A.: A Vlasov equation with Dirac potential used in fusion plasmas. J. Math. Phys. **53**(11), 115621 (2012)
7. Bardos, C., Golse, F., Nguyen, T.T., Sentis, R.: The Maxwell-Boltzmann approximation for ion kinetic modeling. Phys. D **376/377**, 94–107 (2018)
8. Bellan, P.M.: Fundamentals of Plasma Physics. Cambridge University, Cambridge (2008)
9. Berk, H.L., Nielsen, C.E., Roberts, K.V.: Phase space hydrodynamics of equivalent nonlinear systems: experimental and computational observations. Phys. Fluids **13**(4), 980–995 (1970)

10. Bonhomme, G., Pierre, T., Leclert, G., Trulsen, J.: Ion phase space vortices in ion beam-plasma systems and their relation with the ion acoustic instability: numerical and experimental results. Plasma Phys. Controlled Fusion **33**(5), 507–520 (1991)
11. Bossy, M., Fontbona, J., Jabin, P.E., Jabir, J.F.: Local existence of analytical solutions to an incompressible Lagrangian stochastic model in a periodic domain. Comm. Partial Differential Equations **38**(7), 1141–1182 (2013)
12. Bouchut, F.: Global weak solution of the Vlasov-Poisson system for small electrons mass. Comm. Partial Differential Equations **16**(8–9), 1337–1365 (1991)
13. Bouchut, F., Dolbeault, J.: On long time asymptotics of the Vlasov-Fokker-Planck equation and of the Vlasov-Poisson-Fokker-Planck system with Coulombic and Newtonian potentials. Differential Integral Equations **8**(3), 487–514 (1995)
14. Braun, W., Hepp, K.: The Vlasov dynamics and its fluctuations in the $1/N$ limit of interacting classical particles. Comm. Math. Phys. **56**(2), 101–113 (1977)
15. Brenier, Y.: Une formulation de type Vlassov–Poisson pour les équations d'Euler des fluides parfaits incompressibles. [Rapport de recherche] RR-1070, INRIA (1989)
16. Brenier, Y.: Minimal geodesics on groups of volume-preserving maps and generalized solutions of the euler equations. Comm. Pure Appl. Math. **52**(4), 411–452 (1999)
17. Brenier, Y.: Convergence of the Vlasov–Poisson system to the incompressible Euler equations. Comm. Partial Differential Equations **25**(3–4), 737–754 (2000)
18. Brenier, Y., Grenier, E.: Limite singulière du système de Vlasov-Poisson dans le régime de quasi neutralité: le cas indépendant du temps. C. R. Acad. Sci. Paris Sér. I Math. **318**(2), 121–124 (1994)
19. Carles, R., Nouri, A.: Monokinetic solutions to a singular Vlasov equation from a semiclassical perspective. Asymptot. Anal. **102**(1–2), 99–117 (2017)
20. Chen, F.F.: Introduction to Plasma Physics and Controlled Fusion, 3rd edn. Springer, New York (2016)
21. Dobrushin, R.L.: Vlasov equations. Funktsional. Anal. i Prilozhen. **13**(2), 48–58 (1979)
22. Ferriere, G.: Convergence rate in Wasserstein distance and semiclassical limit for the defocusing logarithmic Schrödinger equation (2019). Preprint, arXiv:1903.04309
23. Golse, F.: On the dynamics of large particle systems in the mean field limit. In: Muntean, A., Rademacher, J., Zagaris, A. (eds.) Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity. Lecture Notes in Application Mathematical Mechanical, vol. 3, pp. 1–144. Springer, New York (2016)
24. Golse, F., Saint-Raymond, L.: The Vlasov-Poisson system with strong magnetic field in quasineutral regime. Math. Models Methods Appl. Sci. **13**(5), 661–714 (2003)
25. Grenier, E.: Defect measures of the Vlasov-Poisson system in the quasineutral regime. Comm. Partial Differential Equations **20**(7–8), 1189–1215 (1995)
26. Grenier, E.: Oscillations in quasineutral plasmas. Comm. Partial Differential Equations **21**(3–4), 363–394 (1996)
27. Grenier, E.: Limite quasineutre en dimension 1. In: Journées "Équations aux Dérivées Partielles" (Saint-Jean-de-Monts, 1999), pp. Exp. No. II, 8. University of Nantes, Nantes (1999)
28. Griffin-Pickering, M., Iacobelli, M.: Global well-posedness for the Vlasov-Poisson system with massless electrons in the 3-dimensional torus. ArXiv:1810.06928
29. Griffin-Pickering, M., Iacobelli, M.: A mean field approach to the quasi-neutral limit for the Vlasov–Poisson equation. SIAM J. Math. Anal. **50**(5), 5502–5536 (2018)
30. Griffin-Pickering, M., Iacobelli, M.: Singular limits for plasmas with thermalised electrons. J. Math. Pures Appl. **135**, 199–255 (2020)
31. Gurevich, A.V., Pitaevsky, L.P.: Non-linear dynamics of a rarefied ionized gas. Prog. Aerosp. Sci. **16**(3), 227–272 (1975)
32. Gurevich, A., Pariiskaya, L., Pitaevskii, L.: Self-similar motion of rarefied plasma. Soviet Phys. JETP **22**(2), 449–454 (1966)
33. Gurevich, A., Pariiskaya, L., Pitaevskii, L.: Self-similar motion of a low-density plasma II. Soviet Phys. JETP **27**(3), 476–482 (1968)

34. Han-Kwan, D.: Quasineutral limit of the Vlasov–Poisson system with massless electrons. Comm. Partial Differential Equations **36**(8), 1385–1425 (2011)
35. Han-Kwan, D., Hauray, M.: Stability issues in the quasineutral limit of the one-dimensional Vlasov-Poisson equation. Comm. Math. Phys. **334**(2), 1101–1152 (2015)
36. Han-Kwan, D., Iacobelli, M.: Quasineutral limit for Vlasov-Poisson via Wasserstein stability estimates in higher dimension. J. Differential Equations **263**(1), 1–25 (2017)
37. Han-Kwan, D., Iacobelli, M.: The quasineutral limit of the Vlasov-Poisson equation in Wasserstein metric. Commun. Math. Sci. **15**(2), 481–509 (2017)
38. Han-Kwan, D., Iacobelli, M.: From Newton's second law to Euler's equations of perfect fluids (2020). Preprint, arXiv:2006.14924
39. Han-Kwan, D., Nguyen, T.T.: Ill-posedness of the hydrostatic Euler and singular Vlasov equations. Arch. Ration. Mech. Anal. **221**(3), 1317–1344 (2016)
40. Han-Kwan, D., Rousset, F.: Quasineutral limit for Vlasov-Poisson with Penrose stable data. Ann. Sci. Éc. Norm. Supér. (4) **49**(6), 1445–1495 (2016)
41. Hauray, M.: Mean field limit for the one dimensional Vlasov-Poisson equation. In: Séminaire Laurent Schwartz—Équations aux dérivées partielles et applications. Année 2012–2013, Exp. No. XXI, Sémin. Équ. Dériv. Partielles École Polytechnic, Palaiseau (2014)
42. Hauray, M., Jabin, P.E.: N-particles approximation of the Vlasov equations with singular potential. Arch. Ration. Mech. Anal. **183**(3), 489–524 (2007)
43. Hauray, M., Jabin, P.E.: Particle approximation of Vlasov equations with singular forces: propagation of chaos. Ann. Sci. Éc. Norm. Supér. (4) **48**(4), 891–940 (2015)
44. Herda, M.: On massless electron limit for a multispecies kinetic system with external magnetic field. J. Differential Equations **260**(11), 7861–7891 (2016)
45. Jabin, P.E.: A review of the mean field limits for Vlasov equations. Kinet. Relat. Models **7**(4), 661 (2014)
46. Jabin, P., Nouri, A.: Analytic solutions to a strongly nonlinear Vlasov equation. C.R. Acad. Sci. Paris, Sér. 1 **349**, 541–546 (2011)
47. Lazarovici, D.: The Vlasov-Poisson dynamics as the mean field limit of extended charges. Comm. Math. Phys. **347**(1), 271–289 (2016)
48. Lazarovici, D., Pickl, P.: A mean field limit for the Vlasov-Poisson system. Arch. Ration. Mech. Anal. **225**(3), 1201–1231 (2017)
49. Lions, P.L., Perthame, B.: Propagation of moments and regularity for the 3-dimensional Vlasov-Poisson system. Invent. Math. **105**(2), 415–430 (1991)
50. Loeper, G.: Uniqueness of the solution to the Vlasov-Poisson system with bounded density. J. Math. Pures Appl. (9) **86**(1), 68–79 (2006)
51. Masmoudi, N.: From Vlasov–Poisson system to the incompressible Euler system. Comm. Partial Differential Equations **26**(9–10) (2001)
52. Mason, R.J.: Computer simulation of ion-acoustic shocks. The diaphragm problem. Phys. Fluids **14**(9), 1943–1958 (1971)
53. Medvedev, Y.V.: Ion front in an expanding collisionless plasma. Plasma Phys. Controlled Fusion **53**(12), 125007 (2011)
54. Mouhot, C., Villani, C.: On Landau damping. Acta Math. **207**(1), 29–201 (2011)
55. Neunzert, H., Wick, J.: Die Approximation der Lösung von Integro-Differentialgleichungen durch endliche Punktmengen. In: Numerische Behandlung nichtlinearer Integrodifferential- und Differentialgleichungen. Lecture Notes in Mathematical, vol. 395, pp. 275–290. Springer, Berlin (1974)
56. Penrose, O.: Electrostatic Instabilities of a Uniform Non-Maxwellian Plasma. Phys. Fluids **3**(2), 258–265 (1960)
57. Pfaffelmoser, K.: Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data. J. Differential Equations **95**(2), 281–303 (1992)
58. Sakanaka, P., Chu, C., Marshall, T.: Formation of ion-acoustic collisionless shocks. Phys. Fluids **14**(611) (1971)
59. Serfaty, S.: Mean field limit for Coulomb-type flows. Duke Math. J. **169**(15), 2887–2935 (2020). Appendix with M. Duerinckx

60. Ukai, S., Okabe, T.: On classical solutions in the large in time of two-dimensional Vlasov's equation. Osaka J. Math. **15**(2), 245–261 (1978)
61. Zakharov, V.E.: Benney equations and quasiclassical approximation in the inverse problem method. Funktsional. Anal. i Prilozhen. **14**(2), 15–24 (1980)

A Note on Acoustic Limit for the Boltzmann Equation

Juhi Jang and Chanwoo Kim

Abstract We introduce a new Hilbert-type expansion of the Boltzmann equation with the acoustic scaling. By using recent L^p-L^∞ theory of the Boltzmann equation, we show the validity of the acoustic limit in optimal scaling. In particular, our scheme requires only the second order of the expansion with a remainder, and thereby it gives less restrictions on the initial data.

1 Introduction

We study the rescaled Boltzmann equation

$$\mathcal{S}t\partial_t F + v \cdot \nabla_x F = \frac{1}{\mathcal{K}n} Q(F, F) \tag{1}$$

with dimensionless numbers: *Strouhal number* $\mathcal{S}t$ and the *Knudsen number* $\mathcal{K}n$. Here $F = F(t, x, v) \geq 0$ is the distribution function of the gas particles with the time variable $t \in \mathbb{R}_+ := \{t \geq 0\}$, the space variable $x = (x_1, x_2, x_3) \in \Omega = \mathbb{R}^3$ or $\mathbb{T}^3$ (a periodic box), and the velocity variable $v = (v_1, v_2, v_3) \in \mathbb{R}^3$. We consider the hard sphere model for which the corresponding Boltzmann collision operator $Q(\cdot, \cdot)$ takes the form

$$\begin{aligned} Q(F, G) = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} |(v - v_*) \cdot \mathfrak{u}| \{ & F(v')G(v'_*) + G(v')F(v'_*) \\ & - F(v)G(v_*) - G(v)F(v_*) \} \mathrm{d}\mathfrak{u}\mathrm{d}v_*, \end{aligned} \tag{2}$$

where $v' := v - ((v - v_*) \cdot \mathfrak{u})\mathfrak{u}$ and $v'_* := v_* + ((v - v_*) \cdot \mathfrak{u})\mathfrak{u}$.

J. Jang (✉)
Department of Mathematics, University of Southern California, Los Angeles, CA, USA
e-mail: juhijang@usc.edu

C. Kim
Department of Mathematics, University of Wisconsin-Madison, Madison, WI, USA
e-mail: chanwoo.kim@wisc.edu

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_10

The Boltzmann operator Q satisfies so-called the collision invariance

$$\int_{\mathbb{R}^3} Q(F,G)(v)\Big(1, v, \frac{|v|^2-3}{\sqrt{6}}\Big)\mathrm{d}v = (0,0,0) \tag{3}$$

which represents the local conservation laws of mass, momentum and energy.

An equilibrium, satisfying $Q(\cdot,\cdot)=0$, is given by a local Maxwellian associated with the density $R>0$, the macroscopic velocity $U\in\mathbb{R}^3$ and the temperature $T>0$

$$M_{R,U,T}(v) := \frac{R}{(2\pi T)^{\frac{3}{2}}}\exp\left\{-\frac{|v-U|^2}{2T}\right\}. \tag{4}$$

If (R,U,T) are constant in t and x, it is called a global Maxwellian.

In addition to the Strouhal number and Knudsen number we introduce the *Mach number* $\mathcal{M}a$. By passing some or all of $\mathcal{S}t$, $\mathcal{K}n$, and $\mathcal{M}a$ to zero, one may formally derive PDEs of hydrodynamic variables for the fluctuations around the reference state $(1,0,1)$, which are determined as

$$\Big(\sigma(t,x),u(t,x),\theta(t,x)\Big) = \lim_{\mathcal{M}a\downarrow 0}\frac{1}{\mathcal{M}a}\int_{\mathbb{R}^3}\{F(t,x,v)-M_{1,0,1}(v)\}\Big(1,v,\frac{|v|^2-3}{\sqrt{6}}\Big)\mathrm{d}v. \tag{5}$$

In fact, fundamental fluid equations such as Euler and Navier-Stokes equations can be derived as the hydrodynamic limit of the Boltmzann equation with appropriate scalings and there has been a lot of mathematical progress over the decades to justify various fluid equations in both compressible and incompressible regimes [1–19].

In this note, we are interested in the following acoustic scaling:

$$\mathcal{S}t=1,\ \mathcal{K}n=\varepsilon,\ \mathcal{M}a=\delta \quad \text{with } \delta=\delta(\varepsilon)\downarrow 0 \text{ as } \varepsilon\downarrow 0 \tag{6}$$

where multi-scale parameters (ε,δ) appear. Under the scale (6), the hydrodynamic variables (5) in the limit satisfy the acoustic system:

$$\begin{aligned}\partial_t\sigma_A+\nabla_x\cdot u_A&=0,\\ \partial_t u_A+\nabla_x(\sigma_A+\theta_A)&=0,\\ \tfrac{3}{2}\partial_t\theta_A+\nabla_x\cdot u_A&=0.\end{aligned} \tag{7}$$

As far as the rigorous justification of the acoustic system under (6) is concerned, the relative strength of δ with respect to ε turns out to be playing an important role. In particular, $\delta(\varepsilon)=O(\varepsilon^{\frac{1}{2}})$ is a well-known threshold for the acoustic limit in the framework of renormalized solutions [3, 6, 14], while the optimal scaling of

$\delta = \delta(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$ has been validated in the framework of smooth solutions [9] based on the truncated Hilbert expansion.

The goal of this note is to introduce a new Hilbert-type expansion for the Boltzmann equation with (6) in optimal scaling and to provide another proof of the acoustic limit that requires fewer expansions than the ones used in the previous works. We hope that this new multi-scale Hilbert expansion will be useful for other problems.

In the next section, we introduce a Hilbert-type expansion with the multi-scale parameters and discuss the formal derivation of the acoustic equations under (6).

2 Hilbert Expansion and the Result

We take the scaling (6) for the rescaled Boltzmann equation (1) and let

$$\delta \to 0 \quad \text{and} \quad \frac{\varepsilon}{\delta} \to 0 \text{ as } \varepsilon \to 0. \tag{8}$$

The second condition has been added to address the optimal scaling. The acoustic limit is closely related to the compressible Euler limit, as the acoustic system (7) is the linearization of the compressible Euler system around the trivial state $(1, 0, 1)$. Following the strategy of [9], we will make use local Maxwellians induced by the Euler system to derive the acoustic system. To this end, we first recall that a local Maxwellian $M_{R,U,T}$ satisfies

$$\int_{\mathbb{R}^3} \left\{ \partial_t M_{R,U,T} + v \cdot \nabla_x M_{R,U,T} \right\} \left[1 \;\; v - U \;\; |v-U|^2 \right]^T \mathrm{d}v = 0, \tag{9}$$

if and only if (R, U, T) solves the compressible Euler system

$$\begin{aligned} \partial_t R + \nabla_x \cdot (\rho U) &= 0, \\ \partial_t (RU) + \nabla_x \cdot (RU \otimes U) + \nabla_x p &= 0, \\ \partial_t \left[R(e + \tfrac{1}{2}|U|^2) \right] + \nabla_x \cdot \left[RU(e + \tfrac{1}{2}|U|^2) \right] + \nabla_x \cdot (pU) &= 0, \end{aligned} \tag{10}$$

with the equation of state

$$p = RT = \tfrac{3}{2} Re. \tag{11}$$

With the expectation of $(\sigma_A, u_A, \theta_A)$ satisfying (7) being close to $\frac{1}{\delta}(R-1, U, T-1)$, we introduce a local Maxwellian corresponding to $(R, U, T) = (1 + \delta\sigma(t,x), \delta u(t,x), 1 + \delta\theta(t,x))$ solving the Euler equations (10) with initial condition $(\sigma, u, \theta)|_{t=0} = (\sigma_A, u_A, \theta_A)|_{t=0} = (\sigma_A^0, u_A^0, \theta_A^0)$:

$$\mu_\delta := M_{1+\delta\sigma, \delta u, 1+\delta\theta} \tag{12}$$

and consider a Boltzmann solution F_ε of (1) in a form of

$$F_\varepsilon = \mu_\delta + \delta\varepsilon F_{1\delta} + \delta\varepsilon^2 F_{2\delta} + \varepsilon^{\frac{3}{2}} F_{R\varepsilon,\delta}. \tag{13}$$

A linearized operator with μ_δ is given by

$$L_\delta f := -\frac{2}{\sqrt{\mu_\delta}} Q(\mu_\delta, \sqrt{\mu_\delta} f). \tag{14}$$

From (3) the kernel of L_δ, $\mathrm{Ker}L_\delta = \langle \{\varphi_i \sqrt{\mu_\delta}\}_{i=1}^5 \rangle_{L^2_v(\mathbb{R}^3)}$, has five orthonormal basis

$$\varphi_0 := \frac{1}{\sqrt{1+\delta\sigma}}, \quad \varphi_i := \frac{1}{\sqrt{1+\delta\sigma}} \frac{v_i - \delta u_i}{\sqrt{1+\delta\theta}} \ \text{ for } i = 1,2,3, \quad \varphi_4 := \frac{1}{\sqrt{1+\delta\sigma}} \frac{\left| \frac{v-\delta u}{\sqrt{1+\delta\theta}} \right|^2 - 3}{\sqrt{6}}. \tag{15}$$

We denote an L^2_v-projection $\mathbf{P}_\delta$ on $\mathrm{Ker}L_\delta$ such as

$$\mathbf{P}_\delta g := \sum_{j=0}^{4} (P_j g) \varphi_j \sqrt{\mu_\delta}, \quad P_\delta g := (P_0 g, P_1 g, P_2 g, P_3 g, P_4 g), \tag{16}$$

where $P_j g := \int_{\mathbb{R}^3} g \varphi_j \sqrt{\mu_\delta} \mathrm{d}v$ for $j = 0, 1, \cdots, 4$. We remark that for the purpose of notational convenience, we have omitted the dependence on δ of φ_i, P_i, which should read as $\varphi_i = \varphi_{i\delta}$, $P_i = P_{i\delta}$.

Then the equation of $F_{R\varepsilon,\delta}$ with (13) is given by

$$\partial_t F_{R\varepsilon,\delta} + v\cdot\nabla_x F_{R\varepsilon,\delta} - \frac{2}{\varepsilon} Q(\mu_\delta, F_{R\varepsilon,\delta}) - \frac{1}{\varepsilon^{1-\frac{3}{2}}} Q(F_{R\varepsilon,\delta}, F_{R\varepsilon,\delta}) \tag{17}$$

$$= \frac{\delta}{\varepsilon^{\frac{3}{2}}} \left\{ \frac{-\partial_t \mu_\delta - v\cdot\nabla_x \mu_\delta}{\delta} + 2Q(\mu_\delta, F_{1\delta}) \right\} \tag{18}$$

$$+ \frac{\delta\varepsilon}{\varepsilon^{\frac{3}{2}}} \left\{ -\partial_t F_{1\delta} - v\cdot\nabla_x F_{1\delta} + 2\delta Q(F_{1\delta}, F_{1\delta}) + 2Q(\mu_\delta, F_{2\delta}) \right\} \tag{19}$$

$$+ \frac{\delta\varepsilon^2}{\varepsilon^{\frac{3}{2}}} \left\{ -\partial_t F_{2\delta} - v\cdot\nabla_x F_{2\delta} + 2\delta Q(F_{1\delta}, F_{2\delta}) + \delta\varepsilon Q(F_{2\delta}, F_{2\delta}) \right\} \tag{20}$$

$$+ \frac{\varepsilon^{\frac{3}{2}}}{\varepsilon^{\frac{3}{2}}} \left\{ 2\delta Q(F_{1\delta}, F_{R\varepsilon,\delta}) + 2\delta\varepsilon Q(F_{2\delta}, F_{R\varepsilon,\delta}) \right\}. \tag{21}$$

We first claim that (18) and (19) vanish upon the suitable choice of μ_δ and $F_{1\delta}$.

By the Fredholm and the collision invariance (3), the whole line of (18) vanishes if

$$\int_{\mathbb{R}^3} \{\partial_t \mu_\delta + v\cdot\nabla_x \mu_\delta\}\varphi_j \mathrm{d}v = 0 \ \text{ for all } j = 0, 1, \cdots, 4 \tag{22}$$

and

$$(\mathbf{I} - \mathbf{P}_\delta)(\frac{F_{1\delta}}{\sqrt{\mu_\delta}}) = -{L_\delta}^{-1}\left(\frac{\partial_t \mu_\delta + v\cdot\nabla_x \mu_\delta}{\delta\sqrt{\mu_\delta}}\right) \tag{23}$$

where $\mathbf{I}$ is the identity operator. Realizing that (22) is equivalent to (9), the line of (18) vanishes if $(1 + \delta\sigma, \delta u, 1 + \delta\theta)$ solves the Euler system (10). From the standard theory on the classical solutions of the Euler system, a smooth solution $(1 + \delta\sigma, \delta u, 1 + \delta\theta)$ persists with lifespan bounded from below by

$$\tau_\delta \geq \frac{\mathfrak{C}}{\delta} \ \text{ for some } \mathfrak{C} > 0. \tag{24}$$

On the other hand, the acoustic system (7) is the linearization of (10) around $(1, 0, 1)$. A smooth solution of this linear system persists global in time and it stays close to the perturbation of the Euler solution as

$$\sup_{0\leq t\leq \tau_\delta} \|(\sigma - \sigma_A, u - u_A, \theta - \theta_A)\|_{H^s} \lesssim \delta \tag{25}$$

with initial condition $(\sigma, u, \theta)|_{t=0} = (\sigma_A, u_A, \theta_A)|_{t=0} = (\sigma_A^0, u_A^0, \theta_A^0)$. See Lemmas 3.1 and 3.2 of [9] for the proofs. Therefore we have

$$\mu_\delta = \mu_0 + \delta\left\{\sigma_A + u_A\cdot v + \theta_A\frac{|v|^2 - 3}{2}\right\}\mu_0 + o(\delta)\mu_0^{1-} \ \text{ for } 0 \leq t \leq \tau_\delta. \tag{26}$$

Likewise, the whole line of (19) vanishes if $F_{1\delta}$ satisfies

$$\int_{\mathbb{R}^3} \{\partial_t F_{1\delta} + v\cdot\nabla_x F_{1\delta}\}\varphi_j \mathrm{d}v = 0 \ \text{ for all } j = 0, 1, \cdots, 4. \tag{27}$$

$(\mathbf{I} - \mathbf{P}_\delta)(\frac{F_{1\delta}}{\sqrt{\mu_\delta}})$ is determined by (23) and hence, the condition (27) gives rise to the equations for $P_j(\frac{F_{1\delta}}{\sqrt{\mu_\delta}})$. The standard theory of linear hyperbolic system induces a smooth solution of (27) (see Appendix). Further by (27), the microscopic part of $F_{2\delta}$ is completely determined by $F_{1\delta}$:

$$(\mathbf{I} - \mathbf{P}_\delta)(\frac{F_{2\delta}}{\sqrt{\mu_\delta}}) = -{L_\delta}^{-1}\left(\frac{\partial_t F_{1\delta} + v\cdot\nabla_x F_{1\delta} - 2\delta Q(F_{1\delta}, F_{1\delta})}{\sqrt{\mu_\delta}}\right). \tag{28}$$

We set $\mathbf{P}_\delta(\frac{F_{2\delta}}{\sqrt{\mu_\delta}}) = 0$ for simplicity. With (22) and (27), the equation of F_R is (17), (20), (21) without (18) and (19).

The main result of this note is the rigorous justification of the acoustic limit by using the expansion (13) to the rescaled Boltzmann equation (1) with (6):

Theorem 1 *Assume* $0 < \varepsilon \ll 1$ *and* δ *in (8). Assume* $(\rho_A^0, u_A^0, \theta_A^0) \in H^s(\Omega)$ *and* $P_\delta(\frac{F_{1\delta}}{\sqrt{\mu_\delta}})|_{t=0} \in H^s(\Omega)$ *for* $s > 5$*, and also assume* (23) *holds at* $t = 0$*. Then there exist smooth* $F_{1\delta}$ *and* $F_{2\delta}$ *uniformly bounded in* δ *for* $t \geq \frac{C}{\delta}$*. Suppose the initial datum satisfies (75), then for* $0 < \delta \ll 1$

$$\sup_{0\leq t\leq \frac{C}{\delta}} \left\| \frac{F_\varepsilon(t) - \mu_0}{\delta} - \big(\sigma_A + u_A \cdot v + \theta_A \frac{|v|^2 - 3}{2}\big)\mu_0 \right\|_{L^2_{xv}} \lesssim \delta, \tag{29}$$

where C *does not depend on* δ, t.

Remark 1 The initial conditions are "well-prepared" as in the assumptions of Theorem 1, but they are less restrictive than in the previous setting [9] as the expansion order is 2. And the required initial compatibility conditions avoid initial layers.

Remark 2 Other collision operators for the hard potential and soft potential with an angular cutoff can be treated in the same way as in [9].

Remark 3 Multi-scale Hilbert expansion introduced in this section sheds some light on other hydrodynamic limit problems. For instance, see [20] for a multi-scale Hilbert expansion to the incompressible Euler limit from the Boltzmann equation with diffuse boundary, where the scaling is given $\mathcal{S}t = \varepsilon$, $\mathcal{M}a = \varepsilon$, $\mathcal{K}n = \kappa\varepsilon$ with $\kappa = \kappa(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$ and $\kappa \gg \epsilon$.

With smooth coefficients μ_δ, $F_{1\delta}$ and $F_{2\delta}$ obtained in the Hilbert expansion in the above using (13) and (26), in order to prove Theorem 1, it suffices to derive the uniform bounds of the remainder F_R. To that end, we invoke the $L^p - L^\infty$ theory of the Boltzmann equation and a recent L^6-integrability of $\mathbf{P}_\delta f$ in [21]. Key estimates and ingredients of the proof are presented in the following sections.

3 L^2 Theory

In order to utilize the L^2-theory with symmetric linear operator (14) we introduce

$$F_{R\varepsilon,\delta} = \sqrt{\mu_\delta} f_{R\varepsilon,\delta}, \quad F_{i\delta} = \sqrt{\mu_\delta} f_{i\delta} \text{ for } i = 1, 2. \tag{30}$$

The equation of $f_{R_{\varepsilon,\delta}}$ is given by

$$\begin{aligned}&\left[\partial_t + v\cdot\nabla_x + \varepsilon^{-1}L_\delta\right] f_{R_{\varepsilon,\delta}}\\ &= \varepsilon^{\frac{1}{2}}\Gamma_\delta(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}}) + 2\delta\Gamma_\delta(f_{1\delta}, f_{R_{\varepsilon,\delta}}) + \delta\mathfrak{r}_{\varepsilon,\delta} - \frac{1}{2}\frac{[\partial_t + v\cdot\nabla_x]\mu_\delta}{\mu_\delta} f_{R_{\varepsilon,\delta}},\end{aligned} \tag{31}$$

where

$$\begin{aligned}\mathfrak{r}_{\varepsilon,\delta} :=& 2\varepsilon\Gamma_\delta(f_{2\delta}, f_{R_{\varepsilon,\delta}}) + \delta\varepsilon^{\frac{1}{2}}\{2\Gamma_\delta(f_{1\delta}, f_{2\delta}) + \varepsilon\Gamma_\delta(f_{2\delta}, f_{2\delta})\}\\ &+ \varepsilon^{\frac{1}{2}}\Big\{-\partial_t f_{2\delta} - v\cdot\nabla_x f_{2\delta} - \frac{1}{2}\frac{[\partial_t + v\cdot\nabla_x]\mu_\delta}{\mu_\delta} f_{2\delta}\Big\}.\end{aligned} \tag{32}$$

Here we have used the notation

$$\Gamma_\delta(f, g) := \frac{1}{\sqrt{\mu_\delta}} Q(\sqrt{\mu_\delta} f, \sqrt{\mu_\delta} g).$$

From (3) and (15) we have

$$\int_{\mathbb{R}^3} \Gamma_\delta(f, g) h \mathrm{d}v = \int_{\mathbb{R}^3} \Gamma_\delta(f, g)(\mathbf{I} - \mathbf{P}_\delta) h \mathrm{d}v.$$

The same holds for L:

$$\int_{\mathbb{R}^3} L_\delta f \; f \mathrm{d}v = \int_{\mathbb{R}^3} L_\delta f \; (\mathbf{I} - \mathbf{P}_\delta) f \mathrm{d}v = \int_{\mathbb{R}^3} L_\delta (\mathbf{I} - \mathbf{P}_\delta) f \; (\mathbf{I} - \mathbf{P}_\delta) f \mathrm{d}v.$$

A standard decomposition yields

$$L_\delta f = \nu_\delta f - \int_{\mathbb{R}^3} \mathbf{k}(v, v_*) f(v_*) \mathrm{d}v_*, \tag{33}$$

where $\nu_\delta(v) := \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} B(v - v_*, \omega)\mu_\delta(v_*)\mathrm{d}\omega\mathrm{d}v_* \sim \langle v\rangle$. Here we assume $1 + \delta\sigma > 0$ and $1 + \delta\theta > 0$, which are valid for $0 \le t \le \tau_\delta$.

Our analysis will involve the estimates of $\partial_t f_{R_{\varepsilon,\delta}}$. The equation of $\partial_t f_{R_{\varepsilon,\delta}}$ is given by

$$\begin{aligned}&\left[\partial_t + v\cdot\nabla_x + \varepsilon^{-1}L_\delta\right]\partial_t f_{R_{\varepsilon,\delta}} = -\varepsilon^{-1}L_{\delta t}(\mathbf{I} - \mathbf{P}_\delta) f_{R_{\varepsilon,\delta}} + \varepsilon^{-1}L_\delta(\mathbf{P}_{\delta t} f_{R_{\varepsilon,\delta}})\\ &\quad + 2\varepsilon^{\frac{1}{2}}\Gamma_\delta(f_{R_{\varepsilon,\delta}}, \partial_t f_{R_{\varepsilon,\delta}}) + \varepsilon^{\frac{1}{2}}\Gamma_{\delta t}(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}}) + 2\delta\Gamma_\delta(f_{1\delta}, \partial_t f_{R_{\varepsilon,\delta}})\\ &\quad + 2\delta\Gamma_\delta(\partial_t f_{1\delta}, f_{R_{\varepsilon,\delta}}) + 2\delta\Gamma_{\delta t}(f_{1\delta}, f_{R_{\varepsilon,\delta}})\\ &\quad + \delta\partial_t\mathfrak{r}_{\varepsilon,\delta} - \frac{1}{2}\frac{[\partial_t + v\cdot\nabla_x]\mu_\delta}{\mu_\delta}\partial_t f_{R_{\varepsilon,\delta}} - \frac{1}{2}\partial_t\left(\frac{[\partial_t + v\cdot\nabla_x]\mu_\delta}{\mu_\delta}\right) f_{R_{\varepsilon,\delta}}\end{aligned} \tag{34}$$

where the commutators $L_{\delta t}$, $\mathbf{P}_{\delta t}$, and $\Gamma_{\delta t}$ are given as

$$\begin{aligned} &L_{\delta t} g = \partial_t (L_\delta g) - L_\delta(\partial_t g), \quad \mathbf{P}_{\delta t} g = \partial_t (\mathbf{P}_\delta g) - \mathbf{P}_\delta (\partial_t g), \\ &\Gamma_{\delta t}(g_1, g_2) = \partial_t(\Gamma_\delta(g_1, g_2)) - \Gamma_\delta(\partial_t g_1, g_2) - \Gamma_\delta(g_1, \partial_t g_2). \end{aligned} \tag{35}$$

We now define the energy and dissipation as

$$\mathcal{E}_{\varepsilon,\delta}(t) := \|f_{R_{\varepsilon,\delta}}(t)\|^2_{L^2_{x,v}} + \varepsilon \|\partial_t f_{R_{\varepsilon,\delta}}(t)\|^2_{L^2_{x,v}} \tag{36}$$

and

$$\mathcal{D}_{\varepsilon,\delta}(t) := [\varepsilon^{-\frac{1}{2}} \|\sqrt{\nu_\delta}(\mathbf{I} - \mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}\|_{L^2_{x,v}}]^2 + \|\sqrt{\nu_\delta}(\mathbf{I} - \mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^2_{x,v}}. \tag{37}$$

From the standard spectral gap estimate, we have the following L^2 estimate:

$$\mathcal{E}_{\varepsilon,\delta}(t) + \int_0^t \mathcal{D}_{\varepsilon,\delta}(t) \tag{38}$$

$$\lesssim \mathcal{E}_{\varepsilon,\delta}(0) + \int_0^t \left\| \frac{\varepsilon}{\sqrt{\nu_\delta}} \Gamma_\delta(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}}) \right\|^2_{L^2_{x,v}} + \left\| \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{\nu_\delta}} \Gamma_\delta(f_{R_{\varepsilon,\delta}}, \partial_t f_{R_{\varepsilon,\delta}}) \right\|^2_{L^2_{x,v}} \tag{39}$$

$$+ \int_0^t \left\| \frac{\delta \varepsilon^{\frac{1}{2}}}{\sqrt{\nu_\delta}} \Gamma_\delta(f_{1\delta}, f_{R_{\varepsilon,\delta}}) \right\|^2_{L^2_{x,v}} \tag{40}$$

$$+ \int_0^t \left\| \frac{\delta \varepsilon}{\sqrt{\nu_\delta}} \Gamma_\delta(f_{1\delta}, \partial_t f_{R_{\varepsilon,\delta}}) \right\|^2_{L^2_{x,v}} + \left\| \frac{\delta \varepsilon}{\sqrt{\nu_\delta}} \Gamma_\delta(\partial_t f_{1\delta}, f_{R_{\varepsilon,\delta}}) \right\|^2_{L^2_{x,v}} \tag{41}$$

$$\begin{aligned} &+ \int_0^t \iint_{\Omega \times \mathbb{R}^3} \{ |L_{\delta t}(\mathbf{I} - \mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}| + |L_\delta(\mathbf{P}_{\delta t} f_{R_{\varepsilon,\delta}})| + \varepsilon^{\frac{3}{2}} |\Gamma_{\delta t}(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}})| \\ &\qquad + \delta\varepsilon |\Gamma_{\delta t}(f_{1\delta}, f_{R_{\varepsilon,\delta}})| \} |\partial_t f_{R_{\varepsilon,\delta}}| \end{aligned} \tag{42}$$

$$\begin{aligned} &+ \sup_{0 \le s \le t} \|(\sigma, u, \theta)\|_{C^2_{t,x}} \\ &\qquad \times \int_0^t \iint_{\Omega \times \mathbb{R}^3} \delta \langle v \rangle^3 |f_{R_{\varepsilon,\delta}}|^2 + \delta\varepsilon \langle v \rangle^3 (|f_{R_{\varepsilon,\delta}}||\partial_t f_{R_{\varepsilon,\delta}}| + |\partial_t f_{R_{\varepsilon,\delta}}|^2) \end{aligned} \tag{43}$$

$$+ \delta \int_0^t \iint_{\Omega \times \mathbb{R}^3} |\mathfrak{r}_{\varepsilon,\delta} f_{R_{\varepsilon,\delta}}| + \varepsilon |\partial_t \mathfrak{r}_{\varepsilon,\delta} \partial_t f_{R_{\varepsilon,\delta}}|. \tag{44}$$

4 L^∞ Estimate

The main obstacle arising from the L^2 estimate around a local Maxwellian μ_δ is the unpleasant velocity growth coming from the term $\mu_\delta{}^{-1}(\partial_t + v\cdot\nabla_x)\mu_\delta f_{R_{\varepsilon,\delta}}$, which is a cubic polynomial in v as in (43). To remedy this difficulty, following Caflisch [9, 22], we introduce a global Maxwellian

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

where T_M satisfies the following condition

$$T_M < T(t,x) = 1 + \delta\theta(t,x) < 2T_M \ \text{ for all } (t,x) \in [0,\tau]\times\Omega. \tag{45}$$

This moderate temperature variation condition is achieved for sufficiently small initial perturbations. Note that under the assumption (45), there exist constants c_1 , c_2 such that for some $1/2 < \alpha < 1$

$$c_1\mu_M \le \mu_\delta \le c_2\mu_M^\alpha \ \text{ for all } \ (t\,,x\,,v) \in [0,\tau]\times\Omega\times\mathbb{R}^3. \tag{46}$$

We further define

$$h_{\varepsilon,\delta} := \frac{w}{\sqrt{\mu_M}} F_{R_{\varepsilon,\delta}}, \quad w = \langle v\rangle^\beta \tag{47}$$

for any fixed $\beta \ge 9$.

We also define

$$\mathcal{L}_{\delta M} g = -\frac{2}{\sqrt{\mu_M}} Q(\mu_\delta, \sqrt{\mu_M} g) = \nu_\delta g - K_\delta g \tag{48}$$

where

$$K_\delta g = \int_{\mathbb{R}^3} \mathbf{k}(v, v_*) \frac{\sqrt{\mu_\delta(v)}\sqrt{\mu_M(v_*)}}{\sqrt{\mu_\delta(v_*)}\sqrt{\mu_M(v)}} g(v_*)\mathrm{d}v_*. \tag{49}$$

We may write for any $m > 0$,

$$K_\delta g = K_\delta{}^m g + K_\delta{}^c g \tag{50}$$

where

$$|K_\delta{}^m g(v)| \lesssim m^4 \nu_\delta \|g\|_\infty \tag{51}$$

and

$$K_\delta{}^c g(v) = \int_{\mathbb{R}^3} l(v, v_*) dv_* \text{ where } l(v, v_*) \lesssim_m \frac{e^{-c|v-v_*|^2}}{|v - v_*|} \text{ for some } c > 0. \tag{52}$$

See Lemma 2.3 of [9] for the proof.

Now by letting $K_{\delta w} g = w K_\delta(\frac{g}{w})$,

$$\begin{aligned} &\left[\partial_t + v\cdot\nabla_x + \frac{\nu_\delta}{\varepsilon}\right] h_{\varepsilon,\delta}(t, x, v) \\ &= \frac{1}{\varepsilon} K_{\delta w} h_{\varepsilon,\delta} + \varepsilon^{1/2} \frac{w}{\sqrt{\mu_M}} Q\left(\frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}, \frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}\right) + \delta \tilde{\mathfrak{r}}_{\varepsilon,\delta} \end{aligned} \tag{53}$$

where

$$\begin{aligned} \tilde{\mathfrak{r}}_{\varepsilon,\delta} = &\frac{w}{\sqrt{\mu_M}}\left[2Q\left(F_1, \frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}\right) + 2\varepsilon Q\left(F_{2\delta}, \frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}\right)\right] \\ &+ \varepsilon^{\frac{1}{2}} \frac{w}{\sqrt{\mu_M}}[-\partial_t F_{2\delta} - v\cdot\nabla_x F_{2\delta} + 2\delta Q(F_{1\delta}, F_{2\delta}) + \delta\varepsilon Q(F_{2\delta}, F_{2\delta})]. \end{aligned} \tag{54}$$

Then we may integrate (53) along the trajectory:

$$\begin{aligned} h_{\varepsilon,\delta}(t, x, v) &= \exp\{-\frac{1}{\varepsilon}\int_0^t \nu_\delta d\tau\} h_{\varepsilon,\delta}(0, x - vt, v) \\ &- \int_0^t \exp\{-\frac{1}{\varepsilon}\int_s^t \nu_\delta d\tau\}\left(\frac{1}{\varepsilon} K_\delta{}^m_w h_{\varepsilon,\delta}\right)(s, x - v(t-s), v) ds \\ &- \int_0^t \exp\{-\frac{1}{\varepsilon}\int_s^t \nu_\delta d\tau\}\left(\frac{1}{\varepsilon} K_\delta{}^c_w h_{\varepsilon,\delta}\right)(s, x - v(t-s), v) ds \\ &+ \int_0^t \exp\{-\frac{1}{\varepsilon}\int_s^t \nu_\delta d\tau\}\left(\frac{\varepsilon^{1/2} w}{\sqrt{\mu_M}} Q(\frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}, \frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w})\right)(s, x - v(t-s), v) ds \\ &+ \int_0^t \exp\{-\frac{1}{\varepsilon}\int_s^t \nu_\delta d\tau\}\delta\tilde{\mathfrak{r}}_{\varepsilon,\delta}(s, x - v(t-s), v) ds. \end{aligned} \tag{55}$$

Recall

$$\left|\frac{w}{\sqrt{\mu_M}} Q\left(\frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}, \frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}\right)\right| \lesssim \nu_\delta \|h_{\varepsilon,\delta}\|_\infty^2. \tag{56}$$

From $\mu_M \lesssim \mu_\delta$ and (51), (52), we first obtain for $N \gg 1$,

$$
\begin{aligned}
&|h_{\varepsilon,\delta}(t,x,v)| \\
&\lesssim e^{-\frac{\nu_\delta}{\varepsilon}t}\|hh_{\varepsilon,\delta}(0)\|_{L^\infty_{x,v}} \\
&+\int_0^t \frac{\nu_\delta}{\varepsilon}e^{-\frac{\nu_\delta}{\varepsilon}t}\varepsilon\Big\{o(1)+\sup_{0\le s\le t}\varepsilon^{\frac12}\|e^{\frac{\nu_\delta}{2\varepsilon}s}h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}}\Big\}\sup_{0\le s\le t}\|e^{\frac{\nu_\delta}{2\varepsilon}s}h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}}\mathrm{d}s \\
&+\int_0^t e^{-\frac{\nu_\delta}{\varepsilon}(t-s)}\delta|\tilde{\mathfrak{r}}_{\varepsilon,\delta}|\mathrm{d}s++o(1)e^{-\frac{\nu_0}{2\varepsilon}t}\sup_{0\le s\le t}\|e^{\frac{\nu_\delta}{2\varepsilon}s}h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}} \\
&+\int_0^t\int_{|v_*|\le N,|v-v_*|\ge\frac1N}\frac{e^{-\frac{\nu_\delta}{\varepsilon}(t-s)}}{\varepsilon}l_w(v,v_*)|h_{\varepsilon,\delta}(s,x-(t-s)v,v_*)|\mathrm{d}v_*\mathrm{d}s.
\end{aligned}
\tag{57}
$$

Iterating once again and splitting the time integration in $s_* \in [0, s-o(1)\varepsilon]\cup[s-o(1)\varepsilon, s]$ we bound (57) by

$$
\begin{aligned}
&e^{-\frac{\nu_\delta}{\varepsilon}t}\|h_{\varepsilon,\delta}(0)\|_{L^\infty_{x,v}} \\
&+\varepsilon\Big\{o(1)+\sup_{0\le s\le t}\varepsilon^{\frac12}\|e^{-\frac{\nu_\delta}{2\varepsilon}(t-s)}h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}}\Big\}\sup_{0\le s\le t}\|e^{-\frac{\nu_\delta}{2\varepsilon}(t-s)}h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}} \\
&+\varepsilon\delta\sup_{0\le s\le t}\|e^{-\frac{\nu_\delta}{2\varepsilon}(t-s)}\tilde{\mathfrak{r}}_{\varepsilon,\delta}\|_{L^\infty_{x,v}} \\
&+\int_0^t\frac{e^{-\frac{\nu_\delta(v)}{\varepsilon}(t-s)}}{\varepsilon}\int_{|v_*|\le N,|v-v_*|\ge\frac1N}\int_0^{s-o(1)\varepsilon}\frac{e^{-\frac{\nu_\delta(v_*)}{\varepsilon}(s-s_*)}}{\varepsilon}\int_{|v_{**}|\le N,|v_*-v_{**}|\ge\frac1N} \\
&\quad\times l_w(v,v_*)l_w(v_*,v_{**})w(v_{**})|f_{R\varepsilon,\delta}(s_*,x-(t-s)v-(s-s_*)v_*,v_{**})|\mathrm{d}v_{**}\mathrm{d}s_*\mathrm{d}v_*\mathrm{d}s.
\end{aligned}
\tag{58}
$$

Note that $l_w(v,v_*)l_w(v_*,v_{**})w(v_{**}) \lesssim_N 1$ within the above integration regime. We first split $f_{R\varepsilon,\delta}$ into $\mathbf{P}_\delta f_{R\varepsilon,\delta}+(\mathbf{I}-\mathbf{P}_\delta)f_{R\varepsilon,\delta}$ and then use Hölder inequality for any $1<p_1,p_2<\infty$ to bound the last two whole lines of (58) by

$$
\begin{aligned}
&\int_0^t\frac{e^{-\frac{\nu_0}{\varepsilon}(t-s)}}{\varepsilon}\int_0^{s-o(1)\varepsilon}\frac{e^{-\frac{\nu_0}{\varepsilon}(s-s_*)}}{\varepsilon}\Big\{\|\mathbf{P}_\delta f_{R\varepsilon,\delta}(s_*,X(v_*),v_{**})\|_{L^{p_1}_{v_*,v_{**}}} \\
&\qquad+\|(\mathbf{I}-\mathbf{P}_\delta)f_{R\varepsilon,\delta}(s_*,X(v_*),v_{**})\|_{L^{p_2}_{v_*,v_{**}}}\Big\}\mathrm{d}s_*\mathrm{d}s.
\end{aligned}
\tag{59}
$$

Then applying the change of variables $v_* \mapsto X(v_*) := x-(t-s)v-(s-s_*)v_*$ with $\det\left(\frac{\partial(x-(t-s)v-(s-s_*)v_*)}{\partial v_*}\right) \gtrsim o(1)\varepsilon^3$ for $s_* \in [0, s-o(1)\varepsilon]$, we derive

$$\begin{aligned}\|\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}(s_*, X(v_*), v_{**})\|_{L^{p_1}_{v_*,v_{**}}} &\lesssim \varepsilon^{-\frac{3}{p_1}}\|\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}(s_*)\|_{L^{p_1}_{x,v}},\\ \|(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(s_*, X(v_*), v_{**})\|_{L^{p_2}_{v_*,v_{**}}} &\lesssim \varepsilon^{-\frac{3}{p_2}}\|(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(s_*)\|_{L^{p_2}_{x,v}}.\end{aligned} \tag{60}$$

From (57)–(60) we conclude that

$$\begin{aligned}\|h_{\varepsilon,\delta}(t)\|_{L^\infty_{x,v}} \lesssim\ & \|h_{\varepsilon,\delta}(0)\|_{L^\infty_{x,v}} + \varepsilon\Big\{o(1) + \sup_{0\le s\le t}\varepsilon^{\frac12}\|h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}}\Big\}\sup_{0\le s\le t}\|h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}}\\ &+\varepsilon\delta\|\tilde{\mathfrak{r}}_{\varepsilon,\delta}\|_{L^\infty_{x,v}} + \sup_{0\le s\le t}\Big\{\varepsilon^{-\frac{3}{p_1}}\|\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}(s)\|_{L^{p_1}_{x,v}} + \varepsilon^{-\frac{3}{p_2}}\|(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(s)\|_{L^{p_2}_{x,v}}\Big\}.\end{aligned} \tag{61}$$

Analogously, we have the equation for $\partial_t h_{\varepsilon,\delta}$:

$$\begin{aligned}&\Big[\partial_t + v\cdot\nabla_x + \frac{\nu_\delta}{\varepsilon}\Big]\partial_t h_{\varepsilon,\delta}(t,x,v)\\ =&\frac{1}{\varepsilon}K_{\delta w}\partial_t h_{\varepsilon,\delta} + 2\varepsilon^{1/2}\frac{w}{\sqrt{\mu_M}}Q\left(\frac{h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}, \frac{\partial_t h_{\varepsilon,\delta}\sqrt{\mu_M}}{w}\right) + \delta\partial_t\tilde{\mathfrak{r}}_{\varepsilon,\delta}\\ &+\frac{1}{\varepsilon}(K_{\delta w})_t h_{\varepsilon,\delta} - \frac{(\nu_\delta)_t h_{\varepsilon,\delta}}{\varepsilon}\end{aligned} \tag{62}$$

where the last line represents the commutators typically containing $\delta\langle v\rangle^3$. By a similar trajectory argument, we deduce that for $\tilde{h}_{\varepsilon,\delta} = \langle v\rangle^{-3}\partial_t h_{\varepsilon,\delta}$

$$\begin{aligned}&\|\tilde{h}_{\varepsilon,\delta}\|_{L^2_t L^\infty_{x,v}}\\ \lesssim\ & \|\tilde{h}(0)\|_{L^\infty_{x,v}} + \varepsilon\Big\{o(1) + \sup_{0\le s\le t}\varepsilon^{\frac12}\|h_{\varepsilon,\delta}(s)\|_{L^\infty_{x,v}}\Big\}\|\tilde{h}_{\varepsilon,\delta}\|_{L^2_t L^\infty_{x,v}} + \varepsilon\delta\|\langle v\rangle^{-3}\partial_t\tilde{\mathfrak{r}}_{\varepsilon,\delta}\|_{L^2_t L^\infty_{x,v}}\\ &+\delta\left\{\|h_{\varepsilon,\delta}\|_{L^\infty_{t,x,v}} + \varepsilon^{-\frac{3}{p_1}}\|P_\delta f_{R_{\varepsilon,\delta}}(s)\|_{L^2_t L^{p_1}_{x,v}} + \varepsilon^{-\frac32}\|(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(s)\|_{L^2_{t,x,v}}\right\}\\ &+\varepsilon^{-\frac{3}{p_1}}\|P_\delta\partial_t f_{R_{\varepsilon,\delta}}(s)\|_{L^2_t L^{p_1}_{x,v}} + \varepsilon^{-\frac32}\|(\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}(s)\|_{L^2_{t,x,v}}.\end{aligned} \tag{63}$$

5 $P_\delta f_{R_{\varepsilon,\delta}}$ Estimate in $L^2_t L^3_x$ and $L^\infty_t L^6_x$

In this section we estimate $P_\delta f_{R_{\varepsilon,\delta}}$ in (16). The first estimate is a direct consequence of the Average lemma (e.g. [23]) and the Sobolev embedding $H^{\frac{1}{2}}_x \subset L^3_x$ in 3D:

$$\int_0^t \|\varepsilon^{\frac{1}{2}} P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^3_x} + \int_0^t \|\varepsilon P \partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^3_x} \tag{64}$$

$$\lesssim \int_0^t \mathcal{E}_{\varepsilon,\delta} + \int_0^t \mathcal{D}_{\varepsilon,\delta} + \int_0^t \|\varepsilon^{\frac{1}{2}} \delta \Gamma_\delta(f_{1\delta}, f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} \tag{65}$$

$$+ \int_0^t \|\varepsilon \Gamma_\delta(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} + \int_0^t \|\varepsilon^{\frac{3}{2}} \Gamma_\delta(f_{R_{\varepsilon,\delta}}, \partial_t f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} \tag{66}$$

$$+ \int_0^t \|\varepsilon \delta \Gamma_\delta(f_{1\delta}, \partial_t f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} + \|\varepsilon \delta \Gamma_\delta(\partial_t f_{1\delta}, f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} \tag{67}$$

$$+ \int_0^t \|L_{\delta_t}(\mathbf{I} - \mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}\|^2_{L^2_{x,v}} + \|L_\delta(\mathbf{P}_{\delta_t} f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}}$$
$$+ \|\varepsilon^{\frac{3}{2}} \Gamma_{\delta_t}(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} + \|\delta\varepsilon \Gamma_{\delta_t}(f_{1\delta}, f_{R_{\varepsilon,\delta}})\|^2_{L^2_{x,v}} \tag{68}$$

$$+ \int_0^t \|\varepsilon^{\frac{1}{2}} \delta \langle v\rangle^3 f_{R_{\varepsilon,\delta}}\|^2_{L^2_{x,v}} + \|\varepsilon\delta \langle v\rangle^3 \partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^2_{x,v}} \tag{69}$$

$$+ \int_0^t \|\varepsilon^{\frac{1}{2}} \delta \mathfrak{r}_{\varepsilon,\delta}\|^2_{L^2_{x,v}} + \|\varepsilon\delta \partial_t \mathfrak{r}_{\varepsilon,\delta}\|^2_{L^2_{x,v}}. \tag{70}$$

The second estimate comes from the test function method of [21, 24]. We employ a weak formulation of (31)

$$\iint_{\Omega\times\mathbb{R}^3} -\varepsilon^{\frac{1}{2}} \mathbf{P}_\delta f_{R_{\varepsilon,\delta}} v \cdot \nabla_x \psi$$
$$= \iint_{\Omega\times\mathbb{R}^3} \{-\varepsilon^{\frac{1}{2}} \partial_t f_{R_{\varepsilon,\delta}} - \varepsilon^{-\frac{1}{2}} L_\delta f_{R_{\varepsilon,\delta}} + \varepsilon \Gamma_\delta(f_{R_{\varepsilon,\delta}}, f_{R_{\varepsilon,\delta}}) + \varepsilon^{\frac{1}{2}} \delta \mathfrak{r}_{\varepsilon,\delta} \tag{71}$$
$$- \frac{\varepsilon^{\frac{1}{2}}}{2} \frac{[\partial_t + v\cdot\nabla_x]\mu_\delta}{\mu_\delta} f_{R_{\varepsilon,\delta}}\}\psi + \varepsilon^{\frac{1}{2}}(\mathbf{I} - \mathbf{P}_\delta) f_{R_{\varepsilon,\delta}} v \cdot \nabla_x \psi.$$

with special test functions

$$\begin{aligned}
\psi_j &:= (|v-\delta u|^2-\beta_j)v\cdot\nabla_x\phi_j\sqrt{\mu_\delta} \quad \text{for } j=0,4\\
\psi^1_{i,j} &:= \big((v_i-\delta u_i)^2-\beta\big)\partial_j\phi_j\sqrt{\mu_\delta} \qquad \text{for } i,j\in\{1,2,3\},\\
\psi^2_{i,j} &:= |v-\delta u|^2(v_i-\delta u_i)(v_j-\delta u_j)\partial_i\phi_j\sqrt{\mu_\delta} \quad \text{for } i,j\in\{1,2,3\},\ i\neq j,
\end{aligned} \tag{72}$$

with $-\Delta_x\phi_j=(P_jf_{R_{\varepsilon,\delta}})^5-\frac{1}{|\Omega|}\int_\Omega(P_jf_{R_{\varepsilon,\delta}})^5\mathrm{d}x$ with some condition on β,β_0,β_4.

Then following the strategy of [20, 21], we use these test functions in (71), Hölder inequality and the embedding $W^{1,\frac{6}{5}}(\mathbb{R}^3_x)\subset L^2(\mathbb{R}^3_x)$ with $\frac{1}{2}=\frac{1}{6/5}-\frac{1}{3}$ which implies that

$$\begin{aligned}
&\|\nabla_x(-\Delta_x)^{-1}|\varepsilon^{\frac{1}{2}}\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}|^5\|_2\\
\lesssim\ &\|\nabla_x(-\Delta_x)^{-1}|\varepsilon^{\frac{1}{2}}\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}|^5\|_{W^{1,\frac{6}{5}}(\mathbb{R}^3_x)}\\
\lesssim\ &\||\varepsilon^{\frac{1}{2}}\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}|^5\|_{L^{\frac{6}{5}}(\mathbb{R}^3_x)}\\
\lesssim\ &\|\varepsilon^{\frac{1}{2}}\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}\|^5_{L^6(\mathbb{R}^3_x)}
\end{aligned}$$

to deduce that

$$\begin{aligned}
\|\varepsilon^{\frac{1}{2}}Pf_{R_{\varepsilon,\delta}}(t)\|_6 \lesssim&\|\varepsilon^{-\frac{1}{2}}\sqrt{\nu}(\mathbf{I}-\mathbf{P}_\delta)f_{R_{\varepsilon,\delta}}(t)\|_{L^2_{x,v}}+\sqrt{\mathcal{E}_{\varepsilon,\delta}(t)}\\
&+\|\varepsilon\Gamma_\delta(f_{R_{\varepsilon,\delta}}(t),f_{R_{\varepsilon,\delta}}(t))\|_{L^2_{x,v}}+\varepsilon^{\frac{1}{2}}\delta\|\mathfrak{r}_{\varepsilon,\delta}(t)\|_{L^2_{x,v}},
\end{aligned} \tag{73}$$

as long as $\|(\sigma_A,u_A,\theta_A)\|_{C^1}\lesssim 1$. We refer to [20, 21] for more detail.

