Springer Proceedings in Mathematics & Statistics

Giambattista Giacomin
Stefano Olla
Ellen Saada
Herbert Spohn
Gabriel Stoltz *Editors*

Stochastic Dynamics Out of Equilibrium

Institut Henri Poincaré, Paris, France, 2017

Springer

Springer Proceedings in Mathematics & Statistics

Volume 282

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Giambattista Giacomin · Stefano Olla ·
Ellen Saada · Herbert Spohn ·
Gabriel Stoltz
Editors

Stochastic Dynamics Out of Equilibrium

Institut Henri Poincaré, Paris, France, 2017

Springer

Editors
Giambattista Giacomin
Laboratoire de Probabilités,
Statistiques et Modélisation (UMR 8001)
Université Paris Diderot
Paris, France

Ellen Saada
CNRS UMR 8145, MAP5
Université Paris Descartes
Paris, France

Gabriel Stoltz
CERMICS
Ecole des Ponts
Marne-La-Vallée, France

Inria
Paris, France

Stefano Olla
CEREMADE, CNRS UMR 7534
Université Paris Dauphine - PSL
Paris, France

Herbert Spohn
Zentrum Mathematik
Technische Universität München
Garching, Bayern, Germany

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-15095-2 ISBN 978-3-030-15096-9 (eBook)
https://doi.org/10.1007/978-3-030-15096-9

Mathematics Subject Classification (2010): 60Gxx, 92Bxx, 82C05, 65C30

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

Preface

In statistical mechanics, it is common practice to use models of large interacting assemblies governed by stochastic dynamics. The trimester "Stochastic Dynamics Out of Equilibrium", held at the Institut Henri Poincaré (IHP) in Paris from April to July 2017, focused on the "out-of-equilibrium" aspect. Indeed, non-reversible dynamics have features which cannot occur at equilibrium and for which novel methods have to be developed. The three domains relevant to this trimester were (i) transport in nonequilibrium statistical mechanics; (ii) the design of more efficient simulation methods; (iii) life sciences.

The trimester at IHP brought together physicists, mathematicians from many domains, computer scientists as well as researchers working at the interface between biology, physics and mathematics. Various events were scheduled during the trimester: a pre-school in Marseille-Luminy, three workshops and several series of courses and seminars; see the website of the trimester

https://indico.math.cnrs.fr/e/stoneq17

for complete information. Each chapter in this book corresponds to one of these events.

Part I gathers lecture notes from the pre-school at the Centre International de Recherche Mathématique (CIRM). This one-week event provided an introduction to the domains listed above. It was intended especially for a junior audience (PhD students and post-docs) but also for more senior researchers not familiar with some of these domains.

Part II includes lecture notes for two of the seven mini-courses which took place during the trimester. Each mini-course was a set of three sessions of one hour and a half, with a first lecture sufficiently introductory to be understood by all the participants of the trimester, and then more specialized sessions. A broad spectrum of scientific fields, topics and techniques was covered by the speakers. Indeed, with a balance depending on the speaker's background, all lectures featured a mix of rigorous mathematical arguments and more physically motivated derivations; they

used, from the mathematical perspective, techniques from analysis, partial differential equations, probability theory and dynamical systems.

Part III corresponds to the workshop "Numerical aspects of nonequilibrium dynamics". The scientific motivation for this event was that many successful approaches for the efficient simulation of equilibrium systems cannot be adapted as such to nonequilibrium dynamics. This is the case for instance for standard variance reduction techniques such as importance sampling or stratification. This three-day workshop (held from Tuesday, April 25 to Thursday, April 27) was focused on the developments of original numerical methods specifically dedicated to the simulation of nonequilibrium systems, as well as their certification in terms of error estimates.

Part IV corresponds to the workshop "Life sciences". This three-day workshop (held from Tuesday, May 16 to Thursday, May 18) gathered researchers coming from different fields—mathematics, physics, life sciences—and working with different approaches and tools, ranging from researchers dealing directly with real data to scientists interested in the theoretical aspects of the models. The aim was on one hand to understand the impact that recent advances in nonequilibrium statistical mechanics and PDE analysis can have on life sciences and, on the other hand, to widen the spectrum of models and phenomenologies tackled by mathematicians and physicists.

Part V corresponds to the workshop "Stochastic dynamics out of equilibrium". This one-week workshop (held from Monday, June 12 to Friday, June 16) was oriented towards general aspects of nonequilibrium stochastic dynamics, with a broad audience. The topics concerned interface dynamics and KPZ universality, nonequilibrium fluctuations, thermal conductivity and superconductivity in one dimension, connection to macroscopic thermodynamics and more.

Let us conclude by acknowledging the various institutions and persons who contributed to the success of the trimester we organized, and who helped us in producing this volume. Let us first thank the staff at the Centre Emile Borel of IHP who was in charge of the administrative aspects of the organization and handled them with a spectacular efficiency. The funding from CNRS (Centre National de la Recherche Scientifique), as well as from IHP and CIRM (through labex CARMIN) were crucial for hosting our visitors. We also benefited from additional fundings from various institutions in Paris (Fondation des Sciences Mathématiques de Paris, Institut des Hautes Etudes Scientifiques, Sorbonne Université, Université Paris Sud, Université Paris Dauphine, Université Paris Descartes, Université Paris Diderot, Inria Paris, etc.) and abroad (Technische Universität München, Italian–French agreement LYSM, Portugal–France agreement), as well as individual grants from French or European funding agencies (ANR COSMOS and LSD from Agence Nationale de la Recherche, projects HyLEF and MsMaths funded by the European

Research Council). Finally, we warmly thank the contributors to this volume and the referees of the contributions, as well as the staff of Springer, in particular Elena Griniari, for helping us in the editorial process.

November 2018

Giambattista Giacomin
Stefano Olla
Ellen Saada
Herbert Spohn
Gabriel Stoltz

Contents

Mini-courses of the Pre-school at CIRM

Stochastic Mean-Field Dynamics and Applications to Life Sciences

Paolo Dai Pra(✉)

Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Padua, Italy
daipra@math.unipd.it

1 Introduction

Although we do not intend to give a general, formal definition, the stochastic mean-field dynamics we present in these notes can be conceived as the random evolution of a system comprised by N interacting components which is: (a) invariant in law for permutation of the components; (b) such that the contribution of each component to the evolution of any other is of order $\frac{1}{N}$. The permutation invariance clearly does not allow any freedom in the choice of the geometry of the interaction; however, this is exactly the feature that makes these models analytically treatable, and therefore attractive for a wide scientific community.
Originally designed as toy models in Statistical Mechanics, the emergence of applications in which the interaction is typically of very long range and not determined by fundamental laws, have renewed the interest in models of this sort. Applications include, in particular, *Life Sciences* and *Social Sciences*.
The goal of these lectures is to

- review some of the basic techniques allowing to derive the macroscopic limit of a mean-field model, and provide quantitative estimates on the rate of convergence;
- illustrate, without technical details, some applications relevant to life sciences, in particular for what concerns the study of the properties of the macroscopic limit.

2 Generalities

2.1 The Prototypical Model

Mainly inspired by [46], we introduce the topic by some heuristics on a simple class of models.
Consider a system of N interacting diffusions on $\mathbb{R}^d$ solving the following system of SDE:

$$dX_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} b(X_t^{i,N}, X_t^{j,N})dt + dW_t^i$$

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 3–27, 2019.
https://doi.org/10.1007/978-3-030-15096-9_1

where $b : \mathbb{R}^d \times \mathbb{R}^d$ is a Lipschitz function, $(W^i)_{i\geq 1}$ are independent standard Brownian motions, and we assume $(X_0^{i,N})_{i=1}^N$ to be i.i.d square integrable random variables. In particular, the dynamical equation is well posed.
Note that, for $t > 0$, the random variables $(X_t^{j,N})_{j=1}^N$ will be, by permutation invariance of the model, identically distributed, but the interaction will break the initial independence. The following heuristics is based on the assumption that a *Law of Large Numbers* for these random variables holds also for $t > 0$. Thus, if we consider the evolution of a single component $X^{i,N}$, and let $N \to +\infty$, it is natural to guess that $X^{i,N}$ converges, as $N \to +\infty$, to a limit process $\overline{X}^i$ solving

$$\begin{aligned} d\overline{X}_t^i &= \int b(\overline{X}_t^i, y) q_t(dy)dt + dW_t^i \\ \overline{X}_0^i &= X_0^i \end{aligned} \tag{2.1}$$

where $q_t = Law(\overline{X}_t^i)$. Once the nontrivial problem of well posedness of this last equation is settled, one aims at showing that, for any given $T > 0$ and indicating by $X_{[0,T]} \in \mathcal{C}([0,T])$ the whole trajectory up to time T, the following statement holds: for any $m \geq 1$

$$(X_{[0,T]}^{1,N}, X_{[0,T]}^{2,N}, \ldots, X_{[0,T]}^{m,N}) \to (\overline{X}_{[0,T]}^1, \overline{X}_{[0,T]}^2, \ldots, \overline{X}_{[0,T]}^m)$$

in distribution as $N \to +\infty$. Note that the components of the process

$$(\overline{X}_{[0,T]}^1, \overline{X}_{[0,T]}^2, \ldots, \overline{X}_{[0,T]}^m)$$

are independent. Thus, independence at time 0 propagates in time, at least in the macroscopic limit $N \to +\infty$. This property is referred to as *propagation of chaos*.

2.2 Propagation of Chaos and Law of Large Numbers

Propagation of chaos can be actually rephrased as a *Law of Large Numbers*. To this aim, given a generic vector $\underline{x} = (x_1, x_2, \ldots, x_N)$, denote by $\rho_N(\underline{x}; dy) := \frac{1}{N}\sum_{i=1}^N \delta_{x_i}(dy)$ the corresponding empirical measure. The propagation of chaos property stated above, is equivalent to the fact that the sequence of empirical measures $\rho_N(\underline{X}_{[0,T]}^N)$ converges in distribution to $Q \in \mathcal{P}(\mathcal{C}([0,T]))$, where $\mathcal{P}(\mathcal{C}([0,T]))$ denotes the set of probabilities on the space of continuous functions $[0,T] \to \mathbb{R}^d$, provided with the topology of weak convergence and Q is the law of the solution of (2.1). This is established in the following result (see also [46], Proposition 2.2).

Proposition 1. *Let $(X^{i,N} : N \geq 1, 1 \leq i \leq N)$ be a triangular array of random variables taking values in a topological space E, such that for each N the law of $(X^{i,N})_{1\leq i\leq N}$ is symmetric (i.e. invariant by permutation of components). Moreover let $(\overline{X}^i)_{i\geq 1}$ be a i.i.d. sequence of E-valued random variables. Then the following statements are equivalent:*

(a) for every $m \geq 1$

$$(X^{1,N}, X^{2,N}, \dots X^{m,N}) \to (\overline{X}^1, \overline{X}^2, \dots, \overline{X}^m)$$

in distribution as $N \to +\infty$*;*

(b) the sequence of empirical measures $\rho_N(\underline{X}^N)$ *converges in distribution to* $Q := Law(\overline{X}^1)$ *as* $N \to +\infty$*.*

Proof. Denote by Q_N the joint law of $(X^{1,N}, X^{2,N}, \dots X^{N,N})$ in E^N, and by $\Pi_m Q_N$ its projection on the first m components, i.e. the law of $(X^{1,N}, X^{2,N}, \dots X^{m,N})$. The statements in (a) is equivalent to: for each $m \geq 1$

$$\Pi_m Q_N \to Q^{\otimes m} \tag{2.2}$$

weakly, where $Q^{\otimes m}$ is the m-fold product of Q.
(a) $\Rightarrow$ *(b).*
To begin with, let $F : E \to \mathbb{R}$ be bounded and continuous. Writing $\langle F, \mu \rangle$ for $\int F d\mu$ and denoting by $\mathbb{E}^{Q_N}$ the expectation w.r.t. Q_N:

$$\begin{aligned}
\mathbb{E}^{Q_N}\left(\langle F, \rho_N(\underline{x}) - Q\rangle^2\right) &= \frac{1}{N^2}\sum_{i,j=1}^{N} \mathbb{E}^{Q_N}\left[F(x_i)F(x_j)\right] \\
&\quad - \frac{2}{N}\langle F, Q\rangle \sum_{i=1}^{N} \mathbb{E}^{Q_N}[F(x_i)] + \langle F, Q\rangle^2 \\
&= \frac{1}{N}\mathbb{E}^{Q_N}[F^2(x_1)] + \frac{N-1}{N}\mathbb{E}^{Q_N}[F(x_1)F(x_2)] \\
&\quad - 2\langle F, Q\rangle \mathbb{E}^{Q_N}[F(x_1)] + \langle F, Q\rangle^2,
\end{aligned}$$

where we have used the symmetry of Q_N. By Assumption (a) this last expression goes to zero as $N \to +\infty$.

Now, let $\Phi : \mathcal{P}(E) \to \mathbb{R}$ be continuous and bounded, where $\mathcal{P}(E)$ is the space of probabilities on the Borel subsets of E, provided with the weak topology. By definition of weak topology, given $\epsilon > 0$ one can find $\delta > 0$ and $F_1, \dots F_k : E \to \mathbb{R}$ bounded and continuous such that if

$$U := \{P \in \mathcal{P}(E) : |\langle P - Q, F_j\rangle| < \delta \text{ for } j = 1, \dots, k\}$$

then $P \in U$ implies $|\Phi(P) - \Phi(Q)| < \epsilon$. Thus

$$\left|\mathbb{E}^{Q_N}[\Phi(\rho_N(\underline{x}))] - \Phi(Q)\right| \leq \epsilon Q_N(\rho_N(\underline{x}) \in U) + \|\Phi\|_\infty Q_N(\rho_N(\underline{x}) \notin U).$$

Therefore, to show (b), i.e. $\left|\mathbb{E}^{Q_N}[\Phi(\rho_N(\underline{x})] - \Phi(Q)\right| \to 0$ for every Φ bounded and continuous, it is enough to show that

$$\lim_{N \to +\infty} Q_N(\rho_N(\underline{x}) \notin U) = 0.$$

But, by what seen above and the Markov inequality,

$$Q_N(\rho_N(\underline{x}) \notin U) \le \sum_{j=1}^{k} Q_N(|\langle \rho_N(\underline{x}) - Q, F_j\rangle| \ge \delta)$$
$$\le \sum_{j=1}^{k} \frac{\mathbb{E}^{Q_N}\left(\langle F_j, \rho_N(\underline{x}) - Q\rangle^2\right)}{\delta^2} \to 0.$$

(b) ⇒ (a).
It is enough to show that if $F_1, F_2, \dots, F_m : E \to \mathbb{R}$ are bounded and continuous, then

$$\mathbb{E}^{Q_N}\left[F_1(x_1) \cdot F_2(x_2) \cdots F_m(x_m)\right] \to \prod_{j=1}^{m} \mathbb{E}^{Q}[F_j(x)] \tag{2.3}$$

Observe that

$$\left|\mathbb{E}^{Q_N}\left[F_1(x_1) \cdot F_2(x_2) \cdots F_m(x_m)\right] - \prod_{j=1}^{m} \mathbb{E}^{Q}[F_j(x)]\right|$$
$$\le \left|\mathbb{E}^{Q_N}\left[F_1(x_1) \cdot F_2(x_2) \cdots F_m(x_m)\right] - \mathbb{E}^{Q_N}\left[\prod_{j=1}^{m} \langle \rho_N(\underline{x}), F_j\rangle\right]\right|$$
$$+ \left|\mathbb{E}^{Q_N}\left[\prod_{j=1}^{m} \langle \rho_N(\underline{x}), F_j\rangle\right] - \prod_{j=1}^{m} \mathbb{E}^{Q}[F_j(x)]\right| \tag{2.4}$$

By (b), the last summand converges to 0. Using symmetry

$$\mathbb{E}^{Q_N}\left[\prod_{j=1}^{m} \langle \rho_N(\underline{x}), F_j\rangle\right] = \frac{1}{N^m}\mathbb{E}^{Q_N}\left[\sum_{\tau:\{1,\dots,m\}\to\{1,\dots,N\}} \prod_{j=1}^{m} F_j(x_{\tau(j)})\right]$$
$$= \frac{D_{N,m}}{N^m}\mathbb{E}^{Q_N}\left[F_1(x_1) \cdot F_2(x_2) \cdots F_m(x_m)\right]$$
$$+ \frac{1}{N^m}\mathbb{E}^{Q_N}\left[\sum_{\tau \text{ not injective}} \prod_{j=1}^{m} F_j(x_{\tau(j)})\right],$$

where $D_{N,m} = \frac{N!}{(N-m)!}$ is the number of injective functions

$$\{1, \dots, m\} \to \{1, \dots, N\}.$$

Since $\frac{D_{N,m}}{N^m} \to 1$, we obtain

$$\mathbb{E}^{Q_N}\left[\prod_{j=1}^{m} \langle \rho_N(\underline{x}), F_j\rangle\right] \to \mathbb{E}^{Q_N}\left[F_1(x_1) \cdot F_2(x_2) \cdots F_m(x_m)\right]$$

which, by (2.4), completes the proof.

Going back to the model in Sect. 2.1, once the propagation of chaos

$$(X^{1,N}_{[0,T]}, X^{2,N}_{[0,T]}, \dots, X^{m,N}_{[0,T]}) \to (\overline{X}^{1}_{[0,T]}, \overline{X}^{2}_{[0,T]}, \dots, \overline{X}^{m}_{[0,T]})$$

is shown, Proposition 1 implies that the empirical measure at time t, $\rho_N(\underline{X}^N_t)$ converges in distribution to $q_t = Law(\overline{X}^1_t)$, for every $t \geq 0$. Moreover, being the law of the solution of (2.1), q_t solves the so-called *McKean-Vlasov equation*

$$\frac{\partial}{\partial t} q_t - \nabla \left[q_t \int b(\cdot, y) q_t(dy) \right] + \frac{1}{2} \Delta q_t = 0.$$

2.3 Symmetry and Empirical Measures

Invariance by permutations of components is the main feature of mean-field dynamics. In practice, for most of the models considered in the literature, permutation invariance is obtained by assuming the characteristics of the dynamics, e.g. the drift for diffusions, to be a function of the empirical measure ρ_N. Next result provides sufficient conditions for a function which is invariant by permutation to be asymptotically a function of the empirical measure. The main assumption is that changing a single component produces variations of order $\frac{1}{N}$ in the value of the function.

Proposition 2. *Let $K \subseteq \mathbb{R}$ be a compact set, and, for $N \geq 1$, $f_N : K^N \to \mathbb{R}$. Assume the following conditions hold:*

(i) the functions f_N are invariant by permutations of components;
(ii) the functions f_N are uniformly bounded, i.e. there is $C > 0$ such that $|f_N(x)| \leq C$ for every $N \geq 1$ and $x \in \mathbb{R}^N$;
(iii) there is a constant $C > 0$ such that for every $N \geq 1$, if $x, y \in \mathbb{R}^N$ and $x_j = y_j$ for all $j \neq i$, then

$$|f_N(x) - f_N(y)| \leq \frac{C}{N} |x_i - y_i|.$$

Then there exists a continuous function $U : \mathcal{P}(K) \to \mathbb{R}$ and an increasing sequence n_k such that

$$\lim_{k \to +\infty} \sup_{x \in K^{n_k}} |f_{n_k}(x) - U(\rho_{n_k}(x))| = 0.$$

Proof. Consider the *Wasserstein metric* on $\mathcal{P}(K)$

$$d(\nu, \nu') := \inf \left\{ \int |x - y| \Pi(dx, dy) : \Pi \text{ has marginals } \nu \text{ and } \nu' \right\}$$

which, by compactness of K, induces the weak topology.

We define the function $U_N : \mathcal{P}(K) \to \mathbb{R}$ by

$$U_N(\mu) := \inf_{x \in K^N} \left[f_N(x) + Cd(\mu, \rho_N(x)) \right],$$

where C is a constant for which assumption (iii) holds. We claim that, for each $y \in K^N$

$$U_N(\rho_N(y)) = f_N(y). \tag{2.5}$$

If not, there would be $x \in K^N$ with

$$f_N(x) + Cd(\rho_N(y), \rho_N(x)) < f_N(y),$$

in particular

$$|f_N(y) - f_N(x)| > Cd(\rho_N(y), \rho_N(x)). \tag{2.6}$$

However a basic result in optimal transport states that

$$d(\rho_N(y), \rho_N(x)) = \inf_{\sigma \in S_N} \frac{1}{N} \sum_{i=1}^{N} |x_i - y_{\sigma(i)}|,$$

where S_N denotes the set of permutations of $\{1, 2, \ldots, N\}$. This, the permutation invariance of f_N and assumption (iii) imply

$$|f_N(y) - f_N(x)| \leq Cd(\rho_N(y), \rho_N(x)),$$

which contradicts (2.6), thus proving (2.5).
Now, let $\mu, \nu \in \mathcal{P}(K)$. By definition of U_N, given $\epsilon > 0$ there is $x \in K^N$ such that

$$U_N(\nu) \geq f_N(x) + Cd(\nu, \rho_N(x)) - \epsilon.$$

Thus

$$\begin{aligned} U_N(\mu) &\leq f_N(x) + Cd(\mu, \rho_N(x)) \leq U_N(\nu) + Cd(\mu, \rho_N(x)) - Cd(\nu, \rho_N(x)) + \epsilon \\ &\leq U_N(\nu) + Cd(\mu, \nu) + \epsilon. \end{aligned}$$

By symmetry this implies that

$$|U_N(\mu) - U_N(\nu)| \leq Cd(\mu, \nu).$$

Therefore, the sequence of functions (U_N) is *equicontinuous* and, clearly, bounded uniformly in N. By the Theorem of Ascoli-Arzelà there is a subsequence converging uniformly to a function U. This, together with Claim 1, completes the proof.

3 Propagation of Chaos for Interacting Systems

3.1 The Microscopic Model

In this section we introduce a wide class of $\mathbb{R}^d$-valued interacting dynamics, which includes the prototypical model above. The main aim is to introduce *quenched disorder*, which accounts for inhomogeneities in the system, and jumps in the dynamics; this allows to include processes with discrete state space. The dynamics is determined by the following characteristics.

- "Local" parameters $(h_i)_{i=1}^N$, drawn independently from a distribution μ on $\mathbb{R}^{d'}$ *with compact support.*
- A *drift* $b(x_i, h_i; \rho_N(\underline{x}, \underline{h}))$, where

$$\rho_N(\underline{x}, \underline{h}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{(x_i, h_i)},$$

and

$$b : \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'}) \to \mathbb{R}^d.$$

- A diffusion coefficient $\sigma(x_i, h_i; \rho_N(\underline{x}, \underline{h}))$

$$\sigma : \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'}) \to \mathbb{R}^{d \times n},$$

where n is the dimension of the driving Brownian Motion.
- A jump rate $\lambda(x_i, h_i; \rho_N(\underline{x}, \underline{h}))$ with

$$\lambda : \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'}) \to [0, +\infty).$$

- A distribution for the jump $f(x_i, h_i; \rho_N(\underline{x}, \underline{h}); v)\alpha(dv)$ with

$$f : \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'}) \times [0, 1] \to \mathbb{R}^d$$

and $\alpha(dv)$ is a probability on $[0, 1]$.

The dynamics could be introduced via generator and semigroup, but it will be convenient to use the language of Stochastic Differential Equations (SDE). So let $(W^i)_{i \geq 1}$ be a i.i.d. sequence of n-dimensional Brownian motions; moreover let $(N^i(dt, du, dv))_{i \geq 1}$ be i.i.d. Poisson random measures on $[0, +\infty) \times [0, +\infty) \times [0, 1]$ with characteristic measure $dt \otimes du \otimes \alpha(dv)$. The microscopic model is given as solution of the SDE for *every given realization of the local parameters* (h_i):

$$\begin{aligned} X_t^{i,N} &= X_0^i + \int_0^t b\left(X_s^{i,N}, h_i, \rho(\underline{X}_s^N, \underline{h})\right) ds + \int_0^t \sigma\left(X_s^{i,N}, h_i, \rho(\underline{X}_s^N, \underline{h})\right) dW_s^i \\ &+ \int_{[0,t] \times [0,+\infty) \times [0,1]} f\left(X_{s^-}^{i,N}, h_i; \rho_N(\underline{X}_{s^-}^N, \underline{h}); \alpha\right) \mathbf{1}_{\left[0, l\left(X_{s^-}^{i,N}, h_i, \rho(\underline{X}_{s^-}^N, \underline{h})\right)\right]}(u) N^i(ds, du, dv) \end{aligned} \tag{3.1}$$

It will be assumed, without further notice, that the initial states X_0^i are i.i.d., square integrable, independent of both the local parameters (h_i) and of the driving noises (W^i, N^i).

3.2 The Macroscopic Limit

At heuristic level it is not hard to identify the limit of a given component $X^{i,N}$ of (3.1) subject to a local field h. We omit the apex i on the process and of the driving noises

$$\begin{aligned}\overline{X}_t(h) =& \overline{X}_0 + \int_0^t b\left(\overline{X}_s(h), h, r_s\right) ds + \int_0^t \sigma\left(\overline{X}_s(h), h, r_s\right) dW_s \\ &+ \int_{[0,t]\times[0,+\infty)\times[0,1]} f\left(\overline{X}_{s^-}(h), h; r_s; \alpha\right) \mathbf{1}_{[0,\lambda\left(\overline{X}_{s^-}(h), h, r_s\right)]}(u) N(ds, du, dv)\end{aligned} \tag{3.2}$$

where $r_s = Law(\overline{X}_s(h)) \otimes \mu(dh)$. Choosing $\overline{X}_0 = X_0^i$, and driving noises W^i, N^i, we indicate by $\overline{X}^i$ the corresponding solution (3.2).

3.3 Well Posedness of the Microscopic Model: Lipschitz Conditions

We now give conditions that guarantee well posedness of (3.1) and (3.2); they are far from being optimal, but allow a reasonable economy of notations. Weaker conditions can be found, for instance in [1]. It is useful to work with probability measures possessing mean value:

$$\mathcal{P}^1(\mathbb{R}^d) := \left\{ \nu \in \mathcal{P}(\mathbb{R}^d) : \int |x| \nu(dx) < +\infty \right\}$$

which is provided with the *Wasserstein metric*

$$d(\nu, \nu') := \inf \left\{ \int |x - y| \Pi(dx, dy) : \Pi \text{ has marginals } \nu \text{ and } \nu' \right\}.$$

- [**L1**] The function $b(x, h, r)$ and $\sigma(x, h, r)$, defined in $\mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}^1(\mathbb{R}^d \times \mathbb{R}^{d'})$ are continuous, and globally Lipschitz in (x, r) uniformly in h.
- [**L2**] The Lipschitz condition of the jumps is slightly less obvious. We assume $f : \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}^1(\mathbb{R}^d \times \mathbb{R}^{d'}) \times [0,1] \to \mathbb{R}^d$ and $\lambda : \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^{d'}) \to [0, +\infty)$ are continuous, and obey the following condition

$$\begin{aligned}\int \left| f(x, h, r, v) \mathbf{1}_{[0,\lambda(x,h,r)]}(u) - f(y, h, r', v) \mathbf{1}_{[0,\lambda(x,h,r)]}(u) \right| du\alpha(dv) \\ \leq L\left[|x - x'| + d(r, r')\right]\end{aligned} \tag{3.3}$$

  for all x, y, r, r', h.

Remark 1. The above assumptions imply that when one replaces r by the empirical measure $\rho_N(\underline{x}, \underline{h})$, one recovers a Lipschitz condition in $\underline{x}$. For instance, the function $b(x_i, h_i; \rho_N(\underline{x}, \underline{h}))$ is globally Lipschitz in $\underline{x}$ uniformly in $\underline{h}$.

Remark 2. Continuity, global Lipschitzianity and compactness of the support of μ imply the linear growth conditions

$$\begin{aligned}|b(x, h, r)| &\leq C\left[1 + |x| + \int |y| r(dy, dh)\right] \\ |\sigma(x, h, r)| &\leq C\left[1 + |x| + \int |y| r(dy, dh)\right] \\ \int |f(x, h, r, v)| \lambda(x, h, r) \alpha(dv) &\leq C\left[1 + |x| + \int |y| r(dy, dh)\right].\end{aligned} \tag{3.4}$$

Remark 3. Condition **L2** is satisfied if both f and λ are continuous, bounded and globally Lipschitz in x, r uniformly of the other variables. In the case f does not depend on x, r but on h, v only, unbounded Lipschitz jump rate λ can be afforded.

Using Remark 1, together with standard methods in stochastic analysis, one obtains the following result. A detailed proof can be found e.g. in [30].

Proposition 3. *Under* **L1** *and* **L2**, *the system* (3.1) *admits a unique strong solution.*

3.4 Well Posedness of the Macroscopic Limit

The proof of the convergence of one component of (3.1) toward a solution of (3.2) allows two alternative strategies. One consists in: (a) showing tightness of the sequence of microscopic processes; (b) showing that any limit point solves weakly (3.2); (c) showing that for (3.2) uniqueness in law holds true. We rather follow the following approach, which is somewhat simpler and allows for quantitative error estimates: (a) we show that (3.2) is well posed; (b) by a coupling argument we show L^1-convergence of one component of (3.1) to a solution of (3.2) driven by the *same noise*.

Proposition 4. *Under* **L1** *and* **L2**, *the system* (3.2) *admits a unique strong solution.*

Proof. We sketch the proof of existence. We use a standard Picard iteration. Define $X_t^{(0)}(h) \equiv \overline{X}_0$ and

$$\begin{aligned} X_t^{(k+1)}(h) \;&=\; \overline{X}_0 + \int_0^t b\left(X_s^{(k)}(h), h, r_s^{(k)}\right) ds + \int_0^t \sigma\left(X_s^{(k)}(h), h, r_s^{(k)}\right) dW_s \\ &+ \int_{[0,t]\times[0,+\infty)\times[0,1]} f\left(X_{s^-}^{(k)}(h), h; r_s^{(k)}; \alpha\right) \mathbf{1}_{\left[0,\lambda\left(X_{s^-}^{(k)}(h), h, r_s^{(k)}\right)\right]}(u) N(ds, du, dv) \end{aligned} \tag{3.5}$$

where

$$r_s^{(k)} = Law\left(X_s^{(k)}(h)\right) \otimes \mu(dh).$$

We estimate

$$E_T^{(k)} := \int \mathbb{E}\left[\sup_{t\in[0,T]} \left|X_t^{(k+1)}(h) - X_t^{(k)}(h)\right|\right] \mu(dh). \tag{3.6}$$

If we use (3.5) and subtract the equations for $X^{(k+1)}$ and $X^{(k)}$, take the $\sup_{t\in[0,T]}$ and use the triangular inequality, we obtain the sum of three terms.

(A) The first term comes from the drift.

$$\sup_{t\in[0,T]}\left|\int_0^t b\left(X_s^{(k)}(h),h,r_s^{(k)}\right)ds-\int_0^t b\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)ds\right|$$
$$\leq\int_0^T\left|b\left(X_s^{(k)}(h),h,r_s^{(k)}\right)-b\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)\right|ds$$
$$\leq L\int_0^T\left(\left|X_s^{(k)}(h)-X_s^{(k-1)}(h)\right|+d(r_s^{(k)},r_s^{(k-1)})\right)ds$$
$$\leq L\int_0^T\left(\left|X_s^{(k)}(h)-X_s^{(k-1)}(h)\right|+\int\mathbb{E}\left|X_s^{(k)}(h')-X_s^{(k-1)}(h')\right|\mu(dh')\right)ds$$

where the inequality

$$d(r_s^{(k)},r_s^{(k-1)})\leq\int\mathbb{E}\left|X_s^{(k)}(h')-X_s^{(k-1)}(h')\right|\mu(dh') \tag{3.7}$$

comes directly form the definition of the metric d, and we have used (**L1**). Averaging:

$$\int\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t b\left(X_s^{(k)}(h),h,r_s^{(k)}\right)ds-\int_0^t b\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)ds\right|\right]\mu(dh)$$
$$\leq 2L\int_0^T\int\mathbb{E}\left|X_s^{(k)}(h)-X_s^{(k-1)}(h)\right|\mu(dh)\leq 2LTE_T^{(k-1)}.$$

(B) The second term comes from the diffusion coefficient.

$$\sup_{t\in[0,T]}\left|\int_0^t\sigma\left(X_s^{(k)}(h),h,r_s^{(k)}\right)ds-\int_0^t\sigma\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)dW_s\right|.$$

By the L^1 Burkholder-Davis-Gundy inequality (see e.g. [42])

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t\left[\sigma\left(X_s^{(k)}(h),h,r_s^{(k)}\right)-\sigma\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)\right]dW_s\right|\right]$$
$$\leq C\mathbb{E}\left[\left(\int_0^T\left|\sigma\left(X_s^{(k)}(h),h,r_s^{(k)}\right)-\sigma\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)\right|^2ds\right)^{\frac{1}{2}}\right]$$
$$\leq CL\mathbb{E}\left[\left(\int_0^T\left(\left|X_s^{(k)}(h)-X_s^{(k-1)}(h)\right|+d(r_s^{(k)},r_s^{(k-1)})\right)^2ds\right)^{\frac{1}{2}}\right]$$
$$\leq CL\sqrt{T}\mathbb{E}\left[\sup_{s\in[0,T]}\left(\left|X_s^{(k)}(h)-X_s^{(k-1)}(h)\right|+d(r_s^{(k)},r_s^{(k-1)})\right)ds\right]$$

Averaging over h and using (3.7) as before, we obtain

$$\int \mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t \left[\sigma\left(X_s^{(k)}(h),h,r_s^{(k)}\right)-\sigma\left(X_s^{(k-1)}(h),h,r_s^{(k-1)}\right)\right]dW_s\right|\right]\mu(dh) \leq 2CL\sqrt{T}E_T^{(k-1)}.$$

(C) Finally, we have the term coming from the jumps.

$$\sup_{t\in[0,T]}\left|\int_{[0,t]\times[0,+\infty)\times[0,1]} f\left(X_{s^-}^{(k)}(h),h;r_s^{(k)};v\right)\mathbf{1}_{\left[0,\lambda\left(X_{s^-}^{(k)}(h),h,r_s^{(k)}\right)\right]}(u)N(ds,du,dv) - \int_{[0,t]\times[0,+\infty)\times[0,1]} f\left(X_{s^-}^{(k-1)}(h),h;r_s^{(k-1)};v\right)\mathbf{1}_{\left[0,\lambda\left(X_{s^-}^{(k-1)}(h),h,r_s^{(k-1)}\right)\right]}(u)N(ds,du,dv)\right| \tag{3.8}$$

Let

$$F_s^k := f\left(X_{s^-}^{(k)}(h),h;r_s^{(k)};v\right)\mathbf{1}_{\left[0,\lambda\left(X_{s^-}^{(k)}(h),h,r_s^{(k)}\right)\right]}(u).$$

Since N is a positive measure, (3.8) is bounded above by,

$$\int_0^T |F_s^k - F_s^{k-1}|N(ds,du,dv) = \int_0^T |F_s^k - F_s^{k-1}|dsdu\alpha(dv) + \int_0^T |F_s^k - F_s^{k-1}|\tilde{N}(ds,du,dv), \tag{3.9}$$

where $\int_0^T |F_s^k - F_s^{k-1}|\tilde{N}(ds,du,dv)$ has mean zero, since $dsdu\alpha(dv)$ is the compensator of $N(ds,du,dv)$. Thus averaging, we are only left with the term $\int_0^T |F_s^k - F_s^{k-1}|dsdu\alpha(dv)$, which is dealt with using **(L2)**, and gives an upper bound similar of that of part **(A)**.

Summing up the contributions of **(A)**, **(B)** and **(C)**, we get, for a sufficiently large constant C,

$$E_T^{(k)} \leq C(T+\sqrt{T})E_T^{(k-1)}.$$

We now observe that the processes $X^{(k)}$, $k \geq 0$, $h \in \mathbb{R}^{d'}$ are progressively measurable for the filtration generated by the initial condition and the driving noise W, N, and satisfy

$$\int \mathbb{E}\left[\sup_{t\in[0,T]}\left|X_t^{(k)}(h)\right|\right]\mu(dh) < +\infty.$$

This can be seen by induction on k, replicating the steps above but using, rather than the Lipschitz conditions, the linear growth conditions (3.4). If we denote by $\mathcal{M}$ the space of progressively measurable, *cadlag*, $\mathbb{R}^d$ valued processes such that

$$\|X\| := \mathbb{E}\left[\sup_{t\in[0,T]} |X_t|\right] < +\infty,$$

and we take T sufficiently small, we have shown that

$$\sum_k \int \|X^{(k+1)}(h) - X^{(k)}(h)\|\mu(dh) < +\infty,$$

and therefore for all h in a set F of μ-full measure

$$\sum_k \|X^{(k+1)}(h) - X^{(k)}(h)\| < +\infty.$$

The norm $\|\cdot\|$ is not complete in $\mathcal{M}$, as the sup-norm is not complete in the space of cadlag functions. To get a complete metric, we replace the distance in sup-norm by the Skorohod distance d_S (see [5]), i.e.

$$D_S(X,Y) := \mathbb{E}\left[d_S(X,Y)\right].$$

Since the Skorohod distance is dominated by the distance in sup-norm, a Cauchy sequence for $\|\cdot\|$ is also Cauchy for the metric D_S. Thus, the limit $\overline{X}(h)$ of the sequence $X^{(k)}(h)$ can be defined for all $h \in F$, where F is a set of measure one for μ, and it is not hard to show (using also Proposition 1) that (3.2) holds for the limit. $\overline{X}(h)$ can be then easily defined for $h \notin F$ just by imposing that (3.2) holds.

This establishes existence of solution in $\mathcal{M}$ for T small. Since the condition on T does not involve the initial condition, the argument can be iterated on adjacent time intervals, obtaining a solution on any time interval.

Establishing uniqueness would actually be easy by using similar arguments. For us it is not actually needed, as uniqueness will follow from the convergence result in next section (Theorem 1).

Remark 4. It is more customary to use L^2 norms rather that L^1 norms for constructing solutions to SDE. The main difference is in **(C)**, where we estimate (3.8). When estimating the mean of the *square* of (3.9), the martingale contributes with

$$\int_0^T |F_s^k - F_s^{k-1}|^2 ds du\alpha(dv).$$

To complete the argument one needs a Lipschitz condition of the form

$$\int \left|f(x,h,r,v)\mathbf{1}_{[0,\lambda(x,h,r)]}(u) - f(y,h,r',v)\mathbf{1}_{[0,\lambda(y,h,r')]}(u)\right|^2 du\alpha(dv)$$
$$\leq L\left[|x-x'|^2 + d_2^2(r,r')\right], \quad (3.10)$$

where, in the whole argument, the distance

$$d_2(\nu,\nu') := \left(\inf\{\int |x-y|^2 \Pi(dx,dy) : \Pi \text{ has marginals } \nu \text{ and } \nu'\}\right)^{\frac{1}{2}}$$

would be used. The Lipschitz condition (3.10) is harder to check than (3.3), for the simple reason that "squaring an indicator function does not produce any square".

3.5 Propagation of Chaos

Theorem 1. *Suppose conditions* **L1** *and* **L2** *hold. For $i \geq 1$ denote by $\overline{X}^i(h)$ the solution of (3.2) with the local parameter h and the same initial condition X_0^i of (3.1). Then for each i and $T > 0$*

$$\lim_{N\to+\infty} \int \mathbb{E}\left[\sup_{t\in[0,T]} \left|X_t^{i,N} - \overline{X}_t^i(h_i)\right|\right] \mu^{\otimes N}(d\underline{h}) = 0$$

where $\mu^{\otimes N}$ is the N-fold product of μ.

Proof. As in the proof of Proposition 4 we subtract the two equations for $X^{i,N}$ and $\overline{X}^i$. Using the triangular inequality, we estimate $\sup_{t\in[0,T]} \left|X_t^{i,N} - \overline{X}_t^i(h_i)\right|$ as sum of three terms, corresponding respectively to drift, diffusion and jumps. In this proof we only show how to deal with the drift term. The other two terms, involving stochastic integrals, are reduced to terms with Lebesgue time integrals as in the proof of Proposition 4, and then are estimated as the drift term.

We therefore give estimates for

$$\begin{aligned}
&\int \mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t b\left(X_s^{i,N}, h_i, \rho(\underline{X}_s^N, \underline{h})\right) ds - \int_0^t b\left(\overline{X}_s^i(h_i), h_i, r_s\right) ds\right|\right] \mu^{\otimes N}(d\underline{h}) \\
&\quad \leq \int \mathbb{E}\left[\int_0^T \left|b\left(X_s^{i,N}, h_i, \rho(\underline{X}_s^N, \underline{h})\right) - b\left(\overline{X}_s^i(h_i), h_i, r_s\right)\right|\right] \mu^{\otimes N}(d\underline{h}) \qquad (3.11)
\end{aligned}$$

By **(L1)**

$$\begin{aligned}
&\left|b\left(X_s^{i,N}, h_i, \rho(\underline{X}_s^N, \underline{h})\right) - b\left(\overline{X}_s^i(h_i), h_i, r_s\right)\right| \\
&\qquad \leq L\left[\left|X_s^{i,N} - \overline{X}_s^i(h_i)\right| + d\left(\rho(\underline{X}_s^N, \underline{h}), r_s\right)\right]. \qquad (3.12)
\end{aligned}$$

Now,

$$d\left(\rho(\underline{X}_s^N, \underline{h}), r_s\right) \leq d\left(\rho(\underline{X}_s^N, \underline{h}), \rho(\underline{\overline{X}}_s, \underline{h})\right) + d\left(\rho(\underline{\overline{X}}_s, \underline{h}), r_s\right). \qquad (3.13)$$

We consider the two summands in the r.h.s. of (3.13) separately. By definition of the metric $d(\cdot,\cdot)$

$$d\left(\rho(\underline{X}_s^N, \underline{h}), \rho(\underline{\overline{X}}_s, \underline{h})\right) \leq \frac{1}{N}\sum_{j=1}^N \left|X_s^{j,N} - \overline{X}_s^j\right|,$$

so, by symmetry,

$$\int \mathbb{E}\left[d\left(\rho(\underline{X}_s^N, \underline{h}), \rho(\overline{\underline{X}}_s, \underline{h})\right)\right] \mu^{\otimes N}(d\underline{h}) \leq \int \mathbb{E}\left[\left|X_s^{i,N} - \overline{X}_s^i(h_i)\right|\right] \mu^{\otimes N}(d\underline{h}). \tag{3.14}$$

For the second summand in (3.13) we observe that, under $\mathbb{P} \otimes \mu^{\otimes \infty}$, the random variables $\left(\overline{X}_s^i(h_i), h_i\right)$ are i.i.d. with law $r_s \in \mathcal{P}(\mathbb{R}^{d+d'})$. By a recent version of the Law of Large Number ([27], Theorem 1), there exists a constant $C > 0$, only depending on d and d', and $\gamma > 0$ (any $\gamma < \frac{1}{d+d'}$ does the job) such that

$$\int \mathbb{E}\left[d\left(\rho(\overline{\underline{X}}_s, \underline{h}), r_s\right)\right] \mu^{\otimes N}(d\underline{h}) \leq \frac{C}{N^\gamma}. \tag{3.15}$$

Inserting what obtained in (3.12), (3.13) and (3.14) in (3.11) we get for some $C > 0$, which may also depend on T,

$$\int \mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t b\left(X_s^{i,N}, h_i, \rho(\underline{X}_s^N, \underline{h})\right) ds - \int_0^t b\left(\overline{X}_s^i(h_i), h_i, r_s\right) ds\right|\right] \mu^{\otimes N}(d\underline{h})$$
$$\leq C \int \mathbb{E}\left[\int_0^T \left|X_s^{i,N} - \overline{X}_s^i(h_i)\right|\right] \mu^{\otimes N}(d\underline{h}) + \frac{C}{N^\gamma}.$$

Dealing similarly with all terms arising in $\sup_{t\in[0,T]} \left|X_t^{i,N} - \overline{X}_t^i(h_i)\right|$, if we set

$$E_t := \int \mathbb{E}\left[\sup_{s\in[0,t]} \left|X_s^{i,N} - \overline{X}_s^i(h_i)\right|\right] \mu^{\otimes N}(d\underline{h})$$

we obtain

$$E_t \leq C \int_0^t E_s ds + \frac{C}{N^\gamma},$$

which, by Gromwall's Lemma and the fact that $E_0 = 0$ yields

$$E_T \leq \frac{C_T}{N^\gamma}$$

for some T-dependent constant C_T, and this complete the proof.

4 Applications

In this section we review some classes of models that are relevant for life sciences. Some key results will be stated, but no proofs are given.

4.1 The Stochastic Kuramoto Model

Synchronization phenomena leading to macroscopic rhythms are ubiquitous in science. Most (ab)used examples include

- applauses;
- flashing fireflies;
- protein concentration within cells in a multicellular system (reprissilators).

In these examples the systems are comprised by many units, each unit tending to behave periodically. Under circumstances depending on how units communicate, oscillation may *synchronize*, producing macroscopic pulsing. The (stochastic) Kuramoto model [33] is perhaps the most celebrated stylized model to capture this behavior.

In the Kuramoto model units are *rotators*, i.e. the state variable is an angle. Denoting by $X^{i,N}$ the angular variable (*phase*) of the i-th rotator, with $i = 1, 2, \dots, N$, the evolution is given by

$$dX_t^{i,N} = h_i dt + \frac{\theta}{N} \sum_{j=1}^{N} \sin\left(X_t^{j,N} - X_t^{i,N}\right) dt + dW_t^i. \tag{4.1}$$

Here h_i is the characteristic angular velocity of the i-th rotator. The effect of the interaction term is to favor phases to stay close. We assume the h_i's are i.i.d., drawn from a distribution μ on $\mathbb{R}$ with compact support. By possibly adding a constant speed rotation, there is no further loss of generality to assume that μ has mean zero. We further assume μ is symmetric, i.e. invariant by reflection around zero.

Clearly all results in Sect. 3 apply, and we get the following macroscopic limit:

$$d\overline{X}_t(h) = h dt + \theta \int \sin(y - \overline{X}_t) q_t(dy; h') \mu(dh') dt + dW_t, \tag{4.2}$$

where $q_t(dy; h')$ is the law of $\overline{X}_t(h')$. The flow of measures $q_t(\,\cdot\,, h)$ solves (indeed in the classical sense for the density w.r.t. the Lebesgue measure)

$$\frac{\partial}{\partial t} q_t(x; h) = \frac{1}{2} \frac{\partial^2}{\partial x^2} q_t(x; h) - \frac{\partial}{\partial x} \left[(h + \theta r_{q_t} \sin(\varphi_{q_t} - x)) \, q_t(x, h) \right] =: \mathcal{M}[q_t](h), \tag{4.3}$$

where

$$r_{q_t} e^{i\varphi_{q_t}} := \int e^{ix} q_t(dx; h) \mu(dh).$$

Equation (4.3) describes the collective behavior of the system of rotators. r_{q_t} captures the degree of synchronization of the system: $r_{q_t} = 0$ indicates total lack of synchronization, while a perfectly synchronized systems has $r_{q_t} = 1$.

One is interested in the long time behavior of solutions of (4.3), in particular stable equilibria. Note that, since the model is rotation invariant, if $q(x; h)$ solves $\mathcal{M}[q] = 0$, then also $q(x + x_0; h)$ does; thus there is no loss of generality in looking for equilibria satisfying $\varphi_q = 0$.

The proof of the following statement can be found in [7].

Theorem 2. *q^* is a solution of $\mathcal{M}[q] = 0$ with $\varphi_{q^*} = 0$ if and only if it is of the form*

$$q^*(x;h) = (Z_*)^{-1} \cdot e^{2(hx+\theta r_* \cos x)} \left[e^{4\pi h} \int_0^{2\pi} e^{-2(hx+\theta r_* \cos x)} dx \right. \\ \left. + (1 - e^{4\pi h}) \int_0^x e^{-2(hy+\theta r_* \cos y)} dy \right], \quad (4.4)$$

where Z_* is a normalization factor and r_* satisfies the consistency relation

$$r_* = \int e^{ix} \, q_*(x,h) \, \mu(dh) \, dx. \quad (4.5)$$

$r_* = 0$ is a solution of (4.5), and it corresponds to the incoherent *solution*

$$q^*(x;h) \equiv \frac{1}{2\pi},$$

i.e. the phases of the rotators are uniformly distributed on the torus.

Linear stability of the incoherent solution depends in a highly nontrivial way on θ and on the distribution μ of the local parameters. It is rather well understood in some special cases [7,8,20].

Theorem 3. *Denote by*

$$\theta_c = \left[\int \frac{\mu(dh)}{1 + 4h^2} \right]^{-1}. \quad (4.6)$$

(a) Suppose μ is unimodal, *i.e. it has a (even) density decreasing on $(0, +\infty)$. Then the incoherent solution is linearly stable if and only if $\theta < \theta_c$. At θ_c one (circle of)* synchronized *solution (i.e. with $r_q > 0$ bifurcates for the incoherent solution.*

(b) Suppose $\mu = \frac{1}{2}(\delta_{-h_0} + \delta_{h_0})$ for some $h_0 > 0$. Then the incoherent solution is linearly stable if and only if $\theta < \theta_c \wedge 2$. For $\theta_c < 2$ at $\theta = \theta_c$ one (circle of) synchronized *solution (i.e. with $r_q > 0$) bifurcates. For $\theta_c > 2$ (which occurs for h_0 sufficiently large), at $\theta = 2$ the incoherent solution loses stability via a Hopf bifurcation: it is believed, but not rigorously proved, that stable time-periodic solutions emerge.*

It is not true in general that when the incoherent solution is stable then it is unique. It is believed it is so in the unimodal case, but proved either for θ small, or up to the critical point if μ is sufficiently concentrated around zero [37]. In the binary case, for certain values of the parameters it is known that there are values of θ smaller than the critical value for which *two distinct* circles of synchronized solutions exists [37].

In general, when the support of μ is contained in a sufficiently small interval, then synchronized solutions exist if and only if $\theta > \theta_c$, are unique up to rotation, and are linearly stable [4,28].

4.2 Interacting Fitzhugh-Nagumo Neurons

Designed as reduction of more realistic models (e.g. the Hodgkin-Huxley model), the Fitzhugh-Nagumo model describes the evolution of the membrane potential x_t of a neuron through the following differential equation

$$\begin{aligned} \dot{x}_t &= x_t - \frac{1}{3}x_t^3 + y_t + I_t^{ext} \\ \dot{y}_t &= \epsilon(a + bx_t - \gamma y_t) \end{aligned} \tag{4.7}$$

where

- y_t is a *recovery variable* obtained by reduction of other variables;
- I_t^{ext} is the input current, assumed to be random and stationary. Without loss of generality, choosing a properly, we can assume I_t^{ext} has mean zero.
- b is the interaction strength between x and y, $\gamma \geq 0$ is a dissipation parameter, and a is a kinetic parameter related with input current and synaptic conductance.

The parameter ϵ can be used to separate the time scales of the evolutions of the two variables. In what follows we assume $dI_t^{ext} = \sigma dW_t$ for a Brownian motion W.

To begin with, consider the equation in absence of randomness in the input current ($\sigma = 0$), and set $b = -1$, $\gamma = 0$ to make the analysis simpler. In this case (4.7) has a unique equilibrium in $(a, -a + a^3/3)$, which is globally stable for $|a| < 1$, is has a Hopf bifurcation at $|a| = 1$ and a stable periodic orbit emerges for $|a| > 1$. Thus, the system can be excited by the input, producing, at least for appropriate choice of the parameters, rapid variations of the potential (*spikes*) which occur periodically.

There are various ways to make several neurons interact in a network, even within the mean-field scheme, depending of how we model synapsis (see [2]). The simplest, corresponding to electrical synapsis, leads to the following system. Here $X^{i,N}$ denotes the membrane potential of the i-th neuron. The local parameter h_i may be interpreted as the *macroscopic location* of the neuron, or its *type*.

$$\begin{aligned} dX_t^{i,N} &= \left(X_t^{i,N} - \frac{1}{3}(X_t^{i,N})^3 + Y_t^{i,N} \right) dt \\ &\quad + \frac{1}{N}\sum_{j=1}^{N} J(h_i, h_j) \left(X_t^{i,N} - X_t^{j,N} \right) dt + \sigma dW_t^i \\ dY_t^{i,N} &= \epsilon(h_i) \left[a(h_i) + b(h_i) X_t^{i,N} - \gamma(h_i) Y_t^{i,N} \right] dt, \end{aligned} \tag{4.8}$$

where the coupling parameters $J(h_i, h_j)$ tune the interaction between pairs of neurons.

The model exhibits a richer behavior if one introduces a delay τ in the transmission of informations between different neurons:

$$\begin{aligned} dX_t^{i,N} &= \left(X_t^{i,N} - \frac{1}{3}(X_t^{i,N})^3 + Y_t^{i,N} \right) dt \\ &\quad + \frac{1}{N} \sum_{j=1}^{N} J(h_i, h_j) \left(X_t^{i,N} - X_{t-\tau(h_i,h_j)}^{j,N} \right) dt + \sigma dW_t^i \\ dY_t^{i,N} &= \epsilon(h_i) \left[a(h_i) + b(h_i) X_t^{i,N} - \gamma(h_i) Y_t^{i,N} \right] dt. \end{aligned} \tag{4.9}$$

Delay makes a bit more painful the well posedness analysis for both the model and its macroscopic limit, but for propagation of chaos the same proof carries through (see [48] for details), giving the following macroscopic limit

$$\begin{aligned} d\overline{X}_t(h) &= \left(\overline{X}_t(h) - \frac{1}{3}\overline{X}_t^3(h) + \overline{Y}_t(h) \right) dt \\ &\quad + \int J(h, h') \left(\overline{X}_t(h) - y \right) q_{t-\tau(h,h')}(dy; h') \mu(dh') dt + \sigma dW_t \\ d\overline{Y}_t(h) &= \epsilon(h)(a(h) + b(h)\overline{X}_t(h) - \gamma(h)\overline{Y}_t(h))dt, \end{aligned} \tag{4.10}$$

where $q_t(dx; h)$ denotes the law of $\overline{X}_t(h)$. Not much is known at this level of generality, so we consider the simplest, homogeneous case in which h is constant, $\gamma = 0$, $b = -1$ which gives

$$\begin{aligned} d\overline{X}_t &= \left[\overline{X}_t - \frac{1}{3}\overline{X}_t^3 + \overline{Y}_t + J(\overline{X}_t - \mathbb{E}(\overline{X}_{t-\tau})) \right] dt + \sigma dW_t \\ d\overline{Y}_t &= \epsilon(a - \overline{X}_t)dt \end{aligned} \tag{4.11}$$

A further simplification consists in letting the noise go to zero, in both the diffusion and the initial condition. We obtain the deterministic system with delay

$$\begin{aligned} \dot{x}_t &= x_t - \frac{1}{3}x_t^3 + y_t + J(x_t - x_{t-\tau}) \\ \dot{y} &= \epsilon(a - x_t). \end{aligned} \tag{4.12}$$

This system has been extensively studied in [32]. Here we assume $J \geq 0$

- The point $(a, -a + a^3/3)$ is still the unique fixed point, and it is stable for $|a| > \sqrt{1 + 2J}$ and unstable for $|a| < 1$, no matter what τ is.
- For $1 < |a| < \sqrt{1 + 2J}$ loss of stability via a Hopf bifurcation can be obtained by increasing τ: *interaction and transmission delay may produce oscillations even if single neurons are in the stability region.*

Does noise play any role in exciting the neuronal network?

This question has only partial answers (see e.g. [35,39,41,44]). Consider the simplified system (4.11) and remove the delay.

$$\begin{aligned} d\overline{X}_t &= \left[\overline{X}_t - \frac{1}{3}\overline{X}_t^3 + \overline{Y}_t + J(\overline{X}_t - \mathbb{E}(\overline{X}_t))\right] dt + \sigma dW_t \\ d\overline{Y}_t &= \epsilon(a - \overline{X}_t)dt \end{aligned} \tag{4.13}$$

Some indications on the behavior of this system, confirmed by numerical simulations, are obtained via the following heuristic argument. For a similar model details can be found in [18]

- Writing down the equation for the moments of $(\overline{X}_t, \overline{Y}_t)$ and *pretending* the system is Gaussian, we get at formal level a closed equation for the means and the covariance matrix.
- This equation corresponds to a truly Gaussian process $(\tilde{X}, \tilde{Y})$, which can be shown to be a good approximation of $(\overline{X}, \overline{Y})$ for σ small.

The evolution of the law of $(\tilde{X}, \tilde{Y})$ can be studied at least locally around the fixed point. It can be shown that for $|a| > 1$ but sufficiently close to 1, periodic solutions for the law of $(\tilde{X}_t, \tilde{Y}_t)$ emerge for *moderate* values of σ, i.e. within some interval $0 < \sigma_0 < \sigma < \sigma_1$: we therefore obtain *noise-induced* oscillations. It should be remarked noise-induced oscillations were pointed out in similar Gaussian models long time ago [45].

4.3 Interacting Hawkes Processes

The Fitzhugh-Nagumo model exhibits some qualitative features of neuronal dynamics, in particular excitability. Periodicity of spikes for a single neuron is however unrealistic: spike trains are more effectively modeled by point processes. An appropriate model in this context is obtained by using Hawkes processes [14–16].

Let $Z_t^{i,N}$ be the counting process that counts the spikes of neuron i, having local parameter (position, type...) h_i. It is assumed that $Z_t^{i,N}$ jumps with a *rate* $\lambda_i^N(t)$ of the form

$$\lambda_i^N(t) = f\left(h_i; \frac{1}{N}\sum_{j=1}^{N} J(h_i, h_j) \int_{[0,t]} k(t-s) dZ_s^{j,N}\right)$$

where $f(h; \cdot)$ is a positive, increasing function, and $k(\cdot)$ is a given positive function modeling the *memory* of the system, including possible transmission delay. If $J(h_i, h_j) > 0$ then spikes of neuron j tend to favor future spikes of neuron i (excitatory link), while the opposites holds true (inhibitory link) when $J(h_i, h_j) < 0$.

There are convenient choices for the kernel $k(\cdot)$ which allow a simple "Markovianization" of the system, namely the *Erlang kernels*: $k(r) = c\frac{r^m}{m!}e^{-\lambda r}$, $c, \lambda > 0$.

Note that for $m \geq 1$ the function k attains its maximum at some positive $r^* = \tau$, producing a "smoothed" form of delay. For simplicity, we deal here with the case $k(r) = e^{-\lambda r}$, corresponding to no delay.

Define

$$X_t^{i,N} := \int_{[0,t]} k(t-s) dZ_s^{i,N},$$

the "discounted" number of spikes of neuron i before time t. The exponential form of $k(\cdot)$ yields

$$\begin{aligned} X_t^{i,N} &= -\lambda \int_0^t X_s^{i,N} ds + Z_t^{i,N} \\ &= -\lambda \int_0^t X_s^{i,N} ds + \int_{[0,t]} \mathbf{1}_{[0, f(h_i, \frac{1}{N}\sum_{j=1}^N J(h_i,h_j) X_{s^-}^{j,N}]}(u) N^i(du, ds), \end{aligned} \tag{4.14}$$

where the N^i are i.i.d. Poisson random measures on $[0,+\infty) \times [0,+\infty)$ with characteristic measure $duds$. The system is therefore in the form seen in Sect. 3. Assuming $f(\cdot)$ is Lipschitz, propagation of chaos holds, and we obtain the macroscopic limit

$$\overline{X}_t(h) = -\lambda \int_0^t \overline{X}_s(h) ds + \int_{[0,t]} \mathbf{1}_{[0, f(h, \int J(h,h') \mathbb{E}[\overline{X}_{s^-}(h')] \mu(dh'))]}(u) N(du, ds). \tag{4.15}$$

Letting $m_t(h) := \mathbb{E}[\overline{X}_t(h)]$, we obtain from (4.15) a closed equation for m_t:

$$\dot{m}_t(h) = -\lambda m_t(h) + f\left(h, \int J(h,h') m_s(h') \mu(dh')\right). \tag{4.16}$$

If the support of μ is finite, this is a finite dimensional dynamical system. A case considered recently [25] is that of the so-called *cyclic negative feedback systems.*

Theorem 4. *Suppose μ is supported on the discrete torus $\mathbb{Z}/n\mathbb{Z}$, $J(h,h') = 0$ unless $h' = h+1$ mod n. Set*

$$\delta := \prod_{h \in \mathbb{Z}/n\mathbb{Z}} J(h, h+1).$$

If $n \geq 3$, $\delta < 0$ and $|\delta|$ is large enough, then (4.16) has at least one stable periodic orbit. This orbit is unique for $n = 3$.

Although single neurons have no intrinsic tendency of spiking periodically, the collective spike train may be periodic if

- the macroscopic geometry of the network is circular;
- at macroscopic level there is an odd number of inhibitory links;
- the interaction is sufficiently strong.

These conditions are quite unrealistic for a real network. This result suggests, however, that the topology of the network and the competition between excitatory and inhibitory links are *factors* that may induce rhythmic behavior.

5 Further Reading

These notes on mean field models have been essentially dealing with propagation of chaos and, for what applications are concerned, with the analysis of the attractors of the macroscopic dynamics. We briefly mention here some further developments, well aware of being far from exhaustive.

5.1 Long-Time Behavior of the Microscopic System

Theorem 1 states that if we *fix* the time interval $[0, T]$ then the microscopic and the macroscopic systems are close if N is large enough. How large, for a given error threshold, might indeed depend on T. In other words for a given large N, this "closeness" might deteriorate as time increases: the long time behavior of the microscopic system is not necessarily reflected in the macroscopic one.

Whenever such "deterioration" *does not* occur, we say there is *uniform propagation of chaos*. One consequence of uniform propagation of chaos is that stationary measures for the microscopic system are close to products of stationary measures of the macroscopic one.

Uniform propagation of chaos has been proven in cases in which the microscopic process satisfies very strong ergodicity properties, see e.g. [6,11,24,40,47,49]. When uniform propagation of chaos fails, it is of interest to identify the time scale (possibly diverging with N) in which the limit macroscopic system still approximate the microscopic one, and determine the behavior beyond this time scale. In general this is a very delicate problem. Quite remarkable results for a class of system inspired by the Kuramoto model are obtained in [29].

5.2 Fluctuations

We have seen (Proposition 1) that propagation of chaos is equivalent to a Law of Large Numbers:

$$\rho_N(\underline{X}^N) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{i,N}} \;\longrightarrow\; Q \tag{5.1}$$

as $N \to +\infty$, where Q is the law of the macroscopic dynamics. It is therefore natural to consider a corresponding Central Limit Theorem, describing *fluctuations* around the limit. In particular, one considers the distribution-valued process

$$\Phi_t^N := \sqrt{N}\left[\frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{i,N}} - q_t\right],$$

where q_t is the marginal of Q at time t. One can prove, with remarkable generality, that for any bounded time-interval $[0, T]$, the process Φ^N converges weakly to a distribution valued Gaussian process. Classical results in this direction can be found in [19,20,26,31].

When quenched disorder is present, fluctuations of the disorder compete with state fluctuations, producing phenomena which are not seen if the disorder

is averaged out; the dynamics of state fluctuations are different for different realizations of the disorder. Sharp results have been obtained for the Kuramoto model in [36].

5.3 Critical Fluctuations

All examples we have treated undergo a *phase transition*: in the macroscopic dynamics, the stationary solution that is unique for small interaction, loses its stability as the interaction strength crosses a threshold, and is subject to bifurcation. At the critical point, the fluctuations process Φ^N defined above exhibit, if evaluated on certain observables, a peculiar space-time scaling, that typically leads to non-Gaussian fluctuations. Literature on this subject has a long history, going back to [19,22]. Possible effects of quenched disorder are dealt with in [17]. Recently, examples of mean field dynamics in which critical fluctuations *self-organize*, i.e. do not require tuning parameters to critical values, are provided in [13]

The results just cited apply to cases in which the bifurcation at the critical point is of *pitchfork* type. Various interesting models, including the Kuramoto model with large quenched disorder, undergoes a Hopf bifurcation. Some indications on how critical fluctuations look like in this case can be found in [21].

5.4 Large Deviations

A refinement of the Law of Large Numbers in (5.1) different from the Central Limit Theorems consists in obtaining a Large Deviation Principle, i.e. the exponential decay in N of probabilities of the form

$$\mathbb{P}(\rho_N(\underline{X}^N) \in U) \quad \text{for } U \not\ni Q.$$

The case of mean-field interacting diffusions with a constant diffusion coefficient given by a multiple of the identity matrix dates back to [20,23], where spin-flip dynamics have also been dealt with. Large deviation principles for system as general as those in Sect. 3 of these note require more sophisticated tools, see [9]. In presence of quenched disorder, it would be desirable to obtain a Large deviation Principle that holds *for almost every* realization of the disorder. For interacting diffusions this is done in [38].

5.5 Generalizing Network's Microscopic Geometry

In the models presented in these notes the quenched disorder is introduced via the local parameters h_i, one per each component. An interesting alternative way of introducing disorder is to associate it with *links*, i.e. to pairs of components. For instance, this would modify the prototypical model in Sect. 2 as

$$dX_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} b(h_{ij}, X_t^{i,N}, X_t^{j,N})dt + dW_t^i$$

where the h_{ij} are random parameters, describing the microscopic architecture of the network. Model of this type, motivated by neurosciences, are dealt with in [43]. We remark that these models are reminiscent of mean field spin-glass dynamics (see e.g. [3]), but actually have very different nature: in spin glasses the contribution of each pair scales as $\frac{1}{\sqrt{N}}$ rather that as $\frac{1}{N}$; the thermodynamic limit for spin glasses is in general much harder to analyze, and the resulting dynamic behavior is quite different.

5.6 Mean-Field Games

In many applications, mainly in social science, collective dynamics are the result of a competitive optimization procedure involving several entities (players). Each player controls, to some extent, his own dynamics, and aims at maximizing his utility; he is therefore taking part to a *dynamic game*. Under symmetry conditions of the players, letting the number of players going to infinity, one expect to obtain a macroscopic game, called *mean field game*.

Introduced in the seminal paper [34], the theory of mean field games has had a tremendous development. The actual convergence of the microscopic dynamics to the mean-field game has been however left open for several years, and recently proved, under rather severe conditions, in [10,12]. In [12] fluctuations around the limit and Large Deviations have also been studied.

Acknowledgement. The author is grateful to an anonymous referee for his careful reading and the useful comments and corrections.

References

1. Andreis, L., Dai Pra, P., Fischer, M.: McKean–Vlasov limit for interacting systems with simultaneous jumps. Stochast. Anal. Appl. **36**(6), 960–995 (2018)
2. Baladron, J., Fasoli, D., Faugeras, O., Touboul, J.D.: Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons. J. Math. Neurosci. **2**(1), 10 (2012)
3. Ben Arous, G., Guionnet, A.: Large deviations for Langevin spin glass dynamics. Probab. Theory Relat. Fields **102**(4), 455–509 (1995)
4. Bertini, L., Giacomin, G., Pakdaman, K.: Dynamical aspects of mean field plane rotators and the Kuramoto model. J. Statist. Phys. **138**, 270–290 (2010)
5. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (2013)
6. Bolley, F., Gentil, I., Guillin, A.: Uniform convergence to equilibrium for granular media. Arch. Ration. Mech. Anal. **208**, 429–445 (2013)
7. Bonilla, L.L., Neu, J.C., Spigler, R.: Nonlinear stability of incoherence and collective synchronization in a population of coupled oscillators. J. Statist. Phys. **67**(1–2), 313–330 (1992)
8. Bonilla, L.L., Pérez Vicente, C.J., Spigler, R.: Time-periodic phases in populations of nonlinearly coupled oscillators with bimodal frequency distributions. Phys. D **113**(1), 79–97 (1998)
9. Budhiraja, A., Dupuis, P., Fischer, M.: Large deviation properties of weakly interacting processes via weak convergence methods. Ann. Probab. **40**(1), 1–435 (2012)

10. Cardaliaguet, P., Delarue, F., Lasry, J.M., Lions, P.-L.: The master equation and the convergence problem in mean field games. arXiv preprint arXiv:1509.02505 (2015)
11. Cattiaux, P., Guillin, A., Malrieu, F.: Probabilistic approach for granular media equations in the non-uniformly convex case. Probab. Theory Relat. Fields **140**(1–2), 19–40 (2008)
12. Cecchin, A., Pelino, G.: Convergence, fluctuations and large deviations for finite state mean field games via the master equation. Stochast. Process. Appl. (2018). https://doi.org/10.1016/j.spa.2018.12.002
13. Cerf, R., Gorny, M.: A Curie-Weiss model of self-organized criticality. Ann. Probab. **44**(1), 444–478 (2016)
14. Chevallier, J.: Mean-field limit of generalized Hawkes processes. Stochast. Process. Appl. **127**(12), 3870–3912 (2017)
15. Chevallier, J., Caceres, M.J., Doumic, M., Reynaud-Bouret, P.: Microscopic approach of a time elapsed neural model. Math. Models Methods Appl. Sci. **25**(14), 2669–2719 (2015)
16. Chevallier, J., Duarte, A., Löcherbach, E., Ost, G.: Mean field limits for nonlinear spatially extended Hawkes processes with exponential memory kernels. Stochast. Process. Appl. **129**(1), 1–27 (2019)
17. Collet, F., Dai Pra, P.: The role of disorder in the dynamics of critical fluctuations of mean field models. Electron. J. Probab. **26**, 1–40 (2012)
18. Collet, F., Dai Pra, P., Formentin, M.: Collective periodicity in mean-field models of cooperative behavior. NoDEA Nonlinear Differ. Equat. Appl. **22**(5), 1461–1482 (2015)
19. Comets, F., Eisele, T.: Asymptotic dynamics, noncritical and critical fluctuations for a geometric long-range interacting model. Commun. Math. Phys. **118**, 531–567 (1988)
20. Dai Pra, P., den Hollander, F.: McKean-Vlasov limit for interacting random processes in random media. J. Statist. Phys. **84**(3–4), 735–772 (1996)
21. Dai Pra, P., Tovazzi, D.: The dynamics of critical fluctuations in asymmetric Curie–Weiss models. Stochast. Process. Appl. **129**(3), 1060–1095 (2019)
22. Dawson, D.A.: Critical dynamics and fluctuations for a mean-field model of cooperative behavior. J. Statist. Phys. **31**(1), 29–85 (1983)
23. Dawson, D.A., Gartner, J.: Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. Stochastics **20**, 247–308 (1987)
24. Del Moral, P., Rio, E.: Concentration inequalities for mean field particle models. Ann. Appl. Probab. **21**, 1017–1052 (2011)
25. Ditlevsen, S., Löcherbach, E.: Multi-class oscillating systems of interacting neurons. Stochast. Process. Appl. **127**(6), 1840–1869 (2017)
26. Fernandez, B., Méléard, S.: A Hilbertian approach for fluctuations on the McKean-Vlasov model. Stochast. Process. Appl. **71**(1), 33–53 (1997)
27. Fournier, N., Guillin, A.: On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Relat. Fields **162**(3–4), 707–738 (2015)
28. Giacomin, G., Luçon, E., Poquet, C.: Coherence stability and effect of random natural frequencies in populations of coupled oscillators. J. Dyn. Diff. Equat. **26**(2), 333–367 (2014)
29. Giacomin, G., Poquet, C., Shapira, A.: Small noise and long time phase diffusion in stochastic limit cycle oscillators. J. Diff. Equat. **264**(2), 1019–1049 (2018)
30. Graham, C.: Nonlinear diffusion with jumps. Annales de l'I.H.P. Probabilités et statistiques **28**(3), 393–402 (1992)

31. Jourdain, B., Méléard, S.: Propagation of chaos and fluctuations for a moderate model with smooth initial data. Ann. Inst. H. Poincaré Probab. Statist. **34**(6), 727–766 (1998)
32. Krupa, M., Touboul, J.D.: Complex oscillations in the delayed FitzHugh-Nagumo equation. J. Nonlinear Sci. **26**(1), 43–81 (2016)
33. Kuramoto, Y.: Chemical Oscillations, Waves, and Turbulence. Courier Dover Publications, New York (2003)
34. Lasry, J.M., Lions, P.-L.: Jeux à champ moyen I. le cas stationnaire. C.R. Acad. Sci. Paris **343**(9), 619–625 (2006)
35. Lindner, B., Garca Ojalvo, A., Neiman, A., Schimansky-Geier, L.: Effects of noise in excitable systems. Phys. Rep. **392**(6), 321–424 (2004)
36. Luçon, E.: Quenched limits and fluctuations of the empirical measure for plane rotators in random media. Electron. J. Probab. **16**, 792–829 (2011)
37. Luçon, E.: Oscillateurs couplés, désordre et synchronisation. Diss. Université Pierre et Marie Curie-Paris VI (2012)
38. Luçon, E.: Quenched large deviations for interacting diffusions in random media. J. Stat. Phys. **166**(6), 1405–1440 (2017)
39. Luçon, E., Poquet, C.: Emergence of oscillatory behaviors for excitable systems with noise and mean-field interaction, a slow-fast dynamics approach. arXiv preprint arXiv:1802.06410 (2018)
40. Malrieu, F.: Logarithmic Sobolev inequalities for some nonlinear PDE's. Stochast. Process. Appl. **95**, 109–132 (2001)
41. Mischler, S., Quiñinao, C., Touboul, J.: On a kinetic Fitzhugh-Nagumo model of neuronal network. Commun. Math. Phys. **342**(3), 1001–1042 (2016)
42. Protter, P.: Stochastic differential equations. In: Stochastic Integration and Differential Equations, pp. 187–284. Springer, Heidelberg (2005)
43. Quiñinao, C., Touboul, J.: Limits and dynamics of randomly connected neuronal networks. Acta Applicandae Mathematicae **136**(1), 167–192 (2015)
44. Quiñinao, C., Touboul, J.D.: Clamping and Synchronization in the strongly coupled FitzHugh-Nagumo model. arXiv:1804.06758v3, April 2018
45. Scheutzow, M.: Noise can create periodic behavior and stabilize nonlinear diffusions. Stoch. Proc. Appl. **20**, 323–331 (1985)
46. Sznitman, A.-S.: Topics in propagation of chaos. In: Ecole d'Eté de Probabilités de Saint-Flour XIX–1989, pp. 165–251. Springer (1991)
47. Veretennikov, A.Y.: On ergodic measures for McKean-Vlasov stochastic equations. In: Monte Carlo and Quasi-Monte Carlo Methods, pp. 471–486 (2006)
48. Touboul, J.D.: Limits and dynamics of stochastic neuronal networks with random heterogeneous delays. J. Stat. Phys. **149**(4), 569–597 (2012)
49. Salhi, J., MacLaurin, J., Toumi, S.: On uniform propagation of chaos. Stochastics **90**(1), 49–60 (2018)

Alignment of Self-propelled Rigid Bodies: From Particle Systems to Macroscopic Equations

Pierre Degond[1(✉)], Amic Frouvelle[2], Sara Merino-Aceituno[1,3,4], and Ariane Trescases[5]

[1] Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
pdegond@imperial.ac.uk
[2] CEREMADE, CNRS, Université Paris-Dauphine, Université PSL, 75016 Paris, France
frouvelle@ceremade.dauphine.fr
[3] School of Mathematical and Physical Sciences, University of Sussex, Falmer BN1 9RH, UK
s.merino-aceituno@sussex.ac.uk
[4] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria
sara.merino@univie.ac.at
[5] IMT, UMR5219, Université de Toulouse, CNRS, 31400 Toulouse, France
ariane.trescases@math.univ-toulouse.fr

Abstract. The goal of these lecture notes is to present in a unified way various models for the dynamics of aligning self-propelled rigid bodies at different scales and the links between them. The models and methods are inspired from [17,18], but, in addition, we introduce a new model and apply on it the same methods. While the new model has its own interest, our aim is also to emphasize the methods by demonstrating their adaptability and by presenting them in a unified and simplified way. Furthermore, from the various microscopic models we derive the same macroscopic model, which is a good indicator of its universality.

Keywords: Self-propelled particles · Rotation matrix · Quaternions · Alignment · Velocity jumps · Generalized Collisional Invariants · Self-organized hydrodynamics

1 Introduction

Collective behavior arises ubiquitously in nature: fish schools, flocks of birds, herds, colonies of bacteria, pedestrian dynamics, opinion formation, are just some examples. One of the main challenges in the investigation of collective behavior is to explain its emergent properties, that is, how from the local interactions between a large number of agents, large-scale structures and self-organization

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 28–66, 2019.
https://doi.org/10.1007/978-3-030-15096-9_2

arise at a much larger scale than the agents' sizes. Kinetic theory provides a mathematical framework for the study of emergent phenomena with the rigorous derivation of equations for the large-scale dynamics (called macroscopic equations) from particle or individual-based models. The derivation of macroscopic equations establishes a rigorous link between the particle dynamics and the large-scale dynamics. Moreover, the simulation of macroscopic equations have the advantage of being, generally, computationally far more efficient than particle simulations, especially as the number of agents grows large.

Tools for the derivation of macroscopic equations were first developed in Mathematical Physics, particularly, in the framework of the Boltzmann equation for rarefied gases [11,13,35]. However, compared to the case of classical equations in Mathematical Physics, an additional difficulty arises here in the study of living systems: the lack of conservation laws. In classical physical systems, each macroscopic quantity corresponds to a conservation law (like the conservation of the total mass, momentum and energy). However, in the models that we will consider here, the number of conserved quantities is less than the number of macroscopic quantities to be determined. To overcome this difficulty we will use the methodological breakthrough presented in [21]: the Generalized Collision Invariant (GCI). This new concept relaxes the condition of being a conserved quantity, and has then been used in a lot of works related to alignment of self-propelled particles [8,15–20,22–24,28]. The goal of this exposition is precisely to clearly illustrate the application of this methodology to models for collective dynamics based on alignment of the body position.

Specifically, in the models that will be considered in this exposition each agent is described by its location in the three-dimensional space and the orientation of its body, represented by a three-dimensional frame. Each agent perceives (directly or indirectly) the orientations of the bodies of the neighboring agents and tends to align with them. This type of collective motion can be found, e.g., in sperm dynamics and animals (birds, fish), and it is a stepping stone to modeling more complex agents composed of articulated bodies (corpora [12]). For more examples and applications based on body attitude coordination see [33] and references therein.

Our models are inspired by time-continuous versions of the Vicsek model, introduced in the 90's [37]. The Vicsek model is now a classic in the field of collective motion: self-propelled particles move at constant speed while trying to align their direction of movement with their neighbors up to some noise. We consider time-continuous versions of the Vicsek model since they are more prone to mathematical studies, as pointed out in [21]. However, there is no obvious unique way of writing a time-continuous version. In [21] and then in [24], two different continuous versions have been proposed that differ by the way agents approach the aligned state: in the first one the particles' velocities align gradually over time towards an aligned state, and in the second one the velocities make discontinuous jumps at discrete times towards an aligned state. Interestingly, both models in [21] and in [24] give rise to the same hydrodynamic/macroscopic limit (with different values for the constants in the equations). Inspired by this,

here we will present two models for alignment of rigid bodies, one given by a time-continuous gradual alignment (taken from the references [17,18]), and another one for alignment based on a jump process on the velocities, that we present here for the first time.

The reason for considering here these two types of models is the following. The main difficulty in applying the Generalized Collision Invariant method to obtain the macroscopic equations lays, precisely, on finding the explicit form of the Generalized Collision Invariants. Indeed, in [17,18] that was the main mathematical difficulty. However, we will see that in the jump model it is straightforward to obtain the GCI but, at the same time, the computation of the macroscopic limit keeps the same structure as in the previous results [17,18]. Particularly, we will obtain the same macroscopic equations (though with different values for the coefficients). The jump model constitutes, therefore, an excellent framework for a didactic exposition of the GCI methodology. With this, the proofs in [17,18] will become more accessible to the reader.

Here, to model alignment of the orientations of the agents seen as rigid bodies (and not only the alignment of their velocities as in the original Vicsek model), we represent the body orientation of an agent as a three-dimensional frame, obtained by the rotation of a fixed frame. Therefore, we will represent the orientations of the agents as rotations. But, as we will see in Sect. 2, in the three-dimensional space rotations can be equivalently represented by rotation matrices and unitary quaternions. Using rotation matrices, the modeling at the individual-based level is more natural and intuitive. However, in terms of numerical efficiency, quaternions require less memory usage (it only requires storing 4 entries rather than 9 entries for matrices) and are less costly to renormalize (while obtaining a rotation matrix from an approximate matrix typically requires a polar decomposition, obtaining a unit quaternion from an approximate quaternion only requires dividing by the norm). We will also see that working with quaternions can give rise to a better presentation of the macroscopic equations.

We conclude by noting that the study of collective behavior based on the Vicsek model and its variations is a fertile field. Among many of the existing mathematical works, we highlight [21] where the hydrodynamic limit has been computed as well as [16], where the emergence of phase transitions is investigated. Many refinements have been proposed to incorporate additional mechanism, such as, to cite only a few of them, volume exclusion [23], presence of leaders [27] or polarization of the group [10]. We refer the interested reader to [17,18] and the references therein for more on this topic. The description of microscopic active particles by coarsed-grained macroscopic equations is also of importance in the physics literature, and this has been tackled through various aspects, such as statistical mechanics on collisional models or lattice models, using for instance Chapman-Enskog approaches or large deviation principles. For the reader interested in this perspective we refer for instance to the works [2–5,31,32].

Let us now describe the structure of the document together with the methodology to link the models at the different scales, and the outcomes that we get.

We start by introducing some notations and recalling some useful properties on matrices, rotations and quaternions in Sect. 2. We present the individual-based models in Sect. 3. From there we derive the mesoscopic models in Sect. 4. These are mean-field models obtained in the limit of a large number of individuals, which describe the evolution in time of the local density of individuals at a given position and orientation. While still taking in account the spatial heterogeneity of the population, these models are valid in the regime where the number of particles interacting with one individual is large. Let us also mention that the models we have chosen does not exhibit any phenomenon of phase transition from disorder to collective motion in this regime, and therefore we are only focused on the derivation of macroscopic model of collective motion. It is possible to study a variation of this model which exhibits phase transition at this mesoscopic scale and this is the object of future work [14]. For the reader interested in the physics discussion whether the order of a transition is correctly given by these mean-field limit or if we need to incorporate finite-densities correction, we refer the reader to [26,34]. Finally, from these mesoscopic models, we perform a hydrodynamic scaling in time and space and use the methodology of the Generalized Collisional Invariants to compute the hydrodynamic limit as this scale parameter tends to zero in Sect. 5. We obtain a macroscopic equation describing the evolution of the local density and the local average orientation. Regarding the predictive interest of this hydrodynamic model, the main question that we are interested in is to know whether we can quantify the difference between simple velocity alignment models such as the Vicsek model and models describing alignment of whole oriented bodies. For instance can we say that body alignment introduce genuinely new dynamics, or is it just similar to aligning the direction of motion, with frame dynamics superimposed to this behavior? This is not easy to answer at the level of the Individual-Based Model, but the macroscopic equations we obtain have a term which is new compared to the macroscopic equation for the Vicsek model alone, and we discuss it in the very end of Sect. 5. At the end of the document, we briefly summarize and discuss all the results, in Sect. 6.

2 Preliminaries: Matrices, Rotations and Quaternions

We present in this section some notations and useful properties on matrices, rotations, and quaternions. We first introduce some notations and present some properties on matrices and rotation matrices. In a second subsection we detail the link between rotations in $\mathbb{R}^3$ and unit quaternions.

The main drawback in using quaternion to represent a rotation is that two opposite quaternions represent the same rotation. In analogy with the theory of rodlike polymers [25], where two opposite unit vectors represent the same orientation, we also present in this first subsection the formalism of Q-tensors, which will be helpful for the modeling. Most of the results will be given without detailed proofs, which can be found in Section 5 of [18].

2.1 Matrices, Rotation Matrices and $\mathbb{R}^3$

We start by introducing a few notations. We will use the following matrix spaces:

- $\mathcal{M}$ is the set of three-by-three matrices,
- $\mathcal{S}$ is the set of symmetric three-by-three matrices,
- $\mathcal{A}$ is the set of antisymmetric three-by-three matrices,
- $O_3(\mathbb{R})$ is the orthogonal group in dimension three,
- $SO_3(\mathbb{R})$ is the special orthogonal group in dimension three.

For a matrix $A \in \mathcal{M}$, we denote by A^{T} its transpose, and we write Tr A its trace, Tr $A = \sum_i A_{ii}$. The matrix I is the identity matrix in $\mathcal{M}$. We use the following definition of the dot product on $\mathcal{M}$: for A, $B \in \mathcal{M}$,

$$A \cdot B := \frac{1}{2} \sum_{i,j=1}^{3} A_{ij} B_{ij}.$$

The choice of this dot product (note in particular the factor $\frac{1}{2}$) is motivated by the following property: for any $u = (u_1, u_2, u_3) \in \mathbb{R}^3$, define the antisymmetric matrix $[u]_\times$ such that

$$[u]_\times := \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix},$$

(or equivalently such that for any $v \in \mathbb{R}^3$, we have $[u]_\times v = u \times v$). Then we have for any u, $v \in \mathbb{R}^3$:

$$[u]_\times \cdot [v]_\times = u \cdot v.$$

The following properties will be useful in the sequel. We state them without proof but the interested reader can find them in Ref. [17].

Proposition 1 (Space decomposition in symmetric and antisymmetric matrices). *We have*

$$\mathcal{S} \oplus \mathcal{A} = \mathcal{M} \text{ and } \mathcal{A} \perp \mathcal{S}.$$

Proposition 2 (Tangent space to $SO_3(\mathbb{R})$, and projection). *For a matrix $A \in SO_3(\mathbb{R})$, denote by T_A the tangent space to $SO_3(\mathbb{R})$ at A. Then*

$$M \in T_A \text{ if and only if there exists } P \in \mathcal{A} \text{ s.t } M = AP,$$

or equivalently the same statement with $M = PA$. Consequently, the orthogonal projection of a matrix M on T_A is given by

$$P_{T_A}(M) = \frac{1}{2}(M - AM^{\mathrm{T}}A), \tag{1}$$

and we have that $M \in T_A^{\perp}$ if and only if $M = AS$ (or equivalently $M = SA$), for some $S \in \mathcal{S}$.

We end up by recalling the polar decomposition of a matrix.

Proposition 3. *Let $M \in \mathcal{M}$. There exist $A \in O_3(\mathbb{R})$ and $S \in \mathcal{S}$ such that*

$$M = AS.$$

Furthermore, if $\det M \neq 0$, then A and S are unique. In this case, we write

$$PD(M) := A.$$

2.2 Quaternions, Rotations and Q-Tensors

Besides rotation matrices, another common representation of rotations in $\mathbb{R}^3$ is done through the unit quaternions, which will be denoted by $\mathbb{H}_1$. Recall that any quaternion q can be written as $q = a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$ with $a, b, c, d \in \mathbb{R}$. Quaternions form a four dimensional (non commutative) division algebra, by the rules $\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = \mathrm{ijk} = -1$. The real part $\mathrm{Re}(q)$ of the quaternion q is a and its imaginary part, denoted $\mathrm{Im}(q)$ is $b\mathrm{i}+c\mathrm{j}+d\mathrm{k}$. The three-dimensional space of purely imaginary quaternions is then identified with $\mathbb{R}^3$, therefore whenever in the paper we have a vector in $\mathbb{R}^3$ which is used as a quaternion, it should be understood that it is a purely imaginary quaternion thanks to this identification. For instance, the vector $e_1 \in \mathbb{R}^3$ (resp. e_2, e_3) is identified with the quaternion i (resp. j, k). The conjugate of the quaternion q is given by $q^* = \mathrm{Re}(q) - \mathrm{Im}(q)$, therefore we get $qq^* = |q|^2 = a^2 + b^2 + c^2 + d^2 \geqslant 0$.

We now explain how the group $\mathbb{H}_1$ (the unit quaternions q, such that $|q| = 1$) provides a representation of rotations. Any unit quaternion $q \in \mathbb{H}_1$ can be written in a polar form as $q = \cos(\theta/2)+\sin(\theta/2)n$, where $\theta \in [0, 2\pi)$ and $n \in \mathbb{S}^2$ (a purely imaginary quaternion with the previous identification). With this notation, the unit quaternion q represents the rotation of angle θ around the axis given by the direction n, anti-clockwise. More specifically, for any vector $u \in \mathbb{R}^3$, the vector quq^* (which is indeed a pure imaginary quaternion whenever $q \in \mathbb{H}_1$ and u is a pure imaginary quaternion, so it can be seen as a vector in $\mathbb{R}^3$) is the rotation of u of angle θ around the axis given by the direction n (note that θ and n are uniquely defined except when $q = \pm 1$: in this case the associated rotation is the identity, and any direction $n \in \mathbb{S}^2$ is suitable).

The underlying map from the group of unit quaternions to the group of rotation matrices is then given by

$$\Phi : \begin{array}{l} \mathbb{H}_1 \to SO_3(\mathbb{R}) \\ q \mapsto \Phi(q) : \begin{array}{l} \mathbb{R}^3 \to \mathbb{R}^3 \\ u \mapsto quq^*. \end{array} \end{array} \tag{2}$$

It is then straightforward to get that Φ is a morphism of groups: for any q and $\widetilde{q}$ in $\mathbb{H}_1$, we have $\Phi(q\widetilde{q}) = \Phi(q)\Phi(\widetilde{q})$ and $\Phi(q^*) = \Phi(q)^{\mathrm{T}}$.

An important remark is that two opposite unit quaternions represent the same rotation:

$$\forall q \in \mathbb{H}_1, \quad \Phi(q) = \Phi(-q). \tag{3}$$

More precisely, the kernel of Φ is given by $\{\pm 1\}$, so that Φ induces an isomorphism between $\mathbb{H}_1/\{\pm 1\}$ and $SO_3(\mathbb{R})$.

We finally briefly introduce the notion of Q-tensors. Indeed, since a unitary quaternion and its opposite correspond to the same rotation matrix, we can see an analogy with the theory of suspensions of rodlike polymers [25]. Those polymers are also modeled using unit vectors (in this case, in $\mathbb{R}^3$), and two opposite vectors are describing the same orientation. Their alignment is called nematic. One relevant object in this theory is the so-called Q-tensor associated with the unit quaternion q, given by the matrix $Q = q \otimes q - \frac{1}{4}\mathrm{I}_4$, where q is seen as a unit vector in $\mathbb{R}^4$, and I_4 is the identity matrix of size four. This object is a symmetric and trace free four by four matrix, which is invariant under the transformation $q \mapsto -q$. We denote by $\mathcal{S}_4^0$ the space of symmetric trace free 4×4 matrices (a vector space of dimension 9), and endow it with the dot product known as "contraction of tensors"; more precisely if $Q, \widetilde{Q}$ are in $\mathcal{S}_4^0$, their contraction $Q : \widetilde{Q} = \sum_{i,j} Q_{ij}\widetilde{Q}_{ij}$ is the trace of $Q\widetilde{Q}^{\mathrm{T}}$. We then get a map

$$\Psi : \begin{array}{ccc} \mathbb{H}_1 & \to & \mathcal{S}_4^0 \\ q & \mapsto & q \otimes q - \frac{1}{4}\mathrm{I}_4, \end{array}$$

whose image can also be identified with $\mathbb{H}_1/\{\pm 1\}$. Indeed, the preimage of $\Psi(q)$ is always equal to $\{q, -q\}$. We therefore have two ways to see $\mathbb{H}_1/\{\pm 1\}$ as a submanifold of a nine-dimensional vector space: either as the image of Φ (in $\mathcal{M}$), which is exactly $SO_3(\mathbb{R})$, or as the image of Ψ (in $\mathcal{S}_4^0$). It appears that the dot products on these spaces behave remarkably well, regarding the maps Φ and Ψ, as stated in the following proposition, from which we can also see that the images are submanifolds of the spheres of radii $\sqrt{\frac{3}{2}}$ (in $\mathcal{M}$) and $\frac{\sqrt{3}}{2}$ (in $\mathcal{S}_4^0$).

Proposition 4. *For any unit quaternions q and $\widetilde{q}$, we have*

$$\frac{1}{2}\Phi(q) \cdot \Phi(\widetilde{q}) = (q \cdot \widetilde{q})^2 - \frac{1}{4} = \Psi(q) : \Psi(\widetilde{q}).$$

Proof. For the second equality, recall that for any quaternions q and $\widetilde{q}$ we have by definition $(q \otimes \widetilde{q})_{ii} = q_i\widetilde{q}_i$, therefore $\mathrm{Tr}(q \otimes \widetilde{q}) = q \cdot \widetilde{q}$ (this justifies the fact that $\mathrm{Tr}(q \otimes q - \frac{1}{4}\mathrm{I}_4) = 0$ when q is a unit quaternion). Using the fact that $(q \otimes q)(\widetilde{q} \otimes \widetilde{q}) = (q \cdot \widetilde{q})\, q \otimes \widetilde{q}$, we get, when q and $\widetilde{q}$ are unit quaternions:

$$\begin{aligned} \Psi(q) : \Psi(\widetilde{q}) &= \mathrm{Tr}((q \otimes q - \tfrac{1}{4}\mathrm{I}_4)(\widetilde{q} \otimes \widetilde{q} - \tfrac{1}{4}\mathrm{I}_4)) \\ &= \mathrm{Tr}((q \otimes q)(\widetilde{q} \otimes \widetilde{q} - \tfrac{1}{4}\mathrm{I}_4)) = (q \cdot \widetilde{q})^2 - \frac{1}{4}. \end{aligned}$$

For the first equality, we first prove that $\mathrm{Tr}(\Phi(q)) = 4\mathrm{Re}(q)^2 - |q|^2$ for any quaternion q. Indeed we first have that $\mathrm{Tr}(\Phi(q)) = \sum_{i=1}^{3} e_i \cdot \Phi(q)e_i$. Writing q as $a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$ and using the identifications between $\mathbb{R}^3$ and the purely imaginary quaternions, we get for instance that $e_1 \cdot \Phi(q)e_1 = \mathrm{Re}(\mathrm{i}(q\mathrm{i}q^*)^*)$, which after computations is $a^2 + b^2 - c^2 - d^2$. At the end, with similar computations for e_2 and e_3, we get $\mathrm{Tr}(\Phi(q)) = 3a^2 - b^2 - c^2 - d^2 = 4\mathrm{Re}(q)^2 - |q|^2$. Therefore, if q and $\widetilde{q}$ are unit quaternions, we get

$$\begin{aligned} 2\Phi(q) \cdot \Phi(\widetilde{q}) &= \mathrm{Tr}(\Phi(q)\Phi(\widetilde{q})^T) \\ &= \mathrm{Tr}(\Phi(q\widetilde{q}^*)) = 4\mathrm{Re}(q\widetilde{q}^*)^2 - |q\widetilde{q}^*|^2 = 4(q \cdot \widetilde{q})^2 - 1. \square \end{aligned}$$

We will also make use of the two following properties regarding the differentiability of the map Φ and the volume forms in $SO_3(\mathbb{R})$ and $\mathbb{H}_1$.

Proposition 5. *The map Φ is continuously differentiable on $\mathbb{H}_1$. Denoting its differential at $q \in \mathbb{H}_1$ by $D_q\Phi : q^\perp \longrightarrow T_{\Phi(q)}$, we have that for any $p \in q^\perp$,*

$$D_q\Phi(p) = 2\,[pq^*]_\times\,\Phi(q).$$

Here, we wrote T_A the tangent space of $SO_3(\mathbb{R})$ at $A = \Phi(q)$, and $q^\perp$ the orthogonal of q.

Proposition 6. *Consider a function $g : SO_3(\mathbb{R}) \to \mathbb{R}$, then*

$$\int_{SO_3(\mathbb{R})} g(A)\,\mathrm{d}A = \int_{\mathbb{H}_1} g(\Phi(q))\,\mathrm{d}q,$$

where $\mathrm{d}q$ and $\mathrm{d}A$ are the normalized Lebesgue measures on the hypersphere $\mathbb{H}_1$ and on $SO_3(\mathbb{R})$, respectively. Furthermore, if $B \in SO_3(\mathbb{R})$, then

$$\int_{SO_3(\mathbb{R})} g(A)\,\mathrm{d}A = \int_{SO_3(\mathbb{R})} g(AB)\,\mathrm{d}A = \int_{SO_3(\mathbb{R})} g(BA)\,\mathrm{d}A.$$

3 Individual Based Modeling: Alignment of Self-propelled Rigid Bodies

Our goal is to model a large number N of particles described, for $n = 1, \ldots, N$, by their positions $X_n \in \mathbb{R}^3$ and their orientations as rigid bodies. The most natural way to describe such an orientation is to give three orthogonal unit vectors u_n, v_n, and w_n. For instance, one way to describe the full orientation of a bird, would be to set the first vector u_n as the direction of its movement (from the center to the beak), the second vector v_n as the direction of its left wing (from the center to the wing), and the last one w_n as the direction of the back, so that u_n, v_n and w_n form a direct orthogonal frame. Therefore the matrix A_n whose three columns are exactly u_n, v_n and w_n is a special orthogonal matrix. This rotation matrix A_n represents the rotation that has to be done between a reference particle the orthogonal vectors of which are exactly the canonical basis of $\mathbb{R}^3$ (denoted by $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$), and the particle number n. A particle can then be described by a pair $(X_n, A_n) \in \mathbb{R}^3 \times SO_3(\mathbb{R})$.

In the spirit of the Vicsek model [37], we want to include in the modeling the three following rules:

- particles move at constant speed,
- particles try to align with their neighbors,
- this alignment is subject to some noise.

Up to changing the time units, we will consider that all particles move at speed one. The first rule requires a direction of movement for each particle.

Therefore, in the following, we will suppose that the first vector $u_n = A_n e_1$ of the matrix A_n represents the velocity of the particle number n. Then, the evolution of the position X_n will simply be given by

$$\frac{\mathrm{d}X_n}{\mathrm{d}t} = A_n e_1. \tag{4}$$

In the quaternion framework, if the quaternion q_n represents the orientation of particle number n (meaning that $\Phi(q_n) = A_n$) then the equation corresponding to (4) reads:

$$\frac{\mathrm{d}X_n}{\mathrm{d}t} = q_n e_1 q_n^*. \tag{5}$$

We now want to describe the evolution of q_n (or of the rotation matrix A_n), taking into account the two remaining rules.

3.1 Defining the Target for the Alignment Mechanism

To implement the second rule in the modeling, in the spirit of the Vicsek model, we need to provide for each particle a way to compute the "average orientation" of the neighbors. In the Vicsek model, the idea was to take the sum of all the velocities of the neighbors and to normalize it in order to have a unit target velocity.

In our framework of rotation matrices, to apply the same procedure, if we want the target orientation $\bar{A}_n$ (viewed from the particle number n) to be a rotation matrix, we need a procedure of normalization which from any matrix gives a matrix of $SO_3(\mathbb{R})$. Indeed, the sum of all rotations matrix A_m of the neighbors need not be a rotation matrix (nor a multiple of such a matrix). The choice that had been done in [17] was to take the polar decomposition: we denote

$$\bar{J}_n = \frac{1}{N}\sum_{m=1}^{N} K(X_m - X_n) A_m \tag{6}$$

$$\bar{A}_n = \mathrm{PD}(\bar{J}_n), \tag{7}$$

where $\mathrm{PD}(J)$ (when $\det(J) \neq 0$) denotes the orthogonal matrix in the polar decomposition (see Proposition 3) of a matrix J, and K is an observation kernel, which weights the orientations of neighbors. A simple example is $K(x) = 1$ if $0 \leqslant |x| < R$ and $K(x) = 0$ if $|x| \geqslant R$. In that case all the neighbors located in the ball of radius R and center X_n have the same influence on the computation of the average matrix $\bar{A}_n$, and all individuals located outside this ball have no influence on this computation.

A first difficulty arises here when the polar decomposition is not a rotation matrix, that is to say $\det(\bar{J}_n) \leqslant 0$. Indeed, due to random effects, we can expect that the polar decomposition is almost surely defined (that is to say $\det(\bar{J}_n) \neq 0$, which happens on a negligible set), but we cannot expect that $\det(\bar{J}_n) > 0$ almost surely.

In the framework of unitary quaternions, things are slightly more complicated. Indeed, since a unitary quaternion and its opposite correspond to the same rotation matrix (see Eq. (3)), the expression used to compute an "average orientation" needs to be invariant by the change of sign of any of the quaternions appearing in the formula. Using the Q-tensors as in the theory of suspensions of rodlike polymers [25] is a good option. We are then led to averaging objects of the form $q \otimes q - \frac{1}{4}\mathrm{I}_4$ which are invariant under the transformation $q \mapsto -q$. The average Q-tensor of the neighbors would then take the form

$$\bar{Q}_n = \frac{1}{N} \sum_{m=1}^{N} K(X_m - X_n)(q_m \otimes q_m - \tfrac{1}{4}\mathrm{I}_4). \tag{8}$$

To define now an "average" quaternion from this Q-tensor $\bar{Q}_n$, we need a procedure which provides a unit vector. We expect that if all quaternions q_n are all equal to a given q (or to $-q$), the procedures returns q or $-q$. Therefore, from the form $\alpha(q \otimes q - \frac{1}{4}\mathrm{I}_4)$, with $\alpha > 0$, it should return q or $-q$. These two vectors are precisely the unit eigenvectors associated to the maximal eigenvalue (which is equal to $\frac{3}{4}$, the other eigenvectors, orthogonal to q, being associated to the eigenvalue $\frac{-1}{4}$). Therefore, in [18], we defined

$$\bar{q}_n = \text{one of the unit eigenvectors of } \bar{Q}_n \text{ of maximal eigenvalue.} \tag{9}$$

Note that the direction of $\bar{q}_n$ is uniquely defined when the maximal eigenvalue is simple. Since symmetric matrices with multiple maximal eigenvalues are negligible, we can expect this definition to be well-posed almost surely.

The first unexpected link found in [18] between this framework of quaternions and the previous framework of average matrices, is that these two averaging procedures are actually equivalent (when the polar decomposition in formula (7) actually returns a rotation matrix). This is due to the following observations [18]:

Proposition 7

(i) If $M \in \mathcal{M}$ is such that $\det(M) > 0$, then the polar decomposition of M is a rotation matrix, and it is the unique maximizer of the function $A \mapsto A \cdot M$ among all matrices A in $SO_3(\mathbb{R})$.

(ii) A unit eigenvector corresponding to the maximal eigenvalue of a symmetric matrix Q maximizes the function $q \mapsto q \cdot Qq$ among all unit vectors q of $\mathbb{H}_1$.

(iii) If for all n we have $\Phi(q_n) = A_n$, and $\det(\bar{J}_n) > 0$, then $\Phi(\bar{q}_n) = \bar{A}_n$, where $\bar{J}_n$, $\bar{A}_n$ and $\bar{q}_n$ are given by (6)–(7), and (8)–(9).

We therefore have now a good procedure to compute $\bar{A}_n$ thanks to the following maximization problem (instead of polar decomposition):

$$\bar{A}_n = \operatorname{argmax} \{A \in SO_3(\mathbb{R}) \mapsto A \cdot M_n\}, \tag{10}$$

where $\bar{J}_n$ is defined in Eq. (6), and from now on we use this definition of $\bar{A}_n$, which ensures that it corresponds to the definition (9) of $\bar{q}_n$ in the world of quaternions.

Since the next part of the modeling will include some random effects, we can expect that the configurations for which the average is not well-defined will be of negligible probability.

We now need to have evolution equations for the orientations (either the rotation matrices A_n or the unit quaternions q_n). In the spirit of the time discrete Vicsek model, it would correspond to saying $A_n(t + \Delta t) = \bar{A}_n(t)+$ "noise", or $q_n(t + \Delta t) = \bar{q}_n(t)+$"noise". However, as was pointed out in [21], in this procedure Δt is actually a parameter of the model, and not a time discretization of an underlying process: indeed, this parameter controls the frequency at which particles change their orientation, and changing the value of this frequency leads to drastic changes in the behavior of the model. Regarding the mathematical study of this type of time-discrete models, it is far from being clear how to go beyond observations of numerical simulations. However, it is possible to build models in the same spirit as the Vicsek model which will be much more prone to mathematical study, in particular if we want to derive a kinetic description (when the number of particles is large) and macroscopic limit (when the scale of observation is large). In the next two subsections we present two ways of building such models. The first one corresponds to a time-continuous alignment mechanism as was proposed in [21], in which the orientation of one particle continuously tries to align with its target orientation, up to some noise. This leads to the models presented in [17] (in the framework of rotation matrices) and in [18] (in the framework of unit quaternions). The second one, as in the Vicsek model, corresponds to a process in which orientations undergo jumps as time evolves, but where the jumps are not synchronous: instead of taking place every time step for all particles, they all have independent times at which they change from their orientations to their target orientations, up to some noise. This leads to a new model, which is different at the particle and kinetic levels, but for which the derivation of the macroscopic model gives the same system of evolution equations (up to the values of the constant parameters of the model). This procedure was studied in [24] for the alignment mechanism of the Vicsek model, and they also found that the macroscopic model corresponds to the Self-Organized Hydrodynamic system of [21] (derived from the time continuous alignment process).

3.2 Gradual Alignment Model

We first consider a time-continuous alignment mechanism. We have to take into account the last two rules (particles try to align with their neighbors, and this alignment is subject to some noise). For the sake of simplicity, we first present the alignment dynamics without noise. We will add noise at the end of this subsection.

The alignment is modeled by a gradual alignment of an agent's body orientation towards its local average defined in the previous subsection. We express the evolution towards the average as the gradient of a polar distance between the agent and the average. It takes the form, in the world of rotation matrices

$$\frac{\mathrm{d}A_n}{\mathrm{d}t} = \nabla_{A_n}\left[A_n \cdot \bar{A}_n\right],$$

and in the world of unit quaternions

$$\frac{\mathrm{d}q_n}{\mathrm{d}t} = \nabla_{q_n}\left[\frac{1}{2}(q_n \cdot \bar{q}_n)^2\right],$$

where the strength of alignment (or equivalently the relaxation frequency) has been taken to be one (which can be done without loss of generality by changing time units), and ∇_{A_n} and ∇_{q_n} represent the gradients on $SO_3(\mathbb{R})$ and $\mathbb{H}_1$ respectively. For the quaternions, we took the square of the norm to account for the fact that only the directions of the vectors q_n and $\bar{q}_n$, and not their sign, should influence the alignment dynamics (this is called nematic alignment, and it is analogous to the case of rodlike polymers, as described in Subsect. 3.1).

The alignment forces can be rewritten respectively as

$$\nabla_{A_n}\left[A_n \cdot \bar{A}_n\right] = P_{T_{A_n}}\bar{A}_n,$$

for the matrices, where $P_{T_{A_n}}$ is the orthogonal projection on the tangent space of $SO_3(\mathbb{R})$ at A_n, given by Eq. (1) (see Proposition 2), and

$$\nabla_{q_n}\left[\frac{1}{2}(q_n \cdot \bar{q}_n)^2\right] = P_{q_n^\perp}\left[\left(\bar{q}_n \otimes \bar{q}_n - \tfrac{1}{4}\mathrm{I}_4\right) q_n\right].$$

for the quaternions, where $P_{q_n^\perp} = \mathrm{I}_4 - q_n \otimes q_n$ is the projection on the orthogonal of q_n.

The second link found in [18] between the frameworks of quaternions and rotation matrices, is that these alignment mechanisms are also equivalent.

Proposition 8. *Consider the system, for all $n = 1..N$,*

$$\frac{\mathrm{d}A_n}{\mathrm{d}t} = \nabla_{A_n}\left[A_n \cdot \bar{A}_n\right], \tag{11}$$

$$A(t=0) = A_n^0 \in SO_3(\mathbb{R}), \tag{12}$$

with $\bar{A}_n$ defined in (10), and the system, for all $n = 1..N$,

$$\frac{\mathrm{d}q_n}{\mathrm{d}t} = \nabla_{q_n}\left[\frac{1}{2}(q_n \cdot \bar{q}_n)^2\right], \tag{13}$$

$$q(t=0) = q_n^0 \in \mathbb{H}_1, \tag{14}$$

with $\bar{q}_n$ defined in (9). If $A_n^0 = \Phi(q_n^0)$ for $n = 1..N$, then, for any solution $(q_n)_n$ of the Cauchy problem (13)–(14), the N-tuple $(A_n)_n := (\Phi(q_n))_n$ is a solution of the Cauchy problem (11)–(12).

The proof of this proposition relies on two main properties: the equivalence of the averaging procedures of Proposition 7 on one hand, and, on the other hand, the computation of the differential of Φ in Proposition 5, which allows us to write a link between the gradient operators on $SO_3(\mathbb{R})$ and on $\mathbb{H}_1$.

We finally describe the complete model by adding the third rule (the fact that the alignment is subject to some noise). The natural way to introduce it is to transform the ordinary differential equations (4)–(11) and (5)–(13), into stochastic differential equations, which take the form of the two following systems:

$$\begin{cases} \mathrm{d}X_n = A_n e_1 \,\mathrm{d}t, \\ \mathrm{d}A_n = P_{T_{A_n}} \circ \left[\bar{A}_n \mathrm{d}t + 2\sqrt{D}\,\mathrm{d}B_t^{9,n} \right], \end{cases} \tag{15}$$

and

$$\begin{cases} \mathrm{d}X_n = q_n e_1 q_n^* \,\mathrm{d}t, \\ \mathrm{d}q_n = P_{q_n^\perp} \circ \left[\left(\bar{q}_n \otimes \bar{q}_n - \frac{1}{4}\mathrm{I}_4 \right) q_n \mathrm{d}t + \sqrt{D/2}\,\mathrm{d}B_t^{4,n} \right], \end{cases} \tag{16}$$

where $(B_t^{9,n})_n$ are matrices of $\mathcal{M}$ with coefficients given by standard independent Brownian motions, and $(B_t^{4,n})_n$ are independent standard Brownian motions on $\mathbb{R}^4$, $D > 0$ representing the noise intensity. The stochastic differential equations have to be understood in the Stratonovich sense, which is well adapted to write stochastic processes on manifolds [30].

Theorem 1 (Equivalence in law [18]). *The processes* (15) *and* (16) *are equivalent in law.*

This theorem relies on the properties of the map Φ defined in (2), in the same way as they are used to prove the equivalence of the alignment dynamics alone in Proposition 8. However in that case the trajectories were exactly the same due to the uniqueness of the solution of the Cauchy problem. Here, since the driving Brownian motions do not belong to the same space (one is on a nine-dimensional space, the other one in a four-dimensional one) we cannot easily give a sense to some pathwise equivalence. However, the projection of these driving Brownian motions on the tangent space of the manifold we consider produce process which are actually three-dimensional, in the sense that their trajectories are contained in a three-dimensional manifold. This is why the equivalence is at the level of the law of the trajectories. Working on the partial differential equations satisfied by the densities of the laws of the processes, and relying on the equivalence of measures in Proposition 6, we can make further use of the differential properties of Φ to write a link between the divergence and Laplacian operators on $SO_3(\mathbb{R})$ and on $\mathbb{H}_1$. We obtain that these partial differential equation are equivalent, which give the equivalence in law. More precise details on the law of such a stochastic differential equation is given in Subsect. 4.1 for the case of a single individual evolving in a given orientation field.

3.3 Alignment Model with Orientation Jumps

In this section we describe an alternative alignment mechanism where the orientations of the particles make jumps at random times. For this, we attach a

Poisson point process with parameter 1 to each particle n (for $n = 1, \dots, N$), which corresponds to the times at which this particle updates its orientation. The increasing sequence of positive times will be denoted by $(t_{n,m})_{m \geqslant 1}$, and can be constructed by independent increments between two consecutive times, given by exponential variables of parameter 1. This means that the unit of time has been chosen in order that it corresponds to the average of the time between two jumps of a given particle.

Next we need to define how the orientation (A_n or q_n) of a particle changes when there is a jump. Recall the definition of the averages $\bar{A}_n$ and $\bar{q}_n$ in (10) and (9) respectively. We want the new orientation to be drawn according to a probability "centered" around $\bar{A}_n$ (resp. $\pm\bar{q}_n$) and radially symmetric, that is, it should have a density of the form $A \mapsto M_{\bar{A}_n}(A)$ (resp. $q \mapsto \widetilde{M}_{\bar{q}_n}(q)$), which only depends on the distance between A and $\bar{A}_n$ (resp. the distance between $\pm q$ and $\pm\bar{q}_n$). In the matrix world, the square of the norm of an orthogonal matrix is $\frac{1}{2}\operatorname{Tr}(A^{\mathrm{T}}A) = \frac{3}{2}$, therefore we have $\|A - \bar{A}_n\|^2 = 3 - 2A \cdot \bar{A}_n$, we are thus looking at a probability density only depending on $A \cdot \bar{A}_n$. Thanks to Proposition 4, in the world of quaternions, it corresponds to a function only depending on $(q \cdot \bar{q}_n)^2$.

To fix the ideas, and to see analogies with the gradual alignment model, we will take for $M_{\bar{A}_n}$ the von-Mises distribution centered around $\bar{A}_n$ and with concentration parameter $\frac{1}{D}$. We will indeed see that in this case, the results of the computations for the macroscopic limits that were done in [17,18] can directly be reused. Of course, the method that we present here still applies for a generic smooth function of $A \cdot \bar{A}_n$.

The von-Mises distribution centered in $\Lambda \in SO_3(\mathbb{R})$ and with concentration parameter $\frac{1}{D}$ is defined, for A in $SO_3(\mathbb{R})$, by

$$M_\Lambda(A) = \frac{1}{Z} \exp\left(\frac{1}{D} A \cdot \Lambda\right), \quad \int_{SO_3(\mathbb{R})} M_\Lambda(A)\, dA = 1, \tag{17}$$

where $Z = Z_D < \infty$ is a normalizing constant such that this function is a probability density on $SO_3(\mathbb{R})$.

Analogously, we define the von-Mises distribution on $\mathbb{H}_1$ as

$$M_{\bar{q}}(q) = \frac{1}{Z'} \exp\left(\frac{2}{D}\left((\bar{q} \cdot q)^2 - \frac{1}{4}\right)\right), \qquad \int_{\mathbb{H}_1} M_{\bar{q}}(q)\, dq = 1, \tag{18}$$

where $Z' = Z'_D$ is a normalizing constant. Thanks to Propositions 4 and 6, if q is a random variable on $\mathbb{H}_1$ distributed according to $M_{\bar{q}}$, then $A = \Phi(q)$ is a random variable on $SO_3(\mathbb{R})$ distributed according to M_Λ, where $\Lambda = \Phi(\bar{q})$.

A useful property of $SO_3(\mathbb{R})$ (or $\mathbb{H}_1$) is that the dot product is invariant by multiplication: we have that $M_{\bar{A}_n}(A) = M_{\mathrm{I}_3}(\bar{A}_n^{\mathrm{T}} A)$, since $\mathrm{I}_3 \cdot (\bar{A}_n^{\mathrm{T}} A) = \bar{A}_n \cdot A$. Furthermore, the measure on $SO_3(\mathbb{R})$ is also left-invariant. We therefore only need to be able to draw random variable according to M_{I_3}, thanks to the following proposition.

Proposition 9. *If $B \in SO_3(\mathbb{R})$ is a random variable distributed according to the density M_{I_3}, then $\bar{A}_n B$ is a random variable distributed according to the density $M_{\bar{A}_n}$.*

Analogously, if $r \in \mathbb{H}_1$ is a random variable distributed according to the density M_1, then $\bar{q}_n r$ is a random variable distributed according to the density $M_{\bar{q}_n}$.

Proof. If U is a measurable set of $SO_3(\mathbb{R})$, then, by left invariance of the measure

$$\begin{aligned}\mathbb{P}(\bar{A}_n B \in U) &= \mathbb{P}(B \in \bar{A}_n^{\mathrm{T}} U)\\ &= \int_{\bar{A}_n^{\mathrm{T}} U} M_{\mathrm{I}_3}(A)\mathrm{d}A = \int_U M_{\mathrm{I}_3}(\bar{A}_n^{\mathrm{T}} A)\mathrm{d}A = \int_U M_{\bar{A}_n}(A)\mathrm{d}A.\end{aligned}$$

Notice that this proof does not rely on the particular expression of the von-Mises distribution, and still applies if $M_{\bar{A}_n}(A)$ is a generic function of $\bar{A}_n \cdot A$. The proof is analogous for the quaternion version. □

We are now ready to construct the stochastic process corresponding to the evolution of positions and orientations of the particles.

Definition 1. *We are given:*

- *a probability density M_{I_3} on $SO_3(\mathbb{R})$, with the property that $M_{\mathrm{I}_3}(A)$ only depends on $\mathrm{I}_3 \cdot A = \frac{1}{2}Tr(A)$ (we will take the von-Mises distribution defined in (17) in the following of the paper),*
- *some independent random variables $S_{n,m} > 0$ and $\eta_{n,m} \in SO_3(\mathbb{R})$, such that for $1 \leqslant n \leqslant N$ and $m \in \mathbb{N}$, $S_{n,m}$ is distributed according to an exponential law of parameter 1 and $\eta_{n,m}$ is distributed according to M_{I_3},*
- *some initial positions $X_{n,0} \in \mathbb{R}^3$ and initial body orientations $A_{n,0} \in SO_3(\mathbb{R})$ for $1 \leqslant n \leqslant N$.*

The variables $(S_{n,m})_{m\in\mathbb{N}}$ represent the intervals of time between consecutive jumps for particle number n. Therefore we define $t_{n,m} = \sum_{0\leqslant \ell<m} S_{n,\ell}$, which corresponds to the time at which particle number n changes its orientation for the m-th time. The positions and orientations are then defined inductively (almost surely, all times $t_{n,m}$ are distinct) by

$$\begin{cases} X_n(0) = X_{n,0}, & \\ X_n(t) = X_n(t_{n,m}) + (t - t_{n,m})A_n(t)e_1, & \text{if } t \in [t_{n,m}, t_{n,m+1}), \\ A_n(t) = A_{n,0}, & \text{if } t \in [0, t_{n,1}), \\ A_n(t) = \bar{A}_n(t_{n,m}^-)\eta_{n,m}, & \text{if } t \in [t_{n,m}, t_{n,m+1}), m \geqslant 1, \end{cases} \tag{19}$$

where $\bar{A}_n$ is the maximizer of the function $A \mapsto A \cdot \frac{1}{N}\sum_{l=1}^N K(X_l - X_n)A_l$.

Since all the independent random variables $\eta_{n,m}$ are distributed according to a law which has a density with respect to the Lebesgue measure on SO_3, and the set of configurations for which this maximizer $\bar{A}_n$ is not well defined are included in low dimensional manifolds (compared to the configuration space), we expect that this process is almost surely well defined. We do not give a detailed proof of this fact here since we are interested in derivation of kinetic models which will share the same issues, therefore we will focus on the formal derivation of

these model in the case where this maximizer is well defined everywhere. The rigorous treatment of this issue is outside the scope of these lecture notes. It is even far from being well understood, even in the case of the Vicsek model, for which the only bad configurations are those with a zero average velocity. At the kinetic level, the only known global existence of solutions requires very strong assumptions of non-vanishing average velocity (which are not only assumptions on the initial conditions) [29].

Analogously, we can define this process in the world of quaternions.

Definition 2. *We are given:*

- *a probability density M_1 on $\mathbb{H}_1$, with the property that $M_1(q)$ only depends on $(1 \cdot q)^2 = \mathrm{Re}(q)^2$ (we will take the von-Mises distribution defined in (18) in the following of the paper),*
- *some independent random variables $S_{n,m} > 0$ and $\eta_{n,m} \in \mathbb{H}_1$, such that for $1 \leqslant n \leqslant N$ and $m \in \mathbb{N}$, $S_{n,m}$ is distributed according to an exponential law of parameter 1 and $\eta_{n,m}$ is distributed according to M_1,*
- *some initial positions $X_{n,0} \in \mathbb{R}^3$ and initial body orientations $q_{n,0} \in \mathbb{H}_1$ for $1 \leqslant n \leqslant N$.*

Again, we define $t_{n,m} = \sum_{0 \leqslant \ell < m} S_{n,\ell}$, which corresponds to the time at which particle number n changes its orientation for the m-th time. The positions and orientations are then defined inductively (almost surely, all times $t_{n,m}$ are distinct) by

$$\begin{cases} X_n(0) = X_{n,0}, & \\ X_n(t) = X_n(t_{n,m}) + (t - t_{n,m})\, q_n e_1 q_n^*, & \text{if } t \in [t_{n,m}, t_{n,m+1}), \\ q_n(t) = q_{n,0}, & \text{if } t \in [0, t_{n,1}), \\ q_n(t) = \bar{q}_n(t_{n,m}^-)\eta_{n,m}, & \text{if } t \in [t_{n,m}, t_{n,m+1}), m \geqslant 1, \end{cases} \tag{20}$$

where $\bar{q}_n$ is defined in (8)–(9).

Once again, we expect this process to be defined almost surely, and as we remarked, thanks to Proposition 4, these two definitions give rise to processes which are equivalent in law, through the map Φ. A last remark is that these processes are a particular case of Piecewise Deterministic Markov Processes (PDMP's): between two jumps, the configuration follows an Ordinary Differential Equation (which in our case is nothing else than free transport). More comments on PDMP's will be made in Subsect. 4.2.

4 Derivation of Kinetic Models

The aim of this section is to present a heuristic derivation of kinetic models corresponding to the limit of the particle systems when the number of particles is large. We present this derivation in the framework of rotation matrices, and we will give the corresponding kinetic models in the framework of quaternions at the end of this section.

To this aim, we introduce the so-called empirical distribution f^N of the particles as the measure

$$f^N(x,A,t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i(t)}(x) \otimes \delta_{A_i(t)}(A),$$

that is to say that if φ is a continuous and bounded function from $\mathbb{R}^3 \times SO_3(\mathbb{R})$ to $\mathbb{R}$, the integral of φ with respect to this measure (at time t) is given by

$$\int_{\mathbb{R}^3\times SO_3(\mathbb{R})} \varphi(x,A) f^N(x,A,t) \mathrm{d}x\, \mathrm{d}A = \frac{1}{N}\sum_{i=1}^{N} \varphi(X_i(t), A_i(t)). \tag{21}$$

This function is independent of the change of numbering of particles, we say that particles are indistinguishable.

Notice that the average orientation $\bar{A}_n$ defined in (10) can be constructed through the empirical distribution: if we define, for a given probability density f, the functions J_f^K and Λ_f^K by

$$J_f^K(x) = \int_{\mathbb{R}^3\times SO_3(\mathbb{R})} K(x-y) A\, f(y,A)\, \mathrm{d}y\, \mathrm{d}A, \tag{22}$$

$$\Lambda_f^K(x) \text{ is a maximizer on } SO_3(\mathbb{R}) \text{ of } A \mapsto A \cdot J_f^K, \tag{23}$$

we get that the definition (6) can be written as $\bar{J}_n = J_{f^N}^K(X_n)$. And therefore we get $\bar{A}_n = \Lambda_{f^N}^K(X_n)$. Therefore we obtain that the interaction between particles (which is only due to this target orientation $\bar{A}_n$) corresponds to an interaction, for each particle, with the field generated by the empirical distribution f^N. The type of limit we want to understand is called mean-field limit: when the number of particles is large, correlations between finite numbers of particles tend to vanish, and a kind of law of large numbers gives that the empirical distribution is well approached by the law of one single particle. This phenomenon is linked to the notion of propagation of chaos, and we refer to [36] for an introduction. This type of limit has been rigorously shown to be valid in various models of collective behavior, such as [7] in a regularized Vicsek model, and [6,9] in cases with less regularity. In our model, it is not straightforward to apply this strategy (due to the regularity issues for the definition of Λ_f^K), therefore we only present a heuristic derivation of the mean-field limit one would obtain if the empirical distribution f^N converges to the law of one single particle when N is large.

Let us now focus for the moment on a single particle model aligning with a given "target field" $\Lambda(x,t) \in SO_3(\mathbb{R})$, and subject to some noise, as in the models given in the previous section. This corresponds to replacing $\bar{A}_n(t)$ by $\Lambda(X_n,t)$ in the models given by (15) (for the gradual alignment model) and (19) (for the alignment model by orientation jumps).

4.1 Gradual Alignment of a Single Individual in an Orientation Field

We then consider the following stochastic differential equation, for the evolution of a particle at position X_t and body orientation A_t, in an orientation field $\Lambda(x,t) \in SO_3(\mathbb{R})$:

$$\begin{cases} \mathrm{d}X_t = A_t e_1 \,\mathrm{d}t, \\ \mathrm{d}A_t = P_{T_{A_t}} \Lambda(X_t,t)\,\mathrm{d}t + 2\sqrt{D}\, P_{T_{A_t}} \circ \mathrm{d}B_t^9, \end{cases} \tag{24}$$

where B_t^9 is a matrix with independent coefficients given by 9 standard Brownian motions on $\mathbb{R}$, and the $\circ$ indicates that this has to be understood in the Stratonovich sense. Let us see how this last fact ensures that the orientation $A(t)$ stays on $SO_3(\mathbb{R})$. Thanks to the classical chain rule satisfied by Stratonovich SDE's [30], for a smooth function $\varphi(x,A)$, we have

$$\begin{aligned} \varphi(X_t,A_t) =& \varphi(X_0,A_0) \\ &+ \int_0^t \left(\mathrm{D}_x\varphi(X_s,A_s)[A_s e_1] + \mathrm{D}_A\varphi(X_s,A_s)[P_{T_{A_s}}\Lambda(X_s,s)]\right)\mathrm{d}s \\ &+ 2\sqrt{D}\int_0^t \left(\mathrm{D}_A\varphi(X_s,A_s)[P_{T_{A_s}}(\cdot)]\right)\circ \mathrm{d}B_s^9, \end{aligned} \tag{25}$$

where D_x and D_A are the differentials with respect to $x \in \mathbb{R}^3$ and $A \in \mathcal{M}$. Now, if we take $\varphi(x,A) = A^{\mathrm{T}}A - \mathrm{I}_3$, we get that $\mathrm{D}_A\varphi(x,A)[H] = A^{\mathrm{T}}H + H^{\mathrm{T}}A$. Thanks to the formula (1), we then get that the linear operator $\mathrm{D}_A\varphi(x,A)[P_{T_A}(\cdot)]$ (from $\mathcal{M}$ to $\mathcal{M}$) is given by

$$\begin{aligned} \mathrm{D}_A\varphi(X,A)[P_{T_A}H] &= \frac{1}{2}A^{\mathrm{T}}(H - AH^{\mathrm{T}}A) + \frac{1}{2}(H^{\mathrm{T}} - A^{\mathrm{T}}HA^{\mathrm{T}})A \\ &= \frac{1}{2}(A^{\mathrm{T}}H\varphi(x,A) - \varphi(x,A)H^{\mathrm{T}}A). \end{aligned}$$

Therefore, if we define the linear operator $L(Y,t) : H \mapsto \frac{1}{2}(A_t^{\mathrm{T}}HY - YH^{\mathrm{T}}A_t)$ and the process $Y_t = \varphi(X_t,A_t) = A_t^{\mathrm{T}}A_t - \mathrm{I}_3$, Eq. (25) becomes

$$Y_t = Y_0 + \int_0^t L(Y_s,s)[\Lambda(X_s,s)]\mathrm{d}s + 2\sqrt{D}\int_0^t L(Y_s,s)\circ \mathrm{d}B_s^9,$$

hence the process Y_t satisfies a linear SDE with initial condition 0, therefore it is 0 for all time, which means that A_t stays in $SO_3(\mathbb{R})$ for all time.

Moreover, it is shown in Chapter 3 of [30] that for a manifold $\mathcal{N}$ embedded in the euclidean space $\mathbb{R}^d$, the generator of the SDE equation $\mathrm{d}Z_t = \sigma P_{T_{Z_t}} \circ \mathrm{d}B_t^d$ (where P_{T_y} is the orthogonal projection on the tangent space T_y of $\mathcal{N}$ at y) is $\frac{\sigma^2}{2}\Delta_{\mathcal{N}}$, where $\Delta_{\mathcal{N}}$ is the Laplace-Beltrami operator on $\mathcal{N}$ (the solution of this SDE is called Brownian motion on $\mathcal{N}$). This means that for a smooth function φ on $\mathcal{N}$,

$$\mathbb{E}[\varphi(Z_t)] = \mathbb{E}[\varphi(Z_0)] + \frac{\sigma^2}{2}\mathbb{E}\left[\int_0^t \Delta_{\mathcal{N}}\varphi(Z_s)\,\mathrm{d}s\right].$$

Using the Stratonovich chain rule, this means that

$$\mathbb{E}\left[\sigma\int_0^t \mathrm{D}_Z\varphi(Z_t)[P_{T_{Z_t}}(\cdot)]\circ \mathrm{d}B_t^d\right] = \frac{\sigma^2}{2}\mathbb{E}\left[\int_0^t \Delta_{\mathcal{N}}\varphi(Z_s)\,\mathrm{d}s\right].$$

In our case, if we write $\mathcal{N} = SO_3(\mathbb{R})$, with the metric induced by the euclidean metric in $\mathbb{R}^9$, this would mean that the expectation of the last term of (25) is $2D\int_0^t \Delta_{\mathcal{N}}\varphi(X_s, A_s)\mathrm{d}s$. However, the metric we used for $SO_3(\mathbb{R})$ is induced by the dot product $(A, B) \mapsto \frac{1}{2}\mathrm{Tr}(A^T B)$, which is half of what corresponds to the euclidean dot product in $\mathbb{R}^9$. The Riemannian metric is then divided by 2, and the formula for the Laplace-Beltrami operator gives that it is then multiplied by 2 (recall the condensed form $\Delta_g\varphi = \frac{1}{\sqrt{|g|}}\partial_i(\sqrt{|g|}g^{ij}\partial^i\varphi)$, where g^{ij} are the coefficients of the inverse of the metric tensor $(g_{ij})_{i,j}$). Therefore we get $\Delta_A\varphi = 2\Delta_{\mathcal{N}}\varphi$. Finally, for an arbitrary test function φ with values in $\mathbb{R}$, taking the expectation in Eq. (25), and using the gradient formulation instead of the differential, we get

$$\begin{aligned}\mathbb{E}[\varphi(X_t, A_t)] &= \mathbb{E}[\varphi(X_0, A_0)]\\ &+\mathbb{E}\Bigg[\int_0^t [\nabla_x\varphi(X_s, A_s)\cdot A_s e_1 + \nabla_A\varphi(X_s, A_s)\cdot P_{T_{A_s}}\Lambda(X_s, s)\\ &\qquad + D\Delta_A\varphi(X_s, A_s)]\mathrm{d}s\Bigg].\end{aligned} \tag{26}$$

Finally, we denote by $f(x, A, t)$ the law of such a particle at time t, which is defined by the formula $\mathbb{E}[\varphi(X_t, A_t)] = \int_{\mathbb{R}\times SO_3(\mathbb{R})}\varphi(x, a)f(x, A, t)\mathrm{d}A\mathrm{d}x$. Then, the fact that Eq. (26) holds for any test function φ, corresponds exactly to the fact that f is a weak solution of the following linear evolution equation:

$$\partial_t f + (Ae_1)\cdot\nabla_x f = -\nabla_A\cdot(P_{T_A}\Lambda f) + D\Delta_A f. \tag{27}$$

4.2 Alignment by Orientation Jumps, for a Single Individual in a Field

We now turn to the model of alignment by orientation jumps. We then consider the following process: given some independent random variables $S_m > 0$ and $\eta_m \in SO_3(\mathbb{R})$, such that for $m \in \mathbb{N}$, S_m is distributed according to an exponential law of parameter 1 and η_m is distributed according to M_{I_3}, an initial position $X_0 \in \mathbb{R}^3$ and orientation $A_0 \in SO_3(\mathbb{R})$, we define $t_m = \sum_{0\leqslant\ell<m} S_\ell$, and the position and orientation at time t are then defined inductively (almost surely, all times t_m are distinct) by

$$\begin{cases} X_t = X_{t_m} + (t - t_m)A_t e_1, & \text{if } t \in [t_m, t_{m+1}),\\ A_t = A_0, & \text{if } t \in [0, t_1),\\ A_t = \Lambda(X_{t_m}, t_m^-)\eta_m, & \text{if } t \in [t_m, t_{m+1}),\ \text{with } m \geqslant 1,\end{cases} \tag{28}$$

Another way to describe this process (X_t, A_t) is to say that it is a (non autonomous) Piecewise Deterministic Markov Process (PDMP) with jump rate 1, with flow ϕ given by $\phi((X,A),t) = (X + tAe_1, A)$ and with transition measure $Q_t((X,A),\cdot) = \delta_X \otimes M_{\Lambda(X,t)}$. The only difference with classical description of PDMP's (see for instance [1] for a review of recent results), except from the fact that we work on a manifold rather than an open set of $\mathbb{R}^d$, is that the transition measure depends on time.

Let us explain how to derive the evolution equation for the law of the process (X_t, A_t). We take once again a smooth test function $\varphi(x,A)$, we fix a small time interval δt, and we evaluate the expectation of $\varphi(X_{t+\delta t}, A_{t+\delta t})$. With probability $1 - \delta t + o(\delta t)$, there is no jump in $(t, t+\delta t)$ and therefore $A_{t+\delta t} = A_t$, and $X_{t+\delta t} = X_t + \delta t A_t e_1$. With probability $\delta t + o(\delta t)$, there is exactly one jump at time s in $(t, o(\delta t))$, and therefore $A_{t+\delta t} = A_s$ which follows the distribution $M_{\Lambda(X_s,s)}$. Of course we have $(t-s) = o(1)$. Finally, there are two or more jumps in $(t, t+\delta t)$ with probability $o(\delta t)$. We therefore get

$$\mathbb{E}[\varphi(X_{t+\delta t},A_{t+\delta t})] = (1 - \delta t + o(\delta t))\mathbb{E}[\varphi(X_t + \delta t A_t e_1, A_t)] + \delta t\, \mathbb{E}\left[\int_{SO_3(\mathbb{R})} \varphi(X_t + o(1), A') M_{\Lambda(X_t+o(1),t+o(1))}(A')\mathrm{d}A'\right] + o(\delta t),$$

which gives, by smoothness of φ, and if we assume that Λ and $\Lambda \mapsto M_\Lambda$ are smooth enough, that

$$\frac{1}{\delta t}\Big(\mathbb{E}[\varphi(X_{t+\delta t},A_{t+\delta t})] - \mathbb{E}[\varphi(X_t,A_t)]\Big) = \mathbb{E}[\nabla_x\varphi(X_t,A_t)\cdot A_t e_1] - \mathbb{E}[\varphi(X_t,A_t)] + \mathbb{E}\left[\int_{SO_3(\mathbb{R})} \varphi(X_t,A') M_{\Lambda(X_t,t)}(A')\mathrm{d}A'\right] + o(1),$$

that is to say

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\varphi(X_t,A_t)] = \mathbb{E}\Big[\nabla_x\varphi(X_t,A_t)\cdot A_t e_1 - \varphi(X_t,A_t) + \int_{SO_3(\mathbb{R})} \varphi(X_t,A') M_{\Lambda(X_t,t)}(A')\mathrm{d}A'\Big]. \tag{29}$$

Finally, as in the previous subsection, we denote by $f(x,A,t)$ the law of such a particle at time t, defined by $\mathbb{E}[\varphi(X_t,A_t)] = \int_{\mathbb{R}\times SO_3(\mathbb{R})} \varphi(x,a) f(x,A,t)\mathrm{d}A\mathrm{d}x$. Now, the fact that Eq. (29) holds for any test function φ corresponds exactly to the fact that f is a weak solution of the following linear evolution equation:

$$\partial_t f + (Ae_1)\cdot\nabla_x f = \rho_f M_\Lambda - f, \tag{30}$$

where

$$\rho_f(x,t) = \int_{SO_3(\mathbb{R})} f(x,A,t)\mathrm{d}A. \tag{31}$$

4.3 Kinetic Mean-Field Models of Alignment

Let us summarize the results of the two previous subsections: for the evolution of a particle in an orientation field $\Lambda(x,t)$ according to one of the models (24) or (28), the law f of the particle is evolving according to one of the (linear) kinetic equations (27) or (30) which is of the form:

$$\partial_t f + (Ae_1)\cdot\nabla_x f = \Gamma_\Lambda(f), \tag{32}$$

with

$$\Gamma_\Lambda(f) = \begin{cases} -\nabla_A\cdot(P_{T_A}\Lambda f) + D\Delta_A f & \text{in the gradual alignment model,} \\ (\rho_f M_\Lambda - f) & \text{in the jump model.} \end{cases} \tag{33}$$

We are now ready to provide a formal derivation of the equation satisfied by the law of one particle in the limit of a large number of particles. The heuristic is as follows: if we consider that the empirical distribution f^N of the particles converges to a deterministic law f, either for the gradual alignment process (15) or for the model with orientation jumps (19), then each particle will evolve in the limit $N\to\infty$ as a single particle in a orientation field $\Lambda(x,t)$ corresponding to $\Lambda^K_{f(t,\cdot)}(x)$, given by the formulas (22)–(23). Therefore the evolution of the law of one particle, in the limit $N\to\infty$, is governed by the evolution equation (32) where Λ is replaced by Λ^K_f. This gives the following (now non-linear and non-local) evolution equation:

$$\partial_t f + (Ae_1)\cdot\nabla_x f = \Gamma_{\Lambda^K_f}(f), \tag{34}$$

where $\Gamma_\Lambda(f)$ is defined above in (33), and with

$$J^K_f(x) = \int_{\mathbb{R}^3\times SO_3(\mathbb{R})} K(x-y)A\, f(y,A)\,\mathrm{d}y\,\mathrm{d}A,$$
$$\Lambda^K_f(x) \text{ maximizes } A\mapsto A\cdot J^K_f(x) \text{ on } SO_3(\mathbb{R}).$$

This heuristic can be made rigorous if the map $f\mapsto\Lambda_f$ is regular (which is not the case here, as there are configurations for which it is not even well defined). Let us quickly present the coupling method (see for instance [36]) to understand how we can indeed use the law of large numbers for independent processes. The idea is first to construct a nonlinear process (for one single particle) which is the natural limit of the evolution of one particle in the particle system corresponding to (15) or (19), and for which the law is following the evolution equation (34). For the gradual alignment process, given a Brownian motion B^9_t as in (24), it would be defined as follows

$$\begin{cases} \mathrm{d}\overline{X}_t = \overline{A}_t e_1\,\mathrm{d}t, \\ \mathrm{d}\overline{A}_t = P_{T_{\overline{A}_t}}\Lambda_{f(t,\cdot)}(\overline{X}_t)\,\mathrm{d}t + 2\sqrt{D}\,P_{T_{\overline{A}_t}}\circ\mathrm{d}B^9_t, \\ f(t,\cdot) \text{ is the law of } (\overline{X}_t,\overline{A}_t). \end{cases} \tag{35}$$

For the orientation by jumps, given random variables t_m and η_m as in (28), it would be given by

$$\begin{cases} \overline{X}_t = \overline{X}_{t_m} + (t - t_m)\overline{A}_t e_1, & \text{if } t \in [t_m, t_{m+1}), \\ \overline{A}_t = A_0, & \text{if } t \in [0, t_1), \\ \overline{A}_t = \Lambda_{f(t_m,\cdot)}(\overline{X}_{t_m})\eta_m, & \text{if } t \in [t_m, t_{m+1}), \text{ with } m \geqslant 1, \\ f(t,\cdot) \text{ is the law of } (\overline{X}_t, \overline{A}_t). \end{cases} \tag{36}$$

These constructions can be seen as fixed point problems for the laws of the trajectories, and this is where the regularity of $f \mapsto \Lambda_f$ can be used to prove a contraction property in an appropriate space. Once these processes are well defined, the second idea of the method of couplings is to introduce N of these processes (35) or (36), for which the initial conditions and random variables (Brownian motions B^9_t, n or jump times $t_{n,m}$ and rotations $\eta_{n,m}$) are the same as for the particle systems (15) and (19). By construction, these auxiliary nonlinear processes $(\overline{X}_n(t), \overline{A}_n(t))$ are independent and identically distributed according to the law f, solution of the kinetic equation (34). Therefore the last step of the coupling method is to perform estimates of the differences between the trajectories of the particle system and of the auxiliary process in order to let appear quantities reminiscent of (21), but of the form

$$\frac{1}{N}\sum_{i=1}^{N} \varphi(\overline{X}_i(t), \overline{A}_i(t)), \tag{37}$$

for which the law of large numbers applies.

Let us finish this subsection by presenting the kinetic equation we obtain (exactly in the same manner) when working with unit quaternions instead of rotation matrices. The formal mean-field limit of the particle system (16) or (20) is given by the following evolution equation, for the density $f(t, x, q)$ of finding a particle at position x with orientation given by the unit quaternion q at time t:

$$\partial_t f + \Phi(q)(e_1) \cdot \nabla_x f = \Gamma_{\overline{q}^K_f} f, \tag{38}$$

with

$$\Gamma_{\overline{q}}(f) = \begin{cases} -\nabla_q \cdot (P_{q^\perp}(\overline{q} \otimes \overline{q}) q\, f) + \frac{D}{4}\Delta_q f & \text{in the gradual alignment model,} \\ (\rho_f M_{\overline{q}} - f) & \text{in the jump model,} \end{cases} \tag{39}$$

where

$$\rho_f = \int_{\mathbb{H}_1} f(q)\mathrm{d}q, \tag{40}$$

$$Q^K_f(x) = \int_{\mathbb{R}^3 \times \mathbb{H}_1} K(x - y)(q \otimes q - \tfrac{1}{4}\mathrm{I}_4) f(y, q)\mathrm{d}y\,\mathrm{d}q, \tag{41}$$

$$\overline{q}^K_f(x) \text{ is an eigenvector of } Q^K_f(x) \text{ of maximal eigenvalue.} \tag{42}$$

5 Macroscopic Limit

In this section we derive the macroscopic dynamics for the kinetic equations (34) and (38). This means that we are interested in the dynamics in the large time as well as large-space scale. For this we first introduce a scaling with respect to a small parameter ε. We then determine the local equilibria of the collision operator, which depend on two macroscopic quantities, a density ρ and a local orientation Λ. The final step is then the derivation of the evolution equations of these macroscopic functions ρ and Λ. The first one comes from the conservation of mass (a collision invariant), and the second one needs more work, and can be derived using the concept of Generalized Collisional Invariants introduced in [21]. The subsection presenting this concept and how to use it to obtain the evolution equation for Λ is the main part of this section.

5.1 Scaling

We introduce the macroscopic temporal and spatial variables (t', x') given by

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

where $0 < \varepsilon \ll 1$ is a scale parameter. We also consider the following rescaling for the interaction kernel:

$$K_\varepsilon(x) = \frac{1}{\varepsilon^3} K\left(\frac{x}{\varepsilon}\right).$$

This corresponds to localized interactions as $\varepsilon \to 0$ (see Remark 1 below). Notice that

$$\int_{\mathbb{R}^3} K_\varepsilon(x)\, dx = \int_{\mathbb{R}^3} K(x)\, dx = 1.$$

Define the function f_ε in the macroscopic variables as

$$f_\varepsilon(t', x', A) = f(t, x, A).$$

Our goal is to determine the dynamics for this function as $\varepsilon \to 0$. Firstly, one can check that the evolution equation for f_ε is given by

$$\varepsilon(\partial_t f_\varepsilon + (Ae_1)\cdot\nabla_x f_\varepsilon) = \Gamma_{\Lambda_{f_\varepsilon}^{K_\varepsilon}}(f_\varepsilon), \qquad \text{(matrix formulation)}, \tag{43}$$

$$\varepsilon(\partial_t f_\varepsilon + \Phi(q)(e_1)\cdot\nabla_x f_\varepsilon) = \Gamma_{\bar{q}_{f_\varepsilon}^{K_\varepsilon}}(f_\varepsilon), \qquad \text{(quaternion formulation)}, \tag{44}$$

where the primes have been skipped. We recall that the definition of the operators Γ_Λ and $\Gamma_{\bar{q}}$ are given in (33) and (39) respectively, and the definitions of the average orientations $\Lambda_{f_\varepsilon}^{K_\varepsilon}$ and $\bar{q}_{f_\varepsilon}^{K_\varepsilon}$ are given in (22)–(23) and (41)–(42) respectively.

Next, we expand the collision operators in the parameter ε.

Lemma 1 (Expansion for localized interactions). *The following expansion holds:*

$$J_f^{K_\varepsilon} = K_\varepsilon *_x J_f = J_f + \mathcal{O}(\varepsilon^2),$$

where $J_f(x)$ takes in account the dependence of f on the variable A only:

$$J_f(x) = \int_{SO_3(\mathbb{R})} A\, f(x, A)\, \mathrm{d}A. \tag{45}$$

Consequently, we can recast Eq. (43) as

$$\varepsilon(\partial_t f_\varepsilon + (Ae_1)\cdot \nabla_x f_\varepsilon) = \Gamma_{\Lambda_{f_\varepsilon}}(f_\varepsilon) + \mathcal{O}(\varepsilon^2), \tag{46}$$

where

$$\Lambda_f \text{ maximizes } A \mapsto A \cdot J_f \text{ on } SO_3(\mathbb{R}). \tag{47}$$

Analogously, we have that

$$Q_f^{K_\varepsilon} = K_\varepsilon *_x Q_f = Q_f + \mathcal{O}(\varepsilon^2),$$

where

$$Q_f(x) = \int_{\mathbb{R}^3\times\mathbb{H}_1} (q\otimes q - \tfrac{1}{4}\mathrm{I}_4) f(x,q)\, \mathrm{d}q, \tag{48}$$

and Eq. (44) is recast as

$$\varepsilon(\partial_t f_\varepsilon + \Phi(q)(e_1)\cdot \nabla_x f_\varepsilon) = \Gamma_{\bar{q}_{f_\varepsilon}} f_\varepsilon + \mathcal{O}(\varepsilon^2), \tag{49}$$

where

$$\bar{q}_f \text{ is an eigenvector of } Q_f \text{ of maximal eigenvalue.} \tag{50}$$

Remark 1 (Localized interactions). Notice that in the leading order of the expansion of K_ε, we obtain a delta distribution in x. This is why we say that this kind of rescaling corresponds to localized interactions in the limit $\varepsilon \to 0$.

The proof of the expansions in Lemma 1 is straightforward using the Taylor expansion and the fact that

$$\int_{\mathbb{R}^3} xK(x)\, dx = 0,$$

for more details on this and the fact that Λ_f and $\bar{q}_f$ are indeed defined as in the lemma, the reader is referred to [17, Lem. 4.1] and [18, Lem. 4.2, Prop. 4.3], respectively.

Our goal is to investigate the limit of f_ε as $\varepsilon \to 0$. Firstly, notice that, formally, from Eq. (46) we have that

$$\text{if } f_\varepsilon(t,x,\cdot) \to f_0(t,x,\cdot) \text{ as } \varepsilon \to 0 \text{ then } f_0(t,x,\cdot) \in \ker(\Gamma_{\Lambda_{f_0(t,x,\cdot)}}), \tag{51}$$

in the matrix formulation, or

$$\text{if } f_\varepsilon(t,x,\cdot) \to f_0(t,x,\cdot) \text{ as } \varepsilon \to 0 \text{ then } f_0(t,x,\cdot) \in \ker(\Gamma_{\bar{q}_{f_0(t,x,\cdot)}}), \tag{52}$$

in the quaternion formulation. For this reason, we study next the kernel of the operator Γ_Λ (which is an operator acting on functions of $A \in SO_3(\mathbb{R})$ only) for a fixed $\Lambda \in SO_3(\mathbb{R})$ (analogously $\Gamma_{\bar{q}}$ for fixed $\bar{q} \in \mathbb{H}_1$) in the following section.

5.2 Study of the Collision Operator Γ

The goal of this subsection is to show that both the jump model and the gradual alignment model have the same type of equilibria. More precisely, we show the following proposition.

Proposition 10 (Equilibria, matrix formulation). *Recall the definition of the operator Γ_Λ in Eq. (33), and the definition of the von Mises distribution M_Λ in Eq. (17). Then, for any $f \geq 0$, we have*

$$\Gamma_{\Lambda_f}(f) = 0 \iff f = \rho M_\Lambda, \text{ for some } \rho \geq 0, \Lambda \in SO_3(\mathbb{R}). \tag{53}$$

Furthermore, any element f of the form $f = \rho M_\Lambda$, with $\rho \geq 0$ and $\Lambda \in SO_3(\mathbb{R})$ satisfies the consistency relations

$$\rho_f = \rho, \quad J_f = \rho c_1 \Lambda, \quad \Lambda_f = \Lambda, \tag{54}$$

where $c_1 \in (0,1)$ is an explicit constant, and ρ_f, J_f and Λ_f are defined in Eq. (31), Eq. (45), and Eq. (47), respectively.

As a consequence, both the gradual alignment model and the jump model have the same equilibria, and therefore, the same type of (formal) limit as $\varepsilon \to 0$: we can write

$$f_0(t,x,A) = \rho(t,x) M_{\Lambda(t,x)}(A),$$

for some $\rho(t,x) \geq 0$ and $\Lambda \in SO_3(\mathbb{R})$ satisfying furthermore the consistency relations

$$\rho_{f_0}(t,x) = \rho(t,x), \quad J_{f_0}(t,x) = \rho(t,x) c_1 \Lambda(t,x), \quad \Lambda_{f_0}(t,x) = \Lambda(t,x). \tag{55}$$

Remark 2 (Variants in the jump-based model). In the jump-based model, the result holds true formally if we replace (in the collision operator and in the result) the von-Mises distribution by any distribution. Therefore, the jump-based model can reproduce a great variety of behaviors in terms of equilibria.

The proof of Proposition 10 is done in [17] in the case of the gradual alignment model. We summarize here the main ideas.

We first prove the consistency relations. Let us take a function f of the form

$$f = \rho M_\Lambda, \tag{56}$$

for some $\rho \geq 0$ and $\Lambda \in SO_3(\mathbb{R})$. Now, one can check by direct computation that the following consistency relation holds for the average of M_Λ (see proof in [17, Lem. 4.4]):

$$\int_{SO_3(\mathbb{R})} A\, M_\Lambda(A)\, dA = c_1 \Lambda, \quad \text{for some} \quad c_1 \in (0,1) \quad \text{explicit.}$$

With this, integrating expression (56) against 1 and A in $SO_3(\mathbb{R})$ we obtain the two first equalities of Eq. (54). To conclude the proof of Eq. (54), the last equality is a consequence of the second one.

Now, in the gradual alignment model, it is proved in [17] that the operator Γ_{Λ_f} can be recast as

$$\Gamma_{\Lambda_f}(f) = D\nabla_A \cdot \left(M_{\Lambda_f} \nabla_A \left(\frac{f}{M_{\Lambda_f}} \right) \right). \tag{57}$$

Using expression (57), one can obtain that

$$\ker(\Gamma_{\Lambda_f}) = \{\rho M_{\Lambda_f} \quad \text{for any} \quad \rho = \rho(t,x)\}$$

(see detailed proof of this statement in [17, Lem. 4.3]), which, thanks to the consistency relations proved before, is equivalent to Eq. (53).

In the case of the jump model, from the definition of the operator Γ_{Λ_f} it is straightforward that its kernel is given by the functions f such that

$$f = \rho_f M_{\Lambda_f},$$

that is, since we have taken M_Λ to be the von-Mises distribution, and using again the consistency relations, exactly Eq. (53).

Therefore, for both models, we can use Eq. (51) to see that the limit f_0 must be of the form

$$f_0(t,x,A) = \rho(t,x) M_{\Lambda(t,x)}(A), \tag{58}$$

for some $\rho = \rho(t,x)$ and $\Lambda = \Lambda(t,x) \in SO_3(\mathbb{R})$ to be determined, and which satisfy the consistency relations.

Analogously, we obtain the same kind of results for the formulation with quaternions. We write only the result on the limiting function.

Lemma 2 (Equilibria, quaternion formulation). *Recall the definition of the operator $\Gamma_{\bar{q}}$ in Eq. (39) and of the von Mises distribution $M_{\bar{q}}$ in Eq. (18). Then, both the gradual alignment model and the jump model have the same equilibria, and therefore, the same type of limit as $\varepsilon \to 0$: we can write*

$$f_0(t,x,q) = \rho(t,x) M_{\bar{q}(t,x)}(q),$$

for some $\rho(t,x) \geq 0$ and $\bar{q}(t,x) \in \mathbb{H}_1$ satisfying furthermore the consistency relations

$$\rho_{f_0}(t,x) = \rho(t,x), \quad \bar{q}_{f_0}(t,x) = \bar{q}(t,x), \tag{59}$$

where $c_1 \in (0,1)$ is an explicit constant, and ρ_f and $\bar{q}_f$ are defined in Eq. (40) and Eq. (50), respectively.

The proof of this proposition is done in [18] in the case of the gradual alignment model. We only recall the main ideas here. First, the consistency relations rely on the consistency relation satisfied by the von-Mises distribution on $\mathbb{H}_1$, which is (see [18, Prop 4.4]):

$$\text{the leading eigenvector of } \int_{\mathbb{H}_1} (q \otimes q - \frac{1}{4} I_4) M_{\bar{q}} \, dq \text{ corresponds to } \bar{q}. \tag{60}$$

Therefore, if we take any f of the form

$$f = \rho M_{\bar{q}},$$

multiplying this expression by 1 and $(q \otimes q - \frac{1}{4}\mathrm{I}_4)$ and integrating on $\mathbb{H}_1$ we have that

$$\rho_f = \rho, \qquad Q_f = \rho_f \int_{\mathbb{H}_1} (q \otimes q - \frac{1}{4} I_4) M_{\bar{q}}\, dq,$$

where Q_f is defined in Eq. (48). As a consequence of the last equality,

$$\bar{q}_f = \bar{q}.$$

To compute the kernel of the collision operator, in the gradual alignment model we use that the collision operator can be recast as

$$\Gamma_{\bar{q}_f}(f) = \frac{D}{4} \nabla_q \cdot \left(M_{\bar{q}_f} \nabla_q \left(\frac{f}{M_{\bar{q}_f}} \right) \right),$$

(proved in [18]). In the jump-based model the computation of the kernel is straightforward.

We then use Eq. (52) to conclude the proposition.

In summary, we have seen that, formally, the limit of f_ε will be of the form ρM_Λ (or $\bar{\rho} M_{\bar{q}}$ for the quaternion case). We are left with determining the dynamics of the functions $\rho = \rho(t,x)$, $\bar{\rho} = \bar{\rho}(t,x)$, $\Lambda = \Lambda(t,x)$ and $\bar{q} = \bar{q}(t,x)$ (macroscopic quantities). This is done in the following section.

5.3 The Equation for the Density ρ

We first compute the evolution for the density $\rho = \rho(t,x)$. We integrate the rescaled kinetic equation (46) over $SO_3(\mathbb{R})$ and divide by ε to obtain

$$\partial_t \left(\int_{SO_3(\mathbb{R})} f_\varepsilon \, dA \right) + \nabla_x \cdot \left(\int_{SO_3(\mathbb{R})} A e_1 \, f_\varepsilon \, dA \right) = 0.$$

Importantly, the right-hand side has vanished in the integration. This cancellation reflects the fact that the total mass is conserved, i.e., the number of particles is preserved through the interactions. Now we can take formally the limit $\varepsilon \to 0$ and since we know the limit of f_ε in Eq. (58) we have that

$$\partial_t \left(\int_{SO_3(\mathbb{R})} \rho M_\Lambda(A) \, dA \right) + \nabla_x \cdot \left(\int_{SO_3(\mathbb{R})} (Ae_1) \, \rho M_\Lambda(A) \, dA \right) = 0,$$

which corresponds to

$$\partial_t \rho + c_1 \nabla_x \cdot (\rho \Lambda e_1) = 0, \tag{61}$$

given the consistency relations (55). The equation for the density ρ in (61) corresponds to the continuity equation: the density of particles is transported with a velocity equal to $c_1 \Lambda e_1$.

In the formulation with quaternions, analogous computations give the same equation for $\bar{\rho}$ with $\Phi(\bar{q})(e_1)$ instead of Λe_1, that is,

$$\partial_t \rho + c_1 \nabla_x \cdot (\rho \Phi(\bar{q})(e_1)) = 0.$$

We are left with computing the evolution for $\Lambda = \Lambda(t,x)$ and $\bar{q} = \bar{q}(t,x)$. This is done in the following section.

5.4 The Equation for the Body Orientation Λ

The natural path to obtain an equation for $\Lambda = \Lambda(t,x)$ is to multiply the rescaled kinetic equation (46) by A; integrate this expression in $SO_3(\mathbb{R})$; and use the consistency relations (55) at the limit $\varepsilon = 0$. First multiplying by A and integrating we obtain

$$\partial_t \int_{SO_3(\mathbb{R})} A f_\varepsilon \, dA + \int_{SO_3(\mathbb{R})} [A \, (Ae_1 \cdot \nabla_x) f_\varepsilon] \, dA = \frac{1}{\varepsilon} \int_{SO_3(\mathbb{R})} A \Gamma_{\Lambda_{f_\varepsilon}}(f_\varepsilon) \, dA + \mathcal{O}(\varepsilon),$$

after dividing by ε on both sides. Notice that the limit of the first term indeed will correspond to $c_1 \partial_t(\rho \Lambda)$ thanks to the second consistency relation in Eq. (55). However, it is unclear how to deal with the ε^{-1} term on the right hand side as we do not have enough information on the asymptotics of the integral (and the same difficulty arises in the quaternion framework). In classical kinetic theory (in Mathematical Physics), this difficulty does not arise: typically every macroscopic quantity corresponds to what is called a conserved quantity or collision invariant. We say that a function ψ is a collision invariant if for all f (in a reasonable class of functions)

$$\int_{SO_3(\mathbb{R})} \psi \Gamma_{\Lambda_f} \, dA = 0.$$

We have already seen that $\psi = 1$ is a collision invariant corresponding to the total mass being conserved. However, here the body orientation Λ is not a conserved quantity. The same kind of non-conservative property arises in the Vicsek model for the momentum of the particles. To sort out this problem we will relax the condition of being a collision invariant taking into account the constraint given by the second equality in Eq. (55) for the limiting function. This gives rise to the concept coined as the Generalized Collision Invariant in Ref. [21] and that we explain in the following.

The Generalized Collision Invariant. Consider the following definition:

Definition 3 (Generalised Collision Invariant). *A function ψ_{Λ_0} is a Generalized Collision Invariant (GCI) associated with $\Lambda_0 \in SO_3(\mathbb{R})$ of the operator Γ if it holds that*

$$\int_{SO_3(\mathbb{R})} \Gamma_{\Lambda_0}(f) \psi_{\Lambda_0} = 0, \quad \textit{for all } f \textit{ such that } P_{T_{\Lambda_0}} \left(\int_{SO_3(\mathbb{R})} A \, f \, dA \right) = 0.$$

We denote by $\mathrm{GCI}(\Lambda_0)$ *this set of Generalized Collision Invariants associated with* Λ_0.

In the quaternion formulation, we say that a function ψ_{q_0} *is a Generalized Collision Invariant associated to* $q_0 \in \mathbb{H}_1$ *of the operator* Γ *if it holds that*

$$\int_{\mathbb{H}_1} \Gamma_{q_0}(f)\psi_{q_0} = 0, \text{ for all } f \text{ such that } P_{q_0^\perp}\left(\int_{\mathbb{H}_1} (q \otimes q - \tfrac{1}{4}\mathrm{I}_4) f(q)\, dq\, q_0\right) = 0.$$

We also denote by $\mathrm{GCI}(q_0)$ *this set of Generalized Collision Invariants associated with* q_0.

Remark 3 (On the constraints on the test functions). In the matrix formulation, one can notice that the condition on the test functions f is equivalent to saying that $J_f \in T_{\Lambda_0}^\perp$, which is equivalent to $J_f = \Lambda_0 S$ for some symmetric matrix S (see Proposition 2). Taking $S = \rho c_1 \mathrm{I}_3$ and $\Lambda_0 = \Lambda_{f_0}$, we recover the second equality in (55). That is, the limiting function f_0 is an admissible test function in the definition of the GCI (associated with Λ_{f_0}). Something similar happens in the case of the quaternions: the conditions on the test functions f is equivalent to asking that $P_{q_0^\perp}(Q_f\, q_0) = 0$, which will hold true if q_0 is an eigenvector of Q_f, which is what happens, in particular, for $f = f_0$ and $q_0 = \bar{q}_{f_0}$ by the second equality in (59). Therefore, the limiting function f_0 is an admissible test function in the definition of the GCI (associated with $\bar{q}_{f_0}$).

Remark 4. It is straightforward to see that this notion extends the notion of collision invariant. In particular, the mass $\psi = 1$, which is a collision invariant, is also a GCI (associated with Λ_0 for any $\Lambda_0 \in SO_3(\mathbb{R})$ if we see ψ as a function on $SO_3(\mathbb{R})$, and associated with q_0 for any $q_0 \in \mathbb{H}_1$ if we see ψ as a function on $\mathbb{H}_1$). Note in particular that the definition of the GCI is non-empty.

We explain next how the GCI is useful. Multiplying Eq. (46) by a GCI associated with Λ_{f_ε} and integrating with respect to A we obtain

$$\varepsilon \int \left(\partial_t f_\varepsilon + (Ae_1 \cdot \nabla_x)(f_\varepsilon)\right) \psi_{\Lambda_{f_\varepsilon}}\, dA = 0.$$

Notice that, indeed, the right hand side vanishes since

$$\int_{SO_3(\mathbb{R})} \Gamma_{\Lambda_{f_\varepsilon}}(f_\varepsilon)\, \psi_{\Lambda_{f_\varepsilon}} = 0,$$

given that f_ε satisfies the condition

$$P_{T_{\Lambda_{f_\varepsilon}}}\left(\int_{SO_3(\mathbb{R})} A f_\varepsilon\, dA\right) = P_{T_{\Lambda_{f_\varepsilon}}} J_{f_\varepsilon} = P_{T_{\Lambda_{f_\varepsilon}}}(S_\varepsilon \Lambda_{f_\varepsilon}) = 0,$$

where $J_{f_\varepsilon} = S_\varepsilon \Lambda_{f_\varepsilon}$ is the Polar Decomposition of J_{f_ε} and S_ε is a symmetric matrix, therefore $J_{f_\varepsilon} \in T_{\Lambda_{f_\varepsilon}}^\perp$ (see Proposition 2).

Now, dividing by ε and then making $\varepsilon \to 0$, using (58) we obtain

$$\int_{SO_3(\mathbb{R})} \left(\partial_t(\rho M_{\Lambda_{f_0}}) + (Ae_1)\cdot\nabla_x(\rho M_{\Lambda_{f_0}})\right)\psi_{\Lambda_{f_0}}\, dA = 0. \tag{62}$$

Consequently, if we can compute the Generalized Collision Invariants in an explicit form, then we will be able to make explicit the limit given in Eq. (62) and we will be done. This is done in the following.

Description of the GCI. The explicit description of the GCI is given in the following proposition. For this, we need to introduce $h = h(r)$ the unique solution (see Ref. [18]) of the following differential equation on $(-1, 1)$:

$$\begin{aligned}(1-r^2)^{3/2}\exp\left(\frac{2r^2}{d}\right)\left(\frac{-4}{d}r^2 - 3\right)h(r) + \frac{d}{dr}\left[(1-r^2)^{5/2}\exp\left(\frac{2r^2}{d}\right)h'(r)\right]\\ = r\,(1-r^2)^{3/2}\exp\left(\frac{2r^2}{d}\right).\end{aligned} \tag{63}$$

The function h is *odd*: $h(-r) = -h(r)$, and it satisfies for all $r \geq 0$, $h(r) \leq 0$ (by maximum principle).

Proposition 11 (Description of the GCI). *Let $\Lambda_0 \in SO_3(\mathbb{R})$ and $q_0 \in \mathbb{H}_1$. Then, it holds that*

$$\begin{aligned}\mathrm{GCI}(\Lambda_0) &= \mathrm{span}\left\{1, \cup_{P\in\mathcal{A}}\psi^P_{\Lambda_0}\right\}, \quad &&(\textit{matrix formulation}),\\ \mathrm{GCI}(q_0) &= \mathrm{span}\left\{1, \cup_{\beta\in q_0{}^{\perp}}\psi^\beta_{q_0}\right\}, \quad &&(\textit{quaternion formulation}),\end{aligned}$$

where, for $P \in \mathcal{A}$ and $\beta \in q_0^{\perp}$,

$$\begin{aligned}\psi^P_{\Lambda_0}(A) &= P\cdot(\Lambda_0^T A)\,\bar{k}(\Lambda_0\cdot A), \quad &&(\textit{matrix formulation}),\\ \psi^\beta_{q_0}(q) &:= (\beta\cdot q)\,\bar{h}(q\cdot q_0), &&(\textit{quaternion formulation})\end{aligned} \tag{64}$$

with $\bar{h}$ given by, for $r \in (-1, 1)$,

$$\bar{h}(r) = \begin{cases} h(r) & \textit{in the gradual alignment model},\\ r & \textit{in the jump model},\end{cases}$$

where h is the unique solution of the differential equation (63), and $\bar{k}$ given by, for $r \in (-1/2, 3/2)$,

$$\bar{k}(s) = \frac{\bar{h}(\frac{1}{2}\sqrt{2s+1})}{\frac{1}{2}\sqrt{2s+1}}. \tag{65}$$

The function $\bar{k}$ is designed so that $\bar{k}(\frac{1}{2} + \cos\theta) = \frac{\bar{h}(\cos\frac{\theta}{2})}{\cos\frac{\theta}{2}}$. It is negative in the gradual alignment model and a constant equal to $\bar{k} = 1$ in the jump model.

Remark 5. The relation between the functions $\bar{k}$ and $\bar{h}$ in Eq. (65) is related to the relation between dot products, see Proposition 4.

The first step to prove this proposition is a characterization of the GCI in terms of the adjoint of the collision operator.

Lemma 3 (Characterization of the GCI). *A function $\psi_{\Lambda_0} : SO_3(\mathbb{R}) \to \mathbb{R}$ (resp., $\psi_{q_0} : \mathbb{H}_1 \to \mathbb{R}$) is a GCI associated with $\Lambda_0 \in SO_3(\mathbb{R})$ (resp., associated with $q_0 \in \mathbb{H}_1$) if and only if there exists $P \in \mathcal{A}$ (resp., $\beta \in q_0^\perp$) such that ψ_{Λ_0} (resp., ψ_{q_0}) is solution of*

$$\begin{aligned} &\Gamma^*_{\Lambda_0}\psi_{\Lambda_0}(A) = P \cdot \Lambda_0^T A, \ \textit{for all } A \in SO_3(\mathbb{R}) \ (\textit{matrix formulation}),\\ &\Gamma^*_{q_0}\psi_{q_0}(q) = (\beta \cdot q)(q \cdot q_0), \ \textit{for all } q \in \mathbb{H}_1 (\textit{quaternion formulation}), \end{aligned} \tag{66}$$

where $\Gamma^*_{\Lambda_0}$ denotes the adjoint in $L^2(SO_3(\mathbb{R}))$ of the operator Γ_{Λ_0} (resp., and $\Gamma^*_{q_0}$ denotes the adjoint in $L^2(\mathbb{H}_1)$ of Γ_{q_0}).

Proof. We show the proof here for the matrix formulation. For the quaternion formulation it is done analogously. Given $f : SO_3(\mathbb{R}) \to \mathbb{R}$ and $\Lambda_0 \in SO_3(\mathbb{R})$, we have the following equivalences (in the second equivalence we use Proposition 2):

$$\begin{aligned} P_{T_{\Lambda_0}}\left(\int_{SO_3(\mathbb{R})} A\ f\ dA\right) = 0 &\Leftrightarrow \int_{SO_3(\mathbb{R})} A\ f\ dA \in T^\perp_{\Lambda_0},\\ &\Leftrightarrow (\Lambda_0 P) \cdot \int_{SO_3(\mathbb{R})} A\ f\ dA = 0 \quad \text{for all} \quad P \in \mathcal{A},\\ &\Leftrightarrow \int_{SO_3(\mathbb{R})} P \cdot (\Lambda_0^T A)\ f\ dA = 0 \quad \text{for all} \quad P \in \mathcal{A},\\ &\Leftrightarrow f \in G^\perp, \end{aligned}$$

where

$$G = \{g \in L^2(SO_3(\mathbb{R})) \mid g(A) = P \cdot \Lambda_0^T A, \quad \text{for some} \quad P \in \mathcal{A}\}.$$

Starting from Definition 3, we then get, for $\psi_{\Lambda_0} : SO_3(\mathbb{R}) \to \mathbb{R}$:

$$\begin{aligned} \psi_{\Lambda_0} \in \mathrm{GCI}(\Lambda_0) &\Leftrightarrow \int_{SO_3(\mathbb{R})} \Gamma_{\Lambda_0}(f)\psi_{\Lambda_0} = 0 \quad \text{for all } f \text{ such that } \ f \in G^\perp,\\ &\Leftrightarrow \int_{SO_3(\mathbb{R})} f\Gamma^*_{\Lambda_0}(\psi_{\Lambda_0}) = 0 \quad \text{for all } f \text{ such that } \ f \in G^\perp,\\ &\Leftrightarrow \Gamma^*_{\Lambda_0}(\psi_{\Lambda_0}) \in (G^\perp)^\perp = G, \end{aligned}$$

where $\Gamma^*_{\Lambda_0}$ is the adjoint of $\Gamma^*_{\Lambda_0}$ in $L^2(SO_3(\mathbb{R}))$. The last equality comes from the fact that the space G is a finite-dimensional subspace of L^2. The last equivalence therefore implies that ψ_{Λ_0} is a GCI if and only if there exists $P \in \mathcal{P}$ such that ψ_{Λ_0} is solution of (66). □

One can check that, in the matrix formulation, for $\psi : SO_3(\mathbb{R}) \to \mathbb{R}$, the adjoint is given by

$$\Gamma^*_{\Lambda_0}(\psi) = \begin{cases} D\nabla_A \cdot (M_{\Lambda_0}\nabla_A\psi) & \text{(gradual alignment model)},\\ \int_{SO_3(\mathbb{R})} \psi(A)\, M_{\Lambda_0}(A)\ dA - \psi & \text{(jump model)}, \end{cases} \tag{67}$$

and in the quaternion formulation we have, for $\bar{\psi} : \mathbb{H}_1 \to \mathbb{R}$:

$$\Gamma^*_{q_0}(\bar{\psi}) = \begin{cases} D\nabla_q \cdot (M_{q_0} \nabla_q \bar{\psi}) & \text{(gradual alignment model)}, \\ \int_{\mathbb{H}_1} \bar{\psi}(q) M_{q_0}(q)\, dq - \bar{\psi} & \text{(jump model)}. \end{cases}$$

The end of the proof of Proposition 11 for the gradual alignment model relies on the application of Lax-Milgram theorem. It is done in references [17] (for the matrix formulation) and [18] (for the quaternion formulation). We do not repeat it here.

In the case of the jump model, for the matrix formulation it is a direct check that for any $P' \in \mathcal{A}$, the function $\psi^{P'}_{\Lambda_0}$ defined by

$$\psi^{P'}_{\Lambda_0}(A) = P' \cdot \Lambda_0^T A,$$

satisfies Eq. (66) with $P = -P'$ and is, therefore, a GCI. As noticed in Remark 4, the constant function $\psi = 1$ is also a GCI. Conversely, using the explicit form (67) of the adjoint operator $\Gamma^*_{\Lambda_0}$, it is also direct to see that any solution ψ_{Λ_0} of Eq. (66) for some $P \in \mathcal{A}$ satisfies

$$\psi_{\Lambda_0}(A) = -P \cdot \Lambda_0^T A + \int_{SO_3(\mathbb{R})} \psi_{\Lambda_0}(A')\ M_{\Lambda_0}(A')\ dA' \in \operatorname{span}\left\{1,\ \psi^{-P}_{\Lambda_0}\right\}.$$

Analogously, one can check that for any $\beta' \in q_0^\perp$, the function $\psi = \psi^{\beta'}_{q_0}$ given in Eq. (64) is indeed a GCI using Eq. (66) with $\beta = -\beta'$ and the consistency relation (60). Conversely, one can check that any solution ψ of Eq. (66) for some $\beta \in q_0^\perp$ belongs to span $\left\{1,\ \psi^{-\beta}_{q_0}\right\}$.

Limiting Equation. Now that we have an explicit form for the GCI, we can go back to the limiting equation (62) (in the matrix formulation) and substitute its value. This way we have that for all $P \in \mathcal{A}$ it holds:

$$\int_{SO_3(\mathbb{R})} \Big(\partial_t(\rho M_\Lambda) + (Ae_1 \cdot \nabla_x)(\rho M_\Lambda)\Big)\ (P \cdot \Lambda^T A)\ dA = 0.$$

This is equivalent to:

$$P \cdot \left[\int_{SO_3(\mathbb{R})} \Big(\partial_t(\rho M_\Lambda) + (Ae_1 \cdot \nabla_x)(\rho M_\Lambda)\Big)\ \Lambda^T A\ dA\right] = 0 \quad \text{for all} \quad P \in \mathcal{A},$$

which implies thanks to Proposition 1 that

$$\int_{SO_3(\mathbb{R})} \Big(\partial_t(\rho M_\Lambda) + (Ae_1 \cdot \nabla_x)(\rho M_\Lambda)\Big)\ \Lambda^T A\ dA \in \mathcal{S},$$

or, in other words,

$$\int_{SO_3(\mathbb{R})} \Big(\partial_t(\rho M_\Lambda) + (Ae_1 \cdot \nabla_x)(\rho M_\Lambda)\Big)(\Lambda^T A - A^T \Lambda)\ dA = 0.$$

It remains only to compute this expression. This expression is exactly the same is in Ref. [17, Equation (4.25)] with the function $\bar{\psi}_0$ appearing in this reference to be taken equal to one. Therefore, here we do not repeat again the computation for this expression and put directly the result in Theorem 2 (Eq. (70)) in the following section.

5.5 Main Results

To introduce the results on the matrix formulation we need to introduce first some notation: For a smooth function Λ from $\mathbb{R}^3$ to $SO_3(\mathbb{R})$, and for $x \in \mathbb{R}^3$, we define the matrix $\mathcal{D}_x(\Lambda)$ such that

$$(w \cdot \nabla_x)\Lambda = [\mathcal{D}_x(\Lambda)w]_\times \Lambda, \qquad \text{for any} \quad w \in \mathbb{R}^3.$$

This matrix is well defined (see [17, Sec. 4.5]). With this, we define the following first-order operators

$$\delta_x(\Lambda) = \mathrm{Tr}(\mathcal{D}_x(\Lambda)), \qquad [\mathrm{r}_x(\Lambda)]_\times = \mathcal{D}_x(\Lambda) - \mathcal{D}_x(\Lambda)^T.$$

In order to present the results on the quaternion formulation, we first introduce the (right) relative differential operator on $\mathbb{H}_1$: for a function $q = q(t,x)$ where $q(t,x) \in \mathbb{H}_1$ and for $\partial \in \{\partial_t, \partial_{x_1}, \partial_{x_2}, \partial_{x_3}\}$, let

$$\partial_{\mathrm{rel}} q := (\partial q)q^*, \ \Big(= \mathrm{Im}((\partial q)q^*)\Big), \tag{68}$$

where ∂q belongs to the orthogonal space of q, and the product has to be understood in the sense of quaternions. Notice that, effectively, $\partial_{\mathrm{rel}} q$ is a purely imaginary quaternion, since $\mathrm{Re}((\partial q)q^*) = q \cdot \partial q = 0$ (by the fact that q is a unit quaternion), and it can be identified with a vector in $\mathbb{R}^3$. With this, we define the (right) relative space differential operators

$$\nabla_{x,\mathrm{rel}} q = (\partial_{x_i,\mathrm{rel}} q)_{i=1,2,3} = ((\partial_{x_i} q)q^*)_{i=1,2,3} \in (\mathbb{R}^3)^3 \subset \mathbb{H}^3,$$
$$\nabla_{x,\mathrm{rel}} \cdot q = \sum_{i=1,2,3} (\partial_{x_i,\mathrm{rel}} q)_i = \sum_{i=1,2,3} ((\partial_{x_i} q)q^*)_i \in \mathbb{R},$$

where $((\partial_{x_i} q)q^*)_i$ indicates the i-th component of $(\partial_{x_i} q)q^*$.

With these notations, we can state the main result:

Theorem 2 ((Formal) macroscopic limit). *The following results hold true for both the jump model and the gradual alignment model. When $\varepsilon \to 0$ in the kinetic equations (46) (matrix representation) and (49) (quaternion representation) it holds (formally) that*

$$f_\varepsilon \to f = f(t,x,A) = \rho M_\Lambda(A), \quad \textit{with } \Lambda = \Lambda(t,x) \in SO_3(\mathbb{R}),\ \rho = \rho(t,x) \geq 0,$$
$$f_\varepsilon \to f = f(t,x,q) = \bar{\rho} M_{\bar{q}}(q), \quad \textit{with } \bar{q} = \bar{q}(t,x) \in \mathbb{H}_1,\ \bar{\rho} = \bar{\rho}(t,x) \geq 0,$$

for the matrix representation and the quaternion representation, respectively. Moreover, if the convergence is strong enough and the pair functions (ρ, Λ), $(\bar{\rho}, \bar{q})$ are regular enough, then they satisfy the following systems, respectively:

$$\partial_t \rho + \nabla_x \cdot (c_1 \rho \Lambda e_1) = 0, \tag{69}$$

$$\begin{aligned} &\rho(\partial_t \Lambda + c_2((\Lambda e_1) \cdot \nabla_x)\Lambda) \\ &\quad + [(\Lambda e_1) \times (2c_3 \nabla_x \rho + c_4 \rho \mathrm{r}_x(\Lambda)) + c_4 \rho \delta_x(\Lambda) \Lambda e_1]_\times \Lambda = 0, \end{aligned} \tag{70}$$

and

$$\partial_t \bar{\rho} + \nabla_x \cdot (c_1 e_1(\bar{q}) \bar{\rho}) = 0, \tag{71}$$

$$\begin{aligned} &\bar{\rho}(\partial_t \bar{q} + c_2'(e_1(\bar{q}) \cdot \nabla_x)\bar{q}) \\ &\quad + c_3 \left[e_1(\bar{q}) \times \nabla_x \bar{\rho}\right] \bar{q} + c_4 \bar{\rho} \left[\nabla_{x,rel} \bar{q}\, e_1(\bar{q}) + (\nabla_{x,rel} \cdot \bar{q}) e_1(\bar{q})\right] \bar{q} = 0, \end{aligned} \tag{72}$$

where the (right) relative differential operator $\nabla_{x,rel}$ is defined in Eq. (68); and

$$e_1(\bar{q}) = \mathrm{Im}(\bar{q} e_1 \bar{q}^*),$$

and where c_i, $i = 1, \ldots, 4$ are explicit constants. To define them we use the following notation: for two real functions g, w consider

$$\langle g \rangle_w := \int_0^\pi g(\theta) \frac{w(\theta)}{\int_0^\pi w(\theta') d\theta'} \, d\theta.$$

Then the constants are given by

$$\begin{aligned} c_1 &= \frac{2}{3} \langle 1/2 + \cos\theta \rangle_{m(\theta)\, \sin^2(\theta/2)}, \\ c_2 &= \frac{1}{5} \langle 2 + 3\cos\theta \rangle_{m(\theta)\, \sin^4(\theta/2) \bar{h}(\cos(\theta/2))\cos(\theta/2)}, \\ c_2' &= \frac{1}{5} \langle 1 + 4\cos\theta \rangle_{m(\theta)\, \sin^4(\theta/2) \bar{h}(\cos(\theta/2))\cos(\theta/2)}, \\ c_3 &= \frac{D}{2}, \\ c_4 &= \frac{1}{5} \langle 1 - \cos\theta \rangle_{m(\theta)\, \sin^4(\theta/2) \bar{h}(\cos(\theta/2))\cos(\theta/2)}, \end{aligned}$$

where

$$m(\theta) := \exp\left(\frac{1}{D}\left(\frac{1}{2} + \cos\theta\right)\right),$$

with $\bar{h}$ given by, for $r \in (-1, 1)$,

$$\bar{h}(r) = \begin{cases} h(r) & \text{in the gradual alignment model,} \\ r & \text{in the jump model,} \end{cases}$$

where h is the unique solution of the differential equation (63).

Note that the matrix product in the fourth term of Eq. (72) has to be understood as a matrix product, giving rise to a scalar product in $\mathbb{H}$:

$$\nabla_{x,\mathrm{rel}\bar{q}}\, e_1(\bar{q}) = ((\partial_{x_i,\mathrm{rel}\bar{q}}) \cdot e_1(\bar{q}))_{i=1,2,3}.$$

We now state the equivalence of the matrix formulation and the quaternion formulation:

Theorem 3 (Equivalences of the equations [18]). *Let $\rho_0 = \rho_0(x) \geq 0$. Let $\bar{q}_0 = \bar{q}_0(x) \in \mathbb{H}_1$ and $\Lambda_0 = \Lambda_0(x) \in SO_3(\mathbb{R})$ represent the same rotation, i.e., $\Lambda_0(x) = \Phi(\bar{q}_0(x))$ for all $x \in \mathbb{R}^3$. Then the system (69)–(70) and the system (71)–(72) are equivalent (in the sense that any solution $(\rho, \Lambda = \Phi(\bar{q}))$ of the system (69)–(70) is a solution $(\bar{\rho}, \bar{q})$ of (71)–(72)).*

Therefore the equations in the matrix formulation and in the quaternion formulation are equivalent. For an explicit term-by-term equivalence, the reader is referred to [18, Sec. 5.3.3]. Moreover, we have the following corollary:

Corollary 1. *The jump model and the gradual alignment model give rise to the same macroscopic equations with different constants when the equilibria in the jump model is given by a von-Mises distribution.*

We conclude this section by giving a short interpretation of the macroscopic equations obtained in Theorem 2. For a full description and justification we refer the reader to [17,18]. Since by Theorem 3 we know that the systems (69)–(70) and (71)–(72) are equivalent, we will restrict ourselves to interpreting the matrix formulation (for more details on the quaternion formulation the reader is referred to [18]).

Equation (69) is the continuity equation for ρ and ensures mass conservation. The convection velocity is given by $c_1 \Lambda e_1$ and Λe_1 gives the direction of motion. Equation (70) gives the evolution of the mean orientation Λ. We remark that every term in Eq. (70) belongs to the tangent space at Λ in $SO(3)$; this is true for the first term since $(\partial_t + c_2(\Lambda e_1) \cdot \nabla_x)$ is a differential operator and it also holds for the second term because it is the product of an antisymmetric matrix with Λ (see Proposition 2).

The term corresponding to c_3 in (70) gives the influence of $\nabla_x \rho$ (pressure gradient) on the body attitude Λ. It has the effect of rotating the body around the vector directed by $(\Lambda e_1) \times \nabla_x \rho$ at an angular speed given by $\frac{c_3}{\rho} \|(\Lambda e_1) \times \nabla_x \rho\|$, so as to align Λe_1 with $-\nabla_x \rho$ (for more details on this, see [17]). Therefore, the $\nabla_x \rho$ term has the same effect as a pressure gradient in classical hydrodynamics. In this case the pressure gradient has the effect of rotating the whole body frame.

If we had that $c_3 = c_4 = 0$, then we would recover the Self-Organized Hydrodynamic (SOH) model. The SOH model corresponds to the macroscopic equations of the Vicsek model [21]. The SOH model bears analogies with the compressible Euler equations, where (69) is obviously the mass conservation equation and (70) is akin to the momentum conservation equation. There are however major differences. The first one is that we preserve the constraint $\Lambda(t) \in SO_3(\mathbb{R})$

for all times and so the mass convection speed is $|c_1 \Lambda(t) e_1| = c_1$ for all times, while the velocity in the Euler equations is an arbitrary vector. The second one is that the convection speed c_2 is a priori different from the mass convection speed c_1. This difference is a signature of the lack of Galilean invariance of the system, which is a common feature of all dry active matter models.

The major novelty of the present model are the terms with constants c_3 and c_4. They influence the transport of the direction of motion Λe_1. The overall dynamics tends to align the velocity orientation Λe_1, not opposite to the density gradient $\nabla_x \rho$ but opposite to a composite vector $(c_3 \nabla_x \rho + c_4 \rho \, \mathrm{r}_x)$. The vector r_x gives rise to an effective pressure force which adds up to the usual pressure gradient. In addition to this effective force, spatial inhomogeneities of the body attitude also have the effect of inducing a proper rotation of the frame about the direction of motion. This proper rotation is proportional to δ_x. For an interpretation of r_x, δ_x, see [17].

Finally, we add the following interpretation based on the quaternion formulation. First, note that considering $\partial = \partial_t$ the time derivative, for a function $q = q(t, x)$ with values in $\mathbb{H}_1$ the vector $\partial_{t,\mathrm{rel}} q = \partial_t q \, q^{-1}$ is half of the angular velocity of a solid of orientation represented by q. By analogy, the vector $\partial_{x_i,\mathrm{rel}} q = \partial_{x_i} q \, q^{-1}$ for $i = 1, 2, 3$ is half of the angular variation in space of a solid of orientation represented by q. Now, in the quaternion formulation the evolution equation for the body attitude can be rewritten as

$$\bar{\rho}(\partial_{t,\mathrm{rel}} \bar{q} + c_2 (e_1(\bar{q}) \cdot \nabla_{x,\mathrm{rel}}) \bar{q}) \\ + c_3 e_1(\bar{q}) \times \nabla_x \bar{\rho} + c_4 \bar{\rho} \, [\nabla_{x,\mathrm{rel}} \bar{q} \, e_1(\bar{q}) + (\nabla_{x,\mathrm{rel}} \cdot \bar{q}) e_1(\bar{q})] = 0,$$

simply by multiplying Eq. (72) by $\bar{q}^{-1}$ on the right. This equation lives in $\mathbb{R}^3$ (since $\partial_{\mathrm{rel}} \bar{q}$ lives in $\mathbb{R}^3$), and it only involves the following physical quantities: the macroscopic density ρ (and its space gradient), the macroscopic direction of movement $e_1(\bar{q})$, and the macroscopic angular time/space variations of the body attitude $2\partial_{\mathrm{rel}} \bar{q}$.

6 Conclusion

In these notes, we have formally derived macroscopic models, starting from the description of particle systems, and using an intermediate kinetic model to link the two scales. The two limits ($N \to \infty$ for the particle system, and $\varepsilon \to 0$ for the rescaled kinetic equations) are formal derivations, but some steps towards a rigorous limit can be done. A way to recover a rigorous mean-field limit is to change the model in such a way that the singular behavior of the alignment is removed, as in [7], but it introduces a phenomenon of phase transition as in [15,16]. The study of this phase transition is an ongoing work. Another issue to have a better understanding of the limit $\varepsilon \to 0$ is to have well-posedness of the macroscopic system (69)–(70), so we need to study its hyperbolicity. This is also an ongoing work.

Acknowledgments. P. Degond acknowledges support from the Royal Society and the Wolfson foundation through a Royal Society Wolfson Research Merit Award ref WM130048; the British "Engineering and Physical Research Council" under grants ref: EP/M006883/1 and EP/P013651/1; the National Science Foundation under NSF Grant RNMS11-07444 (KI-Net). P. Degond is on leave from CNRS, Institut de Mathématiques de Toulouse, France.

A. Frouvelle acknowledges support from the EFI project ANR-17-CE40-0030 and the Kibord project ANR-13-BS01-0004 of the French National Research Agency (ANR), as well as from the project Défi S2C3 POSBIO of the interdisciplinary mission of CNRS, and the project SMS co-funded by CNRS and the Royal Society.