6 Closing the Estimates

Define the final energy and dissipation with a parameter $0<a\ll 1$

$$\begin{aligned}
\tilde{\mathcal{E}}_{\varepsilon,\delta}(t) &:= \sup_{0\le s\le t}\Big\{\mathcal{E}_{\varepsilon,\delta}(t)+a\|\varepsilon h_{\varepsilon,\delta}(s)\|^2_{L^\infty_{x,v}}+a\|\varepsilon^{\frac{1}{2}}\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}(t)\|^2_{L^6_x}\Big\},\\
\tilde{\mathcal{D}}_{\varepsilon,\delta}(t) &:= \mathcal{D}_{\varepsilon,\delta}(t)+a\|\varepsilon^{\frac{1}{2}}P_\delta f_{R_{\varepsilon,\delta}}(t)\|^2_{L^3_x}+a\|\varepsilon P_\delta\partial_t f_{R_{\varepsilon,\delta}}(t)\|^2_{L^3_x}+a\|\varepsilon^{\frac{3}{2}}\tilde{h}_{\varepsilon,\delta}(s)\|^2_{L^\infty_{x,v}}.
\end{aligned} \tag{74}$$

We further assume the initial data satisfy with some $0 < b \ll 1$ which will be specified later

$$\tilde{\mathcal{E}}_{\varepsilon,\delta}(0) + \|\varepsilon^{-\frac{1}{2}}(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(0)\|^2_{L^2_{x,v}} < b\delta. \tag{75}$$

Define

$$T_\delta := \sup\{t \geq 0 : \tilde{\mathcal{E}}_{\varepsilon,\delta}(t) < \delta\}. \tag{76}$$

To prove Theorem 1, it suffices to show the following:

Lemma 1 *Assume the same as in Theorem 1. Then there exists a constant $C > 0$ independent of δ such that*

$$T_\delta \geq \frac{C}{\delta}. \tag{77}$$

Proof First from the Hölder inequality we obtain

$$\begin{aligned}
\|\Gamma_\delta(g_1, g_2)\|_{L^2_{x,v}} &\lesssim \|g_1/\mu_0^{0+}\|_{L^6_{x,v}} \|g_2/\mu_0^{0+}\|_{L^3_{x,v}}, \\
\|\Gamma_\delta(g_1, g_2)\|_{L^2_{x,v}} &\lesssim \|\langle v\rangle^{3+} g_1\|_{L^\infty_{x,v}} \|\sqrt{\nu_\delta} g_2\|_{L^2_{x,v}}.
\end{aligned}$$

These yield

$$\begin{aligned}
&\|\Gamma_\delta(\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}, \mathbf{P}_\delta f_{R_{\varepsilon,\delta}})\|_{L^2_{x,v}} \lesssim \|P_\delta f_{R_{\varepsilon,\delta}}\|_{L^6_x} \|P_\delta f_{R_{\varepsilon,\delta}}\|_{L^3_x}, \\
&\|\Gamma_\delta(\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}, \mathbf{P}_\delta \partial_t f_{R_{\varepsilon,\delta}})\|_{L^2_{x,v}} \lesssim \|P_\delta f_{R_{\varepsilon,\delta}}\|_{L^6_x} \|P_\delta \partial_t f_{R_{\varepsilon,\delta}}\|_{L^3_x}, \\
&\|\Gamma_\delta(f_{R_{\varepsilon,\delta}}, (\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}})\|_{L^2_{x,v}} \lesssim \|h_{\varepsilon,\delta}\|_{L^\infty_{x,v}} \|(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}\|_{L^2_{x,v}}, \\
&\|\Gamma_\delta(f_{R_{\varepsilon,\delta}}, (\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}})\|_{L^2_{x,v}} \lesssim \|h_{\varepsilon,\delta}\|_{L^\infty_{x,v}} \|(\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}\|_{L^2_{x,v}}.
\end{aligned} \tag{78}$$

We also note that from $|A(t)|^2 = |A(0)|^2 + \frac{1}{2}\int_0^t \frac{d}{ds}|A(s)|^2 \mathrm{d}s \lesssim |A(0)|^2 + \int_0^t |A|^2 + \int_0^t |\partial_t A|^2$,

$$\|g\|_{L^\infty_t L^p} \lesssim \|g(0)\|_{L^p} + \|g\|_{L^2_t L^p} + \|\partial_t g\|_{L^2_t L^p}. \tag{79}$$

From (31), for $0 \leq t \leq \tau_\delta$

$$\begin{aligned}
\|\mathfrak{r}_{\varepsilon,\delta}(t)\|_{L^2_{x,v}} + \|\partial_t \mathfrak{r}_{\varepsilon,\delta}(t)\|_{L^2_{x,v}} &\lesssim \varepsilon^{\frac{1}{2}} + \varepsilon^{\frac{1}{2}}\sqrt{\mathcal{E}_{\varepsilon,\delta}(t)}, \\
\|\mathfrak{r}_{\varepsilon,\delta}(t)\|_{L^\infty_{x,v}} &\lesssim \|h_{\varepsilon,\delta}\|_{L^\infty_{t,x,v}} + \varepsilon^{\frac{1}{2}}.
\end{aligned} \tag{80}$$

Applying (78), (79), (80)–(61), (64), (73) we derive

$$
\begin{aligned}
&\|\varepsilon h_{\varepsilon,\delta}\|^2_{L^\infty_{t,x,v}} \\
&\lesssim \|\varepsilon h_{\varepsilon,\delta}(0)\|^2_{L^\infty_{x,v}} + \varepsilon\{\varepsilon^{\frac12} + \|\varepsilon h_{\varepsilon,\delta}\|_{L^\infty_{t,x,v}}\}^2\|\varepsilon h_{\varepsilon,\delta}\|^2_{L^\infty_{t,x,v}} + \varepsilon^4\delta^2\|\mathfrak{r}_{\varepsilon,\delta}\|^2_{L^\infty_{t,x,v}} \\
&\quad + \|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^\infty_t L^6_x} + \|\varepsilon^{-\frac12}(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(0)\|^2_{L^2_{x,v}} + \int_0^t \mathcal{D}_{\varepsilon,\delta} \\
&\le C_\infty\left\{\tilde{\mathcal{E}}_{\varepsilon,\delta}(0) + \|\varepsilon^{-\frac12}(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(0)\|^2_{L^2_{x,v}} + \varepsilon^4\delta^2 + \tilde{\mathcal{E}}_{\varepsilon,\delta}(t) + \int_0^t \mathcal{D}_{\varepsilon,\delta}\right\} \\
&\quad \text{for } 0 \le t \le T_\delta,
\end{aligned}
\tag{81}
$$

$$
\begin{aligned}
&\|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^\infty_t L^6_x} \\
&\lesssim \{1+\varepsilon^{\frac12}\|\varepsilon h_{\varepsilon,\delta}\|_{L^\infty_{t,x,v}}\}^2\left\{\|\varepsilon^{-\frac12}(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}(0)\|^2_{L^2_{x,v}} + \int_0^t \mathcal{D}_{\varepsilon,\delta}\right\} \\
&\quad + \mathcal{E}_{\varepsilon,\delta}(t) + \varepsilon\delta^2\|\mathfrak{r}_{\varepsilon,\delta}\|^2_{L^2_{x,v}} \\
&\quad + \{\|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}(0)\|^2_{L^3_x} + \|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x} + \|\varepsilon P_\delta \partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x}\}\|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^\infty_t L^6_x} \\
&\le C_6\left\{\mathcal{E}_{\varepsilon,\delta} + \delta^2\tilde{\mathcal{E}}_{\varepsilon,\delta}(t) + \int_0^t \mathcal{D}_{\varepsilon,\delta} + \delta^2\right\} \quad \text{for } 0 \le t \le T_\delta,
\end{aligned}
\tag{82}
$$

$$
\begin{aligned}
&\|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x} + \|\varepsilon P_\delta \partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x} \\
&\lesssim \{1+\varepsilon^{\frac12}\|\varepsilon h_{\varepsilon,\delta}\|_{L^\infty_{t,x,v}}\}^2\int_0^t \mathcal{D}_{\varepsilon,\delta} + \int_0^t \mathcal{E}_{\varepsilon,\delta} + \int_0^t \|\varepsilon^{\frac12}\delta\mathfrak{r}_{\varepsilon,\delta}\|^2_{L^2_{x,v}} + \|\varepsilon^{\frac12}\delta\partial_t\mathfrak{r}_{\varepsilon,\delta}\|^2_{L^2_{x,v}} \\
&\quad + \|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^\infty_t L^6_x}\{\|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x} + \|\varepsilon P_\delta \partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x}\} + (68) + (69) \\
&\le C_3\left\{\int_0^t \tilde{\mathcal{E}}_{\varepsilon,\delta} + \int_0^t \mathcal{D}_{\varepsilon,\delta} + \varepsilon t\right\} + (68) + (69) \quad \text{for } 0 \le t \le T_\delta.
\end{aligned}
\tag{83}
$$

From (39),

$$
\begin{aligned}
\text{RHS of (39)} &\lesssim \|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^\infty_t L^6_x}\{\|\varepsilon^{\frac12}P_\delta f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x} \\
&\quad + \|\varepsilon P_\delta \partial_t f_{R_{\varepsilon,\delta}}\|^2_{L^2_t L^3_x}\} + \varepsilon\|\varepsilon h_{\varepsilon,\delta}\|_{L^\infty_{t,x,v}}\int_0^t \mathcal{D}_{\varepsilon,\delta} \\
&\le C_1\left\{\frac{\delta^2}{a}\int_0^t \mathcal{D}_{\varepsilon,\delta} + \delta\varepsilon\int_0^t \mathcal{D}_{\varepsilon,\delta}\right\}.
\end{aligned}
\tag{84}
$$

Now we consider (43). From the decomposition $f_{R_{\varepsilon,\delta}} = \mathbf{P}_\delta f_{R_{\varepsilon,\delta}} + (\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}$ and $\partial_t f_{R_{\varepsilon,\delta}} = \mathbf{P}_\delta \partial_t f_{R_{\varepsilon,\delta}} + (\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}$ the terms containing $\mathbf{P}_\delta f_{R_{\varepsilon,\delta}}$ or $\mathbf{P}_\delta \partial_t f_{R_{\varepsilon,\delta}}$ in (43) are bounded as $\delta \|(\sigma_A, u_A, \theta_A)\|_{C^1_{t,x}} \int_0^t \mathcal{E}_{\varepsilon,\delta}$. For the other terms consist of only $(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}$ and $(\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}$ we split $\{v \in \mathbb{R}^3\} = \{|v| \le \varepsilon^{-\kappa}\} \cup \{|v| \ge \varepsilon^{-\kappa}\}$ as in [9] to derive

$$
\begin{aligned}
&\int_0^t \iint_{\Omega\times\mathbb{R}^3} \delta \langle v\rangle^3 \big\{ |(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}||(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}| \\
&\qquad\qquad + \varepsilon |(\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}||(\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}| \big\} \\
&\lesssim \int_0^t \iint_{\Omega\times\mathbb{R}^3} \mathbf{1}_{|v|\le\varepsilon^{-\kappa}} \delta\varepsilon \langle v\rangle^3 \{ |\varepsilon^{-\frac{1}{2}}(\mathbf{I}-\mathbf{P}_\delta) f_{R_{\varepsilon,\delta}}|^2 + |(\mathbf{I}-\mathbf{P}_\delta)\partial_t f_{R_{\varepsilon,\delta}}|^2 \} \\
&\quad + \int_0^t \iint_{\Omega\times\mathbb{R}^3} \mathbf{1}_{|v|\ge\varepsilon^{-\kappa}} \delta \frac{\langle v\rangle^3}{w} \|h_{\varepsilon,\delta}\|_{L^\infty_{x,v}} |f_{R_{\varepsilon,\delta}}| + \delta\varepsilon \frac{\langle v\rangle^6}{w} \|\tilde{h}_{\varepsilon,\delta}\|_{L^\infty_{x,v}} |\partial_t f_{R_{\varepsilon,\delta}}| \\
&\lesssim \delta \int_0^t \mathcal{D}_{\varepsilon,\delta} + \delta \int_0^t \tilde{\mathcal{E}}_{\varepsilon,\delta}
\end{aligned}
$$

for some $1/(\beta-6) \le \kappa \le 1/3$. Hence,

$$
(43) \le C_{\text{cub}} \left\{ \delta \int_0^t \tilde{\mathcal{E}}_{\varepsilon,\delta} + \delta \int_0^t \mathcal{D}_{\varepsilon,\delta} \right\}. \tag{85}
$$

Similar estimates hold for (42), (68), and (69).

Finally by adding $a \times \{(83) + (82) + (81)\}$ to (39)–(44) and using (84) and (85)

$$
\tilde{\mathcal{E}}_{\varepsilon,\delta}(t) + \int_0^t \tilde{\mathcal{D}}_{\varepsilon,\delta} \le \frac{\delta}{2} + C\delta \int_0^t \tilde{\mathcal{E}}_{\varepsilon,\delta},
$$

for some $0 < a, b \ll 1$. Then by the Gronwall's equality

$$
\tilde{\mathcal{E}}_{\varepsilon,\delta}(t) \le \frac{\delta}{2} e^{C\delta t}, \tag{86}
$$

and hence $\delta^{-1} \lesssim T_\delta$ and this finishes the argument. □

7 Solvability of (27)

By the Fredholm we have a unique solution $g \in (\mathrm{Ker}L_\delta)^\perp$ of $L_\delta g = h$ if and only if $\mathbf{P}_\delta h = 0$. We denote

$$L_\delta{}^{-1} h = g \in (\mathrm{Ker}L_\delta)^\perp. \tag{87}$$

From (18) and (22)

$$(\mathbf{I} - \mathbf{P}_\delta) f_{1\delta} = -L_\delta{}^{-1} \left(\frac{\partial_t \mu_\delta + v \cdot \nabla_x \mu_\delta}{\delta \sqrt{\mu_\delta}} \right). \tag{88}$$

We write and equivalent equation of (27) as

$$\begin{aligned} &\frac{\partial}{\partial t} P_j f_{1\delta} + P_j (v \cdot \nabla_x \mathbf{P}_\delta f_{1\delta}) + P_j \left(\frac{\partial_t \mu_\delta + v \cdot \nabla \mu_\delta}{2\sqrt{\mu_\delta}} \mathbf{P}_\delta f_{1\delta} \right) \\ &\qquad = -P_j (v \cdot \nabla_x (\mathbf{I} - \mathbf{P}_\delta) f_{1\delta}) - P_j \left(\frac{\partial_t \mu_\delta + v \cdot \nabla \mu_\delta}{2\sqrt{\mu_\delta}} (\mathbf{I} - \mathbf{P}_\delta) f_{1\delta} \right). \end{aligned} \tag{89}$$

Then

$$\begin{aligned} &P_j (v \cdot \nabla_x \mathbf{P}_\delta f_{1\delta}) \\ &= \sum_{m=1}^{3} \int_{\mathbb{R}^3} v_m \partial_m \Big(\sum_{i=0}^{4} (P_i f_{1\delta}) \varphi_i \sqrt{\mu_\delta} \Big) \varphi_j \sqrt{\mu_\delta} \mathrm{d}v \\ &= \sqrt{1+\delta\theta}\sqrt{1+\delta\sigma} \sum_{m=1}^{3} \sum_{i=0}^{4} \Big(\int_{\mathbb{R}^3} \varphi_m \varphi_i \varphi_j \mu_\delta \mathrm{d}v \Big) \partial_m (P_i f_{1\delta}) + \delta u \cdot \nabla (P_j f_{1\delta}) \\ &\quad + \sqrt{1+\delta\theta}\sqrt{1+\delta\sigma} \sum_{i=0}^{4} \Big(\sum_{m=1}^{3} \int_{\mathbb{R}^3} \varphi_m \partial_m (\varphi_i \sqrt{\mu_\delta}) \varphi_j \sqrt{\mu_\delta} \mathrm{d}v \Big) P_i f_{1\delta}. \end{aligned} \tag{90}$$

By direct computation,

$$\begin{aligned} &\int_{\mathbb{R}^3} \varphi_m \varphi_i \varphi_j \mu_\delta \mathrm{d}v \\ &= \begin{cases} \frac{1}{\sqrt{1+\delta\sigma}} & \text{for } \{i = 0 \,\&\, j = m = 1,2,3\} \text{ or } \{j = 0 \,\&\, i = m = 1,2,3\}, \\ \frac{\sqrt{6}}{\sqrt{1+\delta\sigma}} & \text{for } \{i = 4 \,\&\, j = m = 1,2,3\} \text{ or } \{j = 4 \,\&\, i = m = 1,2,3\}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Now (89) equals

$$\begin{aligned}\frac{\partial}{\partial t}\begin{bmatrix} P_0 f_{1\delta} \\ P_1 f_{1\delta} \\ P_2 f_{1\delta} \\ P_3 f_{1\delta} \\ P_4 f_{1\delta} \end{bmatrix} &= \sum_{i=1}^{3} \begin{bmatrix} \delta u_i & \sqrt{1+\delta\theta}\, e_i^T & 0 \\ \sqrt{1+\delta\theta}\, e_i & \delta u_i \mathbb{I}_{3\times 3} & \sqrt{6}\sqrt{1+\delta\theta}\, e_i \\ 0\sqrt{6}\sqrt{1+\delta\theta}\, e_i^T & \delta u_i & \end{bmatrix} \frac{\partial}{\partial x_i}\begin{bmatrix} P_0 f_{1\delta} \\ P_1 f_{1\delta} \\ P_2 f_{1\delta} \\ P_3 f_{1\delta} \\ P_4 f_{1\delta} \end{bmatrix} \\ &\quad + O(1)\begin{bmatrix} P_0 f_{1\delta} \\ P_1 f_{1\delta} \\ P_2 f_{1\delta} \\ P_3 f_{1\delta} \\ P_4 f_{1\delta} \end{bmatrix} \\ &= O((\mathbf{I}-\mathbf{P}_\delta) f_{1\delta}).\end{aligned}$$

For the existence we need some regularity on the initial datum such as

$$P_\delta f_{1\delta}|_{t=0} \in H^s(\Omega). \tag{91}$$

On the other hand the initial condition of $(\mathbf{I}-\mathbf{P}_\delta) f_{1\delta}$ has to be chosen as

$$(\mathbf{I}-\mathbf{P}_\delta) f_{1\delta}|_{t=0} = -L_\delta{}^{-1}\left(\frac{\partial_t \mu_\delta^0 + v\cdot\nabla_x \mu_\delta^0}{\delta\sqrt{\mu_\delta^0}}\right). \tag{92}$$

Acknowledgments JJ was supported in part by the NSF Grant DMS-2009458 and by the Simons Fellowship (grant # 616364). CK was supported in part by the NSF Grant DMS-1900923, and the Wisconsin Alumni Research Foundation.

References

1. Bardos, C., Golse, F., Levermore, D.: Fluid dynamic limits of kinetic equations. I Formal derivations. J. Statist. Phys. **63**, 323–344 (1991)
2. Bardos, C., Golse, F., Levermore, D.: Fluid dynamic limits of kinetic equations. II convergence proofs for the Boltzmann equation. Comm. Pure Appl. Math. **46**, 667–753 (1993)
3. Bardos, C., Golse, F., Levermore, D.: The acoustic limit for the Boltzmann equation. Arch. Rational. Mech. Anal. **153**, 177–204 (2000)
4. Bardos, C., Ukai, S.: The classical incompressible Navier-Stokes limit of the Boltzmann equation. Math. Models Methods Appl. Sci. **1**(2), 235–257 (1991)
5. A. de Masi, R. Esposito, J. L. Lebowitz, Incompressible Navier-Stokes and Euler Limits of the Boltzmann Equation, CPAM, 1189-1214, 1989
6. Golse, F., Levermore, C.D.: The Stokes-Fourier and acoustic limits for the Boltzmann equation. Comm. Pure and Appl. Math. **55**, 336–393 (2002)

7. Golse, F., Saint-Raymond, L.: The Navier-Stokes limit of the Boltzmann equation for bounded collision kernels. Invent. Math. **155**, 81–161 (2004)
8. Guo, Y.: Boltzmann diffusive limit beyond the Navier-Stokes approximation. Comm. Pure. Appl. Math. **59**, 626–687 (2006)
9. Guo, Y., Jang, J., Jiang, N.: Acoustic Limit for the Boltzmann equation in Optimal Scaling. Comm. Pure Appl. Math. **63**(3), 337–361 (2010)
10. Guo, Y., Jang, J., Jiang, N.: Local Hilbert expansion for the Boltzmann Equation. Kinet. Relat. Models **2**, 205–214 (2009)
11. Guo, Y., Huang, F., Wang, Y.: Hilbert expansion of the Boltzmann equation with specular boundary condition in half-space, preprint (2020)
12. Jang, J.: Vlasov-Maxwell-Boltzmann diffusive limit. Arch. Ration. Mech. Anal. **194**, 531–584 (2009)
13. Jang, J., Jiang, N.: Acoustic Limit of the Boltzmann equation: classical solutions. Discrete Contin. Dyn. Syst. **25**, 869–882 (2009)
14. Jiang, N., Levermore, C.D., Masmoudi, N.: Remarks on the acoustic limit for the Boltzmann equation. Comm. Partial Differential Equations **35**,1590–1609 (2010)
15. Kawashima, S., Matsumura, S., Nishida, T.: On the fluid-dynamical approximation to the Boltzmann equation at the level of the Navier-Stokes equation. Comm. Math. Phys. **70**(2), 97–124 (1979)
16. Lions, P.-L., Masmoudi, N.: From Boltzmann equations to incompressible fluid mechanics equation. I. Arch. Rational. Mech. Anal. **158**, 173–193 (2001)
17. Lions, P.-L., Masmoudi, N.: From Boltzmann equations to incompressible fluid mechanics equation. II. Arch. Rational. Mech. Anal. **158**, 195–211 (2001)
18. Nishida, T., Fluid dynamical limit of the nonlinear Boltzmann equation to the level of the compressible Euler equation. Comm. Math. Phys. **61**(2), 119–148 (1978)
19. Saint-Raymond, L.: Convergence of solutions to the Boltzmann equation in the incompressible Euler limit. Arch. Ration. Mech. Anal. **166**(1), 47–80 (2003)
20. Jang, J., Kim, C.: Incompressible Euler limit from Boltzmann equation with Diffuse Boundary Condition for Analytic data. Ann. PDE **7**, 22 (2021)
21. Esposito, R., Guo, Y., Kim, C., Marra, R.: Stationary solutions to the Boltzmann equation in the Hydrodynamic limit. Ann. PDE **4**, 1 (2018)
22. Caflisch, R.: The fluid dynamic limit of the nonlinear Boltzmann equation. Comm. Pure Appl. Math. **33**(5), 651–666 (1980)
23. Golse, F., Lions, P.L., Perthame, B., Sentis, R.: Regularity of the moments of the solution of a transport equation. J. Funct. Anal. **76**, 110–125 (1988)
24. Esposito, R., Guo, Y., Kim, C., Marra, R.: Non-Isothermal Boundary in the Boltzmann Theory and Fourier Law. Commun. Math. Phys **323**, 177–239 (2013)

Thermal Boundaries in Kinetic and Hydrodynamic Limits

Tomasz Komorowski and Stefano Olla

Abstract We investigate how a thermal boundary, modelled by a Langevin dynamics, affect the macroscopic evolution of the energy at different space-time scales.

1 Introduction

Chains of an-harmonic oscillators are commonly used models in non-equilibrium statistical mechanics, in particular to study macroscopic energy transport. To treat mathematically non-linear dynamics is a very hard task, even for a small non linear perturbation of the harmonic chain, see [15]. In the purely harmonic chain the energy transport is ballistic, see [14]. Numerical evidence, see e.g. [13], shows that non-linear perturbations can cause the transport in a one-dimensional system to become diffusive, in case of optical chains and superdiffusive for acoustic chains. Replacing the non-linearity by a stochastic exchange of momenta between neighboring particles makes the problem mathematically treatable (see the review [2] and the references therein). This stochastic exchange can be modelled in various ways: e.g. for each pair of the nearest neighbor particles the exchange of their momenta can occur independently at an exponential rate (which models their elastic collision). Otherwise, for each triple of consecutive particles, exchange of momenta can be performed in a continuous, diffusive fashion, so that its energy and momentum are preserved. The energy transport proven for such stochastic dynamics is qualitatively similar to the one expected in the case of the non-linear deterministic

T. Komorowski
Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland
e-mail: tkomorowski@impan.pl

S. Olla (✉)
CEREMADE, UMR CNRS, Université Paris Dauphine, PSL Research University, Paris, France

Institute Universitaire de France (IUF), Paris, France

GSSI, L'Aquila, Italie
e-mail: olla@ceremade.dauphine.fr

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_11

dynamics. In particular, for a one-dimensional acoustic chain it could be proved that the macroscopic thermal energy density evolves according to a fractional heat equation corresponding to the fractional laplacian $(-\Delta)^{3/4}$, see [6].

In the recent years we have been interested in the macroscopic effects of a heat bath in contact with the chain at a point.

In Sect. 3 we consider first a purely harmonic chain in contact with a stochastic Langevin thermostat at temperature T, see (6) and (5). In the absence of a thermostat, its dynamics is completely integrable and the energy of each frequency (the Fourier mode k) is conserved. Rescaling space-time by the same parameter, the energy of mode k, localized by Wigner distribution $W(t, y, k)$, evolves according to a linear transport equation

$$\partial_t W + \bar{\omega}'(k)\partial_y W = 0,$$

with velocity $\bar{\omega}'(k) = \omega'(k)/2\pi$, where $\omega(k)$ is the dispersion relation of the harmonic chain. We can interpret $W(t, y, k)$ as the energy density of *phonons* of mode k at time t in the position y. The presence of a Langevin thermostat results in the emergence of a boundary (interface) condition at $y = 0$ ([9]):

$$\begin{aligned} W(t, 0^+, k) &= p_-(k)W(t, 0^+, -k) + p_+(k)W(t, 0^-, k) + \mathfrak{g}(k)T, \qquad \text{for } k > 0 \\ W(t, 0^-, k) &= p_-(k)W(t, 0^-, -k) + p_+(k)W(t, 0^+, k) + \mathfrak{g}(k)T, \qquad \text{for } k < 0. \end{aligned} \tag{1}$$

The coefficients appearing in the boundary condition correspond to probabilities of the phonon transmission $p_+(k)$, reflection $p_-(k)$ and absorption $\mathfrak{g}(k)$. These parameters are non-negative and satisfy $p_+(k) + p_-(k) + \mathfrak{g}(k) = 1$. In addition, $T\mathfrak{g}(k)$ is the intensity of the phonon creations. The transmission, reflection and absorption parameters depend in a quite complicated way on the dispersion relation $\omega(\cdot)$ and the strength of the thermostat $\gamma > 0$ (cf. (39), (42), and (50)). Some of their properties and an explicit calculation for the nearest-neighbor interactions are presented in Appendix 1, the results contained there are original and are not part of [9]).

It is somewhat surprising that an incident *phonon* of mode k after scattering, if not absorbed, can produce only an identical transmitted phonon, or a reflected phonon of mode $-k$ (at least for a unimodal dispersion relation). This stands in contrast with what takes place at the microscopic scale. Then, an incident wave of frequency k scatters and produces waves of all possible frequencies. In the macroscopic limit, all frequencies produced by the scattering on the thermostat, except those corresponding to $\pm k$, are damped by oscillations.

In [7] we have considered the same problem after adding the bulk noise that conserves energy and momentum, see Sect. 4. The noise is properly rescaled in such a way that finite total amount of momentum is exchanged locally in the macroscopic unit time (in analogy to a kinetic limit). The effect of the bulk noise is to add a macroscopic scattering term to the transport equation:

$$\partial_t W + \bar{\omega}'(k)\partial_y W = \gamma_0 \int_{\mathbb{T}} R(k, k') \left(W(k') - W(k)\right) dk', \tag{2}$$

i.e. a phonon of mode k changes to the one of mode k', with intensity $\gamma_0 R(k, k')$, given by (58). The case when no heat bath is present, has been studied in [1]. In [7] we prove that the heat bath adds to (2) the same boundary condition (1), see Sect. 4.

Since in (2) the energies of different modes are mixed up by the bulk scattering, we can further rescale space-time in this equation in order to obtain an autonomous equation for the evolution of the total energy. Without the thermal bath this case has been studied in [5] and the following results have been obtained:

- For *optical chains* the velocity of the phonon behaves like $\bar{\omega}'(k) \sim k$ for small k, while for the total scattering rate $R(k) = \int R(k, k')dk' \sim k^2$. This means that phonons of low frequency rarely scatter, but move very slowly so they have the time to diffuse under an appropriate (diffusive) scaling. The phonons corresponding to other modes also behave diffusively at the respective scales. Consequently, in the optical chain, under diffusive space-time scaling, all modes homogenize equally, contributing to the macroscopic evolution of the energy $e(t, y)$, i.e. $W(t/\delta^2, y/\delta, k) \longrightarrow_{\delta \to 0} e(t, y)$, that follows the linear heat equation $(\partial_t - D\partial_y^2)e(t, y) = 0$ with an explicitly given $D > 0$, see (86).
- In *acoustic chains*, the bulk scattering rate is the same, but $\bar{\omega}'(k) \sim O(1)$ for small k. Consequently low frequency phonons scatter rarely but they still move with velocities of order 1, and the resulting macroscopic limit is superdiffusive. In particular, in this case the low frequency modes are responsible for the macroscopic transport of the energy. The respective superdiffusive space-time scaling limit $W(t/\delta^{3/2}, y/\delta, k) \longrightarrow_{\delta \to 0} e(t, y)$, described by the solution of a fractional heat equation $(\partial_t - \hat{c}|\partial_y^2|^{3/4})e(t, y) = 0$, with $\hat{c} > 0$ given by (94).

The thermal bath adds a boundary condition at $y = 0$ to the diffusive, or superdiffusive equations described above. More precisely the situation is as follows.

- In the optical chain we obtain a Dirichlet boundary condition $e(t, 0) = T$ for the respective heat equation (see (85)): phonons trajectories behave like Brownian motions, and since they can cross the boundary infinitely many times, they are absorbed almost surely. In effect, there is no energy *transfer* through the boundary at the macroscopic scale. This is proven in [3] using analytic techniques, see Sect. 5.1.
- In the acoustic chain, the long wave phonons, are responsible for the macroscopic energy transport. Their trajectories behave in the limit like superdiffusive symmetric, 3/2–stable Levy processes: they can jump over the boundary on the macroscopic scale and there is a positive probability of survival, i.e. of energy macroscopic transmission across the thermal boundary. Since the absorption probability $\mathfrak{g}(k)$ remains strictly positive as $k \to 0$ (as we prove here in Appendix 1, at least for the nearest neighbor acoustic chain, see (119)), the thermal boundary affects the transport. The macroscopic energy evolution is given by a fractional heat equation with boundary defined by (93) and (87). This is proven in [10] using probabilistic techniques, see Sect. 5.2.

In Sect. 2 we review the results of [9] for the harmonic chain with the thermostat attached at a point. Since the calculations for the transmission and reflection scattering of [9] are quite complex, we present their somewhat simplified version (conveying nevertheless their gist) in Appendix 2. We hope that the outline would help the reader to understand how the macroscopic scattering emerges.

In Sect. 3 we review the results of [7] in the presence of the conservative bulk noise. Section 4 contains the review of the diffusive and superdiffusive limits proven in [3, 10]. In Sect. 5 we mention some open problems, in particular the question of the direct hydrodynamic limit, without passing through the kinetic limit, in the spirit of [6].

Appendix 1 contains some original results that are not present in the discussed articles, concerning the properties of the scattering coefficients and their behaviour for $k \to 0$.

2 Notation

Given $a > 0$ by $\mathbb{T}_a$ we denote the torus of size $a > 0$, i.e. the interval $[-a/2, a/2]$ with identified endpoints. When $a = 1$ we shall write $\mathbb{T} := \mathbb{T}_1$ for the unit torus. Let $\mathbb{T}_\pm := [k \in \mathbb{T} : 1/2 > \pm k > 0]$. We also let $\mathbb{R}_+ := (0, +\infty)$, $\mathbb{R}_* := \mathbb{R} \setminus \{0\}$ and $\mathbb{T}_* := \mathbb{T} \setminus \{0\}$.

By $\ell^p(\mathbb{Z})$, $L^p(\mathbb{T})$, where $p \geq 1$, we denote the spaces of all complex valued sequences $(f_x)_{x\in\mathbb{Z}}$ and functions $f : \mathbb{T} \to \mathbb{C}$ that are summable with p-th power, respectively. The Fourier transform of $(f_x)_{x\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$ and the inverse Fourier transform of $\hat{f} \in L^2(\mathbb{T})$ are given by

$$\hat{f}(k) = \sum_{x\in\mathbb{Z}} f_x \exp\{-2\pi i x k\}, \quad f_x = \int_{\mathbb{T}} \hat{f}(k) \exp\{2\pi i x k\} dk, \quad x \in \mathbb{Z}, \quad k \in \mathbb{T}. \tag{3}$$

We use the notation

$$(f \star g)_y = \sum_{y'\in\mathbb{Z}} f_{y-y'} g_{y'}$$

for the convolution of two sequences $(f_x)_{x\in\mathbb{Z}}, (g_x)_{x\in\mathbb{Z}}$ that belong to appropriate spaces $\ell^p(\mathbb{Z})$. In most cases we shall assume that one of the sequences rapidly decays, while the other belongs to $\ell^2(\mathbb{Z})$.

For a function $G : \mathbb{R} \times \mathbb{T} \to \mathbb{C}$ that is either L^1, or L^2-summable, we denote by $\hat{G} : \mathbb{R} \times \mathbb{T} \to \mathbb{C}$ its Fourier transform, in the first variable, defined as

$$\hat{G}(\eta, k) := \int_{\mathbb{R}} e^{-2\pi i \eta x} G(x, k) dx, \quad (\eta, k) \in \mathbb{R} \times \mathbb{T}. \tag{4}$$

Denote by $C_0(\mathbb{R} \times \mathbb{T})$ the class of functions G that are continuous and satisfy $\lim_{|y|\to+\infty} \sup_{k\in\mathbb{T}} |G(y,k)| = 0$.

3 Harmonic Chain in Contact with a Langevin Thermostat

We consider the evolution of an infinite particle system governed by the Hamiltonian

$$\mathcal{H}(\mathfrak{p},\mathfrak{q}) := \frac{1}{2}\sum_{y\in\mathbb{Z}} \mathfrak{p}_y^2 + \frac{1}{2}\sum_{y,y'\in\mathbb{Z}} \alpha_{y-y'}\mathfrak{q}_y\mathfrak{q}_{y'}. \tag{5}$$

Here, the particle label is $y \in \mathbb{Z}$, $(\mathfrak{q}_y, \mathfrak{p}_y)$ is the position and momentum of the y's particle, respectively, and $(\mathfrak{q},\mathfrak{p}) = \{(\mathfrak{q}_y,\ \mathfrak{p}_y),\ y \in \mathbb{Z}\}$ denotes the entire configuration. The coupling coefficients α_y are assumed to have exponential decay and chosen positive definite such that the energy is positive. We couple the particle with label 0 to a Langevin thermostat at temperature T. Then the evolution of the system can be described using the stochastic differential equations:

$$\begin{aligned} &\dot{\mathfrak{q}}_y(t) = \mathfrak{p}_y(t), \\ &d\mathfrak{p}_y(t) = -(\alpha \star \mathfrak{q}(t))_y dt + \big(-\gamma\mathfrak{p}_0(t)dt + \sqrt{2\gamma T}dw(t)\big)\delta_{0,y}, \quad y \in \mathbb{Z}. \end{aligned} \tag{6}$$

Here, $\{w(t),\ t \geq 0\}$ is a standard Wiener process, while $\gamma > 0$ is a coupling parameter with the thermostat.

Assumptions on the Dispersion Relation and Its Basic Properties
We assume (cf [1]) that the coupling constants $(\alpha_x)_{x\in\mathbb{Z}}$ satisfy the following:

(a1) they are real valued and there exists $C > 0$ such that $|\alpha_x| \leq Ce^{-|x|/C}$ for all $x \in \mathbb{Z}$,
(a2) $\hat{\alpha}(k) = \sum_{x\in\mathbb{Z}} \alpha_x e^{-2\pi ikx}$ is also real valued and $\hat{\alpha}(k) > 0$ for $k \neq 0$ and in case $\hat{\alpha}(0) = 0$ we have $\hat{\alpha}''(0) > 0$.

The above conditions imply that both functions $x \mapsto \alpha_x$ and $k \mapsto \hat{\alpha}(k)$ are even. In addition, $\hat{\alpha} \in C^\infty(\mathbb{T})$ and in case $\hat{\alpha}(0) = 0$ we have $\hat{\alpha}(k) = k^2\phi(k^2)$ for some strictly positive $\phi \in C^\infty(\mathbb{T})$. The dispersion relation $\omega : \mathbb{T} \to \bar{\mathbb{R}}_+$, given by

$$\omega(k) := \sqrt{\hat{\alpha}(k)}, \quad k \in \mathbb{T}. \tag{7}$$

is obviously also even. Throughout the paper it is assumed to be unimodal, i.e. increasing on $\bar{\mathbb{T}}_+$ and then, in consequence, decreasing on $\bar{\mathbb{T}}_-$. Its unique minimum and maximum are attained at $k = 0$, $k = 1/2$, respectively. They are denoted by $\omega_{\min} \geq 0$ and $\omega_{\max}$, correspondingly. Denote the two branches of its inverse by $\omega_\pm : [\omega_{\min}, \omega_{\max}] \to \bar{\mathbb{T}}_\pm$.

In order to avoid technical problems with the definition of the dynamics, we assume that the initial conditions are random but with finite energy: $\mathcal{H}(\mathfrak{p}, \mathfrak{q}) < \infty$. This property will be conserved in time. For such configurations we can define the complex wave function

$$\psi_y(t) := (\tilde{\omega} \star \mathfrak{q}(t))_y + i\mathfrak{p}_y(t) \tag{8}$$

where $\left(\tilde{\omega}_y\right)_{y\in\mathbb{Z}}$ is the inverse Fourier transform of the dispersion relation. We have $\mathcal{H}(\mathfrak{p}(t), \mathfrak{q}(t)) = \sum_y |\psi_y(t)|^2$.

The Fourier transform of the wave function is given by

$$\hat{\psi}(t, k) := \omega(k)\hat{\mathfrak{q}}(t, k) + i\hat{\mathfrak{p}}(t, k), \quad k \in \mathbb{T}, \tag{9}$$

so that

$$\hat{\mathfrak{p}}\,(t, k) = \frac{1}{2i}[\hat{\psi}(t, k) - \hat{\psi}^*(t, -k)], \quad \mathfrak{p}_0(t) = \int_{\mathbb{T}} \mathrm{Im}\,\hat{\psi}(t, k)dk.$$

Using (6), it is easy to verify that the wave function evolves according to

$$d\hat{\psi}(t, k) = \big(- i\omega(k)\hat{\psi}(t, k) - i\gamma\mathfrak{p}_0(t)\big)dt + i\sqrt{2\gamma T}dw(t). \tag{10}$$

Introducing a (small) parameter $\varepsilon \in (0, 1)$, we wish to study the behaviour of the distribution of the energy at a large space-time scale, i.e. for the wave function $\psi_{[x/\varepsilon]}(t/\varepsilon)$, $(t, x) \in \mathbb{R}_+ \times \mathbb{R}$, when $\varepsilon \to 0$. In this scaling limit we would like to maintain each particle contribution to the total energy to be of order $O(1)$, on the average, and therefore keep the total energy of the chain to be order ε^{-1}. For this reason, we choose random initial data that is distributed by probability measures μ_ε defined on the phase space $(\mathfrak{p}, \mathfrak{q})$, in such a way that

$$\sup_{\varepsilon\in(0,1)} \varepsilon\langle\mathcal{H}(\mathfrak{p}, \mathfrak{q})\rangle_{\mu_\varepsilon} = \sup_{\varepsilon\in(0,1)} \sum_{y\in\mathbb{Z}} \varepsilon\langle|\psi_y|^2\rangle_{\mu_\varepsilon} = \sup_{\varepsilon\in(0,1)} \varepsilon\langle\|\hat{\psi}\|^2_{L^2(\mathbb{T})}\rangle_{\mu_\varepsilon} < \infty. \tag{11}$$

The symbol $\langle\cdot\rangle_{\mu_\varepsilon}$ denotes, as usual, the average with respect to measure μ_ε. To simplify our calculations we will also assume that

$$\langle\hat{\psi}(k)\hat{\psi}(\ell)\rangle_{\mu_\varepsilon} = 0, \quad k, \ell \in \mathbb{T}. \tag{12}$$

This condition is easily satisfied by local Gibbs measures like

$$\prod_{y\in\mathbb{Z}} \frac{e^{-\beta_y^\varepsilon|\psi_y|^2/2}}{Z_{\beta_y^\varepsilon}} d\psi_y \tag{13}$$

for a proper choice of temperature profiles $(\beta_y^\varepsilon)^{-1} > 0$, decaying fast enough to 0, as $|y| \to +\infty$. Here $Z_{\beta_y^\varepsilon}$ is the normalizing constant.

Wigner Distributions

Wigner distributions provide an effective tool to localize in space energy per frequency, separating microscopic from macroscopic scale. The (averaged) Wigner distribution (or Wigner transform) is defined by its action on a test function $G \in \mathcal{S}(\mathbb{R} \times \mathbb{T})$ as

$$\langle G, W^{(\varepsilon)}(t)\rangle := \frac{\varepsilon}{2} \sum_{y,y' \in \mathbb{Z}} \int_{\mathbb{T}} e^{2\pi i k(y'-y)} \mathbb{E}_\varepsilon \left[\psi_y \left(\frac{t}{\varepsilon}\right) (\psi_{y'})^* \left(\frac{t}{\varepsilon}\right) \right] G^*\Big(\varepsilon \frac{y+y'}{2}, k\Big) dk. \tag{14}$$

The Fourier transform of the Wigner distribution, or the Fourier-Wigner function is defined as

$$\widehat{W}_\varepsilon(t,\eta,k) := \frac{\varepsilon}{2} \mathbb{E}_\varepsilon \left[\hat{\psi}^* \left(\frac{t}{\varepsilon}, k - \frac{\varepsilon\eta}{2}\right) \hat{\psi} \left(\frac{t}{\varepsilon}, k + \frac{\varepsilon\eta}{2}\right) \right], \quad (t,\eta,k) \in [0,\infty) \times \mathbb{T}_{2/\varepsilon} \times \mathbb{T}, \tag{15}$$

so that

$$\langle G, W^{(\varepsilon)}(t)\rangle = \int_{\mathbb{T}\times\mathbb{R}} \widehat{W}_\varepsilon(t,\eta,k) \hat{G}^*(\eta,k) d\eta dk, \quad G \in \mathcal{S}(\mathbb{R} \times \mathbb{T}). \tag{16}$$

Taking $G(x,k) := G(x)$ in (14) we obtain

$$\langle G, W^{(\varepsilon)}(t)\rangle = \frac{\varepsilon}{2} \sum_{y \in \mathbb{Z}} \mathbb{E}_\varepsilon \left[\left| \psi_y \left(\frac{t}{\varepsilon}\right) \right|^2 \right] G(\varepsilon y). \tag{17}$$

In what follows we assume that the initial data, after averaging, leads to a sufficiently fast decaying (in η) Fourier-Wigner function. More precisely, we suppose that there exist $C, \kappa > 0$ such that

$$|\widehat{W}_\varepsilon(0,\eta,k)| \le \frac{C}{(1+\eta^2)^{3/2+\kappa}}, \quad (\eta,k) \in \mathbb{T}_{2/\varepsilon} \times \mathbb{T},\ \varepsilon \in (0,1). \tag{18}$$

In addition, we assume that there exists a distribution $W_0 \in \mathcal{S}'(\mathbb{R} \times \mathbb{T})$ such that for any $G \in \mathcal{S}(\mathbb{R} \times \mathbb{T})$

$$\lim_{\varepsilon \to 0+} \langle G, W^{(\varepsilon)}(0)\rangle = \langle G, W_0\rangle. \tag{19}$$

Note that, thanks to (18), distribution W_0 is in fact a function that belongs to $C_0(\mathbb{R} \times \mathbb{T}) \cap L^2(\mathbb{R} \times \mathbb{T})$.

3.1 The Thermostat Free Case: $\gamma = 0$

If the thermostat is not present ($\gamma = 0$), the equation of motion (10) can be explicitly solved and the soution is $\hat{\psi}(t,k) = \hat{\psi}(k)e^{-i\omega(k)t}$. Defining

$$\delta_\varepsilon\omega(k,\eta) := \frac{1}{\varepsilon}\left[\omega\left(k+\frac{\varepsilon\eta}{2}\right) - \omega\left(k-\frac{\varepsilon\eta}{2}\right)\right], \tag{20}$$

we can compute explicitly the Wigner transform:

$$\widehat{W}_\varepsilon(t,\eta,k) \;=\; e^{-i\delta_\varepsilon\omega(k,\eta)t/\varepsilon}\widehat{W}_\varepsilon(0,\eta,k) \underset{\varepsilon\to 0}{\longrightarrow} e^{-i\omega'(k)\eta t}\widehat{W}_0(\eta,k), \tag{21}$$

assuming the corresponding convergence at initial time, see (19). The inverse Fourier transform gives

$$W(t,y,k) \;=\; W_0(y-\bar{\omega}'(k)t,k), \tag{22}$$

where $\bar{\omega}'(k) := \omega'(k)/(2\pi)$, i.e. it solves the simple linear transport equation

$$\partial_t W(t,y,k) + \bar{\omega}'(k)\partial_y W(t,y,k) = 0, \quad W(0,y,k) = W_0(y,k). \tag{23}$$

We can view this equation as the evolution of the density in independent particles (phonons), labelled by the frequency mode $k \in \mathbb{T}$, and moving with velocity $\bar{\omega}'(k)$.

3.2 The Evolution with the Langevin Thermostat: $\gamma > 0$

We use the mild formulation of (10):

$$\hat{\psi}(t,k) = e^{-i\omega(k)t}\hat{\psi}(0,k) - i\gamma\int_0^t e^{-i\omega(k)(t-s)}\mathfrak{p}_0(s)ds + i\sqrt{2\gamma T}\int_0^t e^{-i\omega(k)(t-s)}dw(t). \tag{24}$$

Integrating both sides in the k-variable and taking the imaginary part in both sides, we obtain a closed equation for $\mathfrak{p}_0(t)$:

$$\mathfrak{p}_0(t) = \mathfrak{p}_0^0(t) - \gamma\int_0^t J(t-s)\mathfrak{p}_0(s)ds + \sqrt{2\gamma T}\int_0^t J(t-s)dw(s), \tag{25}$$

where

$$J(t) = \int_{\mathbb{T}} \cos(\omega(k)t)\, dk, \tag{26}$$

and

$$\mathfrak{p}_0^0(t) = \int_{\mathbb{T}} \mathrm{Im}\left(\hat{\psi}(0,k)e^{-i\omega(k)t}\right) dk, \tag{27}$$

is the momentum at $y = 0$ for the free evolution with $\gamma = 0$ (without the thermostat).

Taking the Laplace transform

$$\tilde{\mathfrak{p}}_0(\lambda) = \int_0^{+\infty} e^{-\lambda t}\mathfrak{p}_0(t)dt, \quad \mathrm{Re}\,\lambda > 0,$$

in (25) we obtain

$$\tilde{\mathfrak{p}}_0(\lambda) = \tilde{g}(\lambda)\tilde{\mathfrak{p}}_0^0(\lambda) + \sqrt{2\gamma T}\tilde{g}(\lambda)\tilde{J}(\lambda)\tilde{w}(\lambda). \tag{28}$$

Here, $\tilde{g}(\lambda)$ is given by

$$\tilde{g}(\lambda) := (1 + \gamma\tilde{J}(\lambda))^{-1}. \tag{29}$$

and

$$\tilde{J}(\lambda) := \int_0^{\infty} e^{-\lambda t}J(t)dt = \int_{\mathbb{T}} \frac{\lambda}{\lambda^2 + \omega^2(k)}dk, \quad \mathrm{Re}\,\lambda > 0. \tag{30}$$

We will show below that $\tilde{g}(\lambda)$ is the Laplace transform of a signed locally finite measure $g(d\tau)$. Then, the term $(\lambda + i\omega(k))^{-1}\tilde{g}(\lambda)\tilde{\mathfrak{p}}_0^0(\lambda)$, that appears in (33), is the Laplace transform of the convolution

$$\int_0^t \phi(t-s,k)\mathfrak{p}_0^0(s)ds, \tag{31}$$

where

$$\phi(t,k) = \int_0^t e^{-i\omega(k)(t-\tau)}g(d\tau). \tag{32}$$

Next, taking the Laplace transform of both sides of (24) and using (28), we arrive at an explicit formula for the Fourier-Laplace transform of $\psi_y(t)$:

$$\begin{aligned}\tilde{\psi}(\lambda,k) &= \frac{\hat{\psi}(0,k) - i\gamma\tilde{\mathfrak{p}}_0(\lambda) + i\sqrt{2\gamma T}\tilde{w}(\lambda)}{\lambda + i\omega(k)} \\ &= \frac{\hat{\psi}(0,k) - i\gamma\tilde{g}(\lambda)(\tilde{\mathfrak{p}}_0^0(\lambda) + \sqrt{2\gamma T}\tilde{J}(\lambda)\tilde{w}(\lambda)) + i\sqrt{2\gamma T}\tilde{w}(\lambda)}{\lambda + i\omega(k)} \\ &= \frac{\hat{\psi}(0,k) - i\gamma\tilde{g}(\lambda)\tilde{\mathfrak{p}}_0^0(\lambda) + i\tilde{g}(\lambda)\sqrt{2\gamma T}\tilde{w}(\lambda)}{\lambda + i\omega(k)}.\end{aligned} \tag{33}$$

The Laplace inversion of (33) yields an explicit expression for $\hat{\psi}(t,k)$:

$$\begin{aligned}\hat{\psi}(t,k) =& e^{-i\omega(k)t}\hat{\psi}(0,k) - i\gamma\int_0^t \phi(t-s,k)\mathfrak{p}_0^0(s)\,ds \\ &+ i\sqrt{2\gamma T}\int_0^t \phi(t-s,k)\,dw(s).\end{aligned} \tag{34}$$

3.3 Phonon Creation by the Heat Bath

Since the contribution to the energy given by the thermal term and the initial energy are completely separate, we can assume first that $\widehat{W}_0 = 0$. In this case $\hat{\psi}(0,k) = 0$ and (34) reduces to a stochastic convolution:

$$\hat{\psi}(t,k) = i\sqrt{2\gamma T}\int_0^t \phi(t-s,k)\,dw(s). \tag{35}$$

To shorten the notation, denote

$$\tilde{\phi}(t,k) = \int_0^t e^{i\omega(k)\tau} g(d\tau) = e^{i\omega(k)t}\phi(t,k),$$

We can compute directly the Fourier-Wigner function

$$\widehat{W}_\varepsilon(t,\eta,k) = \gamma T\int_0^t e^{-i\delta_\varepsilon\omega(k,\eta)s}\tilde{\phi}\left(s/\varepsilon, k+\frac{\varepsilon\eta}{2}\right)\tilde{\phi}^*\left(s/\varepsilon, k-\frac{\varepsilon\eta}{2}\right)ds.$$

Taking its Laplace transform we obtain

$$\begin{aligned}\widehat{w}_\varepsilon(\lambda,\eta,k) &= \gamma T\int_0^\infty dte^{-\lambda t}\int_0^t dse^{-i\delta_\varepsilon\omega(k,\eta)s}\tilde{\phi}\left(\varepsilon^{-1}s,k+\frac{\varepsilon\eta}{2}\right)\tilde{\phi}^*\left(\varepsilon^{-1}s,k-\frac{\varepsilon\eta}{2}\right)\\ &= \frac{\gamma T}{\lambda}\int_0^\infty dse^{-(\lambda+i\delta_\varepsilon\omega(k,\eta))s}\tilde{\phi}\left(\varepsilon^{-1}s,k+\frac{\varepsilon\eta}{2}\right)\tilde{\phi}^*\left(\varepsilon^{-1}s,k-\frac{\varepsilon\eta}{2}\right).\end{aligned} \tag{36}$$

Using the inverse Laplace formula for the product of functions we obtain, for any $c>0$,

$$\begin{aligned}\widehat{w}_\varepsilon(\lambda,\eta,k) &= \frac{\gamma T}{\lambda}\frac{1}{2\pi i}\lim_{\ell\to\infty}\int_{c-i\ell}^{c+i\ell}\left\{\sigma\left(\lambda+i\delta_\varepsilon\omega(k,\eta)-\sigma\right)\right\}^{-1}\\ &\times\tilde{g}\left(\varepsilon\sigma-i\omega(k+\frac{\varepsilon\eta}{2})\right)\tilde{g}^*\left(\varepsilon(\lambda+i\delta_\varepsilon\omega(k,\eta)-\sigma)-i\omega(k-\frac{\varepsilon\eta}{2})\right)d\sigma.\end{aligned} \tag{37}$$

Since $\tilde{g}$ is bounded and $\mathrm{Re}\lambda>0$, we can take the limit as $\varepsilon\to 0$, obtaining

$$\widehat{w}(\lambda,\eta,k)=\frac{\gamma T|\nu(k)|^2}{\lambda\left(\lambda+i\omega'(k)\eta\right)}, \tag{38}$$

where

$$\nu(k):=\lim_{\varepsilon\to 0}\tilde{g}(\varepsilon-i\omega(k)). \tag{39}$$

The limit in (39) is well defined everywhere, see Appendix 1 for details.

The inverse Laplace transform of (38) gives

$$\widehat{W}(t,\eta,k)=\frac{1-e^{-i\omega'(k)\eta t}}{i\omega'(k)\eta}\gamma T|\nu(k)|^2. \tag{40}$$

Performing the inverse Fourier transform, according to (30), we obtain

$$W(t,y,k)=T\mathfrak{g}(k)1_{[[0,\bar{\omega}'(k)t]]}(y) \tag{41}$$

where $\bar{\omega}(k)=\omega(k)/2\pi$,

$$\mathfrak{g}(k):=\frac{\gamma|\nu(k)|^2}{|\bar{\omega}'(k)|} \tag{42}$$

and

$$[[0,a]] := \begin{cases} [0,a], & \text{if } a > 0 \\ [a,0], & \text{if } a < 0. \end{cases}$$

We can interpret (41) as the energy density of k-phonons that are created at the interface $y = 0$ by the heat bath with intensity $T\mathfrak{g}(k)$ and then move with velocity $\bar{\omega}'(k)$. The Wigner function $W(t, y, k)$ can be viewed as a formal solution of

$$\partial_t W(t,y,k) + \bar{\omega}'(k)\partial_y W(t,y,k) = |\bar{\omega}'(k)| T\mathfrak{g}(k)\delta(y), \qquad W(0,y,k) = 0. \tag{43}$$

3.4 Phonon Scattering and Absortion by the Heat Bath

The scattering of incoming waves can be studied at temperature $T = 0$, by looking at the deterministic equation

$$\hat{\psi}(t,k) = e^{-i\omega(k)t}\hat{\psi}(0,k) - i\gamma \int_0^t \phi(t-s,k)\mathfrak{p}_0^0(s)\, ds, \tag{44}$$

Proceeding along the lines of the calculation of the previous section we obtain

$$\widehat{W}_\varepsilon(t,\eta,k) \;=\; \widehat{W}_\varepsilon^0(t,\eta,k) + \widehat{W}_\varepsilon^1(t,\eta,k) + \widehat{W}_\varepsilon^2(t,\eta,k), \tag{45}$$

where $\widehat{W}_\varepsilon(t,\eta,k)$ is given by (15),

$$\begin{aligned}
\widehat{W}_\varepsilon^0(t,\eta,k) &:= e^{-i\delta_\varepsilon\omega(k,\eta)t/\varepsilon}\widehat{W}_\varepsilon(0,\eta,k),\\
\widehat{W}_\varepsilon^1(t,\eta,k) &:= -i\frac{\varepsilon\gamma}{2}\int_0^{t/\varepsilon}\Big\{\mathbb{E}_\varepsilon\Big[\hat{\psi}\left(0,k-\frac{\varepsilon\eta}{2}\right)^*\mathfrak{p}_0^0(s)\Big]e^{i\omega(k-\frac{\varepsilon\eta}{2})\frac{t}{\varepsilon}}\phi\left(\frac{t}{\varepsilon}-s,k+\frac{\varepsilon\eta}{2}\right)\\
&\qquad - \mathbb{E}_\varepsilon\Big[\hat{\psi}\left(0,k+\frac{\varepsilon\eta}{2}\right)\mathfrak{p}_0^0(s)\Big]e^{-i\omega(k+\frac{\varepsilon\eta}{2})\frac{t}{\varepsilon}}\phi\left(\frac{t}{\varepsilon}-s,k-\frac{\varepsilon\eta}{2}\right)^*\Big\}\, ds,\\
\widehat{W}_\varepsilon^2(t,\eta,k) &:= \frac{\varepsilon\gamma^2}{2}\int_0^{t/\varepsilon} ds_1\int_0^{t/\varepsilon} ds_2\mathbb{E}_\varepsilon\left[\mathfrak{p}_0^0(s_1)\mathfrak{p}_0^0(s_2)\right]\\
&\qquad\times\phi\left(\frac{t}{\varepsilon}-s_1,k-\frac{\varepsilon\eta}{2}\right)^*\phi\left(\frac{t}{\varepsilon}-s_2,k-\frac{\varepsilon\eta}{2}\right).
\end{aligned} \tag{46}$$

The limit behavior of $\widehat{W}_\varepsilon^0(t,\eta,k)$ is already described by (21). The calculations for the other two terms $\widehat{W}_\varepsilon^1$ and $\widehat{W}_\varepsilon^2$ are more involved. Their outline is presented in

Appendix 2 below. We have

$$\lim_{\varepsilon\to 0+}\widehat{W}^1_\varepsilon(t,\eta,k)=-\gamma\mathrm{Re}\nu(k)e^{-i\omega'(k)t}\int_{\mathbb{R}}\frac{1-e^{-i\omega'(k)(\eta'-\eta)t}}{i\omega'(k)(\eta'-\eta)}\widehat{W}(0,\eta',k)d\eta' \tag{47}$$

and

$$\lim_{\varepsilon\to 0+}\widehat{W}^2_\varepsilon(t,\eta,k)=\frac{|\nu(k)|^2}{4|\bar\omega'(k)|}\sum_{\iota=\pm}\int_{\mathbb{R}}\frac{d\eta}{\lambda+i\omega'(k)\eta}\int_{\mathbb{R}}\frac{\widehat{W}(0,\eta',\iota k)d\eta'}{\lambda+i\omega'(\iota k)\eta'}. \tag{48}$$

Putting together the limits of the solutions for $T\geq 0$ and $\widehat{W}_0=0$, given by (41), and the solution of the deterministic equation for $T=0$ and a non-vanishing $\widehat{W}_0$, obtained by taking the sum of the limits of $\widehat{W}^j_\varepsilon(t,\eta,k)$, $j=0,1,2$, we conclude the formula for the limit as $\varepsilon\to 0$, of Wigner function $\widehat{W}_\varepsilon(t,\eta,k)$, see (15), for $\hat\phi(t,k)$, given by (34), equals:

$$\begin{aligned}&W(t,y,k)=1_{[[0,\bar\omega'(k)t]]^c}(y)W(0,y-\bar\omega'(k)t,k)\\&+p_+(k)1_{[[0,\bar\omega'(k)t]]}(y)W(0,y-\bar\omega'(k)t,k)\\&+p_-(k)1_{[[0,\bar\omega'(k)t]]}(y)W(0,-y+\bar\omega'(k)t,-k)+T\mathfrak{g}(k)1_{[[0,\bar\omega'(k)t]]}(y),\end{aligned} \tag{49}$$

where the coefficient $\mathfrak{g}(k)$ is given by (42) and

$$p_+(k):=\left|1-\frac{\gamma\nu(k)}{2|\bar\omega'(k)|}\right|^2,\qquad p_-(k):=\left(\frac{\gamma|\nu(k)|}{2|\bar\omega'(k)|}\right)^2. \tag{50}$$

By a direct inspection we can verify that $W(t,y,k)$, given by (49), solves the transport equation

$$\partial_t W(t,y,k)+\bar\omega'(k)\partial_y W(t,y,k)=0,\qquad y\neq 0, \tag{51}$$

with the transmission/reflection and phonon creation boundary condition at $y=0$:

$$\begin{aligned}&W(t,0^+,k)=p_-(k)W(t,0^+,-k)+p_+(k)W(t,0^-,k)+\mathfrak{g}(k)T,\qquad \text{for } k\in\mathbb{T}_+\\&W(t,0^-,k)=p_-(k)W(t,0^-,-k)+p_+(k)W(t,0^+,k)+\mathfrak{g}(k)T,\qquad \text{for } k\in\mathbb{T}_-.\end{aligned} \tag{52}$$

In Appendix 1 below we show that

$$p_+(k)+p_-(k)+\mathfrak{g}(k)=1. \tag{53}$$

Coefficients $p_+(k)$ and $p_-(k)$ can be interpreted therefore as the probabilities of phonon transmission and reflection, respectively. Since $p_+(k) + p_-(k) = 1 - \mathfrak{g}(k)$, the coefficient $\mathfrak{g}(k)$ is the phonon absorption probability at the interface.

4 Harmonic Chain with Bulk Conservative Noise in Contact with Langevin Thermostat

4.1 The Model and the Statement of the Result

In [7] we consider a stochastically perturbed chain of harmonic oscillators thermostatted at a fixed temperature $T \geq 0$ at $x = 0$. Its dynamics is described by the system of Itô stochastic differential equations

$$\begin{aligned} &d\mathfrak{q}_x(t) = \mathfrak{p}_x(t)dt, \qquad x \in \mathbb{Z}, \\ &d\mathfrak{p}_x(t) = \Big[-(\alpha \star \mathfrak{q}(t))_x - \frac{\varepsilon\gamma_0}{2}(\theta \star \mathfrak{p}(t))_x\Big]dt \\ &+ \sqrt{\varepsilon\gamma_0} \sum_{k=-1,0,1} (Y_{x+k}\, \mathfrak{p}_x(t))dw_{x+k}(t) + \Big(-\gamma\mathfrak{p}_0(t)dt + \sqrt{2\gamma T}dw(t)\Big)\delta_{0,x}. \end{aligned} \tag{54}$$

Here the coupling constants $(\alpha_x)_{x\in\mathbb{Z}}$ are as in (5),

$$Y_x := (\mathfrak{p}_x - \mathfrak{p}_{x+1})\partial_{\mathfrak{p}_{x-1}} + (\mathfrak{p}_{x+1} - \mathfrak{p}_{x-1})\partial_{\mathfrak{p}_x} + (\mathfrak{p}_{x-1} - \mathfrak{p}_x)\partial_{\mathfrak{p}_{x+1}} \tag{55}$$

and $(w_x(t))_{t\geq 0}$, $x \in \mathbb{Z}$ with $(w(t))_{t\geq 0}$, are i.i.d. one dimensional independent Brownian motions. In addition,

$$\theta_x = \Delta\theta^{(0)}_x := \theta^{(0)}_{x+1} + \theta^{(0)}_{x-1} - 2\theta^{(0)}_x$$

with

$$\theta^{(0)}_x = \begin{cases} -4, & x = 0 \\ -1, & x = \pm 1 \\ 0, & \text{if otherwise.} \end{cases}$$

Parameters $\varepsilon\gamma_0 > 0$, γ describe the strength of the inter-particle and thermostat noises, respectively. In what follows we shall assume that $\varepsilon > 0$ is small, that corresponds to the low density hypothesis that results in atoms suffering finitely many "collisions" in a macroscopic unit of time (the Boltzmann-Grad limit). Although the noise considered here is continuous we believe that the results extend

to other type of conservative noises, such as e.g. Poisson exchanges of velocities between nearest neighbor particles.

Since the vector field Y_x is orthogonal both to a sphere $\mathfrak{p}_{x-1}^2+\mathfrak{p}_x^2+\mathfrak{p}_{x+1}^2 \equiv \text{const}$ and plane $\mathfrak{p}_{x-1}+\mathfrak{p}_x+\mathfrak{p}_{x+1} \equiv \text{const}$, the inter-particle noise conserves locally the kinetic energy and momentum. Because these conservation laws are common also for chaotic hamiltonian system, this model has been used to understand energy transport in presence of momentum conservation, see [2] and references there.

The case without the Langevin thermostat, i.e. with $\gamma = 0$, was studied in [1], where it is proved that

$$W_\varepsilon(t, y, k) \underset{\varepsilon\to 0}{\longrightarrow} W(t, y, k) \tag{56}$$

where $W(t, y, k)$ is the solution of the transport equation

$$\partial_t W(t, y, k) + \bar{\omega}'(k)\partial_y W(t, y, k) = \gamma_0 \int_{\mathbb{T}} R(k, k') \big(W(t, y, k') - W(t, y, k)\big)\, dk', \tag{57}$$

where

$$R(k, k') = 32 \sin^2(\pi k) \sin^2(\pi k') \left\{\sin^2(\pi k) \cos^2(\pi k') + \sin^2(\pi k') \cos^2(\pi k)\right\}. \tag{58}$$

We have therefore

$$R(k) = \int_{\mathbb{T}} R(k, k')dk' = 4 \sin^2(\pi k)\big(1 + 3 \cos^2(\pi k)\big). \tag{59}$$

In [7] we have proved the following result.

Theorem 1 *Suppose that $\gamma_0, \gamma > 0$ and the initial data satisfies (18), Then,*

$$\begin{aligned} \lim_{\varepsilon\to 0+} \int_0^{+\infty} dt \iint_{\mathbb{T}_{2/\varepsilon}\times\mathbb{T}} W_\varepsilon(t, y, k)G(t, y, k)dydk \\ = \int_0^{+\infty} dt \iint_{\mathbb{R}\times\mathbb{T}} W(t, y, k)G(t, y, k)dydk \end{aligned} \tag{60}$$

for any $G \in C_0^\infty([0, +\infty) \times \mathbb{R} \times \mathbb{T})$, where the limiting Wigner $W(t, y, k)$ function satisfies (57) for $(t, y, k) \in \mathbb{R}_+ \times \mathbb{R}_* \times \mathbb{T}_*$, with the boundary conditions (52), at the interface $y = 0$.

4.2 A Sketch of the Proof of Theorem 1

Consider the wave function $\psi(t)$ that corresponds to the dynamics (54) via (8) and $\hat{\psi}(t)$ its Fourier transform. In contrast with the situation described in Sect. 3.2 (the case $\gamma_0 = 0$) we no longer have an explicit expression for the solution of the equation for $\hat{\psi}(t)$, see (24), so we cannot proceed by a direct calculation of the Wigner distributions as in Sect. 3 and Appendix 2.

In order to close the dynamics of the Fourier-Wigner function, we shall need all the components of the full covariance tensor of the Fourier transform of the wave field. Define therefore the Wigner distribution tensor $\mathbf{W}_\varepsilon(t)$, as a 2×2-matrix tensor, whose entries are distributions, given by their respective Fourier transforms

$$\widehat{\mathbf{W}}_\varepsilon(t,\eta,k) := \begin{bmatrix} \widehat{W}_{\varepsilon,+}(t,\eta,k) & \widehat{Y}_{\varepsilon,+}(t,\eta,k) \\ \widehat{Y}_{\varepsilon,-}(t,\eta,k) & \widehat{W}_{\varepsilon,-}(t,\eta,k) \end{bmatrix}, \quad (\eta,k)\in \mathbb{T}_{2/\varepsilon}\times\mathbb{T}, \tag{61}$$

with

$$\begin{aligned}
\widehat{W}_{\varepsilon,+}(t,\eta,k) &:= \widehat{W}_\varepsilon(t,\eta,k) = \frac{\varepsilon}{2}\mathbb{E}_\varepsilon\left[\hat{\psi}\left(t/\varepsilon, k+\frac{\varepsilon\eta}{2}\right)\hat{\psi}^\star\left(t/\varepsilon, k-\frac{\varepsilon\eta}{2}\right)\right],\\
\widehat{Y}_{\varepsilon,+}(t,\eta,k) &:= \frac{\varepsilon}{2}\mathbb{E}_\varepsilon\left[\hat{\psi}\left(t/\varepsilon, k+\frac{\varepsilon\eta}{2}\right)\hat{\psi}\left(t/\varepsilon, -k+\frac{\varepsilon\eta}{2}\right)\right],\\
\widehat{Y}_{\varepsilon,-}(t,\eta,k) &:= \widehat{Y}^\star_{\varepsilon,+}(t,-\eta,k), \quad \widehat{W}_{\varepsilon,-}(t,\eta,k) := \widehat{W}_{\varepsilon,+}(t,\eta,-k).
\end{aligned}$$

By a direct calculation we show that the following energy bound is satisfied, see Proposition 2.1 of [7]

$$\sup_{\varepsilon\in(0,1]}\frac{\varepsilon}{2}\mathbb{E}_\varepsilon\|\hat{\psi}(t/\varepsilon)\|^2_{L^2(\mathbb{T})} \le \sup_{\varepsilon\in(0,1]}\frac{\varepsilon}{2}\mathbb{E}_\varepsilon\|\hat{\psi}(0)\|^2_{L^2(\mathbb{T})} + \gamma T t, \quad t\ge 0. \tag{62}$$

The above estimate implies in particular that

$$\sup_{\varepsilon\in(0,1]}\|W_{\varepsilon,+}\|_{L^\infty([0,\tau];\mathcal{A}')} < +\infty, \quad \text{for any } \tau>0, \tag{63}$$

where $\mathcal{A}'$ is the dual to $\mathcal{A}$—the Banach space obtained by the completion of $\mathcal{S}(\mathbb{R}\times\mathbb{T})$ in the norm

$$\|G\|_{\mathcal{A}} := \int_{\mathbb{R}}\sup_{k\in\mathbb{T}}|\widehat{G}(\eta,k)|d\eta, \quad G\in\mathcal{S}(\mathbb{R}\times\mathbb{T}). \tag{64}$$

Similar estimates hold also for the remaining entries of $\mathbf{W}_\varepsilon(t)$.

In consequence $(\mathbf{W}_\varepsilon(\cdot))$ is sequentially $\star$-weakly compact in $L^\infty_{\mathrm{loc}}([0,+\infty),\mathcal{A}')$ and the problem of proving its $\star$-weak convergence reduces to the limit identification.

4.2.1 The Case of Zero Temperature at the Thermostat

Suppose first that $T = 0$. We can treat then the microscopic dynamics (54) as a small (stochastic) perturbation of the purely deterministic dynamics when all the terms containing the noises $(w_x(t))$ and $(w(t))$ are omitted. Denote by $\hat{\phi}^{(\varepsilon)}(t,k)$ the Fourier transform of the wave function corresponding to the latter dynamics, cf (9). We consider then the respective Wigner distribution tensor $\mathbf{W}_\varepsilon^{\mathrm{un}}(t,y,k)$ whose Fourier transform is given by an analogue of (61), where the wave function of the "true" (perturbed) dynamics $\hat{\psi}(t,k)$ is replaced by $\hat{\phi}(t,k)$, which corresponds to the deterministic dynamics:

$$\begin{aligned} d\mathfrak{q}_x(t) &= \mathfrak{p}_x(t)dt, \qquad x \in \mathbb{Z}, \\ d\mathfrak{p}_x(t) &= \Big[-(\alpha \star \mathfrak{q}(t))_x - \frac{\varepsilon\gamma_0}{2}(\theta \star \mathfrak{p}(t))_x\Big]dt - \gamma\mathfrak{p}_0(t)\delta_{0,x}dt. \end{aligned} \tag{65}$$

Denote by $\mathcal{L}_{2,\varepsilon}$ the Hilbert space made of the 2×2 matrix valued distributions on $\mathbb{R}\times\mathbb{T}$, such that the Fourier transforms of their entries belong to $L^2(\mathbb{T}_{2/\varepsilon}\times\mathbb{T})$. The Hilbert norm on $\mathcal{L}_{2,\varepsilon}$ is defined in an obvious way using the L^2 norms of the Fourier transforms.

Using the equations for the microscopic dynamics of $\hat{\phi}(t,k)$ we conclude that the tensor $\mathbf{W}_\varepsilon^{\mathrm{un}}(t)$ can be described by an $\mathcal{L}_{2,\varepsilon}$ strongly continuous semigroup $\Big(\mathfrak{W}_\varepsilon^{\mathrm{un}}(t)\Big)$, i.e.

$$\mathbf{W}_\varepsilon^{\mathrm{un}}(t) = \mathfrak{W}_\varepsilon^{\mathrm{un}}(t)\Big(\mathbf{W}_\varepsilon^{\mathrm{un}}(0)\Big), \quad t \geq 0. \tag{66}$$

Using a very similar argument to that used in the case of $\gamma_0 = 0$ (remember no noise is present in the unperturbed dynamics) we can prove, see Theorem 5.7 of [7], that

Theorem 2 *Under the assumptions on the initial data made in* (18) *and* (19)*, we have*

$$\lim_{\varepsilon\to 0+}\langle G, \mathbf{W}_\varepsilon^{\mathrm{un}}(t)\rangle = \langle G, \mathbf{W}^{\mathrm{un}}(t)\rangle, \quad t \geq 0,\ G \in \mathcal{S}(\mathbb{R}\times\mathbb{T})$$

where

$$\mathbf{W}^{\mathrm{un}}(t,y,k) = \begin{bmatrix} W_+^{\mathrm{un}}(t,y,k) & 0 \\ 0 & W_+^{\mathrm{un}}(t,y,-k) \end{bmatrix}, \quad (y,k) \in \mathbb{R}\times\mathbb{T},$$

and

$$\partial_t W^{\mathrm{un}}(t,y,k) + \bar{\omega}'(k)\partial_y W^{\mathrm{un}}(t,y,k) = -\gamma_0 R(k) W^{\mathrm{un}}(t,y,k), \quad (t,y,k) \in \mathbb{R}_+\times\mathbb{R}_*\times\mathbb{T}_*, \tag{67}$$

with the interface conditions (52) *for* $T = 0$.

Similarly to (51) Eq. (67) can be solved explicitly and we obtain $W^{\mathrm{un}}(t) = \mathfrak{W}^{\mathrm{un}}_t(W^{\mathrm{un}}(0))$, where $W^{\mathrm{un}}(t,y,k) = e^{-\gamma_0 R(k)t}\tilde{W}^{\mathrm{un}}(t,y,k)$ and $\tilde{W}^{\mathrm{un}}(t,y,k)$ is given by (49). Consider a semigroup defined by

$$\mathfrak{W}^{\mathrm{un}}_t(W^{\mathrm{un}}(0))\,(y,k) := W^{\mathrm{un}}(t,y,k). \tag{68}$$

One can show that $\left(\mathfrak{W}^{\mathrm{un}}_t\right)_{t\geq 0}$ forms a strongly continuous semigroup of contractions on any $L^p(\mathbb{R}\times\mathbb{T})$, $1\leq p<+\infty$.

We can use the semigroup $\mathfrak{W}^{\mathrm{un}}_\varepsilon(t)$ to write a Duhamel type equation for

$$\mathbf{w}_\varepsilon(\lambda) = \begin{bmatrix} \widehat{w}_{\varepsilon,+}(\lambda,\eta,k) & \widehat{y}_{\varepsilon,+}(\lambda,\eta,k) \\ \widehat{y}_{\varepsilon,-}(\lambda,\eta,k) & \widehat{w}_{\varepsilon,-}(\lambda,\eta,k) \end{bmatrix} = \int_0^{+\infty} e^{-\lambda t}\mathbf{W}_\varepsilon(t)dt$$

the Laplace transform of the Wigner tensor of (61) defined for $\mathrm{Re}\,\lambda > \lambda_0$ and some sufficiently large $\lambda_0 > 0$. It reads

$$\mathbf{w}_\varepsilon(\lambda) = \tilde{\mathfrak{W}}^{\mathrm{un}}_\varepsilon(\lambda)\mathbf{W}_\varepsilon(0) + \frac{\gamma_0}{2}\tilde{\mathfrak{W}}^{\mathrm{un}}_\varepsilon(\lambda)\,\mathbf{v}_\varepsilon(\lambda), \quad \mathrm{Re}\,\lambda > \lambda_0. \tag{69}$$

Here

$$\mathfrak{W}^{\mathrm{un}}_\varepsilon(\lambda) := \int_0^{+\infty} e^{-\lambda t}\mathfrak{W}^{\mathrm{un}}_\varepsilon(t)dt, \tag{70}$$

and $\mathbf{v}_\varepsilon(\lambda) := \mathfrak{R}_\varepsilon\mathbf{w}_\varepsilon(\lambda)$. The operator $\mathfrak{R}_\varepsilon$ acts on 2×2 matrix valued $\mathbf{w}$ whose entries belong to $\mathcal{L}_{2,\varepsilon}$ and whose Fourier transform (in y) is given by

$$\widehat{\mathbf{w}}(\eta,k) := \begin{bmatrix} \widehat{w}_+(\eta,k) & \widehat{y}_+(\eta,k) \\ \widehat{y}_-(\eta,k) & \widehat{w}_{\varepsilon,-}(\eta,k) \end{bmatrix}, \quad (\eta,k)\in\mathbb{T}_{2/\varepsilon}\times\mathbb{T}$$

as follows. The Fourier transform $\widehat{\mathfrak{R}_\varepsilon\mathbf{w}}$ have entries of the form

$$\pm\int_{\mathbb{T}} r_0(k,k',\varepsilon\eta)\left[\widehat{w}_{\varepsilon,+}(\eta,k') + \widehat{w}_{\varepsilon,-}(\eta,k') - \widehat{y}_{\varepsilon,+}(\eta,k') - \widehat{y}_{\varepsilon,-}(\eta,k')\right]dk'.$$

Here $r_0(k,k',\varepsilon\eta)$ is a scattering kernel that satisfies $R(k,k') = r_0(k,k',0) + r_0(k,-k',0)$, with $R(k,k')$ given by (58). Suppose now that we test both sides of (69) against a 2×2-matrix valued smooth function $\mathbf{G}$ whose entries have compactly supported Fourier transforms in the y variable, say in the interval $[-K,K]$ for some $K>0$. Denote by $\left(\mathfrak{W}^{\mathrm{un}}_\varepsilon(\lambda)\right)^*$ and $\mathcal{R}^*_\varepsilon$ the adjoints of the respective operators in $\mathcal{L}_{2,\varepsilon}$.

We already know that from any sequence $\left(\mathbf{w}_{\varepsilon_n}(\lambda)\right)$, where $\varepsilon_n\to 0+$, we can choose a subsequence, that will be denoted by the same symbol, converging $\star$-weakly in $\mathcal{A}'$ to some $\mathbf{w}(\lambda)$. Using Theorem 2 and the strong convergence of the

sequence $1_{[-K,K]}(\eta)\mathcal{R}^*_{\varepsilon_n}\left(\mathfrak{W}^{\rm un}_{\varepsilon_n}(\lambda)\right)^*\mathbf{G}$, $n \to +\infty$ in $L^2(\mathbb{R}\times\mathbb{T})$ we can prove the following, see Theorem 5.7 of [7].

Theorem 3 *Suppose that* $\mathbf{W}$ *is the* $\star$*-weak limit of* $\left(\mathbf{W}_{\varepsilon_n}\right)$ *in* $\left(L^1([0,+\infty);\mathcal{A})\right)'$ *for some sequence* $\varepsilon_n \to 0+$. *Then, it has to be of the form*

$$\mathbf{W}(t,y,k) = \begin{bmatrix} W(t,y,k) & 0 \\ 0 & W(t,y,-k) \end{bmatrix}, \quad (y,k) \in L^2(\mathbb{R}\times\mathbb{T}), \tag{71}$$

where $W(t,y,k)$ *satisfies the equation*

$$W(t) = \mathfrak{W}^{\rm un}_t\Big(W(0)\Big) + \gamma_0 \int_0^t \mathfrak{W}^{\rm un}_{t-s}\Big(\mathcal{R}W_s\Big)ds, \tag{72}$$

and $\mathcal{R} : L^2(\mathbb{R}\times\mathbb{T}) \to L^2(\mathbb{R}\times\mathbb{T})$ *is given by*

$$\mathcal{R}F(y,k) := \int_{\mathbb{T}} R(k,k')F\left(y,k'\right)dk', \quad (y,k) \in \mathbb{R}\times\mathbb{T}, \quad F \in L^2(\mathbb{R}\times\mathbb{T}). \tag{73}$$

The convergence claimed in Theorem 1 is then a direct consequence of Theorems 2 and 3.

It turns out that the microscopic evolution of the Wigner transform given by (61) allows us to define a strongly continuous semigroup on $L^2(\mathbb{T}_{2/\varepsilon}\times\mathbb{T})$ by letting $\mathfrak{W}_\varepsilon(t)\Big(\mathbf{W}_\varepsilon(0)\Big) := \mathbf{W}_\varepsilon(0)$. The norms of the semigroups $L^2(\mathbb{T}_{2/\varepsilon}\times\mathbb{T})$ remain bounded with $\varepsilon \in (0,1]$, see Corollary 4.2 of [7]. Using Theorems 2 and 3 we can show therefore that for any $\mathbf{G} \in L^1([0,+\infty),\mathcal{A})$ we have

$$\lim_{\varepsilon\to 0+}\int_0^{+\infty} \langle\, \mathfrak{W}_\varepsilon(t)\mathbf{W}_\varepsilon(0), \mathbf{G}(t)\rangle dt = \int_0^{+\infty} \langle\, \mathfrak{W}_t\mathbf{W}(0), \mathbf{G}(t)\rangle dt, \tag{74}$$

where $\mathfrak{W}_t\mathbf{W}(0) := \mathbf{W}(t)$ is given by (71).

4.2.2 The Case of Positive Temperature at the Thermostat

Finally we consider the case $T > 0$. Suppose that $\chi \in C^\infty_c(\mathbb{R})$ is an arbitrary real valued, even function satisfying

$$\chi(y) = \begin{cases} 1, & \text{for } |y| \le 1/2, \\ 0, & \text{for } |y| \ge 1, \\ \text{belongs to } [0,1], & \text{if otherwise.} \end{cases} \tag{75}$$

Then its Fourier transform $\widehat{\chi} \in \mathcal{S}(\mathbb{R})$. Let $\widehat{\chi}_\varepsilon \in C^\infty(\mathbb{T}_{2/\varepsilon})$ be given by

$$\widehat{\chi}_\varepsilon(\eta) := \sum_{n\in\mathbb{Z}} \widehat{\chi}\left(\eta + \frac{2n}{\varepsilon}\right), \quad \eta \in \mathbb{T}_{2/\varepsilon}.$$

and

$$\widehat{\mathbf{V}}_\varepsilon(t,\eta,k) := \widehat{\mathbf{W}}_\varepsilon(t,\eta,k) - T\widehat{\chi}_\varepsilon(\eta)\mathbf{I}_2,$$

where $\mathbf{I}_2$ is the 2×2 identity matrix. In fact, $\mathbf{V}_\varepsilon(t)$ is a solution of the equation

$$\mathbf{V}_\varepsilon(t) = \mathfrak{W}_\varepsilon(t)\mathbf{V}_\varepsilon(0) + \int_0^t \mathfrak{W}_\varepsilon(s)\Big(\mathbf{F}_\varepsilon\Big)ds, \tag{76}$$

where

$$\widehat{\mathbf{F}}_\varepsilon(\eta,k) := -\frac{iT\widehat{\chi}_\varepsilon(\eta)}{\varepsilon}\left[\omega\left(k+\frac{\varepsilon\eta}{2}\right) - \omega\left(k-\frac{\varepsilon\eta}{2}\right)\right]\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}.$$

Using the convergence of (74) we conclude that

$$\lim_{\varepsilon\to 0+}\int_0^{+\infty}\langle \mathbf{V}_\varepsilon(t), \mathbf{G}(t)\rangle dt = \int_0^{+\infty}\langle \mathbf{V}(t), \mathbf{G}(t)\rangle dt, \tag{77}$$

where

$$\mathbf{V}(t,y,k) = \begin{bmatrix} V(t,y,k) & 0\\ 0 & V(t,y,-k)\end{bmatrix}$$

and

$$V(t,y,k) := W(t,y,k) + T\int_0^t \mathfrak{W}_s\Big(\bar{\omega}'(k)\chi'(y)\Big)ds. \tag{78}$$

We can identify therefore $W(t,y,k)$ with the solution of the equation

$$W(t,y,k) := \mathfrak{W}_t(\widetilde{W}_0)(y,k) + \int_0^t \mathfrak{W}_s(F)(y,k)ds + T\chi(y), \quad (t,y,k)\in\bar{\mathbb{R}}_+\times\mathbb{R}\times\mathbb{T}. \tag{79}$$

Here

$$F(y,k) := -T\bar{\omega}'(k)\chi'(y), \quad \widetilde{W}_0(y,k) := W_0(y,k) - T\chi(y). \tag{80}$$

This ends the proof of Theorem 1 for an arbitrary $T \geq 0$.

5 Diffusive and Superdiffusive Limit from the Kinetic Equation with Boundary Thermostat

We are now interested in the space-time rescaling of the solution of the Eq. (57) with the boundary condition (52). We should distinguish here two cases:

- the *optical chain*, when $\omega'(k) \sim k$ for small k,
- the *acoustic chain*, when $\omega(k) \sim |k|$ for small k.

In the optical chain, the long-wave phonons (corresponding to small k) have a small velocity, consequently even if the bulk scattering rate is small ($R(k) \sim k^2$), they still have time to diffuse. In fact, all other phonons (i.e. those corresponding to other k) have their non-trivial contribution to the diffusive limit.

In the acoustic chain the long-wave phonons move with the speed that is bounded away from 0 and rarely scatter. Therefore, they are responsible for a superdiffusion of the Levy type arising in the macroscopic limit. In the superdiffusive time-scale all other phonons (corresponding to non-vanishng k) do not yet move, their contribution to the asymptotic limit is therefore negligible.

5.1 The Optical Chain: Diffusive Behavior

This case was studied in [3]. The diffusive rescaling of the solution of (57) is defined by

$$W^\delta(t, y, k) = W(t/\delta^2, y/\delta, k), \tag{81}$$

with an initial condition that varying in the *macroscopic* space scale

$$W^\delta(0, y, k) = W_0(y, k). \tag{82}$$

We assume here that $W_0(y, k) = T + \tilde{W}_0(y, k)$, with $\tilde{W}_0 \in L^2(\mathbb{R} \times \mathbb{T})$. This rescaled solution solves

$$\partial_t W^\delta(t, y, k) + \frac{1}{\delta}\tilde{\omega}'(k)\partial_y W^\delta(t, y, k) = \frac{\gamma_0}{\delta^2}\int_{\mathbb{T}} R(k, k')\left(W^\delta(t, y, k') - W^\delta(t, y, k)\right)dk', \tag{83}$$

with the boundary condition (52) in $y = 0$.

In [3] it is proven that, for any test function $\varphi(t, y, k) \in C_0^\infty([0, +\infty) \times \mathbb{R} \times \mathbb{T})$,

$$\lim_{\delta \to 0} \int_0^{+\infty} dt \iint_{\mathbb{R}\times\mathbb{T}} W^\delta(t, y, k)\varphi(t, y, k)\, dy\, dk = \int_0^{+\infty} dt \iint_{\mathbb{R}\times\mathbb{T}} \rho(t, y)\varphi(t, y, k)\, dy\, dk, \tag{84}$$

where $\rho(t, y)$ is the solution of the heat equation

$$\begin{aligned} \partial_t \rho(t, y) &= D\partial_y^2 \rho(t, y), \qquad y \neq 0, \\ \rho(t, 0) &= T, \qquad \forall t > 0, \\ \rho(0, y) &= \rho_0(y) := \int_{\mathbb{T}} W_0(y, k) dk. \end{aligned} \tag{85}$$

The diffusion coefficient is given by

$$D := \frac{1}{\gamma_0} \int_{\mathbb{T}} \frac{\bar{\omega}'(k)^2}{R(k)} dk. \tag{86}$$

Notice that, under the condition of the optical dispersion relation, $D < +\infty$. The proof in [3] follows a classical Hilbert expansion method, with a modification needed to account for the boundary condition.

Intuitively, the result can be explained in the following way: phonons of all frequencies behave diffusively, under the scaling they converge to Brownian motions with diffusion D, that has continuous path. As they get close to the thermostat boundary, they cross it many times till they get absorbed with probability 1 in the macroscopic time scale. Consequently there is no (macroscopic) trasmission of energy from one side to the other. Phonons are created with intensity T, and this explain the value at the boundary $y = 0$.

5.2 *The Acoustic Chain: Superdiffusive Behavior*

This limit was studied in [10], while the case without thermostat had been previously considered in [5]. In a one dimensional acoustic chain, long wave phonons (small k) move with finite velocities but still scatter very rarely. Consequently these longwaves phonons on the microscopic scale move ballistically with some rare scattering of their velocities. Under the superdiffusive rescaling $\delta^{-3/2}t, \delta^{-1}y$ they converge to corresponding Levy processes, generated by the fractional laplacian $-|\Delta|^{3/4}$. The effect of the thermal boundary is more complex than in the diffusive case, as now the phonons have a positive probability to cross the boundary without absorption and jump at a macroscopic distance on the other side. This causes a particular boundary condition for the fractional laplacian at the interface $y = 0$, that we explain below. Let us define the fractional laplacian $-|\Delta|^{3/4}$, admitting an

interface value T, with absorption $\mathfrak{g}_0$, transmission p_+ and reflection p_-, as the L^2 closure of the singular integral operator

$$\begin{aligned}\Lambda_{3/4}F(y) =& \text{p.v.} \int_{yy'>0} q(y-y')[F(y')-F(y)]dy' \\ &+ \mathfrak{g}_0[T-F(y)] \int_{yy'<0} q(y-y')dy' \\ &+ p_- \int_{yy'<0} q(y-y')[F(-y')-F(y)]dy' \\ &+ p_+ \int_{yy'<0} q(y-y')[F(y')-F(y)]dy', \quad y\neq 0,\ F\in C_0^\infty(\mathbb{R}),\end{aligned} \tag{87}$$

where, cf [12, Theorem 1.1 e)],

$$q(y)=\frac{c_{3/4}}{|y|^{5/2}}, \qquad c_{3/4}=\frac{2^{3/2}\Gamma(5/4)}{\sqrt{\pi}|\Gamma(-3/4)|}=\frac{3}{2^{5/2}\sqrt{\pi}}. \tag{88}$$

The first integral appearing in the right hand side of (87) is understood in the principal value (p.v.) sense. The choice of constant $c_{3/4}$ is made in such a way that the "free" fractional laplacian, defined by the kernel $q(\cdot)$, coincides with the definition using the "usual" Fourier symbol, see [12, Theorem 1.1 a)]. To define $\Lambda_{3/4}F(0)$, note that, due to the fact that $\mathfrak{g}_0>0$, the finiteness of the second integral forces the condition $F(0)=T$ on any function belonging to the domain of the generator. We can define $\Lambda_{3/4}F(0)$ using (87) for any continuous function that satisfies $F(0)=T$, for which the integrals appearing in the right hand side (without the principal value) converge.

Notice that in the case without thermal interface, $\mathfrak{g}_0=0$, $p_-=0$, $p_+=1$, and we recover the usual "free" fractional laplacian on the real line. The absorption, transmission and reflection coefficients that arise here are given by

$$\mathfrak{g}_0=\lim_{k\to 0}\mathfrak{g}(k), \quad p_\pm=\lim_{k\to 0}p_\pm(k). \tag{89}$$

For the nearest neighbor acoustic chain, with the dispersion relation $\omega(k):=\omega_a|\sin(\pi k)|$ (cf (111)) it turns out that, see (119),

$$p_+=\left(\frac{\omega_a}{\omega_a+\gamma}\right)^2, \qquad p_-(k):=\left(\frac{\gamma}{\omega_a+\gamma}\right)^2 \qquad \mathfrak{g}_0=\frac{2\gamma\omega_a}{(\omega_a+\gamma)^2}. \tag{90}$$

The rescaled solution of the kinetic equation, see (57), is defined now by

$$W^\delta(t,y,k)=W(t/\delta^{3/2},y/\delta,k) \tag{91}$$

In [10] it is proven that for any $t > 0$ and $\varphi \in C_0^\infty(\mathbb{R} \times \mathbb{T})$

$$\lim_{\delta \to 0} \iint_{\mathbb{R}\times\mathbb{T}} W^\delta(t, y, k)\varphi(y, k)\, dy\, dk \;=\; \iint_{\mathbb{R}\times\mathbb{T}} \rho(t, y)\varphi(y, k)\, dy\, dk, \tag{92}$$

where $\rho(t, y)$ is the solution of

$$\partial_t \rho(t, y) = \hat{c}\Lambda_{3/4}\rho(t, y), \tag{93}$$

where

$$\hat{c} := \frac{\pi^2 \omega_a^{3/2}}{(2^5\gamma_0)^{1/2}} \int_0^{+\infty} \frac{(1 - \cos\lambda)d\lambda}{\lambda^{5/2}} = \left(\frac{\pi^5 \omega_a^3}{6\gamma_0}\right)^{1/2}, \tag{94}$$

cf [4, formula 3.762, 1, p. 437]

The proof of (92), presented in [10], is based on the probabilistic representation of the phonon trajectory process associated with the kinetic equation (57). It is shown that superdiffusively scaled trajectories of the process converge in law to those of a Levy process, with corresponding probabilities to be absorbed, transmitted or reflected when crossing $y = 0$, with a creation in the same point (its generator is given by (87)).

Remark Notice that the convergence in (92) holds for every time $t > 0$, while in the diffusive case it is only weakly in time (cf (84)). The explanation comes from different methods adopted in the respective proofs. The proof of (92) is of probabilistic nature, and uses the fact that the corresponding limiting transmitted/reflected/absorbed process jumps over the thermostat interface only finitely many times before being absorbed. On the other hand, the proof of (84) is analytic, and it would be difficult to establish, by a probabilistic method, a result for every time, since the corresponding Brownian motion crosses the thermostat infinitely many times before being absorbed by it.

6 Perspectives and Open Problems

6.1 Direct Hydrodynamic Limit

The results presented in the previous sections are obtained in the typical two-step procedure: we first take a kinetic limit (rarefied collisions) and obtain a kinetic equation with a boundary condition for the thermostat, next we rescale (diffusively or superdiffusively) this equation getting a diffusive or superdiffusive equation with an appropriate boundary condition.

It would be interesting to obtain a direct hydrodynamic limit, rescaling diffusively or superdiffusively the microscopic dynamics, without rarefaction of the random collision in the bulk. This means considering the evolution equations (54) with $\varepsilon = 1$, then setting a scale parameter δ (that does not appear in the evolution equations) and define the Wigner distribution by

$$\langle G, W^{(\delta)}(t)\rangle := \frac{\delta}{2}\sum_{y,y'\in\mathbb{Z}}\int_{\mathbb{T}} e^{2\pi ik(y'-y)}\mathbb{E}\left[\psi_y\left(\frac{t}{\delta^\alpha}\right)(\psi_{y'})^*\left(\frac{t}{\delta^\alpha}\right)\right]G^*\Big(\delta\frac{y+y'}{2},k\Big)dk. \tag{95}$$

with $\alpha = 2$, or $\alpha = 3/2$ in the diffusive, or superdiffusive case, respectively. Then one would like to show that, in some sense,

$$W^{(\delta)}(t,y,k) \underset{\delta\to 0}{\longrightarrow} \rho(t,y), \tag{96}$$

where $\rho(t,y)$ is solution of (85) or (93), depending on the scaling. In absence of a thermostat, this has been proved in [6].

6.2 More Thermostats

In non-equilibrium statistical mechanics it is always interesting to put the system in contact with a number of heat baths at various temperatures. If, in the case of dynamics defined by (6) or (54), we add another Langevin thermostat at the site $[\epsilon^{-1}y_0]$ with $y_0 \neq 0$, at a temperature T_1, we expect to obtain the same kinetic equations with added boundary conditions at the point y_0 analogous to (52) but of course the phonon production rate $\mathfrak{g}(k)T_1$. The difficulty in constructing the proof, lies in the fact that we no longer have an explicit formula for a solution in the case the inter-particle scattering is absent, that has been quite essential in our argument.

6.3 Poisson Thermostat

A different model for a heat bath at temperature T is given by a renewal of the velocity $p_0(t)$ at random times given by a Poisson process of intensity γ: each time the Poisson clock rings, the velocity is renewed with value chosen with a Gaussian distribution of variance T, independently of anything else. This mechanism represents the interaction with an infinitely extended reservoir of independent particles in equilibrium at temperature T and uniform density.

From a preliminary calculation (cf [8]) it seems that the scattering rates in the high frequency-kinetic limit are different, implying their dependence on the microscopic model of the thermostat. Obviously, in the hydrodynamic limit, diffusive or

superdiffusive, we expect that there boundary conditions will not depend anymore on the microscopic model of the thermostat.

Appendix 1: Properties of the Scattering Coefficients

Some Properties of the Scattering Coefficient for a Unimodal Dispersion Relation

Recall that $\nu(k)$ is defined by (39). From (30), we have

$$\lim_{\varepsilon\to 0}\tilde{J}(\varepsilon - i\omega(k)) = iG\big(\omega(k)\big) + iH\big(\omega(k)\big), \quad \text{for } \omega(k)\neq 0$$

where

$$G(u) := \int_{\mathbb{T}_+}\frac{d\ell}{u+\omega(\ell)}, \qquad H(u) := \frac{1}{2}\lim_{\varepsilon\to 0}\int_{\mathbb{T}}\frac{d\ell}{i\varepsilon + u - \omega(\ell)}. \tag{97}$$

If $\omega(k) = 0$, then $k = 0$ and, according to (30),

$$\lim_{\varepsilon\to 0}\tilde{J}(\varepsilon) = \frac{\pi}{|\omega'(0+)|}. \tag{98}$$

If the dispersion relation $\omega(k)$ is unimodal and $\omega_{\min} := \omega(0)$, $\omega_{\max} := \omega(1/2)$, then we can write $H(\omega(k)) = H^r(\omega(k)) + iH^i(\omega(k))$, with $H^r(u)$, $H^i(u)$ real valued functions equal

$$H^r(u) := \lim_{\varepsilon\to 0}\int_{\omega_{\min}}^{\omega_{\max}}\frac{(u-v)dv}{|\omega'(\omega_+^{-1}(v))|[\varepsilon^2+(u-v)^2]} \tag{99}$$

and

$$H^i(u) := -\lim_{\varepsilon\to 0}\int_{\omega_{\min}}^{\omega_{\max}}\frac{\varepsilon dv}{|\omega'(\omega_+^{-1}(v))|[\varepsilon^2+(u-v)^2]} = -\frac{\pi}{|\omega'(\omega_+^{-1}(u))|}. \tag{100}$$

Here $\omega_+^{-1} : [\omega_{\min}, \omega_{\max}] \to [0, 1/2]$ is the inverse of the increasing branch of $\omega(\cdot)$. For $u \in (\omega_{\min}, \omega_{\max})$ we can write

$$H^r(u) = \frac{1}{\omega'(\omega_+^{-1}(u))}\log\frac{\omega_{\max}-u}{u-\omega_{\min}} + \int_{\omega_{\min}}^{\omega_{\max}}\frac{\Big[\Big(\omega_+^{-1}\Big)'(v) - \Big(\omega_+^{-1}\Big)'(u)\Big]dv}{u-v}. \tag{101}$$

According to (29) and (39)

$$\nu(k) = \begin{cases} \{1 - \gamma H^i(\omega(k)) + i\gamma[G(\omega(k)) + H^r(\omega(k))]\}^{-1}, & \text{if } \omega(k) \neq 0, \\ \dfrac{2|\bar{\omega}'(0+)|}{2|\bar{\omega}'(0+)| + \gamma}, & \text{if } \omega(k) = 0. \end{cases} \tag{102}$$

Summarizing, from the above argument we conclude the following.

Theorem 4 *For a unimodal dispersion relation $\omega(\cdot)$ the following are true:*

(i) we have

$$|\nu(k)| \leq \frac{2|\bar{\omega}'(k)|}{\gamma + 2|\bar{\omega}'(k)|}, \quad k \in \mathbb{T}, \tag{103}$$

(ii) if k_* is such that $\omega'(k_*) = 0$, then

$$\lim_{k \to k_*} \nu(k) = 0 \quad \textit{and} \quad \lim_{k \to k_*} \mathfrak{g}(k) = 0, \tag{104}$$

(iii)

$$\operatorname{Re} \nu(k) > 0, \quad \textit{for all } k \in \mathbb{T} \setminus \{0, 1/2\}, \tag{105}$$

(iv)

$$p_+(k) > 0 \quad \textit{and } p_-(k) < 1 \textit{ for all } k \textit{ such that } \omega'(k) \neq 0 \tag{106}$$

and

$$p_-(k) > 0 \textit{ for all } k \in \mathbb{T} \setminus \{0, 1/2\}, \tag{107}$$

(v) we have the formula

$$\operatorname{Re} \nu(k) = \left(1 + \frac{\gamma}{2|\bar{\omega}'(k)|}\right) |\nu(k)|^2, \quad k \in \mathbb{T}. \tag{108}$$

Proof Substituting into (102) from (99) and (100) immediately yields (108). Estimate (105) follows directly from (102), formulas (97), (100), and (101).

Statement (109) is a consequence of (100) and (102). Part (ii) follows from part (i), cf (42). Estimates (106) follow directly from (109), while (107) is a straightforward consequence of part (iii), cf (50). □

From part (v) of Theorem 4 we immediately conclude the following.

Corollary 1 *Suppose that $\nu(k) \neq 0$ is real valued. Then,*

$$\nu(k) = \frac{|\omega'(k)|}{|\omega'(k)| + \gamma\pi}. \tag{109}$$

Proof of (53)

To conclude (53) we invoke (50). Then, thanks to (108), we can write

$$p_+(k) + p_-(k) + \mathfrak{g}(k) = 1 + \frac{\gamma}{|\bar{\omega}'(k)|}\left[|\nu(k)|^2\left(1 + \frac{\gamma}{2|\bar{\omega}'(k)|}\right) - \operatorname{Re}\nu(k)\right] = 1$$

and (53) follows.

An Example: Scattering Coefficient $\nu(k)$ for a Nearest Neighbor Interaction Harmonic Chain—Computation of $\tilde{J}(\lambda)$ Using Contour Integration

Assume that $\omega(k)$ is the dispersion relation of a nearest neighbor interaction harmonic chain. We let $\alpha_0 := (\omega_0^2 + \omega_a^2)/2$ and $\alpha_{\pm 1} := -\omega_a^2/4$, and $\omega_0 \geq 0$, $\omega_a > 0$. Then, see Sect. 3,

$$\hat{\alpha}(k) = \frac{\omega_0^2 + \omega_a^2}{2} - \frac{\omega_a^2}{4}\left(e^{2\pi ik} + e^{-2\pi ik}\right) = \frac{\omega_0^2}{2} + \omega_a^2 \sin^2(\pi k) \tag{110}$$

and, according to (7),

$$\omega(k) := \sqrt{\frac{\omega_0^2}{2} + \omega_a^2 \sin^2(\pi k)}. \tag{111}$$

Using the definition of $\tilde{J}(\lambda)$, see (30), and (110) for any $\lambda \in \mathbb{C}$ such that $\operatorname{Re}\lambda > 0$ we can write

$$\tilde{J}(\lambda) = \int_{\mathbb{T}} \frac{\lambda d\ell}{\lambda^2 + \hat{\alpha}(\ell)} = -\frac{4\lambda}{\omega_a^2}\int_{-1/2}^{1/2} \frac{e^{2\pi i\ell} d\ell}{e^{4\pi i\ell} - 2W(\lambda)e^{2\pi i\ell} + 1}, \tag{112}$$

where

$$W(\lambda) = 1 + \left(\frac{\omega_0}{\omega_a}\right)^2 + 2\left(\frac{\lambda}{\omega_a}\right)^2. \tag{113}$$

Note that $W(\lambda) \in \mathbb{C} \setminus [-1, 1]$, if $\operatorname{Re}\lambda > 0$.

The expression for $\tilde{J}(\lambda)$ can be rewritten using the contour integral over the unit circle $C(1)$ on the complex plane oriented counterclockwise and

$$\tilde{J}(\lambda) = -\frac{2\lambda}{i\pi\omega_a^2}\int_{C(1)} \frac{d\zeta}{\zeta^2 - 2W(\lambda)\zeta + 1}. \tag{114}$$

When $w \in \mathbb{C} \setminus [-1, 1]$ the equation

$$z^2 - 2wz + 1 = 0$$

has two roots. They are given by Φ_+, Φ_-, holomorphic functions on $\mathbb{C} \setminus [-1, 1]$, that are the inverse branches of the Joukowsky function $\mathfrak{J}(z) = 1/2(z + z^{-1})$, $z \in \mathbb{C}$ taking values in $\bar{\mathbb{D}}^c$ and $\mathbb{D}$, respectively. Here $\mathbb{D} := [z \in \mathbb{C} : |z| < 1]$ is the unit disc. We have

$$\lim_{\varepsilon\to 0+} \frac{1}{2}\Big(\Phi_+(a - \varepsilon i) - \Phi_-(a - \varepsilon i)\Big) = -i\sqrt{1 - a^2}, \quad \text{for } a \in [-1, 1]. \tag{115}$$

Using the Cauchy formula for contour integrals, from (114) we obtain

$$\tilde{J}(\lambda) = \frac{4\lambda}{\omega_a^2\Big(\Phi_+(W(\lambda)) - \Phi_-(W(\lambda))\Big)}. \tag{116}$$

For the dispersion relation $\omega(k)$ given by (111) and $\varepsilon > 0$ we have, cf (113),

$$W(\varepsilon - i\omega(k)) = \cos(2\pi k) + 2\left(\frac{\varepsilon}{\omega_a}\right)^2 - 4i\frac{\omega(k)\varepsilon}{\omega_a^2}.$$

As a result we get, cf (115) and (116),

$$\lim_{\varepsilon\to 0+} \tilde{J}(\varepsilon - i\omega(k)) = \frac{2}{\omega_a^2 \sin(2\pi|k|)}\sqrt{\frac{\omega_0^2}{2} + \omega_a^2 \sin^2(\pi k)}$$

and the following result holds.

Theorem 5 *For the dispersion relation given by* (111) *we have*

$$\nu(k) = \omega_a^2 \sin(2\pi|k|)\left\{\omega_a^2 \sin(2\pi|k|) + 2\gamma\sqrt{\frac{\omega_0^2}{2} + \omega_a^2 \sin^2(\pi k)}\right\}^{-1}, \quad k \in \mathbb{T}. \tag{117}$$

In particular, if $\omega_0 = 0$ (the acoustic case) we have, cf (42) *and* (50),

$$\nu(k) = \frac{\omega_a \cos(\pi k)}{\omega_a \cos(\pi k) + \gamma}, \quad k \in \mathbb{T} \tag{118}$$

and

$$\begin{aligned} p_+(k) &:= \left(\frac{\omega_a \cos(\pi k)}{\omega_a \cos(\pi k) + \gamma} \right)^2, \qquad p_-(k) := \left(\frac{\gamma}{\omega_a \cos(\pi k) + \gamma} \right)^2 \\ \mathfrak{g}(k) &= \frac{2\gamma\omega_a \cos(\pi k)}{(\omega_a \cos(\pi k) + \gamma)^2}, \quad k \in \mathbb{T}. \end{aligned} \tag{119}$$

Appendix 2: Proofs of (47) and (48)

Proof of (47)

Using (27) and (12) we can write

$$\widehat{W}^1_\varepsilon(t, \eta, k) = -\frac{\gamma}{2} \left\{ \mathcal{I}\left(\frac{t}{\varepsilon}, \eta, k \right) + \mathcal{I}^*\left(\frac{t}{\varepsilon}, -\eta, k \right) \right\}, \tag{120}$$

where

$$\mathcal{I}(t, \eta, k) := \frac{\varepsilon}{2} \int_{\mathbb{T}} dk' \mathbb{E}_\varepsilon \left[\hat{\psi}\left(0, k - \frac{\varepsilon\eta}{2}\right)^* \hat{\psi}\left(0, k'\right) \right] \int_0^t e^{i[\omega(k - \frac{\varepsilon\eta}{2})t - \omega(k')s]} \phi\left(t - s, k + \frac{\varepsilon\eta}{2}\right) ds. \tag{121}$$

The Laplace transform of $\mathcal{I}\left(\frac{t}{\varepsilon}, \eta, k\right)$ equals

$$\begin{aligned} &\tilde{I}_\varepsilon(\lambda, \eta, k) := \frac{\varepsilon^2}{2} \int_0^{+\infty} e^{-\varepsilon\lambda t} \mathcal{I}(t, \eta, k)\, dt \\ &= \frac{\varepsilon^2}{2} \int_0^{+\infty} e^{i\omega(k + \frac{\varepsilon\eta}{2})\tau} g(d\tau) \int_\tau^{+\infty} e^{i[\omega(k') - \omega(k + \frac{\varepsilon\eta}{2})]s} ds \int_s^{+\infty} e^{-\{\varepsilon\lambda + i[\omega(k') - \omega(k - \frac{\varepsilon\eta}{2})]\}t} dt \\ &\times \int_{\mathbb{T}} dk' \mathbb{E}_\varepsilon \left[\hat{\psi}\left(0, k - \frac{\varepsilon\eta}{2}\right)^* \hat{\psi}\left(0, k'\right) \right]. \end{aligned} \tag{122}$$

Performing the integration over the temporal variables we conclude that

$$\tilde{I}_\varepsilon(\lambda,\eta,k)=\int_{\mathbb{T}}\mathbb{E}_\varepsilon\Big[\hat{\psi}\left(0,k-\frac{\varepsilon\eta}{2}\right)^*\hat{\psi}\left(0,k'\right)\Big]\\ \times\frac{\tilde{g}\left(\lambda\varepsilon-i\omega(k-\frac{\varepsilon\eta}{2})\right)dk'}{2\{\lambda+i\varepsilon^{-1}[\omega(k-\frac{\varepsilon\eta}{2})-\omega(k')]\}\{\lambda+i\varepsilon^{-1}[\omega(k+\frac{\varepsilon\eta}{2})-\omega(k-\frac{\varepsilon\eta}{2})]\}}. \tag{123}$$

Using (39) we conclude that for any test function $G\in\mathcal{S}(\mathbb{R}\times\mathbb{T})$

$$\int_{\mathbb{R}\times\mathbb{T}}\tilde{I}_\varepsilon(\lambda,\eta,k)\hat{G}^*(\eta,k)d\eta dk\\ \approx\int_{\mathbb{R}\times\mathbb{T}^2}\frac{\mathbb{E}_\varepsilon\Big[\hat{\psi}\,(0,k)^*\,\hat{\psi}\,\left(0,k'\right)\Big]\hat{G}^*(\eta,k)\nu(k)d\eta dkdk'}{2\{\lambda+i\varepsilon^{-1}[\omega(k')-\omega(k)]\}\{\lambda+i\varepsilon^{-1}[\omega(k+\varepsilon\eta)-\omega(k)]\}}, \tag{124}$$

as $\varepsilon\ll 1$. Changing variables $k:=\ell-\varepsilon\eta'/2$ and $k':=\ell+\varepsilon\eta'/2$ we obtain that

$$\lim_{\varepsilon\to 0+}\int_{\mathbb{R}\times\mathbb{T}}\tilde{I}_\varepsilon(\lambda,\eta,k)\hat{G}^*(\eta,k)d\eta dk\\ =\int_{\mathbb{R}^2\times\mathbb{T}}\frac{\widehat{W}(0,\eta',\ell)\hat{G}^*(\eta,\ell)\nu(\ell)d\eta d\eta' d\ell}{(\lambda+i\omega'(\ell)\eta')(\lambda+i\omega'(\ell)\eta)}. \tag{125}$$

The limit of $\hat{w}^1_\varepsilon(\lambda,\eta,k)$—the Laplace transform of $\widehat{W}^1_\varepsilon(t,\eta,k)$—is therefore given by

$$\hat{w}^1(\lambda,\eta,k)=-\frac{\gamma\mathrm{Re}\nu(k)}{\lambda+i\omega'(k)\eta}\int_{\mathbb{R}}\frac{\widehat{W}(0,\eta',k)d\eta'}{\lambda+i\omega'(k)\eta'}. \tag{126}$$

Therefore

$$\lim_{\varepsilon\to 0+}\widehat{W}^1_\varepsilon(t,\eta,k)=-\gamma\mathrm{Re}\nu(k)e^{-i\omega'(k)t}\int_{\mathbb{R}}\frac{1-e^{-i\omega'(k)(\eta'-\eta)t}}{i\omega'(k)(\eta'-\eta)}\widehat{W}(0,\eta',k)d\eta' \tag{127}$$

and, performing the inverse Fourier transform, (47) follows.

Proof of (48)

Concerning the term $\widehat{W}^2_\varepsilon(t,\eta,k)$, from the third formula of (46), (12) and (27) we obtain

$$\widehat{W}^2_\varepsilon(t,\eta,k)=\frac{\gamma^2}{2}\left\{\mathcal{J}\left(\frac{t}{\varepsilon},\eta,k\right)+\mathcal{R}\left(\frac{t}{\varepsilon},\eta,k\right)\right\}, \tag{128}$$

where

$$\begin{aligned}
\mathcal{J}(t,\eta,k)&:=\frac{\varepsilon}{4}\int_{[0,t]^2}dsds'\int_{\mathbb{T}^2}d\ell d\ell'\phi\left(t-s,k-\frac{\varepsilon\eta}{2}\right)^*\phi\left(t-s',k+\frac{\varepsilon\eta}{2}\right)\\
&\times e^{i[\omega(\ell)s-\omega(\ell')s']}\mathbb{E}_\varepsilon\left[\hat\psi(0,\ell)^*\hat\psi(0,\ell')\right],\\
\mathcal{R}(t,\eta,k)&:=\frac{\varepsilon}{4}\int_{[0,t]^2}dsds'\int_{\mathbb{T}^2}d\ell d\ell'\phi\left(t-s,k-\frac{\varepsilon\eta}{2}\right)^*\phi\left(t-s',k+\frac{\varepsilon\eta}{2}\right)\\
&\times e^{-i[\omega(\ell)s-\omega(\ell')s']}\mathbb{E}_\varepsilon\left[\hat\psi(0,\ell)\hat\psi(0,\ell')^*\right].
\end{aligned} \tag{129}$$

A simple computation shows that

$$\begin{aligned}
\mathcal{J}(t,\eta,k)&=\frac{\varepsilon}{4}\int_{[0,t]^2}dsds'\int_0^s g(d\tau)g(d\tau')\int_{\mathbb{T}^2}d\ell d\ell'\mathbb{E}_\varepsilon\left[\hat\psi(0,\ell)^*\hat\psi(0,\ell')\right]\\
&\times e^{i[\omega(k-\varepsilon\eta/2)(s-\tau)-\omega(k+\varepsilon\eta/2)(s'-\tau')]}e^{-i[\omega(\ell')(t-s')-\omega(\ell)(t-s)]}.
\end{aligned} \tag{130}$$

The respective Laplace transform equals

$$\begin{aligned}
\tilde{\mathcal{J}}_\varepsilon(\lambda,\eta,k)&:=\varepsilon\int_0^{+\infty}e^{-\varepsilon\lambda\tau_0}\mathcal{J}(\tau_0,\eta,k)d\tau_0\\
&=\varepsilon\int_0^{+\infty}\int_0^{+\infty}\delta(\tau_0-\tau_0')e^{-\varepsilon\lambda\tau_0/2}e^{-\varepsilon\lambda\tau_0'/2}\mathcal{J}(\tau_0,\eta,k)\mathcal{J}(\tau_0',\eta,k)d\tau_0d\tau_0'\\
&=\frac{\varepsilon^2}{4}\int_{\bar{\mathbb{R}}^4_+\times\bar{\mathbb{R}}^4_+}d\tau_{0,2}g(d\tau_3)d\tau'_{0,2}g(d\tau'_3)\int_{\mathbb{T}^2}d\ell d\ell'\\
&\times\delta(\tau_0-\tau_0')\delta\left(\tau_0-\sum_{j=1}^3\tau_j\right)\delta\left(\tau_0'-\sum_{j=1}^3\tau_j'\right)e^{-\varepsilon\lambda\sum_{j=0}^3\tau_j/4}e^{-\varepsilon\lambda\sum_{j=0}^3\tau_j'/4}\\
&\times e^{i[\omega(k-\varepsilon\eta/2)\tau_2+\omega(\ell)\tau_1]}e^{-i[\omega(k+\varepsilon\eta/2)\tau_2'+\omega(\ell')\tau_1']}\mathbb{E}_\varepsilon\left[\hat\psi(0,\ell)^*\hat\psi(0,\ell')\right].
\end{aligned} \tag{131}$$

Here, for abbreviation sake we write $d\tau_{0,3} = d\tau_0 d\tau_1 d\tau_2$ and likewise for the prime variables. Using the identity $\delta(t) = (2\pi)^{-1}\int_{\mathbb{R}} e^{i\beta t}d\beta$ and integrating out the τ and τ' variables we obtain

$$\begin{aligned}\tilde{\mathcal{J}}_\varepsilon(\lambda,\eta,k) &= \frac{\varepsilon^2}{2^5\pi^3}\int_{\mathbb{R}^3} d\beta_0 d\beta_1 d\beta_1' \int_{\mathbb{T}^2} d\ell d\ell' \mathbb{E}_\varepsilon\left[\hat{\psi}(0,\ell)^*\hat{\psi}(0,\ell')\right]\\ &\times \frac{\tilde{g}(\varepsilon\lambda/4+i\beta_1)}{[\varepsilon\lambda/4-i(\beta_0+\beta_1)][\varepsilon\lambda/4+i(\beta_1-\omega(\ell))][\varepsilon\lambda/4+i(\beta_1-\omega(k-\varepsilon\eta/2))]}\\ &\times \frac{\tilde{g}(\varepsilon\lambda/4+i\beta_1')}{[\varepsilon\lambda/4+i(\beta_0-\beta_1')][\varepsilon\lambda/4+i(\beta_1'+\omega(\ell'))][\varepsilon\lambda/4+i(\beta_1'+\omega(k+\varepsilon\eta/2))]}.\end{aligned} \tag{132}$$

We integrate β_1 and β_1' variables using the Cauchy integral formula

$$\frac{1}{2\pi}\int_{\mathbb{R}}\frac{f(i\beta)d\beta}{z-i\beta} = f(z), \quad z\in\mathbb{H}, \tag{133}$$

valid for any holomorphic function f on the right half-plane $\mathbb{H} := [z\in\mathbb{C}:\ \mathrm{Re}\, z > 0]$ that belongs to the Hardy class $H^p(\mathbb{H})$ for some $p \geq 1$, see e.g. [11, p. 113]. Performing the above integration and, subsequently, changing variables $\varepsilon\beta_0' := \beta_0 + \omega(k-\varepsilon\eta/2)$ we get

$$\begin{aligned}\tilde{\mathcal{J}}_\varepsilon(\lambda,\eta,k) &= \frac{1}{2^3\pi\varepsilon}\int_{\mathbb{R}}\frac{d\beta_0}{\lambda/2-i\beta_0}\int_{\mathbb{T}^2} d\ell d\ell' \mathbb{E}_\varepsilon\left[\hat{\psi}(0,\ell)^*\hat{\psi}(0,\ell')\right]\\ &\times \frac{|\tilde{g}\big(\varepsilon\lambda/2-i\varepsilon\beta_0+i\omega(k-\varepsilon\eta/2)\big)|^2}{\lambda/2-i\varepsilon^{-1}(\omega(\ell)-\omega(k-\varepsilon\eta/2))-i\beta_0}\\ &\times \frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(\ell')-\omega(k-\varepsilon\eta/2))+i\beta_0}\\ &\times \frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(k+\varepsilon\eta/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}.\end{aligned} \tag{134}$$

Change variables ℓ, ℓ' according to the formulas $\ell := k'-\varepsilon\eta'/2$ and $\ell' := k'+\varepsilon\eta'/2$ and use (cf (39))

$$|\tilde{g}\big(\varepsilon\lambda/2-i\varepsilon\beta_0+i\omega(k-\varepsilon\eta/2)\big)|^2 \approx |\nu(k)|^2, \quad \text{as } \varepsilon \ll 1.$$

We obtain then

$$\begin{aligned}\tilde{\mathcal{J}}_\varepsilon(\lambda,\eta,k) \approx{}& \frac{|\nu(k)|^2}{2^3\pi\varepsilon}\int_{\mathbb{R}}\frac{d\beta_0}{\lambda/2-i\beta_0}\int_{\mathbb{R}\times\mathbb{T}}\widehat{W}(0,\eta',k')d\eta'dk'\\ &\times\frac{1}{\lambda/2-i\varepsilon^{-1}(\omega(k'-\varepsilon\eta'/2)-\omega(k-\varepsilon\eta/2))-i\beta_0}\\ &\times\frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(k'+\varepsilon\eta'/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}\\ &\times\frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(k+\varepsilon\eta/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}.\end{aligned}\tag{135}$$

Since $\omega(k)$ is unimodal we can write

$$\begin{aligned}\tilde{\mathcal{J}}_\varepsilon(\lambda,\eta,k) \approx{}& \frac{|\nu(k)|^2}{2^3\pi\varepsilon}\sum_{\iota=\pm}\int_{\mathbb{R}}\frac{d\beta_0}{\lambda/2-i\beta_0}\int_{\mathbb{R}\times[\iota k-\delta,\iota k+\delta]}\widehat{W}(0,\eta',k')d\eta'dk'\\ &\times\frac{1}{\lambda/2-i\varepsilon^{-1}(\omega(k'-\varepsilon\eta'/2)-\omega(k-\varepsilon\eta/2))-i\beta_0}\\ &\times\frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(k'+\varepsilon\eta'/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}\\ &\times\frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(k+\varepsilon\eta/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}.\end{aligned}\tag{136}$$

for a (small) fixed $\delta > 0$. Changing variables $k' = k+\varepsilon\eta''$ and using the approximations $\varepsilon^{-1}[\omega(k+\varepsilon\xi)-\omega(k)] \approx \omega'(k)\xi$ and $\widehat{W}(0,\eta',\iota k+\varepsilon\eta'') \approx \widehat{W}(0,\eta',\iota k)$ we conclude that

$$\begin{aligned}\tilde{\mathcal{J}}_\varepsilon(\lambda,\eta,k) \approx{}& \frac{|\nu(k)|^2}{2^3\pi}\sum_{\iota=\pm}\int_{\mathbb{R}}\frac{d\beta_0}{\lambda/2-i\beta_0}\int_{\mathbb{R}^2}\widehat{W}(0,\eta',\iota k+\varepsilon\eta'')d\eta'd\eta''\\ &\times\frac{1}{\lambda/2-i\varepsilon^{-1}(\omega(\iota(k+\varepsilon\eta'')-\varepsilon\eta'/2)-\omega(k-\varepsilon\eta/2))-i\beta_0}\\ &\times\frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(\iota(k+\varepsilon\eta'')+\varepsilon\eta'/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}\\ &\times\frac{1}{\lambda/2+i\varepsilon^{-1}(\omega(k+\varepsilon\eta/2)-\omega(k-\varepsilon\eta/2))+i\beta_0}\end{aligned}\tag{137}$$

$$\approx \frac{|\nu(k)|^2}{2^3\pi} \sum_{\iota=\pm} \int_{\mathbb{R}} \frac{d\beta_0}{(\lambda/2 - i\beta_0)(\lambda/2 + i\omega'(k)\eta + i\beta_0)} \int_{\mathbb{R}} \widehat{W}(0, \eta', \iota k) d\eta'$$
$$\times \int_{\mathbb{R}} \frac{1}{\lambda/2 - i\omega'(k)(\eta'' + \eta/2 - \iota\eta'/2) - i\beta_0}$$
$$\times \frac{d\eta''}{\lambda/2 + i\omega'(k)(\eta'' + \eta/2 + \iota\eta'/2) + i\beta_0}.$$

Integrating, first with respect to η'' and then β_0 variables, using e.g. (133), we get

$$\lim_{\varepsilon\to 0+} \tilde{\mathcal{J}}_\varepsilon(\lambda, \eta, k) = \frac{|\nu(k)|^2}{4|\bar{\omega}'(k)|} \sum_{\iota=\pm} \int_{\mathbb{R}} \frac{d\eta}{\lambda + i\omega'(k)\eta} \int_{\mathbb{R}} \frac{\widehat{W}(0, \eta', \iota k) d\eta'}{\lambda + i\iota\omega'(k)\eta'} \tag{138}$$

From the second equality of (129) we can see that formula for $\tilde{\mathcal{R}}_\varepsilon(\lambda, \eta, k)$ can be obtained from (135) by changing $\omega(\ell)$ and $\omega(\ell')$ to $-\omega(\ell)$ and $-\omega(\ell')$ respectively and altering the complex conjugation by the wave functions. It yields

$$\tilde{\mathcal{R}}_\varepsilon(\lambda, \eta, k) \approx \frac{|\nu(k)|^2}{2^3\pi\varepsilon} \sum_{\iota=\pm} \int_{\mathbb{R}} \frac{d\beta_0}{\lambda/2 - i\beta_0} \int_{\mathbb{R}\times[\iota k-\delta, \iota k+\delta]} \widehat{W}(0, \eta', k') d\eta' dk'$$
$$\times \frac{1}{\lambda/2 + i\varepsilon^{-1}(\omega(k' - \varepsilon\eta'/2) + \omega(k - \varepsilon\eta/2)) - i\beta_0}$$
$$\times \frac{1}{\lambda/2 - i\varepsilon^{-1}(\omega(k' + \varepsilon\eta'/2) + \omega(k - \varepsilon\eta/2)) + i\beta_0}$$
$$\times \frac{1}{\lambda/2 + i\varepsilon^{-1}(\omega(k + \varepsilon\eta/2) - \omega(k - \varepsilon\eta/2)) + i\beta_0} \approx 0, \tag{139}$$

as both the second and third lines are of order ε, while the fourth one is of order 1.

Summarizing, we have shown that (see [9] for a rigorous derivation)

$$\frac{1}{2\pi} \lim_{\varepsilon\to 0+} \int_{\mathbb{R}} e^{i\eta y} \widehat{W}_\varepsilon^2(t, \eta, k) d\eta$$
$$= \frac{\gamma^2 |\nu(k)|^2}{4|\bar{\omega}'(k)|^2} 1_{[[0, \bar{\omega}'(k)t]]}(y) \left(W(0, y - \bar{\omega}'(k)t, k) + W(0, -y + \bar{\omega}'(k)t, -k)\right) \tag{140}$$

and (48) follows.

Acknowledgments TK was partially supported by the NCN grant 2016/23/B/ST1/00492, SO by the French Agence Nationale Recherche grant LSD ANR-15-CE40-0020-01.

References

1. Basile, G., Olla, S., Spohn, H.: Energy transport in stochastically perturbed lattice dynamics. Arch. Rat. Mech. **195**(1), 171–203 (2009). https://doi.org/10.1007/s00205-008-0205-6
2. Basile, G., Bernardin, C., Jara, M., Komorowski, T., Olla, S.: Thermal conductivity in harmonic lattices with random collisions. In: Lepri, S. (ed.) Thermal Transport in Low Dimensions: From Statistical Physics to Nanoscale Heat Transfer. Lecture Notes in Physics, vol. 921, chapter 5. Springer, Berlin (2016). https://doi.org/10.10007/978-3-319-29261-8-5
3. Basile, G., Komorowski, T., Olla, S.: Diffusive limits for a kinetic equation with a thermostatted interface. Kinetic and Related Models. AIMS **12**(5), 1185–1196 (2019). https://doi.org/10.3934/krm.2019045
4. Gradshteyn, I.S., Ryzhik, I.M.: Table of integrals, series, and products. Transl. from the Russian. In: Jeffrey, A., Zwillinger, D. (eds.) Translation edited and with a preface, 7th edn. Elsevier/Academic Press, Amsterdam (2007)
5. Jara, M., Komorowski, T., Olla, S.: A limit theorem for an additive functionals of Markov chains. Ann. Appl. Probab. **19**(6), 2270–2300 (2009). https://doi.org/10.1214/09-AAP610
6. Jara, M., Komorowski, T., Olla, S.: Superdiffusion of Energy in a system of harmonic oscillators with noise. Commun. Math. Phys. **339**, 407–453 (2015). https://doi.org/10.1007/s00220-015-2417-6
7. Komorowski, T., Olla, S.: Kinetic limit for a chain of harmonic oscillators with a point Langevin thermostat. J. Funct. Anal. **279**(12), 108764 (2020). https://doi.org/10.1016/j.jfa.2020.108764
8. Komorowski, T., Olla, S.: Asymptotic Scattering by Poissonian Thermostats (2021). https://arxiv.org/abs/2101.04360
9. Komorowski, T., Olla, S., Ryzhik, L., Spohn, H.: High frequency limit for a chain of harmonic oscillators with a point Langevin thermostat. Arch. Rational Mech. An. **237**, 497–543 (2020). https://doi.org/10.1007/s00205-020-01513-7
10. Komorowski, T., Olla, S., Ryzhik, L.: Fractional Diffusion limit for a kinetic equation with an interface. Ann. Probab. **48**(5), 2290–2322, (2020). https://doi.org/10.1214/20-AOP1423
11. Koosis, P.: Introduction to H^p spaces. Cambridge University, Cambridge (1980)
12. Kwaśnicki, M.: Ten equivalent definitions of the fractional Laplace operator. Fract. Calc. Appl. Anal. **20**(1), 751 (2017)
13. Lepri, S., Livi, R., Politi, A.: Thermal conduction in classical low-dimensional lattices. Phys. Rep. **377**, 1–80 (2003)
14. Rieder, Z., Lebowitz, J.L., Lieb, E.: Properties of harmonic crystal in a stationary non-equilibrium state. J. Math. Phys. **8**, 1073–1078 (1967)
15. Spohn, H.: The phonon Boltzmann equation, properties and link to weakly anharmonic lattice dynamics. J. Stat. Phys. **124**(2–4), 1041–1104 (2006)

Control of Collective Dynamics with Time-Varying Weights

Benedetto Piccoli and Nastassia Pouradier Duteil

Abstract This paper focuses on a model for opinion dynamics, where the influence weights of agents evolve in time. We formulate a control problem of consensus type, in which the objective is to drive all agents to a final target point under suitable control constraints. Controllability is discussed for the corresponding problem with and without constraints on the total mass of the system, and control strategies are designed with the steepest descent approach. The mean-field limit is described both for the opinion dynamics and the control problem. Numerical simulations illustrate the control strategies for the finite-dimensional system.

1 Introduction

Social dynamics models are used to describe the complex behavior of large systems of interacting agents. Application areas include examples from biology, such as the collective behavior of animal groups [3, 6, 10, 16], aviation [22], opinion dynamics [13] and other. In most applications, a key phenomenon observed is that of *self-organization*, that is the spontaneous emergence of global patterns from local interactions. Self-organization patterns include *consensus*, *alignment*, *clustering*, or the less studied *dancing equilibrium* [1, 5]. In another direction, the control of such systems was addressed in the control community with a wealth of different approaches, see [4, 14, 21].