A. Trescases acknowledges support from the Kibord project ANR-13-BS01-0004 of the French National Research Agency (ANR).

References

1. Azaïs, R., Bardet, J.B., Génadot, A., Krell, N., Zitt, P.A.: Piecewise deterministic Markov process—recent results. Journées MAS 2012, ESAIM Proc. **44**, 276–290 (2014)
2. Barré, J., Chétrite, R., Muratori, M., Peruani, F.: Motility-induced phase separation of active particles in the presence of velocity alignment. J. Stat. Phys. **158**(3), 589–600 (2015)
3. Baskaran, A., Marchetti, M.C.: Hydrodynamics of self-propelled hard rods. Phys. Rev. E **77**, 011920 (2008)
4. Bertin, E., Droz, M., Grégoire, G.: Boltzmann and hydrodynamic description for self-propelled particles. Phys. Rev. E **74**, 022101 (2006)
5. Bertin, E., Droz, M., Grégoire, G.: Hydrodynamic equations for self-propelled particles: microscopic derivation and stability analysis. J. Phys. A: Math. Theor. **42**(44), 445001 (2009)
6. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Stochastic mean-field limit: non-Lipschitz forces & swarming. Math. Models Methods Appl. Sci. **21**(11), 2179–2210 (2011)
7. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Mean-field limit for the stochastic Vicsek model. Appl. Math. Lett. **3**(25), 339–343 (2012)
8. Bostan, M., Carrillo, J.A.: Reduced fluid models for self-propelled particles interacting through alignment. Math. Models Methods Appl. Sci. **27**(7), 1255–1299 (2017)
9. Carrillo, J.A., Choi, Y., Hauray, M., Salem, S.: Mean-field limit for collective behavior models with sharp sensitivity regions. J. Eur. Math. Soc. (2018, to appear)
10. Cavagna, A., Del Castello, L., Giardina, I., Grigera, T., Jelic, A., Melillo, S., Mora, T., Parisi, L., Silvestri, E., Viale, M., et al.: Flocking and turning: a new model for self-organized collective motion. J. Stat. Phys. **158**(3), 601–627 (2014)
11. Cercignani, C., Illner, R., Pulvirenti, M.: The Mathematical Theory of Dilute Gases, vol. 106. Springer, New York (2013)
12. Constantin, P.: The Onsager equation for corpora. J. Comput. Theor. Nanosci. **7**(4), 675–682 (2010)
13. Degond, P.: Macroscopic limits of the Boltzmann equation: a review. In: Degond, P., Pareschi, L., Russo, G. (eds.) Modeling and Computational Methods for Kinetic Equations, pp. 3–57. Springer (2004)
14. Degond, P., Diez, A., Frouvelle, A., Merino-Aceituno, S.: Phase transitions and macroscopic limits in a BGK model of body-attitude coordination (2018, in preparation)

15. 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)
16. 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)
17. 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)
18. Degond, P., Frouvelle, A., Merino-Aceituno, S., Trescases, A.: Quaternions in collective dynamics. Multiscale Mod. Simul. **16**(1), 28–77 (2018)
19. Degond, P., Liu, J.G., Motsch, S., Panferov, V.: Hydrodynamic models of self-organized dynamics: derivation and existence theory. Methods Appl. Anal. **20**(2), 89–114 (2013)
20. Degond, P., Manhart, A., Yu, H.: A continuum model for nematic alignment of self-propelled particles. Discrete Contin. Dyn. Syst. Ser. B **22**(4), 1295–1327 (2017)
21. Degond, P., Motsch, S.: Continuum limit of self-driven particles with orientation interaction. Math. Models Methods Appl. Sci. **18**, 1193–1215 (2008)
22. Degond, P., Motsch, S.: A macroscopic model for a system of swarming agents using curvature control. J. Stat. Phys. **143**(4), 685–714 (2011)
23. Degond, P., Navoret, L.: A multi-layer model for self-propelled disks interacting through alignment and volume exclusion. Math. Models Methods Appl. Sci. **25**(13), 2439–2475 (2015)
24. Dimarco, G., Motsch, S.: Self-alignment driven by jump processes: macroscopic limit and numerical investigation. Math. Models Methods Appl. Sci. **26**(07), 1385–1410 (2016)
25. Doi, M., Edwards, S.F.: The Theory of Polymer Dynamics. International Series of Monographs on Physics, vol. 73. Oxford University Press, Oxford (1999)
26. Farrell, F.D.C., Marchetti, M.C., Marenduzzo, D., Tailleur, J.: Pattern formation in self-propelled particles with density-dependent motility. Phys. Rev. Lett. **108**, 248101 (2012)
27. Ferdinandy, B., Ozogány, K., Vicsek, T.: Collective motion of groups of self-propelled particles following interacting leaders. Phys. A **479**, 467–477 (2017)
28. Frouvelle, A.: A continuum model for alignment of self-propelled particles with anisotropy and density-dependent parameters. Math. Models Methods Appl. Sci. **22**(7), 1250011 (2012)
29. Gamba, I.M., Kang, M.J.: Global weak solutions for Kolmogorov-Vicsek type equations with orientational interactions. Arch. Ration. Mech. Anal. **222**(1), 317–342 (2016)
30. Hsu, E.P.: Stochastic Analysis on Manifolds. Graduate Series in Mathematics, vol. 38. American Mathematical Society, Providence (2002)
31. Ihle, T.: Kinetic theory of flocking: derivation of hydrodynamic equations. Phys. Rev. E **83**, 030901 (2011)
32. Kourbane-Houssene, M., Erignoux, C., Bodineau, T., Tailleur, J.: Exact hydrodynamic description of active lattice gases. Phys. Rev. Lett. **120**, 268003 (2018)
33. Sarlette, A., Sepulchre, R., Leonard, N.E.: Autonomous rigid body attitude synchronization. Automatica **45**(2), 572–577 (2009)
34. Solon, A.P., Tailleur, J.: Revisiting the flocking transition using active spins. Phys. Rev. Lett. **111**, 078101 (2013)
35. Sone, Y.: Kinetic Theory and Fluid Dynamics. Springer, Berlin (2012)

36. 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)
37. 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)

Fluctuations in Stochastic Interacting Particle Systems

Gunter M. Schütz(✉)

Institute of Complex Systems II, Forschungszentrum Jülich, 52425 Jülich, Germany
g.schuetz@fz-juelich.de

Abstract. We discuss fluctuations in stochastic lattice gas models from a microscopic and mesoscopic perspective by using techniques from algebra, in particular the use of symmetries and time-reversal. First we present a generic method to derive rigorously duality functions. As applications we obtain detailed information about density fluctuations in the symmetric simple exclusion process on any graph and about the microscopic structure and fluctuations of shocks in the one-dimensional asymmetric simple exclusion process. Then we use time reversal to prove a general current fluctuation theorem from which celebrated fluctuation relations such as the Jarzynski relation and the Gallavotti-Cohen symmetry arise as corollaries and which can be straightforwardly generalized to derive other fluctuation relations. Finally, going beyond rigorous results, we describe briefly how nonlinear fluctuating hydrodynamics yields the Fibonacci family of dynamical universality classes which has the diffusive and Kardar-Parisi-Zhang universality classes as its first two members.

Keywords: Interacting particle systems · Duality · Shocks · Fluctuation theorems · Dynamical universality classes

1 Introduction

A major outcome of the research on driven diffusive systems in one dimension is the insight that they exhibit remarkably rich stationary and dynamical properties even when interactions are only short-ranged. One observes in nonequilibrium steady states anomalous transport, boundary-induced phase transitions, spontaneous symmetry breaking, long-range order and phase coexistence which have no equilibrium counterpart in one space dimension. The dynamics exhibit intriguing shock-like discontinuities, universal non-diffusive dynamical scaling, remarkable large deviation properties, and more. From a theoretical perspective the main task is to characterize universal features of these phenomena and to understand how they emerge from the microscopic dynamics, in particular, from conservation laws and other kinetic constraints on the microscopic interactions.

A major contribution to this program has come from the study of stochastic lattice gas models [11,12,26,60,67,68,93,94,100]. These are Markovian processes for classical interacting particles subject to one or both of the following two mechanisms that break time-reversal symmetry

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 67–134, 2019.
https://doi.org/10.1007/978-3-030-15096-9_3

- Action of directed non-Hamiltonian random forces
- Particle exchange with boundary reservoirs at different densities

As a result of these mechanisms the particle system can support macroscopic stationary currents that are odd under time-reversal, which is a hallmark of nonequilibrium behaviour.

Further significant progress on the role of time-reversal has been made by studying large deviations in Markov processes where the transition rates may depend on time. It has been realized that time-reversal implies a relation between the probability of a positive value of some fluctuating quantity to the probability of the negative of that quantity [31,42,44,65,99] arbitrarily far from its mean value. Prominent examples for such relations include the Jarzynski relation for the distribution of nonequilibrium work [49] and the Gallavotti-Cohen theorem [34] for the distribution of the entropy production in deterministic dynamics. These and other fluctuation theorems provide further deep insight into many-body systems far from thermal equilibrium.

Various exact and rigorous results along these lines have been obtained by exploiting a very simple mathematical relation between the Markov generator of such processes and the Hamiltonian operator of certain quantum systems [69,93]. In some case the quantum Hamiltonian is amenable to treatment by mathematical tools coming from linear and multilinear algebra, and the theory of associative algebras. These tools can then be employed to solve probabilistic problems, the quantum physics of the Hamiltonian being completely irrelevant in this context. In these lectures we survey some of these methods which seem somewhat alien to probability theory but are nevertheless useful, particularly on the microscopic level of the model. No reference to actual quantum physics is necessary and none will be made.

The usefulness of the correspondence to quantum Hamiltonians goes beyond the algebraic properties and notably includes in some cases of interest the full integrability of the quantum Hamiltonian where powerful techniques such as the Bethe ansatz and random matrix theory come into play. The full correspondence to quantum Hamiltonians has given rise to the study of integrable probability [16]. The present notes may be useful as a stepping stone to this more advanced application of the quantum Hamiltonian formalism.

1.1 Exclusion Processes

In order make these general ideas more concrete we introduce some basic Markovian lattice gas models called exclusion processes. From a physics perspective the importance of these models can hardly be understated:

Stochastic lattice gas models, in particular, exclusion processes, are mathematical models amenable to rigorous and numerical analysis which are of fundamental importance for understanding nonequilibrium phenomena in driven diffusive systems.

We start with the simplest model which consists of identical conserved particles with hard-core interaction, viz., the asymmetric simple exclusion process (ASEP) in one space dimension [104].

The Asymmetric Simple Exclusion Process: A Short Review. The ASEP has become a paradigmatic example for a driven diffusive system and has attained a status in the study of nonequilibrium systems somewhat similar to the role that the Ising model plays in equilibrium statistical mechanics. The ASEP is a Markov process in continuous time which in its one-dimensional version can be described informally as follows.

(a) State Space: Each site k of the integer lattice Λ is occupied by at most one particle. This occupation is specified by the random variable $\eta_k \in \{0, 1\}$ indicating whether site k is vacant or occupied. The lattice Λ is a finite or infinite contiguous subset of $\mathbb{Z}$. The set of configurations $\boldsymbol{\eta} := \{\eta_k : k \in \Lambda\}$ is therefore $\Omega = \{0, 1\}^{\Lambda}$. Sometimes these configurations are referred to a microstates.

(b) Bulk Dynamics: Particles hop randomly in continuous time to the right neighboring site (clockwise in case of periodic boundary conditions) with rate r and to the left (anticlockwise) with rate ℓ respectively, provided the target site is empty. Otherwise the attempted move is rejected. Hopping attempts take place independently with an exponential waiting time distribution with mean $\tau_w = 1/(r + \ell)$ (Fig. 1). We present this hopping rule as follows:

$$\begin{aligned} A0 &\rightarrow 0A \text{ with rate } r \\ 0A &\rightarrow A0 \text{ with rate } \ell. \end{aligned} \tag{1.1}$$

Here the symbol A represents occupation by a single particle and 0 represents an empty site. For definiteness we shall assume $r \geq \ell$ corresponding to an average drift on positive lattice direction, or clockwise in case of a finite periodic lattice. As a function of time, the occupation numbers $\eta_k(t)$ describe a single history (i.e., realization) of the stochastic dynamics. We shall write out the time-dependence only where necessary for avoiding confusion.

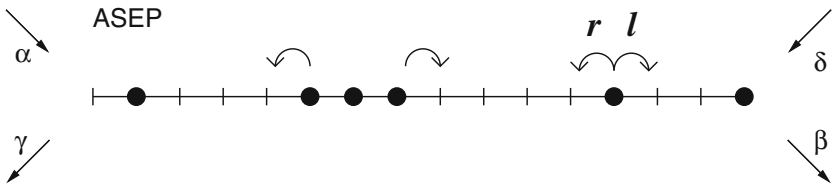


Fig. 1. Pictorial representation of the ASEP with open boundaries and bulk hopping rates r to the right and $l \equiv \ell$ to the left. Some but not all possible jumps are indicated.

(c) Boundary Conditions: For a finite lattice with L sites one has to specify boundary conditions. Most commonly studied are periodic boundary conditions, reflecting boundaries (hopping confined to a box) [67,86], and open boundary conditions [24,62,89] where particles may enter and exit the lattice at the boundary sites 1 and L under the exclusion constraint with rates $\alpha, \beta, \gamma, \delta$ as indicated in Fig. 1. The parametrization $\alpha = r\lambda_- \rho_-$, $\gamma = \ell\lambda_-(1-\rho_-)$ as left boundary rates and $\beta = r\lambda_+(1-\rho_+)$, $\delta = \ell\lambda_+\rho_+$ as right boundary rates may be interpreted as a connection to particle reservoirs with constant density ρ_- at the left boundary and density ρ_+ at the right boundary, respectively. Physically, the parameters $\lambda_\pm$ describe a hopping mechanism between the reservoirs and the chain which differs from the hopping inside the chain unless $\lambda_\pm = 1$.

Terminology: For $\ell = 0$ or $r = 0$ the process is called totally asymmetric simple exclusion process (TASEP). For $r = \ell$ one uses the term symmetric simple exclusion process (SSEP or SEP). If the difference $r - \ell$ is taken to zero in some limiting procedure then one speaks of the weakly asymmetric simple exclusion process (WASEP).

The ASEP was first proposed in an early biophysics context in 1968 as a model to describe the kinetics of protein synthesis through ribosomes moving along m-RNA templates [70,91]. Later it became the "mother" of lattice gas models for automobile traffic [73], see [88] for a thorough discussion of these developments and applications to real biological systems and traffic flow. From a mathematical perspective perhaps the most significant feature is its intimate link to the Kardar-Parisi-Zhang equation [56] for interface growth. Despite being one-dimensional in its simplest formulation it also serves as a model to capture features of driven noisy dynamics in zeolites, carbon nanotubes, artificial narrow channels for colloidal particles, or, via various mappings, for interface dynamics in two dimensions and polymer dynamics and flux lines in three dimensions [93]. As the ASEP has become a paradigmatic reference model we review some of its basic features.

Stationary Distributions. For periodic boundary conditions the total particle number $N = \sum_{k\in\Lambda} \eta_k$ is conserved and the invariant measure for the process with a fixed number of particles is easily seen to be uniform. This fact allows for the construction of a family of invariant measures which are Bernoulli product measures with parameter $\rho = \mathbf{E}_\rho \eta_k$ which is the particle density. We note that this measure is also an invariant measure for the ASEP defined on the infinite lattice $\mathbb{Z}$.

The particle number fluctuations in this product measure are captured by the compressibility

$$K := \sum_{k\in\Lambda} \mathbf{E}_\rho \left[\eta_k(\eta_0 - \rho)\right] = \rho(1-\rho). \tag{1.2}$$

The *instantaneous current*

$$\mathrm{j}(t) := r\eta_k(t)(1-\eta_{k+1}(t)) - \ell(1-\eta_k(t))\eta_{k+1}(t) \tag{1.3}$$

is a time-dependent random number with stationary expectation

$$j(\rho) := \mathbf{E}_\rho \mathrm{j}_k(t) = (r - \ell)\rho(1 - \rho). \tag{1.4}$$

This *stationary current* j tells us the net number of particle that flow across a lattice bond $\langle k, k+1\rangle$ per infinitesimal time interval. The time-integrated current is the random number

$$\mathrm{J}_k(t) := \int_0^t \mathrm{d}s\, \mathrm{j}_k(s). \tag{1.5}$$

Obviously, one has $\mathbf{E}_\rho \mathrm{J}_k(t) = j(\rho)t$.

The open ASEP is ergodic. The exactly known unique invariant measure is non-trivial and exhibits an intriguing phase diagram as a function of the reservoir densities $\rho_\pm$ (Fig. 2). The bulk density ρ undergoes a nonequilibrium discontinuous transition along the line $0 < \rho_- = 1 - \rho_+ < 1/2$ between a low-density phase with bulk density $\rho = \rho_-$ to a high-density phase with bulk density $\rho = \rho_+$. There are nonequilibrium continuous transitions from both phases to a maximal current phase with $\rho = 1/2$, which one has inside the region $1/2 < \rho_- \geq 1, 0 \leq \rho_+ < 1/2$ [23,25,62,66,70,89].

The microscopic density profiles are non-trivial in all phases [25,89]. At the first-order transition line one has phase coexistence with a left domain of density ρ_- and a right domain of density ρ_+, separated by a domain wall. On macroscopic scale this domain wall corresponds to a shock, i.e., a density discontinuity. Inside the maximal current phase the local density decays algebraically from the boundaries to its asymptotic bulk value 1/2. The theory of boundary-induced phase transitions [61,75] explains that these phase transitions arise on microscopic level from the interplay of the shock motion and the flow of local perturbations as described by the so-called dynamical structure function. In this way one understands and extends the hydrodynamic derivation of the phase diagram, first proposed by Krug [62]. Recently, the theory was proved rigorously by Bahadoran [3].

Dynamical Properties. On microscopic scale the local density $\rho_k(t) := \mathbf{E}\eta_k(t)$ starting from some initial measure μ_0 satisfies, due to particle number conservation, away from the boundaries the lattice continuity equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_k(t) = j_{k-1}(t) - j_k(t) \tag{1.6}$$

where $j_k(t) = \mathbf{E}\mathrm{j}_k(t)$ is the expectation of the instantaneous current (1.3). This equation does not allow for an explicit solution on microscopic level. However, it can be proved that on macroscopic Eulerian scale the density profile of the ASEP evolves according to the inviscid Burgers equation [85]

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x} j(\rho) = \frac{\partial}{\partial t}\rho + v_c(\rho)\frac{\partial}{\partial x}\rho = 0 \tag{1.7}$$

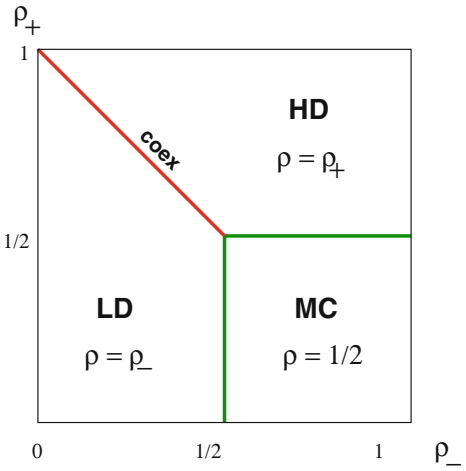


Fig. 2. Phase diagram of the ASEP with open boundaries. LD (HD) denotes the low (high) density phase, and MC the maximal current phase. The red coexistence line marks a discontinuous phase transition between bulk densities ρ_- and ρ_+. The green phase transition lines correspond to a continuous phase transition between bulk densities $\rho_\pm$ and $1/2$.

where here $\rho \equiv \rho(x,t)$ is the coarse-grained local density, $j(\cdot)$ is the stationary current-density relation (1.4) and

$$v_c(\rho) = \frac{\mathrm{d}}{\mathrm{d}\rho} j(\rho) = (r-\ell)(1-2\rho) \tag{1.8}$$

is the collective velocity, also known as speed of the characteristics, that plays a prominent role in the dynamics of the ASEP.

The density develops a travelling shock discontinuity unless the initial density profile is monotonously decreasing. The shock velocity of a shock with constant left density ρ_- and constant right density $\rho_+ > \rho_-$ is given by the Rankine-Hugoniot condition [63]

$$v_s(\rho_+, \rho_-) = \frac{j_+ - j_-}{\rho_+ - \rho_-} \tag{1.9}$$

with the stationary currents $j_\pm = j(\rho_\pm)$ in the two branches of the shock. Looking at diffusive scale into the vicinity of the moving shock position one finds that the shock performs a diffusive motion around its mean position [33] with diffusion coefficient

$$D_s(\rho_+, \rho_-) = \frac{1}{2}\frac{j_+ + j_-}{\rho_+ - \rho_-} \tag{1.10}$$

The shock has been shown to be sharp even on microscopic lattice scale [6,25,32].

More detailed information about fluctuations are captured by the dynamical structure function

$$S_k(t) := \mathbf{E}_\rho\left(\eta_k(t)\eta_0(0)\right) - \rho^2 \tag{1.11}$$

that describes the flow and spreading of local fluctuations (Fig. 3). On large scales it acquires for the ASEP a universal scaling form

$$S(x,t) = \frac{\kappa}{(Et)^{\frac{1}{z}}} f_{PS}\left(\frac{x - v_c t}{(Et)^{\frac{1}{z}}}\right) \tag{1.12}$$

with the universal Prähofer-Spohn scaling function $f_{PS}(\cdot)$ [82], universal dynamical exponent $z = 3/2$ characteristic for the universality of the celebrated Kardar-Parisi-Zhang equation [40,56], the collective velocity $v_c = j'(\rho)$ (1.8) and the non-universal constant $E = |j''|\sqrt{2\kappa}$ where κ is the static compressibility (1.2). This scaling form implies means that the center of mass of a fluctuations travels with collective velocity v_c and spreads with a width that increases in time superdiffusively as $t^{1/z}$. The symmetric process (SSEP) with $r = \ell$ is in the diffusive universality class with dynamical exponent $z = 2$ and Gaussian scaling function and will be discussed in more detail below.

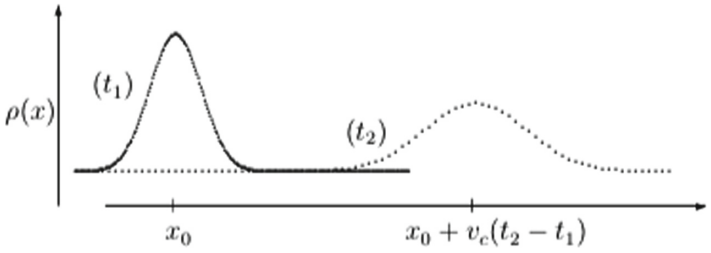


Fig. 3. Schematic plot of the dynamical structure at two times $t_2 > t_1$ with center of mass at lattice x_0 at $t = t_1$.

For completeness, and also because of the important connection with and great current interest in the Kardar-Parisi-Zhang equation, we mention that the time-integrated current, measured in a frame moving with the characteristic velocity, also has a universal scaling form that depends on properties of the initial distribution [2,9,17,18,46,54,64,81,87,107]. Its fluctuations around the mean (1.4) exhibit the KPZ exponent $z = 3/2$.

Multispecies and Multilane Exclusion Processes. Models with more than one conservation law are much less understood, but some exact results on stationary distributions and, in particular, numerical simulations and analytical approximations indicate a wealth of intriguing behaviour. For an older review we refer to [94]. Some recent numerical results will be discussed below and therefore we briefly describe some simple models.

The perhaps simplest particle system with more than one conserved species of particles is a multi-species exclusion process where each lattice site can be

found in three different states: empty, or occupied by either an A-particle or a B-particle. Such an exclusion process is described by the six hopping rates

$$\begin{aligned} A0 &\to 0A \text{ with rate } w_{A0} \\ 0A &\to A0 \text{ with rate } w_{0A} \\ B0 &\to 0B \text{ with rate } w_{B0} \\ 0B &\to B0 \text{ with rate } w_{0B} \\ AB &\to BA \text{ with rate } w_{AB} \\ BA &\to AB \text{ with rate } w_{BA}. \end{aligned} \tag{1.13}$$

Since there are two conservation laws one has two evolution equations for the two local densities

$$\frac{d}{dt}\rho_k^A(t) = j_{k-1}^A(t) - j_k^A(t) \tag{1.14}$$

$$\frac{d}{dt}\rho_k^B(t) = j_{k-1}^B(t) - j_k^B(t) \tag{1.15}$$

where $j_k^{A,B}$ are the expectations of the respective instantaneous currents. The stationary distribution of this process and hence the current-density relations $j_A(\rho^A, \rho^B)$ and $j_B(\rho^A, \rho^B)$ are known only on certain parameter manifolds [12, 94]. For

$$w_{AB} + w_{0A} + w_{0B} = w_{A0} + w_{BA} + w_{B0} \tag{1.16}$$

the canonical measure with fixed particle numbers N_A and N_B is uniform which allows for the construction of a product measure parametrized by densities ρ_A and ρ_B [47, 95].

The stationary currents are then given by

$$j_A(\rho_A, \rho_B) = (w_{A0} - w_{0A})\rho_A(1 - \rho_A) - (w_{B0} - w_{0B})\rho_A\rho_B \tag{1.17}$$

$$j_B(\rho_A, \rho_B) = (w_{B0} - w_{0B})\rho_B(1 - \rho_B) - (w_{A0} - w_{0A})\rho_A\rho_B. \tag{1.18}$$

The hopping asymmetry generates a coupling between the two densities, leading to a non-trivial coupled system

$$\frac{\partial}{\partial t}\rho_A + \frac{\partial}{\partial x} j_A(\rho_A, \rho_B) = 0 \tag{1.19}$$

$$\frac{\partial}{\partial t}\rho_B + \frac{\partial}{\partial x} j_B(\rho_A, \rho_B) = 0 \tag{1.20}$$

of hyperbolic conservation laws for the coarse-grained local densities $\rho_{A,B}(x, t)$. If, however, e.g. $w_{B0} = w_{0B}$ the macroscopic evolution is known: The density of the A-particles evolves autonomously as the in the single-species ASEP and the B-particle density can be integrated straightforwardly [83].

A different way of constructing models with more than one conservation law are coupled multi-lane models where hopping rates on one lane depend on the particle configuration also of other lanes, but no particle exchange between lanes

take place. An interesting class are models where the invariant measure is not changed by the coupling to the other lanes. This can be realized e.g. in a two-lane TASEP by making the rate of jump from site k to site $k+1$ proportional to a linear function of the number of particles on site k and $k+1$ in the adjacent lane, i.e., for $n_k^\alpha := \eta_k^\alpha + \eta_{k+1}^\alpha$ one chooses rates $r_1(n_k^2)$ for lane 1 and $r_2(n_k^1)$ for lane 2 given by [77]

$$r_1(n_k^2) = 1 + \gamma n_k^2/2, \quad r_2(n_k^1) = b + \gamma n_k^1/2, \tag{1.21}$$

see Fig. 4 for illustration. It is easy to see that a product of two Bernoulli product measures is invariant under the stochastic dynamics of this process. One finds the two stationary currents

$$j_1(\rho_1, \rho_2) = \rho_1(1-\rho_1)(1+\gamma\rho_2), \quad j_2(\rho_1, \rho_2) = \rho_2(1-\rho_2)(b+\gamma\rho_1). \tag{1.22}$$

Notice that like (generically) in the single-lane multi-species process described above the currents depend on both densities unless the interaction constant γ vanishes.

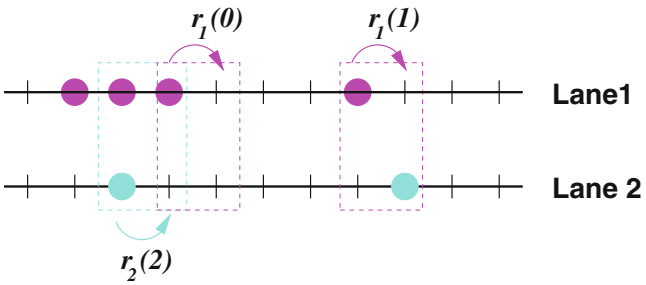


Fig. 4. Two-lane TASEP without hopping between lanes. Some possible jumps and their rates according to (1.21) are shown. The boxes drawn with broken lines indicate on which sites in the neighbouring lane the jump rate depends.

The particle number fluctuations in the Bernoulli product measure are described by the compressibility matrix with matrix elements

$$K_{\lambda\mu} := \frac{1}{L} \langle (N_\lambda - \rho_\lambda L)(N_\mu - \rho_\mu L) \rangle \tag{1.23}$$

Due to the factorized nature of the invariant measure K is diagonal with diagonal elements $K_{\lambda\lambda} = \rho_\lambda(1-\rho_\lambda)$ like in a single-lane TASEP. Two-lane exclusion processes as well as their multilane generalizations serve as prototypical models for studying universal fluctuations which, as will be argued below, are not always diffusive or KPZ.

1.2 Generator of Markov Processes in Matrix Form

We recall the definition of a Markov process η_t with state space Ω and transition rates $w_{\eta',\eta}$ for a transition from a configuration $\eta \in \Omega$ to a configuration $\eta' \in \Omega$ in terms of a generator $\mathcal{L}$ acting on suitably chosen functions $f(\eta)$ through the relation

$$\mathcal{L}f(\eta) = \sum\nolimits_{\eta' \in \Omega \setminus \eta} w_{\eta',\eta}[f(\eta') - f(\eta)]. \tag{1.24}$$

The transition rates can be viewed as matrix elements of the so-called intensity matrix.

Definition 1. *The negative intensity matrix H of the process η_t with state space Ω is the matrix with elements*

$$H_{\eta'\eta} = \begin{cases} -w_{\eta',\eta} & \eta \neq \eta' \\ \displaystyle\sum_{\eta' \in \Omega \setminus \eta} w_{\eta',\eta} & \eta = \eta'. \end{cases} \tag{1.25}$$

Remark 1. The intensity matrix is often represented in transposed form and with opposite sign and also called transition rate matrix. By definition of a transition rate one has $-H_{\eta'\eta} \in \mathbb{R}_0^+$ (positivity of rates) and $\sum_{\eta \in \Omega} H_{\eta'\eta} = 0$ (probability conservation). We shall call any matrix with these properties an intensity matrix.

The link to quantum mechanical condensed matter systems mentioned above is simple: In many cases of interest the intensity matrix is the same object as the quantum Hamiltonian operator of some many-body quantum system. The idea of exploiting this relationship is very simple:

> **One writes the generator of a Markov process in terms of the intensity matrix of transition rates, expresses expectations as bilinear forms, and uses tools from algebra and condensed matter theory for solving probabilistic problems.**

Quantum mechanics as such plays no part in extracting information about properties of the intensity matrix, only the purely mathematical machinery developed for many-body quantum systems comes into play. Here we focus on mathematical techniques from algebra that have proved useful in the treatment of quantum Hamiltonian operators. Other useful techniques from condensed matter theory, in particular the Bethe ansatz, are not discussed here.

The idea of formulating the master equation in terms of a many-body quantum Hamiltonian is not new. Systematic treatments of various aspects of the quantum Hamiltonian formalism go back to [1,28,37,55]. A mathematically rigorous account is given in [69,105] and a detailed (non-rigorous) review is [93]. The extension to infinite systems can usually be made without great difficulty if the state space is countably infinite or by taking appropriate limits of expectation values if the state space of the infinite system is not countable.

Matrix Formulation of the Generator. To exhibit as clearly as possible the essential ideas we shall consider mostly irreducible systems with finite state space Ω. This allows us to straightforwardly adopt the strategy of describing the time evolution of the process by a *master equation* for the probability measure which is the differential form of the Chapman-Kolmogorov equation. Solving the master equation, which is a first-order linear differential equation in the time variable, yields the probability of finding any given state the system may take given that it started from some initial state.

The defining Eq. (1.24) can be written in terms of the negative intensity matrix H as

$$\mathcal{L}f(\eta) = -\sum_{\eta' \in \Omega} f(\eta') H_{\eta'\eta} \tag{1.26}$$

with summation over η on the r.h.s. included. This follows from splitting the sum on the r.h.s. into two terms $-(f(\eta)H_{\eta\eta} + \sum_{\eta' \in \Omega \setminus \eta} f(\eta')H_{\eta'\eta})$ from which one recovers (1.24) by using (1.25). According to (A.2) the r.h.s. of (1.26) represents the left multiplication of the matrix H with a row vector with components $f(\eta')$.

Taking the expectation $\mathbf{E}_\mu$ under a measure μ, one gets from (1.26) (after renaming dummy variables inside the sums)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}_{\mu_t} f = \mathbf{E}_{\mu_t}[\mathcal{L}f] = -\sum_{\eta' \in \Omega} f(\eta') \sum_{\eta'' \in \Omega} H_{\eta'\eta''}\mu_t(\eta'') = \sum_{\eta \in \Omega} f(\eta)\mathcal{L}^T\mu_t(\eta). \tag{1.27}$$

Choosing as $f(\eta)$ the indicator function $1_\eta : \Omega \to \{0,1\}$, $\xi \mapsto 1_\eta(\xi) = \delta_{\eta,\xi}$ the second equality yields

$$\mathcal{L}^T\mu(\eta) = -\sum_{\eta' \in \Omega} H_{\eta\eta'}\mu(\eta') \tag{1.28}$$

where the r.h.s. represents the right multiplication of the matrix H with a column vector with components $\mu(\xi)$. The semigroup property of Markov processes then implies for the time-evolving measure μ_t the *master equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t(\eta) = -\sum_{\eta' \in \Omega} H_{\eta\eta'}\mu_t(\eta') = \sum_{\eta' \in \Omega \setminus \eta} (w_{\eta\eta'}\mu_t(\eta') - w_{\eta'\eta}\mu_t(\eta)) \tag{1.29}$$

which is the adjoint version of (1.26) for functions $f(\eta)$. The quantity

$$j_t(\eta', \eta) := w_{\eta'\eta}\mu_t(\eta) - w_{\eta\eta'}\mu_t(\eta') \tag{1.30}$$

is called the *probability current* from η to η'.

The matrix multiplications (1.26) and (1.28) raise the question of choice of basis for the intensity matrix in computations. We assume Ω to be countable so that to each configuration η one can associate bijectively an integer $\iota(\eta) \in \mathbb{N}$ that enumerates the configurations. We shall call $\iota(\eta)$ the enumeration function. It is natural to choose the canonical basis vectors denoted by $\langle e_i |$ (represented as row vectors with components $(e_i)_j = \delta_{i,j}$ through the bijective map $\eta \mapsto$

$\langle e_{\iota(\eta)}| =: \langle \eta|$ and to define also the column vectors $|\eta\rangle := \langle \eta|^T$. A given enumeration function thus fixes uniquely the matrix H which without explicit enumeration function would be fixed only up to permutations of the canonical basis vectors.

With the canonical basis vectors and an enumeration function at hand we define the function vector

$$\langle f| := \sum_{\eta\in\Omega} f(\eta)\langle \eta| \tag{1.31}$$

and the *probability vector*

$$|\mu(t)\rangle := \sum_{\eta\in\Omega} \mu_t(\eta)|\eta\rangle \tag{1.32}$$

for a time-dependent measure $\mu(t)$.

Observing biorthogonality one realizes that a function f can be expressed as dual pairing $f(\eta) = \langle f|\eta\rangle$ and similarly $\mu_t(\eta) = \langle \eta|\mu(t)\rangle$. These observations allow us to rewrite (1.24) in the form

$$\mathcal{L}f(\eta) = -\langle f|H|\eta\rangle \tag{1.33}$$

and the master equation (1.29) can be written in vector form as

$$\frac{\mathrm{d}}{\mathrm{d}t}|\mu(t)\rangle = -H|\mu(t)\rangle. \tag{1.34}$$

with

$$H = -\sum_{\eta\in\Omega}\sum_{\eta'\in\Omega\setminus\eta} w_{\eta'\eta}\left(E^{\eta'\eta} - \hat{1}_\eta\right) \tag{1.35}$$

where

$$E^{\eta'\eta} := |\eta'\rangle\langle\eta|, \quad \hat{1}_\eta := |\eta\rangle\langle\eta|. \tag{1.36}$$

Integration then expresses the time-dependent measure

$$|\mu(t)\rangle = \mathrm{e}^{-Ht}|\mu\rangle \tag{1.37}$$

in terms of an arbitrary initial measure $\mu = \mu(0)$. In slight abuse of language we shall call also the negative intensity matrix H the generator of the process. We shall call the exponential $\exp(-Ht)$ the transition matrix at time t.[1]

Some comments on the spectrum of H for finite state space Ω are in place. Obviously, $\dim(H) = |\Omega|$. Since H is real all eigenvalues are either real or come in complex conjugate pairs. The negative sign for the off-diagonal elements is by convention. It ensures, by the theorem of Gershgorin [36], that all eigenvalues of H are either 0 or have strictly positive real part. Consequently, the eigenvalues

[1] This equation has the form of a quantum mechanical Schrödinger equation in imaginary time, with H playing formally the role of the quantum Hamiltonian. This fact has given rise to the notion "quantum Hamiltonian formalism".

of the transition matrix $\exp(-Ht)$ are either 1 or strictly inside the unit circle in the complex plane for all times $t \in \mathbb{R}_0^+$. This rules out periodicity of the process. If the process η_t is irreducible then the matrix H is also irreducible and has unique lowest eigenvalue 0. By Perron-Frobenius [72] the corresponding right and left eigenvector can be chosen to have strictly positive real components. More generally, the following two statements on reducible chains are equivalent: (i) Ω has exactly n mutually communicating subsets Ω_α. (ii) The eigenvalue 0 of H is n-fold degenerate. The process restricted to a single communicating subset is ergodic since it is both aperiodic and irreducible.

Expectations. In order to work with this matrix reformulation of the generator we introduce some further key objects. All summations run over the full set Ω unless stated otherwise.

Definition 2 *(a) The summation vector is the constant bra-vector*

$$\langle\, s \,| := \sum_\eta \langle\, \eta \,|. \tag{1.38}$$

(b) The function matrix $\hat{f}$ for a function $f : \Omega \to \mathbb{Z}$ and the measure matrix $\hat{\mu}$ for a probability measure μ are the diagonal matrices

$$\hat{f} := \sum_\eta f(\eta) |\, \eta \,\rangle\langle\, \eta \,|, \quad \hat{\mu} := \sum_\eta \mu(\eta) |\, \eta \,\rangle\langle\, \eta \,|. \tag{1.39}$$

(c) The time-dependent function matrix $\hat{f}(t)$ is defined by

$$\hat{f}(t) := \mathrm{e}^{Ht} \hat{f} \mathrm{e}^{-Ht}. \tag{1.40}$$

If $\hat{f}(t) = \hat{f}(0)$ for all $t \in \mathbb{R}$ we say that f is conserved.

The function matrix for the indicator function 1_η is the projector $\hat{1}_\eta = |\, \eta \,\rangle\langle\, \eta \,|$, i.e. the dyadic product of the canonical basis vector $|\, \eta \,\rangle$ with its transpose. For a strictly positive measure any power $\hat{\mu}^\alpha$ exists. Therefore, in particular, the inverse $\hat{\mu}^{-1}$ exists.

Since by construction in each column of H all matrix elements sum up to zero the summation vector is a left eigenvector of H with eigenvalue 0, i.e.,

$$\langle\, s \,| H = 0. \tag{1.41}$$

This fact expresses conservation of probability since $\langle s| \mu(t) \rangle = \langle\, s \,| \mathrm{e}^{-Ht} |\, \mu \,\rangle = \langle s| \mu \rangle = 1$ with $\mu = \mu_0$. For the function vector $\langle\, f \,|$ (1.31) we have trivially that

$$\langle\, s \,| \hat{f} = \langle\, f \,|. \tag{1.42}$$

This yields for the expectation of a function $f(\eta)$ the various equivalent matrix representations

$$\mathbf{E}_{\mu_t} f \equiv \langle f \rangle_{\mu_t} = \langle\, s \,| \hat{f} |\, \mu(t) \,\rangle = \langle\, s \,| \hat{f} \mathrm{e}^{-Ht} |\, \mu \,\rangle = \langle\, s \,| \hat{f}(t) |\, \mu \,\rangle \equiv \langle f_t \rangle_\mu \tag{1.43}$$

where in the rightmost expression we use the notation $f_t(\eta) = f(\eta_t)$.

The expectation – which we shall denote by angular brackets – is an average both over histories of the process and over the initial distribution μ. Of course, if the initial distribution is concentrated on a particular configuration ξ, the brackets reduce to an average over histories. For a process starting at a configuration ξ the expectation of the indicator function 1_η yields the conditional probability (sometimes called *propagator*)

$$P(\eta, t|\xi, 0) = \langle\, s\,|\hat{1}_\eta \mathrm{e}^{-Ht}|\,\xi\,\rangle = \langle\,\eta\,|\mathrm{e}^{-Ht}|\,\xi\,\rangle = \langle\,\xi\,|\mathrm{e}^{-H^T t}|\,\eta\,\rangle. \tag{1.44}$$

Multi-time expectations can be expressed analogously using the propagator and the Chapman-Kolmogorov equation arising from the Markov property of the process.

Stationarity. One of the most basic questions to ask is the behaviour at late times of the stochastic evolution. If the process is ergodic then the measure in the limit $t \to \infty$ is independent of the initial state and one would like to know for interacting particle systems quantities like the mean density, density fluctuations, or the spatial structure of the particle distribution and its correlations. For transition rates that are constant in time this asymptotic measure is invariant under time translations and hence called stationary. We shall denote any normalized stationary measure by μ^*, its associated probability vector by $|\,\mu^*\,\rangle$ and the diagonal measure matrix by $\hat{\mu}^*$. From the considerations of the previous subsections it is clear that $|\,\mu^*\,\rangle$ is a right eigenvector of H with eigenvalue zero,

$$H|\,\mu^*\,\rangle = 0. \tag{1.45}$$

Trivially, one has

$$|\,\mu^*\,\rangle = \hat{\mu}^*|\,s\,\rangle \tag{1.46}$$

where $|\,s\,\rangle := \langle\, s\,|^T$ has constant components $s_\eta = 1$. If the process is ergodic then $\mu^* = \mu_\infty$ is unique and the diagonal matrix power $(\hat{\mu}^*)^\alpha$ with diagonal elements $(\mu^*(\eta))^\alpha$ exists for every $\alpha \in \mathbb{Z}$.

Symmetry

Definition 3. *Let $S : \Omega \times \Omega \to \mathbb{Z}$ be a function and $\hat{S}$ be a matrix with elements $S_{\eta,\xi} = S(\eta, \xi)$. If $\hat{S}$ satisfies*

$$[H\,,\,\hat{S}] = 0. \tag{1.47}$$

then S is called a symmetry of the process with generator H.

Conservation of f implies the commutation relation $[H\,,\,\hat{f}] = 0$. Therefore a conserved f is a symmetry of the process. Notice that conservation of f implies $\langle\, f\,|H = 0$, which means that f is a harmonic function. The converse, however, is not true: A function may be harmonic, but not conserved. Nevertheless, a non-constant harmonic function implies existence of a conserved function S with the property $\langle\, s\,|\hat{S} = \langle\, f\,|$.

1.3 Time Reversal

For every process with generator H one can define a time-reversed process as follows.

Definition 4 *(Reversed process and ground state transformation). Let μ be a strictly positive stationary solution of the master equation (1.29) for a generator H. Then*

$$H^* := \hat{\mu} H^T \hat{\mu}^{-1} \tag{1.48}$$

is called generator of the time-reversed (or simply reversed) process. The transformation to the matrix $\tilde{H}$ defined by

$$\tilde{H} := \hat{\mu}^{-1/2} H \hat{\mu}^{1/2} \tag{1.49}$$

is called the ground state transformation.

The reversed process evidently has the same invariant measure as the original process since $H^*|\,\mu\,\rangle = \hat{\mu} H^T \hat{\mu}^{-1} |\,\mu\,\rangle = \hat{\mu} H^T |\,s\,\rangle = 0$ where the final equality comes from probability conservation (1.41) of the original process with generator H. Moreover, the reversed process has the same waiting time distribution for all states and the same allowed transitions as the original process H, but with different and often complicated non-local transition rates

$$w^{rev}_{\eta',\eta} = w_{\eta,\eta'} \frac{\mu(\eta')}{\mu(\eta)}. \tag{1.50}$$

The notion of reversibility has its origin in the following property of equilibrium correlation functions.

Proposition 1. *Let H be a generator with strictly positive invariant measure μ and f_1 and f_2 measurable functions of the configurations η and let H^* be the generator of the reversed process. Then*

$$\langle f_1(t) f_2(0) \rangle_\mu = \langle f_2(0) f_1(-t) \rangle^*_\mu \tag{1.51}$$

where the asterisk denotes expectation under the reversed process.

Proof. By definition $\hat{f}_1, \hat{f}_2, \hat{\mu}$ are all diagonal and hence commute and are invariant under transposition. Therefore

$$\begin{aligned}
\langle f_2(t) f_1(0) \rangle_\mu &= \langle\, s\,|\hat{f}_1 \mathrm{e}^{-Ht} \hat{f}_2 |\,\mu\,\rangle \\
&= \langle\, s\,|\hat{f}_1 \mathrm{e}^{-Ht} \hat{\mu} \hat{f}_2 |\,s\,\rangle \\
&= \langle\, s\,|\hat{\mu}^* \hat{f}_1 \mathrm{e}^{-(H^*)^T t} \hat{f}_2 |\,s\,\rangle \\
&= \langle\, s\,|\hat{f}_2 \mathrm{e}^{-H^* t} \hat{f}_1 |\,\mu\,\rangle \\
&= \langle f_2(t) f_1(0) \rangle^*_\mu = \langle f_2(0) f_1(-t) \rangle^*_\mu
\end{aligned} \tag{1.52}$$

where the last equality follows from time-translation invariance of the stationary distribution. □

Time-reversal symmetry can be extended straightforwardly to multi-time correlators.

1.4 Thermal Equilibrium

Often any stationary measure is called an equilibrium measure which is confusing from a physics viewpoint where "equilibrium" has a much more specific meaning than just stationarity. It is linked to the existence of some well-defined energy and time reversal symmetry. In this supplementary section we try to clarify this notion in terms of Markov processes.

In a physical systems defined by an energy $U(\eta)$ of the microstate η the notion of thermal equilibrium at temperature T refers to the situation where stationary measure is the Boltzmann distribution

$$\mu^*(\eta) = \frac{1}{Z} \exp\left(-\beta U(\eta)\right) \tag{1.53}$$

which is proportional to the *Boltzmann weight* $\exp\left(-\beta U(\eta)\right)$. Here $\beta = 1/(k_B T)$ is proportional to the inverse temperature T, k_B is the Boltzmann constant and

$$Z = \sum_{\eta \in \Omega} \exp\left(-\beta U(\eta)\right) \tag{1.54}$$

is the partition function, related to the free energy F by

$$F = -k_B T \ln Z. \tag{1.55}$$

In order to construct a stochastic process for describing an equilibrium setting one therefore has to assure that for a given energy function the Boltzmann distribution (1.53) is stationary.

Stationarity of (1.53), however, is only a sufficient condition to justify the interpretation of a process as an equilibrium process.

The notion of thermal equilibrium is strongly related to symmetry under time reversal, which means that some time-reversed process should be the same as the original forward process. In the following we explain what this reversed process should be.

In classical mechanics governed by Newton's equations of motion the concept of time reversal refers to moving backward in time along a classical trajectory $\Phi(t)$ in the phase space spanned by the coordinates and momenta of the particles, but not along the *same* trajectory as in the forward time evolution, but along a the trajectory with all momenta $p_i(t)$ replaced by $-p_i(t)$. On the other hand, in an overdamped motion where acceleration can be neglected, phase space reduces to the space spanned by the coordinates only. Hence in this case the time-reversed trajectory is the same as the forward trajectory.

These notions have a natural analogue in the context of stochastic dynamics where the phase space corresponds to the state space of the process and a trajectory is a history η_t. Depending on the physical interpretation of the stochastic variables the time-reversed trajectory may then be either the same with time running backwards (if the image of any microstate under time-reversal is the microstate itself like in an overdamped motion) or a different trajectory, if the image of a microstate under time-reversal is some other microstate $\tilde{\eta}$. Clearly

the time-reversal map $C : \Omega \to \Omega$ with $C(\eta) = \tilde{\eta}$ that mimics time reversal on microstates satisfies $C^2(\eta) = \eta$.

This mapping, be it the identity or not, has to be incorporated into the definition of a time-reversed process and thus into time-reversal symmetry necessary for dealing with processes that are models of thermal equilibrium.

Detailed Balance. The simplest way to achieve a time-reversal symmetry is to impose *detailed balance*, defined as follows.

Definition 5 *(Detailed balance). A Markov process η_t with state space Ω and transition rates $w(\eta', \eta)$ is said to satisfy detailed balance (or to be reversible) if there exists a strictly positive measure $\pi(\eta)$ such that*

$$\pi(\eta) w_{\eta',\eta} = \pi(\eta') w_{\eta,\eta'} \ \forall\ \eta, \eta' \in \Omega. \tag{1.56}$$

A measure π with this property is called reversible. If $Z := \sum_\eta \pi(\eta) < \infty$ then Z is called the partition function.

The equilibrium measure entering the detailed balance definition is an invariant measure of the process. This follows immediately from the master equation (1.29) since due to (1.56) each term in the sum over η' on the r.h.s. of (1.29) is equal to zero and therefore the time-derivative of π vanishes. In terms of the time-reversed process H^* defined in (1.48) detailed balance simply means $H^* = H$ so that the time-reversal operation (1.51) of Proposition 1 becomes a symmetry. Therefore, processes satisfying detailed balance are also called equilibrium processes.

Detailed balance means for a Boltzmann distribution that the ratio of transition rates between two microstates η, η' equals the exponential $\exp(-\beta \Delta U)$ of the energy gain $\Delta U = U(\eta') - U(\eta)$ incurred by the transition (Fig. 5). Thus the transition rate ratio is the equilibrium ratio of the probabilities of finding these states.

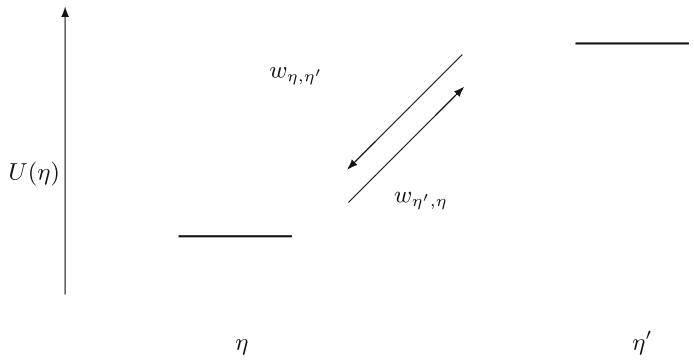


Fig. 5. Stochastic transitions between two states of different equilibrium energies U, U'.

In order to further elaborate on the link between detailed balance and time-reversal symmetry we note the following for the generator H of a reversible process.

Proposition 2. *Let π be a strictly positive measure on the state space Ω. For an ergodic process η_t with generator H the following statements are equivalent:*

(i) The process satisfies detailed balance with reversible measure π.
(ii) $H^T = \hat{\pi}^{-1} H \hat{\pi}$, where H^T is the transpose of H.
(iii) H can be written in the form $H = F\hat{\pi}^{-1}$ for some symmetric negative intensity matrix F.

Proof

(a) Assume (i) is true. By strict positivity π^{-1} exists and the detailed balance condition (1.56) can be recast as $\pi^{-1}(\eta')w(\eta', \eta)\pi(\eta) = w(\eta, \eta')$. This is assertion (ii) in terms of each matrix element.
(b) Assume (ii) is true. Writing out the matrix equation (ii) in terms of each matrix element one gets (i). Moreover, (ii) can be recast as $H\hat{\pi} = \hat{\pi}H^T = (H\hat{\pi})^T$ which implies that $H\hat{\pi}$ is symmetric. That $F := H\hat{\pi}$ is a negative intensity matrix follows from the fact that F has non-positive off-diagonal elements (meaning: non-negative transition rates) and $0 = \langle\, s\,|H\hat{\pi} = \langle\, s\,|F$, which is conservation of probability. Thus (iii) follows from (ii).
(c) Assume (iii) is true. Since F is symmetric it follows that $H^T = \hat{\pi}^{-1}F = \hat{\pi}^{-1}H\hat{\pi}$. Thus (ii) follows from (iii). □

With these notions Proposition 2 has a simple corollary that is worth noting.

Corollary 1. *Let the process η_t with generator H be ergodic and reversible. Then (i) $H^* = H$ and (ii) $\tilde{H}$ is symmetric.*

Thus for a reversible ergodic process the spectrum of H is real and strictly positive except for its unique lowest eigenvalue which is 0.[2]

Remark 2. Detailed balance means that all stationary probability currents (1.30) vanish, thus exposing a direct link between probability currents and reversibility. Notice, however, that a system that does not satisfy detailed balance for the microscopic transition rates may nevertheless be reversible on macroscopic scales. A simple example is a translation-invariant random walk whose increments have zero mean and finite variance. Then by the central limit theorem the large scale behaviour is that of a diffusive particle whose probability distribution satisfies the reversible free diffusion equation, irrespective of whether or not the microscopic increments satisfy detailed balance w.r.t. the stationary uniform measure.

[2] On other words, detailed balance implies that the eigenvalues of the generator are all real and that the related symmetrized generator obtained from the ground state transformation can be interpreted as Hamiltonian of some quantum system. One sees that the use of the term quantum Hamiltonian formalism is justified by more than the formal analogy between Schrödinger equation and master equation.

Generalized Detailed Balance Relation. Mathematically, the reversal of time is encoded in the transposition of the generator. This corresponds to a time-reversal operation where all microstates are mapped onto themselves. In order to describe time-reversal symmetry where the action C on microstates is not simply the identity for all η we must look for other solutions of the master equations that generalize detailed balance. With the matrix formulation of the master equation this problem is straightforward to solve.

Definition 6. *A Markov process η_t with state space Ω and transition rates $w(\eta', \eta)$ is said to satisfy detailed balance (or to be reversible) under conjugation $C : \Omega \to \Omega$ if $C^2(\eta) = \eta$ and if there exists a strictly positive measure $\pi(\eta)$ such that*

$$\pi(\eta) w_{\eta',\eta} = \pi(\eta') w_{C(\eta),C(\eta')} \; \forall \; \eta, \eta' \in \Omega. \tag{1.57}$$

A measure π with this property is called reversible under the conjugation C.

It is easy to see that $\pi(\eta)$ is stationary: In matrix form the generalized detailed balance condition reads

$$H = \hat{\pi} \hat{C} H^T \hat{C} \hat{\pi}^{-1} \tag{1.58}$$

where $\hat{C}$ with matrix elements $\hat{C}_{\eta'\eta} = \delta_{\eta',C(\eta)}$ is the matrix form of the conjugation map. Since $\hat{C}|\, s \,\rangle = |\, s \,\rangle$ it follows that $H|\, \pi \,\rangle = \hat{\pi} \hat{C} H^T \hat{C} |\, s \,\rangle = \hat{\pi} \hat{C} H^T |\, s \,\rangle = 0$. Therefore also processes satisfying the generalized conjugation reversibility may describe thermal equilibrium, provided that the conjugation C has a physical interpretation as time reversal of the microstates.

2 Duality

Duality is a powerful tool in the study of some interacting particles as in some cases it allows for expressing one problem in terms of a much simpler problem. We discuss this property for the SSEP where it was first pointed out by Spitzer in 1970 [104]. Later, by importing known results about quantum spin systems, it was realized that this duality arises from a non-abelian symmetry of the generator [90] known as $SU(2)$ symmetry and eventually extended to the ASEP [92] which has a related symmetry that we shall not discuss in detail. The relationship between symmetries and duality was brought into a neat and systematic form by Giardinà et al. [35] which triggered renewed interest in duality, see also [48] for a survey. The idea we intend to convey in this lecture is summarized as follows.

The explicit form of the generator for lattice gas models in the quantum Hamiltonian formalism, i.e., for a suitable choice of tensor basis of the intensity matrix, often makes explicit non-abelian symmetries that allow for the derivation of non-trivial dualities.

We begin by defining duality and presenting it in matrix form [105].

2.1 Duality and Symmetry

Definition 7. *Let x_t be a Markov process with countable state space Ξ and negative intensity matrix G and η_t be a Markov process with countable state space Ω and negative intensity matrix H. Furthermore, let $D : \Xi \times \Omega \to \mathbb{R}$ be a bounded measurable function. The processes x_t and η_t are said to be dual w.r.t. the duality function D if*

$$\mathbf{E}_x D(x_t, \eta) = \mathbf{E}_\eta D(x, \eta_t). \tag{2.1}$$

The $|\Omega| \times |\Xi|$ matrix

$$\hat{D} := \sum_{x\in\Xi} \sum_{\eta\in\Omega} D(x,\eta) |\, x \,\rangle\langle\, \eta \,| \tag{2.2}$$

with matrix elements $D_{x,\eta} = D(x,\eta)$ is called the duality matrix. For $|\Omega| = |\Xi|$ a duality function of the form $D(x,\eta) = \sum_x d(x)\delta_{x,\eta}$ is called diagonal. If $H = G$ then the process is said to be self-dual w.r.t. D.

Remark 3. In terms of transition probabilities $P(\cdot|\cdot)$ for x_t and $Q(\cdot|\cdot)$ for η_t the defining relation (2.1) reads

$$\sum_{x'\in\Xi} D(x',\eta) P(x',t|x,0) = \sum_{\eta'\in\Omega} D(x,\eta') Q(\eta',t|\eta,0). \tag{2.3}$$

This yields an equivalent formulation of duality in matrix form by taking the time derivative at $t = 0$. With (1.44) one obtains [105]

$$\hat{D}H = G^T \hat{D}. \tag{2.4}$$

Remark 4. A process with strictly positive invariant measure and its reversed are dual w.r.t. the diagonal duality function $D^*(\eta,\eta') = \sum_x \mu^{-1}(\eta)\delta_{\eta,\eta'}$ where $\mu > 0$ is the common invariant measure. This follows directly from the definition (1.48) of the reversed process and the matrix representation $\hat{D} = \hat{\mu}^*$ of the diagonal duality function.

Following [7,35] we show now that symmetries of a generator may lead to non-trivial dualities.

Theorem 1. *Let H be the negative intensity matrix of an ergodic Markov process η_t with countable state space and H^{rev} be the negative intensity matrix of the reversed process x_t. Assume that there exists an intertwiner $\hat{S}$ such that*

$$\hat{S}H = H^{rev}\hat{S}. \tag{2.5}$$

Then H is self-dual with duality function $D(x,\eta) = D_{x,\eta}$ given by the matrix elements of the duality matrix

$$\hat{D} = \hat{\mu}^{-1}\hat{S}. \tag{2.6}$$

with the diagonal stationary distribution matrix of Definition 2.

Proof. Given the hypothesis (2.5), self-duality with duality matrix (2.6) follows from the chain of equalities

$$\hat{D}H = \hat{\mu}^{-1}\hat{S}H = \hat{\mu}^{-1}H^{rev}\hat{S} = \hat{\mu}^{-1}H^{rev}\hat{\mu}\hat{D} = H^T\hat{D}. \tag{2.7}$$

The first and the third equality are the definition (2.6), the second equality is the hypothesis (2.5) of the theorem, and the fourth equality is the definition (1.48) of the reversed process. □

Remark 5. It follows that if H is reversible then the hypothesis (2.5) reads $\hat{S}H = H\hat{S}$, i.e. according to (1.47) $\hat{S}$ is a symmetry of H.

Corollary 2. *Let H be the negative intensity matrix of an ergodic Markov process η_t with countable state space and strictly positive invariant measure μ and S be a symmetry of H. Then H and H^{rev} are dual w.r.t. the duality function $D(\eta, \eta') = \mu^{-1}(\eta)S(\eta, \eta')$.*

2.2 The Symmetric Simple Exclusion Process

Above we have introduced in an informal fashion the ASEP on the one-dimensional integer lattice. For symmetric hopping rates $r = \ell =: w$ the process randomly interchanges the occupation variables of a pair of sites. This has a natural generalization to arbitrary graphs and link-dependent hopping rates and can then informally be described as follows. Let $\Gamma = (\Lambda, \Upsilon)$ be a finite graph with nodes $k \in \Lambda$ and undirected links $\langle k, l\rangle \in \Upsilon$. A configuration of the SSEP is denoted by $\boldsymbol{\eta} := \{\eta_k : k \in \Lambda\}$ with the $L = |\Lambda|$ occupation numbers $\eta_k \in \{0, 1\}$. Each link $\langle k, l\rangle$ carries a "clock" that rings after an exponentially distributed random time with parameter $w_{kl} \equiv w_{lk}$. When the clock rings the occupation numbers η_k and η_l are interchanged, corresponding to a particle jump across bond $\langle k, l\rangle$ if one of the two sites is occupied and the other is empty.

We first derive the intensity matrix process for this general SSEP. From the form of the intensity matrix (2.21) that is obtained in the tensor basis (2.13) used below it becomes evident that this process has a non-Abelian symmetry under the Lie algebra $\mathfrak{su}(2)$. This implies that its generator H commutes with a suitably chosen the representation matrices of the Lie algebra and thus allows for the computation of duality functions.

Generator of the SSEP in Matrix Form

Definition 8. *Let Λ be a finite set of cardinality L, $\eta_j \in \{0, 1\}$ for $j \in \Lambda$ the occupation number of an exclusion process, $\Omega_L = \{0, 1\}^L$, the state space and $\boldsymbol{\eta} = \{\eta_j : j \in \Lambda\}$ be a configuration of an exclusion process. For a pair $\langle k, l\rangle \in \Lambda \times \Lambda$ the $\langle k, l\rangle$-permutation of a configuration $\boldsymbol{\eta} \in \Omega_L$ is the mapping $\pi^{kl} : \Omega_L \to \Omega_L$ such that $\pi^{kl}(\boldsymbol{\eta}) \mapsto \boldsymbol{\eta}^{kl}$ with interchanged occupation numbers*

$$\eta_j^{kl} = \eta_j + (\eta_k - \eta_l)\,(\delta_{j,l} - \delta_{k,l})\,. \tag{2.8}$$

The informal description of the SSEP on the Graph Γ means that the transition rates are given by

$$w_{\boldsymbol{\eta}',\boldsymbol{\eta}} = \sum_{\langle k,l\rangle\in\Upsilon} w_{kl}\left(\eta_k(1-\eta_l)+(1-\eta_k)\eta_l\right)\delta_{\boldsymbol{\eta}',\boldsymbol{\eta}^{kl}} \tag{2.9}$$

for all links $\langle k,l\rangle$. Thus the generator reads

$$\mathcal{L}f(\boldsymbol{\eta}) = \sum_{\langle k,l\rangle\in\Upsilon} w_{kl}\left[f(\boldsymbol{\eta}^{kl})-f(\boldsymbol{\eta})\right] = \sum_{\boldsymbol{\eta}'\in\Omega_L}\sum_{\langle k,l\rangle\in\Upsilon} w_{kl}\left(\delta_{\boldsymbol{\eta}',\boldsymbol{\eta}^{kl}}-\delta_{\boldsymbol{\eta}',\boldsymbol{\eta}}\right)f(\boldsymbol{\eta}') \tag{2.10}$$

from which one reads off the matrix elements

$$H_{\boldsymbol{\eta}'\boldsymbol{\eta}} = -\sum_{\langle k,l\rangle\in\Upsilon} w_{kl}\left(\delta_{\boldsymbol{\eta}',\boldsymbol{\eta}^{kl}}-\delta_{\boldsymbol{\eta}',\boldsymbol{\eta}}\right) \tag{2.11}$$

of the negative intensity matrix H of the SSEP.

In order to fix the canonical basis vectors $\langle\,\boldsymbol{\eta}\,| = \langle\, e_{\iota(\boldsymbol{\eta})}\,|$ and $|\,\boldsymbol{\eta}\,\rangle = \langle\,\boldsymbol{\eta}\,|^T$ for the intensity matrix we choose the enumeration function

$$\iota(\boldsymbol{\eta}) = 1+\sum_{k=1}^{L}\eta_k 2^{L-k}. \tag{2.12}$$

Thus $\iota(\boldsymbol{\eta})$ is the decimal value plus 1 of the binary number $\eta_1,\eta_2,\ldots,\eta_L$. By the definition of the Kronecker product Definition 13 this choice of enumeration function corresponds to the tensor basis

$$\langle\,\boldsymbol{\eta}\,| = \langle\,\eta_1,\ldots,\eta_L\,| \equiv \langle\,\eta_1\,|\otimes\cdots\otimes\langle\,\eta_L\,| \tag{2.13}$$

with the one-site basis vectors

$$\langle\,\eta_k\,| = (1-\eta_k,\eta_k). \tag{2.14}$$

This yields the constant summation vector in the tensor form

$$\langle\,s\,| = (1,1)^{\otimes L}. \tag{2.15}$$

In order to write the generator as a matrix it is useful to introduce the unit matrix $\mathbb{1}$ and three Pauli matrices

$$\sigma^x = \begin{pmatrix}0&1\\1&0\end{pmatrix}, \sigma^y = \begin{pmatrix}0&-i\\i&0\end{pmatrix}, \sigma^z = \begin{pmatrix}1&0\\0&-1\end{pmatrix}, \tag{2.16}$$

where i is the imaginary unit. From these we construct the so-called spin-lowering and raising operator

$$\sigma^+ = \frac{1}{2}\left(\sigma^x+i\sigma^y\right) = \begin{pmatrix}0&1\\0&0\end{pmatrix}, \quad \sigma^- = \frac{1}{2}\left(\sigma^x-i\sigma^y\right) = \begin{pmatrix}0&0\\1&0\end{pmatrix} \tag{2.17}$$

which are nilpotent of degree 2 and the projectors

$$\hat{n} = \frac{1}{2}(\mathbb{1} + \sigma^z) = \begin{pmatrix} 0\,0 \\ 0\,1 \end{pmatrix}, \quad \hat{v} = \frac{1}{2}(\mathbb{1} - \sigma^z) = \begin{pmatrix} 1\,0 \\ 0\,0 \end{pmatrix} \tag{2.18}$$

on a particle and vacancy vector respectively.

From the action of these matrices on the single-site basis vectors one reads off

$$(\mathbf{1} + \sigma_k^+\sigma_l^- + \sigma_k^-\sigma_l^+ - \hat{n}_k\hat{v}_l - \hat{v}_k\hat{n}_l)|\,\boldsymbol{\eta}\,\rangle = |\,\boldsymbol{\eta}^{kl}\,\rangle. \tag{2.19}$$

The orthogonality relations $\langle\boldsymbol{\eta}'|\,\boldsymbol{\eta}^{kl}\rangle = \delta_{\boldsymbol{\eta}',\boldsymbol{\eta}^{kl}}$ and $\langle\boldsymbol{\eta}'|\,\boldsymbol{\eta}\rangle = \delta_{\boldsymbol{\eta}',\boldsymbol{\eta}}$ then yields from the matrix elements (2.11) the matrix representation

$$H_{\boldsymbol{\eta}',\boldsymbol{\eta}} = -\sum_{\langle k,l\rangle} w_{kl}\langle\,\boldsymbol{\eta}'\,|(\sigma_k^+\sigma_l^- + \sigma_k^-\sigma_l^+ - \hat{n}_k\hat{v}_l - \hat{v}_k\hat{n}_l))|\,\boldsymbol{\eta}\,\rangle. \tag{2.20}$$

of the generator of the SSEP in terms of spin operators. In quantum mechanics this matrix is known as the Hamiltonian of the spin-1/2 Heisenberg ferromagnet. We can write

$$H = \sum_{\langle k,l\rangle} w_{kl}h_{kl} \tag{2.21}$$

with the hopping matrices

$$h_{kl} = -\left(\sigma_k^+\sigma_l^- + \sigma_k^-\sigma_l^+ - \hat{n}_k\hat{v}_l - \hat{v}_k\hat{n}_l\right) \tag{2.22}$$

$$= -\frac{1}{2}\left(\sigma_k^x\sigma_l^x + \sigma_k^y\sigma_l^y + \sigma_k^z\sigma_l^z - \mathbf{1}\right). \tag{2.23}$$

Equilibrium Measures. Since H is symmetric it follows that $|\,u\,\rangle = \langle\,s\,|^T$ is a stationary measure. Moreover, the SSEP obviously satisfies detailed balance (1.56) w.r.t. this measure. Thus the uniform measure

$$|\,u^*\,\rangle := \frac{1}{|\Omega_L|}|\,s\,\rangle = 2^{-L}|\,s\,\rangle \tag{2.24}$$

is an equilibrium measure with trivial energy $U(\boldsymbol{\eta})$ that does not depend on the configuration $\boldsymbol{\eta}$. Since particle number is conserved the SSEP defined on Ω is trivially non-ergodic. However, since the dynamics is a sequence of permutations, the SSEP restricted to the state space $\Omega_{L,N} := \{\boldsymbol{\eta} \in \Omega_L : \sum_{k\in\Lambda}\eta_k = N\}$ of fixed particle number $N \in \{0,\ldots,L\}$ is ergodic. Since there are $|\Omega_{L,N}| = \binom{L}{N}$ ways of distributing N exclusion particles on L sites, the *canonical* uniform measure

$$|\,u_{L,N}\,\rangle = \sum_{\boldsymbol{\eta}\in\Omega_{L,N}} |\,\boldsymbol{\eta}\,\rangle, \quad |\,u^*_{L,N}\,\rangle = \frac{1}{Z}\sum_{\boldsymbol{\eta}\in\Omega_{L,N}} |\,\boldsymbol{\eta}\,\rangle \tag{2.25}$$

with canonical partition function

$$Z_{L,N} = \binom{L}{N} \tag{2.26}$$

is the unique equilibrium measure $u^*_{L,N}$ on $\Omega_{L,N}$. One has for $\boldsymbol{\eta} \in \Omega_L$

$$u^*_{L,N}(\boldsymbol{\eta}) = \binom{L}{N}^{-1} \delta_{N,N(\boldsymbol{\eta})} \tag{2.27}$$

where

$$N(\boldsymbol{\eta}) = \sum_{k\in\Lambda} \eta_k \tag{2.28}$$

is the number of particles in the configuration $\boldsymbol{\eta}$. The canonical partition function (2.26) yields the canonical free energy

$$F_{L,N} = -\ln Z_{L,N}. \tag{2.29}$$

Clearly, any normalized convex combination of the unnormalized canonical invariant measure $u_{L,N}(\boldsymbol{\eta}) := \delta_{N,N(\boldsymbol{\eta})}$ defines an equilibrium measure. Of particular importance is the *grandcanonical measure*

$$\pi^*_{L,\phi}(\boldsymbol{\eta}) := \frac{1}{Z_L(\phi)} \sum_{N=0}^{L} \mathrm{e}^{\varphi N} u_{L,N}(\boldsymbol{\eta}) \tag{2.30}$$

with so-called chemical potential ϕ and *grandcanonical partition function*

$$Z_L(\phi) := \sum_{\boldsymbol{\eta}\in\Omega_L} \sum_{N=0}^{L} \mathrm{e}^{\phi N} u_{L,N}(\boldsymbol{\eta}) = \sum_{N=0}^{L} \mathrm{e}^{\phi N} Z_{L,N} = \left(1+\mathrm{e}^{\phi}\right)^L. \tag{2.31}$$

The simple form of this partition function comes from the fact that the grandcanonical measure can be written in product form as

$$\pi^*_{L,\phi}(\boldsymbol{\eta}) := \frac{1}{Z_L(\phi)} \prod_{k\in\Lambda} \left(1-\eta_k+\mathrm{e}^{\phi}\eta_k\right) \tag{2.32}$$

which is a Bernoulli product measure.

By construction the particle number $N(\boldsymbol{\eta})$ in this *grandcanonical ensemble* of configuration is not a fixed number even though for any given realization of the process it is. Instead one has

$$\rho(\phi) := \frac{\langle N\rangle_\phi}{L} = \frac{1}{L} \sum_{\boldsymbol{\eta}\in\Omega_L} N(\boldsymbol{\eta}) \pi_{L,\phi}(\boldsymbol{\eta}) = \frac{1}{L} \frac{\mathrm{d}}{\mathrm{d}\phi} \ln Z_L(\phi) = \frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \tag{2.33}$$

Defining the inverse function

$$\phi(\rho) = \ln\rho - \ln(1-\rho) \tag{2.34}$$

one finds for the composite function $\tilde{Z}_L(\rho) = (Z_L \circ \phi)(\rho)$ the density dependence

$$\tilde{Z}_L(\rho) = Z_L(\phi(\rho)) = (1-\rho)^{-L}. \tag{2.35}$$

of the grandcanonical partition function and the corresponding ρ-parametrization

$$\tilde{\pi}^*_{L,\rho}(\boldsymbol{\eta}) := \pi^*_{L,\phi(\rho)}(\boldsymbol{\eta}) = \sum_{N=0}^{L} (1-\rho)^{L-N} \rho^N u_{L,N}(\boldsymbol{\eta}) \tag{2.36}$$

of the grandcanonical measure (2.30).

One realizes that – as expected – the grandcanonical free energy

$$G(L,\phi) := -\ln Z_{L,\phi} = -L \ln (1 + \mathrm{e}^{\phi}) \tag{2.37}$$

is extensive in L. The associated canonical free energy defined by the Legendre transform

$$F(L,\rho) = G(L,\phi(\rho)) + L\rho\phi(\rho) = L\left[(1-\rho)\ln(1-\rho) + \rho \ln \rho\right] \tag{2.38}$$

is given the thermodynamic limit (2.29)

$$\lim_{L\to\infty} \frac{1}{L} F_{L,\rho L} = (1-\rho)\ln(1-\rho) + \rho \ln \rho \tag{2.39}$$

of the canonical free energy density. This indicates the well-known equivalence of the canonical ensemble with $N = \rho L$ particles and the grandcanonical ensemble at density ρ in the thermodynamic limit.

The grandcanonical probability vector $|\, \pi^*_{L,\phi} \,\rangle$ is obtained from (2.25). Defining the unnormalized canonical stationary probability vector

$$|\, u_{L,N} \,\rangle = \sum_{\boldsymbol{\eta} \in \Omega_{L,N}} |\, \boldsymbol{\eta} \,\rangle \tag{2.40}$$

and the particle number operator

$$\hat{N} = \sum_{\boldsymbol{\eta} \in \Omega_L} N(\boldsymbol{\eta}) |\, \boldsymbol{\eta} \,\rangle\langle\, \boldsymbol{\eta} \,| \tag{2.41}$$

one gets $f(N)|\, u_{L,N} \,\rangle = f(\hat{N})|\, u_{L,N} \,\rangle$ since each component in $|\, \pi_{L,N} \,\rangle$ with non-zero weight has exactly N particles which means that $N|\, u_{L,N} \,\rangle = \hat{N}|\, u_{L,N} \,\rangle$.

Thus

$$
\begin{aligned}
|\,\pi^*_{L,\phi}\,\rangle &= Z_L^{-1}(\phi)\sum_{N=0}^{L} \mathrm{e}^{\phi N}|\,u_{L,N}\,\rangle \\
&= Z_L^{-1}(\phi)\sum_{N=0}^{L} \mathrm{e}^{\phi \hat{N}}|\,u_{L,N}\,\rangle \\
&= Z_L^{-1}(\phi)\mathrm{e}^{\phi \hat{N}}\sum_{N=0}^{L}\sum_{\boldsymbol{\eta}\in\Omega_{L,N}} |\,\boldsymbol{\eta}\,\rangle \\
&= Z_L^{-1}(\phi)\mathrm{e}^{\phi\sum_{k=1}^{L}\hat{n}_k}|\,u\,\rangle \\
&= Z_L^{-1}(\phi)\prod_{k=1}^{L}(\hat{v}_k + \mathrm{e}^{\phi}\hat{n}_k)|\,u\,\rangle \\
&= \left(1+\mathrm{e}^{\phi}\right)^{-L}\left((1,\mathrm{e}^{\phi})^T\right)^{\otimes L} \\
&= \frac{1}{(1+\mathrm{e}^{\phi})^L}\begin{pmatrix}1\\ \mathrm{e}^{\phi}\end{pmatrix}^{\otimes L} = \begin{pmatrix}1-\rho(\phi)\\ \rho(\phi)\end{pmatrix}^{\otimes L}
\end{aligned}
\tag{2.42}
$$

which is an L-fold tensor product.

This tensor structure of the grandcanonical probability vector makes the computation of correlations trivial. From (A.17) one has

$$
\langle \eta_{k_1}\dots\eta_{k_m}\rangle_{\phi} = \rho^m(\phi) \tag{2.43}
$$

when all k_i are mutually different. Therefore one finds the static structure function

$$
C_{k,l} := \langle \eta_k\eta_l\rangle_{\phi} - \rho^2(\phi) = \rho(\phi)(1-\rho(\phi))\delta_{k,l}. \tag{2.44}
$$

This yields the compressibility

$$
K(\rho) = \frac{1}{L}\sum_{k\in\Lambda}\sum_{l\in\Lambda} C_{k,l} = \frac{1}{L}\langle (N-\rho L)^2\rangle = \rho(1-\rho). \tag{2.45}
$$

Of course, this result could directly have been obtained from the usual thermodynamic relation

$$
\tilde{K}(\phi) = \frac{\mathrm{d}}{\mathrm{d}\phi}\rho(\phi) = \frac{\mathrm{e}^{\phi}}{(1+\mathrm{e}^{\phi})^2} \tag{2.46}
$$

and using (2.34).

Duality Functions for the SSEP. From the structure of the hopping matrices in (2.23) it is clear that the generator is symmetric under the action of the Lie-algebra $\mathfrak{su}(2)$ [5], i.e., H satisfies the commutation relations

$$
[H\,,\,S^{\pm}] = [H\,,\,S^{z}] = 0 \tag{2.47}
$$

with the representation matrices

$$S^{\pm} = \sum_{k\in\Lambda} \sigma_k^{\pm}, \quad S^z = \frac{1}{2}\sum_{k\in\Lambda} \sigma_k^z \tag{2.48}$$

which satisfy the $\mathfrak{su}(2)$ commutation relations

$$[S^+\,,\, S^-] = 2S^z, \quad [S^z\,,\, S^{\pm}] = \pm S^{\pm}. \tag{2.49}$$

Using the symmetry approach to duality discussed above, the well-known self-duality of the SSEP [67,104] restated and generalized in the following theorem becomes a trivial corollary of the $\mathfrak{su}(2)$ symmetry.

Theorem 2. *The SSEP on a lattice Λ is selfdual w.r.t. the duality function*

$$D(\boldsymbol{\zeta},\boldsymbol{\eta}) = \prod_{k\in\Lambda} (\alpha + \beta\eta_k)^{\gamma+\delta\zeta_k} \tag{2.50}$$

for configurations $\boldsymbol{\eta},\boldsymbol{\zeta} \in \{0,1\}^{\Lambda}$ and $\alpha,\beta,\gamma,\delta \in \mathbb{R}$, provided that $N(\boldsymbol{\eta}) < \infty$ if $\gamma \neq 0$ and $N(\boldsymbol{\zeta}) < \infty$ if $\delta \neq 0$.

Remark 6. For $\gamma = 0$ the duality function (2.50) can be written in alternative form as follows. Let $\mathbf{x}(\boldsymbol{\zeta}) := \{k : \zeta_k = 1\}$ be the set of occupied sites $x_i \in \Lambda$ of the configuration $\boldsymbol{\zeta}$ and $N(\mathbf{x}) = |\mathbf{x}|$ be the number of particles in the configuration $\mathbf{x}$. This mapping induces an obvious bijection between the state space $\Omega = \{0,1\}^{\Lambda}$ and the coordinate set Ξ of possible distinct occupied sites and thus allows for describing the SSEP in terms of the evolution $\mathbf{x}_t$ of particle coordinates. With $a = \alpha^{\delta}$, $b = (\alpha+\beta)^{\delta} - \alpha^{\delta}$ the duality function (2.50) then becomes

$$\tilde{D}(\mathbf{x},\boldsymbol{\eta}) = \prod_{i=1}^{N(\mathbf{x})} (a + b\eta_{x_i}) \tag{2.51}$$

for all $\mathbf{x} \in \Xi$ and $\boldsymbol{\eta} \in \Omega$. For $\alpha = 0$, $\beta = \delta = 1$ corresponding to $a = 0$ and $b = 1$ one recovers the well-known duality function formulated and proved in a different way in [67] and which goes back to [104].

Proof. The $\mathfrak{su}(2)$-symmetry implies that the L-fold Kronecker product $\hat{D} = A^{\otimes L}$ is a symmetry operator for any $\times 2$ matrix A. Since the SSEP is reversible with uniform invariant measure (2.24) this yields the duality function $D(\boldsymbol{\zeta},\boldsymbol{\eta}) = \langle\, \boldsymbol{\zeta}\,|\hat{D}|\,\boldsymbol{\eta}\,\rangle$. The factorization of the symmetry operator and also of the dual pairing (see (A.17)) yields

$$D(\boldsymbol{\zeta},\boldsymbol{\eta}) = \prod_{k\in\Lambda} \langle\, \zeta_k\,|A|\,\eta_k\,\rangle \tag{2.52}$$

Explicit computation of the two-dimensional bilinear form

$$\langle\, \zeta_k\,|A|\,\eta_k\,\rangle = (1-\zeta_k, \zeta_k) \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} 1-\eta_k \\ \eta_k \end{pmatrix} \tag{2.53}$$

yields $(\alpha+\beta\eta_k)^{\gamma+\delta\zeta_k}$ with $A_{11} = \alpha^{\gamma}$, $A_{12} = (\alpha+\beta)^{\gamma}$, $A_{12} = \alpha^{\gamma+\delta}$, $A_{22} = (\alpha+\beta)^{\gamma+\delta}$. □

Remark 7. The duality function (2.50) is not unique. Any measurable function of the symmetry operators $S^{\pm,z}$ (2.48) yields a duality function.

One realizes that the mapping to the quantum Hamiltonian immediately reveals the well-known $\mathfrak{su}(2)$ symmetry of the generator of the SSEP and therefore provides instantly self-duality functions. Moreover, the matrix formulation reduces to proof of selfduality to elementary multilinear algebra. The $\mathfrak{su}(2)$ symmetry allows for the derivation of similarly strong results for multi-time correlation functions $\langle n_{i_1}(t_1)\dots n_{i_k}(t_k)\rangle$.

Remark 8. Any Markov process whose generator is a function of the hopping matrices $e_{k,l} = \sigma_k^x\sigma_l^x + \sigma_k^y\sigma_l^y + \sigma_k^z\sigma_l^z - \mathbf{1}$ is $\mathfrak{su}(2)$ symmetric and therefore self-dual w.r.t. the same duality functions as the SSEP.

Remark 9. The approach can be straightforwardly generalized to the symmetric partial exclusion process [59,90]. The partial exclusion process is the spin-s version of this model where each lattice site i can be occupied by at most $2s_i$ particles and where single-particle hopping from site i to site j occurs with rate $n_i(2s_j - n_j)$.

Density Profile and Dynamical Structure Function. We focus now on the case $\alpha = \gamma = 0$ and $\beta = \delta = 1$ in the duality function (2.50). The self-duality has the remarkable consequence that for any initial measure with support on configurations with any number of particles the joint expectations of n occupation numbers can be expressed in terms of transition probabilities for initial states with only n particles. In particular, for the density profile $\rho_k(t) = \mathbf{E}_{\mu_t}\eta_k$ one finds by inserting the duality function in the form (2.51) for $N = 1$ into the definition (2.3) of duality. Inserting $D(x,\boldsymbol{\eta}) = \eta_x$ into the r.h.s. of (2.3) and using the propagator representation (1.44) of the transition probability yields for an initial configuration $\boldsymbol{\eta}$

$$\sum_{\boldsymbol{\eta}'}\eta_x\langle\,\boldsymbol{\eta}'\,|\mathrm{e}^{-Ht}|\,\boldsymbol{\eta}\,\rangle = \sum_{\boldsymbol{\eta}'}\langle\,\boldsymbol{\eta}'\,|\hat{n}_x\mathrm{e}^{-Ht}|\,\boldsymbol{\eta}\,\rangle = \langle\,s\,|\hat{n}_x\mathrm{e}^{-Ht}|\,\boldsymbol{\eta}\,\rangle \tag{2.54}$$

since $\langle\,\boldsymbol{\eta}'\,|\hat{n}_x = \eta_x\langle\,\boldsymbol{\eta}'\,|$. On the other hand, the l.h.s. of (2.3) becomes

$$\sum_{x'\in\Lambda}\eta_{x'}P(x',t|x,0) = \sum_{x'\in\Lambda}\langle\,s\,|\hat{n}_{x'}|\,\boldsymbol{\eta}\,\rangle P(x,t|x',0) \tag{2.55}$$

since for a single particle one has $\Xi = \Lambda$ and since the generator for the SSEP is symmetric.

This yields for an arbitrary initial measure μ the density profile

$$\rho_x(t) = \langle\,s\,|\eta_x\mathrm{e}^{-Ht}|\,\mu\,\rangle = \sum_{x'\in\Lambda}\rho_{x'}(0)P(x,t|x',0). \tag{2.56}$$

This means that irrespective of the lattice and of the jump rates between lattice points the time evolution of the local density in a system of any number

of interacting particles is completely determined by the (non-interacting) time evolution of a single particle. Specifically, on the d-dimensional hypercubic lattice $\mathbb{Z}^d$ with translation-invariant nearest-neighbour hopping the single-particle propagator satisfies a discrete diffusion equation which can be solved in explicit form in terms of modified Bessel functions

$$I_n(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}p \; \mathrm{e}^{ipn - t\cos p}. \tag{2.57}$$

On $\mathbb{Z}^d$ with hopping rates w_i in each direction one then has for point $\mathbf{x} = (x_1, \ldots, x_d)$

$$\rho_{\mathbf{x}}(t) = \prod_{j=1}^{d} \sum_{x'_j \in \mathbb{Z}} \rho_{x'_j}(0) \mathrm{e}^{-w_j t} I_{x_j - x'_j}(w_j t). \tag{2.58}$$

As a corollary of (2.56) we note

$$\langle\, s \,|\eta_x \mathrm{e}^{-Ht} = \sum_{x' \in \Lambda} P(x,t|x',0) \langle\, s \,|\hat{n}_{x'}. \tag{2.59}$$

For the dynamical structure function defined by

$$S_{x,y}(t) := \mathbf{E}_\rho \left(\eta_x(t)\eta_y(0)\right) - \rho^2 \tag{2.60}$$

this yields

$$\begin{aligned} S_{x,y}(t) &= \langle\, s \,|\eta_x \mathrm{e}^{-Ht} \eta_y |\, \rho \,\rangle - \rho^2 \\ &= \sum_{x' \in \Lambda} P(x,t|x',0) \langle\, s \,|\hat{n}_{x'} \hat{n}_y |\, \rho \,\rangle - \rho^2 \\ &= \sum_{x' \in \Lambda} P(x,t|x',0) \left(\rho^2 + \rho(1-\rho)\delta_{x',y}\right) - \rho^2 \\ &= \rho(1-\rho) P(x,t|y,0) \end{aligned} \tag{2.61}$$

where we have used reversibility of the SSEP and conservation of probability which gives $\sum_{x' \in \Lambda} P(x,t|x',0) = \sum_{x' \in \Lambda} P(x',t|x,0) = 1$ for the single-particle process.

On the translation-invariant hypercubic lattice with nearest-neighbour jumps with rates w_i in direction i the dynamical structure function $S_{\mathbf{x}}(t) := S_{\mathbf{x},\mathbf{0}}(t)$ becomes

$$S_{\mathbf{x}}(t) = \prod_{j=1}^{d} \mathrm{e}^{-2w_j t} I_{x_j - x'_j}(2w_j t). \tag{2.62}$$

In the scaling limit $x_i(t) = r_i \sqrt{4w_i t}$ and $t \to \infty$ the modified Bessel function becomes a Gaussian. Thus

$$\prod_{j=1}^{d} \sqrt{4\pi w_j} \lim_{t \to \infty} t^{d/2} S_{\mathbf{x}(t)}(t) = \mathrm{e}^{-\sum_{j=1}^{d} r_j^2}. \tag{2.63}$$

We read off the dynamical exponent $z = 2$ and the universal Gaussian scaling function with diagonal diffusion matrix $D_{ij} = 2w_i\delta_{ij}$.

Higher order correlation functions can be studied using the Bethe ansatz [39,58,93]. One finds that *all n-point correlation functions of the symmetric exclusion process are, to leading order in time, identical to the same n-point correlators of non-interacting particles.* Corrections are of order $1/\sqrt{t}$, see [19] for a related rigorous result. Hence diffusive scaling with dynamical exponent $z = 2$ leaves finite-order correlation functions invariant up to an overall amplitude.

2.3 Selfduality of the 1-D ASEP

The ASEP on the graph Γ is the asymmetric generalization of the SSEP with directed hopping rates w_{kl} for jumps from site k to site l and w_{lk} for the reversed jump. Little is known about this process on general graphs where it does not have a symmetry analogous to the $\mathfrak{su}(2)$-symmetry of the SSEP and where not even the invariant measure is known. We restrict our attention to the most-studied one-dimensional finite integer lattice $\Lambda = [L^-, L^+] \setminus \mathbb{Z}$ with nearest-neighbour jumps with rates $r_k \equiv w_{kk+1} > 0, \ell_{k+1} \equiv w_{k+1k} > 0$ for constant hopping bias.

Periodic Boundary Conditions with Constant Rates. For constant bond hopping rates $r_k = r$, $\ell_k = \ell$ it is straightforward to prove that for periodic boundary conditions the product measure (2.32) is a family of stationary distribution of the ASEP [104]. Therefore the stationary distribution is the same as the equilibrium distribution of the SSEP, even though the ASEP does not satisfy detailed balance and hence is not an equilibrium process. The lack of reversibility is reflected in the fact that the stationary current

$$j = r\langle\eta_k\,(1-\eta_{k+1})\rangle - \ell\langle\eta_{k+1}\,(1-\eta_k)\rangle = (r-\ell)\rho(1-\rho) \tag{2.64}$$

is non-zero.

Generator of the ASEP with Reflecting Boundaries. For reflecting boundaries where hopping between the boundary sites L^- and L^+ is not allowed it is convenient to define the parameters

$$q \equiv \mathrm{e}^f = \sqrt{\frac{r_k}{\ell_{k+1}}}, \quad w_k = \sqrt{r_k\ell_{k+1}} \tag{2.65}$$

and define the system size

$$L = L^+ + 1 - L^-. \tag{2.66}$$

With the local hopping rates

$$w_{kk+1}(\boldsymbol{\eta}) = w_k\left(q\eta_k(1-\eta_{k+1}) + q^{-1}(1-\eta_k)\eta_{k+1}\right), \quad k \in \{L^-, \ldots, L^+ - 1\} \tag{2.67}$$

the transition rate from a configuration $\boldsymbol{\eta}$ to a configuration $\boldsymbol{\eta}'$ is given by

$$w_{\boldsymbol{\eta}',\boldsymbol{\eta}} = \sum_{k=L^-}^{L^+-1} w_{kk+1}(\boldsymbol{\eta})\delta_{\boldsymbol{\eta}',\boldsymbol{\eta}^{kk+1}}. \tag{2.68}$$

and the generator reads

$$\mathcal{L}f(\boldsymbol{\eta}) = \sum_{k=L^-}^{L^+-1} w_{kk+1}(\boldsymbol{\eta})[f(\boldsymbol{\eta}^{kk+1}) - f(\boldsymbol{\eta})]. \tag{2.69}$$

Using the Pauli matrices (2.16) one finds

$$H = \sum_{k=L^-}^{L^+-1} w_k h_k \tag{2.70}$$

with non-symmetric hopping matrices

$$h_k = -q\left(\sigma_k^+\sigma_{k+1}^- - \hat{n}_k\hat{v}_{k+1}\right) - q^{-1}\left(\sigma_k^-\sigma_{k+1}^+ - \hat{v}_k\hat{n}_{k+1}\right). \tag{2.71}$$

Grandcanonical Equilibrium Measure. The hopping matrices can be symmetrized by the ground state transformation

$$V := q^{\sum_{k=L^-}^{L^+} k\hat{n}_k}. \tag{2.72}$$

One has

$$h_k^T = V^{-2}h_kV^2 \tag{2.73}$$

$$\tilde{h}_k := V^{-1}h_kV = \tilde{h}_k^T. \tag{2.74}$$

This implies that this ASEP is reversible and together with particle number conservation one concludes that

$$\pi^*_{L,\phi}(\boldsymbol{\eta}) = \frac{1}{Z_{L,\phi}}\langle\,\boldsymbol{\eta}\,|V^2|\,\boldsymbol{\eta}\,\rangle = \frac{1}{Z_{L,\phi}}q^{2\sum_{k=L^-}^{L^+}(k-\kappa(\phi))\eta_k}, \quad \kappa(\phi) = -\phi/(2f) \tag{2.75}$$

is an equilibrium measure[3] for any chemical potential $\phi \in \mathbb{R}$, corresponding to a linear potential energy

$$U(\eta) = -\epsilon\sum_{k=L^-}^{L^+} k\eta_k \tag{2.76}$$

with $\epsilon = 2k_BTf$. This is a product measure with grandcanonical partition function

$$Z_{L,\phi} = \prod_{k=L^-}^{L^+}\left(1 + q^{2k-2\kappa(\phi)}\right). \tag{2.77}$$

[3] The measure (2.75) as well as all related measures and functions introduced below depend both on L^- and L^+. In order to avoid heavy notation we indicate this dependence only by the volume L.

Correspondingly the associated probability vector is a tensor product

$$|\,\pi^*_{L,\phi}\,\rangle = \frac{1}{Z_{L,\phi}} \mathrm{e}^{\phi \hat{N}} V^2 |\,s\,\rangle = |\,\rho_1\,\rangle \otimes \cdots \otimes |\,\rho_L\,\rangle \tag{2.78}$$

with marginals

$$|\,\rho_k\,\rangle = \left(1 + q^{2k-2\kappa(\phi)}\right)^{-1} \begin{pmatrix} 1 \\ q^{2k-2\kappa(\phi)} \end{pmatrix}. \tag{2.79}$$

This is the blocking measure [67] restricted to Λ.

The stationary particle density is not uniform, but given by

$$\rho_k = \frac{q^{2k-2\kappa(\phi)}}{1 + q^{2k-2\kappa(\phi)}} = \frac{1}{2}\left(1 + \tanh\left(f(k - \kappa(\phi))\right)\right), \tag{2.80}$$

see Fig. 6. This means that the stationary local density is approximately equal to $1/2$ near the lattice point $k^* = [\kappa(\phi)]$ provided that $k^* \in \Lambda$. The density approaches 1(0) to the right(left) on a length scale of order $1/f$. Thus on macroscopic scale the density has a shock discontinuity at $x^* = \kappa(\phi)/L \in [b^-, b^+]$ where $b^\pm = \lim_{L\to\infty} L^\pm/L$.

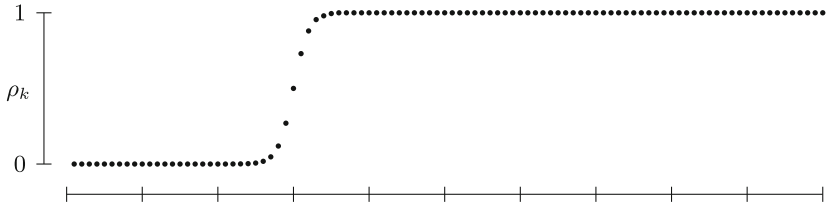


Fig. 6. Stationary density profile of the ASEP with reflecting boundaries with 100 sites. The position of the step is determined by the particle number, its width depends on the driving field. Here we have chosen $\beta f = 1/2$, corresponding to $q = \sqrt{e}$.

Duality Functions for the 1-D ASEP. The process is symmetric under the quantum algebra $U_q[\mathfrak{gl}(2)]$ [74] which is the q-deformed universal enveloping algebra of $\mathfrak{gl}(2)$ [52,53]. This implies that the generator H given by (2.70) commutes with the symmetry operators $S^\pm(q)$ and S^z where [92]

$$S^+(q) = \sum_{k=L^-}^{L^+} q^{\hat{N}_k} \sigma_k^+, \quad S^-(q) = \sum_{k=L^-}^{L^+} q^{-\hat{V}_k} \sigma_k^- \tag{2.81}$$

with the non-local particle balance operators

$$\hat{N}_k = \sum_{j=k+1}^{L^+} \hat{n}_k - \sum_{j=L^-}^{k-1} \hat{n}_k, \quad \hat{V}_k = \sum_{j=k+1}^{L^+} \hat{v}_k - \sum_{j=L^-}^{k-1} \hat{v}_k, \tag{2.82}$$

The particle balance function

$$N_k(\boldsymbol{\eta}) := \sum_{j=k+1}^{L^+} \eta_k - \sum_{j=L^-}^{k-1} \eta_k \tag{2.83}$$

gives the difference between the number of particles to right and left of site k. Hence $N_k(\boldsymbol{\eta}_t) - N_k(\boldsymbol{\eta}_0)$ is the integrated particle current across site k up to time t.

By the duality Theorem 1 the ASEP with reflecting boundary conditions defined by (2.70) is self-dual w.r.t.

$$D^{gen}(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \pi_{L,0}^{-1}(\boldsymbol{\zeta}) \langle\, \boldsymbol{\zeta} \,|F(S^+(q), S^-(q), S^z)|\, \boldsymbol{\eta} \,\rangle \tag{2.84}$$

where $F(S^+(q), S^-(q), S^z)$ is some bounded function of the symmetry operators and $\pi_{L,0}^{-1}(\boldsymbol{\zeta}) = q^{-2\sum_{k\in\Lambda}\zeta_k}$ is the unnormalized equilibrium measure of the ASEP.

The computation of the matrix elements of $F(S^+(q), S^-(q), S^z)$ is less forward than in the $\mathfrak{su}(2)$ case. We note [7].

Proposition 3. *For all $q \in \mathbb{Z} \setminus 0$ the symmetry operator*

$$Y^+(q) = \sum_{r=0}^{L} \frac{(S^+(q))^r}{[r]_q!} \tag{2.85}$$

with the q-numbers

$$[x]_q = \frac{q^x - q^{-x}}{q - q^{-1}}, \quad x \in \mathbb{Z}, \quad [n]_q! = \prod_{k=1}^{n} [k]_q, \quad k \in \mathbb{N} \tag{2.86}$$

has matrix elements

$$\langle\, \boldsymbol{\zeta} \,|Y^+(q)|\, \boldsymbol{\eta} \,\rangle = \prod_{k=L^-}^{L^+} (Q_k(\boldsymbol{\eta}))^{\zeta_k} \tag{2.87}$$

with $Q_k(\boldsymbol{\eta}) = \eta_k q^{-N_k(\boldsymbol{\eta})}$.

This result yields from (2.84) the duality function of [92] in the coordinate representation $\mathbf{x}$ of the configuration $\boldsymbol{\zeta}$:

Corollary 3. *The ASEP defined by (2.70) is selfdual w.r.t. the duality function*

$$D_\omega(\mathbf{x}, \boldsymbol{\eta}) = \prod_{i=1}^{N(\mathbf{x})} \eta_{x_i} q^{-2\sum_{k=L^-}^{x_i-1}(1-\eta_k)+\omega N(\boldsymbol{\eta})}. \tag{2.88}$$

To see this notice first that Proposition 3 together with (2.84) implies that

$$\tilde{D}(\mathbf{x}, \boldsymbol{\eta}) = \prod_{i=1}^{N(\mathbf{x})} q^{-2x_i} \eta_{x_i} q^{-N_{x_i}(\boldsymbol{\eta})} \tag{2.89}$$

is a duality function. Because of particle conservation this duality function can be multiplied with any function of the particle numbers $N(\mathbf{x})$ and $N(\boldsymbol{\eta})$ to obtain a new duality function. So in particular we have that

$$D_\omega(\mathbf{x},\boldsymbol{\eta}) = \prod_{i=1}^{N(\mathbf{x})} q^{(1+\omega)N(\boldsymbol{\eta})+2L^- -1-2x_i}\eta_{x_i} q^{-N_{x_i}(\boldsymbol{\eta})}$$
$$= \tilde{D}(\mathbf{x},\boldsymbol{\eta}) q^{N(\mathbf{x})\left((1+\omega)N(\boldsymbol{\eta})+2L^- -1\right)} \tag{2.90}$$

is also a duality function.

Microscopic Structure of Shocks in the ASEP. This duality function is not local and therefore it cannot be used to compute the dynamical structure of the ASEP. However, it carries non-trivial information about the distribution of the time-integrated current [45] and for constant bond hopping rates $w_k = w$ also about the microscopic structure and dynamics of shocks [6,8]. It turns out that just as in the SSEP the time evolution of an n-point density correlation is given by the transition probabilities of only n particles in the SSEP, the time evolution of a shock measure for the ASEP defined on the infinite integer lattice $\mathbb{Z}$ with n microscopic shocks is given by the transition probabilities of a modified ASEP with only n particles. To be precise, we state the result of [6] for a single microscopic shock where the modified ASEP reduces to a biased random walk.

Definition 9 *(Shock measure). A shock measure ν_x on $\{0,1\}^{\mathbb{Z}}$ indexed by the microscopic shock position $x \in \mathbb{Z}$ is the product measure given by the marginals*

$$\nu_x(\eta_k) = \begin{cases} 1 & k = x \\ \rho_0\delta_{\eta_k,1} + (1-\rho_-)\delta_{\eta_k,0} & k < x \\ \rho_1\delta_{\eta_k,n} + (1-\rho_+)\delta_{\eta_k,0} & k > x \end{cases} \tag{2.91}$$

The restriction to Λ for $x \in \Lambda$

$$\mu^L_{\mathbf{x}}(\boldsymbol{\eta}) := \prod_{k=L^-}^{L^+} \nu_x^k(\eta_k) \tag{2.92}$$

is also called shock measure with microscopic shock position x.

Then one has [4,6,8]:

Theorem 3. *Let $\nu_x(t)$ denote the measure at time t of the ASEP on $\mathbb{Z}$ with constant rates $r = wq > 0$, $\ell = wq^{-1} > 0$, starting from a shock measure ν_x defined in Definition 9 with*

$$\frac{\rho_+(1-\rho_-)}{\rho_-(1-\rho_+)} = q^2. \tag{2.93}$$

Then, for any $x \in \mathbb{Z}$

$$\nu_x(t) = \sum_{y\in\mathbb{Z}} P(y,t|x,0)\,\nu_y \tag{2.94}$$

where $P(y,t|x,0)$ is the transition probability of a biased random walk with jump rates

$$p_\pm = (r-\ell)\frac{\rho_\pm(1-\rho_\pm)}{\rho_+ - \rho_-} \tag{2.95}$$

to the right $(+)$ and left $(-)$ respectively.

Corollary 4. *The shock is microscopically sharp at all times and performs on macroscopic scale a diffusive motion with drift velocity*

$$v_s = p_+ - p_- = (r-\ell)(1-\rho_+-\rho_-) \tag{2.96}$$

and diffusion coefficient

$$D_s = \frac{1}{2}(p_+ + p_-) = (r-\ell)\frac{\rho_-(1-\rho_-)+\rho_+(1-\rho_+)}{\rho_+-\rho_-}. \tag{2.97}$$

Remark 10. The microscopic sharpness follows from the product structure of the shock measure (2.92). One recognizes in the microscopic shock velocity (2.96) the Rankine-Hugoniot velocities (1.9) since $j_\pm := w(q-q^{-1})\rho_\pm(1-\rho_\pm)$ is the expectation of the particle current to the right and to the left of shock. The shock diffusion coefficients (2.97) are consistent with the general result (1.10) of [33] on shock motion in the ASEP on diffusive scale.

We outline the proof and refer for the technical details to [6,8]. Al alternative probabilistic proof is given in [4].

Consider first the finite lattice Λ. We recall (2.3) which reads for the duality function (2.88) with a single-particle configuration $\boldsymbol{\zeta}$

$$\begin{aligned}\sum_{y\in\Lambda} D_\omega(y,\boldsymbol{\eta})P(y,t|x,0) &= \sum_{\boldsymbol{\eta}'\in\Omega} \eta'_x q^{-2\sum_{k=L_-}^{x-1}(1-\eta'_k)+\omega N(\boldsymbol{\eta}')} \langle\, \boldsymbol{\eta}' \,|\mathrm{e}^{-Ht}|\, \boldsymbol{\eta}\,\rangle \\ &= \sum_{\boldsymbol{\eta}'\in\Omega} \langle\, \boldsymbol{\eta}' \,|\hat{n}_x q^{-2\sum_{k=L_-}^{x-1}(1-\hat{n}_k)+\omega\hat{N}}\mathrm{e}^{-Ht}|\, \boldsymbol{\eta}\,\rangle \\ &= \langle\, s \,|\hat{n}_x q^{-2\sum_{k=L_-}^{x-1}(1-\hat{n}_k)+\omega\hat{N}}\mathrm{e}^{-Ht}|\, \boldsymbol{\eta}\,\rangle \\ &= \langle\, \boldsymbol{\eta} \,|\mathrm{e}^{-H^T t}\hat{n}_x q^{-2\sum_{k=L_-}^{x-1}(1-\hat{n}_k)+\omega\hat{N}}|\, s\,\rangle \end{aligned} \tag{2.98}$$

On the other hand, for the l.h.s. we have

$$D_\omega(y,\boldsymbol{\eta}) = \eta_y q^{-2\sum_{k=L_-}^{y-1}(1-\eta_k)+\omega N(\boldsymbol{\eta})} = \langle\, \eta \,|\hat{n}_y q^{-2\sum_{k=L_-}^{y-1}(1-\hat{n}_k)+\omega\hat{N}(\boldsymbol{\eta})}|\, s\,\rangle \tag{2.99}$$

Next we observe that

$$\hat{n}_x q^{-2\sum_{k=L_-}^{x-1}(1-\hat{n}_k)+\omega\hat{N}(\boldsymbol{\eta})}|\, s\,\rangle = Z_x^{-1}|\, \nu_x\,\rangle \tag{2.100}$$

with densities

$$\rho_- = \frac{q^\omega}{q^{-2}+q^\omega}, \quad \rho_+ = \frac{q^\omega}{1+q^\omega} \tag{2.101}$$

and normalization constant

$$Z_x = \left(q^{-2}+q^{\omega}\right)^{x-L^-}\,(1+q^{\omega})^{L^+-x} = \frac{(1+q^{\omega})^{L^+}}{(q^{-2}+q^{\omega})^{L^-}}\left(\frac{\rho_+}{\rho_-}\right)^x \tag{2.102}$$

Thus selfduality yields

$$\sum_{y\in\Lambda} Z_y^{-1}|\,\nu_y\,\rangle P(y,t|x,0) = \mathrm{e}^{-H^T t} Z_x^{-1}|\,\nu_x\,\rangle. \tag{2.103}$$

or equivalently

$$\sum_{y\in\Lambda} |\,\nu_y\,\rangle \left(\frac{\rho_+}{\rho_-}\right)^{x-y} P(y,t|x,0) = \mathrm{e}^{-H^T t}|\,\nu_x\,\rangle. \tag{2.104}$$

Notice now the trivial random walk property that up to a boundary term

$$\left(\frac{\rho_+}{\rho_-}\right)^{x-y} P(y,t|x,0) = \mathrm{e}^{-\lambda t}\tilde{P}(y,t|x,0) \tag{2.105}$$

where $\tilde{P}(y,t|x,0)$ is the transition probability of a random walk with rates $\tilde{p}_\pm = (q\rho_-/\rho_+)^{\pm 1}$ and $\lambda = p_+ + p_- - q - q^{-1}$. Thus

$$\sum_{y\in\Lambda} \tilde{P}(y,t|x,0)|\,\nu_y\,\rangle = \mathrm{e}^{-(H^T-\lambda)t}|\,\nu_x\,\rangle. \tag{2.106}$$

On the other hand, $\tilde{H} = H^T - \lambda$ is, up to another boundary term, the generator of the ASEP with inverse hopping asymmetry q^{-1}. Using coupling arguments it can be shown that in the thermodynamic limit these boundary terms are irrelevant [6]. Thus we arrive at

$$\sum_{y\in\mathbb{Z}\!\!\mathbb{Z}} \tilde{P}(y,t|x,0)|\,\nu_y\,\rangle = |\,\nu_x(t)\,\rangle^{\triangleleft} \tag{2.107}$$

where the upper left-pointing triangle indicates the evolution under the ASEP on $\mathbb{Z}\!\!\mathbb{Z}$ with reversed bias.

The densities satisfy

$$\frac{\rho_+(1-\rho_-)}{\rho_-(1-\rho_+)} = q^{-2}. \tag{2.108}$$

Therefore

$$q - q^{-1} = q(1-q^{-2}) = \frac{p_+}{w}\,\frac{\rho_- - \rho_+}{\rho_+(1-\rho_+)} \tag{2.109}$$

$$= q^{-1}(q^2-1) = \frac{p_-}{w}\,\frac{\rho_- - \rho_+}{\rho_-(1-\rho_-)} \tag{2.110}$$

which yields the transition rates

$$p_\pm = (r-\ell)\frac{\rho_- - \rho_+}{\rho_\pm(1-\rho_\pm)} \tag{2.111}$$

for the random walk. Substituting $q \to q^{-1}$ then proves the theorem.

2.4 Recipe for Constructing the Quantum Hamiltonian of Exclusion Processes

To construct H for a given process without going through the explicit matrix multiplications we note that any changes of the state of the system are represented by offdiagonal matrices. To be precise, they represent attempts rather than actual changes: Acting on a state with an already occupied site with σ^- yields zero, i.e. no change in the probability vector. This reflects the rejection of any attempt at creating a second particle on a given site. Thus the exclusion of double occupancy is encoded in the properties of the Pauli matrices.

Simultaneous events are represented by products of Pauli matrices acting on different sites. E.g. hopping of a particle from site i to site j is equivalent to annihilating a particle at site i and at the same time creating one at site j. Thus it is given by the matrix $\sigma_i^+\sigma_j^-$. The hopping attempt is successful only if site i is occupied and site j is empty. Otherwise acting with $\sigma_i^+\sigma_j^-$ on the state gives zero and hence no change. The rate of hopping (or of any other possible stochastic event) is the numerical prefactor of each hopping matrix (or other attempt matrix). Of course, in principle the rate may depend on the configuration of the complete system. Suppose the hopping rate is given by a function $w(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is the configuration prior to hopping. In this case the hopping matrix is given by $\sigma_i^+\sigma_j^-\hat{w}$ where the diagonal operator $\hat{w}$ is obtained form the rate $w(\boldsymbol{\eta})$ by replacing any η_i by the projector $\hat{n}_i$. If e.g. for some reason hopping from site i to site j should occur only if a third site k is empty, then the hopping matrix would be given by $\sigma_i^+\sigma_j^-(1-\hat{n}_k)$. For a hopping from site i to site j with a rate that is proportional to the number of particles on some set of $\mathbb{S}(i,j)$ sites one finds the matrix $w\sigma_i^+\sigma_j^-\sum_{k\in\mathbb{S}(i,j)}\hat{n}_k)$. The construction of the attempt matrices for other processes or for n-states model is analogous.

For two-states models one notes the useful identities

$$\langle\, s\,|\sigma_i^+ = \langle\, s\,|\hat{n}_i, \quad \langle\, s\,|\sigma_i^- = \langle\, s\,|(\hat{v}_i \tag{2.112}$$

which follow immediately from the tensor structure of the summation vector and the definition of the local Pauli matrices. With these relations it is easy to construct the diagonal part of the quantum Hamiltonian in order ensure conservation of probability. To each off-diagonal attempt matrix one constructs a diagonal matrix by replacing all σ_i^+ by $\hat{n}_i$ and by replacing all σ_i^- by $\hat{v}_i$. E.g. to hopping from i to j with constant rate w represented by $-w\sigma_i^+\sigma_j^-$ one adds $w\hat{n}_i\hat{v}_j$. The (negative) sum of all attempt matrices minus their diagonal counterparts is then the full generator. In the same way one constructs the diagonal parts of n-states models by using the analogues of Eq. (2.112). Conservation of probability (1.41) is then automatically satisfied.

3 Fluctuation Theorems for Currents

Fluctuation theorems relate the probability of a positive value of some observable to the probability of the negative of that quantity [31,42,44,65,99] in a

Markov process that can be time-inhomogeneous, i.e., where the transition rates depend on time. The Gallavotti-Cohen theorem [34] for the distribution of the entropy production in deterministic dynamics and a related result by Lebowitz and Spohn for the particle current in stochastic interacting particle systems [65] have been established in a mathematically rigorous fashion. A vast body of work has been devoted to further develop fluctuation relations and to test them experimentally. For relatively recent reviews we refer to [44,99].

It has become clear that currents which are odd under time-reversal play an important role in understanding the nonequilibrium properties of a particle system. Therefore we consider here fluctuation theorems that concern such time-integrated currents like the current of particles in a particle system like the ASEP (cf. (1.3)) and related time-integrated quantities that can be measured in an experiment. To this end we keep track of the trajectory of a time-inhomogeneous process, i.e., the whole sequence of transitions from an initial configuration η_0 at time 0 to a final state η_t at time $t > 0$. In a nutshell, what this lecture is about can be summarized as follows.

Fluctuations relations for time-integrated currents follow from a single fundamental fluctuation relation that arises from time-reversal of a time-inhomogeneous process with a time-reversed protocol of the time-dependent transition rates.

In order to make this precise we first need to introduce some more tools.

3.1 Tools

Counting Processes. The time-integrated particle current provides an example of what we shall call a *counting process* C_t with state space $\mathbb{R}$, defined informally by the following properties [42].

Property 1: The value of C_t changes only at a transition of an underlying process η_t. It changes by an increment $r_{\eta',\eta} \in \mathbb{R}$ for a transition $\eta \to \eta'$.

Property 2: The transition rates $w_{\eta',\eta}$ of the process η_t do not depend on C_t.

We also introduce the extended counting process by adding "boundary values" that do not depend on C_t:

Property 3: For given functions $r^i : \Omega \to \mathbb{R}$, $r^F : \Omega \to \mathbb{R}$ the extended counting process is the random number $R_t = r^i_{\eta_0} + C_t + r^f_{\eta_t}$.

Notice that in general a counting process C_t is not Markovian. The joint process $\sigma_t = (\eta_t, C_t)$ with state space $\Xi = \Omega \times \mathbb{R}$, however, is Markov with generator

$$\mathcal{S}f(\eta, C) = \sum_{\eta' \in \Omega \setminus \eta} w_{\eta',\eta} \left[f(\eta', C + r_{\eta',\eta}) - f(\eta, C) \right]. \tag{3.1}$$

The values C_t, R_t of a counting process at time t can be regarded as a functional of the trajectories of the underlying process η_t as we note in the following proposition which is an immediate consequence of the definition of the counting process, see Fig. 7 illustration.

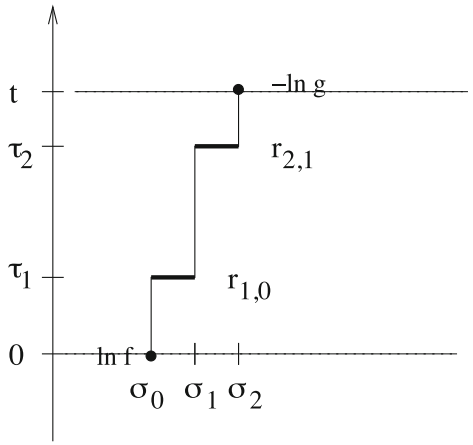


Fig. 7. A stochastic trajectory σ_τ with $0 \le \tau \le t$ with the sequence of configurations $\{\sigma_0, \sigma_1, \sigma_2\}$ where $\sigma_t = \sigma_2$. Time points upwards, the horizontal direction is the abstract space of configurations. The increment $r_{\sigma_i,\sigma_j}(\tau)$ is abbreviated as $r_{i,j}$ and the boundary values are $r^i = \ln f(\sigma_0)$ and $r^f = -\ln g(\sigma_2)$. Therefore $C_t = r_{0,1} + r_{1,2}$ and $R_t = \ln f(\sigma_0) + r_{0,1} + r_{1,2} - \ln g(\sigma_2)$.

Proposition 4. *Let η_t be a Markov process with finite state space Ω and $C_0 = 0$. Then for a trajectory $\eta_{[0,t]}$ of the process with $n_t \ge 0$ transitions at random times $t_k \in (0,t)$, $1 \le k \le n_t$ and $\eta_\tau = \eta_{t_k}$ for $t_k \le \tau < t_{k+1} \le t$, $0 \le k < n_t$ and $\eta_\tau = \eta_t$ for $t_{n_t} \le \tau \le t$ one has*

$$C_t = \sum_{k=1}^{n_t} r_{\eta_{t_k},\eta_{t_{k-1}}} \tag{3.2}$$

and

$$R_t = r^i_{\eta_0} + \sum_{k=1}^{n_t} r_{\eta_{t_k},\eta_{t_{k-1}}} + r^f_{\eta_t}. \tag{3.3}$$

The physical scenario described by a counting process is the following. One imagines that besides the physical random process described by η_t there is some physical property such as the energy of a heat reservoir that can be measured and whose value is C_t. This quantity does not directly depend on the state η_t but changes by an amount $r_{\eta',\eta}$ whenever the underlying transition $\eta \to \eta'$ occurs.

Moreover, it is assumed that this physical property does not perturb the dynamics of the process. The extended counting process then yields a physical property that depends also on the boundary states of the physical random process. Here "boundary" refers to the temporal boundaries of the trajectory at times $\tau = 0$ and $\tau = t$ (Fig. 7). Concrete examples of physical importance will be discussed below.

As a direct consequence of (3.1) we note for a function $g(\eta, C)$ of the form $g(\eta, C) = f(\eta)\mathrm{e}^{-\lambda C}$ the factorization property

$$\mathcal{S}\left(f(\eta)\mathrm{e}^{-\lambda C}\right) = \mathrm{e}^{-\lambda C}\tilde{\mathcal{L}}_\lambda f(\eta) \tag{3.4}$$

with the *tilted generator*

$$\tilde{\mathcal{L}}_\lambda f(\eta) = \sum_{\eta' \in \Omega \setminus \eta} w_{\eta',\eta}\left[\mathrm{e}^{-\lambda r_{\eta',\eta}} f(\eta') - f(\eta)\right]. \tag{3.5}$$

This factorization property of the generator has the important consequence that for a factorized initial measure $\mu(\eta, C) = \mu(\eta)\delta_{C,0}$ one has

$$\langle f_t \mathrm{e}^{-\lambda C_t}\rangle_\mu = \mathrm{e}^{-\lambda C_0} \sum_{\eta \in \Omega} f(\eta)\tilde{\mu}_{\lambda,t}(\eta) \tag{3.6}$$

where the tilted measure $\tilde{\mu}_{\lambda,t}$ with $\tilde{\mu}_{\lambda,0} = \mu$ derives from the time evolution of the initial measure μ under the semigroup generated by the tilted generator $\tilde{\mathcal{L}}_\lambda$ [21].

In matrix form one has

$$\tilde{H}_\lambda = -\sum_{\eta \in \Omega} \sum_{\eta' \in \Omega \setminus \eta} w_{\eta'\eta}\left(\mathrm{e}^{-\lambda r_{\eta',\eta}} E^{\eta'\eta} - \hat{1}_\eta\right) \tag{3.7}$$

with the transition matrix $E^{\eta'\eta}$ defined in (1.36). Even though $\tilde{H}$ is not a stochastic generator the evolution under $\tilde{H}$ has a straightforward stochastic interpretation by appealing to the interpretation of C_t as a trajectory functional. Each stochastic trajectory generated by the underlying process H gets weighted under the evolution of $\tilde{H}$ by a factor $\mathrm{e}^{-\lambda r_{\eta',\eta}}$ whenever a transition $\eta \to \eta'$ occurs. Hence the tilted transition probability, or equivalently the generating function (3.6) of the counter C_t for $f(\cdot) = 1_{\eta'}(\cdot)$ and initial measure concentrated on η,

$$P_\lambda(\eta'; t|\eta; 0) = \langle\, \eta' \,|\mathrm{e}^{-\tilde{H}_\lambda t}|\, \eta \,\rangle \tag{3.8}$$

can be interpreted as a measure for the weighted trajectories from a configuration η to a configuration η' for a time interval of length t.

Thus, taking $f_t(\eta) = 1$ and arbitrary initial measure $\mu(\eta)$ one obtains the generating function

$$G(\lambda, t) := \langle \mathrm{e}^{-\lambda C_t}\rangle_\mu = \langle\, s \,|\mathrm{e}^{-\tilde{H}_\lambda t}|\, \mu \,\rangle \tag{3.9}$$

of the counting function C_t. Likewise, the tilted correlation function

$$C_{12}(\lambda, t) := \langle\, s \,|\hat{f}_2 \mathrm{e}^{-\tilde{H}_\lambda t}\hat{f}_1|\, \mu \,\rangle \tag{3.10}$$

is the measure for the weighted trajectories drawn from an initial distribution μ with *boundary weights* $r^i_{\eta_0} = \ln(f_1(\eta_0))$ and $r^f_{\eta_t} = \ln(f_2(\eta_t))$. Therefore, choosing $f_1(\eta) = f^{-\lambda}(\eta)$, $f_2(\eta) = g^{\lambda}(\eta)$ and initial measure $\mu(\eta) = f(\eta)/Z$ with partition function $Z = \langle s|\, f\rangle$, one thus finds for the extended counting process

$$\langle \mathrm{e}^{-\lambda R_t}\rangle_f = \langle\, s\,|\hat{g}^{\lambda}\mathrm{e}^{-\tilde{H}_\lambda t}\hat{f}^{-\lambda}|\, f\,\rangle \tag{3.11}$$

which means that the generating function of the extended counting process is a tilted correlation function.

Time-Dependent Transition Rates. Above we have tacitly assumed that the transition rates of the Markov process were independent of time. When we make them explicitly time-dependent the finite-time transition matrix is no longer $\exp(-Ht)$, but given by the time-ordered exponential $\mathcal{T}\left[\exp(-\int_0^t \mathrm{d}\tau H(\tau))\right]$ defined for general square matrices as follows.

Definition 10. *Let $H(t)$ be a finite-dimensional square matrix parametrized by time t. The time-ordered exponential of $\int_0^t \mathrm{d}\tau H(\tau)$ is the infinite sum*

$$\mathcal{T}\left[\mathrm{e}^{-\int_0^t \mathrm{d}\tau H(\tau)}\right] = \sum_{n=0}^{\infty}(-1)^n G_n(t) \tag{3.12}$$

where the matrix $G_n(t)$ is defined recursively by

$$G_n(t) := \int_0^t \mathrm{d}\tau H(\tau) G_{n-1}(\tau), \quad n \geq 1 \tag{3.13}$$

and $G_0(t) = \mathbf{1}$.

For illustration we write out explicitly the first few terms:

$$\mathcal{T}\left[\mathrm{e}^{-\int_0^t \mathrm{d}\tau H(\tau)}\right] = \mathbf{1} - \int_0^t \mathrm{d}\tau H(\tau) + \int_0^t \mathrm{d}\tau_1 H(\tau_1)\int_0^{\tau_1}\mathrm{d}\tau_2 H(\tau_2) - \ldots \tag{3.14}$$

Evidently one has $\frac{\mathrm{d}}{\mathrm{d}t}G_n(t) = H(t)G_{n-1}(t)$ for $n \geq 1$ and $\frac{\mathrm{d}}{\mathrm{d}t}G_0(t) = 0$. Thus, with the short-hand notation

$$P(t) = \mathcal{T}\left[\mathrm{e}^{-\int_0^t \mathrm{d}\tau H(\tau)}\right], \tag{3.15}$$

one has

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t) = -H(t)P(t). \tag{3.16}$$

Since also $\lim_{t\searrow 0} P(t) = \mathbf{1}$ we find that

$$|\,\mu(t)\,\rangle = P(t)|\,\mu\,\rangle \tag{3.17}$$

is the time-dependent measure satisfying the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}|\,\mu(t)\,\rangle = -H(t)|\,\mu(t)\,\rangle \tag{3.18}$$

of a time-inhomogeneous Markov process with time-dependent transition rates $w_{\eta',\eta}(t)$. Thus the time-ordered exponential yields the transition matrix of the time-inhomogeneous Markov process. We shall refer to the time-dependence of the rates as *protocol* of the process since we have in mind an experiment where one changes a process in time in some specific way (called protocol) by means of some technical device.

For a similarity transformation (not dependent on t) one has

$$AP(t)A^{-1} = \mathcal{T}\left[\mathrm{e}^{-\int_0^t \mathrm{d}\tau AH(\tau)A^{-1}}\right]. \tag{3.19}$$

Notice that transposition yields

$$P^T(t) = \mathcal{T}\left[\mathrm{e}^{-\int_0^t \mathrm{d}\tau H^T(t-\tau)}\right] \tag{3.20}$$

with transposition *and* time-reversal of the protocol inside the exponential.

Since in defining the time-ordered exponential we have nowhere used that H is the generator of a stochastic process the formulas (3.18)–(3.20) apply also to the tilted generator, including the case where the increments are explicitly time-dependent. We write $r_{\eta',\eta}(t)$ to make such a dependence clear.

3.2 The Fundamental Fluctuation Relation

Loosely speaking, fluctuation relations arise from comparing the probability of a trajectory of a process to the probability of a "time-reversed" trajectory. Here we prove a single master fluctuation theorem from which many fundamental fluctuation relations that have appeared in the literature follow as simple corollaries. It turns out that with the machinery developed above the proof of this master fluctuation theorem itself reduces to a mathematical triviality. The significance of this master fluctuation relation and its famous corollaries is not mathematical depth but lies in the rather general applicability in physics, the validity arbitrarily far from equilibrium, and a unifying description of the various fluctuation theorems available for stochastic dynamics.

We are mostly interested in currents that change sign under time reversal and therefore focus on antisymmetric increments satisfying $r_{\eta',\eta}(t) = -r_{\eta,\eta'}(t)$. We denote the associated counting process by J_t. Since fluctuation theorems arise from time reversal we differentiate between a *forward* process η_t^F and a *backward* process η_t^B which are not to be confused with the definition of the reversed process (1.48).

Definition 11 *(Forward and backward process). Fix an observation time $t > 0$ and let η_τ^F be a Markov process with countable state space Ω and time-dependent transition rates $w^F_{\eta'\eta}(\tau)$ such that for all $\tau \in [0,t]$ and all $\eta, \eta' \in \Omega \times \Omega$ either*

$w^F_{\eta'\eta}(\tau)w^F_{\eta\eta'}(\tau) > 0$ *or* $w_{\eta'\eta}(\tau) = w_{\eta\eta'}(\tau) = 0$. *We say that for* $\tau \in [0,t]$ *the process* η^B_τ *with transition rates* $w^B_{\eta'\eta}(\tau) = w^F_{\eta'\eta}(t-\tau)$ *is the backward process associated to the forward process* η^F_t *and the set of functions* $w^B_{\eta'\eta}(\tau)$ *is the backward protocol associated with the forward protocol* $w^F_{\eta'\eta}(\tau)$.

Expectations for the forward and backward process w.r.t. some initial measure μ are denoted by $\mathbf{E}^X_\mu$ with $X \in \{F,B\}$ indicating the process (forward or backward). For expectations with initial measure $\mu(\eta) = \delta_{\eta,\eta_0}$ concentrated on a fixed initial configuration η_0 we use the notation $\mathbf{E}^X_{\eta_0}$. Sums over such expectations are denoted by

$$\langle\cdot\rangle^X_f := \sum_{\eta_0\in\Omega} f(\eta_0)\mathbf{E}^X_{\eta_0}(\cdot) \tag{3.21}$$

with a function $f : \Omega \to \mathbb{R}$. The central result from which many celebrated fluctuation theorems derive is the following [42].

Theorem 4 *(Fundamental fluctuation relation). Fix* $t > 0$ *and let* η^X_τ *with* $X \in \{F,B\}$ *be forward and backward Markov processes according to Definition 11 with finite state space* Ω *and associated counting processes* J^X_τ *with antisymmetric increments*

$$r^X_{\eta'\eta}(\tau) = \ln\left(\frac{w^X_{\eta'\eta}(\tau)}{w^X_{\eta\eta'}(\tau)}\right) \tag{3.22}$$

for transitions satisfying $w^X_{\eta'\eta}(\tau)w^X_{\eta\eta'}(\tau) > 0$ *and* $r^X_{\eta'\eta}(\tau) = 0$ *otherwise. Furthermore, let*

$$R^F_t := \ln f(\eta^F_0) + J^F_t - \ln g(\eta^F_t), \quad R^B_t := \ln g(\eta^B_0) + J^B_t - \ln f(\eta^B_t) \tag{3.23}$$

with $f(\eta), g(\eta) \neq 0$ *for all* $\eta \in \Omega$ *be the associated extended counting processes at time* t. *Then the generating functions*

$$\Phi^F(\lambda,t) := \langle \mathrm{e}^{-\lambda R^F_t}\rangle^F_f, \quad \Phi^B(\lambda,t) := \langle \mathrm{e}^{-\lambda R^B_t}\rangle^B_g \tag{3.24}$$

of the trajectory functionals $R^{F,B}_t$ *obey the symmetry*

$$\Phi^F(\lambda,t) = \Phi^B(1-\lambda,t) \tag{3.25}$$

for all $\lambda \in \mathbb{R}$.

Proof. First we note the following lemma.

Lemma 1. *Let* H^F *be the generator of the forward process* η^F_τ *and* H^B *be the generator of the backward process* η^B_τ *according to Definition 11 and let* $J^{F,B}_\tau$ *be the associated counting processes with increments (3.22). Then tilted evolution operators satisfy*

$$\left(\tilde{P}^F_\lambda(\tau)\right)^T = \tilde{P}^B_{1-\lambda}(\tau). \tag{3.26}$$

for all $\tau \in [0,t]$ *and* $\lambda \in \mathbb{R}$.

Proof. For brevity we denote in the following

$$\sum_{\eta,\eta'} := \sum_{\eta\in\Omega}\sum_{\eta'\in\Omega\setminus\eta}. \tag{3.27}$$

Recall the matrix representation (3.7) from which one obtains for the time-dependent case

$$\begin{aligned}
\left(\tilde{H}^F_\lambda(t-\tau)\right)^T &= -\sum_{\eta,\eta'} w^F_{\eta'\eta}(t-\tau)\left(\mathrm{e}^{-\lambda r^F_{\eta',\eta}(t-\tau)}E^{\eta\eta'}-\hat{1}_\eta\right)\\
&= -\sum_{\eta,\eta'} w^F_{\eta'\eta}(t-\tau)\left(\left(\frac{w^F_{\eta\eta'}(t-\tau)}{w^F_{\eta'\eta}(t-\tau)}\right)^{\lambda}E^{\eta\eta'}-\hat{1}_\eta\right)\\
&= -\sum_{\eta,\eta'}\left(w^F_{\eta\eta'}(t-\tau)\left(\frac{w^F_{\eta\eta'}(t-\tau)}{w^F_{\eta'\eta}(t-\tau)}\right)^{\lambda-1}E^{\eta\eta'}-w^F_{\eta'\eta}(t-\tau)\hat{1}_\eta\right)\\
&= -\sum_{\eta,\eta'}\left(w^B_{\eta\eta'}(\tau)\left(\frac{w^B_{\eta\eta'}(\tau)}{w^B_{\eta'\eta}(\tau)}\right)^{\lambda-1}E^{\eta\eta'}-w^B_{\eta'\eta}(\tau)\hat{1}_\eta\right)\\
&= -\sum_{\eta,\eta'}\left(w^B_{\eta\eta'}(\tau)\left(\frac{w^B_{\eta'\eta}(\tau)}{w^B_{\eta\eta'}(\tau)}\right)^{1-\lambda}E^{\eta\eta'}-w^B_{\eta'\eta}(\tau)\hat{1}_\eta\right)\\
&= -\sum_{\eta,\eta'} w^B_{\eta'\eta}(\tau)\left(\left(\frac{w^B_{\eta\eta'}(\tau)}{w^B_{\eta'\eta}(\tau)}\right)^{1-\lambda}E^{\eta'\eta}-\hat{1}_\eta\right)\\
&= -\sum_{\eta,\eta'} w^B_{\eta\eta'}(\tau)\left(\mathrm{e}^{-(1-\lambda)r^B_{\eta',\eta}(t-\tau)}E^{\eta'\eta}-\hat{1}_\eta\right).
\end{aligned} \tag{3.28}$$

where we have used the antisymmetry of the increments. Thus

$$\left(\tilde{H}^F_\lambda(t-\tau)\right)^T = \tilde{H}^B_{1-\lambda}(\tau). \tag{3.29}$$

The transposition property (3.20) of the time-ordered exponential then proves the Lemma. □

Continuing with the proof of Theorem 4 we include now the boundary terms. We have

$$\Phi^F(\lambda,t) = \langle\, s\,|\hat{g}^\lambda \tilde{P}^F_\lambda(t)\hat{f}^{-\lambda}|\, f\,\rangle, \quad \Phi^B(\lambda,t) = \langle\, s\,|\hat{f}^\lambda \tilde{P}^B_\lambda(t)\hat{g}^{-\lambda}|\, g\,\rangle \tag{3.30}$$

with $|\, f\,\rangle = \sum_{\eta\in\Omega} f(\eta)|\,\eta\,\rangle$ and $|\, g\,\rangle = \sum_{\eta\in\Omega} g(\eta)|\,\eta\,\rangle$. By transposition we obtain

$$\Phi^F(\lambda,t) = \langle\, f\,|\hat{f}^{-\lambda}\left(\tilde{P}^F_t(\lambda)\right)^T\hat{g}^\lambda|\, s\,\rangle = \langle\, s\,|\hat{f}^{1-\lambda}\left(\tilde{P}^F_t(\lambda)\right)^T\hat{g}^{-(1-\lambda)}|\, g\,\rangle. \tag{3.31}$$

Lemma 1 then concludes the proof. □

Remark 11. We required the functions f and g to be non-vanishing for all η. However, one can generalize Theorem 4 by introducing indicator functions 1_X, 1_Y on subsets $X, Y \in \Omega$. Going through the same steps as above one obtains for

$$\Phi^F_{YX}(\lambda,t) := \langle I_Y(\eta_t)\mathrm{e}^{-\lambda R^F_t} I_X(\eta_0)\rangle^F_f, \quad \Phi^B_{YX}(\lambda,t) := \langle I_X(\eta_t)\mathrm{e}^{-\lambda R^B_t} I_Y(\eta_0)\rangle^B_g \tag{3.32}$$

the extended fluctuation theorem

$$\Phi^F_{YX}(\lambda,t) = \Phi^B_{XY}(1-\lambda,t) \tag{3.33}$$

In particular, choosing $X = |\,\eta_a\,\rangle\langle\,\eta_a\,|$ and $Y = |\,\eta_b\,\rangle\langle\,\eta_b\,|$ one obtains a symmetry relation for trajectories between fixed configurations η_a, η_b. This yields the detailed fluctuation theorems introduced in [50].

3.3 Some Specific Fluctuation Theorems

In applications one thinks of a stochastic transition as being triggered by thermal processes in the physical environment into which the system described by the process is embedded. The choice of increments (3.22) then means that J_t is the change of entropy ΔS_{env} of the physical environment along a trajectory of the process [98]. Since for thermal systems at temperature T, the dissipated heat is given by

$$Q = T\Delta S_{\mathrm{env}} \tag{3.34}$$

we can also think of this current J_t as defining a nonequilibrium heat term. Different physical scenarios are then described an appropriate choice of the boundary terms $r^i(\eta)$ and $r^f(\eta)$ in the extended process R_t. We list some well-known cases.

Integral Fluctuation Relations. Setting $\lambda = 1$ in (3.25) gives the "integral fluctuation relation" [71]

$$\langle \mathrm{e}^{-R^F_t}\rangle^F_f = 1 \tag{3.35}$$

with *any* normalized choice of $r^i = \ln f$ and $r^f = -\ln g$. The specific choice of f and g determines the physical interpretation of R^F_t.

(1) Jarzynski equality

Consider a process in which the rates obey detailed balance (1.56) at all times w.r.t. a time-dependent distribution

$$\mu^*_\tau(\eta) = \mathrm{e}^{-\beta U_\tau(\eta)}/Z_\tau \tag{3.36}$$

with temperature $T = 1/\beta$, internal energy $U_\tau(\eta)$ of the configuration η, partition function

$$Z_\tau = \sum_{\eta\in\Omega} \mathrm{e}^{-\beta U_\tau(\eta)} \tag{3.37}$$

and free energy

$$F_\tau = -T\ln Z_\tau. \tag{3.38}$$

Now we imagine preparing an experiment in which we start with initial distribution $f(\eta) = \mu_0^*(\eta)$ and measure at some fixed time $t > 0$ the quantity

$$g(\eta) = \mu_t^*(\eta). \tag{3.39}$$

Then (3.35) reads

$$\langle \mathrm{e}^{-R_t^F} \rangle_f^F = \frac{Z_0}{Z_t} \langle\, s \,|\mathrm{e}^{-\beta\hat{U}} \tilde{P}_1(t) \mathrm{e}^{\beta\hat{U}} |\, \mu_0^* \,\rangle \tag{3.40}$$

In this case R_t^F is proportional to the dissipated work, which can be seen as follows.

Using the Boltzmann form of the time-dependent pseudo-equilibrium distribution μ_τ^* the boundary part of the functional R_t^F becomes

$$\ln f(\eta_0) - \ln g(\eta_t) = \frac{\Delta U}{T} - \frac{\Delta F}{T} \tag{3.41}$$

with the changes of internal energy $\Delta U := U(\eta_t) - U(\eta_0)$ and free energy $\Delta F := F_t - F_0$ resp. during the experimental time span t. Since the current part J_t is proportional to dissipated heat (3.34) one finds

$$R_t^F = \frac{Q}{T} + \frac{\Delta U}{T} - \frac{\Delta F}{T} \tag{3.42}$$

According to the first law of thermodynamics the work is given by $W = (Q + \Delta U)/T$ and hence R_t^F is the dissipated work.

Thus (3.35) yields the Jarzynski relation [49]

$$\langle \mathrm{e}^{-W/T} \rangle = \mathrm{e}^{-\Delta F/T}. \tag{3.43}$$

Notice that it is *not* assumed that the system during its time evolution is in its time-dependent pseudo-equilibrium state μ_τ^*, not even at the final measurement time $\tau = t$. This is important as it implies one can measure equilibrium free energies from an average of the nonequilibrium work performed. The Jarzynski equality can also be related to some earlier work theorems [13–15]. A discussion of the connections can be found in [51].

(2) Integral fluctuation theorem for entropy

Now consider a different experimental scenario. Prepare experimentally an initial distribution $\mu = f$ and measure a quantity g chosen to correspond to the final probability distribution of the process, i.e.,

$$g(\eta) = \langle\, \eta \,|P_t|\, f \,\rangle. \tag{3.44}$$

Then the boundary term of R_t^F can be written as

$$f(\eta_0) - g(\eta_t) = \ln \mu(\eta_0, 0) - \ln \mu(\eta_t, t) \tag{3.45}$$

where $\mu(\eta_t, t)$ is the solution of the time-dependent master equation (3.18).

Using the general definition

$$S = \mathbf{E}_\mu \ln \mu(\eta) = \sum_{\eta \in \Omega} \mu(\eta) \ln \mu(\eta) \tag{3.46}$$

of entropy these boundary terms can be interpreted as the change in "system" entropy $\Delta S_{\rm sys} := S_{\rm sys}(t) - S_{\rm sys}(0)$ along a trajectory [98]. Hence in this case we have

$$R_t^F = \Delta S_{\rm env} + \Delta S_{\rm sys} =: \Delta S_{\rm tot}, \tag{3.47}$$

and (3.35) becomes an integral relation for the *total* entropy change [98]

$$\langle \mathrm{e}^{-\Delta \mathcal{S}_{\rm tot}} \rangle = 1. \tag{3.48}$$

Jensen's inequality then implies $\langle \Delta \mathcal{S}_{\rm tot} \rangle \geq 0$. In other words, the fluctuation theorem is entirely consistent with the Second Law of Thermodynamics which, properly interpreted, is a statement about averages, not individual trajectories.

Detailed Fluctuation Relations. For general λ the master Theorem 4 leads to various "stronger" fluctuation relation. To fix the idea we write the generating-function relation (3.25) formally as

$$\sum_R \mathrm{Prob}^F(R_t^F = R)\mathrm{e}^{-\lambda R} = \sum_R \mathrm{Prob}^B(R_t^B = R)\mathrm{e}^{-(1-\lambda)R}. \tag{3.49}$$

where $\mathrm{Prob}^F(R_t^F = R)$ denotes the probability $R_t^F = R$ in the forward process (with initial distribution μ^F) and analogously for the backward process. This is trivially equivalent to

$$\sum_R \mathrm{Prob}^F(R_t^F = R)\mathrm{e}^{-\lambda R} = \sum_R \mathrm{Prob}^B(R_t^B = R)\mathrm{e}^{(1-\lambda)R}. \tag{3.50}$$

Validity for all $\lambda \in \mathbb{R}$ implies

$$\frac{\mathrm{Prob}^B(R_t^B = -R)}{\mathrm{Prob}^F(R_t^F = R)} = \mathrm{e}^{-R} \tag{3.51}$$

which is time-reversal symmetry of the extended forward and backward counting processes, or, equivalently, of the generating function of the forward and backward trajectory functionals.

We point that if r^i and r^f are related by reversal of protocol, then R_t^F and R_t^B measure the same physical quantity in forward and reverse processes (with initial distributions f and g respectively). We can then denote this quantity by R_t without subscript and write (3.51) in the simplified form

$$\frac{p_B(-R_t)}{p_F(R_t)} = \mathrm{e}^{-R_t} \tag{3.52}$$

which is known as the transient fluctuation theorem, see [42] for a more detailed discussion. Here $p_F(R_t)$ denotes the probability distribution for the physical quantity $\mathcal{R}$ in the forward process and $p_B(R_t)$ is the corresponding distribution for the backward process. We point out some examples.

(1) Crooks' fluctuation theorem

Choose for rates obeying time-dependent detailed balance,

$$f(\eta_0) = \mu^*(\eta_0) \qquad \text{and} \qquad g(\eta_t) = \mu^*(\eta_t). \tag{3.53}$$

which allows the identification of R_t as proportional to the dissipated work $W_d = Q + \Delta U - \Delta F$. (3.52) then becomes the fluctuation theorem

$$\frac{p_B(-W_d)}{p_F(W_d)} = \mathrm{e}^{-\mathcal{W}_d/T}, \tag{3.54}$$

which is due to [22].

(2) Evans-Searles fluctuation theorem

For constant rates the forward and reverse processes are obviously identical. If we also take $f = g$ then we can drop the subscripts on the probability distributions $p_F(R_t)$ and $p_B(R_t)$ and one obtains the Evans-Searles fluctuation theorem [30, 31, 97].

A special case corresponds to taking $f = g = \mu^*$. Experimentally, this simply means allowing a system (with time-independent rates) to relax to stationarity before starting the measurement. In this case we can identify R_t with the total entropy change. This yields a fluctuation theorem for entropy changes in the steady state [98]

$$\frac{p(-\Delta S_{\mathrm{tot}})}{p(\Delta \mathcal{S}_{\mathrm{tot}})} = \mathrm{e}^{-\Delta S_{\mathrm{tot}}}, \tag{3.55}$$

which is essentially a stochastic form of the original fluctuation theorem proposed by [29].

Gallavotti-Cohen-Theorem for Stochastic Interacting Particle Systems. In this subsection we focus on *time-independent* rates (in which case backward and forward processes are identical) and discuss the limit $t \to \infty$ of the fundamental fluctuation relation (3.33). For finite state space the result is quite simple [65] and the analogue of the Gallavotti-Cohen theorem [34].

Theorem 5 *(Gallavotti-Cohen symmetry). Let η_t be an ergodic Markov process with finite state space Ω and transition rates $w_{\eta'\eta}$ satisfying either $w_{\eta'\eta}w_{\eta\eta'} > 0$ or $w_{\eta'\eta} = w_{\eta\eta'} = 0$. Furthermore, let J_t be the associated counting processes with antisymmetric increments*

$$r_{\eta'\eta} = \ln\left(\frac{w_{\eta'\eta}}{w_{\eta\eta'}}\right) \tag{3.56}$$

for transitions where $w_{\eta'\eta}w_{\eta\eta'} > 0$ *and* $r_{\eta'\eta} = 0$ *otherwise and let*

$$R_t := \ln f(\eta_0^F) + J_t - \ln g(\eta_t^F) \tag{3.57}$$

with $f(\eta), g(\eta) \neq 0$ *for all* $\eta \in \Omega$ *be the associated extended counting process at time* t*. Then for the generating function*

$$\Phi_\mu(\lambda, t) := \langle \mathrm{e}^{-\lambda R_t} \rangle_\mu \tag{3.58}$$

with any initial measure μ *one has the asymptotic behaviour*

$$\lim_{t\to\infty} \frac{1}{t} \ln \Phi_\mu(\lambda, t) = -E_0(\lambda) \quad \forall \lambda \in \mathbb{R} \tag{3.59}$$

and the Gallavotti-Cohen symmetry

$$E_0(\lambda) = E_0(1-\lambda) \quad \forall \lambda \in \mathbb{R} \tag{3.60}$$

where $E_0(\lambda) \in \mathbb{R}$ *is the lowest eigenvalue of the tilted generator* $\tilde{H}_\lambda$*.*

Proof. By definition

$$\Phi_\mu(\lambda, t) = \langle\, s\,|\hat{g}^\lambda \mathrm{e}^{-\tilde{H}_\lambda t} \hat{f}^{-\lambda}|\, \mu\,\rangle. \tag{3.61}$$

The spectral decomposition

$$\mathrm{e}^{-\tilde{H}_\lambda t} = \sum_k \mathrm{e}^{-E_k(\lambda)t} |\, \Phi_k(\lambda)\,\rangle\langle\, \Psi_k(\lambda)\,| \tag{3.62}$$

into the dyadic product of biorthogonal left and right eigenvectors of $\tilde{H}_\lambda$ yields

$$\Phi_\mu(\lambda, t) = \mathrm{e}^{-E_0(\lambda)t} \sum_k \mathrm{e}^{-(E_k(\lambda)-E_0(\lambda))t} a_k(\lambda) b_k(\lambda) \tag{3.63}$$

where $E_0(\lambda)$ denotes the lowest eigenvalue of $\tilde{H}_\lambda$ and

$$a_k(\lambda) := \langle\, s\,|\hat{g}^\lambda|\, \Phi_k(\lambda)\,\rangle, \quad b_k(\lambda) := \langle\, \Psi_k(\lambda)\,|\hat{f}^{-\lambda}|\, \mu\,\rangle. \tag{3.64}$$

By Perron-Frobenius the lowest eigenvalue corresponding to index $k = 0$ in the decomposition is unique. Thus $\Re(E_k(\lambda) - E_0(\lambda)) > 0$ for all $k \neq 0$. Since in finite state space $a_k(\lambda)$ and $b_k(\lambda)$ are bounded (3.59) is proved. The Gallavotti-Cohen symmetry (3.59) then follows from (3.29) which here reduces to $\tilde{H}_\lambda^T = \tilde{H}_{1-\lambda}$. □

Remark 12. The assumption of finite state space is not a minor technicality, but essential for the validity of the theorem. For infinite state space the coefficients $a_0(\lambda)$ and/or $b_0(\lambda)$ may diverge so that (3.59) is not valid. A simple lattice gas model where this happens is the zero-range process where each lattice site can occupied by an arbitrary number of particles, see [41,43,84].

Remark 13. If (3.59) holds then the symmetry relation (3.60) for the lowest eigenvalue implies the more popular (but not precise) version of the Gallavotti-Cohen symmetry

$$\frac{p(J,t)}{p(-J,t)} = \mathrm{e}^{-Jt} \tag{3.65}$$

for the probability density $p(J,t) = \mathrm{Prob}\,[\,J_t = J\,]$ of the entropy production J_t.

Notice the independence of (3.59) and (3.60) of boundary terms. Heuristically this corresponds to the intuition that $J_t \propto t^\alpha$ for large t with some positive power α, while the boundary terms (which depend only one point in time) are bounded as $t \to \infty$.

As pointed out in [65] the existence of the limit (3.59) implies a large deviation property for the probability distribution $p(j,t) := \mathrm{Prob}\,[\,j_t = j\,]$ of the observed "average" current $j_t = J_t/t$. Specifically, the long-time limiting behaviour is given by

$$p(j,t) \sim \mathrm{e}^{-t\hat{E}(j)} \tag{3.66}$$

where the large deviation function $\hat{E}(j)$ is the Legendre transformation, i.e.,

$$\hat{E}(j) = \max_\lambda \{E_0(\lambda) - \lambda j\}. \tag{3.67}$$

of $E_0(\lambda)$.

4 Dynamical Universality Classes

In the SSEP we have seen that local perturbation spread diffusively with dynamical exponent $z = 2$ while in the ASEP the spreading is superdiffusive and the KPZ-universality class with dynamical exponent $z = 3/2$. For a long time, these were the only universal dynamical exponents known to appear in driven diffusive systems. However, based on numerical evidence and analytical results for other types of models also a dynamical exponent $z = 5/3$ has been reported [20,108]. Thus the question arises which dynamical exponents can generally arise and what universal scaling functions describe the dynamical structure function.

Of course, this question is posed rather imprecisely and very generally. We shall narrow down the quest for an answer to lattice gas models and will arrive at a surprisingly simple (albeit non-rigorous) conclusion:

For n locally conserved species of particles and with dynamics whose large scale behaviour is determined by the slow relaxation of these conserved densities the theory of non-linear fluctuating hydrodynamics, analysed non-rigorously with mode coupling theory, predicts an infinite discrete family of dynamical universality classes whose dynamical exponents are the successive Kepler ratios of neighboring Fibonacci numbers, starting either with diffusion ($z = 2 = 2/1$) or KPZ ($z = 3/2$) or containing at least two dynamical exponents which are given by the golden mean $z = (1 + \sqrt{5})/2$. The scaling form of the non-diffusive and non-KPZ-type modes are z-stable Lévy distributions.

These results are believed to be valid beyond lattice gas models. Indeed, the simplest case prediction that does not involve the KPZ universality class, but a Lévy universality class with $z = 3/2$, has been proved rigorously recently for a harmonic chain with a certain type of conservative noise [10].

4.1 Multi-lane Exclusion Processes

We consider a two-lane asymmetric simple exclusion process on two parallel chains with L sites each and periodic boundary conditions. Particles do not change lanes. We denote the particle occupation number on site k in the first (upper) lane by $\eta_k^{(1)} \in \{0,1\}$, and on the second (lower) lane by $\eta_k^{(2)} \in \{0,1\}$. The total particle number is conserved in each lane and denoted N_λ.

The jump rates for particle on lane λ depend on the particle configuration on the adjacent lane, somewhat similar to the two-lane model introduced in the Introduction and illustrated in Fig. 4. Particles on lane λ jump from site k to site $k+1$ with rate $r_\lambda(k,k+1)$ and from site $k+1$ to site k with rate $\ell_\lambda(k+1,k)$ as given by [78]

$$\begin{aligned} r_1(k,k+1) &= p_1 + b_1 n_k^{(2)} + c_1 n_{k+1}^{(2)} + d_1 n_k^{(2)} n_{k+1}^{(2)} \\ \ell_1(k+1,k) &= q_1 + e_1 n_k^{(2)} + f_1 n_{k+1}^{(2)} + g_1 n_k^{(2)} n_{k+1}^{(2)} \\ r_2(k,k+1) &= p_2 + b_2 n_k^{(1)} + c_2 n_{k+1}^{(1)} + d_2 n_k^{(1)} n_{k+1}^{(1)} \\ \ell_2(k+1,k) &= q_2 + e_2 n_k^{(1)} + f_2 n_{k+1}^{(1)} + g_2 n_k^{(1)} n_{k+1}^{(1)}. \end{aligned} \tag{4.1}$$

In order to write the generator we choose the tensor basis for $2L$ sites and introduce the local operators $\sigma_k^{(i)\pm}$, $\hat{n}_k^{(i)}$ and $\hat{v}_k^{(i)} = 1 - \hat{n}_k^{(i)}$ with Pauli matrices acting non-trivially on site k of chain i (chosen to correspond to the factor $2k + 1 - i$ in the $2L$-fold tensor product). The generator H can be written in term of the diagonal matrices $\hat{r}_k^{(1)}, \hat{\ell}_k^{(1)}, \hat{r}_k^{(2)}, \hat{\ell}_k^{(1)}$ where

$$\hat{r}_k^{(1)} := p_1 + b_1 \hat{n}_k^{(2)} + c_1 \hat{n}_{k+1}^{(2)} + d_1 \hat{n}_k^{(2)} \hat{n}_{k+1}^{(2)} \tag{4.2}$$

$$\hat{\ell}_{k+1} := q_1 + e_1 \hat{n}_k^{(2)} + f_1 \hat{n}_{k+1}^{(2)} + g_1 \hat{n}_k^{(2)} \hat{n}_{k+1}^{(2)} \tag{4.3}$$

(and similar for the other diagonal matrices) as

$$H = -\sum_{i=1}^{2}\sum_{k=1}^{L}\Big[\Big(\sigma_k^{(i)+}\sigma_{k+1}^{(i)-} - \hat{n}_k^{(i)}\hat{v}_{k+1}^{(i)}\Big)\hat{r}_k^{(i)} + \Big(\sigma_k^{(i)-}\sigma_{k+1}^{(i)+} - \hat{v}_k^{(i)}\hat{v}_{k+1}^{(i)}\Big)\hat{\ell}_{k+1}^{(i)}\Big]. \tag{4.4}$$

The invariant measures are easy to characterize for certain constraints on the rates.

Theorem 6. *Let $b_1 - e_1 = c_2 - f_2$, $b_2 - e_2 = c_1 - f_1$, $d_1 = g_1$ and $d_2 = g_2$. The Bernoulli product measure*

$$\mu^*(\eta) = \prod_{k=1}^{L}\prod_{i=1}^{2}\Big[(1-\rho_i)\Big(1-\eta_k^{(i)}\Big) + \rho_i\eta_k^{(i)}\Big] \tag{4.5}$$

is invariant under the dynamics generated by H (4.4).

Proof. We need to show that $H|\,\rho_1,\rho_2\,\rangle = 0$ for the probability vector $|\,\rho_1,\rho_2\,\rangle$ corresponding to the product measure (4.5). We observe that $|\,\rho_1,\rho_2\,\rangle$ is a tensor product, see appendix. Moreover, the diagonal matrices $\hat{r}_k^{(i)}, \hat{\ell}_k^{(i)}$ commute with the non-diagonal hopping matrices $\sigma_k^{(i)\pm}\sigma_{k+1}^{(i)\mp}$. Since

$$\sigma_k^{(i)+}|\,\rho_1,\rho_2\,\rangle = \frac{\rho_i}{1-\rho_i}\hat{v}_k^{(i)}|\,\rho_1,\rho_2\,\rangle, \quad \sigma_k^{(i)-}|\,\rho_1,\rho_2\,\rangle = \frac{1-\rho_i}{\rho_i}\hat{n}_k^{(i)}|\,\rho_1,\rho_2\,\rangle \tag{4.6}$$

one finds

$$H|\,\rho_1,\rho_2\,\rangle = -\sum_{i=1}^{2}\sum_{k=1}^{L}\Big[\Big(\hat{v}_k^{(i)}\hat{n}_{k+1}^{(i)} - \hat{n}_k^{(i)}\hat{v}_{k+1}^{(i)}\Big)\hat{r}_k^{(i)} + \Big(\hat{n}_k^{(i)}\hat{v}_{k+1}^{(i)} - \hat{v}_k^{(i)}\hat{v}_{k+1}^{(i)}\Big)\hat{\ell}_{k+1}^{(i)}\Big]\,|\,\rho_1,\rho_2\,\rangle. \tag{4.7}$$

The r.h.s. contains only diagonal matrices which due to the telescopic property of the sum and periodic boundary conditions sum up to 0. □

Remark 14. Using the similar ideas one constructs multilane processes [77], for a quite general three-lane generalization see [79].

Corollary 5. *The canonical stationary distribution with fixed particle numbers N_i in each lane is uniform.*

Corollary 6. *The fluctuation of the total particle number in the grand canonical Bernoulli product measure described by the compressibility matrix K (1.23) where $\lambda, \mu \in \{1, 2\}$ are given by*

$$\kappa_\lambda := K_{\lambda\lambda} = \rho_\lambda(1-\rho_\lambda), \quad \bar{\kappa} := K_{12} = 0. \tag{4.8}$$

The stationary current vector $\mathbf{j}$ *has components*

$$\begin{aligned} j_1(\rho_1,\rho_2) &= \rho_1(1-\rho_1)(a+\gamma\rho_2), \\ j_2(\rho_1,\rho_2) &= \rho_2(1-\rho_2)(b+\gamma\rho_1). \end{aligned} \tag{4.9}$$

where

$$a = p_1 - q_1,\ b = p_2 - q_2,\ \gamma = b_1 + c_1 - e_1 - f_1. \tag{4.10}$$

For $b = 1$ we recover the totally asymmetric two-lane model of [76] which is a special case of the multi-lane model of [77]. Interestingly, the current-density relation (4.9) is of the same form (1.22) as for the totally asymmetric model (1.21), but with constraint on the range of the parameters a, b, γ. We consider $a = 1$, $\gamma \neq 0$.

4.2 Brief Outline of Nonlinear Fluctuating Hydrodynamics

In order to study fluctuations in this process we follow [102] and take the nonlinear fluctuating hydrodynamics approach together with a mode-coupling analysis of the non-linear equation. We summarize here the main ingredients of this well-established description.

Let us denote microscopic time by the symbol τ rather than t as done in the previous section. We begin by describing the large-scale dynamics of the process under Eulerian scaling where the lattice spacing a is taken to zero such that the macroscopic coordinate $x = ka$ remains finite and where the microscopic time τ is taken to infinity such that the macroscopic time $t = \tau a$ is finite. One then assumes the validity of a law of large numbers such that the local distribution of particles can be described by a coarse-grained local density $\rho_\lambda(x,t)$ of the particle component λ. This leads the system of conservation laws [60,100]

$$\frac{\partial}{\partial t}\boldsymbol{\rho}(x,t) + \frac{\partial}{\partial x}\mathbf{j}(x,t) = 0 \tag{4.11}$$

which follow rigorously or heuristically from the microscopic local conservation of the particle number. Here $\rho_\lambda(x,t)$ is a component of the density vector $\boldsymbol{\rho}(x,t)$, and $j_\lambda(x,t)$ is a component of the current vector $\mathbf{j}(x,t)$ which we regard as column vectors.

According to the assumption of local stationarity the current is a function of x and t only through its dependence on the local conserved densities. Therefore

$$\frac{\partial}{\partial t}\boldsymbol{\rho}(x,t) + J\frac{\partial}{\partial x}\boldsymbol{\rho}(x,t) = 0 \tag{4.12}$$

where J is the current Jacobian with matrix elements $J_{\lambda\mu} = \partial j_\lambda/\partial\rho_\mu$. The product JK of the Jacobian with the compressibility matrix (1.23) is symmetric [38] which guarantees hyperbolicity of the system (4.12) [106]. The eigenvalues v_α of J are the characteristic velocities of the system. If $v_1 \neq v_2$ the system is called strictly hyperbolic.

In order to extract information from this non-linear system of PDE's one expands the local densities $\rho_\lambda(x,t) = \rho_\lambda + u_\lambda(x,t)$ around their long-time stationary values ρ_λ. To linear order one gets

$$\partial_t \mathbf{u} = -J\partial_x \mathbf{u}. \tag{4.13}$$

where J is now fixed at the values stationary values ρ_λ. We transform to normal modes $\boldsymbol{\phi} = R\mathbf{u}$ where $RJR^{-1} = \mathrm{diag}(v_\alpha)$ and the transformation matrix R is normalized such that $RKR^T = \mathbb{1}$. Thus we get, since we have a linear system,

$$\phi_\alpha(x,t) = \phi_0(x + v_\alpha t) \tag{4.14}$$

with initial data $\phi_\alpha(x,0) = \phi_0(x)$. This result demonstrates the significance of the eigenvalues of the current Jacobian which are called characteristic velocities v_α. They are the velocities at which perturbations of the flat stationary density profile move.

In order to study the effect of the non-linearity we now expand to second order. This yields

$$\partial_t \mathbf{u} = -\partial_x \left(J\mathbf{u} + \frac{1}{2}\mathbf{u}^T \mathbf{H}\mathbf{u} \right) \tag{4.15}$$

where $\mathbf{H}$ is a column vector whose entries $(\mathbf{H})_\lambda = H^\lambda$ are the Hessians with matrix elements $H^\lambda_{\mu\nu} = \partial^2 j_\lambda/(\partial\rho_\mu \partial\rho_\nu)$. The term $\mathbf{u}^T H^\lambda \mathbf{u}$ denotes the inner product in component space. One recognizes in (4.15) a system of coupled Burgers equations.

Finally, the effect of fluctuations, which occur on finer space-time scales where $t = \tau a^z$ with dynamical exponent $z > 1$, can be captured by adding phenomenological diffusion matrix D and white noise terms $\xi_i(x,t)$. For quadratic nonlinearities (4.12) then yields

$$\partial_t \mathbf{u} = -\partial_x \left(J\mathbf{u} + \frac{1}{2}\mathbf{u}^T \mathbf{H}\mathbf{u} - D\partial_x \mathbf{u} + B\boldsymbol{\xi} \right). \tag{4.16}$$

If the quadratic non-linearity is absent one has diffusive behaviour, up to possible logarithmic corrections that may arise from cubic non-linearities [27]. In normal modes one has

$$\partial_t \phi_\alpha = -\partial_x \left(v_\alpha \phi_\alpha + \boldsymbol{\phi}^T G^\alpha \boldsymbol{\phi} - \partial_x (\tilde{D}\boldsymbol{\phi})_\alpha + (\tilde{B}\boldsymbol{\xi})_\alpha \right) \tag{4.17}$$

with $\tilde{D} = RDR^{-1}$, $\tilde{B} = RB$ and

$$G^\alpha = \frac{1}{2} \sum_\lambda R_{\alpha\lambda} (R^{-1})^T H^\lambda R^{-1} \tag{4.18}$$

are the called the mode coupling matrices.

Consider now the dynamical structure matrix $\bar{S}_k(t)$ of the microscopic model defined on the lattice. Its matrix elements are the dynamical structure functions

$$\bar{S}^{\lambda\mu}_k(t) := \langle (n^{(\lambda)}_k(t) - \rho_\lambda)(n^{(\mu)}_0(t) - \rho_\mu) \rangle \tag{4.19}$$

which measure density fluctuations in the stationary state. In normal modes one then has

$$S_k^{\alpha\beta}(t) = [R\bar{S}_k(t)R^T]_{\alpha\beta} = \langle \phi_k^\alpha(t)\phi_0^\beta(0)\rangle \tag{4.20}$$

where the transformation R acts on the lattice density vector.

We focus from now on strictly hyperbolic systems where all characteristic velocities are different. Then one expects that the off-diagonal elements of S decay quickly and for long times and large distances the diagonal elements which we denote by

$$S_\alpha(x,t) := S^{\alpha\alpha}(x,t) \tag{4.21}$$

with initial value $S_\alpha(x,0) = \delta(x)$ remain significant. One expects the scaling form

$$S_\alpha(x,t) \sim t^{-1/z_\alpha} f_\alpha((x - v_\alpha t)^{z_\alpha}/t) \tag{4.22}$$

with a dynamical exponent z_α that may be different for the different modes. The exponent in the power law prefactor is determined by mass conservation.

The dynamical structure function can interpreted as describing the stationary two-time correlations of the local density fluctuations. Alternatively, it can regarded as the expectation of the local density evolving from an initial distribution that is microscopically peaked around the origin $k = 0$ [78], corresponding to a δ-peak on macroscopic scale. Thus the characteristic velocity describes the velocity at which the center of mass of these peaks move [76] and the dynamical exponent describes the spreading around the center of mass.

4.3 Fibonacci Universality Classes

In order to analyze the system of nonlinear stochastic PDE's in more detail we employ mode coupling theory [80,102]. The starting point for computing the $S_\alpha(x,t)$ are the one-loop mode coupling equations

$$\partial_t S_\alpha(x,t) = (-v_\alpha\partial_x + D_\alpha\partial_x^2)S_\alpha(x,t) + \int_0^t \mathrm{d}s \int_{-\infty}^{\infty} \mathrm{d}y\, S_\alpha(x-y,t-s)\partial_y^2 M_{\alpha\alpha}(y,s) \tag{4.23}$$

with the diagonal element $D_\alpha := \tilde{D}_{\alpha\alpha}$ of the phenomenological diffusion matrix and the memory kernel

$$M_{\alpha\alpha}(y,s) = 2\sum_{\beta,\gamma}(G_{\beta\gamma}^\alpha)^2 S_\beta(y,s)S_\gamma(y,s). \tag{4.24}$$

The strategy is to rewrite this equation in terms of the Fourier transform

$$\hat{S}_\alpha(p,t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-ipx} S_\alpha(x,t). \tag{4.25}$$

and then to plug into this equation the scaling ansatz

$$\hat{S}_\alpha(p,t) \sim \mathrm{e}^{-iv_\alpha pt}\hat{f}_\alpha(p^{z_\alpha}t). \tag{4.26}$$

Remarkably, the coupled system of equation then becomes exactly solvable [79, 80]. One obtains equations for the dynamical exponents arising from requiring non-trivial scaling solutions and using the known results $z = 3/2$ for KPZ and $z = 2$ for diffusion. In a next step one can then solve for the actual scaling functions.

The scaling behaviour of the solutions of (4.23) is turns out to be determined by the diagonal terms $G^\alpha_{\beta\beta}$ of the mode coupling matrices G^α. We define the set

$$\mathbb{I}_\alpha := \{\beta : G^\alpha_{\beta\beta} \neq 0\} \tag{4.27}$$

of non-zero diagonal mode coupling coefficients. This means that $\mathbb{I}_\alpha$ is the set of modes β that give rise to a non-linear term in the time evolution of the mode α whose dynamical exponent and scaling function one wishes to compute.

The equations that determine the dynamical exponents for a system with n modes are then:

$$z_\alpha = \begin{cases} 2 & \text{if } \mathbb{I}_\alpha = \emptyset \\ 3/2 & \text{if } \alpha \in \mathbb{I}_\alpha \\ \min_{\beta\in\mathbb{I}_\alpha}\left[\left(1+\frac{1}{z_\beta}\right)\right] & \text{else} \end{cases} \tag{4.28}$$

and

$$1 < z_\alpha \leq 2 \quad \forall\alpha \tag{4.29}$$

Remarkably the solution to this non-linear recursion yields as possible dynamical exponents the Kepler ratios of neighbouring Fibonacci numbers

$$z_\alpha = 2, 3/2, 5/3, 8/5, \ldots \tag{4.30}$$

or its limiting value which is the golden mean $z_\alpha = (1+\sqrt{5})/2$.

Specifically, if all self-coupling term $G^\alpha_{\alpha\alpha}$ vanish then mode α is diffusive with $z_\alpha = 2$ and Gaussian scaling function (except for possible logarithmic corrections).

If $G^\alpha_{\alpha\alpha} \neq 0$ and there is no diffusive mode β such that $G^\alpha_{\beta\beta} \neq 0$ then the mode is KPZ with $z_\alpha = 3/2$ and Prähofer-Spohn scaling function [81,82].

If $G^\alpha_{\alpha\alpha} \neq 0$, but there is a diffusive mode β such that $G^\alpha_{\beta\beta} \neq 0$ then again $z_\alpha = 3/2$, but the scaling function is unknown [103]. A lattice model with this form of the mode coupling matrix has been proposed recently [96].

If the self-coupling $G^\alpha_{\alpha\alpha} = 0$ but some $G^\alpha_{\beta\beta} \neq 0$ then the mode is a Lévy mode where the scaling function is an asymmetric Lévy distribution [79,80,102]. The lowest Fibonacci mode has $z = 3/2$ like KPZ, but is not KPZ. The scaling function for this mode satisfies a fractional diffusion equation which has been proved rigorously in a system of harmonic oscillators that are perturbed by a conservative noise [10].

Thus non-linear hydrodynamics yields an infinite discrete family of dynamical universality classes. The ubiquitous diffusive universality class and the celebrated KPZ universality class are the lowest members of this family.

4.4 Ballistic Universality Class in Conditioned Dynamics

Finally we consider the question whether one can have a "ballistic" universality class with $z = 1$. Such a universality class indeed exists as shown by Spohn for an exclusion process with long-range interactions [101]. No models with short-range interactions and $z = 1$ are known. However, the model of Spohn arises as "conditioned" dynamics of the usual ASEP, viz. the ASEP conditioned on carrying a large current. This observation then points to the existence of a much larger family of models with $z = 1$ with the conjecture that *all* stationary space-time correlation functions can be predicted from conformal invariance [57]. This conjecture arises from the mapping to quantum spin systems and then using well-established properties of the quantum ground state which is known to be described by conformal field theory.

5 Conclusions

It has been realized in recent years that the stochastic time evolution of many stochastic interacting particle systems can be mapped to quantum spin systems, and in special one-dimensional cases to integrable quantum chains. This insight has made available the tool box of quantum mechanics for these interacting particle systems far from equilibrium. With these methods many new exact results for their dynamical and stationary properties have been derived. It is also amusing to note that the Hamiltonians for such systems are mostly not hermitian and therefore from a quantum mechanical point of view not interesting. Stochastic interacting particle systems which can be described in this way comprise a large variety of phenomena in physics and beyond. In this way one obtains detailed information about the microscopic properties and large-scale fluctuations of lattice gas models with conserved particle species.

Going beyond these exact results we have shown that non-linear fluctuating hydrodynamics predicts an infinite discrete family of dynamical universality classes whose dynamical exponents are the Kepler ratios of neighbouring Fibonacci numbers. This fact encourages the search for other discrete families of nonequilibrium universality classes and asks for a formal mathematical proof at least for some specific class of models.

A Some Linear and Multilinear Algebra

In order to obtain more information about stochastic lattice gas models it will turn out to be convenient to write the generator of the process as a matrix which is called intensity matrix. Many probabilistic operations and notions then have a natural counterpart in linear algebra which we summarize here in elementary form. Indeed, much of what is presented is trivial, but perhaps necessary to write down in order to point out technically important subtleties and to introduce notation for the less common applications such the Kronecker product of

matrices, sometimes also called outer product, which is essential for the choice of basis of the intensity matrix for lattice models.

In the following and throughout this work we use the Kronecker-symbol defined by

$$\delta_{\alpha,\beta} = \begin{cases} 1 \text{ if } \alpha = \beta \\ 0 \text{ else} \end{cases} \tag{A.1}$$

for α, β from any set. Complex conjugation is denoted by a bar as e.g. in $\overline{z}$.

Matrices and Vectors. A $m \times n$ matrix A is a number array

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots \\ A_{21} & A_{22} & A_{23} & \dots \\ A_{31} & A_{32} & A_{33} & \dots \\ A_{41} & A_{42} & A_{43} & \dots \\ A_{51} & A_{52} & A_{53} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

with $m \geq 1$ rows and $n \geq 1$ columns and matrix elements A_{kl} in row k and column l. The matrix elements A_{kl} will be mostly real numbers, but they can also be complex in certain applications. Hence we shall generally assume $A_{kl} \in \mathbb{Z}$. We discuss some special cases.

(a) If $m = n = 1$ the matrix reduces a single number and we shall not differentiate between numbers and 1×1-matrices.
(b) If $n = 1$ and $m > 1$ a matrix Φ is a column array of m numbers. We call such a matrix a *ket-vector* that we denote by the so-called ket-symbol $|\,\Phi\,\rangle$. The matrix elements Φ_{1l} with $1 \leq l \leq m$ will be denoted in simplified form by Φ_l and called components of the ket-vector. Thus

$$|\,\Phi\,\rangle = \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_m \end{pmatrix}.$$

The vector with $\Phi_i = \delta_{ik}$ is a canonical basis vector of the vector space $\mathbb{Z}^m$ denoted by $|\,e_k\,\rangle$. The set $B_m := \{|\,e_k\,\rangle : k \in \{1, \dots, m\}\}$ spans $\mathbb{Z}^m$ and is called the canonical basis.
(c) If $m = 1$ and $n > 1$ a matrix Ψ is a row array of n numbers. We call such a matrix a *bra-vector* that we denote by the so-called bra-symbol $\langle\,\Psi\,|$. The matrix elements Ψ_{k1} with $1 \leq k \leq n$ will be denoted in simplified form as Ψ_k and called components of the bra-vector. Thus

$$\langle\,\Psi\,| = (\Psi_1, \Psi_2, \dots, \Psi_n)\,.$$

Defining $\langle\,e_k\,| = |\,e_k\,\rangle^T$ one realizes that the set $B^* := \{\langle\,e_k\,| : k \in \{1, \dots, n\}\}$ spans $\mathbb{Z}^n$. Since any finite-dimensional vector space is isomorphic to its dual,

we can think of the bra-vectors $\mathbb{B}_n^*$ as representing the canonical basis of the dual space $\mathbb{Z}^{n*} \cong C^n$.

The letter or number inside the ket-symbol $|\cdot\rangle$ or the bra-symbol $\langle\cdot|$ is not to be understood as the argument of some function, but just as a symbol that collectively represents the components of the vector. When we use the term matrix we shall tacitly assume that $m, n \geq 2$. The distinction between "proper" matrices on the one hand and the two types of vectors or simple numbers on the other hand is useful because many fundamental linear algebra operations can be represented as products involving numbers, bra- and ket-vectors and proper matrices with more than one column or row.

We usually denote proper matrices by capital letters or small letters with circumflex accent as e.g. in $\hat{a}$. The unit matrix of dimension $n > 2$ with components $A_{kl} = \delta_{k,l}$ is denoted by $\mathbf{1}$ and for $n = 2$ we use the notation $\mathbb{1}$. Since multiplication of a vector with the unit matrix is the same as multiplication with the scalar unity 1 of the field F we do not usually differentiate between the two operations, i.e., in equations for matrices we often write a multiple $x\mathbf{1}$ of the unity matrix simply as x.

Addition and Multiplication of Matrices. Any two matrices A and B which have the same number of rows and columns can be multiplied by a number and added to form a matrix $C = xA + yB$ with the rule that $C_{kl} = xA_{kl} + yB_{kl}$ where $x, y \in \mathbb{Z}$. Square matrices with $m = n$ form a ring with a multiplication rule that can be generalized to non-square matrices as follows.

Definition 12 *(Matrix product). For $m, n, p \geq 1$ let A be a $m \times p$-matrix and B be a $p \times n$-matrix, both with matrix elements in some field F. The matrix product AB is an $m \times n$ matrix C with matrix elements $C_{kl} \equiv (AB)_{kl} \in F$ given by*

$$C_{kl} = \sum_{j=1}^{p} A_{kj} B_{jl}, \quad 1 \leq k \leq m, \quad 1 \leq l \leq n. \tag{A.2}$$

Square matrices A, B of the same dimension $m = n = p$ satisfying

$$[A\,,\,B] := AB - BA = 0 \tag{A.3}$$

are said to commute.

Notice that unless $m = n$ the reverse product BA is not defined since the number of columns in the first factor must be equal to the number of rows in the second factor of any matrix product. For a square matrix A the p^{th} power of A is denoted A^p and is defined for strictly positive integers $p \in \mathbb{N}$ by iteration of (A.2). By convention $A^0 = \mathbf{1}$.

We discuss separately the special cases where at least one of the three number m, n, p is equal to one.

(a) If $n = 1$ and $p, m > 1$ the we can write the matrix B as a ket-vector $|\Phi\rangle$ with components $\Phi_k := B_{k1}$, $k \in \{1, \ldots, p\}$. Then also the matrix product C is

a ket-vector (with m components given by (A.2)) and the matrix product can be interpreted as a linear mapping $|\,\Phi\,\rangle \mapsto |\,\tilde{\Phi}\,\rangle$ given by $|\,\tilde{\Phi}\,\rangle = A|\,\Phi\,\rangle$, corresponding to the standard right multiplication of a matrix A with the column vector $|\,\Phi\,\rangle$.
(b) Likewise, for $m = 1$ and $p, n > 1$ we can write $A = \langle\,\Psi\,|$ as a bra-vector with p components with components $\Psi_l := A_{1l}$ and find that the matrix product is a linear mapping $\langle\,\Psi\,| \mapsto \langle\,\tilde{\Psi}\,|$ that yields the bra-vector $\langle\,\tilde{\Phi}\,| = \langle\,\Psi\,|B$ with n components given by (A.2), corresponding to the left multiplication of a matrix B with the row vector $\langle\,\Psi\,|$.
(c) If $p = 1$ and $m, n > 1$ then the matrix product actually turns into a product of two vectors. It maps a m-component ket-vector $|\,\Phi\,\rangle$ (=$m \times 1$-matrix A) with components $\Phi_k := A_{k1}$ and an n-component bra-vector $\langle\,\Psi\,|$ (=$1 \times n$-matrix B) with components $\Psi_l := B_{1l}$ into a proper $m \times n$ matrix

$$C = |\,\Phi\,\rangle\langle\,\Psi\,| \tag{A.4}$$

with matrix elements $C_{kl} = \Phi_k\Psi_l$ as given by (A.2). This mapping, called dyadic product, is a special form of the Kronecker product discussed below.
(d) For $m = n = 1$ the matrix product reduces to a single number $C = \langle\,\Psi\,||\,\Phi\,\rangle = C_{11} \in F$ with

$$\langle\,\Psi\,||\,\Phi\,\rangle = \sum_{i=1}^{p} \Psi_i\Phi_i \equiv \langle\Psi|\,\Phi\rangle. \tag{A.5}$$

It defines a bilinear mapping $(\langle\,\Psi\,|, |\,\Phi\,\rangle) \mapsto C_{11}$ which can be interpreted as a dual pairing $d : \mathfrak{V}^* \times \mathfrak{V} \to F$, $(\langle\,\Psi\,|, |\,\Phi\,\rangle) \mapsto \langle\Psi|\,\Phi\rangle$ since it is natural to regard the bra-vector to be an element of the vector space dual to the vector space to which the ket-vector belongs. This motivates the simplified notation $\langle\Psi|\,\Phi\rangle$ of this matrix product with only one vertical bar.

Specifically, for the basis vectors we obtain from (A.5) the biorthogonality relation

$$\langle e_i|\,e_j\rangle = \delta_{ij}. \tag{A.6}$$

Notice the difference between the dual pairing (A.5) and the scalar product $s : \mathfrak{V} \times \mathfrak{V} \to F$ defined by the sesquilinear form $(|\,\Phi'\,\rangle, |\,\Phi\,\rangle) \mapsto \langle\Phi', \Phi\rangle := \sum_{i=1}^{p} \overline{\Phi}'_i\Phi_i$ which is linear in the second argument, but antilinear in the first. When $\langle\,\Phi'\,|$ has only real components (as is the case in most of our applications) this distinction is irrelevant, but should nevertheless be kept in mind.

The Kronecker Product. The Kronecker product $A \otimes B$ is defined for arbitrary rectangular matrices (including vectors and numbers) as follows.

Definition 13 *(Kronecker product). Let A and B be two finite-dimensional matrices with $m_A \geq 1$ ($m_B \geq 1$) rows and $n_A \geq 1$ ($n_B \geq 1$) columns with matrix elements A_{ij} and B_{kl} respectively. The Kronecker product $A \otimes B$ is a $m_Am_B \times n_An_B$-matrix C with matrix elements*

$$C_{(i-1)m_B+k,(j-1)n_B+l} = A_{ij}B_{kl} \tag{A.7}$$

with $i \in \{1, \ldots, m_A\}, j \in \{1, \ldots, n_A\}, k \in \{1, \ldots, m_B\}, l \in \{1, \ldots, n_B\}$.

Alternatively we can write

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & A_{13}B & \dots \\ A_{21}B & A_{22}B & A_{23}B & \dots \\ A_{31}B & A_{32}B & A_{33}B & \dots \\ A_{41}B & A_{42}B & A_{43}B & \dots \\ A_{51}B & A_{52}B & A_{53}B & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Here each matrix "element" is itself a matrix, viz. the matrix B multiplied by the number A_{ij}. In general $A \otimes B \neq B \otimes A$. For $p \in \mathbb{N}_0$ the p-fold Kronecker product of a matrix A with itself is denoted by $A^{\otimes p}$ with the convention that $A^{\otimes 1} := A$ and $A^{\otimes 0} := 1$ where 1 is the unit element of F and not the unit matrix. We discuss special cases.

(a) Consider $n_A = n_B = 1$, i.e., the Kronecker product of ket-vectors $|\Phi^1\rangle, |\Phi^2\rangle$ with components Φ^1_i where $i \in \{1, \dots, m_A\}$ and Φ^2_k where $k \in \{1, \dots, m_B\}$. The tensor product $|\Phi^1\rangle \otimes |\Phi^2\rangle$ is a column vector of dimension $m_A m_B$ denoted by $|\Phi^1, \Phi^2\rangle$ and has factorized components $(|\Phi^1, \Phi^2\rangle)_{(i-1)m_B+k} = \Phi^1_i \Phi^2_k$. Specifically, for the canonical basis vectors one gets $|e_i\rangle \otimes |e_k\rangle \equiv |e_i, e_k\rangle = |e_{(i-1)m_B+k}\rangle$. Thus the Kronecker product of two canonical basis vectors yields a canonical basis vector. The set $\mathbb{B}_{m_A m_B} := \{|e_{(i-1)m_B+k}\rangle : (i,k) \in \{1, \dots, m_A\} \times \{1, \dots, m_B\}\}$ forms the canonical basis of the tensor space $\mathbb{Z}^{m_A} \otimes \mathbb{Z}^{m_B} \cong \mathbb{Z}^{m_A m_B}$.
(b) Similarly, for $m_A = m_B = 1$, i.e., for bra-vectors $\langle\Psi^1|, \langle\Psi^2|$ with components Ψ^1_j where $j \in \{1, \dots, n_A\}$ and Ψ^2_l where $l \in \{1, \dots, n_B\}$ the tensor product $\langle\Psi^1| \otimes \langle\Psi^2|$ is a row vector of dimension $n_A n_B$ denoted by $\langle\Psi^1, \Psi^2|$. It has factorized components $(\langle\Psi^1, \Psi^2|)_{(j-1)n_B+l} = \Psi^1_j \Psi^2_l$ and for the canonical basis vectors one gets $\langle e_j, e_l| = \langle e_{(j-1)n_B+l}|$.
(c) For the Kronecker product of a bra-vector $\langle\Psi|$ and a ket-vector $|\Phi\rangle$ the Definition 13 yields

$$\langle\Psi| \otimes |\Phi\rangle = |\Phi\rangle \otimes \langle\Psi| = |\Phi\rangle\langle\Psi| \tag{A.8}$$

with the dyadic product (A.4).

The Kronecker product is associative. Multiple Kronecker products of matrices define multilinear maps of the multiple tensor product of vector spaces defined by iterating the Kronecker product Definition 13. They satisfy the multiplication rule

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{A.9}$$

where we assume that the matrix products AC and BD are defined by (A.2).

We note an important factorization property of the dual pairing of Kronecker products of vectors which is an immediate consequence of the multilinearity of the Kronecker product encoded in (A.7).

Proposition 5. *Let $\langle\Psi^k|$ ($|\Phi^k\rangle$) be a bra-vector (ket-vector) of dimension d_k with components $\Psi^k_i \in \mathbb{Z}$ ($\Phi^k_i \in \mathbb{Z}$) and $\langle\Psi^1, \Psi^2, \dots, \Psi^L| = \langle\Psi^1| \otimes \langle\Psi^2| \otimes \cdots \otimes$*

$\langle \Psi^L |$ $(|\Phi^1, \Phi^2, \ldots, \Phi^L\rangle = |\Phi^1\rangle \otimes |\Phi^2\rangle \otimes \cdots \otimes |\Phi^L\rangle)$ *be the L-fold Kronecker product of these vectors. Then the dual pairing factorizes as*

$$\langle \Psi^1, \Psi^2, \ldots, \Psi^L | \Phi^1, \Phi^2, \ldots, \Phi^L \rangle = \prod_{k=1}^{L} \langle \Psi^k | \Phi^k \rangle \tag{A.10}$$

with $\langle \Psi^k | \Phi^k \rangle$ *given by (A.5).*

When $\langle \Psi^k | = \langle \Psi |$ for all $k \in \{1, \ldots, L\}$ then we write $\langle \Psi^1, \Psi^2, \ldots, \Psi^L | = \langle \Psi |^{\otimes L}$ and analogously for ket-vectors and proper matrices A.

Finally we introduce *local operators* which act non-trivially only on component k in an L-fold tensor space. For simplicity we assume equal dimensions $d := d_1 = d_2 = \cdots = d_L$.

Definition 14 *(Local operator). Let* $\mathbf{1}$ *be the d-dimensional unit matrix and A be an arbitrary square matrix of dimension* $d \geq 1$*. The local operator* A_k *is the Kronecker product*

$$A_k := \mathbf{1}^{\otimes(k-1)} \otimes A \otimes \mathbf{1}^{\otimes(L-k)}. \tag{A.11}$$

Notice the difference between the number $1 \in \mathbb{Z}$ and the unit matrix $\mathbf{1}$ in this definition. The expression "local operator" come from the fact that when acting on a tensor vector $|\Phi^1, \ldots, \Phi^L\rangle$ only the k^{th} factor is changed by the action of A_k. More precisely,

$$A_k \left(|\Phi^1\rangle \otimes \cdots \otimes |\Phi^k\rangle \otimes \cdots \otimes |\Phi^L\rangle \right) = |\Phi^1\rangle \otimes \cdots \otimes |\tilde{\Phi}^k\rangle \otimes \cdots \otimes |\Phi^L\rangle. \tag{A.12}$$

where $|\tilde{\Phi}^k\rangle = A|\Phi^k\rangle$.

From (A.9) one finds

$$A_k B_k = (AB)_k \tag{A.13}$$

which is equal to $B_k A_k$ if and only if $AB = BA$. On the other hand, by construction one has for two square matrices A, B of dimension k the commutation relation

$$A_k B_l = B_l A_k \text{ for } k \neq l \tag{A.14}$$

even when $AB \neq BA$. In order to avoid confusion concerning the role of the indices we point out that for $L = 2$ and $[A\,,\,B] \neq 0$ we have

$$A \otimes B = A_1 B_2 = B_2 A_1 \neq B \otimes A = B_1 A_2 = A_2 B_1. \tag{A.15}$$

We also note that for matrices $A^{(k)}$ one has

$$A_1^{(1)} A_2^{(2)} \ldots A_L^{(L)} = A^{(1)} \otimes A^{(2)} \otimes \cdots \otimes A^{(L)}. \tag{A.16}$$

The upper index defines the matrix while the lower index defines its position in the L-fold Kronecker product. We stress that $A^{(k)}$ is a matrix of dimension d while $A_k^{(k)}$ is a matrix of dimension d^L.

From Proposition 5 one finds for d-dimensional square matrices $A^{(k)}$ the factorization property

$$\langle \Psi^1, \Psi^2, \ldots, \Psi^L | A_1^{(1)} A_2^{(2)} \ldots A_L^{(L)} | \Phi^1, \Phi^2, \ldots, \Phi^L \rangle = \prod_{k=1}^{L} \langle \Psi^k | A^{(k)} | \Phi^k \rangle. \tag{A.17}$$

We write explicitly two special cases of particular importance:

$$\frac{\langle \Psi^1, \Psi^2, \ldots, \Psi^L | A_k | \Phi^1, \Phi^2, \ldots, \Phi^L \rangle}{\langle \Psi^1, \Psi^2, \ldots, \Psi^L | \Phi^1, \Phi^2, \ldots, \Phi^L \rangle} = \frac{\langle \Psi^k | A | \Phi^k \rangle}{\langle \Psi^k | \Phi^k \rangle} \tag{A.18}$$

$$(\langle \Psi |)^{\otimes L} (| \Phi \rangle)^{\otimes L} = \langle \Psi | \Phi \rangle^L. \tag{A.19}$$

These computational properties of the matrix product (A.2) and of the Kronecker product defined in Definition 13 will be exploited throughout these notes.