This paper focuses on models for opinion dynamics. A long history started back in the 1950s, see [9, 11], then linear models were studied by De Groot [7] and others, while among recent approaches we can mention the bounded-confidence

B. Piccoli
Department of Mathematical Sciences, Rutgers University—Camden, Camden, NJ, USA
e-mail: piccoli@camden.rutgers.edu

N. P. Duteil (✉)
Sorbonne Université, Inria, Université Paris-Diderot SPC, CNRS, Laboratoire Jacques-Louis Lions, Paris, France
e-mail: nastassia.pouradier_duteil@sorbonne-universite.fr

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_12

model by Hegselmann and Krause of [13], see also [12, 15]. In most of the existing models, interactions take place between pairs of individuals (typically referred to as *agents*) and depend only on the distance separating the two agents. More recently, a model was introduced in which the interactions are proportional to the agents' *weights of influence*, which can evolve over time according to their own dynamics [2, 17, 18, 20]. This augmented framework allows us to model opinion dynamics in which an agent's capacity to influence its neighbors depends not only on their proximity but also on an internal time-varying characteristic (such as charisma, popularity, etc.). Four models were proposed in [17] for the time-varying weights: the first model allows agents to gain mass in pairwise interactions depending on midpoint dynamics; the second increases the weights of agents that influence the most the other agents; and the third and fourth focus on the capability to attract the most influential agents. In particular, the developed theory allows to address control problems, which is the focus of the present paper.

The main idea is that an external entity (for instance with global control) may influence the dynamics of agents by increasing the weights of some of them. We thus assume that a central controller is able to act on each agent but possibly influence just a few at a time, thus also looking for *sparse control* strategies. We first formulate the control problem by allowing a direct control of weights but imposing the total sum of weights to be constant, resulting in a linear constraint on allowable controls. Under natural assumptions on the interaction kernel we show that the convex hull of the agents' positions is shrinking, thus we look for control strategies stabilizing to a specific point of the initial convex hull.

The constraints on the control and given by the dynamics (shrinking convex hull) prevent a complete controllability of the system. However, we show that any target position strictly within the initial convex hull of the system can be reached given large enough bounds on the control.

We then look for a greedy policy by maximizing the instantaneous decrease of the distance from the target point. This gives rise to a steepest descent algorithm which is formulated via the linear constraints of the problem. Under generic conditions, the solution is expected to be at a vertex of the convex set determined by constraints.

As customary for multi-agent and multi-particle systems, we consider the mean-field limit obtained when the number of agents tends to infinity. In classical models without mass variation, the limit measure satisfies a transport-type equation with non-local velocity. Here, due to the presence of the weight dynamics, our mean-field equation presents a non-local source term. We formulate a control problem for the mean-field limit and show how to formulate the control constraints in this setting.

In the last section we provide simulations for the finite-dimensional control algorithm and illustrate how the control strategies reach the final target in the various imposed constraints.

2 Control Problems

We consider a collective dynamics system with time-varying weights, introduced in [17]. Let $x^0 \in (\mathbb{R}^d)^N$ represent the N agents' initial positions (or opinions) and $m^0 \in (\mathbb{R}^+)^N$ represent their initial weights of influence. We denote by $a \in C(\mathbb{R}^+, \mathbb{R}^+)$ the interaction function. Lastly, let $M = \sum_{i=1}^N m_i^0$ denote the initial mass of the system. In this model, the evolution of each agents' state variable $x_i(t)$ depends on its interaction with other agents through the interaction function a (as in the classical Hegselmann-Krause dynamics [13]), weighted by the other agents' weights of influence $m_i(t)$. The weights of influence also evolve in time due to their own dynamics. More precisely, the evolution of the N positions and weights is given by the following system:

$$\begin{cases} \dot{x}_i(t) = \dfrac{1}{M} \displaystyle\sum_{j=1}^N m_j(t) a(\|x_i(t) - x_j(t)\|) \left(x_j(t) - x_i(t)\right), \\ \dot{m}_i(t) = m_i(t) \psi_i(x(t), m(t)) \\ x_i(0) = x_i^0, \quad m_i(0) = m_i^0. \end{cases} \tag{1}$$

We have established in [17] the well-posedness of (1) along with the following hypotheses:

Hypothesis 1 The function $s \mapsto a(\|s\|)s$ is locally Lipschitz in $\mathbb{R}^d$, and the function ψ is locally bounded in $(\mathbb{R}^d)^N \times \mathbb{R}^N$.

Hypothesis 2 For all $(x, m) \in (\mathbb{R}^d)^N \times \mathbb{R}^N$,

$$\sum_{i=1}^N m_i \psi_i(x, m) = 0. \tag{2}$$

Note that Hypothesis 2 is not necessary for the well-posedness of (1). It is a modeling choice which enforces conservation of the total mass of the system, so that the weights m_i are allowed to shift continuously between agents, but their sum remains constant. We refer the reader to [17] for a detailed analysis of this system for various choices of the weight dynamics, exhibiting behaviors such as emergence of a single leader, or emergence of two co-leaders.

In the present paper, we aim to study the control of system (1) by acting only on the weights of influence. Let $\Omega(x)$ denote the convex hull of x, defined as follows.

Definition 1 Let $(x_i)_{i \in \{1, \cdots, N\}} \in (\mathbb{R}^d)^N$. Its convex hull Ω is defined by:

$$\Omega = \left\{ \sum_{i=1}^N \xi_i x_i \mid \xi \in [0, 1]^N \text{ and } \sum_{i=1}^N \xi_i = 1 \right\}.$$

It was shown in [17] that for the dynamics (1)–(2), the convex hull $\Omega(x(t))$ is contracting in time, i.e. for all $t_2 \geq t_1 \geq 0$, $\Omega(x(t_2)) \subseteq \Omega(x(t_1))$.

Given $\alpha \in \mathbb{R}^+$ and $A \in \mathbb{R}^+$, we define two control sets U_∞^α and U_1^A:

$$\begin{cases} U_\infty^\alpha = \{u : \mathbb{R}^+ \to \mathbb{R}^N \text{ measurable, s.t. } |u_i| \leq \alpha\} \\ U_1^A = \{u : \mathbb{R}^+ \to \mathbb{R}^N \text{ measurable, s.t. } \sum_{i=1}^N |u_i| \leq A\}. \end{cases}$$

We also define a set of controls that conserve the total mass M of the system: $U_M = \{u : \mathbb{R}^+ \to \mathbb{R}^N \text{ measurable, s.t. } \sum_{i=1}^N m_i u_i = 0\}$. From here onwards, U will stand for a general control set, equal to either U_1^A, U_∞^α, $U_1^A \cap U_M$ or $U_\infty^\alpha \cap U_M$.

We aim to solve the following control problem:

Problem 1 For all $x^* \in \Omega(x^0)$, find $u \in U$ such that the solution to

$$\begin{cases} \dot{x}_i = \dfrac{1}{M} \displaystyle\sum_{j=1}^N m_j a(\|x_i - x_j\|) \left(x_j - x_i\right), \\ \dot{m}_i(t) = m_i \left(\psi_i(m, x) + u_i\right) \\ x_i(0) = x_i^0, \quad m_i(0) = m_i^0, \end{cases} \tag{3}$$

satisfies: for all $i \in \{1, \cdots, N\}$, $\lim_{t\to\infty} \|x_i(t) - x^*\| = 0$.

We also suppose that the interaction function satisfies $a(s) > 0$ for all $s > 0$. Then from [17], if the total mass is conserved, the system converges asymptotically to consensus. Let $\bar{x} := \frac{1}{\sum_{i=1}^N m_i} \sum_{i=1}^N m_i x_i$ denote the weighted barycenter of the system. Then the control problem simplifies to:

Problem 2 Find $u \in U$ such that the solution to (3) satisfies

$$\lim_{t\to\infty} \|\bar{x}(t) - x^*\| = 0.$$

We seek a control that will vary the weights of the system so that its barycenter converges to the target position x^*. In (3), the control u must also compensate for the inherent mass dynamics. Here we will only consider the simpler case in which there is no inherent mass dynamics, i.e. $\psi_i \equiv 0$ for all $i \in \{1, \cdots, N\}$. The control problem re-writes:

Problem 3 For all $x^* \in \Omega(0)$, find $u \in U$ such that the solution to

$$\begin{cases} \dot{x}_i = \dfrac{1}{M} \displaystyle\sum_{j=1}^N m_j a(\|x_i - x_j\|) \left(x_j - x_i\right), \\ \dot{m}_i(t) = m_i u_i \\ x_i(0) = x_i^0, \quad m_i(0) = m_i^0, \end{cases} \tag{4}$$

satisfies: $\lim_{t\to\infty} \|\bar{x}(t) - x^*\| = 0$.

The solution to the more general Problem 2 can be recovered by the feedback transformation $u_i \mapsto u_i - \psi_i$, hence without loss of generality we will focus on Problem 3. It was proven in [17] that without control (i.e. with non-evolving weights), the weighted average $\bar{x}$ is constant. The control strategy will consist of driving $\bar{x}$ to x^*.

3 Control with Mass Conservation

In this section, we explore the controllability of the system when constraining the total mass of the system $\sum_{i=1}^N m_i(t)$ to M, by imposing $u \in U_M$. This amounts to looking for a control that will redistribute the weights of the agents while preserving their sum. It was shown in [17] that this condition implies that the convex hull $\Omega(t)$ is contracting in time. We remind an even stronger property of the system in the case of constant total mass (see [17], Prop. 10):

Proposition 1 *Let (x, m) be a solution to (1)–(2), and let $D(t) := \sup\{\|x_i - x_j\|(t) \mid (i, j) \in \{1, \cdots, N\}^2\}$ be the diameter of the system. If $\inf\{a(s) \mid s \leq D(0)\} := a_{min} > 0$ then the system (1)–(2) converges to consensus, with the rate $D(t) \leq D(0)e^{-a_{min}t}$.*

Remark 1 As a consequence, the convex hull converges to a single point $\Omega_\infty := \cap_{t\geq 0}\Omega(x(t)) = \{\lim_{t\to\infty} \bar{x}(t)\}$.

The properties of contraction of the convex hull and convergence to consensus imply that the target position x^* is susceptible to exit the convex hull in finite time. However, we show that that given sufficiently large upper bounds on the strength of the control, the system is approximately controllable to any target position within the interior of the convex hull, that we denote by $\mathring{\Omega}$. We state and demonstrate the result for the control constraints $u \in U_\infty^\alpha \cap U_M$, but the proof can be easily adapted to the case $u \in U_1^A \cap U_M$.

Theorem 1 *Let $(x_i^0)_{i\in\{1,\cdots,N\}} \in \mathbb{R}^{dN}$, $(m_i^0) \in (0, M)^N$ such that $\sum_{i=1}^N m_i^0 = M$ and let $x^* \in \mathring{\Omega}(x^0)$. Then for all $\varepsilon > 0$, there exists $\alpha > 0$, $t_\varepsilon \geq 0$ and $u \in U_\infty^\alpha \cap U_M$ such that the solution to (4) satisfies: $\|\bar{x}(t_\varepsilon) - x^*\| \leq \varepsilon$.*

Proof First, notice that since $m_i^0 > 0$ for all $i \in \{1, \cdots, N\}$, $\|x_i(t) - x_i^0\| > 0$ for all $t > 0$. Notice also that since the shrinking hull is contracting, we have $\|x_i(t) - x_j(t)\| \leq D_0$ for all $(i, j) \in \{1, \cdots, N\}^2$ and $t \geq 0$, where D_0 denotes the initial diameter of the system. Let

$$\delta := \sup_{s\in[0,D_0]} \{sa(s)\}. \tag{5}$$

From Hypothesis 1, $\delta < \infty$. Then for all $u \in U_M$, $\sum_{j=1}^N m_j \equiv M$, hence for all $t > 0$,

$$\frac{d}{dt}\|x_i - x_i^0\| = \frac{1}{\|x_i - x_i^0\|}\langle x_i - x_i^0, \dot{x}_i\rangle \leq \frac{1}{\|x_i - x_i^0\|}\|x_i - x_i^0\|\frac{1}{M}\sum_{j=1}^N m_j\delta = \delta$$

from which we deduce that for all $t \geq 0$, $\|x_i(t) - x_i^0\| \leq \delta t$. Since $x^* \in \mathring{\Omega}(x^0)$, there exists $\eta > 0$ such that $B(x^*, \eta) \subset \mathring{\Omega}(x^0)$. So for $t \leq \frac{\eta}{\delta}$, $x^* \in \mathring{\Omega}(x(t))$ for any control u. We now look for a control strategy that can drive $\bar{x}$ to a distance ε of x^* in time $t_\varepsilon := \frac{\eta}{\delta}$.

Let us compute the time derivative of the weighted barycenter. For $u \in U_M$, the sum of masses is conserved and $\bar{x} = \frac{1}{M}\sum_{i=1}^N m_i x_i$. Then

$$\frac{d}{dt}\bar{x} = \frac{1}{M}\sum_{i=1}^N(\dot{m}_i x_i + m_i\dot{x}_i) = \frac{1}{M}\sum_{i=1}^N m_i u_i x_i,$$

as the second term vanishes by antisymmetry of the summed coefficient. While $\|\bar{x} - x^*\| > 0$, we have

$$\frac{d}{dt}\|\bar{x} - x^*\| = \frac{1}{M\|\bar{x} - x^*\|}\sum_{i=1}^N\langle\bar{x} - x^*, m_i u_i x_i\rangle = \frac{1}{M\|\bar{x} - x^*\|}\sum_{i=1}^N\langle\bar{x} - x^*, x_i - x^*\rangle m_i u_i$$

since $\sum_{i=1}^N m_i u_i x^* = 0$. Let i_- and i_+ be defined as follows: for all $i \in \{1, \cdots, N\}$,

$$\begin{cases} m_{i_-}\langle\bar{x} - x^*, x_{i_-} - x^*\rangle \leq m_i\langle\bar{x} - x^*, x_i - x^*\rangle \\ m_{i_+}\langle\bar{x} - x^*, x_{i_+} - x^*\rangle \geq m_i\langle\bar{x} - x^*, x_i - x^*\rangle. \end{cases}$$

Note that i_- and i_+ are time-dependent, but we keep the notation $i_- = i_-(t)$ and $i_+ = i_+(t)$ for conciseness. For all $t \leq t_\varepsilon$, $x^* \in \mathring{\Omega}(x(t))$ so necessarily

$$\langle\bar{x} - x^*, x_{i_-} - x^*\rangle \leq 0 \leq \langle\bar{x} - x^*, x_{i_+} - x^*\rangle.$$

Notice also that the following holds (by summing over all indices):

$$m_{i_+}\langle\bar{x} - x^*, x_{i_+} - x^*\rangle \geq \frac{M}{N}\|\bar{x} - x^*\|^2.$$

Let $\tilde{\alpha} > 0$. We now design a control u such that:

$$u_{i_-} = \tilde{\alpha}\frac{m_{i_+}}{m_{i_-}}; \qquad u_{i_+} = -\tilde{\alpha}; \qquad u_i = 0 \text{ for all } i \in \{1, \cdots, N\}, i \neq i_-, i \neq i_+.$$

One can easily check that $u \in U_M$. With this control, we compute:

$$\begin{aligned}\frac{d}{dt}\|\bar{x}-x^*\| =& \frac{1}{M\|\bar{x}-x^*\|}[m_{i_-}\langle \bar{x}-x^*, x_{i_-}-x^*\rangle u_{i_-} + m_{i_+}\langle \bar{x}-x^*, x_{i_+}-x^*\rangle u_{i_+}] \\ \leq& \frac{1}{M\|\bar{x}-x^*\|} m_{i_+}\langle \bar{x}-x^*, x_{i_+}-x^*\rangle(-\tilde{\alpha}) \\ \leq& -\frac{\tilde{\alpha}}{M\|\bar{x}-x^*\|}\frac{M}{N}\|\bar{x}-x^*\|^2 \leq -\tilde{\alpha}\frac{\|\bar{x}-x^*\|}{N}.\end{aligned}$$

Then $\|\bar{x}-x^*\|(t) \leq \|\bar{x}^0-x^*\|e^{-\frac{\tilde{\alpha}}{N}t}$. If $\tilde{\alpha} \geq \frac{N}{t_\varepsilon}\ln\left(\frac{\|\bar{x}^0-x^*\|}{\varepsilon}\right)$, then $\|\bar{x}-x^*\|(t_\varepsilon) \leq \varepsilon$.

It remains to show that there exists $\alpha > 0$ such that $u \in U_\infty^\alpha$. By construction of the control u, for all $t \geq 0$ it holds:

$$\dot{m}_{i_-(t)}(t) = \tilde{\alpha} m_{i_+(t)}(t); \qquad \dot{m}_{i_+(t)}(t) = -\tilde{\alpha} m_{i_+(t)}(t); \qquad \dot{m}_i(t) = 0 \text{ for all } i \neq i_-, i \neq i_+.$$

From the first equation, for all $i \in \{1, \cdots, N\}$, $\dot{m}_i(t) \leq \tilde{\alpha}\max_j\{m_j(t)\}$, which implies that for all $i \in \{1, \cdots, N\}$, $m_i(t) \leq \max_j\{m_j^0\}e^{\tilde{\alpha}t}$.

From the second equation, for all $i \in \{1, \cdots, N\}$, $\dot{m}_i(t) \geq -\tilde{\alpha} m_i(t)$, which implies that $m_i(t) \geq \min_j\{m_j^0\}e^{-\tilde{\alpha}t}$.

We deduce that for all $t \leq t_\varepsilon$,

$$|u_{i_-(t)}(t)| \leq \tilde{\alpha}\frac{\max_j\{m_j^0\}}{\min_j\{m_j^0\}} \leq \tilde{\alpha}\frac{\max_j\{m_j^0\}e^{\tilde{\alpha}t_\varepsilon}}{\min_j\{m_j^0\}e^{-\tilde{\alpha}t_\varepsilon}} = \tilde{\alpha}\frac{\max_j\{m_j^0\}}{\min_j\{m_j^0\}}e^{2\tilde{\alpha}t_\varepsilon} := \alpha,$$

where α depends on δ, η, $(m_i^0)_{i\in\{1,\cdots,N\}}$ and t_ε. Since $|u_{i_+(t)}(t)| = \alpha \leq \tilde{\alpha}$ and for all $i \neq i_+(t), i_-(t)$, $|u_i(t)| = 0$, we deduce that $u \in U_\infty^\alpha$, which concludes the proof. □

Remark 2 The proof can be easily adapted to the case $u \in U_1^A \cap U_M$ by replacing α by A/N.

We have shown that any target position strictly within the initial convex hull of the system can be reached given sufficient control strength. The converse problem of determining the set of reachable positions given a control bound is much more difficult and remains open.

We now focus on designing feedback control strategies. Let us define the functional

$$X : t \mapsto X(t) = \|\bar{x}(t) - x^*\|^2.$$

We propose a gradient-descent control strategy to minimize instantaneously the time-derivative of X, i.e. we define $u \in U$ such that for almost all $t \in [0, T]$,

$$u(t) \in \arg\min_{v \in U} \frac{d}{dt} X^v(t). \tag{6}$$

We have

$$\frac{d}{dt} X = 2\langle \bar{x} - x^*, \dot{\bar{x}} \rangle = 2\langle \bar{x} - x^*, \frac{1}{M} \sum_{i=1}^{N} u_i m_i x_i \rangle = \frac{2}{M} \sum_{i=1}^{N} m_i \langle \bar{x} - x^*, x_i - x^* \rangle u_i \tag{7}$$

since $\sum_{i=1}^{N} u_i m_i x^* = 0$ if $u \in U$. Hence, for all $t \in \mathbb{R}^+$, we seek

$$\min_{u \in U} F_t(u)$$

where we define the linear functional F_t as $F_t : u \mapsto F_t(u) = \sum_{i=1}^{N} m_i(t) \langle \bar{x}(t) - x^*, x_i(t) - x^* \rangle u_i$. We minimize a linear functional on a convex set U. Hence the minimum is achieved at extremal points of U. Notice that the control set $U_\infty^\alpha \cap U_M$ is the intersection of the hypercube U_∞^α and of the hyperplane U_M. Similarly, the control set $U_1^A \cap U_M$ is the intersection of the diamond U_1^A and of the hyperplane U_M. These intersections are non-empty since U_∞^α, U_1^A and U_M contain the origin.

The condition $u \in U_M$ renders even this simple instantaneous-decrease control strategy not straightforward. Notice that despite the condition $u \in U_1^A$ that promotes sparse control, no control satisfying $u \in U_M$ can have just one active component. We will provide illustrations of this phenomenon in Sect. 6.

4 Control with Mass Variation

In this section, we remove the total mass conservation constraint on the control, and consider Problem 3 for $U = U_\infty^\alpha$ or $U = U_1^A$. Remark that this problem can be solved with the controls found in Sect. 3 (thus satisfying the mass conservation constraint). However we purposefully look for a different solution in order to exploit the larger control possibilities that appear due to the fewer constraints.

We first point out a fundamental difference in the behavior of the system compared to that of the previous section: with a varying total mass, one can break free of the convergence property stated in Properties 1.

Proposition 2 *Let (x, m) be a solution to (1). Then there exist mass dynamics ψ that do not satisfy Hypothesis 2, such that the system does not converge to consensus.*

Proof Consider the constant mass dynamics given by: $\psi_i(x, m) \equiv -R$ for all $i \in \{1, \cdots, N\}$. Then for all $i \in \{1, \cdots, N\}$, $m_i(t) = m_i^0 e^{-Rt}$ and we can compute:

$$\frac{d(\|x_i - x_i^0\|^2)}{dt} = \langle x_i - x_i^0, \sum_{j=1}^{N} \frac{2m_j}{M} a(\|x_i - x_j\|)(x_j - x_i)\rangle \leq 2\|x_i - x_i^0\|\delta e^{-Rt},$$

where δ was defined in (5). From this we get: $\|x_i - x_i^0\| \leq \frac{\delta}{R}(1 - e^{-tR})$. Hence for R big enough, each x_i is confined to a neighborhood of its initial position, which prevents convergence to consensus. □

Remark 3 As a consequence, in such cases the convex hull tends to a limit set $\Omega_\infty := \cap_{t\geq 0}\Omega(x(t))$ not restricted to a single point.

The dynamics of the barycenter of the system are now less trivial than in the previous section due to the total mass variation. Nevertheless, as previously, we prove approximate controllability to any target position strictly within the initial convex hull.

Theorem 2 *Let $(x_i^0)_{i\in\{1,\cdots,N\}} \in \mathbb{R}^{dN}$, $(m_i^0) \in (0, M)^N$ such that $\sum_{i=1}^N m_i^0 = M$ and let $x^* \in \mathring{\Omega}(x^0)$. Then for all $\varepsilon > 0$, there exists $\alpha > 0$, $t_\varepsilon \geq 0$ and $u \in U_\infty^\alpha \setminus U_M$ such that the solution to (4) satisfies: $\|\bar{x}(t_\varepsilon) - x^*\| \leq \varepsilon$.*

Proof Let $x^* \in \mathring{\Omega}(x^0)$ and let $\varepsilon > 0$. Then there exists $(\tau_i^0)_{i\in\{1,\cdots,N\}}$ with $\tau_i^0 \in [0, 1]^N$, $\sum_{i=1}^N \tau_i^0 = 1$ and $\tau_i^0 > 0$ for all $i \in \{1, \cdots, N\}$ such that

$$x^* = \sum_{i=1}^{N} \tau_i^0 x_i^0.$$

We will show that we can drive each weight m_i to a multiple $\kappa\tau_i^0$ of its target weight, while maintaining the positions withing close distance of the initial ones, ensuring that the target position remains in the shrinking convex hull. Define

$$\begin{cases} r_{\min} = \min\{\ln\left(\frac{m_i^0}{\tau_i^0}\right) \mid i \in \{1, \cdots, N\}\} \\ r_{\max} = \max\{\ln\left(\frac{m_i^0}{\tau_i^0}\right) \mid i \in \{1, \cdots, N\}\}. \end{cases}$$

Let $\tilde{\alpha} \geq \frac{\delta}{\varepsilon}$, with δ defined in (5) and let $\alpha > \tilde{\alpha} > 0$. Let $T := \frac{r_{\max} - r_{\min}}{\alpha - \tilde{\alpha}}$ and $\kappa := e^{r_{\min} - \tilde{\alpha}T}$. Now consider the constant control defined by: for all $i \in \{1, \cdots, N\}$,

$$u_i = -\frac{1}{T} \ln\left(\frac{m_i^0}{\kappa\tau_i^0}\right).$$

One can easily show that for all $i \in \{1, \cdots, N\}$, $-\alpha \leq u_i \leq -\tilde{\alpha}$, and furthermore, $m_i(T) = \kappa \tau_i^0$. From the proof of Properties 2, for all $t \in [0, T]$, $\|x_i(t) - x_i^0\| \leq \frac{\delta}{\tilde{\alpha}}$, where δ was defined in (5). From this we compute:

$$\|\bar{x}(T) - x^*\| = \left\| \frac{\sum_{i=1}^N m_i(T) x_i(T)}{\sum_{i=1}^N m_i(T)} - \sum_{i=1}^N \tau_i^0 x_i^0 \right\| = \left\| \sum_{i=1}^N \tau_i^0 (x_i(T) - x_i^0) \right\|$$
$$\leq \sum_{i=1}^N \tau_i^0 \left\| x_i(T) - x_i^0 \right\| \leq \frac{\delta}{\tilde{\alpha}} \leq \varepsilon,$$

which proves the theorem. □

Remark 4 As for Theorem 1, the proof can be easily adapted to the case $u \in U_1^A$ by replacing α by $\frac{A}{N}$.

As in the previous section, we design a feedback control strategy that minimizes the time-derivative of the functional X instantaneously. With a total mass now varying in time, we have:

$$\frac{d}{dt} X = \frac{2}{\sum_{i=1}^N m_i} \sum_{i=1}^N m_i \langle \bar{x} - x^*, x_i - \bar{x} \rangle u_i. \tag{8}$$

Since we removed the constraint $u \in U_M$, the control strategy minimizing $\frac{dX}{dt}$ is straightforward. For $u \in U_\infty^\alpha$, we have:

$$\begin{cases} u_i = -\alpha \text{ if } \langle \bar{x} - x^*, x_i - \bar{x} \rangle > 0 \\ u_i = \alpha \text{ if } \langle \bar{x} - x^*, x_i - \bar{x} \rangle < 0. \end{cases} \tag{9}$$

For $u \in U_1^A$, we define the set $I := \arg\max\{|m_i \langle \bar{x} - x^*, x_i - \bar{x} \rangle|, \quad i \in \{1, \cdots, N\}\}$, and we have:

$$\begin{cases} u_i = -\frac{A}{|I|} \text{sgn}(\langle \bar{x} - x^*, x_i - \bar{x} \rangle) \text{ if } i \in I \\ u_i = 0 \text{ otherwise,} \end{cases} \tag{10}$$

where $|\cdot|$ represents the cardinality of a set.

5 Mean-Field Limit

5.1 Mean-Field Limit of Mass-Varying Dynamics Without Control

In this section, we recall the definition of mean-field limit. Consider System (1). The goal of the mean-field limit is to describe the behavior of the system when the number of agents N tends to infinity. Instead of following the individual trajectory of each individual, we aim to describe the group by its limit density μ, which belongs to $\mathcal{M}(\mathbb{R}^d)$, the set of Radon measures with finite mass. We endow $\mathcal{M}(\mathbb{R}^d)$ with the topology of the weak convergence of measures, i.e.

$$\mu_i \rightharpoonup_{i\to\infty} \mu \quad \Leftrightarrow \quad \lim_{i\to\infty} \int f d\mu_i = \int f d\mu$$

for all $f \in C_c^\infty(\mathbb{R}^d)$. Let $\mu_0 \in \mathcal{M}(\mathbb{R}^d)$. We consider the following transport equation for μ:

$$\begin{cases} \partial_t \mu + \nabla \cdot (V[\mu]\mu) = h[\mu] \\ \mu(0) = \mu_0. \end{cases} \tag{11}$$

We recall conditions for well-posedness of (11), see [18]:

Hypothesis 3 The function $V[\cdot] : \mathcal{M}(\mathbb{R}^d) \to C^1(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ satisfies

- $V[\mu]$ is uniformly Lipschitz and uniformly bounded
- V is uniformly Lipschitz with respect to the generalized Wasserstein distance (see [18])

Hypothesis 4 The source term $h[\cdot] : \mathcal{M}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d)$ satisfies

- $h[\mu]$ has uniformly bounded mass and support
- h is uniformly Lipschitz with respect to the generalized Wasserstein distance (see [18])

We now recall the definition of mean-field limit.

Definition 2 Let $(x, m) \in \mathbb{R}^{dN} \times (\mathbb{R}^+)^N$ be a solution to (1). We denote by μ_N the corresponding empirical measure defined by

$$\mu_N(t) = \frac{1}{M} \sum_{i=1}^{N} m_i(t) \delta_{x_i(t)}.$$

The transport equation (11) is the mean-field limit of the collective dynamics (1) if

$$\mu_N(0) \rightharpoonup_{N\to\infty} \mu(0) \quad \Rightarrow \quad \mu_N(t) \rightharpoonup_{N\to\infty} \mu(t)$$

where μ is the solution to (11) with initial data $\mu(0)$.

The definition of empirical measure requires a crucial property of the finite-dimensional system (1): that of **indistinguishability** of the agents. Indeed, notice that there isn't a one-to-one relationship between the set of empirical measures (finite sums of weighted Dirac masses) and the set of coupled positions and weights $(x, m) \in \mathbb{R}^{dN} \times (\mathbb{R}^+)^N$. For instance, two pairs $(x(t), m(t)) \in \mathbb{R}^{dN} \times (\mathbb{R}^+)^N$ and $(y(t), q(t)) \in \mathbb{R}^{d(N-1)} \times (\mathbb{R}^+)^{N-1}$ satisfying $x_1^0 = x_N^0 = y_1^0$, $m_1^0 + m_N^0 = q_1^0$ and $(x_i^0, m_i^0) = (y_i^0, q_i^0)$ for all $i \in \{2, \cdots, N-1\}$ correspond to the same empirical measure. Hence if we want the concept of mean-field limit to make sense, we must consider discrete systems that give the same dynamics to $(x(t), m(t))$ and $(y(t), q(t))$.

Definition 3 Let $t \mapsto (x(t), m(t)) \in \mathbb{R}^{dN} \times (\mathbb{R}^+)^N$ and $t \mapsto (y(t), q(t)) \in \mathbb{R}^{d(N-1)} \times (\mathbb{R}^+)^{N-1}$ be two solutions to system (1). We say that *indistinguishability* is satisfied if

$$\begin{cases} x_1^0 = x_N^0 = y_1^0 \\ m_1^0 + m_N^0 = q_1^0 \\ x_i = y_i, \ i \in \{2, \cdots, N-1\} \\ m_i = q_i, \ i \in \{2, \cdots, N-1\} \end{cases} \Rightarrow \begin{cases} x_1 \equiv x_N \equiv y_1 \\ m_1 + m_N \equiv q_1 \\ x_i \equiv y_i, \ i \in \{2, \cdots, N-1\} \\ m_i \equiv q_i, \ i \in \{2, \cdots, N-1\}. \end{cases}$$

Indistinguishability is a strong property, and it is not necessarily satisfied by the general function ψ defining the weights' dynamics in (1). We refer the reader to [2, 17, 20] for examples of mass dynamics satisfying or not the indistinguishability property. From here onward, we will focus on the following particular form of mass dynamics that does satisfy indistinguishability:

$$\psi_i(x, m) = \frac{1}{M} \sum_{j=1}^{N} m_j S(x_i, x_j), \tag{12}$$

with $S \in C(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$.

In order for a transport equation to be the mean-field limit of a finite-dimensional system, it is sufficient for it to satisfy the following two properties (see [23]):

(i) When the initial data μ^0 is an empirical measure μ_N^0 associated with an initial data $(x^0, m^0) \in \mathbb{R}^{dN} \times \mathbb{R}^N$ of N particles, then the dynamics (11) can be rewritten as the system of ordinary differential equations (1).
(ii) The solution $\mu(t)$ to (11) is continuous with respect to the initial data μ^0.

The following holds:

Proposition 3 *Consider System* (1) *with mass dynamics given by* (12)*, where* $S \in C(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ *is skew-symmetric:* $S(x, y) = -S(y, x)$*. Then its mean-field limit is the transport equation with source* (11) *with the interaction kernel*

$$V[\mu](x) = \int_{\mathbb{R}^d} a(\|x - y\|)(y - x)d\mu(y) \tag{13}$$

and the source term

$$h[\mu](x) = \int_{\mathbb{R}^d} S(x, y)d\mu(y)\mu(x). \tag{14}$$

The proof of this result should consist of proving the two properties (i) and (ii) above. Notice that well-posedness of (11)–(13)–(14) and continuity with respect to the initial data cannot be obtained by applying directly the results of [18] since h does not satisfy Hypothesis 4. Nevertheless, well-posedness and continuity can be proven, using the total conservation of mass coming from the skew-symmetric property of S, see [20]. In the present paper, we focus on proving the first property (i).

Proof We prove that the transport equation (11) with the vector field (13) and the source term (14) satisfies the property (i) above. Let $(x, m) : \mathbb{R}^+ \to \mathbb{R}^{dN} \times \mathbb{R}^N$ be the solution to the system (1) with the weight dynamics given by (12) and initial data $(x^0, m^0) \in \mathbb{R}^{dN} \times \mathbb{R}^N$. We show that the empirical measure $\mu_N(t, x) = \frac{1}{M}\sum_{i=1}^N m_i(t)\delta_{x_i(t)}(x)$ is the solution to the PDE (11)–(13)–(14) with initial data $\mu_N^0(x) = \sum_{i=1}^N m_i^0 \delta_{x_i^0}(x)$. Let $f \in C_c^\infty(\mathbb{R}^d)$. We show that

$$\frac{d}{dt}\int f d\mu_N - \int \nabla f \cdot V[\mu_N]d\mu_N = \int f dh[\mu_N]. \tag{15}$$

We compute each term independently. Firstly, we have:

$$\begin{aligned}\frac{d}{dt}\int f d\mu_N =& \frac{d}{dt}\frac{1}{M}\sum_{i=1}^N m_i f(x_i) = \frac{1}{M}\sum_{i=1}^N (\dot{m}_i f(x_i) + m_i \dot{x}_i \cdot \nabla f(x_i)) \\ =& \frac{1}{M^2}\sum_{i=1}^N\sum_{j=1}^N m_i m_j \Big[S(x_i, x_j)f(x_i) + a(\|x_i - x_j\|)(x_j - x_i)\cdot \nabla f(x_i)\Big].\end{aligned} \tag{16}$$

Secondly,

$$\int \nabla f \cdot V[\mu_N] d\mu_N = \int \nabla f(x) \cdot \int a(\|x-y\|)(x-y) d\mu_N(y) d\mu_N(x)$$
$$= \frac{1}{M^2} \sum_{i=1}^{N} \sum_{j=1}^{N} m_i m_j a(\|x_i - x_j\|)(x_j - x_i) \cdot \nabla f(x_i). \tag{17}$$

Thirdly,

$$\int f dh[\mu_N] = \int f(x) \int S(x,y) d\mu_N(y) d\mu_N(x) = \frac{1}{M^2} \sum_{i=1}^{N} \sum_{j=1}^{N} m_i m_j f(x_i) S(x_i, x_j). \tag{18}$$

Putting together (16), (17), and (18) and using the fact that (x, m) satisfies (1)–(12), we deduce that μ_N satisfies (15)–(13)–(14). □

The general weight dynamics (12) include special cases studied in previous works. Indeed:

- if $S(x,y) := S_0(x)$, the mass dynamics can be simply written as $h[\mu](x) = |\mu| S_0(x) \mu(x)$ (see [18])
- if $S(x,y) := S_1(y-x)$, the mass dynamics can be rewritten as the convolution $h[\mu] = (S_1 * \mu)\mu$ (see [18])
- if $\dot{m}_i = \frac{1}{M} \sum_{j=1}^{N} \sum_{k=1}^{N} m_j m_k S(x_i, x_j, x_k)$, we can show in a similar way that the mean-field limit is the PDE (11) with the source term

$$h[\mu](x) = \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} S(x,y,z) d\mu(y) d\mu(z) \right) \mu(x).$$

  In particular, this applies to the following mass-conserving dynamics, which are a slight modification of Model 2 proposed in [17]:

$$\dot{m}_i = \frac{m_i}{M} \left(\sum_{j=1}^{N} m_j a(\|x_i - x_j\|) \|x_i - x_j\| - \frac{1}{M} \sum_{j=1}^{N} \sum_{k=1}^{N} m_j m_k a(\|x_j - x_k\|) \|x_j - x_k\| \right)$$

where $S(x_i, x_j, x_k) := \frac{1}{M} (a(\|x_i - x_j\|) \|x_i - x_j\| - a(\|x_j - x_k\|) \|x_j - x_k\|)$.

5.2 Control Problem

From the mean-field limit of the system without control, we extract a natural control problem corresponding to the mean-field limit of (3). Consider the controlled PDE:

$$\begin{cases} \partial_t \mu + \nabla \cdot (V[\mu]\mu) = \mu u \\ \mu(0) = \mu_0. \end{cases} \tag{19}$$

We define the kinetic variance $\mathcal{X}(t) = \| \int_{\mathbb{R}^d} (x - x^*) d\mu(t, x) \|^2$. We seek a control function $u : \mathbb{R}^+ \times \mathbb{R}^{dN}$ that minimizes instantaneously $\frac{d}{dt}\mathcal{X}(t)$. Similarly to Sect. 3, we can further restrict the set of controls to functions satisfying

$$\int_{\mathbb{R}^d} u(t, x) d\mu(t, x) = 0 \quad \text{for a.e } t \in \mathbb{R}^+.$$

We can also extend the L^1 and L^∞ bounds on the control to the mean-field setting:

- L^∞ condition: $\|u\|_{L^\infty(\mathbb{R}^+ \times \mathbb{R}^d)} \leq \alpha$
- L^1 condition: $\|u(t, \cdot)\|_{L^1(\mathbb{R}^d)} \leq A$

We can compute:

$$\begin{aligned} \frac{d}{dt}\mathcal{X}(t) =& 2\langle \int_{\mathbb{R}^d} (x - x^*) d\mu(t, x), \frac{d}{dt} \int_{\mathbb{R}^d} (x - x^*) d\mu \rangle = 2\langle \int_{\mathbb{R}^d} (x - x^*) d\mu(t, x), \\ & - \int_{\mathbb{R}^d} (x - x^*) d(\nabla \cdot (V[\mu]\mu)) \rangle + 2\langle \int_{\mathbb{R}^d} (x - x^*) d\mu(t, x), \int_{\mathbb{R}^d} (x - x^*) u(t, x) d\mu \rangle. \end{aligned}$$

6 Numerical Simulations

We now provide simulations of the evolution of System (4) with the various control strategies presented in Sects. 3 ($u \in U_\infty^\alpha \cap U_M$ and $u \in U_1^A \cap U_M$) and 4 ($u \in U_\infty^\alpha$ and $u \in U_1^A$).

Four simulations were run with the same set of initial conditions $x^0 \in \mathbb{R}^{dN}$ for $d = 2$, $N = 10$, and control bounds $\alpha = 2$ and $A = 10$. In each simulation, the control maximizes the instantaneous decrease of the functional X, with one of the various constraints exposed in Sects. 3 and 4. Figure 1 shows that in all cases, the control successfully steers the weighted barycenter $\bar{x}$ to the target position x^*. The evolution of the functional $t \mapsto \|\bar{x}(t) - x^*\|$ (Fig. 4 (right)) shows that the target is reached faster with controls that allow for mass variation than for controls constrained to the set U_M. Figure 2 shows the evolution of each agent's individual weight for each of the four cases of Fig. 1. Interestingly, when mass variation is

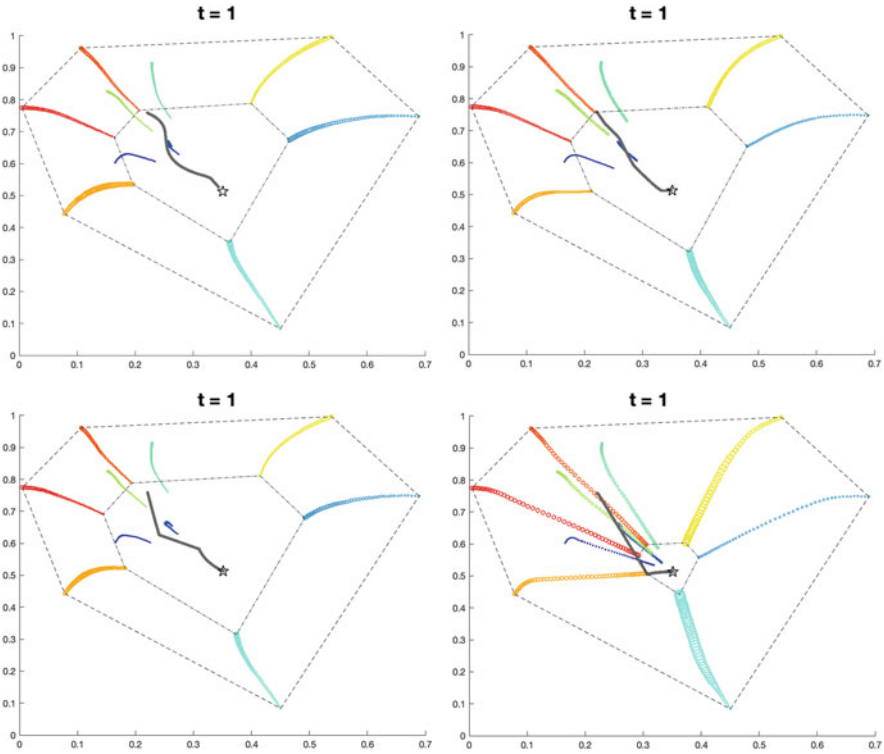

Fig. 1 Trajectories of the positions $x_i(t)$ in $\mathbb{R}^2$ corresponding to the controlled system (4) with $N = 10$ and $a : s \mapsto e^{-s^2}$. The top row corresponds to controls satisfying $u \in U_M$ (Sect. 3) while the second row corresponds to controls allowing total mass variation (Sect. 4). In each row, the left column corresponds to $u \in U_\infty^\alpha$ and the right one corresponds to $u \in U_1^A$. In each plot, different agents are represented by different colors, and the size of each dot is proportional to the weight of the corresponding agent at that time. The gray dotted trajectory represents the weighted barycenter $\bar{x}$. The black star represents the target position, inside the convex hull of the initial positions (dashed polygon). The convex hull of the positions at final time is represented by the dot-dashed polygon

allowed, we observe a general decrease in the total mass of the system in the case $u \in U_\infty^\alpha$ (dotted grey line, Fig. 2-left) and a general increase in the case $u \in U_1^A$ (dotted grey line, Fig. 2-right). Figure 3 shows the control values $u_i(t)$ for each $i \in \{1, \cdots, N\}$ and each $t \in [0, 1]$. Notice that in the case of mass-preserving control $u \in U_M$ (top row), the controls do not saturate the constraints $u \in U_\infty^\alpha$ or $u \in U_1^A$. In the case of varying total mass, as shown in Sect. 4, the control strategies minimizing $\frac{dX}{dt}$ saturate the constraints.

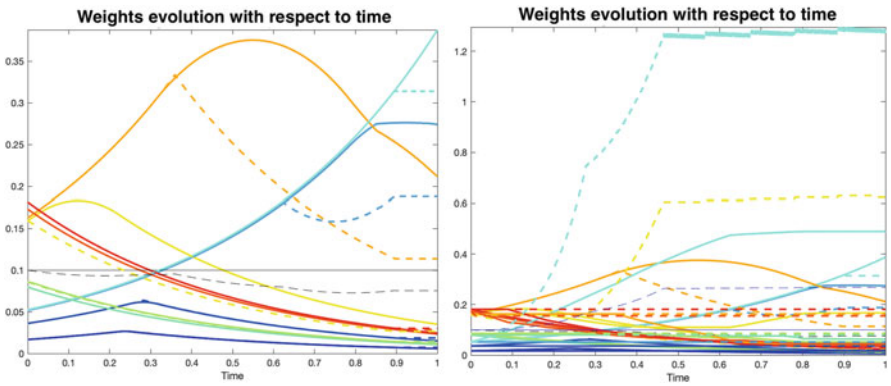

Fig. 2 Evolution of the weights $t \mapsto m_i(t)$ corresponding to control strategies $u \in U_\infty^\alpha$ (left) and $u \in U_1^A$ (right). In each plot, the continuous lines correspond to the mass-preserving control $u \in U_M$ of Sect. 3, and the dashed lines to the controls of Sect. 4. Each colored line (respectively dashed or continuous) shows the evolution of the corresponding colored agent of Fig. 1, and the grey lines represent the evolution of the average weight $\frac{1}{N}\sum_{i=1}^N m_i$

Figure 4 (left) shows that the constraint $u \in U_1^A$ promotes a *sparse* control strategy. A control is said to be sparse if it is active only on a small number of agents. As mentioned in Sect. 3, mass-varying controls cannot be strictly sparse, and need to have at least two non-zero components at each time. Indeed, the control strategy $u \in U_1^A \cap U_M$ has either two or three active components at all time.

7 Conclusion

In this paper we aimed to control to a fixed consensus target a multi-agent system with time-varying influence, by acting only on each agent's weight of influence. We proved approximate controllability of the system to any target position inside the convex hull of the initial positions. We then focused on designing control strategies with various constraints on the control bounds and on the total mass of the system.

We also presented the mean-field limit of the discrete model for general mass dynamics that satisfy the indistinguishability property. The population density satisfies a transport equation with source, where both the source term and the velocity are non-local.

The combination of our analysis with numerical simulations allows us to compare the control performances of the four strategies. Firstly, the control strategies allowing total mass variation are more efficient than the control strategies conserving the total mass, as the weighted barycenter reaches the target position faster.Interestingly, this is not obvious a priori from Eqs. (7) and (8), as the time derivatives of the functional $X = \|\bar{x} - x^*\|^2$ are of the same order of magnitude in

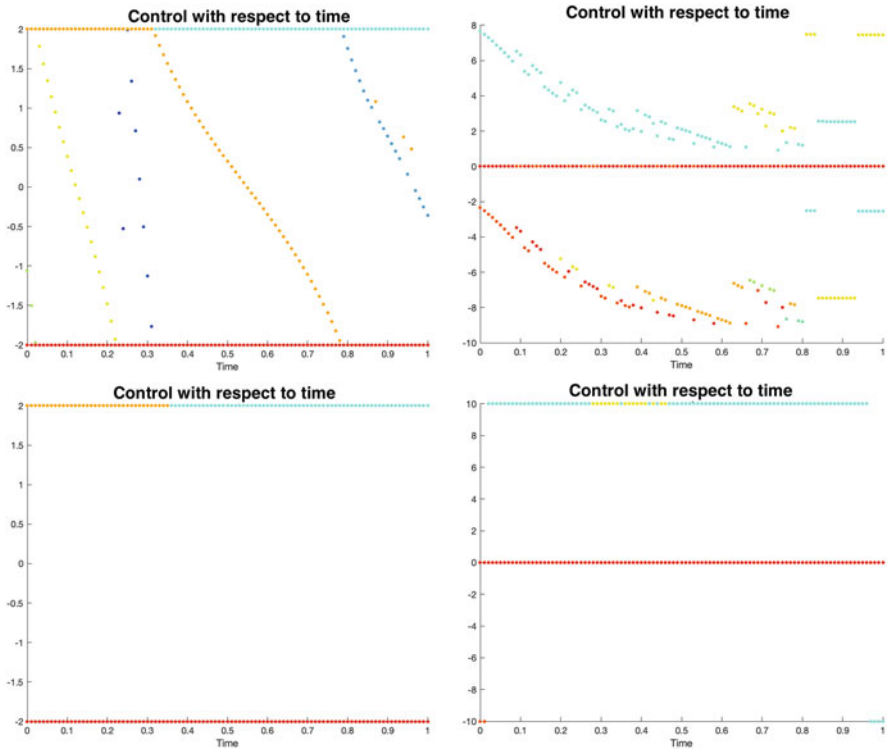

Fig. 3 Evolution of the control functions $t \mapsto u_i(t)$ corresponding to the systems of Fig. 1. The top row corresponds to controls satisfying $u \in U_M$ (Sect. 3) while the second row corresponds to controls allowing total mass variation (Sect. 4). In each row, the left column corresponds to $u \in U_\infty^\alpha$ and the right one corresponds to $u \in U_1^A$. Each control function u_i is colored according to the corresponding agent x_i of Fig. 1

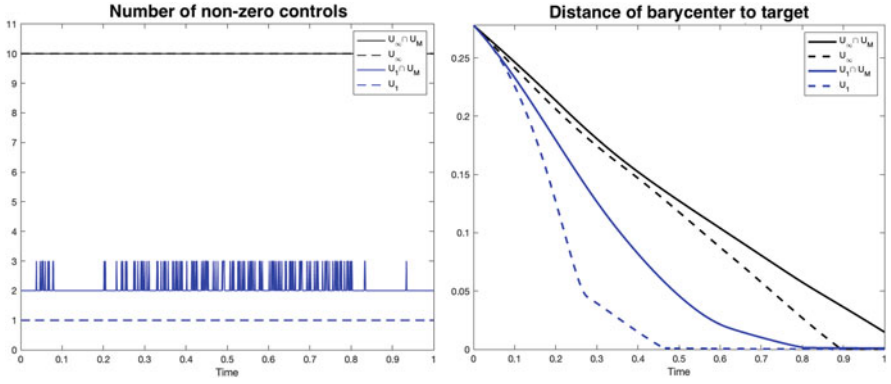

Fig. 4 Left: Evolution of the number of active components of the control with the various strategies corresponding to Fig. 1. Right: Distance of the barycenter to the target position $t \mapsto \|\bar{x}(t) - x^*\|$

the two cases. We also remark that the controls allowing mass variation can either increase or decrease the total mass of the system.

The constraint $u \in U_1^A$ is usually enforced to promote sparsity (see [8, 19]), that is the activation at any given time of as few control components as possible. However, the added constraint $u \in U_M$ renders strict sparsity impossible, and we already remarked that in order to preserve the total mass, the control has to be active on at least two components at any given time. Simulations shows that indeed, the control $u \in U_1^A \cap U_M$ oscillates between two and three active components, whereas the control $u \in U_1^A$ maintains strict sparsity. On the other hand, the controls $u \in U_\infty^\alpha$ and $u \in U_\infty^\alpha \cap U_M$ act simultaneously on all components at all time.

Although in the illustrating simulations, all four controls manage to drive the system's weighted barycenter to the target position x^*, this would not have necessarily been achievable with either a target closer to the initial convex hull boundary or with stricter control bounds α and A. The question of determining the set of achievable targets given an initial distribution of positions and weights and control bounds remains open and is an intriguing future direction of this work, as is the control of the mean-field model obtained as limit of the finite-dimensional one when the number of agents tends to infinity.

References

1. Aydoğdu, A., McQuade, S., Pouradier Duteil, N.: Opinion dynamics on Riemannian manifolds. Netw. Heterog. Media **12**(3), 489–523 2017
2. Ayi, N., Pouradier Duteil, N.: Mean-field and graphs limits for collective dynamics models with time-varying weights, submitted (2021)
3. Bellomo, N., Soler, J.: On the mathematical theory of the dynamics of swarms viewed as complex systems. Math. Models Methods Appl. Sci. **22**, 1140006 (2012)
4. Bullo, F., Cortés, J., Martínez, S.: Distributed control of robotic networks: A mathematical approach to motion coordination algorithms. Princeton University, Princeton (2015)
5. Caponigro, M., Lai, A.C., Piccoli, B.: A nonlinear model of opinion formation on the sphere. Discrete Contin. Dynam. Systems Ser. A **35**(9), 4241–4268 (2015)
6. Couzin, I., Krause, J., James, R., Ruxton, G., Franks, N.: Collective memory and spatial sorting in animal groups. J. Theor. Biol. **218**(1), 1–11 (2002)
7. De Groot, M.H.: Reaching a consensus. J. Am. Stat. Assoc. **69**, 118–121 (1974)
8. Fornasier, M., Piccoli, B., Rossi, F.: Mean-field sparse optimal control. In: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2014)
9. French, J.R.P.: A formal theory of social power. Psychol. Rev. **63**, 181–194 (1956)
10. Giardina, I.: Collective behavior in animal groups: theoretical models and empirical studies. HFSP J. **2**(4), 205–219 (2008)
11. Harary, F.: A criterion for unanimity in french's theory of social power. In: Cartwright, D. (ed.) Studies in Social Power (1959)
12. Hegselmann, R., Flache, A.: Understanding complex social dynamics—a plea for cellular automata based modelling. J. Artif. Soc. Soc. Simul. **1**(3), 1 (1998)
13. Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence models, analysis, and simulation. J. Artif. Soc. Soc. Simul. **5**(3), 1–24 (2002)
14. Justh, E.W., Krishnaprasad, P.S.: Equilibria and steering laws for planar formations. Syst. Control Lett. **52**(1), 25–38 (2004)

15. Krause, U.: A discrete nonlinear and non—autonomous model of consensus formation. In: Elaydi, S., Ladas, G., Popenda, J., Rakowski, J. (eds.) Communications in Difference Equations, pp. 227–236. Gordon and Breach Publication, Amsterdam (2000)
16. Krause, J., Ruxton, G.: Living in groups. In: Oxford Series in Ecology and Evolution. Oxford University, New York (2002)
17. McQuade, S., Piccoli, B., Pouradier Duteil, N.: Social dynamics models with time-varying influence. Math. Models Methods Appl. Sci. **29**(04), 681–716 (2019)
18. Piccoli, B., Rossi, F.: Measure-Theoretic Models for Crowd Dynamics. Springer International Publishing, Berlin (2018)
19. Piccoli, B., Pouradier Duteil, N., Trélat, E.: Sparse control of Helgselmann-Krause models: Black hole and declusterization. SIAM J. Control Optim. **57**(4), 2628–2659 (2019)
20. Pouradier Duteil, N.: Mean-field limit of collective dynamics with time-varying weights. Preprint (2021)
21. Tanner, H.G., Jadbabaie, A., Pappas, G.J.: Flocking in fixed and switching networks. IEEE Trans. Autom. Control **52**(5), 863 (2007)
22. Tomlin, C., Pappas, G.J., Sastry, S.: Conflict resolution for air traffic management: A study in multiagent hybrid systems. IEEE Trans. Autom. Control **43**(4), 509–521 (1998)
23. Villani, C.: Limite de champ moyen. In: Cours de DEA, 2001–2002. ENS Lyon, Lyon

Kinetic Modelling of Autoimmune Diseases

M. Piedade M. Ramos, C. Ribeiro, and Ana Jacinta Soares

Abstract In this paper, we review previous results obtained by the authors, concerning the mathematical modelling of autoimmune diseases when the kinetic theory approach is used in order to describe the microscopic interactions between cells. Three cell populations are considered and the distribution function of each population depends on the biological activity variable defining the functional state relevant for that population. We revisit the wellposedness of the kinetic system and focus our study on the numerical simulations with the kinetic system in view of investigating the sensitivity of the solution to certain parameters of the model with biological significance.

Keywords Mathematical modelling · Kinetic theory · Cellular interactions · Autoimmune diseases

1 Introduction

The main job of the immune system is to protect the organism against disease whether caused by external factors such as bacteria and viruses, or internal aspects such as the existence of cancerous tumour cells in the human body. In order to provide this protection, the main players of the immune system must distinguish between pathogens and healthy tissue.

An autoimmune disease is an illness in which the immune system wrongly attacks healthy cells by reacting to self-antigens. In many cases it is chronic, and patients alternate between periods of relapse, having suffering symptoms, and periods of remittance, in which symptoms are absent.

Autoimmune diseases can affect just about any part of the body, and depending on which part of the body is affected by the such a perverse mechanism, a different

M. P. M. Ramos · C. Ribeiro · A. J. Soares (✉)
Centre of Mathematics (CMAT), Universidade do Minho, Braga, Portugal
e-mail: mpr@math.uminho.pt; cribeiro@math.uminho.pt; ajsoares@math.uminho.pt

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_13

autoimmune disease can be identified. The consequence of this is that over one hundred types of autoimmune diseases exist, some of the most common include type 1 diabetes, rheumatoid arthritis, multiple sclerosis, lupus, psoriasis, thyroid diseases, and inflammatory bowel disease. Although these diseases are not, in general, deadly, they are, in most cases, chronic. The chronic nature of autoimmunity can have serious implications on the quality of life of patients suffering from these diseases. Unfortunately, in spite of a significant increase in the number of patients suffering from these conditions, particularly in the developed world, much about the process of autoimmunity remains a mystery, although environmental changes associated with industrialization have been long suspected as well as genetic factors. See, for example, papers [1–4].

Motivated by the idea of developing a mathematical model in order to describe, in a rigorous way, the complex dynamics of the variables involved in some autoimmune disease, we have initiated a research project with this objective in mind. We have proposed in paper [5] a rather simple, but mathematically robust, model with the aim of describing the immune system interactions in the context of autoimmune disease. The interacting populations are self-antigen presenting cells, self reactive T cells and the set of immunosuppressive cells consisting of Regulatory T (Treg) cells and Natural Killer (NK) cells. In paper [5], we have developed a rather complete qualitative analysis of the model equations and investigated the existence of biologically realistic solutions. Then, in paper [6], a new model has been proposed by considering a further population of IL-2 cytokines and an artificial inlet of external drug therapy with the aim of studying optimal policies for the immunotherapeutic treatment of autoimmune diseases. Paper [6] focus on the macroscopic formulation of this new model, whereas paper [7] introduces the kinetic system approach and exploits the corresponding cellular dynamics. We believe that the kinetic approach, where the model is developed at the cellular scale, can give some insights concerning the biological processes involved in autoimmunity.

In these proceedings, we revisit the model proposed in [5] and summarize the results there obtained. Then we further develop a sensitivity analysis of the parameters involved in the model equations in order to investigate which trends and outcomes, that are common in autoimmune diseases, can be replicated with our numerical simulations. On the one hand, the sensitivity analysis presented here studies the effect of immunotolerance on the evolution of the main populations of cells involved in autoimmunity by, for example, decreasing or increasing certain proliferative parameters defined in the model and on the other hand it shows the effect of immunosupression in the evolution of the same populations by changing certain destructive parameters appearing in the model. A sensitivity analysis of the model to certain conservative parameters is also given, showing the effect of increasing or decreasing these parameters on the number of more active cells participating in the process.

To the best of our knowledge, only few contributions are known on the mathematical modelling of the process of autoimmunity. Some examples of these models prior to our work can be found in [8–10]. On the other hand, several well-known

studies on the mathematical modeling of the tumour-immune system interactions can be found in [11–15].

The content of these proceedings is organized as follows. In Sect. 2 we briefly describe how the immune system can be represented within a mathematical framework, introducing the cellular populations considered in our model and their main role in the dynamics. Then, in Sects. 3 and 4, we revisit the model proposed in [5] and summarize the results concerning the wellposedeness of the kinetic system. Section 5 is devoted to the numerical simulations and their biological interpretation and contains a sensitivity analysis of the parameters involved in the model equations. Finally, in Sect. 6 we state our conclusions and present future ideas in terms of research perspectives.

2 The Mathematical Representation of the Immune System

The immune system can be considered, at the cellular level, as a system constituted by a large number of cells belonging to different interacting populations, and therefore a kinetic theory approach can be used to describe the dynamics of the populations.

In our model, we consider three interacting cell populations p_i, $i = 1, 2, 3$, that are involved in the development of autoimmunity, namely the population p_1 of SAPCs (self-antigen presenting cells), the population p_2 of SRTCs (self-reactive T cells), and the population p_3 of ISCs (immunosuppressive cells).

These populations interact at the cellular level, and the relevant effects that are considered in our description are the following.

- SAPCs transport self-antigens to their encounter with SRTCs.
- SRTCs are activated when they encounter a SAPC that has digested a self-antigen.
- ISCs regulate the activity of SRTCs and SAPCs.

2.1 The Functional Activity at the Cellular Level

The functional state of each population is described by a positive real variable $u \in [0, 1]$, called activation variable or activity, whose biological meaning is characterized as follows.

- The activity u of SAPCs is the ability to stimulate and activate SRTCs. When $u = 0$, SAPCs do not activate SRTCs and, therefore, any autoimmune response is induced in the body.
- The activity u of SRTCs is the ability of promoting the secretion of cytokines which, in turn, can induce an inflammatory process. When $u = 0$, SRTCs do

not produce cytokines, meaning that SRTCs are not sensitive to the stimulus by SAPCs and no inflammatory process is triggered.
- The activity u of ISCs is the ability to inhibit the autoimmune response by either suppressing the activity of SAPCs and SRTCs or eliminating SAPCs or SRTCs. When $u = 0$, the ISCs are neither able to inhibit the activity of SAPCs and SRTCs nor to eliminate SAPCs or SRTCs.

2.2 The Cellular Interactions

The dynamics at the cellular level is modelled under the following assumptions.

(i) Interactions are homogeneous in space and instantaneous modify the state of the participating cells.
(ii) Only binary interactions between cells of different populations are significant for the evolution of the system.
(iii) Interactions among cells of populations p_1 (SAPCs), p_2 (SRTCs) and p_3 (ISCs) can create SAPCs, SRTCs and ISCs (proliferative type), or destroy SAPCs and SRTCs (destructive type), and they can also simply change the activity of SAPCs and SRTCs (conservative type).
(iv) The population p_3 (ISCs) is homogeneous with respect to its biological activity, so that interactions involving ISCs can be only proliferative or destructive type.

Assumption (i) indicates that the interactions occur without time delay. Assumption (ii) is rather natural and common when modelling biological systems, and means that interactions involving more than two cells are not effective in our model. Assumption (iii) is motivated by the immunobiology associated to autoimmune diseases. We consider that interactions among cells of populations p_1 (SAPCs), p_2 (SRTCs) and p_3 (ISCs) can create SAPCs, SRTCs and ISCs (proliferative type), or destroy SAPCs and SRTCs (destructive type), and they can also simply change the activity of SAPCs and SRTCs (conservative type). In fact, during an immune response, a proliferation of both SRTCs and ISCs occurs and an increase of circulating APCs also occurs. Simultaneously, the role of ISCs is to control proliferation of both magenta SRTCs and SAPCs and, decrease their activity. Assumption (iv) results from the fact that we do not consider internal degrees of freedom for ISCs population. In fact, we do not consider the impact of the cellular interactions on the activity of both Treg and NK cells and, therefore, the population of ISCs is considered homogeneous with respect to its biological activity.

The admissible interactions in our model are described as follows.

- Interactions between SAPCs and SRTCs can be of conservative type, increasing the activity of both SAPCs and SRTCs, of proliferative type, enlarging the number of SRTCs and also that of SAPCs.

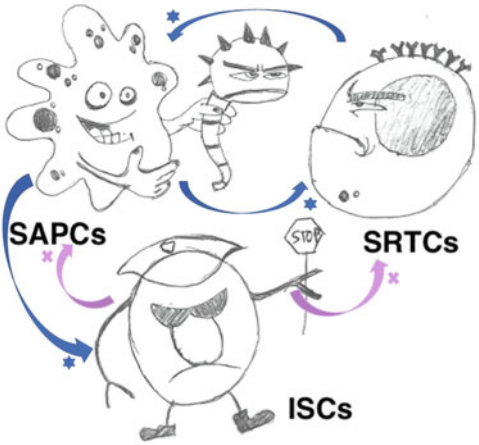

Fig. 1 *Illustration of the immune system interactions among SAPCs, SRTCs and ISCs.* Proliferative interactions are represented by blue starred arrows whereas destructive interactions are represented by purple crossed arrows

- Interactions between SAPCs and ISCs can be of conservative type, decreasing the activity of SAPCs, of proliferative type, enlarging the number of ISCs, as well as of destructive type, decreasing the number SAPCs.
- Interactions between SRTCs and ISCs can be of conservative type, decreasing the activity of SRTCs, and of destructive type, decreasing the number SRTCs.

The populations considered in our biological system and the non-conservative interactions among them are illustrated in Fig. 1. The proliferation of SRTCs by stimulation by SAPCs (blue starred arrow) induces an inflammatory response, in which the immune system mistakenly attacks the body. A cytokine storm produced by SRTCs increases the number of SAPCs (blue starred arrow) which, in turn, will activate more SRTCs. Additionally, ISCs, on the one hand, downgrade the function of both SAPCs (purple crossed arrow) and SRTCs (purple crossed arrow) and, on the other hand, eliminate both SAPCs and SRTCs.

3 The Kinetic Model for Autoimmune Diseases

The overall state of the biological system is described by the distribution functions associated to the populations p_1, p_2, p_3, namely $f_i : [0, \infty] \times [0, 1] \to \mathbb{R}^+$, $i = 1, 2, 3$, such that $f_i(t, u)$ gives the expected number of cells of population p_i with activity u at time t. Integration of each function f_i over the activity variable leads to the number density of p_i population,

$$n_i(t) = \int_0^1 f_i(t, u)\mathrm{d}u, \qquad i = 1, 2, 3, \tag{1}$$

which defines the expected number of cells of population p_i at time t.

Note that, as a consequence of Assumption D introduced in Sect. 2.2, the distribution function of the population p_3 is independent of its functional state, that is $f_3 = f_3(t)$.

The time evolution of the distribution functions f_i is described by the kinetic equations, that require a detailed description of the interaction balance operators, regarding the encounter rates and transition probability densities of cells in conservative interactions, as well as the proliferation rates and destructive rates of cell of different populations. See paper [5], where the complete structure of the kinetic system is explained in detail.

The kinetic system consists of the following coupled integro-differential equations

$$\begin{aligned}\frac{\partial f_1}{\partial t}(t,u) = {} & 2c_{12}\int_0^u (u-v)\,f_1(t,v)\mathrm{d}v\int_0^1 f_2(t,w)\mathrm{d}w - c_{12}(u-1)^2 f_1(t,u)\int_0^1 f_2(t,w)\mathrm{d}w \\ & + 2c_{13}f_3(t)\int_u^1 (v-u)\,f_1(t,v)\mathrm{d}v - c_{13}u^2 f_1(t,u)\,f_3(t) \\ & + p_{12}\,f_1(t,u)\int_0^1 f_2(t,w)\mathrm{d}w - d_{13}\,f_1(t,u)\,f_3(t), \end{aligned} \tag{2}$$

$$\begin{aligned}\frac{\partial f_2}{\partial t}(t,u) = {} & 2c_{21}\int_0^u (u-v)\,f_2(t,v)dv\int_{w^\star}^1 f_1(t,w)dw - c_{21}(u-1)^2 f_2(t,u)\int_{w^*}^1 f_1(t,w)dw \\ & + 2c_{23}f_3(t)\int_u^1 (v-u)\,f_2(t,v)dv - c_{23}u^2 f_2(t,u)\,f_3(t) \\ & + p_{21}f_2(t,u)\int_0^1 f_1(t,w)dw - d_{23}f_2(t,u)\,f_3(t), \end{aligned} \tag{3}$$

$$\frac{df_3}{dt}(t) = p_{31}f_3(t)\int_0^1 f_1(t,w)dw, \tag{4}$$

where parameters p_{ij}, d_{ij} and c_{ij} indicate constant rates of proliferative, destructive and conservative interactions, respectively, and parameter $w^\star \in]0,1[$ describes the tolerance of SRTCs towards self-antigens, in the sense that the greater the value of w^* the less efficient are SAPCs in increasing the activity of SRTCs after encounter. We have considered that during proliferative encounters, cloned cells inherit the same aggressive state as their mother cell, at a constant proliferation rate, and, additionally, that the destructive encounters occur at a constant destruction rate. See paper [5] for more details about the derivation of Eqs. (2)–(4).

The initial conditions for the system (2)–(4) are given by

$$f_1(0,u) = f_1^0(u), \quad f_2(0,u) = f_2^0(u), \quad f_3(0) = f_3^0. \tag{5}$$

The kinetic system (2)–(4) describes the microscopic dynamics at the cellular level starting from the initial data (5). The system reflects how the cellular interactions affect the activity of the various populations and how they contribute

to the evolution of the distribution functions f_i, $i = 1, 2, 3$. This system is used in the numerical simulations presented in Sect. 5.

4 The Mathematical Analysis of the Model

The mathematical analysis of the kinetic system (2)–(4) is in general a complex problem. Conversely, the mathematical analysis of the macroscopic system derived from kinetic equations is obviously an easier task, with the particularity that, under certain assumptions, relevant information on the solution to the kinetic system can be extracted from the mathematical analysis of the macroscopic equations. This is the case of our model. These observations motivate the content of the present section.

4.1 On the Initial Value Problem for the Kinetic System

The existence of a unique local solution to the initial value problem (2)–(4) and (5) can be stated, as follows.

Theorem 1 (Local Existence) *Assume initial data* $f_i^0(u)$ *in* $L^1[0, 1]$. *Then, there exists* $T_0 > 0$ *such that a unique positive solution to the Cauchy problem* (2)–(4) *and* (5) *exists in* $L^1[0, 1]$, *for* $t \in [0, T_0]$.

A general local result has been proven in paper [12] for a rather vast class of kinetic systems with conservative, proliferative and destructive interactions. The solution does not exist globally in time, since a blow-up can occur due to the proliferative interactions. However, a local result is enough when the system is solved numerically and an approximate solution is obtained in the considered biological context.

As it will become clear in the following, Theorem 1, together with the assumption of constant proliferation and destruction rates, assure that the basic information on the kinetic model is contained in the corresponding macroscopic system. Therefore, we introduce now the macroscopic model and present the main results concerning its qualitative analysis.

4.2 The Macroscopic Equations

From the kinetic equations (2)–(4), we formally derive the corresponding macroscopic balance equations describing the time evolution of the number of cells of each population, namely $n_i(t)$, $i = 1, 2, 3$, defined as in (1). These balance equations are obtained by integration of the kinetic equations (2)–(4) over the biological

activity variable $u \in [0, 1]$. As expected, conservative interactions do not give any contribution to the equations for $n_i(t)$, since they do not modify the number of cells of each population and are lost through the integration process. Therefore, the system of ordinary differential equations (ODEs) obtained in this way is

$$\frac{dn_1}{dt}(t) = p_{12}n_1(t)n_2(t) - d_{13}n_1(t)n_3(t), \tag{6}$$

$$\frac{dn_2}{dt}(t) = p_{21}n_2(t)n_1(t) - d_{23}n_2(t)n_3(t), \tag{7}$$

$$\frac{dn_3}{dt}(t) = p_{31}n_3(t)n_1(t). \tag{8}$$

For this system, we consider the following initial data

$$n_1(0) = n_1^0, \quad n_2(0) = n_2^0, \quad n_3(0) = n_3^0, \quad \text{with} \quad n_i^0 > 0 \quad \text{for} \quad i = 1, 2, 3. \tag{9}$$

The description obtained with the balance equations (6)–(8) gives information at a macroscopic scale and only reflects information concerning the changes on the number of cells of each population. All aspects related to the cellular activity are embedded in the macroscopic dynamics but are not directly recognizable in the balance equations.

4.3 The Qualitative Analysis of the Macroscopic Model Equations

The starting point of this analysis is the local existence result stated in Theorem 1. In fact, Theorem 1, together with the assumption of constant proliferation and destruction rates, assure that the boundedness of the solution to the macroscopic system (6)–(8) implies the boundedness of the L^1-norm $||f_i(t, \cdot)||_1$. See also paper [13]. This is an immediate consequence of the positivity of the local L^1-solution stated in Theorem 1. The estimates on the solution to the macroscopic system (6)–(8) provide a priori estimates on the solution to the kinetic system (2)–(4), due to the relationship kinetic-macro given by Eq. (1) of the population densities $n_i(t)$ in terms of the distribution functions $f_i(t, u)$.

Starting from Theorem 1, we prove in paper [5] the following results on the existence of a global, positive solution of the Cauchy problem for the macroscopic system (6)–(8) and (9).

Theorem 2 (Positivity) *Let $\underline{n}(t) = (n_1(t), n_2(t), n_3(t))$ be a solution of the Cauchy problem (6)–(8) and (9) defined on $[0, T]$, $0 < T < +\infty$. Then $n_1(t) > 0$, $n_2(t) > 0$, $n_3(t) > 0$, for $t \in [0, T]$.*

Theorem 3 (Global Solution and Asymptotic Behaviour) *Assume that* $p_{21} < p_{31}$. *Then the Cauchy problem* (6)–(8) *and* (9) *has a unique solution* $\underline{n}(t) = \big(n_1(t), n_2(t), n_3(t)\big)$ *defined on* $\mathbb{R}_+$, *satisfying the conditions*

$$\lim_{t\to+\infty} n_1(t) = 0, \quad \lim_{t\to+\infty} n_2(t) = 0, \quad \lim_{t\to+\infty} n_3(t) = \sigma < +\infty,$$

whatever are the corresponding initial data.

From the biological point of view, condition $p_{21} < p_{31}$, considered in Theorem 3, corresponds to assume that the proliferation of SRTCs resulting from the encounters with SAPCs is dominated by the proliferation of ISCs resulting from the encounters with SAPCs. In this case, the solution of the system does not possess blowups.

Theorems 2 and 3 are crucial to assure the consistency of the model and therefore to validate the numerical simulations to be performed with the kinetic system (2)–(4). These properties are important, not only from the mathematical point of view, but also from the biological point of view, to obtain solutions that are biologically significant. In particular, the positivity and the boundedness of the solution are essential features in the present context.

5 Numerical Simulations for the Biological System

In this section, we perform some numerical simulations with the kinetic system (2)–(4) in order to investigate the sensitivity of the solution to certain parameters of the model. Different scenarios are considered with the aim of analyzing if the solution is capable of describing the behavior of autoimmune diseases. The simulations show the evolution of the number density of the SRTCs, this being biologically the main indicator of an autoimmune reaction.

5.1 The Numerical Scheme

System (2)–(4) is solved numerically by discretizing the integro-differential equations in the activation variable u and using a trapezoidal quadrature rule to perform the numerical integration of the interaction terms.

More specifically, we choose a uniform discrete grid for the activation state variable $u \in [0, 1]$ and introduce the set U of $m + 1$ ($m \in \mathbb{N}$) equidistant grid points $u_k \in [0, 1]$, $k = 0, \ldots, m$, defined by

$$u_k = k\Delta u,$$

where $\Delta u = 1/m$ is the step size. We assume that parameter $w^\star$, describing the tolerance of SRTCs towards self-antigens and appearing in Eq. (3), coincides with the grid-point on the ℓ-position in U, that is $w^\star = u_\ell$.

Grid points u_k are used to approximate both the distribution function $f_i(t,u)$ and the integral collision terms in Eqs. (2)–(4). Therefore, we introduce the notation

$$f_i^k(t) = f_i(t,u_k), \tag{10}$$

where i stands for the population p_i and k indicates the localization of the activation state variable $u \in [0,1]$, with $i = 1,2$ and $k = 0,1,\ldots,m$. Moreover, we consider the integral approximations

$$\int_{u_\alpha}^{u_\beta} g(t,v)dv \approx \mathscr{Q}_\alpha^\beta[g(t,v)], \qquad 0 \le \alpha < \beta \le m, \tag{11}$$

with

$$\mathscr{Q}_\alpha^\beta[g(t,v)] = \frac{g(t,v_\alpha)+g(t,v_\beta)}{2}\Delta v + \sum_{s=\alpha+1}^{\beta-1} g(t,v_s)\Delta v, \qquad 0 \le \alpha < \beta \le m, \tag{12}$$

to obtain the quadrature approximations

$$\begin{aligned}
&\int_0^1 f_j(t,v)dv \approx \mathscr{Q}_0^m[f_j(t,v)], \quad \int_0^1 v f_j(t,v)dv \approx \mathscr{Q}_0^m[v f_j(t,v)], \quad j=1,2,\\
&\int_{u_k}^1 f_j(t,v)dv \approx \mathscr{Q}_k^m[f_j(t,v)], \quad \int_{u_k}^1 v f_j(t,v)dv \approx \mathscr{Q}_k^m[v f_j(t,v)], \quad j=1,2,\\
&\int_0^{u_k} f_j(t,v)dv \approx \mathscr{Q}_0^k[f_j(t,v)], \quad \int_0^{u_k} v f_j(t,v)dv \approx \mathscr{Q}_0^k[v f_j(t,v)], \quad j=1,2,\\
&\int_{w*}^1 f_1(t,v)dv \approx \mathscr{Q}_\ell^m[f_1(t,v)].
\end{aligned} \tag{13}$$

Proceeding in this way, we obtain the following system of $2(m+1)+1$ ODEs,

$$\begin{aligned}
\frac{df_1^k}{dt}(t) = {} & 2c_{13}f_3(t)\Big(\mathscr{Q}_k^m[v f_1(t,v)] - u_k\mathscr{Q}_k^m[f_1(t,v)]\Big) - c_{13}u_k^2 f_1^k(t) f_3(t) \\
& + c_{12}\Big[2\Big(u_k\mathscr{Q}_0^k[f_1(t,v)] - \mathscr{Q}_0^k[v f_1(t,v)]\Big) - (u_k-1)^2 f_1^k(t)\Big]\mathscr{Q}_0^m[f_2(t,v)] \\
& + p_{12}f_1^k(t)\mathscr{Q}_0^m[f_2(t,v)] - d_{13}f_1^k(t)f_3(t), \qquad k = 0,\ldots,m,
\end{aligned} \tag{14}$$

$$\frac{df_2^k}{dt}(t) = 2c_{23} f_3(t)\Big(\mathscr{Q}_k^m[vf_2(t,v)] - u_k \mathscr{Q}_k^m[f_2(t,v)]\Big) - c_{23} u_k^2 f_2^k(t) f_3(t) \tag{15}$$

$$+c_{21}\Big[2\Big(u_k \mathscr{Q}_0^k[f_2(t,v)] - \mathscr{Q}_0^k[vf_2(t,v)]\Big)\mathscr{Q}_\ell^m[f_1(t,v)]$$

$$-(u_k-1)^2 f_2^k(t)\mathscr{Q}_\ell^m[f_1(t,v)]\Big]$$

$$+p_{21} f_2^k(t)\mathscr{Q}_0^m[f_1(t,v)] - d_{23} f_2^k(t) f_3(t), \qquad k = 0, \ldots, m,$$

$$\frac{df_3}{dt}(t) = p_{31} f_3(t)\mathscr{Q}_0^m[f_1(t,v)]. \tag{16}$$

The ODE system (14)–(16) constitutes the numerical scheme to approximate the solution to the full kinetic system (2)–(4).

5.2 The Numerical Solution

We solve system (14)–(16) using the standard Maple dsolve command with the numeric option. A considerable number of simulations have been performed and we have selected a representative sample of figures to show the common features of the evolution of autoimmune diseases. These figures show the evolution of the number density of the SRTCs when different scenarios are considered.

In all simulations, the initial data are taken to be

$$f_i^0 = 10^{-2}, \quad \text{for} \quad i = 1, 2, 3. \tag{17}$$

The parameters that are not investigated in the present simulations are fixed as

$$c_{12} = 2 \quad \text{and} \quad c_{13} = 0.01. \tag{18}$$

They are associated to the SAPCs conservative interactions with SRTCs (c_{12}) and with ISCs (c_{13}).

All other parameters are varied in order to appreciate their influence on the solution to the kinetic system. In particular, parameters

$$w^\star, \quad p_{21}, \quad d_{23}, \quad c_{21} \quad \text{and} \quad c_{23} \tag{19}$$

have a direct influence on the number density of SRTCs, since they represent the tolerance parameter of the SRTCs with respect to SAPCs or, equivalently, the capacity of SAPCs to activate SRTCs ($w^\star$), the proliferative rate of SRTCs after interaction with SAPCs (p_{21}), the destructive rate of SRTCs after interaction with ISCs (d_{23}), the conservative rate of SRTCs after interaction with SAPCs (c_{21}) and the conservative rate of SRTCs after interaction with ISCs (c_{23}). On the other hand,

parameters

$$p_{12}, \quad p_{31} \quad \text{and} \quad d_{13} \tag{20}$$

have an indirect influence on the number density of SRTCs, because they represent the proliferative rate of SAPCs after interaction with SRTCs (p_{12}), the proliferative rate of ISCs after interaction with SAPCs (p_{31}) and the destructive rate of SAPCs after interaction with ISCs (d_{13}).

We underline that the influence of the conservation rates c_{21} and c_{23} on the number density of the SRTCs is quite recognizable, because we are dealing with a kinetic system which retains the conservative cellular interactions in the dynamics. On the other hand, the simulations show that the effect of the other conservation parameters, c_{12} and c_{13}, is not as recognizable in the evolution of the number density of the SRTCs because the related conservative interactions have an indirect impact on the evolution of SRTCs.

We consider different scenarios in view of illustrating the sensitivity of the solution when varying the parameters (19) and (20) that have biological significance in the present modelling of autoimmunity. More specifically, we have a first scenario describing the trend to illness and three other scenarios in which the autoimmune reaction is controlled to a certain extent.

(A) The scenario where there is *development of an autoimmune disease* corresponds to the situation in which the ISCs are unable to regulate the autoimmune reaction, resulting in a full autoimmune cascade and trending to illness. In this scenario, we consider

$$\begin{aligned} &w^\star = 1/30, \quad p_{21} = 19, \quad d_{23} = 0.025, \quad c_{21} = 10, \quad c_{23} = 0.01, \\ &p_{12} = 1, \quad p_{31} = 20, \quad d_{13} = 0.35, \end{aligned} \tag{21}$$

and the corresponding solution is depicted in Fig. 2. We can observe a considerable mass proliferation of very active SRTCs, of the order 10^4 of the

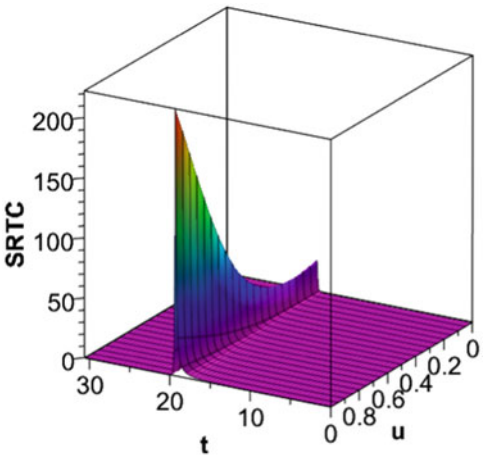

Fig. 2 *Scenario (A)—trend to illness.* The evolution of SRTCs is determined by the approximating solution to the kinetic system (2)–(4), when the parameters are given by (21). The figure shows a considerable mass proliferation of very active SRTCs

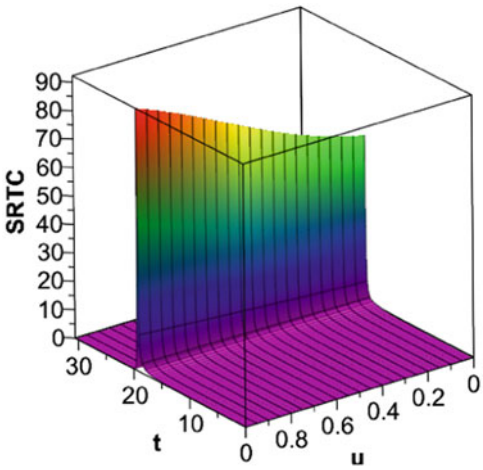

Fig. 3 *Scenario (B)—Immunotolerance.* Effect of increasing the parameter $w^\star$, as described in. The evolution of SRTCs is determined by the approximating solution to the kinetic system (2)–(4), when the parameters are given by (22), and in particular $w^\star = 29/30$

initial data, due to insufficient regulation by ISCs and low tolerance of SRTCs to SAPCs.

(B) The scenario where SRTCs *become more tolerant to* SAPCs corresponds to the situation in which SAPCs are less efficient in increasing the activity of SRTCs. In this scenario, we consider

$$w^\star = 29/30, \quad p_{21} = 19, \quad d_{23} = 0.025, \quad c_{21} = 10, \quad c_{23} = 0.01,$$
$$p_{12} = 1, \quad p_{31} = 20, \quad d_{13} = 0.35, \tag{22}$$

and the corresponding solution is illustrated in Fig. 3. We can observe that, in comparison with Fig. 2, a moderate decrease in the mass proliferation of very active SRTCs is observed, whereas a slight decrease in the mass proliferation of low active SRTCs is recognizable.

(C) The scenario where there is *immunosuppression of the autoimmune reaction* corresponds to the situation in which the biological system is able to abort the autoimmune reaction in an efficient manner, by controlling different proliferative or destructive rates.

In this scenario, we maintain all parameters of scenario (A) with exception of one that is varying once per time. In particular, we consider a lower value of p_{21} or p_{12}, or a greater value of p_{31}, d_{13} or d_{23}. The corresponding solutions are shown in diagrams (a)–(e) of Fig. 4. From the qualitative point of view, the behaviour represented in these diagrams is the same and all pictures exhibit a very low proliferation of active SRTCs.

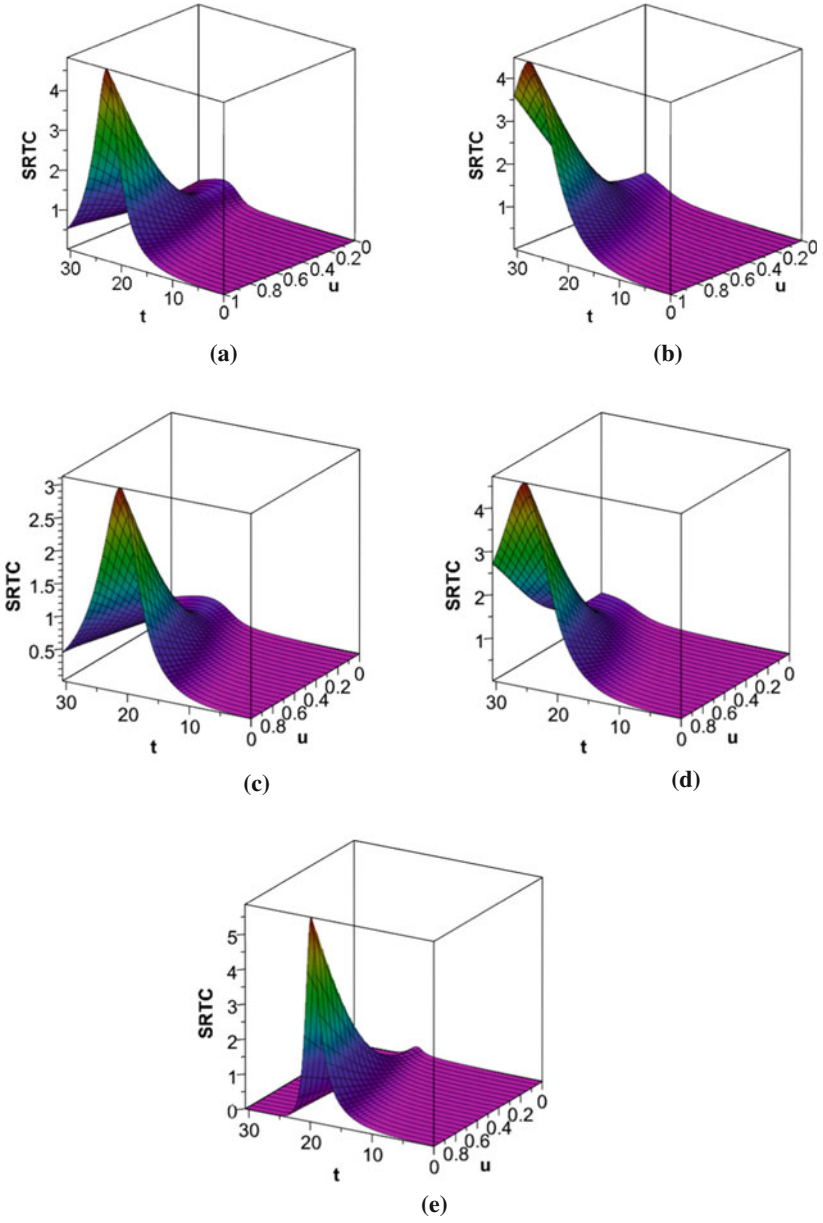

Fig. 4 *Scenario (C)—Immunosupression.* The evolution of SRTCs is determined by the approximating solution to the kinetic system (2)–(4), when the parameters are given by (21), with exception of one parameter. (**a**) Decreasing the proliferative rate p_{21} to $p_{21} = 17$. (**b**) Decreasing the proliferative rate p_{12} to $p_{12} = 0.5$. (**c**) Increasing the proliferative rate p_{31} to $p_{31} = 23$. (**d**) Increasing the destructive rate d_{13} to $d_{13} = 0.7$. (**e**) Increasing the destructive rate d_{23} to $d_{23} = 0.1$. Each diagram shows that, by varying one parameter with respect to the value considered in (21), the biological system is able to reduce considerably the mass proliferation of the SRTCs and therefore to abort the autoimmune reaction in an efficient manner

- In Fig. 4a, the reduction of SRTCs proliferative encounters with SAPCs (lower value of p_{21}) obviously implies a significant impact on the mass production of SRTCs capable of avoiding the trend to illness. The figure shows the effect of p_{21} on the suppression of the autoimmune reaction.
- In Fig. 4b, the reduction of SAPCs proliferative encounters with SRTCs (lower value of p_{12}) has an indirect impact on the mass production of SRTCs since the concentration of SAPCs decreases and the activation of SRTCs by SAPCs is weakened, so that the trend to illness is avoided. The figure illustrates the effect of p_{12} on the suppression of the autoimmune reaction.
- In Fig. 4c, the number of ISCs produced by the biological system is increased by proliferative interactions with SAPCs (greater value of p_{31}), the result being that the trend to illness is avoided in an efficient manner. The figure shows the effect of p_{31} on the suppression of the autoimmune reaction.
- In Fig. 4d, the results show that for the number of SAPCs destroyed as a consequence of their interaction with ISCs (greater value of d_{13}) will ultimately control the proliferation of SRTCs and therefore avoid illness. The figure shows the consequences of d_{13} on the suppression of the autoimmune reaction.
- In Fig. 4e, the results show that the number of SRTCs destroyed as a consequence of their interaction with ISCs (greater value of d_{23}) can definitively avoid a full blown autoimmune reaction. The figure shows the impact of d_{23} on the suppression of the autoimmune reaction.

(D) The scenario where there is *control of the disease* also corresponds to the situation in which the biological system is able to abort the autoimmune reaction in an efficient manner, due to a reduction of the activity of the SRTCs after conservative interactions with SAPCs or ISCs.

In this scenario, we maintain all parameters of scenario (A) with exception of one that is varying once per time. In particular, we consider lower values of c_{21} or greater values of c_{23}. The corresponding solutions are shown in diagrams (**a**)–(**d**) of Fig. 5.

The comparison between this scenario and scenario (A) shows that the total number of SRTCs for $u \in [0, 1]$ is exactly the same, because we only modify the rates of certain conservative encounters. As a consequence, the mass proliferation of SRTCs shows a moderate reduction and the aggressive nature of the autoimmnune reaction is only slightly weakened.

- Diagrams (**a**) and (**b**) of Fig. 5 show that the mass proliferation of very active SRTCs is slightly reduced when the conservative rate c_{21} is decreased. This is a consequence of a lower production of cytokines by SRTC since these encounters reduce the activity of SRTCs and, therefore, control the triggering of an inflammatory process and the development of an autoimmune disease to a certain extent.

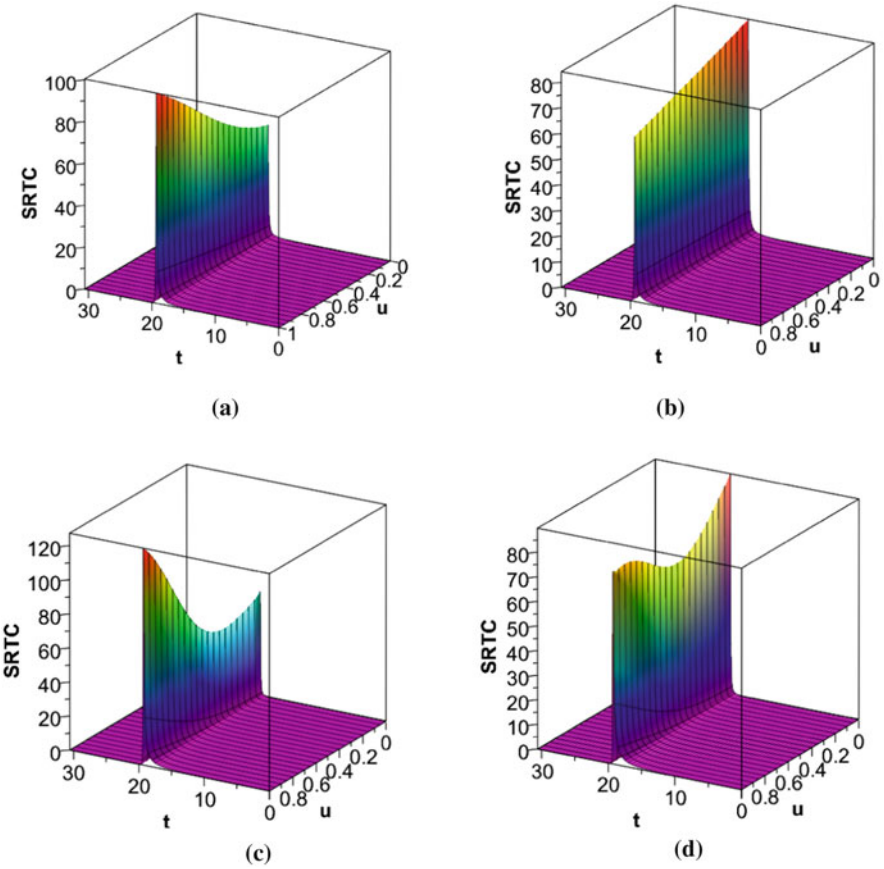

Fig. 5 *Immunosupression. The disease is controlled by decreasing the conservative rate* c_{21}*, diagrams* (***a***) *and* (***b***)*, or by increasing the conservative rate* c_{23}*, diagrams* (***c***) *and* (***d***)*, as described in scenario (D).* (**a**) $c_{21} = 2$. (**b**) $c_{21} = 0.5$. (**c**) $c_{23} = 0.03$. (**d**) $c_{23} = 0.05$. The evolution of SRTCs is determined by the approximating solution to the kinetic system (2)–(4), when the parameters are given by (21) with exception of c_{21} and c_{23}

- Diagrams (**c**) and (**d**) of Fig. 5 also show that the mass proliferation of very active SRTCs is slightly reduced when the conservative rate c_{23} is increased. This is a consequence of a lower production of cytokines by SRTC due to a greater inhibiting effect of ISCs on the SRTC function and, therefore moderating the autoimmune disease.

6 Conclusion and Perspectives

The mathematical model that has been proposed in [5], based on a kinetic theory approach, is here revisited. The mathematical analysis of the model, showing existence, uniqueness, positivity and boundedness of the solution, is also reviewed here.

Starting from the model proposed in [5], we develop here some numerical simulations in order to investigate the sensitivity of the model to certain parameters that are involved in the biological description. We consider different scenarios with the aim of describing different behaviors occurring in autoimmunity. In particular, we study the influence of certain parameters related to immunotolerance and immunosupression on the evolution of the variables characterizing this model for autoimmunity. The conclusion of this study is that increasing the parameters related to immunotolerance and immnunosupression is effective in reducing the production of highly active SRTCs and thefore controlling the progression of an autoimmune episode.

Therefore the numerical simulations developed here and the corresponding biological interpretation of the results constitute a valuable complement of the mathematical model proposed in [5].

Other extensions of the model proposed in [5] have been already considered and others are still open to further developments. We have extended our research work in view of introducing drug therapies on the dynamics and investigating optimal treatment strategies. The results have been submitted for publication, see [6, 7].

Another extension has been considered in order to introduce recurrence in the macroscopic model presented in [5] by considering a constant input by the host environment of self-antigen presenting cells (SAPCs) and the natural death of all cell populations involved. Such a model is able to study the chronic character of the autoimmune diseases. The results are presented in [16].

Other interesting problems that we plan to study is the introduction of delay terms in the equations in order to describe the delay in the reaction to cellular impulses. Memory terms may also be introduced with the aim of describing the ability of cells to retain information related to past experienced cell interactions.

Acknowledgments This work is partially supported by the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM.

References

1. Shi, F., Kaer, L.V.: Reciprocal regulation between natural killer cells and autoreactive T cells. Nat. Rev. Immunol. **6**, 751–760 (2006)
2. Tian, Z., Gershwin, M.E., Zhang, C.: Regulatory NK cells in autoimmune disease. J. Autoimmun **39**, 206–215 (2012)
3. Poggi, A., Zocchi, M.R.: NK cell autoreactivity and autoimmune diseases. Front. Immunol. **5**, 1–15 (2014)

4. Sharabi, A., Tsokos, M.G., Ding, Y., Malek, T.R., Klatzmann, D., Tsokos, G.C.: Regulatory T cells in the treatment of disease. Nat. Rev. Drug Discov. **17**, 823–844 (2018)
5. Ramos, M.P., Ribeiro, C., Soares, A.J.: A kinetic model of T cell autoreactivity in autoimmune diseases. J. Math. Biol. **79**, 2005–2031 (2019)
6. Costa, M.F., Ramos, M.P., Ribeiro, C., Soares, A.J.: Mathematical modeling and optimal control of immunotherapy for autoimmune disease. Math. Meth. Appl. Sci. **44**, 8883–8902 (2021). https://doi.org/10.1002/mma.7318
7. Costa, M.F., Ramos, M.P., Ribeiro, C., Soares, A.J.: Recent developments on the modelling of cell interactions in autoimmune diseases. In: Proceedings of the Conference PSPDE 2019. Springer series in Mathematics and Statistics (2020)
8. Kolev, M., Nikolova, I.: Dynamical properties of autoimmune disease models: Tolerance, flare-up, dormancy. J. Theor. Biol. **246**, 646–659 (2007)
9. Delitala, M., Dianzani, U., Lorenzi, T., Melensi, M.: A mathematical model for immune and autoimmune response mediated by T-cells. Comput. Math. Appl. **66**, 1010–1023 (2013)
10. Kolev, M., Nikolova, I.: A mathematical model of some viral-induced autoimmune diseases Math. Applic. **46**, 97–108 (2018)
11. Bellomo, N., Forni, G.: Dynamics of tumor interaction with the host immune system Mathl. Comput. Modelling **20**, 107–122 (1994)
12. Arlotti, L., Bellomo, N., Latrach, K.: From the Jager and Segel model to kinetic population dynamics nonlinear evolution problems and applications. Mathl. Comput. Modelling **30**, 15–40 (1999)
13. Arlotti, L., Lachowicz, M.: Qualitative analysis of a nonlinear integrodifferential equation modeling tumor-host dynamics. Mathl. Comput. Modelling **23**, 11–29 (1996)
14. Bellouquid, A., De Angelis, E.: From kinetic models of multicellular growing systems to macroscopic biological tissue models. Nonlin. An: Real World Applics. **12**, 1111–1122 (2011)
15. Eftimie, R., Gibelli, L.: A kinetic theory approach for modelling tumour and macrophages heterogeneity and plasticity during cancer progression Math. Mod. Meth. App. Sci. **30**, 659–683 (2020)
16. Della Marca, R., Ramos, M.P., Ribeiro, C., Soares, A.J.: Mathematical modelling of oscillating patterns for chronic autoimmune diseases (submitted in 2021)

A Generalized Slip-Flow Theory for a Slightly Rarefied Gas Flow Induced by Discontinuous Wall Temperature

Satoshi Taguchi and Tetsuro Tsuji

Abstract A system of fluid-dynamic-type equations and their boundary conditions derived from a system of the Boltzmann equation is of great importance in kinetic theory when we are concerned with the motion of a slightly rarefied gas. It offers an efficient alternative to solving the Boltzmann equation directly and, more importantly, provides a clear picture of the flow structure in the near-continuum regime. However, the applicability of the existing slip-flow theory is limited to the case where both the boundary shape and the kinetic boundary condition are smooth functions of the boundary coordinates, which precludes, for example, the case where the kinetic boundary condition has a jump discontinuity. In this paper, we discuss the motion of a slightly rarefied gas caused by a discontinuous wall temperature in a simple two-surface problem and illustrate how the existing theory can be extended. The discussion is based on our recent paper [Taguchi and Tsuji, J. Fluid Mech. 897, A16 (2020)] supported by some preliminary numerical results for the newly introduced kinetic boundary layer (the Knudsen zone), from which a source-sink condition for the flow velocity is derived.

1 Introduction

Let us consider a rarefied gas in contact with a smooth boundary (or boundaries). We are concerned with the steady behavior of the gas. Suppose that the molecular mean free path is small compared with the characteristic system size (the Knudsen number is small). Then, it is often advantageous to solve the fluid-dynamic system derived from the Boltzmann system. This approach is known as the generalized slip-flow theory and was developed notably by Sone and his coworkers [13–16].

The generalized slip-flow theory is based on the asymptotic analysis of the Boltzmann system for small Knudsen numbers. Both the boundary shape and the

S. Taguchi (✉) · T. Tsuji
Department of Advanced Mathematical Sciences, Graduate School of Informatics, Kyoto University, Kyoto, Japan
e-mail: taguchi.satoshi.5a@kyoto-u.ac.jp; tsuji.tetsuro.7x@kyoto-u.ac.jp

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*, Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_14

boundary condition need to be smooth. This smoothness condition is required for the Knudsen-layer problem to be reduced to a half-space problem of a kinetic equation in space one dimension, from which the slip/jump boundary conditions are obtained.

The smoothness condition can be, however, restrictive in some situations. For example, S.T. considered in [17] a rarefied gas flow around a sharp edge with different surface temperatures on each side. But due to the limitation, only a qualitative argument was possible for the flow structure around the edge. Motivated by this, in this article, we discuss the possibility to extend the generalized slip-flow theory to the case where the boundary condition has a jump discontinuity in a simple two-surface problem. That is, we consider a steady rarefied gas flow between two parallel plates with a discontinuous wall temperature in the framework of the generalized slip-flow theory. The discussion is based on our recent paper [18] with some new numerical result, which supports the present theory.

Finally, we remark the following. In our problem (to be stated next), the boundary condition has a jump discontinuity (through the plate's temperature distribution). This induces discontinuities of the velocity distribution function on the boundary, and they propagate into the gas region. This feature is important in a numerical analysis and was taken into account in [2], where a similar temperature-driven flow has been considered (see also [18]). It is also considered in our numerical results shown in Sect. 5, although the numerical approach is different. The propagation of boundary-induced discontinuity in kinetic equations is also a mathematical concern and has been investigated in, e.g., [1, 6–9].

2 Problem and Formulation

2.1 Problem

Let L be the reference length and let ρ_0, T_0, and p_0 be the reference density, temperature, and pressure of the gas, respectively. We consider a monatomic rarefied gas occupying the space between two parallel plates located at $x_1 = -\frac{\pi}{2}$ and $x_1 = \frac{\pi}{2}$, where (Lx_1, Lx_2, Lx_3) is the Cartesian coordinate system, as shown in Fig. 1. The upper halves of the plates ($x_2 > 0$) are kept at temperature $T_0(1 + \tau_w)$, while the lower halves ($x_2 < 0$) at temperature $T_0(1 - \tau_w)$, where τ_w is a constant. Henceforth, we assume $\tau_w > 0$. Therefore, the surfaces' temperature has a step-like distribution, which is discontinuous at $x_2 = 0$ with the jump $2T_0\tau_w$. We also assume that the gas is subject to no pressure gradient nor external force. We investigate the steady behavior of the gas under the following assumptions: (i) the behavior of the gas is described by the Boltzmann equation; (ii) the gas molecules make diffuse reflection on the plates; (iii) τ_w is so small that the equation and boundary conditions can be linearized around the reference equilibrium state at rest with density ρ_0 and temperature T_0; (iv) the Knudsen number defined by the molecular mean free path at the reference state divided by L is small.

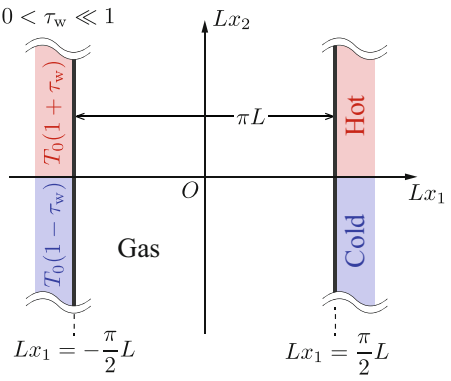

Fig. 1 Schematic of the problem. A rarefied gas between two parallel plates located at $x_1 = \pm\pi/2$ with a step-like temperature distribution is considered. The temperature of the plates is discontinuous at $x_2 = 0$

2.2 Formulation

Let us denote by $(2RT_0)^{1/2}(\zeta_1, \zeta_2, \zeta_3)$ the molecular velocity (R is the specific gas constant) and by $\rho_0(2RT_0)^{-3/2}(1 + \phi(\boldsymbol{x}, \boldsymbol{\zeta}))E$ the velocity distribution function, where $E = \pi^{-3/2}\exp(-|\boldsymbol{\zeta}|^2)$. The time-independent Boltzmann equation reads

$$\zeta_i \partial_i \phi = \frac{1}{\varepsilon}\mathscr{L}(\phi), \tag{1}$$

where $\partial_i = \partial/\partial x_i$, $\mathscr{L}$ is the linearized collision operator [16], and ε is a parameter defined by

$$\varepsilon = \frac{\sqrt{\pi}}{2}\mathrm{Kn} = \frac{\sqrt{\pi}}{2}\frac{\ell_0}{L} \qquad \text{(Kn: Knudsen number)}.$$

Here, ℓ_0 is the mean free path of the gas molecules in the equilibrium state at rest with temperature T_0 and density ρ_0. Note that ε is the Knudsen number multiplied by $\sqrt{\pi}/2$. The operator $\mathscr{L}$ is given by

$$\mathscr{L}(F) = \int_{(\boldsymbol{\zeta}_*, \boldsymbol{e}) \in \mathbf{R}^3 \times \mathbf{S}^2} E_*(F'_* + F' - F_* - F)\, B\, \mathrm{d}\Omega(\boldsymbol{e})\mathrm{d}\boldsymbol{\zeta}_*, \tag{2a}$$

$$F = F(\boldsymbol{\zeta}), \quad F_* = F(\boldsymbol{\zeta}_*), \quad F' = F(\boldsymbol{\zeta}'), \quad F'_* = F(\boldsymbol{\zeta}'_*), \tag{2b}$$

$$\boldsymbol{\zeta}' = \boldsymbol{\zeta} + [(\boldsymbol{\zeta}_* - \boldsymbol{\zeta}) \cdot \boldsymbol{e}]\boldsymbol{e}, \quad \boldsymbol{\zeta}'_* = \boldsymbol{\zeta}_* - [(\boldsymbol{\zeta}_* - \boldsymbol{\zeta}) \cdot \boldsymbol{e}]\boldsymbol{e}, \tag{2c}$$

$$B = B\left(\frac{|\boldsymbol{e} \cdot (\boldsymbol{\zeta}_* - \boldsymbol{\zeta})|}{|\boldsymbol{\zeta}_* - \boldsymbol{\zeta}|}, |\boldsymbol{\zeta}_* - \boldsymbol{\zeta}|\right), \quad E_* = \frac{1}{\pi^{3/2}}e^{-|\boldsymbol{\zeta}_*|^2}, \tag{2d}$$

where $\mathrm{d}\Omega(\boldsymbol{e})$ is the solid angle element in the direction of $\boldsymbol{e}$, B is a non-negative function whose functional form is determined by the designated intermolecular force. For example, $B = \frac{1}{4\sqrt{2\pi}}|\boldsymbol{e} \cdot (\boldsymbol{\zeta}_* - \boldsymbol{\zeta})|$ for a hard-sphere gas. The diffuse

reflection boundary conditions on the plates are summarized as

$$\phi = 2\sqrt{\pi}\int_{\zeta_1<0} |\zeta_1|\phi E \mathrm{d}\boldsymbol{\zeta} \pm (|\boldsymbol{\zeta}|^2 - 2)\tau_{\mathrm{w}}, \quad \zeta_1 > 0 \qquad \left(x_1 = -\frac{\pi}{2},\ x_2 \gtrless 0\right), \tag{3a}$$

$$\phi = 2\sqrt{\pi}\int_{\zeta_1>0} |\zeta_1|\phi E \mathrm{d}\boldsymbol{\zeta} \pm (|\boldsymbol{\zeta}|^2 - 2)\tau_{\mathrm{w}}, \quad \zeta_1 < 0 \qquad \left(x_1 = \frac{\pi}{2},\ x_2 \gtrless 0\right), \tag{3b}$$

where $\mathrm{d}\boldsymbol{\zeta} = \mathrm{d}\zeta_1\mathrm{d}\zeta_2\mathrm{d}\zeta_3$.

The macroscopic quantities of interest, namely, the density, the flow velocity, the temperature, and the pressure of the gas denoted by $\rho_0(1+\omega)$, $(2RT_0)^{1/2}u_i$, $T_0(1+\tau)$, and $p_0(1+P)$, respectively, are defined in terms of ϕ as

$$\omega = \langle\phi\rangle, \quad u_i = \langle\zeta_i\phi\rangle, \quad \tau = \frac{2}{3}\left\langle\left(|\boldsymbol{\zeta}|^2 - \frac{3}{2}\right)\phi\right\rangle, \quad P = \frac{2}{3}\langle|\boldsymbol{\zeta}|^2\phi\rangle = \omega + \tau, \tag{4}$$

where $\langle\cdot\rangle$ designates

$$\langle F\rangle = \int_{\mathbf{R}^3} F(\boldsymbol{\zeta})E\mathrm{d}\boldsymbol{\zeta}. \tag{5}$$

In the present two-dimensional problem, we may assume that ϕ is independent of x_3. Nevertheless, the x_3-dependency has not been precluded in the above formulation for later convenience.

The study on the behavior of a slightly rarefied gas (i.e., the gas with small Knudsen numbers) has a long history (see, e.g., references in [15]). In the case of a smooth boundary, Sone and his coworkers have extensively studied the question both for the steady [13–16] and unsteady [16, 19] settings. It is based on the asymptotic analysis of the Boltzmann system for small Knudsen numbers, and the theory is nowadays known as the generalized slip-flow theory. However, the approach above precludes the discontinuous boundary data. One of the paper's purposes is to show that we can extend Sone's asymptotic theory to include the latter situation.

3 Case of a Smooth Temperature Distribution

Before we discuss the discontinuous surface temperature case, it is useful to review the case of a smooth temperature distribution. Let the temperature of the two plates be given by $T_0(1+\overline{\tau}_{\mathrm{w}})$, where $\overline{\tau}_{\mathrm{w}}$ is a smooth function of (x_2, x_3). Then, assuming

the diffuse reflection condition, the boundary conditions (3a) and (3b) are replaced by

$$\phi = 2\sqrt{\pi}\int_{\zeta_1 \lessgtr 0} |\zeta_1|\phi E \mathrm{d}\boldsymbol{\zeta} + (|\boldsymbol{\zeta}|^2 - 2)\overline{\tau}_{\mathrm{w}}, \quad \zeta_1 \gtrless 0$$
$$\left(x_1 = \mp\frac{\pi}{2},\ -\infty < x_2 < \infty,\ -\infty < x_3 < \infty\right). \tag{6}$$

We consider the asymptotic behavior of the solution ϕ of the linear system (1) and (6) for small ε following Sone's method [15, 16]. It should be noted that for the linearization, $|\partial_i \overline{\tau}_{\mathrm{w}}| \ll 1$ should be assumed.

By the symmetry of the problem, one can assume that the solution is even with respect to $x_1 = 0$. Therefore, in the sequel, we consider the problem only in the left-half domain $D^- = \{(x_1, x_2, x_3) \mid -\frac{\pi}{2} < x_1 < 0, -\infty < x_2 < \infty, -\infty < x_3 < \infty\}$. The solution in the right-half domain is obtained from that of D^- by $\phi(x_1, x_2, x_3, \zeta_1, \zeta_2, \zeta_3) = \phi(-x_1, x_2, x_3, -\zeta_1, \zeta_2, \zeta_3)$.

According to [15], the solution is expressed in the form

$$\phi = \phi_{\mathrm{H}} + \phi_{\mathrm{K}}, \tag{7}$$

where ϕ_{H} is called the Hilbert solution and describes the overall behavior of the gas, while ϕ_{K} is a correction to ϕ_{H} required in the vicinity of the boundary (the Knudsen-layer correction). More precisely, ϕ_{H} is a solution to Eq. (1) subject to the condition $\partial_i \phi_{\mathrm{H}} = O(\phi_{\mathrm{H}})$ (i.e., moderately varying solution). On the other hand, ϕ_{K} is appreciable only in a thin layer (the Knudsen layer) adjacent to the boundary $x_1 = -\frac{\pi}{2}$, whose thickness is of the order of ε. The Knudsen-layer correction ϕ_{K} is subject to the conditions

$$\partial_1 \phi_{\mathrm{K}} = O(\phi_{\mathrm{K}}/\varepsilon), \quad (\delta_{ij} - n_i n_j)\partial_j \phi_{\mathrm{K}} = O(\phi_{\mathrm{K}}), \tag{8}$$

where δ_{ij} is Kronecker's delta and $\boldsymbol{n} = (1, 0, 0)$. The ϕ_{H} and ϕ_{K} are expanded in ε as

$$\phi_{\mathrm{H}} = \phi_{\mathrm{H0}} + \varepsilon\phi_{\mathrm{H1}} + \varepsilon^2\phi_{\mathrm{H2}} + \cdots, \tag{9a}$$
$$\phi_{\mathrm{K}} = \varepsilon\phi_{\mathrm{K1}} + \varepsilon^2\phi_{\mathrm{K2}} + \cdots. \tag{9b}$$

Accordingly, the macroscopic quantities h ($h = \omega,\ u_i,\ \tau,\ P$) are also expressed as

$$h = h_{\mathrm{H}} + h_{\mathrm{K}}, \tag{10a}$$
$$h_{\mathrm{H}} = h_{\mathrm{H0}} + \varepsilon h_{\mathrm{H1}} + \varepsilon^2 h_{\mathrm{H2}} + \cdots, \tag{10b}$$
$$h_{\mathrm{K}} = \varepsilon h_{\mathrm{K1}} + \varepsilon^2 h_{\mathrm{K2}} + \cdots, \tag{10c}$$

where

$$\omega_{\mathrm{H}m} = \langle \phi_{\mathrm{H}m} \rangle, \quad u_{i\mathrm{H}m} = \langle \zeta_i \phi_{\mathrm{H}m} \rangle, \quad \tau_{\mathrm{H}m} = \frac{2}{3} \left\langle \left(|\boldsymbol{\zeta}|^2 - \frac{3}{2} \right) \phi_{\mathrm{H}m} \right\rangle, \tag{11a}$$

$$P_{\mathrm{H}m} = \omega_{\mathrm{H}m} + \tau_{\mathrm{H}m}, \tag{11b}$$

($m = 0, 1, \ldots$), and

$$\omega_{\mathrm{K}m} = \langle \phi_{\mathrm{K}m} \rangle, \quad u_{i\mathrm{K}m} = \langle \zeta_i \phi_{\mathrm{K}m} \rangle, \quad \tau_{\mathrm{K}m} = \frac{2}{3} \left\langle \left(|\boldsymbol{\zeta}|^2 - \frac{3}{2} \right) \phi_{\mathrm{K}m} \right\rangle, \tag{12a}$$

$$P_{\mathrm{K}m} = \omega_{\mathrm{K}m} + \tau_{\mathrm{K}m}, \tag{12b}$$

($m = 1, 2, \cdots$).

Then, it is shown in [15] that $\phi_{\mathrm{H}0}$, $\phi_{\mathrm{H}1}$, and $\phi_{\mathrm{K}1}$ are expressed in the form

$$\phi_{\mathrm{H}0} = \phi_{\mathrm{eH}0}, \tag{13a}$$

$$\phi_{\mathrm{H}1} = \phi_{\mathrm{eH}1} - \zeta_i A(|\boldsymbol{\zeta}|) \partial_i \tau_{\mathrm{H}0} - \frac{1}{2} \zeta_i \zeta_j B(|\boldsymbol{\zeta}|) (\partial_j u_{i\mathrm{H}0} + \partial_i u_{j\mathrm{H}0}), \tag{13b}$$

$$\begin{aligned} \phi_{\mathrm{K}1} = {} & \varphi_1^{(0)}(\eta, \zeta_1, |\overline{\boldsymbol{\zeta}}|) \, (\partial_1 \tau_{\mathrm{H}0})_0 \\ & + \overline{\zeta}_i \left[\varphi_1^{(1)}(\eta, \zeta_1, |\overline{\boldsymbol{\zeta}}|) \, n_j (\partial_j u_{i\mathrm{H}0} + \partial_i u_{j\mathrm{H}0})_0 \right. \\ & \left. + \varphi_2^{(1)}(\eta, \zeta_1, |\overline{\boldsymbol{\zeta}}|) \, (\partial_i \tau_{\mathrm{H}0})_0 \right], \quad \eta = \frac{x_1 + \frac{\pi}{2}}{\varepsilon}. \end{aligned} \tag{13c}$$

Here,

1. $\phi_{\mathrm{eH}m}$ is a linear combination of $(1, \zeta_i, |\boldsymbol{\zeta}|)$ forming the (linearized) local Maxwellian

$$\phi_{\mathrm{eH}m} = P_{\mathrm{H}m} + 2\zeta_i u_{i\mathrm{H}m} + \left(|\boldsymbol{\zeta}|^2 - \frac{5}{2} \right) \tau_{\mathrm{H}m}, \quad m = 0, 1.$$

2. The functions $A(|\boldsymbol{\zeta}|)$ and $B(|\boldsymbol{\zeta}|)$ are the solutions to the integral equations

$$\mathscr{L}(\zeta_i A) = -\zeta_i \left(|\boldsymbol{\zeta}|^2 - \frac{5}{2} \right), \quad \text{with } \langle |\boldsymbol{\zeta}|^2 A \rangle = 0,$$

$$\mathscr{L}(\zeta_{ij} B) = -2\zeta_{ij},$$

   where $\zeta_{ij} = \zeta_i \zeta_j - \frac{|\boldsymbol{\zeta}|^2}{3} \delta_{ij}$.
3. η is a stretched coordinate of x_1 near the boundary $x_1 = -\frac{\pi}{2}$, adequate to describe the Knudsen-layer corrections.

4. $\overline{\boldsymbol{\zeta}}$ is a projection of $\boldsymbol{\zeta}$ onto a plane orthogonal to $\boldsymbol{n} = (1, 0, 0)$, i.e.,

$$\overline{\zeta}_i = \zeta_j(\delta_{ij} - n_i n_j).$$

5. The symbol $(\cdot)_0$ indicates the value on $x_1 = -\frac{\pi}{2}$.
6. The functions $\varphi_1^{(0)} = \varphi_1^{(0)}(\eta, \zeta_1, |\overline{\boldsymbol{\zeta}}|)$ and $\varphi_j^{(1)} = \varphi_j^{(1)}(\eta, \zeta_1, |\overline{\boldsymbol{\zeta}}|)$, $j = 1, 2$, solve the following half-space problems (Knudsen-layer problems):

$$\zeta_1 \partial_\eta \varphi_1^{(0)} = \mathscr{L}(\varphi_1^{(0)}), \tag{14a}$$

$$\varphi_1^{(0)} = -(|\boldsymbol{\zeta}|^2 - 2)c_1^{(0)} + \zeta_1 A(|\boldsymbol{\zeta}|) + 4\int_0^\infty \int_{-\infty}^0 |\zeta_1||\overline{\boldsymbol{\zeta}}|\varphi_1^{(0)} \mathrm{e}^{-|\boldsymbol{\zeta}|^2} \mathrm{d}\zeta_1 \mathrm{d}|\overline{\boldsymbol{\zeta}}|, \quad \zeta_1 > 0, \quad \eta = 0, \tag{14b}$$

$$\varphi_1^{(0)} \to 0, \quad \text{as } \eta \to \infty; \tag{14c}$$

$$\zeta_1 \partial_\eta \varphi_j^{(1)} = \mathscr{L}(\varphi_j^{(1)}), \quad j \in \{1, 2\}, \tag{15a}$$

$$\varphi_j^{(1)} = -2b_j^{(1)} + J_j, \quad \zeta_1 > 0, \quad \eta = 0, \tag{15b}$$

$$\varphi_j^{(1)} \to 0, \quad \text{as } \eta \to \infty, \tag{15c}$$

with

$$J_1 = \zeta_1 B(|\boldsymbol{\zeta}|), \quad J_2 = A(|\boldsymbol{\zeta}|), \tag{16a}$$

$$c_1^{(0)}, \quad b_j^{(1)} \quad (j = 1, 2): \quad \text{constants}. \tag{16b}$$

Note that $|\boldsymbol{\zeta}| = \sqrt{\zeta_1^2 + |\overline{\boldsymbol{\zeta}}|^2}$. It is known that there exists a solution to the problem if and only if the constant $c_1^{(0)}$ or $b_j^{(0)}$ takes a special value and that the solution is unique [3, 5, 15]. It has also been proved that the solution decays exponentially fast as $\eta \to \infty$.

Suppose that the functions A, B, $\varphi_1^{(0)}$, and $\varphi_i^{(1)}$, $i = 1, 2$, are known. Then, the functional dependency of $\phi_{\mathrm{H}m}$ and $\phi_{\mathrm{K}m}$ on the molecular velocity $\boldsymbol{\zeta}$ is prescribed through these auxiliary functions and $\phi_{\mathrm{eH}m}$. On the other hand, the spatial dependency enters through those of $u_{i\mathrm{H}m}(\boldsymbol{x})$, $\tau_{\mathrm{H}m}(\boldsymbol{x})$, and $P_{\mathrm{H}m}(\boldsymbol{x})$ (and their spatial derivatives when $m \geq 1$). The dependency of $u_{i\mathrm{H}m}$, $\tau_{\mathrm{H}m}$, and $P_{\mathrm{H}m}$, and $\omega_{\mathrm{H}m}$ on $\boldsymbol{x}$ are obtained via the fluid-dynamic-type problems stated next.

Stokes Problem The expansion coefficients of the macroscopic quantities $h_{\mathrm{H}m}$ ($h = \omega, u_i, \tau, P$) are described by the following equations and boundary conditions on $x_1 = -\frac{\pi}{2}$. The equations are

$$\partial_i P_{\mathrm{H}0} = 0, \tag{17}$$

$$\partial_i u_{i\mathrm{H}m} = 0, \qquad \text{(continuity equation)} \tag{18a}$$

$$\gamma_1 \Delta u_{i\mathrm{H}m} - \partial_i P_{\mathrm{H}m+1} = 0, \qquad \text{(equation of motion)} \tag{18b}$$

$$\Delta \tau_{\mathrm{H}m} = 0, \qquad \text{(energy equation)} \tag{18c}$$

$$\omega_{\mathrm{H}m} = P_{\mathrm{H}m} - \tau_{\mathrm{H}m}, \qquad \text{(equation of state)} \tag{18d}$$

($m = 0, 1, \ldots$). The boundary conditions on $x_1 = -\frac{\pi}{2}$ are

$$\text{Order } \varepsilon^0: \quad u_{1\mathrm{H}0} = u_{2\mathrm{H}0} = u_{3\mathrm{H}0} = 0, \quad \tau_{\mathrm{H}0} = \overline{\tau}_{\mathrm{w}}, \tag{19a}$$

$$\text{Order } \varepsilon^1: \quad u_{1\mathrm{H}1} = 0, \quad \tau_{\mathrm{H}1} = c_1^{(0)} \partial_1 \tau_{\mathrm{H}0}, \tag{19b}$$

$$u_{j\mathrm{H}1} t_j = b_1^{(1)} t_j n_k (\partial_j u_{k\mathrm{H}0} + \partial_k u_{j\mathrm{H}0}) + b_2^{(1)} t_j \partial_j \tau_{\mathrm{H}0}. \tag{19c}$$

Here, $\Delta = \partial_1^2 + \partial_2^2 + \partial_3^2$ is the Laplacian, the *viscosity* $\gamma_1 > 0$ is defined by

$$\gamma_1 = \frac{2}{15} \langle |\boldsymbol{\zeta}|^4 B \rangle, \tag{20}$$

t_i is any unit vector orthogonal to $\boldsymbol{n} = (1, 0, 0)$, and $b_i^{(1)}$ ($i = 1, 2$) and $c_1^{(0)}$, known as the slip/jump coefficients, are the same constants arising in the Knudsen-layer problem introduced above. The numerical value of γ_1 and those of the slip/jump coefficients for a hard-sphere gas are obtained as $\gamma_1 = 1.270042427$ and $(b_1^{(1)}, b_2^{(1)}, c_1^{(0)}) = (-k_0, -K_1, d_1) = (1.2540, 0.6465, 2.4001)$, where k_0, K_1, and d_1 are the notations used in [15, 16].

It should be noted that, since we are seeking a solution that is symmetric with respect to $x_1 = 0$, the above system should be supplemented by an appropriate reflection condition at $x_1 = 0$. A similar comment applies throughout the paper and will not be repeated in the sequel.

Solution Procedure For a given $\overline{\tau}_{\mathrm{w}}$, the process to obtain the solution ϕ to order ε is as follows:

1. From Eq. (17), $P_{\mathrm{H}0} = C_0$ (constant).
2. Solve Eqs. (18a)–(18c) for $m = 0$ under the condition (19a) to obtain $u_{\mathrm{H}0}$, $P_{\mathrm{H}1}$, and $\tau_{\mathrm{H}0}$. Note that $P_{\mathrm{H}1}$ is determined up to an additive constant (say, C_1). Compute $\omega_{\mathrm{H}0}$ from Eq. (18d) with $m = 0$. The leading-order solution $\phi_{\mathrm{H}0}$ is derived from Eq. (13a).

3. Solve Eqs. (18a)–(18c) for $m = 1$ under the conditions (19b) and (19c) to obtain $u_{\mathrm{H}1}$, $P_{\mathrm{H}2}$, and $\tau_{\mathrm{H}1}$. Note that $P_{\mathrm{H}2}$ is determined up to an additive constant (say, C_2). Compute $\omega_{\mathrm{H}1}$ from Eq. (18d) with $m = 1$. The first order solution $\phi_{\mathrm{H}1}+\phi_{\mathrm{K}1}$ is obtained from Eqs. (13b) and (13c).

In the above procedure, $P_{\mathrm{H}m}$, $\omega_{\mathrm{H}m}$, and $\phi_{\mathrm{H}m}$ are determined up to a (common) additive constant C_m at each m, although $u_{i\mathrm{H}m}$ and $\tau_{\mathrm{H}m}$ are determined without such ambiguities. A physical argument can single out a solution. For example, we can specify the gas pressure at a certain point in the domain or specify the average gas density in the domain. Another possibility to remove the ambiguity might be through a symmetry argument (depending on $\overline{\tau}_{\mathrm{w}}$), as in the next section.