References

1. Alcaraz, F.C., Rittenberg, V.: Reaction-diffusion processes as physical realizations of Hecke algebras. Phys. Lett. B **314**, 377–380 (1993)
2. Amir, G., Corwin, I., Quastel, J.: Probability distribution of the free energy of the continuum directed random polymer in 1 + 1 dimensions. Commun. Pure Appl. Math. **64**, 466 (2011)
3. Bahadoran, C.: Hydrodynamics and hydrostatics for a class of asymmetric particle systems with open boundaries. Commun. Math. Phys. **310**(1), 1–24 (2012)
4. Balázs, M., Farkas, G., Kovács, P., Rákos, A.: Random walk of second class particles in product shock measures. J. Stat. Phys. **139**(2), 252–279 (2010)
5. Baxter, R.J.: Exactly Solved Models in Statistical Mechanics. Academic Press, New York (1982)
6. Belitsky, V., Schütz, G.M.: Diffusion and coalescence of shocks in the partially asymmetric exclusion process. Electron. J. Prob. **7**, 1–21 (2002). Paper No. 11
7. Belitsky, V., Schütz, G.M.: Self-duality for the two-component asymmetric simple exclusion process. J. Math. Phys. **56**, 083302 (2015)
8. Belitsky, V., Schütz, G.M.: Self-duality and shock dynamics in the n-species priority ASEP. Stoch. Proc. Appl. **128**, 1165–1207 (2018)
9. Ben Arous, G., Corwin, I.: Current fluctuations for TASEP: a proof of the PrähoferSpohn conjecture. Ann. Probab. **39**, 104–138 (2011)
10. Bernardin, C., Gonçalves, P., Jara, M.: 3/4-fractional superdiffusion in a system of harmonic oscillators perturbed by a conservative noise. Arch. Ration. Mech. Anal. **220**, 505–542 (2016)
11. Bertini, L., De Sole, A., Gabrielli, D., Jona Lasinio, G., Landim, C.: Macroscopic fluctuation theory. Rev. Mod. Phys. **87**, 593–636 (2015)
12. Blythe, R.A., Evans, M.R.: Nonequilibrium steady states of matrix product form: a solvers guide. J. Phys. A Math. Theor. **40**, R333–R441 (2007)
13. Bochkov, G.N., Kuzovlev, Y.E.: General theory of thermal fluctuations in nonlinear systems. Sov. Phys.—JETP **45**, 125–130 (1977)
14. Bochkov, G.N., Kuzovlev, Y.E.: Fluctuation-dissipation relations for nonequilibrium processes in open systems. Sov. Phys.–JETP **49**, 543–551 (1979)

15. Bochkov, G.N., Kuzovlev, Y.E.: Nonlinear fluctuation-dissipation relations and stochastic models in nonequilibrium thermodynamics I. Generalized fluctuation-dissipation theorem. Phys. A **106**, 443–479 (1981)
16. Borodin, A., Petrov, L.: Lectures on integrable probability: stochastic vertex models and symmetric functions. In: Schehr, G., Altland, A., Fyodorov, Y.V., O'Connell, N., Cugliandolo, L.F. (eds.) Lecture Notes of the Les Houches Summer School, vol. 104, July 2015
17. Calabrese, P., Le Doussal, P.: Exact solution for the Kardar-Parisi-Zhang Equation with flat initial conditions. Phys Rev. Lett. **106**, 250603 (2011)
18. Calabrese, P., Le Doussal, P.: The KPZ equation with flat initial condition and the directed polymer with one free end. J. Stat. Mech. **2012**, P06001 (2012)
19. Cancrini, N., Galves, A.: Approach to equilibrium in the symmetric simple exclusion process. Markov Processes Relat. Fields **1**, 175–184 (1995)
20. Cipriani, P., Denisov, S., Politi, A.: From anomalous energy diffusion to Lévy walks and heat conductivity in one-dimensional systems. Phys. Rev. Lett. **94**, 244301 (2005)
21. Chetrite, R., Touchette, H.: Nonequilibrium Markov processes conditioned on large deviations. Ann. Henri Poincaré **16**(9), 2005–2057 (2015)
22. Crooks, G.E.: Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys. Rev. E **60**, 2721–2726 (1999)
23. Derrida, B., Domany, E., Mukamel, D.: An exact solution of a one-dimensional asymmetric exclusion model with open boundaries. J. Stat. Phys. **69**, 667 (1992)
24. Derrida, B., Evans, M.R., Hakim, V., Pasquier, V.: Exact solution of a 1D asymmetric exclusion model using a matrix formulation. J. Phys. A: Math. Gen. **26**, 1493–1517 (1993)
25. Derrida, B., Janowsky, S.A., Lebowitz, J.L., Speer, E.R.: Exact solution of the totally asymmetric simple exclusion process: shock profiles. J. Stat. Phys. **73**, 813–842 (1993)
26. Derrida, B.: An exactly soluble nonequilibrium system: the asymmetric simple exclusion process. Phys. Rep. **301**, 65–83 (1998)
27. Devillard, P., Spohn, H.: Universality class of interface growth with reflection symmetry. J. Stat. Phys. **66**, 1089–1099 (1992)
28. Doi, M.: Second quantization representation for classical many-particle system. J. Phys. A: Math. Gen. **9**, 1465–1477 (1976)
29. Evans, D.J., Cohen, E.G.D., Morriss, G.P.: Probability of second law violations in shearing steady states. Phys. Rev. Lett. **71**(15), 2401–2404 (1993)
30. Evans, D.J., Searles, D.J.: Equilibrium microstates which generate second law violating steady states. Phys. Rev. E **50**(2), 1645–1648 (1994)
31. Evans, D.J., Searles, D.J.: The fluctuation theorem. Adv. Phys. **51**, 1529–1585 (2002)
32. Ferrari, P.A., Kipnis, C., Saada, E.: Microscopic structure of travelling waves in the asymmetric simple exclusion process. Ann. Probab. **19**(1), 226–244 (1991)
33. Ferrari, P.A., Fontes, L.R.G.: Shock fluctuations in the asymmetric simple exclusion process. Probab. Theory Relat. Fields **99**, 305–319 (1994)
34. Gallavotti, G., Cohen, E.G.D.: Dynamical ensembles in stationary states. J. Stat. Phys. **80**, 931–970 (1995)
35. Giardinà, C., Kurchan, J., Redig, F., Vafayi, K.: Duality and hidden symmetries in interacting particle systems. J. Stat. Phys. **135**, 25–55 (2009)
36. Gradshteyn, I.S., Ryzhik, I.M.: Tables of Integrals, Series and Products. Academic Press, Orlando (1981)

37. Grassberger, P., Scheunert, M.: Fock-space methods for identical classical objects. Fortschr. Phys. **28**, 547–578 (1980)
38. Grisi, R., Schütz, G.M.: Current symmetries for particle systems with several conservation laws. J. Stat. Phys. **145**, 1499–1512 (2011)
39. Gwa, L.H., Spohn, H.: Bethe solution for the dynamical-scaling exponent of the noisy Burgers equation. Phys. Rev. A **46**(2), 844–854 (1992)
40. Halpin-Healy, T., Takeuchi, K.A.: A KPZ cocktail- shaken, not stirred: toasting 30 years of kinetically roughened surfaces. J. Stat. Phys. **160**(4), 794–814 (2015)
41. Harris, R.J., Rákos, A., Schütz, G.M.: Breakdown of Gallavotti-Cohen symmetry for stochastic dynamics. Europhys. Lett. **75**, 227–233 (2006)
42. Harris, R.J., Schütz, G.M.: Fluctuation theorems for stochastic dynamics. J. Stat. Mech. P07020 (2007)
43. Harris, R.J., Popkov, V., Schütz, G.M.: Dynamics of instantaneous condensation in the ZRP conditioned on an atypical current. Entropy **15**, 5065–5083 (2013)
44. Harris, R.J., Schütz, G.M.: Fluctuation theorems for stochastic interacting particle systems. Markov Processes Relat. Fields **20**, 3–44 (2014)
45. Imamura, T., Sasamoto, T.: Current moments of 1D ASEP by duality. J. Stat. Phys. **142**, 919–930 (2011)
46. Imamura, T., Sasamoto, T.: Exact Solution for the Stationary Kardar-Parisi-Zhang equation. Phys. Rev. Lett. **108**, 190603 (2012)
47. Isaev, A.P., Pyatov, P.N., Rittenberg, V.: Diffusion algebras. J. Phys. A: Math. Gen. **34**, 5815–5834 (2001)
48. Jansen, S., Kurt, N.: On the notion(s) of duality for Markov processes. Probab. Surv. **11**, 59–120 (2014)
49. Jarzynski, C.: A non-quilibrium equality for free energy differences. Phys. Rev. Lett. **78**, 2690–2693 (1997)
50. Jarzynski, C.: Hamiltonian derivation of a detailed fluctuation theorem. J. Stat. Phys. **98**, 77–102 (2000)
51. Jarzynski, C.: Comparison of far-from-equilibrium work relations. C. R. Phys. **8**, 495–506 (2007)
52. Jimbo, M.: A q-difference analogue of $\mathrm{U}(\mathfrak{g})$ and the Yang-Baxter equation. Lett. Math. Phys. **10**, 63–69 (1985)
53. Jimbo, M.: A q-analogue of $\mathrm{U}(\mathfrak{gl}(\mathrm{N}+1))$, Hecke algebra, and the Yang-Baxter equation. Lett. Math. Phys. **11**, 247–252 (1986)
54. Johansson, K.: Shape fluctuations and random matrices. Commun. Math. Phys. **209**, 437–476 (2000)
55. Kadanoff, L.P., Swift, J.: Transport coefficients near the critical point: a master-equation approach. Phys. Rev. **165**, 310–322 (1968)
56. Kardar, M., Parisi, G., Zhang, Y.-C.: Dynamic scaling of growing interfaces. Phys. Rev. Lett. **56**, 889–892 (1986)
57. Karevski, D., Schütz, G.M.: Conformal invariance in driven diffusive systems at high currents. Phys. Rev. Lett. **118**, 030601 (2017)
58. Kim, D.: Bethe ansatz solution for crossover scaling functions of the asymmetric XXZ chain and the KPZ-type growth model. Phys. Rev. E **52**, 3512–3524 (1995)
59. Kipnis, C., Landim, C., Olla, S.: Hydrodynamic limit for a nongradient system: the generalized symmetric exclusion process. Commun. Pure Appl. Math. **47**, 1475–1545 (1994)
60. Kipnis, C., Landim, C.: Scaling Limits of Interacting Particle Systems. Springer, Berlin (1999)

61. Kolomeisky, A.B., Schütz, G.M., Kolomeisky, E.B., Straley, J.P.: Phase diagram of one-dimensional driven lattice gases with open boundaries. J. Phys. A: Math. Gen. **31**, 6911–6919 (1998)
62. Krug, J.: Boundary-induced phase transitions in driven diffusive systems. Phys. Rev. Lett. **67**, 1882–1885 (1991)
63. Lax, P.D.: Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves. CBMS 2011. SIAM, Philadelphia (1973)
64. Le Doussal, P.: Crossover from droplet to flat initial conditions in the KPZ equation from the replica Bethe ansatz. J. Stat. Mech. **2014**, P04018 (2014)
65. Lebowitz, J.L., Spohn, H.: A Gallavotti-Cohen-type symmetry in the large deviation functional for stochastic dynamics. J. Stat. Phys. **95**, 333–365 (1999)
66. Liggett, T.M.: Ergodic theorems for the asymmetric simple exclusion process. Trans. Am. Math. Soc. **213**, 237–261 (1975)
67. Liggett, T.M.: Interacting Particle Systems. Springer, Berlin (1985)
68. Liggett, T.M.: Stochastic Interacting Systems: Contact, Voter and Exclusion Processes. Springer, Berlin (1999)
69. Lloyd, P., Sudbury, A., Donnelly, P.: Quantum operators in classical probability theory: I. "Quantum spin" techniques and the exclusion model of diffusion. Stoch. Processes Appl. **61**(2), 205–221 (1996)
70. MacDonald, J.T., Gibbs, J.H., Pipkin, A.C.: Kinetics of biopolymerization on nucleic acid templates. Biopolymers **6**, 1–25 (1968)
71. Maes, C.: On the origin and the use of fluctuation relations for the entropy. Sém. Poincaré **2**, 29–62 (2003)
72. Minc, H.: Nonnegative Matrices. Wiley, New York (1988)
73. Nagel, K., Schreckenberg, M.: A cellular automaton model for freeway traffic. J. Phys. I France **2**, 2221–2229 (1992)
74. Pasquier, V., Saleur, H.: Common structures between finite systems and conformal field theories through quantum groups. Nucl. Phys. B **330**, 523–556 (1990)
75. Popkov, V., Schütz, G.M.: Steady-state selection in driven diffusive systems with open boundaries. Europhys. Lett. **48**, 257–264 (1999)
76. Popkov, V., Schütz, G.M.: Shocks and excitation dynamics in a driven diffusive two-channel system. J. Stat. Phys. **112**, 523–540 (2003)
77. Popkov, V., Salerno, M.: Hydrodynamic limit of multichain driven diffusive models. Phys. Rev. E **69**, 046103 (2004)
78. Popkov, V., Schmidt, J., Schütz, G.M.: Universality classes in two-component driven diffusive systems. J. Stat. Phys. **160**, 835–860 (2015)
79. Popkov, V., Schadschneider, A., Schmidt, J., Schütz, G.M.: Fibonacci family of dynamical universality classes. Proc. Natl. Acad. Sci. U.S.A. **112**(41), 12645–12650 (2015)
80. Popkov, V., Schadschneider, A., Schmidt, J., Schütz, G.M.: Exact scaling solution of the mode coupling equations for non-linear fluctuating hydrodynamics in one dimension. J. Stat. Mech. 093211 (2016)
81. Prähofer, M., Spohn, H.: In and out of equilibrium. In: Sidoravicius, V. (ed.) Progress in Probability, vol. 51. Birkhauser, Boston (2002)
82. Prähofer, M., Spohn, H.: Exact scaling functions for one-dimensional stationary KPZ growth. J. Stat. Phys. **115**, 255–279 (2004)
83. Rákos, A., Schütz, G.M.: Exact shock measures and steady state selection in a driven diffusive system with two conserved densities. J. Stat. Phys. **117**, 55–76 (2004)
84. Rákos, A., Harris, R.J.: On the range of validity of the fluctuation theorem for stochastic Markovian dynamics. J. Stat. Mech. P05005 (2008)

85. Rezakhanlou, F.: Hydrodynamic limit for attractive particle systems on $\mathbb{Z}^d$. Commun. Math. Phys. **140**, 417–448 (1991)
86. Sandow, S., Schütz, G.: On $U_q[SU(2)]$-symmetric driven diffusion. Europhys. Lett. **26**, 7–13 (1994)
87. Sasamoto, T., Spohn, H.: One-dimensional Kardar-Parisi-Zhang equation: an exact solution and its universality. Phys. Rev. Lett. **834**, 523 (2010)
88. Schadschneider, A., Chowdhury, D., Nishinari, K.: Stochastic Transport in Complex Systems. Elsevier, Amsterdam (2010)
89. Schütz, G., Domany, E.: Phase transitions in an exactly soluble one-dimensional asymmetric exclusion model. J. Stat. Phys. **72**, 277–296 (1993)
90. Schütz, G., Sandow, S.: Non-abelian symmetries of stochastic processes: derivation of correlation functions for random vertex models and disordered interacting many-particle systems. Phys. Rev. E **49**, 2726–2744 (1994)
91. Schütz, G.M.: The Heisenberg chain as a dynamical model for protein synthesis - some theoretical and experimental results. Int. J. Mod. Phys. B **11**, 197–202 (1997)
92. Schütz, G.M.: Duality relations for asymmetric exclusion processes. J. Stat. Phys. **86**, 1265–1288 (1997)
93. Schütz, G.M.: Solvable models for many-body systems far from equilibrium. In: Domb, C., Lebowitz, J. (eds.) Phase Transitions and Critical Phenomena, vol. 19, pp. 1–251. Academic, London (2001)
94. Schütz, G.M.: Critical phenomena and universal dynamics in one-dimensional driven diffusive systems with two species of particles. J. Phys. A: Math. Gen. **36**, R339–R379 (2003)
95. Schütz, G.M., Wehefritz-Kaufmann, B.: Kardar-Parisi-Zhang modes in d-dimensional directed polymers. Phys. Rev. E **96**, 032119 (2017)
96. Schütz, G.M.: On the Fibonacci universality classes in nonlinear fluctuating hydrodynamics. In: Gonçalves, P., Soares, A. (eds.) From Particle Systems to Partial Differential Equations. PSPDE V, Braga, Portugal, November 2016. Springer Proceedings in Mathematics & Statistics, vol. 258, pp. 149–167. Springer, Cham (2018)
97. Searles, D.J., Evans, D.J.: Fluctuation theorem for stochastic systems. Phys. Rev. E **60**(1), 159–164 (1999)
98. Seifert, U.: Entropy production along a stochastic trajectory and an integral fluctuation theorem. Phys. Rev. Lett. **95**, 040602 (2005)
99. Seifert, U.: Stochastic thermodynamics, fluctuation theorems, and molecular machines. Rep. Prog. Phys. **75**, 126001 (2012)
100. Spohn, H.: Large Scale Dynamics of Interacting Particles. Springer, Berlin (1991)
101. Spohn, H.: Bosonization, vicinal surfaces, and hydrodynamic fluctuation theory. Phys. Rev. E **60**, 6411–6420 (1999)
102. Spohn, H.: Nonlinear fluctuating hydrodynamics for anharmonic chains. J. Stat. Phys. **154**, 1191–1227 (2014)
103. Spohn, H., Stoltz, G.: Nonlinear fluctuating hydrodynamics in one dimension: the case of two conserved fields. J. Stat. Phys. **160**, 861–884 (2015)
104. Spitzer, F.: Interaction of Markov processes. Adv. Math. **5**, 246–290 (1970)
105. Sudbury, A., Lloyd, P.: Quantum operators in classical probability theory. II: the concept of duality in interacting particle systems. Ann. Probab. **23**(4), 1816–1830 (1995)
106. Tóth, B., Valkó, B.: Onsager relations and Eulerian hydrodynamic limit for systems with several conservation laws. J. Stat. Phys. **112**, 497–521 (2003)

107. Tracy, C.A., Widom, H.: Asymptotics in ASEP with step initial condition. Commun. Math. Phys. **290**, 129–154 (2009)
108. van Beijeren, H.: Exact results for anomalous transport in one-dimensional Hamiltonian systems. Phys. Rev. Lett. **108**, 108601 (2012)

Mini-courses at IHP

Hydrodynamics for Symmetric Exclusion in Contact with Reservoirs

Patrícia Gonçalves[1,2(✉)]

[1] Center for Mathematical Analysis, Geometry and Dynamical Systems, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[2] Institut Henri Poincaré, UMS 839 (CNRS/UPMC), 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05, France
pgoncalves@tecnico.ulisboa.pt
http://patriciamath.wixsite.com/patricia

Abstract. We consider the symmetric exclusion process with jumps given by a symmetric, translation invariant, transition probability $p(\cdot)$. The process is put in contact with stochastic reservoirs whose strength is tuned by a parameter $\theta \in \mathbb{R}$. Depending on the value of the parameter θ and the range of the transition probability $p(\cdot)$ we obtain the hydrodynamical behavior of the system. The type of hydrodynamic equation depends on whether the underlying probability $p(\cdot)$ has finite or infinite variance and the type of boundary condition depends on the strength of the stochastic reservoirs, that is, it depends on the value of θ. More precisely, when $p(\cdot)$ has finite variance we obtain either a reaction or reaction-diffusion equation with Dirichlet boundary conditions or the heat equation with different types of boundary conditions (of Dirichlet, Robin or Neumann type). When $p(\cdot)$ has infinite variance we obtain a fractional reaction-diffusion equation given by the regional fractional laplacian with Dirichlet boundary conditions but for a particular strength of the reservoirs.

Keywords: Symmetric exclusion · Stochastic reservoirs · Heat equation · Regional fractional laplacian · Reaction-diffusion · Boundary conditions

1 Introduction

These notes have been written based on material of the articles [1–3] which was presented on a mini-course that the author gave while visiting Institut Henri Poincaré in Paris in May 2017 for the trimester "Stochastic dynamics out of equilibrium" that held from the 3rd of April to the 7th of July. The slides and the videos of the mini-course can be seen in https://indico.math.cnrs.fr/event/844/page/5.

The content of the notes is to explain how to derive partial differential equations with different types of boundary conditions from varied underlying microscopic stochastic dynamics [15,20]. In the next coming sections we consider a

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 137–205, 2019.
https://doi.org/10.1007/978-3-030-15096-9_4

macroscopic space, namely, the interval $[0,1]$ and we discretize it according to a scaling parameter N giving rise to N intervals of size $\frac{1}{N}$. To each $q \in [0,1]$ belonging to the interval $[\frac{i}{N}, \frac{i+1}{N})$ we associate to it the point $\frac{i}{N}$ and in the discrete set of points $\{1, ..., i, ..., N-1\}$ we will define a microscopic dynamics of exclusion type which is Markovian. The discrete set of points $\{1, ..., i, ..., N-1\}$ will be called the bulk and to it we will add two extra points $x = 0$ and $x = N$ which will act as reservoirs. The exclusion dynamics [19] ensures that there is at most one particle per site in the bulk and the Markovian dynamics comes from the fact that each particle waits for rings of random clocks exponentially distributed and independent, after which the particle jumps from a site x in the bulk to another site y in the bulk according to a probability transition rate $p : \mathbb{Z} \times \mathbb{Z} \to [0,1]$, or the particle leaves the system through one of the reservoirs. The reservoirs will be regulated by a parameter which has the role of slowing or fasting the boundary dynamics. More precisely, particles can be injected in the bulk from the site $x = 0$ (resp. $x = N$) to the site y at rate $\alpha\kappa N^{-\theta} p(y)$ (resp. $\beta\kappa N^{-\theta} p(N-y)$) and can be removed from the bulk at the site y to the site $x = 0$ (resp. $x = N$) at rate $(1-\alpha)\kappa N^{-\theta} p(y)$ (resp. $(1-\beta)\kappa N^{-\theta} p(N-y)$). Above, $\alpha, \beta \in [0,1]$, $\theta \in \mathbb{R}$ and $\kappa > 0$.

The goal in these notes is to derive the partial differential equations which describe the space-time evolution of the density of particles in the system. The type of these equations will depend on the finiteness of the variance of the underlying transition probability $p(\cdot)$ and the type of boundary conditions will depend on the strength of the boundary dynamics, namely, the range of the parameter θ. We note that in [10–13] similar models have been considered evolving on the full line, that is, without the presence of stochastic reservoirs.

The goal is to analyse which type of equation and which type of boundary conditions we can get and what is their dependence on the strength of the reservoirs. For that purpose, we split these notes into two main sections to distinguish the case in which jumps are nearest-neighbor or not. Therefore in Sect. 2, we consider the dynamics described above but with $p : \mathbb{Z} \times \mathbb{Z} \to [0,1]$ which satisfies $p(x,y) = p(y-x) = 0$ if $|x-y| > 1$, $p(0) = 0$ so that $p(1) = p(-1) = \frac{1}{2}$. This means that in the bulk particles can jump to one of their nearest-neighbors and particles can be injected/removed in the bulk/from the bulk through the sites $x = 1$ or $x = N-1$. For these models we will derive the heat equation with three different types of boundary conditions: non-homogeneous Dirichlet boundary conditions when the reservoirs are fast (which corresponds to $\theta < 1$) and Neumann boundary conditions when the reservoirs are slow (which corresponds to $\theta > 1$). Linking the aforementioned two types of boundary conditions, for a particular strength of the boundary dynamics (which corresponds to $\theta = 1$), we will derive the heat equation with a type of linear Robin boundary conditions.

In Sect. 3, we will consider the dynamics described above, but allowing long jumps given by a transition probability $p : \mathbb{Z} \times \mathbb{Z} \to [0,1]$ such that $p(x,y) = p(y-x)$, which is symmetric, namely $p(y-x) = p(x-y)$, and we will distinguish

two cases: the first one where $p(\cdot)$ has finite variance and then the case where $p(\cdot)$ has infinite variance. In the first case, we will obtain an extension of the results of the model with only nearest-neighbor jumps, that is we will derive the heat equation with the three types of boundary conditions mentioned above but for a certain choice of the transition probability two new regimes appear when the reservoirs are fast, namely, a reaction-diffusion equation and a reaction equation, both endowed with non-homogeneous Dirichlet boundary conditions. In the case where $p(\cdot)$ has infinite variance and for a particular strength of the reservoirs (which corresponds to $\theta = 0$), we will derive a collection of fractional reaction-diffusion equations with non-homogeneous Dirichlet boundary conditions. For the interested reader we note that when $p(\cdot)$ has infinite variance and when the strength of the reservoirs is slow (which corresponds to $\theta > 0$), we cannot say anything about the equation nor its boundary conditions. In [2] a similar model has been studied and some conjectures have been presented in the case where the reservoirs are slow. We believe that the same conjecture should be true for this model, but we leave this for a future problem to look at. We also note that it would be very interesting to consider other types of boundary dynamics or even more general type of bulk dynamics than the exclusion in order to obtain other partial differential equations with various boundary conditions.

These notes are organized as follows: in Sect. 2 we derive the hydrodynamic limit for the symmetric exclusion in contact with stochastic reservoirs but only allowing jumps to nearest-neighbors and in Sect. 3 we derive the hydrodynamics in the case where the system exhibits long jumps.

More precisely, in Sects. 2.1, 2.2 and 2.3 we present the dynamics of the model; in Sect. 2.4 we present its stationary measures; in Sect. 2.5 we analyse the empirical profile and the two point correlation function; in Sect. 2.6 we present the hydrodynamic equations and the notion of their weak solutions; in Sect. 2.7 we state the hydrodynamic limit; in Sect. 2.8 we give an heuristic argument to deduce the weak formulation of the solutions by means of auxiliary martingales associated to the process; in Sect. 2.9 we prove tightness of the process of empirical measures; in Sect. 2.10 we characterize the limit point of the tight sequence and in Sect. 2.11 we prove the hydrostatic limit, which is the hydrodynamic limit starting from the invariant measure of the system.

In Sect. 3 we analyse the hydrodynamics for the symmetric exclusion with long jumps given by a transition probability which is symmetric. In Sect. 3.1 we describe the model; in Sect. 3.2 we present the case in which the underlying transition probability has finite variance and in Sect. 3.3 we analyse the case in which the transition probability has infinite variance.

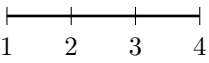
In the Appendix we present some of the technical results that are needed along the proofs regarding the derivation of the weak solution of the corresponding hydrodynamic equations.

2 Symmetric Simple Exclusion in Contact with Reservoirs

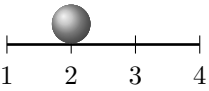
2.1 The Model

In this section we describe the collection of models that we are going to consider in these notes. First we start by fixing the notation which fits all the models and then we particularize our choice of the parameters in such a way that we treat each model, with its special features, separately. For that purpose, let N be a scaling parameter, which will be taken to infinity later on and denote by $\Lambda_N = \{1, ..., N-1\}$ the discrete set of points to which we call the bulk.

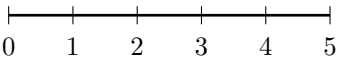
The exclusion process in contact with stochastic reservoirs is a Markov process, denoted by $\{\eta_t : t \geq 0\}$, which has state space $\Omega_N := \{0,1\}^{\Lambda_N}$. The configurations of the state space Ω_N are denoted by η, so that for $x \in \Lambda_N$, $\eta(x) = 0$ means that the site x is vacant while $\eta(x) = 1$ means that the site x is occupied. For an illustration of the dynamics let us first take $N = 5$ so that the bulk is the discrete set of points $\{1, 2, 3, 4\}$:



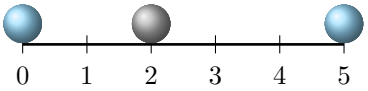
Now, to describe a possible initial configuration we can do the following. Toss a coin, if we get head we put a particle at the site 1 and if we get a tail we leave it empty. Repeat this for each site of the discrete set Λ_5 and suppose that we get at the end to the configuration $\eta_0 = (0, 1, 0, 0)$ which can be represented as:



Now, we start to particularize our choice for the dynamics. We are going to add one reservoir at each end point of the bulk. This means that in our construction, we add the points $x = 0$ and $x = N$ to the bulk. Going back to the picture above, this means that we have now the set $\{0, 1, 2, 3, 4, 5\}$ where particles can be placed, but the sites $x = 0$ and $x = 5$ will act as reservoirs.



Note that the bulk stays unchanged, the role of the boundary points $\{0, N\}$ is to allow particles to get in and out of the bulk. So, for example, in the initial configuration given above, now we have, in the figure below, the sites $x = 0$ and $x = N$ occupied, representing the fact that in $x = 0$ and $x = N$ there are particles that can enter to the bulk and that can be removed from the bulk.

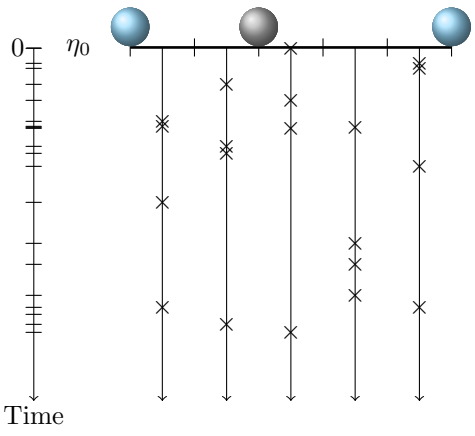


Now we describe the time between jumps. For that purpose, for each pair of sites (x, y) we associate a Poisson process of intensity $p(x, y) = p(y - x)$. The Poisson processes associated to different bonds are independent. Note that the bonds in the bulk are not oriented. In the first dynamics that we are describing, we consider $p(y-x) = 0$ if $|x-y| > 1$, $p(1) = p(-1) = \frac{1}{2}$ so that jumps can only occur to a nearest-neighbor position and for that reason the exclusion process coins the name *simple exclusion process*. At the boundary points we associate two Poisson processes to each bond containing a boundary point. More precisely, to the bond $\{0, 1\}$ (resp. $\{1, 0\}$) we associate a Poisson process of intensity $\alpha\kappa N^{-\theta}$ (resp. $(1-\alpha)\kappa N^{-\theta}$) and to the bond $\{N-1, N\}$ (resp. $\{N, N-1\}$) we associate a Poisson process of intensity $(1 - \beta)\kappa N^{-\theta}$ (resp. $\beta\kappa N^{-\theta}$). Above we fix the parameters $\alpha, \beta \in [0, 1]$, $\theta \in \mathbb{R}$ and $\kappa > 0$. The role of the parameter θ is to regulate the slowness/fastness of the reservoirs. If $\theta > 0$ and θ increases then the reservoirs are slower and if $\theta < 0$ and θ decreases then the reservoirs are faster.

We remark that another interpretation of the previous dynamics at the boundary could be given as follows. Particles can either be created or annihilated at the sites $x = 1$ and $x = N - 1$ according to the following rates:

- at site $x = 1$:
  - creation rate $\alpha\kappa N^{-\theta}$,
  - annihilation rate $(1 - \alpha)\kappa N^{-\theta}$,
- at site $x = N - 1$:
  - creation rate $\beta\kappa N^{-\theta}$,
  - annihilation rate $(1 - \beta)\kappa N^{-\theta}$.

Note that in any case, the exclusion rule has to be respected. At most one particle is allowed at each site of the bulk (recall that the state space is $\{0, 1\}^{\Lambda_N}$) so that particles can only be created (resp. removed) at the sites $x = 1$ or $x = N - 1$ if the corresponding site is empty (resp. occupied), otherwise nothing happens. Before we proceed let us see an illustration of a possible realization of the Poisson processes as given in the figure below.



In the figure at the left hand side we represent by "×" each mark of a possible realization of the Poisson processes associated to the bonds. At the left hand side we put an arrow going down which is representing the evolution of time and each sign "–" means that a clock has rung according to some Poisson clock, so that at the corresponding time, a jump from a particle might have occurred.

We note that in this figure we did not distinguish the marks of the Poisson processes associated to the oriented bonds at the boundary because we believe that it is simpler to analyse the dynamics at the boundary by allowing particles to

get in or get out according to the Poisson marks but also taking into account the exclusion rule.

In order to give an example, let us see now all the configurations that we obtain starting the dynamics from the configuration $\eta_0 = (0, 1, 0, 0)$ represented above and the realization of the Poisson processes given in the previous figure.

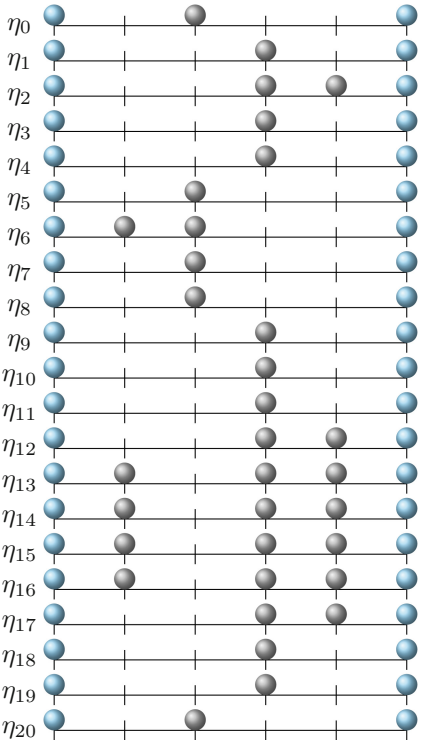


Fig. 1. Possible configurations starting from $(0, 1, 0, 0)$

By abuse of notation, in the figure at the left hand side, we numbered the configurations that we obtained by the number of the marks of the Poisson processes (which in the example are equal to 20) just to make the presentation simple. We note that the configurations are indexed by time t which is continuous and not discrete. Note that the difference between $\eta_0 = (0, 1, 0, 0)$ and $\eta_1 = (0, 0, 1, 0)$ is only at two sites (this is always the case when we compare two configurations which differ by a jump of a particle in the bulk, *a jump in the bulk affects the occupation variables at two sites*) and η_1 is obtained from η_0 by shifting the particle at the site 2 in η_0 to the site 3. This is a consequence of the fact that the first mark of the Poisson process that occurs is associated to the bond $\{2, 3\}$ and that in η_0 there is a particle at the site 2. The next mark we see is associated to the bond $\{4, 5\}$ and since in $\eta_1 = (0, 0, 1, 0)$ there is no particle at the site $x = 4$, a particle is injected in the bulk at the site 4, giving rise to $\eta_2 = (0, 0, 1, 1)$ and so on. Note that the boundary dynamics only changes the configuration at one site (Fig. 1).

We also note that the ring of a clock does not imply that the configuration of the system has changed. In the example above $\eta_3 = \eta_4 = (0, 0, 1, 0)$ since the corresponding Poisson mark is associated to the bond $\{1, 2\}$ and since both sites $x = 1$ and $x = 2$ are empty, nothing happens and particles wait a new ring of a clock.

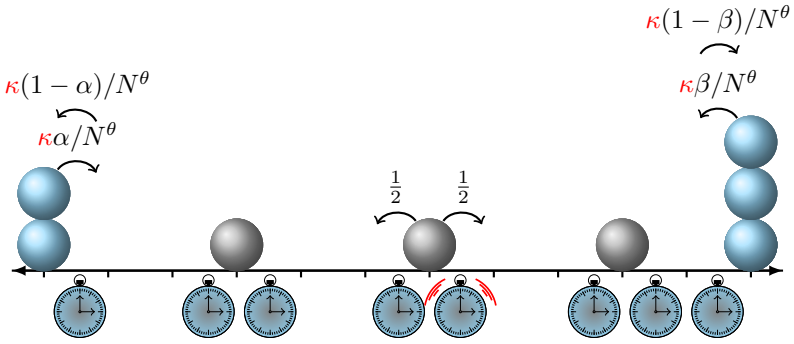
The first dynamics that we are going to consider in these notes, and which is described in this section is completely characterized by now, but we note that in Sect. 3 we are going to generalize the previous dynamics by allowing particles to give long jumps according to some probability transition rate $p : \mathbb{Z} \times \mathbb{Z} \to [0, 1]$ such that $p(x, y) = p(y - x)$ and which is symmetric, that is $p(y - x) = p(x - y)$.

In the latter dynamics, there is only one reservoir at each end point of the bulk but particles can be injected from them to any site of the bulk or they can be removed from any site of the bulk to one of the reservoirs. We will distinguish two cases: when $p(\cdot)$ has finite variance and when $p(\cdot)$ has infinite variance.

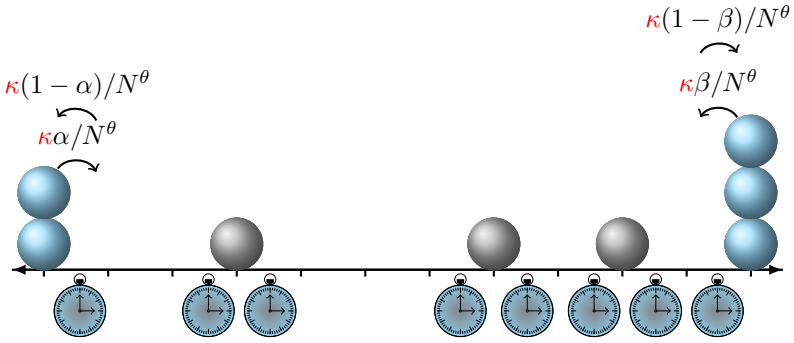
2.2 Illustration of the Dynamics

In this section we draw some pictures to illustrate more easily the dynamics that we defined in the previous subsection. The particles at the bulk are coloured in gray and the particles at the two reservoirs are coloured in blue. We also added the clocks only at the bonds where there are particles but we note that the clocks are present in all bonds of the form $\{x, x+1\}$. Whenever there is a ring of a clock we see some red lines on top of the corresponding clock and the jump rates are indicated above the corresponding jumps which are represented by arrows.

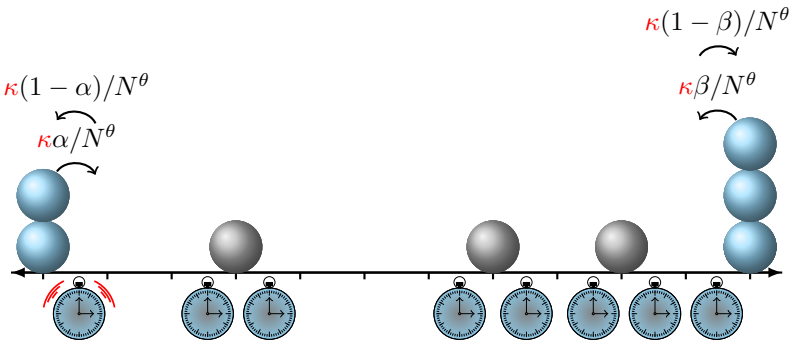
In the first picture below, we take $N = 11$ and the initial configuration is $\eta_0 = (0, 0, 1, 0, 0, 1, 0, 0, 1, 0)$. Note that this initial configuration changes only if one of the clocks associated to bonds containing the sites $x = 3, 6, 9$ rings (which makes the corresponding particle to displace one position to the left or right of it) or if the clocks at the boundary sites $x = 0$ (resp. $x = 11$) ring (which makes a particle get into the system at the site $x = 1$ (resp. $x = 10$).



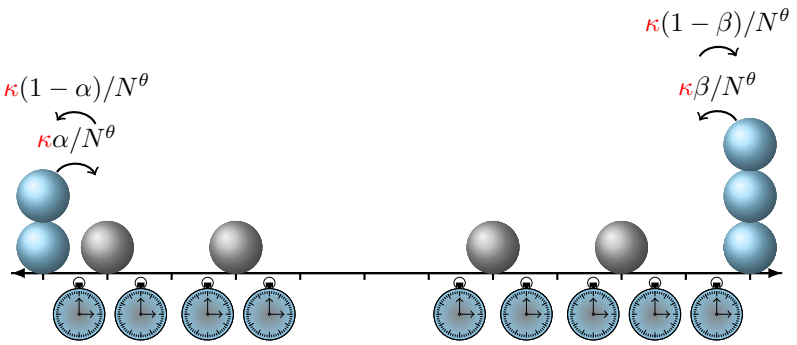
Suppose that the first clock to ring is associated to the bond $\{6, 7\}$. Since there is a particle at the site $x = 6$ it jumps to the site $x = 7$ with rate $1/2$. See the figure below.



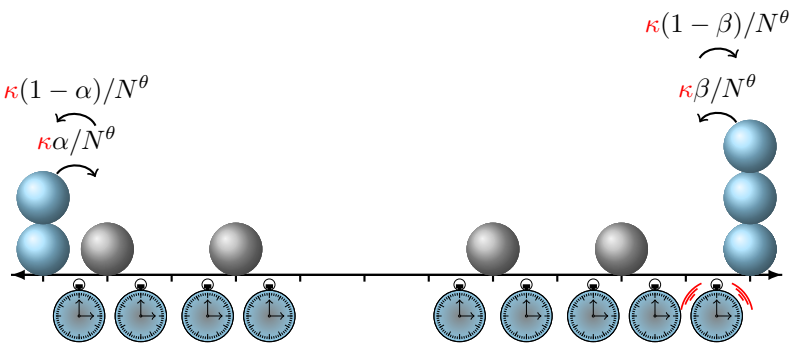
Now let us suppose that the next clock to ring is associated to the oriented bond $\{0,1\}$.



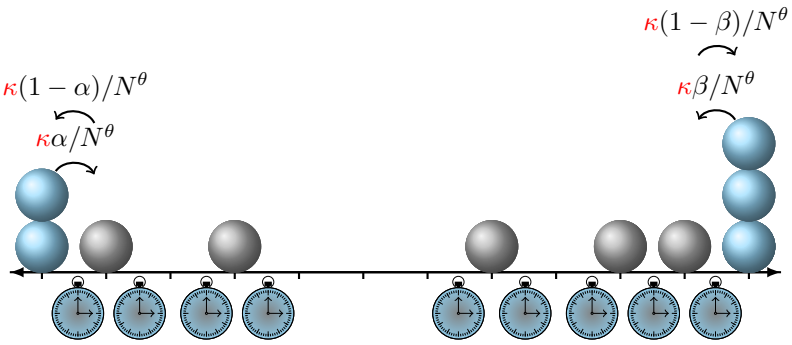
Since there is no particle at the site $x = 1$, a particle is injected into the system at the site $x = 1$ with rate $\alpha\kappa N^{-\theta}$. See the figure below.



Finally let us suppose that the next clock to ring is associated to the oriented bond $\{N, N-1\}$.



Since there is no particle at the site $x = N-1$, a particle is injected into the system at the site $x = N-1$ with rate $\beta\kappa N^{-\theta}$. See the figure below.



We note that the bulk dynamics conserves the total number of particles in the bulk, but the boundary dynamics destroys this quantity since it injects/removes particles in/from the bulk.

2.3 Infinitesimal Generator

The dynamics described above is Markovian and can be completely characterized by mean of its infinitesimal generator, see [17,18]. The Markov process $\{\eta_t : t \geq 0\}$ whose dynamics we have just defined has infinitesimal generator denoted by $\mathcal{L}_N$ which is expressed as

$$\mathcal{L}_N = \mathcal{L}_{N,0} + \mathcal{L}_{N,b}, \tag{1}$$

where $\mathcal{L}_{N,0}$ and $\mathcal{L}_{N,b}$ are given on functions $f : \Omega_N \to \mathbb{R}$ by

$$(\mathcal{L}_{N,0} f)(\eta) \;=\; \sum_{x=1}^{N-2} \frac{1}{2}\Big(f(\eta^{x,x+1}) - f(\eta)\Big),$$

$$\mathcal{L}_{N,b} = \mathcal{L}^1_{N,b} + \mathcal{L}^{N-1}_{N,b}, \tag{2}$$

where for $x \in \{1, N-1\}$

$$(\mathcal{L}^x_{N,b} f)(\eta) \;=\; \frac{\kappa}{N^\theta} c_x(\eta, r(x))\Big(f(\eta^x) - f(\eta)\Big),$$

$r(1) = \alpha$ and $r(N-1) = \beta$,

$$(\eta^{x,y})(z) = \begin{cases} \eta(z), \; z \neq x, y, \\ \eta(y), \; z = x, \\ \eta(x), \; z = y \end{cases}, \quad (\eta^x)(z) = \begin{cases} \eta(z), \; z \neq x, \\ 1 - \eta(x), \; z = x, \end{cases} \tag{3}$$

and for $x \in \{1, N-1\}$

$$c_x(\eta; r(x)) := \frac{1}{2}\left[\eta(x)\left(1 - r(x)\right) + (1 - \eta(x))r(x)\right]. \tag{4}$$

Note that the generator above splits into the sum of the generator $\mathcal{L}_{N,0}$ (which is related to the jumps in the bulk) and $\mathcal{L}_{N,b}$ (which is related to the

jumps from the boundary or from the reservoirs). We will refer to the first one as the *exchange dynamics* and the latter one as the *flip dynamics*, because in $\mathcal{L}_{N,0}$ we exchange the occupation variables $\eta(x)$ and $\eta(x+1)$ and in $\mathcal{L}^x_{N,b}$ we flip the value of the occupation variable at $\eta(x)$.

We consider the Markov process speeded up in the time scale $\Theta(N)$ and we note that the process $\{\eta_{t\Theta(N)} : t \geq 0\}$ has infinitesimal generator given by $\Theta(N)\mathcal{L}_N$. To see this relation, let $\tilde{\mathcal{L}}_N$ be the generator of the process $\{\eta_{t\Theta(N)} : t \geq 0\}$. By definition, for $f : \Omega_N \to \mathbb{R}$, we have that

$$\tilde{\mathcal{L}}_N f = \lim_{s\to 0} \frac{\tilde{S}_s f - f}{s}, \tag{5}$$

where $\tilde{S}_s := S_{s\Theta(N)}$ is the semigroup associated to $\tilde{\mathcal{L}}_N$ and S_s is the semigroup associated to $\mathcal{L}_N$. Then,

$$\Theta(N)\mathcal{L}_N f = \lim_{t\to 0} \Theta(N)\frac{S_t f - f}{t} = \lim_{s\to 0} \Theta(N)\frac{S_{s\Theta(N)} f - f}{s\Theta(N)} = \tilde{\mathcal{L}}_N f, \tag{6}$$

from where we conclude that $\tilde{\mathcal{L}}_N := \Theta(N)\mathcal{L}_N$.

We note that $\eta_{t\theta(N)}$ depends on α, β, θ and κ but we will omit these indexes in order to simplify notation. Fix $T > 0$ and $\theta \in \mathbb{R}$. Let μ_N be a probability measure in Ω_N. We denote by $\mathbb{P}_{\mu_N}$ the probability measure in the Skorohod space $\mathcal{D}([0,T], \Omega_N)$ induced by the Markov process $\{\eta_{t\Theta(N)} : t \geq 0\}$ and the initial probability measure μ_N and we denote by $E_{\mathbb{P}_{\mu_N}}$ the expectation with respect to $\mathbb{P}_{\mu_N}$.

Our goal in these notes is to analyse the impact of changing the strength of the reservoirs (by changing the value of θ) on the macroscopic behavior of the system. More precisely, we want to obtain the hydrodynamic equations of the process which will have different boundary conditions depending on the range of the parameter θ which rules the strength of the reservoirs. Before proceeding, in the next subsection we analyse the invariant measures for this model.

2.4 Stationary Measures

For $\rho \in (0,1)$ we denote by ν^N_ρ the Bernoulli product measure in Ω_N with density ρ, that is, for $x \in \Lambda_N$:

$$\nu^N_\rho\{\eta : \eta(x) = 1\} = \rho. \tag{7}$$

According to this measure the occupation variables $\{\eta(x)\}_{x\in\Lambda_N}$ are independent and for each $x \in \Lambda_N$ the random variable $\eta(x)$ has Bernoulli distribution of parameter ρ. When we restrict the parameters α and β such that $\alpha = \beta = \rho$, then these measures are invariant for the dynamics described above. In fact, a stronger result is true, see the next lemma where we prove that these measures are reversible.

Lemma 1. *For $\alpha = \beta = \rho$ the Bernoulli product measures ν^N_ρ are reversible.*

Proof. Fix two functions $f, g : \Omega_N \to \mathbb{R}$. To prove the lemma, we need to show that

$$\int_{\Omega_N} g(\eta)\mathcal{L}_N f(\eta) d\nu_\rho^N = \int_{\Omega_N} f(\eta)\mathcal{L}_N g(\eta) d\nu_\rho^N. \tag{8}$$

Let us start with the exchange dynamics given by $\mathcal{L}_{N,0}$. In this case we need to check that

$$\sum_{x\in\Lambda_N} \int_{\Omega_N} g(\eta)(f(\eta^{x,x+1}) - f(\eta)) d\nu_\rho^N = \sum_{x\in\Lambda_N} \int_{\Omega_N} f(\eta)(g(\eta^{x,x+1}) - g(\eta)) d\nu_\rho^N.$$

For that purpose note that, for fixed $x \in \Lambda_N$ and performing a change of variables $\xi = \eta^{x,x+1}$, we have that

$$\begin{aligned}\int_{\Omega_N} g(\eta) f(\eta^{x,x+1}) d\nu_\rho^N &= \sum_{\eta\in\Omega_N} g(\eta) f(\eta^{x,x+1}) \nu_\rho^N(\eta) \\ &= \sum_{\xi\in\Omega_N} g(\xi^{x,x+1}) f(\xi) \frac{\nu_\rho^N(\xi^{x,x+1})}{\nu_\rho^N(\xi)} \nu_\rho^N(\xi).\end{aligned}$$

Now note that

$$\nu_\rho^N(\xi) = \prod_{x\in\Lambda_N} \rho^{\xi(x)}(1-\rho)^{1-\xi(x)}$$

so that

- if $\xi(x) = 1$ and $\xi(x+1) = 0$, denoting by $\tilde{\xi}$ the configuration ξ removing its values at x and $x+1$ so that $\xi = (\tilde{\xi}, \xi(x), \xi(x+1))$, then $\nu_\rho^N(\xi) = \nu_\rho^N(\tilde{\xi})\rho(1-\rho)$ and $\nu_\rho^N(\xi^{x,x+1}) = \nu_\rho^N(\tilde{\xi})(1-\rho)\rho$, so that

$$\frac{\nu_\rho^N(\xi^{x,x+1})}{\nu_\rho^N(\xi)} = 1. \tag{9}$$

- if $\xi(x) = 0$ and $\xi(x+1) = 1$, then $\nu_\rho^N(\xi) = \nu_\rho^N(\tilde{\xi})(1-\rho)\rho$ and $\nu_\rho^N(\xi^{x,x+1}) = \nu_\rho^N(\tilde{\xi})\rho(1-\rho)$, so that (9) is also true.

Therefore, we obtain that

$$\int_{\Omega_N} g(\eta) f(\eta^{x,x+1}) d\nu_\rho^N = \sum_{\xi\in\Omega_N} g(\xi^{x,x+1}) f(\xi) \nu_\rho^N(\xi) = \int_{\Omega_N} g(\eta^{x,x+1}) f(\eta) d\nu_\rho^N,$$

which proves (8) for $\mathcal{L}_{N,0}$. For the flip dynamics given by $\mathcal{L}_{N,b}$ we note, for the left boundary, that

$$\begin{aligned}&\int_{\Omega_N} g(\eta) c_1(\eta,\alpha) f(\eta^1) d\nu_\rho^N \\ &\quad= \sum_{\eta\in\Omega_N} g(\eta)(1-\eta(1))\alpha f(\eta^1)\nu_\rho^N(\eta) + \sum_{\eta\in\Omega_N} g(\eta)(1-\alpha) f(\eta^1)\nu_\rho^N(\eta).\end{aligned}$$

By the change of variables $\xi = \eta^1$, the previous expression can be written as

$$\sum_{\xi\in\Omega_N} f(\xi)\Big\{g(\xi^1)\xi(1)\alpha\frac{\nu_\rho^N(\xi^1)}{\nu_\rho^N(\xi)} + g(\xi^1)(1-\xi(1))(1-\alpha)\frac{\nu_\rho^N(\xi^1)}{\nu_\rho^N(\xi)}\Big\}\nu_\rho^N(\xi).$$

A simple computation shows that if $\xi(1) = 1$, then $\frac{\nu_\rho^N(\xi^1)}{\nu_\rho^N(\xi)} = \frac{1-\rho}{\rho}$ so that the previous expression can be written as

$$\frac{\kappa}{N^\theta}\sum_{\xi\in\Omega_N} f(\xi)\Big\{g(\xi^1)\xi(1)\alpha\frac{1-\rho}{\rho} + g(\xi^1)(1-\xi(1))(1-\alpha)\frac{\rho}{1-\rho}\Big\}\nu_\rho^N(\xi),$$

from where we get, for $\alpha = \rho$, that

$$\int_{\Omega_N} g(\eta)c_1(\eta,\alpha)f(\eta^1)d\nu_\rho^N = \int_{\Omega_N} g(\eta^1)c_1(\eta^1,\rho)f(\eta)d\nu_\rho^N.$$

The same computation can be done if $\xi(1) = 0$, from where we conclude. We can repeat the same computation for the right boundary and this proves (8) for $\mathcal{L}_{N,b}$. This ends the proof of the lemma. □

When $\alpha \neq \beta$, the Bernoulli product measures are not reversible nor invariant. A simple way to check the non-invariance is to argue as follows. Suppose that the measures ν_ρ^N are invariant. Then we know that for any function $f : \Omega_N \to \mathbb{R}$ we have that

$$\int_{\Omega_N} \mathcal{L}_N f(\eta)d\nu_\rho^N = 0. \tag{10}$$

But for $f(\eta) = \eta(1)$, a simple computation shows that $\mathcal{L}_{N,0}f(\eta) = \frac{1}{2}(\eta(2)-\eta(1))$ and $\mathcal{L}_{N,b}f(\eta) = \frac{\kappa}{2N^\theta}[\alpha - \eta(1)]$, so that

$$\int_{\Omega_N} \mathcal{L}_N f(\eta)d\nu_\rho^N = \frac{\kappa}{2N^\theta}(\alpha-\rho)$$

and this equals to 0 iff $\alpha = \rho$. Analogously, repeating the same computations as above for $f(\eta) = \eta(N-1)$, we would conclude (10) iff $\beta = \rho$. But this contradicts the fact that $\alpha \neq \beta$.

When $\alpha \neq \beta$, since we have a finite state irreducible Markov process, then there exists a unique stationary measure that we denote by μ_{ss}. A way to get information about this measure is to use the matrix ansatz method introduced in [6,7]. The idea behind the method is the following. Let

$$f_{N-1}(\eta(1),\cdots,\eta(N-1))$$

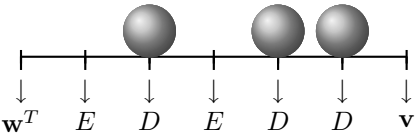
denote the weight of the configuration $\eta := (\eta(1), \cdots, \eta(N-1))$ with respect to the stationary measure μ_{ss} and let us suppose that

$$f_{N-1}(\eta(1), \eta(2), \cdots, \eta(N-1)) = \mathbf{w}^T X_{\eta(1)} X_{\eta(2)} \cdots X_{\eta(N-1)} \mathbf{v},$$

where

$$X_{\eta(x)} = \eta(x) D + (1 - \eta(x)) E,$$

and D, E are matrices (which in general do not commute) and the vectors $\mathbf{w}^T, \mathbf{v}$ are present in order to convert the matrix product into a scalar. In the figure below we take $N = 6$ and we present a possible configuration $\eta = (0, 1, 0, 1, 1)$ whose corresponding weight is given by $f_{N-1}(\eta) = \mathbf{w}^T EDEDD\mathbf{v}$.



Let $P(\eta(1), \eta(2), \cdots, \eta(N-1))$ be the normalized weight of the configuration $\eta := (\eta(1), \cdots, \eta(N-1))$ with respect to the stationary state μ_{ss}, which is given by

$$P(\eta(1), \eta(2), \cdots, \eta(N-1)) = \frac{f_{N-1}(\eta(1), \eta(2), \cdots, \eta(N-1))}{Z_{N-1}},$$

where Z_{N-1} is the sum of the weights of the 2^{N-1} possible configurations in Ω_N:

$$Z_{N-1} = \sum_{\eta(1) \in \{0,1\}} \cdots \sum_{\eta(N-1) \in \{0,1\}} f_{N-1}(\eta(1), \eta(2), \cdots, \eta(N-1)).$$

From the definition of f_{N-1}, we have that

$$P(\eta(1), \eta(2), \cdots, \eta(N-1)) = \frac{\mathbf{w}^T X_{\eta(1)} X_{\eta(2)} \cdots X_{\eta(N-1)} \mathbf{v}}{Z_{N-1}},$$

and the normalization can be written as

$$\begin{aligned} Z_{N-1} &= \sum_{\eta(1) \in \{1,0\}} \cdots \sum_{\eta(N-1) \in \{1,0\}} \mathbf{w}^T X_{\eta(1)} X_{\eta(2)} \cdots X_{\eta(N-1)} \mathbf{v} \\ &= \sum_{\eta(1) \in \{1,0\}} \cdots \sum_{\eta(N-2) \in \{1,0\}} \mathbf{w}^T X_{\eta(1)} X_{\eta(2)} \cdots X_{\eta(N-2)} (D+E) \mathbf{v} \qquad (11) \\ &= \cdots = \mathbf{w}^T (D+E)^{N-1} \mathbf{v}. \end{aligned}$$

Let us now impose conditions on the matrices D and E. For that purpose, let $C = D+E$. The expectation of the occupation variable at the site x, with respect to the stationary state μ_{ss}, is given by

$$\begin{aligned}\rho_{ss}^N(x) &= \int_{\Omega_N} \eta(x) d\mu_{ss} \\ &= \frac{\sum_{\eta(1)\in\{1,0\}} \cdots \sum_{\eta(N-1)\in\{1,0\}} \eta(x) f_{N-1}(\eta(1), \cdots, \eta(N-1))}{Z_{N-1}} \\ &= \frac{1}{Z_{N-1}} \sum_{\eta(1)\in\{1,0\}} \cdots \sum_{\eta(N-1)\in\{1,0\}} \Big[\mathbf{w}^T \prod_{j=1}^{x-1} X_{\eta(j)} D \prod_{j=x+1}^{N-1} X_{\eta(j)} \mathbf{v}\Big] \\ &= \frac{\mathbf{w}^T C^{x-1} D C^{N-1-x} \mathbf{v}}{\mathbf{w}^T C^{N-1} \mathbf{v}}.\end{aligned} \tag{12}$$

The function $\rho_{ss}^N(\cdot)$ is called the stationary empirical density profile since it is the average with respect to the stationary measure μ_{ss}, otherwise we refer to it as the empirical density profile. Note that above the sum does not contain the factor $\eta(x) \in \{1, 0\}$ since the expectation is non-zero iff $\eta(x) = 1$. We can also compute the expectation of the product of two point occupation variables at the sites x and y, with respect to the stationary state μ_{ss}, that is, for $1 \le x < y \le N-1$, we have that

$$\begin{aligned}&\int_{\Omega_N} \eta(x)\eta(y) d\mu_{ss} = \\ &\quad = \frac{\sum_{\eta(1)\in\{0,1\}} \cdots \sum_{\eta(N-1)\in\{0,1\}} \eta(x)\eta(y) f_{N-1}(\eta(1), \cdots, \eta(N-1))}{Z_{N-1}} \\ &\quad = \frac{\mathbf{w}^T C^{x-1} D C^{y-x-1} D C^{N-1-y} \mathbf{v}}{\mathbf{w}^T C^{N-1} \mathbf{v}}.\end{aligned}$$

Therefore, the two point correlation function, with respect to the stationary state μ_{ss}, is given on $1 \le x < y \le N-1$ by

$$\begin{aligned}\varphi_{ss}^N(x,y) &:= \int_{\Omega_N} (\eta(x) - \rho_{ss}^N(x))(\eta(y) - \rho_{ss}^N(y)) d\mu_{ss} \\ &= \frac{\mathbf{w}^T C^{x-1} D C^{y-x-1} D C^{N-1-y} \mathbf{v}}{\mathbf{w}^T C^{N-1} \mathbf{v}} \\ &- \frac{\mathbf{w}^T C^{x-1} D C^{N-1-x} \mathbf{v}}{\mathbf{w}^T C^{N-1} \mathbf{v}} \cdot \frac{\mathbf{w}^T C^{y-1} D C^{N-1-y} \mathbf{v}}{\mathbf{w}^T C^{N-1} \mathbf{v}}.\end{aligned} \tag{13}$$

A simple computation (see [5]) shows that for the dynamics that we are considering in this section, the matrices D, E and the vectors $\mathbf{w}^T, \mathbf{v}$ satisfy the following relations:

$$\begin{aligned}&DE - ED = D + E = C, \\ &\mathbf{w}^T \left[\frac{\kappa\alpha}{2N^\theta} E - \frac{\kappa(1-\alpha)}{2N^\theta} D\right] = \mathbf{w}^T, \\ &\left[\frac{\kappa(1-\beta)}{2N^\theta} D - \frac{\kappa\beta}{2N^\theta} E\right] \mathbf{v} = \mathbf{v}.\end{aligned} \tag{14}$$

We note that the equations above also show that

$$C(D+I) = (D+E)(D+I) = DD + D + ED + E,$$

and that $C(D+I) = DD+DE = DC$. Analogously we have that $CD = (D-I)C$. Using (11), we obtain that Z_{N-1} is given by

$$Z_{N-1} = \frac{1}{(\alpha-\beta)^{N-1}} \frac{\Gamma(2N^\theta + N - 1)}{\Gamma(2N^\theta)},$$

where $\Gamma(\cdot)$ denotes the Gamma function. For the details on these computations we refer the interested reader to [5]. Now, in (12), by writing $DC^{N-1-x} = DCC^{N-2-x}$ and using the fact that $C(D+I) = DC$ we obtain

$$\rho^N_{ss}(x) = \frac{\mathbf{w}^T C^{x-1} C(D+I) C^{N-2-x}\mathbf{v}}{Z_{N-1}} = \frac{\mathbf{w}^T C^x D C^{N-2-x}\mathbf{v}}{Z_{N-1}} + \frac{\mathbf{w}^T C^{N-2}\mathbf{v}}{Z_{N-1}}.$$

Repeating the procedure above and using the explicit expression for Z_{N-1} given above, we obtain a simple expression for $\rho^N_{ss}(x)$ given by

$$\rho^N_{ss}(x) = \beta + (N-x)\frac{\alpha-\beta}{2N^\theta + N - 2} + (N^\theta - 1)\frac{\alpha-\beta}{2N^\theta + N - 2}. \tag{15}$$

In fact last identity can be rewritten as

$$\rho^N_{ss}(x) = \frac{\kappa(\beta-\alpha)x}{2N^\theta + N - 2} + \alpha + \frac{\kappa(\beta-\alpha)x}{2N^\theta + N - 2}\Big(\frac{N^\theta}{\kappa} - 1\Big).$$

Analogously, from a simple, but long computation (see [5]), we have that

$$\int_{\Omega_N} \eta(x)\eta(y)d\mu_{ss} = \beta\rho^N_{ss}(x) + (N - y + N^\theta - 1)\frac{\alpha-\beta}{2N^\theta + N - 2}\rho^{N-1}_{ss}(x),$$

and from (15), we obtain

$$\int_{\Omega_N} \eta(x)\eta(y)d\mu_{ss} = \beta\left[\frac{\beta(x+N^\theta-1) + \alpha(N-x+N^\theta-1)}{2N^\theta + N - 2}\right] + \frac{(N-y+N^\theta-1)(\alpha-\beta)}{2N^\theta + N - 2}\left[\frac{\beta(x+N^\theta-1) + \alpha(N-x+N^\theta-2)}{2N^\theta + N - 3}\right].$$

Putting together last expressions and doing simple, but long, computations we conclude that

$$\varphi^N_{ss}(x,y) = -\frac{(\alpha-\beta)^2(x+N^\theta-1)(N-y+N^\theta-1)}{(2N^\theta + N - 2)^2(2N^\theta + N - 3)}. \tag{16}$$

From the previous identity it follows that

$$\max_{x<y} |\varphi^N_{ss}(x,y)| = \begin{cases} O\left(\frac{N^\theta}{N^2}\right), & \theta < 1, \\ O\left(\frac{1}{N^\theta}\right), & \theta \geq 1, \end{cases} \quad \to_{N\to\infty} 0. \tag{17}$$

This means that as the size of the bulk tens to infinity, the two point correlation function vanishes. In the next subsection we analyse the empirical profile and the two point correlation function for more general initial measures.

2.5 Empirical Profile and Correlations

Before stating the hydrodynamic limit result we explain here how to have a guess on the form of the hydrodynamic equations by using the *empirical profile*, which was defined above in the case of the measure μ_{ss}. Now we generalize its definition. For a measure μ_N in Ω_N and for each $x \in \Lambda_N$ we denote by $\rho_t^N(x)$ the empirical profile at the site x, given by

$$\rho_t^N(x) \;=\; E_{\mathbb{P}_{\mu_N}}[\eta_{tN^2}(x)].$$

We extend this definition to the boundary by setting

$$\rho_t^N(0) \;=\; \alpha \text{ and } \rho_t^N(N) \;=\; \beta\,, \text{ for all } t \geq 0\,.$$

Note that since μ_{ss} is a stationary measure the stationary empirical profile $\rho_{ss}^N(\cdot)$ does not depend on time, but now since μ_N is a general measure the empirical profile $\rho_t^N(\cdot)$ depends on time. From Kolmogorov's backward equation we know that $\rho_t^N(\cdot)$ is a solution of

$$\partial_t \rho_t^N(x) = E_{\mathbb{P}_{\mu_N}}[\mathcal{L}_N \eta_{tN^2}(x)].$$

A simple computation shows that

$$\mathcal{L}_N \eta(x) = j_{x-1,x}(\eta) - j_{x,x+1}(\eta)$$

where for $x \in \Lambda_N$, the quantity $j_{x,x+1}(\eta)$ denotes the microscopic current at the bond $\{x, x+1\}$, which is given by the difference between the jump rate from x to $x+1$ and the jump rate from $x+1$ to x. Note that for $x = 0$ (resp. $x = N-1$) $j_{x,x+1}$ is equal to the creation rate minus the annihilation rate at the site $x = 1$ (resp. $x = N-1$). Therefore

$$\begin{aligned} j_{0,1}(\eta) &= \frac{\kappa}{2N^\theta}(\alpha - \eta(1)), \\ j_{x,x+1}(\eta) &= \frac{1}{2}(\eta(x) - \eta(x+1)), \forall x \in \{1, ..., N-2\} \\ j_{N-1,N}(\eta) &= \frac{\kappa}{2N^\theta}(\eta(N-1) - \beta). \end{aligned} \tag{18}$$

A simple computation shows that $\rho_t^N(\cdot)$ is a solution of the equation

$$\begin{cases} \partial_t \rho_t^N(x) \;=\; \big(N^2 \mathcal{B}_N^\theta \rho_t^N\big)(x)\,, \;\; x \in \Lambda_N\,, \;\; t \geq 0\,, \\ \rho_t^N(0) \;= \alpha\;, \;\; t \geq 0, \\ \rho_t^N(N) \;=\; \beta\,, \;\; t \geq 0, \end{cases} \tag{19}$$

where the operator $\mathcal{B}_N^\theta$ acts on functions $f : \Lambda_N \cup \{0, N\} \to \mathbb{R}$ as

$$\begin{cases} N^2(\mathcal{B}_N^\theta f)(x) \;=\; \frac{1}{2}\Delta_N f(x)\,, \quad \text{for } x \in \{2, ..., N-2\}, \\ N^2(\mathcal{B}_N^\theta f)(1) = \frac{N^2}{2}(f(2) - f(1)) + \frac{\kappa N^2}{2N^\theta}(f(0) - f(1)), \\ N^2(\mathcal{B}_N^\theta f)(N-1) = \frac{N^2}{2}(f(N-2) - f(N-1)) + \frac{\kappa N^2}{2N^\theta}(f(N) - f(N-1)). \end{cases}$$

Above $\Delta_N f$ denotes the discrete Laplacian of $f(\cdot)$ which is given on $x \in \Lambda_N$ by

$$\Delta_N f(x) = f(x+1) + f(x-1) - 2f(x). \tag{20}$$

Note that for $\theta = 0$ the operator $\mathcal{B}_N^\theta$ is basically the discrete laplacian but when $\theta \neq 0$ we see some distortion at the boundary due to the mechanism of creation and annihilation.

A simple computation shows that the stationary solution of (19) is given by

$$\rho_{ss}^N(x) = E_{\mathbb{P}_{\mu_{ss}}}[\eta_{tN^2}(x)] = a_N x + b_N$$

where the coefficients a_N and b_N are equal to

$$a_N = \frac{\kappa(\beta-\alpha)}{2N^\theta + \kappa(N-2)} \quad \text{and} \quad b_N = a_N\Big(\frac{N^\theta}{\kappa} - 1\Big) + \alpha.$$

From this we get that

$$\lim_{N\to\infty} \max_{x\in\Lambda_N} \left|\rho_{ss}^N(x) - \bar{\rho}(\tfrac{x}{N})\right| = 0, \tag{21}$$

where for $q \in (0,1)$

$$\bar{\rho}(q) = \begin{cases} (\beta-\alpha)q + \alpha\,;\, \theta < 1, \\ \frac{\kappa(\beta-\alpha)}{2+\kappa}q + \alpha + \frac{\beta-\alpha}{2+\kappa}\,;\, \theta = 1, \\ \frac{\beta+\alpha}{2}\,;\, \theta > 1. \end{cases} \tag{22}$$

Note that $\bar{\rho}(\cdot)$ will be a stationary solution of the hydrodynamic equation that we are looking for. See Fig. 3 for a representation of $\bar{\rho}(\cdot)$.

Now we obtain information about the two point correlation function. Let

$$V_N = \{(x,y) \in \{0,...,N\}^2 : 0 < x < y < N\},$$

and its boundary

$$\partial V_N = \{(x,y) \in \{0,...,N\}^2 : x = 0 \text{ or } y = N\}.$$

See Fig. 2.

For $x < y \in V_N$, let $\varphi_t^N(x,y)$ denote the two point correlation function between the occupation sites at $x < y \in V_N$ which is defined by

$$\varphi_t^N(x,y) = E_{\mathbb{P}_{\mu_N}}[(\eta_{tN^2}(x) - \rho_t^N(x))(\eta_{tN^2}(y) - \rho_t^N(y))]. \tag{23}$$

Doing some simple, but long, computations we see that φ_t^N is a solution of

$$\begin{cases} \partial_t \varphi_t^N(x,y) = n^2 \mathcal{A}_N^\theta \varphi_t^n(x,y) + g_t^N(x,y), & \text{for } (x,y) \in V_N,\ t > 0, \\ \varphi_t^N(x,y) = 0, & \text{for } (x,y) \in \partial V_N,\ t > 0, \\ \varphi_0^N(x,y) = E_{\mu_N}[\eta_0(x)\eta_0(y)] - \rho_0^N(x)\rho_0^N(y), & \text{for } (x,y) \in V_N \cup \partial V_N, \end{cases} \tag{24}$$

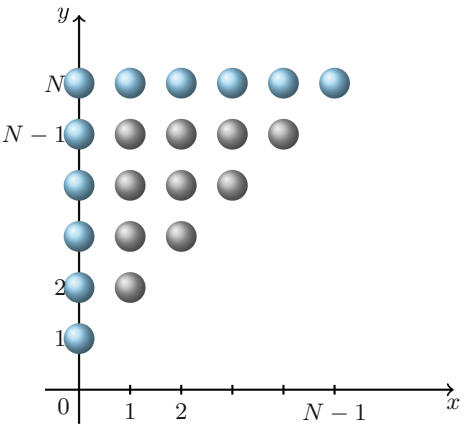


Fig. 2. The set V_N and its boundary ∂V_N

where $\mathcal{A}_N^\theta$ is the linear operator that acts on functions $f : V_N \cup \partial V_N \to \mathbb{R}$ as

$$(\mathcal{A}_N^\theta f)(u) = \sum_{v \in V_N} c_N^\theta(u,v)\big[f(v) - f(u)\big],$$

with

$$c_N^\theta(u,v) \;=\; \begin{cases} 1, & \text{if } \|u-v\| = 1 \text{ and } \; u, v \in V_N, \\ N^{-\theta}, & \text{if } \|u-v\| = 1 \text{ and } u \in V_N, \; v \in \partial V_N, \\ 0, & \text{otherwise,} \end{cases}$$

for $\theta \geq 0$. Note that $\mathcal{A}_N^\theta$ is the generator of a random walk in $V_N \cup \partial V_N$ with jump rates given by $c_N^\theta(u,v)$, which is absorbed at ∂V_N. Above $\|\cdot\|$ denotes the supremum norm,

$$g_t^N(x,y) = -(\nabla_N^+ \rho_t^N(x))^2 \delta_{y=x+1}$$

and

$$\nabla_N^+ \rho_t^N(x) = N(\rho_t^N(x+1) - \rho_t^N(x)). \tag{25}$$

In this case, contrarily to the empirical profile, it is quite complicated to obtain an expression for the stationary solution of (24). Nevertheless, we note that a simple, but long, computation shows that the solution obtained in (16), in the case where the starting measure is the stationary state μ_{ss}, is in fact the stationary solution of (24). We also observe that in [9] it was obtained the following bound on the case $\theta = 1$ for a general initial measure μ_N. There it was proved that there exists a constant $C > 0$ such that

$$\sup_{t \geq 0} \max_{(x,y) \in V_N} |\varphi_t^N(x,y)| \;\leq\; \frac{C}{N}\,, \tag{26}$$

but we note that the bounds on the other regimes of θ are still open, apart the case $\theta = 0$ where the bound above is given by C/N^2, see [16].

2.6 Hydrodynamic Equations

From now on up to the rest of these notes we fix a finite time horizon $[0,T]$. We denote by $\langle\cdot,\cdot\rangle_\mu$ the inner product in $L^2([0,1])$ with respect to a measure μ defined in $[0,1]$ and $\|\cdot\|_{L^2(\mu)}$ is the corresponding norm. We note that when μ is the Lebesgue measure we write $\langle\cdot,\cdot\rangle$ and $\|\cdot\|_{L^2}$ for the corresponding norm.

We denote by $C^{m,n}([0,T]\times[0,1])$ the set of functions defined on $[0,T]\times[0,1]$ that are m times differentiable on the first variable and n times differentiable on the second variable and with continuous derivatives. For a function $G := G(s,q) \in C^{m,n}([0,T]\times[0,1])$ we denote by $\partial_s G$ its derivative with respect to the time variable s and by $\partial_q G$ its derivative with respect to the space variable q. For simplicity of notation we set $\Delta G := \partial_q^2 G$. We will also make use of the set $C_c^{m,n}([0,T]\times[0,1])$ of functions $G \in C^{m,n}([0,T]\times[0,1])$ such that for any time s the function G_s has a compact support included in $(0,1)$ and we denote by $C_c^m(0,1)$ (resp. $C_c^\infty(0,1)$) the set of all m continuously differentiable (resp. smooth) real-valued functions defined on $(0,1)$ with compact support. The supremum norm is denoted by $\|\cdot\|_\infty$. Finally, $C_0^{m,n}([0,T]\times[0,1])$ is the set of functions $G \in C^{m,n}([0,T]\times[0,1])$ such that for any time s the function G_s vanishes at the boundary, that is, $G_s(0) = G_s(1) = 0$.

Now we want to define the space where the solutions of the hydrodynamic equations will live on, namely the Sobolev space $\mathcal{H}_1$ on $[0,1]$. For that purpose, we define the semi inner-product $\langle\cdot,\cdot\rangle_1$ on the set $C^\infty([0,1])$ by

$$\langle G, H\rangle_1 = \int_0^1 (\partial_q G)(q)\,(\partial_q H)(q)\,dq, \tag{27}$$

for $G, H \in C^\infty([0,1])$ and the corresponding semi-norm is denoted by $\|\cdot\|_1$.

Definition 1. *The Sobolev space $\mathcal{H}^1$ on $[0,1]$ is the Hilbert space defined as the completion of $C^\infty([0,1])$ for the norm*

$$\|\cdot\|_{\mathcal{H}^1}^2 := \|\cdot\|_{L^2}^2 + \|\cdot\|_1^2.$$

Its elements elements coincide a.e. with continuous functions.

The space $L^2(0,T;\mathcal{H}^1)$ is the set of measurable functions $f : [0,T] \to \mathcal{H}^1$ such that

$$\int_0^T \|f_s\|_{\mathcal{H}^1}^2 ds < \infty.$$

We can now give the definition of the weak solutions of the hydrodynamic equations that will be derived for the symmetric simple exclusion process in contact with stochastic reservoirs. We start by giving the notion of a weak solution to the heat equation with Dirichlet boundary conditions which will be the notion that we will derive in the regime $\theta \in [0,1)$. In what follows $g : [0,1] \to [0,1]$ is a measurable function and it is the initial condition of all the partial differential equations that we define below, that is $\rho_0(q) = g(q)$, for all $q \in (0,1)$.

Definition 2. *We say that $\rho : [0,T] \times [0,1] \to [0,1]$ is a weak solution of the heat equation with Dirichlet boundary conditions*

$$\begin{cases} \partial_t \rho_t(q) = \frac{1}{2} \Delta \rho_t(q), \quad (t,q) \in [0,T] \times (0,1), \\ \rho_t(0) = \alpha, \quad \rho_t(1) = \beta, \quad t \in (0,T], \end{cases} \tag{28}$$

starting from a measurable function $g : [0,1] \to [0,1]$, if the following two conditions hold:

1. *$\rho \in L^2(0,T;\mathcal{H}^1)$;*
2. *ρ satisfies the weak formulation:*

$$\begin{aligned} F_{Dir} := & \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq \\ & - \int_0^t \int_0^1 \rho_s(q) \Big(\frac{1}{2} \Delta + \partial_s \Big) G_s(q)\, dq\, ds \\ & + \int_0^t \Big\{ \frac{\beta}{2} \partial_q G_s(1) - \frac{\alpha}{2} \partial_q G_s(0) \Big\}\, ds = 0, \end{aligned} \tag{29}$$

for all $t \in [0,T]$ and any function $G \in C_0^{1,2}([0,T] \times [0,1])$.

In the regime $\theta < 0$ we will make use of another notion of weak solution to the heat equation with Dirichlet boundary conditions which uses as input for test functions elements in the set $C_c^{1,2}([0,T] \times [0,1])$. Since functions in that space have compact support, in order to get a proper notion of weak solution we need to add an extra condition to Definition 2 (see 3. in Definition 3).

Definition 3. *We say that $\rho : [0,T] \times [0,1] \to [0,1]$ is a weak solution of the heat equation with Dirichlet boundary conditions given in (28), starting from a measurable function $g : [0,1] \to [0,1]$, if the following three conditions hold:*

1. *$\rho \in L^2(0,T;\mathcal{H}^1)$,*
2. *ρ satisfies the weak formulation:*

$$\begin{aligned} F_{Dir}^c := & \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq \\ & - \int_0^t \int_0^1 \rho_s(q) \Big(\frac{1}{2} \Delta + \partial_s \Big) G_s(q)\, dq\, ds = 0, \end{aligned} \tag{30}$$

for all $t \in [0,T]$ and any function $G \in C_c^{1,2}([0,T] \times [0,1])$,
3. *$\rho_t(0) = \alpha, \quad \rho_t(1) = \beta$ for all $t \in (0,T]$.*

Remark 1. We note that (30) coincides with (29) by taking as input a test function $G \in C_c^{1,2}([0,T] \times [0,1])$, since in this case $\partial_q G_s(0) = \partial_q G_s(1) = 0$, so that the last term in (29) vanishes.

Now we introduce the notion of weak solution of the hydrodynamic equation that we will derive in the case $\theta = 1$. In this regime the boundary reservoirs are so slow and as a consequence, a different boundary condition appears. In the case of Dirichlet boundary conditions, the value of the profile $\rho_t(\cdot)$ is fixed to be equal to α at 0 and β at 1. This is no longer the case when $\theta \geq 1$ as we will see later on.

Definition 4. *We say that* $\rho : [0,T] \times [0,1] \to [0,1]$ *is a weak solution of the heat equation with Robin boundary conditions*

$$\begin{cases} \partial_t \rho_t(q) = \frac{1}{2} \Delta \rho_t(q), \quad (t,q) \in [0,T] \times (0,1), \\ \partial_q \rho_t(0) = \kappa(\rho_t(0) - \alpha), \quad \partial_q \rho_t(1) = \kappa(\beta - \rho_t(1)), \quad t \in (0,T], \end{cases} \tag{31}$$

starting from a measurable function $g : [0,1] \to [0,1]$*, if the following two conditions hold:*

1. $\rho \in L^2(0,T;\mathcal{H}^1)$,
2. ρ *satisfies the weak formulation:*

$$\begin{aligned} F_{Rob} &:= \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq \\ &- \int_0^t \int_0^1 \rho_s(q) \Big(\frac{1}{2} \Delta + \partial_s \Big) G_s(q)\, ds\, dq + \frac{1}{2} \int_0^t \{ \rho_s(1) \partial_q G_s(1) - \rho_s(0) \partial_q G_s(0) \}\, ds \\ &- \frac{\kappa}{2} \int_0^t \{ G_s(0)(\alpha - \rho_s(0)) + G_s(1)(\beta - \rho_s(1)) \}\, ds = 0, \end{aligned} \tag{32}$$

for all $t \in [0,T]$ *and any function* $G \in C^{1,2}([0,T] \times [0,1])$.

In the regime $\theta = 1$ the boundary reservoirs are so slow so that a type of Robin boundary condition appears. In this case it fixes the value of the flux through the system as being proportional to the difference of concentration. Note that, for example at $q = 0$, the value $\partial_q \rho_t(0)$ corresponds to the flux of particles through the left boundary and $\kappa(\rho_t(0) - \alpha)$ corresponds to the difference of the concentration, since in this case, contrarily to what happens in the case of Dirichlet boundary conditions, it is not true that $\rho_t(0) = \alpha$ (the value of the profile at the boundaries is not fixed!)

Remark 2. Observe that in the case $\kappa = 0$ the equation above is the heat equation with Neumann boundary conditions and it is the hydrodynamic equation that we will derive in the case $\theta > 1$.

Remark 3. We observe that all the partial differential equations defined above have a unique weak solution in the sense given above. We do not include the proof of this result in these notes but we refer the interested reader to [2] for the proof of the uniqueness in the case of Dirichlet boundary conditions and to [1] for the proof of the uniqueness in the case of Robin boundary conditions.

Deriving the Weak Formulation: We note that the weak formulation given in all the regimes above can be obtained from the formal expression of the corresponding partial differential equation in the following way. Take a test function $G \in C^{1,2}([0,T] \times [0,1])$ and multiply both sides of the equality

$$\partial_s \rho_s(q) = \frac{1}{2} \Delta \rho_s(q)$$

by G and then integrate in the time interval $[0,t]$ and in the space interval $[0,1]$ to get

$$\int_0^1 \int_0^t \partial_s \rho_s(q) G_s(q)\, ds\, dq = \int_0^1 \int_0^t \frac{1}{2} \Delta\, \rho_s(q) G_s(q)\, ds\, dq. \tag{33}$$

To treat the term at the left hand side of last display, we perform an integration by parts in the time integral and we get to

$$\int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq - \int_0^t \int_0^1 \rho_s(q) \partial_s G_s(q)\, ds\, dq. \tag{34}$$

The term at the right hand side of (33) can be treated by doing an integration by parts in the space integral and we get to

$$\frac{1}{2} \int_0^t \Big\{ \partial_q \rho_s(1) G_s(1) - \partial_q \rho_s(0) G_s(0) \Big\}\, ds - \frac{1}{2} \int_0^t \int_0^1 \partial_q \rho_s(q) \partial_q G_s(q)\, ds\, dq.$$

Now, we do another integration by parts in the integral in space at the term on the right hand side of last expression and we write the previous display as

$$\begin{aligned} &\frac{1}{2} \int_0^t \Big\{ \partial_q \rho_s(1) G_s(1) - \partial_q \rho_s(0) G_s(0) \Big\}\, ds \\ &-\frac{1}{2} \int_0^t \Big\{ \rho_s(1) \partial_q G_s(1) - \rho_s(0) \partial_q G_s(0) \Big\} ds + \frac{1}{2} \int_0^t \int_0^1 \rho_s(q) \Delta G_s(q)\, ds\, dq. \end{aligned} \tag{35}$$

Putting together (35) and (34) we obtain

$$\begin{aligned} \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq = &\int_0^t \int_0^1 \rho_s(q) \Big(\frac{1}{2} \Delta + \partial_s \Big) G_s(q)\, ds\, dq \\ &+ \frac{1}{2} \int_0^t \Big\{ \partial_q \rho_s(1) G_s(1) - \partial_q \rho_s(0) G_s(0) \Big\}\, ds \\ &- \frac{1}{2} \int_0^t \Big\{ \rho_s(1) \partial_q G_s(1) - \rho_s(0) \partial_q G_s(0) \Big\}\, ds. \end{aligned}$$

Now we obtain each one of the weak formulations given above. We start with the case where $G \in C_0^{1,2}([0,T] \times [0,1])$ and we will derive (29). For that purpose note that since G vanishes at the boundary of $[0,1]$ and since $\rho_s(0) = \alpha$ and $\rho_s(1) = \beta$, the expression in the previous display becomes equivalent to $F_{Dir} = 0$.

On the other hand, if $G \in C_c^{1,2}([0,T] \times [0,1])$, then G vanishes at the boundary of $[0,1]$ and $\partial_q G$ also vanishes at the boundary of $[0,1]$, so that for ρ satisfying the Dirichlet boundary conditions of (28), the expression in the display above becomes equivalent to $F^c_{Dir} = 0$.

Finally for $G \in C^{1,2}([0,T] \times [0,1])$ and for $\rho(\cdot)$ satisfying the Robin boundary conditions of (31), the expression in the previous display becomes equivalent to $F_{Rob} = 0$.

Stationary Solutions: Now we deduce the stationary solutions for each one of the equations given above. We start with (28). For that purpose note that, denoting by $\bar{\rho}(\cdot)$ the stationary solution we have that $\Delta\bar{\rho}(t,q) = 0$ implies that $\bar{\rho}(q) = aq + b$ for $a, b \in \mathbb{R}$ and $q \in (0,1)$. Imposing the Dirichlet boundary conditions we arrive at

$$a = (\beta - \alpha) \quad \text{and} \quad b = \beta,$$

so that

$$\bar{\rho}_{Dir}(q) = (\beta - \alpha)q + \alpha. \tag{36}$$

On the other hand, imposing the Robin boundary conditions of (31) we arrive at

$$a = \frac{\kappa(\beta - \alpha)}{2 + \kappa} \quad \text{and} \quad b = \alpha + \frac{\beta - \alpha}{2 + \kappa},$$

so that for $q \in (0,1)$

$$\bar{\rho}_{Rob}(q) = \frac{\kappa(\beta - \alpha)}{2 + \kappa} q + \alpha + \frac{\beta - \alpha}{2 + \kappa}. \tag{37}$$

Finally, if we impose the Neumann boundary conditions, any constant solution is a stationary solution of (31) with $\kappa = 0$ (which corresponds to the Neumann regime). In this case we note that the stationary solution is not unique. Below we draw the graph of these stationary solutions for a choice of $\alpha = 0.2$ and $\beta = 0.8$ (Fig. 3).

Now we give the explicit expression for the solution of each hydrodynamic equation.

Proposition 1. *We have that:*

1. *The solution of* (28) *with initial condition* $g(\cdot)$ *is equal to*

$$\rho_t(q) = \bar{\rho}_{Dir}(q) + \sum_{n=1}^{\infty} e^{-\frac{(n\pi)^2}{2}t} 2\sin(n\pi q).$$

2. *The solution of* (31) *with initial condition* $g(\cdot)$ *is equal to*

$$\rho_t(q) = \bar{\rho}_{Rob}(q) + \sum_{n=1}^{\infty} C_n e^{-\frac{\lambda_n}{2}t} X_n(q),$$

where

$$X_n(q) = A_n \sin(\sqrt{\lambda_n}\, q) + A_n \kappa \sqrt{\lambda_n} \cos(\sqrt{\lambda_n}\, q), \tag{38}$$

A_n is a normalizing constant in such a way that X_n has unitary $L^2([0,1])$-norm and

$$C_n = \int_0^1 (g(q) - \bar{\rho}_{Rob}(q)) X_n(q) dq.$$

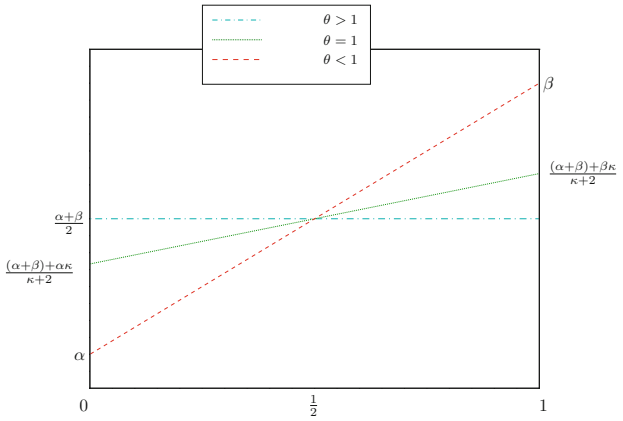


Fig. 3. Stationary solutions of the hydrodynamic equations.

Proof. The solution $\rho(\cdot)$ to (28) starting from a profile $g(\cdot)$ is such that $u = \rho - \bar{\rho}_{Dir}$ is solution to (28) with homogeneous boundary conditions $\alpha = \beta = 0$, i.e.

$$\begin{cases} \partial_t u_t(q) = \frac{1}{2}\Delta u_t(q), \quad (t,q) \in [0,T] \times (0,1), \\ u_t(0) = 0 = u_t(1), \quad t \in [0,T]. \end{cases} \tag{39}$$

It is well known that $u_t(q)$ is given by

$$u_t(q) = \sum_{n=1}^{\infty} e^{-\frac{(n\pi)^2}{2}t} 2\sin(n\pi q).$$

From the previous computations we conclude that the solution $\rho(\cdot)$ of (28) starting from $g(\cdot)$ is given by

$$\rho_t(q) = (\beta - \alpha)q + \alpha + \sum_{n=1}^{\infty} e^{-\frac{(n\pi)^2}{2}t} 2\sin(n\pi q).$$

On the other hand, the solution $\rho(\cdot)$ of (31) starting from $g(\cdot)$ is such that $u = \rho - \bar{\rho}_{Rob}$ is solution to (31) with $\alpha = \beta = 0$, i.e.

$$\begin{cases} \partial_t u_t(q) = \frac{1}{2}\Delta u_t(q), \quad (t,q) \in [0,T] \times (0,1), \\ \partial_q u_t(0) = \kappa u_t(0), \quad \partial_q u_t(1) = -\kappa u_t(1), \quad t \in [0,T]. \end{cases} \tag{40}$$

It is well known that $u_t(q)$ is given by

$$u_t(q) = \sum_{n=1}^{\infty} C_n e^{-\frac{\lambda_n}{2}t} X_n(q),$$

where $X_n(q)$ writes as

$$X_n(q) = A_n \sin(\sqrt{\lambda_n} q) + B_n \cos(\sqrt{\lambda_n} q),$$

for some constants A_n and B_n. Then, the first boundary condition in (40) gives $B_n = \sqrt{\lambda_n}\kappa A_n$. To avoid the null solution we consider $A_n \neq 0$. The second boundary condition in (40) gives

$$\tan(\sqrt{\lambda_n}) = \frac{2\kappa\sqrt{\lambda_n}}{\lambda_n\kappa^2 - 1}, \tag{41}$$

whose solution λ_n satisfying $(n-1)\pi \leq \sqrt{\lambda_n} \leq n\pi$ is such that $\lambda_n \sim n^2\pi^2$ as $n \to \infty$. From the previous computations we get that $X_n(q)$ is given by (38) and there A_n is a normalizing constant in such a way that X_n has unitary $L^2([0,1])$-norm. Moreover

$$C_n = \int_0^1 (g(q) - \bar{\rho}_{Rob}(q))X_n(q)dq.$$

From the previous computations we conclude that the solution $\rho(\cdot)$ of (31) starting from $g(\cdot)$ is given by

$$\rho_t(q) = \frac{\kappa(\beta-\alpha)}{2+\kappa}q + \alpha + \frac{\beta-\alpha}{2+\kappa} + \sum_{n=1}^{\infty} C_n e^{-\frac{\lambda_n}{2}t} X_n(q).$$

2.7 Hydrodynamic Limit

In this section we want to state the hydrodynamic limit of the process $\{\eta_{tN^2} : t \geq 0\}$ with state space Ω_N and with infinitesimal generator $N^2\mathcal{L}_N$ defined in (1). Note that here we are going to take $\Theta(N) = N^2$. Let $\mathcal{M}^+$ be the space of positive measures on $[0,1]$ with total mass bounded by 1 equipped with the weak topology. We can define a metric $d(\cdot,\cdot)$ in the space $\mathcal{M}^+$ by taking a dense countable set $\{f_n\}_{n\geq 1}$ of real valued continuous functions defined in $[0,1]$ through the following expression:

$$d(\mu,\nu) = \sum_{n\geq 1} \frac{1}{2^k} \frac{|\int f_n d\mu - \int f_n d\nu|}{1 + |\int f_n d\mu - \int f_n d\nu|}. \tag{42}$$

For any configuration $\eta \in \Omega_N$ we define the empirical measure $\pi^N(\eta, dq)$ on $[0,1]$ by

$$\pi^N(\eta, dq) = \frac{1}{N-1} \sum_{x\in\Lambda_N} \eta(x)\delta_{\frac{x}{N}}(dq), \tag{43}$$

where δ_a is a Dirac mass on $a \in [0,1]$, and

$$\pi_t^N(\eta, dq) := \pi^N(\eta_{tN^2}, dq).$$

This measure gives weight $\frac{1}{N-1}$ to each occupied site of the configuration η.

Fix $T > 0$ and $\theta \in \mathbb{R}$. Recall that $\mathbb{P}_{\mu_N}$ is the probability measure in the Skorohod space $\mathcal{D}([0,T], \Omega_N)$ induced by the Markov process $\{\eta_{tN^2} : t \geq 0\}$ and the initial probability measure μ_N and we denote by $E_{\mathbb{P}_{\mu_N}}$ the expectation

with respect to $\mathbb{P}_{\mu_N}$. Now let $\{\mathbb{Q}_N\}_{N\geq 1}$ be the sequence of probability measures on $\mathcal{D}([0,T],\mathcal{M}^+)$ induced by the Markov process $\{\pi_t^N : t\geq 0\}$ and by $\mathbb{P}_{\mu_N}$.

At this point we need to fix an initial profile $\rho_0 : [0,1]\to[0,1]$ which is measurable and an initial probability measure $\mu_N \in \Omega_N$. We are going to consider the following set of initial measures:

Definition 5. *A sequence of probability measures $\{\mu_N\}_{N\geq 1}$ in Ω_N is associated to the profile $\rho_0(\cdot)$ if for any continuous function $G : [0,1]\to\mathbb{R}$ and any $\delta > 0$*

$$\lim_{N\to\infty}\mu_N\Big(\eta\in\Omega_N : \Big|\frac{1}{N-1}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\eta(x) - \int_0^1 G(q)\rho_0(q)dq\Big| > \delta\Big) = 0. \quad (44)$$

Note that (44) states that

$$\int G(q)\pi^N(\eta,dq) \longrightarrow_{N\to\infty} \int_0^1 G(q)\rho_0(q)dq, \quad (45)$$

with respect to μ_N, which means that the empirical measure at time $t = 0$ converges, in probability with respect to μ_N, as $N\to\infty$, to the deterministic measure $\rho_0(q)dq$, which is absolutely continuous with respect to the Lebesgue measure and the density is the profile $\rho_0(\cdot)$.

The hydrodynamic limit that we want to derive states that the previous result is also true for any $t\in[0,T]$, that is, the empirical measure at time t converges in probability with respect to the distribution of the system at time t, as $N\to\infty$, to the deterministic measure $\rho_t(q)dq$, where $\rho_t(\cdot)$ is a solution (here in the weak sense) to some partial differential equation, *the hydrodynamic equation*.

The first main result of these notes is summarized in the following theorem (see also Fig. 4).

Theorem 1. *Let $g : [0,1]\to[0,1]$ be a measurable function and let $\{\mu_N\}_{N\geq 1}$ be a sequence of probability measures in Ω_N associated to $g(\cdot)$. Then, for any $t\in[0,T]$,*

$$\lim_{N\to\infty}\mathbb{P}_{\mu_N}\Big(\eta_\cdot : \Big|\frac{1}{N-1}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\eta_{tN^2}(x) - \int_0^1 G(q)\rho_t(q)dq\Big| > \delta\Big) = 0,$$

where $\rho_t(\cdot)$ is the unique weak solution of :

- (28) *as given in Definition 3, if $\theta < 0$;*
- (28) *as given in Definition 2, if $\theta\in[0,1)$;*
- *(31), if $\theta = 1$;*
- *(31) with $\kappa = 0$, if $\theta > 1$.*

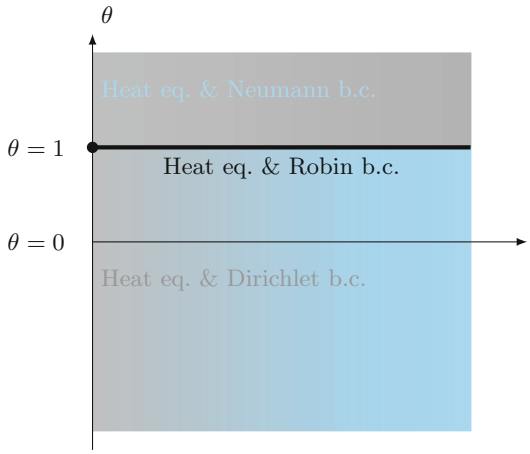


Fig. 4. The three hydrodynamic equations depending on θ.

Remark 4. We note that in [1] it was studied the case where the reservoirs are slowed (which corresponds to the regime $\theta \geq 0$). In the previous theorem we considered also the case where the reservoirs are fast (which corresponds to $\theta < 0$) but we note that the macroscopic behavior of the system is also given by the heat equation with Dirichlet boundary conditions as happens in the case $\theta \in [0, 1)$. To prove this result we note that the notion of weak solution in the case $\theta < 0$ is different from the notion of weak solution in the case $\theta \in [0, 1)$ since it uses as input functions with compact support.

The proof of Theorem 1 proceeds as follows: We split the proof into showing first the tightness of the sequence $\{\mathbb{Q}_N\}_{N\geq 1}$ and then we characterize uniquely the limiting point $\mathbb{Q}$ of this sequence. These two results combined together, imply the convergence of $\{\mathbb{Q}_N\}_{N\geq 1}$ to $\mathbb{Q}$ as $N \to \infty$.

The next section is dedicated to the presentation of an heuristic argument to deduce the hydrodynamic equations from the interacting particle system by means of the Dynkin's formula; in Sect. 2.9 we present the proof of tightness and in Sect. 2.10 we characterize the limit point $\mathbb{Q}$. We note that in order to characterize the limit point $\mathbb{Q}$, we prove in Sect. 2.10 that all limiting points of the sequence $\{\mathbb{Q}_N\}_{N\geq 1}$ are concentrated on trajectories of measures that are absolutely continuous with respect to the Lebesgue measure and that the density $\rho_t(\cdot)$ is a weak solution of the corresponding hydrodynamic equation. From the uniqueness of weak solutions of the hydrodynamic equations, see Remark 3, we conclude that $\{\mathbb{Q}_N\}_{N\geq 1}$ has a unique limit point $\mathbb{Q}$, and therefore we conclude the convergence of the sequence to this limit point.

2.8 Heuristics for Hydrodynamic Equations

In this section we give the main ideas which are behind the identification of limit points as weak solutions of the partial differential equations given in Sect. 2.6.

Now we argue that the density $\rho_t(\cdot)$ is a weak solution of the corresponding hydrodynamic equation for each regime of θ. We remark that we are not going to prove here that the solution $\rho_t(\cdot)$ belongs to the space $L^2(0,T;\mathcal{H}^1)$ but we refer the reader to [1,2] for a complete proof of this fact. In order to prove that $\rho_t(\cdot)$ satisfies the weak formulation we use auxiliary martingales associated to the Markov process $\{\eta_t : t \geq 0\}$. For that purpose, and to make the exposition simpler, we fix a function $G : [0,1] \to \mathbb{R}$ which does not depend on time and which is two times continuously differentiable. If $\theta < 0$ we will assume further that it has a compact support included in $(0,1)$. First we recall Dynkin's formula.

Theorem 2. *Let $\{\eta_t : t \geq 0\}$ be a Markov process with generator $\mathcal{L}$ and with countable state space E. Let $F : \mathbb{R}^+ \times E \to \mathbb{R}$ be a bounded function such that*

- *$\forall \eta \in E, F(\cdot,\eta) \in C^2(\mathbb{R}^+)$,*
- *there exists a finite constant C, such that for $j = 1, 2$*

$$\sup_{(s,\eta)} |\partial_s^j F(s,\eta)| \leq C.$$

For $t \geq 0$, let

$$\begin{aligned} M_t^F =& F(t,\eta_t) - F(0,\eta_0) - \int_0^t (\partial_s + \mathcal{L})F(s,\eta_s)ds, \\ N_t^F =& (M_t^F)^2 - \int_0^t \{\mathcal{L}F(s,\eta_s)^2 - 2F(s,\eta_s)\mathcal{L}F(s,\eta_s)\}ds. \end{aligned}$$

Then, $\{M_t^F\}_{t\geq 0}$ and $\{N_t^F\}_{t\geq 0}$ are martingales with respect to $\mathcal{F}_t = \sigma(\eta_s\,;\, s \leq t)$.

Let us fix a test function $G : [0,1] \to \mathbb{R}$ and apply Dynkin's formula with

$$F(t,\eta_t) = \langle \pi_t^N, G\rangle = \frac{1}{N-1} \sum_{x\in\Lambda_N} \eta_{tN^2}(x) G\Big(\tfrac{x}{N}\Big). \tag{46}$$

Above $\langle \pi_t^N, G\rangle$ represents the integral of G with respect the measure π_t^N. Note that F does not depend on time, only through η_t. A simple computation shows that

$$\begin{aligned} N^2\mathcal{L}_N\langle \pi_s^N, G\rangle =& \langle \pi_s^N, \frac{1}{2}\Delta_N G\rangle \\ &+ \frac{1}{2}\Big(\nabla_N^+ G(0)\eta_{sN^2}(1) - \nabla_N^- G(1)\eta_{sN^2}(N-1)\Big) \\ &+ \frac{\kappa}{2}\frac{N^{2-\theta}}{N-1} G\Big(\tfrac{1}{N}\Big)(\alpha - \eta_{sN^2}(1)) \\ &+ \frac{\kappa}{2}\frac{N^{2-\theta}}{N-1} G\Big(\tfrac{N-1}{N}\Big)(\beta - \eta_{sN^2}(N-1)), \end{aligned} \tag{47}$$

from where we obtain that

$$\begin{aligned} M_t^N(G) = \langle \pi_t^N, G\rangle - \langle \pi_0^N, G\rangle - \int_0^t \langle \pi_s^N, \frac{1}{2}\Delta_N G\rangle\, ds \\ - \frac{1}{2}\int_0^t \nabla_N^+ G(0)\eta_{sN^2}(1) - \nabla_N^- G(1)\eta_{sN^2}(N-1)\, ds \\ - \frac{\kappa}{2}\int_0^t \frac{N^{2-\theta}}{N-1} G\Big(\tfrac{1}{N}\Big)(\alpha - \eta_{sN^2}(1))\, ds \\ - \frac{\kappa}{2}\int_0^t \frac{N^{2-\theta}}{N-1} G\Big(\tfrac{N-1}{N}\Big)(\beta - \eta_{sN^2}(N-1))\, ds, \end{aligned} \tag{48}$$

is a martingale with respect to the natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where for each $t \geq 0$, $\mathcal{F}_t := \sigma(\eta_s : s < t)$. Above, Δ_N is the discrete laplacian defined in (20), ∇_N^+ is defined in (25) and

$$\nabla_N^- f(x) = N(f(x) - f(x-1)).$$

Now we look at the integral terms in (48).

The Case $\boldsymbol{\theta} \in [\mathbf{0}, \mathbf{1})$: In this regime, we take a test function $G : [0,1] \to \mathbb{R}$ two times continuously differentiable such that $G(0) = G(1) = 0$. Then, we can subtract $G(0)$ (resp. $G(1)$) in the fifth term (resp. sixth term) at the right hand side of (48) and then doing a Taylor expansion on G we get that

$$\begin{aligned} M_t^N(G) = \langle \pi_t^N, G\rangle - \langle \pi_0^N, G\rangle - \int_0^t \langle \pi_s^N, \frac{1}{2}\Delta_N G\rangle ds \\ - \frac{1}{2}\int_0^t \nabla_N^+ G(0)\eta_{sN^2}(1) - \nabla_N^- G(1)\eta_{sN^2}(N-1) ds + O(N^{-\theta}). \end{aligned}$$

If we can replace $\eta_{sN^2}(1)$ by α and $\eta_{sN^2}(N-1)$ by β, which will be a consequence of Lemma 9 in Appendix A.4 (see Remark 19), then above we have

$$\begin{aligned} M_t^N(G) = \langle \pi_t^N, G\rangle - \langle \pi_0^N, G\rangle - \int_0^t \langle \pi_s^N, \frac{1}{2}\Delta_N G\rangle ds \\ - \frac{1}{2}\int_0^t \nabla_N^+ G(0)\alpha - \nabla_N^- G(1)\beta ds + O(N^{-\theta}) \end{aligned}$$

plus a term that vanishes as $N \to +\infty$.

Taking the expectation with respect to $\mathbb{P}_{\mu_N}$ in the expression above we get

$$\begin{aligned} \frac{1}{N-1}\sum_{x=1}^{N-1} G\Big(\tfrac{x}{N}\Big)\Big(\rho_t^N(x) - \rho_0^N(x)\Big) - \int_0^t \frac{1}{N-1}\sum_{x=1}^{N-1}\frac{1}{2}\Delta_N G\Big(\tfrac{x}{N}\Big)\rho_s^N(x) ds \\ - \frac{1}{2}\int_0^t \nabla_N^+ G(0)\alpha - \nabla_N^- G(1)\beta ds + O(N^{-\theta}) = 0. \end{aligned}$$

Note that above we used the fact that the average of martingales is constant in time and that $M_0^N(G) = 0$. Now, assuming that $\rho_t^N(x) \sim \rho_t(\frac{x}{N})$ and taking the limit as $N \to \infty$ we get that

$$\int_0^1 \rho_t(q)G(q) - \rho_0(q)G(q)dq - \int_0^t \int_0^1 \frac{1}{2}\Delta G(q)\rho_s(q)dqds$$
$$- \frac{1}{2}\int_0^t \partial_q G(0)\alpha - \partial_q G(1)\beta ds = 0.$$

Note that the restriction $\theta \geq 0$ comes from the fact that the errors, which arise from the Taylor expansion in G, have to vanish as $N \to \infty$ and the restriction $\theta < 1$ comes from the replacement of the occupation variables $\eta(1)$ and $\eta(N-1)$ by α and β, respectively, see Lemma 9 in Appendix A.4. At this point compare the previous expression with the weak formulation given in (29) and note that the test function G does not depend on time.

The Case $\theta < 0$: In this regime we take a function $G : [0,1] \to \mathbb{R}$ with compact support and we note that the last three terms at the right hand side of (48) vanish in this case. From this and the same arguments as above we get that

$$M_t^N(G) = \langle \pi_t^N, G\rangle - \langle \pi_0^N, G\rangle - \int_0^t \langle \pi_s^N, \frac{1}{2}\Delta_N G\rangle ds.$$

Taking the expectation with respect to $\mathbb{P}\mu_N$ in the expression above and assuming that $\rho_t^N(x) \sim \rho_t(\frac{x}{N})$, and then taking the limit as $N \to \infty$ we get that

$$\int_0^1 \rho_t(q)G(q) - \rho_0(q)G(q)dq - \int_0^t \int_0^1 \frac{1}{2}\Delta G(q)\rho_s(q)dqds = 0.$$

Again compare with the weak formulation given in (30) and note that the test function G does not depend on time.

Remark 5. We remark here that in this particular case there is an extra condition in Definition 3 with respect to the other notions of weak solutions where we only have to check the weak formulation and to show that the solution belongs to a Sobolev space. In this case we also need to show that the value of the profile $\rho_t(\cdot)$ is fixed at the boundary. We leave this issue to Appendix A.4.

The Case $\theta = 1$: In this case we consider an arbitrary function $G : [0,1] \to \mathbb{R}$ which is two times continuously differentiable and we get

$$M_t^N(G) = \langle \pi_t^N, G\rangle - \langle \pi_0^N, G\rangle - \int_0^t \langle \pi_s^N, \frac{1}{2}\Delta_N G\rangle ds$$
$$- \frac{1}{2}\int_0^t \nabla_N^+ G(0)\eta_{sN^2}(1) - \nabla_N^- G(1)\eta_{sN^2}(N-1)ds$$
$$- \frac{\kappa}{2}\frac{N}{N-1}\int_0^t G\Big(\tfrac{1}{N}\Big)(\alpha - \eta_{sN^2}(1)) + G\Big(\tfrac{N-1}{N}\Big)(\beta - \eta_{sN^2}(N-1))ds.$$

In this regime Lemma 9 in Appendix A.4 is no longer valid. Nevertheless, by Remark 18 we can replace $\eta_{sN^2}(1)$ (resp. $\eta_{sN^2}(N-1)$) by the average in a box around 1 (resp. $N-1$):

$$\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) := \frac{1}{\varepsilon N}\sum_{x=1}^{1+\varepsilon N}\eta_{sN^2}(x), \quad \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) := \frac{1}{\varepsilon N}\sum_{x=N-1}^{N-1-\varepsilon N}\eta_{sN^2}(x). \quad (49)$$

Here we note that the sum above goes from 1 to $1+\lfloor \varepsilon N \rfloor$ but for sake of simplicity we write $1+\varepsilon N$. By noting that

$$\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) \sim \rho_s(0) \quad (\text{resp. } \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) \sim \rho_s(1)),$$

for details on this approximation see for example [1,2] - and repeating the same arguments as above, we get to

$$\begin{aligned}\int_0^1 \rho_t(q)G(q) - \rho_0(q)G(q)dq - \int_0^t\int_0^1 \frac{1}{2}\Delta G(q)\rho_s(q)dqds \\ - \frac{1}{2}\int_0^t \partial_q G(0)\rho_s(0) - \partial_q G(1)\rho_s(1)ds \\ + \frac{\kappa}{2}\int_0^t G(0)(\alpha-\rho_s(0)) - G(1)(\beta-\rho_s(1))ds = 0.\end{aligned}$$

Again compare with the weak formulation given in (30) and note that the test function G does not depend on time.

The Case $\theta > 1$: This regime is quite similar to the previous one. We consider again an arbitrary function $G : [0,1] \to \mathbb{R}$ which is two times continuously differentiable and we note that the last two terms at the right hand side of (48) vanish since $\theta > 1$. Then, repeating the same arguments as in the previous section and noting that Remark 18 also applies to $\theta > 1$ we obtain at the end that

$$\begin{aligned}\int_0^1 \rho_t(q)G(q) - \rho_0(q)G(q)dq - \int_0^t\int_0^1 \frac{1}{2}\Delta G(q)\rho_s(q)dqds \\ - \frac{1}{2}\int_0^t \partial_q G(0)\rho_s(0) - \partial_q G(1)\rho_s(1)ds = 0.\end{aligned}$$

Again compare with the weak formulation given in (30) and note that the test function G does not depend on time.

Remark 6. Note that the parameter κ that appears in the boundary dynamics is only seen at the macroscopic level in the case $\theta = 1$ which corresponds to the heat equation with Robin boundary conditions.

2.9 Tightness

In this section we show that the sequence of probability measures $\{\mathbb{Q}_N\}_{N\geq 1}$, defined in the beginning of Sect. 2.7, is tight in the Skorohod space $\mathcal{D}([0,T],\mathcal{M}_+)$. In order to do that, we invoke the Aldous's criterium which says that

Lemma 2. *A sequence $\{P_N\}_{N\geq 1}$ of probability measures defined on $\mathcal{D}([0,T], \mathcal{M}_+)$ is tight if these two conditions hold:*

a. For every $t \in [0,T]$ and every $\varepsilon > 0$, there exists $K_\varepsilon^t \subset \mathcal{M}_+$ compact, such that

$$\sup_{N\geq 1} P_N\Big(\pi_t \notin K_\varepsilon^t\Big) \leq \varepsilon,$$

b. For every $\varepsilon > 0$

$$\lim_{\gamma\to 0} \limsup_{N\to\infty} \sup_{\substack{\tau\in\mathcal{T}_T \\ \theta\leq\gamma}} P_N\Big(d(\pi_{\tau+\theta}, \pi_\tau) > \varepsilon\Big) = 0,$$

where $\mathcal{T}_T$ denotes the set of stopping times with respect to the canonical filtration, bounded by T and d is the metric in the space $\mathcal{M}_+$ defined in (42).

By Proposition 1.7 of Chap. 4 in [15] it is enough to show that for every function G in a dense subset of $C([0,1])$, with respect to the uniform topology, the sequence of measures that corresponds to the real processes $\langle \pi_t^N, G\rangle$ is tight.

In our setting case, the first condition **a.** above translates by saying that:

$$\lim_{A\to+\infty} \lim_{N\to+\infty} \mathbb{P}_{\mu_N}\Big(|\langle \pi_t^N, G\rangle| > A\Big) = 0.$$

This is a consequence of Chebychev's inequality and the fact that for the exclusion type dynamics, the number of particles per site is at most one, we leave the details on this to the reader. So, it remains to show condition **b**. In this context and since we are considering the real process $\langle \pi_t^N, G\rangle$, the distance d above is the usual distance in $\mathbb{R}$. Then, we must show that for all $\varepsilon > 0$ and any function G in a dense subset of $C([0,1])$, with respect to the uniform topology, it holds that

$$\lim_{\delta\to 0} \limsup_{N\to\infty} \sup_{\tau\in\mathcal{T}_T, \bar\tau\leq\delta} \mathbb{P}_{\mu_N}\Big(\eta. : \Big|\langle \pi_{\tau+\bar\tau}^N, G\rangle - \langle \pi_\tau^N, G\rangle\Big| > \varepsilon\Big) = 0. \tag{50}$$

Above we assume that all the stopping times are bounded by T, thus, $\tau + \bar\tau$ should be understood as $(\tau + \bar\tau) \wedge T$.

Recall that it is enough to prove the assertion for functions G in a dense subset of $C([0,1])$ with respect to the uniform topology. We will use two different dense sets, namely the space $C^1([0,1])$ in the case $\theta < 1$ and the space $C^2([0,1])$ in the case $\theta \geq 1$, which are both dense in $C([0,1])$ with respect to the uniform topology. For that purpose, we split the proof according to $\theta \geq 1$ and $\theta < 1$. When $\theta \geq 1$ we prove (50) directly for functions $G \in C^2([0,1])$ and we conclude that the sequence is tight. For $\theta < 1$, we prove (50) first for functions $G \in C_c^2(0,1)$ and then we extend it, by a L^1 approximation procedure which is explained below, to functions $G \in C^1([0,1])$.

Recall from (48) that $M_t^N(G)$ is a martingale with respect to the natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Then

$$\begin{aligned}
&\mathbb{P}_{\mu_N}\Big(\eta_\cdot : \big|\langle \pi^N_{\tau+\bar\tau}, G\rangle - \langle \pi^N_\tau, G\rangle\big| > \varepsilon\Big)\\
=&\mathbb{P}_{\mu_N}\Big(\eta_\cdot : \Big|M^N_\tau(G) - M^N_{\tau+\bar\tau}(G) + \int_\tau^{\tau+\bar\tau} N^2\mathcal{L}_N\langle \pi^N_s, G\rangle ds\Big| > \varepsilon\Big)\\
\leq&\mathbb{P}_{\mu_N}\Big(\eta_\cdot : \Big|M^N_\tau(G) - M^N_{\tau+\bar\tau}(G)\Big| > \frac{\varepsilon}{2}\Big)\\
+&\mathbb{P}_{\mu_N}\Big(\eta_\cdot : \Big|\int_\tau^{\tau+\bar\tau} N^2\mathcal{L}_N\langle \pi^N_s, G\rangle ds\Big| > \frac{\varepsilon}{2}\Big).
\end{aligned}$$

Applying Chebychev's inequality (resp. Markov's inequality) in the first (resp. second) term on the right hand side of last inequality, we can bound the previous expression from above by

$$\frac{2}{\varepsilon^2}E_{\mathbb{P}_{\mu_N}}\Big[\Big(M^N_\tau(G) - M^N_{\tau+\bar\tau}(G)\Big)^2\Big] + \frac{2}{\varepsilon}E_{\mathbb{P}_{\mu_N}}\Big[\Big|\int_\tau^{\tau+\bar\tau} N^2\mathcal{L}_N\langle \pi^N_s, G\rangle ds\Big|\Big].$$

Therefore, in order to prove (50) it is enough to show that

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\tau\in\mathcal{T}_T,\bar\tau\leq\delta} E_{\mathbb{P}_{\mu_N}}\Big[\Big|\int_\tau^{\tau+\bar\tau} N^2\mathcal{L}_N\langle \pi^N_s, G\rangle ds\Big|\Big] = 0 \tag{51}$$

and

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\tau\in\mathcal{T}_T,\bar\tau\leq\delta} E_{\mathbb{P}_{\mu_N}}\Big[\Big(M^N_\tau(G) - M^N_{\tau+\bar\tau}(G)\Big)^2\Big] = 0. \tag{52}$$

Let us start by proving (51). Given a test function G, we will show that there exists a constant C such that

$$N^2\mathcal{L}_N(\langle \pi^N_s, G\rangle) \leq C \tag{53}$$

for any $s \leq T$. We start with the case $\theta \geq 1$. For that purpose, recall (47). Note that, since $|\eta_{sN^2}(x)| \leq 1$ for all $s \in [0,t]$ and since $G \in C^2([0,1])$, we have that

$$\Big|\langle \pi^N_s, \Delta_N G\rangle + \nabla_N^+ G(0)\eta_{sN^2}(1) - \nabla_N^- G(1)\eta_{sN^2}(N-1)\Big| \leq 2\|G''\|_\infty + 2\|G'\|_\infty$$

and

$$\begin{aligned}
\Big|\kappa N^{1-\theta}G\Big(\tfrac{1}{N}\Big)(\alpha - \eta_{sN^2}(1)) + \kappa N^{1-\theta}G\Big(\tfrac{N-1}{N}\Big)(\beta - \eta_{sN^2}(N-1))\Big|&\\
\leq 4\kappa N^{1-\theta}\|G\|_\infty&\\
\leq 4\kappa\|G\|_\infty.&
\end{aligned}$$

This proves (53) for the case $\theta \geq 1$. In the case $\theta < 1$, we take $G \in C_c^2([0,1])$ and we see that in this case (47) reduces to $\langle \pi^N_s, \frac{1}{2}\Delta_N G\rangle$ whose absolute value is bounded from above by $\|G''\|_\infty$ and this proves (53) for the case $\theta < 1$.

Let us now prove (52). Applying Dynkin's formula with $F(\cdot,\cdot)$ given by (46) we get that

$$\left(M_t^N(G)\right)^2 - \int_0^t N^2 \left[\mathcal{L}_N\langle\pi_s^N, G\rangle^2 - 2\langle\pi_s^N, G\rangle\mathcal{L}_N\langle\pi_s^N, G\rangle\right] ds, \tag{54}$$

is a martingale with respect to the natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$. A simple computation shows that

$$\begin{aligned} &N^2 \left[\mathcal{L}_{N,0}\langle\pi_s^N, G\rangle^2 - 2\langle\pi_s^N, G\rangle\mathcal{L}_{N,0}\langle\pi_s^N, G\rangle\right] \\ =&\frac{1}{2N^2}\sum_{x=1}^{N-2}(\eta_{sN^2}(x)-\eta_{sN^2}(x+1))^2(\nabla_N^+ G(\tfrac{x}{N}))^2 \end{aligned}$$

and by using the fact that $|\eta_{sN^2}(x)| \leq 1$ for all $s \in [0,t]$ last expression is bounded from above by $\frac{2}{N}\|G'\|_\infty^2$. On the other hand, we also have that

$$\begin{aligned} &N^2 \left[\mathcal{L}_{N,b}\langle\pi_s^N, G\rangle^2 - 2\langle\pi_s^N, G\rangle\mathcal{L}_{N,b}\langle\pi_s^N, G\rangle\right] \\ =&\frac{\kappa}{2N^\theta}\left[c_1(\eta_{sN^2},\alpha)G(\tfrac{1}{N})^2 + c_{N-1}(\eta_{sN^2},\beta)G(\tfrac{N-1}{N})^2\right] \end{aligned}$$

and by using the fact that $|\eta_{sN^2}(x)| \leq 1$ for all $s \in [0,t]$ last expression is bounded from above by $\frac{4\kappa}{N^\theta}\|G\|_\infty^2$.

This ends the proof of tightness in the case $\theta \geq 1$, since $C^2([0,1])$ is a dense subset of $C([0,1])$ with respect to the uniform topology. Nevertheless, for $\theta < 1$, since we considered functions $G \in C_c^2(0,1)$, last display is equal to zero. Therefore, we have proved (51) and (52), and thus (50), but for functions $G \in C_c^2(0,1)$ and, as mentioned above, we need to extend this result to functions in $C^1([0,1])$. To accomplish that, we take a function $G \in C^1([0,1]) \subset L^1([0,1])$, and we take a sequence of functions $\{G_k\}_{k\geq 0} \in C_c^2(0,1)$ converging to G, with respect to the L^1-norm, as $k \to \infty$. Now, since the probability in (50) is less or equal than

$$\begin{aligned} &\mathbb{P}_{\mu_N}\left(\eta_\cdot : \left|\langle\pi_{\tau+\bar\tau}^N, G_k\rangle - \langle\pi_\tau^N, G_k\rangle\right| > \frac{\varepsilon}{2}\right) \\ &+ \mathbb{P}_{\mu_N}\left(\eta_\cdot : \left|\langle\pi_{\tau+\bar\tau}^N, G-G_k\rangle - \langle\pi_\tau^N, G-G_k\rangle\right| > \frac{\varepsilon}{2}\right) \end{aligned}$$

and since G_k has compact support, from the computation above, it remains only to check that the last probability vanishes as $N \to \infty$ and then $k \to \infty$. For that purpose, we use the fact that

$$\left|\langle\pi_{\tau+\bar\tau}^N, G-G_k\rangle - \langle\pi_\tau^N, G-G_k\rangle\right| \leq \frac{2}{N}\sum_{x\in\Lambda_N}\left|(G-G_k)(\tfrac{x}{N})\right|, \tag{55}$$

and we use the estimate

$$\frac{1}{N}\sum_{x\in\Lambda_N}\left|(G-G_k)(\tfrac{x}{N})\right| \le \sum_{x\in\Lambda_N}\int_{\frac{x}{N}}^{\frac{x+1}{N}}\left|(G-G_k)(\tfrac{x}{N})-(G-G_k)(q)\right|dq$$
$$+\int_0^1|(G-G_k)(q)|dq$$
$$\le \frac{1}{N}\|(G-G_k)'\|_\infty+\int_0^1|(G-G_k)(q)|dq.$$

The result follows by first taking $N\to\infty$ and then $k\to\infty$.

2.10 The Limit Point

Here, we prove at first that all limit points $\mathbb{Q}$ of the sequence $\{\mathbb{Q}_N\}_{N\ge1}$ are concentrated on measures absolutely continuous with respect to the Lebesgue measure, that are equal to $g(q)dq$ at the initial time and finally that $\mathbb{Q}$ is concentrated on trajectories of measures satisfying $\pi_t(dq)=\rho_t(q)dq$, where $\rho_t(\cdot)$ is the weak solution of the corresponding hydrodynamic equation. Let $\mathbb{Q}$ be a limit point of $\{\mathbb{Q}_N\}_{N\ge1}$.

Characterization of Absolutely Continuity: We start by showing that $\mathbb{Q}$ is concentrated on measures which are absolutely continuous with respect to the Lebesgue measure. Fix a continuous function $G:[0,1]\to\mathbb{R}$. Since

$$\sup_{t\in[0,T]}|\langle\pi_t^N,G\rangle|\le\frac{1}{N}\sum_{x\in\Lambda_N}|G(\tfrac{x}{N})|,$$

which is a consequence of the fact of having at most one particle per site, the function that associates to each trajectory $\pi_\cdot$, $\sup_{t\in[0,T]}|\langle\pi_t,G\rangle|$ is continuous. As a consequence, all limit points are concentrated in trajectories π_t such that

$$|\langle\pi_t,G\rangle|\le\int_0^1|G(q)|dq.$$

In order to show that the measure π_t is absolutely continuous with respect to the Lebesgue measure, that we denote by Leb, we have to show that for each set A such that $Leb(A)=0$, then $\pi_t(A)=0$. With this purpose, we use last estimate for a sequence of continuous functions $\{G_N\}_{N\ge1}$ that converges to the indicator function over the set A and the result follows. Concluding, we have just proved that

$$\mathbb{Q}\Big(\pi_\cdot:\pi_t(dq)=\pi(t,q)dq,\forall t\in[0,T]\Big)=1$$

i.e. $\pi_t(dq)$ is absolutely continuous with respect to the Lebesgue measure with a density $\pi(t,q)$.

Characterization of the Initial Measure: Here we show that $\mathbb{Q}$ is concentrated on a Dirac measure equal to $g(q)dq$ at time 0. For that purpose, fix $\varepsilon > 0$. From the results of Sect. 2.9, we know, from the weak convergence over a subsequence and Portmanteau's Theorem, that:

$$\begin{aligned}
&\mathbb{Q}\Big(\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G(\tfrac{x}{N})\eta_0(x) - \int_0^1 G(q)g(q)dq\Big| > \varepsilon\Big)\\
&\leq \liminf_{K\to+\infty} \mathbb{Q}_{N_k}\Big(\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G(\tfrac{x}{N})\eta_0(x) - \int_0^1 G(q)g(q)dq\Big| > \varepsilon\Big)\\
&= \liminf_{K\to+\infty} \mu_{N_k}\Big(\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G(\tfrac{x}{N})\eta(x) - \int_0^1 G(q)g(q)dq\Big| > \varepsilon\Big).
\end{aligned}$$

This last limit is equal to zero, by the hypothesis of μ_N being associated to the profile $g(\cdot)$, see Definition 5. This shows that

$$\mathbb{Q}\Big(\pi_\cdot : \pi_0(dq) = g(q)dq\Big) = 1.$$

Characterization of the Density $\pi(t,q)$: Up to here we know that all limit points $\mathbb{Q}$ of the sequence sequence $\{\mathbb{Q}_N\}_{N\geq 1}$ are concentrated on trajectories $\pi_t(dq)$ which are absolutely continuous with respect to the Lebesgue measure, that is, $\pi_t(dq) = \pi(t,q)dq$. Moreover, we also know that all limit points $\mathbb{Q}$ of the sequence $\{\mathbb{Q}_N\}_{N\geq 1}$ are such that the initial trajectory is a Dirac measure equal to $g(q)dq$. Now we prove that all limit points are concentrated on trajectories of measures of the form $\rho_t(q)dq$, that is we are going to show that $\pi(t,q) = \rho_t(q)$ and that $\rho_t(\cdot)$ is a weak solution of the corresponding hydrodynamic equation. For that purpose, let $\mathbb{Q}$ be a limit point of the sequence $\{\mathbb{Q}_N\}_{N\geq 1}$, whose existence follows from the computations of Sect. 2.9 and assume, without lost of generality, that $\{\mathbb{Q}_N\}_{N\geq 1}$ converges to $\mathbb{Q}$, as $N \to +\infty$.

Proposition 2. *If $\mathbb{Q}$ is a limit point of $\{\mathbb{Q}_N\}_{N\in\mathbb{N}}$ then*

$$\mathbb{Q}\left(\pi_\cdot : F_\theta = 0, \forall t \in [0,T],\, \forall G \in C_\theta\right) = 1,$$

where

$$F_\theta = \begin{cases} F^c_{Dir}, \text{ if } \theta < 0,\\ F_{Dir}, \text{ if } \theta \in [0,1),\\ F_{Rob}, \text{ if } \theta \geq 1, \end{cases} \quad \text{and} \quad C_\theta = \begin{cases} C_c^{1,2}([0,T]\times[0,1]), \text{ if } \theta < 0,\\ C_0^{1,2}([0,T]\times[0,1]), \text{ if } \theta \in [0,1),\\ C^{1,2}([0,T]\times[0,1]), \text{ if } \theta \geq 1. \end{cases}$$

Proof. We consider the case $\theta \geq 1$. Note that we need to verify, for $\delta > 0$ and $G \in C^{1,2}([0,T]\times[0,1])$, that

$$\mathbb{Q}\left(\pi_\cdot \in \mathcal{D}([0,T],\mathcal{M}^+) : \sup_{0\leq t\leq T} |F_{Rob}| > \delta\right) = 0, \tag{56}$$

Recall F_{Rob} from (32) and note that, due to the terms that involve $\rho_s(1)$ and $\rho_s(0)$ and that appear in F_{Rob}, the set inside the probability in (56) is not an open set in the Skorohod space, and as a consequence we cannot use directly Portmanteau's Theorem. To avoid this difficulty, we fix $\varepsilon > 0$ and we consider two approximations of the identity given by

$$\iota_\varepsilon^0(q) = \frac{1}{\varepsilon} 1_{(0,\varepsilon)}(q) \quad \text{and} \quad \iota_\varepsilon^1(q) = \frac{1}{\varepsilon} 1_{(1-\varepsilon,1)}(q) \tag{57}$$

and we sum and subtract to $\rho_s(0)$ and to $\rho_s(1)$ the mean

$$\langle \pi_s, \iota_\varepsilon^0 \rangle = \tfrac{1}{\varepsilon} \int_0^\varepsilon \rho_s(q) dq \quad \text{and} \quad \langle \pi_s, \iota_\varepsilon^1 \rangle = \tfrac{1}{\varepsilon} \int_{1-\varepsilon}^\varepsilon \rho_s(q) dq, \tag{58}$$

respectively. Above we used the fact that $\mathbb{Q}$ is concentrated on trajectories $\pi_t(dq)$ which are absolutely continuous with respect to the Lebesgue measure: $\pi_t(dq) = \rho_t(q)dq$. Thus, we bound the probability in (56) from above by the sum of the following terms

$$\begin{aligned} &\mathbb{Q}\Big(\sup_{0\le t\le T} \Big| \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 \rho_0(q) G_0(q)\, dq \\ &- \int_0^t \int_0^1 \rho_s(q) \Big(\frac{1}{2}\Delta + \partial_s \Big) G_s(q)\, dq ds - \frac{\kappa}{2} \int_0^t G_s(0)\alpha + G_s(1)\beta\, ds \\ &+ \frac{1}{2} \int_0^t \langle \pi_s, \iota_\varepsilon^1 \rangle \Big(\partial_q G_s(1) + \kappa G_s(1) \Big)\, ds - \frac{1}{2} \int_0^t \langle \pi_s, \iota_\varepsilon^0 \rangle \Big(\partial_q G_s(0) - \kappa G_s(0) \Big)\, ds \Big| > \frac{\delta}{4} \Big), \end{aligned} \tag{59}$$

$$\mathbb{Q}\left(\Big| \int_0^1 (\rho_0(q) - g(q)) G_0(q)\, dq \Big| > \frac{\delta}{4} \right), \tag{60}$$

$$\sum_{j\in\{0,1\}} \mathbb{Q}\left(\sup_{0\le t\le T} \Big| \frac{1}{2} \int_0^t (\rho_s(j) - \langle \pi_s, \iota_\varepsilon^j \rangle) \left[\kappa G_s(j) - \partial_q G_s(j) \right] ds \Big| > \frac{\delta}{4} \right), \tag{61}$$

and we note that the terms in (61) converge to 0 as $\varepsilon \to 0$ since we are comparing $\rho_s(0)$ and $\rho_s(1)$ with the averages (58) around 0 and 1, respectively. Moreover, (60) is equal to zero since $\mathbb{Q}$ is a limit point of $\{\mathbb{Q}_N\}_{N\ge 1}$ and $\mathbb{Q}_N$ is induced by a measure μ_N which is associated to the profile $g(\cdot)$. Note that in (59) we still cannot use Portmanteau's Theorem, since the functions ι_ε^0 and ι_ε^1 are not continuous. Nevertheless, by approximating each one of these functions by continuous functions in such a way that the error vanishes as $\varepsilon \to 0$ then, from Proposition A.3 of [10] we can use Portmanteau's Theorem and bound (59) from above by

$$\begin{aligned} &\liminf_{N\to\infty} \mathbb{Q}_N\Big(\sup_{0\le t\le T} \Big| \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 \rho_0(q) G_0(q)\, dq \\ &- \int_0^t \int_0^1 \rho_s(q) \Big(\frac{1}{2}\Delta + \partial_s \Big) G_s(q)\, dq ds - \frac{\kappa}{2} \int_0^t G_s(0)\alpha + G_s(1)\beta\, ds \\ &- \frac{1}{2} \int_0^t \langle \pi_s, \iota_\varepsilon^0 \rangle \Big(\partial_q G_s(0) - \kappa G_s(0) \Big)\, ds + \frac{1}{2} \int_0^t \langle \pi_s, \iota_\varepsilon^1 \rangle \Big(\partial_q G_s(1) + \kappa G_s(1) \Big)\, ds \Big| > \frac{\delta}{2^4} \Big). \end{aligned} \tag{62}$$

Summing and subtracting $\int_0^t N^2 \mathcal{L}_N \langle \pi_s^N, G_s \rangle ds$ to the term inside the supremum in (62), recalling (48) and (49), the definition of $\mathbb{Q}_N$, we bound (62) from above by the sum of the next two terms

$$\liminf_{N\to\infty} \mathbb{P}_{\mu_N} \left(\sup_{0\le t\le T} \left| M_t^N(G) \right| > \frac{\delta}{2^5} \right), \tag{63}$$

and

$$\begin{aligned} &\liminf_{N\to\infty} \mathbb{P}_{\mu_N} \Big(\sup_{0\le t\le T} \Big| \int_0^t N^2 \mathcal{L}_N \langle \pi_s^N, G_s \rangle \, ds - \int_0^t \int_0^1 \rho_s(q) \frac{1}{2} \Delta G_s(q) \, dq ds \\ &- \frac{\kappa}{2} \int_0^t \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) \Big(\partial_q G_s(0) - \kappa G_s(0) \Big) \, ds + \frac{1}{2} \int_0^t \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) \Big(\partial_q G_s(1) + \kappa G_s(1) \Big) \, ds \\ &- \frac{\kappa}{2} \int_0^t G_s(0)\alpha + G_s(1)\beta \, ds \Big| > \frac{\delta}{2^5} \Big). \end{aligned} \tag{64}$$

Doob's inequality together with the computations right below (54) show that (63) goes to 0 as $N \to \infty$. Finally, (64) can be rewritten as

$$\begin{aligned} &\liminf_{N\to\infty} \mathbb{P}_{\mu_N} \Big(\sup_{0\le t\le T} \Big| \int_0^t N^2 \mathcal{L}_N \langle \pi_s^N, G_s \rangle \, ds - \int_0^t \langle \pi_s^N, \frac{1}{2} \Delta G_s \rangle \, ds \\ &- \frac{1}{2} \int_0^t \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) \Big(\partial_q G_s(0) - \kappa G_s(0) \Big) \, ds + \frac{1}{2} \int_0^t \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) \Big(\partial_q G_s(1) + \kappa G_s(1) \Big) \, ds \\ &- \frac{\kappa}{2} \int_0^t G_s(0)\alpha + G_s(1)\beta \, ds \Big| > \frac{\delta}{2^5} \Big). \end{aligned} \tag{65}$$

Now, from (47) we can bound from above the probability in (65) by the sum of the following terms

$$\mathbb{P}_{\mu_N} \left(\sup_{0\le t\le T} \Big| \frac{1}{N} \int_0^t \sum_{x\in\Lambda_N} \frac{1}{2} \Delta_N G_s(\tfrac{x}{N}) \eta_{sN^2}(x) ds - \int_0^t \left\langle \pi_s^N, \frac{1}{2} \Delta G_s \right\rangle ds \Big| > \frac{\delta}{2^6} \right), \tag{66}$$

$$\mathbb{P}_{\mu_N} \left(\sup_{0\le t\le T} \Big| \frac{1}{2} \int_0^t \nabla_N^+ G_s(0) \eta_{sN^2}(1) - \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) \partial_q G_s(0) \, ds \Big| > \frac{\delta}{2^6} \right), \tag{67}$$

and

$$\mathbb{P}_{\mu_N} \left(\sup_{0\le t\le T} \Big| \frac{\kappa}{2} \int_0^t N^{1-\theta} G_s\Big(\tfrac{1}{N}\Big)(\alpha - \eta_{sN^2}(1)) - G_s(0)(\alpha - \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)) ds \Big| > \frac{\delta}{2^6} \right) \tag{68}$$

and two other terms which are very similar to the two previous ones but related to the action of the right boundary dynamics given by $\mathcal{L}_{N,b}^{N-1}$. Applying a Taylor expansion on the test function G it is easy to show that (66) goes to 0 as $N \to \infty$. Also by Taylor expansion, (67) can be bounded from above by

$$\mathbb{P}_{\mu_N} \left(\sup_{0\le t\le T} \Big| \int_0^t \partial_q G_s(0) (\eta_{sN^2}(1) - \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)) ds \Big| > \frac{\delta}{2^8} \right), \tag{69}$$

plus a term that vanishes as $N \to \infty$. Using Lemma 7 we see that (69) vanishes as $N \to \infty$. The term (68) can be estimated using exactly the same argument that we just used, that is: Taylor expansion on G plus Lemma 7. For the terms related to the right boundary the argument is the same and with this we finish the proof.

We leave the other case, namely $\theta < 1$ for the reader. This case is even simpler than the previous one and for the interested reader we refer to, for example, [1,2].

2.11 Hydrostatic Limit

In this section we prove that the hydrodynamic limit holds when we start the system from the stationary measure μ_{ss}, see Sect. 2.4. By looking at the statement of Theorem 1 we see that, in fact, to conclude we only need to show the next result.

Proposition 3. *Let μ_{ss} be the stationary measure for the Markov process $\{\eta_{tN^2} : t \geq 0\}$ with generator $N^2\mathcal{L}_N$. Then, μ_{ss} is associated to the profile $\bar{\rho} : [0,1] \to [0,1]$ given on $q \in (0,1)$ by (22), that is*

$$\bar{\rho}(q) = \begin{cases} (\beta - \alpha)q + \alpha \,;\, \theta < 1, \\ \frac{\kappa(\beta-\alpha)}{2+\kappa} q + \alpha + \frac{\beta-\alpha}{2+\kappa} \,;\, \theta = 1, \\ \frac{\beta+\alpha}{2} \,;\, \theta > 1, \end{cases}$$

which is a stationary solution of the corresponding hydrodynamic equation, see (36) and (37).

Proof. Recall from (44), that we need to prove:

$$\lim_{N\to\infty} \mu_{ss}\Big(\eta \in \Omega_N : \Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\eta(x) - \int_0^1 G(q)\bar{\rho}(q)dq\Big| > \delta\Big) = 0$$

for any continuous function $G : [0,1] \to \mathbb{R}$. By Markov's and triangular inequalities, we bound the previous probability from above by

$$\begin{aligned} &\frac{1}{\delta} E_{\mathbb{P}_{\mu_{ss}}}\Big[\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\Big(\eta(x) - \rho^N_{ss}(x)\Big)\Big| \\ &+\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\rho^N_{ss}(x) - \int_0^1 G(q)\bar{\rho}(q)dq\Big|\Big] \\ &\leq \frac{1}{\delta} E_{\mathbb{P}_{\mu_{ss}}}\Big[\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\Big(\eta(x) - \rho^N_{ss}(x)\Big)\Big|\Big] \\ &+\frac{1}{\delta}\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\rho^N_{ss}(x) - \int_0^1 G(q)\bar{\rho}(q)dq\Big|. \end{aligned} \tag{70}$$

The last term can be bounded from above by

$$\frac{1}{\delta}\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\Big(\rho^N_{ss}(x)-\bar\rho\Big(\tfrac{x}{N}\Big)\Big)\Big|$$
$$+\frac{1}{\delta}\Big|\frac{1}{N}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\bar\rho\Big(\tfrac{x}{N}\Big)-\int_0^1 G(q)\bar\rho(q)dq\Big|.$$

The term at the left hand side of last expression is bounded from above by

$$\frac{1}{\delta}\frac{1}{N}\sum_{x\in\Lambda_N}\Big|G\left(\tfrac{x}{N}\right)\Big|\Big|\rho^N_{ss}(x)-\bar\rho\Big(\tfrac{x}{N}\Big)\Big|\le\frac{\|G\|_\infty}{\delta}\max_{x\in\Lambda_N}\Big|\rho^N_{ss}(x)-\bar\rho\Big(\tfrac{x}{N}\Big)\Big|$$

where from (21) it vanishes as $N\to\infty$, while the term at the right hand side also vanishes as $N\to\infty$ since we compare the Riemann sum with the corresponding converging integral.

To finish the proof it remains to analyse the third term in (70). By the Cauchy-Schwarz's inequality the expectation appearing in that term can bounded from above by

$$\Big(\Big|\frac{1}{N^2}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)E_{\mathbb{P}_{\mu_{ss}}}\Big[(\eta(x)-\rho^N_{ss}(x))^2\Big]$$
$$+\frac{2}{N}\sum_{x<y} G\left(\tfrac{x}{N}\right)G\left(\tfrac{y}{N}\right)E_{\mathbb{P}_{\mu_{ss}}}\Big[(\eta(x)-\rho^N_{ss}(x))(\eta(y)-\rho^N_{ss}(y))\Big]\Big)^{\frac12}$$
$$\le\Big(\frac{C\|G\|_\infty}{N}+2\|G\|_\infty\max_{x<y}\varphi^N_{ss}(x,y)\Big)^{\frac12}.$$

From (17) the previous expression vanishes as $N\to\infty$. This finishes the proof.

Note that the proof presented above uses the information about the two point correlation function which is not always easy to obtain. We refer the reader to [8] for another proof of this results without using the knowledge on the correlations.

3 Symmetric Exclusion with Long Jumps in Contact with Reservoirs

3.1 The Model

In this section we want to generalize the results of the previous section to the case where particles can give jumps arbitrarily large. As in the previous section, the bulk consists in the set of points $\Lambda_N=\{1,...,N-1\}$ and we artificially add two end points $x=0$ and $x=N$. Now, we explain the dynamics of the models we consider and we start by describing the conditions on the jump rate. For that purpose, let $p:\mathbb{Z}\times\mathbb{Z}\to[0,1]$ be a transition probability such that $p(x,y)=p(y-x)$ and which is symmetric. We are going to discuss two cases:

the first one, when $p(\cdot)$ has finite variance and the second one when $p(\cdot)$ has infinite variance. Note that since $p(\cdot)$ is symmetric it has mean zero, that is:

$$\sum_{z\in\mathbb{Z}} zp(z) = 0.$$

We denote $m = \sum_{z\geq 1} zp(z)$. As an example we consider $p(\cdot)$ given by $p(0) = 0$ and

$$p(z) = \frac{c_\gamma}{|z|^{\gamma+1}}, \tag{71}$$

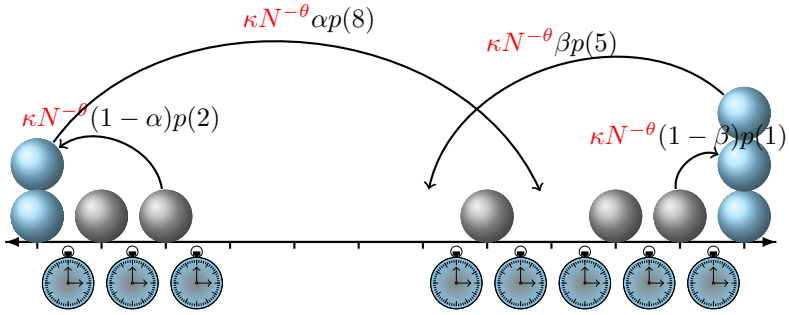
for $z \neq 0$, where c_γ is a normalizing constant. For simplicity of the presentation we stick to this choice of $p(\cdot)$ along this section but we note that many of our results are true, in the case where $p(\cdot)$ has finite variance, in a more general setting where we only assume $p(\cdot)$ to be translation invariant and mean zero.

We consider the process in contact with stochastic reservoirs at the left and the right of the bulk. We fix four parameters $\alpha, \beta \in [0,1]$, $\kappa > 0$ and $\theta \in \mathbb{R}$, so that particles can get in the bulk of the system from the site $x = 0$ to any site $y \in \Lambda_N$ at rate $\alpha\kappa N^{-\theta}p(y)$ or leave the bulk from any site $y \in \Lambda_N$ to the site $x = 0$ at rate $(1-\alpha)\kappa N^{-\theta}p(y)$; and particles can get in the bulk to any site $y \in \Lambda_N$ from the site $x = N$ at rate $\beta\kappa N^{-\theta}p(N-y)$ or leave the bulk from any site $y \in \Lambda_N$ to the site $x = N$ at rate $(1-\beta)\kappa N^{-\theta}p(N-y)$.

We define the dynamics of the process in the following way. We start with the bulk dynamics. Each pair of sites of the bulk $\{x, y\} \subset \Lambda_N$ carries a Poisson process of intensity $p(y-x)/2$. Poisson processes associated to different bonds are independent. If for the configuration η, the clock associated to the bound $\{x, y\}$ rings, then we exchange the value of the occupation variables $\eta(x)$ and $\eta(y)$ at rate $p(y-x)/2$. Now we explain the dynamics at the boundary. Each pair of sites $\{0, x\}$ with $x \in \Lambda_N$ carries two Poisson processes, all of them being independent. If for the configuration η, the clock associated to the Poisson process of the oriented bond $\{0, x\}$ (resp. $\{x, 0\}$) rings, then we change the value $\eta(x)$ into $1-\eta(x)$ with rate $\kappa N^{-\theta}p(x)\alpha(1-\eta(x))$ (resp. $\kappa N^{-\theta}p(x)(1-\alpha)\eta(x)$). At the right boundary the dynamics is similar but instead of α the intensity is given by β. Observe that the reservoirs ($x = 0$ and $x = N$) add and remove particles on all the sites of the bulk Λ_N, and not only at the boundaries $x = 1$ and $x = N-1$ as happened in the model of Sect. 2, but with a rate that decreases as the distance from the corresponding reservoir increases. We remark that as in the previous section, we could do another interpretation of the previous dynamics at the boundary, as follows. Particles can either be created or annihilated at any site $x \in \Lambda_N$ according to the following rates:

- from the left reservoir, from $x = 0$ to $y \in \Lambda_N$:
  - creation rate: $\alpha\kappa N^{-\theta}p(y)$,
  - annihilation rate: $(1-\alpha)\kappa N^{-\theta}p(y)$.
- from the right reservoir, from $x = N-1$ to $y \in \Lambda_N$:
  - creation rate: $\beta\kappa N^{-\theta}p(N-y)$,
  - annihilation rate: $(1-\beta)\kappa N^{-\theta}p(N-y)$.

Let us see an illustration of the dynamics just described with $N = 11$ and the configuration $\eta = (1, 1, 0, 0, 0, 0, 1, 0, 1, 1)$:



The infinitesimal generator of the process is given by

$$\mathcal{L}_N = \mathcal{L}_{N,0} + \mathcal{L}_{N,b}, \tag{72}$$

where $\mathcal{L}_{N,0}$ and $\mathcal{L}_{N,b}$ act on functions $f : \Omega_N \to \mathbb{R}$ as

$$\begin{aligned} (\mathcal{L}_{N,0} f)(\eta) &= \frac{1}{2} \sum_{x,y \in \Lambda_N} p(x-y)[f(\eta^{x,y}) - f(\eta)], \\ (\mathcal{L}_{N,b} f)(\eta) &= \frac{\kappa}{N^\theta} \sum_{y \in \{0,N\}} \sum_{x \in \Lambda_N} p(y-x) c_x(\eta, r(y))[f(\eta^x) - f(\eta)] \end{aligned} \tag{73}$$

where the configurations $\eta^{x,y}$ and η^x have been defined in (3), the rates $c_x(\eta, r(y))$ have been defined in (4) and $r(0) = \alpha$ and $r(N) = \beta$.

We consider the Markov process speeded up in the time scale $t\Theta(N)$ and note that $\{\eta_{t\Theta(N)} : t \geq 0\}$ has infinitesimal generator given by $\Theta(N)\mathcal{L}_N$. Although $\eta_{t\theta(N)}$ depends on α, β and θ, we shall omit these index in order to simplify notation.

As in Sect. 2.4 we can prove that the Bernoulli product measures ν_ρ^N as defined in (7) are reversible when we consider $\alpha = \beta = \rho$. The proof is quite similar to the one given in Lemma 1 and for that reason it is omitted.

In the next section we analyse the case where $p(\cdot)$ has finite variance and we denote it by σ^2, so that

$$\sigma^2 := \sum_{z \in \mathbb{Z}} z^2 p(z) < \infty.$$

As an example we consider $p(\cdot)$ as in (71), that is $p(0) = 0$ and

$$p(z) = \frac{c_\gamma}{|z|^{\gamma+1}},$$

for $z \neq 0$, where c_γ is a normalizing constant and we take $\gamma > 2$, so that $p(\cdot)$ has finite variance. For simplicity of the presentation we stick to this choice of $p(\cdot)$ whenever we mention to the case where $p(\cdot)$ has finite variance but we note that many of our results are true in the more general setting where we just assume $p(\cdot)$ to be translation invariant, mean zero and with finite variance.

Remark 7. We note that for the choice of p with $p(1) = \frac{1}{2} = p(-1)$ the dynamics described above coincides with the one of the first section. In that sense many of the results that we will derive here are a generalization of those obtained before.

In Sect. 3.3 we analyse the case where $p(\cdot)$ is as in (71) but we consider $\gamma \in (1,2)$ so that $p(\cdot)$ has mean zero but infinite variance. We note that in the case $\gamma = 2$ the transition probability $p(\cdot)$ also has mean zero and infinite variance, but in this case the results are similar to those when $p(\cdot)$ has finite variance, see Remark 12.

3.2 The Finite Variance Case

Hydrodynamic Equations: Recall the notation introduced in Sect. 2.6. We can now give the definition of the weak solutions of the hydrodynamic equations that will be derived in this section when $p(\cdot)$ is assumed to have finite variance. In what follows $g : [0,1] \to [0,1]$ is a measurable function and it is the initial condition of all the partial differential equations that we define below, that is $\rho_0(q) = g(q)$, for all $q \in (0,1)$.

Definition 6. *Let $\hat{\sigma} \geq 0$ and $\hat{\kappa} \geq 0$ be some parameters. We say that $\rho : [0,T] \times [0,1] \to [0,1]$ is a weak solution of the reaction-diffusion equation with Dirichlet boundary conditions*

$$\begin{cases} \partial_t \rho_t(q) = \frac{\hat{\sigma}^2}{2} \Delta \rho_t(q) + \hat{\kappa} \Big\{ \frac{\alpha - \rho_t(q)}{q^{\gamma+1}} + \frac{\beta - \rho_t(q)}{(1-q)^{\gamma+1}} \Big\}, \quad (t,q) \in (0,T] \times (0,1), \\ \rho_t(0) = \alpha, \quad \rho_t(1) = \beta, \quad t \in (0,T], \end{cases} \tag{74}$$

starting from a measurable function $g : [0,1] \to [0,1]$, if the following three conditions hold:

1. – $\rho \in L^2(0,T;\mathcal{H}^1)$ *if* $\hat{\sigma} > 0$,
   – $\int_0^T \int_0^1 \Big\{ \frac{(\alpha - \rho_t(q))^2}{q^{\gamma+1}} + \frac{(\beta - \rho_t(q))^2}{(1-q)^{\gamma+1}} \Big\} dq\, dt < \infty$ *if* $\hat{\kappa} > 0$,
2. *ρ satisfies the weak formulation:*

$$\begin{aligned} F_{RD} &:= \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq \\ &- \int_0^t \int_0^1 \rho_s(q) \Big(\frac{\hat{\sigma}^2}{2} \Delta + \partial_s \Big) G_s(q)\, dq\, ds \\ &- \hat{\kappa} \int_0^t \int_0^1 G_s(q) \left(\frac{\alpha - \rho_s(q)}{q^{\gamma+1}} + \frac{\beta - \rho_s(q)}{(1-q)^{\gamma+1}} \right) dq\, ds = 0, \end{aligned} \tag{75}$$

   for all $t \in [0,T]$ and any function $G \in C_c^{1,2}([0,T] \times [0,1])$,
3. *if $\hat{\sigma} > 0$ then $\rho_t(0) = \alpha, \quad \rho_t(1) = \beta$ for all $t \in (0,T]$.*

Remark 8. Observe that in the case $\hat{\sigma} > 0$ and $\hat{\kappa} = 0$ we recover the heat equation with Dirichlet boundary conditions. If $\hat{\sigma} = 0$ the equation does not have the diffusion term.

Definition 7. *Let $\hat{\sigma} > 0$ and $\hat{m} \geq 0$ be some parameters. We say that $\rho : [0,T] \times [0,1] \to [0,1]$ is a weak solution of the heat equation with Robin boundary conditions*

$$\begin{cases} \partial_t \rho_t(q) = \frac{\hat{\sigma}^2}{2} \Delta \rho_t(q), \quad (t,q) \in [0,T] \times (0,1), \\ \partial_q \rho_t(0) = \frac{2\hat{m}}{\hat{\sigma}^2}(\rho_t(0) - \alpha), \quad \partial_q \rho_t(1) = \frac{2\hat{m}}{\hat{\sigma}^2}(\beta - \rho_t(1)), \quad t \in [0,T], \end{cases} \tag{76}$$

starting from a measurable function $g : [0,1] \to [0,1]$, if the following two conditions hold:

1. $\rho \in L^2(0,T;\mathcal{H}^1)$,
2. *ρ satisfies the weak formulation:*

$$\begin{aligned} F_{Rob} := & \int_0^1 \rho_t(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq \\ & - \int_0^t \int_0^1 \rho_s(q) \Big(\frac{\hat{\sigma}^2}{2} \Delta + \partial_s \Big) G_s(q)\, dq\, ds \\ & + \frac{\hat{\sigma}^2}{2} \int_0^t \{ \rho_s(1) \partial_q G_s(1) - \rho_s(0) \partial_q G_s(0) \}\, ds \\ & - \hat{m} \int_0^t \{ G_s(0)(\alpha - \rho_s(0)) + G_s(1)(\beta - \rho_s(1)) \}\, ds = 0, \end{aligned} \tag{77}$$

for all $t \in [0,T]$, any function $G \in C^{1,2}([0,T] \times [0,1])$.

Remark 9. Observe that in the case $\hat{m} = 0$ the equation above is the heat equation with Neumann boundary conditions.

Hydrodynamic Limit: Recall the notion of the empirical measure given in Sect. 2.6 and note that in this case we have

$$\pi_t^N(\eta, dq) := \pi^N(\eta_{t\theta(N)}, dq)$$

and we note that, in this case, the time scale $\theta(N)$ will change with the range of θ, contrarily to what happens in the model of Sect. 2. As before, let $\mathbb{P}_{\mu_N}$ be the probability measure in the Skorohod space $\mathcal{D}([0,T], \Omega_N)$ induced by the Markov process $\{\eta_{t\theta(N)} : t \geq 0\}$ and the initial probability measure μ_N and we denote by $E_{\mathbb{P}_{\mu_N}}$ the expectation with respect to $\mathbb{P}_{\mu_N}$. Let $\{\mathbb{Q}_N\}_{N \geq 1}$ be the sequence of probability measures on $\mathcal{D}([0,T], \mathcal{M}^+)$ induced by the Markov process $\{\pi_t^N ; t \geq 0\}$ and by $\mathbb{P}_{\mu_N}$.

Remark 10. We note that due to the presence of long jumps in the system, we cannot obtain information about the empirical profile nor the two point correlation function in a simple way as we did in Sect. 2.5. We also note that the matrix ansatz method described in Sect. 2.4 in this case does not give us any information about the stationary measures for this model. This study is left for a future work.

Let $g : [0,1] \to [0,1]$ be a measurable function and let $\{\mu_N\}_{N\geq 1}$ be a sequence of probability measures in Ω_N associated to $g(\cdot)$, see (44). The first result in this section is stated in the following theorem (see Fig. 7).

Theorem 3. *Let $g : [0,1] \to [0,1]$ be a measurable function and let $\{\mu_N\}_{N\geq 1}$ be a sequence of probability measures in Ω_N associated to $g(\cdot)$. Then, for any $0 \leq t \leq T$,*

$$\lim_{N\to\infty} \mathbb{P}_{\mu_N}\Big(\eta_\cdot : \Big|\frac{1}{N-1}\sum_{x\in\Lambda_N} G\left(\tfrac{x}{N}\right)\eta_{t\theta(N)}(x) - \int_0^1 G(q)\rho_t(q)dq\Big| > \delta\Big) = 0,$$

where the time scale is given by

$$\Theta(N) = \begin{cases} N^2, & \text{if } \theta \geq 1-\gamma, \\ N^{\gamma+\theta+1}, & \text{if } \theta < 1-\gamma, \end{cases} \tag{78}$$

and $\rho_t(\cdot)$ is the unique weak solution of:

- (74) *with $\hat{\sigma} = 0$ and $\hat{\kappa} = \kappa c_\gamma$, if $\theta < 1-\gamma$;*
- (74) *with $\hat{\sigma} = \sigma$ and $\hat{\kappa} = \kappa c_\gamma$, if $\theta = 1-\gamma$;*
- (74) *with $\hat{\sigma} = \sigma$ and $\hat{\kappa} = 0$, if $\theta \in (1-\gamma, 1)$;*
- (76) *with $\hat{\sigma} = \sigma$ and $\hat{m} = \frac{\kappa}{2}$, if $\theta = 1$;*
- (76) *with $\hat{\sigma} = \sigma$ and $\hat{m} = 0$, if $\theta > 1$.*

Remark 11. We note that for a transition probability $p(\cdot)$ which is symmetric and with finite variance the last three regimes obtained above are in force (however (74) with $\hat{\kappa} = 0$ is obtained for $\theta \in [0,1)$). We note that the two first regimes depend on the specific choice of the transition probability $p(\cdot)$ that we have assumed in (71). We also note that if we impose that the higher moments of $p(\cdot)$ are finite then the regime (74) with $\hat{\kappa} = 0$ can be reached for $\theta \in [v,1)$ where $v < 0$ depends on the finiteness of the moments of $p(\cdot)$.

Remark 12. Despite, in the case $\gamma = 2$, the transition probability $p(\cdot)$ has infinite variance, we obtain a very similar behavior to the one described above but the time scale that one has to consider is $N^2/\log(N)$ instead of N^2. We leave the adaptation of the proof in this case as an exercise to the reader.

Remark 13. We note that the solution of the hydrodynamic equation depends on the parameter κ which appears at the boundary dynamics in two different regimes of θ, namely $\theta = 1-\gamma$ and $\theta = 1$.

Now note that as before, the stationary solutions of the hydrodynamic limits in the case $\theta > 1-\gamma$ are standard and for that reason they are ommited. On the other hand, the form and properties of the stationary solutions in the case $\theta \leq 1-\gamma$ are more complicated to obtain in the case $\theta = 1-\gamma$. This problem is studied in more details in [14] for a slightly different dynamics. In Fig. 6 we only present some graphs of the stationary solutions and refer the interested reader

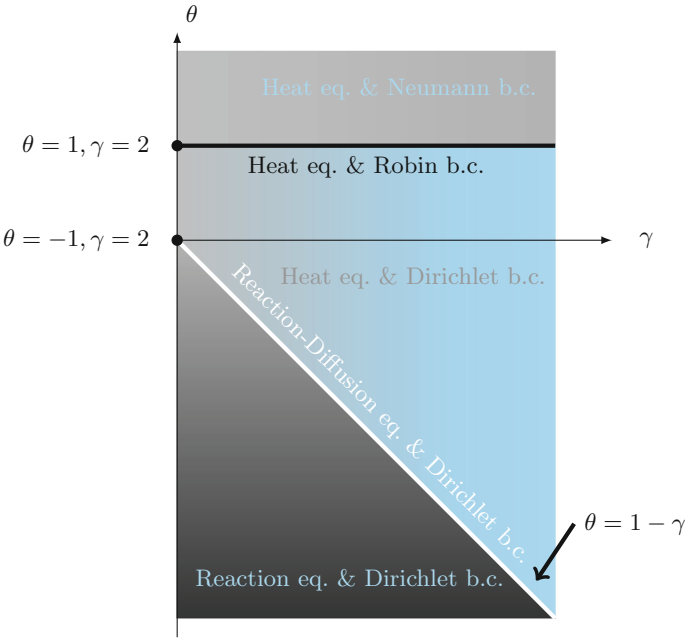


Fig. 5. The five different hydrodynamic regimes in terms of γ and θ.

to [14] for a complete description on the behavior of those solutions. Below we draw the graph of these stationary solutions for a choice of $\alpha = 0.2$ and $\beta = 0.8$.

The proof of Theorem 3 is described in Sect. 2.7 below Fig. 4 and for that reason many steps now are omitted. The proof of tightness of the sequence $\{\mathbb{Q}_N\}_{N\geq 1}$ is quite similar to the one given in Sect. 2.9. The characterization of limit points is also close to the one given in Sect. 2.10, the only difference

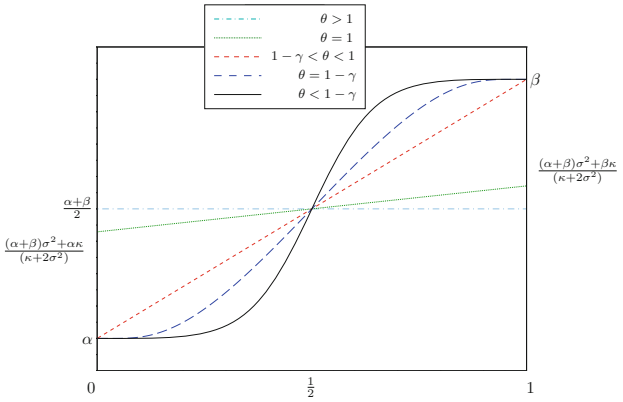


Fig. 6. Stationary solutions of the hydrodynamic equations.

comes at the level of the identification of the density as a weak solution of the corresponding partial differential equation. For that purpose, the next section is dedicated to the presentation of an heuristic argument to deduce the weak formulation for the solution of the corresponding hydrodynamic equation. The adaptation of the rest of the arguments to this new dynamics is left to the reader.

Heuristics for Hydrodynamic Equations: As in Sect. 2.8, the identification of the density $\rho_t(\cdot)$ as a weak solution of the corresponding hydrodynamic equation is obtained by using auxiliary martingales. Fix then a function $G : [0,1] \to \mathbb{R}$ which does not depend on time and which is two times continuously differentiable. As in Sect. 2.8, we use Dynkin's formula and we note that

$$\begin{aligned}\int_0^t \Theta(N)\mathcal{L}_N(\langle \pi_s^N, G\rangle)\, ds = \\ \frac{\Theta(N)}{N-1}\int_0^t \sum_{x\in\Lambda_N} \tilde{\mathcal{L}_N} G(\tfrac{x}{N})\eta_{s\theta(N)}(x)\, ds \\ + \frac{\kappa\Theta(N)}{(N-1)N^\theta}\int_0^t \sum_{y\in\{0,N\}}\sum_{x\in\Lambda_N} G(\tfrac{x}{N})p(y-x)(r(y)-\eta_{s\theta(N)}(x))\, ds,\end{aligned} \tag{79}$$

where for all $x \in \Lambda_N$

$$(\tilde{\mathcal{L}}_N G)(\tfrac{x}{N}) = \sum_{y\in\Lambda_N} p(y-x)\left[G(\tfrac{y}{N}) - G(\tfrac{x}{N})\right]. \tag{80}$$

Now, we extend the first sum in (79) to all the integers so that we extend the function G to $\mathbb{R}$ in such a way that it remains two times continuously differentiable. By the definition of $\tilde{\mathcal{L}_N}$, we get that

$$\begin{aligned}\frac{\Theta(N)}{N-1}\int_0^t \sum_{x\in\Lambda_N} \tilde{\mathcal{L}_N} G(\tfrac{x}{N})\eta_{s\theta(N)}(x)\, ds \\ = \frac{\Theta(N)}{N-1}\int_0^t \sum_{x\in\Lambda_N} (K_N G)(\tfrac{x}{N})\eta_{s\theta(N)}(x)\, ds \\ - \frac{\Theta(N)}{N-1}\int_0^t \sum_{x\in\Lambda_N}\sum_{y\le 0}\left[G(\tfrac{y}{N}) - G(\tfrac{x}{N})\right] p(x-y)\eta_{s\theta(N)}(x)\, ds \\ - \frac{\Theta(N)}{N-1}\int_0^t \sum_{x\in\Lambda_N}\sum_{y\ge N}\left[G(\tfrac{y}{N}) - G(\tfrac{x}{N})\right] p(x-y)\eta_{s\theta(N)}(x)\, ds,\end{aligned} \tag{81}$$

where

$$(K_N G)(\tfrac{x}{N}) = \sum_{y\in\mathbb{Z}} p(y-x)\left[G(\tfrac{y}{N}) - G(\tfrac{x}{N})\right]. \tag{82}$$

Now, we are going to analyse how the different boundary conditions appear on the hydrodynamic equations given in Sect. 3.2 from this dynamics.

The Case $\theta < 1-\gamma$: Take a function $G : (0,1) \to \mathbb{R}$ two times continuously differentiable and with compact support in $(0,1)$, so that we can choose an extension by 0 outside of the support of G. Since $\Theta(N) = N^{\gamma+\theta+1}$ (see the statement of Theorem 3) a simple computation shows that the first term in (81) vanishes for $\theta < 1-\gamma$. Indeed, by a Taylor expansion on G and the fact that $p(\cdot)$ is mean zero, we have that

$$N^{\gamma+\theta+1} \sum_{y\in\mathbb{Z}} (G(\tfrac{y+x}{N}) - G(\tfrac{x}{N}))p(y)$$

is of same order as

$$N^{\gamma+\theta-1} G''(\tfrac{x}{N}) \sum_{y\in\mathbb{Z}} y^2 p(y)$$

and since $\theta < 1-\gamma$ last expression vanishes as $N \to \infty$.

Now, the second and third terms in (81) vanish as $N \to \infty$, since $\Theta(N) = N^{\gamma+\theta+1}$ and $\theta < 1-\gamma$. Note that since G vanishes outside $(0,1)$, those terms can be rewritten as

$$\begin{aligned} &\frac{\Theta(N)}{N-1} \int_0^t \sum_{x\in\Lambda_N} G(\tfrac{x}{N}) r_N^-(\tfrac{x}{N}) \eta_{sN^{\gamma+\theta+1}}(x)\, ds \\ &+ \frac{\Theta(N)}{N-1} \int_0^t \sum_{x\in\Lambda_N} G(\tfrac{x}{N}) r_N^+(\tfrac{x}{N}) \eta_{sN^{\gamma+\theta+1}}(x)\, ds, \end{aligned} \tag{83}$$

where

$$r_N^-(\tfrac{x}{N}) = \sum_{y\geq x} p(y), \quad r_N^+(\tfrac{x}{N}) = \sum_{y\leq x-N} p(y). \tag{84}$$

We observe that, for any $a \in (0,1)$, uniformly in $u \in (a, 1-a)$, as $N \to \infty$:

$$\begin{aligned} &N^\gamma r_N^-([uN]) \to_{N\to+\infty} c_\gamma \gamma^{-1} u^{-\gamma} := r^-(u), \\ &N^\gamma r_N^+([uN]) \to_{N\to+\infty} c_\gamma \gamma^{-1} (1-u)^{-\gamma} := r^+(u). \end{aligned} \tag{85}$$

Now we note that we can bound from above, for example the term at the left hand side in (83) by $N^{\theta+1}$ times

$$\frac{1}{N-1} \int_0^t \sum_{x\in\Lambda_N} N^\gamma r_N^-(\tfrac{x}{N})\, |G(\tfrac{x}{N})|$$

because $|\eta_{sN^{\gamma+\theta}}(x)| \leq 1$ for all $s > 0$. Since $\theta < -1$ and since the previous sum converges to the (finite) integral of $|G| r^-$ on $(0,1)$, by (85), the previous display vanishes as $N \to \infty$. Now we look at the boundary terms in (79), which can be written, for the choice of $\Theta(N) = N^{\gamma+\theta+1}$, as:

$$\frac{\kappa N^{\gamma+1}}{N-1} \int_0^t \sum_{y\in\{0,N\}} \sum_{x\in\Lambda_N} G(\tfrac{x}{N}) p(y-x)(r(y) - \eta_{sN^{\gamma+\theta+1}}(x))\, ds$$

which is equal to

$$\kappa \int_0^t \langle \alpha - \pi_s^N, Gp \rangle + \langle \beta - \pi_s^N, G\tilde{p} \rangle \, ds,$$

where $\tilde{p}(q) = p(1-q)$, and can be replaced, thanks to the fact that G has compact support, by

$$\kappa \int_0^1 G(q) \Big(p(q)(\alpha - \rho_s(q)) + \tilde{p}(q)(\beta - \rho_s(q)) \Big) dq$$

as $N \to \infty$. The last convergence holds because G has compact support included in $(0,1)$ so that Gp and $G\tilde{p}$ are continuous function. From the previous computations we recognize the terms in (75) with $\hat{\kappa} = \kappa c_\gamma$ and $\hat{\sigma} = 0$.

The Case $\boldsymbol{\theta = 1 - \gamma}$: In this case we also take a function $G : (0,1) \to \mathbb{R}$ two times continuously differentiable and with compact support in $(0,1)$, so that we can choose an extension by 0 outside of its support. In this case, since $\Theta(N) = N^2$, by Lemma 3, which we state below, the first term in (81) can be replaced, for N sufficiently big, by

$$\frac{1}{N-1} \int_0^t \sum_{x \in \Lambda_N} \tfrac{\sigma^2}{2} \Delta G(\tfrac{x}{N}) \, \eta_{sN^2}(x) \, ds = \int_0^t \langle \pi_s^N, \tfrac{\sigma^2}{2} \Delta G \rangle \, ds.$$

Moreover, a similar computation to the one above shows that the second and third terms in (81) vanish as $N \to \infty$ (recall that $\Theta(N) = N^2$ and $\gamma > 2$). Finally, the second term in (79) can be rewritten as

$$\frac{\kappa N^{\gamma+1}}{(N-1)} \int_0^t \sum_{y \in \{0,N\}} \sum_{x \in \Lambda_N} G(\tfrac{x}{N}) p(y-x) \, (r(y) - \eta_{sN^2}(x)) \, ds$$

and repeating the analysis we did in the previous case it converges, as $N \to \infty$ to

$$\kappa \int_0^t \int_0^1 G(q) \Big(p(q)(\alpha - \rho_s(q)) + \tilde{p}(q)(\beta - \rho_s(q)) \Big) dq \, ds.$$

As above, from the previous computations we recognize the terms in (75) with $\hat{\kappa} = \kappa c_\gamma$ and $\hat{\sigma} = \sigma$.

The Case $\boldsymbol{\theta \in (1-\gamma, 1)}$: Take again a function $G : (0,1) \to \mathbb{R}$ two times continuously differentiable and with compact support in $(0,1)$ and extend it by 0 outside $(0,1)$. As above, since $\Theta(N) = N^2$, by Lemma 3, which we prove below, the first term in (81) can be replaced, for N sufficiently big, by

$$\int_0^t \langle \pi_s^N, \tfrac{\sigma^2}{2} \Delta G \rangle \, ds.$$

Now, the second term in (79) equals to

$$\frac{\kappa N^{2-\theta}}{N-1} \int_0^t \sum_{y \in \{0,N\}} \sum_{x \in \Lambda_N} G(\tfrac{x}{N}) p(y-x) (r(y) - \eta_{sN^2}(x)) \, ds$$

and vanishes as $N \to \infty$ since $\theta > 1 - \gamma$. Now, the last two terms in (81) also vanish because, for example, the second term in (81) can be written as

$$\int_0^t \frac{N^2}{N-1} \sum_{x \in \Lambda_N} G(\tfrac{x}{N}) r_N^-(\tfrac{x}{N}) \eta_{sN^2}(x)\, ds$$

which can be bounded from above by a constant times $tN^{2-\gamma}$ times a sum converging to the integral of $|G|r^-$ on $(0,1)$, and since $\gamma > 2$ this term vanishes. From this, we see the terms in (75) with $\hat{\kappa} = 0$ and $\hat{\sigma} = \sigma$.

Remark 14. We remark here that in the last three cases, similarly to what we have seen in the case $\theta < 0$ for the models of Sect. 2 (see Remark 5), there is an extra condition in the definition of the weak solution of (74). In this notion of solution we need to show that the value of the profile $\rho_t(\cdot)$ is fixed at the boundary. This issue is analysed in Appendix A.4.

The Case $\theta = 1$: In this case we consider a function $G : [0,1] \to \mathbb{R}$ which is two times continuously differentiable and we extend it on $\mathbb{R}$ in a two times continuously differentiable function with compact support which strictly contains $[0,1]$. Note that in this case G can take non-zero values at 0 and 1. As above, since $\Theta(N) = N^2$, by Lemma 3, which we state below and which holds for this new space of test functions, the first term in (81) can be replaced, for N sufficiently big, by

$$\int_0^t \langle \pi_s^N, \tfrac{\sigma^2}{2} \Delta G \rangle\, ds.$$

Now we look at the terms coming from the boundary, namely the last term in (79). Then, in the term for $y = 0$ of (79) (resp. for $y = N$) we do at first a Taylor expansion on G and then we replace $\eta(x)$ by the average $\overrightarrow{\eta}^{\varepsilon N}(1) = \frac{1}{\varepsilon N} \sum_{x=1}^{1+\varepsilon N} \eta(x)$ (resp. $\eta(x)$ by $\overleftarrow{\eta}^{\varepsilon N}(N-1) = \frac{1}{\varepsilon N} \sum_{x=N-1-\varepsilon N}^{N-1} \eta(x)$), which can be done as a consequence of Lemma 7 as pointed out in Remark 17. Moreover, note that for $y = 0$ and $y = N$ it holds that

$$\sum_{x \in \Lambda_N} p(y-x) \xrightarrow[N \to +\infty]{} \frac{1}{2}. \tag{86}$$

Therefore, we can write the last term in (79) as

$$\frac{\kappa}{2} \int_0^t \{(\alpha - \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)) G(0) + (\beta - \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1)) G(1)\}\, ds,$$

plus terms that vanish as $N \to +\infty$. Since

$$\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) \sim \rho_s(0) \quad \text{and} \quad \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) \sim \rho_s(1)$$

last term writes as

$$\frac{\kappa}{2} \int_0^t \{(\alpha - \rho_s(0)) G(0) + (\beta - \rho_s(1)) G(1)\}\, ds. \tag{87}$$

Now, we analyse the two last terms in (81). Since the function G has been extended into a two times continuously differentiable function on $\mathbb{R}$, by a Taylor expansion on G we can write those terms as

$$\frac{N}{N-1}\int_0^t \sum_{x\in\Lambda_N} G'(\tfrac{x}{N})\Theta_x^-\eta_{sN^2}(x)\,ds - \frac{N}{N-1}\int_0^t \sum_{x\in\Lambda_N} G'(\tfrac{x}{N})\Theta_x^+\eta_{sN^2}(x)\,ds \tag{88}$$

plus terms that vanish as $N\to+\infty$. Above for $x\in\Lambda_N$,

$$\Theta_x^- = \sum_{y\le 0}(x-y)p(x-y) \quad \text{and} \quad \Theta_x^+ = \sum_{y\ge N}(y-x)p(x-y).$$

Note that

$$\frac{1}{N}\sum_{x\in\Lambda_N} x\Theta_x^- \xrightarrow[N\to+\infty]{} 0 \quad \text{and} \quad \frac{1}{N}\sum_{x\in\Lambda_N} x\Theta_x^+ \xrightarrow[N\to+\infty]{} 0. \tag{89}$$

Moreover, note that

$$\begin{aligned}\sum_{x\in\Lambda_N}\Theta_x^- &= \sum_{x\in\Lambda_N}\sum_{y\ge x} yp(y) \xrightarrow[N\to+\infty]{} \tfrac{\sigma^2}{2},\\ \sum_{x\in\Lambda_N}\Theta_x^+ &= \sum_{x\in\Lambda_N}\sum_{y\ge N-x} yp(y) \xrightarrow[N\to+\infty]{} \tfrac{\sigma^2}{2}.\end{aligned} \tag{90}$$

In order to prove the convergence of $\sum_{x\in\Lambda_N}\Theta_x^-$ (or of $\sum_{x\in\Lambda_N}\Theta_x^+$ in (90)) we use Fubini's theorem to get that

$$\begin{aligned}\sum_{x\in\Lambda_N}\Theta_x^- &= \sum_{y\in\Lambda_N}\sum_{x=1}^{y} yp(y) + \sum_{y\ge N}\sum_{x\in\Lambda_N} yp(y)\\ &= \sum_{y\in\Lambda_N} y^2p(y) + (N-1)\sum_{y\ge N} yp(y),\end{aligned}$$

and since $\gamma>2$ the result follows. By another Taylor expansion on G we can write (88) as

$$\frac{N}{N-1}G'(0)\int_0^t \sum_{x\in\Lambda_N}\Theta_x^-\eta_{sN^2}(x)\,ds - \frac{N}{N-1}G'(1)\int_s^t \sum_{x\in\Lambda_N}\Theta_x^+\eta_{sN^2}(x)\,ds \tag{91}$$

plus terms that vanish as $N\to+\infty$. From Lemma 7 we can replace in the term on the left (resp. right) hand side of last expression $\eta_{sN^2}(x)$ by $\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)$ (resp. $\overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1)$). Therefore, (91) can be replaced, for N sufficiently big and for ε sufficiently small, by

$$\int_0^t G'(0)\tfrac{\sigma^2}{2}\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) - G'(1)\tfrac{\sigma^2}{2}\overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1)\,ds.$$

Since $\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) \sim \rho_s(0)$ and $\overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) \sim \rho_s(1)$, last term tends to

$$\int_0^t G'(0)\tfrac{\sigma^2}{2}\rho_s(0) - G'(1)\tfrac{\sigma^2}{2}\rho_s(1)\,ds, \tag{92}$$

as $N \to \infty$.

Putting together (87) and (92) we see the boundary terms that appear at the right hand side of (77).

The Case $\theta > 1$: In this case we consider an arbitrary function $G : [0,1] \to \mathbb{R}$ which is two times continuously differentiable and we extend it on $\mathbb{R}$ in a two times continuously differentiable function with compact support. Its support strictly contains $[0,1]$ since G can take non-zero values at 0 and 1. As in the last case, since $\Theta(N) = N^2$, by Lemma 3, the first term in (81) can be replaced, for N sufficiently big, by

$$\int_0^t \langle \pi_s^N, \tfrac{\sigma^2}{2}\Delta G\rangle\, ds.$$

The last term in (79) vanishes, as $N \to \infty$ since, we can bound it by a constant times

$$N^{1-\theta}\sum_{x\in\Lambda_N} p(x).$$

Since $\gamma > 2$ last display vanishes if $\theta > 1$, as $N \to +\infty$. Thus, we only need to look at the expression (81). Therefore, in order to see the boundary terms that appear in (77), we can use exactly the computations already done in the case $\theta = 1$ from which we obtain (92).

We finish this section with the statement of the lemma which is used above in order to obtain the diffusion term in the equations above in the cases $\theta \geq 1-\gamma$. Its proof can be seen in [2].

Lemma 3. *Let $G : \mathbb{R} \to \mathbb{R}$ be a two times continuously differentiable function with compact support. We have*

$$\limsup_{N\to\infty}\sup_{x\in\Lambda_N}\left|N^2\sum_{y\in\mathbb{Z}}(G(\tfrac{y+x}{N}) - G(\tfrac{x}{N}))p(y) - \frac{\sigma^2}{2}\Delta G(\tfrac{x}{N})\right| = 0.$$

3.3 The Infinite Variance Case

In this section we analyse the case in which $p(\cdot)$ is as in (71) but now $\gamma \in (1,2)$ so that $p(\cdot)$ has mean zero but infinite variance. We also consider only the case where $\theta = -1$, but we note that in the regime $\theta < -1$ the behavior of the system, when we take the time scale $\Theta(N) = N^{\gamma+\theta+1}$ is the same as when $\theta < 1-\gamma$ and when $p(\cdot)$ has finite variance, that is, it is given by the weak solution of (74) with $\hat{\sigma} = 0$ and $\hat{\kappa} = \kappa c_\gamma$. The other regimes are open and seem to be quite challenging. Recall the infinitesimal generator given in (72) and (73) and since

we are restricted to the case $\theta = -1$, we consider the Markov process speeded up in the time scale $\Theta(N) = N^\gamma$, so that $\{\eta_{tN^\gamma} : t \geq 0\}$ has infinitesimal generator given by $N^\gamma \mathcal{L}_N$. As in Sect. 2.4 we can prove that the Bernoulli product measures ν_ρ^N as defined in (7) are reversible when we consider $\alpha = \beta = \rho$. The proof is quite similar to the one given in Lemma 1 and for that reason it is omitted.

Hydrodynamic Equations: We can now give the definition of the weak solution of the hydrodynamic equation that will be derived in this section when $p(\cdot)$ is assumed to have infinite variance.

Recall the notations introduced in the beginning of Sect. 2.6. We recall the definition of the fractional Laplacian operator of exponent $\gamma/2$ denoted by $(-\Delta)^{\gamma/2}$. It is a non-local operator which is defined on the set of functions $G : \mathbb{R} \to \mathbb{R}$ such that

$$\int_{-\infty}^{\infty} \frac{|G(q)|}{(1+|q|)^{1+\gamma}} dq < \infty \tag{93}$$

by

$$(-\Delta)^{\gamma/2} G(q) = c_\gamma \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \mathbf{1}_{|q-v| \geq \varepsilon} \frac{G(q) - G(v)}{|q-v|^{1+\gamma}} dv \tag{94}$$

provided the limit exists, which is the case, for example, if G is in the Schwartz space. Recall that c_γ is fixed in (71). Up to a multiplicative constant, $-(-\Delta)^{\gamma/2}$ is the generator of a γ-Lévy stable process.

We define another operator L whose action is given on functions $G \in C_c^\infty((0,1))$, by

$$\forall q \in (0,1), \quad (LG)(q) = c_\gamma \lim_{\varepsilon \to 0} \int_0^1 \mathbf{1}_{|q-v| \geq \varepsilon} \frac{G(v) - G(q)}{|q-v|^{1+\gamma}} dv.$$

The operator L is called the *regional fractional Laplacian* on $(0,1)$. The semi inner-product $\langle \cdot, \cdot \rangle_{\gamma/2}$ is defined on the set $C_c^\infty((0,1))$ by

$$\langle G, H \rangle_{\gamma/2} = \frac{c_\gamma}{2} \iint_{[0,1]^2} \frac{(H(q) - H(v))(G(q) - G(v))}{|q-v|^{1+\gamma}} dq dv. \tag{95}$$

The corresponding semi-norm is denoted by $\|\cdot\|_{\gamma/2}$. Observe that for any $G, H \in C_c^\infty((0,1))$ we have that

$$-\int_0^1 G(q) L H(q)\, dq = -\int_0^1 LG(q) H(q)\, dq = \langle G, H \rangle_{\gamma/2}$$

and note that for all $q \in (0,1)$,

$$(LG)(q) = -(-\Delta)^{\gamma/2} G(q) + V_1(q) G(q) \tag{96}$$

where $V_1(q) = r^-(q) + r^+(q)$, see (85), that is, $V_1(\cdot)$ is given on $q \in (0,1)$ by:

$$V_1(q) = c_\gamma \gamma^{-1} \Big(\frac{1}{q^\gamma} + \frac{1}{(1-q)^\gamma} \Big). \tag{97}$$

Definition 8. *The Sobolev space $\mathcal{H}^{\gamma/2}$ consists of all square integrable functions $g:(0,1)\to\mathbb{R}$ such that $\|g\|_{\gamma/2}<\infty$. This is a Hilbert space for the norm $\|\cdot\|_{\mathcal{H}^{\gamma/2}}$ defined by*

$$\|g\|^2_{\mathcal{H}^{\gamma/2}}:=\|g\|^2+\|g\|^2_{\gamma/2}.$$

Its elements elements coincide a.e. with continuous functions.

The space $L^2(0,T;\mathcal{H}^{\gamma/2})$ is the set of measurable functions $f:[0,T]\to\mathcal{H}^{\gamma/2}$ such that

$$\int_0^T \|f_t\|^2_{\mathcal{H}^{\gamma/2}}dt<\infty.$$

We now extend the definition of the regional fractional Laplacian on $(0,1)$ to the space $\mathcal{H}^{\gamma/2}$.

Definition 9. *For $\rho\in\mathcal{H}^{\gamma/2}$ we define the distribution $L\rho$ by*

$$\int_0^1 L\rho(q)G(q)\,dq=\int_0^1\rho(q)LG(q)\,dq,\quad G\in C_c^\infty((0,1)).$$

Let L_κ be the regional fractional Laplacian on $[0,1]$ with zero Dirichlet boundary conditions, indexed by κ, and taking the form

$$L_\kappa=L-\kappa\tilde{V}_1, \tag{98}$$

where for $q\in(0,1)$,

$$\tilde{V}_1(q)=p(q)+\tilde{p}(q)=c_\gamma\Big(\frac{1}{q^{\gamma+1}}+\frac{1}{(1-q)^{\gamma+1}}\Big). \tag{99}$$

Above $\tilde{p}(q)=p(1-q)$. Below $g:[0,1]\to[0,1]$ is a measurable function and it is the initial condition of the partial differential equation that we obtain in this section.

Definition 10. *Let $\kappa>0$ be some parameter. We say that $\rho^\kappa:[0,T]\times[0,1]\to[0,1]$ is a weak solution of the regional fractional reaction-diffusion equation with Dirichlet boundary conditions given by*

$$\begin{cases}\partial_t\rho^\kappa_t(q)=L_\kappa\rho^\kappa_t(q)+\kappa\tilde{V}_0(q),\quad (t,q)\in[0,T]\times(0,1),\\ \rho^\kappa_t(0)=\alpha,\quad \rho^\kappa_t(1)=\beta,\quad t\in[0,T],\end{cases} \tag{100}$$

where

$$\tilde{V}_0(q)=\alpha p(q)+\beta\tilde{p}(q)=c_\gamma\Big(\frac{\alpha}{q^{1+\gamma}}+\frac{\beta}{(1-q)^{1+\gamma}}\Big),$$

and starting from a measurable function $g:[0,1]\to[0,1]$, if:

1. $\rho^\kappa\in L^2(0,T;\mathcal{H}^{\gamma/2})$.
2. $\int_0^T\int_0^1\Big\{\frac{(\alpha-\rho^\kappa_t(q))^2}{q^{1+\gamma}}+\frac{(\beta-\rho^\kappa_t(q))^2}{(1-q)^{1+\gamma}}\Big\}\,dq\,dt<\infty$.

3. For all $t \in [0,T]$ and all functions $G \in C_c^{1,\infty}([0,T] \times (0,1))$ we have that

$$\begin{aligned} F_{Dir}^{\kappa} := & \int_0^1 \rho_t^{\kappa}(q) G_t(q)\, dq - \int_0^1 g(q) G_0(q)\, dq \\ & - \int_0^t \int_0^1 \rho_s^{\kappa}(q) \Big(\partial_s + L_{\kappa}\Big) G_s(q)\, dq ds \\ & - \kappa \int_0^t \int_0^1 G_s(q) \tilde{V}_0(q)\, dq\, ds = 0. \end{aligned} \tag{101}$$

Remark 15. We observe that the partial differential equation above has a unique weak solution in the sense defined above. We do not include the proof of this result in these notes but we refer the interested reader to [2] for the proof of the uniqueness for a very similar equation. The same proof gives uniqueness in this case.

Hydrodynamic Limit: Recall the notion of the empirical measure given in Sect. 2.6 and note that in this case we have

$$\pi_t^N(\eta, dq) := \pi^N(\eta_{tN^{\gamma}}, dq)$$

since the time scale now is equal to $\theta(N) = N^{\gamma}$.

The second result of this section is stated in the following theorem.

Theorem 4. *Let $g : [0,1] \to [0,1]$ be a measurable function and let $\{\mu_N\}_{N \geq 1}$ be a sequence of probability measures in Ω_N associated to $g(\cdot)$. Then, for any $0 \leq t \leq T$,*

$$\lim_{N \to \infty} \mathbb{P}_{\mu_N} \left(\eta : \left| \frac{1}{N-1} \sum_{x \in \Lambda_N} G\left(\tfrac{x}{N}\right) \eta_{tN^{\gamma}}(x) - \int_0^1 G(q) \rho_t^{\kappa}(q) dq \right| > \delta \right) = 0,$$

where $\rho_t^{\kappa}(\cdot)$ is the unique weak solution of (100) in the sense of Definition 10.

Heuristics for Hydrodynamic Equations: Fix $G : [0,1] \to \mathbb{R}$ which does not depend on time and has compact support included in $(0,1)$. Recall (79) and (81) and recall that we assumed $\theta = -1$, so that (3.2) now writes as

$$\begin{aligned} \int_0^t N^{\gamma} \mathcal{L}_N(\langle \pi_s^N, G \rangle)\, ds = & \frac{N^{\gamma}}{N-1} \int_0^t \sum_{x \in \Lambda_N} (\tilde{\mathcal{L}}_N G)(\tfrac{x}{N}) \eta_{sN^{\gamma}}(x) \\ & + \frac{\kappa N^{\gamma+1}}{(N-1)} \int_0^t \sum_{y \in \{0,N\}} \sum_{x \in \Lambda_N} G(\tfrac{x}{N}) p(y-x)(r(y) - \eta_{sN^{\gamma}}(x))\, ds. \end{aligned} \tag{102}$$

Note that the first term on the right hand side in last display is equal to

$$\int_0^t \langle \pi_s^N, \tilde{\mathcal{L}}_N G \rangle\, ds.$$

Since from Lemma 3.3 in [4], we can deduce that

$$\lim_{N\to\infty} N^{\gamma}(\tilde{\mathcal{L}}_N G)(q) = (LG)(q) \tag{103}$$

uniformly in $[a, 1-a]$, for all functions G with compact support included in $[a, 1-a]$. Therefore, the first term on the right hand side of (102) can be replaced by

$$\int_0^t \int_0^1 (LG)(q)\rho_s^{\kappa}(q)\, dq\, ds, \tag{104}$$

for N sufficiently big. Now, the second term on the right hand side in (102) is equal to

$$\kappa \int_0^t \langle \alpha - \pi_s^N, Gp\rangle\, ds + \kappa \int_0^t \langle \beta - \pi_s^N, G\tilde{p}\rangle\, ds$$

and converges, as $N \to \infty$, to

$$\begin{aligned} &\kappa \int_0^t \int_0^1 (\alpha - \rho_t^{\kappa}(q))G(q)p(q)du + \kappa \int_0^t \int_0^1 (\beta - \rho_t^{\kappa}(q))G(q)\tilde{p}(q)dq \\ &= -\kappa \int_0^t \int_0^1 \rho_t^{\kappa}(q)G(q)\tilde{V}_1(q)dq + \kappa \int_0^t \int_0^1 G(q)\tilde{V}_0(q)dq. \end{aligned} \tag{105}$$

Putting together (104) and (105) and using (98) we recognize the corresponding terms in (101).

We finish this section by noting that in [3] it was studied a similar dynamics to the one described above. There we considered the same bulk dynamics with long jumps given by $p(\cdot)$ with the choice (71) and $\gamma \in (1, 2)$ but the boundary dynamics was different. In [3] instead of considering just one boundary at each end point of the bulk, it was added infinitely many reservoirs at the left and at the right of the bullk. As in the dynamics described above, particles can be injected and removed from the system at any point of the bulk by any of the reservoirs located at $y \le 0$ or $y \ge N$. We note that in the case of this new dynamics the results obtained in [3] are similar to those presented here, except that the transitions occur for a different value of θ and for that reason, the potential $\tilde{V}_1$ that appears in the definition of L_κ in the reaction-diffusion equation (100) has a different power than the one that appears in the hydrodynamic equation in [3]. It would be very interesting to analyse other types of boundary dynamics superposed to the bulk dynamics that we defined above in order to see if we can come up with new fractional reaction-diffusion equations with more tricky boundary conditions than the Dirichlet boundary conditions that we obtained here. And it would be very interesting to look at the case where $\theta > -1$, the slow boundary regime, when $p(\cdot)$ is given as above with $\gamma \in (1, 2)$, see the area coloured in rose in the figure below. This is a subject to pursue in the near future. In the figure below we summarize the scenario of the hydrodynamic limit for the models of this section.

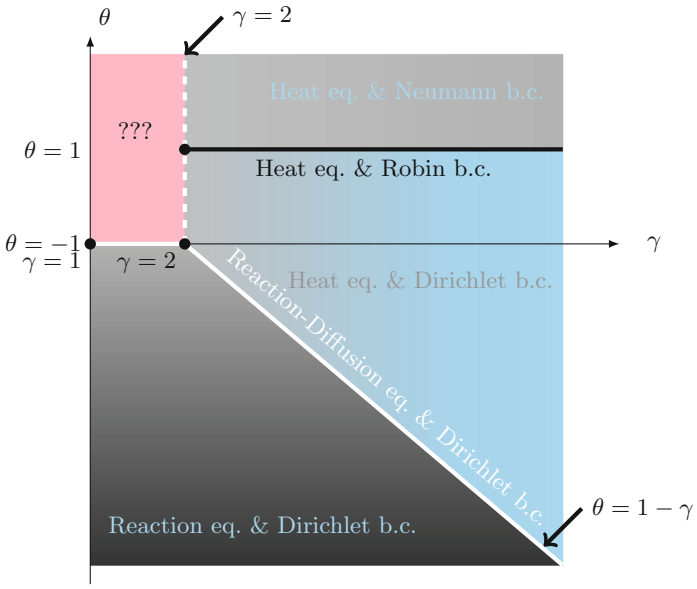


Fig. 7. Hydrodynamical behavior of the symmetric exclusion with long jumps.

Acknowledgements. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovative programme (grant agreement No 715734).

The author would like to express her gratitude to the organizers of the trimester "Stochastic dynamics out of equilibrium", namely Ellen Saada, Gabriel Stoltz, Giambattista Giacomin, Herbert Spohn, Stefano Olla, for the invitation to taught the mini-course and for the financial support to attend the trimester.

Finally the author would like to thank Cédric Bernardin, Byron Jiménez Oviedo and Adriana Neumann for the discussions around the problems described in these notes.

A Auxiliary results

In this section we establish some technical results that are needed in order to prove the hydrodynamic limit for the models discussed in the previous sections.

A.1 Entropy bound

From now on, we suppose that $\alpha \le \beta$. Let $\rho : [0,1] \to [0,1]$ be a function such that $\alpha \le \rho(q) \le \beta$, for all $q \in [0,1]$. Let $\nu^N_{\rho(\cdot)}$ be the Bernoulli product measure on Ω_N with marginals given by

$$\nu^N_{\rho(\cdot)}\{\eta : \eta_x = 1\} = \rho\left(\tfrac{x}{N}\right). \tag{106}$$

Given two functions $f, g : \Omega_N \to \mathbb{R}$ and a probability measure μ on Ω_N, we denote here by $\langle f, g \rangle_\mu$ the scalar product between f and g in $L^2(\Omega_N, \mu)$, that is,

$$\langle f, g \rangle_\mu = \int_{\Omega_N} f(\eta) g(\eta) \, d\mu.$$

Let $H_N(\mu | \nu^N_{\rho(\cdot)})$ be the relative entropy of a probability measure μ on Ω_N with respect to the probability measure $\nu^N_{\rho(\cdot)}$ on Ω_N. We claim that there exists a constant $C_0 := C(\alpha, \beta)$, such that

$$H_N(\mu | \nu^N_{\rho(\cdot)}) \leq C_0 N. \tag{107}$$

For that purpose note that, since $\nu^N_{\rho(\cdot)}$ is product we have that

$$\nu^N_{\rho(\cdot)}(\eta) = \prod_{x=1}^{N-1} \rho(\tfrac{x}{N})^{\eta(x)} (1 - \rho(\tfrac{x}{N}))^{1-\eta(x)} \geq (\alpha \wedge (1-\beta))^N$$

from where we obtain that

$$\begin{aligned} H(\mu | \nu^N_{\rho(\cdot)}) &= \sum_{\eta \in \Omega_N} \mu(\eta) \log \left(\frac{\mu(\eta)}{\nu^N_{\rho(\cdot)}(\eta)} \right) \leq \sum_{\eta \in \Omega_N} \mu(\eta) \log \left(\frac{1}{\nu^N_{\rho(\cdot)}(\eta)} \right) \\ &\leq \log \left(\left[\frac{1}{\alpha \wedge (1-\beta)} \right]^N \right) \sum_{\eta \in \Omega_N} \mu(\eta) \leq N \log \left(\frac{1}{\alpha \wedge (1-\beta)} \right) \leq C_0 N. \end{aligned}$$

We remark here that below when we use as reference measure the Bernoulli product measure given in (106) we have to restrict to $\alpha \neq 0$ and $\beta \neq 1$ since in last estimate the constant $C_0 = -\log(\alpha \wedge (1-\beta))$. We also note that when we use the Bernoulli product measure with a constant parameter we do not need to impose that restriction.