4 Case of a Discontinuous Wall Temperature

Now we return to the original problem. Again, we assume that the solution is symmetric with respect to $x_1 = 0$ and restrict the domain to D^-. Moreover, we seek the solution that is antisymmetric with respect to $x_2 = 0$, i.e.,

$$\phi(x_1, -x_2, x_3, \zeta_1, -\zeta_2, \zeta_3) = -\phi(x_1, x_2, x_3, \zeta_1, \zeta_2, \zeta_3). \tag{21}$$

Henceforth, we assume that the solution is x_3-independent, i.e., $\partial_3 = 0$, and even in ζ_3 (hence, $u_3 = 0$).

First, leaving aside the fact that the boundary condition is discontinuous at $(x_1, x_2) = (-\frac{\pi}{2}, 0)$, we look for a solution to the system (1)–(3) in the form

$$\phi = \phi_{\mathrm{HK}} = \phi_{\mathrm{H}} + \phi_{\mathrm{K}}. \tag{22}$$

Here, ϕ_{H} is the Hilbert solution, ϕ_{K} the Knudsen-layer correction, and ϕ_{HK} their sum. Hereafter, we call ϕ_{HK} the Hilbert-Knudsen (HK) solution. Note that ϕ_{H} and ϕ_{K} are subject to the conditions

$$\partial_i \phi_{\mathrm{H}} = O(\phi_{\mathrm{H}}), \quad i = 1, 2, \quad \partial_1 \phi_{\mathrm{K}} = O(\phi_{\mathrm{K}}/\varepsilon), \quad \partial_2 \phi_{\mathrm{K}} = O(\phi_{\mathrm{K}}). \tag{23}$$

As in the previous section, ϕ_{H} and ϕ_{K}, and thus ϕ_{HK}, are expanded in ε as

$$\phi_{\mathrm{H}} = \phi_{\mathrm{H}0} + \varepsilon\phi_{\mathrm{H}1} + \cdots, \tag{24a}$$

$$\phi_{\mathrm{K}} = \varepsilon\phi_{\mathrm{K}1} + \cdots, \tag{24b}$$

$$\phi_{\mathrm{HK}} = \phi_{\mathrm{HK}0} + \varepsilon\phi_{\mathrm{HK}1} + \cdots, \tag{24c}$$

with

$$\phi_{\mathrm{HK}0} = \phi_{\mathrm{H}0}, \quad \phi_{\mathrm{HK}1} = \phi_{\mathrm{H}1} + \phi_{\mathrm{K}1}. \tag{25}$$

To obtain $\phi_{\mathrm{HK}0}$ and $\phi_{\mathrm{HK}1}$, we apply the solution algorithm given in the previous section.

Step 1 The leading-order pressure is $P_{\mathrm{H}0} = C_0$ (constant). We chose $P_{\mathrm{H}0} = C_0 = 0$ in view of the antisymmetry of the solution.

Step 2 The Stokes problem to determine $u_{i\mathrm{H}0}$ and $\tau_{\mathrm{H}0}$ reads

$$\partial_i u_{i\mathrm{H}0} = 0, \quad \gamma_1 \Delta u_{i\mathrm{H}0} - \partial_i P_{\mathrm{H}1} = 0, \quad \Delta \tau_{\mathrm{H}0} = 0, \quad \omega_{\mathrm{H}0} = -\tau_{\mathrm{H}0}, \quad \text{in } D^-, \tag{26a}$$

$$u_{i\mathrm{H}0} = 0, \quad \tau_{\mathrm{H}0} = \pm\tau_{\mathrm{w}}, \quad \text{on } x_1 = -\frac{\pi}{2}, \ x_2 \gtrless 0. \tag{26b}$$

The solution is given by

$$u_{i\mathrm{H}0} = 0, \quad P_{\mathrm{H}1} = 0, \tag{27a}$$

$$\tau_{\mathrm{H}0} = -\omega_{\mathrm{H}0} = \frac{\tau_{\mathrm{w}}}{\pi} \mathrm{Arg}\left(\frac{1+\sin z}{1-\sin z}\right), \quad z = x_1 + \mathrm{i}\, x_2, \tag{27b}$$

where i is the imaginary unit, and the additive constant in $P_{\mathrm{H}1}$ is chosen to be zero because of the solution's antisymmetry. Hence, we obtain the leading-order HK solution as

$$\phi_{\mathrm{HK}0} = \phi_{\mathrm{H}0} = \left(|\zeta|^2 - \frac{5}{2}\right)\tau_{\mathrm{H}0} = \left(|\zeta|^2 - \frac{5}{2}\right)\frac{\tau_{\mathrm{w}}}{\pi}\mathrm{Arg}\left(\frac{1+\sin z}{1-\sin z}\right). \tag{28}$$

Step 3 The Stokes problem for the first order in ε is reduced to

$$\partial_i u_{i\mathrm{H}1} = 0, \quad \gamma_1 \Delta u_{i\mathrm{H}1} - \partial_i P_{\mathrm{H}2} = 0, \quad \Delta \tau_{\mathrm{H}1} = 0, \quad \omega_{\mathrm{H}1} = -\tau_{\mathrm{H}1}, \quad \text{in } D^-, \tag{29a}$$

$$u_{i\mathrm{H}1} = 0, \quad \tau_{\mathrm{H}1} = -\frac{2\tau_{\mathrm{w}} c_1^{(0)}}{\pi}\frac{1}{\sinh x_2}, \quad \text{on } x_1 = -\frac{\pi}{2}, \ x_2 \neq 0. \tag{29b}$$

The solution is given by

$$u_{i\mathrm{H}1} = 0, \quad P_{\mathrm{H}2} = 0, \tag{30a}$$

$$\tau_{\mathrm{H}1} = -\omega_{\mathrm{H}1} = -\frac{8\tau_{\mathrm{w}} c_1^{(0)}}{\pi^2}\frac{x_2 \cos x_1 \cosh x_2 + x_1 \sin x_1 \sinh x_2}{\cos(2x_1) + \cosh(2x_2)}, \tag{30b}$$

where the additive constant in P_{H2} is chosen to be zero because of the solution's antisymmetry. Hence, we obtain the first-order HK solution ϕ_{HK1} as

$$\begin{aligned}\phi_{\mathrm{H1}} &= \left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right)\tau_{\mathrm{H1}} - \zeta_i A(|\boldsymbol{\zeta}|)\partial_i \tau_{\mathrm{H0}} \\ &= -\frac{8\tau_{\mathrm{w}} c_1^{(0)}}{\pi^2}\left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right)\frac{x_2\cos x_1\cosh x_2 + x_1\sin x_1\sinh x_2}{\cos(2x_1)+\cosh(2x_2)} \\ &\quad - \frac{4\tau_{\mathrm{w}}}{\pi}A(|\boldsymbol{\zeta}|)\frac{\zeta_1\sin x_1\sinh x_2 + \zeta_2\cos x_1\cosh x_2}{\cos(2x_1)+\cosh(2x_2)}, \qquad (31a)\end{aligned}$$

$$\phi_{\mathrm{K1}} = -\frac{2\tau_{\mathrm{w}}}{\pi}\frac{1}{\sinh x_2}\varphi_1^{(0)}\left(\frac{x_1+\frac{\pi}{2}}{\varepsilon}, \zeta_1, |\overline{\boldsymbol{\zeta}}|\right), \qquad (31b)$$

$$\phi_{\mathrm{HK1}} = \phi_{\mathrm{H1}} + \phi_{\mathrm{K1}}. \qquad (31c)$$

Drawbacks We have obtained the first two terms of the HK solution $\phi_{\mathrm{HK}} = \phi_{\mathrm{HK0}} + \varepsilon\phi_{\mathrm{HK1}}$ disregarding the fact that the boundary data are discontinuous at $(x_1, x_2) = (-\frac{\pi}{2}, 0)$. This solution has the following drawbacks.

1. The solution does not produce any non-zero flow velocity, which is not meaningful. Note that a non-uniform surface temperature of a body usually causes a rarefied gas flow such as the thermal creep. This remains true even if the temperature distribution is piecewise uniform with a jump discontinuity [2].
2. Near the point $(x_1, x_2) = (-\frac{\pi}{2}, 0)$, the ϕ_{HK0} and ϕ_{HK1} have the following asymptotic properties:

$$\phi_{\mathrm{HK0}} = \tau_{\mathrm{w}}\left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right)\left(\frac{2}{\pi}\theta + \frac{r^2}{6\pi}\sin(2\theta) + O(r^4)\right), \qquad (32a)$$

$$\begin{aligned}\phi_{\mathrm{HK1}} = &-\frac{2\tau_{\mathrm{w}}}{\pi}\left[\frac{c_1^{(0)}\sin\theta}{r}\left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right) + \frac{\zeta_\theta}{r}A(|\boldsymbol{\zeta}|) + \frac{1}{x_2}\varphi_1^{(0)}\left(\frac{x_1+\frac{\pi}{2}}{\varepsilon}, \zeta_1, |\overline{\boldsymbol{\zeta}}|\right)\right] \\ &+ O(r), \qquad (32b)\end{aligned}$$

as $r \searrow 0$, where

$$r = \sqrt{\left(x_1 + \frac{\pi}{2}\right)^2 + x_2^2}, \quad \theta = \mathrm{Arctan}\left(\frac{x_2}{x_1+\frac{\pi}{2}}\right),$$

and $\zeta_\theta = -\zeta_1\sin\theta + \zeta_2\cos\theta$. Thus, $|\phi_{\mathrm{HK1}}|$ grows indefinitely with the rate r^{-1} as $r \searrow 0$. In other words, the ε-expansion of ϕ_{HK} is meaningful only in the region $r \gg \varepsilon$ in D^-.

4.1 Knudsen Zone

Motivated by the above observation, we now look for a solution in the form

$$\phi = \begin{cases} \phi_{\mathrm{HK}} = \phi_{\mathrm{H}} + \phi_{\mathrm{K}} & \text{in } D^- \cap \left\{ (x_1, x_2) \mid r \gg \varepsilon,\ r = \sqrt{\left(x_1 + \frac{\pi}{2}\right)^2 + x_2^2} \right\}, \\ \phi_{\mathrm{Z}} & \text{in } D^- \cap \left\{ (x_1, x_2) \mid r \ll 1,\ r = \sqrt{\left(x_1 + \frac{\pi}{2}\right)^2 + x_2^2} \right\}, \end{cases} \tag{33}$$

allowing ϕ_{HK} and ϕ_{Z} to overlap in the region $\varepsilon \ll r \ll 1$. Here, ϕ_{Z} replaces ϕ_{HK} in the region close to the point of discontinuity $(x_1, x_2) = (-\frac{\pi}{2}, 0)$ (i.e., the Knudsen zone). In the Knudsen zone, the length scale of variation of ϕ_{Z} is assumed to be of the order of ε, i.e., $\partial_i \phi_{\mathrm{Z}} = O(\phi_{\mathrm{Z}}/\varepsilon)$ $(i = 1, 2)$.

To analyze ϕ_{Z}, we introduce new spatial variables by

$$x_i = -\frac{\pi}{2}\delta_{i1} + \varepsilon y_i, \quad i = 1, 2, \tag{34}$$

and assume that $\phi_{\mathrm{Z}} = \phi_{\mathrm{Z}}(y_1, y_2, \boldsymbol{\zeta})$. Expanding ϕ_{Z} in the form

$$\phi_{\mathrm{Z}} = \phi_{\mathrm{Z0}} + \varepsilon \phi_{\mathrm{Z1}} + \cdots, \tag{35}$$

the zeroth-order term ϕ_{Z0} satisfies the following equation and boundary conditions:

$$\zeta_1 \frac{\partial \phi_{\mathrm{Z0}}}{\partial y_1} + \zeta_2 \frac{\partial \phi_{\mathrm{Z0}}}{\partial y_2} = \mathscr{L}(\phi_{\mathrm{Z0}}), \qquad (y_1 > 0,\ -\infty < y_2 < \infty), \tag{36a}$$

$$\phi_{\mathrm{Z0}} = 2\sqrt{\pi} \int_{\zeta_1 < 0} |\zeta_1| \phi_{\mathrm{Z0}} E \pm (|\boldsymbol{\zeta}|^2 - 2)\tau_{\mathrm{w}}, \quad \zeta_1 > 0, \qquad (y_1 = 0,\ y_2 \gtrless 0), \tag{36b}$$

$$\begin{aligned} \phi_{\mathrm{Z0}} \to\ & \frac{2\tau_{\mathrm{w}} \Gamma_z^{(1)}}{|\boldsymbol{y}|} \zeta_r \sin(2\theta) + \frac{2\tau_{\mathrm{w}}}{\pi} \left(|\boldsymbol{\zeta}|^2 - \frac{5}{2} \right) \left(\theta - \frac{c_1^{(0)}}{|\boldsymbol{y}|} \sin\theta \right) \\ & - \frac{2\tau_{\mathrm{w}}}{\pi} \left(\frac{\zeta_\theta}{|\boldsymbol{y}|} A(|\boldsymbol{\zeta}|) + \frac{1}{y_2} \varphi_1^{(0)}(y_1, \zeta_1, |\overline{\boldsymbol{\zeta}}|) \right), \quad \text{as } |\boldsymbol{y}| \to \infty, \end{aligned} \tag{36c}$$

$$\theta = \operatorname{Arctan}\left(\frac{y_2}{y_1} \right), \quad \zeta_r = \zeta_1 \cos\theta + \zeta_2 \sin\theta, \quad \zeta_\theta = -\zeta_1 \sin\theta + \zeta_2 \cos\theta, \tag{36d}$$

where $\Gamma_z^{(1)}$ is a constant that represents the far-field asymptotic property of ϕ_{Z0}, and should be determined together with the solution. This problem can be viewed as a two-dimensional analog of the thermal creep flow [10–12], and represents a

"reaction" of a rarefied gas to a forced temperature variation in the gas. We give further details on the derivation of (36c) in Appendix.

4.2 A Source-Sink Condition for the Flow Velocity

Let us assume that ϕ_{Z0} is known including $\Gamma_z^{(1)}$. We consider a point in D^- such that $\varepsilon \ll r = \sqrt{(x_1+\frac{\pi}{2})^2+x_2^2} \ll 1$, and consider the asymptotic behavior of ϕ_Z in the limit $\varepsilon \searrow 0$, keeping $r\ (=\varepsilon|\boldsymbol{y}|)$ fixed. With the aid of (36c), this is obtained as

$$\begin{aligned}\phi_Z &= \varepsilon\frac{2\tau_w\Gamma_z^{(1)}}{r}\zeta_r\sin(2\theta)+\frac{2\tau_w}{\pi}\left(|\boldsymbol{\zeta}|^2-\frac{5}{2}\right)\left(\theta-\varepsilon\frac{c_1^{(0)}}{r}\sin\theta\right)\\ &\quad-\varepsilon\frac{2\tau_w}{\pi}\left(\frac{\zeta_\theta}{r}A(|\boldsymbol{\zeta}|)+\frac{1}{x_2}\varphi_1^{(0)}\left(\frac{x_1+\frac{\pi}{2}}{\varepsilon},\zeta_1,|\overline{\boldsymbol{\zeta}}|\right)\right)\\ &=\frac{2\tau_w}{\pi}\left(|\boldsymbol{\zeta}|^2-\frac{5}{2}\right)\theta+\varepsilon\left[\frac{2\tau_w\Gamma_z^{(1)}}{r}\zeta_r\sin(2\theta)-\frac{2\tau_w}{\pi}\left(|\boldsymbol{\zeta}|^2-\frac{5}{2}\right)\frac{c_1^{(0)}}{r}\sin\theta\right.\\ &\quad\left.-\frac{2\tau_w}{\pi}\frac{\zeta_\theta}{r}A(|\boldsymbol{\zeta}|)-\frac{2\tau_w}{\pi}\frac{1}{x_2}\varphi_1^{(0)}\left(\frac{x_1+\frac{\pi}{2}}{\varepsilon},\zeta_1,|\overline{\boldsymbol{\zeta}}|\right)\right],\quad \text{as } \varepsilon\searrow 0 \text{ with } r \text{ fixed},\end{aligned} \tag{37}$$

where $\theta = \mathrm{Arctan}(\frac{x_2}{x_1+\frac{\pi}{2}})$. Hence, ϕ_{HK} is matched to the first two terms of ϕ_Z if

$$\begin{aligned}\phi_{HK1}\to{}&\frac{2\tau_w\Gamma_z^{(1)}}{r}\zeta_r\sin(2\theta)-\frac{2\tau_w}{\pi}\left(|\boldsymbol{\zeta}|^2-\frac{5}{2}\right)\frac{c_1^{(0)}}{r}\sin\theta\\ &-\frac{2\tau_w}{\pi}\frac{\zeta_\theta}{r}A(|\boldsymbol{\zeta}|)-\frac{2\tau_w}{\pi}\frac{1}{x_2}\varphi_1^{(0)}\left(\frac{x_1+\frac{\pi}{2}}{\varepsilon},\zeta_1,|\overline{\boldsymbol{\zeta}}|\right),\quad \text{as } r\to 0.\end{aligned} \tag{38}$$

Separating the Hilbert part from the Knudsen-layer part, we have

$$\phi_{H1}\to\frac{2\tau_w\Gamma_z^{(1)}}{r}\zeta_r\sin(2\theta)-\frac{2\tau_w}{\pi}\left(|\boldsymbol{\zeta}|^2-\frac{5}{2}\right)\frac{c_1^{(0)}}{r}\sin\theta-\frac{2\tau_w}{\pi}\frac{\zeta_\theta}{r}A(|\boldsymbol{\zeta}|), \tag{39}$$

as $r\to 0$. Thus, the radial and circumferential components of the flow velocity $u_{rH1}=\langle\zeta_r\phi_{H1}\rangle$ and $u_{\theta H1}=\langle\zeta_\theta\phi_{H1}\rangle$ near the point of discontinuity behave as

$$u_{rH1}\to\frac{\tau_w\Gamma_z^{(1)}}{r}\sin(2\theta),\quad u_{\theta H1}\to 0,\quad \text{as } r\to 0, \tag{40}$$

with

$$r = \sqrt{\left(x_1 + \frac{\pi}{2}\right)^2 + x_2^2}, \quad \theta = \mathrm{Arctan}\left(\frac{x_2}{x_1 + \frac{\pi}{2}}\right). \tag{41}$$

The condition describes a source-sink pair located at $(x_1, x_2) = (-\frac{\pi}{2}, 0)$ and serves as a "boundary condition" that provokes a non-vanishing flow velocity in the Stokes system. As we will see later (Sect. 5), $\Gamma_z^{(1)}$ is likely to be a positive number. Thus, a sink flow toward the discontinuity point appears in the region $x_2 < 0$ and a source flow in the region $x_2 > 0$.

To summarize, after the consideration of the Knudsen zone, **Step 3** should be replaced by

Step 3' The Stokes problem for the first order in ε is given by

$$\partial_i u_{i\mathrm{H}1} = 0, \quad \gamma_1 \Delta u_{i\mathrm{H}1} - \partial_i P_{\mathrm{H}2} = 0, \quad \Delta \tau_{\mathrm{H}1} = 0, \quad \omega_{\mathrm{H}1} = -\tau_{\mathrm{H}1}, \quad \text{in } D^-, \tag{42a}$$

$$u_{i\mathrm{H}1} = 0, \quad \tau_{\mathrm{H}1} = -\frac{2\tau_{\mathrm{w}} c_1^{(0)}}{\pi} \frac{1}{\sinh x_2}, \quad \text{on } x_1 = -\frac{\pi}{2}, \ x_2 \neq 0, \tag{42b}$$

$$u_{r\mathrm{H}1} \to \frac{\tau_{\mathrm{w}} \Gamma_z^{(1)}}{r} \sin(2\theta), \quad u_{\theta\mathrm{H}1} \to 0, \quad \text{as } r = \sqrt{\left(x_1 + \frac{\pi}{2}\right)^2 + x_2^2} \to 0. \tag{42c}$$

The solution $\tau_{\mathrm{H}1}$ is given by (30b), while $(u_{1\mathrm{H}1}, u_{2\mathrm{H}1})$ can be obtained, for instance, by applying the Fourier transform. With these solutions, the first-order HK solution $\phi_{\mathrm{HK}1}$ is given by

$$\begin{aligned} \phi_{\mathrm{H}1} &= 2\zeta_1 u_{1\mathrm{H}1} + 2\zeta_2 u_{2\mathrm{H}1} \\ &\quad - \frac{8\tau_{\mathrm{w}} c_1^{(0)}}{\pi^2} \left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right) \frac{x_2 \cos x_1 \cosh x_2 + x_1 \sin x_1 \sinh x_2}{\cos(2x_1) + \cosh(2x_2)} \\ &\quad - \frac{4\tau_{\mathrm{w}}}{\pi} A(|\boldsymbol{\zeta}|) \frac{\zeta_1 \sin x_1 \sinh x_2 + \zeta_2 \cos x_1 \cosh x_2}{\cos(2x_1) + \cosh(2x_2)}, \end{aligned} \tag{43a}$$

$$\phi_{\mathrm{K}1} = -\frac{2\tau_{\mathrm{w}}}{\pi} \frac{1}{\sinh x_2} \varphi_1^{(0)}\left(\frac{x_1 + \frac{\pi}{2}}{\varepsilon}, \zeta_1, |\overline{\boldsymbol{\zeta}}|\right), \tag{43b}$$

$$\phi_{\mathrm{HK}1} = \phi_{\mathrm{H}1} + \phi_{\mathrm{K}1}. \tag{43c}$$

Note that $\phi_{\mathrm{K}1}$ has not been changed from (31b).

5 Numerical Results for the Knudsen-Zone Problem

Finally, we show some preliminary results for the Knudsen-zone problem. To simplify the numerical analysis, we employ the Bhatnagar-Gross-Krook (BGK) collision operator [4, 20] instead of the Boltzmann collision operator. The linearized BGK collision operator is well-known and its explicit form is omitted [16]. Figure 2a shows the streamlines of the flow velocity (u_{1Z0}, u_{2Z0}) and the (perturbed) temperature τ_{Z0} in the upper-half domain $y_1 \geq 0$ and $y_2 \geq 0$. Here, u_{iZ0} and τ_{Z0} are defined by

$$u_{iZ0} = \langle \zeta_i \phi_{Z0} \rangle, \quad i = 1, 2, \quad \tau_{Z0} = \frac{2}{3} \left\langle \left(|\boldsymbol{\zeta}|^2 - \frac{3}{2} \right) \phi_{Z0} \right\rangle. \tag{44}$$

Note that the wall temperature is discontinuous at $y_2 = 0$ along $y_1 = 0$ (the plates' temperature is $T_0(1 \pm \tau_w)$ for $y_2 \gtrless 0$). Figure 2b shows the flow-velocity vector (u_{1Z0}, u_{2Z0}) and its absolute value near the origin. As seen from these figures, a flow is induced in the positive y_2 direction, which exhibits a diverging flow pattern in the region far from the origin. Note that, by the antisymmetry, it implies that there is a shrinking flow toward the origin in the region $y_2 < 0$. The flow speed is strongest near the discontinuity point and decreases as $\sqrt{y_1^2 + y_2^2}$ increases (see Fig. 2b). In this way, the flow field obtained by the numerical analysis of the BGK model clearly indicates the presence of a source-sink flow pattern in the far field. This becomes the

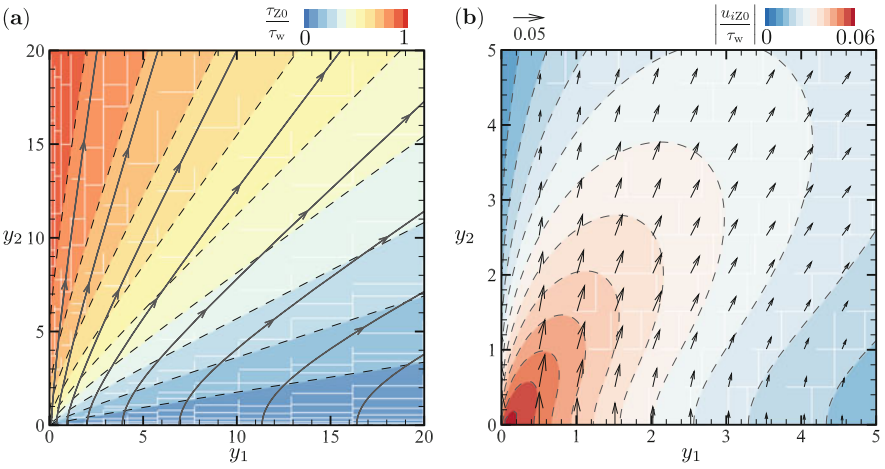

Fig. 2 Numerical results for the Knudsen-zone problem based on the (linearized) BGK collision operator. (**a**) The thick gray curves with arrows show the streamlines of the flow velocity $\tau_w^{-1}(u_{1Z0}, u_{2Z0})$, and the dashed contours show the temperature τ_{Z0}/τ_w. (**b**) A magnified figure near the origin. The arrow indicates the flow-velocity vector $\tau_w^{-1}(u_{1Z0}, u_{2Z0})$ at its starting point, and the contours visualize the absolute value

source-sink condition near the point of discontinuity when rescaled with the spatial variables x_i and the limit $\varepsilon \to 0$ is approached, as discussed in the previous section.

6 Discussions

We have considered a slightly rarefied gas confined between two parallel plates whose common temperature distribution has a jump discontinuity along them. In the case of a smooth temperature distribution without jump discontinuities, the Hilbert expansion and the Knudsen-layer correction yield a practical tool (i.e., the Stokes system) to investigate a thermally-driven flow between the two plates (Sect. 3). On the other hand, the case of the discontinuous surface temperature cannot be handled solely by the Hilbert solution and the Knudsen-layer correction. Indeed, the term ϕ_{HK1} can grow indefinitely near the point of discontinuity, which disproves the validity of the HK solution there (Sect. 4). Given this observation, we have introduced the Knudsen zone near the point $(x_1, x_2) = (-\frac{\pi}{2}, 0)$, in which the solution is allowed to undergo abrupt spatial variations in both x_1 and x_2 directions.

The Knudsen zone is described by the system (36), which is a half-space problem for the linearized Boltzmann equation in two space dimensions. In this problem, the constant $\Gamma_z^{(1)}$ occurring in the far-field asymptotic property (36c) is essential from the macroscopic view points. Indeed, $\Gamma_z^{(1)}$ is inherited to the source-sink condition (42c) in the Stokes system and plays a role to induce a non-zero flow velocity $u_{i\mathrm{H1}}$. In this sense, $\Gamma_z^{(1)}$ is of equal importance as the viscosity or the slip/jump coefficients.

Finally, let us make a brief comment on the global flow structure when ε is small. Since the zeroth-order flow velocity $u_{i\mathrm{H0}}$ is identically zero, the overall flow vanishes as ε tends to zero except in the Knudsen zone. In the Knudsen zone, the nonzero flow of the order $\tau_{\mathrm{w}} O(1)$ is induced as seen from Fig. 2 and remains. However, the Knudsen zone shrinks to $(x_1, x_2) = (-\frac{\pi}{2}, 0)$ with the decrease of ε. Therefore, the strong flow of $\tau_{\mathrm{w}} O(1)$ is gradually localized near $(x_1, x_2) = (-\frac{\pi}{2}, 0)$ as ε becomes smaller. The localized flow affects the global flow at the order ε through the source-sink condition for $u_{i\mathrm{H1}}$ and induces an overall flow with the magnitude $\tau_{\mathrm{w}} O(\varepsilon)$. In this way, a global flow of the order $\tau_{\mathrm{w}} O(\varepsilon)$ is established as a result of the piecewise uniform temperature distribution of the plates. The present analysis successfully provides a clear picture of the flow structure, which is also consistent with the picture inferred in [2].

Appendix

In this appendix, we briefly explain the derivation of the condition (36c). Our stating point is the asymptotic behaviors of the leading order HK solution $\phi_{\mathrm{HK}} = \phi_{\mathrm{HK0}} = \phi_{\mathrm{H0}}$ near $(x_1, x_2) = (-\frac{\pi}{2}, 0)$, i.e.,

$$\phi_{\mathrm{HK0}} = \frac{2\tau_{\mathrm{w}}}{\pi}\left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right)\theta + O(r^2), \quad r \ll 1, \quad \theta = \mathrm{Arctan}\left(\frac{x_2}{x_1 + \frac{\pi}{2}}\right). \tag{45}$$

This suggests that the leading-order term of ϕ_{Z} is of the form

$$\phi_{\mathrm{Z0}} = \frac{2\tau_{\mathrm{w}}}{\pi}\left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right)\theta, \qquad \text{as } |\boldsymbol{y}| \to \infty, \quad y_1 > 0, \quad \theta = \mathrm{Arctan}\left(\frac{y_2}{y_1}\right). \tag{46}$$

Thus, the problem for ϕ_{Z0} consists of (36a), (36b), and (46). We regard this problem as a kind of "scattering problem" and seek a solution with the following asymptotic property [18]:

$$\begin{aligned}\phi_{\mathrm{Z0}} \to{}& \frac{2\tau_{\mathrm{w}}\Gamma_z^{(1)}}{|\boldsymbol{y}|}\zeta_r \sin(2\theta) + \frac{2\tau_{\mathrm{w}}}{\pi}\left(|\boldsymbol{\zeta}|^2 - \frac{5}{2}\right)\left(\theta - \frac{c_1^{(0)}\sin\theta}{|\boldsymbol{y}|}\right) \\ &- \frac{2\tau_{\mathrm{w}}}{\pi}\left(\frac{\zeta_\theta}{|\boldsymbol{y}|}A(|\boldsymbol{\zeta}|) + \frac{1}{y_2}\varphi_1^{(0)}(y_1, \zeta_1, |\overline{\boldsymbol{\xi}}|)\right), \quad \text{as } |\boldsymbol{y}| \to \infty,\end{aligned} \tag{47}$$

where $\Gamma_z^{(1)}$ is a constant. Note that the terms inversely proportional to $|\boldsymbol{y}|$ represent the "reaction" to the imposed external condition (46).

Acknowledgments The present work was supported by JSPS KAKENHI Grant No. 17K06146.

References

1. Aoki, K., Bardos, C., Dogbe, C., Golse, F.: A note on the propagation of boundary induced discontinuities in kinetic theory. Math. Models Methods Appl. Sci. **11**(9), 1581–1595 (2001)
2. Aoki, K., Takata, S., Aikawa, H., Golse, F.: A rarefied gas flow caused by a discontinuous wall temperature. Phys. Fluids **13**(9), 2645–2661 (2001). Erratum: *ibid.* **13**, 3843 (2001)
3. Bardos, C., Caflisch, R.E., Nicolaenko, B.: The Milne and Kramers problems for the Boltzmann equation of a hard sphere gas. Commun. Pure Appl. Math. **39**(3), 323–352 (1986)
4. Bhatnagar, P.L., Gross, E.P., Krook, M.: A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. Phys. Rev. **94**, 511–525 (1954)
5. Coron, F., Golse, F., Sulem, C.: A classification of well-posed kinetic layer problems. Commun. Pure Appl. Math. **41**(4), 409–435 (1988)

6. Esposito, R., Guo, Y., Kim, C., Marra, R.: Non-isothermal boundary in the Boltzmann theory and Fourier law. Commun. Math. Phys. **323**, 177–239 (2013)
7. Guo, Y., Kim, C., Tonon, D., Trescases, A.: BV-regularity of the Boltzmann equation in non-convex domains. Arch. Rat. Mech. Anal. **220**(3), 1045–1093 (2016)
8. Kawagoe, D., Chen, I.K.: Propagation of boundary-induced discontinuity in stationary radiative transfer. J. Stat. Phys. **170**(1), 127–140 (2018)
9. Kim, C.: Formation and propagation of discontinuity for Boltzmann equation in non-convex domains. Commun. Math. Phys. **308**, 641–701 (2011)
10. Loyalka, S.K.: Slip in the thermal creep flow. Phys. Fluids **14**(1), 21–24 (1971)
11. Ohwada, T., Sone, Y., Aoki, K.: Numerical analysis of the shear and thermal creep flows of a rarefied gas over a plane wall on the basis of the linearized Boltzmann equation for hard-sphere molecules. Phys. Fluids A **1**(9), 1588–1599 (1989)
12. Sone, Y.: Thermal creep in rarefied gas. J. Phys. Soc. Jpn. **21**, 1836–1837 (1966)
13. Sone, Y.: Asymptotic theory of flow of rarefied gas over a smooth boundary I. In: Trilling, L., Wachman, H.Y. (eds.) Rarefied Gas Dynamics, vol. 1, pp. 243–253. Academic Press, New York (1969)
14. Sone, Y.: Asymptotic theory of flow of rarefied gas over a smooth boundary II. In: Dini, D. (ed.) Rarefied Gas Dynamics, vol. 2, pp. 737–749. Editrice Tecnico Scientfica, Pisa (1971)
15. Sone, Y.: Kinetic Theory and Fluid Dynamics. Birkhäuser, Boston (2002). Supplementary Notes and Errata: Kyoto University Research Information Repository. http://hdl.handle.net/2433/66099
16. Sone, Y.: Molecular Gas Dynamics: Theory, Techniques, and Applications. Birkhäuser, Boston (2007). Supplementary Notes and Errata: Kyoto University Research Information Repository. http://hdl.handle.net/2433/66098
17. Taguchi, S., Aoki, K.: Rarefied gas flow around a sharp edge induced by a temperature field. J. Fluid Mech. **694**, 191–224 (2012)
18. Taguchi, S., Tsuji, T.: On the motion of slightly rarefied gas induced by a discontinuous surface temperature. J. Fluid Mech. **897**, A16 (2020)
19. Takata, S., Hattori, M.: Asymptotic theory for the time-dependent behavior of a slightly rarefied gas over a smooth solid boundary. J. Stat. Phys. **147**(6), 1182–1215 (2012)
20. Welander, P.: On the temperature jump in a rarefied gas. Ark. Fys. **7**, 507–553 (1954)

A Revisit to the Cercignani–Lampis Model: Langevin Picture and Its Numerical Simulation

Shigeru Takata, Shigenori Akasobe, and Masanari Hattori

Abstract The Cercignani–Lampis (CL) model for the gas–surface interaction is revisited from the Langevin dynamics viewpoint. Starting from a time-independent Fokker–Planck formalism by Cercignani, its time-dependent extension and the corresponding Langevin description are introduced. The Langevin description sheds light on dynamical features of a stochastic process corresponding to the CL model. Numerical simulations on the basis of the Langevin description are performed as well to reproduce the scattering kernel and reflection intensity distribution numerically. Although the noise in the stochastic process is apparently simple, the Milstein scheme rather than the Euler–Maruyama scheme has to be adopted to achieve a satisfactory numerical convergence in time discretisation.

1 Introduction

Gas flows in low pressure and small-scale circumstances, which we generically call rarefied gas flows, require the kinetic theory description rather than the usual fluid dynamics description because the latter is implicitly limited to the local equilibrium states [4, 16]. Inter-molecular collisions inside the gas are not necessarily frequent in such circumstances, and sometimes molecular velocities inside the gas can be traced back without changes to the velocities just after the reflection on a container surface, i.e., a wall. Hence, the velocity distribution of reflected molecules can have a direct impact on the gas behavior in the bulk region. An enough simple but realistic gas–surface interaction model has been desired for a long time.

S. Takata (✉) · M. Hattori
Department of Aeronautics and Astronautics, Kyoto University, Kyoto, Japan

Research Project of Fluid Science and Engineering, Advanced Engineering Research Center, Kyoto University, Kyoto, Japan
e-mail: takata.shigeru.4a@kyoto-u.ac.jp

S. Akasobe
Department of Aeronautics and Astronautics, Kyoto University, Kyoto, Japan

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*, Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_15

Many efforts have been devoted even in rather recent years by Molecular Dynamics (MD), theoretical, and experimental approaches (e.g., [1, 2, 18] and references therein; a very good survey of the gas–surface interaction models before 1990s can be found in [4]). Nevertheless, the progress so far is not necessarily satisfactory, probably due to difficulties of background physics in such interface problems. Even now, the diffuse reflection condition and/or its convex combination with the specular reflection condition, the so-called Maxwell condition, are primarily used and regarded as the standard in the literature [4, 16]. The former implicitly assumes perfect accommodation of incident molecules with the wall and reproduces the Lambert cosine law of the reflection intensity, while the latter introduced by Maxwell takes account of imperfect accommodation. Although the latter reproduces some effects of the imperfect accommodation at a macroscopic level, the specular reflection part induces a spike in the reflection intensity distribution, which is different from observations in molecular beam experiments.

After Maxwell, the concept of accommodation has been developed to introduce different coefficients to represent a possible difference of accommodation in momentum and energy exchanges [4, 11, 15]. Cercignani and Lampis [5] proposed in 1971 a mathematical physical model, which is now called the Cercignani–Lampis (CL) model. A similar model was independently proposed by Kŭscer et al. [12]. Their models have an impact in their capability to reproduce typical features of the reflection intensity distributions experimentally observed.

The CL model has been enjoying successful practical applications, including its extension and easy implementation [13] to the Direct Simulation Monte Carlo (DSMC) algorithm since 1990s. Nevertheless, it seems that the dynamical background is still behind a mysterious veil, though its physical interpretation and alternative derivation were reported in 1970s (e.g., [6, 17]). No further attempts have been made to shed light on the dynamical aspects of the model. It is the main motivation of the present study.

In the present paper, we discuss the CL model mainly along the lines laid by Cercignani in [4]. We, however, modify his original discussions for a time-dependent problem in order to have a stochastic dynamical picture, the Langevin equation description. Results of numerical simulations and scheme accuracy in time discretisation will be presented as well.

2 Scattering Kernel and Cercignani–Lampis (CL) Model

Let us denote by $f(t, \boldsymbol{x}, \boldsymbol{\xi})$ the velocity distribution function of gas molecules, where t is a time, $\boldsymbol{x}$ is a position, and $\boldsymbol{\xi}$ is a molecular velocity. Assuming that a resting solid wall occupies the region $x_1 < 0$, the reflection law for gas molecules on the wall is expressed as

$$f(t, \boldsymbol{x}_\parallel, x_1 = 0, \boldsymbol{\xi}) = \int_{\overline{\xi}_1 < 0} K(\boldsymbol{x}_\parallel, \boldsymbol{\xi}, \overline{\boldsymbol{\xi}}) f(t, \boldsymbol{x}_\parallel, x_1 = 0, \overline{\boldsymbol{\xi}}) d\overline{\boldsymbol{\xi}}, \quad \xi_1 > 0, \tag{1}$$

or equivalently as

$$\xi_1 f(t, \boldsymbol{x}_\parallel, x_1 = 0, \boldsymbol{\xi}) = \int_{\overline{\xi}_1 < 0} \mathcal{R}(\boldsymbol{x}_\parallel, \boldsymbol{\xi}, \overline{\boldsymbol{\xi}}) |\overline{\xi}_1| f(t, \boldsymbol{x}_\parallel, x_1 = 0, \overline{\boldsymbol{\xi}}) d\overline{\boldsymbol{\xi}}, \quad \xi_1 > 0. \tag{2}$$

Here $\boldsymbol{x}_\parallel = (x_2, x_3)$, which will be suppressed mostly in what follows because the discussion is not concerned with the variation of K (or $\mathcal{R}$) in that direction. In the present paper, we shall call K the scattering kernel and $\mathcal{R}$ the reflection probability, respectively.[1] They are related to each other as

$$|\xi_1| K(\boldsymbol{\xi}, \overline{\boldsymbol{\xi}}) = |\overline{\xi}_1| \mathcal{R}(\boldsymbol{\xi}, \overline{\boldsymbol{\xi}}), \tag{3}$$

and are usually supposed to be independent of f both in physics and mathematics. Physically, it implies that the microscopic properties of the wall do not change by the interaction with the gas. We follow this convention, and thus the right-hand sides of (1) and (2) are linear with respect to f. Experiments of mono-collimated molecular beam scattering are performed on the basis of the same convention, though it is not explicitly mentioned. In the case of the diffuse reflection condition, the scattering kernel reads

$$K = \frac{|\overline{\xi}_1|}{2\pi (RT_w)^2} \exp\left(-\frac{|\boldsymbol{\xi}|^2}{2RT_w}\right), \tag{4}$$

where T_w is the wall temperature and R is the specific gas constant (the Boltzmann's constant k_B divided by the mass of a molecule m; $R = k_B/m$). Cercignani and Lampis [5] proposed the following form of the scattering kernel:

$$K = \frac{|\overline{\xi}_1|}{2\pi (RT_w)^2} \frac{1}{\alpha_t (2 - \alpha_t) \alpha_n} I_0\left(\frac{\xi_1 \overline{\xi}_1}{RT_w} \frac{\sqrt{1 - \alpha_n}}{\alpha_n}\right) \exp\left(-\frac{\xi_1^2 + \overline{\xi}_1^2 (1 - \alpha_n)}{2RT_w \alpha_n}\right) \times \exp\left(-\frac{|\boldsymbol{\xi}_\parallel - \overline{\boldsymbol{\xi}}_\parallel (1 - \alpha_t)|^2}{2RT_w \alpha_t (2 - \alpha_t)}\right), \tag{5}$$

where $\boldsymbol{\xi}_\parallel = (\xi_2, \xi_3)$ and I_0 is the modified Bessel function of the first kind and zeroth order:

$$I_0(x) \equiv \frac{1}{2\pi} \int_0^{2\pi} \exp(x \cos \varphi) d\varphi. \tag{6}$$

The boundary condition (1) with the kernel (5) is called the Cercignani–Lampis (CL) model and contains two adjustable parameters: $0 \leq \alpha_n \leq 1$ and $0 \leq \alpha_t \leq 2$. When $\alpha_n = \alpha_t = 1$, it recovers the diffuse reflection condition (4).

[1] We have adopted the terminology in [16]. In [4], $\mathcal{R}$ is called the scattering kernel, which is a flux based terminology like (2). As $\int_{\xi_1 > 0} \mathcal{R} d\boldsymbol{\xi} = 1$ and $\mathcal{R} \geq 0$, $\mathcal{R}$ can be interpreted as the probability density of finding a reflected molecule at a specific value of the velocity.

3 Cercignani's Fokker–Planck (FP) System

In [3], Cercignani introduced a time-independent Fokker–Planck system for the probability density $P(x_1, \boldsymbol{\xi})$ of a molecule at position x_1 with velocity $\boldsymbol{\xi}$. It reproduces the CL model in the parameter range $0 \le \alpha_n \le 1$ and $0 \le \alpha_t \le 1$ and reads

$$\xi_1 \frac{\partial P}{\partial x_1} + \frac{\partial P}{\partial \xi_i} X_i = LP, \quad (-d < x_1 < 0), \tag{7a}$$

$$LP = \frac{\partial^2}{\partial \xi_j \partial \xi_i} \left(D_{ij} P \right) + \frac{\partial}{\partial \xi_i} \left[\left(F_{ij} \xi_j - \frac{\partial D_{ij}}{\partial \xi_j} \right) P \right], \tag{7b}$$

$$\text{b.c. } P(x_1 = 0, \xi_1 < 0, \boldsymbol{\xi}_\parallel) = \delta(\boldsymbol{\xi} - \boldsymbol{\xi}_{\text{in}}), \tag{7c}$$

$$P(x_1 = -d, \xi_1, \boldsymbol{\xi}_\parallel) = P(x_1 = -d, -\xi_1, \boldsymbol{\xi}_\parallel), \quad \xi_1 > 0. \tag{7d}$$

Here $\boldsymbol{\xi}_{\text{in}}$ is the molecular velocity of incidence, the interaction with the wall is supposed to occur in $x_1 < 0$, and $x_1 = -d$ is the position of the potential barrier beyond which a molecule is forbidden to proceed (Fig. 1). The X_i, D_{ij}, and F_{ij} in the above are defined as follows:

$$X_i = 0, \quad D_{11} = \frac{2RT_w}{\ell_n} |\xi_1|, \quad D_{22} = D_{33} = \frac{2RT_w}{\ell_t} |\xi_1|, \tag{7e}$$

$$D_{ij} = 0 \quad (i \ne j), \quad F_{ij} = \frac{1}{RT_w} D_{ij}, \tag{7f}$$

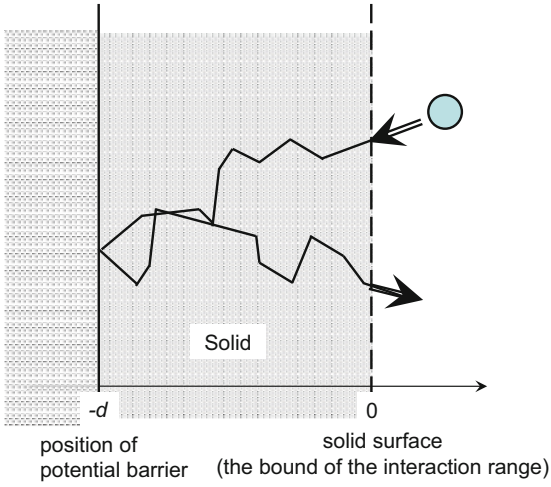

Fig. 1 Schematics of scattering of a gas molecule

where ℓ_t and ℓ_n are a characteristic length of molecular velocity diffusion in the x_2x_3-plane and that in the x_1-direction, respectively. By solving the above system (7), we have the velocity distribution $P(x_1 = 0, \xi_1 > 0, \boldsymbol{\xi}_\parallel)$ of reflected molecules against the incident molecular beam $\delta(\boldsymbol{\xi} - \boldsymbol{\xi}_{\rm in})$. Substitution of $f = \delta(\boldsymbol{\xi} - \boldsymbol{\xi}_{\rm in})$ into (1) or (2) gives the relation

$$\mathcal{R}(\xi_1 > 0, \boldsymbol{\xi}_{\rm in}) = \frac{|\xi_1|}{|\xi_{\rm in1}|} K(\xi_1 > 0, \boldsymbol{\xi}_{\rm in}) = \frac{|\xi_1|}{|\xi_{\rm in1}|} P(x_1 = 0, \xi_1 > 0), \tag{8}$$

(see [4, Sec. III. 2, Eq. (2.12)]). Hence, finding the form of K is identical to finding P at $x_1 = 0$ for $\xi_1 > 0$. Here and in what follows, we suppress $\boldsymbol{\xi}_\parallel$ in the argument of K etc., if no confusion is expected.

4 From Fokker–Planck to Langevin System

The time-independent Fokker–Planck (FP) system in Sect. 3 is the starting point of our discussions. We first introduce its simple but natural extension to the time-dependent situation. Then, we identify the Langevin system, namely the stochastic dynamics of a test particle, that is equivalent to the extended system.

4.1 Time-Dependent Fokker–Planck System

In order to draw out a dynamical picture behind the CL model, we simply add a time derivative term to the left-hand side of (7a), allow the spatial dependence in (x_2, x_3)-directions, and modify the condition (7c) in accordance with the time and spatial localization of the incident molecular beam. Then, we have the following initial- and boundary-value problem:

$$\frac{\partial Q}{\partial t} = -\xi_i \frac{\partial Q}{\partial x_i} + \frac{\partial^2}{\partial \xi_j \partial \xi_i} \left(D_{ij} Q \right) + \frac{\partial}{\partial \xi_i} \left[\left(F_{ij} \xi_j - \frac{\partial D_{ij}}{\partial \xi_j} \right) Q \right], \ (-d < x_1 < 0), \tag{9a}$$

$$\text{b.c. } Q(t, \boldsymbol{x}, \boldsymbol{\xi}) = \delta(t) \delta(\boldsymbol{x}) \delta(\boldsymbol{\xi} - \boldsymbol{\xi}_{\rm in}), \quad \xi_1 < 0, \quad x_1 = 0, \tag{9b}$$

$$Q(t, x_1 = -d, \xi_1) = Q(t, x_1 = -d, -\xi_1), \quad \xi_1 > 0, \tag{9c}$$

where $Q(t, \boldsymbol{x}, \boldsymbol{\xi})$ is the probability density finding a molecule at time t, position $\boldsymbol{x}$, and velocity $\boldsymbol{\xi}$. As Q is the fundamental solution (the Green function) to the initial- and boundary-value problem for the same FP equation, we switch its notation to $G(0, \mathbf{0}, \boldsymbol{\xi}_{\rm in}; t, \boldsymbol{x}, \boldsymbol{\xi})$ from now on. Here, the first three arguments of G indicate that the time, position, and velocity of incidence are $t = 0$, $\boldsymbol{x} = 0$, and $\boldsymbol{\xi} = \boldsymbol{\xi}_{\rm in}$,

respectively. Since the microscopic property of the wall, or the coefficients D_{ij} and F_{ij}, are independent of t and $\boldsymbol{x}_\parallel$, the solution is invariant under the translation both in time and in the x_2x_3-plane:

$$G(s, \boldsymbol{x}_{\rm in}, \boldsymbol{\xi}_{\rm in}; t, \boldsymbol{x}, \boldsymbol{\xi}) = G(0, \boldsymbol{0}, \boldsymbol{\xi}_{\rm in}; t-s, \boldsymbol{x}-\boldsymbol{x}_{\rm in}, \boldsymbol{\xi}), \quad s \le t, \tag{10}$$

where $\boldsymbol{x}_{\rm in} = (0, \boldsymbol{x}_{\rm in\parallel})$. This motivates us to define $\overline{G} \equiv \int_{-\infty}^{t}\int_{\mathbb{R}^2} G(s, \boldsymbol{x}_{\rm in}, \boldsymbol{\xi}_{\rm in}; t, \boldsymbol{x}, \boldsymbol{\xi}) d\boldsymbol{x}_\parallel ds$. The following property holds:

$$\begin{aligned}\overline{G} &= \int_{-\infty}^{t}\int_{\mathbb{R}^2} G(s, \boldsymbol{x}_{\rm in}, \boldsymbol{\xi}_{\rm in}; t, \boldsymbol{x}, \boldsymbol{\xi}) d\boldsymbol{x}_\parallel ds \\ &= \int_{-\infty}^{t}\int_{\mathbb{R}^2} G(0, \boldsymbol{0}, \boldsymbol{\xi}_{\rm in}; t-s, \boldsymbol{x}-\boldsymbol{x}_{\rm in}, \boldsymbol{\xi}) d\boldsymbol{x}_\parallel ds \\ &= \int_{0}^{\infty}\int_{\mathbb{R}^2} G(0, \boldsymbol{0}, \boldsymbol{\xi}_{\rm in}; \tau, \boldsymbol{x}, \boldsymbol{\xi}) d\boldsymbol{x}_\parallel d\tau. \end{aligned} \tag{11}$$

It is seen from the last equality that $\overline{G}$ is a solution of (9a) independent of $\boldsymbol{x}_{\rm in\parallel}$ as well as t and $\boldsymbol{x}_\parallel$; accordingly it will be denoted as $\overline{G}(\boldsymbol{\xi}_{\rm in}; x_1, \boldsymbol{\xi})$. Note that $\overline{G}$ solves (7a) as well. It is readily seen from (9b) and (9c) that $\overline{G}$ satisfies the conditions (7c) and (7d). Thus, $\overline{G}$ is a solution of the system (7).

In Sect. 4.2, we present the Langevin system corresponding to the above system (9). The observation on $\overline{G}$ tells that the scattering kernel K is identical with $\overline{G}(\boldsymbol{\xi}_{\rm in}; x_1 = 0, \boldsymbol{\xi})$ and thus can be constructed from G. This implies that the kernel of the CL model can be reproduced by many samples of a test particle simulation of the Langevin system. We will come back to this issue in Sect. 5.2.

4.2 *Langevin System for the CL Model: A Stochastic Dynamical Picture*

We first consider the following Langevin equation:

$$dx_i = \xi_i dt, \quad d\xi_i = (-\gamma_{ij}\xi_j + F_i)dt + S_{ij}dW_j, \tag{12}$$

where W_j is the Wiener process that satisfies $\langle dW_i dW_j \rangle = dt\,\delta_{ij}$. Just for convenience, let us introduce a six-dimensional vector variable y_α ($\alpha = 1, \ldots, 6$) defined by $y_i = x_i$ and $y_{i+3} = \xi_i$ ($i = 1, \ldots, 3$) and rewrite (12) as follows:

$$dy_\alpha = (A_{\alpha\beta}y_\beta + B_\alpha)dt + \Theta_{\alpha i}dW_i, \tag{13a}$$

where

$$[A_{\alpha\beta}] = \begin{bmatrix} 0\,0\,0 & 1 & 0 & 0 \\ 0\,0\,0 & 0 & 1 & 0 \\ 0\,0\,0 & 0 & 0 & 1 \\ 0\,0\,0 & -\gamma_{11} & -\gamma_{12} & -\gamma_{13} \\ 0\,0\,0 & -\gamma_{21} & -\gamma_{22} & -\gamma_{23} \\ 0\,0\,0 & -\gamma_{31} & -\gamma_{32} & -\gamma_{33} \end{bmatrix}, \tag{13b}$$

$$[B_\alpha] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ F_1 \\ F_2 \\ F_3 \end{bmatrix}, \quad [\Theta_{\alpha i}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix}. \tag{13c}$$

The corresponding Fokker–Planck equation is known to take the following form [7]:

$$\frac{\partial g}{\partial t} = -\frac{\partial}{\partial y_\alpha}([A_{\alpha\beta} y_\beta + B_\alpha] g) + \frac{1}{2}\frac{\partial^2}{\partial y_\alpha \partial y_\beta}(\Theta_{\alpha i}\Theta_{\beta i} g). \tag{14}$$

Using the original pair of three-dimensional vector variables (x_i, ξ_i) in place of y_α, (14) is rewritten as

$$\frac{\partial g}{\partial t} = -\frac{\partial}{\partial x_i}(\xi_i g) - \frac{\partial}{\partial \xi_i}(-\gamma_{ij}\xi_j g + F_i g) + \frac{1}{2}\frac{\partial^2}{\partial \xi_i \partial \xi_j}(S_{ik} S_{jk} g). \tag{15}$$

Now, comparing (15) and (9a) leads us to find the correspondence of coefficients:

$$\frac{1}{2} S_{ik} S_{jk} = D_{ij}, \quad \gamma_{ij}\xi_j - F_i = F_{ij}\xi_j - \frac{\partial D_{ij}}{\partial \xi_j}. \tag{16}$$

Note that, because of the definition (7e),

$$\frac{\partial D_{ij}}{\partial \xi_j} = \frac{\xi_1}{|\xi_1|}\frac{2RT_w}{\ell_n}\delta_{i1}. \tag{17}$$

Thus, γ_{ij}, S_{ij}, and F_i are identified as

$$\gamma_{ij} = F_{ij} = \frac{1}{RT_w} D_{ij} = 2[\frac{|\xi_1|}{\ell_n}\delta_{i1}\delta_{j1} + \frac{|\xi_1|}{\ell_t}(\delta_{i2}\delta_{j2} + \delta_{i3}\delta_{j3})], \tag{18a}$$

$$F_i = \frac{\xi_1}{|\xi_1|}\frac{2RT_w}{\ell_n}\delta_{i1}, \tag{18b}$$

$$S_{ij} = 2\sqrt{RT_w\frac{|\xi_1|}{\ell_n}}\delta_{i1}\delta_{j1} + 2\sqrt{RT_w\frac{|\xi_1|}{\ell_t}}(\delta_{i2}\delta_{j2} + \delta_{i3}\delta_{j3}). \tag{18c}$$

Here, we have chosen S_{ij} to be symmetric.

To summarize, we have identified the Langevin system corresponding to the time-dependent FP system (9):

$$dx_i = \xi_i dt, \quad (i = 1, 2, 3), \tag{19a}$$

$$d\xi_1 = -\frac{2}{\ell_n}(|\xi_1|\xi_1 - \frac{\xi_1}{|\xi_1|}RT_w)dt + 2\sqrt{RT_w\frac{|\xi_1|}{\ell_n}}dW_1, \tag{19b}$$

$$d\xi_2 = -\frac{2}{\ell_t}|\xi_1|\xi_2 dt + 2\sqrt{RT_w\frac{|\xi_1|}{\ell_t}}dW_2, \tag{19c}$$

$$d\xi_3 = -\frac{2}{\ell_t}|\xi_1|\xi_3 dt + 2\sqrt{RT_w\frac{|\xi_1|}{\ell_t}}dW_3, \tag{19d}$$

supplemented by the specular reflection at the potential barrier $x_1 = -d$ and the initial condition

$$\boldsymbol{x}(0) = \boldsymbol{0}, \quad \boldsymbol{\xi}(0) = \boldsymbol{\xi}_{\text{in}}. \tag{19e}$$

5 Discussions

5.1 Dynamical Aspects of the CL Model

The Langevin system (19) tells that, after the incidence, a molecule changes its velocity under two types of interactions with the wall. One is the first term on the right-hand side of (19b)–(19d), which we call a *drift* part. The other is the second term on the same side of (19b)–(19d), which we call a *diffusion* part. Below we discard the spatial translation (19a) and concentrate on the dynamics described by (19b)–(19d). Before going into details, it should be noted that the motion in the

normal direction is seen to be independent of those in tangential directions. The reverse is not true.

Role of the *Drift* Part

In the direction normal to the surface, the *drift* part decelerates a molecule if the kinetic energy $(1/2)m\xi_1^2$ in that direction is beyond the thermodynamic energy $(1/2)k_B T_w$ distributed by the equipartition law [8]. If not, it accelerates the molecule until the kinetic energy $(1/2)m\xi_1^2$ reaches that energy. To see the mechanism more closely, discard the second term in (19b) and integrate it in time. Then, we have for $\xi_1 < 0$

$$\xi_1 = -\sqrt{RT_w}\frac{1 \mp c_- \exp(-\frac{4\sqrt{RT_w}}{\ell_n}t)}{1 \pm c_- \exp(-\frac{4\sqrt{RT_w}}{\ell_n}t)}, \quad -\sqrt{RT_w} \leqq \xi_1 (< 0), \tag{20a}$$

and for $\xi_1 > 0$

$$\xi_1 = \sqrt{RT_w}\frac{1 \pm c_+ \exp(-\frac{4\sqrt{RT_w}}{\ell_n}t)}{1 \mp c_+ \exp(-\frac{4\sqrt{RT_w}}{\ell_n}t)}, \quad \sqrt{RT_w} \leqq \xi_1 (> 0), \tag{20b}$$

where $c_\pm$ is a positive constant not larger than unity. Hence, there is no reversal of motion in the normal direction if neither thermal noise nor potential barrier exist. The *drift* part thus drives ξ_1 towards $\pm\sqrt{RT_w}$ depending on its sign exponentially in time.[2]

In directions tangential to the surface, the *drift* part always decelerates the molecular motion in proportion to the momentum transferred by the incoming molecule, i.e., $-|\xi_1|\boldsymbol{\xi}_\parallel$, where $\boldsymbol{\xi}_\parallel = (\xi_2, \xi_3)$. Thus, it works on the molecule in a similar way to the viscous drag. To see the effect more closely, consider the motion in the x_2-direction. The motion in the x_3-direction follows the same dynamics, as is clear from (19c) and (19d). As before, discarding the second term in (19c) and integrating it in time give

$$\xi_2 = c_0 \exp(-\frac{2}{\ell_t}\int |\xi_1| dt), \tag{21}$$

[2] Using the relation (31) that appears later, the exponential factor can be rewritten as

$$\exp(-\frac{4\sqrt{RT_w}}{\ell_n}t) = (1-\alpha_n)^{\frac{\sqrt{RT_w}}{2d}t}.$$

Hence, if $|\xi_1| < \sqrt{RT_w}$, the molecule stays longer than $2d/\sqrt{RT_w}$ in the interaction region, making the above factor smaller and smaller until leaving. If $|\xi_1| > \sqrt{RT_w}$, the molecule stays shorter, keeping the same factor between $1-\alpha_n$ and unity.

where c_0 is a constant. Hence, as far as $\xi_1 \neq 0$, the *drift*-part force decelerates ξ_2 and makes it vanish if the integration of $|\xi_1|$ in time is not bounded. It is also seen that the larger $|\xi_1|$ is, the larger the decaying rate is.

Role of the *diffusion* Part and Competition with the *drift* Part
In order to see the role of the *diffusion* part and its competition with the *drift* part, we go back to the time-dependent FP system, (9a) without the spatial translation term and (9b).

Let us first single out the *diffusion* part. The FP system without the spatial translation and the *drift* part admits a stationary solution inversely proportional to $|\xi_1|$ in the normal direction, provided that the integrability condition is discarded. In the tangential directions, it is just a usual diffusion process without center shifting, and only its time scale depends on $|\xi_1|$. A couple of examples of particle simulations of (19b) and (19c) without the *drift* part are shown in Fig. 2a,b, which clearly demonstrate those features. The *diffusion* part competes with the *drift* part to form the half-range Maxwellian in the normal direction and the full Maxwellian with the zero mean velocity in the tangential directions as a stationary state; see Fig. 2c,d. The admitted stationary solution under the competition corresponds to the full accommodation situation, namely the diffuse reflection model. In the CL model, however, molecules spatially translate in the interaction region and may leave there before reaching the full accommodation.

5.2 *Langevin System and the Reflection Intensity Distribution*

We first consider the way how to recover $\overline{G}$ from the samples (test particles) of the Langevin system simulation. The base of our discussion is the identity

$$\overline{G}(\overline{\boldsymbol{\xi}}; 0, \boldsymbol{\xi}) = \int_0^\infty \int_{\mathbb{R}^2} G(0, \mathbf{0}, \overline{\boldsymbol{\xi}}; \tau, x_1 = 0, \boldsymbol{x}_\parallel, \boldsymbol{\xi}) d\boldsymbol{x}_\parallel d\tau, \tag{22}$$

which has already appeared in Sect. 4.1. Remind that solving the Langevin system is identical to getting the above integrand G.

Taking account of the time and the spatial integration in (22), let us first count the number of sample molecular velocities at the instance of exit from the interaction region $x_1 < 0$, irrespective of the $\boldsymbol{x}_\parallel$ position and exit time. Then, $\mathcal{N}$ samples for the common velocity of incidence $\overline{\boldsymbol{\xi}}$ yields a normalized distribution in the molecular velocity:

$$\frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta(\boldsymbol{\xi} - \boldsymbol{\xi}^{(i)}), \tag{23}$$

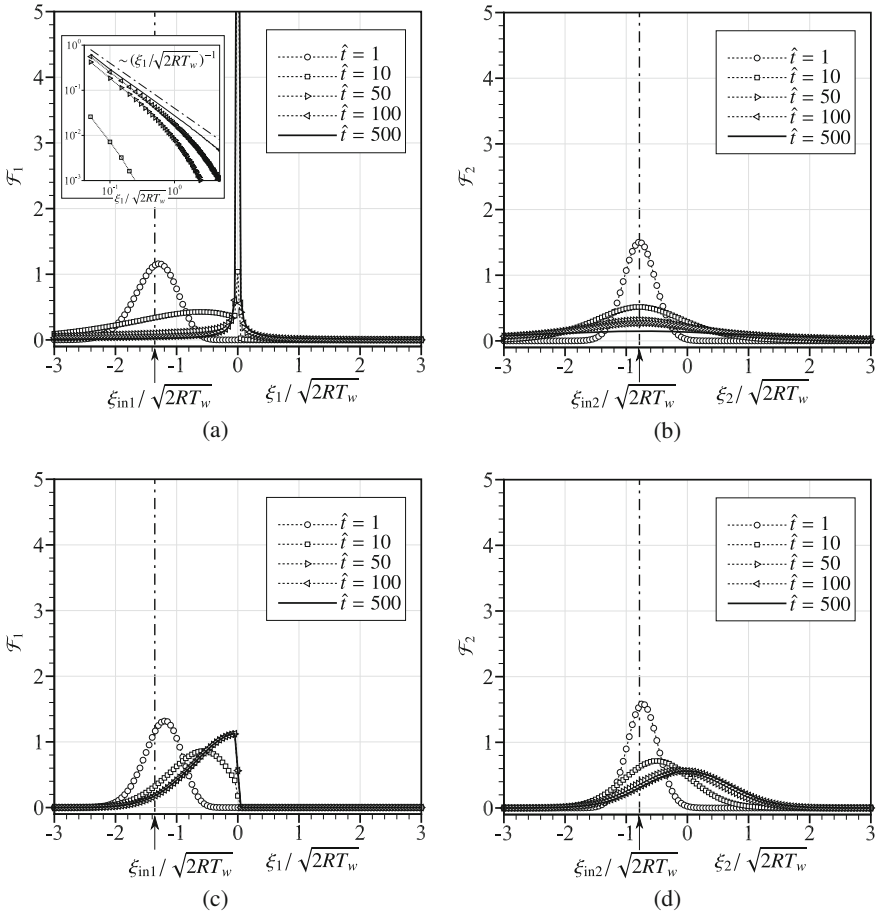

Fig. 2 The *diffusion*-part effect and the competition between the *drift* and the *diffusion* part in the case $|\boldsymbol{\xi}_{\rm in}|/\sqrt{2RT_w} = 1.56718$ with $(\xi_{\rm in2}, \xi_{\rm in3})/\sqrt{2RT_w} = (-0.78359, 0)$. No spatial translation is considered. (**a**) the *diffusion*-part effect in the normal direction, (**b**) the *diffusion*-part effect in the incident tangential direction, (**c**) the competition between the *drift*- and the *diffusion*-part in the normal direction, and (**d**) the competition between the *drift*- and the *diffusion*-part in the incident tangential direction. Here, $\hat{t} = (\sqrt{2RT_w}/\ell)t$ with ℓ being $\ell = -(1/8)(\ln 0.7)\ell_n = -(1/4)(\ln 0.9)\ell_t$ and $\mathcal{F}_\alpha(\xi_\alpha) = \sum_{i=1}^{\mathcal{N}} \chi_{[\xi_\alpha, \xi_\alpha+\Delta\xi_\alpha]}(\xi_\alpha^{(i)})/(\mathcal{N}\Delta\xi_\alpha/\sqrt{2RT_w})$ $(\alpha = 1, 2)$ with $\boldsymbol{\xi}^{(i)}$ and χ_A being the molecular velocity of the i-th sample of simulation and the characteristic function of A, respectively [see the sentence following (29) that appears later]. The Milstein scheme to be explained later is used with the timestep $\Delta t = 0.0002(\ell/\sqrt{2RT_w})$, and the number of sampling and the intervals for the histogram of $\mathcal{F}_\alpha$ in molecular velocity are respectively $\mathcal{N} = 10^7$ and $\Delta\xi_1 = \Delta\xi_2 = 0.05\sqrt{2RT_w}$

where $\boldsymbol{\xi}^{(i)}$ is the molecular velocity at the instance of exit in the i-th sample of simulation. The above simple counting is, however, not directly connected with $\overline{G}$ (or more precisely G) because $\overline{G}$ (or G) is the quantity that is concerned with the small interval $[0, dx_1]$. The time duration for which the molecule is in the small interval should have been taken into account in (23) to have a direct connection with $\overline{G}$. Hence, the counting with weight $dx_1/\xi_1^{(i)}$ should be taken

$$\overline{G}(\overline{\boldsymbol{\xi}}; 0, \boldsymbol{\xi})dx_1 \propto \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \frac{dx_1}{\xi_1^{(i)}} \delta(\boldsymbol{\xi} - \boldsymbol{\xi}^{(i)}). \tag{24}$$

Remember that the left-hand side is nothing else than $K(\boldsymbol{\xi}, \overline{\boldsymbol{\xi}})dx_1$. Thus, from (8) and $\int_{\xi_1>0} \mathcal{R}(\boldsymbol{\xi}, \overline{\boldsymbol{\xi}})d\boldsymbol{\xi} = 1$ (see footnote 1), we arrive at the relation (in the sense of weak formulation) that

$$|\xi_1|\overline{G} = \frac{|\overline{\xi}_1|}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta(\boldsymbol{\xi} - \boldsymbol{\xi}^{(i)}), \tag{25}$$

which establishes the way how to construct the scattering kernel from the Langevin system simulation.

Next, we proceed to the reflection intensity distribution. Introducing the polar coordinates (ξ, θ, φ) of the molecular velocity with positive x_1 being the polar direction, the normalized intensity distribution $I(\theta, \varphi)$ is expressed as

$$I(\theta, \varphi) = \frac{1}{I_{\text{in}}} \int_0^\infty f(\boldsymbol{\xi})\, \xi \cos\theta\, \xi^2 d\xi, \quad \cos\theta > 0, \tag{26a}$$

$$I_{\text{in}} = \int_{\overline{\xi}_1<0} |\overline{\xi}_1| f(\overline{\boldsymbol{\xi}}) d\overline{\boldsymbol{\xi}}, \tag{26b}$$

where f is the velocity distribution function of molecules on the wall. Substitution of (2) into (26a) gives

$$I(\theta, \varphi) = \frac{1}{I_{\text{in}}} \int_0^\infty \Big(\int_{\overline{\xi}_1<0} |\overline{\xi}_1| \mathcal{R}(\boldsymbol{\xi}, \overline{\boldsymbol{\xi}}) f(\overline{\boldsymbol{\xi}}) d\overline{\boldsymbol{\xi}} \Big) \xi^2 d\xi, \quad \xi_1 > 0. \tag{27}$$

In the case of the mono-collimated molecular beam, $f(\overline{\boldsymbol{\xi}}) = \delta(\overline{\boldsymbol{\xi}} - \boldsymbol{\xi}_{\text{in}})$, the intensity distribution is reduced to

$$\begin{aligned} I(\theta, \varphi) = \int_0^\infty \mathcal{R}(\boldsymbol{\xi}, \boldsymbol{\xi}_{\text{in}})\, \xi^2 d\xi &= \frac{1}{|\xi_{\text{in}1}|} \int_0^\infty \xi^2 \xi_1 \overline{G}\, d\xi \\ &= \frac{1}{|\xi_{\text{in}1}|} \int_0^\infty \xi^3 \cos\theta\, \overline{G}\, d\xi, \quad (\cos\theta > 0). \end{aligned} \tag{28}$$

Here it has been used that $I_{\rm in} = \int_{\overline{\xi}_1<0} |\overline{\xi}_1| \delta(\overline{\boldsymbol{\xi}} - \boldsymbol{\xi}_{\rm in}) d\overline{\boldsymbol{\xi}} = |\xi_{\rm in1}|$. Note that I follows the Lambert cosine law, if $\overline{G}$ is isotropic as the diffuse reflection case. The deviation from the cosine law implies the non-isotropy of $\overline{G}$. Now using (25), the intensity distribution can be reproduced by the following sample counting of the Langevin system simulation:

$$
\begin{aligned}
I \, \sin\theta \Delta\theta \Delta\varphi &= \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \int_0^\infty d\xi \, \xi^2 \int_\varphi^{\varphi+\Delta\varphi} d\varphi \int_\theta^{\theta+\Delta\theta} d\theta \, \sin\theta \, \frac{\delta(\xi - \xi^{(i)}) \delta(\theta - \theta^{(i)}) \delta(\varphi - \varphi^{(i)})}{\xi^2 \sin\theta} \\
&= \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \chi_{[\theta,\theta+\Delta\theta]}(\theta^{(i)}) \chi_{[\varphi,\varphi+\Delta\varphi]}(\varphi^{(i)}),
\end{aligned} \tag{29}
$$

where $(\theta^{(i)}, \varphi^{(i)})$ are the polar and the azimuth angle of $\boldsymbol{\xi}^{(i)}$ and $\chi_A(x)$ is the characteristic function: it takes unity when $x \in A$ and zero otherwise. The results of the above sample counting (29) are to be compared with the following intensity distribution $I_{\rm CL}$ for the CL model (5):

$$
\begin{aligned}
I_{\rm CL} =& \frac{1}{|\xi_{\rm in1}|} \int_0^\infty d\xi \, \frac{\xi^2 \xi_1}{2\pi (RT_w)^2} \frac{|\xi_{\rm in1}|}{\alpha_t (2-\alpha_t) \alpha_n} I_0\Big(\frac{\xi_1 \xi_{\rm in1}}{RT_w} \frac{\sqrt{1-\alpha_n}}{\alpha_n}\Big) \\
& \times \exp\Big(-\frac{\xi_1^2 + \xi_{\rm in1}^2 (1-\alpha_n)}{2RT_w \alpha_n}\Big) \exp\Big(-\frac{|\boldsymbol{\xi}_\parallel - \boldsymbol{\xi}_{\rm in\parallel}(1-\alpha_t)|^2}{2RT_w \alpha_t (2-\alpha_t)}\Big) \\
=& \frac{1}{2\pi (RT_w)^2} \frac{\cos\theta}{\alpha_t (2-\alpha_t) \alpha_n} \int_0^\infty d\xi \, \xi^3 I_0\Big(\frac{\xi \xi_{\rm in} \cos\theta \cos\theta_{\rm in}}{RT_w} \frac{\sqrt{1-\alpha_n}}{\alpha_n}\Big) \\
& \times \exp\Big(-\frac{\xi^2 \sin^2\theta + \xi_{\rm in}^2 \sin^2\theta_{\rm in} (1-\alpha_t)^2}{2RT_w \alpha_t (2-\alpha_t)} - \frac{\xi^2 \cos^2\theta + \xi_{\rm in}^2 \cos^2\theta_{\rm in} (1-\alpha_n)}{2RT_w \alpha_n} \\
& \qquad + \frac{\xi \xi_{\rm in} \sin\theta_{\rm in} \sin\theta \cos(\varphi - \varphi_{\rm in})(1-\alpha_t)}{RT_w \alpha_t (2-\alpha_t)}\Big).
\end{aligned} \tag{30}
$$

Figure 3 shows a couple of comparisons of the simulation results of (29) with (30). Good agreement is achieved, telling that the present construction of the Langevin system is appropriate. In the numerical simulations of the Langevin system, the Milstein scheme [10] has been adopted to achieve a sufficient numerical convergence with respect to the time-step size, see Sect. 5.3. Comparisons are made in the figure by using the relation between the parameters (α_n, α_t) and (ℓ_n, ℓ_t, d) in [4]:[3]

$$
\alpha_n = 1 - \exp\Big(-\frac{8d}{\ell_n}\Big), \quad \alpha_t = 1 - \exp\Big(-\frac{4d}{\ell_t}\Big). \tag{31}
$$

[3] There are misprints in (7.23) of [4], probably due to the inconsistent use of the notations ℓ_n and ℓ_t between [3] and [4].

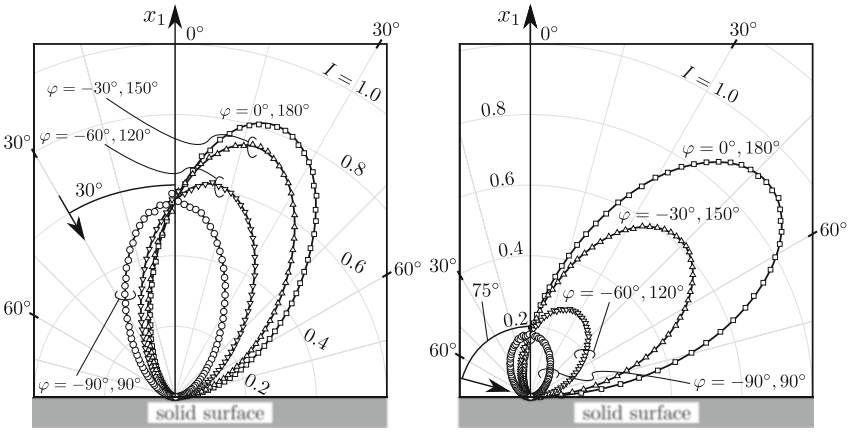

Fig. 3 In-plane and off-plane reflections of the mono-collimated molecular beam obtained by 10^8 particle simulations of the Langevin system (19): the case $\alpha_n = 0.3$ and $\alpha_t = 0.1$. The speed of the incident molecule is set as $|\boldsymbol{\xi}_{\rm in}|/\sqrt{2RT_w} = 0.522394$. (**a**) $\theta_{\rm in}^{\rm C} = 30°$, (**b**) $\theta_{\rm in}^{\rm C} = 75°$, where $\theta_{\rm in}^{\rm C} (\equiv \pi - \theta_{\rm in})$ is the angle of incidence of the velocity of the molecular beam $\boldsymbol{\xi}_{\rm in}$. φ is the azimuth angle measured clockwise from the direction of the projection of $\boldsymbol{\xi}_{\rm in}$ to the ξ_2-ξ_3 plane $(-\pi < \varphi \leq \pi)$. Symbols indicate the simulation results. Corresponding $I_{\rm CL}$'s in (30) are also shown for reference by solid lines. The arrow in each panel indicates the direction of $\boldsymbol{\xi}_{\rm in}$. Note the relation (31). The Milstein scheme with $\Delta t = 0.002\, d/\sqrt{2RT_w}$ and $p = 2$ has been used

5.3 *Some Aspects of the Numerical Method for the Langevin System*

Probably the simplest and widespread numerical algorithm for solving the Langevin equation is the Euler–Maruyama method, the scheme of which reads in the present case

$$x_i^{(n+1)} = x_i^{(n)} + \xi_i^{(n)} \Delta t, \quad (i = 1, 2, 3), \tag{32a}$$

$$\xi_1^{(n+1)} = \xi_1^{(n)} - \frac{2}{\ell_n}\{|\xi_1^{(n)}|\xi_1^{(n)} - \frac{\xi_1^{(n)}}{|\xi_1^{(n)}|} RT_w\}\Delta t + 2\sqrt{RT_w \frac{|\xi_1^{(n)}|}{\ell_n}\Delta t}\,\Delta B_1^{(n)}, \tag{32b}$$

$$\xi_2^{(n+1)} = \xi_2^{(n)} - \frac{2}{\ell_t}|\xi_1^{(n)}|\xi_2^{(n)}\Delta t + 2\sqrt{RT_w \frac{|\xi_1^{(n)}|}{\ell_t}\Delta t}\,\Delta B_2^{(n)}, \tag{32c}$$

$$\xi_3^{(n+1)} = \xi_3^{(n)} - \frac{2}{\ell_t}|\xi_1^{(n)}|\xi_3^{(n)}\Delta t + 2\sqrt{RT_w \frac{|\xi_1^{(n)}|}{\ell_t}\Delta t}\,\Delta B_3^{(n)}. \tag{32d}$$

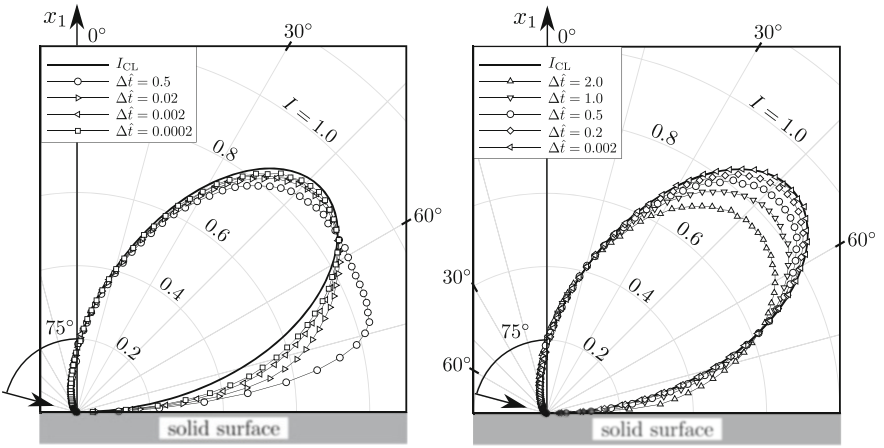

Fig. 4 Numerical convergence: the Euler–Maruyama scheme vs. the Milstein scheme. The results for the in-plane reflection of molecules ($\varphi = 0°,\ 180°$) for the same parameters as Fig. 3b, except for the time step. (**a**) The Euler–Maruyama scheme, (**b**) the Milstein scheme with $p = 2$. Here, $\Delta\hat{t} = (\sqrt{2RT_w}/d)\Delta t$. Symbols indicate the simulation results, while thick solid lines indicate I_{CL} in (30). Note the difference of $\Delta\hat{t}$ between (**a**) and (**b**). Common symbols are used for common values of $\Delta\hat{t}$

Here, $x_i^{(n)} = x_i(t_n)$, $\xi_i^{(n)} = \xi_i(t_n)$, $t_n = n\Delta t$ $(n = 0, 1, 2, \dots)$ is the discretised time, Δt is the size of time step, and $\Delta B_i^{(n)}$ $(i = 1, 2, 3)$ are mutually independent standard Gaussian random variables and are related to W_i as $\sqrt{\Delta t}\Delta B_i^{(n)} = W_i(t_{n+1}) - W_i(t_n)$. The Euler–Maruyama scheme is 1/2-order in the strong-order of convergence. In fortunate cases where S_{ij} is constant, the scheme becomes first-order [9, 10], which does not apply in the present case because of (18c). Indeed, the implementation of the Euler–Maruyama scheme shows a very slow convergence with respect to the size of time discretisation, see Fig. 4a. The difficulty of the slow convergence is, however, resolved dramatically by switching to the Milstein scheme, which is known to be first-order in the strong-order of convergence [10], see Fig. 4b. In the present case, as the noise is not commutative[4] for ξ_2 and ξ_3, the scheme

[4] The noise is said to be commutative, if $\Theta_{\alpha i}$ in (13a) satisfies the condition $\Theta_{\beta i}(\partial\Theta_{\alpha j}/\partial y_\beta) = \Theta_{\beta j}(\partial\Theta_{\alpha i}/\partial y_\beta)$.

becomes rather complicated as[5]

$$\xi_1^{(n+1)} = \xi_1^{(n)} - \frac{2}{\ell_n}(|\xi_1^{(n)}|\xi_1^{(n)} - \frac{\xi_1^{(n)}}{|\xi_1^{(n)}|}RT_w)\Delta t + 2\sqrt{RT_w\frac{|\xi_1^{(n)}|}{\ell_n}\Delta t}\,\Delta B_1^{(n)} + \frac{\xi_1^{(n)}}{|\xi_1^{(n)}|}\frac{2RT_w}{\ell_n}I_{11}^{(n)}, \tag{33a}$$

$$\xi_2^{(n+1)} = \xi_2^{(n)} - \frac{2}{\ell_t}|\xi_1^{(n)}|\xi_2^{(n)}\Delta t + 2\sqrt{RT_w\frac{|\xi_1^{(n)}|}{\ell_t}\Delta t}\,\Delta B_2^{(n)} + \frac{\xi_1^{(n)}}{|\xi_1^{(n)}|}\frac{2RT_w}{\sqrt{\ell_t\ell_n}}I_{12}^{(n)p}, \tag{33b}$$

$$\xi_3^{(n+1)} = \xi_3^{(n)} - \frac{2}{\ell_t}|\xi_1^{(n)}|\xi_3^{(n)}\Delta t + 2\sqrt{RT_w\frac{|\xi_1^{(n)}|}{\ell_t}\Delta t}\,\Delta B_3^{(n)} + \frac{\xi_1^{(n)}}{|\xi_1^{(n)}|}\frac{2RT_w}{\sqrt{\ell_t\ell_n}}I_{13}^{(n)p}, \tag{33c}$$

where

$$I_{11}^{(n)} = \frac{\Delta t}{2}\{(\Delta B_1^{(n)})^2 - 1\}, \tag{34a}$$

$$I_{12}^{(n)p} = \frac{\Delta t}{2}\Delta B_1^{(n)}\Delta B_2^{(n)} + \frac{\Delta t}{2\pi}\sum_{q=1}^{p}\frac{1}{q}\{\zeta_{2q}(\sqrt{2}\Delta B_1^{(n)} - \eta_{1q}) - \zeta_{1q}(\sqrt{2}\Delta B_2^{(n)} - \eta_{2q})\} + \Delta t\sqrt{\rho^{(p)}}(\mu_2^{(p)}\Delta B_1^{(n)} - \mu_1^{(p)}\Delta B_2^{(n)}), \tag{34b}$$

$$I_{13}^{(n)p} = \frac{\Delta t}{2}\Delta B_1^{(n)}\Delta B_3^{(n)} + \frac{\Delta t}{2\pi}\sum_{q=1}^{p}\frac{1}{q}\{\zeta_{3q}(\sqrt{2}\Delta B_1^{(n)} - \eta_{1q}) - \zeta_{1q}(\sqrt{2}\Delta B_3^{(n)} - \eta_{3q})\} + \Delta t\sqrt{\rho^{(p)}}(\mu_3^{(p)}\Delta B_1^{(n)} - \mu_1^{(p)}\Delta B_3^{(n)}), \tag{34c}$$

$$\rho^{(p)} = \frac{1}{12} - \frac{1}{2\pi^2}\sum_{q=1}^{p}\frac{1}{q^2}. \tag{34d}$$

[5] See [10, pp. 346–347] for the details. Unfortunately, there are misprints in the corresponding formula in Sec. 6.4.3 of [9], though the latter reference is an excellent textbook. Incidentally, in [9], the Milstein scheme for the non-commutative noise is referred to as Kloeden and Platen's approximation.

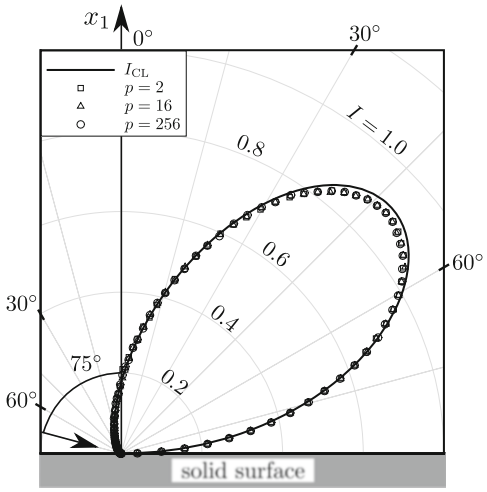

Fig. 5 Influence of the truncation number p in the Milstein scheme: the in-plane reflection of molecules ($\varphi = 0°,\ 180°$) for the same parameters as Fig. 3b, though a coarser time step $\Delta t = 0.2\, d/\sqrt{2RT_w}$ is used here. Symbols indicate the simulation results, while the solid line indicates I_{CL} in (30)

Here, $\mu_i^{(p)}$, η_{iq}, and ζ_{iq} are mutually independent standard Gaussian random variables,[6] and p is the truncation number of the infinite series, which should be chosen so that $p > C/\Delta t$ for a positive constant C. As is clear from (33) and (34), the Milstein scheme requires the generation of $6(p+1)$ standard Gaussian variables at each time step, which is $2(p+1)$-times as many as in the Euler–Maruyama scheme and looks a serious drawback at a glance. Fortunately, however, numerical experiments show that the convergence rate with respect to p is excellent and that the setting $p = 2$ is found to be good enough, see Fig. 5.

5.4 *A Further Observation: Some Features of Time Delay in Exit*

We have so far focused on the way to construct the scattering kernel and/or the reflection intensity distribution without time delay. The scattering model without time delay supposes that the time duration of interaction with the wall is so short that the process may be considered to be instantaneous in the time scale of our

[6] Originally, $\mu_i^{(p)}$ is defined as

$$\mu_i^{(p)} = \frac{1}{\sqrt{\rho^{(p)}\Delta t}} \sum_{q=p+1}^{\infty} \frac{1}{\pi q} \sqrt{\frac{\Delta t}{2}} \zeta_{iq}.$$

According to [10], however, $\mu_i^{(p)}$ thus defined becomes a standard Gaussian random variable. This property is very useful from the actual computational point of view.

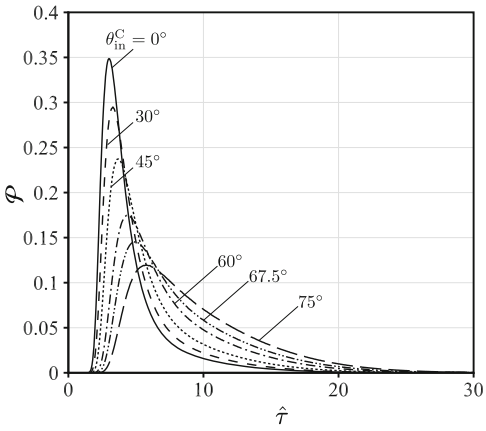

Fig. 6 Distribution of reflected molecules with respect to the exit time τ for various angle of incidence $\theta_{\rm in}^{\rm C}$. The parameters are the same as Fig. 3, except for a part of values of $\theta_{\rm in}^{\rm C}$. Here $\mathcal{P}(\hat{\tau}) = \sum_{i=1}^{\mathcal{N}} \chi_{[\hat{\tau},\hat{\tau}+\Delta\hat{\tau}]}/(\mathcal{N}\Delta\hat{\tau})$ with $\hat{\tau} = \tau(2RT_w)^{1/2}/d$, $\Delta\hat{\tau} = 0.05$, and $\mathcal{N} = 10^8$

interest. However, if we change the sample counting to that at a specified exit time τ, a closer observation of the dynamics is possible. It would also give a hint toward the construction of the scattering kernel with a time-delay effect. Here, we present a few examples of such sample counting.

Figure 6 shows the distribution of the exit time of samples in the same simulation as Fig. 3. As is observed, the larger the angle of incidence is, the longer the time duration of interaction is. We have also observed that there are no test particles that experience the reversal of motion in the normal direction except for the reflection at the potential barrier [see also Fig. 2c]. Hence, they commonly travel $2d$ in depth. These numerical observations suggest that in the CL model molecules of tangential incidence have more chance to remain at a low speed in the normal direction, and thus to need a longer time duration before leaving.

The common travelling distance $2d$ in depth implies that $\int |\xi_1| dt = 2d$ holds, so that the *drift*-part deceleration yields $\boldsymbol{\xi}_{\parallel} = \boldsymbol{\xi}_{{\rm in}\parallel} \exp(-4d/\ell_t) = \boldsymbol{\xi}_{{\rm in}\parallel}(1 - \alpha_t)$ at the exit time; see (21). This coincides with the central velocity of the Gaussian in tangential directions in the kernel of CL model; see (5). Finally, an example of the in-plane reflection intensity distribution in a specified interval of exit time is shown in Fig. 7. The distribution inclines more to the tangential direction for the molecules of larger exit time.

6 Conclusion

In the present paper, we have revisited the Cercignani–Lampis model for the gas–surface interaction, along the lines of Cercignani in [4]. Starting from his time-independent Fokker–Planck system, we have introduced its simple and natural time-dependent extension and have identified the corresponding Langevin system.

In the Langevin system, there are two types of interactions with the wall. One is a stochastic thermal agitation, which we call the *diffusion* part, and the size of

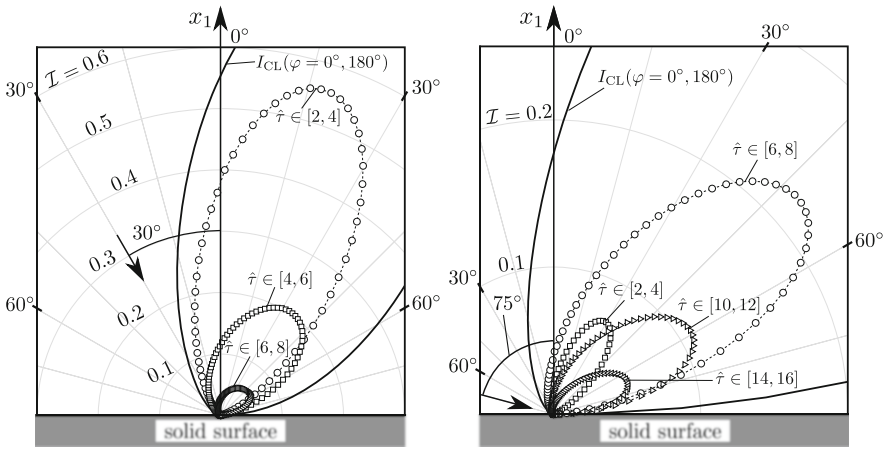

Fig. 7 Time-dependent in-plane reflection intensity distribution $\mathcal{I}$: the same case as Fig. 3. (**a**) $\theta_{\mathrm{in}}^{\mathrm{C}} = 30°$, (**b**) $\theta_{\mathrm{in}}^{\mathrm{C}} = 75°$. $\mathcal{I}$ is computed by the sample counting $\mathcal{I}(\hat{\tau} \in A, \theta, \varphi)\Delta\hat{\tau}_A \sin\theta\Delta\theta\Delta\varphi = (1/\mathcal{N})\sum_{i=1}^{\mathcal{N}} \chi_A(\hat{\tau}^{(i)})\chi_{[\theta,\theta+\Delta\theta]}(\theta^{(i)})\chi_{[\varphi,\varphi+\Delta\varphi]}(\varphi^{(i)})$, where $\mathcal{N} = 10^{10}$, $\Delta\hat{\tau}_A$ is the size of time interval A, and $\hat{\tau}^{(i)}$ is the dimensionless time of exit of the i-th sample

agitation depends on the random variable ξ_1. The other is what we call the *drift* part, which leads $|\xi_1|$ toward the speed of kinetic energy given by the equipartition law. In the tangential directions it decelerates the molecule by the viscous-like drag proportional to the moment transferred by that molecule.

The appropriate sample counting of the Langevin system simulation has also been discussed, and the capability of reproducing the scattering kernel and/or the reflection intensity distribution have been numerically demonstrated. It has also been remarked that the present stochastic noise causes the application of the Euler–Maruyama method to be inefficient and requires the Milstein method.

Finally, we stress that, from a numerical point of view, the Langevin system is advantageous to the FP system in that the incident mono-collimated molecular beam is easily handled to allow a close observation as in Sect. 5.4. Indeed, the sampling there gives a way toward a construction of the time-delay effect in the scattering model. Such an extension has a potential importance for such as an evacuation-speed prediction in vacuum technologies. Modifications of the dynamics by coupling with strong scatterings suggested in [4] will also be possible in the same numerical framework, if desired. Unlike the concise expression of the original CL model, the extensions above suggested might require a data fitting to construct a ready-to-use kinetic boundary condition. Nevertheless, a flexibility of the present simple approach is an advantage of modern computational facilities over the tools/techniques available in 1970s.

Appendix

The numerical simulation of the Langevin system is performed particle by particle. The process of computations for each test particle, say k-th particle ($k = 1, 2, \dots$), is as follows.

Suppose that the size of time step Δt is given. Set the initial position $\boldsymbol{x}^{(0)}$ and velocity $\boldsymbol{\xi}^{(0)}$ of the test particle as $\boldsymbol{x}^{(0)} = \boldsymbol{0}$ and $\boldsymbol{\xi}^{(0)} = \boldsymbol{\xi}_{\rm in}$. Let $\boldsymbol{x}^{(n)}$ and $\boldsymbol{\xi}^{(n)}$ be known, where $x_1^{(n)} \in [-d, 0]$ and $n = 0, 1, 2, \dots$

Step 1. Compute the particle position $\boldsymbol{x}^{(n+1)}$ at time $t^{(n+1)}$ by (32a). If $x_1^{(n+1)} < -d$, discard it and reset $x_1^{(n+1)}$ as $x_1^{(n+1)} = -2d - x_1^{(n)} - \xi_1^{(n)}\Delta t$. This is due to the specular reflection at the potential barrier.

Step 2. Compute the particle velocity $\boldsymbol{\xi}^{(n+1)}$ at time $t^{(n+1)}$ by (33) with (34).

2a. If $x_1^{(n+1)} < -d$ occurs in Step 1, change the sign of $\xi_1^{(n+1)}$; then go to 2c.

2b. If $x_1^{(n+1)} > 0$, put $\Delta t^\sharp = \Delta t - x_1^{(n+1)}/\xi_1^{(n)}$, and compute $\boldsymbol{x}^\sharp$ and $\boldsymbol{\xi}^\sharp$ by (32a) and (33) using $\Delta t^\sharp$ in place of Δt. If $\xi_1^\sharp \geq 0$, which is the case usually, record $n\Delta t + \Delta t^\sharp$, $\boldsymbol{x}^\sharp$, and $\boldsymbol{\xi}^\sharp$ as the exit instance, position, and velocity of the k-th particle, and stop the computation. In case $\xi_1^\sharp < 0$ happens to occur, continue the computation to reset $\boldsymbol{x}^{(n+1)}$ and $\boldsymbol{\xi}^{(n+1)}$ by (32a) and (33) using $(\Delta t - \Delta t^\sharp)$, $\boldsymbol{x}^\sharp$, and $\boldsymbol{\xi}^\sharp$ in place of Δt, $\boldsymbol{x}^{(n)}$, and $\boldsymbol{\xi}^{(n)}$; then go to 2c.

2c. If $x_1^{(n+1)} \leq 0$, go back to Step 1 and shift n to $n + 1$.

Repeat the above steps until an enough number of samples have been collected. In the actual computations, the Mersenne Twister pseudo-random number generator [14] has been used in generating the standard Gaussian variables.

Acknowledgments The present work has been supported in part by JSPS KAKENHI Grant No. 17K18840 and by the Japan-France Integrated Action Program (SAKURA) Grant No. JPJSBP120193219.

References

1. Aoki, K., Charrier, P., Degond, P.: A hierarchy of models related to nanoflows and surface diffusion. Kinet. Relat. Models **4**, 53–85 (2011). https://doi.org/10.3934/krm.2011.4.53
2. Brull, S., Charrier, P., Mieussens, L.: Nanoscale roughness effect on Maxwell-like boundary conditions for the Boltzmann equation. Phys. Fluids **28**, 082004 (2016). https://doi.org/10.1063/1.4960024
3. Cercignani, C.: Scattering kernels for gas–surface interactions. Trans. Theory Stat. Phys. **2**, 27–53 (1972)
4. Cercignani, C.: The Boltzmann Equation and Its Applications. Springer, New York (1988). Chap. III

5. Cercignani, C., Lampis, M.: Kinetic models for gas–surface interactions. Trans. Theory Stat. Phys. **1**, 101–114 (1971). https://doi.org/10.1080/00411457108231440
6. Cowling, T.G.: On the Cercignani–Lampis formula for gas–surface interactions. J. Phys. D: Appl. Phys. **7**, 781–785 (1974)
7. Gardiner, C.W.: Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, 2nd edn. Springer, Berlin (1985). Sec. 4.3
8. Jackson, E.A.: Equilibrium Statistical Mechanics. Dover edition, New York (2000). Sec. 4.6
9. Jacob, K.: Stochastic Processes for Physicists, Understanding Noisy Systems. Cambridge University, Cambridge (2010). Secs. 3.6–3.8
10. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1992)
11. Kogan, M.N.: Rarefied Gas Dynamics. Plenum, New York (1969). Sec. 2.10
12. Kuščer, I., Mozina, J., Krizamic, F.: The Knudsen model of thermal accomodation. In: Dini, D., et al. (eds.) Rarefied Gas Dynamics, vol. I, pp. 97–108. Editrice Tecnico Scientifica, Pisa (1971)
13. Lord, R.G.: Some extensions to the Cercignani–Lampis gas–surface scattering kernel. Phys. Fluids A **3**, 706–710 (1991). https://doi.org/10.1063/1.858076
14. Matsumoto, M., Nishimura, T.: Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. **8**, 3–30 (1998). https://doi.org/10.1145/272991.272995
15. Schaaf, S.A.: Mechanics of rarefied gases. In: Flügge, S. (ed.) Handbuch der Physik, band VIII/2, pp.591–624. Springer, Berlin (1963)
16. Sone, Y.: Molecular Gas Dynamics, Birkhäuser, Boston (2007). Supplement is available from http://hdl.handle.net/2433/66098
17. Williams, M.M.R.: A phenomenological study of gas–surface interactions. J. Phys. D: Appl. Phys. **4**, 1315–1319 (1971)
18. Yamanishi, N., Matsumoto, Y., Shobatake, K.: Multistage gas–surface interaction model for the direct simulation Monte Carlo Method. Phys. Fluids **11**, 3540–3552 (1999). https://doi.org/10.1063/1.870211

On the Accuracy of Gyrokinetic Equations in Fusion Applications

Edoardo Zoni and Stefan Possanner

Abstract This article concerns the asymptotic derivation of equations of motion for gyro-centers in strongly magnetized fusion plasmas. In particular, we focus on the role of the electron–ion mass ratio in the gyrokinetic coordinate transformation. We discuss the ordering assumptions for the ITER and ASDEX Upgrade Tokamaks in detail. A system of generating differential equations is derived and solved by asymptotic expansion to second order for ions and to fourth order for electrons. It is shown that the higher-order expansion for electrons is necessary for achieving first-order accuracy in the gyro-center equations of motion, which is usually desired for gyrokinetic simulations.

1 Gyrokinetic Theory in a Nutshell

The physics application considered in this article is magnetic confinement fusion. In particular, we are interested in the mathematical modeling of high-temperature plasmas created inside experimental fusion reactors, such as, for example, the Tokamaks ASDEX Upgrade [17] and ITER [22]. Such plasmas are macroscopic physical systems composed of a very large number of microscopic particles, of the

E. Zoni (✉)
Max Planck Institute for Plasma Physics, Garching, Germany

Technical University of Munich, Garching, Germany

Lawrence Berkeley National Laboratory, Berkeley, CA, USA
e-mail: ezoni@lbl.gov

S. Possanner
Max Planck Institute for Plasma Physics, Garching, Germany

Technical University of Munich, Garching, Germany
e-mail: stefan.possanner@ipp.mpg.de

F. Salvarani (eds.), *Recent Advances in Kinetic Equations and Applications*,
Springer INdAM Series 48, https://doi.org/10.1007/978-3-030-82946-9_16

order of 10^{20} particles per cubic meter in the case of the two Tokamaks mentioned above. Moreover, we are mostly interested in the study of non-equilibrium physical phenomena, such as plasma turbulence. Based on these observations, we apply the principles of the kinetic theory of gases and derive suitable kinetic models, such as the Vlasov–Maxwell model described in Sect. 2.

However, such models are defined on a six-dimensional phase space (three dimensions for the position of the plasma particles and three dimensions for their velocity) and exhibit a variety of space and time scales that vary within a large range. When the model equations need to be solved through computer simulations, the high-dimensionality of the phase space and the multi-scale nature of the problem become significant computational limitations, resulting in long run times and heavy memory footprints. Therefore, it becomes useful to develop reduced models that retain a physically meaningful description of the system and decrease, at the same time, the computational cost of the model. Gyrokinetic theory [4] is an example of such reduced kinetic models.

The fundamental idea of gyrokinetic theory is to separate the fast time scale associated with the motion of gyration of the charged plasma particles around the field lines of the confining magnetic field from the slower time scales of the problem. This is achieved by choosing new phase-space coordinates, different than the physical positions and velocities of the particles, defined in such a way that the fast motion of gyration is decoupled from the particle dynamics. The resulting dynamical equations govern the evolution of the so-called "gyro-centers". They are defined on a five-dimensional phase space and enable more efficient computer simulations of plasma turbulence in fusion reactors. The basics of this method have been outlined many decades ago [2, 11] and the understanding of gyrokinetics and its application have come a long way since then [3–5, 8, 12, 18, 19, 21, 23]. Indeed, many of the large production computer codes for fusion research are based on gyrokinetics [1, 7, 9, 10].

In this work we aim to investigate the electron gyro-center equations of motion on an equal footing with the ion equations. Electron gyrokinetic modeling has not received a lot of attention on its own, meaning that usually the ion gyro-center equations are used also for electrons, with the adequate adjustment of physical parameters such as mass and charge. This can be dangerous when judging the validity of such a model, or when trying to assess its accuracy. Here, we present a rigorous two-species gyro-center reduction that takes into account the mass ratio between electrons and ions.

2 The Vlasov–Maxwell Model

We consider a non-collisional plasma made of ions and electrons described in terms of particle distribution functions $f_\mathrm{s} : \mathbb{R}^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \ni (t, \mathbf{x}, \mathbf{v}) \mapsto f_\mathrm{s}(t, \mathbf{x}, \mathbf{v}) \in \mathbb{R}^+$ that obey the non-collisional Vlasov equation[1]

$$\frac{\partial f_\mathrm{s}}{\partial t} + \mathbf{v} \cdot \boldsymbol{\nabla} f_\mathrm{s} + \frac{q_\mathrm{s}}{m_\mathrm{s}} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial f_\mathrm{s}}{\partial \mathbf{v}} = 0 \,, \tag{1}$$

where $q_\mathrm{s} \in \mathbb{R}$ and $m_\mathrm{s} \in \mathbb{R}^+$ denote the particle charge and mass, respectively. The electric and magnetic fields $\mathbf{E} : \mathbb{R}^+ \times \mathbb{R}^3 \ni (t, \mathbf{x}) \mapsto \mathbf{E}(t, \mathbf{x}) \in \mathbb{R}^3$ and $\mathbf{B} : \mathbb{R}^+ \times \mathbb{R}^3 \ni (t, \mathbf{x}) \mapsto \mathbf{B}(t, \mathbf{x}) \in \mathbb{R}^3$ satisfy Maxwell's equations

$$\boldsymbol{\nabla} \cdot \mathbf{E} = \frac{\varrho}{\varepsilon_0} \tag{2a}$$

$$\boldsymbol{\nabla} \cdot \mathbf{B} = 0 \tag{2b}$$

$$\boldsymbol{\nabla} \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{2c}$$

$$\boldsymbol{\nabla} \times \mathbf{B} = \mu_0 \mathbf{J} + \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \tag{2d}$$

where ε_0 and μ_0 denote the vacuum electric permittivity and the vacuum magnetic permeability, respectively. The sources $\varrho : \mathbb{R}^+ \times \mathbb{R}^3 \ni (t, \mathbf{x}) \mapsto \varrho(t, \mathbf{x}) \in \mathbb{R}$ and $\mathbf{J} : \mathbb{R}^+ \times \mathbb{R}^3 \ni (t, \mathbf{x}) \mapsto \mathbf{J}(t, \mathbf{x}) \in \mathbb{R}^3$ are expressed in terms of the distribution functions as

$$\varrho = \sum_\mathrm{s} \int \mathrm{d}^3\mathbf{v}\, q_\mathrm{s}\, f_\mathrm{s} \,, \quad \mathbf{J} = \sum_\mathrm{s} \int \mathrm{d}^3\mathbf{v}\, q_\mathrm{s}\, \mathbf{v}\, f_\mathrm{s} \,. \tag{3}$$

The derivation of the Vlasov–Maxwell system (1)–(3) from an action principle was recognized first by Low [16]. Denoting by $\phi : \mathbb{R}^+ \times \mathbb{R}^3 \ni (t, \mathbf{x}) \mapsto \phi(t, \mathbf{x}) \in \mathbb{R}$ and $\mathbf{A} : \mathbb{R}^+ \times \mathbb{R}^3 \ni (t, \mathbf{x}) \mapsto \mathbf{A}(t, \mathbf{x}) \in \mathbb{R}^3$ the electric scalar potential and the magnetic vector potential associated with the electric and magnetic fields via $\mathbf{E} = -\boldsymbol{\nabla}\phi - \partial\mathbf{A}/\partial t$ and $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$, Low's action principle reads

$$\delta \int_{t_0}^{t_1} \mathrm{d}t\, (L_\mathrm{EM} + L_\mathrm{P}) = 0 \,, \tag{4}$$

[1] All equations in this article are written in SI units.

where δ denotes the Fréchet derivative and the Lagrangian is the sum of the electromagnetic free-field Lagrangian

$$L_{\mathrm{EM}}(\phi, \mathbf{A}) = \frac{\varepsilon_0}{2} \int \mathrm{d}^3\mathbf{x} \left| \nabla\phi + \frac{\partial \mathbf{A}}{\partial t} \right|^2 - \frac{1}{2\mu_0} \int \mathrm{d}^3\mathbf{x} \, |\nabla \times \mathbf{A}|^2 \,, \tag{5}$$

and the particle Lagrangian

$$L_{\mathrm{P}}(\mathbf{x}(t), \mathbf{v}(t)) = \sum_{\mathrm{s}} \int \mathrm{d}^3\mathbf{x}_0 \, \mathrm{d}^3\mathbf{v}_0 \, f_{\mathrm{s}0} \, L_{\mathrm{s}}(\mathbf{x}(t), \mathbf{v}(t)) \,. \tag{6}$$

Here, $f_{\mathrm{s}0} := f_{\mathrm{s}}(t_0, \mathbf{x}_0, \mathbf{v}_0)$ and L_{s} denotes the single-particle Lagrangian for the particle species s, which in the phase-space coordinates $(\mathbf{x}, \mathbf{v})$ reads

$$L_{\mathrm{s}}(\mathbf{x}(t), \mathbf{v}(t)) = (m_{\mathrm{s}}\mathbf{v}(t) + q_{\mathrm{s}}\mathbf{A}) \cdot \dot{\mathbf{x}}(t) - H_{\mathrm{s}} \,, \quad H_s = \frac{m_{\mathrm{s}}}{2} |\mathbf{v}(t)|^2 + q_{\mathrm{s}}\phi \,, \tag{7}$$

where H_{s} denotes the particle Hamiltonian and the potentials are evaluated at $(t, \mathbf{x}(t))$. We remark that L_{s} depends implicitly on the potentials ϕ and $\mathbf{A}$ and describes the self-consistent interaction between the plasma particles and the electromagnetic fields.

The variational principle (4) leads to: the characteristics of the Vlasov equation (1), by computing variations of L_{s} with respect to single-particle trajectories $(\mathbf{x}(t), \mathbf{v}(t))$; Coulomb's law (2a), by computing variations of L_{s} with respect to ϕ; Ampère–Maxwell's law (2d), by computing variations of L_{s} with respect to $\mathbf{A}$. We remark that only the non-homogeneous Maxwell's equations, featuring source terms coupling to the plasma particles, can be derived from the variational principle. The homogeneous Maxwell's equations (Faraday's law (2c) and magnetic Gauss law (2b)) follow from the definition of $\mathbf{E}$ and $\mathbf{B}$ through ϕ and $\mathbf{A}$. With appropriate initial and boundary conditions, this results in a well-posed system for $(f_{\mathrm{s}}, \mathbf{E}, \mathbf{B})$, which describes the self-consistent interaction between the plasma particles and the electromagnetic fields.

3 Normalization and Ordering

The formulation of the Vlasov–Maxwell system as a perturbation problem requires the non-dimensionalization of the physical equations, also referred to as scaling or *normalization*. The process of quantifying the size of the non-dimensional coefficients appearing in the normalized equations in terms of a single small perturbation parameter $\varepsilon \ll 1$ is referred to as *ordering*. Different orderings lead to different perturbation theories and to reduced models with different physical content. In other words, an ordering is the mathematical expression of a specific physical scenario. Two such scenarios for magnetic confinement fusion experiments

Table 1 Physical parameters for the Tokamak ASDEX Upgrade [17]

		Ions	Electrons
Major radius	$R_0 = 1.6\,\mathrm{m}$		
Minor radius	$a = 0.8\,\mathrm{m}$		
Toroidal magnetic field	$B_\mathrm{T} = 3.9\,\mathrm{T}$		
Average particle density	$\langle n_\mathrm{s}\rangle$	$2.0 \times 10^{20}/\mathrm{m}^3$	$2.0 \times 10^{20}/\mathrm{m}^3$
Average thermal energy	$k_\mathrm{B}\langle T_\mathrm{s}\rangle$	8.7 keV	8.7 keV
Cyclotron frequency	$\omega_\mathrm{cs} = q_\mathrm{s} B_\mathrm{T}/m_\mathrm{s}$	1.9×10^8 Hz	6.9×10^{11} Hz
Thermal velocity	$v_\mathrm{s} = (k_\mathrm{B}\langle T_\mathrm{s}\rangle/m_\mathrm{s})^{1/2}$	6.4×10^5 m/s	3.9×10^7 m/s
Thermal frequency	$\omega_\mathrm{s} = v_\mathrm{s}/a$	8.0×10^5 Hz	4.9×10^7 Hz
Larmor radius	$\rho_\mathrm{s} = v_\mathrm{s}/\omega_\mathrm{cs}$	3.4×10^{-3} m	5.7×10^{-5} m
Debye length	$\lambda_\mathrm{s} = (\varepsilon_0 k_\mathrm{B}\langle T_\mathrm{s}\rangle/q_\mathrm{s}^2\langle n_\mathrm{s}\rangle)^{1/2}$	4.9×10^{-5} m	4.9×10^{-5} m

Table 2 Physical parameters for the Tokamak ITER [22]

		Ions	Electrons
Major radius	$R_0 = 6.2\,\mathrm{m}$		
Minor radius	$a = 2.0\,\mathrm{m}$		
Toroidal magnetic field	$B_\mathrm{T} = 5.3\,\mathrm{T}$		
Average particle density	$\langle n_\mathrm{s}\rangle$	$1.0 \times 10^{20}/\mathrm{m}^3$	$1.0 \times 10^{20}/\mathrm{m}^3$
Average thermal energy	$k_\mathrm{B}\langle T_\mathrm{s}\rangle$	8.0 keV	8.8 keV
Cyclotron frequency	$\omega_\mathrm{cs} = q_\mathrm{s} B_\mathrm{T}/m_\mathrm{s}$	2.5×10^8 Hz	9.3×10^{11} Hz
Thermal velocity	$v_\mathrm{s} = (k_\mathrm{B}\langle T_\mathrm{s}\rangle/m_\mathrm{s})^{1/2}$	6.2×10^5 m/s	3.9×10^7 m/s
Thermal frequency	$\omega_\mathrm{s} = v_\mathrm{s}/a$	3.1×10^5 Hz	2.0×10^7 Hz
Larmor radius	$\rho_\mathrm{s} = v_\mathrm{s}/\omega_\mathrm{cs}$	2.4×10^{-3} m	4.2×10^{-5} m
Debye length	$\lambda_\mathrm{s} = (\varepsilon_0 k_\mathrm{B}\langle T_\mathrm{s}\rangle/q_\mathrm{s}^2\langle n_\mathrm{s}\rangle)^{1/2}$	6.6×10^{-5} m	7.0×10^{-5} m

are shown in Tables 1 and 2. In this section we present a detailed normalization and ordering analysis, where we consider a plasma consisting of deuterium ions and electrons in a realistic physical scenario relevant for existing and future fusion experimental reactors, such as, for example, the Tokamaks ASDEX Upgrade [17] and ITER [22]. These two scenarios do not differ significantly for modeling purposes. However, from a physics point of view, the ITER scenario is significantly larger in plasma volume and certainly one step closer towards self-sustained fusion energy gains.

In order to write the Vlasov–Maxwell model in non-dimensional form, we introduce reference scales for times, frequencies, lengths, and velocities,

$$t = t'/\omega_\mathrm{i}\,, \quad \omega = \omega_\mathrm{i}\,\omega'\,, \quad \mathbf{x} = a\,\mathbf{x}'\,, \quad \mathbf{v}_\mathrm{s} = v_\mathrm{s}\,\mathbf{v}'\,, \tag{8}$$

where primed quantities are non-dimensional. Here, we choose as characteristic length and time scales of observation the minor radius a of the fusion reactor and the inverse of the ion thermal frequency $\omega_\mathrm{i} = v_\mathrm{i}/a$, that is, the time required for an ion to travel the distance a. The reference velocity v_s denotes the average thermal

velocity per species, defined as $v_\mathrm{s} = (k_\mathrm{B}\langle T_\mathrm{s}\rangle/m_\mathrm{s})^{1/2}$, where $k_\mathrm{B}\langle T_\mathrm{s}\rangle$ represents the average thermal energy per species over the plasma volume.

The ion thermal frequency ω_i is close to the characteristic frequency of micro-turbulence observed in Tokamaks [12, 13, 24]. Since gyrokinetics is ultimately the theory of low-frequency dynamics in strongly-magnetized plasmas, the perturbation parameter ε is typically defined around the ratio between the characteristic ion turbulence frequency and the ion cyclotron frequency $\omega_\mathrm{ci} := q_\mathrm{i} B_\mathrm{T}/m_\mathrm{i}$:

$$\varepsilon := 10^{-3} \approx \frac{\omega_\mathrm{i}}{\omega_\mathrm{ci}}\,. \tag{9}$$

We remark that the numerical values reported here and in the following are computed by taking the mean of the physical parameters given in Tables 1 and 2 for each species.

It is also common to express the magnetic field as the sum of a static background $\mathbf{B}_0$ and dynamic fluctuations $\mathbf{B}_1$. The amplitudes of the corresponding magnetic vector potentials satisfy $A_0/a \sim B_0 \approx B_\mathrm{T}$ and $k_\perp A_1 \sim B_1$, where $k_\perp$ denotes the characteristic wave number of the turbulent fluctuations on the planes perpendicular to the magnetic field, and $A_0 := |\mathbf{A}_0|$ and so forth for the other vectors. With the symbol $\sim$ we mean, for example, $A_0 = O(aB_0)$ as the value of a is changed in the physical setup. Similarly, the amplitude ϕ of the electric potential satisfies $k_\perp\phi \sim E$. Following the standard gyrokinetic ordering [4] we assume $k_\perp\,\rho_\mathrm{i} \sim 1$, where ρ_i denotes the ion Larmor radius.

The single-particle Lagrangians (7) are then normalized with respect to the thermal energy, $L_\mathrm{s} = m_\mathrm{s} v_\mathrm{s}^2\, L'_\mathrm{s}$, which leads to

$$\begin{aligned} L'_\mathrm{s} = &\left(\frac{a\,\omega_\mathrm{i}}{v_\mathrm{s}}\mathbf{v}' + \frac{a\,\omega_\mathrm{i}}{v_\mathrm{s}}\frac{q_\mathrm{s}B_\mathrm{T}}{m_\mathrm{s}}\frac{a}{v_\mathrm{s}}\mathbf{A}'_0 + \frac{a\,\omega_\mathrm{i}}{v_\mathrm{s}}\frac{q_\mathrm{s}B_\mathrm{T}}{m_\mathrm{s}}\frac{a}{v_\mathrm{s}}\frac{B_1}{B_\mathrm{T}}\frac{1}{a\,k_\perp}\mathbf{A}'_1\right)\cdot\dot{\mathbf{x}}' \\ &- \frac{|\mathbf{v}'|^2}{2} - \frac{q_\mathrm{s}\phi}{m_\mathrm{s}v_\mathrm{s}^2}\phi'\,. \end{aligned} \tag{10}$$

The non-dimensional coefficients in (10) can be computed by using the physical parameters given in Tables 1 and 2, combined with the observation that measurements in Tokamaks have shown that fluctuation levels in turbulent plasmas satisfy [4, 13, 24]

$$\frac{B_1}{B_\mathrm{T}} \approx \frac{E}{B_\mathrm{T}v_\mathrm{i}} \approx 10^{-3}\,, \tag{11}$$

which means that fluctuations are small compared to the corresponding background quantities and that the $\mathbf{E}\times\mathbf{B}$ velocity is small compared to the ion thermal velocity. The values of the non-dimensional coefficients in (10) for ions and electrons are summarized in Table 3. The ordering of these coefficients in terms of powers of ε is done by logarithmic comparison: for example, the order p of the coefficient $a\,\omega_\mathrm{i}/v_\mathrm{s}$ is determined by minimizing $|\log_{10}(a\,\omega_\mathrm{i}/v_\mathrm{s}) - p|$, where $p \in \mathbb{Z}$.

Table 3 Non-dimensional coefficients in (10)

	Ions	Electrons
$\frac{a\,\omega_\mathrm{i}}{v_\mathrm{s}}$	1	$1.6 \times 10^{-2} \approx \varepsilon$
$\frac{a\,\omega_\mathrm{i}}{v_\mathrm{s}} \frac{q_\mathrm{s} B_\mathrm{T}}{m_\mathrm{s}} \frac{a}{v_\mathrm{s}}$	$4.9 \times 10^2 \approx \frac{1}{\varepsilon}$	$4.7 \times 10^2 \approx \frac{1}{\varepsilon}$
$\frac{a\,\omega_\mathrm{i}}{v_\mathrm{s}} \frac{q_\mathrm{s} B_\mathrm{T}}{m_\mathrm{s}} \frac{a}{v_\mathrm{s}} \frac{B_1}{B_\mathrm{T}} \frac{1}{a\,k_\perp}$	$10^{-3} \approx \varepsilon$	$9.5 \times 10^{-4} \approx \varepsilon$
$\frac{q_\mathrm{s}\phi}{m_\mathrm{s} v_\mathrm{s}^2}$	$10^{-3} \approx \varepsilon$	$9.5 \times 10^{-4} \approx \varepsilon$

Applying our ordering yields the following normalized Lagrangian (with primes omitted for readability):

$$L = \left(\varepsilon^s \mathbf{v} + \frac{\mathbf{A}_0}{\varepsilon} + \varepsilon\,\mathbf{A}_1\right) \cdot \dot{\mathbf{x}} - \frac{|\mathbf{v}|^2}{2} - \varepsilon\,\phi\,, \tag{12}$$

where $s = 0$ for ions and $s = 1$ for electrons, and the sign of the charge is absorbed into the definition of the potentials. Our ordering is the standard gyrokinetic ordering [4], with the exception that we take the mass ratio $m_\mathrm{e}/m_\mathrm{i}$ into account. The only place in the Lagrangian where this plays a role is the first term of (12), where $a\,\omega_\mathrm{i}/v_\mathrm{e} = v_\mathrm{i}/v_\mathrm{e} \sim (m_\mathrm{e}/m_\mathrm{i})^{1/2} \approx \varepsilon$ appears because $\dot{\mathbf{x}}$ is normalized to the ion thermal velocity for both species. This, in turn, is motivated by the fact that $\dot{\mathbf{x}} \sim a\,\omega_\mathrm{i}$ is not related to the thermal velocity of a species, but rather to the chosen space and time scales.

The Euler–Lagrange equations corresponding to (12), evaluated at the particle trajectory $(\mathbf{x}(t), \mathbf{v}(t))$, yield

$$\varepsilon^s \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{v}\,, \quad \varepsilon^s \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \mathbf{E} + \frac{\mathbf{v} \times \mathbf{B}_0}{\varepsilon} + \mathbf{v} \times \mathbf{B}_1\,. \tag{13}$$

Here, we used the fact that lengths have been normalized to the minor radius a and that we also assumed $k_\perp a \sim 1/\varepsilon$ as well as $k_\parallel a \sim 1$, such that in non-dimensional variables we have

$$\mathbf{E} = \varepsilon \left(\frac{1}{\varepsilon}\nabla_\perp \phi + \mathbf{b}_0 \nabla_\parallel \phi - \mathbf{b}_0 \frac{\partial A_\parallel}{\partial t}\right), \quad \mathbf{B}_1 = \nabla_\perp A_\parallel \times \mathbf{b}_0\,. \tag{14}$$

Here, $\nabla_\perp := -\mathbf{b}_0 \times \mathbf{b}_0 \times \nabla$, $\nabla_\parallel := \mathbf{b}_0 \cdot \nabla$ denote the gradients with respect to the direction perpendicular and parallel to the background magnetic field, respectively, with $\mathbf{b}_0 := \mathbf{B}_0/B_0$. Moreover, we assumed a static and homogeneous background magnetic field $\mathbf{B}_0$, we introduced the parallel vector potential $A_\parallel := \mathbf{A}_1 \cdot \mathbf{b}_0$ and assumed $\mathbf{b}_0 \times \mathbf{A}_1 \times \mathbf{b}_0 = 0$. These assumptions allow us to simplify our calculations in order to focus on the novelty of this work, namely the inclusion of the electron–ion mass ratio in the gyrokinetic reduction.

4 Guiding-Center Reduction

We start with the guiding-center transformation [14, 15] to decouple the rapid gyration around the static and homogeneous background magnetic field $\mathbf{B}_0$ from the slower dynamical variables. We define first a "semi-canonical" set $(\mathbf{x}, \mathbf{p})$ of phase-space coordinates via the transformation $\mathbf{p} := \mathbf{v} + \varepsilon^{1-s} A_\parallel \mathbf{b}_0$. This leads to the single-particle Lagrangian $L = S - H$, where the symplectic part S and the Hamiltonian H read

$$S = \left(\varepsilon^s \mathbf{p} + \frac{\mathbf{A}_0}{\varepsilon}\right) \cdot \dot{\mathbf{x}}\,, \quad H = \frac{|\mathbf{p}|^2}{2} - \varepsilon^{1-s} p_\parallel A_\parallel + \varepsilon^{2-2s} \frac{A_\parallel^2}{2} + \varepsilon\, \phi\,. \tag{15}$$

We then switch to the angle representation $\mathbf{p} = p_\parallel \mathbf{b}_0 + \sqrt{2\mu B_0}\, \mathbf{c}_0$ defined by

$$p_\parallel = \mathbf{p} \cdot \mathbf{b}_0\,, \quad \mu := \frac{|\mathbf{b}_0 \times \mathbf{p} \times \mathbf{b}_0|^2}{2B_0}\,, \quad \theta := \arctan\left(\frac{\mathbf{p} \cdot \mathbf{e}_1}{\mathbf{p} \cdot \mathbf{e}_2}\right), \tag{16}$$

where $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{b}_0)$ represents a local static orthonormal basis of $\mathbb{R}^3$, given an arbitrary unit vector $\mathbf{e}_1$ perpendicular to $\mathbf{b}_0$, and $(\mathbf{a}_0, \mathbf{b}_0, \mathbf{c}_0)$ represents a θ-dependent orthonormal basis, with $\mathbf{c}_0 := -\mathbf{e}_1 \sin\theta - \mathbf{e}_2 \cos\theta$ and $\mathbf{a}_0 := \mathbf{e}_1 \cos\theta - \mathbf{e}_2 \sin\theta$. We remark that the Jacobian determinant of the transformation leading to the angle representation is B_0.