A.2 Estimates on Dirichlet forms

In this section we consider the model described in Sect. 3 since the results for the model of Sect. 2 can be obtained easily from the ones we derive below. In any case we present some remarks along the text about the corresponding results for the model of Sect. 2.

For a probability measure μ on Ω_N, $x, y \in \Lambda_N$ and a density function $f : \Omega_N \to [0, \infty)$ with respect to μ we introduce

$$\begin{aligned} I_{x,y}(\sqrt{f}, \mu) &:= \int_{\Omega_N} \left(\sqrt{f(\eta^{x,y})} - \sqrt{f(\eta)} \right)^2 d\mu, \\ I^{r(y)}_x(\sqrt{f}, \mu) &:= \int_{\Omega_N} c_x(\eta; r(y)) \left(\sqrt{f(\eta^x)} - \sqrt{f(\eta)} \right)^2 d\mu. \end{aligned}$$

In last identity $y \in \{0, N\}$ and $r(0) = \alpha$ and $r(N) = \beta$. We define

$$\mathcal{D}_N(\sqrt{f}, \mu) := (\mathcal{D}_{N,0} + \mathcal{D}_{N,b})(\sqrt{f}, \mu)$$

where

$$\mathcal{D}_{N,0}(\sqrt{f},\mu) := \frac{1}{2}\sum_{x,y\in\Lambda_N} p(y-x)\, I_{x,y}(\sqrt{f},\mu), \tag{108}$$

$$\mathcal{D}_{N,b}(\sqrt{f},\mu) := \frac{\kappa}{N^\theta}\sum_{y\in\{0,N\}}\sum_{x\in\Lambda_N} p(y-x)\, I_x^{r(y)}(\sqrt{f},\mu). \tag{109}$$

Note that for the models of Sect. 2 the expressions above simplify to

$$\mathcal{D}_{N,0}^{NN}(\sqrt{f},\mu) := \sum_{x\in\Lambda_N} I_{x,x+1}(\sqrt{f},\mu), \tag{110}$$

$$\mathcal{D}_{N,b}^{NN}(\sqrt{f},\mu) := \frac{\kappa}{N^\theta}\Big(I_1^\alpha(\sqrt{f},\mu) + I_{N-1}^\beta(\sqrt{f},\mu)\Big). \tag{111}$$

Our first goal is to express, for the measure $\mu = \nu^N_{\rho(\cdot)}$, a relation between the Dirichlet form defined by $-\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}}$ and $\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)})$. We claim that for any positive constant B, there exists a constant $C>0$ such that

$$\begin{aligned}\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \le & -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) \\ & + \frac{C}{BN}\sum_{x,y\in\Lambda_N} p(y-x)\Big(\rho(\tfrac{x}{N})-\rho(\tfrac{y}{N})\Big)^2 \\ & + \frac{C\kappa}{BN^{1+\theta}}\sum_{y\in\{0,N\}}\sum_{x\in\Lambda_N}\Big(\rho(\tfrac{x}{N})-r(y)\Big)^2 p(y-x).\end{aligned} \tag{112}$$

Our aim is then to choose $\rho(\cdot)$ in order to minimize the error term, i.e. the two last terms at the right hand side of the previous inequality.

Remark 16.

1. If $p(\cdot)$ has finite variance σ^2, then:
   - for $\rho(\cdot)$ Lipschitz and such that $\rho(0)=\alpha$ and $\rho(1)=\beta$, we get

$$\begin{aligned}\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} &\le -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{BN^2}\sigma^2 \\ &\quad + \frac{C\kappa}{BN^{3+\theta}}\sum_{y\in\{0,N\}}\sum_{x\in\Lambda_N}(y-x)^2 p(y-x) \\ &\le -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{BN^2}\sigma^2 + \frac{C\kappa}{BN^{3+\theta}}.\end{aligned} \tag{113}$$

   - for $\rho(\cdot)$ such that $\rho(0)=\alpha$, $\rho(1)=\beta$, Hölder of parameter $\frac{\gamma}{2}$ at the boundaries and Lipschitz inside, we get

$$\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \le -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{BN^2}\sigma^2 + \frac{C\kappa\log(N)}{BN^{\gamma+\theta+1}}. \tag{114}$$

– for $\rho(\cdot)$ such that $\rho(0)=\alpha$, $\rho(1)=\beta$, Hölder of parameter $\frac{1+\gamma}{2}$ at the boundaries and Lipschitz inside, we get

$$\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \leq -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{BN^2}\sigma^2 + \frac{C\kappa}{BN^{\gamma+\theta+1}}. \tag{115}$$

– for $\rho(\cdot)$ constant, equal to α or to β, we have

$$\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\alpha}} \leq -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu_{\alpha}) + \frac{C\kappa}{BN^{\theta+1}}. \tag{116}$$

2. If $p(\cdot)$ is such that $p(1)=p(-1)=\frac{1}{2}$, then:
   – for $\rho(\cdot)$ Lipschitz and such that $\rho(0)=\alpha$, $\rho(1)=\beta$ and locally constant at 0 and 1, we get

$$\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \leq -\frac{1}{4BN}\mathcal{D}^{NN}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{BN^2}. \tag{117}$$

   Note that the choice of asking $\rho(\cdot)$ to be locally constant at 0 and 1 turns the errors coming from the boundary dynamics to vanish.
   – for $\rho(\cdot)$ constant, equal to α or to β, then we have exactly the same error as in (116).
3. If $p(\cdot)$ has infinite variance, then:
   – for $\rho(\cdot)$ Lipschitz and such that $\rho(0)=\alpha$ and $\rho(1)=\beta$, we get

$$\begin{aligned}\frac{1}{BN}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} &\leq -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) \\ &+ \frac{C}{BN^3}\sum_{x,y\in\Lambda_N}\frac{1}{|x-y|^{\gamma-1}} \\ &+ \frac{C\kappa}{BN^{3+\theta}}\sum_{y\in\{0,N\}}\sum_{x\in\Lambda_N}(y-x)^2p(y-x) \\ &\leq -\frac{1}{4BN}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{BN^{\gamma}}\sigma^2 + \frac{C\kappa}{BN^{\gamma+\theta+1}}.\end{aligned} \tag{118}$$

In order to prove (112) we need some intermediate results. For that purpose we recall from [2] the following two lemmas.

Lemma 4. *Let $T:\eta\in\Omega_N\to T(\eta)\in\Omega_N$ be a transformation in the configuration space and $c:\eta\in\Omega_N\to c(\eta)$ be a positive local function. Let f be a density with respect to a probability measure μ on Ω_N. Then, we have that*

$$\begin{aligned}&\left\langle c(\eta)[\sqrt{f(T(\eta))}-\sqrt{f(\eta)}]\,,\,\sqrt{f(\eta)}\right\rangle_{\mu} \\ &\leq -\frac{1}{4}\int c(\eta)\left(\left[\sqrt{f(T(\eta))}\right]-\left[\sqrt{f(\eta)}\right]\right)^2 d\mu \\ &+\frac{1}{16}\int\frac{1}{c(\eta)}\left[c(\eta)-c(T(\eta))\frac{\mu(T(\eta))}{\mu(\eta)}\right]^2\left(\left[\sqrt{f(T(\eta))}\right]+\left[\sqrt{f(\eta)}\right]\right)^2 d\mu.\end{aligned} \tag{119}$$

Lemma 5. *There exists a constant $C := C(\rho)$ such that for any $N \geq 1$ and density f be a density with respect to $\nu^N_{\rho(\cdot)}$*

$$\sup_{x \neq y \in \Lambda_N} \int_{\Omega_N} f(\eta^{x,y})\, d\nu^N_{\rho(\cdot)} \;\leq\; C, \qquad \sup_{x \in \Lambda_N} \int_{\Omega_N} f(\eta^{x})\, d\nu^N_{\rho(\cdot)} \;\leq\; C.$$

A simple consequence of the previous lemmas is the next two corollaries. Recall the bulk generator $\mathcal{L}_{N,0}$ given in (73).

Corollary 1. *There exists a constant $C > 0$ (independent of $f(\cdot)$ and N) such that*

$$\Big\langle \mathcal{L}_{N,0}\sqrt{f}, \sqrt{f}\Big\rangle_{\nu^N_{\rho(\cdot)}} \leq -\frac{1}{4}\mathcal{D}_{N,0}(\sqrt{f}, \nu^N_{\rho(\cdot)}) + C \sum_{x,y \in \Lambda_N} p(y-x)\Big(\rho(\tfrac{x}{N}) - \rho(\tfrac{y}{N})\Big)^2$$

for any density $f(\cdot)$ with respect to $\nu^N_{\rho(\cdot)}$.

Now we look at the generator of the boundary dynamics given in (73).

Corollary 2. *Let $\theta \in \mathbb{R}$ be fixed. There exists a constant $C > 0$ (independent of $f(\cdot)$ and N) such that*

$$\begin{aligned} \langle \mathcal{L}_{N,b}\sqrt{f}, \sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \leq & -\frac{1}{4}\mathcal{D}_{N,b}(\sqrt{f}, \nu^N_{\rho(\cdot)}) \\ & + \frac{C\kappa}{N^\theta} \sum_{x \in \Lambda_N} \Big(\rho(\tfrac{x}{N}) - \alpha\Big)^2 p(x) \\ & + \frac{C\kappa}{N^\theta} \sum_{x \in \Lambda_N} \Big(\rho(\tfrac{x}{N}) - \beta\Big)^2 p(N-x) \end{aligned} \tag{120}$$

for any density $f(\cdot)$ with respect to $\nu^N_{\rho(\cdot)}$.

To prove the first corollary take $c \equiv 1$, $T(\eta) = \eta^{x,y}$ and note that

$$|\theta^{x,y}(\eta) - 1|^2 \leq C(\rho(\tfrac{x}{N}) - \rho(\tfrac{y}{N}))^2.$$

To prove the second corollary we take for each $y \in \{0, N\}$, $c(\eta) = c_x(\eta; r(y))$ and $T(\eta) = \eta^x$. From the two previous corollaries the claim (112) follows easily. We leave the details of the gaps to the reader.

A.3 Replacement Lemmas

In this section we prove rigorously all the replacements that were mentioned along the Sects. 2.8 and 3.2. We first recall Lemma 5.5 of [2] adapted to our situation (with just one reservoirs at each end point of the bulk).

Lemma 6. *For any density $f(\cdot)$ with respect to $\nu^N_{\rho(\cdot)}$, any $x \in \Lambda_N$, any $y \in \{0, N\}$ and any positive constant A_x, there exists a constant C such that*

$$\Big|\langle \eta(x) - r(y), f\rangle_{\nu^N_{\rho(\cdot)}}\Big| \;\leq\; \frac{C}{A_x} I_x^{r(y)}(\sqrt{f}, \nu^N_{\rho(\cdot)}) + CA_x + C\Big|\rho(\tfrac{x}{N}) - r(y)\Big|.$$

The first replacement lemma that we prove is the one that is needed for the model of Sect. 3 when $p(\cdot)$ has finite variance for the case $\theta \geq 1$.

Lemma 7. *For any $t > 0$, for $\gamma > 2$ and for any $\theta \geq 1$ we have that*

$$\lim_{\varepsilon \to 0} \lim_{N \to \infty} E_{\mathbb{P}_{\mu_N}} \left[\left| \int_0^t \sum_{x \in \Lambda_N} \Theta_x^- \left(\eta_{sN^2}(x) - \overrightarrow{\eta}_{sN^2}^{\varepsilon N}(1) \right) ds \right| \right] = 0,$$

$$\lim_{\varepsilon \to 0} \lim_{N \to \infty} E_{\mathbb{P}_{\mu_N}} \left[\left| \int_0^t \sum_{x \in \Lambda_N} \Theta_x^+ \left(\eta_{sN^2}(x) - \overleftarrow{\eta}_{sN^2}^{\varepsilon N}(N-1) \right) ds \right| \right] = 0.$$

Proof. Below C is a constant than can change from line to line. Note that since $\theta \geq 1$ we have $\theta(N) = N^2$. We present the proof for the first term, but we note that the proof for the second one is analogous. Here we take as reference measure the Bernoulli product measure with constant parameter (for example α) and we recall (116), from where we see that

$$\frac{N}{B} \langle \mathcal{L}_N \sqrt{f}, \sqrt{f} \rangle_{\nu_\alpha} \leq -\frac{N}{4B} \mathcal{D}_N(\sqrt{f}, \nu_\alpha^N) + \frac{C\kappa}{B} N^{1-\theta} \tag{121}$$

so that the error to change the Dirichlet form vanishes as $N \to \infty$ for $\theta > 1$ and for $\theta = 1$ it vanishes when $B \to +\infty$.

By the entropy and Jensen's inequalities, the first expectation in the statement of the lemma is bounded from above, for any constant $B > 0$, by

$$\frac{H(\mu_N | \nu_\alpha^N)}{BN} + \frac{1}{BN} \log E_{\mathbb{P}_{\nu_\alpha^N}} \left[e^{BN \left| \int_0^t \sum_{x \in \Lambda_N} \Theta_x^- (\eta_{sN^2}(x) - \overrightarrow{\eta}_{sN^2}^{\varepsilon N}(1)) \, ds \right|} \right].$$

We can remove the absolute value inside the exponential since $e^{|x|} \leq e^x + e^{-x}$ and

$$\limsup_{N \to \infty} N^{-1} \log(a_N + b_N) \leq \max \left\{ \limsup_{N \to \infty} N^{-1} \log(a_N), \limsup_{N \to \infty} N^{-1} \log(b_N) \right\}. \tag{122}$$

By (107), the Feynman-Kac's formula and (116), last expression can be estimated from above by

$$\frac{C_0}{B} + t \sup_f \left\{ \sum_{x \in \Lambda_N} \Theta_x^- \langle \eta(x) - \overrightarrow{\eta}^{\varepsilon N}(1), f \rangle_{\nu_\alpha^N} - \frac{N}{4B} \mathcal{D}_N(\sqrt{f}, \nu_\alpha) + \frac{C\kappa}{B} N^{1-\theta} \right\}, \tag{123}$$

where the supremum is carried over all the densities $f(\cdot)$ with respect to ν_α^N.

Now we have to split the sum in x, depending on whether $N - 1 \geq x \geq \varepsilon N$ or $x \leq \varepsilon N - 1$. We start by the first case and we have

$$\begin{aligned} \langle \eta(x) - \overrightarrow{\eta}^{\varepsilon N}(1), f \rangle_{\nu_\alpha^N} &= \frac{1}{\varepsilon N} \sum_{y=1}^{1+\varepsilon N} \int (\eta(x) - \eta(y)) f(\eta) \, d\nu_\alpha^N \\ &= \frac{1}{1+\varepsilon N} \sum_{y=1}^{\varepsilon N} \sum_{z=y}^{x-1} \int (\eta(z+1) - \eta(z)) f(\eta) \, d\nu_\alpha^N. \end{aligned}$$

By writing the previous term as its half plus its half and by performing in one of the terms the change of variables η into $\eta^{z,z+1}$, for which the measure ν_α^N is invariant, we write it as

$$\frac{1}{2\varepsilon N}\sum_{y=1}^{1+\varepsilon N}\sum_{z=y}^{x-1}\int (f(\eta)-f(\eta^{z,z+1}))(\eta(z+1)-\eta(z))\,d\nu_\alpha^N.$$

By using the fact that $(a-b)=(\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})$ for any $a,b\geq 0$ and since $ab\leq \frac{Aa^2}{2}+\frac{b^2}{2A}$ for all $A>0$, we have that

$$\begin{aligned}
&\sum_{x=\varepsilon N}^{N-1}\Theta_x^-\langle \eta(x)-\overrightarrow{\eta}^{\varepsilon N}(1),f\rangle_{\nu_\alpha^N}\\
&\leq \frac{A}{2}\sum_{x=\varepsilon N}^{N-1}\frac{\Theta_x^-}{2\varepsilon N}\sum_{y=1}^{1+\varepsilon N}\sum_{z=y}^{x-1}\int(\sqrt{f(\eta)}-\sqrt{f(\eta^{z,z+1})})^2 d\nu_\alpha^N\\
&+\frac{1}{2A}\sum_{x=\varepsilon N}^{N-1}\frac{\Theta_x^-}{2\varepsilon N}\sum_{y=1}^{1+\varepsilon N}\sum_{z=y}^{x-1}\int(\sqrt{f(\eta)}+\sqrt{f(\eta^{z,z+1})})^2(\eta(z+1)-\eta(z))^2 d\nu_\alpha^N.
\end{aligned}\tag{124}$$

By neglecting the jumps of size bigger than one, we see that

$$\sum_{z\in\Lambda_N}\int\left(\sqrt{f(\eta)}-\sqrt{f(\eta^{z,z+1})}\right)^2 d\nu_\alpha^N\ \leq C\,\mathcal{D}_{N,0}(\sqrt{f},\nu_\alpha^N).$$

Therefore, by using also (89), the first term at the right hand side of (124) can be bounded from above by

$$\frac{A}{4}\sum_{x=\varepsilon N}^{N-1}\Theta_x^-\sum_{z\in\Lambda_N}\int\left(\sqrt{f(\eta)}-\sqrt{f(\eta^{z,z+1})}\right)^2\ \leq CA\mathcal{D}_{N,0}(\sqrt{f},\nu_\alpha^N).\tag{125}$$

Recall (116) and observe that

$$\mathcal{D}_N(\sqrt{f},\nu_\alpha^N)\geq\mathcal{D}_{N,0}(\sqrt{f},\nu_\alpha^N).$$

Then we choose the constant A in the form $A=CN/B$ for some constant C. Moreover, for this choice of A, we can bound from above the last term at the right hand side of (124) by (use Lemma 5)

$$\begin{aligned}
&\frac{B}{N}\sum_{x=\varepsilon N}^{N-1}\Theta_x^-\,\frac{1}{2\varepsilon N}\sum_{y=1}^{\varepsilon N}\sum_{z=y}^{x-1}\int(\sqrt{f(\eta)}+\sqrt{f(\eta^{z,z+1})})^2(\eta(z+1)-\eta(z))^2 d\nu_\alpha^N\\
&\qquad\leq C\frac{B}{N}\sum_{x\in\Lambda_N}x\Theta_x^-
\end{aligned}\tag{126}$$

which vanishes as $N \to \infty$ by (116). Note that the previous result holds for any $\varepsilon > 0$. Now we analyse the case when $x \leq \varepsilon N - 1$. In that case, we write

$$\begin{aligned}\langle \eta(x) - \overrightarrow{\eta}^{\varepsilon N}(1), f\rangle_{\nu_\alpha^N} &= \frac{1}{1+\varepsilon N}\sum_{y=1}^{\varepsilon N}\int (\eta(x)-\eta(y))f(\eta)\, d\nu_\alpha^N \\ &= \frac{1}{\varepsilon N}\sum_{y=1}^{x-1}\sum_{z=y}^{x-1}\int (\eta(z+1)-\eta(z))f(\eta)\, d\nu_\alpha^N \\ &\quad - \frac{1}{\varepsilon N}\sum_{y=x+1}^{1+\varepsilon N}\sum_{z=x}^{y-1}\int (\eta(z+1)-\eta(z))f(\eta)\, d\nu_\alpha^N.\end{aligned}$$

and the same estimates as before give that there exists a constant $C > 0$ such that for any $A > 0$,

$$\sum_{x=1}^{\varepsilon N-1} \Theta_x^- \langle \eta(x) - \overrightarrow{\eta}^{\varepsilon N}(1), f\rangle_{\nu_\alpha^N} \leq C\left[AD_N(\sqrt{f}, \nu_\alpha^N) + \frac{\varepsilon N}{A}\sum_{x=1}^{\varepsilon N-1}\Theta_x^-\right].$$

Recall (116) and (89). Then, we choose $A = N/\,8CB$ and the result follows. □

Remark 17. We note that above, if we change in the statement of the lemma Θ_x^- by r_N^- (resp. Θ_x^+ by r_N^+) then the same result holds by performing exactly the same estimates as above, because what we need is that

$$\sum_{x\in\Lambda_N}\Theta_x^\pm < +\infty \quad \text{and} \quad \frac{1}{N}\sum_{x\in\Lambda_N} x\Theta_x^\pm \to_{N\to+\infty} 0 \tag{127}$$

which also holds for $r_N^\pm$ instead of $\Theta_x^\pm$ since $\gamma > 2$.

Remark 18. Let us see now what the previous lemma says when $p(1) = p(-1) = \frac{1}{2}$. In this case we note that we have the same estimate as in (121), see 2. in Remark 16 and also note that $\Theta_x^- \neq 0$ for $x = 1$ and $\Theta_x^- = 0$ for $x \neq 1$. Moreover, $\Theta_1^- = p(1) = \frac{1}{2}$, so that the result above reads as

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t (\eta_{sN^2}(1) - \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1))\, ds\right|\right] = 0.$$

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t (\eta_{sN^2}(N-1) - \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1))\, ds\right|\right] = 0.$$

A.4 Fixing the Profile at the Boundary

Let $\mathbb{Q}$ be a limit point of the sequence $\{\mathbb{Q}_N\}_{N\geq 1}$ and assume, without lost of generality, that $\{\mathbb{Q}_N\}_{N\geq 1}$ converges to $\mathbb{Q}$, as $N \to +\infty$. In this section we prove that for the model of Sect. 3 if $\theta \in [1-\gamma, 1)$ (and also for the model of Sect. 2 when $\theta < 0$) that the profile satisfies $\rho_t(0) = \alpha$ and $\rho_t(1) = \beta$ for $t \in (0, T]$ a.e. We present the proof for $\rho_t(0) = \alpha$ but the other case is completely analogous.

Recall (49). Observe that

$$E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)-\alpha)\,ds\Big|\right] = E_{\mathbb{Q}_N}\left[\Big|\int_0^t (\langle \pi_s,\iota^0_\varepsilon\rangle-\alpha)\,ds\Big|\right]$$

where $\iota^0_\varepsilon(\cdot)=\varepsilon^{-1}\,\mathbf{1}_{(0,\varepsilon)}(\cdot)$. Therefore we have that for any $\delta>0$,

$$\mathbb{Q}_N\left[\Big|\int_0^t (\langle \pi_s,\iota^0_\varepsilon\rangle-\alpha)\,ds\Big|>\delta\right] \;\le\; \delta^{-1}\,E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)-\alpha)\,ds\Big|\right].$$

Note that ι^0_ε is not a continuous function so the set

$$\Big\{\pi\,;\,\Big|\int_0^t (\langle \pi_s,\iota^0_\varepsilon\rangle-\alpha)\,ds\Big|>\delta\Big\}$$

is not an open set in the Skorohod topology, but, a simple argument as we did in Sect. 2.10 allows to overcome the problem. Therefore, by Portemanteau's Theorem we conclude that

$$\mathbb{Q}\left[\Big|\int_0^t (\langle \pi_s,\iota^0_\varepsilon\rangle-\alpha)\,ds\Big|>\delta\right] \;\le\; \delta^{-1}\,\liminf_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)-\alpha)\,ds\Big|\right].$$

Now, if we are able to prove that the right hand side of the previous inequality is zero, since we have that $\mathbb{Q}$ a.s. $\pi_s(dq)=\rho_s(q)dq$ with $\rho_s(\cdot)$ a continuous function in 0 for a.e. s, by taking the limit $\varepsilon\to 0$, we can deduce that $\mathbb{Q}$ a.s. $\rho_s(0)=\alpha$ for s a.e. The result follows from the next lemma.

Lemma 8. *For any $t\in[0,T]$ we have that*

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)-\alpha)\,ds\Big|\right]=0,$$

$$\lim_{\varepsilon\to 0}\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1)-\beta)\,ds\Big|\right]=0.$$

To prove last lemma we use a two step procedure. First we replace, when integrated in time, $\eta_{sN^2}(1)$ by α and then we replace $\eta_{sN^2}(1)$ by $\overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1)$. This is the content of the next two lemmas.

Lemma 9. *For $\gamma>1$, for $1-\gamma\le\theta<1$ and for $t\in[0,T]$ we have that*

$$\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\eta_{sN^2}(1)-\alpha)\,ds\Big|\right]=0,$$

$$\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\Big|\int_0^t (\eta_{sN^2}(N-1)-\beta)\,ds\Big|\right]=0.$$

Proof. We give the proof for the first display, but we note that for the other one it is similar. Fix a Lipschitz profile $\rho(\cdot)$ such that $\alpha=\rho(0)\le\rho(\cdot)\le\rho(1)=\beta$ and $\rho(\cdot)$ is $\frac{\gamma}{2}$-Hölder at the boundary. From (114) that we know that

$$\frac{N}{B}\langle \mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \le -\frac{N}{4B}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)})+\frac{C}{B}\sigma^2+\frac{C\kappa\log(N)}{BN^{\gamma+\theta+1}}. \tag{128}$$

By the entropy inequality, for any $B > 0$, the previous expectation is bounded from above by

$$\frac{H(\mu_N|\nu^N_{\rho(\cdot)})}{BN} + \frac{1}{BN}\log E_{\mathbb{P}_{\nu^N_{\rho(\cdot)}}}\left[e^{BN\left|\int_0^t(\eta_{sN^2}(1)-\alpha)\,ds\right|}\right].$$

By (107), Jensen's inequality and the Feynman-Kac's formula and noting, as we did in the last proof, that we can remove the absolute value inside the exponential, last display can be estimated from above by

$$\frac{C_0}{B} + t\sup_f\left\{\langle\eta(1)-\alpha, f\rangle_{\nu^N_{\rho(\cdot)}} - \frac{N}{4B}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{B}\sigma^2 + \frac{C\kappa}{BN^{\gamma+\theta-1}}\right\}, \tag{129}$$

where the supremum is carried over all the densities $f(\cdot)$ with respect to $\nu^N_{\rho(\cdot)}$. By Lemma 6, since $\rho(\cdot)$ is $\frac{\gamma}{2}$-Hölder at the boundaries, for any $A > 0$, the first term in the supremum in (129) is bounded from above by

$$C\left[\frac{1}{A}\,I_1^\alpha(\sqrt{f},\nu^N_{\rho(\cdot)}) + A + \frac{1}{N^{\gamma/2}}\right]$$

for some constant $C > 0$ independent of $f(\cdot)$ and A. Moreover from (114), since

$$\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) \geq \mathcal{D}_{N,b}(\sqrt{f},\nu^N_{\rho(\cdot)})$$

and $\gamma+\theta-1 > 0$, by choosing $A = 4C(p(1))^{-1}BN^{\theta-1}$, we get then that the expression inside the brackets in (129) is bounded from above by

$$4C^2\frac{BN^{\theta-1}}{p(1)} + \frac{C}{N^{\gamma/2}} + \frac{C}{B}.$$

Now if $p(1) \neq 0$, then the proof follows by sending first $N \to \infty$ and then $B \to \infty$. For $\gamma+\theta-1 = 0$ the same proof as above holds, the only difference is that we use a Lipschitz profile $\rho(\cdot)$ such that $\alpha = \rho(0) \leq \rho(\cdot) \leq \rho(1) = \beta$ and $\rho(\cdot)$ is $\frac{\gamma+1}{2}$-Hölder at the boundaries. From (115) that we know that

$$\frac{N}{B}\langle\mathcal{L}_N\sqrt{f},\sqrt{f}\rangle_{\nu^N_{\rho(\cdot)}} \leq -\frac{N}{4B}\mathcal{D}_N(\sqrt{f},\nu^N_{\rho(\cdot)}) + \frac{C}{B}\sigma^2 + \frac{C\kappa}{B}, \tag{130}$$

and with last bound and the previous argument the proof ends.

Remark 19. The previous lemma tells us that for the model of Sect. 2 and for $\theta < 1$ and $t \in [0,T]$ we have that

$$\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t(\eta_{sN^2}(1)-\alpha)\,ds\right|\right] = 0,$$
$$\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t(\eta_{sN^2}(N-1)-\beta)\,ds\right|\right] = 0.$$

Note that the previous proof follows since we have the bound (117) and in this model $p(1) = \frac{1}{2}$.

Remark 20. We note that for the case where $p(1) = 0$ above what we have to do is to use the two step procedure with a point z such that $p(z) \neq 0$, from where we get that:

$$\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t (\eta_{sN^2}(z) - \alpha)\, ds\right|\right] = 0$$

and the same result holds by changing α to β.

Now we prove the second part of the two step procedure.

Lemma 10. *For $1 - \gamma \leq \theta < 1$ and $t > 0$ we have that*

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) - \eta_{sN^2}(1)\, ds\right|\right] = 0,\\
&\lim_{\varepsilon\to 0}\lim_{N\to\infty} E_{\mathbb{P}_{\mu_N}}\left[\left|\int_0^t \overleftarrow{\eta}^{\varepsilon N}_{sN^2}(N-1) - \eta_{sN^2}(N-1)\, ds\right|\right] = 0.
\end{aligned} \tag{131}$$

Proof. We present the proof of the first item, but we note that for the second it is exactly the same. When $\gamma + \theta - 1 > 0$, we fix a Lipcshitz profile $\rho(\cdot)$ such that $\alpha = \rho(0) \leq \rho(\cdot) \leq \rho(1) = \beta$, and $\rho(\cdot)$ is $\frac{\gamma}{2}$-Hölder at the boundaries, when $\gamma + \theta - 1 = 0$, the Holder regularity at the boundary is $\frac{\gamma+1}{2}$. Since we imposed the same conditions as in the previous lemma in the profile $\rho(\cdot)$ then in this case (128) and (130) holds. From now on we suppose that $\gamma + \theta - 1 > 0$, the other case is completely analogous. By the entropy and Jensen's inequalities, for any $B > 0$, the previous expectation is bounded from above by

$$\frac{H(\mu_N|\nu^N_{\rho(\cdot)})}{BN} + \frac{1}{BN}\log E_{\mathbb{P}_{\nu^N_{\rho(\cdot)}}}\left[e^{BN\left|\int_0^t \overrightarrow{\eta}^{\varepsilon N}_{sN^2}(1) - \eta_{sN^2}(1)\, ds\right|}\right].$$

By (107), the Feynman-Kac's formula, and using the same argument as in the proof of the previous lemma, the estimate of the previous expression can be reduced to bound

$$\frac{C_0}{B} + t\sup_f\left\{\frac{1}{\ell}\sum_{y=2}^{\ell+1}|\langle\eta(y) - \eta(1), f\rangle_{\nu^N_{\rho(\cdot)}}| - \frac{N}{4B}\mathcal{D}_N(\sqrt{f}, \nu^N_{\rho(\cdot)}) + \frac{C}{B}\sigma^2 + \frac{C\kappa\log(N)}{BN^{\gamma+\theta-1}}\right\}, \tag{132}$$

where $\ell = \varepsilon N$. As above, the supremum is carried over all the densities $f(\cdot)$ with respect to $\nu^N_{\rho(\cdot)}$. Note that since $y \in \Lambda_N$ we know that

$$\eta(y) - \eta(1) = \sum_{z=1}^{y-1}(\eta(z+1) - \eta(z)).$$

Observe now that

$$\begin{aligned}
\int(\eta(z+1) - \eta(z))f(\eta)d\nu^N_{\rho(\cdot)} =& \frac{1}{2}\int(\eta(z+1) - \eta(z))(f(\eta) - f(\eta^{z,z+1}))d\nu^N_{\rho(\cdot)}\\
&+\frac{1}{2}\int(\eta(z+1) - \eta(z))(f(\eta) + f(\eta^{z,z+1}))d\nu^N_{\rho(\cdot)}.
\end{aligned}$$

By using the fact that for any $a, b \geq 0$, $(a-b) = (\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})$ and Young's inequality, we have, for any positive constant A, that

$$\begin{aligned}
\frac{1}{\ell}\sum_{y=2}^{\ell+1} |\langle \eta(y)-\eta(1), f\rangle_{\nu^N_{\rho(\cdot)}}| \\
&\leq \frac{1}{2A\ell}\sum_{y=2}^{\ell+1}\sum_{z=1}^{y-1}\int (\eta(z+1)-\eta(z))^2 \left(\sqrt{f(\eta)}+\sqrt{f(\eta^{z,z+1})}\right)^2 d\nu^N_{\rho(\cdot)} \\
&+ \frac{A}{2\ell}\sum_{y=2}^{\ell+1}\sum_{z=1}^{y-1}\int \left(\sqrt{f(\eta)}-\sqrt{f(\eta^{z,z+1})}\right)^2 d\nu^N_{\rho(\cdot)} \\
&+ \frac{1}{2\ell}\sum_{y=2}^{\ell+1}\left|\sum_{z=1}^{y-1}\int \big(\eta(z+1)-\eta(z)\big)\left(f(\eta)+f(\eta^{z,z+1})\right) d\nu^N_{\rho(\cdot)}\right|.
\end{aligned} \tag{133}$$

Now, we neglect jumps of size bigger than one as we did below (124), from where we get that the second term on the right hand side of (133) is bounded from above by $CA\mathcal{D}_N(\sqrt{f}, \nu^N_{\rho(\cdot)})$ where C is a positive constant independent of A, ℓ, f. Then, for the choice $A = N(4BC)^{-1}$ and since $\gamma+\theta-1 \geq 0$, we can bound from above (132) by

$$\begin{aligned}
&\frac{2BC}{N\ell}\sum_{y=2}^{\ell+1}\sum_{z=1}^{y-1}\int (\eta(z+1)-\eta(z))^2 \left(\sqrt{f(\eta)}+\sqrt{f(\eta^{z,z+1})}\right)^2 d\nu^N_{\rho(\cdot)} \\
&+\frac{1}{2\ell}\sum_{y=2}^{\ell+1}\left|\sum_{z=1}^{y-1}\int \big(\eta(z+1)-\eta(z)\big)\left(f(\eta)+f(\eta^{z,z+1})\right) d\nu^N_{\rho(\cdot)}\right| + \frac{C'}{B} \\
&\leq C\Big(\frac{B\ell}{N}+\frac{1}{B}+\frac{1}{2\ell}\sum_{y=2}^{\ell+1}\Big|\sum_{z=1}^{y-1}\int \big(\eta(z+1)-\eta(z)\big)\left(f(\eta)+f(\eta^{z,z+1})\right) d\nu^N_{\rho(\cdot)}\Big|\Big)
\end{aligned} \tag{134}$$

for some constant C. For the last inequality we used Lemma 5. Observe that $B\ell/N = B\varepsilon$ vanishes as $\varepsilon \to 0$. It remains to estimate the third term on the right hand side of the last inequality. For that purpose we make a similar computation to the one of Lemma 6 from where we get that

$$\sum_{z=1}^{y-1}\left|\int (\eta(z+1)-\eta(z))(f(\eta)+f(\eta^{z,z+1}))d\nu^N_{\rho(\cdot)}\right| \leq C\sum_{z=1}^{y-1}\left|\rho\Big(\tfrac{z+1}{N}\Big)-\rho\Big(\tfrac{z}{N}\Big)\right|.$$

Since $\rho(\cdot)$ is Lipschitz, by (134), this estimate provides an upper bound for (132) which is in the form of a constant times

$$\frac{B\ell}{N}+\frac{1}{B}+\frac{1}{N\ell}\sum_{y=2}^{\ell+1} y \leq B\varepsilon + B^{-1} + \varepsilon$$

which vanishes, as $\varepsilon \to 0$ and then $B \to \infty$. This ends the proof. □

References

1. Baldasso, R., Menezes, O., Neumann, A., Souza, R.: Exclusion process with slow boundary. J. Stat. Phys. **167**(5), 1112–1142 (2017)
2. Bernardin, C., Gonçalves, P., Jiménez Oviedo, B.: Slow to fast infinitely extended reservoirs for the symmetric exclusion process with long jumps. arxiv.org (2017, to appear in Markov Processes and their Applications)
3. Bernardin, C., Gonçalves, P., Jiménez Oviedo, B.: A microscopic model for a one parameter class of fractional laplacians with Dirichlet boundary conditions. arxiv.org (2018, submitted)
4. Bernardin, C., Jiménez Oviedo, B.: Fractional Fick's law for the boundary driven exclusion process with long jumps. ALEA **14**, 473–501 (2017)
5. De Paula, R.: Porous Medium Model in contact with Reservoirs. PUC-Rio Master Thesis in Mathematics (2017)
6. Derrida, B.: Non-equilibrium steady states: fluctuations and large deviations of the density and of the current. J. Stat. Mech. Theory Exp. (2007)
7. Derrida, B., Evans, M.R., Hakim, V., Pasquier, V.: Exact solution of a 1-d asymmetric exclusion model using a matrix formulation. J. Phys. Math. Gen. Phys. (1993)
8. Farfan, J., Landim, C., Mourragui, M.: Hydrostatics and dynamical large deviations of boundary driven gradient symmetric exclusion processes. Stoch. Process. Appl. **121**, 725–758 (2011)
9. Franco, T., Gonçalves, P., Neumann, A.: Non-equilibrium and stationary fluctuations for a slowed boundary symmetric exclusion process. Markov Process. Relat. Fields **129**(4), 1413–1442 (2019)
10. Franco, T., Gonçalves, P., Neumann, A.: Hydrodynamical behavior of symmetric exclusion with slow bonds. Ann. Inst. Henri Poincaré Probab. Stat. **49**(2), 402–427 (2013)
11. Franco, T., Gonçalves, P., Schutz, G.: Scaling limits for the exclusion process with a slow site. Stoch. Process. Appl. **126**(3), 800–831 (2016)
12. Jara, M.: Hydrodynamic limit of particle systems with long jumps. arXiv:0805.1326 (2008)
13. Jara, M.: Nonequilibrium scaling limit for a tagged particle in the simple exclusion process with long jumps. Comm. Pure Appl. Math. **62**(2), 198–214 (2009)
14. Jiménez Oviedo, B., Vavasseur, A.: Hydrostatic limit and Fick's law for the symmetric exclusion with long jumps. In: Particle Systems and Partial Differential Equations V. Springer Proceedings in Mathematics and Statistics (2018)
15. Kipnis, C., Landim, C.: Scaling Limits of Interacting Particle Systems. Springer, New York (1999)
16. Landim, C., Milanés, A., Olla, S.: Stationary and nonequilibrium fluctuations in boundary driven exclusion processes. Markov Processes Related Fields **14**(2), 165–184 (2008)
17. Liggett, T.: Interacting Particles Systems. Springer, New York (1985)
18. Liggett, T.: Stochastic Interacting Systems: Contact, Voter and Exclusion Processes. Springer, New York (1999)
19. Spitzer, F.: Interaction of Markov processes. Adv. Math. **5**(2), 246–290 (1970)
20. Spohn, H.: Large scale Dynamics of Interacting Particles. Springer, Berlin (1991)

Stochastic Solutions to Hamilton-Jacobi Equations

Fraydoun Rezakhanlou(✉)

Department of Mathematics, UC Berkeley, Berkeley, USA
rezakhan@math.berkeley.edu

Abstract. In this expository paper we give an overview of the statistical properties of Hamilton-Jacobi Equations and Scalar Conservation Laws. The first part (Sects. 2–4) is devoted to the recent proof of Menon-Srinivasan Conjecture. This conjecture provides a Smoluchowski-type kinetic equation for the evolution of a Markovian solution of a scalar conservation law with convex flux. In the second part of the paper (Sects. 5 and 6) we discuss the question of homogenization for Hamilton-Jacobi PDEs and Hamiltonian ODEs with deterministic and stochastic Hamiltonian functions.

1 Introduction

The primary goal of these notes is to give an overview of the statistical properties of solutions to the Cauchy problem for the Hamilton-Jacobi Equation

$$\begin{aligned} u_t &= H(x,t,u_x) \quad && \text{in} \quad \mathbb{R}^d \times (0,\infty) \\ u &= u^0 && \text{on} \quad \mathbb{R}^d \times \{t=0\}, \end{aligned} \tag{1.1}$$

or, the scalar conservation law

$$\begin{aligned} \rho_t &= H(x,t,\rho)_x \quad && \text{in} \quad \mathbb{R} \times (0,\infty) \\ \rho &= \rho^0 && \text{on} \quad \mathbb{R} \times \{t=0\}, \end{aligned} \tag{1.2}$$

where either H or $\rho^0 = \rho^0(x)$ is random. Note that if u satisfies (1.1) and $d = 1$, then $\rho = u_x$ satisfies (1.2). As is well-known, the PDE (1.1) or (1.2) does not possess classical solutions even when the initial data is smooth. In the case of Eq. (1.1), we may consider *viscosity solutions* to guarantee the uniqueness under some standard assumptions on the initial data and H. In the case of (1.2) with $d = 1$, we consider the so-called *entropy solutions*.

We will be mostly concerned with the following two scenarios:

(1) $d = 1$, $H(x,t,p) = H(p)$ is convex in p and independent of (x,t), with initial data ρ^0 that is either a white noise, or a Markov process.
(2) $d \geq 1$, and $H(x,t,p)$ is a stationary ergodic process in (x,t), and may not be convex in p.

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 206–238, 2019.
https://doi.org/10.1007/978-3-030-15096-9_5

Our aim is to give an overview of various classical and recent results and formulate a number of open problems. Sections 2–4 are devoted to **(1)**, where we derive an evolution equation for the Markovian law of ρ as a function of x or t. Sections 4 and 5 are devoted to **(2)**, where we address the question of *homogenization* for such Hamiltonian functions.

2 Scalar Conservation Law with Random Initial Data

We first recall the following important features of the solutions to (1.2) when $d = 1$, $H(x, t, p) = H(p)$ is convex in p, and independent of (x, t):

(i) If a discontinuity of ρ occurs at $x = x(t)$, and $\rho_\pm = \rho(x(t)\pm, t)$ represent the left and right limits of ρ at $x(t)$, then for a *weak solution* of (1.2) we must have the Rankin-Hugoniot Equation:

$$\frac{dx}{dt} = -H[\rho_-, \rho_+] =: -\frac{H(\rho_+) - H(\rho_-)}{\rho_+ - \rho_-}.$$

(ii) By an entropy solution, we mean a week solution for which the entropy condition is satisfied. In the case of convex H, the entropy condition is equivalent to the requirement

$$\rho_- < \rho_+.$$

(iii) If ρ^0 has a discontinuity with $\rho_- > \rho_+$, then such a discontinuity disappears instantaneously by inserting a rarefaction wave between ρ_- and ρ_+. That is a solution of the form

$$G\left(\frac{x-c}{t}\right),$$

where $G = (H')^{-1}$.

We next state three results.

(i) *(Burgers Equation with Lévy Initial Data)*

When $H(p) = \frac{1}{2}p^2$, (1.2) is the well-known inviscid Burgers' equation, which has often been considered with random initial data. Burgers studied (1.2) in his investigation of turbulence [5]. Carraro and Duchon [6] defined a notion of *statistical solution* to Burgers' equation and realized that it was natural to consider Lévy process initial data. This statistical solution approach was further developed in 1998 by the same authors [7] and by Chabanol and Duchon [8]. In fact any (random) entropy solution is also a statistical solution, but the converse is not true in general. In 1998, Bertoin [4] proved a closure theorem for Lévy initial data.

Theorem 1. *Consider Burgers' equation with initial data $\rho^0(x)$ which is a Lévy process without negative jumps for $x \geq 0$, and $\rho^0(x) = 0$ for $x < 0$. Assume that the expected value of $\rho^0(1)$ is non-positive, $\mathbb{E}\rho^0(1) \leq 0$. Then, for each fixed $t > 0$, the process $x \mapsto \rho(x,t) - \rho(0,t)$ is also a Lévy process with*

$$\mathbb{E}\exp\big(-s(\rho(x,t)-\rho(0,t))\big) = \exp(x\psi(s,t)),$$

where the exponent ψ solves the following equation:

$$\psi_t + \psi\psi_s = 0. \tag{2.1}$$

Remark 2.1(i) The requirement $\mathbb{E}\rho^0(1) \leq 0$ can be relaxed with minor modifications to the theorem, in light of the following elementary fact. Suppose that $\rho^0(x)$ and $\hat{\rho}^0(x)$ are two different initial conditions for Burgers' equation, which are related by $\hat{\rho}^0(x) = \rho^0(x) + cx$. It is easy to check that the corresponding solutions $\rho(x,t)$ and $\hat{\rho}(x,t)$ are related for $t > 0$ by

$$\hat{\rho}(x,t) = \frac{1}{1+ct}\left[\rho\left(\frac{x}{1+ct}, \frac{t}{1+ct}\right) + cx\right].$$

Using this we can adjust a statistical description for a case where $\mathbb{E}\rho^0(1) > 0$ to cover the case of a Lévy process with general mean drift.

(ii) Sinai [26] and Aurell, Frisch, She [3] considered Burgers equation with Brownian motion initial data, relating the statistics of solutions to convex hulls and addressing pathwise properties of solutions. □

(ii) *(Burgers Equation with white noise initial data)*

Groeneboom [15] considers the white noise initial data. In other words, take two independent Brownian motions $B^{\pm}$, and take a two sided Brownian motion for the initial data

$$u^0(x) = \begin{cases} B^+(x) & \text{if} \quad x \geq 0 \\ B^-(x) & \text{if} \quad x \leq 0, \end{cases} \tag{2.2}$$

Theorem 2. *Let $\rho = u_x$, where u is a viscosity solution of the PDE $u_t = \frac{1}{2}u_x^2$, subject to the initial condition $u(x,0) = u^0(x)$, with u^0 given as in (2.2). Then the process $x \mapsto \rho(x,t)$ is a Markov jump process with drift $-t^{-1}$ and a suitable jump measure $\nu(t,\rho_-,\rho_+)\, d\rho_+$.*

We also refer to [13] for an explicit and simple formula expressing the one-point distribution of ρ in terms of Airy functions.

(iii) A different particular case,

$$-H(p) = \begin{cases} 0 & \text{if} \quad |p| \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

corresponds to the problem of determining Lipschitz minorants, and has been investigated by Abramson and Evans [1].

3 Menon-Srinivasan Conjecture

In 2007 Menon and Pego [19] used the Lévy-Khintchine representation for the Laplace exponent and observed that the evolution according to Burgers' equation in (2.1) corresponds to a Smoluchowski coagulation equation [2], with additive collision kernel, for the jump measure of the Lévy process $\nu(\cdot, t)$. The jumps of $\nu(\cdot, t)$ correspond to shocks in the solution $\rho(\cdot, t)$. Regarding the sizes of the jumps as the usual masses in the Smoluchowski equation, it is plausible that Smoluchowski equation with additive kernel should be relevant.

It is natural to wonder whether this evolution through Markov processes with simple statistical descriptions is specific to the Burgers-Lévy case, or an instance of a more general phenomenon. The biggest step toward understanding the problem for a wide class of H is found in a 2010 paper of Menon and Srinivasan [20]. Here it is shown that when the initial condition ρ^0 is a strong Markov process with positive jumps only, the solution $\rho(\cdot, t)$ remains Markov for fixed $t > 0$. The argument is adapted from that of [4] and both [20] and [4] use the notion of splitting times (due to Getoor [14]) to verify the Markov property according to its bare definition. In the Burgers-Lévy case, the independence and homogeneity of the increments can be shown to survive, from which additional regularity is immediate using standard results about Lévy processes. As [20] points out, without these properties it is not clear whether a Feller process initial condition leads to a Feller process in x at later times. Nonetheless, [20] presents a very interesting conjecture for the evolution of the generator of $\rho(\cdot, t)$, which has a remarkably nice form.

To prepare for the statement of Menon-Srinivasan Conjecture, we first examine the following simple scenario for the solutions of the PDE

$$\rho_t = H(\rho)_x = H'(\rho)\rho_x. \tag{3.1}$$

Imagine that the initial data ρ^0 satisfies an ODE of the form

$$\frac{d\rho^0}{dx}(x) = b^0\big(\rho^0(x)\big), \tag{3.2}$$

for some C^1 function $b^0 : \mathbb{R} \to \mathbb{R}$. We may wonder whether or not this feature of ρ^0 survives at later times. That is, for some function $b(\rho, t)$, we also have

$$\rho_x(x, t) = b\big(\rho(x, t), t\big), \tag{3.3}$$

for $t > 0$. For (3.3) to be consistent with (3.1), observe

$$\rho_t = H'(\rho)\rho_x = H'(\rho)b(\rho, t),$$

and as we calculate mixed derivatives, we arrive at

$$\rho_{xt} = b_\rho(\rho, t)\rho_t + b_t(\rho, t) = b_\rho(\rho, t)H'(\rho)b(\rho, t) + b_t(\rho, t),$$
$$\rho_{tx} = H''(\rho)b(\rho, t)\rho_x + H'(\rho)b_\rho(\rho, t)\rho_x = H''(\rho)b^2(\rho, t) + H'(\rho)b_\rho(\rho, t)b(\rho, t).$$

As a result b must satisfy

$$b_t(\rho,t) = H''(\rho)b^2(\rho,t). \tag{3.4}$$

For a classical solution, all we need to do is solving the ODE (3.3) for the initial data $b(\rho,0) = b^0(\rho)$ for each ρ. When H is convex, the solution may blow up in finite time. More precisely,

- If $b^0(\rho) \le 0$, then $b^0(\rho) \le b(\rho,t) \le 0$ for all t and there would be no blow-up.
- If $b^0(\rho) > 0$, then there exists some finite $T(\rho) > 0$ such that $b(\rho,t)$ is finite in the interval $[0,T(\rho))$, and $b(\rho,T(\rho)) = \infty$.

In fact the Eq. (3.4) is really "the method of characteristics" in disguise, and the blow-up of solutions is equivalent to the occurrence of shock discontinuity.

To go beyond what (3.4) offers, we now take a jump kernel $f^0(\rho,d\rho_*)$ and assume that $\rho^0(x)$ is a realization of a Markov process with infinitesimal generator

$$\mathcal{L}^0 h(\rho) = b^0(\rho)h'(\rho) + \int_\rho^\infty \big(h(\rho_*) - h(\rho)\big)\ f^0(\rho,d\rho_*).$$

In words, ρ^0 solves the ODE (3.3), with some occasional random jumps with rate f^0. We are assuming that the jumps are all positive to avoid rarefaction waves. We may wonder whether the same picture is valid at later times. That is, for fixed $t > 0$, the solution $\rho(x,t)$, as a function of x is a Markov process with the generator

$$\mathcal{L}^t h(\rho) = b(\rho,t)h'(\rho) + \int_\rho^\infty \big(h(\rho_*) - h(\rho)\big)\ f(\rho,d\rho_*,t). \tag{3.5}$$

Menon-Srinivasan Conjecture roughly suggests that if H is convex, and we start with a Markov process with generator $\mathcal{L}^0$, then we have a Markov process at a later time with a generator of the form $\mathcal{L}^t$. Moreover, the drift of the generator satisfies (3.4), and the jump kernel $f(\rho,d\rho_*,t)$ solves an integral equation. Before we derive an equation for the evolution of f, observe that when we assert that $\rho(x,t)$ is a Markov process in x, we are specifying a direction for x. More precisely, we are asserting that if $\rho(a,t)$ is known, then the law of $\rho(x,t)$ can be determined uniquely for all $x > a$. We are doing this for all $t > 0$. In practice, we may try to determine $\rho(x,t)$ for $x > a(t)$, provided that $\rho(a(t),t)$ is specified. For example, we may wonder whether or not we can determine the law of $\rho(x,t)$ with the aid of the following procedure:

- The process $t \mapsto \rho(a(t),t)$ is a Markov process and its generator can be determined. Using this Markov process, we take a realization of $\rho(a(t),t)$, with some initial choice for $\rho(a(0),0)$.
- Once $\rho(a(t),t)$ is selected, we use the generator $\mathcal{L}^t$, to produce a realization of $\rho(x,t)$ for $x \ge a(t)$.

To materialize the above procedure, we need to make sure that for some choice of $a(t)$, the process $\rho(a(t), t)$ is Markovian with a generator that can be described. For a start, we may wonder whether or not we can even choose $a(t) = a$ a constant function. Put it differently, not only $x \mapsto \rho(x, t)$ is a Markov process for fixed $t \geq 0$, the process $t \mapsto \rho(x, t)$ is a Markov process for fixed x. As it turns out, this is the case if H is also increasing. In general, if we can find a negative constant c such that $H'(\rho) > c$, then $\hat{\rho}(x, t) := \rho(x - ct, t)$ satisfies

$$\hat{\rho}_t = \hat{H}(\hat{\rho})_x,$$

for $\hat{H}(\rho) = H(\rho) - c\rho$, which is increasing. Hence, the process $t \mapsto \hat{\rho}(x, t) = \rho(x - ct, t)$ is expected to be Markovian. In summary

- If H is increasing in the range of ρ, then ρ is also Markovian on vertical lines $x = constant$.
- If H' is bounded below by a negative constant c, then ρ is Markovian on straight lines that are titled to the right with the slope $-c$.

To simplify the matter, from now on, we make two assumptions on H:

$$H' > 0, \qquad H'' \geq 0. \tag{3.6}$$

The main consequences of these two assumptions are

- All the jump discontinuities are positive i.e. $\rho_- < \rho_+$.
- The speed of shocks are always negative.

We now argue that in fact the process $t \mapsto \rho(x, t)$ is a (time-inhomogeneous) Markov process with a generator $\mathcal{M}_t$ that is independent of x because the PDE (3.1) is homogeneous (i.e. H is independent of x). Indeed

$$\mathcal{M}_t h(\rho) = H'(\rho) b(\rho, t) h'(\rho) + \int_\rho^\infty \big(h(\rho_*) - h(\rho)\big) H[\rho, \rho_*] f(\rho, d\rho_*, t). \tag{3.7}$$

To explain the form of $\mathcal{M}_t$ heuristically, observe that the ODE $\frac{d\rho}{dx} = b(\rho, t)$ leads to the ODE

$$\frac{d\rho}{dt} = H'(\rho) b(\rho, t).$$

On the other hand, if we fix x, then $\rho(x, t)$ experiences a jump discontinuity when a shock on the right of x crosses x. Given any $t > 0$, a shock would occur at some $s > t$ because all shock speeds are negative; it is just a matter of time for a shock on the right of x to cross x. We can also calculate the rate at which this happens because we have the law of the first shock on the right of x, and its speed. Observe

- The process $x \mapsto \rho(x, t)$ is a homogeneous Markov process with a generator that changes with time.
- The process $t \mapsto \rho(x, t)$ is an inhomogeneous Markov process with a generator that does not depend on x. It is only the initial data $\rho(x, 0)$ that is responsible for the changes of the statistics of $\rho(x, t)$, as x varies.

We are now in a position to derive formally an evolution equation for the generator $\mathcal{L}^t$, under the assumption (3.6). Indeed if we define

$$w(x,t;\rho) = \mathbb{E}^{\rho(0,t)=\rho} h(\rho(x,T)),$$

for $t < T$, then we expect

$$w_t = -\mathcal{M}_t w, \qquad w_x = \mathcal{L}^t w.$$

Differentiating these equations yields

$$w_{tx} = -\mathcal{M}_t w_x = -\mathcal{M}_t \mathcal{L}^t w, \qquad w_{xt} = \frac{d\mathcal{L}^t}{dt} w + \mathcal{L}^t w_t = \frac{d\mathcal{L}^t}{dt} w - \mathcal{L}^t \mathcal{M}_t w.$$

As a result

$$\frac{d\mathcal{L}^t}{dt} = \mathcal{L}^t \mathcal{M}_t - \mathcal{M}_t \mathcal{L}^t. \tag{3.8}$$

As we match the drift parts of both sides of (3.8), we simply get (3.4). Matching the jump parts yields a kinetic-type equation of the form

$$f_t = Q(f,f) + Cf, \tag{3.9}$$

for a quadratic operator Q and a linear operator C. The operator Q is independent of b and is given by

$$\begin{aligned} Q(f,f)(\rho_-, d\rho_+) = &\int_{\rho_-}^{\rho^+} \big(H[\rho_*, \rho_+] - H[\rho_-, \rho_*]\big) f(\rho_-, d\rho_*) f(\rho_*, d\rho_+) \\ &+ \int_{\rho_+}^{\infty} \big(H[\rho_+, \rho_*] - H[\rho_-, \rho_+]\big) f(\rho_+, d\rho_*) f(\rho_-, d\rho_+) \\ &+ \int_{\rho_-}^{\infty} \big(H[\rho_-, \rho_+] - H[\rho_-, \rho_*]\big) f(\rho_-, d\rho_*) f(\rho_-, d\rho_+). \end{aligned}$$

If we set

$$\begin{aligned} \lambda(\rho_-) = \lambda(f)(\rho_-) &= \int_{\rho_-}^{\infty} f(\rho_-, d\rho_+), \\ A(\rho_-) = A(f)(\rho_-) &= \int_{\rho_-}^{\infty} H[\rho_-, \rho_+] f(\rho_-, d\rho_+), \end{aligned}$$

then $Q = Q^+ - Q^-$, with

$$\begin{aligned} Q^+(f,f)(\rho_-, d\rho_+) &= \int_{\rho_-}^{\rho^+} \big(H[\rho_*, \rho_+] - H[\rho_-, \rho_*]\big) f(\rho_-, d\rho_*) f(\rho_*, d\rho_+, t) \\ Q^-(f,f)(\rho_-, d\rho_+) &= \big\{A(\rho_+) - A(\rho_-) - H[\rho_-, \rho_+]\big(\lambda(\rho_+) - \lambda(\rho_-)\big)\big\} f(\rho_-, d\rho_+). \end{aligned} \tag{3.10}$$

To define the operator C we need to assume that $f(\rho_-, d\rho_+) = f(\rho_-, \rho_+) d\rho_+$ has a C^1 density. With a slight abuse of notion, we write $f(\rho_-, \rho_+)$ for the density

of the measure $f(\rho_-, d\rho_+)$, and write C again for the action of the operator C on the density f:

$$\begin{aligned}(Cf)(\rho_-,\rho_+) = b(\rho_-,t)f(\rho_-,\rho_+)\big(H[\rho_-,\rho_+]\big)_{\rho_-} \\ + \big[H[\rho_-,\rho_+] - H'(\rho_-)\big]b(\rho_-,t)f_{\rho_-}(\rho_-,\rho_+) \\ + \big[\big(H[\rho_-,\rho_+] - H'(\rho_+)\big)b(\rho_+,t)f(\rho_-,\rho_+)\big]_{\rho_+}.\end{aligned}$$

Menon-Srinivasan Conjecture has been established in [16] and [17]:

Theorem 3. *Assume H is a C^2 function that satisfies (3.6). Let ρ be an entropic solution of (3.1) such that $\rho(x,0) = 0$, for $x \le 0$, and $\rho(x,0)$ is a Markov process with generator $\mathcal{L}^0$, for $x \ge 0$. Assume that b and f satisfy (3.4) and (3.9) respectively. Then the processes $t \mapsto \rho(0,t)$ and $x \mapsto \rho(x,t)$ are Markov processes with generators $\mathcal{M}_t$ and $\mathcal{L}^t$ respectively.*

The typical situation, for Smoluchowski and other kinetic equations is that we have some (stochastic or deterministic) dynamics defined on a finite system, and these kinetic equations emerge upon passage to a scaling limit. The dynamics might not be definable for the infinite system, and the kinetic equation should describe statistics only approximately for a large but finite system. In the setting of Theorems 1, 2 and 3, the kinetic equations give statistics *exactly* without passage to a rescaled limit. We view this unusual circumstance as demanding an explanation. Further, our treatment in Sect. 4 below (tracking shocks as inelastically colliding particles) seems quite at home in the kinetic context.

4 Heuristics for the Proof of Theorem 3

Let us write $x_i(t)$ for the location of the i-th shock and $\rho_i(t) = \rho(x_i(t)+,t)$. We also write $\phi_x(m_0;t)$ for the flow associated with the velocity b; the function $m(x) = \phi_x(m_0;t)$ satisfies

$$m'(x) = b(m(x),t), \qquad m(0) = m_0.$$

We can readily find the evolution $\mathbf{q} = (x_i, \rho_i : i \in \mathbb{Z})$, and $\hat{\boldsymbol{q}} = (z_i, \rho_i : i \in \mathbb{Z})$, with $z_i = x_{i+1} - x_i$:

– $$\dot{x}_i = -v^i := -H[\hat{\rho}_{i-1}, \rho_i], \qquad \dot{z}_i = -(v^{i+1} - v^i),$$
where $\hat{\rho}_{i-1}(t) = \phi_{z_{i-1}}(\rho_{i-1}(t), t)$.

– $$\dot{\rho}_i = w^i := \big(H'(\rho_i) - H[\hat{\rho}_{i-1}, \rho_i]\big)b(\rho_i, t).$$

– When z_i becomes 0, the pair (ρ_i, z_i) is omitted from $\hat{\boldsymbol{q}}(t)$. The outcome after a relabeling is denoted by $\hat{\boldsymbol{q}}^i(t)$.

Write

$$\Delta = \big\{(z_i, \rho_i : i \in \mathbb{Z}) : \ z_i > 0, \ \rho_i \in \mathbb{R} \quad \text{for all} \quad i \in \mathbb{Z}\big\}.$$

We think of $\hat{\boldsymbol{q}}(t)$ as a deterministic process that has an infinitesimal generator

$$\mathcal{A}G = \sum_{i\in\mathbb{Z}} \big(w^i G_{\rho_i} - (v^{i+1} - v^i) G_{z_i}\big),$$

in the interior of Δ. We only take those G such that on the boundary face of Δ with $z_i = 0$, we have $G\big(\hat{\boldsymbol{q}}\big) = G\big(\hat{\boldsymbol{q}}^i\big)$. This stems from the fact that we are interested in the function $\rho(x) = \rho(x; \hat{\mathbf{q}})$ associated with $\hat{\mathbf{q}}$ (or $\mathbf{q}$) that is defined by

$$\sum_i \phi_{z_i}(x_i; x - x_i) \mathbb{1}\big(x \in [x_i, x_{i+1})\big).$$

Note that $\rho(x; \hat{\mathbf{q}}) = \rho(x; \hat{\mathbf{q}}^i)$ whenever $z_i = 0$.

We make an ansatz that the law of $\hat{\boldsymbol{q}}(t)$ is of the form:

$$\mu\big(d\hat{\boldsymbol{q}}, t\big) = \prod_{i=-\infty}^{\infty} e^{-\int_0^{z_i} \lambda(\phi_y(\rho_i;t),t)dy} \ f\big(\phi_{z_i}(\rho_i; t), \rho_{i+1}, t\big) \ dz_i d\rho_{i+1}.$$

For this to be the case, we need to have

$$\dot{\mu} = \mathcal{A}^* \mu. \tag{4.1}$$

This equation should determine f and λ if our ansatz is correct. To determine $\mathcal{A}^*$, we take a test function G and carry out the following calculation: After some integration by parts, we formally have

$$\int G \ d\mathcal{A}^*\mu = \int \mathcal{A}G \ d\mu = \int G \sum_i \big[w^i \Omega_i^1 - w^i_{\rho_i} + (v^{i+1} - v^i)\Omega_i^2 + v^{i+1}_{z_i} - \Omega_i^3\big] \ d\mu,$$

where

$$\Omega_i^1 = \int_0^{z_i} \big[\lambda\big(\phi_y(\rho_i; t), t\big)\big]_{\rho_i} dy - \frac{\big[f\big(\hat{\rho}_i, \rho_{i+1}, t\big)\big]_{\rho_i}}{f\big(\hat{\rho}_i, \rho_{i+1}, t\big)} - \frac{f_{\rho_+}\big(\hat{\rho}_{i-1}, \rho_i, t\big)}{f\big(\hat{\rho}_{i-1}, \rho_i, t\big)},$$

$$\Omega_i^2 = -\lambda\big(\hat{\rho}_i, t\big) + \frac{\big[f\big(\hat{\rho}_i, \rho_{i+1}, t\big)\big]_{z_i}}{f\big(\hat{\rho}_i, \rho_{i+1}, t\big)},$$

$$\Omega_i^3 = \frac{\int_{\hat{\rho}_{i-1}}^{\rho_i} H\big(\hat{\rho}_{i-1}, \rho_*, \rho_i\big) f\big(\hat{\rho}_{i-1}, \rho_*, t\big) f\big(\rho_*, \rho_i, t\big) \ d\rho_*}{f\big(\hat{\rho}_{i-1}, \rho_i, t\big)},$$

where Ω_i^3 represents the boundary contribution associated with $z_i = 0$, and

$$H(a, b, c) := H[b, c] - H[a, b].$$

To explain the form of Ω_i^3, observe that when $z_i = 0$, we remove the ith-particle and relabel the particles to its right. The expression $f\big(\hat{\rho}_{i-1}, \rho_i, t\big) f\big(\rho_i, \rho_{i+1}, t\big) \ d\rho_i$,

that appears in μ, can be rewritten as $f\big(\hat\rho_{i-1},\rho_*,t\big)f\big(\rho_*,\rho_{i+1},t\big)\,d\rho_*$. The variable ρ_{i+1} becomes ρ_i after our relabeling, and its integral with respect to ρ_* is a function of $\big(\hat\rho_{i-1},\rho_i,t\big)$. If we replace this function with $f\big(\hat\rho_{i-1},\rho_i,t\big)$, we recover the measure μ.

On the other hand

$$\dot\mu=\sum_i\left[\Gamma_i^1+\Gamma_i^2\right]\,\mu=\sum_i\left\{\frac{\big[f\big(\phi_{z_i}(\rho_i;t),\rho_{i+1},t\big)\big]_t}{f\big(\phi_{z_i}(\rho_i;t),\rho_{i+1},t\big)}-\int_0^{z_i}\big[\lambda\big(\phi_y(\rho_i;t),t\big)\big]_t dy\right\}\,\mu.$$

To make the above formal calculation rigorous, we switch from the infinite sum to a finite sum. For this, we restrict the dynamics to an interval, say $[0,L]$. The configuration now belongs to

$$\Delta_L=\cup_{n=0}^\infty\Delta_n^L,$$

with Δ_n^L denoting the set

$$\big\{\mathbf{q}=\big((x_i,\rho_i):i=0,1,\dots,n\big):\ x_0=0<x_1<\cdots<x_n<x_{n+1}=L,\quad \rho_0,\dots,\rho_n\in\mathbb{R}\big\}.$$

Again, what we have in mind is that $\rho_i(t)=\rho(x_i(t)+,t)$ with $x_1,\dots,x_n$ denoting the location of all shocks in $(0,L)$. For our purposes, we need to come up with a candidate for the law $\mu(t,d\mathbf{q})$ of $\mathbf{q}(t)$ in Δ_L. The restriction of μ to Δ_L^n is denoted by μ^n and is given by

$$\ell(d\rho_0,t)\exp\left\{-\sum_{i=0}^n\int_0^{x_{i+1}-x_i}\lambda(\phi_y(\rho_i;t),t)dy\right\}\prod_{i=0}^{n-1}f\big(\phi_{x_{i+1}-x_i}(\rho_i;t),\rho_{i+1},t\big)\,dx_{i+1}d\rho_{i+1},$$

where f solves (3.9) and ℓ is the law of $\rho(0,t)$, which is a Markov process with generator $\mathcal{M}=\mathcal{M}_t$:

$$\dot\ell=\mathcal{M}^*\ell. \tag{4.2}$$

To simplify the presentation, we assume

$$\ell(d\rho_0,t)=\ell(\rho_0,t)\,d\rho_0.$$

As for the dynamics of $\mathbf{q}$, we have the following rules:

(i) So long as x_i remains in (x_{i-1},x_{i+1}), it satisfies

$$\dot x_i=-v^i:=-H[\hat\rho_{i-1},\rho_i],$$

where $\hat\rho_{i-1}(t)=\phi_{z_{i-1}}(\rho_{i-1}(t),t)$.

(ii) We have $\dot\rho_0=w^0:=H'(\rho_0)b(\rho_0,t)$ and for $i>0$,

$$\dot\rho_i=w^i:=\big(H'(\rho_i)-H[\hat\rho_{i-1},\rho_i]\big)b(\rho_i,t).$$

(iii) When $z_i=x_{i+1}-x_i$ becomes 0, then $\mathbf{q}(t)$ becomes $\mathbf{q}^i(t)$, that is obtained from $\mathbf{q}(t)$ by omitting (ρ_i,x_i).

(iv) With rate

$$H[\hat{\rho}_n, \rho_{n+1}] f\big(\hat{\rho}_n, \rho_{n+1}, t\big),$$

the configuration $\mathbf{q}$ gains a new particle (x_{n+1}, ρ_{n+1}), with $x_{n+1} = L$. This new configuration is denoted by $\mathbf{q}(\rho_{n+1})$.

We note that since H is increasing, all velocities are negative. Moreover, when the first particle of location x_1 crosses the origin, a particle is lost.

We wish to establish (4.1). We write G^n for the restriction of a smooth function $G : \Delta^L \to \mathbb{R}$ to Δ_n^L. Recall that we only consider those test functions G that cannot differentiate between $\mathbf{q}$ and $\mathbf{q}^i$ (respectively $\mathbf{q}(\rho_{n+1})$), when $x_i = x_{i+1}$ (respectively $x_{n+1} = L$). We need to verify

$$\dot{\mu}^n = \big(\mathcal{A}^* \mu\big)^n, \tag{4.3}$$

for all $n \geq 0$. Here and below, we write ν^n for the restriction of a measure ν to Δ_n^L. Also, given $H : \Delta_L \to \mathbb{R}$, we write H^n for the restriction of the function H to the set Δ_n^L. To verify (4.3), we show

$$\int G^n \, d\dot{\mu}^n = \int \big(\mathcal{A} G\big)^n \, d\mu^n, \tag{4.4}$$

for every C^1 function G. It is instructive to see why (4.3) (or its integrated version (4.4)) is true when $n = 0$ and 1 before treating the general case. As we will see below, the cases $n = 0, 1$ are already equivalent to the Eq. (3.9). As a warm-up, we first assume that $n = 0$ and $b = 0$. In this case the Eq. (4.3) is equivalent to the fact that the law ℓ of $\rho(0, \cdot)$ is governed by a Markov process with generator $\mathcal{M}_t$. The case $n = 0$ and general b leads to the general form of $\mathcal{M}_t$ for the evolution of $\rho(0, \cdot)$, and an equation for λ that is a consequence of (3.9). The full Eq. (3.9) shows up when we consider the case $n = 1$.

The case $n = 0$ and $b = 0$. As it turns out, the function $\lambda(\rho, t) = \lambda(\rho)$ is independent of time when $b = 0$. We simply have

$$\mu^0(d\rho_0, t) = e^{-L\lambda(\rho_0)} \ell(d\rho_0, t), \qquad \mu_t^0(d\rho_0, t) = e^{-\lambda(\rho_0)L} \ell_t(d\rho_0, t). \tag{4.5}$$

On the other hand, the right-hand side of (4.4) is of the form $\Omega_0^1 + \Omega_0^2$, where Ω_0^1 comes from rule **(i)**, and Ω_0^2 comes from the stochastic boundary dynamics. Indeed

$$\begin{aligned} \Omega_0^1 = &\int H[\rho_0, \rho_1] G^0(0, \rho_1) e^{-\lambda(\rho_1)L} f(\rho_0, \rho_1, t) \, d\rho_1 \, \ell(d\rho_0, t) \\ &- \int H[\rho_0, \rho_1] G^1(0, \rho_0, L, \rho_1) e^{-\lambda(\rho_0)L} f(\rho_0, \rho_1, t) \, d\rho_1 \, \ell(d\rho_0, t), \end{aligned} \tag{4.6}$$

which we get from the boundary terms when we apply an integration by parts to the integral

$$- \int H[\rho_0, \rho_1] G^1_{x_1}(0, \rho_0, x_1, \rho_1) \, e^{-\lambda(\rho_0)x_1 - \lambda(\rho_1)(L - x_1)} f(\rho_0, \rho_1, t) \, d\rho_1 \, \ell(d\rho_0, t) \, dx_1.$$

We note that the other terms of the integration by parts formula contribute to the case $n = 1$ and do not contribute to our $n = 0$ case. Moreover,

$$\Omega_0^2 = \int H[\rho_0, \rho_1] f(\rho_0, \rho_1, t) \big(G^1(0, \rho_0, L, \rho_1) - G^0(0, \rho_0)\big) e^{-\lambda(\rho_0)L} \, d\rho_1 \, \ell(d\rho_0, t).$$

From this and (4.6) we learn

$$\begin{aligned}
\Omega_0^1 + \Omega_0^2 &= \int H[\rho_0, \rho_1] G^0(0, \rho_1) e^{-\lambda(\rho_1)L} f(\rho_0, \rho_1, t) \, d\rho_1 \, \ell(d\rho_0, t) \\
&\qquad - \int H[\rho_0, \rho_1] f(\rho_0, \rho_1, t) G^0(0, \rho_0) e^{-\lambda(\rho_0)L} \, d\rho_1 \, \ell(d\rho_0, t) \\
&= \int H[\rho_1, \rho_0] G^0(0, \rho_0) e^{-\lambda(\rho_0)L} f(\rho_1, \rho_0, t) \, d\rho_0 \, \ell(d\rho_1, t) \\
&\qquad - \int H[\rho_0, \rho_1] f(\rho_0, \rho_1, t) G^0(0, \rho_0) e^{-\lambda(\rho_0)L} \, d\rho_1 \, \ell(d\rho_0, t) \\
&= \int G^0(0, \rho_0) e^{-\lambda(\rho_0)L} \, \big(\mathcal{M}_t^* \ell\big)(d\rho_0, t) = \int G^0(0, \rho_0) e^{-\lambda(\rho_0)L} \, \ell_t(d\rho_0, t),
\end{aligned}$$

as desired. □

The case $n = 0$ and general b. To ease the notation, we write

$$\Gamma(\rho, x, t) = \int_0^x \lambda(\phi_y(\rho; t), t) \, dy.$$

When $n = 0$, the right-hand side of (4.4) equals

$$\begin{aligned}
&\int G^0(0, \rho_0) \left[H'(\rho_0) b(\rho_0, t) \ell(\rho_0, t) \Gamma_\rho(\rho_0, L, t) - \big(H'(\rho_0) b(\rho_0, t) \ell(\rho_0, t)\big)_{\rho_0}\right] \, e^{-\Gamma(\rho_0, L, t)} \, d\rho_0 \\
&\quad + \int H[\rho_0, \rho_1] G^0(0, \rho_1) e^{-\Gamma(\rho_1, L, t)} f(\rho_0, \rho_1, t) \, d\rho_1 \, \ell(d\rho_0, t) \\
&\quad - \int H\big[\phi_L(\rho_0; t), \rho_1\big] f\big(\phi_L(\rho_0; t), \rho_1, t\big) G^1(0, \rho_0, L, \rho_1) e^{-\Gamma(\rho_0, L, t)} \, d\rho_1 \, \ell(d\rho_0, t) \\
&\quad + \int H\big[\phi_L(\rho_0; t), \rho_1\big] f\big(\phi_L(\rho_0; t), \rho_1, t\big) \big(G^1(0, \rho_0, L, \rho_1) - G^0(0, \rho_0)\big) e^{-\Gamma(\rho_0, L, t)} \, d\rho_1 \, \ell(d\rho_0, t)
\end{aligned}$$

This simplifies to

$$\begin{aligned}
&\int G^0(0, \rho_0) \left[H'(\rho_0) b(\rho_0, t) \ell(\rho_0, t) \Gamma_\rho(\rho_0, L, t) - \big(H'(\rho_0) b(\rho_0, t) \ell(\rho_0, t)\big)_{\rho_0}\right] \, e^{-\Gamma(\rho_0, L, t)} \, d\rho_0 \\
&\quad + \int H[\rho_*, \rho_0] G^0(0, \rho_0) e^{-\Gamma(\rho_0, L, t)} f(\rho_*, \rho_0, t) \, d\rho_0 \, \ell(d\rho_*, t) \\
&\quad - \int H\big[\phi_L(\rho_0; t), \rho_1\big] f\big(\phi_L(\rho_0; t), \rho_1, t\big) G^0(0, \rho_0) e^{-\Gamma(\rho_0, L, t)} \, d\rho_1 \, \ell(d\rho_0, t) \\
&\qquad = \int G^0(0, \rho_0) \Lambda(\rho_0, t) \, e^{-\Gamma(\rho_0, L, t)} \, d\rho_0,
\end{aligned}$$

where $\Lambda(\rho_0, t)$ equals

$$\begin{aligned}
&H'(\rho_0) b(\rho_0, t) \ell(\rho_0, t) \Gamma_\rho(\rho_0, L, t) - \big(H'(\rho_0) b(\rho_0, t) \ell(\rho_0, t)\big)_{\rho_0} \\
&\quad + \int H[\rho_*, \rho_0] f(\rho_*, \rho_0, t) \, \ell(d\rho_*, t) - \int H\big[\phi_L(\rho_0; t), \rho_1\big] f\big(\phi_L(\rho_0; t), \rho_1, t\big) \, d\rho_1 \, \ell(\rho_0, t).
\end{aligned}$$

We need to match $\Lambda(\rho_0, t)$ with the corresponding term on left-hand side of (4.4), which, by (4.2) takes the form

$$-\Gamma_t(\rho_0, L, t)\ \ell(\rho_0, t) - \big(H'(\rho_0)b(\rho_0, t)\ell(\rho_0, t)\big)_{\rho_0} + \int H[\rho_*, \rho_0]f(\rho_*, \rho_0, t)\ \ell(d\rho_*, t) - A(\rho_0, t)\ \ell(\rho_0, t),$$

where

$$A(\rho_0, t) = \int H[\rho_0, \rho_*]f(\rho_0, \rho_*, t)\ d\rho_*.$$

We are done if we can verify

$$\Gamma_t(\rho_0, L, t) + H'(\rho_0)b(\rho_0, t)\Gamma_\rho(\rho_0, L, t) = A\big(\phi_L(\rho_0; t), t\big) - A(\rho_0, t). \tag{4.7}$$

Equivalently

$$\int_0^L \big[\lambda(\phi_y(\rho_0; t), t)\big]_t\ dy + H'(\rho_0)b(\rho_0, t)\int_0^L \big[\lambda(\phi_y(\rho_0; t), t)\big]_{\rho_0}\ dy = \int_0^L \big[A(\phi_y(\rho_0; t), t)\big]_y\ dy.$$

For this, it suffices to check

$$\big[\lambda(\phi_y(\rho_0; t), t)\big]_t + H'(\rho_0)b(\rho_0, t)\ \big[\lambda(\phi_y(\rho_0; t), t)\big]_{\rho_0} = \big[A\big(\phi_y(\rho_0; t), t\big)\big]_y.$$

Note that if $u(y, \rho) = A\big(\phi_y(\rho; t), t\big)$, then

$$u_y(y, \rho_0) = b(\rho_0, t)u_\rho(y, \rho_0).$$

Hence for (4.7), it suffices to show

$$\big[\lambda(\phi_y(\rho_0; t), t)\big]_t + H'(\rho_0)b(\rho_0, t)\big[\lambda(\phi_y(\rho_0; t), t)\big]_{\rho_0} = b(\rho_0, t)\big[A\big(\phi_y(\rho_0; t), t\big)\big]_{\rho_0}. \tag{4.8}$$

To have a more tractable formula, let us write $T_y h(m) = h(\phi_y(m; t))$. The family of operators $\{T_y :\ y \in \mathbb{R}\}$, is a group in y. Moreover, if $(\mathcal{B}h)(m) = b(m, t)h'(m)$, then

$$\frac{dT_y}{dy} = \mathcal{B}T_y = T_y\mathcal{B}. \tag{4.9}$$

Using this, we may rewrite (4.8) as

$$\big[\lambda(\phi_y(\rho_0; t), t)\big]_t + H'(\rho_0)b(\phi_y(\rho_0; t), t)\lambda_\rho(\phi_y(\rho_0; t), t) = b(\phi_y(\rho_0; t), t)A_\rho\big(\phi_y(\rho_0; t), t\big). \tag{4.10}$$

This for $y = 0$ takes the form

$$\lambda_t(\rho_0, t) + H'(\rho_0)b(\rho_0, t)\lambda_\rho(\rho_0, t) = b(\rho_0, t)A_\rho(\rho_0, t). \tag{4.11}$$

Because of our choice of λ, namely

$$\lambda(t, \rho-) = \int_{\rho-}^{\infty} f(\rho_-, \rho_+, t)\ d\rho_+, \tag{4.12}$$

we can deduce (4.10) from (3.9) after integrating both sides of (3.9) with respect to ρ_-. On account of (4.11), the claim (4.10) would follow if we can show

$$X(\rho_0, y, t) := \big[\phi_y(\rho_0; t)\big]_t - \big[H'(\phi_y(\rho_0; t)) - H'(\rho_0)\big] b(\phi_y(\rho_0; t), t) = 0. \quad (4.13)$$

This is true for $y = 0$. Differentiating with respect to y yields

$$\begin{aligned} X_y(\rho_0, y, t) &= \big[b(\phi_y(\rho_0; t), t)\big]_t - \big[H'(\phi_y(\rho_0; t))\big]_y b(\phi_y(\rho_0; t), t) \\ &\qquad - \big[H'(\phi_y(\rho_0; t)) - H'(\rho_0)\big]\big[b(\phi_y(\rho_0; t), t)\big]_y \\ &= b_t(\phi_y(\rho_0; t), t) + b_\rho(\phi_y(\rho_0; t), t)\big[\phi_y(\rho_0; t)\big]_t - H''(\phi_y(\rho_0; t))b^2(\phi_y(\rho_0; t), t) \\ &\qquad - \big[H'(\phi_y(\rho_0; t)) - H'(\rho_0)\big](bb_\rho)(\phi_y(\rho_0; t), t) \\ &= b_\rho(\phi_y(\rho_0; t), t)\big[\phi_y(\rho_0; t)\big]_t - \big[H'(\phi_y(\rho_0; t)) - H'(\rho_0)\big](bb_\rho)(\phi_y(\rho_0; t), t) \\ &= b_\rho(\phi_y(\rho_0; t), t) X(\rho_0, y, t), \end{aligned}$$

where we used (3.4) for the third equality. As a result.

$$X(\rho_0, y, t) = X(\rho_0, 0, t) \exp\left[\int_0^y b_\rho(\phi_z(\rho_0; t), t)\, dz\right] = 0.$$

This completes the proof of (4.2), when $n = 0$. □

As we have seen so far, the case $n = 0$ is valid if (4.11), a consequence of the kinetic equation (3.9), is true. On the other hand the case $n = 1$ is equivalent to the kinetic equation. Before embarking on the verification of (4.3) for $n = 1$, let us make some compact notions for some of the expressions that come into the proof. Given a realization $\mathbf{q} = \big(0, \rho_0, x_1, \rho_1, \ldots, x_n, \rho_n\big) \in \Delta_n^L$, we define

$$\begin{aligned} \rho\big(x, t; \mathbf{q}\big) &= \sum_{i=0}^{n} \phi_{x-x_i}\big(\rho_i; t\big) \mathbb{1}\big(x_i \le x < x_{i+1}\big), \\ \Gamma(\mathbf{q}, t) &= \int_0^L \lambda\big(\rho\big(y, t; \mathbf{q}\big)\big)\, dy = \sum_{i=0}^{n} \Gamma(\rho_i, x_{i+1} - x_i, t), \\ \hat{\rho}_{i-1} &= \rho(x_i-, t; \mathbf{q}) = \phi_{x_i - x_{i-1}}(\rho_{i-1}; t), \end{aligned}$$

where λ is defined by (4.12). Note that by (4.13),

$$\frac{d\hat{\rho}_i}{dt} = \big[H'(\hat{\rho}_i) - H'(\rho_i)\big] b(\hat{\rho}_i, t). \quad (4.14)$$

The case $n = 1$**.** We have $\dot{\mu}^1 = X_1 \mu^1$, where

$$X_1(\mathbf{q}, t) = -\Gamma_t(\mathbf{q}, t) + \frac{\ell_t(\rho_0, t)}{\ell(\rho_0, t)} + \frac{\big[f\big(\hat{\rho}_0, \rho_1, t\big)\big]_t}{f\big(\hat{\rho}_0, \rho_1, t\big)}.$$

On the other hand $(\mathcal{A}^*\mu)^1 = Y_1\mu^1$, with

$$Y_1(\mathbf{q},t) = \sum_{j=1}^{7} Y_{1j}(\mathbf{q},t) = \sum_{j=1}^{7} Y_{1j},$$

where

$$Y_{11} = H'(\rho_0)b(\rho_0,t)\left[\Gamma_\rho(\rho_0,x_1,t) - \frac{\big[f\big(\hat\rho_0,\rho_1,t\big)\big]_{\rho_0}}{f\big(\hat\rho_0,\rho_1,t\big)}\right] - \frac{\big(H'(\rho_0)b(\rho_0,t)\ell(\rho_0,t)\big)_{\rho_0}}{\ell(\rho_0,t)}$$

$$Y_{12} = \big(H'(\rho_1) - H[\hat\rho_0,\rho_1]\big)b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t)$$

$$Y_{13} = \frac{\big[\big(H[\hat\rho_0,\rho_1] - H'(\rho_1)\big)b(\rho_1,t)f(\hat\rho_0,\rho_1,t)\big]_{\rho_1}}{f\big(\hat\rho_0,\rho_1,t\big)}$$

$$Y_{14} = \frac{\big(H\big[\hat\rho_0,\rho_1\big]f\big(\hat\rho_0,\rho_1,t\big)\big)_{x_1}}{f\big(\hat\rho_0,\rho_1,t\big)} + H\big[\hat\rho_0,\rho_1\big]\left[\lambda\big(\phi_{L-x_1}(\rho_1;t),t\big) - \lambda\big(\hat\rho_0,t\big)\right]$$

$$Y_{15} = \frac{\int H(\rho_*,\rho_0)f(\rho_*,\rho_0,t)\ \ell(d\rho_*,t)}{\ell(\rho_0,t)}$$

$$Y_{16} = -\int H\big[\phi_{L-x_1}(\rho_1;t),\rho_*\big]f\big(\phi_{L-x_1}(\rho_1;t),\rho_*,t\big)\ d\rho_* = -A\big(\phi_{L-x_1}(\rho_1;t),t\big)$$

$$Y_{17} = \frac{\int \big(H[\rho_*,\rho_1] - H\big[\hat\rho_0,\rho_*\big]\big)f\big(\hat\rho_0,\rho_*,t\big)f(\rho_*,\rho_1,t)\ d\rho_*}{f\big(\hat\rho_0,\rho_1,t\big)}.$$

Here,

- The term Y_{11} comes from an integration by parts with respect to the variable ρ_0. The dynamics of ρ_0 as in rule **(ii)** is responsible for this contribution.
- The terms Y_{12} and Y_{13} come from an integration by parts with respect to the variable ρ_1. The dynamics of ρ_1 as in rule **(ii)** is responsible for these two contributions.
- The term Y_{14} comes from an integration by parts with respect to the variable x_1. The dynamics of x_1 as in rule **(i)** is responsible for this contribution.
- The term Y_{15} comes from the boundary term $x_1 = 0$ in the integration by parts with respect to the variable x_1 when there are two particles at x_1 and x_2. This boundary condition represents the event that x_1 has reached the origin after which ρ_0 becomes ρ_1, and (x_2,ρ_2) is relabeled (x_1,ρ_1).
- The term Y_{16} comes from the boundary term $x_2 = L$ in the integration by parts with respect to the variable x_2, and the stochastic boundary dynamics as in the rule **(iv)**. The boundary term $x_2 = L$ cancels part of the contribution of the boundary dynamics as we have already seen in our calculation in the case $n = 0$.
- The rule **(iii)** is responsible for the term Y_{17}. When $n = 2$, the particles at x_1 and x_2 travel towards each other with speed $H[\hat\rho_1,\rho_2] - H[\hat\rho_0,\rho_1]$. As x_2 catches up with x_1, the particle x_1 disappears and its density $\rho_1 = \hat\rho_1$ is renamed ρ_*, and is integrated out. We then relabel (x_2,ρ_2) as (x_1,ρ_1).

We wish to show that $X_1 = Y_1$. After some cancellation, this simplifies to

$$X_1' = Y_1' := \hat Y_{11} + Y_{12} + Y_{13} + Y_{14} + Y_{16} + Y_{17},$$

where

$$X_1' = -\Gamma_t(\mathbf{q},t) - A(\rho_0,t) + \frac{\big[f(\hat\rho_0,\rho_1,t)\big]_t}{f(\hat\rho_0,\rho_1,t)},$$
$$\hat Y_{11} = H'(\rho_0)b(\rho_0,t)\left[\Gamma_\rho(\rho_0,x_1,t) - \frac{\big[f(\hat\rho_0,\rho_1,t)\big]_{\rho_0}}{f(\hat\rho_0,\rho_1,t)}\right].$$

(The same cancellation led to the Eq. (4.7).) Observe that $\Gamma(\mathbf{q},t) = \Gamma(\rho_0,x_1,t) + \Gamma(\rho_1,L-x_1,t)$. Moreover, by (4.7),

$$\Gamma_t(\rho_0,x_1,t) + H'(\rho_0)b(\rho_0,t)\Gamma_\rho(\rho_0,x_1,t) = A(\hat\rho_0,t) - A(\rho_0,t)$$
$$\Gamma_t(\rho_1,L-x_1,t) + H'(\rho_1)b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t) = A\big(\phi_{L-x_1}(\rho_1;t),t\big) - A(\rho_1,t).$$

As a result,

$$\begin{aligned}-\Gamma_t(\mathbf{q},t) - A(\rho_0,t) = {} & H'(\rho_0)b(\rho_0,t)\Gamma_\rho(\rho_0,x_1,t) + H'(\rho_1)b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t)\\ & - A\big(\phi_{L-x_1}(\rho_1;t),t\big) + A(\rho_1,t) - A(\hat\rho_0,t).\end{aligned}$$

Using this, we learn that the equality $X_1' = Y_1'$ is equivalent to the identity

$$\begin{aligned}\big[f(\hat\rho_0,\rho_1,t)\big]_t = {} & H\big[\hat\rho_0,\rho_1\big]\,\big[\lambda\big(\phi_{L-x_1}(\rho_1;t),t\big) - \lambda\big(\hat\rho_0,t\big)\big]\,f\big(\hat\rho_0,\rho_1,t\big)\\ & + \big[A(\hat\rho_0,t) - A(\rho_1,t)\big]\,f\big(\hat\rho_0,\rho_1,t\big)\\ & + \int \big(H[\rho_*,\rho_1] - H\big[\hat\rho_0,\rho_*\big]\big)f\big(\hat\rho_0,\rho_*,t\big)f(\rho_*,\rho_1,t)\,d\rho_*\\ & + \big[\big(H[\hat\rho_0,\rho_1] - H'(\rho_1)\big)b(\rho_1,t)f(\hat\rho_0,\rho_1,t)\big]_{\rho_1}\\ & - H'(\rho_0)b(\rho_0,t)\big[f\big(\hat\rho_0,\rho_1,t\big)\big]_{\rho_0} - H[\hat\rho_0,\rho_1]b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t)f\big(\hat\rho_0,\rho_1,t\big)\\ & + \big(H\big[\hat\rho_0,\rho_1\big]f\big(\hat\rho_0,\rho_1,t\big)\big)_{x_1}.\end{aligned}$$

By the group property (4.9), we can assert that for any C^1 function h,

$$\big[h(\hat\rho_0)\big]_{x_1} = b(\hat\rho_0,t)h'(\hat\rho_0) = b(\rho_0,t)\big[h(\hat\rho_0)\big]_{\rho_0}.$$

We use this and the definition of the quadratic operator Q in (3.10) to deduce that $X_1' = Y_1'$ is equivalent to the identity

$$\begin{aligned}\big[f(\hat\rho_0,\rho_1,t)\big]_t = {} & Q(f,f)(\hat\rho_0,\rho_1,t) + H\big[\hat\rho_0,\rho_1\big]\,\big[\lambda\big(\phi_{L-x_1}(\rho_1;t),t\big) - \lambda\big(\rho_1,t\big)\big]\,f\big(\hat\rho_0,\rho_1,t\big)\\ & + \big[\big(H[\hat\rho_0,\rho_1] - H'(\rho_1)\big)b(\rho_1,t)f(\hat\rho_0,\rho_1,t)\big]_{\rho_1}\\ & - H'(\rho_0)b(\hat\rho_0,t)f_{\rho_-}\big(\hat\rho_0,\rho_1,t\big) - H[\hat\rho_0,\rho_1]b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t)f\big(\hat\rho_0,\rho_1,t\big)\\ & + b(\hat\rho_0,t)H[\hat\rho_0,\rho_1]f_{\rho_-}\big(\hat\rho_0,\rho_1,t\big) + b(\hat\rho_0,t)H_{\rho_-}[\hat\rho_0,\rho_1]f(\hat\rho_0,\rho_1,t).\end{aligned}$$

Here we are acting the quadratic operator Q on functions because we are assuming that $f(\rho,d\rho_+,t) = f(\rho,\rho_+,t)\ d\rho_+$, is absolutely continuous with respect to the Lebesgue measure. We now use (4.13) to assert that $X_1' = Y_1'$ is equivalent to the identity

$$\begin{aligned} f_t(\hat\rho_0,\rho_1,t) = {} & Q(f,f)(\hat\rho_0,\rho_1,t) + b(\hat\rho_0,t)H_{\rho_-}[\hat\rho_0,\rho_1]f(\hat\rho_0,\rho_1,t) \\ & + H\big[\hat\rho_0,\rho_1\big]\left[\lambda\big(\phi_{L-x_1}(\rho_1;t),t\big) - \lambda\big(\rho_1,t\big)\right] f\big(\hat\rho_0,\rho_1,t\big) \\ & + \big[H[\hat\rho_0,\rho_1] - H'(\hat\rho_0)\big]b(\hat\rho_0,t)f_{\rho_-}\big(\hat\rho_0,\rho_1,t\big) \\ & + \big[\big(H[\hat\rho_0,\rho_1] - H'(\rho_1)\big)b(\rho_1,t)f(\hat\rho_0,\rho_1,t)\big]_{\rho_1} \\ & - H[\hat\rho_0,\rho_1]b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t)f\big(\hat\rho_0,\rho_1,t\big). \end{aligned}$$

On the other hand, by the definition of Γ,

$$\begin{aligned} b(\rho_1,t)\Gamma_\rho(\rho_1,L-x_1,t) &= \int_0^{L-x_1} b(\rho_1,t)\left[\lambda\big(\phi_y(\rho_1;t),t\big)\right]_{\rho_1}\ dy \\ &= \int_0^{L-x_1} \left[\lambda\big(\phi_y(\rho_1;t),t\big)\right]_y\ dy \\ &= \lambda\big(\phi_{L-x_1}(\rho_1;t),t\big) - \lambda(\rho_1,t), \end{aligned} \tag{4.15}$$

where we used the group property (4.9) for the second equality. This leads to

$$\begin{aligned} f_t(\hat\rho_0,\rho_1,t) = {} & Q(f,f)(\hat\rho_0,\rho_1,t) + b(\hat\rho_0,t)f(\hat\rho_0,\rho_1,t)H_{\rho_-}[\hat\rho_0,\rho_1] \\ & + \big[H[\hat\rho_0,\rho_1] - H'(\hat\rho_0)\big]b(\hat\rho_0,t)f_{\rho_-}(\hat\rho_0,\rho_1,t) \\ & + \big[\big(H[\hat\rho_0,\rho_1] - H'(\rho_1)\big)b(\rho_1,t)f(\hat\rho_0,\rho_1,t)\big]_{\rho_1}, \end{aligned}$$

which is exactly our kinetic equation! □

General n. We write $\dot\mu^n = X_n\mu^n$. We have,

$$X_n = -\Gamma_t(\mathbf{q},t) + \frac{\ell_t(\rho_0,t)}{\ell(\rho_0,t)} + \sum_{i=1}^n \frac{[f(\hat\rho_{i-1},\rho_i,t)]_t}{f(\hat\rho_{i-1},\rho_i,t)}. \tag{4.16}$$

By (4.7), and (4.15),

$$\Gamma_t(\mathbf{q},t) = \sum_{i=0}^n \left\{\big(A(\hat\rho_i,t) - A(\rho_i,t)\big) - H'(\rho_i)\big(\lambda(\hat\rho_i,t) - \lambda(\rho_i,t)\big)\right\}.$$

From this, (4.14) and (3.10) we deduce

$$
\begin{aligned}
X_n &= \frac{\ell_t(\rho_0,t)}{\ell(\rho_0,t)} + \sum_{i=1}^{n} \frac{Q^+(f,f)(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} + \sum_{i=1}^{n} \frac{(Cf)(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} \\
&\quad + \sum_{i=0}^{n} \big\{ H'(\rho_i)\big(\lambda(\hat\rho_i,t) - \lambda(\rho_i,t)\big) + A(\rho_i,t) - A(\hat\rho_i,t) \big\} \\
&\quad - \sum_{i=1}^{n} \big\{ A(\rho_i,t) - A(\hat\rho_{i-1},t) - H[\hat\rho_{i-1},\rho_i]\big(\lambda(\rho_i,t) - \lambda(\hat\rho_{i-1},t)\big) \big\} \\
&\quad + \sum_{i=1}^{n} \big[H'(\hat\rho_{i-1}) - H'(\rho_{i-1}) \big] b(\hat\rho_{i-1},t) \frac{f_{\rho_-}(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} \\
&= \frac{\ell_t(\rho_0,t)}{\ell(\rho_0,t)} + \sum_{i=1}^{n} \frac{Q^+(f,f)(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} + \sum_{i=1}^{n} \frac{(Cf)(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} \\
&\quad + H'(\rho_0)\big(\lambda(\hat\rho_0,t) - \lambda(\rho_0,t)\big) + A(\rho_0,t) - A(\hat\rho_n,t) \\
&\quad + \sum_{i=1}^{n} \big\{ H'(\rho_i)\big(\lambda(\hat\rho_i,t) - \lambda(\rho_i,t)\big) + H[\hat\rho_{i-1},\rho_i]\big(\lambda(\rho_i,t) - \lambda(\hat\rho_{i-1},t)\big) \big\} \\
&\quad + \sum_{i=1}^{n} \big[H'(\hat\rho_{i-1}) - H'(\rho_{i-1}) \big] b(\hat\rho_{i-1},t) \frac{f_{\rho_-}(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)}.
\end{aligned}
$$

For the right-hand side of (4.3) we write $(\mathcal{A}^*\mu)^n = Y_n\mu^n$, where

$$
Y_n = Y'_{11} + Y''_{11} + Y'''_{11} + Y_{n2} + Y_{n3} + Y'_{n4} + Y''_{n4} + Y_{n5} + Y_{n6} + Y_{n7}, \tag{4.17}
$$

with Y_{15} independent of n and defined before, $Y_{n6} = -A(\hat\rho_n,t)$, and

$$
\begin{aligned}
Y'_{11} &= H'(\rho_0) b(\rho_0,t) \Gamma_\rho(\rho_0,x_1,t) = H'(\rho_0)\big(\lambda(\hat\rho_0,t) - \lambda(\rho_0,t)\big) \\
Y''_{11} &= -H'(\rho_0) b(\rho_0,t) \frac{\big[f\big(\hat\rho_0,\rho_1,t\big)\big]_{\rho_0}}{f\big(\hat\rho_0,\rho_1,t\big)} \\
Y'''_{11} &= -\frac{\big(H'(\rho_0) b(\rho_0,t) \ell(\rho_0,t)\big)_{\rho_0}}{\ell(\rho_0,t)} \\
Y_{n2} &= \sum_{i=1}^{n} \big(H'(\rho_i) - H[\hat\rho_{i-1},\rho_i]\big) b(\rho_i,t) \Gamma_\rho(\rho_i, x_{i+1} - x_i, t) \\
&= \sum_{i=1}^{n} \big(H'(\rho_i) - H[\hat\rho_{i-1},\rho_i]\big)\big(\lambda(\hat\rho_i,t) - \lambda(\rho_i,t)\big),
\end{aligned}
$$

$$Y_{n3} = \sum_{i=1}^{n-1} \frac{\big[\big(H[\hat\rho_{i-1},\rho_i] - H'(\rho_i)\big)b(\rho_i,t)f(\hat\rho_{i-1},\rho_i,t)f(\hat\rho_i,\rho_{i+1},t)\big]_{\rho_i}}{f(\hat\rho_{i-1},\rho_i,t)f(\hat\rho_i,\rho_{i+1},t)}$$
$$+ \frac{\big[\big(H[\hat\rho_{n-1},\rho_n] - H'(\rho_n)\big)b(\rho_n,t)f(\hat\rho_{n-1},\rho_n,t)\big]_{\rho_n}}{f(\hat\rho_{n-1},\rho_n,t)},$$

$$Y'_{n4} = \sum_{i=1}^{n-1} \frac{\big[H\big[\hat\rho_{i-1},\rho_i\big]f\big(\hat\rho_{i-1},\rho_i,t\big)f\big(\hat\rho_i,\rho_{i+1},t\big)\big]_{x_i}}{f\big(\hat\rho_{i-1},\rho_i,t\big)f\big(\hat\rho_i,\rho_{i+1},t\big)} + \frac{\big[H\big[\hat\rho_{n-1},\rho_n\big]f\big(\hat\rho_{n-1},\rho_n,t\big)\big]_{x_n}}{f\big(\hat\rho_{n-1},\rho_n,t\big)},$$

$$Y''_{n4} = \sum_{i=1}^{n} H\big[\hat\rho_{i-1},\rho_i\big]\left(\lambda\big(\hat\rho_i,t\big) - \lambda\big(\hat\rho_{i-1},t\big)\right),$$

$$Y_{n7} = \sum_{i=1}^{n} \frac{Q^+(f,f)\big(\hat\rho_{i-1},\rho_i,t\big)}{f\big(\hat\rho_{i-1},\rho_i,t\big)},$$

where we used (4.15) for the first and fifth equality. We wish to show that $X_n = Y_n$. From

$$\frac{\ell_t(\rho_0,t)}{\ell(\rho_0,t)} = Y_{15} + Y'''_{11} - A(\rho_0,t),$$

and some cancellation, the equality $X_n = Y_n$ simplifies to $X'_n = Y'_n$, where

$$X'_n(\mathbf{q},t) = \sum_{i=1}^{n} \frac{(Cf)(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} + \sum_{i=1}^{n} \big[H'(\hat\rho_{i-1}) - H'(\rho_{i-1})\big]b(\hat\rho_{i-1},t)\frac{f_{\rho_-}(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)}$$
$$= \sum_{i=1}^{n} \frac{\big[(H[\hat\rho_{i-1},\rho_i] - H'(\rho_i))b(\rho_i,t)f(\hat\rho_{i-1},\rho_i,t)\big]_{\rho_i}}{f(\hat\rho_{i-1},\rho_i,t)}$$
$$+ \sum_{i=1}^{n} \big[H[\hat\rho_{i-1},\rho_i] - H'(\rho_{i-1})\big]b(\hat\rho_{i-1},t)\frac{f_{\rho_-}(\hat\rho_{i-1},\rho_i,t)}{f(\hat\rho_{i-1},\rho_i,t)} + b(\hat\rho_{i-1},t)H_{\rho_-}[\hat\rho_{i-1},\rho_i],$$

and $Y'_n = Y''_{11} + Y_{n3} + Y'_{n4}$. Observe that $Y'_{n4} = Y'_{n41} + Y'_{n42} + Y'_{n43}$, and $Y_{n3} = Y_{n31} + Y_{n32}$, where

$$Y'_{n41} = \sum_{i=1}^{n} H_{\rho_-}\big[\hat\rho_{i-1},\rho_i\big]b(\hat\rho_{i-1},t),$$

$$Y'_{n42} = \sum_{i=1}^{n} H\big[\hat\rho_{i-1},\rho_i\big]\frac{\big[f\big(\hat\rho_{i-1},\rho_i,t\big)\big]_{x_i}}{f\big(\hat\rho_{i-1},\rho_i,t\big)},$$

$$Y'_{n43} = \sum_{i=1}^{n-1} H\big[\hat\rho_{i-1},\rho_i\big]\frac{\big[f\big(\hat\rho_i,\rho_{i+1},t\big)\big]_{x_i}}{f\big(\hat\rho_i,\rho_{i+1},t\big)},$$

$$Y_{n31} = \sum_{i=1}^{n} \frac{\big[\big(H[\hat\rho_{i-1},\rho_i] - H'(\rho_i)\big)b(\rho_i,t)f(\hat\rho_{i-1},\rho_i,t)\big]_{\rho_i}}{f\big(\hat\rho_{i-1},\rho_i,t\big)},$$

$$Y_{n32} = \sum_{i=1}^{n-1} \big(H[\hat\rho_{i-1},\rho_i] - H'(\rho_i)\big)b(\rho_i,t)\frac{\big[f(\hat\rho_i,\rho_{i+1},t)\big]_{\rho_i}}{f(\hat\rho_i,\rho_{i+1},t)}.$$

From these decompositions, we learn that $X'_n = Y'_n$ is equivalent to $X''_n = Y''_n$, where

$$X''_n = \sum_{i=1}^{n} \big[H[\hat\rho_{i-1}, \rho_i] - H'(\rho_{i-1})\big] b(\hat\rho_{i-1}, t) \frac{f_{\rho_-}(\hat\rho_{i-1}, \rho_i, t)}{f(\hat\rho_{i-1}, \rho_i, t)},$$

and $Y''_n = Y''_{11} + Y_{n32} + Y'_{n42} + Y'_{n43}$. By the group property (4.9),

$$\big(h(\hat\rho_{i-1})\big)_{x_i} = b(\rho_{i-1}, t)\big(h(\hat\rho_{i-1})\big)_{\rho_{i-1}}, \qquad \big(h(\hat\rho_i)\big)_{x_i} = -b(\rho_i, t)\big(h(\hat\rho_i)\big)_{\rho_i},$$

This allows us to write

$$Y'_{n42} + Y'_{n43} = \sum_{i=1}^{n-1} H\big[\hat\rho_{i-1}, \rho_i\big] \left\{ b(\rho_{i-1}, t) \frac{\big[f(\hat\rho_{i-1}, \rho_i, t)\big]_{\rho_{i-1}}}{f(\hat\rho_{i-1}, \rho_i, t)} - b(\rho_i, t) \frac{\big[f(\hat\rho_i, \rho_{i+1}, t)\big]_{\rho_i}}{f(\hat\rho_i, \rho_{i+1}, t)} \right\}$$
$$+ H\big[\hat\rho_{n-1}, \rho_n\big] b(\rho_{n-1}, t) \frac{\big[f(\hat\rho_{n-1}, \rho_i, t)\big]_{\rho_{n-1}}}{f(\hat\rho_{n-1}, \rho_n, t)}.$$

Hence

$$\begin{aligned} Y''_n &= -\sum_{i=0}^{n-1} H'(\rho_i) b(\rho_i, t) \frac{\big[f(\hat\rho_i, \rho_{i+1}, t)\big]_{\rho_i}}{f(\hat\rho_i, \rho_{i+1}, t)} + \sum_{i=1}^{n} H\big[\hat\rho_{i-1}, \rho_i\big] b(\rho_{i-1}, t) \frac{\big[f(\hat\rho_{i-1}, \rho_i, t)\big]_{\rho_{i-1}}}{f(\hat\rho_{i-1}, \rho_i, t)} \\ &= \sum_{i=1}^{n} \big[H\big[\hat\rho_{i-1}, \rho_i\big] - H'(\rho_{i-1})\big] b(\rho_{i-1}, t) \frac{\big[f(\hat\rho_{i-1}, \rho_i, t)\big]_{\rho_{i-1}}}{f(\hat\rho_{i-1}, \rho_i, t)} \\ &= \sum_{i=1}^{n} \big[H\big[\hat\rho_{i-1}, \rho_i\big] - H'(\rho_{i-1}\big] b(\hat\rho_{i-1}, t) \frac{f_{\rho_-}(\hat\rho_{i-1}, \rho_i, t)}{f(\hat\rho_{i-1}, \rho_i, t)} = X''_n, \end{aligned}$$

as desired. For the third equality, we have used (4.9). This completes the proof. □

So far, we have been able to formally show that the law of $\mathbf{q}(t)$ is $\mu(d\mathbf{q}, t)$, by verifying the forward Eq. (4.3) for every n (recall that n represents the number of particles/shock discontinuities in the interval $(0, L)$). Our verification of (4.3) is rather tedious but elementary. Our verification is formal at this point because the evolution of $\mathbf{q}$ is governed by a discontinuous deterministic dynamics that is interrupted by stochastic Markovian entrance of new particles at the boundary point L. By selecting a pair ℓ and f that are differentiable with respect to time, it is not hard to justify our calculation for the left-hand side of (4.3), as it appeared in (4.16). It is the justification of the right-hand side as in (4.17) that requires additional work.

Writing $\Phi_s^t(\mathbf{q})$ for $\mathbf{q}(t)$ with initial condition $\mathbf{q}(s) = \mathbf{q}$, it suffices to show that for every nice function $G : \Delta_L \to \mathbb{R}$,

$$\frac{d}{ds} \mathbb{E}\, G\big(\Phi_s^t(\mathbf{q})\big) = \frac{d}{ds} \int G\big(\Phi_s^t(\mathbf{q})\big)\, \mu(d\mathbf{q}, s) = 0. \tag{4.18}$$

Clearly (4.18) implies

$$\mathbb{E}G(\mathbf{q}(t)) = \mathbb{E} \int G\big(\Phi_0^t(\mathbf{q})\big)\, \mu(d\mathbf{q}, 0) = \int G(\mathbf{q})\, \mu(d\mathbf{q}, t),$$

which means that the law of $\mathbf{q}(t)$ is given by our candidate $\mu(d\mathbf{q}, t)$. For a rigorous proof of this, we calculate the left time-derivative of $\mathbb{E}\ G\big(\Phi_s^t(\mathbf{q})\big)$ by hand and show that this left-derivative equals to

$$\sum_{n=0}^{\infty} \int G(X_n - Y_n)\ d\mu^n.$$

The details can be found in [16] and [17].

5 Homogenizations for Hamiltonian ODEs

The Hamilton-Jacobi PDE may be used to model the growth of an interface that is described as a graph of a height function. More precisely, the graph of a solution

$$u : \mathbb{R}^d \times [0, \infty) \to \mathbb{R},$$

of the Hamilton-Jacobi equation

$$u_t + H(x, u_x) = 0, \tag{5.1}$$

describes an interface at time t in microscopic coordinates. If the ratio of micro to macro scale is a large number n, then

$$u^n(x, t) = \frac{1}{n} u(nx, nt),$$

is the corresponding macroscopic height function. In practice n is large and we may obtain a simpler description of our model if the large n limit of u^n exists and satisfies a simple equation. Indeed u^n satisfies

$$u_t^n + H\big(nx, u_x^n\big) = 0,$$

and this equation must be solved for an initial condition of the form $u^n(x, 0) = g(x)$, where g represents the initial macroscopic height function. Let us define

$$(\Gamma_n g)(x) = ng\left(\frac{x}{n}\right);$$

the job of the operator Γ_n is to turn a macroscopic height function to its associated microscopic height function. We also write $T_t = T_t^H$ for the semigroup associated with the PDE (5.1). More precisely, $T_t u^0(x) = u(x, t)$ means

$$\begin{cases} u_t + H(x, u_x) = 0, & t > 0, \\ u(x, t) = u^0(x), \end{cases} \tag{5.2}$$

In terms of the operators T_t and Γ_n, we simply have $u^n = \big(\Gamma_n^{-1} \circ T_{nt} \circ \Gamma_n\big)(g)$. Put it differently,

$$T_t^{H \circ \Gamma_n} = \Gamma_n^{-1} \circ T_{nt}^H \circ \Gamma_n, \tag{5.3}$$

where $\gamma_n(x,p) = (nx,p)$. If we write $T(H)$ for T_1^H, then in particular we have

$$T(H \circ \gamma_n) = \Gamma_n^{-1} \circ T(H)^n \circ \Gamma_n.$$

The hope is that under some assumptions on H, the large n-limit of u^n exists and the limit $\bar{u}$ provides a reduced and simpler description of the growth model under study. For example, when H is 1-periodic in x-variable, the high oscillations of $H \circ \gamma_n$, may result in the convergence of u^n to a function $\bar{u}$, that solves the homogenized equation

$$\bar{u}_t + \bar{H}(\bar{u}_x) = 0. \tag{5.4}$$

When this happens, we write $\mathcal{A}(H) = \bar{H}$.

More generally, write $\mathcal{H}$ for the space of all C^1 Hamiltonian functions and define the natural translation operator

$$\tau_a H(x,p) = H(x+a,p),$$

for every $a \in \mathbb{R}^d$. We then take a probability measure $\mathbb{P}$ on $\mathcal{H}$ that is translation invariant and ergodic. We wish to take advantage of the ergodicity to assert that $T_t^{H\circ\gamma_n} \to T_t^{\bar{H}}$, $\mathbb{P}$-almost surely, as $n \to \infty$. If this happens for a deterministic function $\bar{H}$, then we write $\mathcal{A}(\mathbb{P}) = \bar{H}$. We note

- If $\mathbb{P}$ is supported on the set

  $$A := \left\{\tau_a H^0 : \ a \in \mathbb{R}^d\right\},$$

  for some 1-periodic Hamiltonian function H^0, then A is isomorphic to the d-dimensional torus and we are back to the periodic scenario.
- If $\mathbb{P}$ is supported on the topological closure (with respect to the uniform norm), of the set

  $$A := \left\{\tau_a H^0 : \ a \in \mathbb{R}^d\right\},$$

  for some Hamiltonian function H^0, and this closure is a compact set, then H^0 is almost periodic and the homogenization would allow us to find the large n-limit of $T_t^{H\circ\gamma_n} \to T_t^{\bar{H}}$, for almost all choices of H in the compact support of $\mathbb{P}$. In this case $\bar{A}$ has the structure of an Abelian Lie group and $\mathbb{P}$ is the corresponding Haar measure.

To explore the homogenization question further, we discuss the connection between Hamiltonian ODE and Hamilton-Jacobi PDE. For a classical solution, the method of characteristics suggests that at least for short times, we can solve (5.2) in terms of the flow of the Hamiltonian ODE

$$\begin{aligned} \dot{x} &= H_p(x,p), \\ \dot{p} &= -H_x(x,p). \end{aligned} \tag{5.5}$$

Equivalently we write $\dot{z} = J\nabla H(z)$, where $z = (x,p)$, and

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix},$$

with I denoting the $d \times d$ identity matrix. Writing $\phi_t = \phi_t^H$ for the flow of (5.5), we have

$$\phi_t^H \left\{ \big(x, \nabla u^0(x)\big) : \ x \in \mathbb{R}^d \right\} = \left\{ \big(x, u_x(x,t)\big) : \ x \in \mathbb{R}^d \right\}, \tag{5.6}$$

provided that the left-hand side remains a graph of a function. As we mentioned earlier, the Eq. (5.2) does not possess C^1 solutions in general. This has to do with the fact that if ϕ_t folds the graph of ∇u^0, then the left-hand side of (5.6) is no longer a graph of a function and (5.6) has no chance to be true. One possibility is that we *trim* the left-hand side (5.6) and hope for

$$\phi_t^H \left\{ \big(x, \nabla u^0(x)\big) : \ x \in \mathbb{R}^d \right\} \supseteq \left\{ \big(x, u_x(x,t)\big) : \ x \in \mathbb{R}^d \right\}, \tag{5.7}$$

For this to work, we have to give-up the differentiability of u. This geometric and rather naive idea does not suggest how the trimming should be carried out.

Alternatively, we may add a small viscosity term of the form $\varepsilon \Delta u$ to the right-hand side of (5.1) to guarantee the existence of a unique classical solution, and pass to the limit $\varepsilon \to 0$. The outcome is known as a *viscosity solution* (see [12]). As it turns out, under some coercivity assumption on H, we can guarantee the existence of a solution that is differentiable almost everywhere. We can now modify the right-hand side of (5.7) accordingly and wonder whether or not

$$\phi_t^H \left\{ \big(x, \nabla u^0(x)\big) : \ x \in \mathbb{R}^d \right\} \supseteq \left\{ \big(x, u_x(x,t)\big) : \ x \in \mathbb{R}^d, \quad u_x(x,t) \text{ exists} \right\}, \tag{5.8}$$

is true. The answer is affirmative if H is convex in p. However (5.8) may fail if we drop the convexity assumption. To explain this in the case of piecewise smooth solutions, we recall that if H is convex in p, the only discontinuity we can have is a shock discontinuity. In this case, at every point (a,t), with $t > 0$, we can find a solution $(x(s), p(s) : s \in [0,t])$ (the so-called backward characteristic) such that $x(t) = a$. If $\rho = u_x$ is continuous at a, this backward characteristic is unique and $p(t) = \rho(a,t)$. If ρ is discontinuous at (a,t), then $\rho(a,t)$ is multi-valued and for each possible value p of $\rho(a,t)$, there will be a solution to the Hamiltonian ODE with $(x(t), p(t)) = (a,p)$. In both cases, we still have (5.8).

The situation is far more complex when H is not convex. What may cause the violation of (5.8) is the occurrence of a rarefaction type solutions. To explain this, let us assume that $d = 1$, and H depends on p only. There are three momenta (or densities) $a_1 < a_2 < a_3$ such that

- The graph of H is convex and below its cord in $[a_1, a_2]$.
- The graph of H is concave and above its cord in $[a_2, a_3]$.
- The graph of H is below its cord in the interval $[a_1, a_3]$.

Now imagine that we have two discontinuities at $x(t)$ and $y(t)$ with $x(t) < y(t)$, and both are shock discontinuities. Assume

- The left and right values of ρ at $x(t)$ are $a_2'(t) < a_3'(t)$.
- The left and right values of ρ at $y(t)$ are $a_3'(t) > a_1'(t)$.
- These two shock discontinuities meet at some instant t_0 with $a_i'(t_0) = a_i$.

As a result, at the moment t_0 the two shock discontinuities are replaced with a rarefaction wave. Now if we take a point (x,t) inside the fan of this rarefaction wave (for which necessarily $t > t_0$), then at such (x,t) the connection with the initial data is lost and $(x, u_x(x,t))$ does not belong to the left-hand side of (5.8).

Motivated by the failure of (5.8) for viscosity solutions, we formulate a question.

Question 5.1: Is there a notion of generalized solution for (5.1) for which (5.8) is always true?

Using some ideas from topology and symplectic geometry the notion of *geometric solution* has been developed by Chaperon [9–11], Sikorav [25] and Viterbo [27]. The main features of this solution is as follows:

(i) The geometric solution satisfies (5.8) always.
(ii) The geometric solution satisfies (5.2) at every differentiability point of u.
(iii) The geometric solution coincides with the viscosity solution when H is convex in p.
(iv) Writing $\hat{T}_t u^0$ for the geometric solution of (5.2) with the initial condition u^0, we do not in general have $\hat{T}_t \circ \hat{T}_s = \hat{T}_{t+s}$ (except when H is convex in p).

Needless to say the last feature of the geometric solution is a serious flaw and does not provide a satisfactory answer for Question 5.1. Nonetheless the geometric solution provides a useful notion that helps us to connect the Eq. (5.2) to the Hamiltonian ODEs.

Because of the intimate relation between the Hamilton-Jacobi Equation and the Hamiltonian ODE, we may wonder whether a homogenization phenomenon occurs for the latter. More precisely, does the high-n limit of

$$\phi_t^{H\circ\gamma_n} = \gamma_n^{-1} \circ \phi_{nt}^H \circ \gamma_n,$$

exist in a suitable sense? Note that $H\circ\gamma_n$ has no pointwise limit and the existence of pointwise limit of $\phi_t^{H\circ\gamma_n}$ is not expected either. Writing ϕ_H for ϕ_1^H, we may wonder in what sense, if any, the sequence $\phi_{H\circ\gamma_n}$ has a limit. We note

$$\phi_{H\circ\gamma_n} = \gamma_n^{-1} \circ \phi_H^n \circ \gamma_n =: S_n(\phi_H).$$

We now discuss the existence of some interesting metric on the space $\mathcal{H}$ that is weaker than uniform norm and is closely related to the flow properties of the Hamiltonian ODEs. More importantly, there is a chance that $H \circ \gamma_n$ converges with respect to such metrics.

There are two metrics on $\mathcal{H}$ that are well-suited for our purposes. These metrics were defined by Hofer and Viterbo; the proofs of non-triviality of these metrics are highly non-trivial. Let us write down a wish-list for what our metric should satisfy.

Let us write $\mathcal{D}$ for the space of maps φ such that $\varphi = \phi_H$ for some smooth Hamiltonian function $H : \mathbb{R}^{2d} \times [0,1] \to \mathbb{R}$. (Any such map is *symplectic* as we will see later.) Assume that there exists a function $E : \mathcal{D} \to [0,\infty)$ with the following properties: For $\varphi, \psi, \tau \in \mathcal{D}$,

(i) $E(\varphi) = E(\varphi^{-1})$.
(ii) $E(\varphi) = E\big(\tau^{-1}\varphi\tau\big)$.
(iii) $E(\varphi\psi) \le E(\varphi) + E(\psi)$.
(iv) $E(\varphi) = 0$ if and only if $\varphi = id$.
(v) $E\big(\gamma_\ell^{-1}\varphi\gamma_\ell\big) = \ell^{-1}E(\varphi)$, where $\gamma_\ell(x,p) = (\ell x, p)$ and $\ell \in (0,\infty)$.

Here and below we simply write $\varphi\psi$ for $\varphi \circ \psi$ and think of $\mathcal{D}$ as a group with multiplication given by the map composition.

From E, we build a metric D on $\mathcal{D}$ by $D(\varphi,\psi) = E\big(\varphi\psi^{-1}\big)$. This metric has the following properties:

Proposition 1. **(i)** *$D(\varphi\tau,\psi\tau) = D(\tau\varphi,\tau\psi) = D(\varphi,\psi)$ for $\varphi,\psi,\tau \in \mathcal{D}$.*

(ii) *For $\varphi_1,\psi_1\ldots,\varphi_k,\psi_k$, we have*

$$D(\varphi_1\ldots\varphi_k,\psi_1\ldots\psi_k) \le \sum_{i=1}^{k} D(\varphi_i,\psi_i).$$

(iii) *For $S_n(\varphi) = \gamma_n^{-1}\circ\varphi^n\circ\gamma_n$, we have*

$$D\big(S_n(\varphi),S_n(\psi)\big) \le D(\varphi,\psi).$$

In the case of a homogenization, we expect $S_n(\varphi) \to \bar\varphi$, where $\bar\varphi = \phi_{\bar H}$, for a Hamiltonian function $\bar H$ that is independent of x. Write $\mathcal{D}_0$ for the space of such $\bar\varphi$. We note that $S_n(\bar\varphi) = \bar\varphi$. As a result, for any $\bar\varphi \in \mathcal{D}_0$,

$$D\big(S_n(\varphi),\bar\varphi\big) = D\big(S_n(\varphi),S_n(\bar\varphi)\big) \le D\big(\varphi,\bar\varphi\big), \tag{5.9}$$

by Proposition 1**(iii)**. As was noted by Viterbo [28], (5.9) implies that the set of limit points of the sequence $(S_n(\varphi) :\ n \in \mathbb{N})$ is a singleton: If $\bar\varphi$ and $\bar\psi$ are two limit points, then given $\delta > 0$, we find $n, m \in \mathbb{N}$ such that

$$D(S_n(\varphi),\bar\varphi) \le \delta, \qquad D(S_m(\varphi),\bar\psi) \le \delta.$$

From this and (5.9) we learn,

$$D(S_{nm}(\varphi),\bar\varphi) \le \delta, \qquad D(S_{nm}(\varphi),\bar\psi) \le \delta,$$

because $S_{nm} = S_n \circ S_m$. Hence $D(\bar\varphi,\bar\psi) \le 2\delta$. By sending $\delta \to 0$ we deduce that $\bar\varphi = \bar\psi$.

A natural question is whether we have homogenization with respect to such a metric.

Question 5.2: Given $\varphi \in \mathcal{D}$, does the large n limit of the sequence $\big\{S_n(\varphi)\big\}$ exist with respect to a metric D as above? □

6 Lagrangian Manifolds and Viterbo's Metric

The Question 5.2 has been answered affirmatively by Viterbo [28] when the Hamiltonian H is periodic in x and the metric D is the *Viterbo's metric*. We continue with a brief discussion of Viterbo's metric.

To simplify our presentation, let us assume that H is 1-periodic in x. We may also regard $u(\cdot, t)$ as a function on the d-dimensional torus $\mathbb{T}^d$.

To examine the left-hand side of (5.8), assume that the initially the solution of the ODE (5.5) satisfies the relationship $p = \nabla u^0(x)$, for some smooth function u^0. Whenever (5.6) is true, then at time t we have a similar relationship between the components of $\phi_t(x, p)$. Let us write $M^t := \phi_t(M^0)$, where

$$M^0 = \big\{(x, \nabla u^0(x)) : \ x \in \mathbb{T}^d\big\}.$$

To get a feel for $M^t = \phi_t^H(M^0)$, observe that M^0 is a graph of a an exact derivative. Let us refer to such manifolds as an *exact Lagrangian graph*. In general if

$$M = \big\{(x, X(x)) : \ x \in \mathbb{T}^d\big\},$$

then vectors of the form

$$\hat{a} := \begin{bmatrix} a \\ (DX)(x)a \end{bmatrix},$$

are tangents to M at x. What makes M exact is that if $X = \nabla u$, then the matrix $A = DX = D^2u$ is symmetric. To state this directly in terms of the tangent vectors, observe

$$Aa \cdot b - a \cdot Ab = \begin{bmatrix} Aa \\ -a \end{bmatrix} \cdot \begin{bmatrix} b \\ Ab \end{bmatrix} = J\hat{a} \cdot \hat{b} =: \bar{\omega}(\hat{a}, \hat{b}).$$

Hence the symmetry of A is equivalent to $\bar{\omega} \restriction_M = 0$ identically. (Here $\bar{\omega}$ is the standard symplectic 2-form of $\mathbb{R}^{2d}$.) Motivated by this we call a manifold M *Lagrangian* if the restriction of $\bar{\omega}$ to M is identically 0. The point of this definition is that if M^0 is the graph of an exact derivative, then $\varphi(M^0)$ may not be a graph of a function. However, when φ preserves the form $\bar{\omega}$, then $\varphi(M^0)$ is always a Lagrangian. We say a map φ is *symplectic* if it preserves $\bar{\omega}$ in the following sense:

$$\bar{\omega}\big((D\varphi)(x)a, (D\varphi)(x)b\big) = \bar{\omega}(a, b),$$

for every $x \in \mathbb{T}^d$ and every pair of vectors $a, b \in \mathbb{R}^{2d}$.

It is well-known that the correct topology for the viscosity solution comes from the uniform norm; this has to do with the fact the viscous approximation of Hamilton-Jacobi Equation satisfies a *maximum principle* that survives as we send the viscous term to 0. Since we are now interested in Hamiltonian ODE,

we may try to define some kind of metrics on Lagrangian manifolds of the form $\phi_t^H\big(M^0\big)$, where M^0 is an exact Lagrangian. Let us write $\mathcal{L}_0$ for the set of exact Lagrangian graphs, and define

$$\mathcal{H}_0 = \big\{H : \mathbb{T}^d \times \mathbb{R}^d \times [0,1] \to \mathbb{R} : \ H \ \text{ is } C^1 \quad \text{and 1-periodic in } \ x\big\}$$
$$\mathcal{L} = \big\{\phi_H(M) : \quad H \in \mathcal{H}_0, \quad M \in \mathcal{L}_0\big\}$$

When M is the graph of ∇u, for some C^1 function $u : \mathbb{T}^d \to \mathbb{R}$, we refer to u as the *generating function* of M. When this is the case, we write $\mathcal{G}(M) = u$. We also write

$$L(u) := \big\{(x, \nabla u(x)) : \ x \in \mathbb{T}^d\big\}.$$

Viterbo defines a metric on $\mathcal{L}$ that is a generalization of the L^∞-metric on its generating function. In other words, the metric D is defined in such way that if M^0 and M^1 are two exact Lagrangian graphs, then

$$D\big(M, M'\big) = \|\mathcal{G}(M) - \mathcal{G}(M')\|_\infty,$$

where by $\|\cdot\|_\infty$ we really mean the total oscillation:

$$\|u\|_\infty = \max u - \min u.$$

This definition is quite natural because $\mathcal{L}(u) = \mathcal{L}(u + c)$, for any constant c.

To guess how to extend the definition of this metric to $\mathcal{L}$, we need to develop a better understanding of the Hamiltonian ODEs. First, we claim that there exists a functional $\mathcal{I} = \mathcal{I}^H$ on the space of the paths $z(\cdot) = (x, p)(\cdot)$, such that $\dot{z} = J\nabla H(z,t)$ if and only if $z(\cdot)$ is a critical point of $\mathcal{I}$. Writing the Hamiltonian ODE as $J\dot{z} + \nabla H(z,t) = 0$, it is not hard to come up with an example for $\mathcal{I}$; we use a quadratic term to produce the linear part $J\dot{z}$, and H to produce ∇H. The following function $\mathcal{I} : C^1([0,1]; \mathbb{T}^d \times \mathbb{R}^d) \to \mathbb{R}$, is the integral of the celebrated *Cartan-Poincaré form*:

$$\mathcal{I}(z) = \int_0^1 \left[p(t) \cdot \dot{x}(t) - H(z(t), t)\right] \, dt.$$

Formally, $\partial\mathcal{I}(z) = -J\dot{z} - \nabla H(z,t)$. More precisely, if $\eta : [0,1] \to \mathbb{T}^d \times \mathbb{R}^d$, satisfies $\eta(0) = \eta(1) = 0$, then $\psi(\delta) = \mathcal{I}(z + \delta\eta)$ satisfies

$$\dot{\psi}(0) = -\int_0^1 \big(J\dot{z}(t) + \nabla H(z(t), t)\big) \cdot \eta(t) \, dt.$$

We now use this to come up with a generating-like function for $M^1 = \phi_H\big(M^0\big)$, where $M^0 = \mathcal{G}(u^0)$. To this end, let us define

$$\mathit{\Gamma} := \big\{z : [0,1] \to \mathbb{T}^d \times \mathbb{R}^d : \ z \in C^1\big\}, \qquad \mathit{\Gamma}(a) = \big\{z = (x,p) \in \mathit{\Gamma} \ : \ x(1) = a\big\}.$$

In words, $\mathit{\Gamma}(a)$ consists of position/momentum paths with the position component reaching a at time 1. We note that if $z \in \mathit{\Gamma}(a)$ and $\eta \in \mathit{\Gamma}(0)$, then $z + \delta\eta \in \mathit{\Gamma}(a)$ for all $\delta \in \mathbb{R}$. We then define $\hat{\mathcal{I}} : \mathit{\Gamma}(a) \to \mathbb{R}$ by

$$\hat{\mathcal{I}}(z) = u^0(x(0)) + \mathcal{I}(z) = u^0(x(0)) + \int_0^1 \left[p(t) \cdot \dot{x}(t) - H(z(t), t)\right] \, dt.$$

Since we want to use $\hat{\mathcal{I}}$ to build a generating function for M^1, observe that $\Gamma(0)$ is an infinite dimensional vector space and any $z \in \Gamma(a)$ can be written as

$$z(t) = (a, 0) + \xi(t),$$

with $\xi \in \Gamma(0)$. If M^1 is still a graph of function and has a generating function u^1, then what is happening is that we have a solution z satisfying $\dot{z} = J\nabla H(z,t)$ with

$$z(0) = \big(x(0), \nabla u^0(x(0))\big), \qquad z(1) = \big(x(1), \nabla u^1(x(1))\big).$$

Moreover, if u solves (1.1), then $u^1(x) = u(x,1)$. Note that if $w(t) = u(x(t),t)$, then $\dot{w} = p \cdot \dot{x} - H(z,t)$, or

$$u^1(x(1)) = u^0(x(0)) + \int_0^1 \left[p(t) \cdot \dot{x}(t) - H(z(t),t)\right]\, dt.$$

To separate $x(1)$ from the rest of information in the path $z(\cdot)$, we define $\mathcal{J} : \mathbb{T}^d \times \Gamma(0) \to \mathbb{R}$, by

$$\mathcal{J}(a;\xi) = \hat{\mathcal{I}}\big((a,0) + \xi\big).$$

In other words, if $z = (a,0)+\xi = (x,p)$, and $\xi = (x',p)$, then $x'(t) = x(t) - x(1) = x(t) - a$. Now, if we set

$$\hat{\psi}(\delta) = \hat{\mathcal{I}}(z + \delta\eta) = \mathcal{J}(a; \xi + \delta\eta),$$

for $z \in \Gamma(a)$, and $\eta = (\hat{x}, \hat{p}) \in \Gamma(0)$, then

$$\frac{d\hat{\psi}}{d\delta}(0) = \big(\nabla u^0(x(0)) - p(0)\big) \cdot \hat{x}(0) - \int_0^1 \big(J\dot{z}(t) + \nabla H(z(t),t)\big) \cdot \eta(t)\, dt.$$

We can now assert

$$\partial_\xi \mathcal{J}(a;\xi) = 0 \quad \Longleftrightarrow \quad p(0) = \nabla u^0(x(0)), \quad \text{and} \quad z = (a,0) + \xi \quad \text{satisfies } \dot{z} = J\nabla H(z,t).$$

On the other hand, if we set $\bar{\psi}(\delta) = \hat{\mathcal{I}}\big(z + (\delta b, 0)\big) = \mathcal{J}(a + \delta b; \xi)$, then

$$\partial_a \mathcal{J}(a;\xi) \cdot b = \frac{d\bar{\psi}}{d\delta}(0) = \nabla u^0(x(0)) \cdot b - \int_0^1 H_x(z(t),t) \cdot b\, dt.$$

As a result, if $\partial_\xi \mathcal{J}(a;\xi) = 0$, then

$$\partial_a \mathcal{J}(a;\xi) = \nabla u^0(x(0)) - \int_0^1 H_x(z(t),t)\, dt = \nabla u^0(x(0)) + \int_0^1 \dot{p}(t)\, dt = p(1).$$

From this we deduce

$$\phi^H\big(M^0\big) = \big\{\big(a, \partial_a \mathcal{J}(a,\xi)\big) :\ a \in \mathbb{T}^d, \quad \partial_\xi \mathcal{J}(a,\xi) = 0\big\},$$

where $a = x(1)$ represents the position at time 1. We think of $\mathcal{J}(a;\xi)$ as a *generalized generating function* (or in short GG function) of $M = M^1$.

The Lagrangian M^1 is exact if for every $(a,p) \in \mathbb{T}^d \times \mathbb{R}^d$, there is at most one solution z to the Hamiltonian ODE with $x(1) = a, p(1) = p$. Our aim is to associate a nonnegative number $E(M)$ to $M \in \mathcal{L}$ that in the case of an exact Lagrangian $M = \mathcal{G}(u)$,

$$E(M) = E^+(M) - E^-(M),$$

where $E^\pm(M)$ are two critical values of u, namely the maximum and minimum of u. In the case of a non-exact M, we may use the functional $\mathcal{J} = \mathcal{J}_M$ to select two critical points $z^\pm = (a^\pm, \xi^\pm)$ of the functional $\mathcal{J}_M$ to define

$$E^\pm(M) = \mathcal{J}_M(a^\pm, \xi^\pm) = \hat{\mathcal{I}}(z^\pm).$$

The main question now is how to select the critical paths $z^\pm$. The classical theories of *Morse* and *Lusternik-Schnirelman* would provide us with systematic ways of selecting critical values of a scalar-valued function on a manifold. (See for example Appendix E of [21] for an introduction on LS Theory.) These theories are applicable if the underlying manifold is finite-dimensional and their generalizations to infinite dimensional setting are highly nontrivial. (Floer Theory is a prime example of such generalization.) However in our setting it is possible to approximate the functional $\mathcal{I}$ or $\mathcal{J}$ with a function that is defined on $\mathbb{T}^d \times \mathbb{R}^N$ for a suitable N that depends on H and u^0 and could be large. More precisely, we may try to find a *generating function* $S : \mathbb{T}^d \times \mathbb{R}^N \to \mathbb{R}$ such that

$$M = \big\{(x, S_x(x,\xi)) : \ x \in \mathbb{T}^d, \xi \in \mathbb{R}^N, S_\xi(x,\xi) = 0\big\},$$

In fact any manifold of this form is automatically a Lagrangian manifold, simply because the tangent vectors at a point of the form $(x, S_x(x,\xi))$ are still of the form $\big(v, A(x,\xi)v\big); v \in \mathbb{R}^d$, where $A = S_{xx}$ is a symmetric matrix.

To explain the existence of such finite dimensional generating functions, we need to make another observation about the flows of Hamiltonian ODEs.

We may regard the symplectic property of $\varphi = \phi_1^H$, as saying that its graph

$$Gr(\varphi) : \big\{(x, \varphi(x)) : \ x \in \mathbb{T}^d \times \mathbb{R}^d\big\},$$

is Lagrangian with respect to the 2-form $\omega \oplus (-\omega)$ in $\mathbb{R}^{4d}$. This Lagrangian manifold is an exact graph when the set $Gr(\varphi)$ can be expressed as a graph of the gradient of a scalar-valued function. But now because of the form of the symplectic form $\omega \oplus (-\omega)$, this must be done in a twisted way. More precisely, if $\varphi(x,p) = (X,P)$, then the generating function would depend for example on (X,p). In the case of an exact symplectic map, we may find a scalar-valued function $S(X,p)$ such that

$$\varphi\big(S_p(X,p), p\big) = \big(X, S_X(X,p)\big).$$

The identity map has the generating function $p \cdot X$. This suggests writing $S(X,p) = X \cdot p - w(X,p)$ with w periodic in X. In terms of w,

$$\varphi\big(X - w_p(X,p), p\big) = \big(X, p - w_X(X,p)\big).$$

Now imagine that $M = \varphi(M^0)$, where both M^0 and φ are exact with generating functions u^0 and $S(X,p) = X \cdot p - w(X,p)$. Then

$$\hat{S}(X;x,p) = u^0(x) + p \cdot (X - x) - w(X,p) =: p \cdot (X - x) - \hat{w}(X;x,p),$$

is a GG function for M^1: If $\xi = (x,p)$, then

$$\hat{S}_\xi(X;\xi) = 0 \quad \Longleftrightarrow \quad p = \nabla u^0(x), \quad x = X - w_p(X,p).$$

As a result

$$\hat{S}_\xi(X;\xi) = 0 \quad \Longrightarrow \quad \varphi(x,p) = \big(X, \hat{S}_X(X;\xi)\big),$$

because $\hat{S}_X = p - w_X(X,p) = P$.

As we mentioned earlier, the identity map has a generating function. Using Implicit Function Theorem, it is not hard to show that any symplectic map that is C^1-close to the identity also possesses a generating function. Now if $\varphi = \phi_H$ is the time-one map associated with a smooth Hamiltonian, then we can find $\delta > 0$ sufficiently small, such that the map $\varphi = \phi_\delta^H$ is sufficiently close to the identity map and possesses a generating function. In general, each ϕ^H can be expressed as $\varphi^1 \circ \cdots \circ \varphi^N$ with each φ^i possessing a generating function as above. If each φ^i has a generating function of the form $X \cdot p - w^i(X,p)$, then $M = \varphi(M^0)$ has a generating function of the form

$$\hat{S}\big(x_N;\xi\big) = \hat{S}\big(x_N;x_0,p_0,\ldots,x_{N-1},p_{N-1}\big) := u^0(x_0) + \sum_{i=0}^{N-1} \big[p_i \cdot (x_{i+1} - x_i) - w^i(x_{i+1},p_i)\big].$$

We refer to [22] and Chapter 9 of [21] for more details on generating functions.

So far we know that our Lagrangian manifolds possess finite-dimensional generating functions. The next question to address is that how we can select appropriate critical values $E^\pm(M)$ for $\hat{S}(X;\xi)$.

For the rest of this section, we assume that M is a Lagrangian manifold with a generating function $S(x,\xi)$. More precisely,

$$M = \big\{(x, S_x(x,\xi)) : \ x \in \mathbb{T}^d, \xi \in \mathbb{R}^N, S_\xi(x,\xi) = 0\big\}, \tag{6.1}$$

and $S(x,\xi)$ is a nice perturbation of a quadratic function in ξ. By this we mean that there exists a quadratic function $B(\xi) = A\xi \cdot \xi$ such that A is an invertible symmetric matrix, and

$$\sup_{x,\xi} |S(x,\xi) - B(\xi)|, \qquad \sup_{x,\xi} |S_\xi(x,\xi) - \nabla B(\xi)| < \infty.$$

We wish to put a metric on the space $\mathcal{L}$ of such Lagrangians. For this, we first wish to define the *size* $E(M)$ of a Lagrangian manifold M. If M is an exact Lagrangian graph with generating function u, we simply set

$$E(M) = \max u - \min u.$$

If M can be represented as in (6.1), then $E(M)$ is defined by

$$E(M) = E^+(M) - E^-(M),$$

where $E^-(M)$ and $E^+(M)$ are two critical values of the generating function S that are the analog of $\min u$ and $\max u$. To explain our strategy for selecting $E^{\pm}(M)$, first imagine that $S(x,\xi) = u(x) + B(\xi)$. Then we still have $E^-(M) = \min u = u(x_-)$ and $E^+(M) = \max u = u(x_+)$, because both $(x_\pm, 0)$ are critical points of S. After all 0 is a critical value for B. We may apply *Lusternik-Schnirelman (LS) Theory*, to assert that the function S also has two critical points that are very much the analogs of $(x_\pm, 0)$. (See [28] and Appendix E of [21].) We are now ready to define a metric on the space Lagrangian manifolds who possess generating function as in (6.1). If M and M' are two Lagrangian manifolds with generating functions S and S' respectively, then we define a new generating function

$$\big(S \ominus S'\big)(x,\xi_1,\xi_2) = S(x,\xi_1) - S'(x,\xi_2).$$

This new generating function produces a new Lagrangian manifold

$$\begin{aligned} M \ominus M' &= \big\{(x, S_x(x,\xi_1) - S'_x(x,\xi_2)) :\ x \in \mathbb{T}^d, \xi \in \mathbb{R}^N, \xi \in \mathbb{R}^{N'}, S_\xi(x,\xi) = 0, S'_\xi(x,\xi_2) = 0\big\} \\ &= \big\{(x, p - p') :\ (x,p) \in M,\ (x,p') \in M'\big\}. \end{aligned}$$

This generating function is a bounded perturbation of $\big(B \ominus B'\big)(\xi_1,\xi_2) = B(\xi_1) - B(\xi_2)$. We set

$$D(M,M') = E\big(S \ominus S'\big).$$

We now would like to use the above metric to define a metric for Hamiltonian functions or their corresponding flows that was defined by Viterbo:

$$\mathcal{D}(H,H') = \sup\big\{D\big(\phi_H(M), \phi_{H'}(M)\big) :\ M \in \mathcal{L}\big\}.$$

Theorem 4. *(Viterbo [28]) The large n-limit of $H \circ \gamma_n$ exists with respect to the Viterbo Metric D. Moreover, if the limit is denoted by $\mathcal{B}(H)$, then $\mathcal{B}$ satisfies the following properties*

(i) *For every symplectic $\varphi \in \mathcal{D}$, we have $\mathcal{B}(H \circ \varphi) = \mathcal{B}(H)$.*
(ii) *If $\{H,K\} := J\nabla H \cdot K = 0$, then $\mathcal{B}(H+K) = \mathcal{B}(H) + \mathcal{B}(K)$.*

This should be compared with the Lions-Papanicolaou-Varadhan [18] homogenization result.

Theorem 5. *Assume that $H(x,p)$ is a C^1, x-periodic Hamiltonian function with*

$$\lim_{|p|\to\infty} \inf_x H(x,p) = \infty.$$

Then the large n limit of $T^{H\circ\gamma_n}$ exists. The limit is of the form $T^{\bar H}$, for a Hamiltonian function $\mathcal{A}(H) := \bar H$ that is independent of x.

In fact $\mathcal{A}(H) = \mathcal{B}(H)$ when H is convex in p; otherwise they could be different. Moreover, Theorem 6.2 has been extended to the random ergodic setting when H is convex in p in Rezakhanlou-Tarver [23] and Souganidis [24]. A natural question is whether or not Theorem 6.1 can be extended to the random setting.

Question 6.1: Can we extend Viterbo's metric (or Hofer's metric) to the random setting and does the large n limit of $H \circ \gamma_n$ exist for a stationary ergodic Hamiltonian H? □

Acknowledgments. These notes are based on the mini course that was given by the author at Institut Henri Poincaré, Centre Emile Borel during the trimester *Stochastic Dynamics Out of Equilibrium*. The author thanks this institution for hospitality and support. He is also very grateful to the organizers of the program for invitation, and an excellent research environment. Special thanks to two anonymous referees for their careful reading of these notes and their many insightful comments and suggestions.

References

1. Abramson, J., Evans, S.N.: Lipschitz minorants of Brownian motion and Lévy processes. Probab. Theory Relat. Fields **158**, 809–857 (2014)
2. Aldous, D.J.: Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilist. Bernoulli **5**, 3–48 (1999)
3. Aurell, E., Frisch, U., She, Z.-S.: The inviscid Burgers equation with initial data of Brownian type. Commun. Math. Phys. **148**, 623–641 (1992)
4. Bertoin, J.: The Inviscid Burgers Equation with Brownian initial velocity. Commun. Math. Phys. **193**, 397–406 (1998)
5. Burgers, J.M.: A Mathematics model illustrating the theory of turbulence. In: Von Mises, R., Von Karman, T. (eds.) Advances in Applied Mechanics, vol. 1, pp. 171–199. Elsevier Science (1948)
6. Carraro, L., Duchon, J.: Solutions statistiques intrinsèques de l'équation de Burgers et processus de Lévy. Comptes Rendus de l'Académie des Sciences. Série I. Mathématique **319**, 855–858 (1994)
7. Carraro, L., Duchon, J.: Équation de Burgers avec conditions initiales à accroissements indépendants et homogènes. Annales de l'Institut Henri Poincaré (C) Non Linear Analysis **15**, 431–458 (1998)
8. Chabanol, M.-L., Duchon, J.: Markovian solutions of inviscid Burgers equation. J. Stat. Phys. **114**, 525–534 (2004)
9. Chaperon, M.: Une idée du type géodésiques brisées pour les systèmes hamiltoniens. C. R. Acad. Sci. Paris Sér. I Math. **298**, 293–296 (1984)
10. Chaperon, M.: Familles Génératrices. Cours donné à l'école d'été Erasmus de Samos (1990)
11. Chaperon, M.: Lois de conservation et géométrie symplectique. C. R. Acad. Sci. Paris Sér. I Math. **312**, 345–348 (1991)
12. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics. American Mathematical Society, USA (2010)
13. Frachebourg, L., Martin, Ph.A.: Exact statistical properties of the Burgers equation. J. Fluid Mech. **417**, 323–349 (1992)

14. Getoor, R.K.: Splitting times and shift functionals. Z. Wahrscheinlichkeitstheorie verw. Gebiete **47**, 69–81 (1979)
15. Groeneboom, P.: Brownian motion with a parabolic drift and airy functions. Probab. Theory Relat. Fields **81**, 79–109 (1989)
16. Kaspar, D., Rezakhanlou, F.: Scalar conservation laws with monotone pure-jump Markov initial conditions. Probab. Theory Related Fields **165**, 867–899 (2016)
17. Kaspar, D., Rezakhanlou, F.: Kinetic statistics of scalar conservation laws with piecewise-deterministic Markov process data (preprint)
18. Lions, P.L., Papanicolaou, G., Varadhan, S.R.S.: Homogenization of Hamilton-Jacobi equations (unpublished)
19. Menon, G., Pego, R.L.: Universality classes in Burgers turbulence. Commun. Math. Phys. **273**, 177–202 (2007)
20. Menon, G., Srinivasan, R.: Kinetic theory and Lax equations for shock clustering and Burgers turbulence. J. Statist. Phys. **140**, 1–29 (2010)
21. Rezakhanlou, F.: Lectures on Symplectic Geometry. https://math.berkeley.edu/rezakhan/symplectic.pdf
22. Rezakhanlou, F.: Hamiltonian ODE, Homogenization, and Symplectic Topology. https://math.berkeley.edu/rezakhan/WKAM.pdf
23. Rezakhanlou, F., Tarver, J.E.: Homogenization for stochastic Hamilton-Jacobi equations. Arch. Ration. Mech. Anal. **151**, 277–309 (2000)
24. Souganidis, P.E.: Stochastic homogenization of Hamilton-Jacobi equations and some applications. Asymptot. Anal. **20**(1), 1–11 (1999)
25. Sikorav, J.-C.: Sur les immersions lagrangiennes dans un fibré cotangent admettant une phase génératrice globale. C. R. Acad. Sci. Paris Sér. I Math. **302**, 119–122 (1986)
26. Sinai, Y.G.: Statistics of shocks in solutions of inviscid Burgers equation. Commun. Math. Phys. **148**, 601–621 (1992)
27. Viterbo, C.: Solutions of Hamilton-Jacobi equations and symplectic geometry. Addendum to: Séminaire sur les Equations aux Dérivées Partielles. 1994–1995 [école Polytech., Palaiseau, 1995]
28. Viterbo, C.: Symplectic Homogenization (2014). arXiv:0801.0206v3

Workshop 1: Numerical Aspects of Nonequilibrium Dynamics

On Optimal Decay Estimates for ODEs and PDEs with Modal Decomposition

Franz Achleitner, Anton Arnold(✉), and Beatrice Signorello

Institute for Analysis and Scientific Computing, TU Wien,
Wiedner Hauptstraße 8–10, 1040 Vienna, Austria
{franz.achleitner,anton.arnold,beatrice.signorello}@tuwien.ac.at
https://www.asc.tuwien.ac.at/~achleitner/
https://www.asc.tuwien.ac.at/~arnold/

Abstract. We consider the Goldstein-Taylor model, which is a 2-velocity BGK model, and construct the "optimal" Lyapunov functional to quantify the convergence to the unique normalized steady state. The Lyapunov functional is optimal in the sense that it yields decay estimates in L^2-norm with the sharp exponential decay rate and minimal multiplicative constant. The modal decomposition of the Goldstein-Taylor model leads to the study of a family of 2-dimensional ODE systems. Therefore we discuss the characterization of "optimal" Lyapunov functionals for linear ODE systems with positive stable diagonalizable matrices. We give a complete answer for optimal decay rates of 2-dimensional ODE systems, and a partial answer for higher dimensional ODE systems.

Keywords: Lyapunov functionals · Sharp decay estimates · Goldstein-Taylor model

1 Introduction

This note is concerned with optimal decay estimates of hypocoercive evolution equations that allow for a modal decomposition. The notion *hypocoercivity* was introduced by Villani in [16] for equations of the form $\frac{\mathrm{d}}{\mathrm{d}t} f = -Lf$ on some Hilbert space H, where the generator L is not coercive, but where solutions still exhibit exponential decay in time. More precisely, there should exist constants $\lambda > 0$ and $c \geq 1$, such that

$$\|e^{-Lt} f^I\|_{\widetilde{H}} \leq c\, e^{-\lambda t} \|f^I\|_{\widetilde{H}} \qquad \forall\, f^I \in \widetilde{H}, \tag{1.1}$$

where $\widetilde{H}$ is a second Hilbert space, densely embedded in $(\ker L)^\perp \subset H$.

The large-time behavior of many hypocoercive equations have been studied in recent years, including Fokker-Planck equations [3,4,16], kinetic equations [12] and BGK equations [1,2]. Determining the sharp (i.e. maximal) exponential decay rate λ was an issue in some of these works, in particular [1,2,4]. But finding at the same time the smallest multiplicative constant $c \geq 1$, is so far

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 241–264, 2019.
https://doi.org/10.1007/978-3-030-15096-9_6

an open problem. And this is the topic of this note. For simple cases we shall describe a procedure to construct the "optimal" Lyapunov functional that will imply (1.1) with the sharp constants λ and c.

For illustration purposes we shall focus here only on the following 2-velocity BGK-model (referring to the physicists Bhatnagar, Gross and Krook [7]) for the two functions $f_\pm(x,t) \geq 0$ on the one-dimensional torus $x \in \mathbb{T}$ and for $t \geq 0$. It reads

$$\begin{cases} \partial_t f_+ &= -\partial_x f_+ + \frac{1}{2}(f_- - f_+), \\ \partial_t f_- &= \partial_x f_- - \frac{1}{2}(f_- - f_+). \end{cases} \tag{1.2}$$

This system of two transport-reaction equations is also called *Goldstein-Taylor model*.

For initial conditions normalized as $\int_0^{2\pi} \left[f_+^I(x) + f_-^I(x)\right] \mathrm{d}x = 2\pi$, the solution $f(t) = (f_+(t), f_-(t))^\top$ converges to its unique (normalized) steady state with $f_+^\infty = f_-^\infty = \frac{1}{2}$. The operator norm of the propagator for (1.2) can be computed explicitly from the Fourier modes, see [14]. By contrast, the goal of this paper and of [1,12] is to refrain from explicit computations of the solution and to use Lyapunov functionals instead. Following this strategy, an explicit exponential decay rate of this two velocity model was shown in [12, §1.4]. The sharp exponential decay estimate was found in [1, §4.1] via a refined functional, yielding the following result:

Theorem 1.1 ([1, Th. 6])**.** *Let $f^I \in L^2(0,2\pi;\mathbb{R}^2)$. Then the solution to (1.2) satisfies*

$$\|f(t) - f^\infty\|_{L^2(0,2\pi;\mathbb{R}^2)} \leq c\, e^{-\lambda t} \|f^I - f^\infty\|_{L^2(0,2\pi;\mathbb{R}^2)}, \qquad t \geq 0,$$

with the optimal constants $\lambda = \frac{1}{2}$ and $c = \sqrt{3}$.

Remark 1.2. (a) Actually, the optimal c was not specified in [1], but will be the result of Theorem 3.7 below.

(b) As we shall illustrate in Sect. 5, it does *not* make sense to optimize these two constants at the same time. The optimality in Theorem 1.1 refers to first maximizing the exponential rate λ, and then to minimize the multiplicative constant c.

The proof of Theorem 1.1 is based on the spatial Fourier transform of (1.2), cf. [1,12]. We denote the Fourier modes in the discrete velocity basis $\{\binom{1}{1}, \binom{1}{-1}\}$ by $u_k(t) \in \mathbb{C}^2$, $k \in \mathbb{Z}$. They evolve according to the ODE systems

$$\frac{\mathrm{d}}{\mathrm{d}t} u_k = -\mathbf{C}_k\, u_k, \quad \mathbf{C}_k = \begin{pmatrix} 0 & ik \\ ik & 1 \end{pmatrix}, \quad k \in \mathbb{Z}, \tag{1.3}$$

and their (normalized) steady states are

$$u_0^\infty = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad u_k^\infty = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad k \neq 0.$$

In the main body of this note we shall construct appropriate Lyapunov functionals for such ODEs, in order to obtain sharp decay rates of the form (1.1). In the context of the BGK-model (1.2), combining such decay estimates for all modes u_k then yields Theorem 1.1, as they are uniform in k. We remark that the construction of Lyapunov functionals to reveal optimal decay rates in ODEs was already included in the classical textbook [6, §22.4], but optimality of the multiplicative constant c was not an issue there.

In this article we shall first review, from [1,2], the construction of Lyapunov functionals for linear first order ODE systems that reveal the sharp decay rate. They are quadratic functionals represented by some Hermitian matrix $\mathbf{P}$. As these functionals are not uniquely determined, we shall then discuss a strategy to find the "best Lyapunov" functional in Sect. 3—by minimizing the condition number $\kappa(\mathbf{P})$. The method of Sect. 3 always yields an upper bound for the minimal multiplicative constant c and the sharp constant in certain subcases (see Theorem 3.7). The refined method of Sect. 4 covers another subclass (see Theorem 4.1). Overall we shall determine the optimal constant c for 2-dimensional ODE systems, and give estimates for it in higher dimensions. In the final Sect. 5 we shall illustrate how to obtain a whole family of decay estimates—with suboptimal decay rates, but improved constant c. For small time this improves the estimate obtained in Sect. 3. After the completion of this article, we found out that our results give insights to an open problem in system and control theory [8, Problem 6.3].

2 Lyapunov Functionals for Hypocoercive ODEs

In this section we review decay estimates for linear ODEs with constant coefficients of the form

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} f = -\mathbf{C} f, \quad t \geq 0, \\ f(0) = f^I \in \mathbb{C}^n, \end{cases} \tag{2.1}$$

for some (typically non-Hermitian) matrix $\mathbf{C} \in \mathbb{C}^{n\times n}$. To ensure that the origin is the unique asymptotically stable steady state, we assume that the matrix $\mathbf{C}$ is *hypocoercive* (i.e. positive stable, meaning that all eigenvalues have positive real part). Since we shall *not* require that $\mathbf{C}$ is coercive (meaning that its Hermitian part would be positive definite), we *cannot* expect that all solutions to (2.1) satisfy for the Euclidean norm: $\|f(t)\|_2 \leq e^{-\widetilde{\lambda} t}\|f^I\|_2$ for some $\widetilde{\lambda} > 0$. However, such an exponential decay estimate does hold in an adapted norm that can be used as a Lyapunov functional.

The construction of this Lyapunov functional is based on the following lemma:

Lemma 2.1 **([1, Lemma 2], [4, Lemma 4.3]).** *For any fixed matrix $\mathbf{C} \in \mathbb{C}^{n\times n}$, let $\mu := \min\{\Re(\lambda) | \lambda$ is an eigenvalue of $\mathbf{C}\}$. Let $\{\lambda_j | 1 \leq j \leq j_0\}$ be all the*

eigenvalues of $\mathbf{C}$ *with* $\Re(\lambda_j) = \mu$. *If all* λ_j $(j = 1, \dots, j_0)$ *are non-defective*[1], *then there exists a positive definite Hermitian matrix* $\mathbf{P} \in \mathbb{C}^{n\times n}$ *with*

$$\mathbf{C}^*\mathbf{P} + \mathbf{P}\mathbf{C} \geq 2\mu\mathbf{P}\,, \tag{2.2}$$

but $\mathbf{P}$ *is not uniquely determined.*

Moreover, if all eigenvalues of $\mathbf{C}$ *are non-defective, examples of such matrices* $\mathbf{P}$ *satisfying* (2.2) *are given by*

$$\mathbf{P} := \sum_{j=1}^{n} b_j\, w_j \otimes w_j^*, \tag{2.3}$$

where $w_j \in \mathbb{C}^n$ $(j = 1, \dots, n)$ *denote the (right) normalized eigenvectors of* $\mathbf{C}^*$ *(i.e.* $\mathbf{C}^* w_j = \bar{\lambda}_j w_j$*), and* $b_j \in \mathbb{R}^+$ $(j = 1, \dots, n)$ *are arbitrary weights.*

For $n = 2$ all positive definite Hermitian matrices $\mathbf{P}$ satisfying (2.2) have the form (2.3), but for $n \geq 3$ this is not true (see Lemma 3.1 and Example 3.2, respectively).

In this article, for simplicity, we shall only consider the case when all eigenvalues of $\mathbf{C}$ are non-defective. For the extension of Lemma 2.1 and of the corresponding decay estimates to the defective case we refer to [3, Prop. 2.2] and [5].

Due to the positive stability of $\mathbf{C}$, the origin is the unique and asymptotically stable steady state $f^\infty = 0$ of (2.1): Due to Lemma 2.1, there exists a positive definite Hermitian matrix $\mathbf{P} \in \mathbb{C}^{n\times n}$ such that $\mathbf{C}^*\mathbf{P} + \mathbf{P}\mathbf{C} \geq 2\mu\mathbf{P}$ where $\mu = \min \Re(\lambda_j) > 0$. Thus, the time derivative of the adapted norm $\|f\|_{\mathbf{P}}^2 := \langle f, \mathbf{P}f\rangle$ along solutions of (2.1) satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\|f(t)\|_{\mathbf{P}}^2 \leq -2\mu\|f(t)\|_{\mathbf{P}}^2.$$

Hence the evolution becomes a contraction in the adapted norm:

$$\|f(t)\|_{\mathbf{P}}^2 \leq e^{-2\mu t}\|f^I\|_{\mathbf{P}}^2, \qquad t \geq 0. \tag{2.4}$$

Clearly, this procedure can yield the sharp decay rate μ, only if $\mathbf{P}$ satisfies (2.2).

Next we translate this decay in $\mathbf{P}$-norm into a decay in the Euclidean norm:

$$\|f(t)\|_2^2 \leq (\lambda_{\min}^{\mathbf{P}})^{-1}\|f(t)\|_{\mathbf{P}}^2 \leq (\lambda_{\min}^{\mathbf{P}})^{-1}e^{-2\mu t}\|f^I\|_{\mathbf{P}}^2 \leq \kappa(\mathbf{P})\, e^{-2\mu t}\|f^I\|_2^2\,, \quad t \geq 0, \tag{2.5}$$

where $0 < \lambda_{\min}^{\mathbf{P}} \leq \lambda_{\max}^{\mathbf{P}}$ are, respectively, the smallest and largest eigenvalues of $\mathbf{P}$, and $\kappa(\mathbf{P}) = \lambda_{\max}^{\mathbf{P}}/\lambda_{\min}^{\mathbf{P}}$ is the (numerical) condition number of $\mathbf{P}$ with respect to the Euclidean norm. While (2.4) is sharp, (2.5) is not necessarily sharp: Given the spectrum of $\mathbf{C}$, the exponential decay rate in (2.5) is optimal, but the multiplicative constant not necessarily. For the optimality of the chain of inequalities in (2.5) we have to distinguish two scenarios: Does there exist an

[1] An eigenvalue is defective if its geometric multiplicity is strictly less than its algebraic multiplicity.

initial datum f^I such that each inequality will be (simultaneously) an equality for some *finite* $t_0 \geq 0$? Or is this only possible asymptotically as $t \to \infty$? We shall start the discussion with the former case, which is simpler, and defer the latter case to Sect. 4. The first scenario allows to find the optimal multiplicative constant for $\mathbf{C} \in \mathbb{R}^{2\times 2}$, based on (2.5). But in other cases it may only yield an explicit upper bound for it, as we shall discuss in Sect. 4.

Concerning the first inequality of (2.5), a solution $f(t_0)$ will satisfy $\|f(t_0)\|_2^2 = (\lambda_{\min}^{\mathbf{P}})^{-1}\|f(t_0)\|_{\mathbf{P}}^2$ for some $t_0 \geq 0$ only if $f(t_0)$ is in the eigenspace associated to the eigenvalue $\lambda_{\min}^{\mathbf{P}}$ of $\mathbf{P}$. Moreover, the initial datum f^I satisfies $\|f^I\|_{\mathbf{P}}^2 = \lambda_{\max}^{\mathbf{P}}\|f^I\|_2^2$ if f^I is in the eigenspace associated to the eigenvalue $\lambda_{\max}^{\mathbf{P}}$ of $\mathbf{P}$. Finally we consider the second inequality of (2.5): If the matrix $\mathbf{C}$ satisfies, e.g., $\Re\lambda_j = \mu > 0$; $j = 1, ..., n$, with all eigenvalues non-defective, then we always have

$$\|f(t)\|_{\mathbf{P}}^2 = e^{-2\mu t}\|f^I\|_{\mathbf{P}}^2 \qquad \forall t \geq 0, \tag{2.6}$$

since (2.2) is an equality then. This is the case for our main example (1.3) with $k \neq 0$.

Since the matrix $\mathbf{P}$ is not unique, we shall now discuss the choice of $\mathbf{P}$ as to minimize the multiplicative constant in (2.5). To this end we need to find the matrix $\mathbf{P}$ with minimal condition number that satisfies (2.2). Clearly, the answer can only be unique up to a positive multiplicative constant, since $\widetilde{\mathbf{P}} := \tau\mathbf{P}$ with $\tau > 0$ would reproduce the estimate (2.5).

As we shall prove in Sect. 3, the answer to this minimization problem is very easy in 2 dimensions: The best $\mathbf{P}$ corresponds to equal weights in (2.3), e.g. choosing $b_1 = b_2 = 1$.

3 Optimal Constant via Minimization of the Condition Number

In this section, we describe a procedure towards constructing "optimal" Lyapunov functionals: For solutions $f(t)$ of ODE (2.1) they will imply

$$\|f(t)\|_2 \leq c\, e^{-\mu t}\|f^I\|_2 \tag{3.1}$$

with the sharp constant μ and partly also the sharp constant c.

We shall describe the procedure for ODEs (2.1) with positive stable matrices $\mathbf{C}$. For simplicity we confine ourselves to diagonalizable matrices $\mathbf{C}$ (i.e. all eigenvalues are non-defective). In this case, Lemma 2.1 states that there exist positive definite Hermitian matrices $\mathbf{P}$ satisfying the matrix inequality (2.2). Following (2.5), $\sqrt{\kappa(\mathbf{P})}$ is always an upper bound for the constant c in (3.1).

Our strategy is now to minimize $\kappa(\mathbf{P})$ on the set of all admissible matrices $\mathbf{P}$. We shall prove that this actually yields the minimal constant c in certain cases (see Theorem 3.7). In 2 dimensions this minimization problem can be solved very easily thanks to Lemmas 3.1 and 3.3:

Lemma 3.1. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix. Then all matrices $\mathbf{P}$ satisfying (2.2) are of the form (2.3).*

Proof. We use again the matrix $\mathbf{W}$ whose columns are the normalized (right) eigenvectors of $\mathbf{C}^*$ such that

$$\mathbf{C}^*\mathbf{W} = \mathbf{W}\mathbf{D}^*, \tag{3.2}$$

with $\mathbf{D} = \operatorname{diag}(\lambda_1^{\mathbf{C}}, \lambda_2^{\mathbf{C}})$ where $\lambda_j^{\mathbf{C}}$ $(j \in \{1,2\})$ are the eigenvalues of $\mathbf{C}$. Since $\mathbf{W}$ is regular, $\mathbf{P}$ can be written as

$$\mathbf{P} = \mathbf{W}\mathbf{B}\mathbf{W}^*,$$

with some positive definite Hermitian matrix $\mathbf{B}$. Then the matrix inequality (2.2) can be written as

$$2\mu\mathbf{W}\mathbf{B}\mathbf{W}^* \le \mathbf{C}^*\mathbf{W}\mathbf{B}\mathbf{W}^* + \mathbf{W}\mathbf{B}\mathbf{W}^*\mathbf{C} = \mathbf{W}(\mathbf{D}^*\mathbf{B} + \mathbf{B}\mathbf{D})\mathbf{W}^*.$$

This matrix inequality is equivalent to

$$0 \le (\mathbf{D}^* - \mu\mathbf{I})\mathbf{B} + \mathbf{B}(\mathbf{D} - \mu\mathbf{I}). \tag{3.3}$$

Next we order the eigenvalues $\lambda_j^{\mathbf{C}}$ $(j \in \{1,2\})$ of $\mathbf{C}$ increasingly with respect to their real parts, such that $\Re(\lambda_1^{\mathbf{C}}) = \mu$. Moreover, we consider

$$\mathbf{B} = \begin{pmatrix} b_1 & \beta \\ \overline{\beta} & b_2 \end{pmatrix}$$

where $b_1, b_2 > 0$ and $\beta \in \mathbb{C}$ with $|\beta|^2 < b_1 b_2$. Then the right hand side of (3.3) is

$$(\mathbf{D}^* - \mu\mathbf{I})\mathbf{B} + \mathbf{B}(\mathbf{D} - \mu\mathbf{I}) = \begin{pmatrix} 0 & (\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})\beta \\ \overline{(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})\beta} & 2b_2\Re(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}}) \end{pmatrix} \tag{3.4}$$

with $\operatorname{Tr}[(\mathbf{D}^* - \mu\mathbf{I})\mathbf{B} + \mathbf{B}(\mathbf{D} - \mu\mathbf{I})] = 2b_2\Re(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})$ and

$$\det[(\mathbf{D}^* - \mu\mathbf{I})\mathbf{B} + \mathbf{B}(\mathbf{D} - \mu\mathbf{I})] = -\left|\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}}\right|^2 |\beta|^2 .$$

Condition (3.3) is satisfied if and only if $\operatorname{Tr}[(\mathbf{D}^* - \mu\mathbf{I})\mathbf{B} + \mathbf{B}(\mathbf{D} - \mu\mathbf{I})] \ge 0$ which holds due to our assumptions on $\lambda_2^{\mathbf{C}}$ and b_2, and $\det[(\mathbf{D}^* - \mu\mathbf{I})\mathbf{B} + \mathbf{B}(\mathbf{D} - \mu\mathbf{I})] \ge 0$. The last condition holds if and only if

$$\lambda_2^{\mathbf{C}} = \lambda_1^{\mathbf{C}} \quad \text{or } \beta = 0.$$

In the latter case $\mathbf{B}$ is diagonal and hence $\mathbf{P}$ is of the form (2.3). In the former case, (3.2) shows that $\mathbf{C} = \lambda_1^{\mathbf{C}}\mathbf{I}$, and the inequality (2.2) is trivial. Now any positive definite Hermitian matrix $\mathbf{P}$ has a diagonalization $\mathbf{P} = \mathbf{V}\mathbf{E}\mathbf{V}^*$, with a diagonal real matrix $\mathbf{E}$ and an orthogonal matrix $\mathbf{V}$, whose columns are –of course– eigenvectors of $\mathbf{C}$. Thus, $\mathbf{P}$ is again of the form (2.3). □

In contrast to this 2D result, in dimensions $n \geq 3$ there exist matrices $\mathbf{P}$ satisfying (2.2) which are not of form (2.3):

Example 3.2. Consider the matrix $\mathbf{C} = \mathrm{diag}(1, 2, 3)$. Then, all matrices

$$\mathbf{P}(b_1, b_2, b_3, \beta) = \begin{pmatrix} b_1 & 0 & 0 \\ 0 & b_2 & \beta \\ 0 & \beta & b_3 \end{pmatrix} \tag{3.5}$$

with positive b_j ($j \in \{1, 2, 3\}$) and $\beta \in \mathbb{R}$ such that $8b_2b_3 - 9\beta^2 \geq 0$, are positive definite Hermitian matrices and satisfy (2.2) for $\mathbf{C} = \mathrm{diag}(1, 2, 3)$ and $\mu = 1$. But the eigenvectors of $\mathbf{C}^*$ are the canonical unit vectors. Hence, matrices of form (2.3) would all be diagonal. □

Restricting the minimization problem to admissible matrices $\mathbf{P}$ of form (2.3) we find: Defining a matrix $\mathbf{W} := (w_1|\dots|w_n)$ whose columns are the (right) normalized eigenvectors of $\mathbf{C}^*$ allows to rewrite formula (2.3) as

$$\begin{aligned} \mathbf{P} &= \sum_{j=1}^{n} b_j\, w_j \otimes w_j^* = \mathbf{W}\,\mathrm{diag}(b_1, b_2, \dots, b_n)\mathbf{W}^* \\ &= \big(\mathbf{W}\,\mathrm{diag}(\sqrt{b_1}, \sqrt{b_2}, \dots, \sqrt{b_n})\big)\big(\mathbf{W}\,\mathrm{diag}(\sqrt{b_1}, \sqrt{b_2}, \dots, \sqrt{b_n})\big)^* \end{aligned} \tag{3.6}$$

with positive constants b_j ($j = 1, \dots, n$). The identity

$$\mathbf{W}\,\mathrm{diag}(\sqrt{b_1}, \sqrt{b_2}, \dots, \sqrt{b_n}) = (\sqrt{b_1}w_1|\dots|\sqrt{b_n}w_n)$$

shows that the weights are just rescalings of the eigenvectors. Finally, the condition number of $\mathbf{P}$ is the squared condition number of $(\mathbf{W}\,\mathrm{diag}(\sqrt{b_1}, \sqrt{b_2}, \dots, \sqrt{b_n}))$. Hence, to find matrices $\mathbf{P}$ of form (3.6) with minimal condition number, is equivalent to identifying (right) precondition matrices among the positive definite diagonal matrices which minimize the condition number of $\mathbf{W}$. This minimization problem can be formulated as a convex optimization problem [10] based on the result [15]. Due to [11, Theorem 1], the minimum is attained (i.e. an optimal scaling matrix exists) since our matrix $\mathbf{W}$ is non-singular. (Note that its column vectors form a basis of $\mathbb{C}^n$.) The convex optimization problem can be solved by standard software providing also the exact scaling matrix which minimizes the condition number of $\mathbf{P}$, see the discussion and references in [10]. For more information on convex optimization and numerical solvers, see e.g. [9].

We return to the minimization of $\kappa(\mathbf{P})$ in 2 dimensions:

Lemma 3.3. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix. Then the condition number of the associated matrix $\mathbf{P}$ in (2.3) is minimal by choosing equal weights, e.g. $b_1 = b_2 = 1$.*

Proof. A diagonalizable matrix $\mathbf{C}$ has only non-defective eigenvalues. Up to a unitary transformation, we can assume w.l.o.g. that the eigenvectors of $\mathbf{C}^*$ are

$$w_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad w_2 = \begin{pmatrix} \alpha \\ \sqrt{1-\alpha^2} \end{pmatrix} \quad \text{for some } \alpha \in [0, 1). \tag{3.7}$$

This unitary transformation describes the change of the coordinate system. To construct the new basis, we choose one of the normalized eigenvectors w_1 as first basis vector, and recall that the second normalized eigenvector w_2 is only determined up to a scalar factor $\gamma \in \mathbb{C}$ with $|\gamma| = 1$. The right choice for the scalar factor γ allows to fulfill the above restriction on α.

We use the representation of the positive definite matrix $\mathbf{P}$ in (3.6):

$$\mathbf{P} = \mathbf{W}\,\mathrm{diag}(b_1, b_2)\mathbf{W}^* \quad \text{with } \mathbf{W} = \begin{pmatrix} 1 & \alpha \\ 0 & \sqrt{1-\alpha^2} \end{pmatrix}. \tag{3.8}$$

Since $\mathbf{P}$ and $\tau\mathbf{P}$ have the same condition number, we consider w.l.o.g. $b_1 = 1/b$ and $b_2 = b$. Thus, we have to determine the positive parameter $b > 0$ which minimizes the condition number of

$$\mathbf{P}(b) = \mathbf{W}\,\mathrm{diag}(1/b, b)\mathbf{W}^* = \begin{pmatrix} \frac{1}{b} + b\alpha^2 & b\alpha\sqrt{1-\alpha^2} \\ b\alpha\sqrt{1-\alpha^2} & b(1-\alpha^2) \end{pmatrix}. \tag{3.9}$$

The condition number of matrix $\mathbf{P}(b)$ is given by

$$\kappa(\mathbf{P}(b)) = \lambda_+^{\mathbf{P}}(b)/\lambda_-^{\mathbf{P}}(b) \geq 1,$$

where

$$\lambda_\pm^{\mathbf{P}}(b) = \frac{\mathrm{Tr}\,\mathbf{P}(b) \pm \sqrt{(\mathrm{Tr}\,\mathbf{P}(b))^2 - 4\det\mathbf{P}(b)}}{2}$$

are the (positive) eigenvalues of $\mathbf{P}(b)$. We notice that $\mathrm{Tr}\,\mathbf{P}(b) = b + 1/b$ is independent of α and is a convex function of $b \in (0, \infty)$ which attains its minimum for $b = 1$. Moreover, $\det\mathbf{P}(b) = 1 - \alpha^2$ is independent of b. This implies that the condition number

$$\kappa(\mathbf{P}(b)) = \frac{\lambda_+^{\mathbf{P}}(b)}{\lambda_-^{\mathbf{P}}(b)} = \frac{1 + \sqrt{1 - \frac{4\det\mathbf{P}(b)}{(\mathrm{Tr}\,\mathbf{P}(b))^2}}}{1 - \sqrt{1 - \frac{4\det\mathbf{P}(b)}{(\mathrm{Tr}\,\mathbf{P}(b))^2}}}$$

attains its unique minimum at $b = 1$, taking the value

$$\kappa_{\min} = \frac{1+\alpha}{1-\alpha}. \tag{3.10}$$

□

This 2D-result does not generalize to higher dimensions. In dimensions $n \geq 3$ there exist diagonalizable positive stable matrices $\mathbf{C}$, such that the matrix $\mathbf{P}$ with equal weights b_j does not yield the lowest condition number among all matrices of form (2.3). We give a counterexample in 3 dimensions:

Example 3.4. For some $\mathbf{C}^*$, consider its eigenvector matrix

$$\mathbf{W} := \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathrm{diag}\left(1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}\right), \tag{3.11}$$

which has normalized column vectors. We define the matrices $\mathbf{P}(b_1, b_2, b_3) := \mathbf{W}\,\mathrm{diag}(b_1, b_2, b_3)\mathbf{W}^*$ for positive parameters b_1, b_2 and b_3, which are of form (2.3) and hence satisfy the inequality (2.2). In case of equal weights $b_1 = b_2 = b_3$ the condition number is $\kappa(\mathbf{P}(b_1, b_1, b_1)) \approx 15.12825876$. But using [13, Theorem 3.3], the minimal condition number $\min_{b_j} \kappa(\mathbf{P}(b_1, b_2, b_3)) \approx 13.92820324$ is attained for the weights $b_1 = 2$, $b_2 = 4$ and $b_3 = 3$. □

Combining Lemmas 3.1 and 3.3 we have

Corollary 3.5. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix. Then the condition number is minimal among all matrices $\mathbf{P}$ satisfying (2.2), if $\mathbf{P}$ is of form (2.3) with equal weights, e.g. $b_1 = b_2 = 1$.*

This 2D-result does not generalize to higher dimensions. Extending the conclusion of Example 3.4, we shall now show that $\mathbf{P}$ does not necessarily have to be of form (2.3), if its condition number should be minimal:

Example 3.6. We consider a special case of Example 3.4, with

$$\widetilde{\mathbf{C}} = (\mathbf{W}^*)^{-1}\mathrm{diag}(1, 2, 3)\mathbf{W}^*$$

with $\mathbf{W}$, the eigenvector matrix of $\widetilde{\mathbf{C}}^*$, given by (3.11). Then the matrices $\widetilde{\mathbf{C}}$ and

$$\widetilde{\mathbf{P}}(b_1, b_2, b_3, \beta) := \mathbf{W}\mathbf{P}(b_1, b_2, b_3, \beta)\mathbf{W}^*$$

with matrix $\mathbf{P}(b_1, b_2, b_3, \beta)$ in (3.5) satisfy the matrix inequality (2.2) with $\mu = 1$. But $\widetilde{\mathbf{P}}$ is not of form (2.3) if $\beta \neq 0$. Nevertheless, the condition number $\kappa(\widetilde{\mathbf{P}}(b_1, b_2, b_3, \beta)) \approx 5.82842780720132$ for the weights $b_1 = 2$, $b_2 = 4$, $b_3 = 3$, and $\beta = -2.45$, is much lower than with $\beta = 0$ (i.e. $\kappa(\widetilde{\mathbf{P}}(2, 4, 3, 0)) \approx 13.92820324$, cf. Example 3.4). □

Lemma 3.3 and inequality (2.5) show that $\sqrt{\kappa_{\min}}$ from (3.10) is an *upper bound* for the best constant in (3.1) for the 2D case. For matrices with eigenvalues that have the same real part it actually yields the minimal multiplicative constant c, as we shall show now. Other cases will be discussed in Sect. 4.

For a diagonalizable matrix $\mathbf{C} \in \mathbb{C}^{2\times 2}$ with $\lambda_1^{\mathbf{C}} = \lambda_2^{\mathbf{C}}$ it holds that $\|f(t)\|_2 = e^{-\Re\lambda_1^{\mathbf{C}} t}\|f^I\|_2$. And for the general case we have:

Theorem 3.7. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix with eigenvalues $\lambda_1^{\mathbf{C}} \neq \lambda_2^{\mathbf{C}}$, and associated eigenvectors v_1 and v_2, resp. If the eigenvalues have identical real parts, i.e. $\Re\lambda_1^{\mathbf{C}} = \Re\lambda_2^{\mathbf{C}}$, then the condition number of the associated matrix $\mathbf{P}$ in (2.3) with equal weights, e.g. $b_1 = b_2 = 1$, yields the minimal constant in the decay estimate (3.1) for the ODE (2.1):*

$$c = \sqrt{\kappa(\mathbf{P})} = \sqrt{\frac{1+\alpha}{1-\alpha}} \quad \textit{where } \alpha := \left|\left\langle \frac{v_1}{\|v_1\|}, \frac{v_2}{\|v_2\|} \right\rangle\right|. \tag{3.12}$$

Proof. With the notation from the proof of Lemma 3.3 we have

$$\mathbf{P}(1) = \begin{pmatrix} 1+\alpha^2 & \alpha\sqrt{1-\alpha^2} \\ \alpha\sqrt{1-\alpha^2} & 1-\alpha^2 \end{pmatrix},$$

with the eigenvectors $y_+^{\mathbf{P}} = (\sqrt{1-\alpha^2}, 1-\alpha)^\top$, $y_-^{\mathbf{P}} = (\sqrt{1-\alpha^2}, -1-\alpha)^\top$. According to the discussion after (2.5) we choose the initial condition $f^I = y_+^{\mathbf{P}}$. From the diagonalization (3.2) of $\mathbf{C}$ we get

$$f(t) = (\mathbf{W}^*)^{-1} e^{-\mathbf{D}t} \mathbf{W}^* f^I.$$

Using (3.8) and $\mathbf{W}^* y_\pm^{\mathbf{P}} = \sqrt{1-\alpha^2}\binom{1}{\pm 1}$ we obtain directly that

$$f(t_0) = e^{-\lambda_1^{\mathbf{C}} t_0} y_-^{\mathbf{P}} \quad \text{with} \quad t_0 = \frac{\pi}{|\Im(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})|}.$$

Hence, also the first inequality in (2.5) is sharp at t_0. Sharpness of the whole chain of inequalities then follows from (2.6), and this finishes the proof. □

This theorem now allows us to identify the minimal constant c in Theorem 1.1 on the Goldstein-Taylor model: The eigenvalues of the matrices $\mathbf{C}_k$, $k \neq 0$ from (1.3) are $\lambda = \frac{1}{2} \pm i\sqrt{k^2 - \frac{1}{4}}$.The corresponding transformation matrices $\mathbf{P}_k$ with $b_1 = b_2 = 1$ are given by $\mathbf{P}_0 = \mathbf{I}$ and

$$\mathbf{P}_k = \begin{pmatrix} 1 & -\frac{i}{2k} \\ \frac{i}{2k} & 1 \end{pmatrix}, \quad \text{with} \quad \kappa(\mathbf{P}_k) = \frac{2|k|+1}{2|k|-1}, \qquad k \neq 0.$$

Combining the decay estimates for all Fourier modes $u_k(t)$ shows that the minimal multiplicative constant in Theorem 1.1 is given by $c = \sqrt{\kappa(\mathbf{P}_{\pm 1})} = \sqrt{3}$. For a more detailed presentation how to recombine the modal estimates we refer to §4.1 in [1].

4 Optimal Constant for 2D Systems

The optimal constant c in (3.1) for $\mathbf{C} \in \mathbb{C}^{2\times 2}$ with $\Re\lambda_1^{\mathbf{C}} = \Re\lambda_2^{\mathbf{C}}$ was determined in Theorem 3.7. In this section we shall discuss the remaining 2D cases. We start to derive the minimal multiplicative constant c for matrices $\mathbf{C}$ with eigenvalues that have distinct real parts but identical imaginary parts.

Theorem 4.1. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix with eigenvalues $\lambda_1^{\mathbf{C}}$ and $\lambda_2^{\mathbf{C}}$, and associated eigenvectors v_1 and v_2, resp. If the eigenvalues have distinct real parts $\Re\lambda_1^{\mathbf{C}} < \Re\lambda_2^{\mathbf{C}}$ and identical imaginary parts $\Im\lambda_1^{\mathbf{C}} = \Im\lambda_2^{\mathbf{C}}$, then the minimal multiplicative constant c in (3.1) for the ODE (2.1) is given by*

$$c = \frac{1}{\sqrt{1-\alpha^2}} \quad \text{where } \alpha := \left| \left\langle \frac{v_1}{\|v_1\|}, \frac{v_2}{\|v_2\|} \right\rangle \right|. \tag{4.1}$$

Proof. We use again the unitary transformation as in the proof of Lemma 3.3, such that the eigenvectors w_1 and w_2 of $\mathbf{C}^*$ are given in (3.7). If $f(t)$ is a solution of (2.1), then $\tilde{f}(t) := e^{i\Im\lambda_1^{\mathbf{C}}t} f(t)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{f}(t) = -\widetilde{\mathbf{C}}\tilde{f}(t)\,, \quad \tilde{f}(0) = f^I, \tag{4.2}$$

with

$$\widetilde{\mathbf{C}} := (\mathbf{C} - i\Im\lambda_1^{\mathbf{C}}\mathbf{I}) = (\mathbf{W}^*)^{-1}\begin{pmatrix} \Re\lambda_1^{\mathbf{C}} & 0 \\ 0 & \Re\lambda_2^{\mathbf{C}} \end{pmatrix}\mathbf{W}^*.$$

The multiplication with $e^{i\Im\lambda_1^{\mathbf{C}}t}$ is another unitary transformation and does not change the norm, i.e. $\|f(t)\|_2 = \|\tilde{f}(t)\|_2$. Therefore, we can assume w.l.o.g. that matrix $\mathbf{C}$ has real coefficients and distinct real eigenvalues. Then, the solution $f(t)$ of the ODE (2.1) satisfies $\Re f(t) = f_{re}(t)$ and $\Im f(t) = f_{im}(t)$ where $f_{re}(t)$ and $f_{im}(t)$ are the solutions of the ODE (2.1) with initial data $\Re f^I$ and $\Im f^I$, resp. Altogether, we can assume w.l.o.g. that all quantities are real valued: Considering a matrix $\mathbf{C} \in \mathbb{R}^{2\times 2}$ with two distinct real eigenvalues $\lambda_1 < \lambda_2$ and real eigenvectors v_1 and v_2, then the associated eigenspaces $\mathrm{span}\{v_1\}$ and $\mathrm{span}\{v_2\}$ dissect the plane into four sectors

$$\mathcal{S}^{\pm\mp} := \{z_1 v_1 + z_2 v_2 \mid z_1 \in \mathbb{R}^{\pm}\,,\ z_2 \in \mathbb{R}^{\mp}\}, \tag{4.3}$$

see Fig. 1. A solution $f(t)$ of ODE (2.1) starting in an eigenspace will approach the origin in a straight line, such that

$$\|f(t)\|_2^2 = e^{-2\lambda_j^{\mathbf{C}}t}\|f^I\|_2^2 \qquad \forall t \geq 0. \tag{4.4}$$

If a solution starts instead in one of the four (open) sectors $\mathcal{S}^{\pm\mp}$, it will remain in that sector while approaching the origin. In fact, since $\lambda_1^{\mathbf{C}} < \lambda_2^{\mathbf{C}}$, if $f^I = z_1(v_1 + \gamma v_2)$ for some $z_1 \in \mathbb{R}\setminus\{0\}$ and $\gamma \in \mathbb{R}$, then the solution

$$f(t) = z_1\big(e^{-\lambda_1^{\mathbf{C}}t}v_1 + \gamma e^{-\lambda_2^{\mathbf{C}}t}v_2\big) = z_1 e^{-\lambda_1^{\mathbf{C}}t}\big(v_1 + \gamma e^{-(\lambda_2^{\mathbf{C}}-\lambda_1^{\mathbf{C}})t}v_2\big)$$

of the ODE (2.1) will remain in the sector

$$\mathcal{S}_\gamma^{\pm} := \{z_1(v_1 + z_2 v_2) \mid z_1 \in \mathbb{R}^{\pm}\,,\ z_2 \in [\min(0,\gamma), \max(0,\gamma)]\}, \tag{4.5}$$

see Fig. 1. For a fixed $f^I = z_1(v_1 + \gamma v_2)$, let $\mathcal{S}$ be the corresponding sector $\mathcal{S}_\gamma^{\pm}$. Then estimate (2.5) can be improved as follows

$$\|f(t)\|_2^2 \leq \frac{1}{\lambda_{\min,\mathcal{S}}^{\mathbf{P}}}\|f(t)\|_{\mathbf{P}}^2 \leq \frac{e^{-2\mu t}}{\lambda_{\min,\mathcal{S}}^{\mathbf{P}}}\|f^I\|_{\mathbf{P}}^2 \leq c_{\mathcal{S}}(\mathbf{P})\, e^{-2\mu t}\|f^I\|_2^2, \quad t \geq 0, \tag{4.6}$$

where

$$\lambda_{\min,\mathcal{S}}^{\mathbf{P}} := \inf_{x\in\mathcal{S}} \frac{\langle x,\, \mathbf{P}x\rangle}{\langle x,\, x\rangle}, \quad \lambda_{init,\mathcal{S}}^{\mathbf{P}} := \frac{\langle f^I,\, \mathbf{P}f^I\rangle}{\langle f^I,\, f^I\rangle}, \quad c_{\mathcal{S}}(\mathbf{P}) := \frac{\lambda_{init,\mathcal{S}}^{\mathbf{P}}}{\lambda_{\min,\mathcal{S}}^{\mathbf{P}}}. \tag{4.7}$$

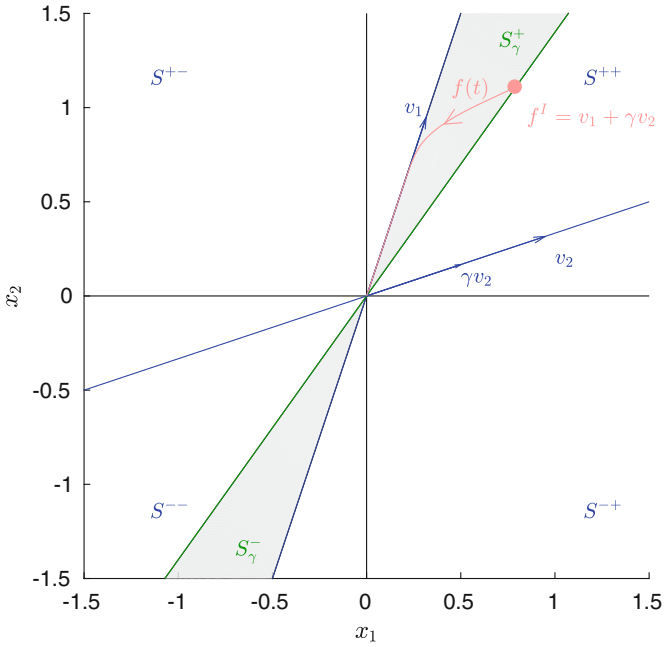


Fig. 1. The blue (black) lines are the eigenspaces span$\{v_1\}$ and span$\{v_2\}$ of matrix $\mathbf{C}$. The red (grey) curve is a solution $f(t)$ of the ODE (2.1) with initial datum f^I. The *shaded regions* are the sectors S^+_γ, S^-_γ with the choice $\gamma = 1/2$. Note: The curves are colored only in the electronic version of this article.

Note that, in the definition of $\lambda^{\mathbf{P}}_{init,\,\mathcal{S}}$ the sector $\mathcal{S} \in \{\mathcal{S}^{\pm}_\gamma | \gamma \in \mathbb{R}\}$ also determines corresponding initial conditions $f^I \in \partial\mathcal{S}$ via $f^I = z_1(v_1 + \gamma v_2)$ (up to the constant $z_1 \neq 0$ which drops out in $\lambda^{\mathbf{P}}_{init,\,\mathcal{S}}$).

For (4.6) to hold for all trajectories and one fixed constant on the right hand side, we have to take the supremum over all initial conditions or, equivalently, over all sectors $\mathcal{S} \in \{\mathcal{S}^{\pm}_\gamma | \gamma \in \mathbb{R}\}$. Although $f^I = z_2 v_2$ is not included in any sector $\mathcal{S}^+_\gamma$, its corresponding multiplicative constant 1 (see (4.4)) is still covered. Then, the minimal multiplicative constant in (3.1) using (4.6) is

$$\widetilde{c} = \sqrt{\inf_{\mathbf{P}} \sup_{\mathcal{S}} c_{\mathcal{S}}(\mathbf{P})}, \tag{4.8}$$

where $\mathbf{P}$ ranges over all matrices of the form (2.3).

Step 1 (computation of $\lambda^{\mathbf{P}}_{\min,\,\mathcal{S}^+_\gamma}$ for γ fixed): To find an explicit expression for this minimal constant c, we first determine $c_{\mathcal{S}}(\mathbf{P})$ for a given admissible matrix $\mathbf{P}$.

As an example of sectors, we consider only $\mathcal{S}_\gamma^+$ for fixed $\gamma \le 0$ and compute

$$\lambda^{\mathbf{P}}_{\min,\,\mathcal{S}_\gamma^+} = \inf_{x\in\mathcal{S}_\gamma^+} \frac{\langle x,\, \mathbf{P}x\rangle}{\|x\|^2} = \inf_{z_1\in\mathbb{R}^+,\, z_2\in[\gamma,0]} \frac{\langle z_1(v_1+z_2v_2),\, \mathbf{P}(z_1(v_1+z_2v_2))\rangle}{\|z_1(v_1+z_2v_2)\|^2}$$
$$= \inf_{z_2\in[\gamma,0]} \frac{\langle v_1+z_2v_2,\, \mathbf{P}(v_1+z_2v_2)\rangle}{\|v_1+z_2v_2\|^2}.$$

This also shows that $\lambda^{\mathbf{P}}_{\min,\,\mathcal{S}_\gamma^+} = \lambda^{\mathbf{P}}_{\min,\,\mathcal{S}_\gamma^-}$ for any fixed $\gamma \in \mathbb{R}$. Next, we use the result of Lemma 3.1 and (3.6), stating that the only admissible matrices are $\mathbf{P} = \mathbf{W}\,\mathrm{diag}(b_1, b_2)\mathbf{W}^*$ for $b_1, b_2 > 0$. Since $c_{\mathcal{S}}(b\mathbf{P}) = c_{\mathcal{S}}(\mathbf{P})$ for all $b > 0$, we consider w.l.o.g. $b_1 = 1/b$ and $b_2 = b$ for $b > 0$. Then, we deduce

$$\lambda^{\mathbf{P}}_{\min,\,\mathcal{S}_\gamma^+} = \inf_{z\in[\gamma,0]} \frac{\langle v_1+zv_2,\, \mathbf{P}(v_1+zv_2)\rangle}{\|v_1+zv_2\|^2}$$
$$= \inf_{z\in[\gamma,0]} \frac{\langle \mathbf{W}^*(v_1+zv_2),\, \mathrm{diag}(1/b, b)\mathbf{W}^*(v_1+zv_2)\rangle}{\|v_1+zv_2\|^2}.$$

In our case of a real matrix $\mathbf{C}$ with distinct real eigenvalues, the left and right eigenvectors are related as follows: Up to a change of orientation, $\langle w_j,\, v_k\rangle = \delta_{jk}$ $(j, k \in \{1, 2\})$. Considering $\langle w_j,\, v_j\rangle = 1$ for $j = 1, 2$, implies that the vectors w_j and v_j can be normalized simultaneously only if matrix $\mathbf{C}$ is symmetric. Therefore, using a coordinate system such that the normalized eigenvectors of $\mathbf{C}^*$ are given as (3.7) and $\mathbf{V} := (v_1|v_2) = (\mathbf{W}^*)^{-1}$ yields

$$v_1 = \frac{1}{\sqrt{1-\alpha^2}}\begin{pmatrix}\sqrt{1-\alpha^2}\\ -\alpha\end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{1-\alpha^2}}\begin{pmatrix}0\\ 1\end{pmatrix} \quad \text{for } \alpha \text{ in (3.7)}.$$

Finally, we obtain

$$\lambda^{\mathbf{P}}_{\min,\,\mathcal{S}_\gamma^+} = \inf_{z\in[\gamma,0]} \frac{\langle \mathbf{W}^*(v_1+zv_2),\, \mathrm{diag}(1/b, b)\mathbf{W}^*(v_1+zv_2)\rangle}{\|v_1+zv_2\|^2} = \inf_{z\in[\gamma,0]} g(z)$$

and $\lambda^{\mathbf{P}}_{init,\,\mathcal{S}_\gamma^+} = g(\gamma)$ with

$$g(z) := \frac{(1-\alpha^2)\,(\frac{1}{b} + bz^2)}{1 - 2\alpha z + z^2}. \tag{4.9}$$

Step 2 (extrema of the function g): The function g has local extrema at

$$z_\pm = \frac{1}{2\alpha b}\Big(b - \frac{1}{b} \pm \sqrt{\big(b-\frac{1}{b}\big)^2 + 4\alpha^2}\Big)$$

which satisfy $z_- < 0 < z_+$. Writing $g'(z) = h_1(z)/h_2(z)$ with $h_1(z) := \big(-2\alpha b z^2 + 2\big(b - \frac{1}{b}\big)z + \frac{2}{b}\alpha\big)$ and $h_2(z) := (1 - 2\alpha z + z^2)^2/(1-\alpha^2) > 0$, we derive

$$g''(z_\pm) = \frac{h_1'(z_\pm)}{h_2(z_\pm)} = \mp 2\frac{1}{h_2(z_\pm)}\sqrt{\big(b-\frac{1}{b}\big)^2 + 4\alpha^2}\,.$$

In fact, the function g attains its global minimum on $\mathbb{R}$ (and on $\mathbb{R}_0^-$) at z_-, and its global maximum on $\mathbb{R}$ at z_+. The global supremum of $g(z)$ on $\mathbb{R}^-$ exists and satisfies

$$\sup_{z\in\mathbb{R}^-} g(z) = \begin{cases} g(0) = (1-\alpha^2)/b & \text{if } b \in (0,1), \\ g(0) = \lim_{z\to-\infty} g(z) = 1-\alpha^2 & \text{if } b = 1, \\ \lim_{z\to-\infty} g(z) = (1-\alpha^2)b & \text{if } b \in (1,\infty). \end{cases}$$

Step 3 (optimization of $c_{\mathcal{S}_\gamma^\pm}(\mathbf{P})$ w.r.t. γ): We obtain

$$c_{\mathcal{S}_\gamma^\pm}(\mathbf{P}(b)) = \frac{g(\gamma)}{\lambda_{\min,\,\mathcal{S}_\gamma^+}^{\mathbf{P}(b)}} = \begin{cases} 1 & \text{if } z_- \le \gamma < 0, \\ g(\gamma)/g(z_-) & \text{if } \gamma \le z_-. \end{cases}$$

Finally, we derive

$$\sup_{\gamma\in\mathbb{R}^-} c_{\mathcal{S}_\gamma^\pm}(\mathbf{P}(b)) = \lim_{\gamma\to-\infty} \frac{g(\gamma)}{g(z_-)} = \frac{(1-\alpha^2)b}{g(z_-)}, \tag{4.10}$$

and in a similar way,

$$\sup_{\gamma\in\mathbb{R}^+} c_{\mathcal{S}_\gamma^\pm}(\mathbf{P}(b)) = \frac{g(z_+)}{g(0)} = \frac{bg(z_+)}{1-\alpha^2}. \tag{4.11}$$

To finish this analysis we note that $c_{\mathcal{S}_0^\pm}(\mathbf{P}(b)) = 1$, due to (4.4) and $f^I = z_1 v_1$.

Step 4 (minimization of $\sup_{\mathcal{S}} c_{\mathcal{S}}(\mathbf{P})$ w.r.t. $\mathbf{P}$): We obtain

$$\inf_{\mathbf{P}} \sup_{\mathcal{S}} c_{\mathcal{S}}(\mathbf{P}) = \inf_{b\in(0,\infty)} \sup_{\gamma\in\mathbb{R}} c_{\mathcal{S}_\gamma^\pm}(\mathbf{P}(b)) = \inf_{b\in(0,\infty)} \max\Big\{\frac{(1-\alpha^2)b}{g(z_-)},\, 1,\, \frac{bg(z_+)}{1-\alpha^2}\Big\}.$$

Taking into account the b-dependence of $z_\pm$, the functions $\frac{(1-\alpha^2)b}{g(z_-)}$ and $\frac{bg(z_+)}{1-\alpha^2}$ are monotone increasing in b, since

$$\frac{\partial}{\partial b}\frac{(1-\alpha^2)b}{g(z_-)} > 0, \qquad \frac{\partial}{\partial b}\frac{bg(z_+)}{1-\alpha^2} > 0.$$

Therefore we have to study their limits as $b \to 0$: We derive

$$\begin{aligned} &\lim_{b\to0} \frac{(1-\alpha^2)b}{g(z_-)} = 1 \qquad \text{using} \quad \lim_{b\to0} z_-(b) = -\infty, \\ &\lim_{b\to0} \frac{bg(z_+)}{1-\alpha^2} = \frac{1}{1-\alpha^2} > 1 \qquad \text{using} \quad \lim_{b\to0} z_+(b) = \alpha. \end{aligned} \tag{4.12}$$

Hence, $\inf_{b\in(0,\infty)} \sup_{\gamma\in\mathbb{R}} c_{\mathcal{S}_\gamma^\pm}(\mathbf{P}(b))$ is realized by the sector $\mathcal{S}_\gamma^\pm$ with $\gamma = z_+(b) > 0$ and in the limit $b \to 0$. Altogether we obtain

$$\widetilde{c} = \sqrt{\inf_{\mathbf{P}} \sup_{\mathcal{S}} c_{\mathcal{S}}(\mathbf{P})} = \frac{1}{\sqrt{1-\alpha^2}},$$

where the first equality holds since we discussed all solutions. This finishes the proof.

Step 5: Finally we have to verify that $\widetilde{c}$ is minimal in (3.1). We shall show that it is attained asymptotically (as $t \to \infty$) for a concrete trajectory: For fixed $b \in (0, \infty)$, the minimal multiplicative constant in (4.6) is attained for the solution with initial datum $f^I = v_1 + z_+(b)v_2 = y_+^{\mathbf{P}(b)}$, which is the eigenvector pertaining to the largest eigenvalue of $\mathbf{P}(b)$ (cp. to the proof of Theorem 3.7). The formula for f^I holds since $\sup_{\mathcal{S}} c_{\mathcal{S}}(\mathbf{P}(b)) = bg(z_+(b))/(1-\alpha^2)$. This can be verified by a direct comparison of (4.10) and (4.11). For b small it also follows from (4.12). In the limit $b \to 0$, $\mathbf{P}(b)$ in (3.9) approaches a multiple of $w_1 \otimes w_1^*$ and

$$f^I = v_1 + z_+(b)v_2 \longrightarrow v_1 + \alpha v_2 = w_1.$$

The solution $f(t)$ of the ODE (2.1) with $f^I = w_1$ satisfies

$$f(t) = e^{-\mathbf{C}t}w_1 = \mathbf{V}\begin{pmatrix} e^{-\lambda_1 t} & 0 \\ 0 & e^{-\lambda_2 t} \end{pmatrix}\mathbf{W}^* w_1 = e^{-\lambda_1 t}v_1 + \alpha e^{-\lambda_2 t}v_2. \tag{4.13}$$

This implies

$$e^{\Re\lambda_1 t}\frac{\|f(t)\|_2}{\|f^I\|_2} \le \|v_1 + \alpha e^{-\Re(\lambda_2-\lambda_1)t}v_2\|_2 \overset{t\to\infty}{\longrightarrow} \|v_1\|_2 = \tfrac{1}{\sqrt{1-\alpha^2}}$$

and it finishes the proof. □

After the analysis in Theorems 3.7 and 4.1, we are left with the case of a matrix $\mathbf{C} \in \mathbb{C}^{2\times 2}$ with eigenvalues λ_1 and λ_2 such that the real and imaginary parts are distinct. This case can not occur for real matrices $\mathbf{C}$. The proof of Lemma 3.3 gives an upper bound $\sqrt{\frac{1+\alpha}{1-\alpha}}$ for the multiplicative constant in (3.1). On the other hand, the solution $f(t)$ of the ODE (2.1) with $f^I = w_1$ satisfies (4.13), hence,

$$\begin{aligned}\|f(t)\|_2^2 &= e^{-2\Re\lambda_1 t}\|v_1 + \alpha e^{-(\lambda_2-\lambda_1)t}v_2\|_2^2 \\ &= \tfrac{1}{1-\alpha^2}e^{-2\Re\lambda_1 t}\Big(1 - 2\alpha^2 e^{-\Re(\lambda_2-\lambda_1)t}\cos\big(\Im(\lambda_2-\lambda_1)t\big) + \alpha^2 e^{-2\Re(\lambda_2-\lambda_1)t}\Big).\end{aligned}$$

The expression in the bracket is bigger than 1, e.g. at time $t = \pi/\Im(\lambda_2-\lambda_1)$. Thus the minimal multiplicative constant c is definitely bigger than $\frac{1}{\sqrt{1-\alpha^2}}$, which is the best constant for $\Im\lambda_1 = \Im\lambda_2$ (see Theorem 4.1).

Next, we derive the upper and lower envelopes for the norm of solutions $f(t)$ of ODE (2.1) in order to determine the sharp constant c. For a diagonalizable matrix $\mathbf{C} \in \mathbb{C}^{2\times 2}$ with $\lambda_1^{\mathbf{C}} = \lambda_2^{\mathbf{C}}$ it holds that $\|f(t)\|_2 = e^{-\Re\lambda_1^{\mathbf{C}} t}\|f^I\|_2$. And for the general case we have:

Proposition 4.2. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix with eigenvalues $\lambda_1^{\mathbf{C}} \neq \lambda_2^{\mathbf{C}}$, and associated eigenvectors v_1 and v_2, resp. Then the norm of solutions $f(t)$ of ODE (2.1) satisfies*

$$h_-(t)\|f^I\|_2^2 \le \|f(t)\|_2^2 \le h_+(t)\|f^I\|_2^2\,, \qquad \forall t \ge 0,$$

where the envelopes $h_\pm(t)$ are given by

$$h_\pm(t) := e^{-2\Re\lambda_1^{\mathbf{C}} t} m_\pm(t)$$

with

$$m_\pm(t) := \pm e^{-\gamma t}\Big(\sqrt{\frac{\big(\cosh(\gamma t) - \alpha^2\cos(\delta t)\big)^2}{(1-\alpha^2)^2} - 1} \pm \frac{\big(\cosh(\gamma t) - \alpha^2\cos(\delta t)\big)}{1-\alpha^2}\Big),$$

where $\gamma := \Re(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})$, $\delta := \Im(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})$, $\alpha := \left|\left\langle \frac{v_1}{\|v_1\|}, \frac{v_2}{\|v_2\|}\right\rangle\right|$ *and* $\alpha \in [0,1)$.

While the rest of the article is based on estimating Lyapunov functionals, the following proof will use the explicit solution formula of the ODE.

Proof. We use again the unitary transformation as in the proof of Lemma 3.3, such that the eigenvectors w_1 and w_2 of $\mathbf{C}^*$ are given in (3.7). If $f(t)$ is a solution of (2.1), then $\tilde f(t) = e^{\lambda_1^{\mathbf{C}} t} f(t)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde f(t) = -\widetilde{\mathbf{C}}\tilde f(t), \quad \tilde f(0) = f^I, \tag{4.14}$$

with

$$\widetilde{\mathbf{C}} = (\mathbf{C} - \lambda_1^{\mathbf{C}}\mathbf{I}) = (\mathbf{W}^*)^{-1}\begin{pmatrix} 0 & 0 \\ 0 & \lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}} \end{pmatrix}\mathbf{W}^*.$$

The explicit solution $\tilde f(t)$ of (4.14) is

$$\tilde f(t) = (\mathbf{W}^*)^{-1}\begin{pmatrix} 1 & 0 \\ 0 & e^{-(\gamma+i\delta)t} \end{pmatrix}\mathbf{W}^* f^I = \begin{pmatrix} f_1^I \\ \frac{\alpha}{\sqrt{1-\alpha^2}}(e^{-(\gamma+i\delta)t} - 1)f_1^I + e^{-(\gamma+i\delta)t} f_2^I \end{pmatrix},$$

where $\gamma = \Re(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})$ and $\delta = \Im(\lambda_2^{\mathbf{C}} - \lambda_1^{\mathbf{C}})$. If the initial data f^I lies in $\mathbb{R}\times\mathbb{C}$ then the solution will satisfy $\tilde f(t) \in \mathbb{R}\times\mathbb{C}$ for all $t \geq 0$. The multiplication with $\overline{f_1^I}/|f_1^I|$ is another unitary transformation and does not change the norm. Therefore, to compute the envelope for the norm of solutions $\tilde f(t)$ of ODE (4.14) we assume w.l.o.g. that

$$f^I_{\phi,\theta} = \begin{pmatrix} \cos(\phi) \\ \sin(\phi)e^{i\theta} \end{pmatrix} \in \mathbb{R}\times\mathbb{C}, \qquad \text{where } \phi,\theta \in [0, 2\pi), \tag{4.15}$$

such that $\|f^I_{\phi,\theta}\| = 1$. We consider the solution $\tilde f_{\phi,\theta}(t)$ for (4.14) with $f^I = f^I_{\phi,\theta}$. To compute the envelopes (for fixed t), we solve $\partial_\phi\|\tilde f_{\phi,\theta}\|^2 = 0$ and $\partial_\theta\|\tilde f_{\phi,\theta}\|^2 = 0$ in terms of ϕ and θ. Evaluating $\|\tilde f_{\phi,\theta}(t)\|^2$ at $\phi = \phi(t)$ and $\theta = \theta(t)$ yields the envelopes for the norm of solutions $\tilde f(t)$ of ODE (4.14). Consequently, we derive the envelopes $h_\pm(t)\|f^I\|^2$ for the original problem, since $\|f(t)\|_2 = e^{-\Re\lambda_1^{\mathbf{C}} t}\|\tilde f(t)\|_2$. □

Corollary 4.3. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable, positive stable matrix. Then the minimal multiplicative constant c in (3.1) for the ODE (2.1) is given by*

$$c = \sqrt{\sup_{t\geq 0} m_+(t)}, \tag{4.16}$$

where $m_+(t)$ is the function given in Proposition 4.2.

In general we could not find an explicit formula for $\sup_{t\geq 0} m_+(t)$.

5 A Family of Decay Estimates for Hypocoercive ODEs

In this section we shall illustrate the interdependence of maximizing the decay rate λ and minimizing the multiplicative constant c in estimates like (3.1). For the ODE-system (2.1), the procedure described in Remark 1.2(b) yields the optimal bound for large time, with the sharp decay rate $\mu := \min\{\Re(\lambda)|\lambda \text{ is an eigenvalue of } \mathbf{C}\}$. But for non-coercive $\mathbf{C}$ we must have $c > 1$. Hence, such a bound cannot be sharp for short time. As a counterexample we consider the simple energy estimate (obtained by premultiplying (2.1) with f^*)

$$\|f(t)\|_2 \leq e^{-\mu_s t}\|f^I\|_2\,, \quad t \geq 0,$$

with $\mathbf{C}_s := \frac{1}{2}(\mathbf{C} + \mathbf{C}^*)$ and $\mu_s := \min\{\lambda|\lambda \text{ is an eigenvalue of } \mathbf{C}_s\}$.

The goal of this section is to derive decay estimates for (2.1) with rates in between this weakest rate μ_s and the optimal rate μ from (2.5). It holds that $\mu_s \leq \mu$. At the same time we shall also present *lower bounds* on $\|f(t)\|_2$. The energy method again provides the simplest example of it, in the form

$$\|f(t)\|_2 \geq e^{-\nu_s t}\|f^I\|_2\,, \quad t \geq 0,$$

with $\nu_s := \max\{\lambda|\lambda \text{ is an eigenvalue of } \mathbf{C}_s\}$. Clearly, estimates with decay rates outside of $[\mu_s, \nu_s]$ are irrelevant.

We present our main result only for the two-dimensional case, as the best multiplicative constant is not yet known explicitly in higher dimensions (cf. Sect. 3):

Proposition 5.1. *Let $\mathbf{C} \in \mathbb{C}^{2\times 2}$ be a diagonalizable positive stable matrix with spectral gap $\mu := \min\{\Re(\lambda_j^{\mathbf{C}})|\, j = 1, 2\}$. Then, all solutions to (2.1) satisfy the following upper and lower bounds:*

(a)

$$\|f(t)\|_2 \leq c_1(\tilde{\mu})\, e^{-\tilde{\mu} t}\|f^I\|_2\,, \quad t \geq 0\,, \quad \mu_s \leq \tilde{\mu} \leq \mu, \tag{5.1}$$

with

$$c_1^2(\tilde{\mu}) = \kappa_{\min}(\beta(\tilde{\mu}))$$

given explicitly in (5.8) below. There, $\alpha \in [0, 1)$ is the cos *of the (minimal) angle of the eigenvectors of $\mathbf{C}^*$ (cf. the proof of Lemma 3.3), and $\beta(\tilde{\mu}) = \max(-\alpha, -\beta_0)$, with β_0 defined in (5.6), (5.7) below.*

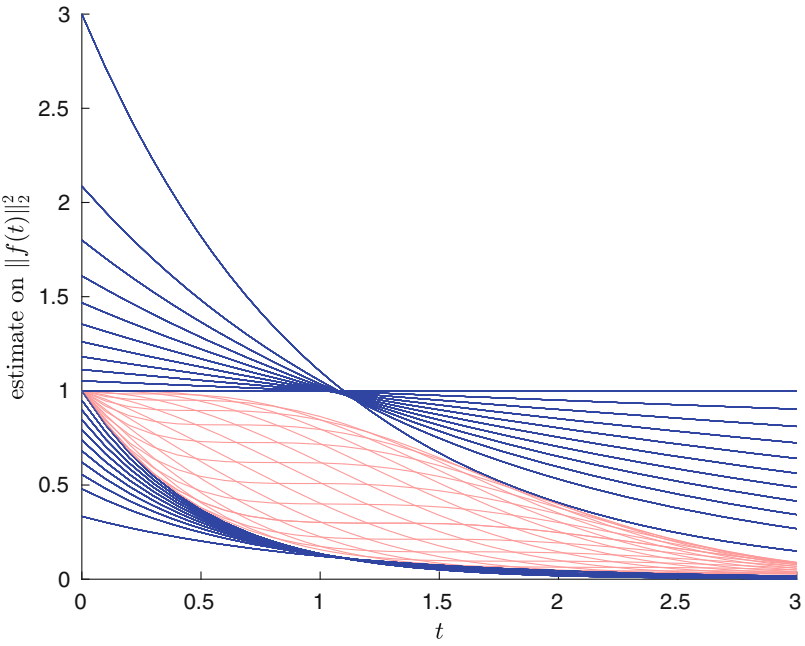


Fig. 2. The red (grey) curves are the squared norm of solutions $f(t)$ for ODE (2.1) with matrix $\mathbf{C} = [1, -1; 1, 0]$ and various initial data f^I with norm 1. The blue (black) curves are the lower and upper bounds for the squared norm of solutions. Note: The curves are colored only in the electronic version of this article.

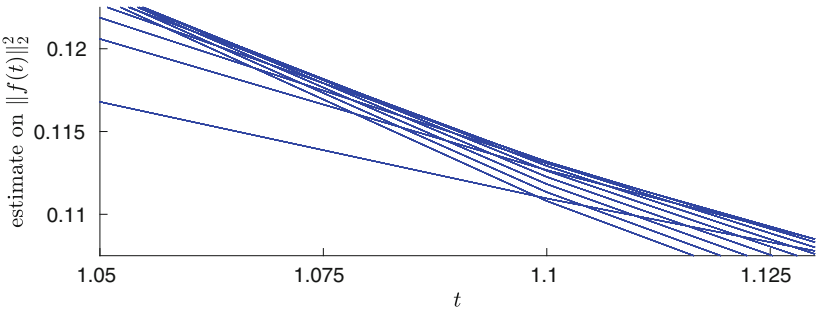


Fig. 3. Zoom of Fig. 2: The curves are the lower bounds for the squared norm of solutions for ODE (2.1) with matrix $\mathbf{C} = [1, -1; 1, 0]$ and various initial data f^I with norm 1.This plot shows that these lower bounds do not intersect in a single point.

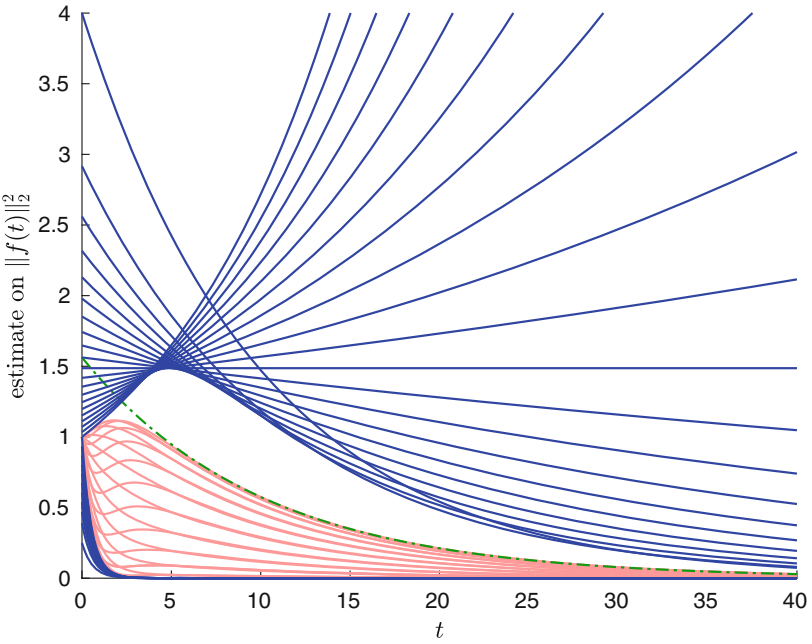


Fig. 4. The red (grey) curves are the squared norm of solutions $f(t)$ for ODE (2.1) with matrix $\mathbf{C} = [19/20, -3/10; 3/10, -1/20]$ and various initial data f^I with norm 1. The blue (black) curves are the lower and upper bounds for the squared norm of solutions derived from Proposition 5.1. The green (black) *dash-dotted curve* is the upper bound for the squared norm of solutions derived from Theorem 4.1. Note: The curves are colored only in the electronic version of this article.

(b)

$$\|f(t)\|_2 \geq c_2(\tilde{\mu})\, e^{-\tilde{\mu}t} \|f^I\|_2\,, \quad t \geq 0\,, \quad \nu \leq \tilde{\mu} \leq \nu_s, \tag{5.2}$$

with $\nu := \max\{\Re(\lambda_j^{\mathbf{C}}) \,|\, j = 1, 2\}$. *The maximal constant*

$$c_2^2(\tilde{\mu}) = \kappa_{\min}(\beta(\tilde{\mu}))^{-1}$$

is given again by (5.8)*, with* α, $\beta(\tilde{\mu})$ *defined as in Part (a).*

Proof. Part (a): For a fixed $\tilde{\mu} \in [\mu_s, \mu]$ we have to determine the smallest constant c_1 for the estimate (5.1), following the strategy of proof from Sect. 3. To this end, we use a unitary transformation of the coordinate system and write $\mathbf{P}(\tilde{\mu}) = \mathbf{W}\mathbf{B}_u\mathbf{W}^*$ with

$$\mathbf{W} = \begin{pmatrix} 1 & \alpha \\ 0 & \sqrt{1-\alpha^2} \end{pmatrix}, \quad \mathbf{B}_u = \begin{pmatrix} 1/b & \beta(\tilde{\mu}) \\ \bar{\beta}(\tilde{\mu}) & b \end{pmatrix}, \tag{5.3}$$

where we set w.l.o.g. $b_1 = 1/b$, $b_2 = b$ with $b > 0$. Moreover, $|\beta|^2 < 1$ has to hold. Now, we have to find the positive definite Hermitian matrix $\mathbf{B}_u$, such that

the analog of (3.3), (3.4) holds, i.e.:

$$\mathbf{A} := \begin{pmatrix} 2(\Re(\lambda_1^{\mathbf{C}}) - \tilde{\mu})/b & (\bar{\lambda}_1^{\mathbf{C}} + \lambda_2^{\mathbf{C}} - 2\tilde{\mu})\beta \\ (\lambda_1^{\mathbf{C}} + \bar{\lambda}_2^{\mathbf{C}} - 2\tilde{\mu})\bar{\beta} & 2(\Re(\lambda_2^{\mathbf{C}}) - \tilde{\mu})b \end{pmatrix} \geq 0, \tag{5.4}$$

As in the proof of Lemma 3.1, we assume that the eigenvalues of $\mathbf{C}$ are ordered as $\Re(\lambda_2^{\mathbf{C}}) \geq \Re(\lambda_1^{\mathbf{C}}) = \mu \geq \tilde{\mu}$. Hence, $\operatorname{Tr}\mathbf{A} \geq 0$. For the non-negativity of the determinant to hold, i.e.

$$\det \mathbf{A} = 4\big(\Re(\lambda_1^{\mathbf{C}}) - \tilde{\mu}\big)\big(\Re(\lambda_2^{\mathbf{C}}) - \tilde{\mu}\big) - |\lambda_1^{\mathbf{C}} + \bar{\lambda}_2^{\mathbf{C}} - 2\tilde{\mu}|^2|\beta|^2 \geq 0, \tag{5.5}$$

we have the following restriction on β:

$$|\beta|^2 \leq \beta_0^2 := \frac{4\big(\Re(\lambda_1^{\mathbf{C}}) - \tilde{\mu}\big)\big(\Re(\lambda_2^{\mathbf{C}}) - \tilde{\mu}\big)}{|\lambda_1^{\mathbf{C}} + \bar{\lambda}_2^{\mathbf{C}} - 2\tilde{\mu}|^2}. \tag{5.6}$$

If $\lambda_1^{\mathbf{C}} + \bar{\lambda}_2^{\mathbf{C}} - 2\tilde{\mu} = 0$, we conclude $\lambda_1^{\mathbf{C}} = \lambda_2^{\mathbf{C}}$ and that we have chosen the sharp decay rate $\tilde{\mu} = \mu$. As the associated, minimal condition number $\kappa(\mathbf{P})$ was already determined in Lemma 3.3, we shall not rediscuss this case here. But to include this case into the statement of the theorem, we set

$$\beta_0 := 1\,, \quad \text{if } \lambda_1^{\mathbf{C}} = \lambda_2^{\mathbf{C}} \text{ and } \tilde{\mu} = \mu. \tag{5.7}$$

From (5.6) we conclude that $\beta_0 \in [0, 1]$. Note that $\beta_0 = 1$ is only possible for $\tilde{\mu} = \mu$ and $\lambda_1^{\mathbf{C}} = \lambda_2^{\mathbf{C}}$, i.e. the case that we just sorted out. For the rest of the proof we hence assume that condition (5.6) holds with $\beta_0 \in [0, 1)$.

For admissible matrices $\mathbf{B}_u$ (i.e. with $b > 0$ and $|\beta| \leq \beta_0$) it remains to determine the matrix

$$\mathbf{P}(b, \beta) = \mathbf{W}\mathbf{B}_u\mathbf{W}^* = \begin{pmatrix} \frac{1}{b} + 2\alpha\Re\beta + b\alpha^2 & (\beta + b\alpha)\sqrt{1-\alpha^2} \\ (\bar{\beta} + b\alpha)\sqrt{1-\alpha^2} & b(1-\alpha^2) \end{pmatrix},$$

(with $\mathbf{W}$ and $\mathbf{B}_u$ given in (5.3)), having the minimal condition number $\kappa\big(\mathbf{P}(b, \beta)\big) = \lambda_+^{\mathbf{P}}(b, \beta)/\lambda_-^{\mathbf{P}}(b, \beta)$. Here

$$\lambda_\pm^{\mathbf{P}}(b, \beta) = \frac{\operatorname{Tr}\mathbf{P}(b, \beta) \pm \sqrt{(\operatorname{Tr}\mathbf{P}(b, \beta))^2 - 4\det\mathbf{P}(b, \beta)}}{2}$$

are the (positive) eigenvalues of $\mathbf{P}(b, \beta)$.

As a first step we shall minimize $\kappa\big(\mathbf{P}(b, \beta)\big)$ w.r.t. b (and for β fixed), since $\operatorname{argmin}_{b>0}\kappa\big(\mathbf{P}(b, \beta)\big)$ will turn out to be independent of β. We notice that $\operatorname{Tr}\mathbf{P}(b, \beta) = b + 2\alpha\Re\beta + 1/b$ is a convex function of $b \in (0, \infty)$ which attains its minimum for $b = 1$. Moreover, $\det\mathbf{P}(b, \beta) = (1-\alpha^2)(1-|\beta|^2) > 0$ is independent of b. This yields the condition number

$$\kappa_{\min}(\beta) = \frac{\lambda_+^{\mathbf{P}}(1, \beta)}{\lambda_-^{\mathbf{P}}(1, \beta)} = \frac{1 + \sqrt{1 - \frac{(1-\alpha^2)(1-|\beta|^2)}{(1+\alpha\Re\beta)^2}}}{1 - \sqrt{1 - \frac{(1-\alpha^2)(1-|\beta|^2)}{(1+\alpha\Re\beta)^2}}}.$$

As a second step we minimize $\kappa_{\min}(\beta)$ on the disk $|\beta| \le \beta_0$. To this end, the quotient $\frac{(1-\alpha^2)(1-|\beta|^2)}{(1+\alpha\Re\beta)^2}$ should be as large as possible. For any fixed $|\beta| \le \beta_0$, this happens by choosing $\beta = -|\beta|$, since $\alpha \in [0,1)$. Hence it remains to maximize the function $g(\beta) := \frac{1-\beta^2}{(1+\alpha\beta)^2}$ on the interval $[-\beta_0, 0]$. It is elementary to verify that g is maximal at $\tilde{\beta} := \max(-\alpha, -\beta_0)$. Then, the minimal condition number is

$$\kappa_{\min}(\tilde{\beta}) = \kappa\big(\mathbf{P}(1,\tilde{\beta})\big) = \frac{1+\sqrt{1-\frac{(1-\alpha^2)(1-\tilde{\beta}^2)}{(1+\alpha\tilde{\beta})^2}}}{1-\sqrt{1-\frac{(1-\alpha^2)(1-\tilde{\beta}^2)}{(1+\alpha\tilde{\beta})^2}}}. \tag{5.8}$$

Part (b): Since the proof of the lower bound is very similar to Part (a), we shall just sketch it. For a fixed $\tilde{\mu} \in [\nu, \nu_s]$ we have to determine the largest constant c_2 for the estimate (5.2). To this end we need to satisfy the inequality

$$\mathbf{C}^*\mathbf{P} + \mathbf{P}\mathbf{C} \le 2\tilde{\mu}\mathbf{P}$$

with a positive definite Hermitian matrix $\mathbf{P}$ with minimal condition number $\kappa(\mathbf{P})$. In analogy to Sect. 2 this would imply

$$\frac{\mathrm{d}}{\mathrm{d}t}\|f(t)\|_{\mathbf{P}}^2 \ge -2\tilde{\mu}\|f(t)\|_{\mathbf{P}}^2,$$

and hence the desired lower bound

$$\|f(t)\|_2^2 \ge (\lambda_{\max}^{\mathbf{P}})^{-1}\|f(t)\|_{\mathbf{P}}^2 \ge (\lambda_{\max}^{\mathbf{P}})^{-1}e^{-2\tilde{\mu}t}\|f^I\|_{\mathbf{P}}^2 \ge (\kappa(\mathbf{P}))^{-1}\, e^{-2\tilde{\mu}t}\|f^I\|_2^2.$$

For minimizing $\kappa(\mathbf{P})$, we again use a unitary transformation of the coordinate system and write $\mathbf{P}$ as $\mathbf{P}(\tilde{\mu}) = \mathbf{W}\mathbf{B}_l\mathbf{W}^*$, with $\mathbf{W}$ from (5.3) and the positive definite Hermitian matrix

$$\mathbf{B}_l = \begin{pmatrix} 1/b & \beta(\tilde{\mu}) \\ \bar{\beta}(\tilde{\mu}) & b \end{pmatrix},$$

with $b > 0$ and $|\beta|^2 < 1$. Then, the matrix $\mathbf{A}$ from (5.4) has to satisfy $\mathbf{A} \le 0$. Since we chose the eigenvalues of $\mathbf{C}$ to be ordered as $\Re(\lambda_1^{\mathbf{C}}) \le \Re(\lambda_2^{\mathbf{C}}) = \nu \le \tilde{\mu}$, we have $\operatorname{Tr}\mathbf{A} \le 0$. The necessary non-negativity of its determinant again reads as (5.5).

In the special case $\lambda_1^{\mathbf{C}} + \bar{\lambda}_2^{\mathbf{C}} - 2\tilde{\mu} = 0$, we conclude again $\lambda_1^{\mathbf{C}} = \lambda_2^{\mathbf{C}}$ and $\tilde{\mu} = \nu$. Hence $\mathbf{A} = 0$. Since β is then only restricted by $|\beta| < 1$, we can again set $\beta_0 = 1$ and obtain the minimal $\kappa(\mathbf{P})$ for $\tilde{\beta}(\nu) = -\alpha$, as in Part (a).

In the generic case, the minimal $\kappa(\mathbf{P})$ is obtained for $\tilde{\beta} = \max(-\alpha, -\beta_0)$ with β_0 given in (5.6). Hence, the maximal constant in the lower bound (5.2) is $c_2^2(\tilde{\mu}) = \kappa_{\min}(\tilde{\beta})^{-1}$ where $\kappa_{\min}$ is given by (5.8). This finishes the proof. □

We illustrate the results of Proposition 5.1 with two examples.

Example 5.2. We consider ODE (2.1) with the matrix

$$\mathbf{C} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$$

which has eigenvalues $\lambda_\pm = (1 \pm i\sqrt{3})/2$, and some normalized eigenvectors of $\mathbf{C}^*$ are, e.g.

$$w_+ = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ \lambda_- \end{pmatrix}, \qquad w_- = \frac{1}{\sqrt{2}} \begin{pmatrix} -\lambda_- \\ 1 \end{pmatrix}. \tag{5.9}$$

The optimal decay rate is $\mu = 1/2$, whereas the minimal and maximal eigenvalues of $\mathbf{C}_s$ are $\mu_s = 0$ and $\nu_s = 1$, respectively. To bring the eigenvectors of $\mathbf{C}^*$ in the canonical form used in the proof of Proposition 5.1, we fix the eigenvector w_+, and choose the unitary multiplicative factor for the second eigenvector w_- as in (5.9) such that $\langle w_+, w_- \rangle$ is a real number. Finally, we use the Gram-Schmidt process to obtain a new orthonormal basis such that the eigenvectors of $\mathbf{C}^*$ in the new orthonormal basis are of the form (3.7) with $\alpha = 1/2$. Then, the upper and lower bounds for the Euclidean norm of a solution of (2.1) are plotted in Figs. 2 and 3. For both the upper and lower bounds, the respective family of decay curves does *not* intersect in a single point (see Fig. 3). Hence, the whole family of estimates provides a (slightly) better estimate on $\|f(t)\|_2$ than if just considering the two extremal decay rates. For the upper bound this means

$$\|f(t)\|_2 \leq \min_{\tilde{\mu} \in [\mu_s, \mu]} c_1(\tilde{\mu})\, e^{-\tilde{\mu} t}\, \|f^I\|_2 \leq \min\{1\,, c_1(\mu)\, e^{-\mu t}\}\, \|f^I\|_2\,, \quad t \geq 0,$$

and for the lower bound

$$\|f(t)\|_2 \geq \max_{\tilde{\nu} \in [\nu, \nu_s]} c_2(\tilde{\nu})\, e^{-\tilde{\nu} t}\, \|f^I\|_2 \geq \max\{c_2(\nu)\, e^{-\nu t}\,, c_2(\nu_s)\, e^{-\nu_s t}\}\, \|f^I\|_2\,, \;\; t \geq 0.$$

Note that the upper bound $\sqrt{3} e^{-t/2}$ with the sharp decay rate $\mu = \frac{1}{2}$ carries the optimal multiplicative constant $c = \sqrt{3}$, as it touches the set of solutions (see Fig. 2). But this is not true for the estimates with smaller decay rates (except of $\tilde{\mu} = 0$). The reason for this lack of sharpness is the fact that the inequality $\|f(t)\|_{\mathbf{P}}^2 \leq e^{-2\tilde{\mu} t} \|f^I\|_{\mathbf{P}}^2$ used in the proof of Proposition 5.1 is, in general, not an equality (in contrast to (2.6)). □

In the next example we consider a matrix $\mathbf{C} \in \mathbb{R}^{2\times 2}$ with $\Re\lambda_1 \neq \Re\lambda_2$, which corresponds to the case analyzed in Theorem 4.1. For such cases the strategy of Proposition 5.1 (based on minimizing $\kappa(\mathbf{P})$) could be improved in the spirit of Theorem 4.1, but we shall not carry this out here. Hence, the estimates of the following example will not be sharp, see Fig. 4.

Example 5.3. We consider ODE (2.1) with the matrix

$$\mathbf{C} = \begin{pmatrix} 19/20 & -3/10 \\ 3/10 & -1/20 \end{pmatrix}$$

which has the eigenvalues $\lambda_1 = 1/20$ and $\lambda_2 = 17/20$, and some normalized eigenvectors of $\mathbf{C}^*$ are, e.g.

$$w_1 = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ -3 \end{pmatrix}, \qquad w_2 = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

The optimal decay rate is $\mu = 1/20$, whereas the minimal and maximal eigenvalues of $\mathbf{C}_s$ are $\mu_s = -1/20$ and $\nu_s = 19/20$, respectively. Since the matrix $\mathbf{C}$ and its eigenvalues are real valued, the eigenvectors of $\mathbf{C}^*$ are already in the canonical form used in the Gram-Schmidt process to obtain a new orthogonal basis such that the eigenvectors of $\mathbf{C}^*$ in the new basis are of the form (3.7) with $\alpha = 3/5$. Then, the upper and lower bounds for the Euclidean norm of a solution of (2.1) are plotted in Fig. 4. Since $\mu_s < 0$, solutions $f(t)$ to this example may initially increase in norm. □

Acknowledgments. All authors were supported by the FWF-funded SFB #F65. The second author was partially supported by the FWF-doctoral school W1245 "Dissipation and dispersion in nonlinear partial differential equations". We are grateful to the anonymous referee who led us to better distinguish the different cases studied in §3 and §4.

References

1. Achleitner, F., Arnold, A., Carlen, E.A.: On linear hypocoercive BGK models. In: Gonçalves, P., Soares, A. (eds.) From Particle Systems to Partial Differential Equations III. Springer Proceedings in Mathematics & Statistics, vol. 162, pp. 1–37. Springer, Cham (2016)
2. Achleitner, F., Arnold, A., Carlen, E.A.: On multi-dimensional hypocoercive BGK models. Kinet. Relat. Models **11**, 953–1009 (2018)
3. Achleitner, F., Arnold, A., Stürzer, D.: Large-Time Behavior in Non-Symmetric Fokker-Planck Equations, vol. 6, pp. 1–68 . Riv. Math. Univ. Parma (N.S.) (2015)
4. Arnold, A., Erb, J.: Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift. arXiv preprint, arXiv:1409.5425 (2014)
5. Arnold, A., Jin, S., Wöhrer, T.: Sharp Decay Estimates in Local Sensitivity Analysis for Evolution Equations with Uncertainties: from ODEs to Linear Kinetic Equations. arXiv preprint, arXiv:1904.01190 (2019)
6. Arnold, V.I.: Ordinary Differential Equations. MIT Press, Cambridge-Mass (1978)
7. 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)
8. Blondel, V.D., Megretski, A. (eds.) Unsolved Problems in Mathematical Systems and Control Theory. Princeton University Press, Princeton (2004)
9. Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Braatz, R.D., Morari, M.: Minimizing the Euclidean condition number. SIAM J. Control Optim. **32**, 1763–1768 (1994)
11. Businger, P.A.: Matrices which can be optimally scaled. Numer. Math. **12**, 346–348 (1968)

12. Dolbeault, J., Mouhot, C., Schmeiser, C.: Hypocoercivity for linear kinetic equations conserving mass. Trans. Amer. Math. Soc. **367**, 3807–3828 (2015)
13. Kolotilina, L.Yu.: Solution of the problem of optimal diagonal scaling for quasireal Hermitian positive definite 3×3 matrices. J. Math. Sci. (N.Y.) **132**, 190–213 (2006)
14. Miclo, L., Monmarché, P.: Étude spectrale minutieuse de processus moins indécis que les autres. In: Donati-Martin, C., Lejay, A., Rouault, A. (eds.) Séminaire de Probabilités XLV, Lecture Notes in Mathematics, vol. 2078, pp. 459–481. Springer, Heidelberg (2013). A summary in English is available at https://www.ljll.math.upmc.fr/~monmarche
15. Sezginer, R.S., Overton, M.L.: The largest singular value of $e^X A_0 e^{-X}$ is convex on convex sets of commuting matrices. IEEE Trans. Automat. Control **35**, 229–230 (1990)
16. Villani, C.: Hypocoercivity. Mem. Amer. Math. Soc. **202**(950), iv+141 (2009)

Adaptive Importance Sampling with Forward-Backward Stochastic Differential Equations

Omar Kebiri[1,2], Lara Neureither[2], and Carsten Hartmann[2(✉)]

[1] Laboratory of Statistics and Random Modeling, University of Abou Bekr Belkaid, Tlemcen, Algeria
[2] Brandenburgische Technische Universität Cottbus-Senftenberg, Cottbus, Germany
{omar.kebiri,lara.neureither,carsten.hartmann}@b-tu.de

Abstract. We describe an adaptive importance sampling algorithm for rare events that is based on a dual stochastic control formulation of a path sampling problem. Specifically, we focus on path functionals that have the form of cumulate generating functions, which appear relevant in the context of, e.g. molecular dynamics, and we discuss the construction of an optimal (i.e. minimum variance) change of measure by solving a stochastic control problem. We show that the associated semi-linear dynamic programming equations admit an equivalent formulation as a system of uncoupled forward-backward stochastic differential equations that can be solved efficiently by a least squares Monte Carlo algorithm. We illustrate the approach with a suitable numerical example and discuss the extension of the algorithm to high-dimensional systems.

Keywords: Importance sampling · Rare events · Path sampling · Forward-backward SDE · Least squares Monte Carlo

1 Introduction

The simulation of rare events is among the key challenges in computational statistical mechanics which involves fields such as molecular dynamics [16], material science [11] or climate modelling [28]. Concrete examples include the study of critical phase transitions in many-particle systems or the estimation of small transition probabilities in protein folding. Estimating small probabilities by Monte Carlo is tricky, because the standard deviation of the corresponding statistical estimator is typically larger than the quantity to be estimated. One technique to improve the efficiency of estimators for small probabilities is importance sampling. Here the idea is to sample from another distribution under which the rare event is no longer rare and then correct (i.e. reweight) the estimator with the appropriate likelihood ratio. Designing such a change of measure so that the variance of the reweighted estimator stays bounded is not at all a trivial task, and several methods have been developed to cope with this issue; for an overview, we refer to the standard textbooks [1,24] and the references therein.

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 265–281, 2019.
https://doi.org/10.1007/978-3-030-15096-9_7

Here we consider adaptive importance sampling strategies where the change of measure is mediated by an exponential tilting of the reference probability measure. For stochastic differential equations, this exponential tilting can be interpreted as a control that changes the drift of the stochastic dynamics. Adaptive importance sampling has been predominantly studied in the context of small noise diffusions, for which the optimal control can be computed from the zero viscosity limit of the corresponding dynamic programming equation [8,9,27]. In this case the value function of the zero viscosity (deterministic) control problem is equal to the large deviations rate function that describes the exponential tails of the rare events under consideration, and as a consequence, the change of measure captures the rare events statistics and results in estimators that, under certain assumptions, have uniformly bounded relative error.

Here we follow a different route, in that we do not resort to large deviations asymptotics, but rather try to compute the zero-variance change of measure from a suitable approximation of the dynamic programming equation that is underlying the stochastic control problem; in contrast to our previous works [16,29], in which the change of measure has been obtained by solving the corresponding variational problem directly, we here focus on the reformulation of the underlying dynamic programming equation as a system of forward-backward stochastic differential equations (see, e.g. [21,25]) that is solved by a least squares Monte Carlo algorithm [4,13]. Our approach is partly inspired by related duality techniques in financial mathematics [17,23], but exploits the specific duality structure of the change of measure problem; see also [22] for a survey of related approaches in financial mathematics.

The paper is organised as follows: In Sect. 2 we introduce our stochastic dynamics, the corresponding path space free energy and its dual variational characterisation. Section 3 deals with the formulation of the free energy sampling problem and the (dual) optimal control problem as a forward-backward stochastic differential equation (FBSDE, in short). The numerical solution of the FBSDE that can be used to either directly compute the free energy or to approximate the optimal control that generates the minimum variance importance sampling scheme is the topic of Sect. 4, with a simple numerical illustration presented in Sect. 5. The article concludes in Sect. 6 with a short summary and a discussion of open problems and future work.

2 Importance Sampling in Path Space

Let $X = (X_s)_{s \geqslant 0}$ be the solution of

$$dX_s = b(X_s, s)\, ds + \sigma(X_s) dB_s\,, \quad X_0 = x\,, \tag{1}$$

where $X_s \in \mathbb{R}^d$, b and σ are smooth drift and noise coefficients, and B is an m-dimensional standard Brownian motion where in general $m \leqslant d$. Our standard example will be a non-degenerate diffusion in an energy landscape,

$$dX_s = -\nabla U(X_s) ds + \sigma dB_s\,, \quad X_0 = x\,, \tag{2}$$

with smooth potential energy function U and $\sigma > 0$ constant. We assume throughout this paper that the functions b, σ, U are such that Eqs. (1) or (2) have unique strong solutions for all $s \geqslant 0$. Now let W be a continuous functional

$$W_\tau(X) = \int_0^\tau f(X_s, s)\, ds + g(X_\tau)\,, \tag{3}$$

of X up to some bounded stopping time τ where f, g are bounded and sufficiently smooth, real valued functions.

Definition 1 (Path space free energy). *Let X be the solution of Eq. (1) and let $W_\tau = W_\tau(X)$ be defined by Eq. (3). The quantity*

$$\gamma = -\log \mathbf{E}\left[\exp(-W_\tau)\right] \tag{4}$$

is called the free energy of W_τ where the expectation is understood with respect to the realisations of (1) for given a initial condition $X_0 = x$.

2.1 Donsker–Varadhan Variational Formula for the Free Energy

The adaptive importance sampling strategy described below is based on a variational characterization of (4) in terms of a change of measure. To make it precise, we define P to be the probability measure on the space $\Omega = C([0,\infty), \mathbb{R}^n)$ of continuous trajectories that is induced by the Brownian motion B in (1). We denote the expectation with respect to P by $\mathbf{E}[\cdot]$. In abstract form, the Donsker–Varadhan variational principle [10] states

$$\gamma = \inf_{Q \ll P} \left\{ \mathbf{E}_Q[W_\tau] + D(Q|P) \right\}\,, \tag{5}$$

where $Q \ll P$ stands for absolute continuity of Q with respect to P, and

$$D(Q|P) = \begin{cases} \displaystyle\int_\Omega \log \frac{dQ}{dP}(\omega)\, dQ(\omega) & \text{if } Q \ll P \\ +\infty & \text{else}\,. \end{cases} \tag{6}$$

denotes the *relative entropy* or *Kullback–Leibler divergence* between Q and P. Note that $D(Q|P) = \infty$ when Q is not absolutely continuous with respect to P, therefore it is sufficient to take the infimum in (5) over all path space measures $Q \ll P$. If $W_\tau \geqslant 0$, it is a simple convexity argument (see, e.g., [7]), which shows that the minimum in Eq. (5) is attained at Q^* given by

$$\left.\frac{dQ^*}{dP}\right|_{\mathcal{F}_\tau} = \exp(\gamma - W_\tau)\,, \tag{7}$$

where $\varphi|_{\mathcal{F}_\tau}$ denotes the restriction of the path space density $\varphi = dQ^*/dP$ to the σ-algebra $\mathcal{F}_\tau \subset \mathcal{E}$ that is generated by the Brownian motion B up to time τ.[1]

[1] More precisely, $\varphi|_{\mathcal{F}_\tau}$ is understood as the restriction of the measure Q^* defined by $dQ^* = \varphi dP$ to the σ-algebra $\mathcal{F}_\tau$ that contains all measurable sets $E \in \mathcal{E}$, with the property that for every $t \geqslant 0$ the set $E \cap \{\tau \leqslant t\}$ is an element of the σ-algebra $\mathcal{F}_t = \sigma(X_s : 0 \leqslant s \leqslant t)$ that is generated by all trajectories $(X_s)_{0 \leqslant s \leqslant t}$ of length t.

By the strict convexity of the exponential function, it holds that Q^*-a.s. [15]

$$\mathbf{E}\left[\exp(-W_\tau)\right] = \exp(-W_\tau)\left(\left.\frac{dQ^*}{dP}\right|_{\mathcal{F}_\tau}\right)^{-1} \tag{8}$$

or, equivalently,

$$\gamma = W_\tau + \log\left(\left.\frac{dQ^*}{dP}\right|_{\mathcal{F}_\tau}\right). \tag{9}$$

That is, Q^* defines a zero-variance change of measure. (Note that the inverse of the Radon–Nikodym derivative in (8) exists since W_τ is bounded.)

2.2 Related Stochastic Control Problem

The only admissible change of measure from P to Q such that $D(Q|P) < \infty$ results in a change of the drift in Eq. (1). Specifically, let u be a process with values in $\mathbb{R}^m$ that is adapted to B and that satisfies

$$\mathbf{E}\left[\exp\left(\frac{1}{2}\int_0^\tau |u_s|^2\, ds\right)\right] < \infty\,. \tag{10}$$

Further define the auxiliary process

$$B_t^u = B_t - \int_0^t u_s\, ds\,,$$

so that (1) can be expressed as

$$dX_s = (b(X_s, s) + \sigma(X_s)u_s)\, ds + \sigma(X_s)dB_s^u\,, \quad X_0 = x\,. \tag{11}$$

By construction, B^u is not a Brownian motion under P, but by Girsanov's Theorem (see, e.g., [19], Theorem 8.6.4) there exists a measure Q defined by

$$\left.\frac{dQ}{dP}\right|_{\mathcal{F}_\tau} = \exp\left(\int_0^t u_s \cdot dB_s^u + \frac{1}{2}\int_0^t |u_s|^2\, ds\right) \tag{12}$$

so that B^u is a standard Brownian motion under Q. (The Novikov condition (10) guarantees that Q is a probability measure.) Inserting (12) into (5), using that B^u is a Brownian motion with respect to Q, it follows that (cf. [6,7]):

$$\gamma = \inf_u \mathbf{E}_Q\left[\int_0^\tau f(X_s, s) + \frac{1}{2}|u_s|^2\, ds + g(X_\tau)\right], \tag{13}$$

with X being the solution of Eq. (11). Since the distribution of B^u under Q is the same as the distribution of B under P, an equivalent representation of the last equation is

$$\gamma = \inf_u \mathbf{E}\left[\int_0^\tau f(X_s^u, s) + \frac{1}{2}|u_s|^2\, ds + g(X_\tau^u)\right]. \tag{14}$$

where X^u is the solution of the controlled SDE

$$dX_s^u = (b(X_s^u, s) + \sigma(X_s^u)u_s)\,ds + \sigma(X_s^u)dB_s\,, \quad X_0^u = x\,, \tag{15}$$

with B being a standard m-dimensional Brownian motion under the probability measure P. The Donsker–Varadhan variational principle (5) and zero-variance property (8) of the probability measure Q^*, for which equality in (5) is attained, have the following stochastic control analogue (see [15, Thm. 3.1]):

Theorem 1. *Let $T > 0$ and $\tau_O = \inf\{s > 0 \colon X_s^u \notin O\}$ for an open and bounded set $O \subset \mathbb{R}^n$ with smooth boundary ∂O. Further define $\tau = \tau_O \wedge T$ and*

$$\Psi(x,t) = \mathbf{E}\left[\exp\left(-\int_t^\tau f(X_s, s)\,ds - g(X_\tau)\right)\middle| X_t = x\right] \tag{16}$$

as the exponential of the negative free energy, considered as a function of the initial condition $X_t = x$ with $0 \leqslant t \leqslant \tau \leqslant T$. Then, the path space measure Q^* induced by the feedback control

$$u_s^* = \sigma(X_s^{u^*})^T \nabla_x \log \Psi(X_s^{u^*}, s) \tag{17}$$

and (15) yields a zero variance estimator, i.e.,

$$\Psi(x,0) = \exp\left(-\int_0^\tau f(X_s^{u^*}, s)\,ds - g(X_\tau^{u^*})\right)\left(\frac{dQ^*}{dP}\bigg|_{\mathcal{F}_\tau}\right)^{-1} \quad Q^* - a.s. \tag{18}$$

3 From Dynamic Programming to Forward-Backward SDE

Following the route taken by [6], it can be shown that the control u^* in (17) is the unique minimiser of the following stochastic control problem: minimise

$$J(u) = \mathbf{E}\left[\int_0^\tau f(X_s^u, s) + \frac{1}{2}|u_s|^2\,ds + g(X_\tau^u)\right] \tag{19}$$

over all measurable and square integrable Markovian controls u, such that the controlled SDE (15) has a unique strong solution. Now let

$$V(x,t) = \min_u \mathbf{E}\left[\int_t^\tau f(X_s^u, s) + \frac{1}{2}|u_s|^2\,ds + g(X_\tau^u)\middle| X_t^u = x\right] \tag{20}$$

be the associated value function (or: optimal cost-to-go). Further define $E = O \times [0,T)$ and let $\partial E^+ = (\partial O \times [0,T)) \cup (O \times \{T\})$ be the terminal set of the augmented process $(X_s^u, s)_{s \geqslant 0}$, such that $\tau = \tau_O \wedge T$ can be recast as

$$\tau = \inf\{s > 0 \colon (X_s^u, s) \notin E\}\,. \tag{21}$$

Assuming sufficient regularity of the coefficients b, σ, f, g and ∂O, a necessary and sufficient condition for $u = u^*$ being optimal is that (see [12, Sec. VI.5])

$$u_s^* = -\sigma(X_s^{u^*})^T \nabla_x V(X_s^{u^*}, s) \tag{22}$$

where $V \in C^{2,1}(E) \cap C(\partial E^+)$ solves the dynamic programming equation

$$\begin{aligned} \partial_t V + LV + h(s, x, V, \sigma^T \nabla_x V) = 0 \quad & \text{in } E \\ V = g \quad & \text{on } \partial E^+ , \end{aligned} \tag{23}$$

with nonlinearity

$$h(s, x, y, z) = -\frac{1}{2}|z|^2 + f(x, s) \tag{24}$$

and the infinitesimal generator of the control-free process X_t,

$$L = \frac{1}{2}\sigma\sigma^T : \nabla_x^2 + b \cdot \nabla_x . \tag{25}$$

For the derivation of (22)–(23) from the Feynman–Kac representation formula for the free energy (4), we refer to [14, Sec. 6].

3.1 FBSDE Representation of the Dynamic Programming Equation

We will now recast the semi-linear, parabolic boundary value problem for $V \in C^{2,1}(E) \cap C(\partial E^+)$. To this end, define the processes

$$Y_s = V(X_s, s) , \quad Z_s = \sigma(X_s)^T \nabla_x V(X_s, s) \tag{26}$$

with X denoting the solution of the uncontrolled SDE (1) with infinitesimal generator (25). Applying Ito's formula to Y, using that V is a classical solution to (23), we obtain the following backward SDE (BSDE)

$$dY_s = -h(s, X_s, Y_s, Z_s)ds + Z_s \cdot dB_s , \quad Y_\tau = g(X_\tau) \tag{27}$$

for the pair (Y, Z). Note that, by definition, Y is continuous and adapted to X, and Z is predictable and a.s. square integrable, i.e.,

$$\int_0^\tau |Z_s|^2 \, ds < \infty , \tag{28}$$

in accordance with the interpretation of Z_s as a control variable. Further note that (27) must be understood as a *backward* SDE rather than a *time-reversed* SDE, since, by definition, Y_s at time $s < \tau$ is measurable with respect to the filtration generated by the Brownian motion $(B_r)_{0 \leqslant r \leqslant s}$, whereas a time-reversed version of Y_s would depend on B_τ via the terminal condition $Y_\tau = g(X_\tau)$, which would require a larger filtration.

By exploiting the specific form of the nonlinearity (24) that appears as the driver h in the backward SDE (27) and the fact that the forward process X is

independent of (Y, Z), we obtain the following representation of the solution to the dynamic programming Eq. (23):

$$\begin{aligned} dX_s &= b(X_s, s)ds + \sigma(X_s)\, dB_s\,, \quad X_t = x \\ dY_s &= -f(X_s, s)ds + \frac{1}{2}|Z_s|^2 + Z_s \cdot dB_s\,, \quad Y_\tau = g(X_\tau)\,. \end{aligned} \tag{29}$$

The solution to (29) now is a triplet (X, Y, Z), and since Y is adapted, it follows that Y_t is a deterministic function of the initial data (x, t) only. Since g is bounded, the results in [18] entail existence and uniqueness of (27); see also [2,3]. As a consequence (see e.g. [20] or [5, Prop. 3.1]),

$$Y_t = V(x, t) \quad \text{(a.s.)} \tag{30}$$

equals the value function of our control problem. Recalling Theorem 1, a straight consequence of Eqs. (14) and (20) therefore is:

Proposition 1. *The free energy (4) is equal to*

$$\gamma = \mathbf{E}[Y_0]\,, \tag{31}$$

where the expectation is over the initial conditions X_0 in $Y_0 = V(X_0, 0)$.

Remark 1. A remark on the role of the control variable Z_s in the BSDE is in order. In (27), let $h = 0$ and consider a random variable ξ that is square-integrable and $\mathcal{F}_\tau$-measurable where $\mathcal{F}_s$ is the σ-Algebra generated by $(B_r)_{0 \leqslant r \leqslant s}$. Ignoring the measurability for a second, a pair of processes (Y, Z) satisfying

$$dY_s = Z_s \cdot dB_s\,, \quad Y_\tau = \xi\,. \tag{32}$$

is $(Y, Z) \equiv (\xi, 0)$, but then Y is not adapted unless the terminal condition ξ is a.s. constant, because Y_t for any $t < \tau$ is not measurable with respect to $\mathcal{F}_s \subset \mathcal{F}_\tau$. An adapted version of Y can be obtained by replacing $Y_t = \xi$ by its best approximation in L^2, i.e. by the projection $Y_t = \mathbf{E}[\xi|\mathcal{F}_t]$. Since the thus defined process Y is a martingale with respect to our filtration, the martingale representation theorem asserts that Y_t must be of the form

$$Y_t = \mathbf{E}[\xi] + \int_0^t \tilde{Z}_s \cdot dB_s\,, \tag{33}$$

for some unique, predictable process $\tilde{Z}$. Subtracting the last equation from $Y_\tau = \xi$ yields

$$Y_t = \xi - \int_t^\tau \tilde{Z}_s \cdot dB_s\,, \tag{34}$$

or, equivalently,

$$dY_t = \tilde{Z}_s \cdot dB_s\,, \quad Y_\tau = \xi\,. \tag{35}$$

Hence $Z_s = \tilde{Z}_s$ in (32) is indeed a control variable that makes Y adapted.

Remark 2. The forward-backward SDE (or: FBSDE) (29) is called *uncoupled* since the forward SDE does not depend on the solution to the associated BSDE, a property that will be exploited in various ways later on.

3.2 Importance Sampling in Path Space, Cont'd

The role of the process Z in the FBSDE representation of the dynamic programming equation is not only to guarantee that Y in (29) is adapted, so that $Y_t = V(x,t)$ is the value function, but it can be literally interpreted as a control since $Z_t = \sigma(X_t)^T \nabla_x V(X_t,t)$. We could compute the optimal control for the zero-variance importance sampling estimator (18) by solving (29) with initial condition $X_t = X_t^u$ on-the-fly, in which case one has to compute the solution of (29) in parallel to the solution of (15). Depending on the nature of the system (in particular the state space dimension) this on-the fly-computation, though computationally demanding, may be nonetheless a sensible alternative to numerical schemes that seek to approximate the value function by globally supported basis functions, which may be an ill-conditioned problem, e.g. if the majority of the trajectories are known to reside inside a small set.

As an alternative that we discuss in detail later on, we suggest to define a feedback control for the controlled SDE (15) by

$$u_t = -\sigma(X_t^u)^T \nabla_x V_K(X_t^u, t)\,, \tag{36}$$

where

$$V_K(x,t) = \sum_{k=1}^{K} \alpha_k(t)\phi_k(x) \tag{37}$$

with $\alpha_k \in \mathbb{R}$ and continuously differentiable (e.g. radial) basis functions ϕ_k is an approximation ansatz for the value function. Then, by Girsanov's Theorem,

$$\mathbf{E}\left[\exp\left(-W_\tau\right)\right] = \mathbf{E}_Q[\exp(-L_\tau^u - W_\tau^u)] \tag{38}$$

where $L_\tau^u = \log(dQ/dP)$ is the log likelihood of the change of measure from P to Q on $\mathcal{F}_\tau$, as given by (12). By continuity of the functional (38), we expect that any unbiased estimator of the right hand side of (38) will have a considerably smaller variance than the plain vanilla estimator (based on independent draws from P), provided that $V_K \approx V$ approximates the value function.

4 Least-Squares Monte Carlo

In this section we discuss the numerical discretisation of the uncoupled FBSDE (29), following an approach that was first suggested by Gobet et al. [13] and later on refined by several authors; here we suggest a semi-parametric approach with radial basis functions based on the work by Bender and Steiner [4].

4.1 Time Stepping Scheme

The fact that the FBSDE (29) is decoupled implies that it can be discretised by an explicit time-stepping algorithm. Here we utilise a variant of the least-squares Monte Carlo algorithm proposed in [13]. The convergence of the numerical schemes for an FBSDE with quadratic nonlinearities in the driver has been

analysed in [26]. The least-squares Monte Carlo scheme is based on the Euler discretisation of (29), specifically,

$$\begin{aligned} \hat{X}_{n+1} &= \hat{X}_n + \Delta t\, b(\hat{X}_n, t_n) + \sqrt{\Delta t}\, \sigma(\hat{X}_n)\xi_{n+1} \\ \hat{Y}_{n+1} &= \hat{Y}_n - \Delta t\, h(\hat{X}_n, \hat{Y}_n, \hat{Z}_n) + \sqrt{\Delta t}\, \hat{Z}_n \cdot \xi_{n+1}\,, \end{aligned} \tag{39}$$

where $(\hat{X}_n, \hat{Y}_n, \hat{Z}_n)$ denotes the numerical discretisation of the joint process (X_s, Y_s, Z_s), where we set $X_s \equiv X_{\tau_O}$ for $s \in (\tau_O, T]$ when $\tau_O < T$, and $(\xi_i)_{i\geqslant 1}$ is an i.i.d. sequence of normalised Gaussian random variables. Now let

$$\mathcal{F}_n = \sigma\big(\{\hat{B}_k : 0 \leqslant k \leqslant n\}\big)$$

be the σ-algebra generated by the discrete Brownian motion $\hat{B}_n := \sqrt{\Delta t}\sum_{i\leqslant n}\xi_i$. By definition, the continuous-time process (X_s, Y_s, Z_s) is adapted to the filtration generated by $(B_r)_{0\leqslant r\leqslant s}$. For the discretised process, this implies

$$\hat{Y}_n = \mathbf{E}\big[\hat{Y}_n|\mathcal{F}_n\big] = \mathbf{E}\big[\hat{Y}_{n+1} + \Delta t\, h(\hat{X}_n, \hat{Y}_n, \hat{Z}_n)|\mathcal{F}_n\big]\,, \tag{40}$$

using that $\hat{Z}_n$ is independent of ξ_{n+1}. In order to compute $\hat{Y}_n$ from $\hat{Y}_{n+1}$, it is convenient to replace $(\hat{Y}_n, \hat{Z}_n)$ on the right hand side by $(\hat{Y}_{n+1}, \hat{Z}_{n+1})$, so that we end up with the fully explicit time stepping scheme

$$\hat{Y}_n = \mathbf{E}\big[\hat{Y}_{n+1} + \Delta t\, h(\hat{X}_n, \hat{Y}_{n+1}, \hat{Z}_{n+1})|\mathcal{F}_n\big]\,. \tag{41}$$

Note that we can use the identification of Z with the optimal control (36) and replace $\hat{Z}_{n+1}$ in the last equation by

$$\hat{Z}_{n+1} = \sigma(\hat{X}_{n+1})^T \nabla V_K(\hat{X}_{n+1}, t_{n+1})\,, \tag{42}$$

where V_K is given by the parametric ansatz (37).

Remark 3. If an explicit representation of $\hat{Z}_n$ such as (42) is not available, it is possible to derive a time stepping scheme for $(\hat{Y}_n, \hat{Z}_n)$ in the following way: multiplying the second equation in (39) by $\xi_{n+1} \in \mathbb{R}^m$ from the left, taking expectations and using the fact that $\hat{Y}_n$ is adapted, it follows that

$$0 = \mathbf{E}\Big[\xi_{n+1}\big(Y_{n+1} - \sqrt{\Delta t}\hat{Z}_n \cdot \xi_{n+1}\big)\Big|\mathcal{F}_n\Big] \tag{43}$$

or, equivalently,

$$\hat{Z}_n = \frac{1}{\sqrt{\Delta t}}\mathbf{E}\big[\xi_{n+1}Y_{n+1}\big|\mathcal{F}_n\big]\,. \tag{44}$$

Together with (41) or, alternatively, with

$$\hat{Y}_n = \mathbf{E}\big[\hat{Y}_{n+1} + \Delta t\, h(\hat{X}_n, \hat{Y}_{n+1}, \hat{Z}_n)|\mathcal{F}_n\big]\,, \tag{45}$$

we have a fully explicit scheme for $(\hat{Y}_n, \hat{Z}_n)$.

4.2 Conditional Expectation

We next address the question how to compute the conditional expectations with respect to $\mathcal{F}_n$. To this end, we recall that the conditional expectation can be characterised as a best approximation in L^2:

$$\mathbf{E}\big[S|\mathcal{F}_n\big] = \underset{Y\in L^2,\,\mathcal{F}_n\text{-measurable}}{\operatorname{argmin}} \mathbf{E}[|Y-S|^2]\,.$$

(Hence the name *least-squares Monte Carlo*.) Here measurability with respect to $\mathcal{F}_n$ means that $(\hat{Y}_n, \hat{Z}_n)$ can be expressed as functions of $\hat{X}_n$. In view of the ansatz (37) and Eq. (41), this suggests the approximation scheme

$$\hat{Y}_n \approx \underset{Y=Y(\hat{X}_n)}{\operatorname{argmin}} \frac{1}{M}\sum_{m=1}^{M}\Big|Y-\hat{Y}_{n+1}^{(m)}-\Delta t\, h\big(\hat{X}_n^{(m)},\hat{Y}_{n+1}^{(m)},\hat{Z}_{n+1}^{(m)}\big)\Big|^2\,, \tag{46}$$

where the data at time t_{n+1} is given in form of M independent realisations of the forward process, $\hat{X}_n^{(m)}$, $m=1,\dots,M$, the resulting values for $\hat{Y}_{n+1}$,

$$\hat{Y}_{n+1}^{(m)} = \sum_{k=1}^{K}\alpha_k(t_{n+1})\phi_k\big(\hat{X}_{n+1}^{(m)}\big)\,, \tag{47}$$

and

$$\hat{Z}_{n+1}^{(m)} = \sigma\big(\hat{X}_{n+1}^{(m)}\big)^T\sum_{k=1}^{K}\alpha_k(t_{n+1})\nabla\phi_k\big(\hat{X}_{n+1}^{(m)}\big)\,. \tag{48}$$

At time $T := N\Delta t$, the data are determined by the terminal cost:

$$\hat{Y}_N^{(m)} = g\big(X_N^{(m)}\big)\,,\quad \hat{Z}_N^{(m)} = \sigma\big(\hat{X}_N^{(m)}\big)^T\nabla g\big(X_N^{(m)}\big) \tag{49}$$

Note that we have defined the forward process so that all trajectories have length T, but the realisations may be constant between τ_O and the terminal time T.

The unknowns that have to be computed in every iteration step are the coefficients α_k, which makes them functions of time, i.e. $\alpha_k = \alpha_k(t_{n+1})$. We call $\hat{\alpha} = (\alpha_1,\dots,\alpha_K)$ the vector of the unknowns, so that the least-squares problem that has to be solved in the n-th step of the backward iteration is of the form

$$\hat{\alpha}(t_n) = \underset{\alpha\in\mathbb{R}^K}{\operatorname{argmin}} \|A_n\alpha - b_n\|^2\,, \tag{50}$$

with coefficients

$$A_n = \Big(\phi_k\big(\hat{X}_n^{(m)}\big)\Big)_{m=1,\dots,M;k=1,\dots,K} \tag{51}$$

and data

$$b_n = \Big(\hat{Y}_{n+1}^{(m)} + \Delta t\, h\big(\hat{X}_n^{(m)},\hat{Y}_{n+1}^{(m)},\hat{Z}_{n+1}^{(m)}\big)\Big)_{m=1,\dots,M}\,. \tag{52}$$

Assuming that the coefficient matrix $A_n \in \mathbb{R}^{M\times K}$, $K \leqslant M$ defined by (51) has maximum rank K, then the solution to (50) is given by

$$\hat{\alpha}(t_n) = \big(A_n^T A_n\big)^{-1} A_n^T b_n\,. \tag{53}$$

Algorithm 1. Least-squares Monte Carlo

Define K, M, N and $\Delta t = T/M$.
Set initial condition $x \in \mathbb{R}^d$.
Choose radial basis functions $\{\phi_k \in C^1(\mathbb{R}^d, \mathbb{R})\colon k = 1, \dots, K\}$.
Generate M independent realisations $\hat{X}^{(1)}, \dots, \hat{X}^{(M)}$ of length N from

$$\hat{X}_{n+1} = \hat{X}_n + \Delta t\, b(\hat{X}_n, t_n) + \sqrt{\Delta t}\, \sigma(\hat{X}_n)\xi_{n+1}\,,\ \hat{X}_0 = x\,.$$

Initialise BSDE by

$$\hat{Y}_N^{(m)} = g\big(\hat{X}_N^{(m)}\big)\,, \quad \hat{Z}_N^{(m)} = \sigma\big(\hat{X}_N^{(m)}\big)^T \nabla g\big(\hat{X}_N^{(m)}\big)\,.$$

for $n = N-1\colon 1$ **do**
Assemble linear system $A_n \hat{\alpha}(t_n) = b_n$ according to (50)–(52).
Evaluate $\hat{Y}_n^{(m)}$ and $\hat{Z}_n^{(m)}$ according to

$$\hat{Y}_n^{(m)} = \sum_{k=1}^{K} \alpha_k(t_n)\phi_k\big(\hat{X}_n^{(m)}\big)\,, \quad \hat{Z}_n^{(m)} = \sigma\big(\hat{X}_n^{(m)}\big)^T \sum_{k=1}^{K} \alpha_k(t_n)\nabla\phi_k\big(\hat{X}_n^{(k)}\big)\,.$$

If necessary, adapt basis functions ϕ_k.
end for

The thus defined scheme that is summarised in Algorithm 1 is strongly convergent of order $1/2$ as $\Delta t \to 0$ and $M, K \to \infty$; see [13]. Controlling the approximation quality for finite values $\Delta t, M, K$, however, requires a careful adjustment of the simulation parameters and appropriate basis functions, especially with regard to the condition number of the matrix A_n, and we will discuss suitable strategies to determine a good basis in the next section.

Remark 4. The accuracy of the solution to the backward SDE depends on wether the distribution of the terminal condition $g(X_\tau)$ is accurately sampled. If the forward process is metastable, however, it may happen that $g(X_\tau)$ is poorly sampled. In this case, it is possible to change the drift of the forward SDE from b to, say, b_0 where b_0 is chosen such that the forward trajectories densely sample the statistic $g(X_\tau)$, without affecting the value function or the resulting optimal control: Assuming that the noise coefficient σ is square and invertible, it is easy to see that the dynamic programming PDE (23) can be recast as

$$\begin{aligned} \partial_t V + \tilde{L}V + \tilde{h}(s, x, V, \sigma^T \nabla_x V) = 0 \quad &\text{in } E \\ V = g \quad &\text{on } \partial E^+\,, \end{aligned}$$

where

$$\tilde{L} = L - (b - b_0)\cdot\nabla$$

is the generator of a forward SDE with drift b_0, and

$$\tilde{h}(x, y, z) = h(x, y, z) + \sigma(x)^{-1}(b(x) - b_0(x)) \cdot z$$

is the driver of the corresponding backward SDE. Hence we can change the drift of the forward SDE at the expense of modifying the running cost, without affecting the optimal control. Changing the drift may be moreover advantageous in connection with the martingale basis approach of Bender and Steiner [4] who have suggested to use basis functions that are defined as conditional expectations of certain linearly independent candidate functions over the forward process, which makes the basis functions martingales. Computing the martingale basis, however, comes with a large computational overhead, which is why the authors consider only cases in which the conditional expectations can be computed analytically. Changing the drift of the forward SDE may thus be used to simplify the forward dynamics so that its distribution becomes analytically tractable.

5 Numerical Illustration

We shall illustrate the previous considerations with a standard example. To this end, we consider a one dimensional diffusion in the double-well potential $U(x) = (x^2 - 1)^2$ that is governed by the equation

$$dX_s = -\nabla U(X_s)ds + \sigma dB_s\,, \quad X_0 = x\,, \tag{54}$$

and want to compute the probability of exiting from the left well $O = \{x < 0\}$ before time $T < \infty$. More specifically, we set $f \equiv 0$ and $g(x) = -\log(\mathbf{1}_{\partial O}(x))$ in Eq. (3) and define the bounded stopping time $\tau = \tau_O \wedge T$ to be the minimum of the first exit time τ_O of the set O and the terminal time T. Note that τ_O is a.s. finite since the potential U is growing sufficiently fast at infinity, so that $(X_s)_{s\geqslant 0}$ is Harris recurrent.

For the equivalent stochastic control problem with the cost

$$J(u) = \mathbf{E}\left[\frac{1}{2}\int_0^\tau |u_s|^2\, ds - \log\left(\mathbf{1}_{\partial O}(X_\tau^u)\right)\right] \tag{55}$$

and the controlled process

$$dX_s^u = (\sigma u_s - \nabla U(X_s^u))\, ds + \sigma dB_s\,, \quad X_0^u = x\,, \tag{56}$$

this means that the control u seeks to push the process towards the set boundary ∂O when $s \approx T$ and the process has not yet left the set O, for otherwise there will an infinite cost to pay.

Since such an infinite terminal cost is numerically difficult to handle, we consider a regularised control problem and replace g by $g^\varepsilon = -\log(\mathbf{1}_{\partial O}(x) + \varepsilon)$; for the numerical calculations, we choose $\varepsilon = 0.01$. The duality relation (5) between the control value $\gamma^\varepsilon = \min_u J(u)$ for fixed initial data $X_0 = x$ and the transition probability $P(\tau_O < T)$ then reads

$$P(\tau_O < T | X_0 = x) = \exp(-\gamma^\varepsilon) - \varepsilon\,. \tag{57}$$

We will compare the results from the FBSDE solution for γ^ε with a reference solution that is obtained from numerically solving the linear PDE

$$\left(\frac{\partial}{\partial t} - L\right)\psi(x,t) = 0\,, \quad (x,t) \in O \times [0,T) \tag{58}$$

together with the boundary conditions[2]

$$\begin{aligned} \psi(0,t) &= 1\,, \quad t \in [0,T) \\ \psi(x,0) &= 0\,, \quad x \in O\,. \end{aligned} \tag{59}$$

Then

$$\psi(x,T) = P(\tau_O < T | X_0 = x)\,. \tag{60}$$

Table 1 below shows the reference value $V^\varepsilon_{ref}(0,x) := -\log(\psi(x,T)+\varepsilon)$, together with the corresponding FBSDE solution. The procedure to obtain the FBSDE solution is described in Algorithm 1, and the table displays the results for different values of $K, M, N = \lfloor T/\Delta t \rfloor$. As basis functions we choose

$$\phi^{\mu_k,\delta}_{k,n}(x) = \exp\left(-\frac{(\mu_k - x)^2}{2\delta}\right)\,, \tag{61}$$

where $\delta = 0.1$ is fixed but $\mu_k = \mu_k(n)$ varies with time such that the forward process can be well covered by the basis functions. More precisely, the centres of the basis functions are chosen by simulating K additional independent forward trajectories $X^{(k)}, k = 1, \ldots, K$ and letting $\mu_k(n) = X^{(k)}_n$. We let the whole algorithm run 20 times and compute empirical mean and variance of V^ε, denoted by $\bar{V}^\varepsilon$ and $S^2(V^\varepsilon)$. The results are shown in the table.

Table 1. Numerical results for the FBSDE scheme described in Algorithm 1.

	$V^\varepsilon_{ref}(0,x)$	$\bar{V}^\varepsilon(0,x)$	$S^2(V^\varepsilon(0,x))$
$K = 8$, $M = 300$, $T = 5$, $\Delta t = 10^{-3}$, $x = -1$, $\sigma = 1$	0.3949	0.3748	10^{-3}
$K = 5$, $M = 300$, $T = 1$, $\Delta t = 10^{-3}$, $x = -1$, $\sigma = 1$	1.7450	1.6446	0.0248
$K = 5$, $M = 400$, $T = 1$, $\Delta t = 10^{-4}$, $x = -1$, $\sigma = 0.6$	4.3030	4.5779	10^{-3}
$K = 6$, $M = 450$, $T = 1$, $\Delta t = 10^{-4}$, $x = -1$, $\sigma = 0.5$	4.5793	4.6044	$5 \cdot 10^{-4}$

[2] For the numerical computation, we add reflecting boundary conditions at $x = -L$ for some $L > 0$, the precise value of which does not affect the results (assuming that it is sufficiently large, say, $L > 3$) since the potential has a 4-th order growth.

Overall we find that the FBSDE scheme results in a fairly good approximation of the value function and, as a consequence of the smoothness of the basis functions, of the optimal control. Moreover, due to the adaptive choice of the basis functions $\{\phi^{\mu_k,\delta}\}$, the results do not seem to be very sensitive to the noise intensity σ or the time horizon T. Speaking of which, we stress that increasing the number of basis functions K is not always advisable, since the matrix A in (51) can easily become rank deficient, especially if σ is small and the trajectories stay close together. Therefore it is crucial to check the rank of A in the simulation and to set K to the value of the maximally observed rank.

5.1 Computational Issues

Let us also discuss the fact that we set $X_s \equiv X_{\tau_O}$ for $s \in (\tau_O, T]$ when $\tau_O < T$ again in more detail. Setting the forward trajectories constant from the exit time on, allows to include the terminal condition $g(X_{\tau\wedge T})$ into the least squares problem at time T, i.e. into the initialisation step b_N, for all backward trajectories. It seems that this stabilises the solution of the backward trajectory $\hat{Y}$. Another approach, following the equations more closely, would be to start each backward trajectories individually from either τ or T depending on whether the corresponding forward trajectory $\hat{X}$ has made an exit or not. This approach induces numerical problems, though, because the data vector (52)—that would normally be dominated by the positive term $\hat{Y}_{n+1}$ when all backward trajectories were starting from T—is now perturbed at the different exit times by the negative value $-\log(\varepsilon)$. This renders the solution $\hat{\alpha}_n$ of the linear Eq. (53) rougher, which in turn leads to fluctuations in the solution of $\hat{Y}_n$ and $\hat{Z}_n$ which can build up and eventually lead to an explosion of the solutions.

Let us further make suggestions how to efficiently treat the case when T is large. We will resort to the ideas of Remark 4 here, which suggests to modify the drift b to b_0 such that under the new drift the event which determines the stopping time τ is not rare anymore. Assume now, that for all trajectories $\hat{X}^{(m)}, m = 1, \ldots, M$ the family of stopping times

$$\tau_O^m = \left\{ s > 0 : X_s^{(m)} \notin O \right\} \tag{62}$$

is dominated by T in the sense that

$$\tilde{T} := \max\{\tau_O^m : m = 1, \ldots, M\} \ll T\,. \tag{63}$$

Then the terminal condition g is essentially known at time $\widetilde{T}$ and the same is true for the backward dynamics. Hence, we suggest in case that T is large to modify the drift such that $\widetilde{T}$ will be small and run the algorithm only up to time $\tilde{T}$. In this case we propose to start each backward trajectory individually from the corresponding exit time on. The matrix A_n is then of size $K \times M_n$ where

$$M_n = \left| \left\{ m : \hat{X}_{n-1}^{(m)} \in O \right\} \right| \tag{64}$$

is the number of trajectories which have not left the set O up to time step n. This ensures that A is not rank deficient at these times which would be the case if we set all trajectory constant after the exit, due to the definition of A_n with

$$(A_n)_{k,m} = \phi_{k,n}^{\mu_k,\delta}(\hat{X}_n^{(m)}) \tag{65}$$

because the basis functions are evaluated at the same constant value for all these trajectories. To the best of our knowledge, the approximation error of the least squares Monte Carlo algorithms with random stopping times has not been analysed so far, and we leave this topic for future work.

We want to add that in contrast to the complexity of numerically solving the HJB equation, which grows exponentially in the dimension d, the complexity of solving the FBSDE is determined by solving the SDE and linear equations, i.e. is at most cubic in d and in the number K of basis functions.

6 Conclusion and Outlook

We have presented a numerical method to compute the free energy of path space functionals of a diffusion process where the functionals may depend on paths having a random length. Free energies of path space functionals appear in connection with rare event simulation and, as a guiding example for this article, we have considered exit probabilities that are relevant in the context of molecular dynamics or risk analysis.

The approach for efficiently computing path space free energies is based on a variational characterisation of the free energy as the value function of an optimal control problem or, equivalently, as an adaptive importance sampling strategy that is based on the optimal control of the aforementioned stochastic control problem; as we have argued, the importance sampling estimator for the free energy enjoys a minimum variance property under the optimal control. Our numerical strategy for solving the underlying stochastic control problem is based on the reformulation of the corresponding semi-linear dynamic programming equation as a forward-backward stochastic differential equation, which can be solved quite efficiently using a least squares Monte Carlo method. For our guiding example, the reformulation of the adaptive importance sampling algorithm as a forward-backward SDE showed promising results.

We have discussed several options that can help to improve the convergence of the least squares algorithm. For example, we have discussed the option of changing the drift of the forward SDE by modifying the cost functional of the corresponding control problem; while this does not change the dynamic programming equation of the underlying control problem, the corresponding forward-backward stochastic differential equations are different, and it is possible to control the speed of convergence of the numerical method in this way, by controlling the random length of the forward trajectories.

Another aspect that we have only briefly touched upon is the choice of the basis functions for the least squares algorithm. A convenient choice are martingale basis functions that, by definition, are non-parametric and adaptive. Eval-

uating the martingale basis requires to compute on-the-fly conditional expectations and it is possible to change the drift of the forward SDE so as to avoid numerically expensive computations of the conditional expectations. In this article we used a semi-parametric approach, and future research should address the non-parametric one. Another interesting topic concerns sampling problems on an infinite time horizon, which can be represented by a stopping time for hitting an *impossible set*, a set which the dynamics can never reach.

We believe that forward-backward SDE are an interesting numerical and analytical tool for applications in computational statistical mechanics that connects such diverse topics as control, filtering and estimation. A specific feature of the proposed method is that the corresponding forward-backward SDE are decoupled, which leaves room for combining the aforementioned tasks with coarse-graining and model reduction techniques. We leave all this for future work.

Acknowledgement. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through the grant CRC 1114 "Scaling Cascades in Complex Systems", Project A05 "Probing scales in equilibrated systems by optimal nonequilibrium forcing". Omar Kebiri received funding from the EU-METALIC II Programme.

References

1. Asmussen, A., Glynn, P.: Stochastic Simulation: Algorithms and Analysis. Springer, New York (2007)
2. Bahlali, K., Gherbal, B., Mezerdi, B.: Existence of optimal controls for systems driven by FBSDEs. Syst. Control Lett. **60**, 344–349 (2011)
3. Bahlali, K., Kebiri, O., Mtiraoui, A.: Existence of an optimal control for a system driven by a degenerate coupled forward-backward stochastic differential equations. C. R. Acad. Sci. Paris, Ser. I **355**(1), 84–89 (2017)
4. Bender, C., Steiner, J.: Least-squares Monte Carlo for BSDEs. In: Carmona et al. (Eds.) Numerical Methods in Finance, pp. 257–289. Springer (2012)
5. Bensoussan, A., Boccardo, L., Murat, F.: Homogenization of elliptic equations with principal part not in divergence form and Hamiltonian with quadratic growth. Commun. Pure Appl. Math. **39**, 769–805 (1986)
6. Boué, M., Dupuis, P.: A variational representation for certain functionals of Brownian motion. Ann. Probab. **26**(4), 1641–1659 (1998)
7. Dai Pra, P., Meneghini, L., Runggaldier, W.J.: Connections between stochastic control and dynamic games. Math. Control Signals Systems **9**, 303–326 (1996)
8. Dupuis, P., Wang, H.: Importance sampling, large deviations, and differential games. Stoch. Int. J. Probab. Stoch. Proc. **76**, 481–508 (2004)
9. Dupuis, P., Wang, H.: Subsolutions of an Isaacs equation and efficient schemes for importance sampling. Math. Oper. Res. **32**, 723–757 (2007)
10. Ellis, R.S.: Entropy, Large Deviations and Statistical Mechanics. Grundlehren der mathematischen Wissenschaften, vol. 271. Springer, New York (1985)
11. Engelund, S., Rackwitz, R.: A benchmark study on importance sampling techniques in structural reliability. Struct. Saf. **12**, 255–276 (1993)
12. Fleming, W.H., Mete Soner, H.: Controlled Markov Processes and Viscosity Solutions. Applications of mathematics, 2nd edn. Springer, New York (2006)

13. Gobet, E., Lemor, J.-P., Warin, X.: A regression-based Monte Carlo method to solve backward stochastic differential equations. Ann. Appl. Probab. **15**, 2172–2202 (2005)
14. Hartmann, C., Banisch, R., Sarich, M., Badowski, Th., Schütte, Ch.: Characterization of rare events in molecular dynamics. Entropy **16**, 350–376 (2014)
15. Hartmann, C., Richter, L., Schütte, Ch., Zhang, W.: Variational characterization of free energy: theory and algorithms. Entropy **19**, 626–653 (2017)
16. Hartmann, C., Schütte, Ch.: Efficient rare event simulation by optimal nonequilibrium forcing. J. Stat. Mech. Theor. Exp. **2012**, 11004 (2012)
17. Haugh, M.B., Kogan, L.: Pricing American options: a duality approach. Oper. Res. **52**, 258–270 (2004)
18. Kobylanski, M.: Backward stochastic differential equations and partial differential equations with quadratic growth. Ann. Probab. **28**(2), 558–602 (2000)
19. Oksendal, B.: Stochastic Differential Equations: An Introduction with Applications, 6th edn. Springer (2010)
20. Pardoux, E., Peng, S.: Backward stochastic differential equations and quasilinear parabolic partial differential equations. In: Rozovskii, B.L., Sowers, R.B. (eds.) Stochastic Partial Differential Equations and their Applications. Lecture Notes in Control and Information Sciences, vol. 176, pp. 200–217. Springer, Berlin (1992)
21. Peng, S.: Backward stochastic differential equations and applications to optimal control. Appl. Math. Optim. **27**, 125–144 (1993)
22. Pham, H.: Continuous-time stochastic control and optimization with financial applications. Stochastic modelling and applied probability. Springer, Heidelberg (2009)
23. Rogers, L.C.G.: Monte Carlo valuation of American options. Math. Finance **12**, 271–286 (2002)
24. Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo Method. Wiley, Hoboken (2008)
25. Touzi, N.: Optimal stochastic control, stochastic target problem, and backward differential equation. Lecture Notes of a course at the Fields Institute (2010). www.cmap.polytechnique.fr/~touzi/Fields-LN.pdf
26. Turkedjiev, P.: Numerical methods for backward stochastic differential equations of quadratic and locally Lipschitz type, Dissertation, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II (2013)
27. Vanden-Eijnden, E., Weare, J.: Rare event simulation of small noise diffusions. Commun. Pure Appl. Math. **65**, 1770–1803 (2012)
28. Wouters, J., Bouchet, F.: Rare event computation in deterministic chaotic systems using genealogical particle analysis. J. Phys. A **49**, 374002 (2016)
29. Zhang, W., Wang, H., Hartmann, C., Weber, M., Schütte, Ch.: Applications of the cross-entropy method to importance sampling and optimal control of diffusions. SIAM J. Sci. Comput. **36**, A2654–A2672 (2014)

Ergodic Properties of Quasi-Markovian Generalized Langevin Equations with Configuration Dependent Noise and Non-conservative Force

Benedict Leimkuhler[1(✉)] and Matthias Sachs[2,3(✉)]

[1] The School of Mathematics and the Maxwell Institute of Mathematical Sciences, James Clerk Maxwell Building, University of Edinburgh, Edinburgh EH9 3FD, UK
b.leimkuhler@ed.ac.uk

[2] Department of Mathematics, Duke University, Box 90320, Durham, NC 27708, USA
msachs@math.duke.edu

[3] The Statistical and Applied Mathematical Sciences Institute (SAMSI), Durham, NC 27709, USA

Abstract. We discuss the ergodic properties of quasi-Markovian stochastic differential equations, providing general conditions that ensure existence and uniqueness of a smooth invariant distribution and exponential convergence of the evolution operator in suitably weighted L^∞ spaces, which implies the validity of central limit theorem for the respective solution processes. The main new result is an ergodicity condition for the generalized Langevin equation with configuration-dependent noise and (non-)conservative force.

Keywords: Generalized Langevin equation · Heat-bath · Quasi-Markovian model · Sampling · Molecular dynamics · Ergodicity · Central limit theorem · Non-equilibrium · Mori-Zwanzig formalism · Reduced model

1 Introduction

Generalized Langevin equations (GLE) arise from model reduction and have many applications such as sampling of molecular systems [4–6,42,60], atom-surface scattering [8], anomalous diffusion in fluids [19], modeling of polymer melts [34], chromosome segmentation in e coli [29], and the modelling of coarse grained particle dynamics [15,33]. The GLE is a non-Markovian formulation, meaning that the evolution of the current state depends not only on the state itself but on the state history. The system is typically formulated with memory terms describing friction with the environment and stochastic forcing. The presence of memory complicates both the analysis of the equation and its numerical solution. In this article, we recall the derivation of the GLE as the result of

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 282–330, 2019.
https://doi.org/10.1007/978-3-030-15096-9_8

Mori-Zwanzig reduction of a large system to model the dynamics of a subset of the variables. We consider the ergodicity of the equation (existence of a unique invariant distribution and exponential convergence of the associated semigroup in a suitably weighted L^∞ space), providing conditions for its validity in case the coefficients of friction and noise depend directly on the reduced position variables.

1.1 The Generalized Langevin Equation

Consider the situation of an open system exchanging energy with a heat bath. If there is a strong time scale separation between the dynamics of the heat bath and the explicitly modelled degrees of freedom, the exchange of energy between these two systems is well modelled by a Markovian process, i.e., dynamic observables such as transport coefficients and first passage times can be well reproduced by a simple Markovian approximation of the heat bath.

By contrast, if we consider a system consisting of a distinguished particle surrounded by a collection of particles of approximately the same mass, then a reduced model where the interaction between the distinguished particle and the solvent particles is replaced by a simple Langevin equation would lead to a poor approximation of the dynamics of the distinguished particle.

In such modelling situations it is necessary to explicitly incorporate memory effects, i.e., non-Markovian random forces and history dependent dissipation. The framework in which such models are typically formulated is that of the generalized Langevin equation. In this article we consider two different types of generalized Langevin equations, both of which are of the form of a stochastic integro differential equation and as such can be viewed as non-Markovian stochastic differential equation (SDE) models.

Let $\Omega_{\boldsymbol{q}} \in \{\mathbb{R}^n, \mathbb{T}^n\}$, where $\mathbb{T}^n = (\mathbb{R}/\mathbb{Z})^n$ denotes the n-dimensional standard torus.[1] We first consider a generalized Langevin equation of the form

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} &= \boldsymbol{F}(\boldsymbol{q}) - \int_0^t \boldsymbol{K}(t-s)\boldsymbol{M}^{-1}\boldsymbol{p}(s)\mathrm{d}s + \boldsymbol{\eta}(t). \end{aligned} \tag{1}$$

where the dynamic variables $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}, \boldsymbol{p} \in \mathbb{R}^n$ denote the configuration variables and conjugate momenta of a Hamiltonian system with energy function

$$H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}, \tag{2}$$

where the mass tensor $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is required to be symmetric positive definite and $U \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R})$ is a smooth potential function so that $\boldsymbol{F} = -\nabla U$ constitutes a conservative force. $\boldsymbol{K} : [0, \infty) \to \mathbb{R}^{n\times n}$ is a matrix-valued function of t, which

[1] The assumption that configurations are restricted to the torus eliminates several technical complications and is motivated by the frequent applications of GLEs in molecular modelling, where such a formulation is commonly used.

is referred to as the memory kernel, and $\boldsymbol{\eta}$ is a stationary Gaussian process taking values in $\mathbb{R}^n$ and which (in equilibrium) is assumed to be statistically independent of $\boldsymbol{q}$ and $\boldsymbol{p}$. We refer to $\boldsymbol{\eta}$ as the noise process or random force. We further assume that a fluctuation-dissipation relation between the random force $\boldsymbol{\eta}$ and the memory kernel holds so that

(i) the random force $\boldsymbol{\eta}$ is unbiased, i.e.,

$$\mathbb{E}[\boldsymbol{\eta}(t)] = \mathbf{0},$$

for all $t \in [0, \infty)$.

(ii) the auto-covariance function of the random force and the memory kernel $\boldsymbol{K}$ coincide up to a constant prefactor, i.e.,

$$\mathbb{E}[\boldsymbol{\eta}(s+t)\boldsymbol{\eta}^\top(s)] = \beta^{-1}\boldsymbol{K}(t), \quad \beta > 0,$$

where the constant $\beta > 0$ corresponds to the inverse temperature of the system under consideration.

Position Dependent Memory Kernels and Non-conservative Forces. To broaden the range of applications for our model, we also consider instances of the generalized Langevin equation where:

(i) the force $\boldsymbol{F}$ is allowed to be non-conservative, i.e., it does not necessarily correspond to the gradient of a potential function,

(ii) the random force is a non-stationary process.

More specifically, we consider the case where the strength of the random force depends on the value of the configurational variable $\boldsymbol{q}$, i.e.,

$$\begin{aligned} \dot{\boldsymbol{q}}(t) &= \boldsymbol{M}^{-1}\boldsymbol{p}(t), \\ \dot{\boldsymbol{p}}(t) &= \boldsymbol{F}(\boldsymbol{q}(t)) - \widetilde{\boldsymbol{K}}(\boldsymbol{q}, t) * \boldsymbol{p} + \widetilde{\boldsymbol{\eta}}(t). \end{aligned} \tag{3}$$

where $\boldsymbol{F} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^n)$ is a smooth vector field, and the random force $\widetilde{\boldsymbol{\eta}}$ is assumed to be of the form

$$\widetilde{\boldsymbol{\eta}}(t) = g^T(\boldsymbol{q}(t))\boldsymbol{\eta}(t),$$

with $\boldsymbol{\eta}$ again satisfying (i) and (ii) and the convolution term, $\widetilde{\boldsymbol{K}}(\boldsymbol{q}, t) * \boldsymbol{p}$, is of the form

$$\widetilde{\boldsymbol{K}}(\boldsymbol{q}, t) * \boldsymbol{p} = g^T(\boldsymbol{q}(t)) \int_0^t \boldsymbol{K}(t-s) g(\boldsymbol{q}(s)) \boldsymbol{p}(s) \mathrm{d}s,$$

with $g \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R}^{n \times n})$ and $\boldsymbol{K}$ as specified above. We motivate the above described type of non-stationary random force and position dependent dissipation term at the end of the following section.

The generic form of the above described GLEs can be derived using a Mori-Zwanzig reduction of the combined Hamiltonian dynamics of an explicit heat bath representation and the system of interest [40,62,63]. In what follows, we briefly outline the Mori-Zwanzig formalism in a simplified setup following the presentation in [15]. We will then consider the particular case of the Kac-Zwanzig model and demonstrate how the above instances of the GLE can be derived from this model.

1.2 Formal Derivation of the Generalized Langevin Equation via Mori-Zwanzig Projection

Consider an ordinary differential equation of the form

$$\begin{aligned} \dot{\boldsymbol{u}} &= f(\boldsymbol{u}, \boldsymbol{v}), \\ \dot{\boldsymbol{v}} &= g(\boldsymbol{u}, \boldsymbol{v}), \end{aligned} \tag{4}$$

subject to the initial condition

$$(\boldsymbol{u}(0), \boldsymbol{v}(0)) = (\boldsymbol{u}_0, \boldsymbol{v}_0), \tag{5}$$

where f, g are smooth functions, i.e., $f \in \mathcal{C}^\infty(\mathbb{R}^{n_u \times n_v}, \mathbb{R}^{n_u})$, $g \in \mathcal{C}^\infty(\mathbb{R}^{n_u \times n_v}, \mathbb{R}^{n_v})$, with $n_{\boldsymbol{v}}, n_{\boldsymbol{u}}$ being positive integers. Also, assume that there is a probability measure $\mu(\mathrm{d}\boldsymbol{u}, \mathrm{d}\boldsymbol{v}) = \rho(\boldsymbol{u}, \boldsymbol{v})\mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v}$ with smooth density $\rho \in \mathcal{C}^\infty(\mathbb{R}^{n_u \times n_v}, [0, \infty))$, which can be associated with a stationary state[2] of the system (4). Consider now the projection operator $\mathscr{P}$, which maps observables $w(\,\cdot\,,\,\cdot\,)$ onto the conditional expectation $\mathscr{P}\boldsymbol{u} \mapsto \mathbb{E}_\mu[w(\boldsymbol{u}, \boldsymbol{v}) \mid \boldsymbol{v}]$, i.e.,

$$(\mathscr{P}w)(\boldsymbol{u}) = \frac{\int_{\mathbb{R}^{n_v}} \rho(\boldsymbol{u}, \boldsymbol{v}) w(\boldsymbol{u}, \boldsymbol{v}) \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v}}{\int_{\mathbb{R}^{n_v}} \rho(\boldsymbol{u}, \boldsymbol{v}) \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v}}.$$

The Mori-Zwanzig projection formalism allows to recast the system (4) as an integro-differential equation (IDE) of the generic form

$$\dot{\boldsymbol{u}}(t) = \bar{f}(\boldsymbol{u}(t)) + \int_0^t K(\boldsymbol{u}(t-s), s)\mathrm{d}s + \eta(\boldsymbol{u}(0), \boldsymbol{v}(0), t), \tag{6}$$

where $\bar{f} = \mathscr{P}f$, $K : \mathbb{R}^{n_u} \times [0, \infty) \to \mathbb{R}^{n_u}$ is a memory kernel, and η is a function of the initial values of $\boldsymbol{u}, \boldsymbol{v}$ and the time variable t. It is important to note that while η depends on the initial condition of both $\boldsymbol{u}$ and $\boldsymbol{v}$ in (4), the remaining terms in the IDE (6) only depend explicitly on the dynamic variable $\boldsymbol{u}$. Similarly as in the stochastic IDEs (1) and (3) the convolution term in (6) can, under appropriate conditions on f, g, be considered as a dissipation term. Likewise, under the assumptions that $\boldsymbol{u}, \boldsymbol{v}$ are initialized randomly according to μ, the term $\eta(\boldsymbol{u}(0), \boldsymbol{v}(0), t)$ in (6) can be interpreted as a random force.

A particularly well studied case is the situation where the functions f and g are such that $(f^T, g^T)^T$ is a Hamiltonian vector field and (4) corresponds to the equation of motion of a Hamiltonian system. In this case a natural choice for μ is the Gibbs-Boltzmann distribution associated with the Hamiltonian. This choice of μ allows us to interpret the degrees of freedom represented by the dynamical variable $\boldsymbol{v}$ as a heat bath or energy reservoir. For example, let $\boldsymbol{u} = (\boldsymbol{q}, \boldsymbol{p}) \in \mathbb{R}^{2n}$, $\boldsymbol{v} = (\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}) \in \mathbb{R}^{2m}$ with $2n = n_{\boldsymbol{u}}, 2m = n_{\boldsymbol{v}}$. We may consider the case where f and g are derived from the Hamiltonian

$$H(\boldsymbol{q}, \boldsymbol{p}, \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}) = V(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{M}^{-1} \boldsymbol{p} + V_c(\boldsymbol{q}, \tilde{\boldsymbol{q}}) + V_h(\tilde{\boldsymbol{q}}) + \frac{1}{2}\tilde{\boldsymbol{p}}^T \widetilde{\boldsymbol{M}}^{-1} \tilde{\boldsymbol{p}}, \tag{7}$$

[2] In the sense that $\mathcal{L}\rho = 0$, with $\mathcal{L}$ being the Liouville operator associated with (4).

where V, V_c, V_h are smooth potential functions such that $V + V_c + V_h$ is confining and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, $\widetilde{\boldsymbol{M}} \in \mathbb{R}^{m\times m}$ are symmetric positive definite matrices. In view of (6) the variables $(\boldsymbol{q}, \boldsymbol{p})$ correspond to the explicitly resolved part of the system; the variables $(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}})$ correspond to the part of the system which is "projected out" and is replaced by the dissipation term and the fluctuation term, thus it functions as the heat bath in the reduced model. The coupling between heat bath and explicitly resolved degrees of freedom is encoded in the form of the coupling potential V_c, and the statistical properties of the heat bath are determined both by the form of the mass matrix $\widetilde{\boldsymbol{M}}$ and the form of the potential V_h.

Let P denote the projection $(\boldsymbol{u}, \boldsymbol{v}) \mapsto \boldsymbol{u}$. The first step in the derivation of the IDE (6) is to rewrite the first line in (4) as

$$\dot{\boldsymbol{u}}(t) = (\mathscr{P}f)\left(P(\boldsymbol{u}(t), \boldsymbol{v}(t))\right) + \left[f(\boldsymbol{u}(t), \boldsymbol{v}(t)) - (\mathscr{P}f)\left(P(\boldsymbol{u}(t), \boldsymbol{v}(t))\right)\right]. \tag{8}$$

Obviously, the first term in (8) corresponds exactly to $\bar{f}(\boldsymbol{u}(t))$ in (6). Let

$$\mathcal{L} = f(\boldsymbol{u}, \boldsymbol{v}) \cdot \nabla_{\boldsymbol{u}} + g(\boldsymbol{u}, \boldsymbol{v}) \cdot \nabla_{\boldsymbol{v}}$$

denote the Liouville operator associated with (4). Noting that

$$\mathcal{L}\left(P(\boldsymbol{u}, \boldsymbol{v})\right) = f(\boldsymbol{u}, \boldsymbol{v}),$$

the term in the square brackets in (8) can be rewritten in semi-group notation as

$$\begin{aligned} f(\boldsymbol{u}(t), \boldsymbol{v}(t)) - (\mathscr{P}f)\left(\boldsymbol{u}(t), \boldsymbol{v}(t)\right) &= e^{t\mathcal{L}}(\boldsymbol{I} - \mathscr{P})f(\boldsymbol{u}(0), \boldsymbol{v}(0)) \\ &= e^{t\mathcal{L}}(\boldsymbol{I} - \mathscr{P})\mathcal{L}P(\boldsymbol{u}(0), \boldsymbol{v}(0)), \end{aligned} \tag{9}$$

where $e^{t\mathcal{L}}$ denotes the flow-map operator associated with the solution of (4), which is defined so that $e^{t\mathcal{L}} w(\boldsymbol{u}(0), \boldsymbol{v}(0)) = w(\boldsymbol{u}(t), \boldsymbol{v}(t))$. The integro-differential form (6) then follows by applying the operator identity

$$e^{t\mathcal{L}} = \int_0^t e^{(t-s)\mathcal{L}} \mathscr{P}\mathcal{L} e^{s(\boldsymbol{I}-\mathscr{P})\mathcal{L}} \mathrm{d}s + e^{t(\boldsymbol{I}-\mathscr{P})\mathcal{L}},$$

which is known as Dyson's formula [41], to the last line in (9) yielding

$$\begin{aligned} e^{t\mathcal{L}}(\boldsymbol{I} - \mathscr{P})\mathcal{L}P(\boldsymbol{u}(0), \boldsymbol{v}(0)) = &\int_0^t e^{(t-s)\mathcal{L}} \mathscr{P}\mathcal{L} e^{s(\boldsymbol{I}-\mathscr{P})\mathcal{L}} (\boldsymbol{I} - \mathscr{P})\mathcal{L}P(\boldsymbol{u}(0), \boldsymbol{v}(0))\mathrm{d}s \\ &+ e^{t(\boldsymbol{I}-\mathscr{P})\mathcal{L}}(\boldsymbol{I} - \mathscr{P})\mathcal{L}P(\boldsymbol{u}(0), \boldsymbol{v}(0)), \end{aligned} \tag{10}$$

where the second term on the right hand side can be identified with η in (6), and the first term in (10) corresponds to the integral term in (6). The form of the last term in (10) suggests that η can be formally written as the solution of a differential equation

$$\begin{aligned} \frac{\partial}{\partial t}\eta(\boldsymbol{u}(0), \boldsymbol{v}(0), t) &= (\boldsymbol{I} - \mathscr{P})\mathcal{L}\eta(\boldsymbol{u}(0), \boldsymbol{v}(0), t), \\ \eta(\boldsymbol{u}(0), \boldsymbol{v}(0), 0) &= f(\boldsymbol{u}(0), \boldsymbol{v}(0)) - (\mathscr{P}f)(\boldsymbol{u}(0)), \end{aligned} \tag{11}$$

which is commonly referred to as the *orthogonal dynamics equation* [7,15].

A couple of remarks are in order. First, we reiterate that the above calculations are purely formal, i.e., the above expressions for the memory kernel K and the fluctuation term η in general do not possess a closed form solution and are therefore often considered as intractable. Moreover, the well-posedness of the orthogonal dynamics Eq. (11) is not obvious and care needs to be taken regarding the existence of solutions and the interpretation of the differential operator $\mathcal{L}$ therein. We refer here to [14] for a rigorous treatment of this equation. We also mention that the above choice of the projection operator $\mathscr{P}$ as a linear operator which maps functions of $(\boldsymbol{u}, \boldsymbol{v})$ into the space of functions of $\boldsymbol{u}$ constitutes a special case of the Mori-Zwanzig formalism. More general forms of the projection operator $\mathscr{P}$ can be considered within the Mori-Zwanzig formalism. For example, the Mori-Zwanzig formalism can be used to derive an IDE for the dynamics of reaction coordinates (collective variables). The corresponding projection operator $\mathscr{P}$ is typically nonlinear in these cases, which can drastically complicate the derivation and the form of the IDE. For a more general presentation of the Mori-Zwanzig projection formalism we refer to the above mentioned papers [7,15] and the references therein as well as the original papers by Mori [40] and Zwanzig [62,63]. In particular the latter paper by Zwanzig considers nonlinear forms of the projection operator $\mathscr{P}$.

Secondly, we point out that in order to derive the stochastic IDEs (1) and (3) an additional step is required. While (1) and (3) are of the form of a stochastic IDE, i.e., they are IDEs driven by a (non-Markovian) stochastic process, the Eq. (6) constitutes an IDE with random initial data, i.e., the system follows a deterministic trajectory after initialization. In the physics literature it is common, in the situation where f, g define a Hamiltonian vector field, to establish equivalence of these systems by virtue of an averaging argument which is considered valid when the system is in equilibrium and $n_{\boldsymbol{v}}$ is sufficiently large (see e.g. [25]).

Drawing a mathematically rigorous connection between (6) and a stochastic IDE which resembles the form of (1) or (3) requires substantial work. As we discuss in the section below, weak convergence as $n_{\boldsymbol{v}} \to \infty$ of the trajectory of $\boldsymbol{u}$ on finite time intervals to the solution of a stochastic integro-differential has been shown in [27,28] for instances of the Ford-Kac model.

The Ford-Kac Model. We consider the Mori-Zwanzig projection formalism in the situation where the ODE (4) corresponds to the equation of motion derived from the Hamiltonian (7). We already mentioned above that the memory kernel K and the fluctuation term in the IDE (6) in general do not possess a closed form solution. A notable exception, however, is the situation of a linearly coupled harmonic heat bath, e.g.,

$$V_c(\boldsymbol{q}, \tilde{\boldsymbol{q}}) = \boldsymbol{q}^T \boldsymbol{A}_c \tilde{\boldsymbol{q}}, \tag{12}$$

with $\boldsymbol{A}_c \in \mathbb{R}^{n \times m}$, and

$$V_h(\tilde{\boldsymbol{q}}) = \frac{1}{2} \tilde{\boldsymbol{q}}^T \boldsymbol{A}_h \tilde{\boldsymbol{q}}, \tag{13}$$

with $\boldsymbol{A}_h \in \mathbb{R}^{m\times m}$ being a symmetric positive (semi-)definite matrix. Under this choice of the potential functions V_c and V_h, the equations of motion associated with (7) are of the form

$$\begin{aligned}
\dot{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\boldsymbol{p},\\
\dot{\boldsymbol{p}} &= -\nabla_{\boldsymbol{q}} V(\boldsymbol{q}) + \boldsymbol{A}_c\tilde{\boldsymbol{q}},\\
\dot{\tilde{\boldsymbol{q}}} &= \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{p}},\\
\dot{\tilde{\boldsymbol{p}}} &= -\boldsymbol{A}_h\tilde{\boldsymbol{q}} + \boldsymbol{A}_c^T\boldsymbol{q}.
\end{aligned} \tag{14}$$

The system (14) was first studied in [13] and is commonly referred to as *Ford-Kac model*. Integrating the 3rd and 4th line of (14) we obtain

$$\begin{pmatrix}\tilde{\boldsymbol{q}}(t)\\ \tilde{\boldsymbol{p}}(t)\end{pmatrix} = e^{t\boldsymbol{R}}\begin{pmatrix}\tilde{\boldsymbol{q}}(0)\\ \tilde{\boldsymbol{p}}(0)\end{pmatrix} + \int_0^t e^{(t-s)\boldsymbol{R}}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{A}_c^T\boldsymbol{q}(s)\end{pmatrix}\mathrm{d}s, \tag{15}$$

where by $\boldsymbol{R}\in\mathbb{R}^{2m\times 2m}$ we denote the matrix

$$\boldsymbol{R} = \begin{pmatrix}\boldsymbol{0} & \widetilde{\boldsymbol{M}}^{-1}\\ -\boldsymbol{A}_h & 0\end{pmatrix}.$$

Partial integration of the integral term in (15) yields

$$\begin{pmatrix}\tilde{\boldsymbol{q}}(t)\\ \tilde{\boldsymbol{p}}(t)\end{pmatrix} = e^{t\boldsymbol{R}}\begin{pmatrix}\tilde{\boldsymbol{q}}(0)\\ \tilde{\boldsymbol{p}}(0)\end{pmatrix} + \boldsymbol{R}^{-1}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{A}_c^T\boldsymbol{q}(t)\end{pmatrix} - \boldsymbol{R}^{-1}e^{t\boldsymbol{R}}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{A}_c^T\boldsymbol{q}(0)\end{pmatrix} + \int_0^t e^{(t-s)\boldsymbol{R}}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{A}_c^T\boldsymbol{p}(s)\end{pmatrix}\mathrm{d}s.$$

Substituting $\tilde{\boldsymbol{q}}$ in the 2nd line by this expression we obtain an IDE of the form (6) with the deterministic vector field $\bar{f}$ being of the form

$$\bar{f}(\boldsymbol{q},\boldsymbol{p}) = \begin{pmatrix}\boldsymbol{M}^{-1}\boldsymbol{p}\\ -\nabla_{\boldsymbol{q}}V(\boldsymbol{q}) - \boldsymbol{A}_c\boldsymbol{A}_h\boldsymbol{A}_c^T\boldsymbol{q}\end{pmatrix},$$

the memory kernel K being of the form

$$K(\boldsymbol{p}(t-s),s) = -\begin{pmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}_c^{-1}\end{pmatrix}e^{(t-s)\boldsymbol{R}}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{A}_c^T\boldsymbol{p}(s)\end{pmatrix}, \tag{16}$$

and the fluctuation term being of the form

$$\eta(\tilde{\boldsymbol{q}}(0),\tilde{\boldsymbol{p}}(0),\boldsymbol{q}(0),t) = e^{t\boldsymbol{R}}\begin{pmatrix}\tilde{\boldsymbol{q}}(0)\\ \tilde{\boldsymbol{p}}(0)\end{pmatrix} - \boldsymbol{R}^{-1}e^{t\boldsymbol{R}}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{A}_c^T\boldsymbol{q}(0)\end{pmatrix}. \tag{17}$$

The Thermodynamic Limit of the Ford-Kac Model. A detailed analysis of the thermodynamic limit $m\to\infty$ of an instance of the Ford-Kac model can be found in [28]; see also [15,27]. The Hamiltonian of the system considered in [28] comprises a single distinguished particle of unit mass, which is subject to an external force associated with the confining potential function $U\in\mathcal{C}^\infty(\mathbb{R},\mathbb{R})$.

The heat bath is modeled by m particles. Each of the heat bath particles is attached by a linear spring to the distinguished particle. The heat bath particles are not subject to any additional force apart from the coupling force. The corresponding Hamiltonian can be written[3]

$$H(\boldsymbol{q},\boldsymbol{p},\tilde{\boldsymbol{q}},\tilde{\boldsymbol{p}}) = \frac{1}{2}\boldsymbol{p}^2 + U(\boldsymbol{q}) + \frac{1}{2}\sum_{j=1}^{m}\frac{\tilde{\boldsymbol{p}}_j^2}{\tilde{m}_j} + \frac{1}{2}\sum_{j=1}^{m} k_j(\tilde{\boldsymbol{q}}_j - \boldsymbol{q}), \tag{18}$$

where $k_j > 0$ corresponds to the stiffness constant of the spring attached to the j-th heat bath particle and $\tilde{m}_j > 0$ is the mass of the j-th heat bath particle. For this system one finds that the terms (16) and (17) take a particular simple form, so that the corresponding IDE can be written as

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{p}, \\ \dot{\boldsymbol{p}} &= -\partial_{\boldsymbol{q}} U(\boldsymbol{q}) - \int_0^t K^{(m)}(t-s)\boldsymbol{p}(s)\mathrm{d}s + \eta^{(m)}(\tilde{\boldsymbol{q}}_i,\tilde{\boldsymbol{p}}_i,t), \end{aligned} \tag{19}$$

where the memory kernel is of the form

$$K^{(m)}(t) = \sum_{i=1}^{m} k_i \cos(\omega_i t),$$

and the fluctuation term is of the form

$$\eta^{(m)}(\tilde{\boldsymbol{q}}_i,\tilde{\boldsymbol{p}}_i,t) = \sum_{i=1}^{m}\sqrt{\frac{k_i}{\beta}}\Big(\tilde{\boldsymbol{q}}_i(0)\cos(\omega_i t) + \tilde{\boldsymbol{p}}_i(0)\sin(\omega_i t)\Big),$$

with $\omega_j = \sqrt{k_j/\tilde{m}_j}$. If the initial conditions of the heat bath particles are assumed to be distributed according to the Gibbs measure associated with (18) and the statistical distribution of the values of k_j and $\tilde{m}_j$ are controlled in a certain way as $m \to \infty$, it can been shown that for any finite $T > 0$ the trajectories of the solution of (19) converge weakly within the interval $[0,T]$ to solutions of a stochastic IDE of the form (1); for a precise statement see [28, Theorem 4.1].

The Kac-Zwanzig Model. The Kac-Zwanzig model (see [63]) is a generalization of the Ford-Kac model, the heat bath is still harmonic, i.e., V_h has the general form (13), but the coupling potential is such that the coupling force is linear in $\tilde{\boldsymbol{q}}$ but non-linear in $\boldsymbol{q}$, i.e.,

$$V_c(\boldsymbol{q},\tilde{\boldsymbol{q}}) = \boldsymbol{G}(\boldsymbol{q})\tilde{\boldsymbol{q}},$$

where $\boldsymbol{G} \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R}^{n\times m})$. For such a system a closed form solution of the terms in the Mori-Zwanzig projection (6) can still be derived (see [63] or [17] for a

[3] One easily verifies that this Hamiltonian corresponds to a parametrization of (7) as $\boldsymbol{M} = 1$, $\widetilde{\boldsymbol{M}} = \mathrm{diag}(\tilde{m}_1,\ldots,\tilde{m}_m)$, $V(\boldsymbol{q}) = U(\boldsymbol{q}) + \frac{1}{2}\sum_{i=1}^{m} k_i \boldsymbol{q}^2$, $V_c(\boldsymbol{q},\tilde{\boldsymbol{q}}) = \sum_{i=1}^{m} k_i \boldsymbol{q}\tilde{\boldsymbol{q}}_i$, $V_h(\tilde{\boldsymbol{q}}) = \frac{1}{2}\sum_{i=1}^{m} k_i \tilde{\boldsymbol{q}}_i^2$.

detailed derivation). However, unlike in the situation of the Ford-Kac model the closed form solution of the memory kernel K and the fluctuation term η are functions of $\boldsymbol{q}$. This observation motivates the study of GLEs of the form (3). Instances of (3) which are derived from such a Kac-Zwanzig heat bath model can be found for example in [25,43,44,56].

We note that an elegant alternative derivation of the GLE can be obtained beginning from a model of an infinite-dimensional heat-bath. Such models have been extensively studied in [21–23], and in a (non-equilibrium) context by Rey-Bellet and coworkers in [11,12,49,51].

1.3 Main Results and Organization of the Paper

In this article we focus on instances of the GLEs (1) and (3) (or, more precisely, (26)), which can be represented in an extended phase space as an Itô diffusion process. We refer to such GLEs, which possess a Markovian representation in an extended phase space as quasi-Markovian generalized Langevin equations (QGLEs). We specify the extended variable formalism, i.e., the particular form of the Itô diffusion processes which we consider for a Markovian representation of GLEs, in the following Sect. 2. In that section we also review results from the literature on the Markovian representation and approximation of generalized Langevin equations. The main results of this article are contained in Theorems 1 to 4 which we present in Sect. 3. In these theorems we provide criteria which ensure geometric ergodicity for the Markovian representation of GLEs of the form (1) and (3). Since the extended variable formalism which we consider in this article is in various ways more general than the extended variable formalisms considered for ergodicity proofs in previous works in the literature our results cover a wide class of GLEs, which have previously not been shown to be (geometrically) ergodic and which are of high interest in applications (for a detailed discussion see the notes at the end of Sect. 3.1). In particular, showing (geometric) ergodicity for QGLEs with non-conservative forces and/or stated dependent memory kernels is a novel contribution of this paper. As a consequence of the geometric ergodicity we can derive in a generic way the validity of a central limit theorem (see Corollary 1) for the solution processes of the respective GLEs. For the proofs of the Theorems 1 to 4 suitable Lyapunov functions must be constructed and the validity of a minorization condition ensured; see Sects. 3.3 and 3.4 for details, and Appendix B for a general overview of the employed framework. For the proof on the existence of suitable Lyapunov functions we use a similar ansatz as in previous works (compare in particular with [36,46]), but we require additional linear algebra arguments due to increased generality of our extended variable formalism. The proof of the validity of the minorization condition in the case of position dependent coefficients requires a non-standard alternation of the common techniques. We show the existence of a minorizing measure by virtue of a Girsanov transformation.

2 Markovian Representation of Generalized Langevin Equations with Configuration Dependent Noise

In this section we derive a Markovian representation of the GLEs introduced in Sect. 1. We start with an Itô diffusion process of the form

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} &= \boldsymbol{F}(\boldsymbol{q}) - \widetilde{\boldsymbol{\Gamma}}_{1,1}(\boldsymbol{q})\boldsymbol{M}^{-1}\boldsymbol{p} - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\boldsymbol{s} + \beta^{-1/2}\widetilde{\boldsymbol{\Sigma}}_1(\boldsymbol{q})\,\dot{\boldsymbol{W}}, \\ \dot{\boldsymbol{s}} &= -\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q})\boldsymbol{M}^{-1}\boldsymbol{p} - \widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q})\boldsymbol{s} + \beta^{-1/2}\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q})\,\dot{\boldsymbol{W}}, \\ \text{with } &\big(\boldsymbol{q}(0), \boldsymbol{p}(0), \boldsymbol{s}(0)\big) \sim \mu_0, \end{aligned} \tag{20}$$

where $\boldsymbol{M}, \boldsymbol{F}, \beta$ are as previously defined. In particular $\boldsymbol{F}$ may correspond to the negative gradient of a smooth and confining potential function $U \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R})$, i.e., $\boldsymbol{F} = -\nabla U$. Furthermore,

(i) the auxiliary variable $\boldsymbol{s}(t)$ takes values in $\mathbb{R}^m$ with $m \geq n$,

(ii) $\dot{\boldsymbol{W}} = [\dot{W}_i]_{1\leq i\leq n+m}$ is a vector of $(n+m)$ independent Gaussian white-noise components, i.e., $\dot{W}_i \sim \mathcal{N}(0,1)$ and $\mathbb{E}[\dot{W}_i(t)\dot{W}_j(s)] = \delta_{ij}\delta(t-s)$.

(iii) $\widetilde{\boldsymbol{\Gamma}}_{i,j}, \widetilde{\boldsymbol{\Sigma}}_i$, $i = 1, 2$ are matrix valued functions so that for $m \geq n$,

$$\widetilde{\boldsymbol{\Gamma}} = \begin{pmatrix} \widetilde{\boldsymbol{\Gamma}}_{1,1} & \widetilde{\boldsymbol{\Gamma}}_{1,2} \\ \widetilde{\boldsymbol{\Gamma}}_{2,1} & \widetilde{\boldsymbol{\Gamma}}_{2,2} \end{pmatrix} \in \mathcal{C}^\infty\ (\Omega_{\boldsymbol{q}}, \mathbb{R}^{(n+m)\times(n+m)}\ ).$$

and

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \widetilde{\boldsymbol{\Sigma}}_{1,1} & \widetilde{\boldsymbol{\Sigma}}_{1,2} \\ \widetilde{\boldsymbol{\Sigma}}_{2,1} & \widetilde{\boldsymbol{\Sigma}}_{2,2} \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{\Sigma}}_1 \\ \widetilde{\boldsymbol{\Sigma}}_2 \end{pmatrix} \in \mathcal{C}^\infty\ (\Omega_{\boldsymbol{q}}, \mathbb{R}^{(n+m)\times(n+m)}\ ),$$

i.e.,

$$\widetilde{\boldsymbol{\Gamma}}_{1,1} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{n\times n}),\ \widetilde{\boldsymbol{\Gamma}}_{2,1}^T, \widetilde{\boldsymbol{\Gamma}}_{1,2} \in (\Omega_{\boldsymbol{q}}, \mathbb{R}^{n\times m}),\ \widetilde{\boldsymbol{\Gamma}}_{2,2} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{m\times m}),$$

and

$$\widetilde{\boldsymbol{\Sigma}}_1 \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{n\times(n+m)}),\ \widetilde{\boldsymbol{\Sigma}}_2 \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{m\times(n+m)}).$$

(vi) The probability measure μ_0 is such that $(\boldsymbol{q}(0), \boldsymbol{p}(0), \boldsymbol{s}(0))$ has finite first and second moments. In particular,

$$\int_{\Omega_{\boldsymbol{q}}\times\mathbb{R}^{n+m}} \|\boldsymbol{q}\|_2^2 + \|\boldsymbol{p}\|_2^2 + \|\boldsymbol{s}\|_2^2\ \mu_0(\mathrm{d}\boldsymbol{q}, \mathrm{d}\boldsymbol{p}, \mathrm{d}\boldsymbol{s}) < \infty.$$

Notation. In the sequel, we write $\boldsymbol{x}^T := (\boldsymbol{q}^T, \boldsymbol{p}^T, \boldsymbol{s}^T)$, as well as $\boldsymbol{z}^T := (\boldsymbol{p}^T, \boldsymbol{s}^T)$ as shorthand notation for the phase space and auxiliary variables, and we use $\Omega_{\boldsymbol{x}} := \Omega_{\boldsymbol{q}} \times \Omega_{\boldsymbol{p}} \times \Omega_{\boldsymbol{s}}$, and $\Omega_{\boldsymbol{z}} := \Omega_{\boldsymbol{p}} \times \Omega_{\boldsymbol{s}}$, where $\Omega_{\boldsymbol{p}} = \mathbb{R}^n, \Omega_{\boldsymbol{s}} = \mathbb{R}^m$, as shorthand notation for the corresponding domains. With some abuse of notation we also denote points in $\Omega_{\boldsymbol{x}}, \Omega_{\boldsymbol{z}}, \Omega_{\boldsymbol{q}}, \Omega_{\boldsymbol{p}}, \Omega_{\boldsymbol{s}}$ by $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}$, respectively.

Associated Generator. We denote the generator of (20) by

$$\mathcal{L}_{\mathrm{GLE}} = \mathcal{L}_H + \mathcal{L}_O, \tag{21}$$

where $\mathcal{L}_H$ and $\mathcal{L}_O$, which when considered as operators on $\mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R})$, have the form

$$\mathcal{L}_H := \boldsymbol{F}(\boldsymbol{q}) \cdot \nabla_{\boldsymbol{p}} + \boldsymbol{M}^{-1}\boldsymbol{p} \cdot \nabla_{\boldsymbol{q}},$$

and

$$\mathcal{L}_O := -\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q}) \begin{pmatrix} \boldsymbol{M}^{-1}\boldsymbol{p} \\ \boldsymbol{s} \end{pmatrix} \cdot \nabla_{\boldsymbol{z}} + \frac{\beta^{-1}}{2} \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q}) \widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q}) : \nabla_{\boldsymbol{z}}^2,$$

where

$$\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q}) \widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q}) : \nabla_{\boldsymbol{z}}^2 = \sum_{i=1}^{M} \sum_{j=1}^{M} \left[\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q}) \widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q}) \right]_{i,j} \partial_{\boldsymbol{z}_i} \partial_{\boldsymbol{z}_j}, \ M = n + m.$$

Derivation of the Associated Stochastic IDE. In what follows we relate the system (20) to a non-Markovian stochastic IDE. Consider the following convolution functional

$$\begin{aligned} \widetilde{\boldsymbol{K}}_{\widetilde{\Gamma}}(\boldsymbol{q}, t) * \boldsymbol{p} &= \widetilde{\boldsymbol{\Gamma}}_{1,1}(\boldsymbol{q}(t)) \boldsymbol{M}^{-1} \boldsymbol{p}(t) \\ &\quad - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t)) \int_0^t \exp\left(- \int_s^t \widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q}(r)) \mathrm{d}r \right) \widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}(s)) \boldsymbol{M}^{-1} \boldsymbol{p}(s) \mathrm{d}s, \end{aligned} \tag{22}$$

and a random force of the form

$$\widetilde{\boldsymbol{\eta}}(t) = \widetilde{\boldsymbol{\eta}}_w(t) + \widetilde{\boldsymbol{\eta}}_c(t),$$

where

$$\widetilde{\boldsymbol{\eta}}_w(t) := \beta^{-1/2} \widetilde{\boldsymbol{\Sigma}}_1(\boldsymbol{q}(t)) \, \dot{\boldsymbol{W}}(t), \tag{23}$$

and

$$\widetilde{\boldsymbol{\eta}}_c(t) := -\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t)) \boldsymbol{\eta}_c(t), \tag{24}$$

with $\boldsymbol{\eta}_c$ being the solution of the linear SDE

$$\dot{\boldsymbol{\eta}}_c(t) = -\widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q}(t)) \boldsymbol{\eta}_c(t) + \beta^{-1/2} \widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}(t)) \, \dot{\boldsymbol{W}}(t), \quad \boldsymbol{\eta}_c(0) = \boldsymbol{s}(0). \tag{25}$$

As shown in the following proposition, under this assumption, the SDE (20) can be rewritten as a stochastic IDE of the form

$$\begin{aligned} \dot{\boldsymbol{q}}(t) &= \boldsymbol{M}^{-1} \boldsymbol{p}(t), \\ \dot{\boldsymbol{p}}(t) &= \boldsymbol{F}(\boldsymbol{q}(t)) - \widetilde{\boldsymbol{K}}_{\widetilde{\Gamma}}(\boldsymbol{q}, t) * \boldsymbol{p} + \widetilde{\boldsymbol{\eta}}(t). \end{aligned} \tag{26}$$

Proposition 1. *If a (weak) solution of* $(\boldsymbol{q}(t), \boldsymbol{p}(t), \boldsymbol{s}(t))$ *of* (20) *exists for all times* $t \geq 0$*, the SDE* (20) *can be rewritten in the form* (26)*.*

Proof. The solution for $\boldsymbol{s}$ in (20) can be written as

$$\begin{aligned}\boldsymbol{s}(t) = \boldsymbol{\Phi}(t,0,\boldsymbol{q})\boldsymbol{s}(0) - \int_0^t \boldsymbol{\Phi}(t,s,\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}(s))\boldsymbol{M}^{-1}\boldsymbol{p}(s)ds \\ + \int_0^t \boldsymbol{\Phi}(t,s,\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}(s))\mathrm{d}\,\boldsymbol{W}(s),\end{aligned} \tag{27}$$

with

$$\boldsymbol{\Phi}(t,s,\boldsymbol{q}) = \exp\left(-\int_s^t \boldsymbol{\Gamma}_{2,2}(\boldsymbol{q}(r))dr\right). \tag{28}$$

Substituting $\boldsymbol{s}(t)$ in the second equation of (20) by the right hand side of (27) we obtain

$$\begin{aligned}\dot{\boldsymbol{p}}(t) = \boldsymbol{F}(\boldsymbol{q}(t)) - \widetilde{\boldsymbol{\Gamma}}_{1,1}(\boldsymbol{q}(t))\boldsymbol{M}^{-1}\boldsymbol{p}(t) \\ + \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t))\int_0^t \boldsymbol{\Phi}(t,s,\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}(s))\boldsymbol{M}^{-1}\boldsymbol{p}(s)\mathrm{d}s - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t))\boldsymbol{\Phi}(t,0,\boldsymbol{q})\boldsymbol{s}(0) \\ - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t))\int_0^t \boldsymbol{\Phi}(t,s,\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}(s))\mathrm{d}\,\boldsymbol{W}(s) + \widetilde{\boldsymbol{\Sigma}}_1(\boldsymbol{q}(t))\mathrm{d}\,\boldsymbol{W}(t).\end{aligned}$$

As the solution of (25), $\boldsymbol{\eta}_c(t)$ can be written as

$$\boldsymbol{\eta}_c(t) = \boldsymbol{\Phi}(t,0,\boldsymbol{q})\boldsymbol{s}(0) - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t))\int_0^t \boldsymbol{\Phi}(t,s,\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}(s))\mathrm{d}\,\boldsymbol{W}(s),$$

and we find:

$$\begin{aligned}\dot{\boldsymbol{p}}(t) &= \boldsymbol{F}(\boldsymbol{q}(t)) - \widetilde{\boldsymbol{K}}_{\widetilde{\boldsymbol{\Gamma}}}(\boldsymbol{q},t) * \boldsymbol{p} + \widetilde{\boldsymbol{\eta}}_w(t) - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t))\boldsymbol{\eta}_c(t) \\ &= \boldsymbol{F}(\boldsymbol{q}(t)) - \widetilde{\boldsymbol{K}}_{\widetilde{\boldsymbol{\Gamma}}}(\boldsymbol{q},t) * \boldsymbol{p} + \widetilde{\boldsymbol{\eta}}(t).\end{aligned}$$

□

Example 1 (Quasi-Markovian GLE with constant coefficients). If we consider the case where $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are constant, i.e., $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}$ and $\widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$ with $\boldsymbol{\Gamma}, \boldsymbol{\Sigma} \in \mathbb{R}^{(n+m)\times(n+m)}$, one finds that the convolution term simplifies to

$$\widetilde{\boldsymbol{K}}_{\widetilde{\boldsymbol{\Gamma}}}(\boldsymbol{q},t) * \boldsymbol{p} = -\boldsymbol{\Gamma}_{1,1}\boldsymbol{M}^{-1}\boldsymbol{p}(t) + \int_0^t \boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}(t-s)}\boldsymbol{\Gamma}_{2,1}\boldsymbol{M}^{-1}\boldsymbol{p}(s)\mathrm{d}s,$$

and the noise terms become

$$\widetilde{\boldsymbol{\eta}}_w(t) = \boldsymbol{\Sigma}_1\,\dot{\boldsymbol{W}}(t),\ \widetilde{\boldsymbol{\eta}}_c(t) = -\boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}t}\boldsymbol{s}(0) - \boldsymbol{\Gamma}_{1,2}\int_0^t e^{-\boldsymbol{\Gamma}_{2,2}(t-s)}\boldsymbol{\Sigma}_2\mathrm{d}\,\boldsymbol{W}(s), \tag{29}$$

so that the stochastic IDE (26) resembles the form of the GLE (1) with

$$\boldsymbol{K}(t) = \delta(t)\boldsymbol{\Gamma}_{1,1} + \boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}(t-s)}\boldsymbol{\Gamma}_{2,1}. \tag{30}$$

Example 2 (Quasi-Markovian GLE with position dependent noise strength). If we consider the case where $\widetilde{\boldsymbol{\Gamma}}_{2,2}$ and $\widetilde{\boldsymbol{\Sigma}}_{2,2}$ are constant, i.e., $\widetilde{\boldsymbol{\Gamma}}_{2,2} \equiv \boldsymbol{\Gamma}_{2,2}$ and $\widetilde{\boldsymbol{\Sigma}}_{2,2} \equiv \boldsymbol{\Sigma}_{2,2}$ with $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{m\times m}$, the convolution term simplifies to

$$\widetilde{\boldsymbol{K}}_{\widetilde{\Gamma}}(\boldsymbol{q},t) * \boldsymbol{p} = \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t)) \int_0^t e^{-\boldsymbol{\Gamma}_{2,2}(t-s)} \widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}(s)) \boldsymbol{M}^{-1} \boldsymbol{p}(s) \mathrm{d}s, \tag{31}$$

and the random force terms $\widetilde{\boldsymbol{\eta}}_w$ and $\widetilde{\boldsymbol{\eta}}_c$ become

$$\widetilde{\boldsymbol{\eta}}_w(t) = \widetilde{\boldsymbol{\Sigma}}_1(\boldsymbol{q}(t))\, \dot{\boldsymbol{W}}(t), \tag{32}$$

and

$$\widetilde{\boldsymbol{\eta}}_c(t) = -\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t)) e^{-\boldsymbol{\Gamma}_{2,2}t} \boldsymbol{s}(0) - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}(t)) \int_0^t e^{-\boldsymbol{\Gamma}_{2,2}(t-s)} \widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}(s)) \mathrm{d}\, \boldsymbol{W}(s). \tag{33}$$

so that for $m = n$ and $\widetilde{\boldsymbol{\Gamma}}_{1,2} = -\widetilde{\boldsymbol{\Gamma}}_{2,1}^T$, $\widetilde{\boldsymbol{\Sigma}}_{1,2} = \widetilde{\boldsymbol{\Sigma}}_{2,1}^T \equiv \boldsymbol{0}$, the stochastic IDE (26) resembles the form of the GLE (3) with $\boldsymbol{K}(t) = e^{-\boldsymbol{\Gamma}_{2,2}t}$.

Remark 1 (Existence of solutions of (20)). A sufficient condition for (20) to possess a unique strong solution $\boldsymbol{x}(t)$ for all times $t \geq 0$, is that the right hand side of the SDE (20) is Lipschitz in $\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}$. Provided that the initial state μ_0 is as specified in (iv), it directly follows by standard existence and uniqueness results for SDEs (see e.g. [45, Theorem 5.2.1.]) that for any $T > 0$ there exists a unique strong solution $\boldsymbol{x}(t), t \in [0,T]$ of (20), which is continuous in t and

$$\mathbb{E}\left[\int_0^T \|\boldsymbol{x}(t)\|_2^2 \mathrm{d}t\right] < \infty.$$

Since $\boldsymbol{F}, \widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ are assumed to be smooth the Lipschitz condition is obviously satisfied for $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$. Similarly, for an unbounded configurational domain, i.e., $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$, the Lipschitz condition for the right hand side of (20) follows directly if the spectra of $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})$ and $\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q})$ are uniformly bounded in $\boldsymbol{q}$ and $\boldsymbol{F}$ satisfies certain asymptotic growths conditions (e.g., Assumption 3). We also note that the existence of suitable Lyapunov functions as derived in, e.g., Lemma 3 is sufficient (see e.g. [50]) to ensure the existence of a weak solution $(\boldsymbol{x}(t))_{t\geq 0}$ under less strict asymptotic growth conditions on the force $\boldsymbol{F}$.

2.1 Fluctuation-Dissipation Relation for Quasi-Markovian Generalized Langevin Equations

The following assumption can be understood as a fluctuation dissipation relation for the SDE (20):

Assumption 1. *There exists a symmetric positive definite matrix $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ such that for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$,*

$$\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q}) \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q} \end{pmatrix} + \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q} \end{pmatrix} \widetilde{\boldsymbol{\Gamma}}^T(\boldsymbol{q}) = \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q}) \widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q}). \tag{34}$$

As shown in Proposition 2, below, for a quasi-Markovian GLE with constant coefficients (see Example 1), Assumption 1 implies that the random force is stationary with covariance function $\boldsymbol{K}$ as specified in (30).

Proposition 2. *Let as in Example 1 $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ be constant, i.e., $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}$ and $\widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$ with $\boldsymbol{\Gamma}, \boldsymbol{\Sigma} \in \mathbb{R}^{(n+m)\times(n+m)}$. If Assumption 1 is satisfied and μ_0 such that $\boldsymbol{s}(0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$, where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ as specified in Assumption 1, then $\widetilde{\boldsymbol{\eta}}$ is a stationary Gaussian process with vanishing expectation and covariance function $\boldsymbol{K}$ as defined in (30).*

Proof. Let

$$\boldsymbol{G}(r) = \boldsymbol{\Gamma}_{1,2} \int_0^r e^{-\boldsymbol{\Gamma}_{2,2}(r-s)} \boldsymbol{\Sigma}_2 d\boldsymbol{W}(s).$$

Without loss of generality we assume that $t \geq t'$, and we find that the covariance of $\widetilde{\boldsymbol{\eta}}$ is indeed of the form (30):

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\boldsymbol{\eta}}(t)\widetilde{\boldsymbol{\eta}}^T(t')\right] &= \mathbb{E}\left[\boldsymbol{\Sigma}_1 \dot{\boldsymbol{W}}(t)\dot{\boldsymbol{W}}(t')^T\boldsymbol{\Sigma}_1^T\right] - \mathbb{E}[\boldsymbol{G}(t)\dot{\boldsymbol{W}}^T(t')\boldsymbol{\Sigma}_1^T] \\
&\quad + \mathbb{E}\left[\boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}t}\boldsymbol{s}(0)\boldsymbol{s}(0)^T e^{-\boldsymbol{\Gamma}_{2,2}^T t'}\boldsymbol{\Gamma}_{1,2}^T\right] \\
&\quad + \mathbb{E}\left[\left(\boldsymbol{\Gamma}_{1,2}\int_0^{t'} e^{-\boldsymbol{\Gamma}_{2,2}(t-s)}\boldsymbol{\Sigma}_2 d\boldsymbol{W}(s)\right)\boldsymbol{G}^T(t')\right] \\
&= \delta(t-t')(\boldsymbol{\Gamma}_{1,1}+\boldsymbol{\Gamma}_{1,1}^T) - \boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}(t-t')}(\boldsymbol{\Gamma}_{2,1}+\boldsymbol{Q}\boldsymbol{\Gamma}_{1,2}^T) \\
&\quad + \boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}t}\boldsymbol{Q}e^{-\boldsymbol{\Gamma}_{2,2}^T t'}\boldsymbol{\Gamma}_{1,2}^T \\
&\quad + \boldsymbol{\Gamma}_{1,2}\int_0^{t'} e^{-\boldsymbol{\Gamma}_{2,2}(t-s)}(\boldsymbol{\Gamma}_{2,2}\boldsymbol{Q}+\boldsymbol{Q}\boldsymbol{\Gamma}_{2,2}^T)e^{-\boldsymbol{\Gamma}_{2,2}^T(t'-s)}\boldsymbol{\Gamma}_{1,2}^T ds \\
&= \delta(t-t')(\boldsymbol{\Gamma}_{1,1}+\boldsymbol{\Gamma}_{1,1}^T) - \boldsymbol{\Gamma}_{1,2}e^{-\boldsymbol{\Gamma}_{2,2}(t-t')}\boldsymbol{\Gamma}_{2,1},
\end{aligned}$$

where expectations are taken over both μ_0 and the path measure of the Wiener process $\boldsymbol{W}$. The last equality follows by partial integration of the integral term. □

In the absence of a white-noise component in the random force, i.e., $\widetilde{\boldsymbol{\Gamma}}_{1,1}, \widetilde{\boldsymbol{\Sigma}}_{1,1} \equiv \boldsymbol{0}$, together with the requirement of $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})$ to be stable for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$, Assumption 1 imposes a constraint on the form $\widetilde{\boldsymbol{\Gamma}}_{1,2}$ and $\widetilde{\boldsymbol{\Gamma}}_{2,1}$ as shown in the following Proposition 3.

Proposition 3. *Let $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{Q}$ be such that the conditions of Proposition 4 are satisfied. $\widetilde{\boldsymbol{\Gamma}}_{1,1} \equiv \boldsymbol{0}$ implies*

$$\forall\, \boldsymbol{q} \in \Omega_{\boldsymbol{q}}:\ \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\boldsymbol{Q} = -\widetilde{\boldsymbol{\Gamma}}_{2,1}^T(\boldsymbol{q}). \tag{35}$$

Proof. Writing (34) in terms of the sub-blocks of $\widetilde{\boldsymbol{\Gamma}}$ we find

$$\begin{pmatrix} \boldsymbol{0} & \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\boldsymbol{Q} + \widetilde{\boldsymbol{\Gamma}}_{2,1}^T(\boldsymbol{q}) \\ \boldsymbol{Q}\widetilde{\boldsymbol{\Gamma}}_{1,2}^T(\boldsymbol{q}) + \widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}) & \widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q})\boldsymbol{Q} + \boldsymbol{Q}\widetilde{\boldsymbol{\Gamma}}_{2,2}^T(\boldsymbol{q}) \end{pmatrix} = \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q}). \tag{36}$$

By Lemma A.1 (iii) it follows that the left hand side of (36) is a positive semi-definite matrix for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$ if and only if (35) holds. □

Equilibrium Generalized Langevin Equation. In the particular case of a conservative force, i.e., $\boldsymbol{F} = -\nabla U$, one can easily derive a closed form solution for an invariant measure of the SDE (20) if Assumption 1 holds:

Proposition 4. *Let $\boldsymbol{F} = -\nabla U$, and let Assumption 1 hold. The SDE (20) conserves the probability measure $\mu_{\boldsymbol{Q},\beta}(d\boldsymbol{x})$ with density*

$$\rho_{\boldsymbol{Q},\beta}(\boldsymbol{x}) \propto e^{-\beta[U(\boldsymbol{q})+\frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}+\frac{1}{2}\boldsymbol{s}^T\boldsymbol{Q}^{-1}\boldsymbol{s}]}. \tag{37}$$

Proof. The statement follows by inspection of the stationary Fokker-Planck equation associated with the SDE (20). □

2.2 Non-equilibrium Quasi-Markovian Generalized Langevin Equations Without Fluctuation-Dissipation Relation

In general one might also consider instances of (20), where a fluctuation dissipation relation in the form of Assumption 1 does not hold. Such situations might appear in the modelling of temperature gradients or swarming/flocking phenomena; see, e.g., [54] for Markovian variants of such models. For example, one may consider an instance of (20), where $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are of the form

$$\widetilde{\boldsymbol{\Gamma}} = \begin{pmatrix} \hat{\boldsymbol{\Gamma}}_{1,1}^{(1)} & \hat{\boldsymbol{\Gamma}}_{1,2}^{(1)} & \hat{\boldsymbol{\Gamma}}_{1,2}^{(2)} \\ \hat{\boldsymbol{\Gamma}}_{2,1}^{(1)} & \hat{\boldsymbol{\Gamma}}_{2,2}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{\Gamma}}_{2,2}^{(2)} \end{pmatrix}, \quad \widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{1,1}^{(2)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{\Sigma}}_{2,2}^{(2)} \end{pmatrix}, \tag{38}$$

where

$$\hat{\boldsymbol{\Gamma}}_{1,1}^{(1)}, \hat{\boldsymbol{\Sigma}}_{1,1}^{(1)} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{n\times n}), \quad \hat{\boldsymbol{\Gamma}}_{1,2}^{(1)}, \left(\hat{\boldsymbol{\Gamma}}_{2,1}^{(1)}\right)^T, \hat{\boldsymbol{\Gamma}}_{1,2}^{(2)} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{n\times \hat{m}}),$$

and

$$\hat{\boldsymbol{\Gamma}}_{2,2}^{(1)}, \hat{\boldsymbol{\Gamma}}_{2,2}^{(2)}, \hat{\boldsymbol{\Sigma}}_{2,2}^{(2)} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{\hat{m}\times \hat{m}}),$$

with $\hat{m} \in \mathbb{N}$ such that $m = 2\hat{m}$ and $\hat{m} \geq n$. One can easily verify that in the view of the corresponding non-Markovian form (26), the coefficients $\hat{\boldsymbol{\Gamma}}_{i,j}^{(1)}, 1 \leq i,j \leq 2$ determine the statistical properties of the dissipation, i.e., the form of the convolution functional $\widetilde{\boldsymbol{K}}_{\widetilde{\boldsymbol{\Gamma}}}(\boldsymbol{q},t) * \boldsymbol{p}$, and the coefficients $\hat{\boldsymbol{\Gamma}}_{1,2}^{(2)}, \hat{\boldsymbol{\Gamma}}_{2,2}^{(2)}$ and $\hat{\boldsymbol{\Sigma}}_{1,1}^{(2)}, \hat{\boldsymbol{\Sigma}}_{2,2}^{(2)}$ determine the statistical properties of the random force $\widetilde{\boldsymbol{\eta}}$. As a simple example we mention the case where the coefficients $\hat{\boldsymbol{\Gamma}}_{i,j}^{(k)}$ and $\hat{\boldsymbol{\Sigma}}_{i,j}^{(k)}$ are constant, i.e.,

$$\hat{\boldsymbol{\Gamma}}_{1,1}^{(1)}, \hat{\boldsymbol{\Sigma}}_{1,1}^{(1)} \in \mathbb{R}^{n\times n}, \hat{\boldsymbol{\Gamma}}_{1,2}^{(1)}, \left(\hat{\boldsymbol{\Gamma}}_{2,1}^{(1)}\right)^T, \hat{\boldsymbol{\Gamma}}_{1,2}^{(2)} \in \mathbb{R}^{n\times m}, \hat{\boldsymbol{\Gamma}}_{2,2}^{(1)}, \hat{\boldsymbol{\Gamma}}_{2,2}^{(2)}, \hat{\boldsymbol{\Sigma}}_{2,2}^{(2)} \in \mathbb{R}^{m\times m},$$

with $\hat{\boldsymbol{\Sigma}}^{(2)}_{2,2} = \hat{\boldsymbol{\Gamma}}^{(2)}_{2,2} + \left(\hat{\boldsymbol{\Gamma}}^{(2)}_{2,2}\right)^T$. Under suitable conditions on these matrices (compare with the respective conditions stated in the preceding sections), it can then be easily shown that the SDE (20) can be rewritten as

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\boldsymbol{p}, \\ \dot{\boldsymbol{p}} &= \boldsymbol{F}(\boldsymbol{q}) - \int_0^t \boldsymbol{K}_1(t-s)\boldsymbol{p}(s)\mathrm{d}s + \widetilde{\boldsymbol{\eta}}, \end{aligned} \tag{39}$$

where

$$\boldsymbol{K}_1(t) = \delta(t)\hat{\boldsymbol{\Gamma}}^{(1)}_{1,1} - \hat{\boldsymbol{\Gamma}}^{(1)}_{1,2} e^{-t\hat{\boldsymbol{\Gamma}}^{(1)}_{2,2}} \hat{\boldsymbol{\Gamma}}^{(1)}_{2,1}, \tag{40}$$

and $\widetilde{\boldsymbol{\eta}}$ is a stationary Gaussian process with covariance function $\boldsymbol{K}_2$ of the form

$$\boldsymbol{K}_2(t) = 2\delta(t)\hat{\boldsymbol{\Sigma}}^{(2)}_{1,1} + \hat{\boldsymbol{\Gamma}}^{(2)}_{1,2} e^{-t\hat{\boldsymbol{\Gamma}}^{(2)}_{2,2}} \left(\hat{\boldsymbol{\Gamma}}^{(2)}_{1,2}\right)^T. \tag{41}$$

2.3 Markovian Representations of the GLE in the Literature

In the special case of $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ being constant (see Example 1), the Markovian representation (20) is of similar generality to that presented in [6,30] and the steps in the derivation are essentially the same (see also [47, Chapter 8]). Likewise, a derivation of a Markovian representation of the form (20) can for example be found in a slightly less general setup in [35]. We point out that besides the above mentioned generic frameworks, there are many Markovian representations of the GLE mentioned in the literature which are derived in the context of a particular physical model or application. For example, the Markovian representations of the GLE derived in [1,8,27,51] can be considered as special instances of the SDE (20) with constant coefficients $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$. Similarly, some of the non-equilibrium models studied in [10–12,49,51] can be represented in the form of (20) with constant coefficients $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$. Markovian representations of the GLE with position dependent memory kernels, which can be viewed as instances of the SDE (20) can be found in [25,34,43,44].

Sufficient Condition for the Existence of a Markovian Representation. Let $\boldsymbol{\eta}$ be a real-valued stationary Gaussian process with vanishing mean and covariance function $\boldsymbol{K} \in \mathcal{C}(\mathbb{R}, \mathbb{R})$, i.e.,

$$\forall s, t \in \mathbb{R},\ \mathbb{E}[\boldsymbol{\eta}(t)] = 0,\ \boldsymbol{K}(t) = \mathbb{E}[\boldsymbol{\eta}(s+t)\boldsymbol{\eta}(s)].$$

We denote by $\widehat{\mu}_{\boldsymbol{K}}$ the spectral measure of $\boldsymbol{K}$, i.e.,

$$\boldsymbol{K}(t) = \int_{\mathbb{R}} e^{ikt}\mathrm{d}\widehat{\mu}_{\boldsymbol{K}}(k).$$

Note that the existence of the spectral measure is a direct consequence of the following proposition, which is an adapted (and simplified) version of what is commonly referred to as Bochner's theorem.

Proposition 5. *A complex-valued function C with domain $\mathbb{R}$ is the covariance function of a continuous weakly stationary*[4] *random process on $\mathbb{R}^n$ with finite first and second moments, if and only if it can be represented as*

$$C(t) = \int_{\mathbb{R}} e^{itk} \mathrm{d}\mu(k),$$

where μ is a positive finite measure.

The above Proposition 5 is a simplified version of [55]. For a proof of the theorem we refer to any standard text book in Fourier analysis, such as [53, Chapter 1].

Assume that $\widehat{\mu}_{\boldsymbol{K}}$ possesses a density with respect to the Lebesgue measure, i.e.,

$$\widehat{\mu}_{\boldsymbol{K}}(\mathrm{d}k) = \widehat{\rho}_{\boldsymbol{K}}(k)\mathrm{d}k.$$

It has been observed in [49] (see also [12,51] for similar results), that $(\widehat{\rho}_{\boldsymbol{K}}(k))^{-1}$ being polynomial implies that $\boldsymbol{\eta}$ can be rewritten as a Markov process in an extended phase space. This can be seen as a consequence of the following criteria for Markovianity:

Proposition 6. *If $p(k) = \sum_{m=1} c_m(-ik)^m$ is a polynomial with real coefficients and roots in upper half plane then the Gaussian process with spectral density $|p(k)|^{-2}$ is the solution of the stochastic differential equation*

$$p\left(-i\frac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{\eta}(t)\mathrm{d}t = \mathrm{d}\,\boldsymbol{W}(t)$$

The above proposition is quoted from [51]. A simple and self-contained proof is also provided in this reference. For a more comprehensive discussion, we refer to [9].

As detailed in [51] the inverse density $(\widehat{\rho}_{\boldsymbol{K}}(k))^{-1}$ being a polynomial indeed implies the applicability of Proposition 6, as positivity of the measure $\widehat{\mu}_{\boldsymbol{K}}$ follows from Bochner's theorem. Therefore $\widehat{\rho}_{\boldsymbol{K}}$ must be a positive function, i.e., a positive polynomial of even degree, which in turn implies the existence of a suitable polynomial $p(k) = \sum_{m=1} c_m(-ik)^m$ with properties as stated in Proposition 6.

Proposition 6 has been used extensively to derive finite dimensional Markovian representations of the type of heat bath models used in [11,12,49,51]. Similarly, Proposition 6 can also be used to derive suitable distributions for the spring constants and the heat bath particle masses in the Ford-Kac model which ensure that in the thermodynamic limit the path of the distinguished particle converges weakly to the solution of a stochastic IDE which can be represented in a Markovian form; see [15,27,28].

[4] A stochastic process $(X(t))_{t\in\mathbb{R}}$ with associated covariance function C is said to be weakly stationary if $\mathbb{E}[X(t)] = \mathbb{E}[X(t+s)] = 0$ and $C(0,s) = C(t,t+s)$ for all $t, s \in \mathbb{R}$. Since Gaussian processes are fully characterized by the mean and covariance function, a Gaussian processes is weakly stationary if and only if it is stationary.

3 Ergodicity Properties

Let $e^{t\mathcal{L}_{\mathrm{GLE}}}$ denote the associated evolution operator of the process (20), i.e.,

$$\forall \varphi \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R}) : e^{t\mathcal{L}_{\mathrm{GLE}}}\varphi(x) = \mathbb{E}[\varphi(\boldsymbol{x}(t)) \mid \boldsymbol{x}(0) = x], \tag{42}$$

where the expectation is taken with respect to the Brownian motion $\boldsymbol{W}$. In this section we derive criteria for exponential convergence of $e^{t\mathcal{L}_{\mathrm{GLE}}}$ in some weighted L^∞ space as $t \to \infty$. More precisely, define for a prescribed $\mathcal{K} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, [1, \infty))$ with the property that $\mathcal{K}(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$ the set

$$L^\infty_{\mathcal{K}}(\Omega_{\boldsymbol{x}}) := \left\{\varphi \text{ measurable } : \|\varphi\|_{L^\infty_{\mathcal{K}}} < \infty\right\}, \tag{43}$$

where

$$\|\varphi\|_{L^\infty_{\mathcal{K}}} := \left\| \frac{\varphi}{\mathcal{K}} \right\|_\infty, \ \varphi : \Omega_{\boldsymbol{x}} \to \mathbb{R} \text{ measureable}, \tag{44}$$

so that $\left(L^\infty_{\mathcal{K}}(\Omega_{\boldsymbol{x}}), \|\cdot\|_{L^\infty_{\mathcal{K}}}\right)$ can be verified to define a Banach space. Furthermore, denote by

$$\mathbb{E}_\mu \varphi := \int \varphi(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x}), \tag{45}$$

the expectation of an observable φ with respect to the probability measure μ.

We show under certain conditions on the coefficients $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ and the force $\boldsymbol{F}$ that there exists a unique probability measure with smooth density $\mu(\mathrm{d}\boldsymbol{x}) = \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, such that

$$\exists\, \kappa > 0, C > 0, \ \forall\, \varphi \in L^\infty_{\mathcal{K}}, \ \left\|\mathbb{E}_\mu \varphi - e^{t\mathcal{L}_{\mathrm{GLE}}}\varphi\right\|_{L^\infty_{\mathcal{K}}} \leq Ce^{-\kappa t}\left\|\mathbb{E}_\mu \varphi - \varphi\right\|_{L^\infty_{\mathcal{K}}}, \tag{46}$$

for all $t \geq 0$, and

$$\int_{\Omega_x} \mathcal{K}(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x}) < \infty, \tag{47}$$

where $\mathcal{K}$ is a suitable Lyapunov function whose exact properties we specify below. In particular, if $\boldsymbol{F} = -\nabla U$ and Assumption 1 holds, then

$$\mu(\mathrm{d}\boldsymbol{x}) = \mu_{\boldsymbol{Q},\beta}(\mathrm{d}\boldsymbol{x}),$$

where $\mu_{\boldsymbol{Q},\beta}$ is as defined in Proposition 4. If the process (20) satisfies (46) for all $t \geq 0$, we say in the sequel that it is *geometrically ergodic*.

All results are derived using standard Lyapunov techniques (see e.g. [36, 38, 39, 50]), which we summarize in Appendix B. That is, we show that (i) the minorization condition (Assumption B.2) is satisfied and (ii) a suitable Lyapunov function exists which satisfies Assumption B.1 (or more generally the existence of a suitable class of Lyapunov functions of which each instance satisfies Assumption B.1). We treat the cases $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$ and $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$ separately. In the situation $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$, we show geometric ergodicity for the case of constant coefficients, i.e., $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}$, and $\widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$, which in the non-Markovian form (26) corresponds to the situation of a stationary random force. For the case of a bounded domain $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$ we can show geometric ergodicity also for the case where $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are not constant in $\boldsymbol{q}$, i.e., the random force, $\widetilde{\boldsymbol{\eta}}$, in the corresponding non-Markovian form (26) is non-stationary. In order to simplify presentation we assume for the remainder of this article $\boldsymbol{M} = \boldsymbol{I}_n$.

3.1 Summary of Main Results

Let in the sequel $g(x) = \Theta(f(x))$ indicate that the function f is bounded both above and below by g asymptotically as $\|x\| \to \infty$, i.e., there exist $c_1, c_2 > 0$ and $\tilde{x} \geq 0$, such that $c_1 g(x) \leq f(x) \leq c_2 g(x)$ for all $\|x\| \geq \tilde{x}$.

Results for Stationary Noise. We first present results for the constant coefficient case, i.e., $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}$, and $\widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$. Let for the remainder of this subsection $\boldsymbol{\Gamma}, \boldsymbol{\Sigma}$ be such that

(i) $-\boldsymbol{\Gamma}$ is a stable matrix, i.e., the real parts of all eigenvalues of $\boldsymbol{\Gamma}$ are positive.
(ii) the SDE (20) satisfies the parabolic Hörmander condition both in the presence of the force term $\boldsymbol{F}$ as well as in absence of a force term, i.e., $\boldsymbol{F} \equiv 0$. We provide algebraic conditions on $\boldsymbol{\Gamma}, \boldsymbol{\Sigma}$ which imply the parabolic Hörmander condition in Sect. 3.2.
(iii) Assumption 1 is satisfied so that for $\boldsymbol{F} = -\nabla U$ the measure $\mu_{\boldsymbol{Q},\beta}(\mathrm{d}\boldsymbol{x}) = \rho_{\boldsymbol{Q},\beta}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ with $\rho_{\boldsymbol{Q},\beta}$ as defined in (37) is an invariant measure of (20).

Theorem 1. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$, and $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ as specified above. There is a unique invariant measure μ such that for any $l \in \mathbb{N}$ there exists $\mathcal{K}_l \in \mathcal{C}^\infty(\mathbb{T}^n \times \mathbb{R}^{n+m})$ with*

$$\mathcal{K}_l(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}) = \Theta(\|\boldsymbol{z}\|^{2l}), \quad \text{as } \|\boldsymbol{z}\| \to \infty, \ \boldsymbol{z} = \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{s} \end{pmatrix},$$

so that (46) and (47) hold for $\mathcal{K} = \mathcal{K}_l$. In particular, if $\boldsymbol{F} = -\nabla U$, then $\mu = \mu_{\boldsymbol{Q},\beta}$.

Proof. The validity of the minorization condition follows from Lemma 2. The existence of a suitable class of Lyapunov functions is shown in Lemma 1. □

In the case of an unbounded configurational domain, i.e., $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$, we require an additional assumption on the force $\boldsymbol{F}$ in order to construct a suitable class of Lyapunov functions.

Assumption 2. *There exists a potential function $V \in \mathcal{C}^2(\Omega_{\boldsymbol{q}}, \mathbb{R})$ with the following properties*

(i) there exists $G \in \mathbb{R}$ such that

$$\langle \boldsymbol{q}, \boldsymbol{F}(\boldsymbol{q}) \rangle \leq -\langle \boldsymbol{q}, \nabla_{\boldsymbol{q}} V(\boldsymbol{q}) \rangle + G.$$

for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$.
(ii) the potential function is bounded from below, i.e., there exists $u_{\min} > -\infty$ such that

$$\forall \boldsymbol{q} \in \Omega_{\boldsymbol{q}}, \ V(\boldsymbol{q}) \geq u_{\min}.$$

(iii) there exist constants $D, E > 0$ and $F \in \mathbb{R}$ such that

$$\forall \boldsymbol{q} \in \Omega_{\boldsymbol{q}}, \ \langle \boldsymbol{q}, \nabla_{\boldsymbol{q}} V(\boldsymbol{q}) \rangle \geq DV(\boldsymbol{q}) + E\|\boldsymbol{q}\|_2^2 + F. \tag{48}$$

Theorem 2. *Let $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$, $\boldsymbol{F}$ satisfies Assumption 2, $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ as specified above with* $\operatorname{rank}(\boldsymbol{\Sigma}) = n+m$ *and* $\operatorname{rank}(\boldsymbol{\Gamma}_{1,1}) = n$. *There is a unique invariant measure μ such that for any $l \in \mathbb{N}$ there exists $\mathcal{K}_l \in \mathcal{C}^\infty(\mathbb{R}^{2n+m}, [1,\infty))$ with*

$$\mathcal{K}_l(\boldsymbol{x}) = \Theta(\|\boldsymbol{x}\|^{2l}), \quad \text{as } \|\boldsymbol{x}\| \to \infty,$$

such that (46) and (47) hold for $\mathcal{K} = \mathcal{K}_l$. In particular, if $\boldsymbol{F} = -\nabla U$, then $\mu = \mu_{\boldsymbol{Q},\beta}$.

Proof. The validity of a minorization condition follows from Lemma 4. The existence of a suitable class of Lyapunov functions is shown in Lemma 3. □

The above theorem covers instances of the GLE with a non-degenerated white noise component. In order to derive geometric ergodicity for GLEs without a white noise component, i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{0}$ which is implied by $\boldsymbol{\Gamma}_{1,1} = \boldsymbol{0}$ (see Proposition 3), we require the force $\boldsymbol{F}$ to satisfy the following assumption:

Assumption 3. *Let the force $\boldsymbol{F}$ be such that*

$$\boldsymbol{F}(\boldsymbol{q}) = \boldsymbol{F}_1(\boldsymbol{q}) + \boldsymbol{F}_2(\boldsymbol{q}),$$

where $\boldsymbol{F}_1 \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R}^n)$ is uniformly bounded in $\Omega_{\boldsymbol{q}}$, i.e.,

$$\sup_{\boldsymbol{q} \in \Omega_{\boldsymbol{q}}} \|\boldsymbol{F}_1(\boldsymbol{q})\|_\infty < \infty,$$

and

$$\boldsymbol{F}_2(\boldsymbol{q}) = -\boldsymbol{H}\,\boldsymbol{q},$$

with $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ being a positive definite matrix, i.e., $\min \sigma(\boldsymbol{H}) = \lambda_{\boldsymbol{H}} > 0$.

Remark 2. Assumption 3 implies that there is $\overline{H} > 0$ and $\overline{h} \in \mathbb{R}$ so that

$$|\langle \boldsymbol{g}, \boldsymbol{F}(\boldsymbol{q})\rangle| \le \overline{H}|\langle \boldsymbol{g}, \boldsymbol{q}\rangle| + \overline{h},$$

for all $\boldsymbol{q}, \boldsymbol{g} \in \mathbb{R}^n$. Moreover, Assumption 3 implies Assumption 2 with the potential function V in Assumption 2 being of the form of a quadratic potential function, i.e.,

$$V(\boldsymbol{q}) = \frac{1}{2}\boldsymbol{q}^T\left(\boldsymbol{H} - \varepsilon \boldsymbol{I}_n\right)\boldsymbol{q},$$

with sufficiently small $\varepsilon > 0$.

The following theorem provides a sufficient condition for geometric ergodicity of (20) for constant coefficients and $\boldsymbol{\Gamma}_{1,1} = \boldsymbol{0}$.

Theorem 3. *Let $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$, $\boldsymbol{F}$ satisfy Assumption 3, and $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ be as specified above with $\boldsymbol{\Gamma}_{1,1} = \boldsymbol{0}$. There exists a unique probability measure $\mu(\mathrm{d}\boldsymbol{x})$ such that for any $l \in \mathbb{N}$ there exists $\mathcal{K}_l \in \mathcal{C}^\infty(\mathbb{R}^{2n+m}, [0,\infty))$ with*

$$\mathcal{K}_l(\boldsymbol{x}) = \Theta(\|\boldsymbol{x}\|^{2l}), \quad \text{as } \|\boldsymbol{x}\| \to \infty,$$

such that (46) and (47) hold for $\mathcal{K} = \mathcal{K}_l$. In particular, if $\boldsymbol{F} = -\nabla U$, then $\mu = \mu_{\boldsymbol{Q},\beta}$.

Proof. The validity of the minorization condition follows from Lemma 5. The existence of a suitable class of Lyapunov functions is shown in Lemma 3. □

Results for Non-stationary Noise. For the case of a periodic configurational domain $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$ we show geometric ergodicity for the SDE (20) for the general case where $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ may not be constant. We focus on the case

$$\widetilde{\boldsymbol{\Gamma}}(\cdot) = \begin{pmatrix} \boldsymbol{0} & \widetilde{\boldsymbol{\Gamma}}_{1,2}(\cdot) \\ \widetilde{\boldsymbol{\Gamma}}_{2,1}(\cdot) & \widetilde{\boldsymbol{\Gamma}}_{2,2}(\cdot) \end{pmatrix} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathbb{R}^{2n\times 2n}).$$

where all non-vanishing sub-blocks are assumed to be invertible, i.e.,

$$\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}), \widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}), \widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q}), \widetilde{\boldsymbol{\Sigma}}_{2,2}(\boldsymbol{q}) \in \mathrm{GL}_n(\mathbb{R}),$$

for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$, where by $\mathrm{GL}_n(\mathbb{R}) \subset \mathbb{R}^{n\times n}$ we denote the set of all invertible $n \times n$-matrices with real valued coefficients. Furthermore, we assume that $-\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})$ is a stable matrix for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$ and that $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ are such that Assumption 1 is satisfied, i.e., since $\widetilde{\boldsymbol{\Gamma}}_{1,1} \equiv \boldsymbol{0}$, it follows by Proposition 3 that

$$\forall \boldsymbol{q} \in \Omega_{\boldsymbol{q}}, \ \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}) = -\boldsymbol{Q}\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}), \tag{49}$$

holds. Moreover we assume

$$\exists\, \boldsymbol{C} \in \mathbb{R}^{(n+m)\times(n+m)} \text{ s.p.d.}, \ \forall\, \boldsymbol{q} \in \Omega_{\boldsymbol{q}} : \ \widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})\boldsymbol{C} + \boldsymbol{C}\widetilde{\boldsymbol{\Gamma}}^T(\boldsymbol{q}) \text{ s.p.d.}, \tag{50}$$

where the notation "s.p.d." stands for "symmetric positive definite". We expect that our result can be easily extended to more general forms of $\widetilde{\boldsymbol{\Gamma}}$, i.e., to the case where $\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q}) \in \mathbb{R}^{m\times m}$ with $m \neq n$; see note N.5 at the end of this subsection. We also point out that the case $\widetilde{\boldsymbol{\Gamma}}_{1,1} \neq \boldsymbol{0}$ would not cause any additional difficulties in the proof of the result as long as the identity (49) holds. (See e.g. [54] for ergodicity results for under-damped Langevin equation with non-constant coefficients.)

Theorem 4. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$. Under the assumptions on $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ described in the preceding paragraph, there is a unique invariant measure μ such that there exists for any $l \in \mathbb{N}$ a function $\mathcal{K}_l \in \mathcal{C}^\infty(\mathbb{T}^n \times \mathbb{R}^{2n}, [1,\infty))$ with*

$$\mathcal{K}_l(\boldsymbol{q},\boldsymbol{p},\boldsymbol{s}) = \Theta(\|\boldsymbol{z}\|^{2l}), \quad \text{as } \|\boldsymbol{z}\| \to \infty, \ \boldsymbol{z} = \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{s} \end{pmatrix},$$

such that (46) and (47) hold for $\mathcal{K} = \mathcal{K}_l$. In particular, if $\boldsymbol{F} = -\nabla U$, then $\mu = \mu_{\boldsymbol{Q},\beta}$.

Proof. The validity of the minorization condition follows from Lemma 7. The existence of a suitable class of Lyapunov functions is shown in Lemma 10. □

We provide a simple example of an instance of (20), which satisfies the condition of Theorem 4:

Example 3. Let $m = n = 1$ and let $\Omega_{\boldsymbol{q}} = \mathbb{T}$. Consider the matrix-valued functions $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ defined by

$$\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q}) = \begin{pmatrix} 0 & -(2+\cos(2\pi\boldsymbol{q})) \\ (2+\cos(2\pi\boldsymbol{q})) & 1 \end{pmatrix}, \quad \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q}) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Obviously, a valid choice for $\boldsymbol{Q}$ in Proposition 4 is

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Moreover,

$$\boldsymbol{C} = \begin{pmatrix} 19/18 & -(1/6) \\ -(1/6) & 1 \end{pmatrix}.$$

satisfies (50). This follows by virtue of Lemma A.1. We provide a plot of the eigenvalues of the matrix

$$\boldsymbol{R}(\boldsymbol{q}) = \widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})\boldsymbol{C} + \boldsymbol{C}\widetilde{\boldsymbol{\Gamma}}^{T}(\boldsymbol{q}), \tag{51}$$

as a function of $\boldsymbol{q}$ in Fig. 1.

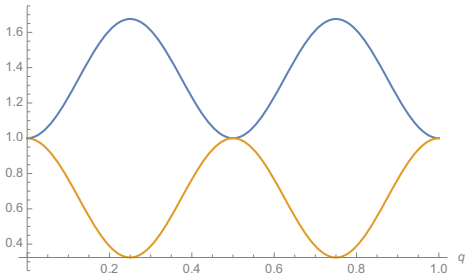


Fig. 1. $\boldsymbol{q}$ vs. the eigenvalues of the matrix $\boldsymbol{R}(\boldsymbol{q})$ which is defined in (51).

Central Limit Theorem for Quasi-Markovian GLE Dynamics. A direct consequence of the geometric ergodicity of the dynamics (20) is the validity of a central limit theorem for certain observables. This result is of practical importance as it justifies the use of GLE dynamics for sampling purposes as, e.g., in [5,6,60].

Define the projection operator

$$\Pi\varphi = \varphi - \mathbb{E}_{\mu}\varphi,$$

and let $L^\infty_{\mathcal{K},0} := \Pi L^\infty_{\mathcal{K}} \subset L^\infty_{\mathcal{K}}$, be the subspace of $L^\infty_{\mathcal{K}}$ which is comprised of observables with vanishing mean. Denote by $\|\cdot\|_{\mathcal{B}(L^\infty_{\mathcal{K}})}$ the operator norm

$$\|A\|_{\mathcal{B}(L^\infty_{\mathcal{K}})} := \sup_{\varphi \in L^\infty_{\mathcal{K}}} \frac{\|A\varphi\|_{L^\infty_{\mathcal{K}}}}{\|\varphi\|_{L^\infty_{\mathcal{K}}}}.$$

induced by the norm $\|\cdot\|_{L^\infty_{\mathcal{K}}}$ for operators $A : L^\infty_{\mathcal{K}} \to L^\infty_{\mathcal{K}}$. The validity of (46) for all $t \geq 0$ immediately implies the inequality

$$\|\Pi e^{t\mathcal{L}_{\mathrm{GLE}}}\|_{\mathcal{B}(L^\infty_{\mathcal{K}})} \leq Ce^{t\kappa}. \tag{52}$$

By [32, Proposition 2.1], $\mathcal{L}_{\mathrm{GLE}}$ considered as an operator on $L^\infty_{\mathcal{K},0}$ is invertible with bounded spectrum. By [2] this implies a central limit theorem for observables contained in $\varphi \in L^\infty_{\mathcal{K}}$ as summarized in the following Corollary 1.

Corollary 1. *Let the conditions of one of the Theorems 1 to 4 be satisfied and let $\mathcal{K}_l$ for $l \in \mathbb{N}$ be a suitable Lyapunov function as specified therein. The spectrum of $\mathcal{L}^{-1}_{\mathrm{GLE}}\Pi$ is bounded in $\|\cdot\|_{\mathcal{B}(L^\infty_{\mathcal{K}_l})}$, i.e.,*

$$\|\mathcal{L}^{-1}_{\mathrm{GLE}}\Pi\|_{\mathcal{B}(L^\infty_{\mathcal{K}_l})} \leq \frac{C_l}{\kappa_l}, \tag{53}$$

where $C_l, \kappa_l > 0$ are such that (46) holds for $\mathcal{K} = \mathcal{K}_l, \kappa = \kappa_l, C = C_l$. In particular, a central limit theorem holds for the solution of (20), i.e.,

$$T^{-1/2} \int_0^T [\mathbb{E}_\mu \varphi - \varphi(\boldsymbol{x}(t))]\, \mathrm{d}t \sim \mathcal{N}(0, \sigma^2_\varphi), \quad \text{as } T \to \infty, \tag{54}$$

for any $\varphi \in L^\infty_{\mathcal{K}_l}$, where μ denotes the unique invariant measure of $\boldsymbol{x}$ and

$$\sigma^2_\varphi = -2 \int \left(\mathcal{L}^{-1}_{\mathrm{GLE}} \Pi\varphi(\boldsymbol{x})\right) \Pi\varphi(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x}).$$

Notes on Theorems 1 to 4:

N.1. Theorems 1 to 4 imply path-wise ergodicity in the sense that

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \varphi(\boldsymbol{x}(t))\mathrm{d}t = \mathbb{E}_\mu \varphi, \tag{55}$$

almost surely for μ-almost all initializations of $\boldsymbol{x}(0)$ an almost all realizations of the Wiener process $\boldsymbol{W}$. We note that in the case that $\boldsymbol{F} = -\nabla U$ and Assumption 1 is satisfied it is sufficient to show that the generator $\mathcal{L}_{\mathrm{GLE}}$ is hypoelliptic in order to conclude uniqueness of the invariant measure and path-wise ergodicity in the above sense. This follows directly from the arguments in [26] as in this case the form of the invariant measure is known and has a smooth positive density.

N.2. The Lyapunov-based techniques on which the proofs of our ergodicity results rely have been studied in the context of stochastic differential equations (see [36,38,50,58]) as well as in the context of discrete time Markov chains (see e.g. [16,18,37,39]). In particular, we mention the application of these techniques to prove geometric ergodicity of solutions of the under-damped Langevin equation in [36,50,58]. As discussed in Sect. 2, the structure of the SDE (20) resembles the structure of the under-damped Langevin equation and it is therefore not surprising that also the structure of the Lyapunov functions constructed in the proofs of [36] resemble the structure of the Lyapunov functions presented in the latter two references.

N.3. In [46] the authors construct a Lyapunov function for a Markovian reformulation of the GLE with conservative force which in the representation (20) corresponds to the case where $\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Sigma}}$ are constant with $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}_{1,1} = \mathbf{0}$ and $\boldsymbol{\Gamma}_{1,2}, \boldsymbol{\Gamma}_{2,1}, \boldsymbol{\Gamma}_{2,2} \in \mathbb{R}^{n\times n}$ are diagonal matrices. In the same article exponential convergence of the law to a unique invariant distribution μ in relative entropy is shown and exponential decay estimates for the semigroup operator $e^{t\mathcal{L}_{\mathrm{GLE}}}$ in weighted Sobolev space $H^1(\mu)$ are derived using the hypocoercivity framework by Villani (see [59]).

N.4. Ergodic properties of non-equilibrium systems which have a similar structure as the QGLE models considered here have been studied in a series of papers [10–12,49,52]. These systems consist of a chain of a finite number of oscillators whose ends are coupled to two different heat baths. In a simplified version these systems can be written in the form

$$\begin{aligned}
\dot{\boldsymbol{r}}_1 &= -\gamma_1 \boldsymbol{r}_1 + \lambda_1 \boldsymbol{p}_1 + \sqrt{2\beta^{-1}\gamma_1}\dot{W}_1, \\
\dot{\boldsymbol{q}}_1 &= \boldsymbol{p}_1, \\
\dot{\boldsymbol{p}}_1 &= -\partial_{\boldsymbol{q}_1} U(\boldsymbol{q}) - \lambda_1 \boldsymbol{r}_1, \\
\dot{\boldsymbol{q}}_i &= \boldsymbol{p}_i, && i = 2,3,\dots,n-1, \\
\dot{\boldsymbol{p}}_i &= -\partial_{\boldsymbol{q}_i} U(\boldsymbol{q}), && i = 2,3,\dots,n-1, \\
\dot{\boldsymbol{q}}_n &= \boldsymbol{p}_n, \\
\dot{\boldsymbol{p}}_n &= -\partial_{\boldsymbol{q}_n} U(\boldsymbol{q}) - \lambda_2 \boldsymbol{r}_2, \\
\dot{\boldsymbol{r}}_2 &= -\gamma_2 \boldsymbol{r}_2 + \lambda_2 \boldsymbol{p}_n + \sqrt{2\beta^{-1}\gamma_2}\dot{W}_2,
\end{aligned} \tag{56}$$

where

$$U(\boldsymbol{q}) = U_1(\boldsymbol{q}_1) + U_n(\boldsymbol{q}_n) + \sum_{i=2}^{n} \tilde{U}(\boldsymbol{q}_i - \boldsymbol{q}_{i-1}),$$

with $U_1, U_2, \tilde{U} \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$, $\gamma_i > 0, \lambda_i > 0$ for $i = 1,2$, and W_1, W_2 are two independent Wiener processes taking values in $\mathbb{R}$. Under certain conditions on the potential functions U_1, U_n and $\tilde{U}$, the existence of an invariant measure (stationary non-equilibrium state) has been shown in [12]. Uniqueness conditions were derived in [10,11], and exponential convergence to the invariant state was shown in [52] (see also the review paper [51] and [3]). In the latter reference slightly more general heat bath models are considered than

above in (56)). Exponential convergence towards a unique invariant measure is proven in [52] by showing the existence of a suitable Lyapunov function and by showing hypoellipticity and controllability in the sense of Assumption B.4. The construction of a suitable control in the proof provided therein relies on $\tilde{U}$ being strictly convex. We expect that the techniques which are used in [52] to prove the existence of a suitable Lyapunov function and the controllability of the SDE can be extended/modified to prove geometric ergodicity for a wide range of GLEs which can be represented in the form (20) with constant coefficients. In fact it has been demonstrated in [51] that controllability in the sense of Assumption B.4 of a system consisting of a chain of oscillators which are coupled to a single heat bath, can be proven by the same techniques as used in [52].

N.5. We expect that Theorem 4 can be generalized to cover instances of (20), where $\widetilde{\boldsymbol{\Gamma}}$ is of a form such that in the non-Markovian reformulation (26) the memory kernel is of the form

$$\boldsymbol{K}_{\widetilde{\boldsymbol{\Gamma}}}(\boldsymbol{q},t) = \widetilde{\boldsymbol{\Gamma}}_{1,1}(\boldsymbol{q})\delta(t) - \sum_{i=1}^{K} \widetilde{\boldsymbol{\Gamma}}_{1,2}^{(i)}(\boldsymbol{q}) e^{-t\boldsymbol{\Gamma}_{2,2}^{(i)}} \widetilde{\boldsymbol{\Gamma}}_{2,1}^{(i)}(\boldsymbol{q}),\ K \in \mathbb{N},$$

where each $\widetilde{\boldsymbol{\Gamma}}^{(i)}$,

$$\widetilde{\boldsymbol{\Gamma}}^{(i)}(\boldsymbol{q}) = \begin{pmatrix} \mathbf{0} & \widetilde{\boldsymbol{\Gamma}}_{1,2}^{(i)}(\boldsymbol{q}) \\ \widetilde{\boldsymbol{\Gamma}}_{2,1}^{(i)}(\boldsymbol{q}) & \boldsymbol{\Gamma}_{2,2}^{(i)} \end{pmatrix}$$

satisfies the same conditions as $\widetilde{\boldsymbol{\Gamma}}$ in Theorem 4.

3.2 Conditions for Hypoellipticity

Consider the case of constant coefficients in (20), i.e., $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}, \widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$. In this subsection we provide criteria in the form of algebraic conditions on $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ which ensure that (20) satisfies the parabolic Hörmander condition, which by Proposition B.2, implies that the differential operators

$$\mathcal{L}_{\mathrm{GLE}},\ \mathcal{L}_{\mathrm{GLE}}^{\dagger},\ \partial_t - \mathcal{L}_{\mathrm{GLE}},\ \partial_t - \mathcal{L}_{\mathrm{GLE}}^{\dagger},$$

are hypoelliptic. Let in the following Proposition 7 $\Sigma_i, 1 \leq i \leq n+m$ denote the column vectors of $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{\Sigma} = [\Sigma_1, \ldots \Sigma_{m+n}] \in \mathbb{R}^{(n+m)\times(n+m)}.$$

Proposition 7. *Let $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma} \in \mathbb{R}^{(n+m)\times(n+m)}$ such that $-\boldsymbol{\Gamma}$ is stable and $\widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma} \in \mathbb{R}^{(n+m)\times(n+m)}$. Any of the following conditions is sufficient for (20) to satisfy the parabolic Hörmander condition.*

(i) $\boldsymbol{F} = \boldsymbol{H}\boldsymbol{q} + \boldsymbol{h}$, *where* $\boldsymbol{H} \in \mathbb{R}^{n\times n}, \boldsymbol{h} \in \mathbb{R}$, *and for all* $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$

$$\mathbb{R}^{2n+m} = \operatorname{lin}\left(\left\{\boldsymbol{S}^k \begin{pmatrix} \boldsymbol{0} \\ \Sigma_i \end{pmatrix} : \ k \in \mathbb{N},\ 1 \le i \le n+m \right\}\right), \tag{57}$$

where

$$\boldsymbol{S} := -\begin{pmatrix} \boldsymbol{0} & -\boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{H} & \boldsymbol{\Gamma}_{1,1} & \boldsymbol{\Gamma}_{1,2} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{2,1} & \boldsymbol{\Gamma}_{2,2} \end{pmatrix} \in \mathbb{R}^{(2n+m)\times(2n+m)}.$$

(ii)

$$\mathbb{R}^{n+m} = \operatorname{lin}\left(\bigcup_{1\le i\le n+m} \left\{\boldsymbol{\Gamma}^k \Sigma_i : \ k \le k_i\right\}\right), \tag{58}$$

where $k_i,\ 1 \le i \le n+m$ *are defined as*

$$k_i := \underset{k\in\mathbb{N}}{arg\ max}\, \boldsymbol{S}_0^k \begin{pmatrix} \boldsymbol{0} \\ \Sigma_i \end{pmatrix} \in \{\boldsymbol{0}\} \times \mathbb{R}^{n+m}, \tag{59}$$

with

$$\boldsymbol{S}_0 := -\begin{pmatrix} \boldsymbol{0} & -\boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{1,1} & \boldsymbol{\Gamma}_{1,2} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{2,1} & \boldsymbol{\Gamma}_{2,2} \end{pmatrix} \in \mathbb{R}^{(2n+m)\times(2n+m)}.$$

(iii) $\operatorname{rank}(\boldsymbol{\Sigma}_2) = m$, *and* $\operatorname{rank}(\boldsymbol{\Gamma}_{1,2}) = n$.