The guiding-center transformation reads $\mathbf{x} = \overline{\boldsymbol{X}} + \overline{\boldsymbol{\rho}}$, where $\overline{\boldsymbol{X}}$ denotes the guiding-center position and $\overline{\boldsymbol{\rho}}$ the generating vector field of the transformation, to be determined. Substituting the guiding-center transformation into S, expanding $\mathbf{A}_0(\overline{\boldsymbol{X}} + \overline{\boldsymbol{\rho}})$ around $\overline{\boldsymbol{X}}$ as $\mathbf{A}_0(\overline{\boldsymbol{X}}) + \overline{\boldsymbol{\rho}} \cdot \nabla \mathbf{A}_0(\overline{\boldsymbol{X}})$, using the equivalence of Lagrangians under the addition of total differentials (denoted by $\sim$ here and in the following), namely $\mathbf{A}_0(\overline{\boldsymbol{X}}) \cdot \dot{\overline{\boldsymbol{\rho}}} \sim -\dot{\mathbf{A}}_0(\overline{\boldsymbol{X}}) \cdot \overline{\boldsymbol{\rho}} = -\dot{\overline{\boldsymbol{X}}} \cdot \nabla \mathbf{A}_0(\overline{\boldsymbol{X}}) \cdot \overline{\boldsymbol{\rho}}$, and the vector identity $\nabla \mathbf{A}_0 \cdot \overline{\boldsymbol{\rho}} - \overline{\boldsymbol{\rho}} \cdot \nabla \mathbf{A}_0 = \overline{\boldsymbol{\rho}} \times (\nabla \times \mathbf{A}_0)$, finally yields

$$\begin{aligned} S = &\left(\varepsilon^s p_\parallel \mathbf{b}_0 + \varepsilon^s \sqrt{2\mu B_0}\, \mathbf{c}_0 + \frac{\mathbf{A}_0(\overline{\boldsymbol{X}})}{\varepsilon} - \frac{\overline{\boldsymbol{\rho}} \times \mathbf{B}_0}{\varepsilon}\right) \cdot \dot{\overline{\boldsymbol{X}}} \\ &+ \left(\varepsilon^s p_\parallel \mathbf{b}_0 + \varepsilon^s \sqrt{2\mu B_0}\, \mathbf{c}_0 + \frac{\overline{\boldsymbol{\rho}} \cdot \nabla \mathbf{A}_0(\overline{\boldsymbol{X}})}{\varepsilon}\right) \cdot \dot{\overline{\boldsymbol{\rho}}}\,. \end{aligned} \tag{17}$$

Defining the generating vector field as

$$\overline{\boldsymbol{\rho}} = \varepsilon^{1+s} \frac{\sqrt{2\mu B_0}}{B_0^2} \mathbf{B}_0 \times \mathbf{c}_0 = \varepsilon^{1+s} \sqrt{\frac{2\mu}{B_0}} \mathbf{a}_0 \tag{18}$$

removes the θ-angle dependence from S and yields the well-known guiding-center Lagrangian [15]

$$L = \left(\varepsilon^s p_\parallel \mathbf{b}_0 + \frac{\mathbf{A}_0(\overline{\boldsymbol{X}})}{\varepsilon} \right) \cdot \dot{\overline{\boldsymbol{X}}} \pm \varepsilon^{1+2s} \mu \, \dot{\theta} - H \, . \tag{19}$$

Here, the positive sign is to be used for ions and the negative sign for electrons. We also note that $\overline{\boldsymbol{\rho}}$ is positive for ions and negative for electrons, due to the term $\mathbf{B}_0$ in (18) and our convention that the fields carry the sign of the particle species.

5 Gyrokinetic Reduction

As a result of the guiding-center transformation, the problem of gyrokinetic reduction has been reduced to removing the θ-angle dependence from the Hamiltonian H in (19). This is usually accomplished by canonical Lie-transforms, which leave the symplectic part of the Lagrangian unchanged [3, 4, 6, 23]. In this work we follow a different approach, which closely resembles the guiding-center formalism and has been applied recently in the drift-kinetic regime [19, 21] and in the gyrokinetic regime in the electrostatic case [18], and apply it to the electromagnetic case.

5.1 Generating Differential Equations (GDEs)

We introduce first the "preliminary" gyro-center coordinates $(\mathbf{X}, P_\parallel, \widehat{\mu}, \Theta)$ via the transformation

$$\overline{\boldsymbol{X}} = \mathbf{X} + \boldsymbol{\rho} \, , \quad p_\parallel = P_\parallel + G_\parallel \, , \quad \mu = \widehat{\mu} + G_\mu \, , \quad \theta = \Theta + G_\Theta \, , \tag{20}$$

where $(\boldsymbol{\rho}, G_\parallel, G_\mu, G_\Theta)$ is the generating vector field of the transformation, to be determined as a function of the new gyro-center coordinates. Moreover, we shall make use of the equivalence $L \sim L + \dot{\mathcal{S}}$, for a total differential $\dot{\mathcal{S}}$ of the form

$$\dot{\mathcal{S}} = \left(\frac{1}{\varepsilon} \nabla_\perp \mathcal{S} + \nabla_\parallel \mathcal{S} \, \mathbf{b}_0 \right) \cdot \dot{\mathbf{X}} + \frac{\partial \mathcal{S}}{\partial P_\parallel} \dot{P}_\parallel + \frac{\partial \mathcal{S}}{\partial \widehat{\mu}} \dot{\widehat{\mu}} + \frac{\partial \mathcal{S}}{\partial \Theta} \dot{\Theta} + \frac{\partial \mathcal{S}}{\partial t} \, . \tag{21}$$

Here, the factor $1/\varepsilon$ in front of $\nabla_\perp$ appears because we assume $\mathcal{S}$ with the same scale lengths as the potentials ϕ and $\mathbf{A}_1$, namely $k_\perp a \approx 1/\varepsilon$ and $k_\parallel a \approx 1$ as in (14). We will check the validity of this assumption a posteriori. Substituting (20) in the

guiding-center Lagrangian (19) and adding (21) leads to

$$
\begin{aligned}
L = &\left[\varepsilon^s (P_\parallel + G_\parallel)\mathbf{b}_0 + \frac{\mathbf{A}_0(\mathbf{X}+\boldsymbol{\rho})}{\varepsilon}\right]\cdot(\dot{\mathbf{X}}+\dot{\boldsymbol{\rho}}) \\
&\pm \varepsilon^{1+2s}\,(\widehat{\mu}+G_\mu)\,(\dot{\Theta}+\dot{G}_\Theta) - H + \frac{\partial \mathcal{S}}{\partial t} \\
&+ \left(\frac{1}{\varepsilon}\nabla_\perp \mathcal{S} + \nabla_\parallel \mathcal{S}\,\mathbf{b}_0\right)\cdot\dot{\mathbf{X}} + \frac{\partial \mathcal{S}}{\partial P_\parallel}\dot{P}_\parallel + \frac{\partial \mathcal{S}}{\partial \widehat{\mu}}\dot{\widehat{\mu}} + \frac{\partial \mathcal{S}}{\partial \Theta}\dot{\Theta}\,,
\end{aligned}
\tag{22}
$$

where the Hamiltonian reads

$$
H = \frac{(P_\parallel + G_\parallel)^2}{2} + (\widehat{\mu}+G_\mu)B_0 - \varepsilon^{1-s}(P_\parallel+G_\parallel)A_\parallel + \varepsilon^{2-2s}\frac{A_\parallel^2}{2} + \varepsilon\,\phi\,. \tag{23}
$$

The next step is to identify in this Lagrangian the components of the symplectic form and the complete Hamiltonian in the new coordinates. In other words, we aim to express $\dot{\boldsymbol{\rho}}$ and $\dot{G}_\Theta$ in terms of the "basis vectors" $(\dot{\mathbf{X}},\ \dot{P}_\parallel,\ \dot{\widehat{\mu}},\ \dot{\Theta})$ and a Hamiltonian part (corresponding to $\dot{t}$), as in (21).

Let us briefly describe the steps that lead to the desired form of the Lagrangian (22). First, we can use $(P_\parallel + G_\parallel)\mathbf{b}_0\cdot\dot{\boldsymbol{\rho}} \sim -(\dot{P}_\parallel + \dot{G}_\parallel)\mathbf{b}_0\cdot\boldsymbol{\rho}$ and introduce the notation $\rho_\parallel := \mathbf{b}_0\cdot\boldsymbol{\rho}$. Moreover, we use $\dot{\widehat{\mu}}\,\dot{G}_\Theta \sim -\dot{\widehat{\mu}}\,G_\Theta$ and apply the same manipulations to the term $\mathbf{A}_0(\mathbf{X}+\boldsymbol{\rho})\cdot(\dot{\mathbf{X}}+\dot{\boldsymbol{\rho}})$, as in the guiding-center transformation. This leads to the Lagrangian

$$
\begin{aligned}
L = &\left[\varepsilon^s\left(P_\parallel + G_\parallel + \frac{1}{\varepsilon^s}\nabla_\parallel\mathcal{S}\right)\mathbf{b}_0 + \frac{\mathbf{A}_0(\mathbf{X})}{\varepsilon} + \frac{(\nabla_\perp\mathcal{S} - \boldsymbol{\rho}\times\mathbf{B}_0)}{\varepsilon}\right]\cdot\dot{\mathbf{X}} \\
&+ \left(\frac{\partial\mathcal{S}}{\partial P_\parallel} - \varepsilon^s\rho_\parallel\right)\dot{P}_\parallel + \left(\frac{\partial\mathcal{S}}{\partial\widehat{\mu}} \mp \varepsilon^{1+2s}\,G_\Theta\right)\dot{\widehat{\mu}} \\
&\pm \varepsilon^{1+2s}\left(\widehat{\mu}+G_\mu \pm \frac{1}{\varepsilon^{1+2s}}\frac{\partial\mathcal{S}}{\partial\Theta}\right)\dot{\Theta} \pm \varepsilon^{1+2s}\,G_\mu\dot{G}_\Theta - \varepsilon^s\,\rho_\parallel\dot{G}_\parallel \\
&- H + \frac{\partial\mathcal{S}}{\partial t}\,.
\end{aligned}
\tag{24}
$$

An additional term quadratic in $\boldsymbol{\rho}$, which can be shown to be zero up to second order, is neglected in the $\dot{\mathbf{X}}$ component for simplicity. Other scalar terms quadratic in the generators are instead kept in order to illustrate the method. The terms $\dot{G}_\Theta$ and $\dot{G}_\parallel$ can be expanded as in (21). From (24) we can define a system of "generating differential equations" (GDEs) for the functions G introduced in the transformation (20). The GDEs follow from the simple principle of eliminating as many terms as possible from L by suitably choosing the generating functions G and the auxiliary function $\mathcal{S}$. The former are the degrees of freedom in the change of coordinates

(20) and the latter is a degree of freedom due to the invariance of Lagrangian dynamics under the addition of total time derivatives. We can now see why the particular representation (24) of L is favorable for our purpose. Each component of the symplectic form contains one degree of freedom that can be used to eliminate terms:

- the component $\dot{\mathbf{X}} \cdot \mathbf{b}_0$ contains $G_\parallel$;
- the component $\dot{\mathbf{X}}_\perp$ contains $\boldsymbol{\rho}_\perp$;
- the component $\dot{P}_\parallel$ contains $\rho_\parallel$;
- the component $\dot{\widehat{\mu}}$ contains G_Θ;
- the component $\dot{\Theta}$ contains $\partial \mathcal{S}/\partial \Theta$;
- the Hamiltonian terms contain G_μ.

The component $\dot{\Theta}$ is particularly important because it contains the partial derivative of $\mathcal{S}$ with respect to Θ. This means that only gyro-fluctuations can be absorbed via $\partial \mathcal{S}/\partial \Theta$, whereas gyro-averaged terms cannot be removed without introducing secularities. Hence, these gyro-averaged terms in front of $\dot{\Theta}$ stay in the Lagrangian and yield eventually higher-order corrections to the gyro-center magnetic moment. In order to compute these corrections we need to introduce the gyro-average $\langle G \rangle$ and gyro-fluctuations $\widetilde{G}$ of a function $G(\Theta)$ via

$$\langle G \rangle := \frac{1}{2\pi} \int_0^{2\pi} G(\Theta)\, \mathrm{d}\Theta\,, \quad \widetilde{G} := G - \langle G \rangle\,. \tag{25}$$

The gyro-center motion will be determined by the gyro-averaged potentials $\langle \phi \rangle$ and $\langle A_\parallel \rangle$. We therefore write the Hamiltonian (23) as the sum $H = H_{\mathrm{gc}} + \delta H$, where H_{gc} contains the gyro-averaged potentials,

$$H_{\mathrm{gc}} = \frac{P_\parallel^2}{2} + \widehat{\mu}\, B_0 - \varepsilon^{1-s} P_\parallel \langle A_\parallel \rangle + \varepsilon^{2-2s} \frac{\langle A_\parallel^2 \rangle}{2} + \varepsilon \langle \phi \rangle\,, \tag{26}$$

and δH contains the remainders,

$$\delta H = P_\parallel G_\parallel + \frac{G_\parallel^2}{2} + G_\mu B_0 - \varepsilon^{1-s} (P_\parallel \widetilde{A}_\parallel + G_\parallel A_\parallel) + \varepsilon^{2-2s} \frac{\widetilde{A}_\parallel^2}{2} + \varepsilon \widetilde{\phi}\,. \tag{27}$$

In order to remove all dependencies on Θ from the Lagrangian (24), we demand that the generating functions G and the gauge function $\mathcal{S}$ satisfy the following system of equations, each of which can be attributed to a particular component of the symplectic form in (24):

- component $\dot{\mathbf{X}} \cdot \mathbf{b}_0$:

$$G_\parallel + \frac{1}{\varepsilon^s} \nabla_\parallel \mathcal{S} \pm \varepsilon^{1+s} G_\mu \nabla_\parallel G_\Theta - \rho_\parallel \nabla_\parallel G_\parallel = 0\,; \tag{28}$$

- component $\dot{\mathbf{X}}_\perp$:

$$\nabla_\perp \mathcal{S} - \boldsymbol{\rho} \times \mathbf{B}_0 \pm \varepsilon^{1+2s} G_\mu \nabla_\perp G_\Theta - \varepsilon^s \rho_\parallel \nabla_\perp G_\parallel = 0\,; \tag{29}$$

- component $\dot{P}_\parallel$:

$$\frac{\partial \mathcal{S}}{\partial P_\parallel} - \varepsilon^s \rho_\parallel \pm \varepsilon^{1+2s} G_\mu \frac{\partial G_\Theta}{\partial P_\parallel} - \varepsilon^s \rho_\parallel \frac{\partial G_\parallel}{\partial P_\parallel} = 0\,; \tag{30}$$

- component $\dot{\widehat{\mu}}$:

$$\frac{\partial \mathcal{S}}{\partial \widehat{\mu}} \mp \varepsilon^{1+2s} G_\Theta \pm \varepsilon^{1+2s} G_\mu \frac{\partial G_\Theta}{\partial \widehat{\mu}} - \varepsilon^s \rho_\parallel \frac{\partial G_\parallel}{\partial \widehat{\mu}} = 0\,; \tag{31}$$

- component $\dot{\Theta}$:

$$\widetilde{G_\mu} \pm \frac{1}{\varepsilon^{1+2s}} \frac{\partial \mathcal{S}}{\partial \Theta} \pm \widetilde{\left(G_\mu \frac{\partial G_\Theta}{\partial \Theta} \right)} - \frac{1}{\varepsilon^{1+s}} \widetilde{\left(\rho_\parallel \frac{\partial G_\parallel}{\partial \Theta} \right)} = 0\,; \tag{32}$$

- Hamiltonian (component $\dot{t}$):

$$-\delta H + \frac{\partial \mathcal{S}}{\partial t} \pm \varepsilon^{1+2s} G_\mu \frac{\partial G_\Theta}{\partial t} - \varepsilon^s \rho_\parallel \frac{\partial G_\parallel}{\partial t} = 0\,. \tag{33}$$

The system (28)–(33) is a non-linear system of PDEs for $(\boldsymbol{\rho}, G_\parallel, G_\mu, G_\Theta, \mathcal{S})$, termed the GDEs. Assuming that the generating functions satisfy the GDEs, the Lagrangian (24) reads

$$\begin{aligned} L = &\left[\varepsilon^s P_\parallel \mathbf{b}_0 + \frac{\mathbf{A}_0(\mathbf{X})}{\varepsilon} \right] \cdot \dot{\mathbf{X}} - H_{\mathrm{gc}} \\ &\pm \varepsilon^{1+2s} \left(\widehat{\mu} + \langle G_\mu \rangle + \left\langle G_\mu \frac{\partial G_\Theta}{\partial \Theta} \right\rangle - \frac{1}{\varepsilon^{1+s}} \left\langle \rho_\parallel \frac{\partial G_\parallel}{\partial \Theta} \right\rangle \right) \dot{\Theta}\,. \end{aligned} \tag{34}$$

This is the most general gyro-center Lagrangian, containing all possible corrections to the gyro-center magnetic moment $\widehat{\mu}$, at any order in ε. In order to leverage absence of the coordinate Θ in L one usually adopts as one of the coordinates the conjugate momentum

$$\mu := \frac{\partial L}{\partial \dot{\Theta}} = \widehat{\mu} + \langle G_\mu \rangle + \left\langle G_\mu \frac{\partial G_\Theta}{\partial \Theta} \right\rangle - \frac{1}{\varepsilon^{1+s}} \left\langle \rho_\parallel \frac{\partial G_\parallel}{\partial \Theta} \right\rangle, \tag{35}$$

which in the literature is typically referred to as gyro-center magnetic moment. Then, since L is independent of the gyro-angle Θ, the coordinate μ is a constant

of the motion. The corresponding Lagrangian in the final gyro-center coordinates $(\mathbf{X}, P_\parallel, \mu, \Theta)$ reads

$$L = \left[\varepsilon^s P_\parallel \mathbf{b}_0 + \frac{\mathbf{A}_0(\mathbf{X})}{\varepsilon}\right] \cdot \dot{\mathbf{X}} \pm \varepsilon^{1+2s}\, \mu\, \dot{\Theta} - H_{\mathrm{gc}}\,, \tag{36}$$

with the Hamiltonian

$$H_{\mathrm{gc}} = \frac{P_\parallel^2}{2} + \widehat{\mu}(\mu) B_0 - \varepsilon^{1-s} P_\parallel \left\langle A_\parallel \right\rangle + \varepsilon^{2-2s} \frac{\left\langle A_\parallel^2 \right\rangle}{2} + \varepsilon \left\langle \phi \right\rangle\,, \tag{37}$$

where $\widehat{\mu}(\mu)$ is the inverse of (35). We remark the following comments:

1. The Lagrangian (34) is equivalent to the exact Lagrangian if and only if the GDE system (28)–(33) has a solution. It is known, and we will show again in the next section, that it has at least asymptotic solutions as $\varepsilon \to 0$. Such solutions can be computed via Lie-transforms for the generating functions, but also with a simple power series expansions, as shown below. Depending on the order of truncation of these asymptotic solutions, gyrokinetic theories of different accuracy can be obtained.
2. A particular complication for solving the GDEs arises from the fact that the dynamical potentials ϕ and $A_\parallel$ are evaluated at the true particle position:

$$\frac{\mathbf{x}}{\varepsilon} = \frac{\mathbf{X} + \boldsymbol{\rho}}{\varepsilon} \pm \varepsilon^s \sqrt{\frac{2(\widehat{\mu} + G_\mu)}{B_0}}\, \mathbf{a}_0(\Theta + G_\Theta)\,. \tag{38}$$

   Here, we indicate the strong variations in the perpendicular direction,[2] $k_\perp a \sim 1/\varepsilon$. It will be shown below that $|\boldsymbol{\rho}| \sim \varepsilon^2$ such that a Taylor expansion around $\mathbf{X}/\varepsilon$ is possible for electrons ($s = 1$) but not for ions ($s = 0$). Therefore, our gyrokinetic reduction will result in a drift-kinetic Lagrangian for electrons (without gyro-average operators) and a gyrokinetic Lagrangian for ions (with gyro-average operators).
3. The Lagrangian (36) is equivalent to the exact Lagrangian if and only if system (28)–(33) has a solution, and moreover if the mapping $\widehat{\mu} \mapsto \mu$ in (35) is invertible. Again, the inverse mapping can be found asymptotically as $\varepsilon \to 0$.
4. The symplectic part of the gyro-center Lagrangian L in (36) is the same as in the guiding-center Lagrangian (19). Hence, with the ansatz (20) we have obtained the same result as with canonical Lie transforms, which leave the symplectic part unchanged by construction.

[2] The argument $\mathbf{x}/\varepsilon$ is a consequence of our normalization of the spatial coordinate with respect to a. We note however that $k_\parallel a \sim 1$ is still assumed. A more rigorous notation would be $\mathbf{x}/\varepsilon = (x_\parallel, \mathbf{x}_\perp/\varepsilon)$.

5. The Euler–Lagrange equations corresponding to (36) read

$$\begin{cases} \dfrac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \dfrac{1}{\varepsilon^s}\dfrac{\partial H}{\partial P_\parallel}\mathbf{b}_0 + \varepsilon\,\dfrac{\mathbf{b}_0 \times \nabla H}{B_0}\,, & \dfrac{\mathrm{d}\mu}{\mathrm{d}t} = 0\,, \\ \dfrac{\mathrm{d}P_\parallel}{\mathrm{d}t} = -\dfrac{1}{\varepsilon^s}\nabla_\parallel H\,, & \dfrac{\mathrm{d}\Theta}{\mathrm{d}t} = \dfrac{1}{\varepsilon^{1+2s}}\dfrac{\partial H}{\partial \mu}\,. \end{cases} \tag{39}$$

The explicit form of the Hamiltonian is determined by the inverse $\widehat{\mu}(\mu)$ of (35). Hence, H is an asymptotic series in ε, as it depends on the generating functions G. Moreover, H is not necessarily convergent when the number of terms in the expansion is increased, and thus needs to be truncated. This results in an error with respect to the exact dynamics. The truncation error is larger for electrons. In particular, the phase $\Theta(t)$ has an error that is by a factor ε^{-2} larger for electrons ($s = 1$) than for ions ($s = 0$), if both Hamiltonians are truncated at the same order.[3] The same is true for the parallel coordinates $\mathbf{X}(t) \cdot \mathbf{b}_0$ and $P_\parallel(t)$, where the error amplification for electrons is ε^{-1} larger than for ions. We conclude that in order to have the same truncation error for ions and electrons with respect to their exact slow-manifold dynamics, the electron Hamiltonian should contain one order more in the ε-expansion. If truncated at the same order, the electron dynamics will be less accurate and possibly lead to erroneous dynamics in the long time limit.

5.2 *Asymptotic Solution of the GDEs*

We now aim to solve the system (28)–(33) for the generating vector field, which we denote by $\mathbf{F} = (\boldsymbol{\rho}, G_\parallel, G_\mu, G_\Theta)$.[4] Since we are not able to solve the GDEs exactly, we try an asymptotic expansion of the solution in ε. There exist different approaches to constructing such an asymptotic expansion, the most popular in the context of gyrokinetics being Lie transforms [2, 11, 23]. In this work we use a simple power series for the unknowns $\mathbf{F}$ and $\mathcal{S}$,

$$\mathbf{F} = \sum_{n=0}^{\infty} \mathbf{F}_n\,\varepsilon^n\,, \quad \mathcal{S} = \sum_{n=0}^{\infty} \mathcal{S}_n\,\varepsilon^n\,, \tag{40}$$

[3] Since the phase $\Theta(t)$ is decoupled from the rest of the gyro-center motion, the phase error is not relevant when looking at gyro-tropic particle distributions.

[4] The function $\mathcal{S}$ plays the role of an auxiliary function.

as well as for the dynamical potentials ϕ and $A_\parallel$,

$$\phi = \sum_{n=0}^{\infty} \phi_n \, \varepsilon^n \,, \quad A_\parallel = \sum_{n=0}^{\infty} A_{\parallel,n} \, \varepsilon^n \,. \tag{41}$$

We compute the coefficients $\mathbf{F}_n, \mathcal{S}_n$ by substituting this ansatz into system (28)–(33) and comparing coefficients of the same power in ε. By convention, coefficients with a negative index vanish. From the component $\dot{\mathbf{X}} \cdot \mathbf{b}_0$ in (28) we obtain

$$G_{\parallel,n-s} + \nabla_\parallel \mathcal{S}_n \pm \sum_{\ell=0}^{n-1-2s} G_{\mu,\ell} \nabla_\parallel G_{\Theta,n-1-2s-\ell} - \sum_{\ell=0}^{n-s} \rho_{\parallel,\ell} \nabla_\parallel G_{\parallel,n-s-\ell} = 0 \,. \tag{42}$$

From the component $\dot{\mathbf{X}}_\perp$ in (29) we obtain

$$\nabla_\perp \mathcal{S}_n - \boldsymbol{\rho}_n \times \mathbf{B}_0 \pm \sum_{\ell=0}^{n-1-2s} G_{\mu,\ell} \nabla_\perp G_{\Theta,n-1-2s-\ell} - \sum_{\ell=0}^{n-s} \rho_{\parallel,\ell} \nabla_\perp G_{\parallel,n-s-\ell} = 0 \,. \tag{43}$$

From the component $\dot{P}_\parallel$ in (30) we obtain

$$\frac{\partial \mathcal{S}_n}{\partial P_\parallel} - \rho_{\parallel,n-s} \pm \sum_{\ell=0}^{n-1-2s} G_{\mu,\ell} \frac{\partial G_{\Theta,n-1-2s-\ell}}{\partial P_\parallel} - \sum_{\ell=0}^{n-s} \rho_{\parallel,\ell} \frac{\partial G_{\parallel,n-s-\ell}}{\partial P_\parallel} = 0 \,. \tag{44}$$

From the component $\dot{\widehat{\mu}}$ in (31) we obtain

$$\frac{\partial \mathcal{S}_n}{\partial \widehat{\mu}} \mp G_{\Theta,n-1-2s} \pm \sum_{\ell=0}^{n-1-2s} G_{\mu,\ell} \frac{\partial G_{\Theta,n-1-2s-\ell}}{\partial \widehat{\mu}} - \sum_{\ell=0}^{n-s} \rho_{\parallel,\ell} \frac{\partial G_{\parallel,n-s-\ell}}{\partial \widehat{\mu}} = 0 \,. \tag{45}$$

From the component $\dot{\Theta}$ in (32) we obtain

$$\widetilde{G_{\mu,n-1-2s}} \pm \frac{\partial \mathcal{S}_n}{\partial \Theta} \pm \left(\widetilde{G_\mu \frac{\partial G_\Theta}{\partial \Theta}} \right)_{n-1-2s} - \left(\widetilde{\rho_\parallel \frac{\partial G_\parallel}{\partial \Theta}} \right)_{n-s} = 0 \,. \tag{46}$$

From the Hamiltonian component $\dot{t}$ in (33) we obtain

$$
\begin{aligned}
&- P_\parallel G_{\parallel,n} - \sum_{\ell=0}^{n} \frac{G_{\parallel,\ell} G_{\parallel,n-\ell}}{2} - G_{\mu,n} B_0 \\
&+ P_\parallel \widetilde{A_{\parallel,n-1+s}} + \sum_{\ell=0}^{n-1+s} G_{\parallel,\ell} A_{\parallel,n-1+s-\ell} - \sum_{\ell=0}^{n-2+2s} \frac{(A_{\parallel,\ell} \widetilde{A_{\parallel,n-2+2s-\ell}})}{2} \\
&- \widetilde{\phi_{n-1}} + \frac{\partial \mathcal{S}_n}{\partial t} \pm \sum_{\ell=0}^{n-1-2s} G_{\mu,\ell} \frac{\partial G_{\Theta,n-1-2s-\ell}}{\partial t} - \sum_{\ell=0}^{n-s} \rho_{\parallel,\ell} \frac{\partial G_{\parallel,n-s-\ell}}{\partial t} = 0 \,.
\end{aligned}
\tag{47}
$$

Proposition 1 *(Order $n = 0$) For electrons, let $A_{\parallel,0} = \langle A_{\parallel,0} \rangle$. Then a solution of the Eqs. (42)–(47) at order $n = 0$ for ions ($s = 0$) and electrons ($s = 1$) is given by*

$n = 0$	Ions	Electrons
$\rho_{\parallel,0}$	0	Undetermined
$\boldsymbol{\rho}_{\perp,0}$	0	0
$G_{\parallel,0}$	0	Undetermined
$G_{\mu,0}$	0	0
$G_{\Theta,0}$	Undetermined	Undetermined

The auxiliary function is $\mathcal{S}_0 = 0$ for both ions and electrons in this case.

Proof Setting $n = 0$ in (42)–(47) yields

$$0 = (1-s) G_{\parallel,0} + \nabla_\parallel \mathcal{S}_0 - (1-s) \rho_{\parallel,0} \nabla_\parallel G_{\parallel,0} \,, \tag{48a}$$

$$0 = \nabla_\perp \mathcal{S}_0 - \boldsymbol{\rho}_0 \times \mathbf{B}_0 - (1-s) \rho_{\parallel,0} \nabla_\perp G_{\parallel,0} \,, \tag{48b}$$

$$0 = \frac{\partial \mathcal{S}_0}{\partial P_\parallel} - (1-s) \rho_{\parallel,0} - (1-s) \rho_{\parallel,0} \frac{\partial G_{\parallel,0}}{\partial P_\parallel} \,, \tag{48c}$$

$$0 = \frac{\partial \mathcal{S}_0}{\partial \widehat{\mu}} - (1-s) \rho_{\parallel,0} \frac{\partial G_{\parallel,0}}{\partial \widehat{\mu}} 0 \,, \tag{48d}$$

$$0 = \frac{\partial \mathcal{S}_0}{\partial \Theta} - (1-s) \left(\rho_{\parallel,0} \widetilde{\frac{\partial G_{\parallel,0}}{\partial \Theta}} \right) , \tag{48e}$$

$$0 = -P_\parallel G_{\parallel,0} - \frac{G_{\parallel,0}^2}{2} - G_{\mu,0}B_0 + s\left(P_\parallel \widetilde{A_{\parallel,0}} + \frac{\widetilde{A_{\parallel,0}^2}}{2}\right) + \frac{\partial \mathcal{S}_0}{\partial t} - (1-s)\rho_{\parallel,0}\frac{\partial G_{\parallel,0}}{\partial t}\,. \tag{48f}$$

By assuming $\rho_{\parallel,0} = 0$ for the ions, Eq. (48e) yields $\mathcal{S}_0 = 0$ for both ions and electrons. The other results can then be deduced easily from the remaining equations. In Eq. (48f) for electrons ($s = 1$) we use the assumption $A_{\parallel,0} = \langle A_{\parallel,0}\rangle$ (or equivalently $\widetilde{A_{\parallel,0}} = 0$). □

Proposition 2 *(Order $n = 1$) For electrons, let $\phi_0 = \langle\phi_0\rangle$. Then a solution of the Eqs. (42)–(47) at order $n = 1$ for ions ($s = 0$) and electrons ($s = 1$) is given by*

$n = 1$	Ions ($s = 0$)	Electrons ($s = 1$)
$\rho_{\parallel,1-s}$	0	0
$\boldsymbol{\rho}_{\perp,1}$	0	0
$G_{\parallel,1-s}$	0	0
$G_{\mu,1}$	$\frac{1}{B_0}(P_\parallel \widetilde{A_{\parallel,0}} - \widetilde{\phi_0})$	$\frac{1}{B_0}(P_\parallel \widetilde{A_{\parallel,1}} - \widetilde{A_{\parallel,0}A_{\parallel,1}})$
$G_{\Theta,0}$	0	Undetermined

The auxiliary function is $\mathcal{S}_1 = 0$ for both ions and electrons in this case. The generators $G_{\Theta,1}$ remain undetermined at this order and the generator $\rho_{\parallel,1}$ remains undetermined for electrons at this order.

Proof Setting $n = 1$ in (42)–(47) and substituting the results obtained from $n = 0$ yields

$$0 = G_{\parallel,1-s} + \nabla_\parallel \mathcal{S}_1 - s\,\rho_{\parallel,0}\nabla_\parallel G_{\parallel,1-s}\,, \tag{49a}$$

$$0 = \nabla_\perp \mathcal{S}_1 - \boldsymbol{\rho}_1 \times \mathbf{B}_0\,, \tag{49b}$$

$$0 = \frac{\partial \mathcal{S}_1}{\partial P_\parallel} - \rho_{\parallel,1-s}\,, \tag{49c}$$

$$0 = \frac{\partial \mathcal{S}_1}{\partial \widehat{\mu}} - (1-s)G_{\Theta,0}\,, \tag{49d}$$

$$0 = \frac{\partial \mathcal{S}_1}{\partial \Theta} - s\left(\widetilde{\rho_{\parallel,0}\frac{\partial G_{\parallel,1}}{\partial \Theta}}\right), \tag{49e}$$

$$0 = -P_\parallel G_{\parallel,1} - G_{\mu,1} B_0 \; + \; P_\parallel \widetilde{A_{\parallel,s}} + s\left(G_{\parallel,1} A_{\parallel,0} - \widetilde{A_{\parallel,0} A_{\parallel,1}}\right) - \widetilde{\phi_0} + \frac{\partial \mathcal{S}_1}{\partial t}\,. \tag{49f}$$

By assuming $\rho_{\parallel,0} = 0$ for the electrons, Eq. (49e) yields $\mathcal{S}_1 = 0$ for both ions and electrons. The other results can then be deduced easily from the remaining equations. In Eq. (49f) for electrons ($s = 1$) we use the assumption $\phi_0 = \langle \phi_0 \rangle$ (or equivalently $\widetilde{\phi_0} = 0$). □

Proposition 3 *(Order $n = 2$) A solution of the Eqs. (42)–(47) at order $n = 2$ for ions and electrons is given by*

$n = 2$	Ions ($s = 0$)	Electrons ($s = 1$)
$\rho_{\parallel,2-s}$	$\dfrac{\partial \mathcal{S}_2}{\partial P_\parallel}$	0
$\boldsymbol{\rho}_{\perp,2}$	$\dfrac{\mathbf{b}_0}{B_0} \times \nabla_\perp \mathcal{S}_2$	0
$G_{\parallel,2-s}$	$-\nabla_\parallel \mathcal{S}_2$	0
$G_{\mu,2}$	$\dfrac{1}{B_0}\left(- P_\parallel G_{\parallel,2} + P_\parallel \widetilde{A_{\parallel,1}} - \widetilde{\phi_1} - \dfrac{\widetilde{A^2_{\parallel,0}}}{2} + \dfrac{\partial \mathcal{S}_2}{\partial t}\right)$	$\dfrac{1}{B_0}\left(P_\parallel \widetilde{A_{\parallel,2}} - \widetilde{\phi_1} - \widetilde{A_{\parallel,0} A_{\parallel,2}} - \dfrac{\widetilde{A^2_{\parallel,1}}}{2}\right)$
$G_{\Theta,1}$	$\dfrac{\partial \mathcal{S}_2}{\partial \widehat{\mu}}$	Undetermined

The auxiliary function $\mathcal{S}_2$ is given by

$$\mathcal{S}_2(\Theta) = \frac{(1-s)}{B_0} \int_{-\infty}^{\Theta} \left(\widetilde{\phi_0} - P_\parallel \widetilde{A_{\parallel,0}}\right) \mathrm{d}\Theta'\,, \tag{50}$$

and therefore vanishes for electrons. The generators $G_{\Theta,2}$ remain undetermined at this order and the generator $\rho_{\parallel,2}$ remains undetermined for electrons at this order. Moreover, the generators $G_{\Theta,0}$ and $G_{\Theta,1}$ also remain undetermined at this order for electrons.

Proof Setting $n = 2$ in (42)–(47) and substituting the results obtained from $n = 0$ and $n = 1$ leads to

$$0 = G_{\parallel,2-s} + \nabla_\parallel \mathcal{S}_2\,, \tag{51a}$$

$$0 = \nabla_\perp \mathcal{S}_2 - \boldsymbol{\rho}_2 \times \mathbf{B}_0\,, \tag{51b}$$

$$0 = \frac{\partial \mathcal{S}_2}{\partial P_\parallel} - \rho_{\parallel,2-s}\,, \tag{51c}$$

$$0 = \frac{\partial \mathcal{S}_2}{\partial \widehat{\mu}} - (1-s) G_{\Theta,1}\,, \tag{51d}$$

$$0 = (1-s)\widetilde{G_{\mu,1}} + \frac{\partial \mathcal{S}_2}{\partial \Theta}\,, \tag{51e}$$

$$\begin{aligned} 0 = &-P_\parallel G_{\parallel,2} - G_{\mu,2} B_0 \;+\; P_\parallel \widetilde{A_{\parallel,1+s}} - (1-s)\frac{\widetilde{A_{\parallel,0} A_{\parallel,0}}}{2} \\ &+ s\left(G_{\parallel,2} A_{\parallel,0} - \widetilde{A_{\parallel,0} A_{\parallel,2}} - \frac{\widetilde{A_{\parallel,1} A_{\parallel,1}}}{2} \right) - \widetilde{\phi}_1 + \frac{\partial \mathcal{S}_2}{\partial t}\,. \end{aligned} \tag{51f}$$

Equation (51e) yields

$$\mathcal{S}_2(\Theta) = \frac{(1-s)}{B_0} \int_{-\infty}^{\Theta} \left(\widetilde{\phi_0} - P_\parallel \widetilde{A_{\parallel,0}} \right) \mathrm{d}\Theta'\,. \tag{52}$$

The other results can then easily be deduced from the remaining equations. □

Proposition 4 *(Order $n = 3$ for electrons) For electrons ($s = 1$), a solution of the Eqs. (42)–(47) at order $n = 3$ is given by*

$n = 3$	Electrons ($s = 1$)
$\rho_{\parallel,2}$	0
$\boldsymbol{\rho}_{\perp,3}$	0
$G_{\parallel,2}$	0
$G_{\mu,3}$	$\frac{1}{B_0}\left(P_\parallel \widetilde{A_{\parallel,3}} - \widetilde{\phi_2} - \widetilde{A_{\parallel,0} A_{\parallel,3}} - \widetilde{A_{\parallel,1} A_{\parallel,2}} \right)$
$G_{\Theta,0}$	0

The auxiliary function is $\mathcal{S}_3 = 0$ and the generators $G_{\Theta,1}$, $G_{\Theta,2}$, $G_{\Theta,3}$ and $\rho_{\parallel,3}$ remain undetermined at this order for electrons.

Proof Setting $n = 3$ and $s = 1$ in (42)–(47) and substituting the results obtained from $n = 0$, $n = 1$ and $n = 2$ leads to

$$0 = G_{\parallel,2} + \nabla_\parallel \mathcal{S}_3\,, \tag{53a}$$

$$0 = \nabla_\perp \mathcal{S}_3 - \boldsymbol{\rho}_3 \times \mathbf{B}_0\,, \tag{53b}$$

$$0 = \frac{\partial \mathcal{S}_3}{\partial P_\parallel} - \rho_{\parallel,2}\,, \tag{53c}$$

$$0 = \frac{\partial \mathcal{S}_3}{\partial \widehat{\mu}} + G_{\Theta,0}\,, \tag{53d}$$

$$0 = \frac{\partial \mathcal{S}_3}{\partial \Theta}\,, \tag{53e}$$

$$\begin{aligned} 0 = &-P_\parallel G_{\parallel,3} - G_{\mu,3} B_0 \; + \; P_\parallel \widetilde{A_{\parallel,3}} \\ &+ \left(G_{\parallel,3} A_{\parallel,0} - \widetilde{A_{\parallel,0} A_{\parallel,3}} - \widetilde{A_{\parallel,1} A_{\parallel,2}} \right) - \widetilde{\phi_2} + \frac{\partial \mathcal{S}_3}{\partial t}\,. \end{aligned} \tag{53f}$$

We try a solution with $G_{\Theta,0} = 0$. Substituting the assumption $A_{\parallel,0} = \langle A_{\parallel_0} \rangle$ (or equivalently $\widetilde{A_{\parallel,0}} = 0$) into Eq. (53e) yields $\mathcal{S}_3 = 0$. The other results can then be deduced easily from the remaining equations. □

Proposition 5 *(Order $n = 4$ for electrons) For the electrons ($s = 1$), a solution of the Eqs. (42)–(47) at order $n = 4$ is given by*

$n = 4$	Electrons ($s = 1$)
$\rho_{\parallel,3}$	$\dfrac{\partial \mathcal{S}_4}{\partial P_\parallel}$
$\boldsymbol{\rho}_{\perp,4}$	$\dfrac{\mathbf{b}_0}{B_0^2} \times \nabla_\perp \mathcal{S}_4$
$G_{\parallel,3}$	$-\nabla_\parallel \mathcal{S}_4$
$G_{\mu,4}$	$\dfrac{1}{B_0}\left(- P_\parallel G_{\parallel,4} + P_\parallel \widetilde{A_{\parallel,4}} + G_{\parallel,4} A_{\parallel,0} - \widetilde{A_{\parallel,0} A_{\parallel,4}} - \widetilde{A_{\parallel,1} A_{\parallel,3}} - \dfrac{\widetilde{A_{\parallel,2}^2}}{2} \right) - \widetilde{\phi_3} + \dfrac{\partial \mathcal{S}_4}{\partial t}$
$G_{\Theta,1}$	$-\dfrac{\partial \mathcal{S}_4}{\partial \widehat{\mu}}$

The auxiliary function $\mathcal{S}_4$ for the electrons is given by

$$\mathcal{S}_4(\Theta) = \frac{1}{B_0} \int_{-\infty}^{\Theta} \left(P_\parallel \widetilde{A_{\parallel,1}} - \widetilde{A_{\parallel,0} A_{\parallel,1}} \right) \mathrm{d}\Theta'\,. \tag{54}$$

The generators $G_{\Theta,2}$, $G_{\Theta,3}$, $G_{\Theta,4}$ and $\rho_{\parallel,4}$ remain undetermined at this order for electrons.

Proof Setting $n = 4$ and $s = 1$ in (42)–(47) and substituting the results obtained from $n = 0$, $n = 1$, $n = 2$ and $n = 3$ leads to

$$0 = G_{\parallel,3} + \nabla_\parallel \mathcal{S}_4\,, \tag{55a}$$

$$0 = \nabla_\perp \mathcal{S}_4 - \boldsymbol{\rho}_4 \times \mathbf{B}_0\,, \tag{55b}$$

$$0 = \frac{\partial \mathcal{S}_4}{\partial P_\parallel} - \rho_{\parallel,3}\,, \tag{55c}$$

$$0 = \frac{\partial \mathcal{S}_4}{\partial \widehat{\mu}} + G_{\Theta,1}\,, \tag{55d}$$

$$0 = \frac{1}{B_0}\left(P_\parallel \widetilde{A_{\parallel,1}} - \widetilde{A_{\parallel,0}A_{\parallel,1}}\right) - \frac{\partial \mathcal{S}_4}{\partial \Theta}\,, \tag{55e}$$

$$\begin{aligned} 0 = &-P_\parallel G_{\parallel,4} - G_{\mu,4}B_0 \;+\; P_\parallel \widetilde{A_{\parallel,4}} \\ &+ \left(G_{\parallel,4}A_{\parallel,0} - \widetilde{A_{\parallel,0}A_{\parallel,4}} - \widetilde{A_{\parallel,1}A_{\parallel,3}} - \frac{\widetilde{A^2_{\parallel,2}}}{2}\right) - \widetilde{\phi}_3 + \frac{\partial \mathcal{S}_4}{\partial t}\,. \end{aligned} \tag{55f}$$

Equation (55e) yields

$$\mathcal{S}_4(\Theta) = \frac{1}{B_0}\int_{-\infty}^{\Theta}\left(P_\parallel \widetilde{A_{\parallel,1}} - \widetilde{A_{\parallel,0}A_{\parallel,1}}\right)\mathrm{d}\Theta'\,. \tag{56}$$

The other results can then be deduced easily from the remaining equations. □

Let us make the following remarks on the above propositions:

1. It is not hard to conclude from the above propositions that all generating functions can be determined up to arbitrary order n for both ions and electrons from the GDEs (42)–(47).
2. The derivation presented here and the resulting system of Eqs. (42)–(47) for the gyro-center generators at arbitrary order represents a good starting point for the future implementation of computer algebra software for gyrokinetic reductions, such as the ones mentioned in [5, 20].

5.3 Ion and Electron Gyrokinetic Hamiltonians

In order to get truncation errors of first order $O(\varepsilon)$ for both species in the slow-manifold dynamics ($\dot{\mathbf{X}}$ and $\dot{P}_\parallel$) of (39), it is sufficient to use the first-order ion Hamiltonian and the full second-order electron Hamiltonian. The second-order contributions to the ion Hamiltonian can be used to determine polarization and magnetization of the particles [4]. Based on the preceding propositions, we can compute the explicit form of the gyro-center Hamiltonian (37) up to second order in ε for both ions and electrons. This will be done in three steps:

1. Computation of the conjugate momentum (35);
2. Expansion of the averaged potentials $\langle\phi\rangle$ and $\langle A_\parallel\rangle$ in ε for ions and electrons;
3. Collection of the results to determine the second-order Hamiltonians for both species.

For both ions and electrons, the conjugate momentum (35) can be expanded as

$$\mu = \widehat{\mu} + \varepsilon^2 \left\langle G_{\mu,1} \frac{\partial G_{\Theta,1}}{\partial \Theta} \right\rangle + O(\varepsilon^3) \,, \tag{57}$$

where we already used the results from the preceding propositions: $\rho_{\parallel,0} = \rho_{\parallel,1} = 0$, $G_{\parallel,0} = G_{\parallel,1} = 0$, $G_{\mu,0} = 0$, $\langle G_{\mu,1} \rangle = \langle G_{\mu,2} \rangle = 0$, $G_{\Theta,0} = 0$, and $G_{\parallel,2} = 0$ (for electrons only). By substituting (57) and expanding the potentials as $\phi = \phi_0 + \varepsilon \, \phi_1 + \ldots$ in (37), the second-order gyro-center Hamiltonian reads

$$\begin{aligned} H_{\rm gc} = \frac{P_\parallel^2}{2} + \mu B_0 - \varepsilon^2 \, B_0 \left\langle G_{\mu,1} \frac{\partial G_{\Theta,1}}{\partial \Theta} \right\rangle + \varepsilon \Big(\langle \phi_0 \rangle + \varepsilon \, \langle \phi_1 \rangle \Big) \\ - \varepsilon^{1-s} P_\parallel \Big(\langle A_{\parallel,0} \rangle + \varepsilon \, \langle A_{\parallel,1} \rangle + \varepsilon^2 \langle A_{\parallel,2} \rangle \Big) \\ + \varepsilon^{2-2s} \left(\frac{\left\langle A_{\parallel,0}^2 \right\rangle}{2} + \varepsilon \, \langle A_{\parallel,0} A_{\parallel,1} \rangle + \varepsilon^2 \frac{\left\langle A_{\parallel,1}^2 \right\rangle}{2} + \varepsilon^2 \langle A_{\parallel,0} A_{\parallel,2} \rangle \right) . \end{aligned} \tag{58}$$

It follows that for the ions ($s = 0$) we need first-order expansions of ϕ and $A_\parallel$; for the electrons ($s = 1$) we need additionally $A_{\parallel,2}$. Given that the potentials are evaluated at the particle position (38), using that $\boldsymbol{\rho}_0 = \boldsymbol{\rho}_1 = 0$ for ions and electrons and $\boldsymbol{\rho}_2 = 0$ for electrons, and introducing the guiding-center displacement vector

$$\boldsymbol{\rho}_{\rm gc} = \sqrt{\frac{2\widehat{\mu}}{B_0}} \, \mathbf{a}_0(\Theta) \,, \tag{59}$$

we have

$$\begin{aligned} \frac{\mathbf{x}}{\varepsilon} = \frac{\mathbf{X}}{\varepsilon} \pm \varepsilon^s \boldsymbol{\rho}_{\rm gc} + \varepsilon \, (1 - s) \boldsymbol{\rho}_2 + \varepsilon^2 \boldsymbol{\rho}_3 \\ \pm \varepsilon^{1+s} \left(G_{\mu,1} \frac{\partial \boldsymbol{\rho}_{\rm gc}}{\partial \widehat{\mu}} + G_{\Theta,1} \frac{\partial \boldsymbol{\rho}_{\rm gc}}{\partial \Theta} \right) + \cdots \end{aligned} \tag{60}$$

Taylor expansion of the potentials then leads to

	Ions	Electrons
ϕ_0	$\phi\left(\mathbf{X}/\varepsilon + \boldsymbol{\rho}_{\mathrm{gc}}\right)$	$\phi\left(\mathbf{X}/\varepsilon\right)$
$A_{\parallel,0}$	$A_\parallel\left(\mathbf{X}/\varepsilon + \boldsymbol{\rho}_{\mathrm{gc}}\right)$	$A_\parallel\left(\mathbf{X}/\varepsilon\right)$
ϕ_1	$\left(\boldsymbol{\rho}_2 + G_{\mu,1}\frac{\partial \boldsymbol{\rho}_{\mathrm{gc}}}{\partial \widehat{\mu}} + G_{\Theta,1}\frac{\partial \boldsymbol{\rho}_{\mathrm{gc}}}{\partial \Theta}\right)\cdot \nabla\phi\left(\mathbf{X}/\varepsilon + \boldsymbol{\rho}_{\mathrm{gc}}\right)$	$-\boldsymbol{\rho}_{\mathrm{gc}}\cdot\nabla\phi\left(\mathbf{X}/\varepsilon\right)$
$A_{\parallel,1}$	$\left(\boldsymbol{\rho}_2 + G_{\mu,1}\frac{\partial \boldsymbol{\rho}_{\mathrm{gc}}}{\partial \widehat{\mu}} + G_{\Theta,1}\frac{\partial \boldsymbol{\rho}_{\mathrm{gc}}}{\partial \Theta}\right)\cdot \nabla A_\parallel\left(\mathbf{X}/\varepsilon + \boldsymbol{\rho}_{\mathrm{gc}}\right)$	$-\boldsymbol{\rho}_{\mathrm{gc}}\cdot\nabla A_\parallel\left(\mathbf{X}/\varepsilon\right)$
$A_{\parallel,2}$	not needed	$\frac{1}{2}(\boldsymbol{\rho}_{\mathrm{gc}}\cdot\nabla)^2 A_\parallel\left(\mathbf{X}/\varepsilon\right)$ $+\left(\rho_{3,\parallel}\mathbf{b}_0 - G_{\mu,1}\frac{\partial \boldsymbol{\rho}_{\mathrm{gc}}}{\partial \widehat{\mu}}\right.$ $\left. - G_{\Theta,1}\frac{\partial \boldsymbol{\rho}_{\mathrm{gc}}}{\partial \Theta}\right)\cdot\nabla A_\parallel\left(\mathbf{X}/\varepsilon\right)$

In particular, for ions we can write more elegantly

$$\phi_1 = \boldsymbol{\rho}_2\cdot\nabla\phi + G_{\mu,1}\frac{\mathrm{d}}{\mathrm{d}\widehat{\mu}}\phi + G_{\Theta,1}\frac{\mathrm{d}}{\mathrm{d}\Theta}\phi\,, \tag{61}$$

$$A_{\parallel,1} = \boldsymbol{\rho}_2\cdot\nabla A_\parallel + G_{\mu,1}\frac{\mathrm{d}}{\mathrm{d}\widehat{\mu}}A_\parallel + G_{\Theta,1}\frac{\mathrm{d}}{\mathrm{d}\Theta}A_\parallel\,, \tag{62}$$

which leads to

$$-P_\parallel\left\langle A_{\parallel,1}\right\rangle + \left\langle\phi_1\right\rangle = -B_0\left\langle\boldsymbol{\rho}_2\cdot\nabla G_{\mu,1}\right\rangle - B_0\frac{\mathrm{d}}{\mathrm{d}\widehat{\mu}}\left\langle G_{\mu,1}^2\right\rangle\,. \tag{63}$$

Here, we used $\langle\boldsymbol{\rho}_2\rangle = 0$, $\left\langle G_{\mu,1}\right\rangle = 0$ and $\left\langle G_{\Theta,1}\right\rangle = 0$. For electrons we find

$$\left\langle\phi_1\right\rangle = \left\langle A_{\parallel,1}\right\rangle = 0\,, \tag{64}$$

$$\left\langle A_{\parallel,2}\right\rangle = \frac{\widehat{\mu}}{2B_0}\,\Delta_\perp A_{\parallel,0} - \frac{1}{B_0^2}\left|\nabla_\perp A_{\parallel,0}\right|^2(A_{\parallel,0} - P_\parallel)\,, \tag{65}$$

$$\left\langle\widetilde{A_{\parallel,1}}^2\right\rangle = \frac{\widehat{\mu}}{B_0}\left|\nabla_\perp A_{\parallel,0}\right|^2\,, \tag{66}$$

where we substituted

$$\left\langle (\boldsymbol{\rho}_{\mathrm{gc}} \cdot \nabla^2) A_{\parallel,0} \right\rangle = \frac{\widehat{\mu}}{B_0} \Delta_\perp A_{\parallel,0}\,, \quad \left\langle \boldsymbol{\rho}_{\mathrm{gc}} \boldsymbol{\rho}_{\mathrm{gc}}^\top \right\rangle = \frac{\widehat{\mu}}{B_0} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{67}$$

and $\Delta_\perp$ denotes the Laplace operator in the perpendicular direction to the background field. With regards to the generating functions, from the preceding propositions we deduce

	Ions	Electrons
$G_{\mu,1}$	$\frac{1}{B_0}\left(P_\parallel \widetilde{A_{\parallel,0}} - \widetilde{\phi_0}\right)$	$\frac{1}{B_0}\left(P_\parallel \widetilde{A_{\parallel,1}} - \widetilde{A_{\parallel,0} A_{\parallel,1}}\right)$
$\frac{\partial G_{\Theta,1}}{\partial \Theta}$	$-\frac{\partial G_{\mu,1}}{\partial \widehat{\mu}}$	$-\frac{\partial G_{\mu,1}}{\partial \widehat{\mu}}$
$\left\langle G_{\mu,1} \frac{\partial G_{\Theta,1}}{\partial \Theta} \right\rangle$	$-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\widehat{\mu}} \left\langle G_{\mu,1}^2 \right\rangle$	$-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\widehat{\mu}} \left\langle G_{\mu,1}^2 \right\rangle$

With this we can state our final result by inserting the above findings in the general Lagrangian (36) and the Hamiltonian (58), respectively. The second-order gyrocenter Lagrangian for both ions ($s = 0$) and electrons ($s = 1$) then reads

$$L = \left[\varepsilon^s P_\parallel \mathbf{b}_0 + \frac{\mathbf{A}_0(\mathbf{X})}{\varepsilon} \right] \cdot \dot{\mathbf{X}} \pm \varepsilon^{1+2s} \mu\, \dot{\Theta} - H_{\mathrm{gc}}^0 - H_{\mathrm{gc}}^1 - H_{\mathrm{gc}}^2\,, \tag{68}$$

with the Hamiltonians

	Ions	Electrons
H_{gc}^0	$P_\parallel^2/2 + \mu B_0$	$P_\parallel^2/2 + \mu B_0 - P_\parallel A_{\parallel,0} + A_{\parallel,0}^2/2$
H_{gc}^1	$-P_\parallel \left\langle A_{\parallel,0} \right\rangle + \left\langle \phi_0 \right\rangle$	ϕ_0
H_{gc}^2	$\frac{1}{2}\left\langle A_{\parallel,0}^2 \right\rangle - \frac{B_0}{2}\frac{\mathrm{d}}{\mathrm{d}\mu}\left\langle G_{\mu,1}^2 \right\rangle$ $- B_0 \left\langle \boldsymbol{\rho}_2 \cdot \nabla G_{\mu,1} \right\rangle$	$\frac{B_0}{2}\frac{\mathrm{d}}{\mathrm{d}\mu}\left\langle G_{\mu,1}^2 \right\rangle + \frac{\mu}{B_0}\lvert\nabla_\perp A_{\parallel,0}\rvert^2$ $+ (A_{0,\parallel} - P_\parallel)\left\langle A_{\parallel,2} \right\rangle$

Written fully explicitly, the second-order gyro-center Hamiltonian for ions reads

$$
\begin{aligned}
H^2_{\mathrm{gc}} = {} & \frac{1}{2}\left\langle A^2_{\parallel,0}\right\rangle - \frac{1}{2B_0}\frac{\mathrm{d}}{\mathrm{d}\mu}\left\langle \left(P_\parallel \widetilde{A_{\parallel,0}} - \widetilde{\phi_0}\right)^2\right\rangle \\
& - \left\langle \left[\frac{\mathbf{b}_0}{B_0^2}\times\nabla_\perp \int_{-\infty}^{\Theta} \left(\widetilde{\phi_0} - P_\parallel \widetilde{A_{\parallel,0}}\right)\mathrm{d}\Theta'\right]\cdot\nabla\left(P_\parallel \widetilde{A_{\parallel,0}} - \widetilde{\phi_0}\right)\right\rangle \\
& + \left\langle \left[\frac{\mathbf{b}_0}{B_0}\int_{-\infty}^{\Theta} \widetilde{A_{\parallel,0}}\,\mathrm{d}\Theta'\right]\cdot\nabla\left(P_\parallel \widetilde{A_{\parallel,0}} - \widetilde{\phi_0}\right)\right\rangle .
\end{aligned}
\tag{69}
$$

This result corresponds to the findings in [23], where canonical Lie transforms were applied for the asymptotic expansion. The explicit version of the second-order gyro-center Hamiltonian for electrons reads

$$
H^2_{\mathrm{gc}} = \frac{1}{B_0}|\nabla_\perp A_{\parallel,0}|^2\left[\mu - \frac{(P_\parallel - A_{\parallel,0})^2}{2B_0}\right] - (P_\parallel - A_{0,\parallel})\frac{\mu}{2B_0}\,\Delta_\perp A_{\parallel,0}\,. \tag{70}
$$

According to our analysis of the truncation errors in the gyro-center Eqs. (39), this term should be included in the electron equations to obtain an accuracy that is consistent with the ion equations with first-order Hamiltonian.

Future work will address the implementation of the new electron model derived in this section within existing gyrokinetic codes, for the purpose of verifying numerically the accuracy of the truncation error estimates presented here.

6 Conclusions

In this article we studied the asymptotic derivation of gyro-center equations of motion in strongly magnetized fusion plasmas. We presented a consistent normalization procedure that takes into account the role of the electron–ion mass ratio in the gyrokinetic coordinate transformation and we applied a physical ordering relevant for realistic magnetic confinement fusion devices, such as the Tokamaks ITER and ASDEX Upgrade. We derived a system of generating differential equations for the generating vector field of the gyrokinetic coordinate transformation and presented its solution by means of asymptotic expansions. We finally discussed the accuracy of the gyrokinetic transformation for both ions and electrons necessary to achieve overall first-order accuracy in the gyro-center equations of motion. As a consequence of the ordering assumptions used in the early steps of our gyrokinetic reduction, we showed that higher-order asymptotic expansions for electrons, in particular the term (70), are necessary for achieving consistent accuracy overall with respect to ions.

Acknowledgments We would like to thank Eric Sonnendrücker for supporting our research work and Bruce Scott, Roman Hatzky and Cesare Tronci for insightful discussions about many of the topics treated in this article. This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training program 2014-2018 and 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

1. Bottino, A., Scott, B., Brunner, S., McMillan, B.F., Tran, T.M., Vernay, T., Villard, L., Jolliet, S., Hatzky, R., Peeters, A.G.: Global nonlinear electromagnetic simulations of tokamak turbulence. IEEE Trans. Plasma Sci. **9**, 2129–2135 (2010)
2. Brizard, A.J.: Nonlinear gyrokinetic Maxwell-Vlasov equations using magnetic co-ordinates. J. Plasma Phys. **41**(3):541–559 (1989)
3. Brizard, A.J.: Variational principle for the parallel-symplectic representation of electromagnetic gyrokinetic theory. Phys. Plasmas **24**(8), 081201 (2017). https://doi.org/10.1063/1.4997484
4. Brizard, A.J., Hahm, T.S.: Foundations of nonlinear gyrokinetic theory. Rev. Mod. Phys. **79**, 421–468 (2007). https://doi.org/10.1103/RevModPhys.79.421
5. Burby, J.W., Squire, J., Qin, H.: Automation of the guiding center expansion. Phys. Plasmas **20**, 072105 (2013). https://doi.org/10.1063/1.4813247
6. Cary, J.R.: Lie transform perturbation theory for Hamiltonian systems. Phys. Rep. **79**(2), 129–159 (1981). https://doi.org/10.1016/0370-1573(81)90175-7
7. Chang, C.S., Ku, S., Weitzner, H.: Numerical study of neoclassical plasma pedestal in a tokamak geometry. Phys. Plasmas **11**(5), 2649–2667 (2004)
8. Garbet, X., Idomura, Y., Villard, L., Watanabe, T.H.: Gyrokinetic simulations of turbulent transport. Nucl. Fusion **50**(4), 043002 (2010)
9. Görler, T., Lapillonne, X., Brunner, S., Dannert, T., Jenko, F., Merz, F., Told, D.: The global version of the gyrokinetic turbulence code GENE. J. Comput. Phys. **230**(18), 7053–7071 (2011)
10. Grandgirard, V., Sarazin, Y., Garbet, X., Dif-Pradalier, G., Ghendrih, P., Crouseilles, N., Latu, G., Sonnendrücker, E., Besse, N., Bertrand, P.: GYSELA, a full-f global gyrokinetic Semi-Lagrangian code for ITG turbulence simulations. AIP Conf. Proc. **871**(1), 100–111 (2006)
11. Hahm, T.S.: Nonlinear gyrokinetic equations for tokamak microturbulence. Phys. Fluids **31**(9), 2670–2673 (1988)
12. Krommes, J.A.: The Gyrokinetic Description of Microturbulence in Magnetized Plasmas. Annu. Rev. Fluid Mech. **44**(1), 175–201 (2012). https://doi.org/10.1146/annurev-fluid-120710-101223
13. Liewer, P.C.: Measurements of microturbulence in tokamaks and comparisons with theories of turbulence and anomalous transport. Nucl. Fusion **25**(5), 543–621 (1985). https://doi.org/10.1088/0029-5515/25/5/004
14. Littlejohn, R.G.: A guiding center Hamiltonian: A new approach. J. Math. Phys. **20**(12), 2445–2458 (1979)
15. Littlejohn, R.G.: Variational principles of guiding center motion. J. Plasma Phys. **29**, 111–125 (1983)
16. Low, F.E.: A Lagrangian formulation of the Boltzmann-Vlasov equation for plasmas. In: Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, vol. 248(1253), pp. 282–287 (1958). https://doi.org/10.1098/rspa.1958.0244
17. Meyer, H., Angioni, C., Albert, C.G., Arden, N., Parra, R.A., Asunta, O., De Baar, M., Balden, M., Bandaru, V., Behler, K., Bergmann, A.: Overview of physics studies on ASDEX Upgrade. Nucl. Fusion **59**(11), 112014 (2019)

18. Parra, F.I., Calvo, I.: Phase-space Lagrangian derivation of electrostatic gyrokinetics in general geometry. Plasma Phys. Controlled Fusion **53**(4), 045001 (2011)
19. Possanner, S.: Gyrokinetics from variational averaging: Existence and error bounds. J. Math. Phys. **59**(8), 082702 (2018). https://doi.org/10.1063/1.5018354
20. Qin, H., Tang, W.M., Rewoldt, G.: Gyrokinetic theory for arbitrary wavelength electromagnetic modes in tokamaks. Phys. Plasmas **5**, 1035 (1998). https://doi.org/10.1063/1.872633
21. Scott, B.D.: Gyrokinetic field theory as a Gauge transform or: Gyrokinetic theory without Lie transforms. arXiv:1708.06265 (2017)
22. Sips, A.C.C., For the Steady State Operation, and the Transport Physics topical group Activity: Advanced scenarios for ITER operation. Plasma Phys. Controlled Fusion **47**(5A), A19–A40 (2005). https://doi.org/10.1088/0741-3335/47/5A/003
23. Tronko, N., Chandre, C.: Second-order nonlinear gyrokinetic theory: from the particle to the gyrocenter. J. Plasma Phys. **84**(3):925840301 (2018). https://doi.org/10.1017/S0022377818000430
24. Wootton, A.J., Carreras, B.A., Matsumoto, H., McGuire, K., Peebles, W.A., Ritz, C.P., Terry, P.W., Zweben, S.J.: Fluctuations and anomalous transport in tokamaks. Phys. Fluids B: Plasma Phys. **2**(12), 2879–2903 (1990). https://doi.org/10.1063/1.859358