Proof. In relation to Proposition B.2 the coefficients $\boldsymbol{b}_i$ are

$$\boldsymbol{b}_0(\boldsymbol{x}) = -\boldsymbol{G}\begin{pmatrix} -\boldsymbol{F}(\boldsymbol{q}) \\ \boldsymbol{z} \end{pmatrix},$$

and

$$\boldsymbol{b}_i = \beta^{-\frac{1}{2}} \begin{pmatrix} \boldsymbol{0} \\ \Sigma_i \end{pmatrix} \in \mathbb{R}^{2n+m}, 1 \le i \le n+m,$$

with $\boldsymbol{G} \in \mathbb{R}^{(2n+m)\times(2n+m)}$ as defined in (70). Since for $i > 0$ the coefficients $\boldsymbol{b}_i$ are constant in $\boldsymbol{x}$, we find $[\boldsymbol{b}_i, \boldsymbol{b}_j] = \boldsymbol{0}$ and $[\boldsymbol{b}_0, \boldsymbol{b}_i] = -\nabla_{\boldsymbol{x}}\boldsymbol{b}_0\, \boldsymbol{b}_i$ for $i, j > 0$, where $\nabla_{\boldsymbol{x}}\boldsymbol{b}_0$ denotes the Jacobian matrix of $\boldsymbol{b}_0$, i.e.,

$$\nabla_{\boldsymbol{x}}\boldsymbol{b}_0 = -\begin{pmatrix} \boldsymbol{0} & -\boldsymbol{I}_n & \boldsymbol{0} \\ -\nabla\boldsymbol{F}(\boldsymbol{q}) & \boldsymbol{\Gamma}_{1,1} & \boldsymbol{\Gamma}_{1,2} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{2,1} & \boldsymbol{\Gamma}_{2,2} \end{pmatrix},$$

and $\nabla\boldsymbol{F}(\boldsymbol{q})$ denotes the Jacobian of the force $\boldsymbol{F}$. Therefore,

$$\mathscr{V}_1 = \{-\nabla_{\boldsymbol{x}}\boldsymbol{b}_0\, \boldsymbol{v} : \boldsymbol{v} \in \mathscr{V}_0\} \cup \mathscr{V}_0, \tag{60}$$

where

$$\left\{\begin{pmatrix} \boldsymbol{0} \\ \Sigma_i \end{pmatrix}\right\}_{i=1}^{n+m},$$

– In the case of (i) it follows that $\nabla_{\boldsymbol{x}} \boldsymbol{b}_0(\boldsymbol{x}) = \boldsymbol{S}$. In particular, since $\nabla_{\boldsymbol{x}} \boldsymbol{b}_0$ is constant in $\boldsymbol{x}$, (60) generalizes to

$$\mathscr{V}_{i+1} = \{\boldsymbol{S}\,\boldsymbol{v} : \boldsymbol{v} \in \mathscr{V}_i\} \cup \mathscr{V}_i,\ i \in \mathbb{N}. \tag{61}$$

Since $\mathscr{V}_i$ consists only of constant functions, we have $\mathrm{lin}(\mathscr{V}_i(\boldsymbol{x})) \equiv \mathrm{lin}(\mathscr{V}_i)$ for all $\boldsymbol{x} \in \Omega_{\boldsymbol{x}}, i \in \mathbb{N}$, thus (61) implies that (57) is a sufficient condition for the SDE (20) to satisfy the parabolic Hörmander condition.

– Regarding (ii): Let $k_{\max} = \max_{1 \le i \le n+m} k_i$. k_i being as defined in (59) together with (58) ensures that there is $\widetilde{\mathscr{V}} \subset \mathscr{V}_{k_{\max}}$ such that all elements in $\widetilde{\mathscr{V}}$ are constant and

$$\mathrm{lin}\left(\widetilde{\mathscr{V}}\right) \equiv \begin{pmatrix} \boldsymbol{0} \\ \mathbb{R}^{n+m} \end{pmatrix}.$$

Therefore,

$$\mathscr{V}_{k_{\max}+1} \supset \{-\nabla \boldsymbol{b}_0\,\boldsymbol{v} : \boldsymbol{v} \in \widetilde{\mathscr{V}}\} \cup \widetilde{\mathscr{V}}, \tag{62}$$

thus for all $\boldsymbol{x} \in \Omega_{\boldsymbol{x}}$

$$\mathrm{lin}\ (\mathscr{V}_{k_{\max}+1}(\boldsymbol{x})\ ) = \mathrm{lin}\left(\{-\nabla \boldsymbol{b}_0(\boldsymbol{x})\,\boldsymbol{v}(\boldsymbol{x}) : \boldsymbol{v} \in \widetilde{\mathscr{V}}\} \cup \widetilde{\mathscr{V}}(\boldsymbol{x})\right) = \mathbb{R}^{2n+m},$$

where the latter equivalence is due to the fact that

$$\mathrm{lin}\,(\{-\nabla \boldsymbol{b}_0(\boldsymbol{x})\boldsymbol{v}\ : \boldsymbol{v} \in B\} \cup B) = \mathbb{R}^{2n+m},$$

for all $\boldsymbol{x} \in \Omega_{\boldsymbol{x}}$ and any basis $B \subset \mathbb{R}^{2n+m}$ of $\{\boldsymbol{0}\} \times \mathbb{R}^{n+m}$.

– Regarding (iii): Since $\mathrm{lin}(\boldsymbol{\Sigma}_2) = \mathbb{R}^m$ and $\mathrm{rank}\,(\boldsymbol{\Gamma}_{1,2}) = n$ it follows that

$$\{\boldsymbol{0}\} \times \mathbb{R}^{n+m} = \mathrm{lin}\left(\left\{\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\Gamma}\Sigma_i \end{pmatrix}, 1 \le i \le n+m\right\}\right),$$

thus the result follows by (ii).

□

3.3 Technical Lemmas Required in the Proofs of Ergodicity of (20) with Stationary Random Force

In this subsection we provide the necessary technical lemmas to which we refer in the proofs of Theorems 1 to 3, thus in the remainder of this subsection we assume $\widetilde{\boldsymbol{\Gamma}} \equiv \boldsymbol{\Gamma}, \widetilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}$. We begin by showing the existence of a class of suitable Lyapunov functions in the case of a bounded configurational domain, i.e., $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$.

Lemma 1. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$, $-\boldsymbol{\Gamma} \in \mathbb{R}^{(n+m)\times(n+m)}$ stable, then*

$$\mathcal{K}_l(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}) = \left(\boldsymbol{z}^T \boldsymbol{C}\boldsymbol{z}\right)^l + 1,\ l \in \mathbb{N},$$

where $\boldsymbol{C} \in \mathbb{R}^{n+m}$ is a symmetric positive definite matrix such that $\boldsymbol{\Gamma}^T\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\Gamma}$ is positive definite, defines a family of Lyapunov functions for the differential operator $\mathcal{L}_{\mathrm{GLE}}$, i.e., for each $l \in \mathbb{N}$ there exist constants $a_l > 0$, $b_l \in \mathbb{R}$, such that for $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}, \mathcal{K} = \mathcal{K}_l$, Assumption B.1 holds with $a = a_l, b = b_l$.

Proof. We show the existence of suitable constants $\widetilde{a}_l, \widetilde{b}_l$ so that the inequality (93) is satisfied for $\mathcal{K} = \widetilde{\mathcal{K}}_l := \mathcal{K}_l - 1$, and $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}, a = \widetilde{a}_l, b = \widetilde{b}_l$, which directly implies the statement of Lemma 1 for $a_l = \widetilde{a}_l$ and $b_l = \widetilde{b}_l + \widetilde{a}_l$. $-\boldsymbol{\Gamma}$ being a stable matrix ensures that there indeed exists a symmetric positive definite matrix $\boldsymbol{C}$ such that $\boldsymbol{\Gamma}^T\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\Gamma}$ is positive definite. Without loss of generality let $\min\sigma(\boldsymbol{C}) = 1$, so that

$$\|\boldsymbol{z}\|_2^2 \leq \boldsymbol{z}^T\boldsymbol{C}\boldsymbol{z} = \mathcal{K}_1(\boldsymbol{x}) - 1. \tag{63}$$

Furthermore,

$$\lambda = \sup_{z\in\Omega_z, \|\boldsymbol{z}\|_2=1} \frac{\boldsymbol{z}^T(\boldsymbol{\Gamma}^T\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\Gamma})\boldsymbol{z}}{\boldsymbol{z}^T\boldsymbol{C}\boldsymbol{z}},$$

so that

$$2\boldsymbol{z}^T\Gamma^T\boldsymbol{C}\boldsymbol{z} \geq \lambda\boldsymbol{z}^T\boldsymbol{C}\boldsymbol{z} = \lambda(\mathcal{K}_1(\boldsymbol{x}) - 1). \tag{64}$$

We first consider the case $l = 1$:

$$\begin{aligned}(\mathcal{L}_H + \mathcal{L}_O)\widetilde{\mathcal{K}}_1(\boldsymbol{x}) &= [2\boldsymbol{p}^T\boldsymbol{C}_{1,1} + 2\boldsymbol{s}^T\boldsymbol{C}_{1,2} - \boldsymbol{p}^T]\boldsymbol{F}(\boldsymbol{q}) - 2\boldsymbol{z}^T\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{z} \\ &\quad + \beta^{-1}\sum_{i,j}\left[\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j} \\ &\leq c_1||\boldsymbol{z}||_2 - 2\boldsymbol{z}^T\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{z} + \beta^{-1}\sum_{i,j}\left[\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j} \\ &\leq \frac{c_1}{\epsilon_1} + \epsilon_1||\boldsymbol{z}||_2^2 - 2\boldsymbol{z}^T\boldsymbol{\Gamma}^T\boldsymbol{C}\boldsymbol{z} + \beta^{-1}\sum_{i,j}\left[\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j},\end{aligned}$$

where

$$c_1 = \max_{\boldsymbol{q}\in\Omega_q, \|\boldsymbol{z}\|_2\leq 1}[2\boldsymbol{p}^T\boldsymbol{C}_{1,1} + 2\boldsymbol{s}^T\boldsymbol{C}_{1,2} - \boldsymbol{p}^T]\boldsymbol{F}(\boldsymbol{q}).$$

Thus, by (63) and (64),

$$\begin{aligned}(\mathcal{L}_H + \mathcal{L}_O)\widetilde{\mathcal{K}}_1(\boldsymbol{x}) &\leq \frac{c_1}{\epsilon_1} + \epsilon_1\widetilde{\mathcal{K}}_1(\boldsymbol{x}) - \lambda\widetilde{\mathcal{K}}_1(\boldsymbol{x}) + \beta^{-1}\sum_{i,j}\left[\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j} \\ &= -\widetilde{a}_1\widetilde{\mathcal{K}}_1(\boldsymbol{x}) + \widetilde{b}_1,\end{aligned}$$

with

$$\widetilde{a}_1 := (\lambda - \epsilon_1),\ \widetilde{b}_1 := \frac{c_1}{\epsilon_1} + (\lambda + \epsilon_1) + \beta^{-1}\sum_{i,j}\left[\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j},$$

so that $\widetilde{a}_1 > 0$ for sufficiently small $\epsilon_1 > 0$.

For $l > 1$ we find:

$$\begin{aligned}(\mathcal{L}_H + \mathcal{L}_O)\widetilde{\mathcal{K}}_l(\boldsymbol{x}) = l\widetilde{\mathcal{K}}_{l-1}(\boldsymbol{x})\big[\mathcal{L}_H\widetilde{\mathcal{K}}_1(\boldsymbol{x}) + (-(\boldsymbol{\Gamma}\boldsymbol{z})\cdot\nabla_z\widetilde{\mathcal{K}}_1(\boldsymbol{x})) + \beta^{-1}\sum_{i,j}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j}\big] \\ + 2l(l-1)\beta^{-1}\boldsymbol{z}^T\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\boldsymbol{z}\widetilde{\mathcal{K}}_{l-2}(\boldsymbol{x}).\end{aligned} \tag{65}$$

Let

$$\widetilde{\lambda} := \sup_{\boldsymbol{x}\in\Omega_{\boldsymbol{x}},\|\boldsymbol{z}\|_2=1} \left(\frac{\boldsymbol{z}^T \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \boldsymbol{C}\boldsymbol{z}}{\widetilde{\mathcal{K}}_1(\boldsymbol{x})} \right),$$

so that

$$\forall \boldsymbol{x} \in \Omega_{\boldsymbol{x}},\ \boldsymbol{z}^T \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \boldsymbol{C}\boldsymbol{z}\widetilde{\mathcal{K}}_{l-2}(\boldsymbol{x}) \leq \widetilde{\lambda}\widetilde{\mathcal{K}}_{l-1}(\boldsymbol{x}).$$

Thus, with

$$c_l := \min\left(0, -\beta^{-1}\sum_{i,j}\left[\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j} + \beta^{-1}\sum_{i,j}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{C}\right]_{i,j} + 2(l-1)\beta^{-1}\widetilde{\lambda}\right),$$

we find

$$\begin{aligned}(\mathcal{L}_H + \mathcal{L}_O)\widetilde{\mathcal{K}}_l(\boldsymbol{x}) &\leq l\widetilde{\mathcal{K}}_{l-1}(\boldsymbol{x})\left((\mathcal{L}_H + \mathcal{L}_O)\widetilde{\mathcal{K}}_1(\boldsymbol{x}) + c_l\right)\\ &\leq l\widetilde{\mathcal{K}}_{l-1}(\boldsymbol{x})\left(-\widetilde{a}_1\widetilde{\mathcal{K}}_1(\boldsymbol{x}) + \widetilde{b}_1 + c_l\right)\\ &\leq l\left(-\widetilde{a}_1\mathcal{K}_l(\boldsymbol{x}) + \frac{\widetilde{b}_1 + c_l}{\epsilon_l^{l-1}} + \epsilon_l\mathcal{K}_l(\boldsymbol{x})\right) = -\widetilde{a}_l\mathcal{K}_l(\boldsymbol{x}) + \widetilde{b}_l,\end{aligned} \tag{66}$$

with

$$\widetilde{a}_l := l(\widetilde{a}_1 - \epsilon_l),\ \widetilde{b}_l := l\frac{\widetilde{b}_1 + c_l}{\epsilon_l^{l-1}},$$

where $\epsilon_l > 0$ is chosen sufficiently small so that $\widetilde{a}_l > 0$. □

We next show the existence of a minorization condition in the case of $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$. The idea of the proof is to decompose the diffusion process into an Ornstein-Uhlenbeck process and a bounded remainder term, which then enables us to conclude the existence of a minorizing measure by virtue of the fact that the solution of Fokker-Planck equation associated with the Ornstein-Uhlenbeck process is a non-degenerate Gaussian at all times $t > 0$ and thus has full support. The idea of this approach is borrowed from [31] where it was used to show the minorization condition for a discretized version of the under-damped Langevin equation. Other applications of this trick can be found in [24,48].

Lemma 2. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$. If $\boldsymbol{\Gamma} \in \mathbb{R}^{(n+m)\times(n+m)}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{(n+m)\times(n+m)}$ are as in Theorem 1, then Assumption B.2 (minorization condition) holds for the SDE (20).*

Proof (Proof of Lemma 2). Let $\boldsymbol{q}(0) = \boldsymbol{q}_0$ and $\boldsymbol{z}(0) = \boldsymbol{z}_0$ with

$$(\boldsymbol{q}_0, \boldsymbol{z}_0) \in \Omega_{\boldsymbol{q}} \times \mathcal{C}_r,$$

where

$$C_r = \{z \in \Omega_{\boldsymbol{z}} : \|z\| < r\},$$

for arbitrary but fixed $r > 0$.

We can write the solution of (20) as

$$\boldsymbol{z}(t) = \boldsymbol{z}_0 + \mathscr{D}_{\boldsymbol{z}}(t) + \mathscr{G}_{\boldsymbol{z}}(t), \quad \boldsymbol{q}(t) = \boldsymbol{q}_0 + \mathscr{D}_{\boldsymbol{q}}(t) + \mathscr{G}_{\boldsymbol{q}}(t), \tag{67}$$

with

$$\mathscr{D}_{\boldsymbol{z}}(t) = \int_0^t e^{-(t-s)\boldsymbol{\Gamma}} \begin{pmatrix} \boldsymbol{F}(\boldsymbol{q}(s)) \\ \boldsymbol{0} \end{pmatrix} \mathrm{d}s, \quad \mathscr{G}_{\boldsymbol{z}}(t) = \int_0^t e^{-(t-s)\boldsymbol{\Gamma}} \boldsymbol{\Sigma} \mathrm{d}\boldsymbol{W}(s),$$

and

$$\mathscr{D}_{\boldsymbol{q}}(t) = \int_0^t \Pi_{\boldsymbol{p}} \mathscr{D}_{\boldsymbol{z}}(s) \mathrm{d}s, \quad \mathscr{G}_{\boldsymbol{q}}(t) = \int_0^t \Pi_{\boldsymbol{p}} \mathscr{G}_{\boldsymbol{z}}(s) \mathrm{d}s.$$

The variables $\mathscr{G}_{\boldsymbol{q}}(t)$ and $\mathscr{G}_{\boldsymbol{z}}(t)$ are correlated and Gaussian, i.e.,

$$\begin{pmatrix} \mathscr{G}_{\boldsymbol{q}}(t) \\ \mathscr{G}_{\boldsymbol{z}}(t) \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}_t, \mathcal{V}_t),$$

with some $\boldsymbol{\mu}_t \in \Omega_{\boldsymbol{x}}$ and $\mathcal{V}_t \in \mathbb{R}^{(2n+m)\times(2n+m)}$. More specifically, $\tilde{\boldsymbol{z}}(t) = \boldsymbol{z}(0) + \mathscr{G}_{\boldsymbol{z}}(t)$ and $\boldsymbol{q}(0) + \mathscr{G}_{\boldsymbol{q}}(t)$ corresponds to the solution of of the linear SDE

$$\begin{aligned} \dot{\tilde{\boldsymbol{q}}} &= \tilde{\boldsymbol{p}}, \\ \dot{\tilde{\boldsymbol{z}}} &= -\boldsymbol{\Gamma}\tilde{\boldsymbol{z}} + \boldsymbol{\Sigma}\dot{\boldsymbol{W}}, \end{aligned} \tag{68}$$

where $\tilde{\boldsymbol{z}}(t) = (\tilde{\boldsymbol{p}}(t), \tilde{\boldsymbol{s}}(t)) \in \Omega_{\boldsymbol{p}} \times \Omega_{\boldsymbol{s}}$. The law of $\tilde{\boldsymbol{q}}(t), \tilde{\boldsymbol{z}}(t)$ has full support for all $t > 0$, provided that the covariance matrix $\mathcal{V}_t$ is invertible. This is indeed the case since $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are required to be such that (20) satisfies the parabolic Hörmander condition. It follows that the system (68) satisfies the parabolic Hörmander condition. By Proposition B.2, we conclude that the law of $(\tilde{\boldsymbol{q}}(t), \tilde{\boldsymbol{z}}(t))$ has a density with respect to the Lebesgue measure for any $t > 0$, which rules out the possibility of $\mathcal{V}_t$ being singular.

Let $\boldsymbol{C} \in \mathbb{R}^{(n+m)\times(n+m)}$ be symmetric positive definite such that $\boldsymbol{\Gamma}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{\Gamma}^T$ is positive definite as well, and consider the norm $\|\cdot\|_{\boldsymbol{C}}$,

$$\|\cdot\|_{\boldsymbol{C}} := \boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}, \ \boldsymbol{z} \in \mathbb{R}^{n+m}.$$

The increment $\mathscr{D}_{\boldsymbol{z}}(t)$ is uniformly bounded since

$$\|\mathscr{D}_{\boldsymbol{z}}(t)\|_{\boldsymbol{C}} \leq \|\boldsymbol{\Gamma}^{-1}\|_{\mathcal{B}(\boldsymbol{C})} \|\boldsymbol{F}\|_{L^\infty} < \infty,$$

where

$$\|\boldsymbol{\Gamma}^{-1}\|_{\mathcal{B}(\boldsymbol{C})} := \max_{v \in \mathbb{R}^{2n}} \frac{\|\boldsymbol{\Gamma}^{-1} v\|_{\boldsymbol{C}}}{\|v\|_{\boldsymbol{C}}} = \frac{1}{2} \min \sigma \left(\boldsymbol{\Gamma}^T \boldsymbol{C} + \boldsymbol{C}\boldsymbol{\Gamma}\right),$$

denotes the operator norm of $\boldsymbol{\Gamma}^{-1}$ induced by $\|\cdot\|_{\boldsymbol{C}}$. It follows that also $\mathscr{D}_{\boldsymbol{q}}(t)$ is bounded since

$$\|\mathscr{D}_{\boldsymbol{q}}(t)\| \leq t \|\mathscr{D}_{\boldsymbol{z}}(t)\|_{\boldsymbol{C}} < \infty.$$

Let $\mu_{x_0,t}$ denote the law of $(\boldsymbol{q}(t), \boldsymbol{z}(t))$ and $\rho_{x_0,t}$ be the associated density. For fixed $t > 0$, the terms $\mathscr{D}_{\boldsymbol{q}}(t)$ and $\mathscr{D}_{\boldsymbol{z}}(t)$ are bounded and the law of

$(\boldsymbol{q}(0)+\mathscr{G}_{\boldsymbol{q}}(t), \boldsymbol{z}(0)+\mathscr{G}_{\boldsymbol{z}}(t))$ has full support, in particular the measure $\mu_{x_0,t}(\mathrm{d}\boldsymbol{x}) = \rho_{x_0,t}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ of the superposition

$$(\boldsymbol{q}(t), \boldsymbol{z}(t)) = (\boldsymbol{q}(0) + \mathscr{D}_{\boldsymbol{q}}(t) + \mathscr{G}_{\boldsymbol{q}}(t), \boldsymbol{z}(0) + \mathscr{D}_{\boldsymbol{z}}(t) + \mathscr{G}_{\boldsymbol{z}}(t))$$

has full support. Now define $\rho \in \mathcal{C}(\Omega_{\boldsymbol{x}}, \mathbb{R}_+)$ as

$$\rho(x) := \min_{\boldsymbol{x}_0 \in \mathcal{C}_r} \rho_{x_0,t}(x).$$

By construction the associated probability measure satisfies the properties of ν in Assumption B.2. □

We next consider the case $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$. The following Lemma 3 shows the existence of a suitable class of Lyapunov functions.

Lemma 3. *Let $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$. If*

(i) $-\boldsymbol{\Gamma} \in \mathbb{R}^{(n+m)\times(n+m)}$ is a stable matrix and $\boldsymbol{\Sigma} \in \mathbb{R}^{(n+m)\times(n+m)}$ such that

$$\boldsymbol{\Gamma}_{2,2}\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{\Gamma}_{2,2}^T$$

is positive definite with $\boldsymbol{Q}$ as specified in Assumption 1,
(ii) the force $\boldsymbol{F} \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R}^n)$ satisfies Assumption 2.

Furthermore, if either

(iii) $\boldsymbol{\Gamma}_{1,1}$ is positive definite,

or

(iv) the force $\boldsymbol{F}$ satisfies Assumption 3,

then

$$\mathcal{K}_l(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}) = \left(\boldsymbol{z}^T \boldsymbol{C}_{A,B}\boldsymbol{z} + \|\boldsymbol{q}\|_2^2 + 2\langle \boldsymbol{p}, \boldsymbol{q}\rangle + BD(V(\boldsymbol{q}) - u_{\min}) + 1\right)^l, \ l \in \mathbb{N}, \tag{69}$$

where

$$\boldsymbol{C}_{A,B} = \begin{pmatrix} B\boldsymbol{I}_n & A\boldsymbol{\Gamma}_{2,1}^T \\ A\boldsymbol{\Gamma}_{2,1} & B\boldsymbol{Q}^{-1} \end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)},$$

is a symmetric positive definite matrix for suitably chosen scalars $A, B > 0$, and $V \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ as specified in Assumption 2, defines a family of Lyapunov functions for the differential operator $\mathcal{L}_{\mathrm{GLE}}$, i.e., for each $l \in \mathbb{N}$ there exist constants $a_l > 0$, $b_l \in \mathbb{R}$, such that for $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}, \mathcal{K} = \mathcal{K}_l$, Assumption B.1 holds for $a = a_l, b = b_l$.

Proof. Rewriting $\mathcal{K}_l$ as

$$\mathcal{K}_l(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}) = \left(\boldsymbol{x}^T \hat{\boldsymbol{C}}_{A,B}\boldsymbol{x} + BD(V(\boldsymbol{q}) - u_{\min}) + 1\right)^l, \ l \in \mathbb{N},$$

where

$$\hat{\boldsymbol{C}}_{A,B} = \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{I}_n & B\boldsymbol{I}_n & A\boldsymbol{\Gamma}_{2,1}^T \\ \boldsymbol{0} & A\boldsymbol{\Gamma}_{2,1} & B\boldsymbol{Q}^{-1} \end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)},$$

we find by successive application of Lemma A.1, that for any $A' \geq 0$ there exists $B' > 0$ so that for $A = A'$ and $B \geq B'$ the matrix $\hat{\boldsymbol{C}}_{A,B}$ is positive definite and thus $\mathcal{K}_l \geq 1$ and $\mathcal{K}_l(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$. We first consider the case $l = 1$. Define

$$\boldsymbol{G} := \begin{pmatrix} \boldsymbol{0} & -\boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{I}_n & \boldsymbol{\Gamma}_{1,1} & \boldsymbol{\Gamma}_{1,2} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{2,1} & \boldsymbol{\Gamma}_{2,2} \end{pmatrix} \in \mathbb{R}^{(2n+m)\times(2n+m)}, \tag{70}$$

and

$$\tilde{\boldsymbol{Q}} := \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q} \end{pmatrix},$$

we find

$$\begin{aligned} \mathcal{L}_{\mathrm{GLE}}\mathcal{K}_1(\boldsymbol{x}) = &- \left(-\left[\boldsymbol{F}(\boldsymbol{q})\right]^T, \boldsymbol{p}^T, \boldsymbol{s}^T\right) \boldsymbol{G}^T \hat{\boldsymbol{C}}_{A,B}\boldsymbol{x} + DB\boldsymbol{I}_n\boldsymbol{p}\cdot\nabla_{\boldsymbol{q}}V(\boldsymbol{q}) \\ &+ \frac{\beta^{-1}}{2}\nabla_{\boldsymbol{z}}\cdot\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\nabla_{\boldsymbol{z}}(\boldsymbol{z}\,\tilde{\boldsymbol{Q}}^{-1}\boldsymbol{z})\right), \end{aligned}$$

with

$$\boldsymbol{G}^T\boldsymbol{C} = -\begin{pmatrix} \boldsymbol{I}_n & B\boldsymbol{I}_n & \boldsymbol{\Gamma}_{2,1} \\ -\boldsymbol{I}_n + \boldsymbol{\Gamma}_{1,1} & -\boldsymbol{I}_n + B\boldsymbol{\Gamma}_{1,1} + A\boldsymbol{\Gamma}_{2,1}^T\boldsymbol{\Gamma}_{2,1} & B\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}_{2,1}^T \\ \boldsymbol{\Gamma}_{1,2}^T & \boldsymbol{\Gamma}_{2,1}\boldsymbol{\Gamma}_{2,2} + B\boldsymbol{\Gamma}_{1,2}^T & \boldsymbol{\Gamma}_{2,1}\boldsymbol{\Gamma}_{1,2} + B\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}_{2,2}^T \end{pmatrix}.$$

Hence, by virtue of (48) and Assumption 2(i),

$$\begin{aligned} &\mathcal{L}_{GLE}\mathcal{K}_1(\boldsymbol{x}) \leq \\ &\underbrace{-\boldsymbol{x}^T\begin{pmatrix} E\boldsymbol{I}_n & \boldsymbol{0} & \boldsymbol{0} \\ (-\boldsymbol{I}_n + \boldsymbol{\Gamma}_{1,1}) & -\boldsymbol{I}_n + B\boldsymbol{\Gamma}_{1,1} + A\boldsymbol{\Gamma}_{2,1}^T\boldsymbol{\Gamma}_{2,1} & B\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}_{2,1}^T \\ \boldsymbol{\Gamma}_{1,2}^T & A\boldsymbol{\Gamma}_{2,1}\boldsymbol{\Gamma}_{2,2} + B\boldsymbol{\Gamma}_{1,2}^T & A\boldsymbol{\Gamma}_{2,1}\boldsymbol{\Gamma}_{1,2} + B\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}_{2,2}^T \end{pmatrix}\boldsymbol{x}}_{=:\,\widehat{\boldsymbol{R}}_{A,B}} \\ &- A\nabla_{\boldsymbol{q}}V(\boldsymbol{q})^T\boldsymbol{\Gamma}_{2,1}^T\boldsymbol{s} + F + \frac{\beta^{-1}}{2}\sum_{i,j}[\tilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\tilde{\boldsymbol{Q}}^{-1}]_{i,j}. \end{aligned} \tag{71}$$

In order to show the existence of constants a_1 and b_1 such that the respective Lyapunov inequality satisfied, one needs to show that the right hand side of the above inequality (71) can be bounded from above by a negative definite quadratic form.

Case $\operatorname{rank}(\boldsymbol{\Gamma}_{1,1}) = n$: Let $A = 0$. In this case it is sufficient to show that the symmetric part

$$\widehat{\boldsymbol{R}}^S_{A,B} = \frac{1}{2}\left(\widehat{\boldsymbol{R}}_{A,B} + \widehat{\boldsymbol{R}}^T_{A,B}\right)$$

of $\widehat{\boldsymbol{R}}_{A,B}$ is positive definite. The lower right block

$$\left[\widehat{\boldsymbol{R}}^S_{A,B}\right]_{(n+1):(2n+m),(n+1):(2n+m)} = -\boldsymbol{I}_n + \frac{B}{2}\left(\boldsymbol{\Gamma}\tilde{\boldsymbol{Q}} + \tilde{\boldsymbol{Q}}\boldsymbol{\Gamma}^T\right) \in \mathbb{R}^{(n+m)\times(n+m)},$$

of $\widehat{\boldsymbol{R}}^s_{0,B}$ is positive definite for sufficiently large $B > 0$. In particular

$$\min\sigma\left(\left[\widehat{\boldsymbol{R}}^S_{A,B}\right]_{(n+1):(2n+m),(n+1):(2n+m)}\right) = O(B),$$

as $B \to \infty$. Thus, by virtue of Lemma A.1 for $E > 0$ there is a $B' > 0$ such that $\widehat{\boldsymbol{R}}^s_{0,B}$ is indeed positive definite for all $B \geq B'$.

Case $\boldsymbol{\Gamma}_{1,1} = \boldsymbol{0}$*:* If Assumption 3 holds, then by Remark 2 this implies that there are values $\overline{H} > 0$ and $\overline{h} \in \mathbb{R}$ so that

$$|\langle \boldsymbol{g}, \boldsymbol{F}(\boldsymbol{q})\rangle| \leq \overline{H}|\langle \boldsymbol{g}, \boldsymbol{q}\rangle| + \overline{h}.$$

Therefore, it is sufficient to show that there are constants A, B, E so that the function

$$\begin{aligned}\varphi(\boldsymbol{x}) &= \max\left(-\boldsymbol{x}^T\widehat{\boldsymbol{R}}_{A,B}\boldsymbol{x} - A\overline{H}\boldsymbol{q}^T\boldsymbol{\Gamma}^T_{2,1}\boldsymbol{s},\ \ -\boldsymbol{x}^T\widehat{\boldsymbol{R}}_{A,B}\boldsymbol{x} + A\overline{H}\boldsymbol{q}^T\boldsymbol{\Gamma}^T_{2,1}\boldsymbol{s}\right)\\ &= \max_{i=1,2} -\boldsymbol{x}^T\tilde{\boldsymbol{R}}^{(i)}_{A,B,E}\boldsymbol{x},\end{aligned} \tag{72}$$

can be bounded from above by a negative definite quadratic form. This means that we have to show that for suitable constants $A, B, E > 0$ the symmetric part of the matrix

$$\tilde{\boldsymbol{R}}^{(i)}_{A,B,E} = \begin{pmatrix} E\boldsymbol{I}_n & \boldsymbol{0} & (-1)^i A\overline{H}\boldsymbol{\Gamma}^T_{2,1} \\ -\boldsymbol{I}_n & -\boldsymbol{I}_n + A\boldsymbol{\Gamma}^T_{2,1}\boldsymbol{\Gamma}_{2,1} & \boldsymbol{0} \\ \boldsymbol{\Gamma}^T_{1,2} & A\boldsymbol{\Gamma}_{2,1}\boldsymbol{\Gamma}_{2,2} & A\boldsymbol{\Gamma}_{2,1}\boldsymbol{\Gamma}_{1,2} + B\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}^T_{2,2} \end{pmatrix},$$

is positive definite for $i \in \{0, 1\}$. (Note that we used $\boldsymbol{\Gamma}^T_{1,2} - \boldsymbol{Q}^{-1}\boldsymbol{\Gamma}_{2,1} = \boldsymbol{0}$ in the derivation of the form of $\tilde{\boldsymbol{R}}^i_{A,B}$.) Since $\boldsymbol{\Gamma}^T_{2,1}\boldsymbol{\Gamma}_{2,1}$ is positive definite we can choose A sufficiently large so that $-\boldsymbol{I}_n + A\boldsymbol{\Gamma}^T_{2,1}\boldsymbol{\Gamma}_{2,1}$ is positive definite. The positive definiteness of the symmetric part of $\tilde{\boldsymbol{R}}^{(i)}_{A,B,E}$, $i \in \{0, 1\}$ follows for sufficiently large $B > 0$ and $E > 0$ by successive application of Lemma A.1.

For $l > 1$ we find:

$$\begin{aligned}(\mathcal{L}_H + \mathcal{L}_O)\mathcal{K}_l(\boldsymbol{x}) &= l\mathcal{K}_{l-1}(\boldsymbol{x})\mathcal{L}_H\mathcal{K}_1(\boldsymbol{x}) + l\mathcal{K}_{l-1}(\boldsymbol{x})(-\boldsymbol{\Gamma}^T\boldsymbol{z}\cdot\nabla_{\boldsymbol{z}}\mathcal{K}_1(\boldsymbol{x})) \\ &\quad + l\frac{\beta^{-1}}{2}\nabla_{\boldsymbol{z}}\cdot\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\nabla_{\boldsymbol{z}}\mathcal{K}_1(\boldsymbol{x})\mathcal{K}_{l-1}(\boldsymbol{x})\right) \\ &= -l\mathcal{K}_{l-1}(\boldsymbol{x})(\boldsymbol{z}^T\boldsymbol{\Gamma}^T\tilde{\boldsymbol{Q}}\boldsymbol{z}) + l\beta^{-1}\sum_{i,j}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\tilde{\boldsymbol{Q}}\right]_{i,j}\mathcal{K}_{l-1}(\boldsymbol{x}) \\ &\quad + 2l(l-1)\beta^{-1}\boldsymbol{z}^T\tilde{\boldsymbol{Q}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\tilde{\boldsymbol{Q}}\boldsymbol{z}\mathcal{K}_{l-2}(\boldsymbol{x}) \\ &\leq -l\mathcal{K}_{l-1}(\boldsymbol{x})\left((\mathcal{L}_H + \mathcal{L}_O)\mathcal{K}_1(\boldsymbol{x}) + c_2\right) \\ &\leq l\mathcal{K}_{l-1}(\boldsymbol{x})\left(-a_1\mathcal{K}_1(\boldsymbol{x}) + b_1 + c_2\right) \\ &\leq l\left(-a_1\mathcal{K}_l(\boldsymbol{x}) + \frac{b_1+c_2}{\epsilon_l^{l-1}} + \epsilon_l\mathcal{K}_l\right) = -a_l\mathcal{K}_l(\boldsymbol{x}) + b_l,\end{aligned} \tag{73}$$

with

$$c_2 = -\beta^{-1}\sum_{i,j}\left[\tilde{\boldsymbol{Q}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\tilde{\boldsymbol{Q}}\right]_{i,j} + \beta^{-1}\sum_{i,j}\left[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\tilde{\boldsymbol{Q}}\right]_{i,j}$$

and

$$a_l := l(a_1 - \epsilon_l),\ b_l := l\frac{b_1+c_2}{\epsilon_l^{l-1}}$$

where $\epsilon_l > 0$ sufficiently small so that $a_l > 0$. □
□

We mention that Assumption 2 is commonly also required for the construction of suitable Lyapunov functions in the case of the underdamped Langevin equation if $\Omega_{\boldsymbol{q}}$ is unbounded. Assumption 3 is an additional constraint on the force function $\boldsymbol{F}$, which is not required in the case of the underdamped Langevin equation. It is therefore not surprising that this assumption can be dropped if the noise process $\boldsymbol{\eta}$ in the GLE contains a nondegenerate white noise component.

If $\boldsymbol{\Sigma}$ has full rank the minorization can be demonstrated using a simple control argument.

Lemma 4. *Let $\Omega_{\boldsymbol{q}} = \mathbb{R}^n$. If* $\operatorname{rank}(\boldsymbol{\Sigma}) = n+m$, *then* (20) *satisfies a minorization condition (Assumption B.2).*

Proof. Note that by Proposition 7, (ii) $\operatorname{rank}(\boldsymbol{\Sigma}) = n+m$ immediately implies that the SDE satisfies the parabolic Hörmander condition. Since $\boldsymbol{\Sigma}$ is invertible, we can easily solve the associated control problem which then by Lemma B.1 implies that a minorization condition is satisfied. The proof of the existence of a suitable control is essentially the same as in the case of the under-damped Langevin equation (see e.g. [36]): Let $T > 0$ and $(\boldsymbol{q}^-, \boldsymbol{p}^-, \boldsymbol{s}^-), (\boldsymbol{q}^+, \boldsymbol{p}^+, \boldsymbol{s}^+) \in \mathbb{R}^{2n+m}$. We need to show that there exists $u \in L^1([0,T], \mathbb{R}^m)$, solving the control problem

$$\begin{aligned}\dot{\boldsymbol{q}} &= \boldsymbol{p}, \\ \dot{\boldsymbol{p}} &= \boldsymbol{F}(\boldsymbol{q}) - \boldsymbol{\Gamma}_{1,1}\boldsymbol{p} + \boldsymbol{\Gamma}_{1,2}\boldsymbol{s} + \boldsymbol{\Sigma}_1\boldsymbol{u}, \\ \dot{\boldsymbol{s}} &= -\boldsymbol{\Gamma}_{2,1}\boldsymbol{p} + \boldsymbol{\Gamma}_{2,2}\boldsymbol{s} + \boldsymbol{\Sigma}_2\boldsymbol{u},\end{aligned} \tag{74}$$

subject to

$$(\boldsymbol{q}(0), \boldsymbol{p}(0), \boldsymbol{s}(0)) = (\boldsymbol{q}^-, \boldsymbol{p}^-, \boldsymbol{s}^-),\ (\boldsymbol{q}(T), \boldsymbol{p}(T), \boldsymbol{s}(T)) = (\boldsymbol{q}^+, \boldsymbol{p}^+, \boldsymbol{s}^+).$$

It is easy to verify that there exists a smooth path $\tilde{\boldsymbol{q}} \in \mathcal{C}^2([0,T], \mathbb{R}^n)$ and $\tilde{\boldsymbol{s}} \in \mathcal{C}^2([0,T], \mathbb{R}^m)$ such that

$$(\tilde{\boldsymbol{q}}(0), \dot{\tilde{\boldsymbol{q}}}(0)) = (\boldsymbol{q}^-, \boldsymbol{p}^-),\ (\tilde{\boldsymbol{q}}(T), \dot{\tilde{\boldsymbol{q}}}(T)) = (\boldsymbol{q}^+, \boldsymbol{p}^+),$$

and

$$\tilde{\boldsymbol{s}}(0) = \boldsymbol{s}^-,\ \tilde{\boldsymbol{s}}(T) = \boldsymbol{s}^+.$$

Rewrite (74) as a second order differential equation in $\boldsymbol{q}$ and $\boldsymbol{s}$:

$$\begin{aligned} \ddot{\boldsymbol{q}} &= -\nabla_q U(\boldsymbol{q}) - \boldsymbol{\Gamma}_{1,2}\dot{\boldsymbol{q}} - \boldsymbol{\Gamma}_{1,2}\boldsymbol{s} + \boldsymbol{\Sigma}_1 \boldsymbol{u}, \\ \dot{\boldsymbol{s}} &= -\boldsymbol{\Gamma}_{2,1}\dot{\boldsymbol{q}} - \boldsymbol{\Gamma}_{2,2}\boldsymbol{s} + \boldsymbol{\Sigma}_2 \boldsymbol{u}, \end{aligned}$$

thus,

$$\mathbf{u}(t) = \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \ddot{\tilde{\boldsymbol{q}}}(t) + \nabla_q U(\tilde{\boldsymbol{q}}(t)) + \boldsymbol{\Gamma}_{1,1}\dot{\tilde{\boldsymbol{q}}}(t) + \boldsymbol{\Gamma}_{1,2}\tilde{\boldsymbol{s}}(t) \\ \dot{\tilde{\boldsymbol{s}}}(t) + \boldsymbol{\Gamma}_{2,1}\dot{\tilde{\boldsymbol{q}}}(t) + \boldsymbol{\Gamma}_{2,2}\tilde{\boldsymbol{s}}(t) \end{pmatrix}, \tag{75}$$

is a solution of (74). □

The following Lemma 5 shows that the minorization condition is satisfied in the case of a GLE with unbounded configurational domain and $\boldsymbol{\Gamma}_{1,1} = \mathbf{0}$.

Lemma 5. *Under the same conditions as Theorem 3 it follows that Assumption B.2 is satisfied for* (20).

Proof. By Assumption 3 the force $\boldsymbol{F}$ can be decomposed as

$$\boldsymbol{F}(\boldsymbol{q}) = \boldsymbol{F}_1(\boldsymbol{q}) + \boldsymbol{F}_2(\boldsymbol{q}),$$

where $\|\boldsymbol{F}_1(\boldsymbol{q})\|_\infty$ is uniformly bounded in $\boldsymbol{q} \in \mathbb{R}$ and

$$\boldsymbol{F}_2(\boldsymbol{q}) = \boldsymbol{H}\boldsymbol{q},$$

with $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ being a positive definite matrix. Consider the dynamics

$$\begin{aligned} \dot{\boldsymbol{q}}^a &= \boldsymbol{p}^a, \\ \dot{\boldsymbol{p}}^a &= -\boldsymbol{H}\boldsymbol{q}^a - \boldsymbol{\Gamma}_{1,2}\boldsymbol{s}^a, \\ \dot{\boldsymbol{g}}^a &= -\boldsymbol{\Gamma}_{2,1}\boldsymbol{p}^a - \boldsymbol{\Gamma}_{2,2}\boldsymbol{s}^a + \frac{\beta^{-1}}{2}\boldsymbol{\Sigma}_2 \dot{\boldsymbol{W}}, \\ &\text{with}\ \ (\boldsymbol{q}^a(0), \boldsymbol{p}^a(0), \boldsymbol{s}^a(0)) = \boldsymbol{x}_0, \end{aligned} \tag{76}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^{2n+m}$. The solution of (76) is Gaussian hence

$$\mu_t^a(\mathrm{d}\boldsymbol{x}) = \mathcal{N}(\mathrm{d}\boldsymbol{x}; \boldsymbol{\mu}_t, \mathscr{V}_t),$$

where $\boldsymbol{\mu}_t \in \mathbb{R}^{2n+m}$ and $\mathscr{V}_t \in \mathbb{R}^{(2n+m)\times(2n+m)}$. Moreover, by Proposition 7, (iii), the SDE (76) is hypoelliptic, hence $\mathscr{V}_t$ is non-singular for all $t > 0$. As a consequence

$$\operatorname{supp}(\mu_t^a) = \Omega_{\boldsymbol{x}}$$

for all $t > 0$. Moreover, we notice that

$$\boldsymbol{F}_1(\boldsymbol{q}) = \boldsymbol{u}(\boldsymbol{q})\boldsymbol{\Sigma}_2,$$

with

$$\boldsymbol{u}(\boldsymbol{q}) = -\boldsymbol{F}_1(\boldsymbol{q})\boldsymbol{I}_{n,m}\boldsymbol{\Sigma}_2^{-1},$$

where

$$\boldsymbol{I}_{n,m} = (\boldsymbol{I}_n, \boldsymbol{0}) \in \mathbb{R}^{n\times m}.$$

Using Lemma 9 it follows by the same chain of arguments as in the proof of Lemma 8, that $\boldsymbol{u}$ satisfies Novikov's condition and by virtue of Girsanov's theorem the support of the law μ_t of the solution of (20) with initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$ coincides with the law of $\mu^a_{\boldsymbol{x}_0,t}$, i.e., $\operatorname{supp}(\mu_t) = \Omega_{\boldsymbol{x}}$. Let $\mu_{\boldsymbol{x}_0,t}(\mathrm{d}\boldsymbol{x}) = \rho_{\boldsymbol{x}_0,t}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. As in the proof of Lemma 2 we can construct a minoring measure $\eta(\mathrm{d}x) = \rho(x)\mathrm{d}x$, as

$$\rho(x) := \min_{\boldsymbol{x}_0\in\mathcal{C}_r} \rho_{x_0,t}(x).$$

where $\mathcal{C}_r \subset \mathbb{R}^{2n+m}$ is a sufficiently large compact set. □

Lemma 9 allows to conclude that Novikov's condition is satisfied under the assumptions of the preceding Lemma 5.

Lemma 6. *Let*

$$\widehat{\mathcal{K}}_\theta(\boldsymbol{x}) = e^{\frac{\theta}{2}\mathcal{K}_l(\boldsymbol{x})},\ l = 1,$$

with $\mathcal{K}_1$ *as defined in* (69). *Under the same conditions as in Lemma 3, and provided that Assumption B.1 holds for* $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}$, $\mathcal{K} = \mathcal{K}_1$, *then also* $\widehat{\mathcal{K}}_\theta$ *satisfies Assumption B.1 for* $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}$ *and sufficiently small* $\theta > 0$.

Proof. A simple calculation shows

$$\mathcal{L}_{\mathrm{GLE}}\widehat{\mathcal{K}}_\theta(\boldsymbol{x}) = \left(\theta\mathcal{L}_{\mathrm{GLE}}\mathcal{K}_1(\boldsymbol{x}) + \frac{\beta^{-1}}{2}\left(\theta\sum_{i,j}\left[(\tilde{\boldsymbol{Q}} - \boldsymbol{I}_{n+m})\widetilde{\boldsymbol{C}}\right]_{i,j} + \theta^2\boldsymbol{z}^T\widetilde{\boldsymbol{C}}\boldsymbol{z}\right)\right)\widehat{\mathcal{K}}_\theta(\boldsymbol{x}),$$

with

$$\widetilde{\boldsymbol{C}} = \tilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\tilde{\boldsymbol{Q}}^{-1}.$$

From Lemma 3 we know $\mathcal{L}_{\mathrm{GLE}}\mathcal{K}_1(\boldsymbol{x}) = \Theta\left(-\|\boldsymbol{x}\|^2\right)$, thus

$$\mathcal{L}_{\mathrm{GLE}}\widehat{\mathcal{K}}_\theta(\boldsymbol{x}) = \left(-\Theta\left(\theta\|\boldsymbol{x}\|^2\right) + \Theta\left((1+\theta)\|\boldsymbol{z}\|\right) + \Theta\left(\theta^2\|\boldsymbol{z}\|^2\right)\right)\mathcal{K}_\theta(\boldsymbol{x}),$$

thus for sufficiently small $\theta > 0$ and suitable $b \in \mathbb{R}$,

$$\mathcal{L}_{\mathrm{GLE}}\widehat{\mathcal{K}}_\theta(\boldsymbol{x}) < -\widehat{\mathcal{K}}_\theta(\boldsymbol{x}) + b.$$

□

3.4 Technical Lemmas Required in the Proofs of Ergodicity of (20) with Non-stationary Random Force

We first show that under the assumptions of Theorem 4 a minorization condition is satisfied for (20). For $r > 0$ let in the following $C_r := \{(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{s}) : \|\boldsymbol{p}, \boldsymbol{s}\|_2 < r\}$.

Lemma 7. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$ and $\widetilde{\boldsymbol{\Gamma}}_{1,2}, \widetilde{\boldsymbol{\Gamma}}_{2,1}, \widetilde{\boldsymbol{\Gamma}}_{2,2}, \widetilde{\boldsymbol{\Sigma}}_2 \in \mathcal{C}^\infty(\Omega_{\boldsymbol{q}}, \mathrm{GL}_n(\mathbb{R}))$, such that $-\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})$ is stable for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$. Let $r > 0$ and $\boldsymbol{x}_0 \in C_r$. For any $t > 0$ the law $\mu_t^{\boldsymbol{x}_0} := e^{t\mathcal{L}^\dagger}\delta_{\boldsymbol{x}_0}$ of the solution $\boldsymbol{x}(t)$ of (20) with initial condition $\boldsymbol{x}(t) = \boldsymbol{x}_0$ has full support. In particular, Assumption B.2 (minorization condition) holds.*

Proof. Let $\boldsymbol{x}_0 = (\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{s}_0) \in C_r$ and $\tilde{\boldsymbol{x}}_0 = (\boldsymbol{q}_0, \boldsymbol{p}_0, \boldsymbol{g}_0)$ with $\boldsymbol{g}_0 = \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{s}_0$. Consider the following cascade of modifications of (20):

$$
\begin{aligned}
\dot{\boldsymbol{q}}^c &= \boldsymbol{p}^c, \\
\dot{\boldsymbol{p}}^c &= \boldsymbol{F}(\boldsymbol{q}) - \boldsymbol{g}^c \\
\dot{\boldsymbol{g}}^c &= \sum_{i=1}^{n} \boldsymbol{p}_i^c \left(\partial_{\boldsymbol{q}_i} \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}^c)\right) \boldsymbol{g}^c - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}^c)\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}^c)\boldsymbol{p}^c \\
&\quad - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}^c)\widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q}^c)\widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q}^c)\boldsymbol{g}^c + \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}^c)\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}^c)\,\dot{\boldsymbol{W}}_t, \\
&\text{with} \quad (\boldsymbol{q}^c(0), \boldsymbol{p}^c(0), \boldsymbol{g}^c(0)) = \tilde{\boldsymbol{x}}_0,
\end{aligned}
\tag{77}
$$

and

$$
\begin{aligned}
\dot{\boldsymbol{q}}^b &= \boldsymbol{p}^b, \\
\dot{\boldsymbol{p}}^b &= \boldsymbol{F}(\boldsymbol{q}^b) - \boldsymbol{g}^b, \\
\dot{\boldsymbol{g}}^b &= \boldsymbol{p}^b - \boldsymbol{g}^b + \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}^b)\,\dot{\boldsymbol{W}}_t, \\
&\quad \text{with} \quad (\boldsymbol{q}^b(0), \boldsymbol{p}^b(0), \boldsymbol{g}^b(0)) = \tilde{\boldsymbol{x}}_0,
\end{aligned}
\tag{78}
$$

and

$$
\begin{aligned}
\dot{\boldsymbol{q}}^a &= \boldsymbol{p}^a, \\
\dot{\boldsymbol{p}}^a &= \boldsymbol{F}(\boldsymbol{q}^a) - \boldsymbol{g}^a, \\
\dot{\boldsymbol{g}}^a &= \boldsymbol{p}^a - \boldsymbol{g}^a + \dot{\boldsymbol{W}}, \\
&\quad \text{with} \quad (\boldsymbol{q}^a(0), \boldsymbol{p}^a(0), \boldsymbol{g}^a(0)) = \tilde{\boldsymbol{x}}_0.
\end{aligned}
\tag{79}
$$

Let $\mu_t^a, \mu_t^b, \mu_t^c$ denote the law of the solution of (79), (78) and (77), respectively. We show that for any $t > 0$

(i) $\mathrm{supp}(\mu_t^a) = \Omega_{\boldsymbol{x}}$,
(ii) $\mathrm{supp}(\mu_t^b) = \mathrm{supp}(\mu_t^a)$,
(iii) $\mathrm{supp}(\mu_t^c) = \mathrm{supp}(\mu_t^b)$,
(vi) $\mathrm{supp}(\mu_t) = \mathrm{supp}(\mu_t^c)$,

which then immediately implies that $\mathrm{supp}(\mu_t) = \Omega_{\boldsymbol{x}}$ for $t > 0$ and the minorization condition follows by the same arguments as in the proof of Lemma 2.

- Regarding (i): the system (79) satisfies the condition of Lemma 2, hence for sufficiently large $t' > 0$ the law of (79) at times $t \geq t'$ has full support.
- Regarding (ii): since $\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q})$ is invertible, the controllability properties of (78) are identical to the controllability properties of (79), hence as a consequence of the Strook-Varadhan support theorem [57] the law of (78) and the law of (79) at time t' coincide. In particular, together with (i) $\mathrm{supp}(\mu_t^c) = \mathrm{supp}(\mu_t^b) = \Omega_{\boldsymbol{x}}$.
- Regarding (iii): We show this using Proposition B.3 (Girsanov's theorem). The difference of the drift terms in (78) and (77) can be written as

$$\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}^c)\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q}^c)\boldsymbol{u}(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g}),$$

  with $\boldsymbol{u}(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g})$ as defined in (82). By Lemma 8 the function $\boldsymbol{u}$ satisfies Novikov's condition (101), which means that Proposition B.3 (Girsanov's theorem) is applicable and it follows that the support of the solution of (78) at t' coincides with the support of the solution of (77) at t', i.e., $\mathrm{supp}(\mu_t^c) = \mathrm{supp}(\mu_t^b) = \Omega_{\boldsymbol{x}}$.
- Regarding (iv): We first note that since (i)–(iii) holds, it trivially follows that $\mu_t^c = \Omega_{\boldsymbol{x}}$. Applying the change of variables $\boldsymbol{s} = \widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q})\boldsymbol{g}$ to (77) we obtain (20), which means that μ_t is the push-forward of μ_t^c under the map,

$$f : \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \\ \boldsymbol{g} \end{pmatrix} \mapsto \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \\ \widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q})\boldsymbol{g} \end{pmatrix},$$

  i.e.,

$$\mu_t(A) = f(\mu_t^c)(A) = \mu_t^c\left(f^{-1}(A)\right), \ A \in \mathcal{B}(\Omega_{\boldsymbol{x}}).$$

  Since f is a smooth one-to-one mapping, in particular surjective, and $\mathrm{supp}(\mu_t^c) = \Omega_{\boldsymbol{x}}$ we have

$$\mathrm{supp}(\mu_t) = \mathrm{supp}\left(f(\mu_t^c)\right) = \Omega_{\boldsymbol{x}}.$$

□

The following lemma, Lemma 8, shows that Novikov's condition is satisfied for the function u required for the application of Girsanov's theorem in the above proof of Lemma 7.

Lemma 8. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$ and $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ as in Lemma 7. Define*

$$\begin{aligned} \boldsymbol{u}_1(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g}) &= \left(\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q})\right)^{-1}\left(\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q})\boldsymbol{p} - \boldsymbol{p} - \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q})\boldsymbol{g} + \boldsymbol{g}\right) \\ &= \boldsymbol{G}(\boldsymbol{q})\begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{g} \end{pmatrix}, \end{aligned} \tag{80}$$

with

$$\boldsymbol{G}(\boldsymbol{q}) := \left(\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q})\right)^{-1}\left(\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{2,1}(\boldsymbol{q}) - \boldsymbol{I}_n \ -\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{2,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q}) + \boldsymbol{I}_n\right) \in \mathbb{R}^{n\times 2n},$$

and

$$\boldsymbol{u}_2(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g}) = -\left(\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}_2(\boldsymbol{q})\right)^{-1}\sum_{i=1}^{n}\boldsymbol{p}_i\left(\partial_{\boldsymbol{q}_i}\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q})\right)\boldsymbol{g}, \tag{81}$$

The function

$$\boldsymbol{u}(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g}) = \boldsymbol{u}_1(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g}) + \boldsymbol{u}_2(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g}) \tag{82}$$

satisfies Novikov's condition (101).

Proof (Proof of Lemma 8). Since

$$\|\boldsymbol{u}_1+\boldsymbol{u}_2\|_2^2 \le 2\|\boldsymbol{u}_1\|_2^2 + 2\|\boldsymbol{u}_2\|_2^2,$$

it is sufficient to show that Novikov's condition holds for $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. We only show the validity of Novikov's condition explicitly for $\boldsymbol{u}_1$. [5]

Since $\widetilde{\boldsymbol{\Gamma}}_{1,2}, \widetilde{\boldsymbol{\Gamma}}_{2,1}, \widetilde{\boldsymbol{\Gamma}}_{2,2}$ and $\widetilde{\boldsymbol{\Sigma}}_2$ are smooth functions of $\boldsymbol{q}$ and since $\Omega_{\boldsymbol{q}}$ is compact, the spectrum of $\boldsymbol{G}^T(\boldsymbol{q})\boldsymbol{G}(\boldsymbol{q})$ is uniformly bounded from above in $\boldsymbol{q}$, hence there is $\lambda_{\max} > 0$ such that

$$\lambda_{\max}^2(\|\boldsymbol{p}\|_2^2 + \|\boldsymbol{g}\|^2) \ge (\boldsymbol{p}^T,\boldsymbol{g}^T)\boldsymbol{G}^T(\boldsymbol{q})\boldsymbol{G}(\boldsymbol{q})\begin{pmatrix}\boldsymbol{p}\\ \boldsymbol{g}\end{pmatrix} = \|\boldsymbol{u}_1(\boldsymbol{q},\boldsymbol{p},\boldsymbol{g})\|^2, \tag{83}$$

and therefore

$$\mathbb{E}\left[\exp(\int_0^T \|\boldsymbol{u}_1(\boldsymbol{q}(t),\boldsymbol{p}(t),\boldsymbol{g}(t))\|\mathrm{d}t)\right] \le \mathbb{E}\left[\exp(\int_0^T \lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2)\mathrm{d}t)\right],$$

for any $T > 0$. Let $\epsilon < 2\tilde{\theta}/\lambda_{\max}^2$, with $\tilde{\theta} = \theta/\tilde{\lambda}_{\max}$ and $\theta > 0, \tilde{\lambda}_{\max}$ as defined in Lemma 9. We find

$$\begin{aligned}\exp(\int_0^T \lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2)\mathrm{d}t) &= \exp(\frac{1}{\epsilon}\int_0^T \epsilon\lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2)\mathrm{d}t)\\ &\le \frac{1}{\epsilon}\int_0^T \exp(\epsilon\lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2))\mathrm{d}t,\end{aligned}$$

by Jensen's inequality, thus

$$\begin{aligned}\mathbb{E}\left[\exp(\int_0^T \|\boldsymbol{u}_1(\boldsymbol{q}(t),\boldsymbol{p}(t),\boldsymbol{g}(t))\|\mathrm{d}t)\right] &\le \mathbb{E}\left[\frac{1}{\epsilon}\int_0^T \exp(\epsilon\lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2))\mathrm{d}t\right]\\ &= \frac{1}{\epsilon}\int_0^T \mathbb{E}\left[\exp(\epsilon\lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2))\right]\mathrm{d}t,\end{aligned}$$

[5] The respective proof for $\boldsymbol{u}_2$ is essentially the same with the only difference that in (83) we need to bound $\|\boldsymbol{u}_2\|_2^2$ by a term proportional to $\|\boldsymbol{p}\|_2^4 + \|\boldsymbol{g}\|_2^4$ instead of bounding u_2 by a term which is proportional to $\|\boldsymbol{p}\|_2^2 + \|\boldsymbol{g}\|_2^2$ as we do in the proof for $\boldsymbol{u}_1$. By choosing $l = 2$ in (84) the remaining steps of the proof are then exactly the same as for u_1.

by Tonelli's theorem. Let for $\alpha > 0$,

$$\mathcal{K}_\alpha := \mathcal{K}_{\alpha,l},\ l = 1, \tag{84}$$

with $\mathcal{K}_{\alpha,l}$ as defined in (86). Using

$$\exp(\epsilon\lambda_{\max}^2(\|\boldsymbol{p}\|^2 + \|\boldsymbol{g}\|^2)) \le \mathcal{K}_{\tilde{\theta}}(\boldsymbol{z}), \tag{85}$$

we conclude using Lemma 9, (87)

$$\begin{aligned}\frac{1}{\epsilon}\int_0^T \mathbb{E}\left[\exp(\epsilon\lambda_{\max}^2(\|\boldsymbol{p}(t)\|^2 + \|\boldsymbol{g}(t)\|^2))\right]\mathrm{d}t &\le \frac{1}{\epsilon}\int_0^T \mathbb{E}\left[\mathcal{K}_{\tilde{\theta}}(\boldsymbol{z}(t))\right]\mathrm{d}t \\ &\le \frac{1}{\epsilon}\int_0^T e^{-t}\mathcal{K}_\theta(\boldsymbol{p}_0, \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{g}_0) + b(1-e^{-t})\mathrm{d}t \\ &< \infty.\end{aligned}$$

with $b > 0$ as specified in Lemma 9. □

Lemma 9. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$ and $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Sigma}}$ as in Lemma 7 and let $\boldsymbol{C} \in \mathbb{R}^{2n\times 2n}$ with*

$$\min\sigma(\boldsymbol{C}) = 1,$$

be a symmetric positive definite matrix such that

$$\widetilde{\boldsymbol{\Gamma}}^T(\boldsymbol{q})\boldsymbol{C} + \boldsymbol{C}\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q}),$$

is positive definite for all $\boldsymbol{q} \in \Omega_{\boldsymbol{q}}$. For $\alpha > 0$ and $l \in \mathbb{N}$ define

$$\mathcal{K}_{\alpha,l}(\boldsymbol{p}, \boldsymbol{s}) = e^{\frac{\alpha}{2}\left(\boldsymbol{z}^T\boldsymbol{C}\boldsymbol{z}\right)^l}. \tag{86}$$

There exists $\theta > 0$ such that Assumption B.1 is satisfied with $\mathcal{K} = \mathcal{K}_{\theta,l}$ and $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}$. Moreover, for $\tilde{\theta} = \theta/\tilde{\lambda}_{\max}$ with

$$\tilde{\lambda}_{\max} := \max_{\boldsymbol{q}\in\Omega_{\boldsymbol{q}}}\left\{|\lambda| \mid \lambda \in \sigma\left(\widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q})\right)\right\}$$

the expectation of $\mathcal{K}_{\tilde{\theta},l}$ as function of the solution $(\boldsymbol{q}^c, \boldsymbol{p}^c, \boldsymbol{g}^c)$ of (77) can be bounded as

$$\mathbb{E}\left[\mathcal{K}_{\tilde{\theta},l}(\boldsymbol{p}^c, \boldsymbol{g}^c) \mid (\boldsymbol{p}^c(0), \boldsymbol{g}^c(0)) = (\boldsymbol{p}_0, \boldsymbol{g}_0)\right] \le e^{-t}\mathcal{K}_{\theta,l}(\boldsymbol{p}_0, \widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{g}_0) + b(1-e^{-t}) + c(l,t), \tag{87}$$

where $b > 0$ as above and $c(l,t)$ is a finite nonnegative constant which depends on l and t with $c(l,t) = 0$ for $l = 1$ and all $t \ge 0$.

Proof. We recall that the generator of (20) is of the form

$$\mathcal{L}_{\mathrm{GLE}} = \boldsymbol{F}(\boldsymbol{q})\cdot\nabla_{\boldsymbol{p}} + \boldsymbol{p}\cdot\nabla_{\boldsymbol{q}} - \widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})\boldsymbol{z}\cdot\nabla_{\boldsymbol{z}} + \frac{1}{2}\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q}) : \nabla_{\boldsymbol{p}}^2,$$

We show the result only for the case $l = 1$. For $l > 1$ the result follows by induction. Let $\mathcal{K}_\theta = \mathcal{K}_{\theta,1}$. Applying the generator on $\mathcal{K}_\theta$ we obtain

$$\begin{aligned}\mathcal{L}\mathcal{K}_\theta(\boldsymbol{p},\boldsymbol{s}) &= (\theta \boldsymbol{F}(\boldsymbol{q}) \cdot (\boldsymbol{C}_{1,1}\boldsymbol{p} + \boldsymbol{C}_{1,2}\boldsymbol{s}))\,\mathcal{K}_\theta(\boldsymbol{p},\boldsymbol{s}) \\ &\quad + \left(-\theta\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{q})\boldsymbol{z}\cdot\boldsymbol{C}\boldsymbol{z} + \frac{1}{2}\left(\theta \mathrm{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q})\boldsymbol{C}\right) + \theta^2 \boldsymbol{z}^T\boldsymbol{C}\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{q})\widetilde{\boldsymbol{\Sigma}}^T(\boldsymbol{q})\boldsymbol{C}\boldsymbol{z}\right)\right)\mathcal{K}_\theta(\boldsymbol{p},\boldsymbol{s}) \\ &= \left(-\Theta\left(\theta\|z\|^2\right) + \Theta\left((1+\theta)\|z\|\right) + \Theta\left(\theta^2\|z\|^2\right)\right)\mathcal{K}_\theta(\boldsymbol{p},\boldsymbol{s}) \\ &< -\mathcal{K}_\theta(\boldsymbol{p},\boldsymbol{s}) + b,\end{aligned}$$

for sufficiently small $\theta > 0$ and sufficiently large $b > 0$. Consequently, for $\tilde{\theta} = \theta/\tilde{\lambda}_{\max}$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\mathcal{K}_{\tilde{\theta}}(\boldsymbol{p}^c(t),\boldsymbol{g}^c(t)) \mid (\boldsymbol{p}^c(0),\boldsymbol{g}^c(0)) = (\boldsymbol{p}_0,\boldsymbol{g}_0)\right] \\ =&\,\mathbb{E}\left[\mathcal{K}_{\tilde{\theta}}(\boldsymbol{p}(t),\widetilde{\boldsymbol{\Gamma}}_{1,2}^{-1}(\boldsymbol{q}(t))\boldsymbol{s}(t)) \mid (\boldsymbol{p}(0),\boldsymbol{s}(0)) = (\boldsymbol{p}_0,\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{g}_0)\right] \\ \le&\,\mathbb{E}\left[\mathcal{K}_{\tilde{\theta}}(\tilde{\lambda}_{\max}\boldsymbol{p}(t),\tilde{\lambda}_{\max}\boldsymbol{s}(t)) \mid (\boldsymbol{p}(0),\boldsymbol{s}(0)) = (\boldsymbol{p}_0,\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{g}_0)\right] \\ =&\,\mathbb{E}\left[\mathcal{K}_\theta(\boldsymbol{p}(t),\boldsymbol{s}(t)) \mid (\boldsymbol{p}(0),\boldsymbol{s}(0)) = (\boldsymbol{p}_0,\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{g}_0)\right] \\ \le&\, e^{-t}\mathcal{K}_\theta(\boldsymbol{p}_0,\widetilde{\boldsymbol{\Gamma}}_{1,2}(\boldsymbol{q}_0)\boldsymbol{g}_0) + b(1-e^{-t}).\end{aligned}$$

□

The last Lemma 10 of this section provides conditions for the existence of suitable Lyapunov functions with polynomial growth for (20).

Lemma 10. *Let $\Omega_{\boldsymbol{q}} = \mathbb{T}^n$, $-\boldsymbol{\Gamma} \in \mathbb{R}^{(m+n)\times(n+m)}$ stable, and $U \in \mathcal{C}^\infty(\mathbb{T}^n,\mathbb{R})$. Moreover, assume that (50) holds and let $\boldsymbol{C}$ be as specified therein.*

$$\mathcal{K}_l(\boldsymbol{q},\boldsymbol{p},\boldsymbol{s}) = \left(\boldsymbol{z}^T\boldsymbol{C}\boldsymbol{z} + U(\boldsymbol{q}) - U_{min} + 1\right)^l,\ l \in \mathbb{N},$$

defines a family of Lyapunov functions for the differential operator $\mathcal{L}_{\mathrm{GLE}}$, i.e., for each $l \in \mathbb{N}$ there exist constants $a_l > 0$, $b_l \in \mathbb{R}$, such that for $\mathcal{L} = \mathcal{L}_{\mathrm{GLE}}, \mathcal{K} = \mathcal{K}_l$, Assumption B.1 holds for $a = a_l, b = b_l$.

Proof. The proof is very similar to the proof Lemma 1. The existence of a suitable matrix $\boldsymbol{C}$ as specified in (50) allows to extend all arguments in that proof with only some very small adaptations. For this reason we skip a details of the proof here. □

4 Conclusion

In this article we have presented an integrated perspective on ergodic properties of the generalized Langevin equation, for systems that can be written in the quasi-Markovian form. Although the GLE was well studied in the case of constant friction and damping and for conservative forces, our results indicate that these can often be extended to nonequilibrium models with non-gradient forces and non-constant friction and noise, thus providing a foundation for using GLEs in a much broader range of applications.

Acknowledgements. The authors wish to thank Greg Pavliotis (Imperial), Jonathan Mattingly (Duke) and Gabriel Stoltz (ENPC) for their generous assistance in providing comments at various stages of this project. In particular, the authors thank Jonathan Mattingly for pointing out the possibility of using Girsanov's theorem in the proof of Lemma 7. Both authors acknowledge the support of the European Research Council (Rule Project, grant no. 320823). BJL further acknowledges the support of the EPSRC (grant no. EP/P006175/1) during the preparation of this article. The work of MS was supported by the National Science Foundation under grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute.

A Auxiliary Material on Linear Algebra

The following Lemma A.1 is repeatedly used in the proofs of Proposition 3 and Lemma 3, as well as in Example 3 to show the positive (semi-)definiteness of symmetric matrices.

Lemma A.1. *Let A be a symmetric block structured matrix of the form*

$$\boldsymbol{A} := \begin{pmatrix} \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} \\ \boldsymbol{A}_{1,2}^T & \boldsymbol{A}_{2,2} \end{pmatrix} \in \mathbb{R}^{n+m \times n+m}$$

(i) If $\boldsymbol{A}_{2,2}$ is positive definite, then $\boldsymbol{A}$ is positive (semi-)definite if and only if

$$\boldsymbol{A}_{1,1} - \boldsymbol{A}_{1,2}\boldsymbol{A}_{2,2}^{-1}\boldsymbol{A}_{1,2}^T$$

is positive (semi-)definite

(ii) If $\boldsymbol{A}_{1,1}$ is positive definite, then $\boldsymbol{A}$ is positive (semi-)definite if and only if

$$\boldsymbol{A}_{2,2} - \boldsymbol{A}_{1,2}^T\boldsymbol{A}_{1,1}^{-1}\boldsymbol{A}_{1,2}$$

is positive (semi-)definite

(iii) Let $\boldsymbol{A}_{2,2}^g$ denote a generalised inverse of $\boldsymbol{A}_{2,2}$, i.e., $\boldsymbol{A}_{2,2}^g$ is a $m \times m$ matrix which satisfies

$$\boldsymbol{A}_{2,2}\boldsymbol{A}_{2,2}^g\boldsymbol{A}_{2,2} = \boldsymbol{A}_{2,2}.$$

The matrix $\boldsymbol{A}$ is positive semi-definite if and only if the matrices $\boldsymbol{A}_{2,2}$ and $\boldsymbol{A}_{1,1} - \boldsymbol{A}_{1,2}\boldsymbol{A}_{2,2}^g\boldsymbol{A}_{1,2}^T$ are positive semi-definite, and

$$(\boldsymbol{I} - \boldsymbol{A}_{2,2}\boldsymbol{A}_{2,2}^g)\boldsymbol{A}_{1,2}^T = \boldsymbol{0},$$

i.e., the span of the column vectors of $\boldsymbol{A}_{1,2}$ is contained in the span of the column vectors of $\boldsymbol{A}_{1,1}$.

Proof. The statements (i) and (ii) follow from Theorem 1.12 in [61]. Statement (iii) corresponds to Theorem 1.20 in the same reference. □

B Auxiliary Material on Stochastic Analysis

In this section we provide a brief overview of the general framework used in the ergodicity proofs and derivation of convergence rate in Sect. 3. For a comprehensive overview we refer to the review articles [32,36,50].

Consider an SDE defined on the domain $\Omega_{\boldsymbol{x}} = \mathbb{T}^{n_1} \times \mathbb{R}^{n_2}, n = n_1 + n_2 \in \mathbb{N}$ which is of the form

$$\mathrm{d}X = \boldsymbol{a}(X)\mathrm{d}t + \boldsymbol{b}(X)\mathrm{d}\,\boldsymbol{W},\ X(0) \sim \mu_0, \tag{88}$$

with smooth coefficients $\boldsymbol{a} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R}^n)$, $\boldsymbol{b} = [\boldsymbol{b}_i]_{1\leq i\leq n} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R}^{n\times n})$, and initial distribution μ_0. In order to simplify the presentation we further assume that the diffusion coefficient $\boldsymbol{b}$ is such that the Itô and Stratonovich interpretation of (88) coincide, i.e.,

$$\nabla \cdot \left(\boldsymbol{b}\,\boldsymbol{b}^T\right) - \boldsymbol{b}\,\nabla \cdot \boldsymbol{b}^T \equiv \boldsymbol{0}.$$

Let further $\mathcal{L}$ denote the associated infinitesimal generator of (88), i.e.,

$$\mathcal{L} = \boldsymbol{a}(X) \cdot \nabla + \boldsymbol{b}(X) : \nabla^2, \tag{89}$$

when considered as an operator on the core $\mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R})$, and let $\mathcal{L}^\dagger$ denote the formal adjoint of $\mathcal{L}$, i.e., the Fokker-Planck operator associated with the SDE (88). Furthermore, let $e^{t\mathcal{L}}, e^{t\mathcal{L}^\dagger}$ denote the associated semigroup operators of $\mathcal{L}$, and $\mathcal{L}^\dagger$, respectively, i.e.,[6]

$$\forall \varphi \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R}) : e^{t\mathcal{L}}\varphi(x) = \mathbb{E}[\varphi(X(t)) \mid X(0) = x], \tag{90}$$

for (Lebesgue-)almost all $x \in \mathbb{R}^n$, and

$$\int \left(e^{t\mathcal{L}}\varphi\right)(x)\mu_0(\mathrm{d}x) = \int \varphi(x)\left(e^{t\mathcal{L}^\dagger}\mu_0\right)(\mathrm{d}x).$$

Definition 1. *For a given function $\mathcal{K} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, [1,\infty))$ which is such that $\mathcal{K}(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$, define*

$$\|\varphi\|_{L^\infty_{\mathcal{K}}} := \left\|\frac{\varphi}{\mathcal{K}}\right\|_\infty,\ \varphi : \Omega_{\boldsymbol{x}} \to \mathbb{R}\ \textit{measureable}. \tag{91}$$

We denote by

$$L^\infty_{\mathcal{K}}(\Omega_{\boldsymbol{x}}) := \left\{\varphi\ \textit{measurable}\ : \|\varphi\|_{L^\infty_{\mathcal{K}}} < \infty\right\} \tag{92}$$

the set of measurable functions for which the ratio $\frac{\varphi}{\mathcal{K}}$ is bounded.

[6] The expectation is taken with respect to the Brownian motion $\boldsymbol{W}$.

It can be easily verified that $\|\varphi\|_{L^\infty_\mathcal{K}}$ defines a norm and that $L^\infty_\mathcal{K}(\Omega_{\boldsymbol{x}})$ equipped with the norm $\|\varphi\|_{L^\infty_\mathcal{K}}$ can be associated with a Banach space, which we denote by $\left(L^\infty_\mathcal{K}(\Omega_{\boldsymbol{x}}), \|\cdot\|_{L^\infty_\mathcal{K}}\right)$.

Throughout this article we use Lyapunov function techniques to show (geometric) ergodicity of SDEs of the generic form (88). More specifically, we follow the standard recipe for proofs of exponential convergences of the semigroup operator $e^{t\mathcal{L}}$ in weighted L^∞ spaces as outlined, e.g., in [32,36,38,50], that is we show that a suitable Lyapunov condition (Assumption B.1) and a minorization condition (Assumption B.2) are satisfied:

Assumption B.1 (Infinitesimal Lyapunov condition). *There is a function* $\mathcal{K} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, [1,\infty))$ *with* $\lim_{\|{\boldsymbol{x}}\|\to\infty} \mathcal{K}(x) = \infty$*, and real numbers* $a \in (0,\infty), b \in \mathbb{R}$ *such that,*

$$\mathcal{L}\mathcal{K} \leq -a\mathcal{K} + b. \tag{93}$$

Assumption B.2 (Minorization condition). *For some* $t' > 0$ *there exists a constant* $\eta \in (0,1)$ *and a probability measure* ν *such that*

$$\inf_{x\in\mathcal{C}} e^{t'\mathcal{L}^\dagger}\delta_x(\mathrm{d}y) \geq \eta\nu(\mathrm{d}y)$$

where $\mathcal{C} = \{x \in \Omega_{\boldsymbol{x}} :\ \mathcal{K}(x) \leq \mathcal{K}_{\max}\}$ *for some* $\mathcal{K}_{\max} > 1 + 2b/a$*, where* a, b *are the same constants as in* (93).

If the above assumptions are satisfied, then the following proposition, which follows from the arguments in [32] (see also the other above mentioned references), allows to derive exponential decay estimates in the respective weighted L^∞ space associated with the Lyapunov function $\mathcal{K}$.

Proposition B.1 (Geometric ergodicity, [32]). *Let Assumptions B.1 and B.2 hold. The solution of the SDE* (88) *admits a unique invariant probability measure* π *such that*

(i) there exist positive constant $\lambda, \widetilde{C}$ *so that for any* $\varphi \in L^\infty_\mathcal{K}(\Omega_{\boldsymbol{x}})$

$$\left\|e^{t\mathcal{L}}\varphi - \mathbb{E}_\pi\varphi\right\|_{L^\infty_\mathcal{K}} \leq \widetilde{C}e^{-t\lambda}\left\|\varphi - \mathbb{E}_\pi\varphi\right\|_{L^\infty_\mathcal{K}}. \tag{94}$$

(ii)

$$\int_{\Omega_{\boldsymbol{x}}} \mathcal{K}\mathrm{d}\pi < \infty. \tag{95}$$

If for the solution of (88) the implications of Proposition B.1 hold we also say that the solution X of (88) is *geometrically ergodic*. In the main body of this article we use Proposition B.1 to derive exponential decay estimates of the form (46) in Theorems 1 to 4. In these theorems Assumption B.1 can be directly shown to hold by explicitly constructing a suitable Lyapunov function $\mathcal{K}$ satisfying (93) (see Lemmas 1, 3 and 10). A very common way to show Assumption B.2 is by showing (i) that the transition kernel associated with the SDE (88) is smooth as specified in Assumption B.3, and (ii) that the SDE (88) is controllable as specified in Assumption B.4. By virtue Lemma B.1 it then follows that a minorization condition holds.

Assumption B.3. *For any $t > 0$ the transition kernel associated with the SDE (88) possesses a density $p_t(x, y)$, i.e.,*

$$\forall\, x \in \Omega_{\boldsymbol{x}} :\ (e^{t\mathcal{L}^\dagger}\delta_x)(A) = \int_A p_t(x, y)dy,\ \ A \subset \Omega_{\boldsymbol{x}},\ A \text{ measurable.}$$

and $p_t(x, y)$ is jointly continuous in $(x, y) \in \Omega_{\boldsymbol{x}} \times \Omega_{\boldsymbol{x}}$.

Assumption B.4. *There is a $t_{\max} > 0$ so that for any $x^-, x^+ \in \Omega_{\boldsymbol{x}}$, there is a $t > 0$, with $t \le t_{\max}$, so that the control problem*

$$\dot{\tilde{X}} = \boldsymbol{a}(\tilde{X}) + \boldsymbol{b}(\tilde{X})u, \tag{96}$$

subject to

$$\tilde{X}(0) = x^-,\ \text{and}\ \tilde{X}(t) = x^+,$$

has a smooth solution $u \in \mathcal{C}^1([0, t_{\max}], \Omega_{\boldsymbol{x}})$.

Lemma B.1 ([36]). *If Assumptions B.3 and B.4 are satisfied, then also Assumption B.2 holds.*

Assumption B.3 follows directly from hypoellipticity of the operator $\partial_t - \mathcal{L}^\dagger$ (see e.g. [47,50], for a precise definition of hypoellipticity). A common way to establish hypoellipticity of a differential operators is via Hörmander's theorem ([20], Theorem 22.2.1, on p. 353). The following proposition is an adaption of Hörmander's theorem to the parabolic differential operator $\partial_t - \mathcal{L}^\dagger$:

Proposition B.2. *Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be the drift coefficient and the diffusion coefficient of the SDE (88), respectively. Let $\boldsymbol{b}_0 := \boldsymbol{a}$. Iteratively define a collection of vector fields by*

$$\mathscr{V}_0 = \{\boldsymbol{b}_i : i \ge 1\},\quad \mathscr{V}_{k+1} = \mathscr{V}_k \cup \{[\boldsymbol{v}, \boldsymbol{b}_i] : \boldsymbol{v} \in \mathscr{V}_k, 0 \le i \le n\}. \tag{97}$$

where

$$[\boldsymbol{X}, \boldsymbol{Y}] = (\nabla \boldsymbol{Y})\boldsymbol{X} - (\nabla \boldsymbol{X})\boldsymbol{Y},$$

denotes the commutator of vector fields $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{C}^\infty(\Omega_{\boldsymbol{x}}, \mathbb{R}^n)$ and $(\nabla \boldsymbol{X}), (\nabla \boldsymbol{Y})$ their Jacobian matrices. If

$$\forall \boldsymbol{x} \in \mathbb{R}^n,\quad \operatorname{lin}\left\{\boldsymbol{v}(\boldsymbol{x}) : \boldsymbol{v} \in \bigcup_{k\in\mathbb{N}} \mathscr{V}_k\right\} = \mathbb{R}^n, \tag{98}$$

we say that the SDE (88) satisfies the parabolic Hörmander condition, and it follows that the operator $\partial_t - \mathcal{L}^\dagger$ is hypoelliptic.

We use Lemma B.1 in the proof of Lemma 4 in Theorem 2. For some instances of (20) it is not easy to construct a suitable control u such that Assumption B.4 is satisfied. In these cases we either show a minorization condition by explicitly constructing the minorizing measure ν in Assumption B.2 if the right hand side of (20) can be decomposed into a linear and a bounded part (see Theorem 1), or

by inferring the existence of a suitable minorizing measure by showing that the support of the SDE under consideration is equivalent to the support of another SDE satisfying a minorization condition via Girsanov's theorem (Lemmas 5 and 7). Girsanov's theorem provides conditions under which the path measures of two Itô processes are mutually absolutely continuous, which in particular implies that at any time $t \geq 0$ the laws of these Itô processes are equivalent. We will use Girsanov's theorem in Sect. 3 in order to prove the minorization condition for GLEs which in a Markovian representation possess coefficients which depend on the configurational variable. Here we provide a version of Girsanov's theorem which is adapted to Itô-diffusion processes.

Proposition B.3 (Girsanov's theorem, [45]). *Consider the two Itô diffusion processes*

$$\mathrm{d}X(t) = \boldsymbol{a}_x(X)\mathrm{d}t + \boldsymbol{b}(X)\mathrm{d}\,\boldsymbol{W}(t);\ X(0) = x_0, \tag{99}$$

$$\mathrm{d}Y(t) = \boldsymbol{a}_y(Y)\mathrm{d}t + \boldsymbol{b}(Y)\mathrm{d}\,\boldsymbol{W}(t);\ Y(0) = x_0, \tag{100}$$

where $x_0 \in \Omega_{\boldsymbol{x}}$, $\boldsymbol{W}$ is a standard Wiener process in $\mathbb{R}^n$, and $\boldsymbol{a}_x, \boldsymbol{a}_y : \Omega_{\boldsymbol{x}} \to \mathbb{R}^n$ and $\boldsymbol{b} : \Omega_{\boldsymbol{x}} \to \mathbb{R}^{n\times m}, m \in \mathbb{N}$, are such that there exist unique strong solutions X, Y for (99) and (100), respectively. If there is a function $\boldsymbol{u} \in \mathcal{C}(\Omega_{\boldsymbol{x}}, \mathbb{R}^n)$ such that

$$\boldsymbol{a}_x - \boldsymbol{a}_y = \boldsymbol{b}\boldsymbol{u}$$

and $\boldsymbol{u}$ satisfies Novikov's *condition*

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \|\boldsymbol{u}(X(t))\|_2^2 \mathrm{d}t\right)\right] < \infty. \tag{101}$$

then the path measures of X and Y on any finite time interval are equivalent. In particular, the support of the law of $X(t)$ and the support of the law of $Y(t)$ coincide for any $t > 0$.

References

1. Adelman, S., Doll, J.: Generalized Langevin equation approach for atom/solid-surface scattering: general formulation for classical scattering off harmonic solids. J. Chem. Phys. **64**(6), 2375–2388 (1976)
2. Bhattacharya, R.N.: On the functional central limit theorem and the law of the iterated logarithm for Markov processes. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **60**(2), 185–201 (1982)
3. Carmona, P.: Existence and uniqueness of an invariant measure for a chain of oscillators in contact with two heat baths. Stochast. Process. Appl. **117**(8), 1076–1092 (2007)
4. Ceriotti, M.: GLE4MD. http://gle4md.org
5. Ceriotti, M., Bussi, G., Parrinello, M.: Langevin equation with colored noise for constant-temperature molecular dynamics simulations. Phys. Rev. Lett. **102**(2), 020601 (2009)

6. Ceriotti, M., Bussi, G., Parrinello, M.: Colored-noise thermostats à la carte. J. Chem. Theory Comput. **6**(4), 1170–1180 (2010)
7. Darve, E., Solomon, J., Kia, A.: Computing generalized Langevin equations and generalized Fokker-Planck equations. Proc. Nat. Acad. Sci. **106**(27), 10884–10889 (2009)
8. Doll, J.D., Dion, D.R.: Generalized Langevin equation approach for atom/solid-surface scattering: numerical techniques for Gaussian generalized Langevin dynamics. J. Chem. Phys. **65**(9), 3762–3766 (1976)
9. Dym, H., McKean, H.P.: Gaussian Processes, Function Theory, and the Inverse Spectral Problem. Courier Corporation (2008)
10. Eckmann, J.-P., Hairer, M.: Non-equilibrium statistical mechanics of strongly anharmonic chains of oscillators. Commun. Math. Phys. **212**(1), 105–164 (2000)
11. Eckmann, J.-P., Pillet, C.-A., Rey-Bellet, L.: Entropy production in nonlinear, thermally driven Hamiltonian systems. J. Stat. Phys. **95**(1), 305–331 (1999)
12. Eckmann, J.-P., Pillet, C.-A., Rey-Bellet, L.: Non-equilibrium statistical mechanics of anharmonic chains coupled to two heat baths at different temperatures. Commun. Math. Phys. **201**(3), 657–697 (1999)
13. Ford, G., Kac, M., Mazur, P.: Statistical mechanics of assemblies of coupled oscillators. J. Math. Phys. **6**(4), 504–515 (1965)
14. Givon, D., Kupferman, R., Hald, O.H.: Existence proof for orthogonal dynamics and the Mori-Zwanzig formalism. Isr. J. Math. **145**(1), 221–241 (2005)
15. Givon, D., Kupferman, R., Stuart, A.: Extracting macroscopic dynamics: model problems and algorithms. Nonlinearity **17**(6), R55–R127 (2004)
16. Hairer, M., Mattingly, J.C.: Yet another look at Harris Ergodic theorem for Markov chains. In: Seminar on Stochastic Analysis, Random Fields and Applications VI, vol. 63, pp. 109–117. Springer, Heidelberg (2011)
17. Hänggi, P.: Generalized Langevin equations: a useful tool for the perplexed modeller of nonequilibrium fluctuations? In: Lutz, S.-G., Thorsten, P. (eds.) Stochastic Dynamics, pp. 15–22. Springer, Heidelberg (1997)
18. Harris, T.E.: The existence of stationary measures for certain Markov processes. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, pp. 113–124 (1956)
19. Hohenegger, C., McKinley, S.A.: Fluid-particle dynamics for passive tracers advected by a thermally fluctuating viscoelastic medium. J. Comput. Phys. **340**, 688–711 (2017)
20. Hörmander, L.: The analysis of linear partial differential operators III. Grundlehren der Mathematischen Wissenschaften [fundamental principles of mathematical sciences], vol. 274 (1985)
21. Jakšić, V., Pillet, C.-A.: Ergodic properties of the non-Markovian Langevin equation. Lett. Math. Phys. **41**(1), 49–57 (1997)
22. Jakšić, V., Pillet, C.-A.: Spectral theory of thermal relaxation. J. Math. Phys. **38**(4), 1757–1780 (1997)
23. Jakšić, V., Pillet, C.-A.: Ergodic properties of classical dissipative systems I. Acta Math. **181**(2), 245–282 (1998)
24. Joubaud, R., Pavliotis, G., Stoltz, G.: Langevin dynamics with space-time periodic nonequilibrium forcing. J. Stat. Phys. **158**(1), 1–36 (2015)
25. Kantorovich, L.: Generalized Langevin equation for solids. I. Rigorous derivation and main properties. Phys. Rev. B **78**(9), 094304 (2008)
26. Kliemann, W.: Recurrence and invariant measures for degenerate diffusions. Ann. Probab. **15**, 690–707 (1987)

27. Kupferman, R.: Fractional kinetics in Kac-Zwanzig heat bath models. J. Stat. Phys. **114**(1), 291–326 (2004)
28. Kupferman, R., Stuart, A., Terry, J., Tupper, P.: Long-term behaviour of large mechanical systems with random initial data. Stoch. Dyn. **2**(4), 533–562 (2002)
29. Lampo, T.J., Kuwada, N.J., Wiggins, P.A., Spakowitz, A.J.: Physical modeling of chromosome segregation in *Escherichia coli* reveals impact of force and DNA relaxation. Biophys. J. **108**(1), 146–153 (2015)
30. Lei, H., Baker, N.A., Li, X.: Data-driven parameterization of the generalized Langevin equation. Proc. Nat. Acad. Sci. **113**(50), 14183–14188 (2016)
31. Leimkuhler, B., Matthews, C., Stoltz, G.: The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. IMA J. Numer. Anal. **36**(1), 13–79 (2016)
32. Lelièvre, T., Stoltz, G.: Partial differential equations and stochastic methods in molecular dynamics. Acta Numer. **25**, 681–880 (2016)
33. Li, Z., Bian, X., Li, X., Karniadakis, G.E.: Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. J. Chem. Phys. **143**(24), 243128 (2015)
34. Li, Z., Lee, H.S., Darve, E., Karniadakis, G.E.: Computing the non-Markovian coarse-grained interactions derived from the Mori-Zwanzig formalism in molecular systems: application to polymer melts. J. Chem. Phys. **146**(1), 014104 (2017)
35. Lim, S.H., Wehr, J.: Homogenization for a class of generalized Langevin equations with an application to thermophoresis. J. Stat. Phys. **174**, 656–691 (2017)
36. Mattingly, J.C., Stuart, A.M., Higham, D.J.: Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. Stoch. Process. Appl. **101**(2), 185–232 (2002)
37. Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes I: criteria for discrete-time chains. Adv. Appl. Probab. **24**(3), 542–574 (1992)
38. Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes II: continuous-time processes and sampled chains. Adv. Appl. Probab. **25**(3), 487–517 (1993)
39. Meyn, S.P., Tweedie, R.L.: Markov Chains and Stochastic Stability. Springer Science & Business Media, Heidelberg (2012)
40. Mori, H.: A continued-fraction representation of the time-correlation functions. Prog. Theor. Phys. **34**(3), 399–416 (1965)
41. Morriss, G.P., Evans, D.J.: Statistical Mechanics of Nonequilbrium Liquids. ANU Press, Canberra (2013)
42. Morrone, J.A., Markland, T.E., Ceriotti, M., Berne, B.: Efficient multiple time scale molecular dynamics: using colored noise thermostats to stabilize resonances. J. Chem. Phys. **134**(1), 014103 (2011)
43. Ness, H., Stella, L., Lorenz, C., Kantorovich, L.: Applications of the generalized Langevin equation: towards a realistic description of the baths. Phys. Rev. B **91**(1), 014301 (2015)
44. Ness, H., Genina, A., Stella, L., Lorenz, C.D., Kantorovich, L.: Nonequilibrium processes from generalized Langevin equations: realistic nanoscale systems connected to two thermal baths. Phys. Rev. B **93**(17), 174303 (2016)
45. Øksendal, B.: Stochastic differential equations. In: Stochastic Differential Equations, pp. 65–84. Springer, Heidelberg (2003)
46. Ottobre, M., Pavliotis, G.: Asymptotic analysis for the generalized Langevin equation. Nonlinearity **24**(5), 1629–1653 (2011)
47. Pavliotis, G.A.: Stochastic Processes and Applications. Springer, Heidelberg (2016)
48. Redon, S., Stoltz, G., Trstanova, Z.: Error analysis of modified Langevin dynamics. J. Stat. Phys. **164**(4), 735–771 (2016)

49. Rey-Bellet, L.: Statistical mechanics of anharmonic lattices. In: Advances in Differential Equations and Mathematical Physics (Birmingham, AL, 2002), vol. 327, pp. 283–293 (2003)
50. Rey-Bellet, L.: Ergodic properties of Markov processes. In: Open Quantum Systems II. Lecture Notes in Mathematics, vol. 1881, pp. 1–39. Springer, Heidelberg (2006)
51. Rey-Bellet, L.: Open classical systems. In: Open Quantum Systems II. Lecture Notes in Mathematics, vol. 1881, pp. 41–78. Springer, Heidelberg (2006)
52. Rey-Bellet, L., Thomas, L.E.: Exponential convergence to non-equilibrium stationary states in classical statistical mechanics. Commun. Math. Phys. **225**(2), 305–329 (2002)
53. Rudin, W.: Fourier Analysis on Groups. Courier Dover Publications, New York (2017)
54. Sachs, M., Leimkuhler, B., Danos, V.: Langevin dynamics with variable coefficients and nonconservative forces: from stationary states to numerical methods. Entropy **19**(12), 647 (2017)
55. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media, Heidelberg (2012)
56. Stella, L., Lorenz, C., Kantorovich, L.: Generalized Langevin equation: an efficient approach to nonequilibrium molecular dynamics of open systems. Phys. Rev. B **89**(13), 134303 (2014)
57. Stroock, D.W., Varadhan, S.R.: On the support of diffusion processes with applications to the strong maximum principle. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (University of California, Berkeley, California, 1970/1971), vol. 3, pp. 333–359 (1972)
58. Talay, D.: Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. Markov Process. Related Fields **8**(2), 163–198 (2002)
59. Villani, C.: Hypocoercivity. American Mathematical Society, Providence (2009)
60. Wu, X., Brooks, B.R., Vanden-Eijnden, E.: Self-guided Langevin dynamics via generalized Langevin equation. J. Comput. Chem. **37**(6), 595–601 (2016)
61. Zhang, F. (ed.): The Schur Complement and Its Applications. Numerical Methods and Algorithms, vol. 4. Springer Science & Business Media, Heidelberg (2006)
62. Zwanzig, R.: Memory effects in irreversible thermodynamics. Phys. Rev. **124**(4), 983–992 (1961)
63. Zwanzig, R.: Nonlinear generalized Langevin equations. J. Stat. Phys. **9**(3), 215–220 (1973)

Exit Event from a Metastable State and Eyring-Kramers Law for the Overdamped Langevin Dynamics

Tony Lelièvre[1], Dorian Le Peutrec[2], and Boris Nectoux[1(✉)]

[1] École des Ponts, Université Paris-Est, Inria, 77455 Champs-sur-Marne, France
{tony.lelievre,boris.nectoux}@enpc.fr, boris.nectoux@asc.tuwien.ac.at
[2] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France
dorian.lepeutrec@math.u-psud.fr

Abstract. In molecular dynamics, several algorithms have been designed over the past few years to accelerate the sampling of the exit event from a metastable domain, that is to say the time spent and the exit point from the domain. Some of them are based on the fact that the exit event from a metastable region is well approximated by a Markov jump process. In this work, we present recent results on the exit event from a metastable region for the overdamped Langevin dynamics obtained in [22,23,56]. These results aim in particular at justifying the use of a Markov jump process parametrized by the Eyring-Kramers law to model the exit event from a metastable region.

Keywords: Exit event · Metastability · Eyring-Kramers · Overdamped Langevin

The objective of this note is to give motivations (Sect. 1) and outlines of the proofs (Sect. 2) of results recently obtained in [22,23,56]. These results justify the use of the Eyring-Kramers formulas together with a kinetic Monte Carlo model to model the exit event from a metastable state for the overdamped Langevin dynamics. Such results are particularly useful to justify algorithms and models which use such formulas to build reduced description of the overdamped Langevin dynamics.

1 Exit Event from a Metastable Domain and Markov Jump Process

1.1 Overdamped Langevin Dynamics and Metastability

Let $(X_t)_{t\geq 0}$ be the stochastic process solution to the overdamped Langevin dynamics in $\mathbb{R}^d$:

$$dX_t = -\nabla f(X_t)dt + \sqrt{h}\, dB_t, \tag{1}$$

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 331–363, 2019.
https://doi.org/10.1007/978-3-030-15096-9_9

where $f \in C^\infty(\mathbb{R}^d, \mathbb{R})$ is the potential function, $h > 0$ is the temperature and $(B_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. The overdamped Langevin dynamics can be used for instance to describe the motion of the atoms of a molecule or the diffusion of impurities in a crystal (see for instance [51, Sections 2 and 3] or [10]). The term $-\nabla f(X_t)$ in (1) sends the process towards local minima of f, while thanks to the noise term $\sqrt{h}\, dB_t$, the process X_t may jump from one basin of attraction of the dynamics $\dot{x} = -\nabla f(x)$ to another one. If the temperature is small (i.e. $h \ll 1$), the process $(X_t)_{t\geq 0}$ remains during a very long period of time trapped around a neighborhood of a local minimum of f, called a metastable state, before going to another region. For that reason, the process (1) is said to be metastable. More precisely, a domain $\Omega \subset \mathbb{R}^d$ is said to be metastable for the probability measure μ supported in Ω if, when $X_0 \sim \mu$, the process (1) reaches a local equilibrium in Ω long before escaping from it. This will be made more precise below using the notion of quasi-stationary distribution (see Sect. 1.5). The move from one metastable region to another is typically related to a macroscopic change of configuration of the system. Metastability implies a separation of timescales which is one of the major issues when trying to have access to the macroscopic evolution of the system using simulations made at the microscopic level. Indeed, in practice, many transitions cannot be observed by integrating directly the trajectories of the process (1). To overcome this difficulty, some algorithms use the fact that the exit event from a metastable region can be well approximated by a Markov jump process with transition rates computed with the Eyring-Kramers formula, see for example the Temperature Accelerated Dynamics method [61] that will be described below.

1.2 Markov Jump Process and Eyring-Kramers Law

Kinetic Monte Carlo Methods. Let $\Omega \subset \mathbb{R}^d$ be a domain of the configuration space and let us assume that the process (1) is initially distributed according to the probability measure μ (i.e. $X_0 \sim \mu$) which is supported in Ω and for which the exit event from Ω is metastable. Let us denote by $(\Omega_i)_{i=1,\ldots,n}$ the surrounding domains of Ω (see Fig. 1), each of them corresponding to a macroscopic state of the system. Many reduced models and algorithms rely on the fact that the exit event from Ω, i.e. the next visited state by the process (1) among the Ω_i's as well as the time spent by the process (1) in Ω, is efficiently approximated by a Markov jump process using kinetic Monte Carlo methods [8,25,59,60,66,67]. Kinetic Monte Carlo methods simulate a Markov jump process in a discrete state space. To use a kinetic Monte Carlo algorithm in order to sample the exit event from Ω, one needs for $i \in \{1, \ldots, n\}$ the transition rate k_i to go from the state Ω to the state Ω_i. A kinetic Monte Carlo algorithm generates the next visited state Y among the Ω_i's and the time T spent in Ω for the process (1) as follows:

1. First sample T as an exponential random variable with parameter $\sum_{i=1}^n k_i$, i.e.:

$$T \sim \mathcal{E}\Big(\sum_{i=1}^n k_i\Big). \tag{2}$$

2. Then, sample the next visited state Y independently from T, i.e

$$Y \perp\!\!\!\perp T \tag{3}$$

using the following law: for all $i \in \{1, \dots, n\}$,

$$\mathbb{P}\big[Y = i\big] \;=\; \frac{k_i}{\sum_{\ell=1}^{n} k_\ell}. \tag{4}$$

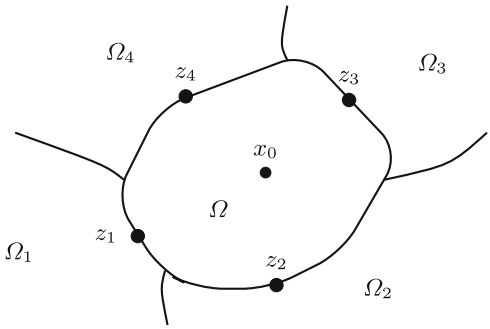


Fig. 1. Representation of the domain Ω, the surrounding domains $(\Omega_i)_{i=1,\dots,4}$ of Ω, the global minimum x_0 of f in Ω and $\{z_i\} = \operatorname{argmin}_{\partial\Omega\cap\Omega_i} f$ $(i \in \{1,2,3,4\})$.

Remark 1. Let us give an equivalent way to sample T and Y in a Monte Carlo method. Let $(\tau_i)_{i\in\{1,\dots,n\}}$ be n independent random variables such that for all $i \in \{1, \dots, n\}$, τ_i is exponentially distributed with parameter k_i. Then, the couple (T, Y) has the same law as $(\min_{j\in\{1,\dots,n\}} \tau_j, \operatorname{argmin}_{j\in\{1,\dots,n\}} \tau_j)$.

Eyring-Kramers Law. In practice, the transition rates $(k_i)_{i\in\{1,\dots,n\}}$ are computed using the Eyring-Kramers formula [29,66]:

$$k_i = A_i \, e^{-\frac{2}{h}(f(z_i)-f(x_0))}, \tag{5}$$

where $x_0 \in \Omega$ is the unique global minimum of f in $\overline{\Omega}$ and $\{z_i\} = \operatorname{argmin}_{\partial\Omega\cap\partial\Omega_i} f$, see Fig. 1. We here assume for simplicity that the minimum is attained at one single point z_i but the results below can be generalized to more general settings. If Ω is the basin of attraction of x_0 for the dynamics $\dot{x} = -\nabla f(x)$ so that z_i is a saddle point of f (i.e. a critical point of index 1), then, for the overdamped Langevin dynamics (1), the prefactor A_i writes:

$$A_i = \frac{|\lambda(z_i)|}{2\pi} \frac{\sqrt{\det \operatorname{Hess} f(x_0)}}{\sqrt{|\det \operatorname{Hess} f(z_i)|}}, \tag{6}$$

where $\lambda(z_i)$ is the negative eigenvalue of the Hessian matrix of f at z_i. Notice that the formula (6) requires that x_0 and z_i are non degenerate critical points of f. The formulas (5) and (6) have been first obtained in the small temperature regime by Kramers [42] (see the review of the literature [29]).

Remark 2. In the Physics literature, the approximation of the macroscopic evolution of the system with a Markov jump process with transition rates computed with the Eyring-Kramers formula (5)–(6) is sometimes called the Harmonic Transition State Theory [47,63].

1.3 The *Temperature Accelerated Dynamics* algorithm

The temperature accelerated dynamics (TAD) algorithm proposed by Sørensen and Voter [61] aims at efficiently approximating the exit event from a metastable domain for the dynamics (1) in order to have access to the macroscopic evolution of the system. We also refer to [1] for a mathematical analysis of this algorithm in a one-dimensional setting.

The basic idea of the TAD algorithm is the following: the exit time from the metastable domain Ω increases exponentially with the inverse of the temperature, see indeed (2)–(5). The idea is then to simulate the process at higher temperature to accelerate the simulation of the exit event. Let us assume that the process $(X_t)_{t\geq 0}$, evolving at the temperature h_{low} is at some time $t_0 \geq 0$ in the domain $\Omega \subset \mathbb{R}^d$ which is metastable for the initial condition $X_{t_0} \in \Omega$. Following [61], let us assume that the process instantaneously reaches the local equilibrium in Ω, i.e. that X_{t_0} is distributed according to this local equilibrium. The existence and the uniqueness of the local equilibrium in Ω as well as the convergence toward this local equilibrium is made more precise in Sect. 1.5 using the notion of quasi-stationary distribution. To ensure the convergence towards the local equilibrium in Ω, a decorrelation step may be used before running the TAD algorithm, see step (M1) in [1, Section 2.2].

As in the previous section, one denotes by $(\Omega_i)_{i=1,\ldots,n}$ the surrounding domains of Ω (see Fig. 1), each of them corresponding to a macroscopic state of the system and, for $i \in \{1,\ldots,n\}$, $\{z_i\} = \operatorname{argmin}_{\partial\Omega\cap\partial\Omega_i} f$. To sample the next visited state among the Ω_i's as well as the time T spent in Ω for the process (1), the TAD algorithm proceeds as follows. Let us introduce $T_{sim} = 0$ (which is the simulation time) and $T_{stop} = +\infty$ (which is the stopping time), and iterate the following steps.

1. Let $(Y_t)_{t\geq T_{sim}}$ be the solution to the evolution equation (1) but for the temperature $h_{high} > h_{low}$, starting from the local equilibrium in Ω at temperature h_{high}. Let $(Y_t)_{t\geq T_{sim}}$ evolve until it leaves Ω and denote by

$$T_{sim} + \tau$$

the first exit time from Ω for the process $(Y_t)_{t\geq T_{sim}}$. Let $j \in \{1,\ldots,n\}$ be such that $Y_{T_{sim}+\tau} \in \partial\Omega_j \cap \partial\Omega$. Then, set $T_{sim} = T_{sim} + \tau$. If it is the first time an exit from Ω through z_j for the process $(Y_t)_{t\geq 0}$ is observed (else one goes directly to the next step), set $\tau_j(h_{high}) = T_{sim}$ and extrapolate the time to $\tau_j(h_{low})$ with the formula

$$\tau_j(h_{low}) = \tau_j(h_{high})\, e^{2\left(\frac{1}{h_{low}} - \frac{1}{h_{high}}\right)(f(z_j)-f(x_0))}, \tag{7}$$

where we recall $x_0 \in \Omega$ is the unique global minimum of f in $\overline{\Omega}$. Then, update the minimum exit time $\tau_{min}(h_{low})$ among the $\tau_j(h_{low})$'s which have been observed so far. Finally, compute a new time T_{stop} so that there is a very small probability (say $\alpha \ll 1$) to observe an exit event from Ω at the temperature h_{high} which, using (7), would change the value of $\tau_{min}(h_{low})$. We refer to [61] or [1] for the computation of T_{stop}.

2. If $T_{sim} \leq T_{stop}$ then go back to the first step starting from the local equilibrium in Ω at time T_{sim}, else go to the next step.
3. Set $T = \tau_{min}(h_{low})$ and $Y = \ell$ where ℓ is such that $\tau_\ell(h_{low}) = \tau_{min}(h_{low})$. Finally, send X_{t_0+T} to Ω_ℓ and evolve the process (1) with the new initial condition X_{t_0+T}.

Remark 3. In [61], when the process $(Y_t)_{t \geq T_{sim}}$ leaves Ω, it is reflected back in Ω and it is then assumed that it reaches instantaneously the local equilibrium in Ω at temperature h_{high}.

Remark 4. One can use a decorrelation step before running the TAD algorithm and the sampling of $Y_{T_{sim}}$ according to the local equilibrium in Ω at the beginning of the step 1 to ensure that the underlying Markov jump process is justified, see [1].

The extrapolation formula (7) which is at the heart of the TAD algorithm relies on the properties of the underlying Markov jump process used to accelerate the exit event from a metastable state and where transition times are exponentially distributed with parameters computed with the Eyring-Kramers formula, see Remark 1 and Eq. (5). In the algorithm TAD, it is indeed assumed that the exit event from Ω can be modeled with a kinetic Monte Carlo method where the transition rates are computed with the Eyring-Kramers law (5)–(6). Then, at high temperature, one checks that under this assumption, each $\tau_i(h_{high})$ ($i \in \{1, \ldots, n\}$) is an exponential law of parameter $A_i\, e^{-\frac{2}{h_{high}}(f(z_i)-f(x_0))}$ (see Remark 1). The formula (7) allows to construct for all $i \in \{1, \ldots, n\}$, an exit time $\tau_i(h_{low})$ which is an exponential law of parameter $A_i\, e^{-\frac{2}{h_{low}}(f(z_i)-f(x_0))}$. By considering the couple $(\min_{i\in\{1,\ldots,n\}} \tau_i(h_{low}), \operatorname{argmin}_{i\in\{1,\ldots,n\}} \tau_i(h_{low}))$, one has access to the exit event from Ω (see Remark 1).

Remark 5. There are other algorithms which use the properties of the underlying Markov jump process to accelerate the simulation of the exit event from a metastable state, see for instance [64,65].

Our objective is to justify rigorously that a Markov jump process with transition rates computed with the Eyring-Kramers formula (5) can be used to model the exit event from a metastable domain Ω for the overdamped Langevin process (1). Before, let us recall mathematical contributions on the exit event from a domain and on the Eyring-Kramers formula (5).

1.4 Mathematical Literature on the Exit Event from a Domain and on the Eyring-Kramers Formulas

In the mathematical literature, there are mainly two approaches to the study of the asymptotic behaviour of the exit event from a domain when $h \to 0$: the global approaches and the local approaches.

Global Approaches. The global approaches study the asymptotic behaviours in the limit $h \to 0$ of the eigenvalues of the infinitesimal generator

$$L_{f,h}^{(0)} = -\frac{h}{2}\Delta + \nabla f \cdot \nabla \tag{8}$$

of the diffusion (1) on $\mathbb{R}^d$. Let us give for example a result obtained in [6,7]. To this end, let us assume that the potential $f : \mathbb{R}^d \to \mathbb{R}$ is a Morse function, has m local minima $\{x_1, \ldots, x_m\}$ and that for h small enough $\int_{\mathbb{R}^d} e^{-\frac{2}{h}f} < +\infty$. Let us recall that $\phi : \mathbb{R}^d \to \mathbb{R}$ is a Morse function if all its critical points are non degenerate. For a Morse function $\phi : \mathbb{R}^d \to \mathbb{R}$, we say that x is a saddle point of ϕ if x is a critical point of ϕ such that the Hessian matrix of ϕ at x has exactly one negative eigenvalue (i.e. x is a critical point of ϕ of index 1). Then, from [35], the operator $L_{f,h}^{(0)}$ has exactly m exponentially small eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ when $h \to 0$ with $\lambda_1 = 0 < \lambda_2 \leq \ldots \leq \lambda_m$ (i.e., when $h \to 0$, for all $i \in \{1, \ldots, m\}$, $\lambda_i = O(e^{-\frac{c}{h}})$ for some $c > 0$ independent of h). Moreover, sharp asymptotic estimates can be derived for the eigenvalues $\{\lambda_2, \ldots, \lambda_m\}$. In [6,7], the following results are obtained. Let us assume that $\{x_1\} = \mathrm{argmin}_{\mathbb{R}^d} f$. For $k \in \{2, \ldots, m\}$ and $B_k = \{x \in \{x_1, \ldots, x_m\} \setminus \{x_k\}, f(x) \leq f(x_k)\}$ (i.e. B_k is the set of local minima of f which are lower in energy than x_k), one denotes by $\mathcal{P}(x_k, B_k)$ the set of curves $\gamma \in C^0([0,1], \mathbb{R}^d)$ such that $\gamma(0) = x_k$ and $\gamma(1) \in B_k$. Let us finally assume that:

1. For all $k \in \{2, \ldots, m\}$, there exists a unique saddle point z_k (i.e. a critical point of f of index 1) such that $f(z_k) = \inf_{\gamma \in \mathcal{P}(x_k, B_k)} \sup_{t \in [0,1]} f(\gamma(t))$.
2. The values $\big(f(z_k) - f(x_k)\big)_{k \in \{2, \ldots, m\}}$ are all distinct.

These assumptions imply that the map $x_k \in \{x_2, \ldots, x_m\} \mapsto z_k$ is injective. The set $\{x_2, \ldots, x_m\}$ is then labeled such that the sequence $\big(f(z_k) - f(x_k)\big)_{k \in \{2, \ldots, m\}}$ is strictly decreasing. The previous assumptions also imply the existence of a cascade of events, which occur with different timescales, to go from one local minimum x_k of f to the global minimum x_1 of f in $\mathbb{R}^d$, see for instance Fig. 2. Then, one has for $k \in \{2, \ldots, m\}$, in the limit $h \to 0$:

$$\lambda_k = \frac{|\lambda(z_k)|}{2\pi} \frac{\sqrt{\det \mathrm{Hess} f(x_k)}}{|\sqrt{\det \mathrm{Hess} f(z_k)}|} e^{-\frac{2}{h}(f(z_k) - f(x_k))}(1 + o(1)), \tag{9}$$

where $\lambda(z_k)$ is the negative eigenvalue of the Hessian matrix of f at z_k. In the articles [6,7], using a potential-theoretic approach, the sharp equivalent (9) is

obtained and each of the eigenvalues λ_k (for $k \in \{2, \ldots, m\}$) is shown to be the inverse of the average time it takes for the process (1) to go from x_k to B_k. We also refer to [24] for similar results. In [30], another proof of (9) is given using tools from semi-classical analysis. Let us also mention [53] for a generalization of the results obtained in [30]. Notice that the results presented above do not provide any information concerning the average time it takes for the process (1) to go from the global minimum of f to a local minimum of f when $h \to 0$. One also refers to [43] for generalization of [6,7] for a class of non reversible processes when f has two local minima, and to [11–13,37,54] for related results.

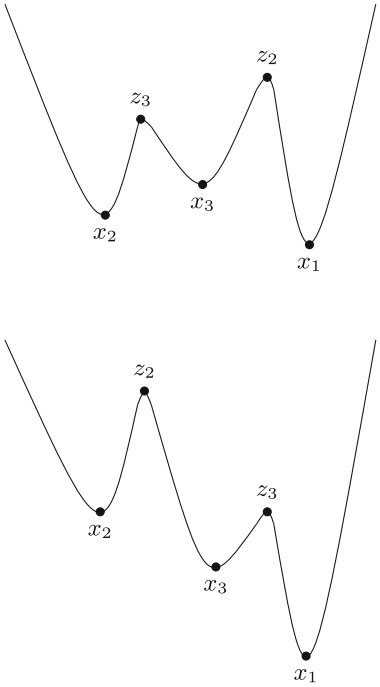


Fig. 2. Examples of two labelings of the local minima $\{x_1, x_2, x_3\}$ of f in dimension one.

Remark 6. The global approaches have been used in [59,60] to construct a Markovian dynamics by projecting the infinitesimal generator $L_{f,h}^{(0)}$ of the diffusion (1) with a Galerkin method onto the vector space associated with the m small eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$. This projection leads to a very good approximation of $L_{f,h}^{(0)}$ in the limit $h \to 0$. The question is then how to relate the transition events (or the trajectories) of the obtained Markov process to the exit events (or the trajectories) of the original one.

Local Approaches. The local approaches consist in studying the asymptotic behaviour when $h \to 0$ of the exit event $(\tau_\Omega, X_{\tau_\Omega})$ from a domain $\Omega \subset \mathbb{R}^d$, where $\tau_\Omega := \inf\{t \geq 0, X_t \notin \Omega\}$.

One of the most well-known approaches is the large deviation theory developed by Freidlin and Wentzell in the 1970s. We refer to the book [26] which summarizes their main contributions. This theory is based on the study of small pieces of the trajectories of the process defined with a suitable increasing sequence of stopping times. The rate function is fundamental in this theory: it quantifies the cost of deviating from a deterministic trajectory when $h \to 0$. The rate functional was first introduced by Schilder [58] for a Brownian motion. Some typical results from [26] (see Theorem 2.1, Theorem 4.1, and Theorem 5.1 there) are the following. Let Ω be a C^∞ open and connected bounded subset of $\mathbb{R}^d$. Let us assume that $\partial_n f > 0$ on $\partial\Omega$ (where ∂_n is the outward normal derivative to Ω) and that f has a unique non degenerate critical point x_0 in Ω such that $f(x_0) = \min_{\overline{\Omega}} f$. Then, for all $x \in \Omega$:

$$\lim_{h\to 0} h \ln \mathbb{E}_x\big[\tau_\Omega\big] = 2\big(\inf_{\partial\Omega} f - f(x_0)\big).$$

The notation $\mathbb{E}_x$ stands for the expectation given the fact that $X_0 = x$. Moreover, let $x \in \Omega$ such that $f(x) < \inf_{\partial\Omega} f$. Then, for any $\gamma > 0$ and $\delta_0 > 0$, there exist $\delta \in (0, \delta_0]$ and $h_0 > 0$ such that for all $h \in (0, h_0)$ and for all $y \in \partial\Omega$:

$$e^{-\frac{2}{h}(f(y)-\inf_{\partial\Omega} f)} e^{-\frac{\gamma}{h}} \leq \mathbb{P}_x\big[|X_{\tau_\Omega} - y| < \delta\big] \leq e^{-\frac{2}{h}(f(y)-\inf_{\partial\Omega} f)} e^{\frac{\gamma}{h}}.$$

The notation $\mathbb{P}_x$ stands for the probability given the fact that $X_0 = x$. Lastly, if the infimum of f on $\partial\Omega$ is attained at one single point $y_0 \in \partial\Omega$, then for all $\delta > 0$:

$$\lim_{h\to 0} \mathbb{P}_x\big[|X_{\tau_\Omega} - y_0| < \delta\big] = 1.$$

A result due to Day [14] (see also [48,49]) concerning the law of τ_Ω is the following. When $h \to 0$, the exit time τ_Ω converges in law to an exponentially distributed random variable and for all $x \in \Omega$

$$\lim_{h\to 0} \lambda_h \mathbb{E}_x\big[\tau_\Omega\big] = 1,$$

where λ_h is the principal eigenvalue of the infinitesimal generator of the diffusion (1) associated with Dirichlet boundary conditions on $\partial\Omega$ (see Proposition 2 below). The interest of this approach is that it can be applied to very general dynamics. However, when it is used to prove that the Eyring-Kramers formulas (5) can be used to study the exit distribution from Ω, it only provides the exponential rates (not the prefactor A_i in (5)) and does not give error bounds when $h \to 0$.

There are also approaches which are based on techniques developed for partial differential equations. In [50,51], using formal computations, when $\partial_n f > 0$ on $\partial\Omega$ and f has a unique non degenerate critical point x_0 in Ω such that $f(x_0) = \min_{\overline{\Omega}} f$, the following formula is derived: for any $F \in C^\infty(\partial\Omega, \mathbb{R})$ and

$x \in \Omega$, one has when $h \to 0$:

$$\mathbb{E}_x\big[F\big(X_{\tau_\Omega}\big)\big] = \frac{\int_{\partial\Omega} F(z)\partial_n f(z)\, e^{-\frac{2}{h}f(z)}\, dz}{\int_{\partial\Omega} \partial_n f\, e^{-\frac{2}{h}f}\, d\sigma} + o(1). \tag{10}$$

The formal asymptotic estimate (10) implies that the law of X_{τ_Ω} concentrates on points where f attains its minimum on $\partial\Omega$. Moreover, an asymptotic equivalent of $\mathbb{E}_x\big[\tau_\Omega\big]$ when $h \to 0$ is also formulated in [55] through formal computations. These results are obtained injecting formal asymptotic expansions in powers of h in the partial differential equations satisfied by $x \in \Omega \mapsto \mathbb{E}_x\big[F\big(X_{\tau_\Omega}\big)\big]$ and $x \in \Omega \mapsto \mathbb{E}_x\big[\tau_\Omega\big]$. We also refer to [51], where using formal computations, asymptotic formulas are obtained concerning both the concentration of the law of X_{τ_Ω} on $\mathrm{argmin}_{\partial\Omega} f$ and $\mathbb{E}_x\big[\tau_\Omega\big]$ when Ω is the union of basins of attraction of the dynamics $\frac{d}{dt}\gamma(t) = -\nabla f(\gamma(t))$. When $\partial_n f > 0$ on $\partial\Omega$ and f has a unique non degenerate critical point x_0 in Ω such that $f(x_0) = \min_{\overline{\Omega}} f$, the formula (10) is proved rigorously by Kamin in [40], and is extended to a non reversible diffusion process $(Y_t)_{t\geq 0}$ solution to $dY_t = b(Y_t)\, dt + \sqrt{h}\, dB_t$ in [15,16,39,57] when Ω contains one attractor of the dynamics $\frac{d}{dt}\gamma(t) = b(\gamma(t))$ and $b(x) \cdot n < 0$ for all $x \in \partial\Omega$. However, the results [15,16,39,40,57] do not provide any information on the probability to leave Ω through a point which is not a global minimum of f on $\partial\Omega$.

Finally, let us mention [20,21,31,37,45,48,49] for a study of the asymptotic behaviour in the limit $h \to 0$ of λ_h and u_h (see Proposition 2 below). The reader can also refer to [19] for a review of the different techniques used to study the asymptotic behaviour of X_{τ_Ω} when $h \to 0$ and to [2] for a review of the different techniques used to study the asymptotic behaviour of τ_Ω when $h \to 0$.

Remark 7. Some authors proved the convergence to a Markov jump process in some specific geometric settings and after a rescaling in time. We refer to [41] for a one-dimensional diffusion in a double well and [27,49] for a study in higher dimension. In [62], assuming that all the saddle points of f are at the same height, it is proved that a suitable rescaling of the time leads to a convergence of the diffusion process to a Markov jump process between the global minima of f.

The results presented in this work (see [22,23]) follow a local approach. The quasi-stationary distribution of the process (1) on Ω is the cornerstone of the analysis. They state that, under some geometric assumptions, the Eyring-Kramers formulas (with prefactors) can be used to model the exit event from a metastable state, and provide explicit error bounds.

1.5 Quasi-Stationary Distribution and Transition Rates

Local Equilibrium. Let Ω be a C^∞ open bounded connected subset of $\mathbb{R}^d$ and $f \in C^\infty(\overline{\Omega}, \mathbb{R})$. Let us recall that $\tau_\Omega := \inf\{t \geq 0, X_t \notin \Omega\}$ denotes the first exit time from Ω. The quasi-stationary distribution of the process (1) on Ω is defined as follows.

Definition 1. *A probability measure ν_h on Ω is a quasi-stationary distribution of the process (1) on Ω if for all $t > 0$ and any measurable set $A \subset \Omega$,*

$$\mathbb{P}_{\nu_h}\big[X_t \in A \big| t < \tau_\Omega\big] = \nu_h(A).$$

The notation $\mathbb{P}_\mu$ stands for the probability given the fact that the process (1) is initially distributed according to μ i.e. $X_0 \sim \mu$. The next proposition [9,44] shows that the law of the process (1) at time t conditioned not to leave Ω on the interval $(0, t)$ converges to the quasi-stationary distribution.

Proposition 1. *Let Ω be a C^∞ open connected and bounded subset of $\mathbb{R}^d$ and $f \in C^\infty(\overline{\Omega}, \mathbb{R})$. Then, there exist a unique probability measure ν_h on Ω and $c > 0$ such that for any probability measure μ on Ω, there exist $C(\mu) > 0$ and $t(\mu) > 0$ such that for all $t \geq t(\mu)$ and all measurable set $A \subset \Omega$:*

$$\big|\mathbb{P}_\mu\big[X_t \in A \big| t < \tau_\Omega\big] - \nu_h(A)\big| \leq C(\mu) e^{-ct}. \tag{11}$$

Moreover, ν_h is the unique quasi-stationary distribution of the process (1) on Ω.

Proposition 1 indicates that the quasi-stationary distribution ν_h can be seen as a local equilibrium of the process (1) in Ω.

The quasi-stationary distribution ν_h can be expressed with the principal eigenfunction of the infinitesimal generator $L^{(0)}_{f,h}$ (see (8)) of the diffusion (1) associated with Dirichlet boundary conditions on $\partial\Omega$. To this end, let us introduce the following Hilbert spaces $L^2_w(\Omega) = \big\{u : \Omega \to \mathbb{R},\ \int_\Omega u^2 e^{-\frac{2}{h}f} < \infty\big\}$ and for $q \in \{1, 2\}$,

$$H^q_w(\Omega) = \big\{u \in L^2_w(\Omega),\ \forall \alpha \in \mathbb{N}^d,\ |\alpha| \leq q,\ \partial_\alpha u \in L^2_w(\Omega)\big\}. \tag{12}$$

The subscript w in the notation $L^2_w(\Omega)$ and $H^q_w(\Omega)$ refers to the fact that the weight function $x \in \Omega \mapsto e^{-\frac{2}{h}f(x)}$ appears in the inner product. Moreover, let us denote by $H^1_{0,w}(\Omega) = \{u \in H^1_w(\Omega),\ u = 0 \text{ on } \partial\Omega\}$. Let us recall the following result [44].

Proposition 2. *Let Ω be a C^∞ open connected and bounded subset of $\mathbb{R}^d$ and $f \in C^\infty(\overline{\Omega}, \mathbb{R})$. Then, the operator $L^{(0)}_{f,h}$ with domain $H^1_{0,w}(\Omega) \cap H^2_w(\Omega)$ on $L^2_w(\Omega)$, which is denoted by $L^{D,(0)}_{f,h}$, is self-adjoint, positive and has compact resolvent. Furthermore, the smallest eigenvalue λ_h of $L^{D,(0)}_{f,h}$ is non degenerate and any eigenfunction associated with λ_h has a sign on Ω.*

In the following, one denotes by u_h an eigenfunction associated with λ_h. The smallest eigenvalue λ_h of $L^{D,(0)}_{f,h}$ is called the principal eigenvalue of $L^{D,(0)}_{f,h}$ and u_h a principal eigenfunction of $L^{D,(0)}_{f,h}$. Without loss of generality, one assumes that

$$u_h > 0 \text{ on } \Omega \quad \text{and} \quad \int_\Omega u_h^2\, e^{-\frac{2}{h}f} = 1. \tag{13}$$

Then, the quasi-stationary distribution ν_h of the process (1) in Ω is given by (see [44]):

$$\nu_h(dx) = \frac{u_h(x)\ e^{-\frac{2}{h}f(x)}}{\int_\Omega u_h\, e^{-\frac{2}{h}f}}\, dx. \tag{14}$$

Moreover, the following result shows that when $X_0 \sim \nu_h$, the law of the exit event $(\tau_\Omega, X_{\tau_\Omega})$ is explicitly known in terms of λ_h and u_h (see [44]).

Proposition 3. *Let us assume that $X_0 \sim \nu_h$, where ν_h is the quasi-stationary distribution of the process (1) in Ω. Then, τ_Ω and X_{τ_Ω} are independent. Moreover, τ_Ω is exponentially distributed with parameter λ_h and for any open set $\Sigma \subset \partial\Omega$, one has:*

$$\mathbb{P}_{\nu_h}\left[X_{\tau_\Omega} \in \Sigma\right] = -\frac{h}{2\lambda_h} \frac{\displaystyle\int_\Sigma \partial_n u_h(z) e^{-\frac{2}{h}f(z)} \sigma(dz)}{\displaystyle\int_\Omega u_h e^{-\frac{2}{h}f}}, \tag{15}$$

where $\sigma(dz)$ is the Lebesgue measure on $\partial\Omega$.

Approximation of the Exit Event with a Markov Jump Process. Let us now provide justifications to the use of a Markov jump process with transition rates computed with the Eyring-Kramers formula (5) to model the exit event from a metastable domain Ω. In view of (11), one can be more precise on the definition of the metastability of a domain Ω given in Sect. 1.1. For a probability measure μ supported in Ω, the domain Ω is said to be metastable if, when $X_0 \sim \mu$, the convergence to the quasi-stationary distribution ν_h in (1) is much quicker than the exit from Ω. Since the process (1) is a Markov process, it is then relevant to study the exit event from Ω starting from the quasi-stationary distribution ν_h, i.e. $X_0 \sim \nu_h$. As a consequence of Proposition 3, the exit time is exponentially distributed and is independent of the next visited state. These two properties are the fundamental features of kinetic Monte Carlo methods, see indeed (2) and (3). It thus remains to prove that the transition rates can be computed with the Eyring-Kramers formula (5). For that purpose, let us first give an expression of the transition rates. Recall that $(\Omega_i)_{i=1,\dots,n}$ denotes the surrounding domains of Ω (see Fig. 1). For $i \in \{1, \dots, n\}$, we define the transition rate to go from Ω to Ω_i as follows:

$$k_i^L := \frac{1}{\mathbb{E}_{\nu_h}\left[\tau_\Omega\right]} \mathbb{P}_{\nu_h}\left[X_{\tau_\Omega} \in \partial\Omega \cap \partial\Omega_i\right], \tag{16}$$

where we recall, ν_h is the quasi-stationary distribution of the process (1) in Ω. The superscript L in (16) indicates that the microscopic evolution of the system is governed by the overdamped Langevin process (1). Notice that, using Proposition 3, it holds for all $i \in \{1, \dots, n\}$:

$$\mathbb{P}_{\nu_h}\left[X_{\tau_\Omega} \in \partial\Omega \cap \partial\Omega_i\right] \ = \ \frac{k_i^L}{\sum_{\ell=1}^n k_\ell^L}.$$

Thus, the expressions (16) are compatible with the use of a kinetic Monte Carlo algorithm, see (2) and (4). Indeed, starting from the quasi-stationary distribution ν_h, the exit event from Ω can be exactly modeled using the rates (16): the exit time is exponentially distributed with parameter $\sum_{\ell=1}^n k_\ell^L$, independent of the exit point, and the exit point is in $\partial\Omega_i \cap \partial\Omega$ with probability $k_i^L / \sum_{\ell=1}^n k_\ell^L$. The remaining question is thus following: does the transition rate (16) satisfy the Eyring-Kramers law (5) in the limit $h \to 0$?

Notice that, using Proposition 3, for $i \in \{1, \dots, n\}$, the transition rate defined by (16) writes:

$$k_i^L = -\frac{h}{2} \frac{\displaystyle\int_{\partial\Omega\cap\partial\Omega_i} \partial_n u_h(z)\, e^{-\frac{2}{h}f(z)} \sigma(dz)}{\displaystyle\int_\Omega u_h\, e^{-\frac{2}{h}f}}, \tag{17}$$

where we recall, u_h is the eigenfunction associated with the principal eigenvalue λ_h of $L_{f,h}^{D,(0)}$.

The remainder of this work is dedicated to the presentation of recent results in [22,23,56] which aim at studying the asymptotic behaviour of the exit event $(\tau_\Omega, X_{\tau_\Omega})$ from a metastable domain Ω in the limit $h \to 0$. In particular, the results give a sharp asymptotic formula of the transition rates (17) when $h \to 0$.

Remark 8. If one wants to recover the expression of the prefactor (6), one has to multiply by $\frac{1}{2}$ the expression (16). This can be explained as follows. Once the process (1) reaches $\partial\Omega \cap \partial\Omega_i$, it has, in the limit $h \to 0$, a one-half probability to come back in Ω and a one-half probability to go in Ω_i. If z_i is a non degenerate saddle point of f, this result is not difficult to prove in dimension 1. Indeed, it is proved in [56, Section A.1.2.2], that when reaching $\partial\Omega \cap \partial\Omega_i$, the probability that the process (1) goes in Ω_i is $\frac{1}{2} + o(1)$ in the limit $h \to 0$. To extend this result to higher dimensions, one can use a suitable set of coordinates around z_i.

2 Main Results on the Exit Event

In all this section, $\Omega \subset \mathbb{R}^d$ is C^∞ open, bounded and connected, and $f \in C^\infty(\overline{\Omega}, \mathbb{R})$[1]. The purpose of this section is to present recent results obtained in [22,23]. Both [22] and [23] are mainly concerned with studying the asymptotic behaviour when $h \to 0$ of the exit law of a domain Ω of the process (1). In [22], when Ω only contains one local minimum of f and $\partial_n f > 0$ on $\partial\Omega$, we obtain sharp asymptotic equivalents when $h \to 0$ of the probability that the process (1) leaves Ω through a subset Σ of $\partial\Omega$ starting from the quasi-stationary distribution or from a deterministic initial condition in Ω. Then, these asymptotic equivalents are used to compute the asymptotic behaviour of the transition rates (16). In [23],

[1] Actually, all the results presented in this section are proved in [22,23] in the more general setting: $\overline{\Omega} = \Omega \cup \partial\Omega$ is a C^∞ oriented compact and connected Riemannian manifold of dimension d with boundary $\partial\Omega$.

we explicit a more general setting than the one considered in [22] where we identify the most probable places of exit of Ω as well as their relative probabilities starting from the quasi-stationary distribution or deterministic initial conditions in Ω. More precisely, we consider in [23] the case when Ω contains several local minima of f and $|\nabla f| \neq 0$ on $\partial\Omega$.

2.1 Sharp Asymptotic Estimates on the Exit Event from a Domain

In this section, we present the results of [22] which give sharp asymptotic estimates on the law of X_{τ_Ω} and on the expectation of τ_Ω when $h \to 0$. These results give in particular the asymptotic estimates of the transition rates $(k_j^L)_{j=1,\dots,n}$ defined in (16).

Geometric Setting. Let us give the geometric setting which is considered in this section:

- **[H1].** The function $f : \overline{\Omega} \to \mathbb{R}$ and the restriction of f to Ω, denoted by $f|_{\partial\Omega}$, are Morse functions. Moreover, $|\nabla f|(x) \neq 0$ for all $x \in \partial\Omega$.
- **[H2].** The function f has a unique global minimum x_0 in $\overline{\Omega}$ and

$$\min_{\partial\Omega} f > \min_{\overline{\Omega}} f = \min_{\Omega} f = f(x_0).$$

  The point x_0 is the unique critical point of f in $\overline{\Omega}$. The function $f|_{\partial\Omega}$ has exactly $n \geq 1$ local minima which are denoted by $(z_i)_{i=1,\dots,n}$. They are ordered such that

$$f(z_1) \leq \dots \leq f(z_n).$$

- **[H3].** $\partial_n f(x) > 0$ for all $x \in \partial\Omega$.

Under the assumption **[H2]**, one denotes by $n_0 \in \{1, \dots, n\}$ the number of global minima of $f|_{\partial\Omega}$, i.e.:

$$f(z_1) = \dots = f(z_{n_0}) < f(z_{n_{n_0+1}}) \leq \dots \leq f(z_n).$$

On Fig. 3, one gives a schematic representation in dimension 2 of a function f satisfying the assumptions **[H1]**, **[H2]**, and **[H3]**, and of its restriction to $\partial\Omega$, in the case $n = 4$ and $n_0 = 2$.

Remark 9. The assumption **[H1]** implies that f does not have any saddle point (i.e. critical point of index 1) on $\partial\Omega$. Actually, under **[H1]**, **[H2]**, and **[H3]**, the points $(z_i)_{i=1,\dots,n}$ play geometrically the role of saddle points and are called *generalized saddle points* of f on $\partial\Omega$, see [31, Section 5.2]. This can be explained by the fact that, under **[H1]**, **[H2]**, **[H3]** and when f is extended by $-\infty$ outside $\overline{\Omega}$, the points $(z_i)_{i=1,\dots,n}$ are geometrically saddle points of f (the extension of f by $-\infty$ is consistent with the Dirichlet boundary conditions used to define $L_{f,h}^{D,(0)}$) in the following sense. For all $i \in \{1, \dots, n\}$, z_i is a local minimum of $f|_{\partial\Omega}$ and a local maximum of $f|_{D_i}$, where D_i is the straight line passing through z_i and orthogonal to $\partial\Omega$ at z_i.

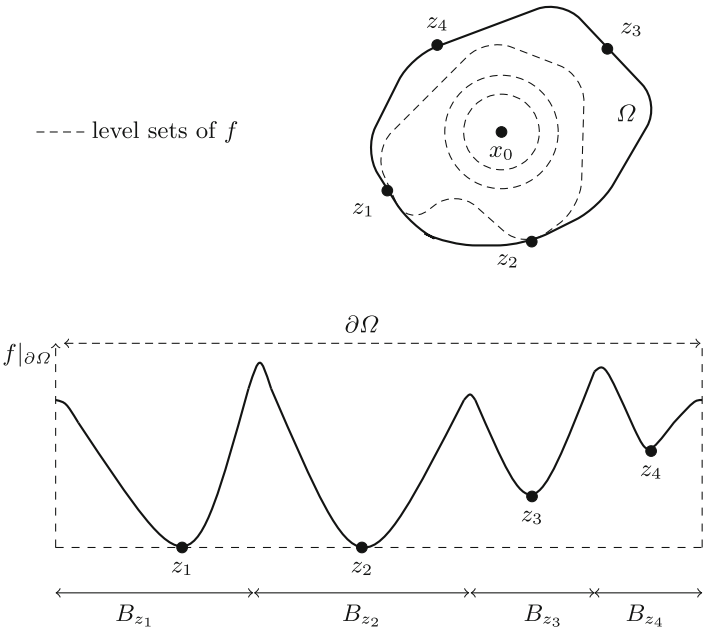


Fig. 3. Schematic representation in dimension 2 of a function f satisfying the assumptions **[H1]**, **[H2]**, and **[H3]**, and of its restriction $f|_{\partial\Omega}$ to $\partial\Omega$. On the figure, $n = 4$ and $n_0 = 2$.

Remark 10. Notice that under **[H1]**, **[H2]**, and **[H3]**, extending f by reflection outside $\overline{\Omega}$ in a neighborhood of z_i also implies that z_i is a geometric saddle point of f as defined in Remark 9. In dimension one, such a construction was considered by Kramers in [42] to derive formulas for transition rates, as explained in [52]. Moreover, as in Remark 8, it can be proved in dimension 1 (exactly as in [56, Section A.1.2.2]), that when reaching $\partial\Omega \cap \partial\Omega_i$, the probability that the process (1) goes in Ω_i is $\frac{1}{2} + O(h)$ when $h \to 0$. To extend this result to higher dimensions, one can use a suitable set of coordinates around z_i.

Let us now define $g : \overline{\Omega} \to \mathbb{R}^+$ by

$$g(x) = \left|\nabla f(x)\right| \text{ when } x \in \Omega \quad \text{and} \quad g(x) = \left|\nabla_T f(x)\right| \text{ when } x \in \partial\Omega, \tag{18}$$

where $\nabla_T f$ is the tangential gradient of f in $\partial\Omega$. Let us recall that for $x \in \partial\Omega$, $\nabla_T f(x)$ is defined by $\nabla_T f(x) = \nabla f(x) - (\nabla f(x) \cdot n)\, n$, where n is the unit outward normal to $\partial\Omega$ at x. The assumptions one needs to state the results in this section depend on the Agmon distance in $\overline{\Omega}$ between the points $(z_i)_{i=1,\ldots,n}$. The Agmon distance is defined as follows: for any $x \in \overline{\Omega}$ and $y \in \overline{\Omega}$,

$$d_a(x, y) := \inf_{\gamma \in \mathrm{Lip}(x,y)} L(\gamma, (0, 1)), \tag{19}$$

where $\mathrm{Lip}(x,y)$ is the set of Lipschitz curves $\gamma : [0,1] \to \overline{\Omega}$ which are such that $\gamma(0) = x$ and $\gamma(1) = y$, and where for $\gamma \in \mathrm{Lip}(x,y)$,

$$L(\gamma,(0,1)) = \int_0^1 g(\gamma(t))|\gamma'(t)|dt.$$

Remark 11. Let us give some common points and differences between the quasipotential V introduced in [26, Section 2] and the Agmon distance (19). Contrary to the quasipotential V, the Agmon distance (19) is symmetric. Moreover, let us consider $x \neq y \in \overline{\Omega}$ such that there exists a curve $\gamma : [0,1] \to \overline{\Omega}$ with $\frac{d}{dt}\gamma(t) = -\nabla f(\gamma(t))$, $\gamma(0) = x$ and $\gamma(1) = y$. Then, the Agmon distance (19) between x and y equals $f(x) - f(y) = V(y,x) > 0$ but $V(x,y) = 0 \neq d_a(x,y)$.

Finally, let us define the following sets. For $i \in \{1,\dots,n\}$, B_{z_i} is the basin of attraction of z_i for the dynamics $\frac{d}{dt}x(t) = -\nabla_T f\big(x(t)\big)$ in $\partial\Omega$, i.e. $B_{z_i} = \{y \in \partial\Omega,\ \lim_{t\to\infty} x(t) = z_i \text{ if } x(0) = y\}$ (see for instance Fig. 3). Moreover, one defines for $i \in \{1,\dots,n\}$:

$$B_{z_i}^c := \partial\Omega \setminus B_{z_i}.$$

Main Results. Let us now give the main results of this section.

Proposition 4. *Let u_h be the eigenfunction associated with the principal eigenvalue λ_h of $L_{f,h}^{D,(0)}$ which satisfies normalization (13). Let us assume that the hypotheses **[H1]**, **[H2]**, **[H3]** are satisfied. Then, in the limit $h \to 0$, one has:*

$$\lambda_h = \frac{\sqrt{\det \mathrm{Hess} f(x_0)}}{\sqrt{\pi h}} \sum_{i=1}^{n_0} \frac{\partial_n f(z_i)}{\sqrt{\det \mathrm{Hess} f|_{\partial\Omega}(z_i)}}\, e^{-\frac{2}{h}(f(z_1)-f(x_0))}\,(1+O(h)) \tag{20}$$

and

$$\int_\Omega u_h(x)\, e^{-\frac{2}{h}f(x)}dx = \frac{\pi^{\frac{d}{4}}}{(\det \mathrm{Hess} f(x_0))^{1/4}}\, h^{\frac{d}{4}}\, e^{-\frac{1}{h}f(x_0)}(1+O(h)). \tag{21}$$

Furthermore, one obtains the following theorem on the asymptotic behaviour of $\partial_n u_h$, which is one of the main results of [22].

Theorem 1. *Let us assume that **[H1]**, **[H2]**, and **[H3]** are satisfied and that the following inequalities hold:*

$$f(z_1) - f(x_0) > f(z_n) - f(z_1) \tag{22}$$

and for all $i \in \{1,\dots,n\}$,

$$d_a(z_i, B_{z_i}^c) > \max[f(z_n) - f(z_i),\ f(z_i) - f(z_1)]. \tag{23}$$

Let $i \in \{1,\dots,n\}$ and $\Sigma_i \subset \partial\Omega$ be an open set containing z_i and such that $\overline{\Sigma_i} \subset B_{z_i}$. Let u_h be the eigenfunction associated with the principal eigenvalue of $L_{f,h}^{D,(0)}$ which satisfies (13). Then, in the limit $h \to 0$:

$$\int_{\Sigma_i} \partial_n u_h\, e^{-\frac{2}{h}f} = C_i(h)\, e^{-\frac{2f(z_i)-f(x_0)}{h}}\,(1+O(h))\,, \tag{24}$$

where $C_i(h) = -\dfrac{(\det \mathrm{Hess} f(x_0))^{1/4} \partial_n f(z_i) 2\pi^{\frac{d-2}{4}}}{\sqrt{\det \mathrm{Hess} f|_{\partial\Omega}(z_i)}}\, h^{\frac{d-6}{4}}$.

These results have the following consequences.

Corollary 1. *Let us assume that all the assumptions of Theorem 1 are satisfied. Let* $i \in \{1, \ldots, n\}$ *and* $\Sigma_i \subset \partial\Omega$ *be an open set containing* z_i *and such that* $\overline{\Sigma_i} \subset B_{z_i}$. *Then, in the limit* $h \to 0$*:*

$$\mathbb{P}_{\nu_h}\left[X_{\tau_\Omega} \in \Sigma_i\right] = \frac{\partial_n f(z_i)}{\sqrt{\det \mathrm{Hess} f|_{\partial\Omega}(z_i)}} \left(\sum_{k=1}^{n_0} \frac{\partial_n f(z_k)}{\sqrt{\det \mathrm{Hess} f|_{\partial\Omega}(z_k)}} \right)^{-1} \times e^{-\frac{2}{h}(f(z_i)-f(z_1))}(1 + O(h)), \tag{25}$$

where ν_h *is the quasi-stationary distribution of the process* (1) *in* Ω *(see* (14)*). Moreover, if* Σ_i *is the common boundary between the state* Ω *and a state* Ω_i*, then, when* $h \to 0$

$$k_i^L = \frac{1}{\sqrt{\pi h}} \partial_n f(z_i) \frac{\sqrt{\det \mathrm{Hess} f(x_0)}}{\sqrt{\det \mathrm{Hess} f|_{\partial\Omega}(z_i)}} e^{-\frac{2}{h}(f(z_i)-f(x_0))}(1 + O(h)), \tag{26}$$

where k_i^L *is the transition rate* (16) *to go from* Ω *to* Ω_i.

Notice that since z_i is not a saddle point of f, the prefactor in (26) is not the prefactor $\frac{1}{2}A_i$ (see Remark 10 for the explanation of the multiplicative term $\frac{1}{2}$), where A_i is defined by (6), but it is actually the expected prefactor for a generalized saddle point of f (see Remarks 9 and 10).

The asymptotic estimate (25) is a consequence of Proposition 4, Theorem 1 together with (15), and (26) is a consequence of Proposition 4, Theorem 1 and (17). The main difficulty is to prove (24) which requires a sharp equivalent of the quantity $\int_{\Sigma_i} \partial_n u_h \, e^{-\frac{2}{h}f}$ when z_i is not a global minimum of f on $\partial\Omega$, i.e. when $i \in \{n_0 + 1, \ldots, n\}$.

In [22], numerical simulations are provided to check that (25) holds and to discuss the necessity of the assumptions (23) to obtain (25). Furthermore, in [22], the results (24) and (25) are generalized to sets $\Sigma \subset \partial\Omega$ which do not necessarily contain a point $z \in \{z_1, \ldots, z_n\}$: this is the other main results of [22] which is not presented here. Moreover, with the help of "leveling" results on the function $x \mapsto \mathbb{E}_x[F(X_{\tau_\Omega})]$, we generalized (25) to deterministic initial conditions in Ω (i.e. when $X_0 = x \in \Omega$) which are the initial conditions considered in the theory of large deviations [26].

The proofs of Proposition 4 and Theorem 1 are based on tools from semi-classical analysis and more precisely, they are based on techniques developed in [31–35, 45].

Starting Points of the Proofs of Proposition 4 and Theorem 1. Let us recall that u_h is the eigenfunction associated with the principal eigenvalue λ_h of $L_{f,h}^{D,(0)}$ which satisfies normalization (13). In view of (15) and in order to

obtain (25), one wants to study the asymptotic behaviour when $h \to 0$ of ∇u_h on $\partial\Omega$. The starting point of the proofs of Proposition 4 and Theorem 1 is the fact that ∇u_h is solution to an eigenvalue problem for the same eigenvalue λ_h. Indeed, recall that u_h is solution to $L^{(0)}_{f,h} u_h = \lambda_h u_h$ in Ω and $u_h = 0$ on $\partial\Omega$. If one differentiates this relation, ∇u_h is solution to

$$\begin{cases} L^{(1)}_{f,h} \nabla u_h = \lambda_h \nabla u_h \text{ in } \Omega, \\ \nabla_T u_h = 0 \text{ on } \partial\Omega, \\ \left(-\dfrac{h}{2}\mathrm{div} + \nabla f \cdot\right) \nabla u_h = 0 \text{ on } \partial\Omega, \end{cases} \tag{27}$$

where $L^{(1)}_{f,h} = -\frac{h}{2}\Delta + \nabla f \cdot \nabla + \mathrm{Hess}\, f$ is an operator acting on 1-forms (namely on vector fields). In the following the operator $L^{(1)}_{f,h}$ with tangential boundary conditions (27) is denoted by $L^{D,(1)}_{f,h}$. From (27), ∇u_h is therefore an eigenform of $L^{D,(1)}_{f,h}$ associated with λ_h. For $p \in \{0, 1\}$, let us denote, by $\pi^{(p)}_h$ the orthogonal projector of $L^{D,(p)}_{f,h}$ associated with the eigenvalues of $L^{D,(p)}_{f,h}$ smaller than $\frac{\sqrt{h}}{2}$. Another crucial ingredient for the proofs of Proposition 4 and Theorem 1 is the fact that, from [31, Chapter 3],

$$\mathrm{Ran}\, \pi^{(0)}_h = \mathrm{Span}\, u_h \text{ and } \dim \mathrm{Ran}\, \pi^{(1)}_h = n. \tag{28}$$

Therefore, from (27), it holds

$$\nabla u_h \in \mathrm{Ran}\, \pi^{(1)}_h, \tag{29}$$

and from (13) and the fact that $\langle L^{(0)}_{f,h} u_h, u_h \rangle_{L^2_w} = \frac{h}{2} \|\nabla u_h\|^2_{L^2_w}$, one has

$$\lambda_h = \frac{h}{2} \|\nabla u_h\|^2_{L^2_w}. \tag{30}$$

Thus, to study the asymptotic behaviour when $h \to 0$ of λ_h, u_h and ∇u_h, we construct a suitable orthonormal basis of $\mathrm{Ran}\, \pi^{(1)}_h$. This basis is constructed using so-called quasi-modes.

Sketch of the Proofs of Proposition 4 and Theorem 1. Let us give the sketch of the proof of (25) which is the main result of [22]. Recall that from Proposition 2, one works in the Hilbert space $L^2_w(\Omega)$. The spaces $L^2_w(\Omega)$ and $H^1_w(\Omega)$ (see (12)) extend naturally on 1-forms as follows

$$\Lambda^1 L^2_w(\Omega) := \left\{ u = {}^t(u_1, \dots, u_d) : \Omega \to \mathbb{R}^d,\ \forall k \in \{1, \dots, d\},\ \int_\Omega u_k^2 e^{-\frac{2}{h} f} < \infty \right\},$$

and

$$\Lambda^1 H^1_w(\Omega) := \left\{ u = {}^t(u_1, \dots, u_d) : \Omega \to \mathbb{R}^d,\ \forall (i,k) \in \{1, \dots, d\}^2,\ \partial_i u_k \in L^2_w(\Omega) \right\}.$$

In the following, one denotes by $\|.\|_{L^2_w}$ (resp. $\|.\|_{H^1_w}$) the norm of $L^2_w(\Omega)$ and of $\Lambda^1 L^2_w(\Omega)$ (resp. $H^1_w(\Omega)$ and $\Lambda^1 H^1_w(\Omega)$). Finally, $\langle .,. \rangle_{L^2_w}$ stands for both the scalar product associated with the norm of $L^2_w(\Omega)$ and with the norm of $\Lambda^1 L^2_w(\Omega)$. In view of (29) and (28), one has for all orthonormal basis $(\psi_j)_{j\in\{1,\dots,n\}}$ of $\mathrm{Ran}\,\pi_h^{(1)}$, in $L^2_w(\Omega)$:

$$\nabla u_h = \sum_{j=1}^{n} \langle \nabla u_h, \psi_j \rangle_{L^2_w} \psi_j, \tag{31}$$

and from (30), it holds

$$\lambda_h = \frac{h}{2} \sum_{j=1}^{n} \left| \langle \nabla u_h, \psi_j \rangle_{L^2_w} \right|^2. \tag{32}$$

In particular, one has for all $k \in \{1, \dots, n\}$,

$$\int_{\Sigma_k} \partial_n u_h \, e^{-\frac{2}{h}f} = \sum_{j=1}^{n} \langle \nabla u_h, \psi_j \rangle_{L^2_w} \int_{\Sigma_k} \psi_j \cdot n \, e^{-\frac{2}{h}f}, \tag{33}$$

where we recall that Σ_k is an open set of $\partial\Omega$ such that $z_k \in \Sigma_k$ and $\overline{\Sigma_k} \subset B_{z_k}$.

Step 1: Approximation of u_h. Under **[H1]**, **[H2]**, and **[H3]**, it is not difficult to find a good approximation of u_h. Indeed, let us consider,

$$\tilde{u} := \frac{\chi}{\|\chi\|_{L^2_w}}, \tag{34}$$

where $\chi \in C_c^\infty(\Omega, \mathbb{R}^+)$ and $\chi = 1$ on $\{x \in \Omega, d(x, \partial\Omega) \geq \varepsilon\}$ where $\varepsilon > 0$. In particular, for ε small enough, $\chi = 1$ in a neighborhood of x_0 (which is assumed in the following). Let us explain why $\tilde{u}$ is a good approximation of u_h. Since $L^{D,(0)}_{f,h}$ is self adjoint on $L^2_w(\Omega)$, one has

$$\left\| (1 - \pi_h^{(0)})\tilde{u} \right\|^2_{L^2_w} \leq \frac{C}{\sqrt{h}} \left\langle L^{D,(0)}_{f,h} \tilde{u}, \tilde{u} \right\rangle_{L^2_w} = \frac{Ch}{2\sqrt{h}} \frac{\int_\Omega |\nabla\chi|^2 e^{-\frac{2}{h}f}}{\int_\Omega \chi^2 e^{-\frac{2}{h}f}}.$$

Since $f(x_0) = \min_\Omega f < \min_{\partial\Omega} f$ and x_0 is the unique global minimum of f on $\overline{\Omega}$ (see **[H2]**), one has using Laplace's method (x_0 is a non degenerate critical point of f and $\chi(x_0) = 1$):

$$\int_\Omega \chi^2 e^{-\frac{2}{h}f} = \frac{(\pi h)^{\frac{d}{2}}}{\sqrt{\det \mathrm{Hess} f(x_0)}} e^{-\frac{2}{h}f(x_0)} (1 + O(h)).$$

Therefore, for any $\delta > 0$, choosing ε small enough, it holds when $h \to 0$:

$$\left\| (1 - \pi_h^{(0)})\tilde{u} \right\|^2_{L^2_w} = O(e^{-\frac{2}{h}(f(z_1) - f(x_0) - \delta)}),$$

and thus:

$$\pi_h^{(0)} \tilde{u} = \tilde{u} + O(e^{-\frac{1}{h}(f(z_1) - f(x_0) - \delta)}) \text{ in } L^2_w(\Omega).$$

From (28) and since $\chi \geq 0$, one has for any $\delta > 0$ (choosing ε small enough), when $h \to 0$

$$u_h = \frac{\pi_h^{(0)} \tilde{u}}{\|\pi_h^{(0)} \tilde{u}\|_{L^2_w}} = \tilde{u} + O(e^{-\frac{1}{h}(f(z_1)-f(x_0)-\delta)}) \text{ in } L^2_w(\Omega). \tag{35}$$

Since $\|\tilde{u}\|_{L^2_w} = 1$, this last relation justifies that $\tilde{u}$ is a good approximation of u_h in $L^2_w(\Omega)$. Notice that (35) implies (21).

Step 2: Construction of a Basis of $\operatorname{Ran} \pi_h^{(1)}$ *to Prove Theorem* 1. In view of (33), the idea is to construct a family of 1-forms $(\widetilde{\psi}_j)_{j\in\{1,\dots,n\}}$ which forms, when projected on $\operatorname{Ran} \pi_h^{(1)}$, a basis of $\operatorname{Ran} \pi_h^{(1)}$ and which allows to obtain sharp asymptotic estimates on $\partial_n u_h$ on all the Σ_j's when $h \to 0$. In the literature, such a 1-form $\widetilde{\psi}_j$ is called a quasi-mode (for $L^{D,(1)}_{f,h}$). A quasi-mode for $L^{D,(1)}_{f,h}$ is a smooth 1-form w such that for some norm, it holds when $h \to 0$:

$$\pi_h^{(1)} w = w + o(1), \tag{36}$$

To prove Theorem 1, one of the major issues is the construction of a basis $(\widetilde{\psi}_j)_{j\in\{1,\dots,n\}}$ so that the remainder term $o(1)$ in (36), when $w = \widetilde{\psi}_k$, is of the order (see (23))

$$\big\|(1-\pi_h^{(1)})\widetilde{\psi}_k\big\|_{H^1_w} = O\big(e^{-\frac{1}{h}\max[f(z_n)-f(z_k),\, f(z_k)-f(z_1)]}\big). \tag{37}$$

This implies that $\big(\pi_h^{(1)}\widetilde{\psi}_j\big)_{j\in\{1,\dots,n\}}$ is a basis of $\operatorname{Ran} \pi_h^{(1)}$ and above all, after a Gram-Schmidt procedure on $\big(\pi_h^{(1)}\widetilde{\psi}_j\big)_{j\in\{1,\dots,n\}}$, when $h \to 0$, that for all $k \in \{1,\dots,n\}$ (see (33)):

$$\int_{\Sigma_k} \partial_n u_h \, e^{-\frac{2}{h}f} = \sum_{j=1}^{n} \langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w} \int_{\Sigma_k} \widetilde{\psi}_j \cdot n \, e^{-\frac{2}{h}f} + O\big(e^{-\frac{2f(z_k)-f(x_0)+c}{h}}\big) \tag{38}$$

and (see (32))

$$\lambda_h = \frac{h}{2} \sum_{j=1}^{n} |\langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w}|^2 + O\big(e^{-\frac{2}{h}(f(z_1)-f(x_0)+c)}\big) \tag{39}$$

for some $c > 0$ independent of h. Here, we recall, $\tilde{u}$ (see (34)) is a good approximation of u_h (see (35)). Let us now explain how we will construct the family $\big(\widetilde{\psi}_j\big)_{j\in\{1,\dots,n\}}$ in order to obtain (38) and (39). Then, we explain how the terms $\Big(\displaystyle\int_{\Sigma_j} \widetilde{\psi}_j \cdot n \, e^{-\frac{2}{h}f}\Big)_{j\in\{1,\dots,n\}}$ and $\Big(\langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w}\Big)_{j\in\{1,\dots,n\}}$ appearing in (38) and (39) are computed.

Step 2a: Construction of the Family $(\widetilde{\psi}_j)_{j\in\{1,\dots,n\}}$. To construct each 1-form $\widetilde{\psi}_j$, the idea is to construct an operator $L^{(1)}_{f,h}$ with mixed tangential Dirichlet and

Neumann boundary conditions on a domain $\dot{\Omega}_j \subset \Omega$ which is such that $\big(\{z_1,\dots,z_n\}\cup\{x_0\}\big)\cap\dot{\Omega}_j = \{z_j\}$. For $j\in\{1,\dots,n\}$, $\widetilde{\psi}_j$ is said to be associated with the generalized saddle point z_j. The goal of the boundary conditions is to ensure that when $h\to 0$, each of these operators has only one exponentially small eigenvalue (i.e. this eigenvalue is $O\big(e^{-\frac{c}{h}}\big)$ for some $c>0$ independent of h), the other eigenvalues being larger than $\sqrt{h}$. Then, we show that each of these small eigenvalues actually equals 0 using the Witten complex structure associated with these boundary conditions on $\partial\dot{\Omega}_j$. To construct such operators $L^{(1)}_{f,h}$ with mixed boundary conditions on $\dot{\Omega}_j$, the recent results of [28,38] are used. The 1-form $\widetilde{\psi}_j$ associated with z_j is then defined using an eigenform $v^{(1)}_{h,j}$ associated with the eigenvalue 0 of the operator $L^{(1)}_{f,h}$ associated with mixed boundary conditions on $\dot{\Omega}_j$:

$$\widetilde{\psi}_j := \frac{\chi_j\, v^{(1)}_{h,j}}{\|\chi_j\, v^{(1)}_{h,j}\|_{L^2_w}}, \tag{40}$$

where χ_j is a well chosen cut-off function with support in $\overline{\dot{\Omega}_j}$. Notice that for $j\in\{1,\dots,n\}$, the quasi-mode $\widetilde{\psi}_j$ is not only constructed in a neighbourhood of z_j: it has a support as large as needed in Ω. This is a difference with previous construction in the literature, such as [31]. We need such quasi-modes for the following reasons. Firstly, we compute the probability that the process (1) leaves Ω through open sets Σ_j which are arbitrarily large in B_{z_j}. Secondly, we use the fact that the quasi-mode $\widetilde{\psi}_j$ decreases very fast away from z_j to get (37). This is needed to state the hypothesis (23) in terms of Agmon distances, see next step.

Step 2b: Accuracy of the Quasi-mode $\widetilde{\psi}_j$ for $j\in\{1,\dots,n\}$. To obtain a sufficiently small remainder term in (36) (to get (37) and then (38)), one needs to quantify the decrease of the quasi-mode $\widetilde{\psi}_j$ outside a neighborhood of z_j. This decrease is obtained with Agmon estimates on $v^{(1)}_{h,j}$ which allow to localize $\widetilde{\psi}_j$ in a neighborhood of z_j. For $j\in\{1,\dots,n\}$, we prove the following Agmon estimate on $v^{(1)}_{h,j}$:

$$\big\|\chi_j\, v^{(1)}_{h,j} e^{\frac{1}{h}d_a(.,z_j)}\big\|_{H^1_w} = O(h^{-N}), \tag{41}$$

for some $N\in\mathbb{N}$ and where d_a is the Agmon distance defined in (19). To obtain (41), we study the properties of this distance. The boundary of Ω introduces technical difficulties. The Agmon estimate (41) is obtained adapting to our case techniques developed in [31,45]. For all $j\in\{1,\dots,n\}$, using the fact that $\big\|(1-\pi^{(1)}_h)\widetilde{\psi}_j\big\|^2_{L^2_w} \le \frac{C}{\sqrt{h}}\big\langle L^{D,(1)}_{f,h}\widetilde{\psi}_j,\widetilde{\psi}_j\big\rangle_{L^2_w}$ and (41), one shows that

$$\big\|(1-\pi^{(1)}_h)\widetilde{\psi}_j\big\|^2_{L^2_w} \le C\, h^{-q}\, e^{-\frac{2}{h}\inf_{\mathrm{supp}\nabla\chi_j} d_a(.,z_j)},$$

for some $q>0$. Thus, in order to get (37), the support of $\nabla\chi_j$ has to be arbitrarily close to x_0 and $B^c_{z_j}$. This explains the assumptions (22) and (23), and

the fact that the quasi-mode $\widetilde{\psi}_j$ is not constructed in a neighborhood of z_j but in a domain $\dot{\Omega}_j$ arbitrarily large in Ω. This is one of the main differences compared with [31]. At the end of this step, one has a family $(\widetilde{\psi}_j)_{j\in\{1,\dots,n\}}$ which satisfies (37). This allows us to obtain, in the limit $h \to 0$ (see (38)), for some $c > 0$ independent of h and for all $k \in \{1, \dots, n\}$:

$$\int_{\Sigma_k} \partial_n u_h \, e^{-\frac{2}{h}f} = \sum_{j=1}^{n} \langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w} \int_{\Sigma_k} \widetilde{\psi}_j \cdot n \, e^{-\frac{2}{h}f} + O\big(e^{-\frac{2f(z_k)-f(x_0)+c}{h}}\big).$$

Etape 3: Computations of $\Big(\int_{\Sigma_j} \widetilde{\psi}_j \cdot n \, e^{-\frac{2}{h}f}\Big)_{j\in\{1,\dots,n\}}$ and $\Big(\langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w}\Big)_{j\in\{1,\dots,n\}}$. In view of (38) and (39), for all $j \in \{1, \dots, n\}$, one needs to compute the terms

$$\int_{\Sigma_j} \widetilde{\psi}_j \cdot n \, e^{-\frac{2}{h}f} \text{ and } \langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w}.$$

To do that, we use for all $j \in \{1, \dots, n\}$ a WKB approximation of $v^{(1)}_{h,j}$, denoted by $v^{(1)}_{z_j,wkb}$. In the literature we follow, $v^{(1)}_{z_j,wkb}$ is constructed in a neighborhood of z_j (see [31,45]). To prove Theorem 1, we extend the construction of $v^{(1)}_{z_j,wkb}$ to neighbourhoods in $\overline{\Omega}$ of arbitrarily large closed sets included in B_{z_j} (indeed, there is no restriction on the size of Σ_j in B_{z_j}). Then, the comparison between $v^{(1)}_{h,j}$ and $v^{(1)}_{z_j,wkb}$ is also extended to neighbourhoods in $\overline{\Omega}$ of arbitrarily large closed sets included in B_{z_j}. Once the terms $\Big(\int_{\Sigma_j} \widetilde{\psi}_j \cdot n \, e^{-\frac{2}{h}f}\Big)_{j\in\{1,\dots,n\}}$ and $\Big(\langle \nabla \tilde{u}, \widetilde{\psi}_j \rangle_{L^2_w}\Big)_{j\in\{1,\dots,n\}}$ are computed, one concludes the proof of (20) using (39) and the proof of (24) using (38).

2.2 Most Probable Exit Points from a Bounded Domain

Setting and Motivation. In this section, we present recent results from [23] on the concentration of the law of X_{τ_Ω} on a subset of $\operatorname{argmin}_{\partial\Omega} f = \{z \in \partial\Omega,\ f(z) = \min_{\partial\Omega} f\}$ when $h \to 0$ in a more general geometric setting than the one of Theorem 1. The main purpose of these results is to prove an asymptotic formula when $h \to 0$ for the concentration of the law of X_{τ_Ω} on a set of points of $\operatorname{argmin}_{\partial\Omega} f$ when Ω contains several local minima of f and when $\partial_n f$ is not necessarily positive on $\partial\Omega$.

Let $\mathcal{Y} \subset \partial\Omega$. We say that the law of X_{τ_Ω} concentrates on $\mathcal{Y}$ if for all neighborhood $\mathcal{V}_\mathcal{Y}$ of $\mathcal{Y}$ in $\partial\Omega$, one has

$$\lim_{h\to 0} \mathbb{P}\left[X_{\tau_\Omega} \in \mathcal{V}_\mathcal{Y}\right] = 1,$$

and if for all $x \in \mathcal{Y}$ and all neighborhood $\mathcal{V}_x$ of x in $\partial\Omega$, it holds:

$$\lim_{h\to 0} \mathbb{P}\left[X_{\tau_\Omega} \in \mathcal{V}_x\right] > 0.$$

In [50,51,55], when $\partial_n f(x) = 0$ for all $x \in \partial\Omega$ or when $\partial_n f(x) > 0$ for all $x \in \partial\Omega$ (and with additional assumptions on f), it has been shown that the law of X_{τ_Ω} concentrates on points where f attains its minimum on $\partial\Omega$ (see (10)). Later on, it has been proved in [15,16,39,40,57] when $\partial_n f > 0$ on $\partial\Omega$ and f has a unique non degenerate critical point in Ω (which is necessarily its global minimum in $\overline{\Omega}$). Tools developed in semi-classical analysis allow us to generalize this geometric setting. For instance, we consider several critical points of f in Ω and we drop the assumptions $\partial_n f > 0$ on $\partial\Omega$ (however we do not consider the case when f has saddle points on $\partial\Omega$). Assuming that f and $f|_{\partial\Omega}$ are Morse functions, and $|\nabla f| \neq 0$ on $\partial\Omega$, we raise the following questions:

- What are the geometric conditions ensuring that, when $X_0 \sim \nu_h$, the law of X_{τ_Ω} concentrates on points where f attains its minimum on $\partial\Omega$ (or a subset of these points)?
- What are the conditions which ensure that these results extend to some deterministic initial conditions in Ω?

The results of [23] aim at answering these questions. Let us recall that when f and $f|_{\partial\Omega}$ are Morse functions and when $|\nabla f| \neq 0$ on $\partial\Omega$, the elements of the set

$$\{z \text{ is a local minimim of } f|_{\partial\Omega}\} \cap \{z \in \partial\Omega,\, \partial_n f(z) > 0\} \tag{42}$$

are the generalized saddle points of f on $\partial\Omega$ and play the role of saddle points of f on $\partial\Omega$, see Remark 9. Before stating the main results of [23], let us discuss the two questions above with one-dimensional examples.

Remark 12. The assumption that the drift term b in (1) is of the form $b = -\nabla f$ is essential here to the existence of a limiting exit distribution of Ω when $h \to 0$. If it is not the case and when for instance the boundary of Ω is a periodic orbit of the dynamics $\frac{d}{dt}x(t) = b\big(x(t)\big)$, the phenomenon of cycling discovered by Day in [17,18] prevents the existence of a limiting exit distribution when $h \to 0$. We also refer to [3–5] for the study of this phenomenon of cycling.

One-Dimensional Examples. To discuss the two questions raised in the previous section, one considers two one-dimensional examples.

Example 1. The goal is here to construct a one-dimensional example for which, starting from the global minimum of f in Ω or from the quasi-stationary distribution ν_h, the law of X_{τ_Ω} does not concentrate on points where f attains its minimum on $\partial\Omega$. To this end, let us consider the function f represented in Fig. 4 for which one has the following result.

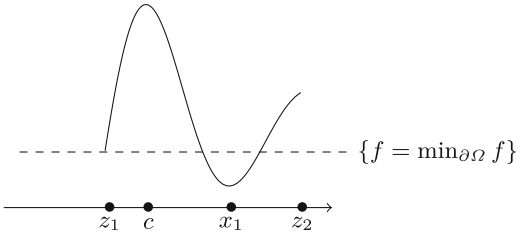


Fig. 4. Example of a function f such that, starting from the global minimum x_1 of f in Ω or from the quasi-stationary distribution ν_h, the law of X_{τ_Ω} concentrates on z_2 whereas $f(z_2) > \min_{\partial\Omega} f = f(z_1)$.

Proposition 5. *Let $z_1 < z_2$ and $f \in C^\infty([z_1, z_2], \mathbb{R})$ be a Morse function. Let us assume that $f(z_1) < f(z_2)$, $\{x \in [z_1, z_2], f'(x) = 0\} = \{c, x_1\}$ with $z_1 < c < x_1 < z_2$ and $f(x_1) < f(z_1) < f(z_2) < f(c)$ (see Fig. 4). Then, for all $x \in (c, z_2]$, there exists $\varepsilon > 0$ such that when $h \to 0$:*

$$\mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_1] = O(e^{-\frac{\varepsilon}{h}}) \text{ and thus } \mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_2] = 1 + O(e^{-\frac{\varepsilon}{h}}).$$

Moreover, there exists $\varepsilon > 0$ such that when $h \to 0$:

$$\mathbb{P}_{\nu_h}[X_{\tau_{(z_1,z_2)}} = z_1] = O(e^{-\frac{\varepsilon}{h}}) \quad \text{and thus} \quad \mathbb{P}_{\nu_h}[X_{\tau_{(z_1,z_2)}} = z_2] = 1 + O(e^{-\frac{\varepsilon}{h}}),$$

where ν_h is the quasi-stationary distribution of the process (1) in (z_1, z_2).

The proof of Proposition 5 is based on the fact that in one dimension, explicit formulas can be written for $x \mapsto \mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_j]$ ($j \in \{1, 2\}$), see [56, Section A.5.3.1] or [23]. According to Proposition 5, when $h \to 0$ and when $X_0 = x \in (c, z_2)$ or $X_0 \sim \nu_h$, the process (1) leaves $\Omega = (z_1, z_2)$ through z_2. However, the generalized saddle point z_2 (see (42)) is not the global minimum of f on $\partial\Omega$. This fact can be explained as follows: the potential barrier $f(c) - f(x_1)$ is larger than the potential barrier $f(z_2) - f(x_1)$. Thus, the law of X_{τ_Ω} when $X_0 = x \in (c, z_2)$ cannot concentrate on z_1 since it is less costly to leave Ω through z_2 rather than to cross the barrier $f(c) - f(x_1)$ to exit through z_1. Moreover, it can be proved that the quasi-stationary distribution ν_h concentrates in any neighborhood of x_1 in the limit $h \to 0$, which explains why the law of X_{τ_Ω} when $X_0 \sim \nu_h$ also concentrates on z_2. Concerning the two questions raised in the previous section, this example indicates that in the small temperature regime, there exist cases for which the process (1), starting from the global minimum of f in Ω or from ν_h, leaves Ω through a point which is not a global minimum of $f|_{\partial\Omega}$.

This example also suggests the following. If one wants the law of X_{τ_Ω} to concentrate when $h \to 0$ on points in $\partial\Omega$ where f attains its minimum, one should exclude cases when the largest timescales for the diffusion process in Ω are not related to energetic barriers involving points of $\partial\Omega$ where $f|_{\partial\Omega}$ attains its minimum. In order to exclude such cases, we will assume in the following that the closure of each of the connected components of $\{f < \min_{\partial\Omega} f\}$ intersects $\partial\Omega$.

Notice that if one modifies the function f in the vicinity of z_1 such that $\partial_n f(z_1) > 0$ and $\operatorname{argmin}_{\overline{\Omega}} f = \{x_1\}$, z_1 is then a generalized order one saddle point and the previous conclusions remain unchanged.

Example 2. Let us construct a one-dimensional example for which the concentration of the law of X_{τ_Ω} on $\operatorname{argmin}_{\partial\Omega} f$ is not the same starting from the global minima of f in Ω or from the quasi-stationary distribution ν_h. For this purpose, let us consider $z_1 > 0$, $z_2 := -z_1$, $z = 0$ and $f \in C^\infty([z_1, z_2], \mathbb{R})$ such that

$$f \text{ is a Morse and even function, } \{x \in [z_1, z_2],\, f'(x) = 0\} = \{x_1, z, x_2\}, \quad (43)$$

where

$$z_1 < x_1 < z < x_2 < z_2,\ f(z_1) = f(z_2) > f(x_1) = f(x_2),\ f(z_1) < f(z). \quad (44)$$

A function f satisfying (43) and (44) is represented in Fig. 5. One has the following result.

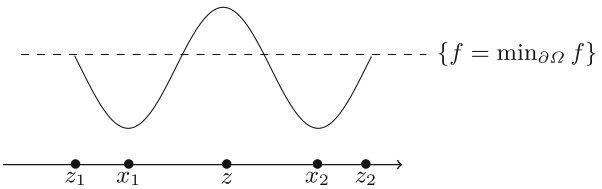


Fig. 5. One-dimensional example where (43) and (44) are satisfied.

Proposition 6. *Let $z_1 > 0$, $z_2 := -z_1$, $z = 0$ and $f \in C^\infty([z_1, z_2], \mathbb{R})$ which satisfies (43) and (44). Then, one has for all $h > 0$,*

$$\mathbb{P}_{\nu_h}[X_{\tau_{(z_1,z_2)}} = z_1] = \frac{1}{2} \quad \text{and} \quad \mathbb{P}_{\nu_h}[X_{\tau_{(z_1,z_2)}} = z_2] = \frac{1}{2}, \quad (45)$$

where ν_h is the quasi-stationary distribution of the process (1) in (z_1, z_2). Moreover, for all $x \in (z_1, z)$, there exists $c > 0$ such that when $h \to 0$,

$$\mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_1] = 1 + O(e^{-\frac{c}{h}}) \quad \text{and} \quad \mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_2] = O(e^{-\frac{c}{h}}), \quad (46)$$

and for all $x \in (z, z_2)$, there exists $c > 0$ such that when $h \to 0$

$$\mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_1] = O(e^{-\frac{c}{h}}) \quad \text{and} \quad \mathbb{P}_x[X_{\tau_{(z_1,z_2)}} = z_2] = 1 + O(e^{-\frac{c}{h}}). \quad (47)$$

The asymptotic estimate (45) is a consequence of the fact that f is an even function (see [23, Section 1]). The asymptotic estimates (46) and (47) are proved exactly as Proposition 5, see [23, Section 1]. Let us also mention that Proposition 6 is a consequence of the results [46]. Concerning the two questions raised

in the previous section, Proposition 6 shows that, when f satisfies (43) and (44), the concentration of the law of X_{τ_Ω} on $\{z_1, z_2\}$ is not the same starting from $x \in (z_1, z_2) \setminus \{z\}$ or from ν_h. This is due to the fact that in this case the quasi-stationary distribution ν_h has an equal repartition in all disjoint neighborhoods of x_1 and x_2, i.e. for every $(a_1, b_1) \subset (z_1, z)$ and $(a_2, b_2) \subset (z, z_2)$ such that $a_1 < x_1 < b_1$ and $a_2 < x_2 < b_2$, it holds for any $j \in \{1, 2\}$, $\lim_{h\to 0} \int_{a_j}^{b_j} \nu_h = \frac{1}{2}$ (see [46]). When $X_0 = x \in (z_1, z_2) \setminus \{z\}$, the asymptotic estimates (46) and (47) can be explained by the existence of a barrier $f(z) - f(x_1)$ which is larger than $f(z_1) - f(x_1)$. In order to exclude such cases, we will assume in the following that there exists a connected component C of $\{f < \min_{\partial\Omega} f\}$, such that $\operatorname{argmin}_{\overline{\Omega}} f \subset C$.

Main Results on the Exit Point Distribution. In this section, a simplified version of the results of [23] is presented. The aim is to exhibit a simple geometric setting for which, on the one hand, the law of X_{τ_Ω} concentrates on the same points of $\partial\Omega$ when $X_0 \sim \nu_h$ or $X_0 = x \in \Omega$ for some $x \in \{f < \min_{\partial\Omega} f\}$ and, on the other hand, this concentration occurs on generalized saddle points of f which belong to $\operatorname{argmin}_{\partial\Omega} f$. To this end, let us define the two following assumptions:

- **[H-Morse].** The function $f : \overline{\Omega} \to \mathbb{R}$ is C^∞. The functions $f : \overline{\Omega} \to \mathbb{R}$ and $f|_{\partial\Omega}$ are Morse functions. Moreover, $|\nabla f|(x) \neq 0$ for all $x \in \partial\Omega$.
- **[H-Min].** The open set $\{f < \min_{\partial\Omega} f\}$ is nonempty, contains all the local minima of f in Ω and the closure of each of the connected components of $\{f < \min_{\partial\Omega} f\}$ intersects $\partial\Omega$. Furthermore, there exists a connected component C of $\{f < \min_{\partial\Omega} f\}$ such that $\operatorname{argmin}_{\overline{\Omega}} f \subset C$.

Notice that under **[H-Morse]** and **[H-Min]**, it holds $\min_{\partial\Omega} f > \min_{\overline{\Omega}} f = \min_{\Omega} f$. Under the assumptions **[H-Morse]** and **[H-Min]**, one defines the set of points $\{z_1, \ldots, z_{k_0}\}$ by

$$\overline{C} \cap \partial\Omega = \{z_1, \ldots, z_{k_0}\}. \tag{48}$$

Remark 13. As already explained, the points $z_1, \ldots, z_{k_0}$ are generalized saddle points of f on $\partial\Omega$ (see (42)) since they satisfy

$$\{z_1, \ldots, z_{k_0}\} \subset \{z \in \partial\Omega,\, \partial_n f(z) > 0\} \cap \operatorname{argmin}_{\partial\Omega} f. \tag{49}$$

Remark 14. Under **[H-Min]**, the normal derivative of f can change sign and the function f can have saddle points in Ω higher than $\min_{\partial\Omega} f$, see for instance Fig. 6.

As shown in the following theorem, the assumption **[H-Min]** ensures that the quasi-stationary distribution ν_h concentrates in neighborhoods of the global minima of f in C and, starting from $x \in C$ or from ν_h, that the concentration of the law of X_{τ_Ω} when $h \to 0$ occurs on the set of generalized saddle points $\{z_1, \ldots, z_{k_0}\}$ (see (48)). Notice that the assumption **[H-Min]** is not satisfied in the two examples given in the previous section (see Figs. 4 and 5).

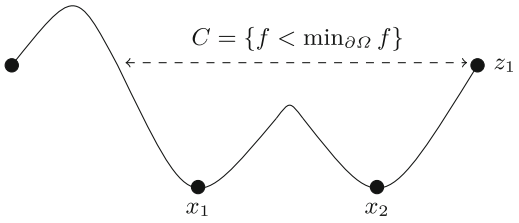


Fig. 6. A one-dimensional example where [**H-Morse**] and [**H-Min**] are satisfied, the normal derivative of f changes sign and the function f has a saddle point in Ω higher than $\min_{\partial\Omega} f$. In this example, $\{f < \min_{\partial\Omega} f\}$ is connected and thus $C = \{f < \min_{\partial\Omega} f\}$. Moreover, $\overline{C} \cap \partial\Omega = \{z_1\}$.

Theorem 2. *Let us assume that the hypotheses **[H-Morse]** and **[H-Min]** are satisfied. Let ν_h be the quasi-stationary distribution of the process (1) in Ω (see (14)). Let $\mathcal{V}$ be an open subset of Ω. Then, if $\mathcal{V} \cap \operatorname{argmin}_C f \neq \emptyset$, one has in the limit $h \to 0$:*

$$\nu_h(\mathcal{V}) = \frac{\sum_{x \in \mathcal{V} \cap \operatorname{argmin}_C f} \left(\det \operatorname{Hess} f(x)\right)^{-\frac{1}{2}}}{\sum_{x \in \operatorname{argmin}_C f} \left(\det \operatorname{Hess} f(x)\right)^{-\frac{1}{2}}} \left(1 + O(h)\right).$$

When $\overline{\mathcal{V}} \cap \operatorname{argmin}_C f = \emptyset$, there exists $c > 0$ such that when $h \to 0$:

$$\nu_h(\mathcal{V}) = O\left(e^{-\frac{c}{h}}\right).$$

In addition, let $F \in C^\infty(\partial\Omega, \mathbb{R})$. Then, when $h \to 0$:

$$\mathbb{E}_{\nu_h}\left[F\left(X_{\tau_\Omega}\right)\right] = \sum_{i=1}^{k_0} F(z_i)\, a_i + O(h^{\frac{1}{4}}), \tag{50}$$

where for $i \in \{1, \ldots, k_0\}$,

$$a_i = \frac{\partial_n f(z_i)}{\sqrt{\det \operatorname{Hess} f|_{\partial\Omega}(z_i)}} \left(\sum_{j=1}^{k_0} \frac{\partial_n f(z_j)}{\sqrt{\det \operatorname{Hess} f|_{\partial\Omega}(z_j)}} \right)^{-1}. \tag{51}$$

Finally, (50) holds when $X_0 = x \in C$.

Remark 15. In [23], one also gives sharp asymptotic estimates of λ_h and $\partial_n u_h$ in a more general setting than the one of Theorem 2 (for instance, we study the case when f has local minima higher than $\min_{\partial\Omega} f$). However, in [23], we do not study the precise asymptotic behaviour of X_{τ_Ω} when $h \to 0$ near generalized saddle points z of f on $\partial\Omega$ which are such that $f(z) > \min_{\partial\Omega} f$ as we did in [22] (see Corollary 1). Finally, in [23], the optimality of the remainder term $O(h^{\frac{1}{4}})$ in (50) is discussed and improved in some situations.

Ideas and Sketch of the Proof of Theorem 2. In this section, one gives the sketch of the proof of (50) which is the main result of Theorem 2. Recall that from (15), for $F \in C^\infty(\partial\Omega, \mathbb{R})$

$$\mathbb{E}_{\nu_h}\left[F(X_{\tau_\Omega})\right] = -\frac{h}{2\lambda_h}\frac{\displaystyle\int_\Sigma F\,\partial_n u_h e^{-\frac{2}{h}f}}{\displaystyle\int_\Omega u_h e^{-\frac{2}{h}f}},$$

where u_h is the eigenfunction associated with the principal eigenvalue λ_h of $L^{D,(0)}_{f,h}$. Therefore, to prove (50), one studies the asymptotic behaviour when $h \to 0$ of the following quantities

$$\lambda_h,\ \partial_n u_h \text{ and } \int_\Omega u_h e^{-\frac{2}{h}f}. \tag{52}$$

Under the assumptions [**H-Morse**] and [**H-Min**], one defines

$$m_0 := \text{Card}\left(\{z \in \Omega, z \text{ is a local minimum of } f\}\right)$$

and

$$\begin{aligned} m_1 :=& \text{Card}\left(\{z \text{ is a local minimum of } f|_{\partial\Omega}\} \cap \{z \in \partial\Omega,\ \partial_n f(z) > 0\}\right) \\ &+ \text{Card}\left(\{z \text{ is saddle point of } f\}\right). \end{aligned} \tag{53}$$

The integer m_1 is the number of generalized saddle points of f in $\overline{\Omega}$ (see [31, Section 5.2]). To study the asymptotic behaviour when $h \to 0$ of the quantities involved in (52), the starting point is to again observe that ∇u_h is solution to an eigenvalue problem for the same eigenvalue λ_h (as already explained at the end of Sect. 2.1). Indeed, ∇u_h is solution to (see (27))

$$\begin{cases} L^{(1)}_{f,h}\nabla u_h = \lambda_h \nabla u_h \text{ in } \Omega, \\ \nabla_T u_h = 0 \text{ on } \partial\Omega, \\ \left(-\frac{h}{2}\text{div} + \nabla f \cdot\right)\nabla u_h = 0 \text{ on } \partial\Omega, \end{cases} \tag{54}$$

where we recall that $L^{(1)}_{f,h} = -\frac{h}{2}\Delta + \nabla f \cdot \nabla + \text{Hess}\, f$ is an operator acting on 1-forms. Let us also recall that the operator $L^{(1)}_{f,h}$ with tangential boundary conditions (54) is denoted by $L^{D,(1)}_{f,h}$. From (54), ∇u_h is an eigenform of $L^{D,(1)}_{f,h}$ associated with λ_h.

The second ingredient is the following result: under the assumptions [**H-Morse**] and [**H-Min**] and when $h \to 0$, the operator $L^{D,(0)}_{f,h}$ has exactly m_0 eigenvalues smaller than $\frac{\sqrt{h}}{2}$ and $L^{D,(1)}_{f,h}$ has exactly m_1 eigenvalues smaller than $\frac{\sqrt{h}}{2}$ (see [31, Chapter 3]). Actually, all theses small eigenvalues are exponentially small when $h \to 0$, i.e. they are all $O\left(e^{-\frac{c}{h}}\right)$ for some $c > 0$ independent of h.

In particular λ_h is an exponentially small eigenvalue of $L^{D,(0)}_{f,h}$ and of $L^{D,(1)}_{f,h}$. Let us denote by $\pi^{(0)}_h$ (resp. $\pi^{(1)}_h$) the orthogonal projector in $L^2_w(\Omega)$ onto the m_0 (resp. m_1) smallest eigenvalues of $L^{D,(0)}_{f,h}$ (resp. $L^{D,(1)}_{f,h}$). Then, according to the foregoing, one has when $h \to 0$:

$$\dim \operatorname{Ran} \pi^{(0)}_h = m_0, \quad \dim \operatorname{Ran} \pi^{(1)}_h = m_1$$

and

$$\nabla u_h \in \operatorname{Ran} \pi^{(1)}_h.$$

Let us now explain how we prove Theorem 2. To this end, let us introduce the set of local minima of f in Ω,

$$\mathsf{U}_0^{\Omega} := \{x \in \Omega,\ x \text{ is a local minimum of } f\},$$

and the set of generalized saddle points of f in $\overline{\Omega}$,

$$\mathsf{U}_1^{\overline{\Omega}} = \Big(\{z \text{ is a local minimum of } f|_{\partial\Omega}\} \cap \{z \in \partial\Omega,\ \partial_n f(z) > 0\}\Big) \bigcup \{z \text{ is a saddle point of } f\}.$$

Let us recall that $m_0 = \operatorname{Card}\big(\mathsf{U}_0^{\Omega}\big)$ and, from (53), that $m_1 = \operatorname{Card}\big(\mathsf{U}_1^{\overline{\Omega}}\big)$. The first step to prove Theorem 2 consists in constructing two maps $\widetilde{\mathbf{j}}$ and $\mathbf{j}$. The goal of the map $\mathbf{j}$ is to associate each local minimum x of f with a set of generalized saddle points $\mathbf{j}(x) \subset \mathsf{U}_1^{\overline{\Omega}}$ such that

$$\forall z, y \in \mathbf{j}(x),\ f(z) = f(y),$$

and such that, in the limit $h \to 0$, there exists at least one eigenvalue of $L^{D,(0)}_{f,h}$ whose exponential rate of decay is $2\big(f(\mathbf{j}(x)) - f(x)\big)$ i.e.

$$\exists \lambda \in \sigma\big(L^{D,(0)}_{f,h}\big) \text{ such that } \lim_{h\to 0} h \ln \lambda = -2\big(f(\mathbf{j}(x)) - f(x)\big).$$

The aim of the map $\widetilde{\mathbf{j}}$ is to associate each local minimum x of f with the connected component of $\{f < f(\mathbf{j}(x))\}$ which contains x.

The second step consists in constructing bases of $\operatorname{Ran} \pi^{(0)}_h$ and $\operatorname{Ran} \pi^{(1)}_h$. To this end, one constructs two families of quasi-modes, denoted by $(\widetilde{u}_k)_{k\in\{1,\dots,m_0\}}$ and $(\widetilde{\psi}_j)_{j\in\{1,\dots,m_1\}}$, which are then respectively projected onto $\operatorname{Ran} \pi^{(0)}_h$ and $\operatorname{Ran} \pi^{(1)}_h$. To construct the family of 1-forms $(\widetilde{\psi}_j)_{j\in\{1,\dots,m_1\}}$, we proceed as follows. For each saddle point z of f in Ω, following the procedure of [30], one constructs a 1-form supported in a neighborhood of z in Ω. For a local minimum z of $f|_{\partial\Omega}$ such that $\partial_n f(z) > 0$, one constructs a 1-form supported in a neighborhood of z in $\overline{\Omega}$ as made in [31]. To construct the family of functions $(\widetilde{u}_k)_{k\in\{1,\dots,m_0\}}$, one constructs for each local minimum x of f a smooth function whose support is almost $\widetilde{\mathbf{j}}(x)$ (this construction is close to the one made in [30,31,36,45,53]).

The next step consists in finding a sharp asymptotic equivalent for λ_h when $h \to 0$. The quantity $\frac{2}{h}\lambda_h$ equals the square of the smallest singular values of the finite dimensional operator

$$\nabla : \mathrm{Ran}\,\pi_h^{(0)} \to \mathrm{Ran}\,\pi_h^{(1)}.$$

To study the asymptotic behaviour when $h \to 0$ of this smallest singular value, one uses the bases of $\mathrm{Ran}\,\pi_h^{(0)}$ and $\mathrm{Ran}\,\pi_h^{(1)}$ which have been constructed previously. The analysis of this finite dimensional problem is inspired by [36] and also yields the asymptotic equivalent of $\int_\Omega u_h\, e^{-\frac{2}{h}f}$ when $h \to 0$.

Then, we study the asymptotic behaviour of the normal derivative of u_h on $\partial\Omega$ when $h \to 0$ to deduce that the law of X_{τ_Ω} concentrates when $h \to 0$ on $\overline{C} \cap \partial\Omega = \{z_1, \ldots, z_{k_0}\}$ when $X_0 \sim \nu_h$.

Lastly, one proves "leveling" results on the function

$$x \mapsto \mathbb{E}_x[F(X_{\tau_\Omega})]$$

to obtain that when $X_0 = x \in C$, the law of X_{τ_Ω} also concentrates when $h \to 0$ on $\{z_1, \ldots, z_{k_0}\}$.

To conclude, the main results of [23] are the following:

1. One uses techniques from semi-classical analysis to study the asymptotic behaviours of λ_h and $\partial_n u_h$ when $h \to 0$, and then, the concentration of the law of X_{τ_Ω} on a subset of $\mathrm{argmin}_{\partial\Omega} f$ when $X_0 \sim \nu_h$.
2. One identifies the points of $\mathrm{argmin}_{\partial\Omega} f$ where the law of X_{τ_Ω} concentrates when $X_0 \sim \nu_h$: this set of points is $\{z_1, \ldots, z_{k_0}\}$. Moreover, explicit formulas for their relative probabilities are given (see indeed (51)) as well as precise remainder terms.
3. One extends the previous results on the law of X_{τ_Ω} to a deterministic initial condition in Ω: $X_0 = x$ where $x \in C$.
4. These results hold under weak assumptions on the function f and one-dimensional examples are given to explain why the geometric assumptions are needed to get them.

Conclusion. We presented recent results which justify the use of a kinetic Monte Carlo model parametrized by Eyring-Kramers formulas to model the exit event from a metastable state Ω for the overdamped Langevin dynamics (1). Our analysis is for the moment limited to situations where $|\nabla f| \neq 0$ on $\partial\Omega$, which does not allow to consider order one saddle points on $\partial\Omega$. The extensions of [22] and [23] which are currently under study are the following: the case when f has saddle points on $\partial\Omega$ and the case when the diffusion process $X_t = (q_t, p_t)$ is solution to the Langevin stochastic differential equation

$$\begin{cases} dq_t = p_t dt, \\ dp_t = -\nabla f(q_t)dt - \gamma\, p_t dt + \sqrt{h\gamma}\, dB_t, \end{cases}$$

where $(q_t, p_t) \in \Omega \times \mathbb{R}^d$, Ω being a bounded open subset of $\mathbb{R}^d$.

Acknowledgements. This work is supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement number 614492.

References

1. Aristoff, D., Lelièvre, T.: Mathematical analysis of temperature accelerated dynamics. Multiscale Model. Simul. **12**(1), 290–317 (2014)
2. Berglund, N.: Kramers' law: validity, derivations and generalisations. Markov Process. Relat. Fields **19**, 459–490 (2013)
3. Berglund, N.: Noise-induced phase slips, log-periodic oscillations, and the Gumbel distribution. Markov Process. Relat. Fields **22**, 467–505 (2016)
4. Berglund, N., Gentz, B.: On the noise-induced passage through an unstable periodic orbit I: two-level model. J. Stat. Phys. **114**(5–6), 1577–1618 (2004)
5. Berglund, N., Gentz, B.: On the noise-induced passage through an unstable periodic orbit II: general case. SIAM J. Math. Anal. **46**(1), 310–352 (2014)
6. Bovier, A., Eckhoff, M., Gayrard, V., Klein, M.: Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. J. Eur. Math. Soc. (JEMS) **6**, 399–424 (2004)
7. Bovier, A., Gayrard, V., Klein, M.: Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. J. Eur. Math. Soc. (JEMS) **7**, 69–99 (2005)
8. Cameron, M.: Computing the asymptotic spectrum for networks representing energy landscapes using the minimum spanning tree. Netw. Heterog. Media **9**(3), 383–416 (2014)
9. Champagnat, N., Villemonais, D.: General criteria for the study of quasi-stationarity. ArXiv preprint arXiv:1712.08092 (2017)
10. Chandrasekhar, S.: Stochastic problems in physics and astronomy. Rev. Mod. Phys. **15**(1), 1 (1943)
11. Davies, E.B.: Dynamical stability of metastable states. J. Funct. Anal. **46**(3), 373–386 (1982)
12. Davies, E.B.: Metastable states of symmetric Markov semigroups I. Proc. London Math. Soc. **45**(3), 133–150 (1982)
13. Davies, E.B.: Metastable states of symmetric Markov semigroups II. J. London Math. Soc. **26**(3), 541–556 (1982)
14. Day, M.V.: On the exponential exit law in the small parameter exit problem. Stochastics **8**(4), 297–323 (1983)
15. Day, M.V.: On the asymptotic relation between equilibrium density and exit measure in the exit problem. Stoch. Int. J. Probab. Stoch. Process. **12**(3–4), 303–330 (1984)
16. Day, M.V.: Recent progress on the small parameter exit problem. Stoch. Int. J. Probab. Stoch. Process. **20**(2), 121–150 (1987)
17. Day, M.V.: Conditional exits for small noise diffusions with characteristic boundary. Ann. Probab. **20**(3), 1385–1419 (1992)
18. Day, M.V.: Exit cycling for the van der Pol oscillator and quasipotential calculations. J. Dyn. Diff. Equat. **8**(4), 573–601 (1996)
19. Day, M.V.: Mathematical approaches to the problem of noise-induced exit. In: McEneaney, W.M., Yin, G.G., Zhang, Q. (eds.) Stochastic Analysis, Control, Optimization and Applications, pp. 269–287. Birkhäuser, Boston (1999)

20. Devinatz, A., Friedman, A.: Asymptotic behavior of the principal eigenfunction for a singularly perturbed Dirichlet problem. Indiana Univ. Math. J. **27**, 143–157 (1978)
21. Devinatz, A., Friedman, A.: The asymptotic behavior of the solution of a singularly perturbed Dirichlet problem. Indiana Univ. Math. J. **27**(3), 527–537 (1978)
22. Di Gesù, G., Lelièvre, T., Le Peutrec, D., Nectoux, B.: Sharp asymptotics of the first exit point density. Ann. PDE **5**(1) (2019). https://link.springer.com/journal/40818/5/1
23. Di Gesù, G., Lelièvre, T., Le Peutrec, D., Nectoux, B.: The exit from a metastable state: concentration of the exit point distribution on the low energy saddle points. ArXiv preprint arXiv:1902.03270 (2019)
24. Eckhoff, M.: Precise asymptotics of small eigenvalues of reversible diffusions in the metastable regime. Ann. Probab. **33**(1), 244–299 (2005)
25. Fan, Y., Yip, S., Yildiz, B.: Autonomous basin climbing method with sampling of multiple transition pathways: application to anisotropic diffusion of point defects in hcp Zr. J. Phys. Condens. Matter **26**, 365402 (2014)
26. Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems. Springer, Heidelberg (1984)
27. Galves, A., Olivieri, E., Vares, M.E.: Metastability for a class of dynamical systems subject to small random perturbations. Ann. Probab. **15**(4), 1288–1305 (1987)
28. Gol'dshtein, V., Mitrea, I., Mitrea, M.: Hodge decompositions with mixed boundary conditions and applications to partial differential equations on Lipschitz manifolds. J. Math. Sci. **172**(3), 347–400 (2011)
29. Hänggi, P., Talkner, P., Borkovec, M.: Reaction-rate theory: fifty years after Kramers. Rev. Mod. Phys. **62**(2), 251–342 (1990)
30. Helffer, B., Klein, M., Nier, F.: Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. Mat. Contemp. **26**, 41–85 (2004)
31. Helffer, B., Nier, F.: Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach: the case with boundary. Mémoire de la Société mathématique de France **105**, 1–89 (2006)
32. Helffer, B., Sjöstrand, J.: Multiple wells in the semi-classical limit I. Comm. Partial Diff. Equat. **9**(4), 337–408 (1984)
33. Helffer, B., Sjöstrand, J.: Multiple wells in the semi-classical limit III-Interaction through non-resonant wells. Mathematische Nachrichten **124**(1), 263–313 (1985)
34. Helffer, B., Sjöstrand, J.: Puits multiples en limite semi-classique. II. Interaction moléculaire. Symétries. Perturbation. Annales de l'IHP Physique théorique **42**(2), 127–212 (1985)
35. Helffer, B., Sjöstrand, J.: Puits multiples en mecanique semi-classique iv etude du complexe de Witten. Comm. Partial Diff. Equat. **10**(3), 245–340 (1985)
36. Hérau, F., Hitrik, M., Sjöstrand, J.: Tunnel effect and symmetries for Kramers-Fokker-Planck type operators. J. Inst. Math. Jussieu **10**(3), 567–634 (2011)
37. Holley, R.A., Kusuoka, S., Stroock, D.W.: Asymptotics of the spectral gap with applications to the theory of simulated annealing. J. Funct. Anal. **83**(2), 333–347 (1989)
38. Jakab, T., Mitrea, I., Mitrea, M.: On the regularity of differential forms satisfying mixed boundary conditions in a class of Lipschitz domains. Indiana Univ. Math. J. **58**(5), 2043–2071 (2009)
39. Kamin, S.: Elliptic perturbation of a first order operator with a singular point of attracting type. Indiana Univ. Math. J. **27**(6), 935–952 (1978)

40. Kamin, S.: On elliptic singular perturbation problems with turning points. SIAM J. Math. Anal. **10**(3), 447–455 (1979)
41. Kipnis, C., Newman, C.M.: The metastable behavior of infrequently observed, weakly random, one-dimensional diffusion processes. SIAM J. Appl. Math. **45**(6), 972–982 (1985)
42. Kramers, H.A.: Brownian motion in a field of force and the diffusion model of chemical reactions. Physica **7**(4), 284–304 (1940)
43. Landim, C., Mariani, M., Seo, I.: Dirichlet's and Thomson's principles for non-selfadjoint elliptic operators with application to non-reversible metastable diffusion processes. Arch. Ration. Mech. Anal. **231**(2), 887–938 (2019)
44. Le Bris, C., Lelièvre, T., Luskin, M., Perez, D.: A mathematical formalization of the parallel replica dynamics. Monte Carlo Methods Appl. **18**(2), 119–146 (2012)
45. Le Peutrec, D.: Small eigenvalues of the Neumann realization of the semiclassical Witten Laplacian. Ann. Fac. Sci. Toulouse Math. (6) **19**(3–4), 735–809 (2010)
46. Le Peutrec, D., Nectoux, B.: Repartition of the quasi-stationary distribution and first exit point density for a double-well potential. ArXiv preprint arXiv:1902.06304 (2019)
47. Marcelin, R.: Contribution à l'étude de la cinétique physico-chimique. Ann. Phys. **3**, 120–231 (1915)
48. Mathieu, P.: Zero white noise limit through Dirichlet forms, with application to diffusions in a random medium. Probab. Theory Relat. Fields **99**(4), 549–580 (1994)
49. Mathieu, P.: Spectra, exit times and long time asymptotics in the zero-white-noise limit. Stochastics **55**(1–2), 1–20 (1995)
50. Matkowsky, B.J., Schuss, Z.: The exit problem for randomly perturbed dynamical systems. SIAM J. Appl. Math. **33**(2), 365–382 (1977)
51. Matkowsky, B.J., Schuss, Z.: The exit problem: a new approach to diffusion across potential barriers. SIAM J. Appl. Math. **36**(3), 604–623 (1979)
52. Matkowsky, B.J., Schuss, Z., Ben-Jacob, E.: A singular perturbation approach to Kramers diffusion problem. SIAM J. Appl. Math. **42**(4), 835–849 (1982)
53. Michel, L.: About small eigenvalues of Witten Laplacian. Pure Appl. Anal. (2017, to appear)
54. Miclo, L.: Comportement de spectres d'opérateurs de Schrödinger à basse température. Bulletin des sciences mathématiques **119**(6), 529–554 (1995)
55. Naeh, T., Klosek, M.M., Matkowsky, B.J., Schuss, Z.: A direct approach to the exit problem. SIAM J. Appl. Math. **50**(2), 595–627 (1990)
56. Nectoux, B.: Analyse spectrale et analyse semi-classique pour la métastabilité en dynamique moléculaire. Ph.D. thesis, Université Paris Est (2017)
57. Perthame, B.: Perturbed dynamical systems with an attracting singularity and weak viscosity limits in Hamilton-Jacobi equations. Trans. Am. Math. Soc. **317**(2), 723–748 (1990)
58. Schilder, M.: Some asymptotic formulas for Wiener integrals. Trans. Am. Math. Soc. **125**(1), 63–85 (1966)
59. Schütte, C.: Conformational dynamics: modelling, theory, algorithm and application to biomolecules. Habilitation dissertation, Free University Berlin (1998)
60. Schütte, C., Sarich, M.: Metastability and Markov State Models in Molecular Dynamics. Courant Lecture Notes, vol. 24. American Mathematical Society, Providence (2013)
61. Sorensen, M.R., Voter, A.F.: Temperature-accelerated dynamics for simulation of infrequent events. J. Chem. Phys. **112**(21), 9599–9606 (2000)
62. Sugiura, M.: Metastable behaviors of diffusion processes with small parameter. J. Math. Soc. Jpn. **47**(4), 755–788 (1995)

63. Vineyard, G.H.: Frequency factors and isotope effects in solid state rate processes. J. Phys. Chem. Solids **3**(1), 121–127 (1957)
64. Voter, A.F.: A method for accelerating the molecular dynamics simulation of infrequent events. J. Chem. Phys. **106**(11), 4665–4677 (1997)
65. Voter, A.F.: Parallel replica method for dynamics of infrequent events. Phys. Rev. B **57**(22), R13985 (1998)
66. Voter, A.F.: Introduction to the kinetic Monte Carlo method. In: Sickafus, K.E., Kotomin, E.A., Uberuaga, B.P. (eds.) Radiation Effects in Solids. Springer, NATO Publishing Unit, Dordrecht (2005)
67. Wales, D.J.: Energy Landscapes. Cambridge University Press, Cambridge (2003)

Collisional Relaxation and Dynamical Scaling in Multiparticle Collisions Dynamics

Stefano Lepri[1,5(✉)], Hugo Bufferand[3], Guido Ciraolo[3], Pierfrancesco Di Cintio[2,5], Philippe Ghendrih[3], and Roberto Livi[1,4,5]

[1] Consiglio Nazionale delle Ricerche, Istituto dei Sistemi Complessi, via Madonna del piano 10, 50019 Sesto Fiorentino, Italy
stefano.lepri@isc.cnr.it
[2] Consiglio Nazionale delle Ricerche, Istituto di Fisica Applicata "Nello Carrara", via Madonna del piano 10, 50019 Sesto Fiorentino, Italy
p.dicintio@ifac.cnr.it
[3] CEA, IRFM, 13108 Saint-Paul-lez-Durance, France
{hugo.bufferand,guido.ciraolo,philippe.ghendrih}@cea.fr
[4] Dipartimento di Fisica e Astronomia and CSDC, Universitá di Firenze, via G. Sansone 1, 50019 Sesto Fiorentino, Italy
[5] Istituto Nazionale di Fisica Nucleare, Sezione di Firenze, via G. Sansone 1, 50019 Sesto Fiorentino, Italy
livi@fi.infn.it

Abstract. We present the Multi-Particle-Collision (MPC) dynamics approach to simulate properties of low-dimensional systems. In particular, we illustrate the method for a simple model: a one-dimensional gas of point particles interacting through stochastic collisions and admitting three conservation laws (density, momentum and energy). Motivated from problems in fusion plasma physics, we consider an energy-dependent collision rate that accounts for the lower collisionality of high-energy particles. We study two problems: (i) the collisional relaxation to equilibrium starting from an off-equilibrium state and (ii) the anomalous dynamical scaling of equilibrium time-dependent correlation functions. For problem (i), we demonstrate the existence of long-lived population of suprathermal particles that propagate ballistically over a quasi-thermalized background. For (ii) we compare simulations with the predictions of nonlinear fluctuating hydrodynamics for the structure factors of density fluctuations. Scaling analysis confirms the prediction that such model belong to the Kardar-Parisi-Zhang universality class.

Keywords: Multi-particle collision simulation · Anomalous transport

1 Introduction

Simulation of many-particle systems can be computationally very demanding, even for simple models. This is challenging especially when trying to measure

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 364–374, 2019.
https://doi.org/10.1007/978-3-030-15096-9_10

asymptotic properties like the celebrated long-time tails of correlation functions in the thermodynamic limit [1]. Although molecular dynamics is the most natural choice, alternative approaches based on effective stochastic processes have been proposed both for computational efficiency and also to get some insight in the general properties of non-equilibrium systems. In this contribution we will briefly review the Multi-Particle-Collision (MPC) approach which was originally proposed by Malevanets and Kapral [2–4] in the context of mesoscopic dynamics of complex fluids (e.g. polymers in solution, colloidal fluids). In essence, it is based on a stochastic and *local* protocol that redistributes particle velocities, while preserving the global conserved quantities such as total energy, momentum and angular momentum.

In this contribution, we will illustrate the method referring to the simple case of a one-dimensional fluid. Since we are interested to explore possible application of the method as a tool to investigate fusion plasma, we will introduce an energy-dependent collision rate that mimics Coulombian interaction in a simple manner. We will consider two problems: (i) the relaxation to equilibrium from a non-equilibrium initial state and (ii) the demonstration or dynamical scaling of time-dependent correlation functions.

Thermalization of many-particle system is a classic problem of non-equilibrium statistical mechanics and kinetic theory. In the context of fusion plasma, the question is relevant in the low-collisionality regime where non-equilibrium condition generate populations of suprathermal electrons and heavy tails in the velocity distribution function [5,6]. These fast particles modify heat and charge transport and thus the overall performance of magnetic confinement devices [7].

On the other hand, transport and dynamical scaling in low-dimensional models have been long investigated in the recent literature [8–10]. The main findings is that many-particle systems with one or two spatial degrees of freedom show anomalous transport properties signaled by the divergence of transport coefficients, like the thermal conductivity in the thermodynamic limit [8–11]. A way to detect this anomalous feature is to study dynamical scaling of equilibrium correlation functions and the corresponding dynamical scaling exponent z (defined below) and seek for deviations from the usual diffusive behavior. More recently, a complete description has been put forward within the Nonlinear Fluctuating Hydrodynamics (NFH) approach, proposed independently by van Beijeren [12] and Spohn [13,14]. These authors have shown that the statistical properties of 1D nonlinear hydrodynamics with three conservation laws (e.g. total energy, momentum and number of particles) are essentially described by the fluctuating Burgers equation which can be mapped onto the well-known Kardar-Parisi-Zhang (KPZ) equation for the stochastic growth of interfaces [15]. As a consequence, correlations of spontaneous fluctuations are characterized by the KPZ dynamical exponent $z = 3/2$ in one-dimension. The origin of the nontrivial dynamical exponents are to be traced back to the nonlinear interaction of long-wavelength modes. The results depends on the fact that the isolated system admits three conserved quantities whose fluctuations are coupled. Models with a different number of conservation laws (e.g. two like in Ref. [16]) may belong to other universality classes characterized by

different dynamic exponents. A generalization to a arbitrary number of conserved quantities has been discussed recently [17].

2 Multi-Particle-Collision Method

The MPC simulation scheme (see Refs. [18,19] for a detailed review) consists essentially in partitioning the system of N_p particles in N_c cells where the local center of mass coordinates and velocity are computed and rotating particle velocities in the cell's center of mass frame are around a random axis. The rotation angles are assigned in a way that the invariant quantities are locally preserved (see e.g. [18,20]). All particles are then propagated freely, or under the effect of an external force, if present.

In the case of the one-dimensional fluid we are interested in, the above steps can be carried on as follows [21]. Let us denote by m_j and v_j the mass and velocity of the j-th particle and by N_i and the instantaneous number of particles inside each cell i on which the system is coarse grained. The collision step amounts to assign random values to the velocities inside each cell, under the constraint of conserving, besides the particle number, the linear momentum P_i and the kinetic energy K_i. In practice, we extract random samples w_j from a Maxwellian distribution at the kinetic temperature of each cell, and let $v_{j,\mathrm{old}} \to v_{j,\mathrm{new}} = a_i w_j + b_i$, where a_i and b_i are the unknown cell-dependent coefficients determined by the conditions

$$\begin{aligned} P_i &= \sum_{j=1}^{N_i} m_j v_{j,\mathrm{old}} = \sum_{j=1}^{N_i} m_j v_{j,\mathrm{new}} = \sum_{j=1}^{N_i} m_j (a_i w_j + b_i); \\ K_i &= \sum_{j=1}^{N_i} m_j \frac{v_{j,\mathrm{old}}^2}{2} = \sum_{j=1}^{N_i} m_j \frac{v_{j,\mathrm{new}}^2}{2} = \sum_{j=1}^{N_i} m_j \frac{(a_i w_j + b_i)^2}{2}, \end{aligned} \tag{1}$$

Equations (1) constitute a system that can be solved for a_i and b_i analytically [20,21]. Finally, the propagation step on the positions r_j for a preassigned time interval Δt is operated and the procedure repeats.

The above collision procedure assumes implicitly that the velocity exchange is an instantaneous process that is not mediated by an effective potential. A further physical ingredient can be added assuming that the collision occurs at a given rate chosen to mimic some feature of the microscopic interaction. An interesting example is encountered in the modelization of plasmas of charged particles were the rate can be fixed to capture the essence of the Coulombian scattering at low impact parameters (i.e. of the order of the cell size) [22]. In the simulations presented here, we perform the above interaction step with a cell-dependent Coulomb-like interaction probability [20,21,23]

$$\mathcal{P}_i = \frac{1}{1 + \Gamma_i^{-2}}, \tag{2}$$

where Γ_i is the plasma coupling parameter computed in cell i, relating the average Coulomb energy and the thermal energy $N_i k_B T_i = 2K_i$, defined by

$$\Gamma_i = \frac{q^2}{4\pi\epsilon_0 a k_B T_i}. \tag{3}$$

In the expression above q is the particles charge, and a a mean inter-particle distance related to the inverse of average number density $\bar{n}$ and ϵ_0 is the vacuum permittivity.

Since the scope of this paper is to study and compare transport in low-dimensional models, we mostly limit ourselves to consider only one dimensional plasmas in a static neutralizing background with charge density ρ_b. In conditions where the neutrality is violated (e.g. when the number density n is no longer uniform), the self-consistent electrostatic potential Φ can be included by simultaneously solving the 1D Poisson equation

$$\nabla^2 \Phi(r) = -(qn(r) + \rho_b(r))/\epsilon_0 \tag{4}$$

by some standard finite-differences method. The resulting electric field is used to propagate the particles between each collision step. The dynamics can be further generalized to higher-dimensional charged fluids in a straightforward manner. For instance, in Ref. [24] a study of two-dimensional case has been considered in detail. Moreover, the effects of the electromagnetic fields in higher dimensions can be implemented via particle-mesh schemes solving self-consistently the Maxwell equations on the grid.

3 Relaxation to Equilibrium

In this section we present simulations of collisional relaxation from non-thermal initial states towards equilibrium. In a first set of numerical experiments we study the evolution of systems characterized by so-called waterbag initial conditions, whereby, positions and velocities r and v are initially distributed according to a phase-space distribution function of the form

$$f_0(r, v) = \mathcal{C} n \Theta(v_m - |v|). \tag{5}$$

In the expression above, $\Theta(x)$ is the Heaviside step function, n is the particle number density constant over the periodic simulation domain $[0, L]$, and the normalization constant $\mathcal{C}$ is defined by the condition

$$\int_0^L \int_{-v_m}^{v_m} f_0(r, v) \mathrm{d}v \mathrm{d}r = 1. \tag{6}$$

In all simulations presented hereafter the times are expressed in units of $t_* = 2\pi/\Omega_P$, where $\Omega_P = \sqrt{q^2 n/\epsilon_0 m}$ is the plasma frequency of the system and we have used units such that $q = m = \epsilon_0 = k_B = 1$. To quantify the collisionality

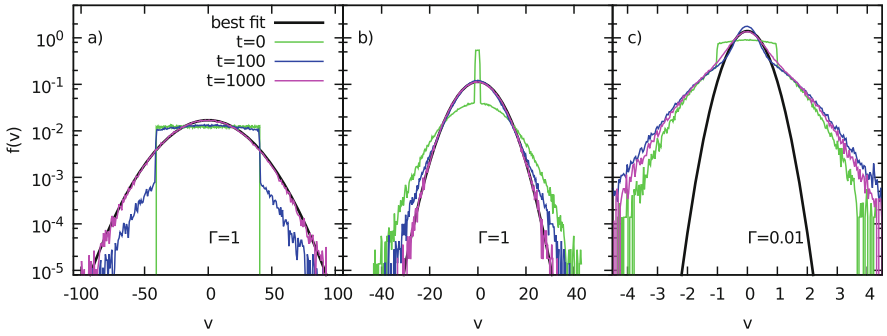


Fig. 1. Velocity distribution $f(v)$ at different times (thin lines) for a system of $N_p = 2.5 \times 10^5$ particles with waterbag initial conditions and $\Gamma = 1$ (panel a), a system initially represented by the sum of a thermal and a waterbag distribution with an average $\Gamma = 1$ (panel b), and a system initially represented by the sum of a thermal and a waterbag distribution with an average $\Gamma = 0.01$ (panel c). In all cases the heavy solid line marks the best fit thermal distribution in the final state (in all cases $t = 1000$).

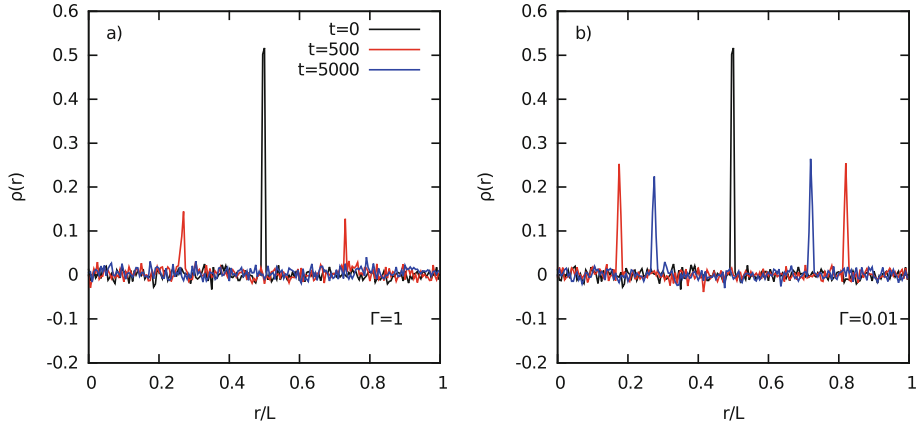


Fig. 2. Evolution of the charge density $\rho(r)$ for models characterized by an initially bunched supra-thermal population in a thermalized background with $\Gamma = 1$ (panel a), and $\Gamma = 0.01$ (panel b).

level within the fluid and compare different simulation protocols, we define the global parameter Γ as the average of the Γ_i over all cells evaluated at $t = 0$.

In Fig. 1, panel (a) we show the evolution of the velocity distribution $f(v)$ for a system with a combination of density and kinetic energy yielding an average coupling parameter $\Gamma = 1$, and starting with a waterbag distribution. The (multi-particle) collisions gradually evolve $f(v)$ towards a Gaussian distribution (marked in figure by the heavy solid) while, due to the imposed neutrality, the tiny fluctuations of Φ play virtually no role. Remarkably, at intermediate times

(here, $t = 100$) $f(v)$ is characterized by high velocity thermal tails, while the bulk of the distribution for $-40 < v < 40$ still bears memory of the step-like initial $f_0(v)$. For larger values of Γ (not shown here), $f(v)$ converges more and more rapidly to a thermal distribution. On the contrary, in the low-collisionality regime, for $\Gamma < 0.01$, the equilibrium is hardly reached on the simulation time. Indeed, particles in the high-velocity tails tend to decouple those belonging to the rest of the distribution and perform an almost ballistic motion.

In a second set of experiments we have studied the evolution of non-thermal populations in an already thermalized background system. In panels (b) and (c) of Fig. 1 we again show the evolution of $f(v)$ starting from an initial state constituted by two components one with f_0 given by Eq. (3) and another with

$$f_0(r, v) = \mathcal{C} n \exp(-v^2/2\sigma^2). \tag{7}$$

While in the moderately coupled case (panel b, $\Gamma = 1$) the two populations rapidly equilibrate, in the weakly coupled case (panel c, $\Gamma = 0.1$) the final ($t = 1000$) total velocity distribution features a low velocity region well fitted by a Gaussian distribution (heavy solid line) and high velocity power-law fat tails.

Moreover, we have also tested the stability of an initially localized bunch of mono-energetic particles in a thermalized background. In this set of numerical experiments, the initial conditions for the two populations were sampled from a thermal distribution like that of Eq. (7), and a spatially bunched distribution of the form

$$f_0(r, v) = \mathcal{C} \exp\left[-(r - r_*)^2/2s^2\right] [\delta(v - v_*) + \delta(v + v_*)], \tag{8}$$

where $\delta(x)$ is the Dirac delta function, r_* is the centroid of the bunch, s its width, and v_* its velocity.

In Fig. 2 we show the evolution of the charge density profile for an initially localized charge bunch placed in a periodic system of 10^6 particles with $\Gamma = 1$ (panel a) and 0.01 (panel b). In both cases, half of the 10^4 bunch particles are initialized according to Eq. (8) with $v = v_* = 5\sigma$ and the other half with $v = -v_*$ in a gaussian bunch with $s = L/100$. As expected, in the less collisional cases ($\Gamma = 0.01$, panel b), the bunch particles do not mix with the thermal background and (the two halves of) the bunch remain essentially coherent (at least for $t < 10^4$, the simulation time), while for a more collisional system ($\Gamma = 1$) the bunch is already completely dispersed at $t = 5000$ by the interplay of collisions and mean field effects.

4 Dynamical Scaling

As mentioned in the introduction, we are also interested in the scaling properties of time-dependent correlation functions evaluated in some equilibrium ensemble (typically the microcanonical one). In the simulations aimed at computing equilibrium correlation functions, the initial conditions on position and velocity are extracted from Eq. (7) and a uniform neutralizing background is assumed.

For such distribution the local coupling parameters, Eq. (2), are basically uniform over the whole system $\Gamma_i \approx \Gamma$. For this reason, and in order to save computational time, Γ is evaluated at the beginning of the simulation and used as the single control parameter. Moreover, as in the limit of neutrality the electrostatic field vanishes, we do not solve Eq. (4) and we simply impose $\nabla\Phi(r) = 0$.

The observables we will focus are the dynamical structure factors of the conserved quantities defined at the resolution set by the cell partition. Denoting ξ_l as a shorthand notation for energy, momentum or density in the l-th cell ($\mathcal{E}, P, \rho$ respectively). It is defined by first performing the discrete space-Fourier transform

$$\hat{\xi}(k,t) = \frac{1}{N_c}\sum_{l=1}^{N_c} \xi_l \exp(-\imath k l). \tag{9}$$

The dynamical structure factors $S_\xi(k,\omega)$ are defined as the modulus squared of the subsequent temporal Fourier transform

$$S_\xi(k,\omega) = \langle |\hat{\xi}(k,\omega)|^2 \rangle. \tag{10}$$

Since we are working with periodic boundary conditions, the allowed values of the wave number k are always integer multiples of $2\pi/N$, therefore in the rest of the paper we will sometimes refer to the (integer) normalized wave number $\tilde{k} = kN/2\pi$.

To connect with transport problems, we also considered the correlation function of the currents J_ξ, associated to the conserved quantity ξ. As above we choose to define the currents on the simulation grid

$$J_\xi(t) = \sum_{i=1}^{N_c} \left[\xi_i'(t) - \xi_{i-1}'(t - \Delta t)\right]. \tag{11}$$

Here, the prime is a shorthand notation to remind that only particles who moved from cell $i-1$ to i between successive time steps must be considered in each term of the sum. We thus computed $C_\xi = \langle |\tilde{J}_\xi(\omega)|^2 \rangle$ where the tilde denotes the Fourier transform in the time domain.

According to the NFH theory [13], long-wavelength fluctuations are described in terms of hydrodynamic modes: in a system with three conserved quantities like chains of coupled oscillators with momentum conservation, the linear theory would yield two propagating sound modes and one diffusing heat mode, all of the three diffusively broadened. Nonlinear terms can be added and treated within the mode-coupling approximation [13,25] that predicts that, at long times, the sound mode correlations satisfy Kardar-Parisi-Zhang scaling, while the heat mode correlations follow a Lévy-walk scaling. As a consequence, it is expected that S_ξ should be a combination of three modes correlations. For instance, for $k \to 0$, $S_\rho(k,\omega)$ should display sharp peaks at $\omega = \pm\omega_{\max}(k)$ that correspond to the propagation of sound modes and for $\omega \approx \pm\omega_{\max}$ it should behave as

$$S_\rho(k,\omega) \sim f_{\mathrm{KPZ}}\left(\frac{\omega \pm \omega_{\max}}{\lambda_s k^{3/2}}\right). \tag{12}$$

Remarkably, the scaling function f_{KPZ} is universal and known exactly [13] albeit not known in a closed form so that one has to be evaluate it numerically [26]. The nonuniversal coefficients λ_s are model-dependent and, in principle, can be evaluated in terms of static correlators [13].

Another relevant signature of anomalous transport is the presence of long-time tails in the correlations or, equivalently, of a low frequency singularity. For instance, it is expected that $C_{\mathcal{E}}$ should diverge, in the large-size and low-frequency limits, as $\omega^{-1/3}$ [8–10].

For a chain of coupled anharmonic oscillators with three conserved quantities like the Fermi-Pasta-Ulam chain, such theoretical predictions have been successfully compared with the numerics [21,27]. Other positive tests have been reported in Ref. [28]. We have performed a series of numerical test for the MPC dynamics presented above. Figure 3 shows the structure factors of density and energy for two strongly collisional cases with $\Gamma = 10$ and 2, corresponding to relatively strong collisionality. Within statistical fluctuations the data display a good data collapse and the lineshape fits with the KPZ-scaling function as predicted by Eq. (12). It should be also mentioned that the same type of agreement has been shown to hold also for quasi-one-dimensional MPC dynamics, namely in the case of a fluid confined in a box with a relatively large aspect ratio [24]. Another prediction of NFH is that the energy structure factors should display a so-called Lévy peak at zero frequency [14]. However, the data reported in [21] (see in particular Fig. 6) show that the contribution of the sound modes is pretty large, thus hindering the direct test of the prediction at least on the timescales of such simulations.

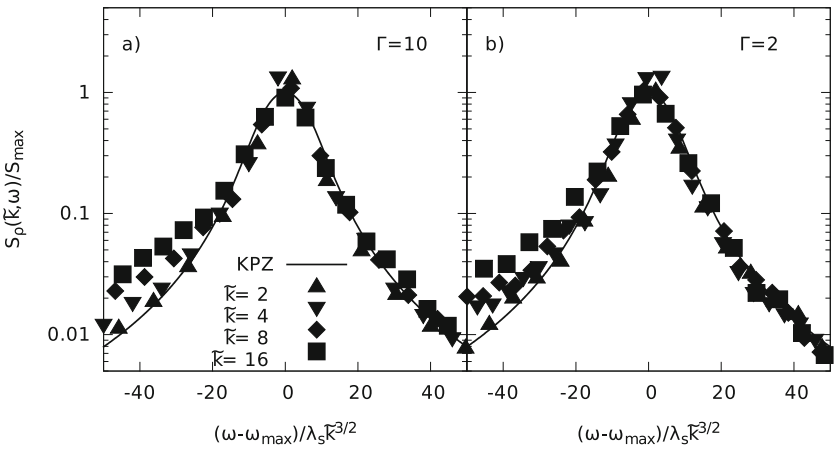


Fig. 3. Data collapse of the number density structure factors to the KPZ scaling function (solid line) of the Fourier spectra of the density profile modes with normalized wave number $\tilde{k} = 2$, 4, 8, and 16, for $\Gamma = 10$ (panel a), and 2 (panel b).

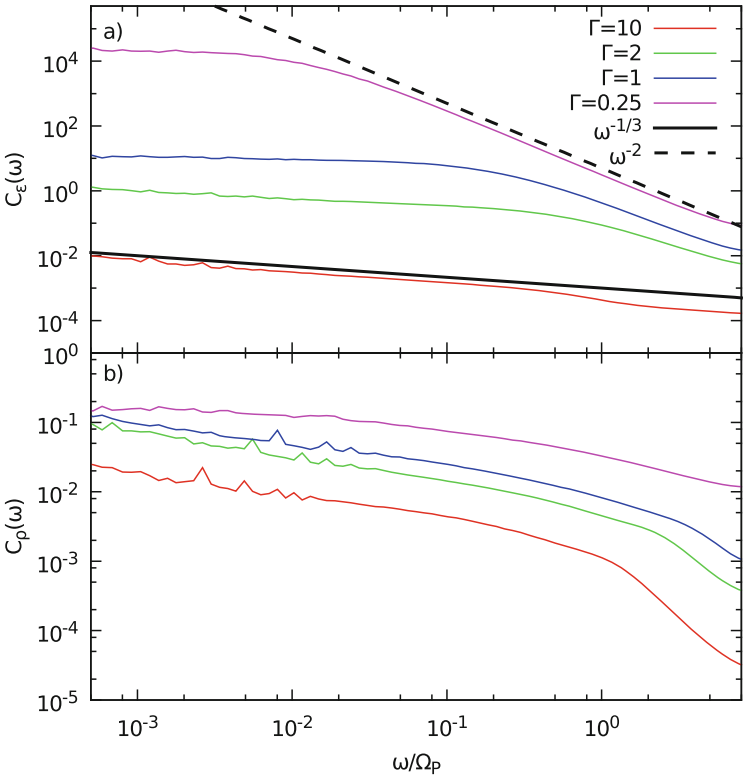


Fig. 4. For thermalized systems with $\Gamma = 10$, 2, 1 and 0.25: Fourier spectra $C_{\mathcal{E}}$ of the energy current (panel a) and Fourier spectra of the charge density current C_ρ (panel b). The curves are averaged over 200 independent realizations. In all cases, the frequency ω is rescaled to the plasma frequency. The cross-over from the $\omega^{-1/3}$ to the ω^{-2} behavior of $C_{\mathcal{E}}$ at around $\Gamma = 2$ is evident. To guide the eye, the dashed and solid black curves with the two slopes $-1/3$ and -2 have been added to the plot.

In Fig. 4 we present the Fourier spectra $C_{\mathcal{E}}$ (panel a), and C_ρ (panel b) of the energy and density currents, respectively, for four typical values of the ratio $\Gamma = 10$, 2, 1 and 0.25, and for $N_p = 12000$ particles distributed on $N_c = 1200$ cells. For strongly interacting systems (i.e. $\Gamma \geq 10$) one recovers the $\omega^{-1/3}$ behavior of the energy correlator $C_{\mathcal{E}}$. Increasing the particle specific kinetic $k_B T$ energy at fixed density n (i.e. reducing Γ and the collisionality of the system), $C_{\mathcal{E}}$ shows a more and more prominent flat region at low frequencies departing form the $\omega^{-1/3}$ trend, and a high frequency tail with slope ω^{-2}. The cross-over from the $\omega^{-1/3}$ to the ω^{-2} behavior of $C_{\mathcal{E}}$ is evident at around $\Gamma = 2$. A different behavior is instead found for the density correlator C_ρ, showing instead a $\omega^{-0.45}$ slope in the central part and a ω^{-2} tail at large ω.

The presence of the flat portion in $C_{\mathcal{E}}$ for $\omega \to 0$, could be naively interpreted as the restoration of normal conductivity. A similar regime where the decay of current correlations is faster (exponential) than the expected power-law decay has been reported for arrays of coupled oscillators [29] and it was argued that thermal conductivity could turn to a normal behavior in the low-energy regimes. Later studies [30] actually showed that this may be rather due to strong finite-size effects. We thus argue that also our results, should be interpreted as such, although the physical origin of the effect is yet unexplained. It is also puzzling that structure factors exhibit the scaling predicted by NFH over a wide range of values of the control parameter Γ whereby a clear crossover is seen in the current spectra upon reducing the collisionality of the particles (see again the panel a of Fig. 4).

5 Conclusions

We have shown that the MPC method is a computationally convenient tool to study nonequilibrium properties of many-particle systems. From the point of view of statistical mechanics, the models are relatively simple to allow for a detailed studies of basic problems like the ones discussed above. Despite its efficiency, the one-dimensional models is still affected by sizeable finite-size effects, particularly close to almost-integrable limits of weak collisionality.

Another attractive feature is that, introducing a suitable energy-dependent collision probability allows to study, at least at a phenomenological level, some interesting issues of confined plasmas, like the effect of suprathermal particles. As a further development, interaction with external reservoirs exchanging energy and particles can be included easily, thus allowing to study genuine nonequilibrium steady states.

Acknowledgements. 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 under grant agreement No 633053 for the project WP17-ENR- CEA-01 ESKAPE. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

1. Pomeau, Y., Résibois, P.: Phys. Rep. **19**(2), 63 (1975)
2. Malevanets, A., Kapral, R.: EPL (Europhys. Lett.) **44**(5), 552 (1998)
3. Malevanets, A., Kapral, R.: J. Chem. Phys. **110**, 8605 (1999). https://doi.org/10.1063/1.478857
4. Malevanets, A., Kapral, R.: In: Karttunen, M., Lukkarinen, A., Vattulainen, I. (eds.) Novel Methods in Soft Matter Simulations. Lecture Notes in Physics, vol. 640, pp. 116–149. Springer, Heidelberg (2004). https://doi.org/10.1007/b95265
5. Dreicer, H.: Phys. Rev. **115**, 238 (1959). https://doi.org/10.1103/PhysRev.115.238
6. Dreicer, H.: Phys. Rev. **117**, 329 (1960). https://doi.org/10.1103/PhysRev.117.329

7. Zhou, R.J., Hu, L.Q., Zhang, Y., Zhong, G.Q., Lin, S.Y., The EAST Team: Nucl. Fusion **57**(11), 114002 (2017). https://doi.org/10.1088/1741-4326/aa7c9d
8. Lepri, S., Livi, R., Politi, A.: Phys. Rep. **377**, 1 (2003). https://doi.org/10.1016/S0370-1573(02)00558-6
9. Dhar, A.: Adv. Phys. **57**, 457 (2008)
10. Lepri, S. (ed.): Thermal Transport in Low Dimensions: From Statistical Physics to Nanoscale Heat Transfer. Lect. Notes Phys., vol. 921. Springer, Heidelberg (2016)
11. Basile, G., Delfini, L., Lepri, S., Livi, R., Olla, S., Politi, A.: Eur. Phys. J. Spec. Top. **151**, 85 (2007)
12. van Beijeren, H.: Phys. Rev. Lett. **108**(18), 180601 (2012). https://doi.org/10.1103/PhysRevLett.108.180601
13. Spohn, H.: J. Stat. Phys. **154**, 1191 (2014). https://doi.org/10.1007/s10955-014-0933-y
14. Spohn, H.: In: Lepri, S. (ed.) Thermal Transport in Low Dimensions: From Statistical Physics to Nanoscale Heat Transfer. Lecture Notes Physics, vol. 921. Springer, Heidelberg (2016)
15. Kardar, M., Parisi, G., Zhang, Y.C.: Phys. Rev. Lett. **56**, 889 (1986). https://doi.org/10.1103/PhysRevLett.56.889
16. Spohn, H., Stoltz, G.: J. Stat. Phys. (2015). https://doi.org/10.1007/s10955-015-1214-0
17. Popkov, V., Schadschneider, A., Schmidt, J., Schütz, G.M.: Proc. Nat. Acad. Sci. **112**(41), 12645 (2015)
18. Gompper, G., Ihle, T., Kroll, D.M., Winkler, R.G.: Multi-particle collision dynamics: a particle-based mesoscale simulation approach to the hydrodynamics of complex fluids. In: Holm, C., Kremer, K. (eds.) Advanced Computer Simulation Approaches for Soft Matter Sciences III. Advances in Polymer Science, vol 221. Springer, Heidelberg (2009)
19. Kapral, R.: Multiparticle collision dynamics: simulation of complex systems on mesoscales. In: Advances in Chemical Physics, pp. 89–146 (2008). https://doi.org/10.1002/9780470371572.ch2
20. Ciraolo, G., Bufferand, H., Di Cintio, P., Ghendrih, P., Lepri, S., Livi, R., Marandet, Y., Serre, E., Tamain, P., Valentinuzzi, M.: Fluid and kinetic modelling for non-local heat transport in magnetic fusion devices. Contrib. Plasma Phys. 1–8 (2018). https://doi.org/10.1002/ctpp.201700222
21. Di Cintio, P., Livi, R., Bufferand, H., Ciraolo, G., Lepri, S., Straka, M.J.: Phys. Rev. E **92**, 062108 (2015)
22. Bufferand, H., Ciraolo, G., Ghendrih, P., Tamain, P., Bagnoli, F., Lepri, S., Livi, R.: J. Phys. Conf. Ser. **260**, 012005 (2010)
23. Bufferand, H., Ciraolo, G., Ghendrih, P., Lepri, S., Livi, R.: Phys. Rev. E **87**(2), 023102 (2013). https://doi.org/10.1103/PhysRevE.87.023102
24. Di Cintio, P., Livi, R., Lepri, S., Ciraolo, G.: Phys. Rev. E **95**, 043203 (2017). https://doi.org/10.1103/PhysRevE.95.043203. https://link.aps.org/doi/10.1103/PhysRevE.95.043203
25. Delfini, L., Lepri, S., Livi, R., Politi, A.: J. Stat. Mech. Theory Exp., P02007 (2007)
26. Mendl, C., Spohn, H.: (private communication)
27. Das, S.G., Dhar, A., Saito, K., Mendl, C.B., Spohn, H.: Phys. Rev. E **90**(1), 012124 (2014)
28. Mendl, C.B., Spohn, H.: Phys. Rev. Lett. **111**, 230601 (2013)
29. Chen, S., Zhang, Y., Wang, J., Zhao, H.: Phys. Rev. E **87**(3), 032153 (2013). https://doi.org/10.1103/PhysRevE.87.032153
30. Das, S., Dhar, A., Narayan, O.: J. Stat. Phys. **154**(1–2), 204 (2014)

A Short Introduction to Piecewise Deterministic Markov Samplers

Pierre Monmarché(✉)

Sorbonne Universités, LJLL, Paris, France
pierre.monmarche@upmc.fr
https://www.ljll.math.upmc.fr/~monmarche/

Abstract. The use of velocity jump Markov processes in MCMC algorithms have recently drawn attention in various fields, such as statistical physics or Bayesian statistics. The aim of this paper is to introduce these processes and to give a few justifications on their interest.

Keywords: MCMC · PDMP · Lifted Markov chain

1 Introduction

1.1 Non-reversible MCMC

The sampling problem is the following: given a target probability measure μ on a space E (say $E = \mathbb{R}^d$), it is necessary in many applied fields (Bayesian statistics, molecular dynamics...) to compute quantities of the form $\int f(x)\mu(\mathrm{d}x)$, where f is a given observable. When μ admits a density proportional to $\exp(-U)$ where the potential $U : E \to \mathbb{R}$ is known but the normalization constant $\int \exp(-U(x))\mathrm{d}x$ is intractable, the MCMC method relies on the simulation of a Markov process $(X_t)_{t\geqslant 0}$ which is ergodic with respect to μ, namely such that

$$\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s \underset{t\to\infty}{\longrightarrow} \int f(x)\mu(\mathrm{d}x). \tag{1}$$

It turns out that several Markov processes do this job. A classical one would be the overdamped Langevin (or Fokker-Planck) diffusion, which is the process solving the SDE

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{2}$$

where B is a Brownian motion. Another process comes from the general Metropolis-Hastings procedure, which is the following: suppose you are given a Markov kernel q on E (say the Gaussian kernel $q(x,y) \propto \exp\left(-\frac{1}{2\sigma^2}|x-y|^2\right)$ for some $\sigma > 0$). Starting from a point $X_n \in \mathbb{R}^d$, draw a random variable Y_n according to the law $q(X_n,\cdot)$ (in the Gaussian case, this means $Y_n = X_n + \sigma G$

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 375–390, 2019.
https://doi.org/10.1007/978-3-030-15096-9_11

where G is a standard Gaussian variable independent from X_n). Then, with probability

$$\min\left(1, \frac{q(y,x)e^{-U(y)}}{q(x,y)e^{-U(x)}}\right),$$

set $X_{n+1} = Y_n$ (we say we accepted the proposal Y_n). Else, set $X_{n+1} = X_n$ (we say we rejected the proposal). This defines a Markov Chain $(X_k)_{k\geqslant 0}$ such that, under mild assumptions on U, (1) holds for discrete times $t \in \mathbb{N}$.

Since there are several possibilities, there is a choice to be made. Which μ-ergodic Markov process should be used in practice? What are the criteria to decide, between two such processes, which one is the best? Ultimately, we want, at a given numerical cost, the error (in some quantified sense) made by approximating $\int f\mathrm{d}\mu$ by the ergodic mean in (1), to be as small as possible. Usually this is not easily tractable, so that alternative criteria may be considered, such as the two following examples:

– Given that a Central Limit Theorem holds together with (1), we get that

$$\sqrt{t}\left(\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s - \int f(x)\mu(\mathrm{d}x)\right) \underset{t\to\infty}{\overset{law}{\longrightarrow}} \mathcal{N}(0,\sigma_f^2) \tag{3}$$

where $\mathcal{N}(a,\sigma^2)$ is the normal distribution with mean a and variance σ^2, and the so-called asymptotic variance σ_f^2 depends on the observable f and on the process X. Then, one wants to use a Markov process X such that σ_f^2 is small.

– At a given time $t \geqslant 0$, the bias of the estimator $\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s$ is given by

$$\mathbb{E}\left(\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s\right) - \int f(x)\mu(\mathrm{d}x) = \frac{1}{t}\int_0^t \left(\mathbb{E}\left(f(X_s)\right) - \int f(x)\mu(\mathrm{d}x)\right)\mathrm{d}s,$$

which would vanish if we were able to sample $X_0 \sim \mu$, in which case we would have $X_t \sim \mu$ for all $t \geqslant 0$. Since we are not able to do so, one wants to use a Markov process X such that the convergence

$$X_t \underset{t\to\infty}{\overset{law}{\longrightarrow}} \mu$$

is as fast as possible (which may be quantified by different (pseudo-)distances over probability measure: total variation, L^2 or Wasserstein distances, relative entropy...).

Other criteria include: the correlation length of the process (see [39] for instance), which measures how far the chain is from an i.i.d. sequence; continuous scalings (such as in [38]), which measure in some sense how many discrete steps are needed to cover a given distance.

These different criteria may not be compatible; optimizing one of them may lead to an inefficient process with respect to another. Nevertheless, in a sense, they all tend to deal with the same general difficulty: the convergence in (1) only occurs once a statistically significant part of the space have been explored.

The problem is thus to explore efficiently the unknown landscape (E, U) with a local, memoryless (by definition of Markov processes) explorer X. Such an amnesic explorer tends to go back repeatedly to places it has already seen, which is inefficient.

To deal with this problem, several long-term memory process have been developed, in particular in molecular dynamics (see [26] and references within). The present paper is concerned with a different (and complementary) direction of research, non-reversible processes.

Let us first recall the definition of a reversible process. Let X be a Markov Chain on E with transition kernel q, and μ be a law on E. Suppose that μ and, for all $x \in E$, $q(x, \cdot)$ admit a density with respect to a reference measure $\mathrm{d}x$ (the counting measure if E is finite, the Lebesgue measure if $E = \mathbb{R}^d$, etc.), still denoted by μ and $q(x, \cdot)$. We say that X is reversible with respect to μ if the following detailed balance condition is met:

$$\forall x, y \in E, \qquad \mu(x)q(x, y) = \mu(y)q(y, x),$$

which implies the global balance condition

$$\forall y \in E, \qquad \int q(x, y)\mu(x)\mathrm{d}x = \mu(y),$$

which means μ is invariant for X.

Formally, the detailed balance condition is equivalent to the fact that the operators

$$\begin{aligned} Qf(x) &= \int f(y)q(x, y)\mathrm{d}y \\ Lf(x) &= Qf(x) - f(x) \end{aligned}$$

are self-adjoint in $L^2(\mu)$. Recall L is called the infinitesimal generator of the continuous-time chain $Z_t = X_{N_t}$ (where N_t is a standard Poisson process independent from X), and satisfies

$$Lf(z) = \lim_{t \to 0} \frac{\mathbb{E}\left(f(Z_t) \mid Z_0 = z\right) - f(z)}{t} \tag{4}$$

whenever this limit exists. The limit (4), in fact, defines the generator L for more general continuous-time processes, like diffusions. For instance, for the overdamped Langevin diffusion (2),

$$Lf(x) = -\nabla U(x) \cdot \nabla f(x) + \Delta f(x).$$

Again, the process is said to be reversible when L is self-adjoint in $L^2(\mu)$ (which is formally the case in the overdamped Langevin case).

In the framework of reversible Markov chain, Peskun's theorem [36] states the following: given two irreducible μ-reversible Markov chains with transition kernels q_1 and q_2 on a finite space E such that $q_1(x, x) \leqslant q_2(x, x)$ for all $x \in E$,

then for any observable f the asymptotic variance in the CLT (3) is smaller for q_1 than for q_2. This is a concrete application of the fact one wants to explore the space more efficiently, since staying at the same place is the worst way to explore.

After ensuring that the process does not stay at the same position, the next step is to ensure that it does not backtrack too much, meaning that for all n, X_{n+1} is unlikely to be equal to X_{n-1}. But then, there is no room of improvement among reversible processes: indeed, the detailed balance condition implies that, if there has been a probability to go from x to y, then there is a given probability to go back from y to x. As a consequence, reversible chains typically have a diffusive behaviour, taking N^2 steps to cover a distance N. To decrease the trend to backtrack, one should necessarily leave the reversible realm. Indeed, Neal [35] proved that, given an irreducible reversible chain on a finite space E, then there exist an irreducible non-reversible chain with the same invariant measure, which backtrack less and such that the asymptotic variance in the CLT is smaller than for the reversible chain.

There is also a spectral argument in favour of non-reversible processes: denoting $P_t f(x) = \mathbb{E}\left(f(Z_t) \mid Z_0 = x\right)$ the semi-group associated to the generator L (formally, $P_t = e^{tL}$), the distance of the law of X_t toward its equilibrium μ may be quantified by the operator norm

$$\|P_t - \mu\|_{L^2(\mu)} := \sup_{f \in L^2(\mu)\setminus\{0\}} \frac{\|P_t f - \int f \mathrm{d}\mu\|_{L^2(\mu)}}{\|f\|_{L^2(\mu)}}.$$

When L is self-adjoint, it admits an orthonormal eigenbasis, so that

$$\|P_t - \mu\|_{L^2(\mu)} = e^{-\lambda_1 t},$$

with $\lambda_1 = \min \sigma(-L) \setminus \{0\}$ the spectral gap of L.

Now, set $L_2 = L + L'$ where L is skew adjoint in $L^2(\mu)$, namely $\int f L' g \mathrm{d}\mu = -\int g L' f \mathrm{d}\mu$ for all $f, g \in L^2(\mu)$. Denoting $P_t' = e^{t(L+L')}$ and $f_t = P_t' f - \int f \mathrm{d}\mu$, then

$$\partial_t \left(\|f_t\|_{L^2(\mu)}\right) \ = \ 2\int f_t (L + L') f_t \mathrm{d}\mu \ = \ 2\int f_t L f_t \mathrm{d}\mu \ \leqslant \ -2\lambda_1 \|f_t\|_{L^2(\mu)}.$$

Grönwall's lemma yields

$$\|P_t' - \mu\|_{L^2(\mu)} \ \leqslant \ e^{-\lambda_1 t} = \|P_t - \mu\|_{L^2(\mu)}.$$

In other words, the spectral gap can only be improved when an anti-adjoint part is added to a reversible chain.

This general argument has motivated studies in different directions. Concerning, for instance, the question to define non-reversible diffusions with a given target measure (for instance, by adding a skew adjoint part to the diffusion (2)), we refer to [21,22,25]. Piecewise deterministic Markov samplers, which are the topic of the present paper, have stemmed from another idea, popularized in [12], which is the definition of lifted Markov chains on finite graphs.

The general idea of lifted chains is the following: instead of considering a Markov chain X on a space E, define a chain (X, Y) on an extended space $E' = E \times F$, such that the image of the invariant measure of (X, Y) by the map $(x, y) \in E \times F \mapsto x \in E$ is the target μ. A particular case is given by second order Markov chains [13,35], where $E' = E \times E$ and the chain (X, Y) is such that for all $n \in \mathbb{N}$, $Y_{n+1} = X_n$. In other words, in this case, the process $(X_n)_{n \geqslant 0}$ on E is not a Markov process, but (X_{n-1}, X_n) on E^2 is. This is the simplest way to add a memory in the exploration, and to avoid backtracking.

1.2 Scaling Limit of the Persistent Walk

Let us recall the definition of the persistent walk introduced in [13, Section 4]. The target μ is the uniform law on $E = \mathbb{Z}/N\mathbb{Z}$, $N \geqslant 2$. For $\alpha \in [0, 1]$, define the Markov chain (X, Y) on $\{(x, y) \in E^2, |x - y| \leqslant 1\}$ by the transitions

$$\mathbb{P}\left((X_{n+1}, Y_{n+1}) = (x, y) \mid X_n, Y_n\right) = \begin{cases} \frac{1+\alpha}{2} & \text{if } y = X_n \text{ and } x - X_n = X_n - Y_n, \\ \frac{1-\alpha}{2} & \text{if } y = X_n \text{ and } x - X_n = -(X_n - Y_n), \\ 0 & \text{else.} \end{cases}$$

In other words, $Y_n = X_{n-1}$ for all $n \in \mathbb{N}$, so that $(X_n)_{n \geqslant 0}$ is a second order Markov chain which is more likely to repeat its previous step than to backtrack (since $\alpha \geqslant 0$). Alternatively, we can consider the chain $(X_n, V_n) = (X_n, X_n - X_{n-1})$ on $\mathbb{Z}/N\mathbb{Z} \times \{-1, +1\}$, whose transitions are given by

$$\mathbb{P}\left(V_{n+1} = V_n \mid X_n, V_n\right) = \frac{1 + \alpha}{2}$$
$$\mathbb{P}\left(V_{n+1} = -V_n \mid X_n, V_n\right) = \frac{1 - \alpha}{2}$$

and almost surely $X_{n+1} = X_n + V_{n+1}$. Seeing X as a position on the discrete circle $\mathbb{Z}/N\mathbb{Z}$, then the evolution V of the position is the velocity of the process. For $\alpha = 0$, V_{n+1} is independent from (X_n, V_n), and X is a simple random walk. For all odd $N \in \mathbb{N}$ and $\alpha \in [0, 1]$, the chain is non-periodic and irreducible with equilibrium μ_N the uniform law on $\mathbb{Z}/N\mathbb{Z} \times \{-1, +1\}$. Denote $\lambda_N(\alpha)$ its spectral gap. As studied in [13,32], the maximal value of λ_N is reached at a positive value

$$\alpha_N = \frac{1 - \sin(\pi/N)}{1 + \sin(\pi/N)},$$

in which case $\lambda_N(\alpha_N) = 1 - \sqrt{\alpha_N}$, which is of order N^{-1}. By comparison, $\lambda_N(0)$ is of order N^{-2}. In other words, it takes $\mathcal{O}\left(N^2\right)$ steps to get close to equilibrium with the simple reversible random walk, and only $\mathcal{O}(N)$ steps with an optimally scaled persistent walk. Note that, in this simple case, the deterministic computation of an integral with respect to the uniform measure on $\mathbb{Z}/N\mathbb{Z}$ is done in exactly N steps.

Seeing $\mathbb{Z}/N\mathbb{Z}$ as N points equally distributed on the continuous torus $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$, the simple random walk converges to the Brownian motion on $\mathbb{T}$ when time is rescaled by a factor N^2. Similarly, a properly scaled persistent walk converges, when time is rescaled by a factor N, toward a continuous process on $\mathbb{T}$ (see [32] for details). More precisely, note that, with the optimal choice $\alpha = \alpha_N$, the number of steps between two changes of velocity is a r.v. with geometric law of parameter $(1-\alpha_N)/2$, which is of order N^{-1}. Since the distance on $\mathbb{T}$ between two points of $\mathbb{Z}/N\mathbb{Z}$ is also of order N^{-1}, the mean distance covered between two changes of the velocities is independent from N, and the law of the time between these changes of velocity converges toward an exponential law.

As a consequence, the continuous process obtained in the limit is the so-called (circular) telegraph process, whose study traces back to [20,23]. It is a continuous time process $(X_t, V_t)_{t\geqslant 0}$ on $E = \mathbb{T} \times \{-1, +1\}$, with $V_t = (-1)^{N_t}$ where N is a homogeneous Poisson process with a given intensity $a > 0$, and $X_t = X_0 + \int_0^t V_s \mathrm{d}s$. In other words, (X, V) is a so-called Piecewise Deterministic Markov process (PDMP): between two random jumps, it follows a deterministic flow. More precisely, here, it is a velocity jump process, in the following sense: first, the deterministic flow is $(x, v)' = (v, 0)$, which means X is the position and V the (piecewise constant) velocity. Second, the jumps only affect the velocity (here, v is changed to $-v$), and not the position.

The circular telegraph is ergodic with respect to μ the uniform measure on E. It is not reversible, since its generator

$$Lf(x, v) = v \cdot \nabla_x f(x, v) + a\,(f(x, -v) - f(x, v))$$

is not self-adjoint, but it is still simple enough for a spectral study to be conducted in [32]. It is proven that $P_t = e^{tL}$ converges exponentially fast toward μ with a rate which is maximal for $a = 1$. In particular, for $a = 1$,

$$\|P_t - \mu\| = e^{-t}\sqrt{1 + \frac{2}{\sqrt{1 + \frac{1}{t^2}} - 1}}.$$

The prefactor is such that $\|P_t - \mu\| \simeq 1 - t^3/3$ when $t \simeq 0$. In other words, for small times, the convergence is slower than in the reversible case, for which $\|P_t - \mu\|$ scales as $1 - t/\lambda_1$. This may be understood in regard of the lack of regularization properties of the dynamics. Nevertheless, for large times, $\|P_t - \mu\| \simeq 2te^{-t}$. It is more delicate to compare this result with the speed of convergence of the Brownian motion on the circle, than it was to compare the persistent walk with the symmetric one.

The generalization in [33,34] of the circular telegraph process with a non-constant rate of jump lead to velocity jump processes with arbitrary invariant measure (see Sect. 2 below), suitable for sampling algorithms. Independently, a similar scaling limit of a lifted Markov chain introduced for the Curie-Weiss spin model [40] lead Bierkens and Roberts to the same kind of dynamics [5].

They also appeared in the physics literature: indeed, the use of so-called event-driven MCMC, which may be seen as lifted chains, had been developed for

the study of hard-sphere systems [1,30]. The general idea is to build Metropolis chains for which, when a move is rejected, then an event (say, a collision) occurs, instead of nothing. From this, Peters and de With [37] obtained a rejection-free chain and, reasoning through infinitesimal steps, ended up with a similar velocity jump process.

In parallel, these dynamics gained interest in the field of bio-mathematics, where they model the motion of a bacterium [10,16,19].

Note that the present introduction does not pretend to be an exhaustive or balanced review of all these works. It focuses on the few theoretical justifications established so far of the interest of velocity jump sampler, with a bias toward the works of the author.

2 Definition of the Processes

Let $E = \mathbb{R}^d \times \mathcal{V}$, where $\mathcal{V} \subset \mathbb{R}^d$ is the set of admissible velocities. The general construction of a velocity jump process (X, V) on E depends on two ingredients: first, the jump rate $\lambda : E \to \mathbb{R}_+$ and the jump kernel $q : E \to \mathcal{P}(\mathcal{V})$, where $\mathcal{P}(F)$ denotes the set of probability measures on F.

Given an initial condition (X_0, V_0), suppose the process $(X_t, V_t)_{t\in[0,t_n]}$ has been defined for some $t_n \geqslant 0$. The next jump time is defined by

$$t_{n+1} = t_n + \inf\left\{t > 0,\ S \leqslant \int_0^t \lambda\left(X_{t_n} + sV_{t_n}, V_{t_n}\right) \mathrm{d}s\right\},$$

where S is a random variable with standard exponential law, independent from $(X_t, V_t)_{t\in[0,t_n]}$. For $t \in (t_n, t_{n+1}]$, set $X_t = X_{t_n} + (t - t_n)V_{t_n}$. For $t \in (t_n, t_{n+1})$, set $V_t = V_{t_n}$. Finally, set $V_{t_{n+1}} = W$, where W is a r.v. with law $q\left(X_{t_{n+1}}, V_{t_n}\right)$, independent from the past. That way, the process is defined up to t_{n+1} and, by induction, up to any jump time t_k, $k \in \mathbb{N}$.

Then, in order to get that the process is defined up to all time, one need to prove that, almost surely, there isn't infinitely many jumps in a finite time interval. This is equivalent to say that the jump rate $\lambda\left(X_t, V_t\right)$ is bounded on finite time intervals. This is true for all the velocity jump processes introduced up to now for sampling purpose. Indeed, for these processes, λ is continuous so that, as long as the velocity is bounded, since in a finite time the position stays in a compact ball, then the jump rate is bounded. When $\mathcal{V}$ is not compact, if the rate at which the norm of the velocity is modified is bounded, the same conclusion holds since, on a finite time interval, the velocity is bounded (by a random but finite bound, see [14] for a precise proof).

Denoting $Z = (X, V)$, the generator L of Z, defined by (4) for smooth functions, is

$$Lf(x, v) = v \cdot \nabla_x f(x, v) + \lambda(x, v)\left(Qf(x, v) - f(x, v)\right),$$

where $Qf(x, v) = \int f(x, w)q(x, v)(\mathrm{d}w)$ (remark that, for $(x, v) \in E$, it is not required that $q(x, v)$ admits a density w.r.t the Lebesgue measure). Then, a given target measure μ is invariant for Z if and only

$$\int Lf(x,v)\mu(\mathrm{d}x,\mathrm{d}v) = 0 \tag{5}$$

for all f in a core of L. Unfortunately, for PDMP's, which lack regularization properties, it is not easy to construct a core of the generator. In particular, when λ is only continuous, sets of smooth functions are not left invariant by the semi-group P_t. Nevertheless, by approximation (truncation and regularization) arguments, in order to show that μ is invariant for Z, it is enough to check that (5) holds for all $f \in \mathcal{C}^1(E)$ with compact support (see [14] for details).

Choose a target measure $\mu(\mathrm{d}x,\mathrm{d}v) = e^{-U(x)}\mathrm{d}x \otimes p(\mathrm{d}v)$, where $p \in \mathcal{P}(\mathcal{V})$. In that case, integrating by part, (5) reads

$$\forall x \in \mathbb{R}^d,\ \ (v\cdot\nabla U(x) - \lambda(x,v))\, p(\mathrm{d}v) + \int \lambda(x,w)q(x,w)(\mathrm{d}v)p(\mathrm{d}w) = 0 \tag{6}$$

(this in an equality in the sense of measures on $\mathcal{V}$). There are many choices of λ, q and p which solves this equation. We refer to [41] for a longer discussion on that matter. In fact, there are even more choices if we don't restrict to velocity jump processes, namely if we consider a PDMP for which the deterministic flow between two random jumps is not simply $(x,v)' = (v,0)$, in which case v is no more a velocity, but a more general auxiliary variable. From an applied perspective, all we need is this deterministic flow to have analytical solutions.

We now give two particular examples of velocity jump processes whose invariant measure is μ, based on different decompositions of $\nabla U(x)$.

2.1 The Bouncy Particle Sampler

The Bouncy Particle Sampler (BPS) has been introduced in [34,37]. In that case, $\mathcal{V}$ is a rotation-invariant subset of $\mathbb{R}^d$, and p is a rotation-invariant probability measure on $\mathcal{V}$. For instance, $\mathcal{V} = \mathbb{R}^d$ and p is a Gaussian law, or $\mathcal{V} = \mathbb{S}_{d-1}$ and p is the uniform law. The jump rate is $\lambda(x,v) = (v\cdot\nabla U(x))_+ + r$, where $(\cdot)_+ = \min(0,\cdot)$ denotes the positive part, and $r \geqslant 0$ is a constant. The jump kernel is

$$Qf(x,v) = \frac{(v\cdot\nabla U(x))_+}{\lambda(x,v)} f\left(x, R(x,v)\right) + \frac{r}{\lambda(x,v)}\int f(x,w)p(\mathrm{d}w),$$

where

$$R(x,v) = v - 2\frac{v\cdot\nabla U(x)}{|\nabla U(x)|^2}\nabla U(x)$$

is the orthogonal reflection of v with respect to ∇U. In other words, the generator of the BPS is $L = L_1 + L_2$ where

$$L_1 f(x,v) = v\cdot\nabla_x f(x,v) + (v\cdot\nabla U(x))_+ \left(f(x,R(x,v)) - f(x,v)\right) \tag{7}$$

$$L_2 f(x,v) = r\left(\int f(x,w)p(\mathrm{d}w) - f(x,v)\right). \tag{8}$$

For $\mu(\mathrm{d}x, \mathrm{d}v) = e^{-U(x)}\mathrm{d}x \otimes p(\mathrm{d}v)$, $\int L_i f \mathrm{d}\mu = 0$ for all suitable f for both $i = 1, 2$. The jump mechanism is a superposition of two mechanisms: at rate $(v \cdot \nabla U(x))_+$, v jumps to $R(x, v)$, which means the process bounces (i.e. undergoes an elastic collision) on the level set U it has reached. Independently, at constant rate $r > 0$, the velocity is refreshed with a whole new one with law p, independent from both the current position and the previous velocity.

2.2 The Zig-Zag Process

The Zig-Zag process (ZZP) has been introduced in [4,5]. In that case, $\mathcal{V} = \{-1, 1\}^d$ and p is the uniform measure on $\mathcal{V}$. The generator is

$$Lf(x, v) = v \cdot \nabla_x f(x, v) + \sum_{i=1}^{d} \left((v_i \cdot \nabla_{x_i} U(x))_+ + r \right) \left(f\left(x, v^{(i)}\right) - f(x, v) \right) \tag{9}$$

where $v^{(i)} = (v_1, \ldots, v_{i-1}, -v_i, v_{i+1}, \ldots, v_d)$ for all $i \in [\![1, d]\!]$, and $r \geqslant 0$ is a constant rate. In other words, the total jump rate is

$$\lambda(x, v) = rd + \sum_{i=1}^{d} (v_i \cdot \nabla_{x_i} U(x))_+$$

and the jump kernel is

$$Qf(x, v) = \sum_{i=1}^{d} \frac{(v_i \cdot \nabla_{x_i} U(x))_+ + r}{\lambda(x, v)} f\left(x, v^{(i)}\right).$$

Again, (6) holds, so that $\mu(\mathrm{d}x, \mathrm{d}v) = e^{-U(x)}\mathrm{d}x \otimes p(\mathrm{d}v)$ is invariant.

Note that, in dimension 1, the ZZP coincides with the BPS with unit scalar velocity (for which $\mathcal{V} = \mathbb{S}_0 = \{-1, 1\}$).

2.3 Practical Implementation

The question of the efficient implementation of velocity jump processes is discussed in [11]. One of the main features of velocity jump processes is that, contrary to diffusions processes which have to be discretized, they can be exactly simulated (see [2] for the exact simulation of a skeleton chain of a diffusion). This is important since an additional discretization step would alter the invariant measure, which would induce a bias in (1). This could be corrected by a supplementary Metropolis-Hastings accept/reject step (such as in the MALA algorithm [15] based on an Euler discretization of (2)), but then the resulting process would be reversible.

Even though velocity jump processes are continuous-time processes, computations are only needed at jump times. Between two jumps, the deterministic flow $(x, v) \mapsto (x + tv, v)$ is explicit. The remaining difficulty is thus to sample

the jump times, which are first jump times of non-homogeneous Poisson process. This is done through the thinning method [27,28]: starting from an initial condition (x_0, v_0), let T be the first jump time, whose survival function is

$$\mathbb{P}(T \geqslant t) = e^{-\int_0^t \lambda(x_0+sv_0,v_0)\mathrm{d}s}.$$

Suppose that we are able to compute, from estimates on ∇U, an upper bound $\psi(t)$ of the jump rate $\lambda(x+tv, v)$. Suppose moreover that we are able to simulate exactly a random variable S with survival function

$$\mathbb{P}(S \geqslant t) = e^{-\int_0^t \psi(s)\mathrm{d}s} \ := \ h(t).$$

This is for instance the case when $\psi(t) = (a + tb)_+$ for some $a, b \in \mathbb{R}$. Define jump time proposal $(t_n)_{n\geqslant 0}$ as follows: $t_0 = 0$ and for all $n \in \mathbb{N}$, $t_{n+1} = t_n + S_n$ where S_n is a r.v. with survival function h, independent from S_k, $k < n$. Let $(U_k)_{k\geqslant 0}$ be an i.i.d. sequence of variables with uniform law on $[0, 1]$, independent from $(S_k)_{k\geqslant 0}$ and

$$K = \inf\left\{k \geqslant 1, \frac{\lambda(x + t_k v, v)}{\psi(t_k)} \geqslant U_k\right\}.$$

Then t_K has the same law as T.

This method is in fact an accept/reject procedure, but contrary to the Metropolis-Hastings algorithm, when a jump proposal is rejected, the process is not frozen, it keeps moving and exploring the space.

Another practical interest with velocity jump processes is exact subsampling [4,41]. In Bayesian statistics, given N observations, the potential U is of the form $U(x) = \sum_{i=1}^N U_i(x)$ where U_i depends on the i^{th} observation. Hence, computing ∇U demands $\mathcal{O}(N)$ computations. A so-called local version of the BPS (the same goes for the Zig-Zag process), with generator $L_1' + L_2$ where L_2 is (8) and

$$L_1' f(x, v) = v \cdot \nabla_x f(x, v) + \sum_{i=1}^N (v \cdot \nabla U_i)_+ \left(f\left(x, R_i(x, v)\right) - f(x, v)\right)$$

(where $R_i(x, v)$ is the reflection of v with respect to $\nabla U_i(x)$) may be simulated only by considering one observation at a time. The invariant measure is unchanged.

It is not completely obvious that this subsampling method improves the computational cost: the cost to compute one jump time is now $\mathcal{O}(1)$, but the number of jumps in a given time is $\mathcal{O}(N)$. Nevertheless, as studied in [4], when associated with efficient bounds on the jump rates, this method may indeed yield a significant acceleration.

3 Long-Time Behaviour

3.1 In Dimension 1

The convergence of the law of the ZZP in dimension 1 toward its equilibrium $\mu(x, v) = e^{-U(x)} \otimes \{-1, 1\}$ has been established in several works. It is proven

by coupling arguments in [18,19] that, when U is convex, convergence in total variation distance is exponentially fast. The exponential convergence in $L^2(\mu)$ is established under more general assumptions on U in [34] thanks to hypocoercive-type methods (see also [10]). The exponential rates obtained are explicit but not sharp, which makes it unsuitable for efficiency comparison with respect to the overdamped Langevin diffusion 2.

In [3], a Central Limit Theorem is obtained, in the case of a unimodal target measure, with an explicit asymptotic variance. The following things are observed: first, the asymptotic variance increases with the parameter r introduced in (9). Second, there exist heavy tailed distributions for which a CLT holds with the ZZP but not with the Langevin diffusion (2).

Asymptotically precise results may be obtained in the low temperature regime. Suppose the target measure is $\exp(-\beta U_0)$ with a large $\beta > 0$ (called the inverse temperature). When U_0 has several local minima, both the Langevin diffusion and the Zig-Zag process are metastable: transitions from the vicinity of a local minima to the vicinity of another one are rare events, since potential barrier of size $\mathcal{O}(\beta)$ have to be overcome. When $\beta \to \infty$, this energetic metastability (by contrast with entropic metastability, see [24] for more details) is the leading cause of slow convergence.

More precisely, suppose U_0 is decreasing from $+\infty$ to a minimum x_0, then increasing up to a maximum $x_1 > 0$. Suppose $X_0 = x_0$ in the Langevin case and $Z_0 = (X_0, V_0) = (x_0, -1)$ in the Zig-Zag case. In both cases, let $\tau = \inf\{t > 0, X_t = x_1 + \delta\}$, for a small $\delta > 0$, be the escape time from the catchment area of x_0. The small-temperature behaviour of τ is given by so-called Eyring-Kramers formulas. For the overdamped Langevin diffusion with potential $U = \beta U_0$, according to [8,9],

$$\mathbb{E}[\tau] = \frac{2\pi}{\beta\sqrt{|U_0''(x_1)|U_0''(x_0)}} e^{\beta(U_0(x_1)-U_0(x_0))}\left(1 + \underset{\varepsilon\to 0}{o}(1)\right). \tag{10}$$

In the Zig-Zag case with potential $U = \beta U_0$ and $r = 0$, according to [34],

$$\mathbb{E}[\tau] = \sqrt{\frac{8\pi}{\beta U_0''(x_0)}} e^{\beta(U_0(x_1)-U_0(x_0))}\left(1 + \underset{\varepsilon\to 0}{o}(1)\right). \tag{11}$$

In both cases, for a fixed $t > 0$,

$$\mathbb{P}(\tau \geq t\mathbb{E}[\tau]) \underset{\beta\to\infty}{\longrightarrow} e^{-t},$$

which is typical of metastability: at a large scale, the processes behave as a Markov chain over the set of local minima of U_0.

It is delicate to use (10) and (11) to compare the processes. Different choices of normalization would change the prefactors. Two things may be noted: first, the Hessian $U''(x_1)$ does not intervene in the Zig-Zag case. This is consistent with the fact the process has some inertia so that, in a flat landscape, it deterministically crosses an interval from one edge to the other, which is not the

case of the Brownian motion. Second, the leading term in both (10) and (11) is the same exponential factor. Hence, at the exponential scale, both processes are comparable.

It is not surprising that the addition of a short-term memory (the velocity) is not enough to get rid of metastability, which is a large-time problem. This is not in contradiction with the arguments in Sect. 1.2 in favour of the persistent walk, which suggest that the ZZP may be more efficient in the local exploration around each minima.

3.2 Irreducibility

In dimension larger than 1, it is not clear that the processes are irreducible. This is essentially a deterministic control question: assuming that we can choose the jump times as long as they occur when the jump rate is positive (i.e. we can choose the exponential variables which are used to define the jump times), is it possible, starting from a given $z_1 \in E$, to reach any $z_2 \in E$ in a finite time?

This is simple to study for both the BPS and the ZZP when there is a non-zero refreshment rate, i.e. if $r > 0$ in (8) or (9). Indeed, in that case, any velocity $v \in \mathcal{V}$ may be chosen at any position $x \in \mathbb{R}^d$, and thus the control problem is easy to solve. Nevertheless, the case $r > 0$ corresponds to a more "diffusive" behaviour of the process, so that one may want to avoid it.

When $r = 0$, in fact, the BPS may not be irreducible, as noted in [7]. Indeed, the jump rate is zero as long as $U(X_t)$ is decreasing along the trajectory. Suppose the target measure is the standard Gaussian one (or more generally, is log-concave and invariant by rotation) in dimension 2, so that the level sets of the potential are the circle centered at the origin. Then, the BPS with $r = 0$, namely the process with generator L_1 defined by (7) will never enter a given disk (except in the degenerate case where the initial condition (x, v) is such that x and v are colinear, in which case both the position and the velocity will forever stay in $span(x)$), see Fig. 1.

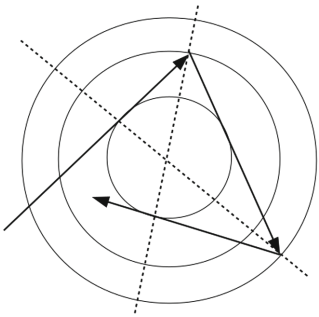


Fig. 1. Trajectory of the BPS with $r = 0$ and the standard Gaussian measure target.

There exist versions of the BPS [31,41,42], slightly different from the process introduced in Sect. 2.1, called BPS with randomized bounce, for which irreducibility is true, even with $r = 0$. When these processes bounce, the velocity, instead of jumping deterministically toward its reflexion with respect to $\nabla U(x)$, takes a random value, according to an *ad hoc* law which admits a Lebesgue density. This additional randomness is enough to be able to reach any state (x, v) from any initial condition (x_0, v_0).

The behaviour of the ZZP with $r = 0$ is different: Bierkens, Roberts and Zitt proved in [6] that it is irreducible under mild assumption on U (it should be $\mathcal{C}^3$ and grow at infinity at least like $\log(|x|)$). The proof that every point (x, v) may be reached from any other (x', v') is far from simple, due to the very degenerate behaviour of the process. Some smoothness for U is necessary: for instance, the ZPP is not irreducible for $U(x_1, x_2) = \max(|x_1|, |x_2|)$. Indeed, from some starting points, it is impossible to reach (x, v) if $x \in \{x_1 > |x_2|\}$ and $v = (-1, -1)$ (see [6] for details and more discussions on the difficulties of the proof).

3.3 Geometric Ergodicity

Up to now, the only results establishing geometric ergodicity for velocity jump samplers in dimension larger than 1 have used the classical Meyn-Tweedie approach [29], which gives an estimate of the form

$$\|\mathcal{L}aw(Z_t) - \mu\|_1 \leqslant C\,(\mathcal{L}aw(Z_0))\, e^{-\rho t}.$$

It relies on two ingredients: first, the construction of a suitable Lyapunov function implies that the process tends to go back to compact sets. Second, a minoration condition ensures that, starting from two different conditions in a compact set, then two trajectories may couple in a finite time with positive probability.

The first result of geometric ergodicity for the BPS in [34], was restricted to the compact case $\mathbb{T}^d \times \mathbb{S}_{d-1}$. In that case, only a minoration condition is necessary, which is more or less obtained from a reachability argument.

To extend this result to a non-compact setting, namely to find a Lyapunov function, is not trivial. This has been done for the ZZP in [6], under the condition that $\nabla U = o(U)$, $\nabla^2 U = o(U)$ and $\nabla U \to \infty$ at infinity. The BPS case has been tackled in [11], but under somehow non completely satisfactory conditions, and later in [14] (see also [17]).

Let us briefly explain the difficulties in the case of the BPS. Let $L = L_1 + L_2$ as defined in (7) and (8), with $r > 0$.

We want to construct a function $W : \mathbb{R}^{2d} \to \mathbb{R}$ with $W \geqslant 1$ and, outside of a compact, $LW \leqslant -cW$ for some $c > 0$, which means that, on average, away from a compact, W should tend to decrease along the trajectory. Remark that, because of the operator $f \mapsto \int f(\cdot, v)p(\mathrm{d}v)$, the construction of W at a point (x, y) influences the value of LW at all points $\{(x, v),\, v \in \mathbb{R}^d\}$. Similarly, the term $f(x, R(x, y))$ is non-local. This is different from the classical case of diffusions, which are local Markov processes.

On the other hand, the reason why W should decrease is different depending on the region the process is visiting, which accounts for the fact there are different reasons to leave these regions: for instance, it is unlikely for the process to have a very large velocity, because of refreshments; it is unlikely that $v \cdot \nabla U(x)$ stays large, because of bouncing. This leads to a definition of W as a combination of several parts but, due to the non-local terms, the balance between these different parts has to be tuned carefully. More precisely, the Lyapunov functions in [6,11] are of the form

$$W(x,v) \;=\; e^{\frac{1}{2}U(x)} g\left(v \cdot \nabla U(x)\right),$$

where the function g is increasing. That way, as long as $V_t \cdot \nabla U(X_t) \leqslant 0$, $U(x)$ decreases along the trajectory, and so does W (provided g does not grow too much in parallel). If $V_t \cdot \nabla U(X_t) \geqslant 0$, then the process has a positive probability to jump, in which case $g\left(v \cdot \nabla U(x)\right)$ decreases, and so does W (provided U hasn't grown too much in the meantime). In [14,17], the mechanism is similar.

References

1. Bernard, E.P., Krauth, W., Wilson, D.B.: Event-chain Monte Carlo algorithms for hard-sphere systems. Phys. Rev. E **80**(5), 056704 (2009)
2. Beskos, A., Roberts, G.O.: Exact simulation of diffusions. Ann. Appl. Probab. **15**(4), 2422–2444 (2005)
3. Bierkens, J., Duncan, A.: Limit theorems for the Zig-Zag process. Adv. Appl. Probab. **49**(3), 791–825 (2017)
4. Bierkens, J., Fearnhead, P., Roberts, G.: The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. arXiv e-prints, arXiv:1607.03188 (2016)
5. Bierkens, J., Roberts, G.: A piecewise deterministic scaling limit of lifted Metropolis-Hastings in the Curie-Weiss model. Ann. Appl. Probab. **27**(2), 846–882 (2017)
6. Bierkens, J., Roberts, G., Zitt, P.-A.: Ergodicity of the zigzag process. arXiv e-prints, arXiv:1712.09875 (2017)
7. Bouchard-Côté, A., Vollmer, S.J., Doucet, A.: The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. J. Am. Stat. Assoc. **113**(522), 855–867 (2018)
8. Bovier, A., Eckhoff, M., Gayrard, V., Klein, M.: Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. J. Eur. Math. Soc. (JEMS) **6**(4), 399–424 (2004)
9. Bovier, A., Gayrard, V., Klein, M.: Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. J. Eur. Math. Soc. (JEMS) **7**(1), 69–99 (2005)
10. Calvez, V., Raoul, G., Schmeiser, C.: Confinement by biased velocity jumps: aggregation of *Escherichia coli*. Kinet. Relat. Models **8**(4), 651–666 (2015)
11. Deligiannidis, G., Bouchard-Côté, A., Doucet, A.: Exponential ergodicity of the bouncy particle sampler. Ann. Stat. **47**(3), 1268–1287 (2019)
12. Diaconis, P., Holmes, S., Neal, R.M.: Analysis of a nonreversible Markov chain sampler. Ann. Appl. Probab. **10**(3), 726–752 (2000)
13. Diaconis, P., Miclo, L.: On the spectral analysis of second-order Markov chains. Ann. Fac. Sci. Toulouse Math. (6) **22**(3), 573–621 (2013)

14. Durmus, A., Guillin, A., Monmarché, P.: Geometric ergodicity of the bouncy particle sampler. arXiv e-prints, arXiv:1807.05401 (2018)
15. Durmus, A., Moulines, É.: Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm. Stat. Comput. **25**(1), 5–19 (2015)
16. Erban, R., Othmer, H.G.: From individual to collective behavior in bacterial chemotaxis. SIAM J. Appl. Math. **65**(2), 361–391 (2004)
17. Fétique, N.: Long-time behaviour of generalised Zig-Zag process. arXiv e-prints, arXiv:1710.01087 (2017)
18. Fontbona, J., Guérin, H., Malrieu, F.: Quantitative estimates for the long-time behavior of an ergodic variant of the telegraph process. Adv. Appl. Probab. **44**(4), 977–994 (2012)
19. Fontbona, J., Guérin, H., Malrieu, F.: Long time behavior of telegraph processes under convex potentials. Stoch. Process. Appl. **126**(10), 3077–3101 (2016)
20. Goldstein, S.: On diffusion by discontinuous movements, and on the telegraph equation. Quart. J. Mech. Appl. Math. **4**, 129–156 (1951)
21. Guillin, A., Monmarché, P.: Optimal linear drift for an hypoelliptic diffusion. Electron. Commun. Probab. **21**, 74 (2016)
22. Hwang, C.R., Hwang-Ma, S.Y., Sheu, S.J.: Accelerating Gaussian diffusions. Ann. Appl. Probab. **3**, 897–913 (1993)
23. Kac, M.: A stochastic model related to the telegrapher's equation. Rocky Mt. J. Math. **4**, 497–509 (1974)
24. Lelièvre, T.: Two mathematical tools to analyze metastable stochastic processes. In: Cangiani, A., Davidchack, R., Georgoulis, E., Gorban, A., Levesley, J., Tretyakov, M. (eds.) Numerical Mathematics and Advanced Applications 2011, pp. 791–810. Springer, Heidelberg (2013)
25. Lelièvre, T., Nier, F., Pavliotis, G.A.: Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. J. Stat. Phys. **152**(2), 237–274 (2013)
26. Lelièvre, T., Rousset, M., Stoltz, G.: Long-time convergence of an adaptive biasing force method. Nonlinearity **21**(6), 1155–1181 (2008)
27. Lemaire, V., Thieullen, M., Thomas, N.: Exact simulation of the jump times of a class of piecewise deterministic Markov processes. J. Sci. Comput. **75**(3), 1776–1807 (2018)
28. Lewis, P.A.W., Shedler, G.S.: Simulation of nonhomogeneous Poisson processes by thinning. Naval Res. Logist. Quart. **26**(3), 403–413 (1979)
29. Meyn, S., Tweedie, R.L.: Markov Chains and Stochastic Stability, 2nd edn. Cambridge University Press, Cambridge (2009)
30. Michel, M., Durmus, A., Sénécal, S.: Forward event-chain Monte Carlo: fast sampling by randomness control in irreversible Markov chains. arXiv e-prints, arXiv:1702.08397 (2017)
31. Michel, M., Kapfer, S.C., Krauth, W.: Generalized event-chain Monte Carlo: constructing rejection-free global-balance algorithms from infinitesimal steps. J. Chem. Phys. **140**(5), 054116 (2014)
32. Miclo, L., Monmarché, P.: Étude spectrale minutieuse de processus moins indécis que les autres. In: Séminaire de Probabilités XLV. Lecture Notes in Mathematics, vol. 2078, pp. 459–481. Springer, Cham (2013)
33. Monmarché, P.: Hypocoercive relaxation to equilibrium for some kinetic models. Kinet. Relat. Models **7**(2), 341–360 (2014)
34. Monmarché, P.: Piecewise deterministic simulated annealing. ALEA Lat. Am. J. Probab. Math. Stat. **13**(1), 357–398 (2016)

35. Neal, R.M.: Improving asymptotic variance of MCMC estimators: non-reversible chains are better. arXiv Mathematics e-prints, arXiv:math/0407281 (2004)
36. Peskun, P.H.: Optimum Monte-Carlo sampling using Markov chains. Biometrika **60**, 607–612 (1973)
37. Peters, E.A.J.F., de With, G.: Rejection-free Monte Carlo sampling for general potentials. Phys. Rev. E **85**, 026703 (2012)
38. Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. J. R. Stat. Soc. B **60**, 255–268 (1997)
39. Scemama, A., Lelièvre, T., Stoltz, G., Caffarel, M.: An efficient sampling algorithm for variational Monte Carlo. J. Chem. Phys. **125**(11), 114105 (2006)
40. Turitsyn, K.S., Chertkov, M., Vucelja, M.: Irreversible Monte Carlo algorithms for efficient sampling. Physica D **240**, 410–414 (2011)
41. Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., Doucet, A.: Piecewise-deterministic Markov chain Monte Carlo. arXiv e-prints, arXiv:1707.05296 (2017)
42. Wu, C., Robert, C.P.: Generalized bouncy particle sampler. arXiv e-prints, arXiv:1706.04781 (2017)

Time Scales and Exponential Trend to Equilibrium: Gaussian Model Problems

Lara Neureither(✉) and Carsten Hartmann

Institut für Mathematik, Brandenburgische Technische Universität Cottbus-Senftenberg, Konrad-Wachsmann-Allee 1, 03046 Cottbus, Germany
{neurelar,hartmanc}@b-tu.de

Abstract. We review results on the exponential convergence of multidimensional Ornstein-Uhlenbeck processes and discuss notions of characteristic time scales by means of concrete model systems. We focus, on the one hand, on exit time distributions and provide explicit expressions for the exponential rate of the distribution in the small-noise limit. On the other hand, we consider relaxation time scales of the process to its equilibrium measure in terms of relative entropy and discuss the connection with exit probabilities. Along these lines, we study examples which illustrate specific properties of the relaxation and discuss the possibility of deriving a simulation-based, empirical definition of slow and fast degrees of freedom which builds upon a partitioning of the relative entropy functional in connection with the observed relaxation behaviour.

Keywords: Multidimensional Ornstein-Uhlenbeck process · Exponential convergence · Relative entropy · Large deviations · Small noise asymptotics

1 Introduction

The characteristic time scales of a random dynamical system, e.g. a diffusion process are often associated with the speed at which the dynamics reaches an equilibrium state or samples its invariant measure. For dynamical systems with several metastable equilibria (also: "metastable states"), such as molecular systems [45], chemical reaction networks [26], or earth and climate systems [28,31], the speed of convergence is often related to the characteristic time scale of transitions between these equilibria.

In this work we are concerned with the speed of convergence of linear ergodic diffusion processes to their unique stationary probability distribution, and we review different concepts of the associated characteristic time scales in terms of exponential estimates for entropy decay, exit probabilities and exit rates. Looking at exit probabilities and exit rates is motivated by the observation that the speed of convergence at which the dynamics reaches a generic multimodal invariant distribution depends on the probability that the process leaves its basins of

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 391–410, 2019.
https://doi.org/10.1007/978-3-030-15096-9_12

attraction. We focus on studying the convergence behavior on the basis of concrete case studies and on deriving explicit expressions for the exit probabilities. Specifically, we confine our considerations to Ornstein-Uhlenbeck processes that can be seen as local linearisations of a more complicated dynamics with multimodal invariant distributions about a metastable equilibrium. Understanding the characteristic time scales of a diffusion process is not only important in statistical physics (to which the aforementioned applications belong), but it is also relevant to assess the asymptotic properties of Markov chain Monte Carlo (MCMC) algorithms [33], or failure probabilities in system reliability and risk analysis [38], to mention just two more examples.

Even though this article is essentially a survey of well-known results, our own contribution lies in putting these results into context with each other, with the twofold aim of (a) revealing some relations between the aforementioned time scale concepts and of (b) making a first step towards understanding these concepts in the case of non-reversible and degenerate diffusions.

Relevant Work. The analysis of characteristic time scales and exponential convergence to equilibrium is system specific, and various different approaches have been developed in the past. We refrain from giving a complete list of references (which would be difficult anyway), but focus on approaches that are most relevant for statistical mechanics applications. Specifically, for reversible and metastable Markov chains and diffusion processes that are relevant for the modeling and the simulation of many-particle systems and critical phase transitions, the analysis of the eigenvalues of the infinitesimal generator has become a standard tool; see e.g. [6,24] and the references therein. For a certain class of non-reversible diffusions, spectral properties have furthermore been analysed in connection with small-noise limits; see e.g. [40,47] or [8] for a tutorial review. Despite the limitation to reversible systems (satisfying detailed balance) or perturbations of such systems, the spectral approach is appealing, since it allows for a hierarchical decomposition of the dynamics, based on the eigenvalues and eigenfunctions of the associated semigroup or its generator [45]. The key observation here is that the eigenvalues close to the principal eigenvalue $\lambda = 0$ represent characteristic time scales (sometimes called "implied time scales") that can be associated with the transitions between metastable sets [10,23]; related results for small-noise diffusions that establish a link between dominant eigenvalues of a diffusion and exit times are discussed in, e.g., [14,17].

A more global perspective to the relaxation dynamics is provided by entropy estimates that can be used to prove exponential convergence to the stationary distribution. These approaches are based on certain functional inequalities like the Poincaré and the (logarithmic) Sobolev inequality, and provide bounds for the convergence to the stationary distribution in the L^1 norm in terms of relative entropy. These bounds utilise the celebrated Csiszár-Kullback-Pinsker inequality, and for reversible systems with potentials that are growing quadratically at infinity, the use of logarithmic Sobolev constants and relative entropy (or: Kullback-Leibler divergence) can be attributed to Bakry and Émery [7].

These results have then been generalised to nonlinear [39], non-reversible [1] and linear diffusions with degenerate noise [2], including generalisations of the Csiszár-Kullback-Pinsker inequality to relative entropies beyond the Kullback-Leibler divergence (see e.g. [3, Ch. 2.2]). For a survey of entropy techniques, functional inequalities and exponential convergence estimates, with a special focus on applications in molecular dynamics, we refer to [35]. An attempt to relate entropy estimates and exit times in a hierarchical way, like it is done in spectral approaches, has been undertaken in [37], but a truly hierarchical approach is, to our knowledge, yet missing.

One motivation for studying exponential rates for the convergence to equilibrium is to devise MCMC methods that either sample the stationary distribution at a higher exponential rate (e.g. [30]) or reduce the variance of certain statistical estimators (e.g. [25]). Importance sampling and related variance reduction methods are naturally connected to large deviations principles, in that they are often applied in the context of small-noise diffusions [16,48], for which MCMC methods are known to converge poorly, or ergodic sampling problems [13,43] that can benefit from faster convergence to equilibrium; see also the seminal articles [19,20] for a discussion of the theoretical connection between large deviations and stochastic control from the viewpoint of viscosity solutions, and [29,46] for applications in molecular dynamics.

Outline. The rest of the article is structured as follows: In Sect. 2 we introduce the multidimensional Ornstein-Uhlenbeck (OU) process and briefly discuss its asymptotic properties for large times. Section 3 is devoted to a review of relevant entropy estimates for reversible, non-reversible and degenerate OU processes, which is contrasted and linked with the Donsker-Varadhan and Freidlin-Wentzell large deviations approaches to the corresponding exit problem in Sect. 4. Numerical examples that illustrate the theoretical results are shown in Sect. 5, with a special focus on degenerate diffusions of Langevin type and slow-fast systems. The discussion is summarised in Sect. 6, and an outlook to possible future research directions is given.

2 Linear Systems

In this section we review some known results about linear stochastic differential equations based on the works [2] and [49] and discuss some concrete specifications for the problem at hand. For this we assume the following setting. Consider an Ornstein-Uhlenbeck process $(X_t)_{(t\geq 0)}$ where for each t, $X_t \in \mathbb{R}^n$ which is described by the SDE

$$\mathrm{d}X_t = AX_t\,\mathrm{d}t + C\,\mathrm{d}W_t. \tag{1}$$

Here $A \in \mathbb{R}^{n\times n}$ is referred to as drift matrix, $C \in \mathbb{R}^{n\times m}$ as diffusion matrix and W_t is a standard m-dimensional Brownian motion and $m \leq n$.

The corresponding Fokker-Planck equation, which describes the time evolution of the probability density function ρ_t according to the dynamics given by

(1), reads

$$\partial_t \rho_t = -\nabla \cdot (A\rho_t) + \frac{1}{2}\nabla^2 : (CC^T \rho_t), \tag{2}$$

where ∇^2 is the Hessian matrix and $A : B$ is the Frobenius inner product of two matrices A and B.

In order to guarantee a unique invariant distribution for our process, we impose the following assumptions on the matrices A and C.

Assumption 1. *The drift matrix A is stable (Hurwitz), i.e. all eigenvalues of A have strictly negative real part.*

Assumption 2. *No eigenvector v of A^T is in the kernel of C^T.*

Assumption 1 guarantees asymptotic stability of the dynamics and entails positive recurrence, whereas Assumption 2 guarantees spreading of the noise in every direction of state space (irreducibility). Assumption 2 has many equivalent formulations, one of which is the complete controllability of the matrix pair (A, C); given a controlled ODE $\dot{x}(t) = Ax(t) + Cu(t)$ with A, C as before, controllability means that there exists a bounded and measurable control u such that the origin $x = 0$ can be reached from any point $y \in \mathbb{R}^n$ in finite time. For further details we refer to [51, Thm. 1.2] or [4, Thm. 4.5].

With these assumptions at hand we are now ready to characterize the unique invariant distribution of (1).

Proposition 1. *Assume that Assumptions 1 and 2 hold true. Then the Fokker-Planck Eq. (2) has a unique stationary state ρ_∞, given by the probability density function of the normal distribution $\mathcal{N}(0, \Sigma_\infty)$, with mean zero and covariance Σ_∞. Furthermore the covariance matrix Σ_∞ is the unique positive definite solution to the Lyapunov equation*

$$A\Sigma_\infty + \Sigma_\infty A^T = -CC^T. \tag{3}$$

We will also refer to ρ_∞ as the invariant distribution of the process $(X_t)_{(t\geq 0)}$ described by (1). For a detailed proof we refer the reader to [2, Thm. 3.1].

Here, we aim at giving an intuitive explanation of the result. Consider the analytic solution to the SDE (1) for a deterministic initial condition $X_0 = x_0$, i.e. the initial covariance Σ_0 fulfils $\Sigma_0 = 0$, which reads

$$X_t = e^{At}x_0 + \int_0^t e^{A(t-s)}C \, \mathrm{d}W_s.$$

The mean μ_t and covariance Σ_t of the process at time t are easily calculated. They fully characterize the distribution of the process at time t, since the dynamics are linear and hence the distributions will be Gaussian at all times. Mean and covariance are calculated easily:

$$\mu_t = \mathbb{E}(X_t) = e^{At}x_0,$$
$$\Sigma_t = \mathbb{E}\left((X_t - \mu_t)(X_t - \mu_t)^T\right) = \int_0^t e^{A(t-s)}CC^T e^{A^T(t-s)}\,\mathrm{d}s.$$

Clearly, $\mu_t \to 0$ as $t \to \infty$, since all eigenvalues of A have negative real part according to Assumption 1. For Σ_t we first note that, again by Assumption 1, $\lim_{t\to\infty} \Sigma_t = \Sigma_\infty$ is well-defined. Further we observe that $\lim_{t\to\infty} \Sigma_t = \Sigma_\infty$ is equivalent to $\lim_{t\to\infty} \dot{\Sigma}_t = 0$. Now,

$$\dot{\Sigma}_t = A\Sigma_t + \Sigma_t A^T + CC^T$$

and thus

$$\lim_{t\to\infty} \Sigma_t = \Sigma_\infty \;\Leftrightarrow\; A\Sigma_\infty + \Sigma_\infty A^T + CC^T = 0.$$

The same calculation goes through when the initial condition is not deterministic (i.e. when $\Sigma_0 \neq 0$), but follows an absolutely continuous probability distribution. Uniqueness of the solution is not hard to prove either, and we refer to [51, Thm. 2.7] for the details. We will refer to the solution of (1) under the Assumptions 1 and 2 as *ergodic Ornstein-Uhlenbeck (OU) process*.

3 Entropy Decay

In this section we mainly review results by [2] who proved exponential convergence to equilibrium for densities evolving according to (2) under Assumptions 1 and 2. To this end, we first sketch the general procedure how to prove such results for the case of non-degenerate Fokker-Planck equations, which means $rank(C) = n$ in our case. Afterwards we explain how the procedure is modified for the degenerate case, i.e. $rank(C) < n$ and state the final result at the end of the section.

3.1 Non-degenerate Case

We start with the non-degenerate case which corresponds to an SDE of the form

$$\mathrm{d}X_t = AX_t\,\mathrm{d}t + C\,\mathrm{d}W_t. \tag{4}$$

where A fulfils Assumption 1 and $rank(C) = n$. We assume the Bakry-Émery criterion

$$\Sigma_\infty^{-1} \geq 2\lambda\, D^{-1} \tag{5}$$

to hold, where we introduced the shorthand notation $D = CC^T$.

Exponential decay to equilibrium in relative entropy then follows from the steps described in the sequel. In the first step the time derivative of relative entropy is computed:

$$-I(\rho_t|\rho_\infty) := \frac{\mathrm{d}}{\mathrm{d}t}H(\rho_t|\rho_\infty) = -\frac{1}{2}\int \left(\nabla \log \frac{\rho_t}{\rho_\infty}\right)^T D\left(\nabla \log \frac{\rho_t}{\rho_\infty}\right)\,\rho_t\,\mathrm{d}x \leq 0. \tag{6}$$

The functional I is called *Fisher information*. One would like to find an estimate of form $-I(\rho_t|\rho_\infty) \leq -\lambda H(\rho_t|\rho_\infty)$ since integration of this inequality yields exponential convergence of $H(\rho_t|\rho_\infty)$ to zero with rate $\lambda > 0$. In the second step—aiming at finding such an estimate—the time derivative of the Fisher information is computed (for details see e.g. [1,3])

$$\frac{\mathrm{d}}{\mathrm{d}t} I(\rho_t|\rho_\infty) = -\frac{1}{2} \int \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right)^T D\Sigma_\infty^{-1} D \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right) \rho_t \, \mathrm{d}x - F, \tag{7}$$

where $F \geq 0$. The third step consists of applying the Bakry-Émery condition (5) to (7) which yields

$$\frac{\mathrm{d}}{\mathrm{d}t} I(\rho_t|\rho_\infty) \leq -\lambda \int \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right)^T D \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right) \rho_t = -2\lambda I(\rho_t|\rho_\infty). \tag{8}$$

Integrating the last inequality in time from 0 to t and using Gronwall's Lemma, we get exponential decay of the Fisher information $I(\rho_t|\rho_\infty) \leq e^{-2\lambda t} I(\rho_0|\rho_\infty)$. Integrating instead from t to ∞, using $-I = \mathrm{d}H/\,\mathrm{d}t$, we find

$$-I(\rho_t|\rho_\infty) \leq -2\lambda H(\rho_t|\rho_\infty), \tag{9}$$

which is the sought inequality. Integration of (9) from 0 to t then yields

$$H(\rho_t|\rho_\infty) \leq H(\rho_0|\rho_\infty) e^{-2\lambda t}. \tag{10}$$

3.2 Degenerate Case

In the degenerate case, i.e. $rank(D) < n$ the usual Bakry-Émery condition (5) cannot hold since D is not invertible. Also, due to the rank deficiency of D, I is not strictly positive anymore but only positive semidefinite. Hence the decay in relative entropy may not be strictly monotone, but can also exhibit plateaus.

In order to achieve strict monotonicity in the decay of relative entropy, the Fisher information I is replaced by a modified Fisher information S where the degenerate diffusion matrix D is replaced by a non-degenerate matrix $P > 0$

$$S(\rho_t|\rho_\infty) = \int \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right)^T P \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right) \rho_t \, \mathrm{d}x.$$

The key ingredients in order to obtain exponential decay in relative entropy are then to prove exponential decay of the functional $S(\rho_t|\rho_\infty)$ and to see that $P \geq \frac{c_P}{2} D$ for some positive constant c_P, by which the exponential decay of the Fisher information follows and hence (8).

In order to establish exponential decay of $S(\rho_t|\rho_\infty)$ its time derivative is computed, yielding (cf. [2, Prop. 4.5])

$$\frac{\mathrm{d}}{\mathrm{d}t} S(\rho_t|\rho_\infty) = -\int \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right)^T \left[QP + PQ^T \right] \left(\nabla \log \frac{\rho_t}{\rho_\infty} \right) \rho_t \, \mathrm{d}x - F_P, \tag{11}$$

where $Q := \Sigma_\infty A^T \Sigma_\infty^{-1}$ and $F_P \geq 0$. The result which replaces the Bakry-Émery criterion (5) is given in [2, Lem. 4.3] and is indispensable for the proof. It yields the existence of a positive definite matrix P such that

$$QP + PQ^T \geq 2\lambda P \tag{12}$$

where either $\lambda = \nu = \min\{\Re(\lambda) : -Av = \lambda v\} > 0$ if A is diagonalizable (i.e. if all eigenvalues have the same geometric and algebraic multiplicity) or $\lambda = \nu - \varepsilon$ for some $\varepsilon > 0$ if at least one eigenvalue is defective (i.e. if geometric and algebraic multiplicity do not agree). Equations (11) and (12) then take the role of (7) and (5), which yields the exponential decay of the functional S. Noting that we can find a constant c_P such that $P \geq \frac{c_P}{2} D$, it follows that the Fisher information decays exponentially, which entails the exponential decay of the relative entropy. The results are summarized in the following Theorem (cf. [2, Thm. 4.9]).

Theorem 1. *Consider the SDE (1) with associated Fokker-Planck Eq. (2), and let Assumptions 1 and 2 hold. Define $\nu = \min\{\Re(\lambda) : -Av = \lambda v\} > 0$ to be the smallest eigenvalue of $-A$ and suppose that $H(\rho_0|\rho_\infty) < \infty$.*

(i) If all eigenvalues of A are non-defective, then there exists a constant $c \geq 1$ such that

$$H(\rho_t|\rho_\infty) \leq cH(\rho_0|\rho_\infty)e^{-2\nu t} \quad \forall t \geq 0\,.$$

(ii) If one or more eigenvalues are defective, then there exists a constant $c_\varepsilon > 1$ for all $\varepsilon \in (0, \nu)$, such that

$$H(\rho_t|\rho_\infty) \leq c_\varepsilon H(\rho_0|\rho_\infty)e^{-2(\nu-\varepsilon)t} \quad \forall t \geq 0\,.$$

The actually observed relaxation behaviour is explored in Sect. 5 where we investigate the influence of temperature and the choice of initial conditions. Further, we study the occurrence of plateaus in the decay and processes with multiple time scales.

4 Exit Probabilities

If the principal eigenvalue of the drift matrix A in the SDE (1) is simple, then the Csiszár-Kullback-Pinsker inequality together with the entropy estimate in Theorem 1 implies that (see e.g. [11] and the references therein)

$$\|\rho_t - \rho_\infty\|_{L^1(\mathbb{R}^n)} \leq \sqrt{2cH(\rho_0|\rho_\infty)}\, e^{-\nu t}, \tag{13}$$

where ν is minus the real part of the principal eigenvalue of A and $c \geq 1$ is a constant. Note, however, that at low temperature (i.e. for small noise) the stationary distribution ρ_∞ of (1) shrinks to a point mass δ_0 concentrated at the origin $x = 0$, and as a consequence the upper bound in (13) degrades for most initial data. In this case the above estimate may not be so informative, and other techniques come into play, such as couplings based on Wasserstein distances

[9], or spectral estimates [10]. Spectral estimates for reversible systems play a huge role in analysing reversible molecular dynamics [45], climate modelling [31], or computational statistics [12], and the reason for this is that the principal eigenvalue of the generator of a reversible diffusion is conventionally associated with the characteristic time scale of the corresponding process; the rationale is that, for many nonlinear reversible systems at low noise, the principal eigenvalue is approximately inversely proportional to the mean first exit time from the deepest energy well, which defines the slowest time scale in the system.

The idea here is to discuss some connections between the dominant spectrum of the generator, entropy decay rates and "local" quantities such as covariance matrices or exit times for non-reversible ergodic OU processes that can be seen as linearisations of more complicated dynamics.

Under the Assumptions 1 and 2 it is a known result from [36] that the infinitesimal generator

$$L = \frac{1}{2} CC^T : \nabla^2 + (Ax) \cdot \nabla \tag{14}$$

associated with (1) has a compact resolvent and therefore a discrete spectrum in $L^p(\mathbb{R}^n, \rho_\infty)$ for $1 < p < \infty$ that can be completely characterised in terms of the eigenvalues of the matrix A; in particular, all eigenvalues of L have multiplicity 1 if and only if A is diagonalisable, and the eigenvalues are independent of p for $1 < p < \infty$. (For $p = 1$, the spectrum of L is the closed left-half plane [36]).

For reversible systems with $A = A^T$ and C being a scalar multiple of the identity $I_{n\times n}$, the spectral properties of L imply exponential convergence of the weighted density $\eta_t = \rho_t/\rho_\infty$ to $\eta_\infty = \mathbf{1}$ in $L^2(\mathbb{R}^n, \rho_\infty)$. It is easy to see, however, that this setting requires that $\eta_0 \in L^2(\mathbb{R}^n, \rho_\infty)$ which is equivalent to the assumption that ρ_0 is in $L^2(\mathbb{R}^n, \rho_\infty^{-1})$; even in the simple situation at hand, this is quite restrictive in that it excludes many standard cases, such as sharp Gaussian or point-like initial conditions, besides that the arguments are restricted to reversible systems only.

4.1 Large Deviations: Exit from a Set

Here we describe an alternative characterisation of the speed of convergence that is based on large deviations arguments and that includes non-reversible systems. To this end, we scale the diffusion matrix according to

$$C \mapsto \sqrt{\beta^{-1}} C\,, \quad \beta > 0. \tag{15}$$

We are interested in the situation $\beta \gg 1$, and, specifically, we want to study the probability that the process $X_t = X_t^\beta$ leaves a bounded and open set O that contains the unique stable fixed point $x = 0$. To this end let

$$\tau = \inf\{t > 0 \colon X_t \notin O\} \tag{16}$$

be the first exit time from the set O. As a consequence of the Donsker-Varadhan large deviations principle [15], the quantity

$$\gamma = -\lim_{t\to\infty} \frac{1}{t} \log P(\tau > t \,|\, X_0 = x) \tag{17}$$

is the principal eigenvalue of $-L$ equipped with homogeneous Dirichlet boundary values on ∂O that we assume to be smooth. That is, $\gamma > 0$ is the smallest eigenvalue such that

$$\begin{aligned} -L\varphi(x) &= \gamma\varphi(x)\,, \quad x \in O \\ \varphi(x) &= 0\,, \qquad x \in \partial O, \end{aligned} \tag{18}$$

where it follows from, e.g. [5, Thm. 1.3] and the fact that L satisfies a weak maximum principle that the principal eigenvalue is real; see also [21].

The interpretation of (17)–(18) is straightforward: the closer the principal eigenvalue $\gamma > 0$ is to zero, the smaller is the probability to observe an exit from the set O before time t, where the dependence is exponential in γ; in other words, the exit time for large t is exponentially distributed with parameter γ.

The relationship between (17) and (18) can be formally derived using the Feynman-Kac theorem for parabolic boundary value problems (e.g. [41, Chapters 8–9]), together with the separation ansatz (cf. [46, Sec. 5.1])

$$P(\tau > t | X_0 = x) \simeq \varphi(x) \exp(-\gamma t) \quad \text{as} \quad t \to \infty, \tag{19}$$

for some non-negative function φ. The ansatz is suggested by the asymptotic formula (17) and the symbol "$\simeq$" should be understood likewise; separation of variables in the Feynman-Kac formula then shows that φ solves the eigenvalue problem (18) where $\varphi > 0$ in the interior of the domain, as a consequence of the Perron-Frobenius theorem and Assumptions 1 and 2.

Remark 1. For non-degenerate, reversible systems, the exit probability is *exactly* exponential when the initial probability density for X_0 is the solution of the eigenvalue equation, with L being replaced by its formal L^2 adjoint L^*. The corresponding eigenfunction is called the *quasi-stationary distribution*, and it has the property that exit times are exponentially distributed, which is relevant in the context of parallelised molecular sampling algorithms [32].

4.2 Small-Noise Approximation of the Principal Eigenvalue

We seek a computable and easily interpretable expression for γ, and we will argue that γ can be computed from the stationary covariance matrix Σ_∞. To this end, we exploit a specific stochastic control interpretation of the principal eigenvalue that is along the lines of related work on non-degenerate diffusions by Fleming and co-workers [18–20]. Specifically, using that the function φ in (18) is strictly positive in the interior of the domain, it follows that $v = -\beta^{-1} \log \varphi$ solves the nonlinear boundary value problem

$$Lv - \frac{1}{2}|\nabla v|^2_{CC^T} = \gamma/\beta \tag{20}$$

with $|w|_{CC^T} = \sqrt{w^T CC^T w}$ denoting a weighted Euclidean pseudo-norm with weight $CC^T \geq 0$ and the specification

$$v(x) \to \infty \quad \text{as} \quad \operatorname{dist}(x, \partial O) \to 0 \tag{21}$$

for the function v when its arguments approach the boundary of O. Noting that

$$-\frac{1}{2}|w|^2_{CC^T} = \min_{a\in\mathbb{R}^n}\left\{\frac{1}{2}|a|^2 + (Ca)\cdot w\right\}, \tag{22}$$

we observe that (20) is the dynamic programming equation of an ergodic stochastic control problem, which implies the following result:

Proposition 2. *Under the previous assumptions, it holds that a.s.*

$$\gamma = \min_{u\in\mathcal{U}}\lim_{T\to\infty}\frac{\beta}{T}\mathbb{E}\left(\frac{1}{2}\int_0^T |u_t|^2\,\mathrm{d}t - \log \mathbf{1}_{\{\tau>T\}}\right) \tag{23}$$

where $\tau = \tau^u$ is the first exit time of the set O under the controlled process

$$\mathrm{d}X_t^u = (Cu_t + AX_t^u)\mathrm{d}t + \sqrt{\beta^{-1}}C\,\mathrm{d}W_t. \tag{24}$$

and the minimisation is over all Markovian controls $u \in \mathcal{U}$ such that (24) has a unique strong solution. Furthermore the minimum is unique and attained at $u_t^* = \beta^{-1}C^T\nabla\log\varphi(X_t^{u^*})$ with $\varphi \in C^2(O)\cap C(\overline{O})$ being the solution of (18).

Proposition 2 can be proved using a minor modification of the arguments in [46, Sec. 3.1] and we refer the reader to this article; see also [27, Thm. 2.3] for an existence and uniqueness theorem under more general assumptions.

What is important for us here is that, in the limit $\beta\to\infty$, the corresponding dynamic programming Eq. (20) can be explicitly solved. Let

$$\Phi(x) = \lim_{T\to\infty}\min_u\left\{\frac{1}{2}\int_0^T |u(t)|^2\,\mathrm{d}t : y(0)=0,\, y(T)=x\right\} \tag{25}$$

with $y(t) = y(t;t_0,y_0)$ being the solution of

$$\dot{y}(t) = Ay(t) + Cu(t)\,,\quad y(t_0) = y_0. \tag{26}$$

Equations (25)–(26) are the deterministic counterpart of the stochastic control problem (23)–(24). The corresponding dynamic programming equation that can be formally derived from (20) by letting $\beta\to\infty$ reads

$$(Ax)\cdot\nabla\Phi - \frac{1}{2}|\nabla\Phi|^2_{CC^T} = 0\,,\quad \Phi(0)=0, \tag{27}$$

where Φ is—in contrast to the solution of the dynamic programming Eq. (20)—bounded on $\overline{O}$, as a consequence of the complete controllability of the control system (26). A simple calculation shows that

$$\Phi(x) = \frac{1}{2}x^T\Sigma_\infty^{-1}x, \tag{28}$$

with $\Sigma_\infty\in\mathbb{R}^{n\times n}$ being the unique symmetric and positive definite solution of the Lyapunov equation $A\Sigma_\infty + \Sigma_\infty A^T + CC^T = 0$. The next statement is a straight consequence of the previous considerations and [50, Thm. 6]:

Corollary 1. *If* $O = \{x \in \mathbb{R}^n : |x| < 1\}$, *then*

$$\lim_{\beta \to \infty} \beta^{-1} \log \gamma = -(2\Lambda)^{-1}, \tag{29}$$

where $\Lambda > 0$ *is the largest eigenvalue of the asymptotic covariance matrix* Σ_∞.

We can interpret this result as follows: Recalling the Donsker-Vardhan large deviations principle (17), we can conclude that the probability of observing an exit from the n-dimensional unit sphere before time t behaves like

$$P(\tau \leq t | X_0 = x) \simeq 1 - \exp\left(-t \exp\left(-\frac{\beta}{2\Lambda}\right)\right), \quad t, \beta \to \infty \tag{30}$$

in the low-temperature regime where it can be readily seen that the limits $t \to \infty$ and $\beta \to \infty$ commute (cf. [50]). In other words, the probability of observing an exit before time t is large whenever the system is "easily controllable" (i.e. has large variance in some direction), whereas it is small if the system is "hardly controllable" (i.e. has uniformly small variance in all degrees of freedom).

Remark 2. If one accepts the underlying small-noise assumption, then Corollary 1 can be seen as a rationalisation of the usual interpretation of the principal eigenvalue of a reversible diffusion as a characteristic time scale, beyond the reversible setting. The fact that the exit from the set O follows an exponential distribution implies that the mean first exit time satisfies (cf. [50, Thm. 6])

$$\mathbb{E}[\tau] \simeq \exp\left(\frac{\beta}{2\Lambda}\right), \quad \beta \to \infty. \tag{31}$$

For reversible, non-degenerate Ornstein-Uhlenbeck processes with $A = A^T$ and $C = I_{n \times n}$, the largest non-zero eigenvalue λ_1 of the operator L with Dirichlet boundary data satisfies [10, Thm. 1.2]

$$\lambda_1 = -(\mathbb{E}(\tau))^{-1}(1 + \mathcal{O}(e^{-M\beta})), \quad \beta \to \infty, \tag{32}$$

for some constant $M > 0$, and, by comparing (29) and (31), we observe that this is consistent with the situation for general ergodic Ornstein-Uhlenbeck processes.

Remark 3. The asymptotic formula (29) is furthermore consistent with the related results by Kifer (e.g. [17, Thm. 2.1]) or Freidlin and Wentzell (e.g. [22, Thms. 7.1 and 7.4]) for nonlinear, non-degenerate diffusions that state that

$$\gamma \simeq e^{-\beta R} \quad \text{as} \quad \beta \to \infty, \tag{33}$$

where $R = \min\{\Phi(x) : x \in \partial O\}$, with Φ being given by the solution of (27) or an appropriate generalisation of it in the nonlinear setting.

4.3 Relation to Entropy Decay Rates

We shall briefly discuss the relation between the two exponential time scales $\nu = \min\{\Re(\lambda) : -Av = \lambda v\}$ and $\Lambda = \max\{\lambda : \Sigma_\infty v = \lambda v\}$ that are determined by the eigenvalues of the matrix A and the asymptotic covariance matrix Σ_∞.

Proposition 3. *Let A and C fulfil Assumptions 1 and 2, and let Σ_∞ solve the corresponding Lyapunov Eq. (3) for $\beta = 1$. Let w be the normalised eigenvector of $-A^T$ which corresponds to ν, i.e. $-A^T w = \nu w$. Introduce the splitting of $w = w_{\mathrm{Ker}} + w_{\mathrm{Im}}$, where $w_{\mathrm{Ker}} \in ker(D)$, $D = CC^T$ and $w_{\mathrm{Im}} \in Im(D)$. Further denote by $\lambda_{\min}(D)$ the smallest non-zero eigenvalue of D. Then*

$$\nu \geq \frac{\lambda_{\min}(D)}{2\Lambda} |w_{\mathrm{Im}}|. \tag{34}$$

Proof. First note, that due to Assumption 2 we have $w_{\mathrm{Im}} \neq 0$, but $w_{\mathrm{Ker}} = 0$ is possible. Multiplying the Lyapunov equation $A\Sigma_\infty + \Sigma_\infty A^T = -D$ from the left and right by w^T and w we find that $2\nu w^T \Sigma_\infty w = w^T D w$.

Now, $w^T \Sigma_\infty w \leq \Lambda$ and $w^T D w = w_{\mathrm{Im}}^T D w_{\mathrm{Im}} \geq \lambda_{\min}(D)|w_{\mathrm{Im}}|$ which together yields the assertion.

The inequality (34) is sharp. In the reversible case with $A = A^T$ negative definite and $C = I_{n\times n}$, the matrix $A = -\nabla^2 V$ can be interpreted as the Hessian of a quadratic potential

$$V(x) = -\frac{1}{2} x^T A x, \tag{35}$$

such that the stationary distribution ρ_∞ of (1) has a density proportional to $\exp(-V/2)$. As a consequence, the Lyapunov equation $A\Sigma_\infty + \Sigma_\infty A + I_{n\times n} = 0$ has an explicit solution with $2\Sigma_\infty = -A^{-1}$, and thus

$$\nu = \frac{1}{2\Lambda}. \tag{36}$$

Remark 4. We stress that, even though the relation (34) is independent of the parameter β, the interpretation of Λ as a characteristic time scale is *not*. In particular, (36) implies that the mean first exit time from the set O for a reversible system grows exponentially with the "stiffness" ν, i.e. $\mathbb{E}(\tau) \simeq e^{\nu\beta}$ as $\beta \to \infty$.

5 Numerical Examples

We will restrict ourselves to two and three dimensional examples, since the intention here is to only illustrate certain characteristics of the convergence behaviour of the system. As in the previous section we consider processes described by a SDE of the general form

$$\mathrm{d}X_t = AX_t\,\mathrm{d}t + \sqrt{\beta^{-1}} C\,\mathrm{d}W_t \tag{37}$$

where A and C fulfil the necessary Assumptions 1 and 2.

We will first discuss general dependencies for a given system (A, C) with respect to the temperature and initial conditions. Also, we shortly discuss the occurrence of plateaus and at the end of the section the case where the system has multiple time scales and suggest a purely data-based identification of slow and fast degrees of freedom.

5.1 Dependencies on the Temperature and Initial Conditions

In order to study the influence which temperature has on our system we split up the relative entropy into three different terms of which the first two correspond to the relaxation of the covariance and the last one to the relaxation of the mean.

$$\begin{aligned} H(t) &:= \int \log\left(\frac{\rho_t}{\rho_\infty}\right)\rho_t \,\mathrm{d}x \\ &= \frac{1}{2}\left[\underbrace{\mathrm{Tr}(\Sigma_t\Sigma_\infty^{-1}) - n}_{=a(t)} + \underbrace{-(\log\det(\Sigma_t\Sigma_\infty^{-1}))}_{=b(t)} + \underbrace{\mu_t^T\Sigma_\infty^{-1}\mu_t}_{=c(t)}\right]. \end{aligned}$$

Specifically, we can interpret these terms in the following sense. The term a embodies the relaxation of the covariance Σ_t to Σ_∞, whereas b ensures the normalization of the densities and finally c comprises the relaxation of the mean μ_t to 0.

We consider the following example where drift and diffusion are given by

$$A = -\begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

In Fig. 1 we illustrate temperature-effects. Recall that for a process given by (37) and $\Sigma_0 = 0$, the covariance Σ_t at time t is given by

$$\Sigma_t = \beta^{-1}\int_0^t e^{A(t-s)}CC^T e^{A^T(t-s)}\,\mathrm{d}s. \tag{38}$$

Hence, the only term which is temperature dependent is $c(t)$, due to Σ_∞^{-1}. Furthermore $c(t)$ grows as $\beta \to \infty$, i.e. for small temperatures $c(t)$ dominates the relaxation behaviour. The terms $a(t)$ and $b(t)$ do not have any temperature dependence because of the multiplication by another temperature dependent term Σ_t, such that β and β^{-1} cancel. This means that for larger temperatures the relaxation is governed by a and b, which describe the equilibration of the covariance (see Fig. 1 right panel and Fig. 2 upper right panel). Note that a and b always have opposite signs, and they change sign depending on whether $\Sigma_t > \Sigma_\infty$ or $\Sigma_t < \Sigma_\infty$. In the first case, i.e. when $\Sigma_t > \Sigma_\infty$, $a(t)$ is strictly positive and contributes the most, in the second case it is $b(t)$ that dominates. For small temperature $c(t)$ plays this role (see Fig. 1 left panel), and the overall relaxation is mainly determined by the relaxation of the mean.

Figure 2 shows the strong influence of the initial conditions on the relaxation behaviour, which is the only parameter varied in this figure. If we choose

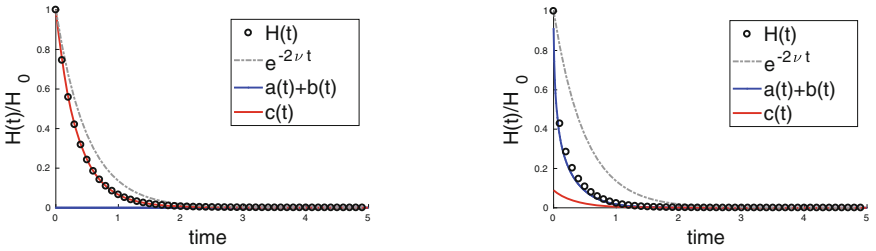


Fig. 1. Temperature-effects: the initial conditions are fixed. Left: low temperature ($\beta = 10^3$), right: high temperature ($\beta = 10^{-2}$).

an eigenvector of the drift matrix A as a deterministic initial condition, this will yield exponential decay with the corresponding eigenvalue (left panel). The initial conditions can also be chosen such that one observes a plateau where $\dot{H}(t) = 0$ (right panel). This also leads to the constant c of Theorem 1 being strictly greater than 1. Which term contributes the most to the total relaxation behaviour then depends on the choice of the initial covariance Σ_0. If $\Sigma_0 > \Sigma_\infty$ then $a(t)$ is the governing term (if the temperature is not too low), otherwise $c(t)$ will take this role.

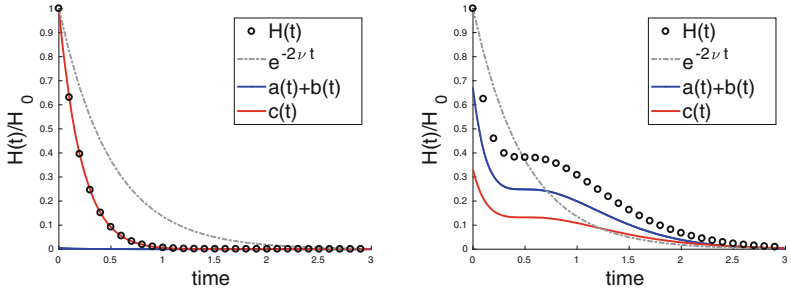


Fig. 2. Influence of the initial conditions when the temperature is fixed $\beta = 20$. Left: $x_0 = (\frac{20}{3}, \frac{10}{3})^T$ (eigenvector of A corresponding to $\lambda = 2$); Right: $x_0 \sim \mathcal{N}((0.3513, -0.5496)^T, \Sigma_0)$, $\Sigma_0 > 0$.

5.2 Multiple Time Scales: Partitioning into Slow and Fast

We now split up the relative entropy into two terms, where one depends on conditional distributions and the other on marginal ones. More specifically, consider a process $(Z_t)_{(t\geq 0)}$ which consists of two components $Z = (X, Y)$. We will think of X being the slow component and Y the fast one. Denote by $\rho(z)$ the density of the joint process, by $\bar{\rho}(x)$ the marginal density of X and by $\hat{\rho}(y; x)$ the conditional density of Y where $X = x$ is given. We can always do the following

computation which yields a partition of the relative entropy into conditional and marginal terms:

$$\begin{aligned} H_Z(t) :&= H(\rho_t|\rho_\infty) \\ &= \int\int \bar{\rho}_t\ \hat{\rho}_t\ \log\left(\frac{\bar{\rho}_t}{\bar{\rho}_\infty}\right) \mathrm{d}y\,\mathrm{d}x + \int\int \bar{\rho}_t\ \hat{\rho}_t\ \log\left(\frac{\hat{\rho}_t}{\hat{\rho}_\infty}\right) \mathrm{d}y\,\mathrm{d}x \\ &= H(\bar{\rho}_t|\bar{\rho}_\infty) + \int H(\hat{\rho}_t|\hat{\rho}_\infty)\bar{\rho}_t\,\mathrm{d}x \\ &= H_X(t) + \mathbb{E}_{\bar{\rho}_t}(H_{Y|X=x}(t)). \end{aligned}$$

In the example of this section we investigate the contribution of the two terms, namely the conditional and the marginal term, to the overall decay in relative entropy. Note that, obviously a splitting into a marginal term H_Y and a conditional term $H_{X|Y=y}$ is possible in the same way.

We consider our previous example, but introduce a timescale parameter $0 < \varepsilon \leq 1$ such that the coefficients now read

$$A = -\begin{pmatrix} 1 & 3\varepsilon^{-1} \\ 0 & 2\varepsilon^{-2} \end{pmatrix}, \quad C = \begin{pmatrix} 0 \\ \varepsilon^{-1} \end{pmatrix}.$$

Note that for $\varepsilon \to 0$, the first component of the dynamics approaches an SDE with effective coefficients $\bar{A} = 1$ and $\bar{C} = \frac{3}{2}$; cf. [42].

Before we come to the numerical results, let us give the constants ν and Λ that correspond to the convergence and exit time behavior of the system as described in Sects. 3 and 4. The smallest eigenvalue of A is $\nu = 1$. The covariance matrix of the stationary solution which solves (3) is given by

$$\Sigma_\infty = \begin{pmatrix} \frac{9}{4(2+\varepsilon^2)} & -\frac{3\varepsilon}{4(2+\varepsilon^2)} \\ -\frac{3\varepsilon}{4(2+\varepsilon^2)} & \frac{1}{4} \end{pmatrix},$$

and its largest eigenvalue is $\Lambda = \frac{11+\varepsilon^2+\sqrt{49+22\varepsilon^2+\varepsilon^4}}{8(2+\varepsilon^2)} \leq \frac{21}{8}$ for $\varepsilon \in (0, 1]$ and we see that the statement of Proposition 3 is fulfilled.

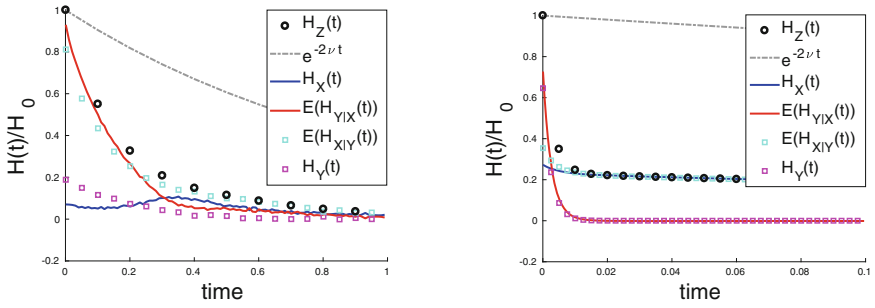


Fig. 3. Decay of the relative entropy in terms of conditional and marginal terms without time scale separation, i.e. $\varepsilon = 1$ (left) and when there is a time scale separation $\varepsilon = 0.1$ (right) and initial condition $z_0 = (1, -0.0655)^T$.

No Time Scale Separation. This case is depicted in Fig. 3 in the left panel. We display both, the splitting into the H_X, $\mathbb{E}(H_{Y|X})$ (solid lines) and H_Y, $\mathbb{E}(H_{X|Y})$ (squares). The marginal term H_X is not monotonically decreasing in time, but can in fact increase. This is due to the fact that when computing the time derivative of e.g. $H(\bar{\rho}_t|\bar{\rho}_\infty)$ one finds as usual the Fisher information, but additionally another term appears which can be estimated by the *empirical measure large deviations rate functional*; see [44, Ch. 2]. The empirical measure large deviations rate functional of $\bar{\rho}_t$ and $\bar{\rho}_\infty$ will in general be non-zero, since the time evolution of $\bar{\rho}_t$ is still described by the Fokker-Planck equation of the full process which differs from the Fokker-Planck equation which has $\bar{\rho}_\infty$ as equilibrium.

Note that the increase of the relative entropy in time cannot be traced back to the irreversibility of the process, but can also be observed for reversible processes with appropriate initial conditions.

We want to emphasise that in the case of no time scale separation no clear judgement is possible as to which of the terms displays a fast or slow relaxation behaviour. This is true for both splittings and will be contrasted with the case of a clear time scale separation below.

Time Scale Separation. We introduce a time scale separation by setting $\varepsilon = 0.1$ (see Fig. 3 right panel) and now refer to X as the slow process and Y as the fast one. The *a priori* assignment of slow and fast degrees of freedom agrees with the observation in the plots: for $\varepsilon \to 0$ the conditional term $H_{Y|X=x}(t)$ relaxes almost instantaneously to its equilibrium. Accordingly, the marginal term $H_X(t)$ governs the long term behaviour of the overall relaxation. For the other splitting, we observe the same behaviour. The marginal of Y converges very fast whereas the conditional of X dominates the long term relaxation. This observation suggests that we can use the partitioning of relative entropy into conditional and marginal terms as a definition for fast and slow degrees of freedom. Both splittings seem reasonable in our setting, but this is due to the linearity of our system. In regard to applying this idea to nonlinear diffusions we propose to use the marginal term for the slow and the conditional term for the fast variable. To be more precise let us consider the nonlinear example given by the SDE

$$\begin{aligned} dZ &= -\nabla V(Z)dt + dB_t, \quad Z = (X, Y) \in \mathbb{R}^2, \\ V(x, y) &= (x^2 - 1)^2 + \varepsilon^{-1} \left(1 + e^x\right)^{-1} y^2. \end{aligned}$$

This SDE describes the diffusive motion of a particle in the potential energy landscape V. In x direction there are two metastable states, given by the domains around the minima at $x = \pm 1$ and in between there is a barrier to overcome. In y direction the motion is confined by a quadratic potential with differing growth which is minimal at $x = 0$. We expect that for each fixed x the conditional distribution of Y will quickly approach its equilibrium, contrary to the marginal of Y which needs the slow variable X to cross the barrier of the potential at $x = 0$ in order to converge to its equilibrium distribution.

Furthermore, we observe that as $\varepsilon \to 0$ all terms become monotonically decreasing.

6 Outlook and Discussion

In the previous section we have seen that relative entropy may be used as a tool to define fast and slow degrees of freedom. That is, the fast variable is defined via an almost immediate relaxation of its conditional density $\hat{\rho}_t$, with the slow variable being fixed. At the same time the slow variable is defined via the relatively slow relaxation of its marginal density $\bar{\rho}_t$. Furthermore, the slow variable will govern the collective relaxation after very short time once the fast one has relaxed. This definition of fast and slow agrees with the coarse-graining concepts of averaging and homogenization in the reversible case and the conditional expectation (cf. [34]) in the general case. These methods seek low dimensional effective dynamics which are built by computing expectations of the slow variables' dynamics with respect to the conditional invariant distributions of the fast variables given the slow ones. The underlying idea is that the fast variables relax almost instantaneously such that their force on the dynamics is well captured by the statistics of their invariant distribution.

In order to identify slow and fast sub-processes, one could argue that for OU processes it is also possible to resort to the spectral decomposition of the associated generator L defined in (14) which is explicitly known in this case [36]. But even here, if the eigenvalues of the drift matrix A are complex and hence the spectrum is not well interpretable, or else, in a more general setting, if the generator or its spectrum is not known, other methods are needed. For this we propose to use relative entropy as a purely data-based tool, which can detect slow and fast degrees of freedom and furthermore might lead towards an understanding between the different concepts of time scales for general non-reversible and degenerate diffusions. To make this idea more precise recall that in the case of a reversible diffusion described by

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{\beta^{-1}}\,\mathrm{d}W_t \tag{39}$$

with V being a confinement potential which grows sufficiently fast at infinity, it is known that, in the small temperature limit, the slowest processes—given by mean first passage times across the highest energy barriers—can be associated with the eigenvalues of the generator in a hierarchical manner [10]. Furthermore, the eigenvalues describe the convergence to equilibrium in the ρ_∞^{-1} weighted L^2 norm. This analysis is inherent to the reversible case but a natural generalisation of convergence to equilibrium in L^2 is given by convergence in relative entropy. Hence, the question of whether one is able to formulate a hierarchical ordering of the systems processes according to the relaxation time scales determined by the convergence in relative entropy, for reversible as well as irreversible processes, is relevant (cf. [37, Remark 2.16]). We leave these questions for future research.

Acknowledgements. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through the grant CRC 1114 "Scaling Cascades in Complex Systems", Projects A05 "Probing scales in equilibrated systems by optimal nonequilibrium forcing" and B05 "Origin of the scaling cascades in protein dynamics".

References

1. Arnold, A., Carlen, E., Ju, Q.: Large-time behavior of non-symmetric Fokker-Planck type equations. Commun. Stoch. Anal. **2**(1), 153–175 (2008)
2. Arnold, A., Erb, J.: Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift. arXiv preprint arXiv:1409.5425 (2014)
3. Arnold, A., Markowich, P., Toscani, G., Unterreiter, A.: On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. Commun. Part. Differ. Equ. **26**(1–2), 43–100 (2001)
4. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia (2005)
5. Birindelli, I., Demengel, F.: First eigenvalue and maximum principle for fully nonlinear singular operators. Adv. Differ. Equ. **11**(1), 91–119 (2006)
6. Bovier, A., den Hollander, F.: Metastability: A Potential-Theoretic Approach. Grundlehren der mathematischen Wissenschaften, vol. 351. Springer, New York (2015)
7. Bakry, D., Émery, M.: Diffusions hypercontractives. In: Azéma, J., Yor, M. (eds.) Séminaire de Probabilités XIX 1983/84. Lecture Notes in Mathematics, vol. 1123, pp. 177–206. Springer, Heidelberg (1985)
8. Berglund, N.: Kramers' law: validity, derivations and generalisations. Markov Process. Relat. **19**(3), 459–490 (2013)
9. Bolley, F., Gentil, I., Guillin, A.: Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. J. Funct. Anal. **263**(8), 2430–2457 (2012)
10. Bovier, A., Gayrard, V., Klein, M.: Metastability in reversible diffusion processes II: precise asymptotics for small eigenvalues. J. Eur. Math. Soc. **7**(1), 69–99 (2005)
11. Bolley, F., Villani, C.: Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. Annales de la Faculté des Sciences de Toulouse: Mathématiques **14**(3), 331–352 (2005)
12. Cowles, M.K., Carlin, B.P.: Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. **91**(434), 883–904 (1996)
13. Chetrite, R., Touchette, H.: Nonequilibrium Markov processes conditioned on large deviations. Annales Henri Poincaré **16**(9), 2005–2057 (2015)
14. Day, M.V.: Recent progress on the small parameter exit problem. Stoch. Int. J. Probab. Stoch. Process. **20**(2), 121–150 (1987)
15. Donsker, M.D., Srinivasa Varadhan, S.R.: On a variational formula for the principal eigenvalue for operators with maximum principle. Proc. Natl. Acad. Sci. **72**(3), 780–783 (1975)
16. Dupuis, P., Wang, H.: Importance sampling, large deviations, and differential games. Stoch. Stoch. Rep. **76**(6), 481–508 (2004)
17. Eizenberg, A., Kifer, Y.: The asymptotic behavior of the principal eigenvalue in a singular perturbation problem with invariant boundaries. Probab. Theory Relat. Fields **76**(4), 439–476 (1987)
18. Fleming, W.H.: Exit probabilities and optimal stochastic control. Appl. Math. Optim. **4**, 329–346 (1977)
19. Fleming, W.H., McEneaney, W.M.: Risk-sensitive control on an infinite time horizon. SIAM J. Control. Optim. **33**(6), 1881–1915 (1995)
20. Fleming, W.H., Sheu, S.-J.: Asymptotics for the principal eigenvalue and eigenfunction of a nearly first-order operator with large potential. Ann. Probab. **25**, 1953–1994 (1997)

21. Fleming, W.H., Sheu, S.J., Soner, H.M.: A remark on the large deviations of an ergodic Markov process. Stochastics **22**(3–4), 187–199 (1987)
22. Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems, vol. 260. Springer, Heidelberg (2012)
23. Huisinga, W., Meyn, S., Schütte, C.: Phase transitions and metastability in Markovian and molecular systems. Ann. Appl. Probab. **14**(1), 419–458 (2004)
24. Huisinga, W., Schmidt, B.: Metastability and dominant eigenvalues of transfer operators. In: Leimkuhler, B., Chipot, C., Elber, R., Laaksonen, A., Mark, A., Schlick, T., Schütte, C., Skeel, R. (eds.) New Algorithms for Macromolecular Simulation, pp. 167–182. Springer, Heidelberg (2006)
25. Hartmann, C., Schütte, C., Zhang, W.: Model reduction algorithms for optimal control and importance sampling of diffusions. Nonlinearity **29**(8), 2298–2326 (2016)
26. Hänggi, P., Talkner, P., Borkovec, M.: Reaction-rate theory: fifty years after Kramers. Rev. Mod. Phys. **62**, 251–341 (1990)
27. Ichihara, N.: Large time asymptotic problems for optimal stochastic control with superlinear cost. Stoch. Proc. Appl. **122**(4), 1248–1275 (2012)
28. Imkeller, P., Von Storch, J.-S.: Stochastic Climate Models. Progress in Probability. Springer, New York (2001)
29. Jack, R.L., Sollich, P.: Large deviations of the dynamical activity in the east model: analysing structure in biased trajectories. J. Phys. A **47**(1), 015003 (2014)
30. Kaiser, M., Jack, R.L., Zimmer, J.: Acceleration of convergence to equilibrium in Markov chains by breaking detailed balance. J. Stat. Phys. **168**(2), 259–287 (2017)
31. Lucarini, V., Bódai, T.: Edge states in the climate system: exploring global instabilities and critical transitions. Nonlinearity **30**(7), R32 (2017)
32. Le Bris, C., Lelièvre, T., Luskin, M., Perez, D.: A mathematical formalization of the parallel replica dynamics. Monte Carlo Methods Appl. **18**(2), 119–146 (2012)
33. Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, New York (2004)
34. Legoll, F., Lelièvre, T.: Effective dynamics using conditional expectations. Nonlinearity **23**(9), 2131–2163 (2010)
35. Lelièvre, T., Stoltz, G.: Partial differential equations and stochastic methods in molecular dynamics. Acta Numer. **25**, 681–880 (2016)
36. Metafune, G., Pallara, D., Priola, E.: Spectrum of Ornstein-Uhlenbeck operators in L^p spaces with respect to invariant measures. J. Funct. Anal. **196**(1), 40–60 (2002)
37. Menz, G., Schlichting, A.: Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. Ann. Probab. **42**(5), 1809–1884 (2014)
38. Martorell, S., Soares, C.G., Barnett, J.: Safety, Reliability and Risk Analysis: Theory. Methods and Applications. CRC Press, Boca Raton (2014)
39. Markowich, P.A., Villani, C.: On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. Mat. Contemp. **19**, 1–29 (2000)
40. Nier, F., Helffer, B.: Hypoelliptic Estimates and Spectral Theory for Fokker-Planck Operators and Witten Laplacians. Lecture Notes in Mathematics, vol. 1862. Springer, Heidelberg (2005)
41. Øksendal, B.K.: Stochastic Differential Equations: An Introduction With Applications. Springer, New York (2003)
42. Pavliotis, G.A.: Stochastic processes and applications. Texts in Applied Mathematics, vol. 60. Springer, New York (2014)

43. Rey-Bellet, L., Spiliopoulos, K.: Irreversible Langevin samplers and variance reduction: a large deviations approach. Nonlinearity **28**(7), 2081–2104 (2015)
44. Sharma, U.: Coarse-graining of Fokker-Planck equations. Ph.D. thesis, Technische Universiteit Eindhoven (2017)
45. Schütte, C., Sarich, M.: Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches. Courant Lecture Notes, vol. 24. American Mathematical Society, Providence (2013)
46. Schütte, C., Winkelmann, S., Hartmann, C.: Optimal control of molecular dynamics using Markov state models. Math. Prog. Ser. B **134**, 259–282 (2012)
47. Ventsel', A.D.: Formulae for eigenfunctions and eigenmeasures associated with a Markov process. Theory Probab. Appl. **18**(1), 1–26 (1973)
48. Vanden-Eijnden, E., Weare, J.: Rare event simulation of small noise diffusions. Commun. Pure Appl. Math. **65**(12), 1770–1803 (2012)
49. Ventsel', A.D., Freidlin, M.I.: On small random perturbations of dynamical systems. Russ. Math. Surv. **25**(1), 1–55 (1970)
50. Zabczyk, J.: Exit problem and control theory. Syst. Control. Lett. **6**(3), 165–172 (1985)
51. Zabczyk, J.: Mathematical Control Theory: An Introduction. Springer, New York (2009)

Workshop 2: Life Sciences

Stochastic Models of Blood Vessel Growth

Luis L. Bonilla(✉), Manuel Carretero, and Filippo Terragni

G. Millan Institute, Department of Materials Science and Engineering,
Universidad Carlos III de Madrid, Leganés, Spain
bonilla@ing.uc3m.es,
{manuel.carretero,filippo.terragni}@uc3m.es

Abstract. Angiogenesis is a complex multiscale process by which diffusing vessel endothelial growth factors induce sprouting of blood vessels that carry oxygen and nutrients to hypoxic tissue. There is strong coupling between the kinetic parameters of the relevant branching - growth - anastomosis stochastic processes of the capillary network, at the microscale, and the family of interacting underlying biochemical fields, at the macroscale. A hybrid mesoscale tip cell model involves stochastic branching, fusion (anastomosis) and extension of active vessel tip cells with reaction-diffusion growth factor fields. Anastomosis prevents indefinite proliferation of active vessel tips, precludes a self-averaging stochastic process and ensures that a deterministic description of the density of active tips holds only for ensemble averages over replicas of the stochastic process. Evolution of active tips from a primary vessel to a hypoxic region adopts the form of an advancing soliton that can be characterized by ordinary differential equations for its position, velocity and a size parameter. A short review of other angiogenesis models and possible implications of our work is also given.

Keywords: Angiogenesis · Active vessel tip model · Stochastic differential equations · Reinforced random walk · Branching process · History-dependent killing process · Cellular Potts models · Integrodifferential equation for active tip density

1 Introduction

The growth of blood vessels out of a primary vessel or *angiogenesis* is a complex multiscale process responsible for organ growth and regeneration, tissue repair, wound healing and many other natural operations in living beings [1–5]. Angiogenesis is triggered by lack of oxygen (hypoxia) experienced by cells in some tissue. Such cells secrete growth factors that diffuse and reach a nearby primary blood vessel. In response, the vessel wall opens and issues endothelial cells that move towards the hypoxic region, build capillaries and bring blood, oxygen and

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 413–436, 2019.
https://doi.org/10.1007/978-3-030-15096-9_13

nutrients to it. Once blood and oxygen have reached the hypoxic region, secretion of growth factors stops, anti-angiogenic substances may be secreted and a regular vessel network may have been put in place, after pruning capillaries with insufficient blood flow. In normal functioning, angiogenic and anti-angiogenic activities balance. Imbalance may result in many diseases including cancer [6]. In fact, after a tumor installed in tissue reaches some 2 mm size, it needs additional nutrients and oxygen to continue growing. Its hypoxic cells secrete growth factors and induce angiogenesis. Unlike normal cells, cancerous ones continue issuing growth factors and attracting blood vessels, which also supply them with a handy transportation system to reach other organs in the body.

Tumor-induced angiogenesis research started with Folkman's pioneering work in 1971 [6]. In addition to vast experimental research [7], models and theory [8] substantially contribute to understanding angiogenesis and developing therapies. In angiogenesis, events happening in cellular and subcellular scales unchain endothelial cell motion and proliferation and build millimeter scale blood sprouts and networks thereof [2–5]. Models range from very simple to extraordinarily complex and often try to illuminate some particular mechanism; see the review [8]. Realistic microscopic models involve postulating mechanisms and a large number of parameters that cannot be directly estimated from experiments, but they often yield qualitative predictions that can be tested. An important challenge is to extract mesoscopic and macroscopic descriptions of angiogenesis from the diverse microscopic models.

During angiogenesis, the relevant branching, growth and anastomosis (vessel fusion) stochastic processes of the capillary network at the microscale are strongly coupled to the interacting underlying biochemical and mechanical fields at the macroscale. In Sect. 2, we consider a hybrid mesoscale tip cell model that involves stochastic branching, anastomosis and extension of active vessel tip cells with reaction-diffusion growth factor fields [9]. Numerical simulations of the model show that anastomosis prevents indefinite proliferation of active vessel tips [10]. Then fluctuations about the mean of the density of active tips are not small and the stochastic process is not self-averaging. However, as shown in Sect. 3, it is possible to obtain a deterministic description of the density of active tips for ensemble averages over replicas of the stochastic process. The deterministic description consists of an integro-partial differential equation for the density of active vessel tips coupled to a reaction-diffusion equation for the growth factor [9,10]. As shown in Sect. 4, the evolution of active tips from a primary vessel to a hypoxic region adopts the form of an advancing soliton-like wave that can be characterized by ordinary differential equations for its position, velocity and a size parameter [11,12]. These results may pave the way to assess optimal control of angiogenesis and therapies based on it.

What are the implications of our work? As described in Sect. 5, there are other models related to ours in which the vessel extension is described by random walks [13,14], and our methodology may be used to extract deterministic descriptions for the density of active tips amenable to analysis. We could also seek to extend microscopic cellular Potts models (described in Sect. 6) to mesoscales and study

them using our methods. The role of blood flow in remodeling vascular networks is briefly considered in Sect. 7. A quite different approach is presented in Sect. 8. Reaction-diffusion equations for growth factors are coupled to a Cahn-Hilliard type equation for a phase field that is fourth order in space. The phase field has a potential with two minima corresponding to the extracellular matrix and to the advancing blood vessels. There are conditions for the velocity of the capillaries and to create new ones. Further remarks are included in our conclusions in Sect. 9.

2 Langevin Tip Cell Models

Tip cell models assume that the tip cells are motile and non-proliferating whereas stalk cells build the blood vessel following the trajectories of the former. Assuming that the tip cells form point particles, their trajectories constitute the blood vessels advancing toward the hypoxic region. In 1991, Stokes and Lauffenburger considered the capillary sprouts as particles of unit mass subject to chemotactic, friction and white noise forces [15,16]. The distribution of vessel endothelial growth factors (VEGF) issuing from a small circular tumor (or from a small circular hypoxic region) is a known stationary non-uniform function. Associated to each sprout, its cell density satisfies a rate equation that takes into account proliferation, elongation, redistribution of cells from the parent vessel, branching and anastomosis. They did not consider the depletion effect that advancing sprouts would have on the VEGF concentration. Later tip cell models combined a continuum description of fields influencing cell motion (chemotaxis, haptotaxis, ...) with random walk motion of individual sprouts that experience branching and anastomosis. Capasso and Morale [17] used ideas from these approaches to propose a hybrid model of Langevin-Ito stochastic equations for the sprouts undergoing chemotaxis, haptotaxis, branching and anastomosis coupled to reaction-diffusion equations for the continuum fields. In this model, the evolution of the continuum fields is influenced by the growing capillary network through smoothed (or mollified) versions thereof [18]. Capasso and Morale also attempted to derive a continuum equation for the density of moving tip cells from the stochastic equations but could not account for branching and anastomosis [17]. In what follows, we present a simplified hybrid model that ignores haptotaxis and derive a deterministic description for the density of active tips [9,10,19]. As in the Capasso-Morale model, the influence of haptotaxis can be included by adding reaction-diffusion equations for fibronectin and matrix-degrading enzymes [20]. The influence of blood circulation through the newly created blood vessels and secondary branching therefrom can be modeled as in [21].

We shall consider a slab geometry as indicated in Fig. 1, which is the result of a numerical simulation of the stochastic model. The extension of the ith capillary sprout with position $\mathbf{X}^i(t)$ and velocity $\mathbf{v}^i(t)$ is given by the nondimensional Langevin-Ito stochastic equation

$$\begin{aligned} d\mathbf{X}^i(t) &= \mathbf{v}^i(t)\,dt \\ d\mathbf{v}^i(t) &= \beta\left[-\mathbf{v}^i(t) + \mathbf{F}\big(C(t,\mathbf{X}^i(t))\big)\right]dt + \sqrt{\beta}\,d\mathbf{W}^i(t) \end{aligned} \tag{1}$$

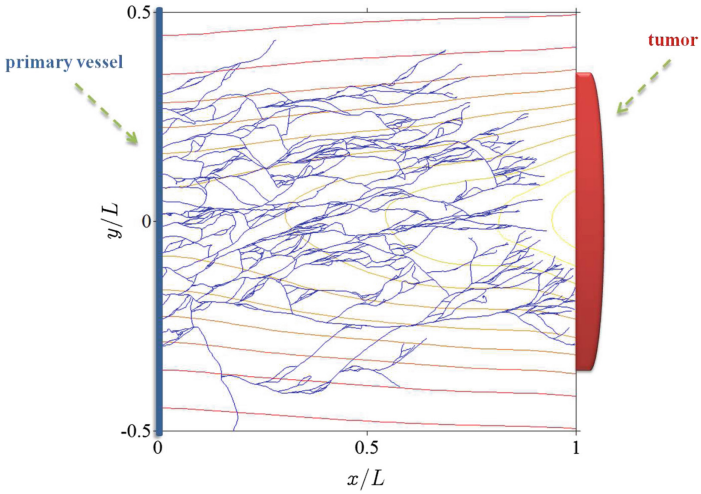


Fig. 1. Network of blood vessels simulated by the stochastic model of tumor induced angiogenesis. The level curves of the density of the tumor angiogenic factor (vessel endothelial growth factor) are also depicted, [11].

for $t > T^i$ (T^i is the random birth time of the ith tip). Here $C(t, \mathbf{x})$ is the VEGF concentration. At time T^i , the velocity of the newly created tip is selected out of a normal distribution with mean $\mathbf{v}_0$ and variance σ_v^2, while the probability that a tip branches from one of the existing ones during an infinitesimal time interval $(t, t + dt]$ is proportional to

$$\sum_{i=1}^{N(t,\omega)} \alpha(C(t, \mathbf{X}^i(t)))dt. \tag{2}$$

Here $N(t, \omega)$ is the number of tips at time t for a realization ω of the stochastic process and

$$\alpha(C) = \frac{A\,C}{C+1}, \tag{3}$$

where A is a positive constant. We ignore secondary angiogenesis from newly formed capillaries [21]. The tip i disappears at a later random time Θ^i, either by reaching the hypoxic region or by anastomosis, i.e., by meeting another capillary. At time t, anastomosis for the ith tip occurs at a point $\mathbf{x}$ such that $\mathbf{X}^i(t) = \mathbf{x}$ and $\mathbf{X}^j(s) = \mathbf{x}$ for another tip that was at $\mathbf{x}$ previously, at time $s < t$. Anastomosis reduces the importance of secondary angiogenesis, because: (i) newly formed capillaries need some time to mature and issue tip cells from their walls, and (ii) secondary branches appear in a crowded environment and their life before they

anastomose is typically short. In (1), $\mathbf{W}^i(t)$ are i.i.d. Brownian motions, and β (friction coefficient) is a positive parameter [9,10,12]. The chemotactic force $\mathbf{F}$ controlling tip cell migration in response to the VEGF released by hypoxic cells is

$$\mathbf{F}(C) = \frac{\delta_1}{1 + \Gamma_1 C} \nabla_x C, \tag{4}$$

where δ_1, and Γ_1 are positive parameters. The VEGF diffuses and is consumed by advancing vessel tips according to [10]

$$\frac{\partial C}{\partial t}(t, \mathbf{x}) = \kappa_c \Delta_x C(t, \mathbf{x}) - \chi_c C(t, \mathbf{x}) \left| \sum_{i=1}^{N(t,\omega)} \mathbf{v}^i(t)\, \delta_{\sigma_x}(\mathbf{x} - \mathbf{X}^i(t)) \right|. \tag{5}$$

Here κ_c and χ_c are positive parameters, while δ_{σ_x} is a regularized delta function (e.g., a Gaussian with standard deviation σ_x). We are assuming that extending the vessel consumes VEGF. As the vessel extends a length $|\mathbf{v}^i(t)|\, dt$ during the time interval between t and $t + dt$, the consumption should be proportional to $|\mathbf{v}^i(t)|$. The resulting equation for the VEGF is then

$$\frac{\partial C}{\partial t}(t, \mathbf{x}) = \kappa_c \Delta_x C(t, \mathbf{x}) - \tilde{\chi}_c C(t, \mathbf{x}) \sum_{i=1}^{N(t,\omega)} |\mathbf{v}^i(t)|\, \delta_{\sigma_x}(\mathbf{x} - \mathbf{X}^i(t)). \tag{6}$$

The difference between the more appropriate model equation (6) and (5) could be considerable for situations where tip cells are moving in all directions. However, for the parameters and the slab geometry considered in the numerical simulations presented in this paper, this difference is negligible (it amounts to having $\tilde{\chi}_c = 1.28\chi_c$ in the previous equation). Initial and boundary conditions for the VEGF field C have been proposed in [9,10].

The concentration of all vessels per unit volume in the physical space, at time t (i.e., the vessel network $\mathbf{X}(t, \omega)$) is [10]

$$\delta(\mathbf{x} - \mathbf{X}(t, \omega)) = \int_0^t \sum_{i=1}^{N(s,\omega)} \delta_{\sigma_x}(\mathbf{x} - \mathbf{X}^i(s, \omega))\, ds. \tag{7}$$

3 Deterministic Description

We shall see that we can understand the results of numerical simulations of the stochastic process described in the previous section by first finding a deterministic description of the density of active tips. The latter evolves in the form of a slowly varying soliton-like wave that we can analyze. Without performing numerical simulations of the stochastic process, we could guess that such a

deterministic description could hold whenever the number of active tips arising from branching becomes very large. In such a case, we could use the law of large numbers to achieve such a description. This was the point of view adopted in the papers [9,17]. However, anastomosis kills off so many active vessel tips that their number hardly grows to a hundred. Then we need a different point of view in order to derive a deterministic description. The alternative is the Gibbsian idea of considering an ensemble of replicas of the original stochastic process and carrying out arithmetic averages over the number of replicas.

We can find a deterministic description of the stochastic model for the densities of active vessel tips and the vessel tip flux, defined as ensemble averages over a sufficient number $\mathscr{N}$ of replicas (realizations) ω of the stochastic process:

$$p_{\mathscr{N}}(t,\mathbf{x},\mathbf{v}) = \frac{1}{\mathscr{N}} \sum_{\omega=1}^{\mathscr{N}} \sum_{i=1}^{N(t,\omega)} \delta_{\sigma_x}(\mathbf{x}-\mathbf{X}^i(t,\omega))\, \delta_{\sigma_v}(\mathbf{v}-\mathbf{v}^i(t,\omega)), \tag{8}$$

$$\tilde{p}_{\mathscr{N}}(t,\mathbf{x}) = \frac{1}{\mathscr{N}} \sum_{\omega=1}^{\mathscr{N}} \sum_{i=1}^{N(t,\omega)} \delta_{\sigma_x}(\mathbf{x}-\mathbf{X}^i(t,\omega)), \tag{9}$$

$$\mathbf{j}_{\mathscr{N}}(t,\mathbf{x}) = \frac{1}{\mathscr{N}} \sum_{\omega=1}^{\mathscr{N}} \sum_{i=1}^{N(t,\omega)} \mathbf{v}^i(t,\omega)\delta_{\sigma_x}(\mathbf{x}-\mathbf{X}^i(t,\omega)). \tag{10}$$

As $\mathscr{N} \to \infty$, these ensemble averages tend to the tip density $p(t,\mathbf{x},\mathbf{v})$, the marginal tip density $\tilde{p}(t,\mathbf{x})$, and the tip flux $\mathbf{j}(t,\mathbf{x})$, respectively.

Figures 2 and 3 show the outcomes of typical simulations of ensemble averaged marginal densities: The two-dimensional lump shown in Fig. 2 is created at the primary vessel at $x = 0$ and marches to the hypoxic region at $x = 1$. Its profile along the x axis is the soliton-like wave shown in Fig. 3.

Reference [10] shows that the angiogenesis model has a deterministic description based on the following equation for the density of vessel tips, $p(t,\mathbf{x},\mathbf{v})$,

$$\frac{\partial p}{\partial t}(t,\mathbf{x},\mathbf{v}) = \alpha(C(t,\mathbf{x}))\, p(t,\mathbf{x},\mathbf{v})\delta_{\sigma_v}(\mathbf{v}-\mathbf{v}_0) - \Gamma\, p(t,\mathbf{x},\mathbf{v}) \int_0^t \tilde{p}(s,\mathbf{x})\, ds$$
$$-\mathbf{v}\cdot\nabla_x p(t,\mathbf{x},\mathbf{v}) - \beta\nabla_v\cdot[(\mathbf{F}(C(t,\mathbf{x})) - \mathbf{v})p(t,\mathbf{x},\mathbf{v})] + \frac{\beta}{2}\Delta_v p(t,\mathbf{x},\mathbf{v}), \tag{11}$$

$$\tilde{p}(t,\mathbf{x}) = \int p(t,\mathbf{x},\mathbf{v}')\, d\mathbf{v}'. \tag{12}$$

The two first terms on the right hand side of (11) correspond to vessel tip branching – from Eqs. (2) and (3) – and anastomosis, respectively. While the branching term follows from (2) and (3) in a straightforward manner, deducing the anastomosis integral term is the real breakthrough from past work achieved in [9]. The anastomosis coefficient, Γ, has to be fitted by comparison of the numerical solution of the deterministic equations and ensemble averages of the stochastic description, [10]. The other terms on the right hand side of (11) are in the Fokker-Planck equation that corresponds to the Langevin equation (1) in the usual manner [22]. While the branching term follows directly from the

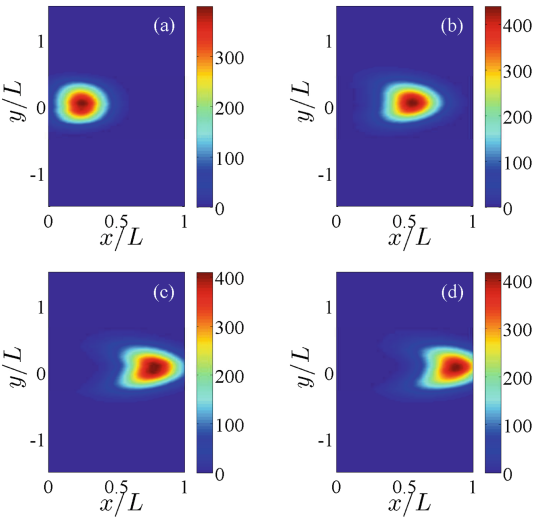


Fig. 2. Marginal density of active vessel tips resulting from an average over 400 replicas of the stochastic process according to Eq. (9) at four different times: (a) 12 h, (b) 24 h, (c) 32 h, and (d) 36 h. At these times, the numbers of active tips are (a) 56, (b) 69, (c) 72, and (d) 66, [10].

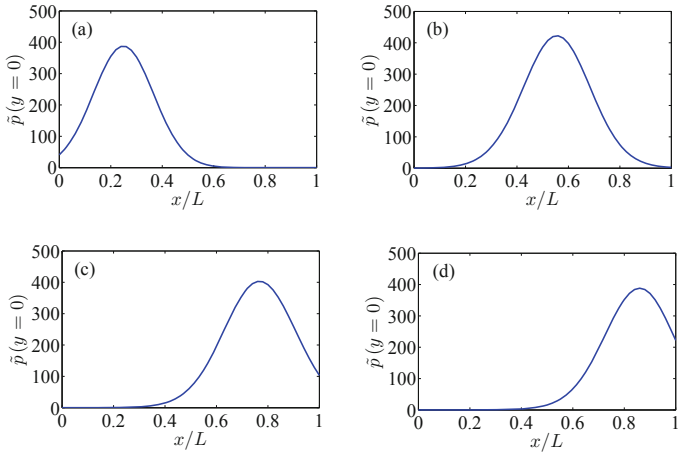


Fig. 3. Marginal density of active vessel tips at the x axis resulting from an average over 400 replicas of the stochastic process as in Fig. 2. The primary vessel at $x = 0$ issues a pulse that marches toward the hypoxic region at $x = 1$, [10].

stochastic branching process, anastomosis occurs when a moving vessel tip at time $t > 0$ encounters a preexisting vessel whose tip was at the same place at an earlier time $s < t$. At time t, a moving vessel tip can reach an area $d\mathbf{x}$ about $\mathbf{x}$ that is either unoccupied or occupied by another vessel. In the latter case, it anastomoses. The occupation time density of the area $d\mathbf{x}$ about $\mathbf{x}$ is proportional to $\int_0^t \tilde{p}(s,\mathbf{x})\,ds$ - the ensemble average of the vessel network density (7). Then the rate of anastomosis should be proportional to $p(t,\mathbf{x},\mathbf{v})$ times this occupation time density [10]. Equation (5) becomes

$$\frac{\partial C}{\partial t}(t,\mathbf{x}) = \kappa_c \Delta_x C(t,\mathbf{x}) - \chi_c C(t,\mathbf{x})\,|\mathbf{j}(t,\mathbf{x})|, \tag{13}$$

where $\mathbf{j}(t,\mathbf{x})$ is the current density (flux) vector at any point $\mathbf{x}$ and any time $t \geq 0$,

$$\mathbf{j}(t,\mathbf{x}) = \int \mathbf{v}' p(t,\mathbf{x},\mathbf{v}')\,d\mathbf{v}'. \tag{14}$$

Equation (6) becomes (13) in which $\int |\mathbf{v}'|\,p(t,\mathbf{x},\mathbf{v}')\,d\mathbf{v}'$ replaces $|\mathbf{j}(t,\mathbf{x})|$.

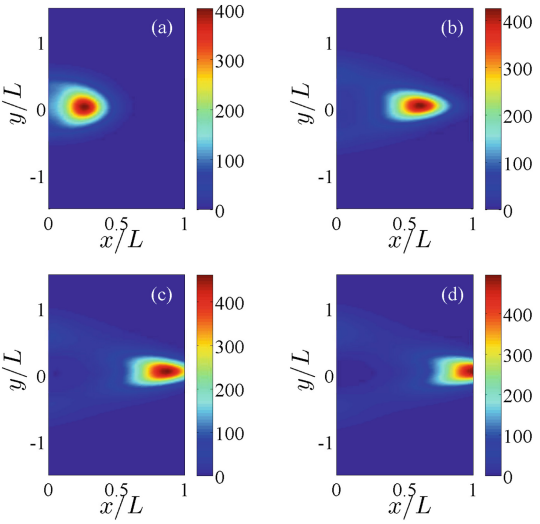


Fig. 4. Marginal density of active vessel tips resulting from a numerical simulation of the deterministic equations with appropriate boundary conditions for the same times as in Fig. 2 [9,10]. Better agreement between both descriptions requires fine tuning of the boundary conditions.

Figure 4 shows that the outcome of a numerical simulation of the deterministic description is similar to that of the stochastic process.

Carpio and collaborators have shown that the deterministic system of Eqs. (11)–(13) together with appropriate boundary and initial conditions has a unique solution that depends smoothly on parameters [23,24]. The proof that the deterministic description (11) follows from ensemble averages of the stochastic process

as described here is an important *open problem*. Anastomosis is a random event that depends on the past history of each realization of the stochastic process. Killing process with memory of this type have been studied formally before. However, the densities (8)–(10) are ensemble averages over infinitely many different realizations, which are, by definition, independent from each other. This could be important in a mathematical investigation of these processes.

In a recent paper [25], Capasso and Flandoli have proved an important convergence result for the deterministic description. They consider an appropriately modified stochastic process for d-dimensional angiogenesis on the whole space that also includes secondary branching at random points of existing capillaries. In the limit as the initial number of tips N_0 tends to infinity, they prove that a relative tip density (scaled with the initial number of tips) converges in probability. The limiting relative tip density satisfies in a weak sense a deterministic integro partial differential equation. In this equation, integrals over time also appear at the source term due to secondary angiogenesis. As explained before, the memory source term due to secondary angiogenesis is likely to be small compared to the local source term considered in (11). Capasso and Flandoli also prove that the number of tips at any given time $t \in [0, T]$ is bounded by a factor $e^{\lambda T} N_0$, with $\lambda > 0$. It would be interesting to see whether the limit as $N_0 \to \infty$ can be replaced by ensemble averages at least in the 2D case. Similarly, comparison of numerical solutions of the deterministic description on an appropriate geometry and averages of the stochastic process would help understanding the implications of the rigorous results in [25].

4 Soliton and Collective Coordinates

In the overdamped limit of negligible inertia in (1), we obtain the simpler Langevin-Ito equation: $d\mathbf{X}^i(t) \approx \mathbf{F}(C(t, \mathbf{X}^i(t)))\, dt + \beta^{-1/2} d\mathbf{W}^i(t)$ [11]. By using the Chapman-Enskog perturbation method whose details are explained in [12], it is then possible to derive the following reduced equation for the marginal tip density,

$$\frac{\partial \tilde{p}}{\partial t} + \nabla_x \cdot (\mathbf{F}\tilde{p}) - \frac{1}{2\beta} \Delta_x \tilde{p} = \mu\, \tilde{p} - \Gamma \tilde{p} \int_0^t \tilde{p}(s, \mathbf{x})\, ds, \tag{15}$$

$$\mu = \frac{\alpha}{\pi} \left[1 + \frac{\alpha}{2\pi\beta(1 + \sigma_v^2)} \ln\left(1 + \frac{1}{\sigma_v^2} \right) \right]. \tag{16}$$

The drift terms in Eq. (15) are those corresponding to the simpler Langevin-Ito equation for $\mathbf{X}^i(t)$ that results in the overdamped limit. The birth and death terms are obtained by integration of the corresponding ones on right hand side of (11) over velocity. However, the perturbation procedure changes the coefficient $\alpha(C)$ to the related function $\mu(C)$ in (16) [12]. Equation (15) has the following soliton-like solution for constant $\mathbf{F} = (F_x, F_y)$, μ, and zero diffusion, $1/\beta = 0$:

$$\tilde{p}_s = \frac{(2K\Gamma + \mu^2)c}{2\Gamma(c - F_x)} \operatorname{sech}^2 \left[\frac{\sqrt{2K\Gamma + \mu^2}}{2(c - F_x)} (x - X(t)) \right], \quad \dot{X} \equiv \frac{dX}{dt} = c, \tag{17}$$

where K is a constant. In fact [11], consider $\tilde{p}_s = \partial P(x-ct)/\partial t = -c\,P'(\xi)$, $\xi = x - ct$, which, inserted in (15) with $1/\beta = 0$, yields

$$(F_x - c)P'' = \mu P' - \Gamma P P' \Longrightarrow (c - F_x)P' = \frac{\Gamma}{2}P^2 - K - \mu P.$$

Setting $P = \nu\tanh(\lambda\xi)+\mu/\Gamma$, we find $\nu^2 = (\mu^2+2K\Gamma)/\Gamma^2$ and $2\nu\lambda(c-F_x)/\Gamma = -\nu^2$, thereby obtaining

$$P = \frac{\mu}{\Gamma} - \frac{\sqrt{2K\Gamma+\mu^2}}{\Gamma}\tanh\left[\frac{\sqrt{2K\Gamma+\mu^2}}{2(c-F_x)}(\xi-\xi_0)\right].$$

Here ξ_0 is a constant of integration. Thus $\tilde{p}_s = \partial P/\partial t = -cP'$ is given by (17).

Note that the source terms (branching and anastomosis) in Eqs. (11) and (16) are crucial for the soliton solution (17) to exist. Their absence in all developments previous to [9] explains that they could not go beyond numerical simulations of the stochastic process.

Numerical simulations on a slab geometry show that the marginal tip density evolves toward (17) after an initial stage [11,12]. *It is an open problem to prove this stability result even for a one-dimensional version of Eq. (15) on the whole real line and having constant values of* $\mathbf{F}$ *and* μ.

A small diffusion and slowly varying continuum field C produce a moving soliton whose shape and speed are slowly changing. We can find them by deducing evolution equations for the *collective coordinates* K, c, and X [11,12]. Then the marginal density profile at $y = 0$ can be reconstructed from (17) with spatially averaged F_x and μ [12]. Note that $\tilde{p}_s$ is a function of $\xi = x - X$ and also of $\mathbf{x}$ and t through $C(t,\mathbf{x})$,

$$\tilde{p}_s = \tilde{p}_s\left(\xi; K, c, \mu(C), F_x\left(C, \frac{\partial C}{\partial x}\right)\right). \tag{18}$$

We assume that the time and space variations of C, which appear when $\tilde{p}_s$ is differentiated with respect to t or x, produce terms that are small compared to $\partial\tilde{p}_s/\partial\xi$. As explained in [12], we shall consider that $\mu(C)$ is approximately constant, ignore $\partial C/\partial t$ because the VEGF concentration varies slowly (the dimensionless coefficients κ_c and χ_c appearing in the VEGF equation (13) are very small according to Table 2 of [12]) and ignore $\partial^2\tilde{p}_s/\partial i\partial j$, where $i, j = K$, F_x. We now insert (17) into (15), thereby obtaining

$$\left(F_x - \dot{X}\right)\frac{\partial\tilde{p}_s}{\partial\xi} + \frac{\partial\tilde{p}_s}{\partial K}\dot{K} + \frac{\partial\tilde{p}_s}{\partial c}\dot{c} - \frac{1}{2\beta}\left(\frac{\partial^2\tilde{p}_s}{\partial\xi^2} + 2\frac{\partial^2\tilde{p}_s}{\partial\xi\partial F_x}\frac{\partial F_x}{\partial x} + \frac{\partial\tilde{p}_s}{\partial F_x}\Delta_x F_x\right)$$
$$+\tilde{p}_s\nabla_x\cdot\mathbf{F} + \frac{\partial\tilde{p}_s}{\partial F_x}\left(\frac{\partial F_x}{\partial t} + \mathbf{F}\cdot\nabla_x F_x\right) = \mu\tilde{p}_s - \Gamma\tilde{p}_s\int_0^t \tilde{p}_s dt. \tag{19}$$

Equation (15) with $1/\beta = 0$ and constant $\mathbf{F}$ and μ has the soliton solution (17). Using this fact, we can eliminate the first term on the left hand side of (19) and

also the right hand side thereof. Equation (19) then becomes

$$\frac{\partial \tilde{p}_s}{\partial K}\dot{K} + \frac{\partial \tilde{p}_s}{\partial c}\dot{c} = \mathscr{A}, \tag{20}$$

$$\mathscr{A} = \frac{1}{2\beta}\frac{\partial^2 \tilde{p}_s}{\partial \xi^2} - \tilde{p}_s \nabla_x \cdot \mathbf{F} - \frac{\partial \tilde{p}_s}{\partial F_x}\left(\mathbf{F}\cdot\nabla_x F_x - \frac{1}{2\beta}\Delta_x F_x\right) + \frac{1}{\beta}\frac{\partial^2 \tilde{p}_s}{\partial \xi \partial F_x}\frac{\partial F_x}{\partial x}. \tag{21}$$

We now find collective coordinate equations (CCEs) for K and c. As the lump-like angiton moves on the x axis, we set $y = 0$ to capture the location of its maximum. On the x axis, the profile of the angiton is the soliton (17). We first multiply (20) by $\partial \tilde{p}_s/\partial K$ and integrate over x. We consider a fully formed soliton far from primary vessel and hypoxic region. As it decays exponentially for $|\xi| \gg 1$, the soliton is considered to be localized on some finite interval $(-\mathscr{L}/2, \mathscr{L}/2)$. The coefficients in the soliton formula (17) and the coefficients in (21) depend on the VEGF concentration at $y = 0$, therefore they are functions of x and time and get integrated over x. The VEGF concentration varies slowly on the support of the soliton, and therefore we can approximate the integrals over x by [12]

$$\int_{\mathscr{I}} F(\tilde{p}_s(\xi; x, t), x)dx \approx \frac{1}{\mathscr{L}}\int_{\mathscr{I}}\left(\int_{-\mathscr{L}/2}^{\mathscr{L}/2} F(\tilde{p}_s(\xi; x, t), x)d\xi\right)dx. \tag{22}$$

The interval $\mathscr{I}$ over which we integrate should be large enough to contain most of the soliton, of extension $\mathscr{L}$. Thus the CCEs hold only after the initial soliton formation stage. Near the primary vessel and near the hypoxic region, the boundary conditions affect the soliton and we should exclude intervals near them from $\mathscr{I}$. We shall specify the integration interval $\mathscr{I}$ below. Acting similarly, we multiply (20) by $\partial \tilde{p}_s/\partial c$ and integrate over x. From the two resulting formulas, we then find $\dot{K}$ and $\dot{c}$ as fractions. The factors $1/\mathscr{L}$ cancel out from their numerators and denominators. As the soliton tails decay exponentially to zero, we can set $\mathscr{L} \to \infty$ and obtain the following CCEs [12]

$$\dot{K} = \frac{\int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial K}\mathscr{A}d\xi \int_{-\infty}^{\infty}\left(\frac{\partial \tilde{p}_s}{\partial c}\right)^2 d\xi - \int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial c}\mathscr{A}d\xi \int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial K}\frac{\partial \tilde{p}_s}{\partial c}d\xi}{\int_{-\infty}^{\infty}\left(\frac{\partial \tilde{p}_s}{\partial K}\right)^2 d\xi \int_{-\infty}^{\infty}\left(\frac{\partial \tilde{p}_s}{\partial c}\right)^2 d\xi - \left(\int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial c}\frac{\partial \tilde{p}_s}{\partial K}d\xi\right)^2}, \tag{23}$$

$$\dot{c} = \frac{\int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial c}\mathscr{A}d\xi \int_{-\infty}^{\infty}\left(\frac{\partial \tilde{p}_s}{\partial K}\right)^2 d\xi - \int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial K}\mathscr{A}d\xi \int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial K}\frac{\partial \tilde{p}_s}{\partial c}d\xi}{\int_{-\infty}^{\infty}\left(\frac{\partial \tilde{p}_s}{\partial K}\right)^2 d\xi \int_{-\infty}^{\infty}\left(\frac{\partial \tilde{p}_s}{\partial c}\right)^2 d\xi - \left(\int_{-\infty}^{\infty}\frac{\partial \tilde{p}_s}{\partial c}\frac{\partial \tilde{p}_s}{\partial K}d\xi\right)^2}. \tag{24}$$

In these equations, all terms varying slowly in space have been averaged over the interval $\mathscr{I}$. The last term in (21) is odd in ξ and does not contribute to the integrals in (23) and (24) whereas all other terms in (21) are even in ξ and do contribute. The integrals appearing in (23) and (24) are calculated in [12]. The resulting CCEs are

$$\dot{K} = \frac{(2K\Gamma+\overline{\mu}^2)^2}{4\Gamma\beta(c-\overline{F_x})^2}\,\frac{\frac{4\pi^2}{75}+\frac{1}{5}+\left(\frac{2\overline{F_x}}{5c}-\frac{2\pi^2}{75}-\frac{9}{10}\right)\frac{\overline{F_x}}{c}}{\left(1-\frac{4\pi^2}{15}\right)\left(1-\frac{\overline{F_x}}{2c}\right)^2}$$
$$-\frac{2K\Gamma+\overline{\mu}^2}{\Gamma c\left(2-\frac{\overline{F_x}}{c}\right)}\left(c\overline{\nabla_x\cdot\mathbf{F}}+\overline{\mathbf{F}\cdot\nabla_x F_x}-\frac{\overline{\Delta_x F_x}}{2\beta}\right), \tag{25}$$

$$\dot{c} = -\frac{7(2K\Gamma+\overline{\mu}^2)}{20\beta(c-\overline{F_x})}\,\frac{1-\frac{4\pi^2}{105}}{\left(1-\frac{4\pi^2}{15}\right)\left(1-\frac{\overline{F_x}}{2c}\right)}+\frac{\overline{\mathbf{F}\cdot\nabla_x F_x}-(c-\overline{F_x})\overline{\nabla_x\cdot\mathbf{F}}-\frac{\overline{\Delta_x F_x}}{2\beta}}{2-\frac{\overline{F_x}}{c}}, \tag{26}$$

$$\overline{g(x,y)} = \frac{1}{\mathscr{I}}\int_{\mathscr{I}} g(x,0)\,dx, \tag{27}$$

in which the functions of $C(t,x,y)$ have been averaged over the interval $\mathscr{I}$ after setting $y=0$. We expect the CCEs (25)–(26) to describe the mean behavior of the soliton whenever it is far from primary vessel and hypoxic region.

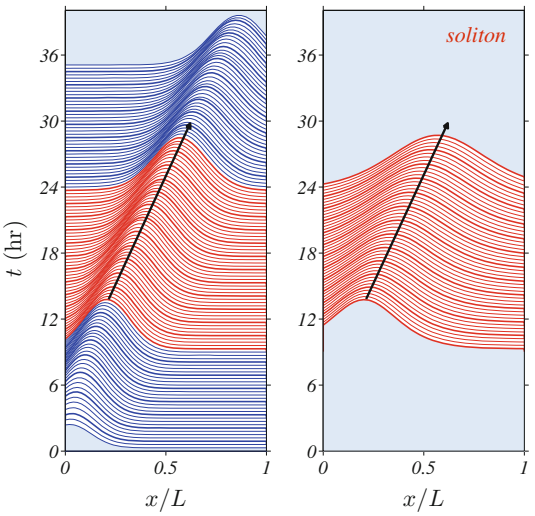


Fig. 5. Comparison of the marginal tip density profile $\tilde{p}(t,x,0)$ (obtained from the stochastic description averaged over 400 replicas) to that of the moving soliton, [11].

Both deterministic or stochastic simulations show that the soliton is formed after some time $t_0 = 0.2$ (10 h) following angiogenesis initiation. To find the soliton evolution afterwards, we need to solve the CCEs (25)–(26), in which the spatial averages depend on an interval $x \in \mathscr{I}$, which should exclude regions affected by boundaries. We calculate the spatially averaged coefficients in (25)–(26) by: (i) approximating all differentials by second order finite differences, (ii) setting $y = 0$, and (iii) averaging the coefficients from $x = 0$ to 0.6 by taking the arithmetic mean of their values at all grid points in the interval

$\mathscr{I} = (0, 0.6]$. For $x > 0.6$, the boundary condition at $x = 1$ influences the outcome and therefore we leave values for $x > 0.6$ out of the averaging [12]. The initial conditions for the CCEs are set as follows. $X(t_0) = X_0$ is the location of the marginal tip density maximum, $\tilde{p}(t_0, x = X_0, 0)$. We find $X_0 = 0.2$ from the stochastic description. We set $c(t_0) = c_0 = X_0/t_0$. $K(t_0) = K_0$ is determined so that the maximum marginal tip density at $t = t_0$ coincides with the soliton peak. This yields $K_0 = 39$. Solving the CCEs (25)–(26) with these initial conditions and using (17), we obtain the curves depicted in Fig. 5.

5 Random Walk Tip Cell Models

These models describe the extension of blood vessels by random walks biased by chemotaxis or haptotaxis instead of using Langevin equations. The first such model, due to Anderson and Chaplain [13], is based on a reaction-diffusion description of angiogenesis. They consider a continuity equation for the density of endothelial cells (ECs) n (with zero-flux boundary conditions) coupled to equations for the VEGF and fibronectin densities, C and f, respectively. In nondimensional form, these equations are [13]:

$$\frac{\partial n}{\partial t} = D\Delta n - \nabla\cdot\left(\frac{\chi}{1+\alpha C} n\nabla C\right) - \nabla\cdot(\rho n\nabla f), \tag{28}$$

$$\frac{\partial f}{\partial t} = \beta n - \gamma n f, \tag{29}$$

$$\frac{\partial C}{\partial t} = -\eta n C. \tag{30}$$

Here all parameters are positive. The three terms on the right hand side of (28) correspond to diffusion of ECs, chemotaxis and haptotaxis, respectively. Note that chemotaxis has the same form in this equation as in (11) with p replaced by n. Haptotaxis follows the gradient of fibronectin in the extracellular matrix. Note that proliferation and death of ECs are not contemplated by (28). In the next step, these equations are solved by an explicit Euler method in time and finite differences. The resulting equation for $n(t, x, y) \approx n^q_{l,m}$,

$$n^{q+1}_{l,m} = n^q_{l,m} W_0 + n^q_{l+1,m} W_1 + n^q_{l-1,m} W_2 + n^q_{l,m+1} W_3 + n^q_{l,m-1} W_4, \tag{31}$$

has the same form as a master equation for a random walk [22], except that the "transition probabilities" W_0 (staying), W_1 (moving to the left), W_2 (moving to the right), W_3 (moving downwards), and W_4 (moving upwards) are not normalized. However, this is easily fixed by defining

$$\mathscr{W}_i = \frac{W_i}{\sum_{j=0}^{4} W_j}, \quad i = 0, 1, \ldots, 4, \tag{32}$$

as new transition probabilities. The random walk associated to these transition probabilities represents extension of vessel tips and replaces the Langevin equation (1). Branching and anastomosis are introduced as in the Langevin tip cell

model, except that the tips have to wait some *maturity* time after branching before they are allowed to branch again. It should be straightforward to find equations for the density of active vessel tips by using the theory described in previous sections.

The Anderson-Chaplain idea is easy to implement starting from continuum models of angiogenesis (and therefore it can be immediately generalized by including more taxis mechanisms, influence of antiangiogenic factors [26], etc.), but it has the drawback of having to rely on the finite difference grid or lattice. Another drawback is that the transition probabilities extracted from a finite difference code may not always be non-negative. A few years later, Plank and Sleeman fixed both these drawbacks. They proposed non-lattice models independent of the grid [14] using biased circular random walk models previously introduced by Hill and Häder for swimming microorganisms [27]. If $\theta(t)$ is a continuous random walk on the unit circle biased by chemo and haptotaxis [14], the trajectory of the corresponding tip cell is

$$\frac{d\mathbf{x}}{dt} = v_0\,(\cos\theta(t), \sin\theta(t)). \tag{33}$$

Thus the tip cells have the same speed v_0, directions given by $\theta(t)$ and their trajectories do not have to follow points on a lattice. While branching and anastomosis are modeled as in Sect. 2, the extensions of vessel tips are described by (33) and the biased circular random walk instead of Langevin equations. The master equation for the circular random walk is [14]

$$\frac{dP_n}{dt} = \hat{\tau}^+_{n-1}P_{n-1} + \hat{\tau}^-_{n+1}P_{n+1} - (\hat{\tau}^+_n + \hat{\tau}^-_n)P_n, \tag{34}$$

$$\hat{\tau}^\pm_n = 2\lambda\frac{\tau\left(n\delta \pm \frac{\delta}{2}\right)}{\tau\left(n\delta + \frac{\delta}{2}\right) + \tau\left(n\delta - \frac{\delta}{2}\right)}. \tag{35}$$

As $\delta \to 0$ and $n \to \infty$ so that $n\delta = \theta$, the master equation (34) becomes the Fokker-Planck equation [14]

$$\frac{\partial P}{\partial t}(t,\theta) = D\frac{\partial}{\partial\theta}\left[P(t,\theta)\frac{\partial}{\partial\theta}\left(\ln\frac{P(t,\theta)}{\tau(\theta)}\right)\right], \tag{36}$$

with $D = \lambda\delta^2$ for $P(t,\theta) = P(t,n\delta) = P_n(t)$. Chemo and haptotaxis are included in the model through the transition probability

$$\tau(\theta) = \frac{\exp[d_C\cos(\theta-\theta_C) + d_f\cos(\theta-\theta_f)]}{\int_{-\pi}^{\pi}\exp[d_C\cos(s-\theta_C) + d_f\cos(s-\theta_f)]\,ds}, \tag{37}$$

$$\tan\theta_C = \frac{\nabla C}{|\nabla C|}, \quad \tan\theta_f = \frac{\nabla f}{|\nabla f|}. \tag{38}$$

Here $\tau(\theta)$ is the stationary probability density of the Fokker-Planck equation (36).

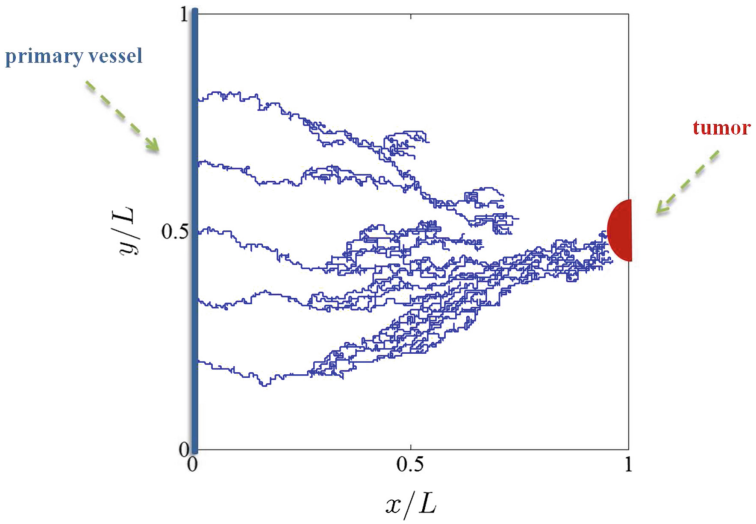


Fig. 6. Sketch of the geometry for angiogenesis from a primary blood vessel to a circular tumor calculated by using the Anderson-Chaplain model.

An extension of these ideas to 2D random walks produces a system with non-negative transition probabilities [14]. Instead of (34), we may write the 2D master equation

$$\frac{dP_{n,m}}{dt} = \hat{\tau}^{H+}_{n-1,m}P_{n-1,m} + \hat{\tau}^{H-}_{n+1,m}P_{n+1,m} + \hat{\tau}^{V+}_{n,m-1}P_{n,m-1} + \hat{\tau}^{V-}_{n,m+1}P_{n,m+1}$$
$$-(\hat{\tau}^{H+}_{n,m} + \hat{\tau}^{H-}_{n,m} + \hat{\tau}^{V+}_{n,m} + \hat{\tau}^{V-}_{n,m})P_{n,m}, \tag{39}$$

$$\hat{\tau}^{H\pm}_{n,m} = 4\lambda \frac{\tau(w_{n\pm\frac{1}{2},m})}{\tau(w_{n+\frac{1}{2},m}) + \tau(w_{n-\frac{1}{2},m}) + \tau(w_{n,m+\frac{1}{2}}) + \tau(w_{n,m-\frac{1}{2}})}, \tag{40}$$

$$\hat{\tau}^{V\pm}_{n,m} = 4\lambda \frac{\tau(w_{n,m\pm\frac{1}{2}})}{\tau(w_{n+\frac{1}{2},m}) + \tau(w_{n-\frac{1}{2},m}) + \tau(w_{n,m+\frac{1}{2}}) + \tau(w_{n,m-\frac{1}{2}})}. \tag{41}$$

Here $w = (C, f)$ and $\tau(w) = \tau_1(C)\tau_2(f)$, with

$$\tau_1(C) = (1 + \alpha C)^{\frac{\chi}{\alpha D}}, \quad \tau_2(f) = e^{\rho f/D}. \tag{42}$$

Clearly, these transition probabilities are positive and it can be proved that the master equation (39) has (28) as a continuum limit [14]. Active tips, branching and anastomosis are treated as in the Anderson-Chaplain paper [13]. Comparisons between numerical simulations of the Anderson-Chaplain and Plank-Sleeman models are carried out in [14]. Figure 6 shows one realization of the Anderson-Chaplain stochastic process that includes vessel extension, branching and anastomosis.

The random walk models of this Section get their input from continuum equations for ECs, VEGF and fibronectin densities, but the moving vessel tips characterized by the random walks do not affect the continuum fields. Their outcomes are numerical simulations of the stochastic processes, without further elaboration. In contrast to this somewhat artificial setting, the Langevin tip cell model of Sect. 2 is a hybrid model in which active vessel tips and continuum fields are fully coupled. Furthermore, we can derive an equivalent deterministic description from the Langevin tip cell model and analyze it in terms of a soliton-like attractor. This latter elaboration has also been carried out for a Langevin tip cell model that includes chemotaxis and haptotaxis [20]. Now, the master equation becomes a Fokker-Planck equation (corresponding to a Langevin-Ito equation) in the continuum limit [22]. Then we may expect that the master equation with two added source terms similar to those in Eq. (11) describes the stochastic process comprising random walk, branching and anastomosis. This seems to be the case [28].

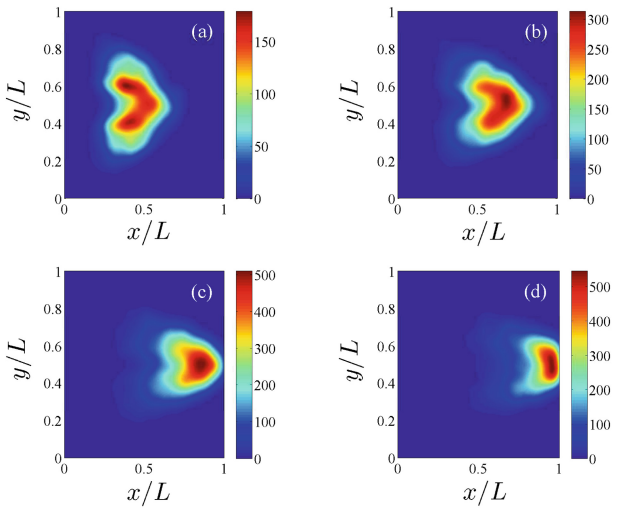


Fig. 7. Density of active vessel tips resulting from an average over 800 replicas of the stochastic process corresponding to reinforced random walk, branching and anastomosis with transition probabilities (40)–(41) at four different times: (a) 5 days, (b) 6 days, (c) 7 days, and (d) 8 days.

When we add source terms to the master equation (39), it becomes the following equation for the density of active vessel tips $\rho_{n,m}(t)$:

$$\frac{d\rho_{n,m}}{dt} = \hat{\tau}^{H+}_{n-1,m}\rho_{n-1,m} + \hat{\tau}^{H-}_{n+1,m}\rho_{n+1,m} + \hat{\tau}^{V+}_{n,m-1}\rho_{n,m-1} + \hat{\tau}^{V-}_{n,m+1}\rho_{n,m+1}$$
$$-(\hat{\tau}^{H+}_{n,m} + \hat{\tau}^{H-}_{n,m} + \hat{\tau}^{V+}_{n,m} + \hat{\tau}^{V-}_{n,m})\rho_{n,m} + \alpha_{n,m}\,\rho_{n,m} - \Gamma_{n,m}\,\rho_{n,m}\int_0^t \rho_{n,m}\,dt. \qquad (43)$$

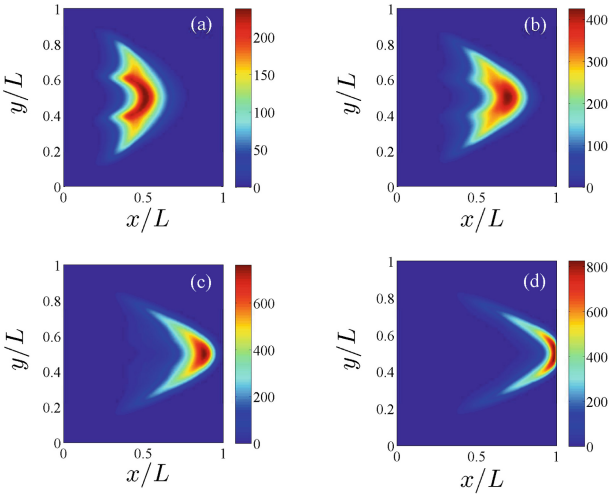


Fig. 8. Density of active vessel tips calculated from the master equation (43) at times: (a) 5 days, (b) 6 days, (c) 7 days, and (d) 8 days.

Figure 7 depicts the active vessel density (9) calculated from ensemble average over replicas of the stochastic process (reinforced random walk, branching and anastomosis) at four different times after angiogenesis starts. Figure 8 shows the solution of the master equation (43) at the same times as in Fig. 7. Both stochastic and deterministic descriptions produce similar results. In particular, the velocity of the patch where most active tips are concentrated is about the same in both descriptions. See [28] for details.

As in the case of the stochastic process including Langevin-Ito equations for vessel extension of Sect. 2, it is an important *open problem* to deduce the master equation (43) from a reinforced random walk process with added branching and anastomosis.

6 Cellular Potts Models

In all the previous models, the cells are treated as point particles. For a more precise view of haptotaxis, i.e., the motion of ECs over the extracellular matrix (ECM), we need to consider adhesion and deformation of the cells. This requires a more microscopic view than that offered by tip cell models or by more complicated models that distinguish between tip and stalk ECs and add extra dynamics for them [8].

Often times, ECs and ECM are modeled by a cellular Potts model (CPM) with Monte Carlo dynamics coupled to continuum fields (elastic fields, VEGF, ...) [29]. Space in these models consists of a lattice whose cells (lattice sites) may be in finitely many different states, denoted by type τ and representing ECs, matrix fibers, tissue cells and interstitial fluid. To account for individual entities

(ECs, fibers, etc), each entity is further associated with a unique identifying number, denoted by σ, that is assigned to every lattice site occupied by it. At every Monte Carlo time step, the cell surface (represented by connected lattice vertices) is updated according to a set of cell behavior rules (e.g., target cell shape and size) that are translated in an energy change. Typically, we select randomly a cell $\mathbf{x}$, assign its type, $\tau(\mathbf{x})$, to a randomly chosen neighbor $\mathbf{x}'$, and update accordingly the total energy of the system, H. Using the Metropolis algorithm, a given update is accepted with probability one if the change in the total energy of the system, ΔH, is reduced and it is accepted with probability $e^{-\beta\Delta H}$ otherwise ($1/\beta$ is the Monte Carlo temperature). The energy in [29] is

$$H = \sum_{\text{sites}} J_{\tau,\tau'}(1-\delta_{\sigma\sigma'}) + \sum_{\text{cells}} \gamma_\tau (a_\sigma - A_\sigma)^2 - \sum_{\text{cells}} \sum_{\text{sites}} \mu_\sigma C(t,\mathbf{x}). \quad (44)$$

The first term in Eq. (44) is the contribution to total energy resulting from cell-cell and cell-medium adhesion. The second term allows deformation of cells with volume a_σ about a target volume (area in 2D space) A_σ, depending on the Potts parameters γ_τ. The target volume is twice that of the initial volume and it corresponds to the volume at which a cell undergoes mitosis, thereby creating a new cell. Thus cell proliferation is contemplated in this CPM. A variation of the last term in (44) is

$$\Delta H_{\text{chem}} = -\mu_\sigma [C(t,\mathbf{x}) - C(t,\mathbf{x}')], \quad (45)$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two randomly picked neighboring lattice cells, $\mu_\sigma > 0$ is the chemical potential, and Eq. (45) represents chemotaxis favoring motion directed along the VEGF gradient. The VEGF concentration satisfies a reaction-diffusion equation [29]. The parameters appearing in the model are chosen in such a way that the progression of blood vessels occurs in the time scale observed in experiments [29].

Under this framework, each entity (ECs, ECM, ...) has a finite volume, a deformable shape and competes for space. ECs proliferate. Intercellular interactions occur only at the cells surface and have a cell-type-dependent surface (or adhesion) energy $J_{\tau,\tau'}$, which is a measure of the coupling strength between the entities τ and τ'. Other CPMs include an ECM strain-dependent term that favors cell extension in the direction of principal strain (durotaxis). The force exerted by the ECs on the ECM is calculated by finite elements [30]. In more complicated models, each cell contains agents that signal to other cells and adhesion is modeled by a CPM [31].

As in the case of random walk tip cell models, there is a connection between CPM and a deterministic formulation for a density. In [32], Alber *et al.* have written a discrete time master equation for the probability density $P(t,\mathbf{r},\mathbf{L})$ that a cell with its center of mass at $\mathbf{r}$ occupy a rectangle with sides $\mathbf{L} = (l_x, l_y)$ at time t. It is based on a CPM with energy given by (44), but with a target perimeter instead of the target area. The corresponding term in the energy is

$$H_{\text{perim}} = \sum_{\text{cells}} [\gamma_x (l_x - L_x)^2 + \gamma_y (l_y - L_y)^2]. \quad (46)$$

Here cells are always rectangles and do not proliferate nor die. Assuming that cells contain many lattice sites, they change little at each Monte Carlo step. Assuming further that cell-cell interactions are always binary, the authors derive a Fokker-Planck equation for $P(t, \mathbf{r}, \mathbf{L})$. These formulations would have to be extended to CPMs that include cell proliferation and be connected to mesoscopic angiogenesis models: from cell densities to densities of active vessel tip cells. It would be interesting to study whether the concept of active vessel tips and related ones can be used to derive deterministic descriptions in the spirit of Sects. 3 and 5.

7 Blood Flow and Vascular Network

Once a vascular network is being created, blood flows through the capillaries, anastomosis enhances flow in some of them and secondary angiogenesis may start in new vessels. Pries and coworkers have modeled blood flow in a vascular network and the response thereof to changing conditions such as pressure differences and wall stresses [33,34]. This response may remodel the vascular network by changing the radii of certain capillaries, and altering the distribution of blood flow [33,34]. McDougall, Anderson and Chaplain [35] have used this formulation to add secondary branching from new capillaries induced by wall shear stress to the original random walk tip cell model [13]. Blood flows according to Poiseuille's law, mass is conserved, there are empirical expressions for blood viscosity and for the wall shear stresses, and radii of capillaries adapt to local conditions. Secondary vessel branching may occur after the new vessel has reached a certain level of maturation and before a basal lamina has formed about it [21,35]. During such a time interval, the probability of secondary branching increases with both the local VEGF concentration and the magnitude of the shear stress affecting the vessel wall. McDougall *et al.*'s model can be used to figure out how drugs could be transported through the blood vessels and eventually reach a tumor [21,35]. In dense vessel networks, secondary branching may have little effect on the number of active tips at a given time, as anastomosis could eliminate secondary branches quickly. Thus we may ignore secondary branching when considering the density of active tips in such networks. Of course we cannot ignore it when describing blood flow and network remodeling.

One missing feature of angiogenesis models that take blood flow into account seems to be pruning. It is known that capillaries with insufficient blood circulation may atrophy and disappear. Pruning such blood vessels is an important mechanism to achieve a hierarchical vascular network such as that observed in retinal vascularization during development [3,4]. Global optimization and adaptation in developing networks has been recently shown to lead to highly optimized transport vascular systems [36,37]. It would be interesting to adapt these studies to angiogenesis.

8 Phase Field Models

Phase field models are continuum models able to represent vascular networks. For example, Travasso et al. [39,40] consider a reaction-diffusion equation for the VEGF $C(t,\mathbf{x})$ coupled to a continuum equation for the phase field $\phi(t,\mathbf{x})$:

$$\frac{\partial C}{\partial t}(t,\mathbf{x}) = \kappa_c \Delta_x C(t,\mathbf{x}) - \chi_c C(t,\mathbf{x})\,\phi(t,\mathbf{x})\,\Theta(\phi(t,\mathbf{x})), \tag{47}$$

$$\frac{\partial \phi}{\partial t}(t,\mathbf{x}) = M\Delta_x[-\phi(t,\mathbf{x}) + \phi^3(t,\mathbf{x}) - \varepsilon^2\Delta_x\phi(t,\mathbf{x})] + \alpha_\phi(C(t,\mathbf{x}))\,\phi(t,\mathbf{x})\,\Theta(\phi(t,\mathbf{x})). \tag{48}$$

Here M is the mobility coefficient for the endothelial cells, the proliferation rate is $\alpha_\phi(C) = \alpha_\phi[C\Theta(C_p - C) + C_p\Theta(C - C_p)]$, ε is the width of the capillary wall, and $\Theta(x)$ is the Heaviside unit step function. Proliferative and non-activated cells are described by an order parameter ϕ which is equal to -1 at the ECM outside the capillary and $+1$ inside it. Areas of high proliferation of endothelial cells have $\phi > 1$, which will lead to the widening of the capillary. The position of the capillary wall made out of stalk cells is given by the level set $\phi(t,\mathbf{x}) = 0$.

In addition to the continuum equations, there are discrete equations for activated tip endothelial cells and criteria to distinguish them. The angiogenic factor at the tip cell is only consumed at its surface receptors, therefore $\chi_C = 0$ is set in Eq. (47) at all points inside the tip cell. A tip cell moves chemotactically with velocity $\mathbf{v}$ (proportional to the VEGF gradient $\nabla_x C$ measured at the tip cell center, $\mathbf{x}_t(t)$):

$$\mathbf{v}(\mathbf{x}_t(t)) = \chi_v(|\nabla_x C(t,\mathbf{x_t})|)\,\nabla_x C(t,\mathbf{x}_t), \tag{49}$$

$$\chi_v(g) = \chi_v\left[\Theta(g - g_m) + \left(\frac{g_M}{g} - 1\right)\Theta(g - g_M)\right], \tag{50}$$

where χ_v is the chemotactic response of the endothelial cells (having radius R_c), g_M is the maximum VEGF gradient and $\chi_v g_M$ is the maximum tip speed. An activated cell moves only if $g_m < |\nabla_x C(t,\mathbf{x}_t)|$, with $0 < g_m < g_M$. When these conditions are met at the center of an endothelial stalk cell and $C > C_c$ there, it acquires the tip cell phenotype, with the caveat that cell-cell contact dependent mechanisms (the Notch pathway) prevent the activation of two neighboring cells. Only points for which there is a minimum distance of $4R_c$ to the centers of all already existing tip cells can become centers of activated tip cells. As in the biological system, when the chemotactic signal is small, $C < C_c$ or $|\nabla_x C(t,\mathbf{x}_t)| < g_m$, the endothelial cell returns to the stalk cell state. Simulations show that tip cell velocity and stalk cell proliferation play important roles in vascular network morphology [39]. An increase in stalk cell proliferation leads to a more branched network constituted by thicker vessels, while a higher tip cell migration velocity leads to a more branched network with thinner vessels [39].

More general phase field models incorporate force at the vessel tip and elasticity [38] and haptotaxis [41]. They are included in the review paper [42].

A study of the relation between morphology of the blood vessel network generated by phase field models, blood supply and obstructions can be found in [43]. Phase field models are thus a deterministic alternative to stochastic models.

9 Conclusions

Angiogenesis is a complex multiscale process by which diffusing vessel endothelial growth factors induce sprouting of blood vessels that carry oxygen and nutrients to hypoxic tissue. Cancerous tumor cells profit from this process to prosper, grow and eventually migrate to other organs. Mathematical models contemplate different aspects of angiogenesis. Here we have reviewed recent work on a simple tip cell model that encompasses vessel extension driven by chemotaxis and described by Langevin equations, stochastic tip branching and vessel fusion (anastomosis). From the stochastic description, we have derived a deterministic integropartial differential equation for the density of active tip cells coupled with a reaction-diffusion equation for the growth factor. The associated initial-boundary value problem is well posed. It is important to note that anastomosis prevents proliferation of active tips and therefore the deterministic description is based on ensemble averages over replicas of the stochastic process. Numerical simulations of both (deterministic and stochastic) descriptions show that the density of active tips adopts the shape of a two-dimensional soliton-like wave (angiton) after a formation stage. We have found an analytical formula for the one-dimensional projection of the soliton and ordinary differential equations for variables that provide its velocity, position and size. These equations also characterize the advance of the vessel network for single replicas. Much more work needs to be carried out to solve mathematical issues arising from our results, both from analysis of the deterministic description and from establishing more precise conditions for its validity. The description of the soliton should be extended to the true two-dimensional soliton (angiton) that appears in the numerical simulations and to the case of a more general geometry than that of the slab. Fluctuations cannot be ignored in the case of ensemble averages, and future work predicting the evolution of a real vessel network should include confidence bands about averages. Anti-angiogenic treatments need to be improved [1,2], and, in this respect, having better models and theories about their solutions should help. Therapies are related to optimal control of angiogenesis and they require accurate mathematical models, validated by comparison with real data (inverse problems - statistics of random geometric structures).

We have also related the specific model we study to other tip cells models in the literature that describe vessel extension by reinforced random walks instead of stochastic differential equations. Our methodology may be adapted to these other models as Langevin equations arise from reinforced random walks in appropriate limits. All these models describe mesoscales in which cells are just point particles, thereby ignoring their shapes and a microscopic description thereof. Other models consider the evolution of individual endothelial cells of variable shape and extension through cellular Potts models, but the continuation of these

models toward the mesoscale has barely begun. Extending the analysis carried out for our mesoscopic stochastic tip cell model to microscopic models is a challenge for the future. Blood circulation through the angiogenic network favors certain vessels, others that do not have enough perfusion shrink and disappear and secondary branching may occur. Future work could delve deeper in the topics of vessel remodeling, pruning, formation of optimal vascular networks and transport of medicals through them.

Apart from the specific application to angiogenesis, we have presented in this paper methodological contributions for a sound mathematical modeling of stochastic vessel networks: (a) the use of stochastic distributions, and their mean densities, describing the vessels, which are random objects of Hausdorff dimension one, cf (7); (b) reduction of vessel distributions to integrals over time of active tip distributions, which are random objects of zero Hausdorff dimension, cf (8)–(10); (c) characterization of the attractor of the density of active tips as a soliton whose position, velocity and size are given as solutions of ordinary differential equations, cf (17), (23)–(24). In our system, which is strongly out of equilibrium, this attractor plays a similar role to the stable stationary equilibrium distribution of many physical systems.

Acknowledgements. The authors thank V. Capasso and D. Morale from the Department of Mathematics of Università degli Studi di Milano, Milan, Italy, and B. Birnir from the Department of Mathematics of University of California at Santa Barbara, USA, for fruitful discussions and contributions. We also thank A. Lasanta from Universidad Carlos III de Madrid for useful comments on the manuscript. This work has been supported by the Ministerio de Economía y Competitividad grants MTM2014-56948-C2-2-P and MTM2017-84446-C2-2-R.

References

1. Carmeliet, P.F.: Angiogenesis in life, disease and medicine. Nature **438**, 932–936 (2005)
2. Carmeliet, P., Jain, R.K.: Molecular mechanisms and clinical applications of angiogenesis. Nature **473**, 298–307 (2011)
3. Gariano, R.F., Gardner, T.W.: Retinal angiogenesis in development and disease. Nature **438**, 960–966 (2005)
4. Fruttiger, M.: Development of the retinal vasculature. Angiogenesis **10**, 77–88 (2007)
5. Carmeliet, P., Tessier-Lavigne, M.: Common mechanisms of nerve and blood vessel wiring. Nature **436**, 193–200 (2005)
6. Folkman, J.: Tumor angiogenesis: therapeutic implications. N. Engl. J. Med. **285**(21), 1182–1186 (1971)
7. Folkman, J.: Angiogenesis. Annu. Rev. Med. **57**, 1–18 (2006)
8. Heck, T., Vaeyens, M.M., Van Oosterwyck, H.: Computational models of sprouting angiogenesis and cell migration: towards multiscale mechanochemical models of angiogenesis. Math. Model. Nat. Phen. **10**, 108–141 (2015)
9. Bonilla, L.L., Capasso, V., Alvaro, M., Carretero, M.: Hybrid modeling of tumor-induced angiogenesis. Phys. Rev. E **90**, 062716 (2014)

10. Terragni, F., Carretero, M., Capasso, V., Bonilla, L.L.: Stochastic model of tumor-induced angiogenesis: ensemble averages and deterministic equations. Phys. Rev. E **93**, 022413 (2016)
11. Bonilla, L.L., Carretero, M., Terragni, F., Birnir, B.: Soliton driven angiogenesis. Sci. Rep. **6**, 31296 (2016)
12. Bonilla, L.L., Carretero, M., Terragni, F.: Solitonlike attractor for blood vessel tip density in angiogenesis. Phys. Rev. E **94**, 062415 (2016)
13. Anderson, A.R.A., Chaplain, M.A.J.: Continuous and discrete mathematical models of tumor-induced angiogenesis. Bull. Math. Biol. **60**, 857–900 (1998)
14. Plank, M.J., Sleeman, B.D.: Lattice and non-lattice models of tumour angiogenesis. Bull. Math. Biol. **66**, 1785–1819 (2004)
15. Stokes, C.L., Lauffenburger, D.A.: Analysis of the roles of microvessel endothelial cell random motility and chemotaxis in angiogenesis. J. Theor. Biol. **152**, 377–403 (1991)
16. Stokes, C.L., Lauffenburger, D.A., Williams, S.K.: Migration of individual microvessel endothelial cells: stochastic model and parameter measurement. J. Cell Sci. **99**, 419–430 (1991)
17. Capasso, V., Morale, D.: Stochastic modelling of tumour-induced angiogenesis. J. Math. Biol. **58**, 219–233 (2009)
18. Sun, S., Wheeler, M.F., Obeyesekere, M., Patrick Jr., C.W.: Multiscale angiogenesis modeling using mixed finite element methods. Multiscale Mod. Simul. **4**(4), 1137–1167 (2005)
19. Bonilla, L.L., Capasso, V., Alvaro, M., Carretero, M., Terragni, F.: On the mathematical modelling of tumour induced driven angiogenesis. Math. Biosci. Eng. **14**, 45–66 (2017)
20. Bonilla, L.L., Carretero, M., Terragni, F.: Ensemble averages, soliton dynamics and influence of haptotaxis in a model of tumor-induced angiogenesis. Entropy **19**, 209 (2017)
21. Stéphanou, A., McDougall, S.R., Anderson, A.R.A., Chaplain, M.A.J.: Mathematical modelling of the influence of blood rheological properties upon adaptative tumour-induced angiogenesis. Math. Comput. Model. **44**, 96–123 (2006)
22. Gardiner, C.W.: Stochastic Methods. A Handbook for the Natural and Social Sciences, 4th edn. Springer, Berlin (2010)
23. Carpio, A., Duro, G.: Well posedness of an angiogenesis related integrodifferential diffusion model. Appl. Math. Model. **40**, 5560–5575 (2016)
24. Carpio, A., Duro, G., Negreanu, M.: Constructing solutions for a kinetic model of angiogenesis in annular domains. Appl. Math. Model. **45**, 303–322 (2017)
25. Capasso, V., Flandoli, F.: On the mean field approximation of a stochastic model of tumor-induced angiogenesis. Eur. J. Appl. Math. (2018). https://doi.org/10.1017/S0956792518000347
26. Levine, H.A., Pamuk, S., Sleeman, B.D., Nilsen-Hamilton, M.: Mathematical modeling of the capillary formation and development in tumor angiogenesis: penetration into the stroma. Bull. Math. Biol. **63**, 801–863 (2001)
27. Hill, N.A., Häder, D.P.: A biased random walk model for the trajectories of swimming micro-organisms. J. Theor. Biol. **186**, 503–526 (1997)
28. Bonilla, L.L., Carretero, M., Terragni, F.: Integrodifference master equation describing actively growing blood vessels in angiogenesis. Preprint (2019)
29. Bauer, A.L., Jackson, T.L., Jiang, Y.: A cell-based model exhibiting branching and anastomosis during tumor-induced angiogenesis. Biophys. J. **92**, 3105–3121 (2007)

30. Van Oers, R.F.M., Rens, E.G., La Valley, D.J., Reinhart-King, C.A., Merks, R.M.H.: Mechanical cell-matrix feedback explains pairwise and collective endothelial cell behavior in vitro. PLoS Comput. Biol. **10**(8), el003774 (2014)
31. Bentley, K., Franco, C.A., Philippides, A., Blanco, R., Dierkes, M., Gebala, V., Stanchi, F., Jones, M., Aspalter, I.M., Cagna, G., Weström, S., Claesson-Welsh, L., Vestweber, D., Gerhardt, H.: The role of differential VE-cadherin dynamics in cell rearrangement during angiogenesis. Nat. Cell Biol. **16**(4), 309–321 (2014)
32. Alber, N., Chen, N., Lushnikov, P.M., Newman, S.A.: Continuous macroscopic limit of a discrete stochastic model for interaction of living cells. Phys. Rev. Lett. **99**, 168102 (2007)
33. Pries, A.R., Secomb, T.W., Gaehtgens, P.: Structural adaptation and stability of microvascular netwoks: theory and simulation. Am. J. Physiol. Heart Circ. Physiol. **275**(44), H349–H360 (1998)
34. Pries, A.R., Secomb, T.W.: Control of blood vessel structure: insights from theoretical models. Am. J. Physiol. Heart Circ. Physiol. **288**(3), H1010–H1015 (2005)
35. McDougall, S.R., Anderson, A.R.A., Chaplain, M.A.J.: Mathematical modelling of dynamic adaptive tumour-induced angiogenesis: clinical implications and therapeutic targeting strategies. J. Theor. Biol. **241**, 564–589 (2006)
36. Ronellenfitsch, H., Katifori, E.: Global optimization, local adaptation, and the role of growth in distribution networks. Phys. Rev. Lett. **117**, 138301 (2016)
37. Ronellenfitsch, H., Lasser, J., Daly, D.C., Katifori, E.: Topological phenotypes constitute a new dimension in the phenotypic space of leaf venation networks. PLOS Comput. Biol. **11**(12), e1004680 (2016)
38. Santos-Oliveira, P., Correia, A., Rodrigues, T., Ribeiro-Rodrigues, T.M., Matafome, P., Rodríguez-Manzaneque, J.C., Seią, R., Girão, H., Travasso, R.D.M.: The force at the tip - modelling tension and proliferation in sprouting angiogenesis. PLoS Comput. Biol. **11**(8), e1004436 (2015)
39. Travasso, R.D.M., Corvera Poiré, E., Castro, M., Rodríguez-Manzaneque, J.C., Hernández-Machado, A.: Tumor angiogenesis and vascular patterning: a mathematical model. PLoS ONE **6**(5), e19989 (2011)
40. Travasso, R.D.M., Castro, M., Oliveira, J.C.R.E.: The phase-field model in tumor growth. Phil. Mag. **91**(1), 183–206 (2011)
41. Vilanova, G., Colominas, I., Gomez, H.: Coupling of discrete random walks and continuous modeling for three-dimensional tumor-induced angiogenesis. Comput. Mech. **53**, 449–464 (2014)
42. Vilanova, G., Colominas, I., Gomez, H.: Computational modeling of tumor-induced angiogenesis. Arch. Computat. Methods Eng. **24**, 1071–1102 (2017)
43. Torres-Rojas, A., Meza Romero, A., Pagonabarraga, I., Travasso, R.D.M., Corvera Poiré, E.: Obstructions in vascular networks. critical vs non-critical topological sites for blood supply. PLoS ONE **10**, e0128111 (2015)

Survival Under High Mutation

Rinaldo B. Schinazi(✉)

University of Colorado, Colorado Springs, CO 80918, USA
rinaldo.schinazi@uccs.edu

Abstract. We consider a stochastic model for an evolving population. We show that in the presence of genotype extinctions the population dies out for a low mutation probability but may survive for a high mutation probability. This turns upside down the widely held belief that above a certain mutation threshold a population cannot survive.

Keywords: Stochastic model · Evolution · Mutation · Random environment

1 A Model with Genotype Extinctions

There seems to be a consensus in theoretical biology that mutations are helpful for the survival of a population but that too many mutations are not, see [1–3]. We propose to use a stochastic model to challenge this belief. In fact, we will show that there are situations where the population survives for large mutation probability but dies out for small mutation probability! That is, we propose to turn upside down the idea that too many mutations are necessarily bad for survival.

We now describe our model. Let μ be a fixed continuous probability distribution with support contained in $[0, \infty)$ and let $r \in [0, 1]$. Start with one individual at time 0, and sample a birth rate λ from the distribution μ. Individuals give birth at rate λ and die at rate 1. Every time there is a birth there are two possibilities,

(i) with probability $1 - r$ the new individual keeps the same birth rate λ as its parent
or
(ii) with probability r the new individual is given a new birth rate λ', sampled independently of everything else from the distribution μ.

Furthermore, every time a new genotype appears (i.e. a new λ) we associate the genotype to a time T (independently of everything else) sampled from a fixed distribution ν. At time T all the individuals of this genotype are killed and the genotype disappears from the population.

We think of r as the mutation probability and the birth rate of an individual as representing the genotype of the individual. Since μ is assumed to be

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 437–442, 2019.
https://doi.org/10.1007/978-3-030-15096-9_14

continuous, a genotype cannot appear more than once in the evolution of the population. Genotype extinctions happen when all the individuals with the same genotype die. These so called background extinctions have been going on since the beginning of life, see for instance [4].

We say that the population survives if there is a strictly positive probability of having at all times at least one individual alive. Note that no genotype can survive forever so the population may only survive if it generates infinitely many genotypes.

Our first result is a necessary and sufficient condition for survival. We start the population with a single individual.

Theorem 1. *The population survives if and only if*

$$m(r) = rE[\Lambda \int_0^T \exp\left((\Lambda(1-r)-1)\, s\right) ds] > 1,$$

where Λ has distribution μ and T has distribution ν.

We now use Theorem 1 to compute two limits.

Corollary 1. *Assume that*

$$E[\Lambda \int_0^T \exp\left((\Lambda - 1)s\right) ds] < +\infty. \tag{1}$$

Then,

$$\lim_{r \to 0^+} m(r) = 0,$$

$$\lim_{r \to 1^-} m(r) = E[\Lambda(1 - e^{-T})].$$

Note that hypothesis (1) holds true if for instance Λ and T have bounded support.

Proof. For r in [0, 1] we have

$$\exp\big((\Lambda(1-r)-1)s\big) \leq \exp\big((\Lambda - 1)s\big).$$

For fixed Λ and T the r.h.s. is Lebesgue integrable on $[0, T]$. Hence, by the Dominated Convergence Theorem

$$\lim_{r \to 0^+} \int_0^T \exp\big((\Lambda(1-r)-1)s\big) ds = \int_0^T \exp\left((\Lambda - 1)s\right) ds$$

and

$$\lim_{r \to 1^-} \int_0^T \exp\big((\Lambda(1-r)-1)s\big) ds = \int_0^T \exp(-s) ds = 1 - \exp(-T).$$

Observe now that

$$\Lambda \int_0^T \exp\big((\Lambda(1-r)-1)s\big) ds \leq \Lambda \int_0^T \exp\big((\Lambda - 1)s\big) ds.$$

By (1) the r.h.s. is integrable. Hence, the Dominated Convergence Theorem applies. We may interchange the limits in r and the expectation. This yields the two limits and completes the proof of Corollary 1.

We have the following consequences of Corollary 1.

- Under hypothesis (1) there exists r_c in $(0,1)$ such that if $r < r_c$ then $m(r) < 1$. Hence, by Theorem 1 survival is not possible for r small.
- Assume hypothesis (1) holds and also that

$$E[\Lambda(1 - e^{-T})] > 1. \tag{2}$$

By Corollary 1, there exists r_c' in $(0,1)$ such that if $r > r_c'$ then $m(r) > 1$. Hence, survival is possible for $r > r_c'$.

Remark 1. Under hypotheses (1) and (2) survival is not possible for a low mutation probability but is possible for a high mutation probability. This goes exactly opposite to what is widely believed in theoretical biology. This belief has important practical consequences. One of the strategies to fight HIV has been to develop drugs that increase the mutation rate of the virus, see [3]. This can be futile or even counter productive if the virus can survive with increased mutation rate.

Remark 2. Hypotheses (1) and (2) hold for a wide range of distributions. If for instance Λ and T are uniformly distributed on $(0,a)$ and $(0,b)$, respectively, then (1) is true and (2) holds if and only if

$$\frac{a}{2}[1 - \frac{1}{b}(1 - e^{-b})] > 1.$$

In particular, for any $a > 2$ we can find b large enough so that the inequality above holds.

Remark 3. If hypothesis (1) fails survival may be possible for r small enough. Assume for instance that Λ is the constant λ and T is exponentially distributed with rate δ. Moreover, assume that $\lambda > 1 + \delta$. Then, the expected value in (1) is infinite. Furthermore, for any

$$0 < r < 1 - \frac{\delta + 1}{\lambda}$$

we have $m(r) = +\infty$. In particular, survival is possible for r small enough.

2 No Genotype Extinctions

In this section we consider the model without genotype extinctions, everything else remains the same. That is, the population starts with a single individual, a birth rate is sampled from a fixed distribution μ and the death rate is 1. For every new individual the same birth rate is kept with probability $1 - r$ or a new birth rate is sampled with probability r.

With no genotype extinction a genotype can survive forever. In fact, if the birth rate is λ for a particular genotype it will survive forever with positive probability if and only if $\lambda(1-r) > 1$. This is so because the number of individuals with a fixed genotype is a birth and death process with birth rate $\lambda(1-r)$ and death rate 1. We will use this fact in the following result.

Theorem 2. *Consider a population with no genotype extinction. If the population has a positive probability of surviving for some probability mutation $r > 0$ then there exists r_c in $(0, 1]$ such that the population survives for all $r < r_c$.*

Comparing Theorem 2 to the results of Sect. 1 we see that genotype extinctions change the behavior of the model in a drastic way.

Proof. Note first that if

$$\mu\{\lambda : \lambda \leq 1\} = 1$$

then the population dies out for all r in [0, 1]. This is so because we can couple our population to a birth and death chain with constant birth rate equal to 1 and death rate equal to 1. Since all the birth rates we sample for the population are below 1 it has less individuals than the birth and death chain with constant rates. Since the birth and death chain is critical it dies out and so does the population.

Assume now that our population has a positive probability of surviving for some $r > 0$. Then we must have

$$\mu(\{\lambda : \lambda > 1\}) > 0$$

and therefore for some $s > 0$

$$\mu(\{\lambda : \lambda(1-s) > 1\}) > 0.$$

Let r_c be the supremum of all such s. As observed above if a genotype has a birth rate λ such that $\lambda(1-r) > 1$ then this genotype has a positive probability of surviving when the mutation probability is r. Hence, the population has a positive survival probability for any $r < r_c$.

This completes the proof of Theorem 2.

Remark 4. The model with no genotype extinctions was introduced in [5]. In the particular case where the distribution μ is uniform the existence of a unique critical value for r was proved. Below the critical value there is survival and above it there is extinction. The existence of a unique critical value for a general μ is unclear. We conjecture that the survival probability for both models (with and without genotype extinctions) is not monotone in r. Hence, results such as Corollary 1 are not enough to prove the existence of a unique critical value for r for general μ.

3 Proof of Theorem 1

We use ideas from [5]. However, there the computation is done for the model with no genotype extinctions. The main idea is the use of the genealogy tree of genotypes from [6]. We now define this tree.

We say that the (unique) individual present at time zero has genotype 1, and the kth type to appear will be called genotype k. Each vertex in the tree will be labeled by a positive integer. There will be a vertex labeled k if and only if an individual of genotype k is born at some time. We draw a directed edge from j to k if the first individual of genotype k to be born had an individual of genotype j as its parent. This construction gives a tree whose root is labeled 1 because all genotypes are descended from genotype 1 that is present at time zero. Since every genotype is eliminated eventually, the population survives if and only if the genealogy tree just described has infinitely many vertices.

We claim that this genealogy tree of genotypes is a discrete time Galton-Watson tree. This is so because offsprings of different individuals in the tree are independent and have the same distribution. Let $m(r)$ be the mean offspring of a given genotype in this tree. We know this genealogy tree is infinite with positive probability if and only if $m(r) > 1$, see for instance [7].

We now compute $m(r)$. Recall that we start the population with a single individual with a genotype that we label 1. We associate a birth rate Λ and a death time T to this genotype. Recall that Λ and T are independent and sampled from fixed distributions μ and ν, respectively.

Let Y_t be the number of individuals born up to time t that are offspring of genotype 1 individuals but whose genotype is not 1. Hence, Y_T is the total number of genotypes that genotype 1 individuals gave birth to before genotype 1 died out. That is, Y_T is the offspring of genotype 1 in the genealogy tree. Hence,

$$m(r) = E(Y_T).$$

By our assumption that T is independent of everything else in the process we have

$$E(Y_T|T = t, \lambda) = E(Y_t|\lambda).$$

Let X_t be the number of genotype 1 individuals present at time t. Genotype 1 individuals produce other genotype individuals at rate λr. Thus,

$$\frac{d}{dt}E(Y_t|\lambda) = \lambda r E(X_t|\lambda).$$

Since X_t is a birth and death process with birth rate $\lambda(1-r)$ (genotype 1 individuals produce genotype 1 individuals at rate $\lambda(1-r)$) and death rate 1,

$$E(X_t|\lambda) = \exp\left((\lambda(1-r)-1)t\right).$$

Hence,

$$E(Y_t|\lambda) = r\lambda \int_0^t E(X_s|\lambda)ds = r\lambda \int_0^t \exp((\lambda(1-r)-1)s)ds.$$

Since $E(Y_T) = \int \int E(Y_t|\lambda) d\mu(\lambda) d\nu(t)$, we have

$$E(Y_T) = \int r\lambda \int_0^t E(X_s|\lambda) ds d\mu(\lambda) d\nu(t) = \int r\lambda \int_0^t \exp\left((\lambda(1-r)-1)s\right) ds d\mu(\lambda) d\nu(t).$$

Hence,

$$m(r) = E(Y_T) = rE[\Lambda \int_0^T \exp\left((\Lambda(1-r)-1)s\right) ds].$$

This completes the proof of Theorem 1.

References

1. Nowak, M.A., May, R.M.: Virus Dynamics. Oxford University Press, Oxford (2000)
2. Eigen, M.: Error catastrophe and antiviral strategy. PNAS **99**, 13374–13376 (2002)
3. Manrubia, S.C., Domingo, E., Lazaro, E.: Pathways to extinction: beyond the error threshold. Phil. Trans. R. Soc. B **365**, 1943–1952 (2010)
4. Mayr, E.: What Evolution Is. Basic Books, New York (2001)
5. Cox, J.T., Schinazi, R.B.: A branching process for virus survival. J. Appl. Probab. **49**, 888–894 (2012)
6. Schinazi, R.B., Schweinsberg, J.: Spatial and non spatial stochastic models for immune response. Markov Process. Relat. Fields **14**, 255–276 (2008)
7. Schinazi, R.B.: Classical and Spatial Stochastic Processes, 2nd edn. Birkhauser, New York (2014)

Particle Transport in a Confined Ratchet Driven by Colored Noise

Yong Xu[1,2,3(✉)], Ruoxing Mei[1], Yongge Li[1], and Jürgen Kurths[2,3]

[1] Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an 710129, China
hsux3@nwpu.edu.cn, ruoxing_mei@yahoo.com, liyonge@163.com
[2] Potsdam Institute for Climate Impact Research, 14412 Potsdam, Germany
juergen.kurths@pik-potsdam.de
[3] Department of Physics, Humboldt University Berlin, 12489 Berlin, Germany

Abstract. In this paper, we study particle transport in a confined ratchet which is constructed by combining a periodic channel with a ratchet potential under colored Gaussian noise excitation. Due to the interaction of colored noise and confined ratchet, particles host remarkably different properties in the transporting process. By means of the second-order stochastic Runge-Kutta algorithm, effects of the system parameters, including the noise intensity, colored noise correlation time and ratchet potential parameters are investigated by calculating particle current. The results reveal that the colored noise correlation time can lead to an increase of particle current. The increase of noise intensity along the horizontal or vertical direction can accelerate the particle transport in the corresponding direction but slow down the particle transport when there are the same noise intensities in both directions. For potential parameters, an increase of the slope parameter results into an increase of particle currents. The interactions of potential parameters and correlation time can induce complex particle transport phenomena, i.e. particle current increases with the increase of the potential depth parameter for a smaller asymmetric parameter and non-zero correlation time, while the tendency changes for a larger asymmetric parameter. Accordingly, suitable system parameters can be chosen to accelerate the particle transport and used to design new devices for particle transport in microscale.

Keywords: Particle current · Confined ratchet · Colored noise · Particle transport

1 Introduction

Directed particle transport is a common phenomenon in many fields ranging from physics to life science [1–4], where the transport in a ratchet potential is one of the most interesting problems. In recent years, many papers have appeared to demonstrate noise-induced transport. Some works focused on the structure of the

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 443–456, 2019.
https://doi.org/10.1007/978-3-030-15096-9_15

ratchet potentials [5–9], including symmetric and asymmetric [5–7], rough and smooth potentials [8,9]. Others mainly investigated the influences of different kinds of noises, such as Gaussian white noise, correlated noise [10], Lévy noise with large jumps [11], etc. Quantities like mean first passage time [3], stationary probability distribution [6] and particle current [7] have been considered to evaluate the transport. Phenomena like negative mobility, multiple current reversals under nonequilibrium state and stochastic resonance [12] have been observed in the theoretical studies.

However, previous works are mostly limited to an open area without confined boundaries. In fact, like ion channels, pores and zeolites [13–16], one has to take the confined geometry into consideration, which will reduce the available motion space and regulate the transport properties. For simplification and idealization, a periodic channel is one of the popular models, based on which the Fick-Jacobs (FJ) equation is developed to describe the particle transport [17–19]. The particle current could be utilized in symmetric periodic channels where an external force is applied on the particles [20,21]. Particle current also could be induced in asymmetric channels without external force because of the non-equilibrium state caused by the channel asymmetry [22,23]. Later, more meaningful factors were introduced, such as periodic channel pore sizes [24], particle radios [25] and shapes [26]. A number of interesting particle transport phenomena have been discovered, for instance, directional transport [27] and current reversals [28]. The related implications were used in catalysis, particle separations [29–31] and directed transport [2,32,33]. In addition, the investigated microscopic levels are getting finer. It is possible to add the semicro-nano systems into confined areas whose boundaries would have influences on the system behavior. Considering this, the particle transport in a confined ratchet which consists of a ratchet potential and a periodic channel under Gaussian white noise was studied [34]. Due to the interaction between the channel and the potential, particle current may be induced, although both of them cannot induce the particle movement when acting individually. This confined ratchet provided new ideas for the study of particle transport and preparation of related devices.

Most of particle transport in the confined environment described the random fluctuation as Gaussian white noise, which is considered as an ideal model. However, perturbations in complex systems, such as ion channels in living bodies, photon statistics of a dye-laser [35] and the motional narrowing of magnetic resonance line shapes [36], are correlated. This correlation can affect the particle transport. In these cases, idealized Gaussian white noise cannot describe the system noises and the effects of noise correlation time appropriately. Gaussian colored noise with a non-vanishing correlation time becomes a proper tool to well understand these disturbances. So, particle transport in a confined ratchet by colored noise is essential. In the case of Gaussian white noise, particle transport can be described by a FJ equation quite accurately [37]. However, in the presence of correlation time, there is no explicit FJ equation to model the particle dynamics in confined channels and potentials. Thus, in this paper, we consider random fluctuations in a periodic channel as colored Gaussian noise with exponential correlation time, to investigate the particle transport numerically.

The influences of Gaussian colored noise on particle current in a confined ratchet consisting of a potential and a periodic channel will be explored thoroughly. This paper is arranged as follows. In Sect. 2, the model of a particle in a periodic channel and a ratchet potential, driven by Gaussian colored noise is introduced. In Sect. 3, effects of system parameters, including noise intensity, correlation time, potential slope parameter, depth parameter and asymmetric parameter, on the particle current are presented. Finally, conclusions are given in Sect. 4.

2 Model Description

We consider the particle transport in a ratchet potential within a periodic channel driven by Gaussian colored noise, and the model in this paper consists of two parts.

The first part governs the dynamical equation of a particle with a ratchet potential. In general, the dynamics of the particle under the assumption [20] can be regarded as a Langevin equation, which reads

$$\eta \frac{d\hat{x}}{d\hat{t}} = -U'(\hat{x}) + \sqrt{\eta K_B T}\varepsilon_1(\hat{t}), \tag{1}$$

$$\eta \frac{d\hat{y}}{d\hat{t}} = \sqrt{\eta K_B T}\varepsilon_2(\hat{t}), \tag{2}$$

where $(\hat{x}, \hat{y}) \in R^2$ is the particle position, $\hat{t} \in R$ the real time, η the friction coefficient, K_B the Boltzmann constant and T the temperature. The prime on $U(\hat{x})$ is the derivation with respect to $\hat{x}$. $\varepsilon_1(\hat{t})$ and $\varepsilon_2(\hat{t})$ are two independent noises. $\varepsilon_1(\hat{t})$ denotes an exponentially correlated Gaussian colored noise with zero-mean and the correlation function $\langle \varepsilon_1(\hat{t})\varepsilon_1(\hat{s})\rangle = D_1 exp(|\hat{t}-\hat{s}|/\tau_1)/\tau_1$ with correlation time $\tau_1 > 0$. $\varepsilon_2(\hat{t})$ represents a zero-mean Gaussian white noise with the correlation function $\langle \varepsilon_2(\hat{t})\varepsilon_2(\hat{s})\rangle = 2D_2\delta(\hat{t}-\hat{s})$. $D_i\,(i=1,2)$ is the noise intensity, and $\delta(\cdot)$ denotes the Dirac function. $U(\hat{x})$ is a ratchet potential, which is given by

$$U(\hat{x}) = \begin{cases} -a\hat{x}/L - b\cos(\pi \bmod (\hat{x}/L, L)/L_1), & 0 \le \hat{x} < L_1, \\ -a\hat{x}/L + b\cos(\pi(\bmod (\hat{x}/L, L) - L_1)/L_2), & L_1 \le \hat{x} < L, \end{cases} \tag{3}$$

where $L > 0$ is the period of $U(\hat{x})$, $L_1 = (1+q)L/2 > 0$ and $L_2 = (1-q)L/2 > 0$ are lengths of left and right parts in one potential period satisfying $L_1 + L_2 = L$, in which $q \in (-1, 1)$ can be viewed as the potential asymmetric parameter, and $\bmod(\cdot)$ is the modulo function. The potential is symmetric for $q = 0$ and asymmetric for $q \neq 0$. a, $b \in R$ are the slope and depth parameters to modify the shape and depth of $U(\hat{x})$, respectively.

The second part is the two-dimensional periodic channel, denoted by $w(\hat{x})$. It consists of the upper channel wall $w_+(\hat{x})$ and the lower channel wall $w_-(\hat{x})$,

which is symmetric about the $\hat{x}$ axis, i.e. $w_+(\hat{x}) = -w_-(\hat{x})$. The upper wall $w_+(\hat{x})$ satisfying

$$w_+(\hat{x}) = m\sin(2\pi\hat{x}/L) + n, \tag{4}$$

where m and n are parameters to control the channel shape with $m < n \in R$.

To simplify the model Eqs. (1–3), several dimensionless variables $x = \hat{x}/L$, $y = \hat{x}/L$, $\tau_c = L^2\eta/K_B T_{room}$ are introduced [20], where T_{room} is a constant room temperature. With these variables, Eqs. (1–3) can be rewritten in the following non-dimensional form

$$\frac{dx}{dt} = -\frac{T}{T_{room}}U'(x) + \sqrt{\frac{T}{T_{room}}}\varepsilon_1(t), \tag{5}$$

$$\frac{dy}{dt} = \sqrt{\frac{T}{T_{room}}}\varepsilon_2(t), \tag{6}$$

$$U(x) = \begin{cases} -ax - b\cos(\pi \bmod(x,1)/L_1), & 0 \le \bmod(x,1) < L_1, \\ -ax + b\cos(\pi(\bmod(x,1) - L_1)/L_2), & L_1 \le \bmod(x,1) < 1, \end{cases} \tag{7}$$

where t represents the dimensionless time. $\varepsilon_1(t)$ is the dimensionless colored noise with the correlation time $\tau = \tau_1/\tau_c$. $U(x)$ is the dimensionless potential with period $L = 1$, $L_1 = (1+q)/2$ and $L_2 = (1-q)/2$. The periodic channel $w(\hat{x})$ becomes $w(x)$, with upper wall $w_+(x)$ satisfying

$$w_+(x) = m\sin(2\pi x) + n, \tag{8}$$

and the corresponding lower channel wall $w_-(x)$ satisfying $w_-(x) = -w_+(x)$. The dimensionless channel $w(x)$ has maximum width $2(m+n)$ and minimum width $2(n-m)$, with $m = 1/(2\pi)$, $n = 1.02/(2\pi)$ in this paper [21]. The channel $w(x)$ and potential $U(x)$ work together to affect particle transport. Diagram of $w(x)$ and $U(x)$ is illustrated in Fig. 1.

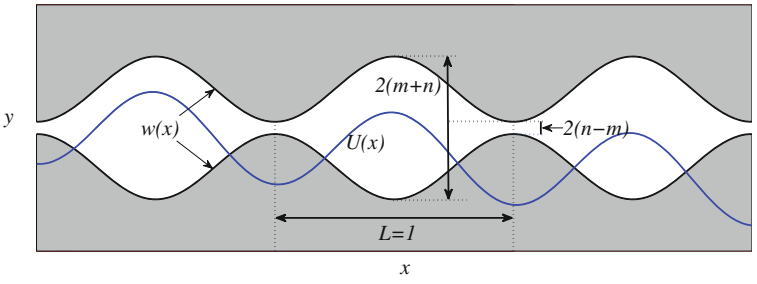


Fig. 1. Sketch of the periodic channel $w(x)$ with period $L = 1$, the maximum width $2(m+n)$, and the minimum width $2(n-m)$. The blue line is $U(x)$, with $q = 0$, $a = 0.1$, $b = 0.2$.

When transporting in $w(x)$, particles will definitely hit the channel walls sometimes. In the present work, we suppose that collisions between particles and channel walls are elastic, i.e. the reflecting boundary condition is fulfilled at the channel walls. Then, the available spaces for particles reduce to the inside of the channel. The reduction of available spaces eventually results into changes of particle transport properties. To explore these distinct transport phenomena, particle current is calculated numerically. Moreover, considering the reflecting conditions of channel walls, the particle transport in the y axis will always be confined, so we only consider the particle current along the x direction, which is defined as [25]

$$J = \lim_{t\to\infty} \frac{\langle x(t) - x(0)\rangle}{t},$$

where $x(t)$ is the particle position at time t, the bracket $\langle\cdot\rangle$ represents the mean value of samples.

3 Results and Discussion

Next, the influences of the parameters a, b, q, D_1 and τ on J are discussed by numerical simulation where the numerical results come from 5000 samples. Each sample consists of 10^7 sample points. To ensure the influences of τ on J are not ignored, the time step is fixed as $\Delta t = 0.0001$, much smaller than τ. Numerical results are obtained by using the second-order stochastic Runge-Kutta algorithm [38] as

$$x(i+1) = x(i) + \Delta t(F_1 + F_2)/2,$$

$$\varepsilon_1(i+1) = \varepsilon_1(i) + \Delta t(H_1 + H_2)/2 + \sqrt{2D_1\Delta t/{\tau_1}^2}\psi_1,$$

$$y(i+1) = y(i) + \sqrt{2D_2\Delta t}\psi_2, \psi_2 \sim N(0,1),$$

where

$$F_1 = f(x(i), \varepsilon_1(i)), \quad H_1 = h(\varepsilon_1(i)),$$

$$F_2 = f\left(x(i) + \Delta t F_1, \varepsilon_1(i) + \Delta t H_1 + \sqrt{2D_1\Delta t/{\tau_1}^2}\psi_1\right),$$

$$H_2 = h\left(\varepsilon_1(i) + \Delta t H_1 + \sqrt{2D_1\Delta t/{\tau_1}^2}\psi_1\right), \quad \psi_1 \sim N(0,1),$$

$$f(x,\varepsilon_1) = -TU'(x)/T_{room} + \sqrt{T/T_{room}}\varepsilon_1, \quad h(\varepsilon_1) = -\varepsilon_1/\tau.$$

3.1 Particle Current *vs* Correlation Time

First, the influences of τ on J are discussed. When $\tau = 0$, the particle is driven by the Gaussian white noise. In this case, its probability density function $p(x, y, t)$ in an open area without confined boundaries satisfies the FPK equation

$$\frac{\partial p\left(x,y,t\right)}{\partial t}=\frac{T}{T_{room}}\left(\frac{\partial U'\left(x\right)}{\partial t}+\frac{\partial^2 D_1}{\partial x^2}+\frac{\partial^2 D_2}{\partial y^2}\right)p\left(x,y,t\right).$$

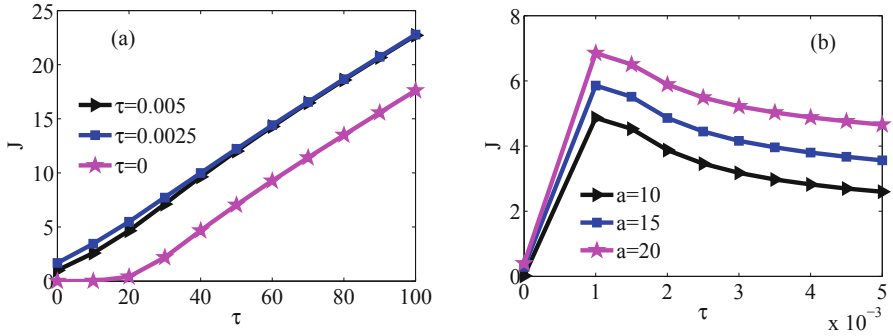


Fig. 2. Dependences between J and τ. Other parameters are $T/T_{room} = 0.2$, $b = 5$, $q = 0$, $D_1 = D_2 = 1$.

Then, taking the confinement and structure of $w(x)$ that its length in the x-axis is much larger than the y-axis into consideration, it is reasonable to assume that the particles in the y-axis direction have fast equilibrium. The FPK equation can then be reduced into

$$\frac{\partial p\left(x,y,t\right)}{\partial t}=\frac{T}{T_{room}}\left(\frac{\partial U'\left(x\right)}{\partial t}+\frac{\partial^2 D_1}{\partial x^2}\right)p\left(x,y,t\right).$$

Then, by integrating the reduced FPK equation about y, FJ equation to describe the particle dynamics in $w(x)$ is obtained, which is

$$\frac{\partial p\left(x,t\right)}{\partial t}=\frac{\partial}{\partial x}\frac{D_1}{\left(1+w'(x)^2\right)^{1/3}}e^{-A(x)}\frac{\partial}{\partial x}e^{A(x)}p\left(x,t\right).$$

From the FJ equation, we obtain the analytical expression of J, satisfying

$$J=\frac{TD_1\left(1-e^{a/D_1}\right)}{T_{room}\int_0^1\int_x^{x+1}\frac{w(x)}{w(z)}\cdot\left(1+w'(x)^2\right)^{1/3}\cdot e^{(U(z)-U(x))/D_1}dzdx}.$$

This analytical J can be calculated by the Gaussian quadrature formula. When $\tau \neq 0$, the particle is driven by colored noise. In this case, J is calculated by numerical simulation.

J under Gaussian white noise and colored case are compared in Fig. 2a. First, we see that J is larger under Gaussian colored noise than white case. This indicates that τ can accelerate the particle transport in the periodic channel and enlarge J. Another interesting phenomenon is that J for $\tau = 0.005$ is smaller than the case $\tau = 0.0025$ for $a < 40$. So, one can guess, the larger the correlation time, the weaker the promoting effect on the particle transport. To prove this conclusion, relations between J and τ for the cases $a = 10$, $a = 15$ and $a = 20$ are plotted in Fig. 2(b). As expected, in all cases of Fig. 2, with the increase of non-zero τ, J decreases.

Thus, the following conclusions about the influences of τ on J are obtained. First, the colored noise correlation time can promote particle transport compared to the white case, lead to a bigger J. Second, promoting effects of τ decrease with the increase of τ for $a < 40$.

3.2 Particle Current *vs* Slope Parameter and Noise Intensity

This subsection is to investigate the influences of potential slope parameter a and noise intensities $D_i\,(i = 1, 2)$ on J.

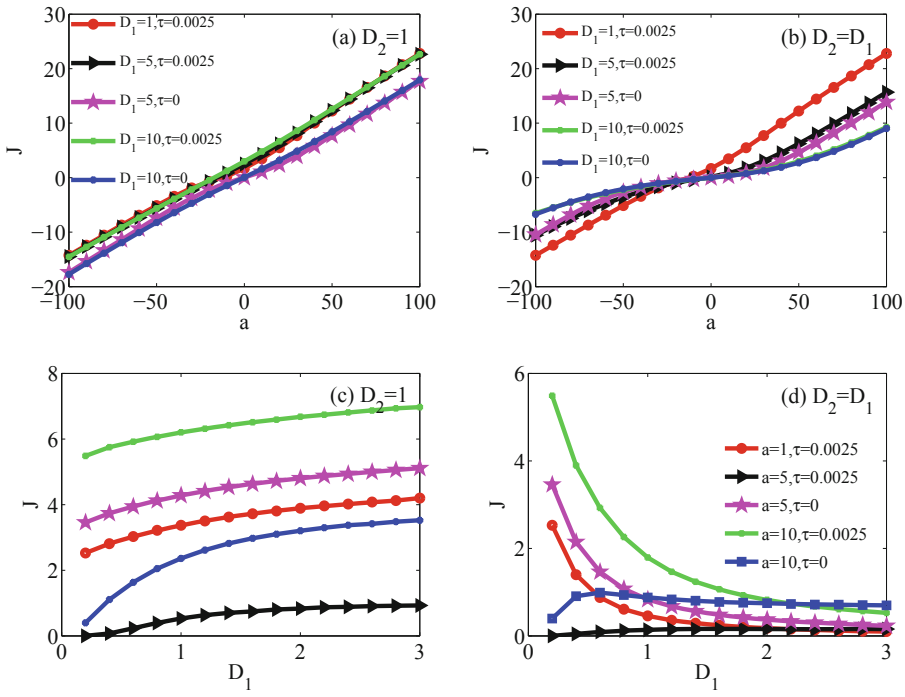


Fig. 3. Dependences between J and τ under different D_1 are plotted at (a): $D_2 = 1$; (b): $D_2 = D_1$. Dependences between J and D_1 under different a are plotted at (c): $D_2 = 1$; (d): $D_2 = D_1$. Other parameters are $T/T_{room} = 0.2$, $b = 5$, $q = 0$.

Note that a has several impacts on the potential $U(x)$. When $a \neq 0$, $U(x)$ can be viewed as a titled potential, which can produce a net current [7]. Also, a determines the descent rate of $U(x)$ and affects the potential well depth to a certain degree. These impacts regulate the particle transport directly or indirectly. This leads us to consider the influence of a on the particle transport, as shown in Fig. 3(a–b). Our results in Fig. 3(a–b) indicate that the larger a is, the faster J increases. As is well-known, when a is larger, for a particle driven by a greater external force, its velocity is faster, which leads to a larger J. Figure 3(a–b) also show that, J has the same direction with a. When a is positive, $U(x)$ goes down along the right direction, thus the stable left point in $U(x)$ is always higher than the stable right point in each of the two adjacent potential wells. In this way, particles in potential wells are more likely to move along the right direction, and thus J is positive.

The influences of $D_i\,(i = 1, 2)$ on J are studied in detail in Fig. 3(c–d). Influences for fixed $D_2 = 1$ indicate that J increases with the increase of D_1, i.e. the increase of D_1 can promote the particle transport in Fig. 3(c). However, when D_2 is not fixed, changing with D_1, J shows different phenomena, just as shown in Fig. 3(d) where the influences of the noise intensity are plotted. In this case, J decreases faster and tends to zero eventually with the increase of D_1. Phenomena in Fig. 3(d) are caused by the increase of D_2. For particles, the increase of D_2 leads to a larger movement along the y axis direction. It is easier for a particle to collide with the channel walls in this case. These collisions result into the decrease of the movement in the x direction consequently. Above all, one can say that the increase of either D_1 or D_2 is able to enlarge J in their corresponding directions. However, when D_1=D_2, the increase of the noise intensities in both directions slows down J.

3.3 Particle Current *vs* Asymmetric Parameter

The asymmetric parameter q determines the asymmetry of $U(x)$. With $q = 0$, $L_1 = L_2$, $U(x)$ is symmetric in each period, and with $q < 0$, $L_1 < L_2$, $U(x)$ is asymmetric in one period. Particles in such an asymmetric potential suffer different forces in the left and right side. It has been confirmed that particles in an asymmetric potential can produce a net current [7]. So, we conclude that q plays an important role to J in our present work.

The influences of q for different a on J are plotted in Fig. 4(a) with $\tau = 0.0025$. The interesting one is that J does not equal to 0 for $a = 0$ and $q = 0$. The occurrence of this phenomenon is related to the particle initial position. In our simulations, the initial position is set to be the potential well, i.e. $(0, 0)$ here, not in the channel cavity center, relatively close to the left channel pore and far from the right pore. Thus, the particle has more chances to transport along the right direction, which lead to positive, non-zero J, just as shown in the purple curve in Fig. 4(a). Another phenomenon is, with the increase of q, J reaches to the maximum value for $q < 0$, while it turns to slow down for smaller positive q and eventually approaches to stabilize for larger positive values of q. At the same time, J tends to the same value, although the values of a are different. Detailed analysis about the changing of J at larger q is displayed in Fig. 5.

The influences of q on J for different b are presented in Fig. 4(b). Now we deal with the case $b < 0$ and the original position of the particle is located at $(0.5\,(1+q)\,,0)$. When $q = 0$, the original position becomes $(0.5, 0)$, at the right side of the channel cavity. The available space on the left is larger than the right in the channel cavity for the particle, which makes it easier to transport along the left direction and brings about a left-directed J eventually, just as shown in the red and black lines in Fig. 4(b). For $q < 0$, the particle still starts from the right side of the channel cavity, which similarly leads to a particle current in the left direction, and the potential is asymmetric. The left part length L_1 and the right part length L_2 in one period satisfy $L_1 < L_2$, i.e. the left side is steeper. This asymmetric structure affects the particle transport, reduces the probability for particles to transport along the left direction and produces a right-directed particle current. J is the superposition of the left-directed and the right-directed current mentioned above. With the increase of q, the initial position of the particle moves to the right, which leads to a decrease of left directed current. This part is viewed as the effect of the channel on particles. The increase of q also leads to the increase of the left part length L_1. Thus, the steep degree of the potential left side decreases, which results in the decrease of right-directed current. This part is the influences of the potential. For the particle, the channel effect is the main aspect to regulate its transport. Thus, J is left-directed and decreases.

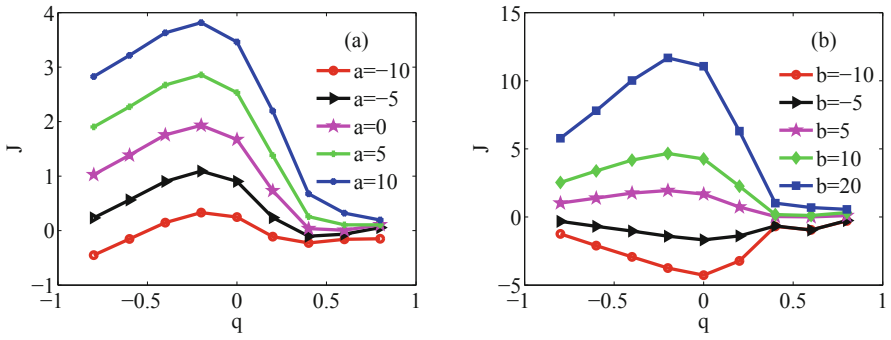


Fig. 4. J with respect to different parameters for $D_1 = D_2 = 1$, $T/T_{room} = 0.2$. (a): J as functions of q for different a, at $b = 5$. (b): J with respect to q for different b, at $a = 0$.

Then the case $q > 0$ is considered. In this situation, the original particle position is in the right and closer to the right channel pore than for $q \leq 0$. The bigger the q is, the smaller the distance between the particle and the right channel pore is. This induces a left-directed particle current. However, there are two other items to affect the particle transport. One is the collisions between the

particle and the channel walls. When positive q increases, the smaller distance makes more chances for the particle collide with the periodic channel walls, which leads to the decrease of J. The other is the potential asymmetry. The right side of the potential well is steep which results into a current directed to the left direction. With the increase of q, these three parts regulate the particle transport together, and lead to the decrease of J eventually. Another meaningful phenomenon in Fig. 4(b) is that, when $q > 1/2$, J tends near to 0. In this case, the initial position is almost in the channel pore, which makes great difficulties for the particle to escape from the pore, although the asymmetric $U(x)$ can help the particle to move along the left direction. Thus, J is almost unchanged ultimately at a larger q.

Moreover, Fig. 4(b) also demonstrates the influences of q for $b > 0$. Similar to the case $b < 0$, with the increase of q, J possesses complex behaviors. As q changes from negative to positive, J increases, then decreases, and tends to 0 at last. From Fig. 4(b), we can infer that, for given q and $b > 0$, when b takes a larger value, J is bigger.

3.4 Particle Current *vs* Interactions of Depth Parameter, Asymmetric Parameter and Correlation Time

The system parameters a, b and q determine the properties of $U(x)$ and come into play with particle transport. Thus, we focus our attention on the influences of interactions among b with respect to different q and τ in Fig. 5. In Fig. 6, the relationships between J and b under different $D_i\,(i = 1, 2)$ are presented.

J as functions of a are plotted in Fig. 5(a–c). There are three similarities in them: (i) For any q and τ, J always increases with the increase of a. (ii) J has the same direction as a. (iii) driven by Gaussian white noise is smaller than in the colored Gaussian one. These phenomena are the same as Fig. 3, and reasons have been explained in the corresponding Sect. 3.2. So, we do not repeat here.

There are also some differences between Fig. 5(a–b) and (c) when the particle is driven by a Gaussian colored noise. In Fig. 5(a–b), when b changes from 1 to 10, J becomes bigger, i.e. a larger b leads to a bigger J for colored noise for $q = -0.7$ and $q = 0$. However, J is smaller for $b = 10$ than in the cases $b = 1$ and $b = 5$ in Fig. 5(c). This leads to a speculation that, when $\tau \neq 0$, the influences of b on J for $q = 0.7$ is opposite to the cases $q = -0.7$ and $q = 0$. To understand the different influences in detail, Fig. 5(d–f) are plotted then. In Fig. 5(d–e), J is proportional to b with $\tau = 0.0025$. While in Fig. 5(f), the tendency of J changes for $\tau = 0.0025$. The interactions of q and $\tau \neq 0$ bring about complicated or opposite effects to J compared with white noise where $\tau = 0$.

However, when $\tau = 0$, i.e. the particle is driven by Gaussian white noise, the results are relatively simple and J always decreases with the increase of b.

Next, we impose on cases of noise intensities $D_i (i = 1, 2)$ interacting with the potential parameter b. The results are plotted in Fig. 6. In Fig. 6(a), the variation tendency of J with respect to b is presented for different noise intensities at $a = 5$. In Fig. 6(a), for a fixed D_1, with the increase of b, J increases for $\tau = 0.0025$, while it decreases at $\tau = 0$, although the variation range is so small to be ignored.

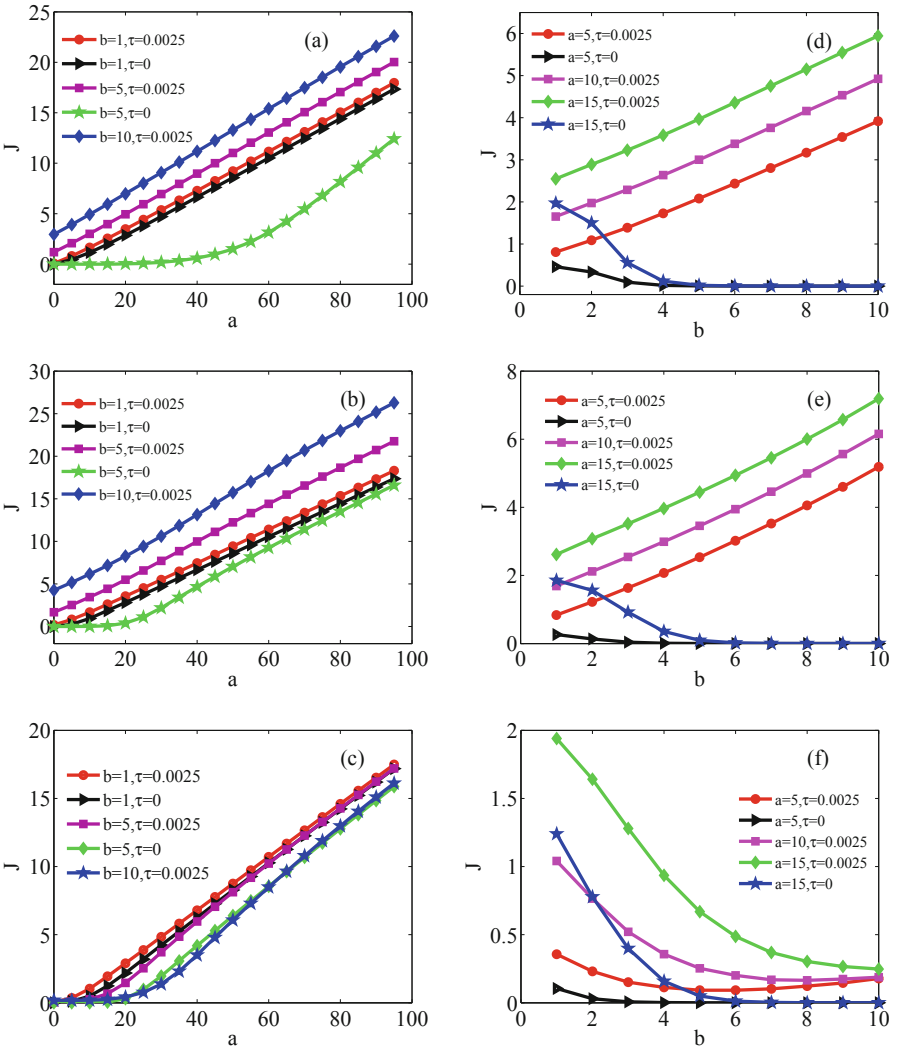


Fig. 5. J with respect to a and b for different q at $D_1 = D_2 = 1$, $T/T_{room} = 0.2$. J as functions of a for different b, at (a): $q = -0.7$, (b): $q = 0$, (c): $q = 0.7$. J with respect to b for different a, at (d): $q = -0.7$, (e): $q = 0$, (f): $q = 0.7$.

This indicates that the influences of the interactions between $D_i (i = 1, 2)$ and $\tau \neq 0$ are different from the white case. In Fig. 6(b), the influences of $D_i (i = 1, 2)$ on J are plotted for different b and τ. We find that J decreases fast and tends to a value near zero with increasing intensity D_1 for $\tau = 0.0025$. But as τ decreases to 0, variation of J becomes very slow.

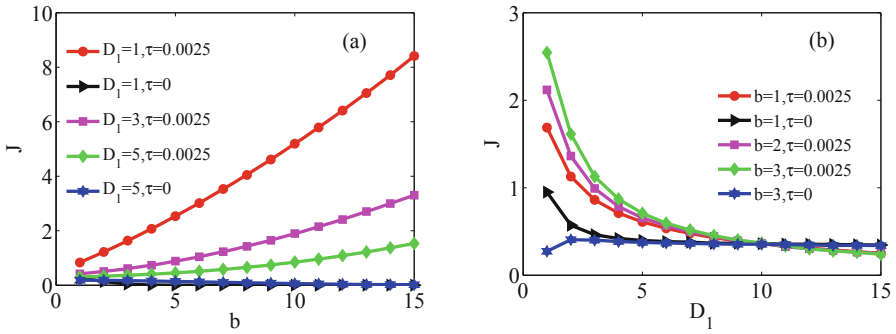


Fig. 6. J with respect to D_1 and b, for $q = 0$, $D_1 = D_2$, $T/T_{room} = 0.2$.

4 Conclusions

In this paper, transport properties of a particle in a confined ratchet driven by a Gaussian colored noise are examined. The particle current is calculated by numerical simulations. The influences of model parameters, like the noise intensity, correlation time, potential slope and depth parameters and asymmetric parameter, on the particle current are analyzed.

Our results indicate that the increase of the Gaussian colored noise intensity enlarges the particle transport and induces a larger particle current. However, with the increase of the colored and white noise intensities simultaneously, the particle current decreases. In addition, influences of the correlation time on particle current are discussed in detail. The existence of a correlation time can promote the particle transport, and lead to a larger particle current than the Gaussian white case. Promoting effects of the Gaussian colored noise decreases with the increase of the correlation time. For the slope parameter, it can promote the particle transport. A larger slope parameter induces a larger particle current. What is more, we also find that the depth parameter has different effects on the particle current in different cases under the Gaussian colored noise. For small asymmetric parameters, the particle current increases with the increase of the depth parameter. However, the relationship changes for a larger asymmetric parameter. The particle current increases and then decreases with the increase of the asymmetric parameter. Based on these transport phenomena, suitable system parameters can be chosen and implemented in the design of new transport control devices.

Acknowledgement. This work was supported by the NSF of China (11772255), the Fundamental Research Funds for the Central Universities, the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University. Y. Xu thanks to the Alexander von Humboldt Foundation.

References

1. Reimann, P.: Introduction to the physics of Brownian motors. Appl. Phys. A **75**(2), 169–178 (2002)
2. Angelani, L.: Active ratchets. Europhys. Lett. **96**(6), 68002 (2011)
3. Xu, Y.: The estimates of the mean first exit time of a bistable system excited by Poisson white noise. J. Appl. Mech. **84**(9), 091004 (2017)
4. Wu, J.: Information-based measures for logical stochastic resonance in a synthetic gene network under Lévy flight superdiffusion. Chaos **27**(6), 063105 (2017)
5. Cao, L.: Fluctuation-induced transport in a spatially symmetric periodic potential. Phys. Rev. E **62**(5), 7478–7871 (2000)
6. Reimann, P.: Brownian motors: noisy transport far from equilibrium. Phys. Rep. **361**(2), 57–265 (2002)
7. Reimann, P.: Giant acceleration of free diffusion by use of tilted periodic potentials. Phys. Rev. Lett. **87**(1), 010602 (2001)
8. Li, Y.: Lévy noise induced transport in a rough triple-well potential. Phys. Rev. E **94**(4), 042222 (2016)
9. Li, Y.: Transports in a rough ratchet induced by Lévy noises. Chaos **27**(10), 103102 (2017)
10. Jung, P.: Colored noise in dynamical systems: some exact solutions. In: Stochastic Dynamics. Lecture Notes in Physics, vol. 484, pp. 23–31 (1997)
11. Xu, Y.: The switch in a genetic toggle system with Lévy noise. Sci. Rep. **6**, 31505 (2016)
12. Wang, Z.: Lévy noise induced stochastic resonance in an FHN model. Sci. China Technol. Sci. **59**(3), 371–375 (2016)
13. Kullman, L.: Transport of maltodextrins through maltoporin: a single-channel study. Biophys. J. **82**(2), 803–812 (2002)
14. Kosinska, I.: Rectification in synthetic conical nanopores: a one-dimensional Poisson-Nernst-Planck model. Phys. Rev. E **77**(3), 031131 (2008)
15. Berezhkovskii, A.: Optimizing transport of metabolites through large channels: molecular sieves with and without binding. Biophys. J. **88**(3), L17–L19 (2005)
16. Siwy, Z.: Asymmetric diffusion through synthetic nanopores. Phys. Rev. Lett. **94**(4), 048102 (2005)
17. Zwanzig, R.: Diffusion past an entropy barrier. J. Phys. Chem. **96**(10), 3926–3930 (1992)
18. Kalinay, P.: Corrections to the Fick-Jacobs equation. Phys. Rev. E **74**(4), 041203 (2006)
19. Dorfman, K.: Assessing corrections to the Fick-Jacobs equation. J. Chem. Phys. **141**(4), 044118 (2014)
20. Reguera, D.: Entropic transport: kinetics, scaling, and control mechanisms. Phys. Rev. Lett. **96**(13), 130603 (2006)
21. Burada, P.: Biased diffusion in confined media: test of the Fick-Jacobs approximation and validity criteria. Phys. Rev. E **75**(5), 051111 (2007)
22. Reguera, D.: Entropic splitter for particle separation. Phys. Rev. Lett. **108**(2), 020604 (2012)
23. Li, Y.: Fine separation of particles via the entropic splitter. Phys. Rev. E **96**(2), 022152 (2017)
24. Ghosh, P.: Detectable inertial effects on Brownian transport through narrow pores. Europhys. Lett. **98**(5), 50002 (2012)

25. Riefler, W.: Entropic transport of finite size particles. J. Phys. Condens. Matter **22**(45), 454109 (2010)
26. Ghosh, P.: Self-propelled Janus particles in a ratchet: numerical simulations. Phys. Rev. Lett. **110**(26), 268301 (2013)
27. Li, F.: Current control in a two-dimensional channel with nonstraight midline and varying width. Phys. Rev. E **87**(6), 062128 (2013)
28. Ao, X.: Active Brownian motion in a narrow channel. Eur. Phys. J.-Spec. Top. **223**(14), 3227–3242 (2014)
29. Volkmuth, W.: DNA electrophoresis in microlithographic arrays. Nature **358**(6387), 600–602 (1992)
30. Han, J.: Separation of long DNA molecules in a microfabricated entropic trap array. Science **288**(5468), 1026–1029 (2000)
31. Chang, R.: Dynamics of chain molecules in disordered materials. Phys. Rev. Lett. **96**(10), 107802 (2006)
32. Pineda, I.: Diffusion in two-dimensional conical varying width channels: comparison of analytical and numerical results. J. Chem. Phys. **137**(17), 174103 (2012)
33. Ai, B.: Rectified Brownian transport in corrugated channels: fractional Brownian motion and Lévy flights. J. Chem. Phys. **137**(17), 174101 (2012)
34. Malgaretti, P.: Confined Brownian ratchets. J. Chem. Phys. **138**(19), 194906 (2013)
35. Short, R.: Correlation functions of a dye laser: comparison between theory and experiment. Phys. Rev. Lett. **49**(9), 647–650 (1982)
36. Kubo, R.: Fluctuation, Relaxation and Resonance in Magnetic Systems. Oliver and Boyd, Edinburgh (1962)
37. Reguera, D.: Kinetic equations for diffusion in the presence of entropic barriers. Phys. Rev. E **64**(6), 061106 (2001)
38. Honeycutt, R.: Stochastic Runge-Kutta algorithms. II. Colored noise. Phys. Rev. A **45**(2), 604–610 (1992)

Long-Time Dynamics for a Simple Aggregation Equation on the Sphere

Amic Frouvelle[1(✉)] and Jian-Guo Liu[2]

[1] CEREMADE, CNRS, Université Paris-Dauphine, Université PSL, 75016 Paris, France
frouvelle@ceremade.dauphine.fr

[2] Department of Physics and Department of Mathematics, Duke University, Durham, NC 27708, USA
jliu@phy.duke.edu

Abstract. We give a complete study of the asymptotic behavior of a simple model of alignment of unit vectors, both at the level of particles, which corresponds to a system of coupled differential equations, and at the continuum level, under the form of an aggregation equation on the sphere. We prove unconditional convergence towards an aligned asymptotic state. In the cases of the differential system and of symmetric initial data for the partial differential equation, we provide precise rates of convergence.

Keywords: Alignment · Unit vectors · Aggregation equation

1 Introduction and Main Results

We are interested in a model of alignment of unit vectors. Our interest comes from the mechanism of alignment of self-propelled particles presented by Degond and Motsch in [9], which is a time-continuous model inspired from the Vicsek model [17] (in which the alignment process is discrete in time). In these models, the velocities of the particles, considered as unit vectors, try to align towards the average orientation of their neighbors and are subject to some angular noise. We want to study the simple case without spatial dependence and without noise. More precisely, at the level of the particle dynamics, we consider the deterministic part of the spatially homogeneous model of [6], which corresponds to a regularized version of [9]: the particles align with the average velocity of the others (instead of dividing this average vector by its norm to get a averaged orientation). It reads as

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = P_{v_i^\perp} J, \qquad \text{with } J = \frac{1}{N}\sum_{j=1}^{N} v_j, \tag{1}$$

where $(v_i)_{1\leqslant i\leqslant N}$ are N unit vectors belonging to $\mathbb{S}$, the unit sphere of $\mathbb{R}^n$, and $P_{v^\perp}$ is the projection on the orthogonal of a unit vector $v \in \mathbb{S}$, given

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 457–479, 2019.
https://doi.org/10.1007/978-3-030-15096-9_16

by $P_{v^\perp} u = u - (v \cdot u)v$ for $u \in \mathbb{R}^n$. This projection ensures that the velocities stay of norm one for all positive times. This system of equations can be seen as alignment towards the unit vector pointing in the same direction as J (the average of all velocities). Indeed the term $P_{v^\perp} J$ is equal to $\nabla_v (J \cdot v)$, where ∇_v is the gradient operator on the unit sphere $\mathbb{S}$. Therefore the dynamics of a particle following the equation $\frac{\mathrm{d}v}{\mathrm{d}t} = \nabla_v (v \cdot J)$ corresponds to the maximization of this quantity $v \cdot J$, which is maximal when v is aligned in the same direction as J.

At the kinetic level, we are interested in the evolution of a probability measure $f(t, \cdot)$ on $\mathbb{S}$ given by

$$\partial_t f + \nabla_v \cdot (f P_{v^\perp} J_f) = 0, \qquad \text{with } J_f = \int_{\mathbb{S}} v f \mathrm{d}v, \tag{2}$$

where $\nabla_v \cdot$ is the divergence operator on the sphere $\mathbb{S}$. The link between this evolution equation and the system of ordinary differential equations (1), is that if the measure f is the so-called empirical distribution of the particles $(v_i)_{1 \leqslant i \leqslant N}$, given by $f = \frac{1}{N} \sum_{i=1}^N \delta_{v_i}$, then it is a weak solution of the kinetic equation (2) if and only if the vectors $(v_i)_{1 \leqslant i \leqslant N}$ are solutions of the system (1) (see Remark 2). This kinetic equation (2) corresponds to the spatially homogeneous version of the mean-field limit of [6] in which the diffusion coefficient has been set to zero. The case with a positive diffusion has been treated in detail in [12] by the authors of the present paper, and it presents a phenomenon of phase transition: when the diffusion coefficient is greater than a precise threshold, all the solutions converge exponentially fast towards the uniform measure on the sphere $\mathbb{S}$, and when it is smaller, all solutions except those for which J_f is initially zero converge exponentially fast to a non-isotropic steady-state (a von Mises distribution). When the diffusion coefficient tends to zero, the von Mises distributions converge to Dirac measures concentrated at one point of $\mathbb{S}$. Therefore, we can expect that the solutions of (2) converge to a Dirac measure. The main object of this paper is to make this statement precise, in proving the following theorem:

Theorem 1. *Let f_0 be a probability measure on $\mathbb{S}$ of $\mathbb{R}^n$, and $f \in C(\mathbb{R}_+, \mathcal{P}(\mathbb{S}))$ be the solution of (2) with initial condition $f(0, v) = f_0(v)$.*

If $J_f(0) \neq 0$, then $t \mapsto |J_f(t)|$ is nondecreasing, so $\Omega(t) = \frac{J_f(t)}{|J_f(t)|} \in \mathbb{S}$ is well-defined for all times $t \geqslant 0$. Furthermore there exists $\Omega_\infty \in \mathbb{S}$ such that $\Omega(t)$ converges to Ω_∞ as $t \to +\infty$.

Finally, there exists a unique $v_{\mathrm{back}} \in \mathbb{S}$ such that the solution of the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp} J_f(t)$ with initial condition $v(0) = v_{\mathrm{back}}$ is such that $v(t) \to -\Omega_\infty$ as $t \to \infty$. Then, if we denote by m the mass of the singleton $\{v_{\mathrm{back}}\}$ with respect to the measure f_0, we have $m < \frac{1}{2}$ (which means that we cannot have too much mass at the "back") and $f(t, \cdot)$ converges weakly as $t \to \infty$ towards the measure $(1 - m)\delta_{\Omega_\infty} + m\delta_{-\Omega_\infty}$.

In particular, this theorem shows that if the initial condition f_0 has no atoms (or only one atom of mass bigger than one half) and satisfies $J_{f_0} \neq 0$, then the measure f converges weakly to a Dirac mass at some $\Omega_\infty \in \mathbb{S}$. Let us mention that there is no rate of convergence in this theorem. In general, there is no

hope to have such a rate for an arbitrary initial condition (see Proposition 6), but under regularity assumptions, one can expect to have an exponential rate of convergence (this is the case when the initial condition has some symmetries implying that $\Omega(t)$ is constant, see Proposition 7).

We will also study in detail the system of ordinary differential equations (1). Since this is a particular case of (2) in the case where $f = \frac{1}{N}\sum_{i=1}^{N}\delta_{v_i}$ (see Remark 2), we can apply the main theorem, but now the measure f has atoms, and actually we will see that working directly with the differential equations allows to have more precise results such as exponential rates of convergence. For instance the quantity $\Omega(t)$ plays the role as a nearly conserved quantity, as it converges to Ω_∞ at a higher rate than the convergence of the $(v_i)_{1\leqslant i\leqslant n}$. More precisely, we will prove the following theorem:

Theorem 2. *Given N positive real numbers $(m_i)_{1\leqslant i\leqslant N}$ with $\sum_{i=1}^{N} m_i = 1$, and N unit vectors $v_i^0 \in \mathbb{S}$ (for $1 \leqslant i \leqslant N$) such that $v_i^0 \neq v_j^0$ for all $i \neq j$, let $(v_i)_{1\leqslant i\leqslant N}$ be the solution of the following system of ordinary differential equations:*

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = P_{v_i^\perp} J, \text{ with } J(t) = \sum_{i=1}^{N} m_i v_i(t), \tag{3}$$

with the initial conditions $v_i(0) = v_i^0$ for $1 \leqslant i \leqslant N$, and where $P_{v_i^\perp}$ denotes the projection on the orthogonal of v_i.

If $J(0) \neq 0$, then $t \mapsto |J(t)|$ is nondecreasing, so $\Omega(t) = \frac{J(t)}{|J(t)|} \in \mathbb{S}$ is well-defined for all times $t \geqslant 0$. Furthermore there exists $\Omega_\infty \in \mathbb{S}$ such that $\Omega(t)$ converges to $\Omega_\infty(t)$ as $t \to +\infty$, and there are only two types of possible asymptotic regimes, which are described below.

(i) All the vectors v_i are converging to Ω_∞. Then this convergence occurs at an exponential rate 1, and Ω is converging to Ω_∞ at an exponential rate 3. More precisely, there exists $a_i \in \{\Omega_\infty\}^\perp \subset \mathbb{R}^n$, for $1 \leqslant i \leqslant N$ such that $\sum_{i=1}^{N} m_i a_i = 0$ and that, as $t \to +\infty$,

$$\begin{aligned} v_i(t) &= (1 - |a_i|^2 e^{-2t})\Omega_\infty + e^{-t} a_i + O(e^{-3t}) \quad \text{for } 1 \leqslant i \leqslant N, \\ \Omega(t) &= \Omega_\infty + O(e^{-3t}). \end{aligned}$$

(ii) There exists i_0 such that v_{i_0} converges to $-\Omega_\infty$. Then $m_{i_0} < \frac{1}{2}$ (once again, we cannot have too much mass on this "back" particle), and if we denote $\lambda = 1 - 2m_{i_0}$, the vector v_{i_0} converges to $-\Omega_\infty$ at an exponential rate 3λ. Furthermore, all the other vectors v_i for $i \neq i_0$ converge to Ω_∞ at a rate λ, and the vector Ω converges to Ω_∞ at a rate 3λ. More precisely, there exists $a_i \in \{\Omega_\infty\}^\perp \subset \mathbb{R}^n$, for $i \neq i_0$ such that $\sum_{i\neq i_0} m_i a_i = 0$ and that, as $t \to +\infty$,

$$\begin{aligned} v_i(t) &= (1 - |a_i|^2 e^{-2\lambda t})\Omega_\infty + e^{-\lambda t} a_i + O(e^{-3\lambda t}) \quad \text{for } i \neq i_0, \\ v_{i_0}(t) &= -\Omega_\infty + O(e^{-3\lambda t}), \\ \Omega(t) &= \Omega_\infty + O(e^{-3\lambda t}). \end{aligned}$$

Notice that the original system (1) can be put as (3) with $m_i = \frac{1}{N}$, but the assumption $v_i^0 \neq v_j^0$ for $i \neq j$ may not be satisfied. Up to renumbering particles and grouping those starting in the same position by setting $m_i = \frac{k}{N}$ where k is the number of particles sharing the same initial condition, we can always fall into the framework of (3) with distinct initial conditions. We can finally remark that this system (3) is still a particular case of the kinetic equation (2) for a measure given by $f = \sum_{i=1}^{N} m_i \delta_{v_i}$ (see once again Remark 2).

Let us conclude this introduction by saying that these models have also been introduced and studied in different contexts from the one of self-propelled particles. Alignment on the sphere has been introduced as a model of opinion formation in [3,7]. Let us also mention some more evolved consensus mechanisms on the sphere, such as with partial influence graphs [15]. The kinetic equation (2) with a diffusion term corresponds to the evolution of rodlike polymers with dipolar potential [10]. Finally the two-dimensional case, where $\mathbb{S}$ is the unit circle, can correspond to the evolution of identical Kuramoto oscillators. The results we present here were first exposed in detail (with the same proofs as in the present paper) by the first author in the CIMPA Summer School "Mathematical Modeling in Biology and Medicine" in June 2016. They are somewhat similar to those of [5] in dimension two, in the context of Kuramoto oscillators, a work that has been raised to us during the presentation of Bastien Fernandez in the workshop "Life Sciences" of the trimester "Stochastic Dynamics out of equilibrium" in May 2017. Very recently, a work [13] on generalization of Kuramoto oscillators in higher dimensions, the so-called Lohe oscillators, recovers the same kind of results, although not using exactly the same techniques and not obtaining the precise estimates of Theorem 2. The estimates given by Proposition 7 are also new, as far as we know.

This paper is divided in two main parts. After this introduction, Sect. 2 is devoted to the kinetic equation (2). It is divided in two subsections, the first one being dedicated to the proof of Theorem 1, and the second one giving more precise estimates of convergence in case of symmetries in the initial condition. Section 3 concerns the system of differential equations (3) and the proof of Theorem 2. Even if some conclusions can be drawn using Theorem 1 thanks to Remark 2, we try to make the two parts independent and the proofs self-contained, so the reader interested in Theorem 2 can directly jump to this last section.

2 The Continuum Model

2.1 Proof of Theorem 1

We start with a proposition about well-posedness of the kinetic equation (2). We proceed for instance as in [16]. We denote by $\mathcal{P}(\mathbb{S})$ the set of probability measures on $\mathbb{S}$. In this set we consider the Wasserstein distance W_1 (also called bounded Lipschitz distance) given by $W_1(\mu,\nu) = \inf_{\varphi \in \mathrm{Lip}_1(\mathbb{S})} \left| \int_{\mathbb{S}} \varphi \, \mathrm{d}\mu - \int_{\mathbb{S}} \varphi \, \mathrm{d}\nu \right|$ for μ and ν in $\mathcal{P}(\mathbb{S})$, where Lip_1 is the set of functions φ such that for all u, v

in $\mathbb{S}$, we have $|\varphi(u) - \varphi(v)| \leqslant |v - u|$. This distance corresponds to the weak convergence of probability measures: $W_1(\mu_n, \mu) \to 0$ if and only if for any continuous function $\varphi : \mathbb{S} \to \mathbb{R}$, we have $\int_{\mathbb{S}} \varphi \, \mathrm{d}\mu_n \to \int_{\mathbb{S}} \varphi \, \mathrm{d}\mu$. The well-posedness result is stated in the space $C(\mathbb{R}_+, \mathcal{P}(\mathbb{S}))$ of family of probability measures weakly continuous with respect to time:

Proposition 1. *Given $T > 0$ and $f_0 \in \mathcal{P}(\mathbb{S})$, there exists a unique weak solution $f \in C([0,T], \mathcal{P}(\mathbb{S}))$ to the Eq. (2) with initial condition f_0, in the sense that for all $t \in [0,T]$, and for all $\varphi \in C^1(\mathbb{S})$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{S}} \varphi(v) f(t,v) \, \mathrm{d}v = \int_{\mathbb{S}} J_{f(t,\cdot)} \cdot \nabla_v \varphi(v) f(t,v) \, \mathrm{d}v, \tag{4}$$

were we use the notation $f(t,v)\,\mathrm{d}v$ even if $f(t,\cdot)$ is not absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}$, and $J_{f(t,\cdot)} = \int_{\mathbb{S}} v f(t,v) \, \mathrm{d}v$.

Proof. Notice that the term $P_{v^\perp} J_f \cdot \nabla_v \varphi$ that we obtain when doing a formal integration by parts of (2) against a test function φ is replaced by $J_f \cdot \nabla_v \varphi$ in the weak formulation (4), since the gradient on the sphere at a point v is already orthogonal to v. The proof of this proposition relies on the fact that the linear equation corresponding to (2) when replacing J_f by an external given "alignment field" $\mathcal{J} \in C(\mathbb{R}_+, \mathbb{R}^n)$ is also well-posed. Indeed the solution to this linear equation, namely

$$\partial_t f + \nabla_v \cdot (P_{v^\perp} \mathcal{J}(t) f) = 0 \quad \text{with} \quad f(0,\cdot) = f_0, \tag{5}$$

is given by the image measure of f_0 by the flow Φ_t of the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp} \mathcal{J}(t)$. In detail, if Φ_t is the solution of

$$\begin{cases} \frac{\mathrm{d}\Phi_t}{\mathrm{d}t} = P_{\Phi_t^\perp} \mathcal{J}(t), \\ \Phi_0(v) = v, \end{cases} \tag{6}$$

then the solution $f(t,\cdot) = \Phi_t \# f_0$ is characterized by the fact that

$$\forall \varphi \in C(\mathbb{S}), \int_{\mathbb{S}} \varphi(v) f(t,v) \, \mathrm{d}v = \int_{\mathbb{S}} \varphi(\Phi_t(v)) f_0(v) \, \mathrm{d}v. \tag{7}$$

Since the differential equation (6) satisfies the assumptions for which the Cauchy-Lipschitz theorem applies, it is well-known (see for instance [1]) that the solution of (5) is unique and given by $\Phi_t \# f_0$.

Therefore, if, given $\mathcal{J} \in C([0,T], \mathbb{R}^n)$, we denote by $\Psi(\mathcal{J})$ the solution of the linear equation (5), solving the nonlinear kinetic equation (2) corresponds to finding a fixed point of the map $f \in C([0,T], \mathcal{P}(\mathcal{S})) \mapsto \Psi(J_f)$, or equivalently of the map $\mathcal{J} \in C([0,T], B) \mapsto J_{\Psi(\mathcal{J})}$, where B is the closed unit ball of $\mathbb{R}^n$ (recall that if $f \in \mathcal{P}(\mathbb{S})$, then $|J_f| \leqslant 1$). The space $E = C([0,T], B)$ is a complete metric space if the distance is given by $d_T(\mathcal{J}, \bar{\mathcal{J}}) = \sup_{t \in [0,T]} |\mathcal{J}(t) - \bar{\mathcal{J}}(t)| e^{-\beta t}$, for an arbitrary $\beta > 0$. Using the fact that $|(P_{v^\perp} - P_{\bar{v}^\perp})u| \leqslant 2|v - \bar{v}|$ if $|u| \leqslant 1$, by a

simple Grönwall estimate, if $\mathcal{J}, \bar{\mathcal{J}} \in E$ and $\Phi_t, \bar{\Phi}_t$ are the associated flow given by (6), we obtain

$$|\Phi_t - \bar{\Phi}_t| \leqslant \int_0^t |\mathcal{J}(s) - \bar{\mathcal{J}}(s)| e^{2(t-s)} \mathrm{d}s.$$

Finally, we get (using the notation $J_f(t) = J_{f(t,\cdot)}$)

$$\begin{aligned} |J_{\Psi(\mathcal{J})}(t) - J_{\Psi(\bar{\mathcal{J}})}(t)| &= \left| \int_{\mathbb{S}} v\, \Psi(\mathcal{J})(t,v)\, \mathrm{d}v - \int_{\mathbb{S}} v\, \Psi(\bar{\mathcal{J}})(t,v)\, \mathrm{d}v \right| \\ &= \left| \int_{\mathbb{S}} [\Phi_t(v) - \bar{\Phi}_t(v)] f_0(v)\, \mathrm{d}v \right| \\ &\leqslant \int_0^t |\mathcal{J}(s) - \bar{\mathcal{J}}(s)| e^{2(t-s)} \mathrm{d}s \leqslant d_t(\mathcal{J}, \bar{\mathcal{J}}) \int_0^t e^{2(t-s)+\beta s} \mathrm{d}s. \end{aligned}$$

Therefore when $\beta > 2$ we get $|J_{\Psi(\mathcal{J})}(t) - J_{\Psi(\bar{\mathcal{J}})}(t)| e^{-\beta t} \leqslant \frac{1}{\beta - 2} d_t(\mathcal{J}, \bar{\mathcal{J}})$, so if we take $\beta > 3$, we get that the map $\mathcal{J} \mapsto J_{\Psi(\mathcal{J})}$ is indeed a contraction mapping from E to E, which gives the existence and uniqueness of the fixed point. □

Remark 1. The well-posedness of the kinetic equation (2) can also be established in Sobolev spaces, by means of harmonic analysis on the sphere and standard Galerkin method (see [12]).

Remark 2. Using the weak formulation (4) and the definition of the pushforward measure (7), it is possible to show that a convex combination of Dirac masses, of the form $f(t,\cdot) = \sum_{i=1}^N m_i \delta_{v_i}(t)$ with $m_i \geqslant 0$ for $1 \leqslant i \leqslant N$ and $\sum_{i=1}^N m_i = 1$ is a weak solution of (2) if and only if the $(v_i)_{1 \leqslant i \leqslant N}$ are solutions of the system of differential equations (3).

We are now ready to prove some qualitative properties of the solution to the kinetic equation (2). Without further notice, we will denote by f this solution, and by Φ_t the flow (6) associated to $\mathcal{J} = J_f$. The first property is a simple lemma related to the monotonicity of $|J_f|$.

Lemma 1. *If f is a solution of (2), then $|J_f|$ is nondecreasing in time. Therefore if $J_{f_0} \neq 0$, the "average orientation" $\Omega(t) = \frac{J_f(t)}{|J_f(t)|}$ is well defined and smooth. Furthermore its time derivative $\dot{\Omega}$ tends to 0 as $t \to \infty$.*

Proof. Notice that if $J_{f_0} = 0$, then $f(t,\cdot) = f_0$ for all t. To compute the evolution of J_f, we use (4) with $\varphi(v) = v \cdot e$ for an arbitrary vector e in $\mathbb{R}^n$. We obtain, using the fact that $\nabla_v (v \cdot e) = P_{v^\perp} e$:

$$e \cdot \frac{\mathrm{d}J_f}{\mathrm{d}t} = J_f \cdot \int_{\mathbb{S}} P_{v^\perp} e f(t,v)\, \mathrm{d}v = e \cdot M_f J_f,$$

where M_f is the matrix given by $\int_{\mathbb{S}} P_{v^\perp} f(t,v)\, \mathrm{d}v$ (it is a symmetric matrix with eigenvalues in $[0,1]$, as convex combination of orthogonal projections). Since M_f

is continuous in time, then J_f is C^1, and by the same procedure we can compute the evolution of M_f, which will depend on higher moments of f, to get that J_f is smooth. More precisely, since any moment is uniformly bounded (the sphere is compact and $f(t,\cdot)$ is a probability density for all t), we get that all derivatives of J_f are uniformly bounded in time. Since

$$\frac{1}{2}\frac{\mathrm{d}|J_f|^2}{\mathrm{d}t} = J_f \cdot M_f J_f = \int_{\mathbb{S}} [|J_f|^2 - (v \cdot J_f)^2] f(t,v)\,\mathrm{d}v \geqslant 0,$$

we get the first part of the proposition.

From now on we suppose that $J_{f_0} \neq 0$, therefore $\Omega(t)$ is well defined. The function $\frac{1}{2}\frac{\mathrm{d}|J_f|^2}{\mathrm{d}t} = |J_f|^2 \Omega \cdot M_f \Omega$ being nonnegative, smooth, integrable in $\mathbb{R}_+$ (since $|J_f|$ is bounded by 1), and with bounded derivative, it is a classical exercise to show that it must converge to 0 as $t \to \infty$ (this is known as Barbălat's Lemma, see [4]). This gives us that $\Omega \cdot M_f \Omega \to 0$ as $t \to \infty$. Let us now compute the evolution of Ω. We get

$$\dot{\Omega} = \frac{1}{|J_f|}\frac{\mathrm{d}J_f}{\mathrm{d}t} - \frac{\mathrm{d}|J_f|}{\mathrm{d}t}\frac{J_f}{|J_f|^2} = M_f\Omega - (\Omega \cdot M_f \Omega)\Omega = P_{\Omega^\perp}(M_f \Omega). \tag{8}$$

Since M_f has eigenvalues in $[0,1]$, we get that $|M_f\Omega|^2 = \Omega \cdot M_f^2 \Omega \leqslant \Omega \cdot M_f \Omega$, therefore $M_f\Omega \to 0$ as $t \to 0$. So we get that $\dot{\Omega} \to 0$ as $t \to \infty$. □

Remark 3. The fact that $|J_f|$ is nondecreasing can be enlightened by the theory of gradient flow in probability spaces [2]. Indeed, the kinetic equation (2) corresponds to the gradient flow of the functional $-\frac{1}{2}|J_f|^2$ for the Wasserstein distance W_2. Therefore the evolution amounts to minimizing in time this quantity. We also remark that since $|J_f|$ is nondecreasing, by an appropriate change of time, we can recover the equation $\partial_t f + \nabla_v \cdot (f P_{v^\perp} \Omega)$ which corresponds to the spatial homogeneous version of [9] without noise. This equation can also be interpreted as a gradient flow [11].

The fact that $\dot{\Omega} \to 0$ is not sufficient to prove that Ω converges to some Ω_∞, we would need $\dot{\Omega} \in L^1(\mathbb{R}_+)$ and we only have up to now $\dot{\Omega} \in L^2(\mathbb{R}_+)$ (since we have seen in the proof of Lemma 1 that $|J_f|^2 \Omega \cdot M_f \Omega$ is integrable in time). To fill this gap, one solution is to compute the second derivative of Ω, and more precisely, to obtain an estimate on $|\dot{\Omega}|$ corresponding to the assumption of the following lemma, which mainly says that if g is integrable, then any bounded solution of the differential equation $y' = y + g$ has to be integrable.

Lemma 2. *Let $y : \mathbb{R}_+ \to \mathbb{R}$ be a nonnegative function such that y^2 is C^1 and bounded. We suppose that there exists a function $g \in L^1(\mathbb{R}_+)$ such that for all $t \in \mathbb{R}$, we have*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} y^2 = y^2 + y\,g. \tag{9}$$

Then $y \in L^1(\mathbb{R}_+)$.

Proof. Let $t \geqslant 0$ such that $y(t) > 0$. We set $T = \sup\{s \geqslant t, y > 0 \text{ on } [t,s]\}$ (we may have $T = +\infty$).

We have that y is C^1, positive and bounded on $[t,T)$, and satisfies the differential equation $y' = y+g$, therefore by Duhamel's formula we have, for $s \in [t,T)$:

$$y(s)e^{-s} - y(t)e^{-t} = \int_t^s g(u)e^{-u}\mathrm{d}u.$$

Letting $s = T$ (resp. $s \to +\infty$ if $T = +\infty$), since $y(T) = 0$ (resp. y is bounded), we obtain

$$y(t) = -\int_t^T g(u)e^{t-u}\mathrm{d}u \leqslant \int_t^\infty |g(u)|e^{t-u}\mathrm{d}u.$$

This equality being true for any $t \in \mathbb{R}_+$ (even if $y(t) = 0$), we have by Fubini's theorem that

$$\int_0^\infty y(t)\mathrm{d}t \leqslant \int_0^\infty \int_t^\infty |g(u)|e^{t-u}\mathrm{d}u\,\mathrm{d}t = \int_0^\infty |g(u)|(1-e^{-u})\mathrm{d}u,$$

which is finite by integrability of g. □

We are now ready to prove the convergence of Ω.

Proposition 2. *If $J_{f_0} \neq 0$, then $\dot{\Omega} \in L^1(\mathbb{R}_+)$, and therefore there exists $\Omega_\infty \in \mathbb{S}$ such that $\Omega \to \Omega_\infty$ as $t \to \infty$.*

Proof. We first compute the derivative of M_f. For convenience, we use the notation $\langle\varphi(v)\rangle_f$ for $\int_{\mathbb{S}} \varphi(v)f(t,v)\,\mathrm{d}v$. Therefore we have $J_f = \langle v\rangle_f$ and $M_f = \langle P_{v^\perp}\rangle_f$, and the weak formulation (4) reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\varphi(v)\rangle_f = J_f \cdot \langle\nabla_v\varphi(v)\rangle_f.$$

We have, for fixed $e_1, e_2 \in \mathbb{R}^n$:

$$e_1 \cdot M_f e_2 = \langle e_1 \cdot P_{v^\perp} e_2\rangle_f = e_1 \cdot e_2 - \langle(e_1\cdot v)(e_2\cdot v)\rangle_f.$$

Therefore, since $\nabla_v(e\cdot v) = P_{v^\perp}e$, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(e_1 \cdot M_f e_2) &= -J_f \cdot \langle(e_2\cdot v)P_{v^\perp}e_1 + (e_1\cdot v)P_{v^\perp}e_2\rangle_f \\ &= e_1 \cdot [-\langle(e_2\cdot v)P_{v^\perp}J_f\rangle_f + \langle J_f \cdot P_{v^\perp}e_2\, v\rangle_f],\end{aligned}$$

so the term in between the brackets is the derivative of $M_f e_2$. We then get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(M_f\Omega) &= M_f\dot{\Omega} - |J_f|\langle(\Omega\cdot v)P_{v^\perp}\Omega\rangle_f - |J_f|\langle\Omega\cdot P_{v^\perp}\Omega\, v\rangle_f \\ &= M_f\dot{\Omega} + 2|J_f|\langle(\Omega\cdot v)^2 v\rangle_f - |J_f|[\langle(\Omega\cdot v)\Omega + v\rangle_f] \\ &= M_f\dot{\Omega} + 2|J_f|\langle(\Omega\cdot v)^2 v\rangle_f - 2|J_f|^2\Omega. \qquad (10)\end{aligned}$$

Thanks to (8), we finally have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\dot{\Omega} &= \frac{\mathrm{d}}{\mathrm{d}t}(M_f\Omega) - (\Omega\cdot M_f\Omega)\dot{\Omega} - (\dot{\Omega}\cdot M_f\Omega)\Omega - \Omega\cdot\frac{\mathrm{d}}{\mathrm{d}t}(M_f\Omega)\,\Omega\\ &= P_{\Omega^\perp}\frac{\mathrm{d}}{\mathrm{d}t}(M_f\Omega) - (\Omega\cdot M_f\Omega)\dot{\Omega} - (\dot{\Omega}\cdot M_f\Omega)\Omega.\end{aligned}$$

Since Ω and $\dot{\Omega}$ are orthogonal, we have some simplifications by taking the dot product with $\dot{\Omega}$ and using (10):

$$\begin{aligned}\dot{\Omega}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\dot{\Omega} &= \dot{\Omega}\cdot\frac{\mathrm{d}}{\mathrm{d}t}(M_f\Omega) - (\Omega\cdot M_f\Omega)|\dot{\Omega}|^2.\\ &= \dot{\Omega}\cdot M_f\dot{\Omega} - 2|J_f|[\langle(\Omega\cdot v)^2\,\dot{\Omega}\cdot v\rangle_f] - (\Omega\cdot M_f\Omega)|\dot{\Omega}|^2\\ &= |\dot{\Omega}|^2 - \langle(\dot{\Omega}\cdot v)^2\rangle_f - (\Omega\cdot M_f\Omega)|\dot{\Omega}|^2 - 2|J_f|[\langle(\Omega\cdot v)^2\,\dot{\Omega}\cdot v\rangle_f]. \quad (11)\end{aligned}$$

If we define u to be the unit vector $\frac{\dot{\Omega}}{|\dot{\Omega}|}$ when $|\dot{\Omega}|\neq 0$ and to be zero if $|\dot{\Omega}|=0$, and we set

$$g(t) = -|\dot{\Omega}|[\langle(u\cdot v)^2\rangle_f + (\Omega\cdot M_f\Omega)] - 2|J_f|\langle(\Omega\cdot v)^2)u\cdot v\rangle_f, \quad (12)$$

we get that the formula (11) is written under the following form, corresponding to (9) with $y=|\dot{\Omega}|$:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\dot{\Omega}|^2 = |\dot{\Omega}|^2 + |\dot{\Omega}|g(t).$$

Our goal is to show that $g\in L^1(\mathbb{R}_+)$ in order to apply Lemma 2. Indeed, thanks to (8), we have that $|\dot{\Omega}|\leqslant 1$ (recall that M_f is a symmetric matrix with eigenvalues in $[0,1]$), and $|\dot{\Omega}|^2$ is C^1.

As was remarked before in the proof of Lemma 1, the quantity $|J_f|^2\Omega\cdot M_f\Omega$ is integrable in time, which gives that $\Omega\cdot M_f\Omega = \langle 1-(\Omega\cdot v)^2\rangle_f$ is integrable. Since u is colinear to $\dot{\Omega}$, which is orthogonal to Ω, we have that $P_{\Omega^\perp}u=u$, and therefore we get (using the fact that $|u|\leqslant 1$, since $|u|$ is 1 or 0)

$$\langle(u\cdot v)^2\rangle_f = \langle(u\cdot P_{\Omega^\perp}v)^2\rangle_f \leqslant \langle|P_{\Omega^\perp}v|^2\rangle_f = \langle 1-(\Omega\cdot v)^2\rangle_f.$$

This gives that the first term in the definition (12) of g is integrable in time. Finally, since $u\cdot\Omega=0$, we have that $\langle u\cdot v\rangle_f=0$, and we get

$$|\langle(\Omega\cdot v)^2\,u\cdot v\rangle_f| = |\langle(1-(\Omega\cdot v)^2)u\cdot v\rangle_f| \leqslant \langle 1-(\Omega\cdot v)^2\rangle_f,$$

since $1-(\Omega\cdot v)^2\geqslant 0$ and $|u\cdot v|\leqslant 1$ for all $v\in\mathbb{S}$. This gives that the last term in the definition (12) of g is also integrable in time. In virtue of Lemma 2, we then get that $|\dot{\Omega}|$ is integrable. Therefore $\Omega(t)=\Omega(0)+\int_0^t\dot{\Omega}(s)\mathrm{d}s$ converges as $t\to+\infty$. □

In order to control the distance between f and δ_{Ω_∞}, we now need to understand the properties of the flow of the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp}J_f$.

Proposition 3. *Let $\mathcal{J}$ be a continuous function $\mathbb{R}_+ \to \mathbb{R}^n$ such that $t \mapsto |\mathcal{J}(t)|$ is positive, bounded and nondecreasing, and $\Omega(t) = \frac{\mathcal{J}(t)}{|\mathcal{J}(t)|}$ converges to $\Omega_\infty \in \mathbb{S}$ as $t \to \infty$.*

Then there exists a unique $v_{\text{back}} \in \mathbb{S}$ such that the solution of the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp}\mathcal{J}$ with initial condition $v(0) = v_{\text{back}}$ satisfies $v(t) \to -\Omega_\infty$ as $t \to +\infty$. Furthermore, for all $v_0 \neq v_{\text{back}}$, the solution of this differential equation with initial condition $v(0) = v_0$ converges to Ω_∞ as $t \to +\infty$.

Proof. The outline of the proof is the following: we first show that any solution satisfies either $v(t) \to -\Omega_\infty$ or $v(t) \to \Omega_\infty$, then we construct v_{back}, and finally we prove that it is unique. We still denote by Φ_t the flow of the differential equation (6).

We first notice that $|\mathcal{J}(t)|$ converges to some $\lambda > 0$, therefore $\mathcal{J}(t)$ converges to $\lambda\Omega_\infty$ as $t \to \infty$. Therefore the solution of the equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp}\mathcal{J}$ with initial condition $v(0) = v_0$ is also the solution of a differential equation of the form

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \lambda P_{v^\perp}\Omega_\infty + r_{v_0}(t), \tag{13}$$

where the remainder term $r_{v_0}(t)$ converges to 0 as $t \to \infty$, uniformly in $v_0 \in \mathbb{S}$. Let us suppose that $v(t)$ does not converge to $-\Omega_\infty$ (that is to say $v(t) \cdot \Omega_\infty$ does not converge to -1), and let us prove that in this case $v(t) \to \Omega_\infty$. Taking the dot product with Ω_∞ in (13), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(v \cdot \Omega_\infty) = \lambda[1 - (v \cdot \Omega_\infty)^2] + \Omega_\infty \cdot r_{v_0}(t), \tag{14}$$

so we can use a comparison principle with the one-dimensional differential equation $y' = \lambda(1 - y^2) - \varepsilon$. Since $\lambda(1 - y^2) - \varepsilon$ is positive for $|y| < \sqrt{1 - \frac{\varepsilon}{\lambda}}$ and negative for $|y| > \sqrt{1 - \frac{\varepsilon}{\lambda}}$, any solution starting with $y(t_0) > -\sqrt{1 - \frac{\varepsilon}{\lambda}}$ converges to $\sqrt{1 - \frac{\varepsilon}{\lambda}}$ as $t \to +\infty$. Since $v(t) \cdot \Omega_\infty$ does not converge to -1, there exists $\delta > 0$ such that $v(t) \cdot \Omega_\infty > -1 + \delta$ for arbitrarily large times t. For any $\varepsilon > 0$ sufficiently small (such that $-\sqrt{1 - \frac{\varepsilon}{\lambda}} < -1 + \delta$), there exists $t_0 \geqslant 0$ such that $v(t_0) \cdot \Omega_\infty > -1 + \delta$ and $|\Omega_\infty \cdot r_{v_0}(t)| \leqslant \varepsilon$ for all $t \geqslant t_0$. By comparison principle, we then get that $\liminf_{t \to +\infty} v(t) \cdot \Omega_\infty \geqslant \sqrt{1 - \frac{\varepsilon}{\lambda}}$. Since this is true for any $\varepsilon > 0$ sufficiently small, we then get that $v(t) \cdot \Omega_\infty$ converges to 1, that is to say $v(t) \to \Omega_\infty$ as $t \to +\infty$.

Let us now prove that if $v(t)$ converges to Ω_∞, then there exists a neighborhood of v_0 such that the convergence to Ω_∞ of solutions starting in this neighborhood is uniform in time. This is done thanks to the same comparison principle. We fix $\delta > 0$ and $\varepsilon > 0$ such that $-1 + \delta > -\sqrt{1 - \frac{\varepsilon}{\lambda}}$. We take $t_0 \geqslant 0$ such that $v(t_0) \cdot \Omega_\infty > -1 + \delta$ and $|\Omega_\infty \cdot r_{\tilde{v}_0}(t)| \leqslant \varepsilon$ for any $\tilde{v}_0 \in \mathbb{S}$ and $t \geqslant t_0$. By continuity of the flow of the equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp}\mathcal{J}$, there exists a neighborhood B of v_0 in $\mathbb{S}$ such that for any $\tilde{v}_0 \in B$, the solution $\tilde{v}(t) = \Phi_t(\tilde{v}_0)$ of this equation with initial condition $\tilde{v}_0$ satisfies $\tilde{v}(t_0) \cdot \Omega_\infty > -1 + \delta$. We now look at the equation $y' = \lambda(1 - y^2) - \varepsilon$ starting with $y(t_0) = -1 + \delta$, which converges to $\sqrt{1 - \frac{\varepsilon}{\lambda}} > 1 - \delta$. There exists T such that $y(t) \geqslant 1 - \delta$ for all $t \geqslant T$. Therefore,

by comparison principle with (14) (where v_0 is replaced by $\tilde{v}_0$), we get that for all $\tilde{v}_0 \in B$, the solution $\tilde{v}$ satisfies $\tilde{v}(t) \cdot \Omega_\infty \geqslant 1 - \delta$ for all $t \geqslant T$.

We are now ready to construct v_{back}. We take (t_n) a sequence of increasing times such that $t_n \to +\infty$ and define v^n_{back} as the solution at time $t = 0$ of the backwards in time differential equation $\frac{\mathrm{d}v^n}{\mathrm{d}t} = P_{(v^n)^\perp}\mathcal{J}$ with terminal condition $v^n(t_n) = -\Omega_\infty$, that is to say $v^n_{\text{back}} = \Phi^{-1}_{t_n}(-\Omega_\infty)$. Up to extracting a subsequence, we can assume that v^n_{back} converges to some $v_{\text{back}} \in \mathbb{S}$ and we set $v(t) = \Phi_t(v_{\text{back}})$. By the first part of the proof, we have that either $v(t) \to \Omega_\infty$ or $v(t) \to -\Omega_\infty$ as $t \to +\infty$. The first case is incompatible with the uniform convergence in time. Indeed, in that case, we would have a neighborhood B of v_{back} and a time T such that for all $t \geqslant T$ and all $\tilde{v} \in B$, $\Phi_t(\tilde{v}) \cdot \Omega_\infty \geqslant 0$ (by taking $\delta = 1$ in the previous paragraph). Since we can take n such that $t_n \geqslant T$ and $v^n_{\text{back}} \in B$, this is in contradiction with the fact that $\Phi_{t_n}(v^n_{\text{back}}) = -\Omega_\infty$.

It remains to prove that v_{back} is unique (which implies that $\Phi_t^{-1}(-\Omega_\infty)$ actually converges to v_{back} as $t \to +\infty$, thanks to the previous paragraph). This is due to a phenomenon of repulsion of two solutions $v(t)$ and $\tilde{v}(t)$ when they are close to $-\Omega(t)$. Indeed, they satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t} v \cdot \tilde{v} = v \cdot P_{\tilde{v}^\perp}\mathcal{J} + \tilde{v} \cdot P_{v^\perp}\mathcal{J} = \mathcal{J} \cdot (v + \tilde{v})(1 - v \cdot \tilde{v}),$$

which can be written, since $\|v - \tilde{v}\|^2 = 2\,(1 - v \cdot \tilde{v})$ as

$$\frac{\mathrm{d}}{\mathrm{d}t} \|v - \tilde{v}\|^2 = \gamma(t) \|v - \tilde{v}\|^2, \tag{15}$$

where $\gamma(t) = -\mathcal{J}(t) \cdot (v(t) + \tilde{v}(t))$. Let us suppose that both $v(t) = \Phi_t(v_0)$ and $\tilde{v}(t) = \Phi_t(\tilde{v}_0)$ converge to $-\Omega_\infty$ as $t \to +\infty$. Since $\mathcal{J}(t) \to \lambda\Omega_\infty$ as $t \to +\infty$, we have $\gamma(t) \to 2\lambda > 0$ as $t \to +\infty$. Therefore the only bounded solution of the linear differential equation (15) is the constant 0, therefore we have $v = \tilde{v}$, and thus $v_0 = \tilde{v}_0$. □

We are now ready to prove the last part of Theorem 1.

Proposition 4. *Let v_{back} be given by Proposition 3 with $\mathcal{J} = J_f$ (we suppose $J_{f_0} \neq 0$). We denote by $m = \int_{\mathbb{S}} \mathbf{1}_{v = v_{\text{back}}} f_0(v)\mathrm{d}v$ the initial mass of $\{v_{\text{back}}\}$. Then $m < \frac{1}{2}$ and $W_1(f(t,\cdot), (1-m)\delta_{\Omega_\infty} + m\delta_{-\Omega_\infty}) \to 0$ as $t \to +\infty$.*

Proof. We write $f_\infty = (1-m)\delta_{\Omega_\infty} + m\delta_{-\Omega_\infty}$. Let $\varphi \in \mathrm{Lip}_1(\mathbb{S})$. We have

$$\begin{aligned}\int_{\mathbb{S}} \varphi(v) f_\infty(v)\,\mathrm{d}v &= m\varphi(-\Omega_\infty) + (1-m)\varphi(\Omega_\infty) \\ &= m\varphi(-\Omega_\infty) + \int_{\mathbb{S}} \mathbf{1}_{v \neq v_{\text{back}}} \varphi(\Omega_\infty) f_0(v)\,\mathrm{d}v,\end{aligned}$$

and $\int_{\mathbb{S}} \varphi(v) f(t,v)\,\mathrm{d}v = \int_{\mathbb{S}} \varphi(\Phi_t(v)) f_0(v)\,\mathrm{d}v$ (recall that $f(t,\cdot) = \Phi_t \# f_0$ is characterized by (7), where Φ_t, defined in (6) is the flow of the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = P_{v^\perp}\mathcal{J}$). Therefore we get

$$\int_{\mathbb{S}} \varphi(v) f(t,v)\mathrm{d}v = m\varphi(\Phi_t(v_{\text{back}})) + \int_{\mathbb{S}} \mathbf{1}_{v \neq v_{\text{back}}} \varphi(\Phi_t(v)) f_0(v)\,\mathrm{d}v. \tag{16}$$

We then obtain

$$\begin{aligned}\Big|\int_{\mathbb{S}}\varphi(v)f(t,v)\,\mathrm{d}v-\int_{\mathbb{S}}\varphi(v)f_\infty(v)\,\mathrm{d}v\Big|&\\ \leqslant m|\varphi(\Phi_t(v_{\mathrm{back}}))-\varphi(-\Omega_\infty)|+\int_{\mathbb{S}}\mathbf{1}_{v\neq v_{\mathrm{back}}}|\varphi(\Phi_t(v))-\varphi(\Omega_\infty)|f_0(v)\,\mathrm{d}v&\\ \leqslant m|\Phi_t(v_{\mathrm{back}})+\Omega_\infty|+\int_{\mathbb{S}}\mathbf{1}_{v\neq v_{\mathrm{back}}}|\Phi_t(v)-\Omega_\infty|f_0(v)\,\mathrm{d}v,&\end{aligned}$$

since $\varphi\in\mathrm{Lip}_1(\mathbb{S})$. We finally get

$$W_1(f(t,\cdot),f_\infty)\leqslant m|\Phi_t(v_{\mathrm{back}})+\Omega_\infty|+\int_{\mathbb{S}}\mathbf{1}_{v\neq v_{\mathrm{back}}}|\Phi_t(v)-\Omega_\infty|f_0(v)\,\mathrm{d}v. \quad (17)$$

Now, by Proposition 3, as $t\to+\infty$ we have $\Phi_t(v)\to\Omega_\infty$ for all $v\neq v_{\mathrm{back}}$, and $\Phi_t(v_{\mathrm{back}})\to-\Omega_\infty$. Therefore by the dominated convergence theorem, the estimate (17) gives that $W_1(f(t,\cdot),f_\infty)\to 0$ as $t\to+\infty$. It remains to prove that $m>\frac{1}{2}$, which comes from Proposition 2, which gives that $\frac{J_f}{|J_f|}\to\Omega_\infty$ as $t\to+\infty$. Indeed, applying (16) with $\varphi(v)=v$, we get

$$J_f(t)=m\Phi_t(v_{\mathrm{back}})+\int_{\mathbb{S}}\mathbf{1}_{v\neq v_{\mathrm{back}}}\Phi_t(v)f_0(v)\,\mathrm{d}v,$$

which gives by dominated convergence that, as $t\to+\infty$, we have

$$J_f(t)\to-m\Omega_\infty+\int_{\mathbb{S}}\mathbf{1}_{v\neq v_{\mathrm{back}}}\Omega_\infty f_0(v)\,\mathrm{d}v=(1-2m)\Omega_\infty.$$

Since $\frac{J_f(t)}{|J_f(t)|}\to\Omega_\infty$ as $t\to+\infty$, we get $1-2m>0$. □

2.2 Symmetries and Rates of Convergence

This subsection is dedicated to the study of rates of convergence, based on somewhat explicit solutions in the case where Ω is constant in time, which is the case when the initial condition has some symmetries.

Proposition 5. *Let G be a group of orthogonal transformations under which f_0 is invariant (that is to say $f_0\circ g=f_0$ and all $g\in G$) and such that the only fixed points on $\mathbb{S}$ of every element of G are two opposite unit vectors that we call $\pm e_n$. Then the solution $f(t,\cdot)$ of the partial differential equation (2) is also invariant under all elements of g. Furthermore if $J_{f_0}\neq 0$, then $J_f(t)=\alpha(t)e_n$ with α positive (up to exchanging e_n and $-e_n$), and $\Omega(t)$ is constantly equal to e_n.*

Proof. The first part of the proposition comes from the fact that $t\mapsto f(t,\cdot)\circ g$ is also a solution of (2) (which is well-posed) with the same initial condition. Then, we have by invariance that $gJ_{f(t,\cdot)}=\int_{\mathbb{S}}gvf_0(v)\,\mathrm{d}v=\int_{\mathbb{S}}gvf_0(gv)\,\mathrm{d}v=J_{f(t,\cdot)}$, for all $g\in G$, and therefore $\Omega(t)$ is a fixed point of every element of g and must be equal to $\pm e_n$. □

Let us mention two simple examples of these kind of symmetries: when $f_0(v)$ only depends on $v \cdot e_n$ (G is then the set of isometries having e_n as fixed point), or when $f(\sin\theta w + \cos\theta e_n) = f(-\sin\theta w + \cos\theta e_n)$ (G is reduced to identity and to $v \mapsto 2e_n \cdot v\, e_n - v$).

Let us now do some preliminary computations in the case where Ω is constant in time. We work in an orthogonal base $(e_1, \dots, e_n)$ of $\mathbb{R}^n$ for which $\Omega = e_n$ is the last vector, and we write write $J_f(t) = \alpha(t)e_n$, with $t \mapsto \alpha(t)$ positive and nondecreasing. We will use the stereographic projection

$$\begin{array}{rccc} s: & \mathbb{S} \setminus \{-e_n\} & \to & \mathbb{R}^{n-1} \\ & v & \mapsto & s(v) = \frac{1}{1+v\cdot e_n} P_{e_n^\perp} v, \end{array} \tag{18}$$

where we identify $P_{e_n^\perp} v$ with its first $n-1$ coordinates. This is a diffeomorphism between $\mathbb{S} \setminus \{-e_n\}$ and $\mathbb{R}^{n-1}$, and its inverse is given by

$$\begin{array}{rccc} p: & \mathbb{R}^{n-1} & \to & \mathbb{S} \setminus \{-e_n\} \subset \mathbb{R}^{n-1} \times \mathbb{R} \\ & z & \mapsto & p(z) = (\frac{2}{1+|z|^2}\, z, \frac{1-|z|^2}{1+|z|^2}). \end{array} \tag{19}$$

If φ is an integrable function on $\mathbb{S}$, the change of variable for this diffeomorphism reads

$$\int_{\mathbb{S}} \varphi(v)\, \mathrm{d}v = c_n^{-1} \int_{\mathbb{R}^{n-1}} \frac{\varphi(p(z))}{(1+|z|^2)^{n-1}}\, \mathrm{d}z, \tag{20}$$

where the normalization constant is $c_n = \int_{\mathbb{R}^{n-1}} \frac{\mathrm{d}z}{(1+|z|^2)^{n-1}}$. If v is a solution to the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = \alpha(t) P_{v^\perp} e_n$ with $v \neq -e_n$, a simple computation shows that $z = s(v)$ satisfies the differential equation $\frac{\mathrm{d}z}{\mathrm{d}t} = -\alpha(t) z$. Therefore, if we write $\lambda(t) = \int_0^t \alpha(\tau)\, \mathrm{d}\tau$, we have an explicit expression for the solution f of the aggregation equation (5): the pushforward formula (7) is given, when f_0 has no atom at $-e_n$, by

$$\forall \varphi \in C(\mathbb{S}), \int_{\mathbb{S}} \varphi(v) f(t,v)\, \mathrm{d}v = c_n^{-1} \int_{\mathbb{R}^{n-1}} \frac{\varphi(p(ze^{-\lambda(t)})) f_0(p(z))}{(1+|z|^2)^{n-1}}\, \mathrm{d}z. \tag{21}$$

In particular, we have

$$\begin{aligned} 1 - \alpha(t) = 1 - J_f(t) \cdot e_n &= \int_{\mathbb{S}} (1 - v \cdot e_n) f(t,v)\, \mathrm{d}v \\ &= c_n^{-1} \int_{\mathbb{R}^{n-1}} \frac{2|z|^2 e^{-2\lambda(t)} f_0(p(z))}{(1+|z|^2 e^{-2\lambda(t)})(1+|z|^2)^{n-1}}\, \mathrm{d}z. \end{aligned} \tag{22}$$

We are now ready to state the first proposition regarding the rate of convergence towards Ω_∞: in the framework of Theorem 1, there is no hope to have a rate of convergence of $f(t,\cdot)$ with respect to the W_1 distance without further assumption on the regularity of f_0, even if it has no atoms (in this case $f(t,\cdot) \to \delta_{\Omega_\infty}$ as $t \to +\infty$). More precisely the following proposition gives the construction of a solution decaying arbitrarily slowly to δ_{Ω_∞}, in contrast with results of local stability of Dirac masses for other models of alignment on the sphere [8], for which as long as the initial condition is close enough to δ_{Ω_∞}, the solution converges exponentially fast in Wasserstein distance.

Proposition 6. *Given a smooth decreasing function $t \mapsto g(t)$ converging to 0 (slowly) as $t \mapsto +\infty$, and such that $g(0) < \frac{1}{2}$, there exists a probability density function f_0 such that the solution $f(t,\cdot)$ of (2) converges weakly to δ_{Ω_∞}, but such that $W_1(f(t,\cdot),\delta_{\Omega_\infty}) \geqslant g(t)$ for all $t \geqslant 0$.*

Proof. We will construct f_0 as a function of the form $f_0(v) = h(|s(v)|)$, where the stereographic projection s is defined in (18). Let us prove that the following choice of h works, for $\varepsilon > 0$ sufficiently small:

$$h(r) = b_n \frac{(1+r^2)^{n-1}}{r^{n-2}} \left[\frac{1-g(0)}{\varepsilon} \mathbf{1}_{0<r<\varepsilon} - \frac{g'(\ln r)}{r} \mathbf{1}_{r\geqslant 1} \right],$$

where the normalization constant is $b_n = \int_{\mathbb{R}_+} \frac{r^{n-2}\,\mathrm{d}r}{(1+r^2)^{n-1}}$. First of all, f_0 is a probability density, since we have, thanks to (20)

$$\begin{aligned}\int_{\mathbb{S}} f_0(v)\mathrm{d}v &= \frac{\int_{\mathbb{R}^{n-1}} \frac{h(|z|)\,\mathrm{d}z}{(1+|z|^2)^{n-1}}}{\int_{\mathbb{R}^{n-1}} \frac{\mathrm{d}z}{(1+|z|^2)^{n-1}}} = \frac{\int_0^{+\infty} \frac{h(r) r^{n-2}\,\mathrm{d}r}{(1+r^2)^{n-1}}}{\int_0^{+\infty} \frac{r^{n-2}\,\mathrm{d}r}{(1+r^2)^{n-1}}} = b_n^{-1} \int_0^{+\infty} \frac{h(r) r^{n-2}\,\mathrm{d}r}{(1+r^2)^{n-1}} \\ &= \int_0^{\varepsilon} \frac{1-g(0)}{\varepsilon}\,\mathrm{d}r - \int_1^{+\infty} \frac{g'(\ln r)}{r}\,\mathrm{d}r = 1 - g(0) - [g(\ln r)]_1^{+\infty} = 1.\end{aligned}$$

By symmetry, we have that $J_f(t) = \alpha(t) e_n$. Let us check that $\alpha(0) > 0$. We do as in formula (22):

$$1 - \alpha(t) = b_n^{-1} \int_0^{+\infty} \frac{2r^2 e^{-2\lambda(t)} h(r) r^{n-2}\,\mathrm{d}r}{(1 + r^2 e^{-2\lambda(t)})(1+r^2)^{n-1}}.$$

We therefore get

$$\begin{aligned}1 - \alpha(0) &= \int_0^{\varepsilon} \frac{2(1-g(0))r^2 \mathrm{d}r}{(1+r^2)\varepsilon} - \int_1^{+\infty} g'(\ln r) \frac{2r}{1+r^2} \mathrm{d}r \\ &\leqslant \frac{2\varepsilon^2}{3}(1-g(0)) - 2\int_1^{\infty} g'(\ln r)\frac{\mathrm{d}r}{r} = 2g(0) + \frac{2\varepsilon^2}{3}(1-g(0)),\end{aligned}$$

which is strictly less than 1 as long as $g(0) < \frac{1}{2}$ and ε is sufficiently small. Therefore in this case we have $\alpha(0) > 0$ (this shows that the restriction $g(0) < \frac{1}{2}$ is somehow optimal, we cannot have $W_1(f(0,\cdot),\delta_{\Omega_\infty}) \geqslant \frac{1}{2}$ and $f(t,\cdot)$ weakly converging to δ_{Ω_∞} for this class of functions). This means that $\Omega(t) = e_n = \Omega_\infty$ for all time t, and thanks to Theorem 1, since f_0 has no atoms, the solution $f(t,\cdot)$ converges weakly to δ_{Ω_∞} as $t \to +\infty$.

Let us also remark that $W_1(f(t,\cdot),\delta_{e_n}) = \int_{\mathbb{S}} |v - e_n| f(t,v)\mathrm{d}v$ (see the proof of the forthcoming Proposition 7), and since we have $1 - v\cdot e_n \leqslant |v - e_n|$, we obtain $1 - \alpha(t) \leqslant W_1(f(t,\cdot),\delta_{e_n})$. Therefore, to prove that the convergence of f towards δ_{Ω_∞} is as slow as $g(t)$, it only remains to prove that $1 - \alpha(t) \geqslant g(t)$. We have $\lambda(t) \leqslant t$, and so when $r \geqslant e^t$, we get $re^{-\lambda(t)} \leqslant 1$. Since $x \mapsto \frac{2x}{1+x}$ is increasing, we get $\frac{2r^2 e^{-2\lambda(t)}}{1+r^2 e^{-2\lambda(t)}} \geqslant 1$. We therefore get

$$1 - \alpha(t) \geqslant -\int_{e^t}^{+\infty} g'(\ln r) \frac{2re^{-2\lambda(t)}}{(1+r^2 e^{-2\lambda(t)})} \mathrm{d}r \geqslant -\int_{e^t}^{+\infty} \frac{g'(\ln r)\mathrm{d}r}{r} = g(t),$$

which ends the proof. □

We conclude this subsection by more precise estimates of the rate of convergence in various Wasserstein distances when Ω is constant in time and when the initial condition has a density with respect to the Lebesgue measure which is bounded above and below. We write $a(t) \asymp b(t)$ whenever there exists two positive constants c_1, c_2 such that $c_1 b(t) \leqslant a(t) \leqslant c_2 b(t)$ for all $t \geqslant 0$. We recall the definition of the Wasserstein distance W_2, for two probability measures μ and ν on $\mathbb{S}$:

$$W_2^2(\mu,\nu) = \inf_{\pi} \int_{\mathbb{S}\times\mathbb{S}} |v-w|^2 \mathrm{d}\pi(v,w),$$

where the infimum is taken over the probability measures π on $\mathbb{S}\times\mathbb{S}$ with first and second marginals respectively equal to μ and ν.

Proposition 7. *Suppose that f_0 has a density with respect to the Lebesgue measure satisfying $m \leqslant f_0(v) \leqslant M$ for all v (for some $0 < m < M$), with $J_{f_0} \neq 0$ and such that $\Omega(t) = e_n$ is constant in time. Then we have*

$$W_1(f(t,\cdot),\delta_{e_n}) \asymp \begin{cases} (1+t)e^{-t} & \text{if } n=2, \\ e^{-t} & \text{if } n \geqslant 3, \end{cases}$$

$$W_2(f(t,\cdot),\delta_{e_n}) \asymp \begin{cases} e^{-\frac{1}{2}t} & \text{if } n=2, \\ \sqrt{1+t}\, e^{-t} & \text{if } n=3, \\ e^{-t} & \text{if } n \geqslant 4. \end{cases}$$

Proof. Let us first give explicit formulas for $W_1(f(t,\cdot),\delta_{e_n})$ and $W_2(f(t,\cdot),\delta_{e_n})$. If $\varphi \in \mathrm{Lip}_1(\mathbb{S})$, we have

$$\left| \int_{\mathbb{S}} \varphi(v) f(t,v)\,\mathrm{d}v - \varphi(e_n) \right| \leqslant \int_{\mathbb{S}} |\varphi(v)-\varphi(e_n)| f(t,v)\,\mathrm{d}v \leqslant \int_{\mathbb{S}} |v-e_n| f(t,v)\,\mathrm{d}v.$$

Therefore, by taking the supremum, we get $W_1(f(t,\cdot),\delta_{e_n}) \leqslant \int_{\mathbb{S}} |v-e_n| f(t,v)\,\mathrm{d}v$. Furthermore, by taking $\varphi(v) = |v-e_n|$, we get that this inequality is an equality. The explicit expression of $W_2(f(t,\cdot),\delta_{e_n})$ comes from the fact that the only probability measure on $\mathbb{S}\times\mathbb{S}$ with marginals $f(t,\cdot)$ and δ_{e_n} is the product measure $\mu \otimes \delta_{v_0}$, and therefore we have $W_2^2(f(t,\cdot),\delta_{e_n}) = \int_{\mathbb{S}} |v-e_n|^2 f(t,v)\,\mathrm{d}v$. Using the fact that $|v-e_n|^2 = 2 - 2v\cdot e_n$ and the definition (19) of p, we get $|p(z)-e_n| = \frac{2|z|}{\sqrt{1+|z|^2}}$. Finally, using (21), we obtain

$$W_1(f(t,\cdot),\delta_{e_n}) = c_n^{-1} \int_{\mathbb{R}^{n-1}} \frac{2|z| e^{-\lambda(t)} f_0(p(z))}{\sqrt{1+|z|^2 e^{-2\lambda(t)}}(1+|z|^2)^{n-1}}\,\mathrm{d}z, \tag{23}$$

and, as in (22):

$$W_2^2(f(t,\cdot),\delta_{e_n}) = 2(1-\alpha(t)) = c_n^{-1} \int_{\mathbb{R}^{n-1}} \frac{4|z|^2 e^{-2\lambda(t)} f_0(p(z))\,\mathrm{d}z}{(1+|z|^2 e^{-2\lambda(t)})(1+|z|^2)^{n-1}}. \tag{24}$$

Thanks to the assumptions on f_0, from (23) we immediately get

$$W_1(f(t,\cdot),\delta_{e_n}) \asymp \int_0^{+\infty} \frac{r^{n-1}e^{-\lambda(t)}\,\mathrm{d}r}{\sqrt{1+r^2e^{-2\lambda(t)}}(1+r^2)^{n-1}},$$

and for $n \geqslant 3$, since $\lambda(t) \geqslant 0$, we get

$$\begin{aligned} 0 < \int_0^{+\infty} \frac{r^{n-1}\,\mathrm{d}r}{\sqrt{1+r^2}(1+r^2)^{n-1}} &\leqslant \int_0^{+\infty} \frac{r^{n-1}\,\mathrm{d}r}{\sqrt{1+r^2e^{-2\lambda(t)}}(1+r^2)^{n-1}} \\ &\leqslant \int_0^{+\infty} \frac{r^{n-1}\,\mathrm{d}r}{(1+r^2)^{n-1}} < +\infty, \end{aligned}$$

which gives $W_1(f(t,\cdot),\delta_{e_n}) \asymp e^{-\lambda(t)}$. For $n=2$, we have

$$\begin{aligned} \int_0^{+\infty} \frac{re^{-\lambda(t)}\,\mathrm{d}r}{\sqrt{1+r^2e^{-2\lambda(t)}}(1+r^2)} &= \left[\frac{e^{-\lambda(t)}}{2\sqrt{1-e^{-2\lambda(t)}}} \ln\left(\frac{\sqrt{1+r^2e^{-2\lambda(t)}}-\sqrt{1-e^{-2\lambda(t)}}}{\sqrt{1+r^2e^{-2\lambda(t)}}+\sqrt{1-e^{-2\lambda(t)}}}\right)\right]_0^{+\infty} \\ &= \frac{e^{-\lambda(t)}}{2\sqrt{1-e^{-2\lambda(t)}}} \ln\left(\frac{1+\sqrt{1-e^{-2\lambda(t)}}}{1-\sqrt{1-e^{-2\lambda(t)}}}\right). \end{aligned}$$

Since this last expression is equivalent to $\lambda(t)e^{-\lambda(t)}$ as $\lambda(t) \to +\infty$ and converges to 1 as $\lambda(t) \to 0$, we then get $W_1(f(t,\cdot),\delta_{e_n}) \asymp (1+\lambda(t))e^{-\lambda(t)}$.

We proceed similarly for the distance W_2. From the assumptions on f_0 and (24) we get

$$W_2^2(f(t,\cdot),\delta_{e_n}) \asymp 1-\alpha(t) \asymp \int_0^{+\infty} \frac{r^n e^{-2\lambda(t)}\,\mathrm{d}r}{(1+r^2e^{-2\lambda(t)})(1+r^2)^{n-1}}.$$

By the same argument of integrability, when $n \geqslant 4$, since $\int_0^{+\infty} \frac{r^n\,\mathrm{d}r}{(1+r^2)^{n-1}} < +\infty$, we obtain $1-\alpha(t) \asymp e^{-2\lambda(t)}$. For $n=2$ we have

$$\begin{aligned} \int_0^{+\infty} \frac{r^2e^{-2\lambda(t)}\,\mathrm{d}r}{(1+r^2e^{-2\lambda(t)})(1+r^2)} &= \left[\frac{e^{-\lambda(t)}\tan^{-1}(e^{-\lambda(t)}r)-e^{-2\lambda(t)}\tan^{-1}(r)}{1-e^{-2\lambda(t)}}\right]_0^{+\infty} \\ &= \frac{\pi\, e^{-\lambda(t)}}{2(1+e^{-\lambda(t)})}, \end{aligned}$$

which gives $1-\alpha(t) \asymp e^{-\lambda(t)}$. For $n=3$ we have

$$\begin{aligned} \int_0^{+\infty} \frac{r^2e^{-2\lambda(t)}\,\mathrm{d}r}{(1+r^2e^{-2\lambda(t)})(1+r^2)^2} &= \frac{e^{-2\lambda(t)}}{2(1-e^{-2\lambda(t)})^2}\left[\ln\left(\frac{1+r^2}{1+r^2e^{-2\lambda(t)}}\right)+\frac{1-e^{-2\lambda(t)}}{1+r^2}\right]_0^{+\infty} \\ &= \frac{e^{-2\lambda(t)}}{2(1-e^{-2\lambda(t)})^2}(2\lambda(t)-1+e^{-2\lambda(t)}). \end{aligned}$$

Since this last expression is equivalent to $\lambda(t)e^{-2\lambda(t)}$ as $\lambda(t) \to +\infty$ and converges to $\frac{1}{4}$ as $\lambda(t) \to 0$, we then get $1-\alpha(t) \asymp (1+\lambda(t))e^{-2\lambda(t)}$.

In all dimensions, we have, since $\lambda(t) = \int_0^t \alpha(\tau)\mathrm{d}\tau \geqslant \alpha(0)t$, that there exists $C > 0$ such that $1 - \alpha(t) \geqslant Ce^{-\alpha(0)t}$. Therefore, integrating in time, we obtain $t - \lambda(t) \geqslant \widetilde{C}e^{-\alpha(0)t}$. This gives, since $\lambda(t) \leqslant t$, that $e^{-\lambda(t)} \sim e^{-t}$ and $1+\lambda(t) \asymp 1+t$. Combining this with all the estimates we obtain so far (and reminding that $W_2(f(t,\cdot),\delta_{e_n}) \asymp \sqrt{1-\alpha(t)}$ ends the proof. □

Interestingly, the estimates given by Proposition 7 depend on the dimension and on the chosen distance. We expect that these estimates still hold when Ω depends on time, and, as in the result of Theorem 2, we expect to have an even better rate of convergence of Ω towards Ω_∞.

3 The Particle Model

The object of this section is to prove Theorem 2, and we divide it into several propositions. We take N positive real numbers $(m_i)_{1\leqslant i\leqslant N}$ with $\sum_{i=1}^N m_i = 1$, and N unit vectors $v_i^0 \in \mathbb{S}$ (for $1 \leqslant i \leqslant N$) such that $v_i^0 \neq v_j^0$ for all $i \neq j$. We denote by $(v_i)_{1\leqslant i\leqslant N}$ the solution of the system of differential equation (3):

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = P_{v_i^\perp} J, \text{ with } J(t) = \sum_{i=1}^N m_i v_i(t),$$

with the initial conditions $v_i(0) = v_i^0$ for $1 \leqslant i \leqslant N$.

Proposition 8. *If $J(0) \neq 0$, then $|J|$ is nondecreasing, so $\Omega(t) = \frac{J(t)}{|J(t)|} \in \mathbb{S}$ is well-defined for all times $t \geqslant 0$. We have one of the two following possibilities:*

- *For all $1 \leqslant i \leqslant N$, $v_i(t)\cdot\Omega(t) \to 1$ as $t \to +\infty$,*
- *There exists i_0 such that $v_i(t)\cdot\Omega(t) \to -1$ as $t \to +\infty$, and for all $i \neq i_0$, we have $v_i(t)\cdot\Omega(t) \to 1$ as $t \to +\infty$.*

Furthermore, if we denote by $\lambda > 0$ the limit of $|J(t)|$ as $t \to +\infty$, we have for all i, j in the first possibility (resp. for all $i \neq i_0$, $j \neq i_0$ in the second possibility), $\|v_i(t) - v_j(t)\| = O(e^{-(\lambda-\varepsilon)t})$ (for any $\varepsilon > 0$).

Proof. Let us see the differential system as a kind of gradient flow of the following interaction energy (this is reminiscent of the gradient flow structure of the kinetic equation (2), see Remark 3):

$$\mathcal{E} = \frac{1}{2}\sum_{i,j=1}^N m_i m_j \|v_i - v_j\|^2 = \sum_{i,j=1}^N m_i m_j (1 - v_i\cdot v_j) = 1 - |J|^2 \geqslant 0$$

Indeed, we then get $\nabla_{v_i}\mathcal{E} = -2\sum_{j=1}^N m_i m_j P_{v_i^\perp} v_j = -2m_i P_{v_i^\perp} J$ (using the formula $\nabla_v(u\cdot v) = P_{v^\perp}u$). We therefore have $\frac{\mathrm{d}v_i}{\mathrm{d}t} = -\frac{1}{2m_i}\nabla_{v_i}\mathcal{E}$, and we obtain

$$\frac{\mathrm{d}|J|^2}{\mathrm{d}t} = -\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = -\sum_{i=1}^N \nabla_{v_i}\mathcal{E}\cdot\frac{\mathrm{d}v_i}{\mathrm{d}t} = 2\sum_{i=1}^N m_i \left|\frac{\mathrm{d}v_i}{\mathrm{d}t}\right|^2 \geqslant 0. \tag{25}$$

This gives that $|J|$ is nondecreasing in time. So we can define $\Omega(t) = \frac{J(t)}{|J(t)|}$ and rewrite (25) as

$$\frac{\mathrm{d}|J|^2}{\mathrm{d}t} = 2\sum_{i=1}^{N} m_i |P_{v_i^\perp} J|^2 = 2|J|^2 \sum_{i=1}^{N} m_i (1 - (v_i \cdot \Omega)^2). \tag{26}$$

We can compute the time derivative of this quantity and observe that all terms are uniformly bounded in time. Therefore, since it is an integrable function of time (since $|J|^2 \leqslant 1$) with bounded derivative, it must converge to 0 as $t \to +\infty$. Therefore we obtain that $(v_i(t) \cdot \Omega(t))^2 \to 1$ for all $1 \leqslant i \leqslant N$. Let us now take $1 \leqslant i, j \leqslant N$ and estimate $\|v_i - v_j\|$. We have

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_i - v_j\|^2 &= -\frac{\mathrm{d}}{\mathrm{d}t}(v_i \cdot v_j) = -|J|(v_j \cdot P_{v_i^\perp}\Omega + v_i \cdot P_{v_j^\perp}\Omega) \\ &= -|J|\,(\Omega \cdot v_i + \Omega \cdot v_j)(1 - v_i \cdot v_j) \\ &= -|J|\,\Omega \cdot \frac{v_i + v_j}{2}\|v_i - v_j\|^2. \end{aligned} \tag{27}$$

Therefore if $v_i \cdot \Omega \to 1$ and $v_j \cdot \Omega \to 1$, we get $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_i - v_j\|^2 \leqslant -(\lambda - \varepsilon)\|v_i - v_j\|^2$ for t sufficiently large, and therefore we obtain $\|v_i - v_j\|^2 = O(e^{-2(\lambda-\varepsilon)t})$.

Finally if $v_i \cdot \Omega \to -1$ and $v_j \cdot \Omega \to -1$, for t sufficiently large (say $t \geqslant t_0$) we obtain $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_i - v_j\|^2 \geqslant (\lambda - \varepsilon)\|v_i - v_j\|^2$. This is the same phenomenon of repulsion as (15) in the previous part, and the only bounded solution to this differential inequality is when $v_i(t_0) = v_j(t_0)$, which means, by uniqueness that $v_i^0 = v_j^0$ and therefore $i = j$. This means that if there is an index i_0 such that $v_{i_0}(t) \cdot \Omega(t) \to -1$, then for all $i \neq i_0$, we have $v_i(t) \cdot \Omega(t) \to 1$ as $t \to \infty$, and this ends the proof. □

Let us now study the first possibility more precisely.

Proposition 9. *Suppose that $v_i(t) \cdot \Omega(t) \to 1$ as $t \to \infty$ for all $1 \leqslant i \leqslant N$. Then there exists $\Omega_\infty \in \mathbb{S}$ and $a_i \in \{\Omega_\infty\}^\perp \subset \mathbb{R}^n$, for $1 \leqslant i \leqslant N$ such that $\sum_{i=1}^N m_i a_i = 0$ and that, as $t \to +\infty$,*

$$\begin{aligned} &v_i(t) = (1 - |a_i|^2 e^{-2t})\Omega_\infty + e^{-t} a_i + O(e^{-3t}) \quad \text{for } 1 \leqslant i \leqslant N, \\ &\Omega(t) = \Omega_\infty + O(e^{-3t}). \end{aligned}$$

Proof. We first have $|J(t)| = J(t) \cdot \Omega(t) = \sum_i m_i v_i(t) \cdot \Omega(t) \to 1$ as $t \to \infty$. Therefore $\lambda = 1$, and thanks to the estimates of Proposition 8 (first possibility), for all i, j we have $1 - v_i \cdot v_j = \frac{1}{2}\|v_i - v_j\|^2 = O(e^{-2(1-\varepsilon)t})$. Summing with weights m_j, we obtain $1 - v_i \cdot J = O(e^{-2(1-\varepsilon)t})$. Plugging back this into (27), we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_i - v_j\|^2 = -\big(1 + O(e^{-2(1-\varepsilon)t})\big)\|v_i - v_j\|^2.$$

We therefore obtain $\|v_i - v_j\|^2 = \|v_i^0 - v_j^0\|^2 e^{-\int_0^t (1 + O(e^{-2(1-\varepsilon)\tau}))\mathrm{d}\tau} = O(e^{-2t})$. This is the same estimate as previously without the ε. Therefore, similarly, we

get $1 - v_i \cdot J = O(e^{-2t})$, which gives $1 - |J|^2 = O(e^{-2t})$ by summing with weights m_i. We finally obtain $1 - v_i \cdot \Omega = 1 - v_i \cdot J + (|J| - 1)v_i \cdot \Omega = O(e^{-2t})$, therefore $|P_{v_i^\perp}\Omega|^2 = |P_{\Omega^\perp} v_i|^2 = 1 - (v_i \cdot \Omega)^2 = O(e^{-2t})$.

Let us now compute the evolution of Ω, as in (8). Since $\frac{\mathrm{d}J}{\mathrm{d}t} = \sum_i m_i P_{v_i^\perp} J$, we use (26) to get $\frac{\mathrm{d}|J|}{\mathrm{d}t} = |J| \sum_i m_i |P_{v_i^\perp}\Omega|^2 = O(e^{-2t})$, and we obtain

$$\begin{aligned}\frac{\mathrm{d}\Omega}{\mathrm{d}t} &= \frac{1}{|J|}\frac{\mathrm{d}J}{\mathrm{d}t} - \frac{\mathrm{d}|J|}{\mathrm{d}t}\frac{J}{|J|^2} = \sum_i m_i P_{v_i^\perp}\Omega - \sum_i m_i |P_{v_i^\perp}\Omega|^2 \Omega \\ &= -\sum_i m_i (v_i \cdot \Omega)(v_i - (v_i \cdot \Omega)\Omega) = -\sum_i m_i (v_i \cdot \Omega) P_{\Omega^\perp} v_i.\end{aligned}$$

Since $\sum_i m_i P_{\Omega^\perp} v_i = P_{\Omega^\perp} J = 0$, we can then add this quantity to the previous identity to get

$$\frac{\mathrm{d}\Omega}{\mathrm{d}t} = \sum_i m_i (1 - v_i \cdot \Omega) P_{\Omega^\perp} v_i. \tag{28}$$

We therefore get $|\frac{\mathrm{d}\Omega}{\mathrm{d}t}| \leqslant \sum_i m_i (1 - v_i \cdot \Omega)|P_{\Omega^\perp} v_i| = O(e^{-3t})$. Therefore Ω converges towards $\Omega_\infty \in \mathbb{S}$ and we have $\Omega = \Omega_\infty + O(e^{-3t})$.

Finally, to get the precise estimates for the v_i, we compute their second derivative.

$$\frac{\mathrm{d}^2 v_i}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t} P_{v_i^\perp} J = P_{v_i^\perp}\frac{\mathrm{d}J}{\mathrm{d}t} - \frac{\mathrm{d}v_i}{\mathrm{d}t} \cdot J\, v_i - v_i \cdot J \frac{\mathrm{d}v_i}{\mathrm{d}t}. \tag{29}$$

We have $P_{v_i^\perp}\frac{\mathrm{d}J}{\mathrm{d}t} = \frac{\mathrm{d}|J|}{\mathrm{d}t} P_{v_i^\perp}\Omega + |J| P_{v_i^\perp}\frac{\mathrm{d}\Omega}{\mathrm{d}t} = O(e^{-3t})$, since $P_{v_i^\perp}\Omega = O(e^{-t})$ and $\frac{\mathrm{d}|J|}{\mathrm{d}t} = O(e^{-2t})$ thanks to (26). Then we notice that $\frac{\mathrm{d}v_i}{\mathrm{d}t} \cdot J = J \cdot P_{v_i^\perp} J = |\frac{\mathrm{d}v_i}{\mathrm{d}t}|^2$ and that $v_i \cdot J \frac{\mathrm{d}v_i}{\mathrm{d}t} = \frac{\mathrm{d}v_i}{\mathrm{d}t} - (1 - v_i \cdot J) P_{v_i^\perp} J = \frac{\mathrm{d}v_i}{\mathrm{d}t} + O(e^{-3t})$. At the end we obtain

$$\frac{\mathrm{d}^2 v_i}{\mathrm{d}t^2} = -\frac{\mathrm{d}v_i}{\mathrm{d}t} - \left|\frac{\mathrm{d}v_i}{\mathrm{d}t}\right|^2 v_i + O(e^{-3t}). \tag{30}$$

Considering first that $|\frac{\mathrm{d}v_i}{\mathrm{d}t}|^2 = O(e^{-2t})$, the resolution of this differential equation gives $\frac{\mathrm{d}v_i}{\mathrm{d}t} = -a_i e^{-t} + O(e^{-2t})$ with $a_i \in \mathbb{R}^n$. Integrating in time, we therefore obtain $v_i(t) = \Omega_\infty + a_i e^{-t} + O(e^{-2t})$, (we already know that $v_i(t)$ converges to Ω_∞ since $v(t) \cdot \Omega(t) \to 1$). The fact that $|v_i(t)| = 1$ gives us $a_i \cdot \Omega_\infty e^{-t} = O(e^{-2t})$ and therefore $a_i \in \{\Omega_\infty\}^\perp$. Summing all these estimations with weights m_i and using the fact that $J - \Omega_\infty = O(e^{-2t})$, we obtain $\sum_i m_i a_i = 0$.

Finally, the more precise estimate for $v_i(t)$ up to order $O(e^{-3t})$ given in the proposition is obtained by plugging back $|\frac{\mathrm{d}v_i}{\mathrm{d}t}|^2 v_i = |a_i|^2 e^{-2t}\Omega_\infty + O(e^{-3t})$ into (30) and solving it again. □

Let us finally study the second possibility.

Proposition 10. *Suppose there exists i_0 such that $v_{i_0}(t) \cdot \Omega(t) \to -1$ as $t \to \infty$. Then we have $\lambda = 1 - 2m_{i_0}$ (which gives $m_{i_0} < \frac{1}{2}$), and there exists $\Omega_\infty \in \mathbb{S}$*

and $a_i \in \{\Omega_\infty\}^\perp \subset \mathbb{R}^n$ *for* $i \neq i_0$ *such that* $\sum_{i\neq i_0} m_i a_i = 0$ *and that, as* $t \to +\infty$,

$$\begin{aligned} v_i(t) &= (1 - |a_i|^2 e^{-2\lambda t})\Omega_\infty + e^{-\lambda t} a_i + O(e^{-3\lambda t}) \quad \text{for } i \neq i_0, \\ v_{i_0}(t) &= -\Omega_\infty + O(e^{-3\lambda t}), \\ \Omega(t) &= \Omega_\infty + O(e^{-3\lambda t}). \end{aligned}$$

Proof. First of all we have $|J(t)| = \Omega(t)\cdot J(t) = \sum_i m_i v_i(t)\cdot\Omega(t)$ which converges as $t \to \infty$ towards $\lambda = \sum_{i\neq i_0} m_i - m_{i_0} = 1 - 2m_{i_0}$. The proof then follows closely the one of Proposition 9, except for the case of v_{i_0}. Indeed, Proposition 8 only gives estimates on $\|v_i - v_j\|$ (and therefore on $v_i \cdot v_j$) when $i \neq i_0$ and $j \neq i_0$. To estimate more precisely the quantity $v_{i_0}\cdot v_i$, let us prove that $-v_{i_0}$ must be in the convex cone spanned by 0 and all the v_i, $i \neq i_0$. The idea is that a configuration which is in a convex cone stays in it for all time.

Let us suppose that all the v_i (including $i = i_0$) satisfy $e \cdot v_i(t_0) \geqslant c$ for some $c > 0$, $t_0 \geqslant 0$ and $e \in \mathbb{S}$ (the direction of the cone). We want to prove that $e \cdot v_i(t) \geqslant c$ for all i and for all $t \geqslant t_0$. If not, we denote by $t_1 > t_0$ a time such that $e \cdot v_i(t) \geqslant 0$ for all i on $[t_0, t_1]$, but with $e \cdot v_j(t_1) < c$ for some j. On $[t_0, t_1]$, we have

$$\frac{\mathrm{d}(e \cdot v_i)}{\mathrm{d}t} = e \cdot J - (e \cdot v_i)(v_i \cdot J) \geqslant e \cdot J - (e \cdot v_i), \tag{31}$$

since $v_i \cdot J \leqslant |J| \leqslant 1$ and $e \cdot v_i \geqslant 0$ on $[t_0, t_1]$. Summing with weights m_i, we obtain $\frac{\mathrm{d}(e\cdot J)}{\mathrm{d}t} \geqslant 0$. Therefore, since $e \cdot J(t_0) \geqslant c$, we obtain $e \cdot J(t) \geqslant c$ on $[t_0, t_1]$, and the estimation (31) becomes $\frac{\mathrm{d}(e\cdot v_i)}{\mathrm{d}t} \geqslant c - (e \cdot v_i)$. By comparison principle, this tells us that $e \cdot v_i \geqslant c$ on $[t_0, t_1]$ for all i, which is a contradiction.

Let us now fix $t_0 \geqslant 0$. We want to prove that there exists $\alpha_i \geqslant 0$ for $i \neq i_0$ such that $-v_{i_0} = \sum_{i\neq i_0} \alpha_i v_i$ (this means that $-v_{i_0}$ is in the convex cone spanned by all other v_i's). This is the typical case where we will apply Farkas' Lemma (see for instance [14]): its precise conclusion is that it is equivalent to prove that this is not possible to find $e \in \mathbb{S}$ such that $e\cdot v_i(t_0) \geqslant 0$ for all $i \neq i_0$ and $e\cdot(-v_{i_0}) < 0$ (which means separating the generators of the cone and the vector $-v_{i_0}$ by a linear hyperplane).

By contradiction, if such a e exists, we would have $e\cdot J(t_0) \geqslant m_{i_0} e\cdot v_{i_0} > 0$ and for $i \neq i_0$, as in (31), if $e \cdot v_i(t_0) = 0$ we get $\frac{\mathrm{d}(e\cdot v_i)}{\mathrm{d}t}|_{t=t_0} = e \cdot J(t_0) > 0$. Therefore for $\delta > 0$ sufficiently small, we have $e \cdot v_i(t_0 + \delta) > 0$ for all i (including i_0, and those for which $e \cdot v_i(t_0) > 0$). Therefore there exists $c > 0$ such that for all i, $e\cdot v_i(t_0+\delta) \geqslant c$, and by the previous paragraph, we get that $e\cdot v_i(t) \geqslant c$ for all $t \geqslant t_0 + \delta$. We therefore get $e \cdot \Omega(t) \geqslant \frac{1}{|J(t)|} e \cdot J(t) \geqslant \frac{c}{|J(0)|}$ for all $t \geqslant t_0 + \delta$. Finally, since $\|v_{i_0}(t) + \Omega(t)\|^2 = 2(1 + v_{i_0}(t) \cdot \Omega(t)) \to 0$ as $t \to \infty$, this is in contradiction with the fact that $e\cdot(v_{i_0}+\Omega(t)) \geqslant (1+\frac{1}{|J|(0)})c > 0$ for all $t \geqslant t_0+\delta$.

In conclusion we have that for all $t \geqslant 0$, there exists $\alpha_i(t) \geqslant 0$ for $i \neq i_0$ such that $-v_{i_0}(t) = \sum_{i\neq i_0} \alpha_i(t) v_i(t)$. We thus obtain, for $i \neq i_0$

$$v_i(t) \cdot v_{i_0}(t) = -\sum_{i\neq i_0} \alpha_i + \sum_{j\neq i_0} \alpha_i (1 - v_j(t) \cdot v_i(t)) \leqslant -1 + O(e^{-2(\lambda-\varepsilon)t}), \tag{32}$$

since $1 = \|v_{i_0}(t)\| \leqslant \sum_{i\neq i_0} \alpha_i \|v_i(t)\| = \sum_{i\neq i_0} \alpha_i$, and thanks to Proposition 8. Since $v_i(t)\cdot v_{i_0}(t) \geqslant -1$, this gives $v_i(t)\cdot v_{i_0}(t) = -1 + O(e^{-2(\lambda-\varepsilon)t})$. From there, we have, if $i \neq i_0$,

$$\begin{aligned} v_i \cdot J &= \sum_{i\neq i_0} m_j\, v_i \cdot v_j - m_{i_0} v_i \cdot v_{i_0} \\ &= \sum_{i\neq i_0} (m_j + O(e^{-2(\lambda-\varepsilon)t}) - m_{i_0} + O(e^{-2(\lambda-\varepsilon)t}) = \lambda + O(e^{-2(\lambda-\varepsilon)t}). \end{aligned}$$

Plugging this into (27), for $i \neq i_0$ and $j \neq i_0$, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_i - v_j\|^2 = -\big(\lambda + O(e^{-2(\lambda-\varepsilon)t})\big)\|v_i - v_j\|^2.$$

We therefore obtain, as in the proof of Proposition 9, $1 - v_i \cdot v_j = O(e^{-2\lambda t})$. As in (32), we now get $v_i \cdot v_{i_0} = -1 + O(e^{-2\lambda t})$. Finally, by summing with weights m_j, we obtain $v_i \cdot J = \lambda + O(e^{-2\lambda t})$ for $i \neq i_0$ and $v_{i_0}\cdot J = -\lambda + O(e^{-2\lambda t})$. Therefore, by summing once again with weights m_i, we get $|J|^2 = \lambda^2 + O(e^{-2\lambda t})$. This allows to get $1 - v_i \cdot \Omega = O(e^{-2\lambda t})$ and $|P_{v_i^\perp}\Omega| = O(e^{-\lambda t})$ when $i \neq i_0$, and $1 + v_{i_0}\cdot\Omega = O(e^{-2\lambda t})$. Unfortunately this is not enough to use (28) to obtain a decay at rate 3λ: we obtain

$$\left|\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right| \leqslant O(e^{-3\lambda t}) + m_{i_0}(1 - v_{i_0}\cdot\Omega)|P_{v_{i_0}^\perp}\Omega|. \tag{33}$$

However, since $|P_{v_{i_0}^\perp}\Omega|^2 = 1 - (v_{i_0}\cdot\Omega)^2 = (1 - v_{i_0}\cdot\Omega)(1 + v_{i_0}\cdot\Omega) = O(e^{-2\lambda t})$, we obtain at least $\left|\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right| \leqslant O(e^{-\lambda t})$, which gives the existence of $\Omega_\infty \in \mathbb{S}$ such that $\Omega(t) = \Omega_\infty + O(e^{-\lambda t})$. To get the rate 3λ, we have to be a little bit more careful, and use the same kind of trick as in Lemma 2 of the first part: if we have a differential equation of the form $y' = y + O(e^{-\beta t})$, and furthermore that y is bounded, then we must have $y = O(e^{-\beta t})$. Indeed, by Duhamel's formula, we get $y = y_0 e^t + O(e^{-\beta t})$ and the only bounded solution corresponds to $y_0 = 0$. We apply this to $y = \frac{\mathrm{d}v_{i_0}}{\mathrm{d}t}$. We have, as in (29)

$$\begin{aligned} \frac{\mathrm{d}^2 v_{i_0}}{\mathrm{d}t^2} &= P_{v_{i_0}^\perp}\frac{\mathrm{d}J}{\mathrm{d}t} - \frac{\mathrm{d}v_{i_0}}{\mathrm{d}t}\cdot J\, v_{i_0} - v_{i_0}\cdot J\,\frac{\mathrm{d}v_{i_0}}{\mathrm{d}t} \\ &= P_{v_{i_0}^\perp}\frac{\mathrm{d}J}{\mathrm{d}t} - \left|\frac{\mathrm{d}v_{i_0}}{\mathrm{d}t}\right|^2 v_{i_0} + \lambda\frac{\mathrm{d}v_{i_0}}{\mathrm{d}t} + O(e^{-3\lambda t}). \end{aligned} \tag{34}$$

We have

$$P_{v_{i_0}^\perp}\frac{\mathrm{d}J}{\mathrm{d}t} = P_{v_{i_0}^\perp}\Big[J - \sum_{i=1}^N m_i (v_i\cdot J) v_i\Big] = (1-\lambda)P_{v_{i_0}^\perp} J + \sum_{i=1}^N m_i(\lambda - v_i\cdot J)P_{v_{i_0}^\perp} v_i.$$

The term for $i = i_0$ in this last sum vanishes and we have $\lambda - v_i\cdot J = O(e^{-2\lambda t})$ for $i \neq i_0$, as well as $|P_{v_{i_0}^\perp} v_i|^2 = 1 - (v_{i_0}\cdot v_i)^2 = O(e^{-2\lambda t})$. We therefore

obtain $P_{v_{i_0}^\perp}\frac{\mathrm{d}J}{\mathrm{d}t} = (1-\lambda)P_{v_{i_0}^\perp}J + O(e^{-3\lambda t})$, and writing $y = P_{v_{i_0}^\perp}J = \frac{\mathrm{d}v_{i_0}}{\mathrm{d}t}$, the formula (34) becomes $y' = y - |y|^2\, v_{i_0} + O(e^{-3\lambda t})$. We of course have that y is bounded, and we even know that $y = \frac{1}{|J|}P_{v_{i_0}^\perp}\Omega = O(e^{-\lambda t})$. We can then apply the result once by replacing $|y|^2$ with $O(e^{-2\lambda t})$ to get $y = O(e^{-2\lambda t})$, and then apply it a second time to obtain $y = O(e^{-3\lambda t})$. This already provides the result $v_{i_0}(t) = -\Omega_\infty + O(e^{-3\lambda t})$, and looking back at (33), we get that $\frac{\mathrm{d}\Omega}{\mathrm{d}t} = O(e^{-3\lambda t})$ and therefore $\Omega(t) = -\Omega_\infty + O(e^{-3\lambda t})$.

It remains to prove the more precise estimates for v_i when $i \neq i_0$, and this is done exactly as in the proof of Proposition 9, from formula (29) to the end of the proof, now we know that $\frac{\mathrm{d}\Omega}{\mathrm{d}t} = O(e^{-3\lambda t})$. The only difference is that $v_i \cdot J$ converges to λ instead of 1, together with the fact that all rates are multiplied by λ. For instance, the main estimate (30) becomes

$$\frac{\mathrm{d}^2 v_i}{\mathrm{d}t^2} = -\lambda\frac{\mathrm{d}v_i}{\mathrm{d}t} - \left|\frac{\mathrm{d}v_i}{\mathrm{d}t}\right|^2 v_i + O(e^{-3\lambda t}),$$

and the rest of the proof does not change. □

Acknowledgments. The authors want to thank the hospitality of Athanasios Tzavaras and the University of Crete, back in 2012, where this work was done and supported by the EU FP7-REGPOT project "Archimedes Center for Modeling, Analysis and Computation".

They also want to thank the anonymous referee for his fast, careful, and efficient reading, despite their very late submission.

A.F. acknowledges support from the EFI project ANR-17-CE40-0030 and the Kibord project ANR-13-BS01-0004 of the French National Research Agency (ANR), from the project Défi S2C3 POSBIO of the interdisciplinary mission of CNRS, and the project SMS co-funded by CNRS and the Royal Society.

J.-G. L. acknowledges support from the National Science Foundation under the NSF Research Network Grant no. RNMS11-07444 (KI-Net) and the grant DMS-1812573.

References

1. Ambrosio, L., Crippa, G.: Existence, uniqueness, stability and differentiability properties of the flow associated to weakly differentiable vector fields. In: Transport Equations and Multi-D Hyperbolic Conservation Laws. Lecture Notes of the Unione Matematica Italiana, vol. 5, pp. 3–57. Springer, Berlin (2008)
2. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Lectures in Mathematics ETH Zürich, 2nd edn. Birkhäuser Verlag, Basel (2008)
3. Aydoğdu, A., McQuade, S.T., Pouradier Duteil, N.: Opinion dynamics on a general compact Riemannian manifold. Netw. Heterog. Media **12**(3), 489–523 (2017)
4. Barbălat, I.: Systèmes d'équations différentielles d'oscillations non linéaires. Rev. Math. Pures Appl. **4**, 267–270 (1959)
5. Benedetto, D., Caglioti, E., Montemagno, U.: On the complete phase synchronization for the Kuramoto model in the mean-field limit. Commun. Math. Sci. **13**(7), 1775–1786 (2015)

6. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Mean-field limit for the stochastic Vicsek model. Appl. Math. Lett. **3**(25), 339–343 (2012)
7. Caponigro, M., Lai, A.C., Piccoli, B.: A nonlinear model of opinion formation on the sphere. Discrete Contin. Dyn. Syst. **35**(9), 4241–4268 (2015)
8. Degond, P., Frouvelle, A., Raoul, G.: Local stability of perfect alignment for a spatially homogeneous kinetic model. J. Stat. Phys. **157**(1), 84–112 (2014)
9. Degond, P., Motsch, S.: Continuum limit of self-driven particles with orientation interaction. Math. Models Methods Appl. Sci. **18**, 1193–1215 (2008)
10. Fatkullin, I., Slastikov, V.: Critical points of the Onsager functional on a sphere. Nonlinearity **18**, 2565–2580 (2005)
11. 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)
12. 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)
13. Ha, S.Y., Ko, D., Ryoo, S.W.: On the relaxation dynamics of Lohe oscillators on some Riemannian manifolds. J. Stat. Phys. **172**, 1427–1478 (2018)
14. Hiriart-Urruty, J.B., Lemaréchal, C.: Fundamentals of Convex Analysis. Grundlehren Text Editions. Springer, Berlin (2001)
15. Markdahl, J., Thunberg, J., Gonçalves, J.: Almost global consensus on the n-sphere. IEEE Trans. Autom. Control **63**(6), 1664–1675 (2018)
16. Spohn, H.: Large Scale Dynamics of Interacting Particles. Texts and Monographs in Physics. Springer, Heidelberg (1991)
17. 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)

Workshop 3: Stochastic Dynamics Out of Equilibrium

Tracy-Widom Asymptotics for a River Delta Model

Guillaume Barraquand(✉) and Mark Rychnovsky

Department of Mathematics, Columbia University, 2990 Broadway, New York, NY 10027, USA
barraquand@math.columbia.edu, mrychnov@gmail.com

Abstract. We study an oriented first passage percolation model for the evolution of a river delta. This model is exactly solvable and occurs as the low temperature limit of the beta random walk in random environment. We analyze the asymptotics of an exact formula from [13] to show that, at any fixed positive time, the width of a river delta of length L approaches a constant times $L^{2/3}$ with Tracy-Widom GUE fluctuations of order $L^{4/9}$. This result can be rephrased in terms of particle systems. We introduce an exactly solvable particle system on the integer half line and show that after running the system for only finite time the particle positions have Tracy-Widom fluctuations.

Keywords: KPZ universality · First passage percolation · Exclusion processes · Tracy-Widom distribution · Integrable probability

1 Model and Results

1.1 Introduction

First passage percolation was introduced in 1965 to study a fluid spreading through a random environment [37]. This model has motivated many tools in modern probability, most notably Kingman's sub-additive ergodic theorem (see the review [5] and references therein); it has attracted attention from mathematicians and physicists alike due to the simplicity of its definition, and the ease with which fascinating conjectures can be stated.

The Kardar-Parisi-Zhang (KPZ) universality class has also become a central object of study in recent years [27]. Originally proposed to explain the behavior of growing interfaces in 1986 [39], it has grown to include many types of models including random matrices, directed polymers, interacting particle systems, percolation models, and traffic models. Much of the success in studying these has come from the detailed analysis of a few exactly solvable models of each type.

We study an exactly solvable model at the intersection of percolation theory and KPZ universality: Bernoulli-exponential first passage percolation (FPP). Here is a brief description (see Definition 1 for a more precise definition). Bernoulli-exponential FPP models the growth of a river delta beginning at the

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 483–522, 2019.
https://doi.org/10.1007/978-3-030-15096-9_17

origin in $\mathbb{Z}^2_{\geq 0}$ and growing depending on two parameters $a, b > 0$. At time 0, the river is a single up-right path beginning from the origin chosen by the rule that whenever the river reaches a new vertex it travels north with probability $a/(a+b)$ and travels east with probability $b/(a+b)$ (thick black line in Fig. 1). The line with slope a/b can be thought of as giving the direction in which the expected elevation of our random terrain decreases fastest.

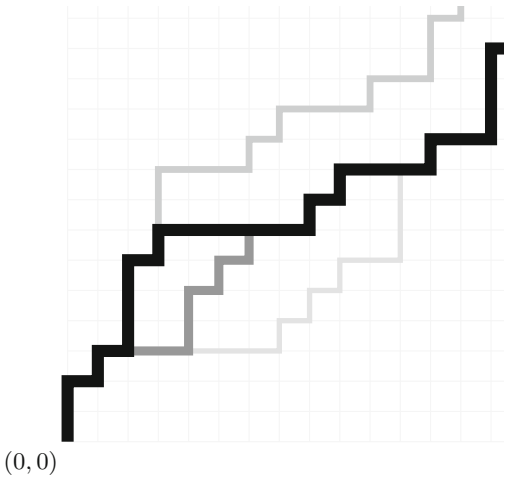


Fig. 1. A sample of the river delta (Bernoulli-exponential FPP percolation cluster) near the origin. The thick black random walk path corresponds to the river (percolation cluster) at time 0. The other thinner and lighter paths correspond to tributaries added to the river delta (percolation cluster) at later times.

As time passes, the river erodes its banks creating forks. At each vertex which the river leaves in the rightward (respectively upward) direction, it takes an amount of time distributed as an exponential random variable with rate a (resp. b) for the river to erode through its upward (resp. rightward) bank. Once the river erodes one of its banks at a vertex, the flow at this vertex branches to create a tributary (see gray paths in Fig. 1). The path of the tributary is selected by the same rule as the path of the time 0 river, except that when the tributary meets an existing river it joins the river and follows the existing path. The full path of the tributary is added instantly when the river erodes its bank.

In this model the river is infinite, and the main object of study is the set of vertices included in the river at time t, i.e. the percolation cluster. We will also refer to the shape enclosed by the outermost tributaries at time t as the river delta (see Fig. 2 for a large scale illustration of the river delta).

The model defined above can also be seen as the low temperature limit of the beta random walk in random environment (RWRE) model [13], an exactly solvable model in the KPZ universality class. Bernoulli-exponential FPP is particularly amenable to study because an exact formula for the distribution of the

percolation cluster's upper border (Theorem 3 below) can be extracted from an exact formula for the beta RWRE [13]. We perform an asymptotic analysis on this formula to prove that at any fixed time, the width of the river delta satisfies a law of large numbers type result with fluctuations converging weakly to the Tracy-Widom GUE distribution (see Theorem 2). Our law of large numbers result was predicted in [13] by taking a heuristic limit of [13, Theorem 1.19]; we present this non-rigorous computation in Sect. 1.4. We also give other interpretations of this result. In Sect. 1.6 we introduce an exactly solvable particle system and show that the position of a particle at finite time has Tracy-Widom fluctuations.

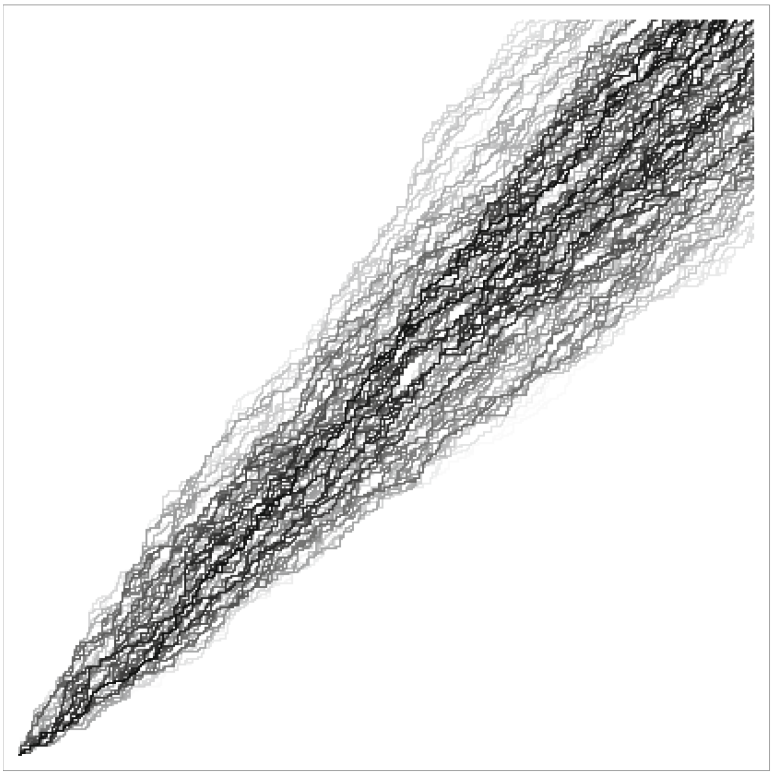


Fig. 2. The percolation cluster for 400×400 Bernoulli-exponential FPP at time 1 with $a = b = 1$. Paths occurring earlier are shaded darker, so the darkest paths occur near $t = 0$ and the lightest paths occur near $t = 1$.

1.2 Definition of the Model

We now define the model more precisely in terms of first passage percolation following [13].

Definition 1 (Bernoulli-exponential first passage percolation). *Let E_e be a family of independent exponential random variables indexed by the edges e of the lattice $\mathbb{Z}^2_{\geq 0}$. Each E_e is distributed as an exponential random variable with parameter a if e is a vertical edge, and with parameter b if e is a horizontal edge. Let $(\zeta_{i,j})$ be a family of independent Bernoulli random variables with parameter $b/(a+b)$. We define the passage time t_e of each edge e in the lattice $\mathbb{Z}^2_{\geq 0}$ by*

$$t_e = \begin{cases} \zeta_{i,j} E_e & \text{if } e \text{ is the vertical edge } (i,j) \to (i,j+1), \\ (1-\zeta_{i,j}) E_e & \text{if } e \text{ is the horizontal edge } (i,j) \to (i+1,j). \end{cases}$$

We define the point to point passage time $T^{\mathrm{PP}}(n,m)$ by

$$T^{\mathrm{PP}}(n,m) = \min_{\pi:(0,0)\to(n,m)} \sum_{e\in\pi} t_e.$$

where the minimum is taken over all up-right paths from $(0,0)$ to (n,m). We define the percolation cluster $C(t)$, at time t, by

$$C(t) = \left\{(n,m) : T^{\mathrm{PP}}(n,m) \leq t\right\}.$$

At each time t, the percolation cluster $C(t)$ is the set of points visited by a collection of up-right random walks in the quadrant $\mathbb{Z}^2_{\geq 0}$. $C(t)$ evolves in time as follows:

- At time 0, the percolation cluster contains all points in the path of a directed random walk starting from $(0,0)$, because at any vertex (i,j) we have passage time 0 to either $(i,j+1)$ or $(i+1,j)$ according to the independent Bernoulli random variables $\zeta_{i,j}$.
- At each vertex (i,j) in the percolation cluster $C(t)$, with an upward (resp. rightward) neighbor outside the cluster, we add a random walk starting from (i,j) with an upward (resp. rightward) step to the percolation cluster with exponential rate (a) (resp. b). This random walk will almost surely hit the percolation cluster after finitely many steps, and we add to the percolation cluster only those points that are in the path of the walk before the first hitting point (see Fig. 1).

Define the height function $H_t(n)$ by

$$H_t(n) = \sup\{m \in \mathbb{Z}_{\geq 0} | T^{\mathrm{PP}}(n,m) \leq t\}, \tag{1}$$

so that $(n, H_t(n))$ is the upper border of $C(t)$.

1.3 History of the Model and Related Results

Bernoulli-exponential FPP was first introduced in [13], which introduced an exactly solvable model called the beta random walk in random environment (RWRE) and studied Bernoulli-exponential FPP as a low temperature limit of

this model (see also the physics works [49,50] further studying the Beta RWRE and some variants). The beta RWRE was shown to be exactly solvable in [13] by viewing it as a limit of q-Hahn TASEP, a Bethe ansatz solvable particle system introduced in [44]. The q-Hahn TASEP was further analyzed in [20,28, 54], and was recently realized as a degeneration of the higher spin stochastic six vertex model [2,15,25,31], so that Bernoulli-exponential FPP fits as well in the framework of stochastic spin models.

Tracy-Widom GUE fluctuations were shown in [13] for Bernoulli-exponential FPP (see Theorem 1) and for Beta RWRE. In the Beta RWRE these fluctuations occur in the quenched large deviation principle satisfied by the random walk and for the maximum of many random walkers in the same environment.

The connection to KPZ universality was strengthened in subsequent works. In [30] it was shown that the heat kernel for the time reversed Beta RWRE converges to the stochastic heat equation with multiplicative noise. In [9] it was shown using a stationary version of the model that a Beta RWRE conditioned to have atypical velocity has wandering exponent 2/3 (see also [26]), as expected in general for directed polymers in $1+1$ dimensions. The stationary structure of Bernoulli-exponential FPP was computed in [48] (In [48] Bernoulli-exponential FPP is referred to as the Bernoulli-exponential polymer).

The first occurrence of the Tracy-Widom distribution in the KPZ universality class dates back to the work of Baik, Deift and Johansson on longest increasing subsequences of random permutations [7] (the connection to KPZ class was explained in e.g. [45]) and the work of Johansson on TASEP [38]. In the past ten years, following Tracy and Widom's work on ASEP [51–53] and Borodin and Corwin's Macdonald processes [16], a number of exactly solvable $1+1$ dimensional models in the KPZ universality class have been analyzed asymptotically. Most of them can be realized as more or less direct degenerations of the higher-spin stochastic six-vertex model. This includes particle systems such as exclusion processes (q-TASEP [10,22,33,43] and other models [6,12,36,54]), directed polymers ([17,18,21,32,40,42]), and the stochastic six-vertex model [1,3,11,19,24].

1.4 Main Result

The study of the large scale behavior of passage times $T^{\mathrm{PP}}(n,m)$ was initiated in [13]. At large times, the fluctuations of the upper border of the percolation cluster (described by the height function $H_t(n)$) has GUE Tracy-Widom fluctuations on the scale $n^{1/3}$.

Theorem 1 ([13, Theorem 1.19]). *Fix parameters $a,b>0$. For any $\theta>0$ and $x\in\mathbb{R}$,*

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{H_{\tau(\theta)n}-\kappa(\theta)n}{\tilde{\rho}(\theta)n^{1/3}}\le x\right)=F_{\mathrm{GUE}}(x), \tag{2}$$

where F_{GUE} is the GUE Tracy-Widom distribution (see Definition 3) and $\kappa(\theta)$, $\tau(\theta)$, $\tilde{\rho}(\theta) = \frac{\kappa'(\theta)}{\tau'(\theta)}\rho(\theta)$ are functions defined in [13] by

$$\kappa(\theta) := \frac{\frac{1}{\theta^2} - \frac{1}{(a+\theta)^2}}{\frac{1}{(a+\theta)^2} - \frac{1}{(a+b+\theta)^2}},$$
$$\tau(\theta) := \frac{1}{a+\theta} - \frac{1}{\theta} + \kappa(\theta)\left(\frac{1}{a+\theta} - \frac{1}{a+b+\theta}\right) = \frac{a(a+b)}{\theta^2(2a+b+2\theta)},$$
$$\rho(\theta) := \left[\frac{1}{\theta^3} - \frac{1}{(a+\theta)^3} + \kappa(\theta)\left(\frac{1}{(a+b+\theta)^3} - \frac{1}{(a+\theta)^3}\right)\right]^{1/3}.$$

Note that as θ ranges from 0 to ∞, $\kappa(\theta)$ ranges from $+\infty$ to a/b and $\tau(\theta)$ ranges from $+\infty$ to 0.

Remark 1. In [13] the limit theorem is incorrectly stated as

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\min_{i\le n} T^{\mathrm{PP}}(i, \kappa(\theta)n) - \tau(\theta)n}{\rho(\theta)n^{1/3}} \le x\right) = F_{\mathrm{GUE}}(x),$$

but following the proof in [13, Section 6.1], we can see that the inequality and the sign of x should be reversed. Further, we have reinterpreted the limit theorem in terms of height function $H_t(n)$ instead of passage times $T^{\mathrm{PP}}(n, m)$ using the relation (1).

In this paper, we are interested in the fluctuations of $H_t(n)$ for large n but fixed time t. Let us scale θ in (2) above as

$$\theta = \left(\frac{na(a+b)}{2t}\right)^{1/3},$$

so that

$$\tau(\theta)n = t + O(n^{-1/3}).$$

Let us introduce constants

$$\lambda = \left(\frac{a(a+b)}{2t}\right)^{1/3}, \quad d = \frac{3a(a+b)}{2b\lambda}, \quad \sigma = \left(\frac{3a(a+b)\lambda}{2b^3}\right)^{1/3}. \tag{3}$$

Then, we have the approximations

$$\kappa(\theta)n = \frac{a}{b}n + dn^{2/3} + o(n^{4/9}),$$
$$\tilde{\rho}(\theta)n^{1/3} = \sigma n^{4/9} + o(n^{4/9}).$$

Thus, formally letting θ and n go to infinity in (2) suggests that for a fixed time t, it is natural to scale the height function as

$$H_t(n) = \frac{a}{b}n + dn^{2/3} + \sigma n^{4/9}\chi_n,$$

and study the asymptotics of the sequence of random variables χ_n.

Our main result is the following.

Theorem 2. *Fix parameters $a, b > 0$. For any $t > 0$ and $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{H_t(n) - \frac{a}{b}n - dn^{2/3}}{\sigma n^{4/9}} \leq x\right) = F_{GUE}(x),$$

where F_{GUE} is the GUE Tracy-Widom distribution.

Note that the heuristic argument presented above to guess the scaling exponents and the expression of constants d and σ is not rigorous, since Theorem 1 holds for fixed θ. Theorem 1 could be extended without much effort to a weak convergence uniform in θ for θ varying in a fixed compact subset of $(0, +\infty)$. However the case of θ and n simultaneously going to infinity requires more careful analysis. Indeed, for θ going to infinity very fast compared to n, Tracy-Widom fluctuations would certainly disappear as this would correspond to considering the height function at time $\tau(\theta)n \approx 0$, that is a simple random walk having Gaussian fluctuations on the $n^{1/2}$ scale. We explain in the next section how we shall prove Theorem 2.

The scaling exponents in Theorem 2 might seem unusual, although the preceding heuristic computation explains how they result from rescaling a model which has the usual KPZ scaling exponents. A similar situation occurs for scaling exponents of the height function of directed last passage percolation in thin rectangles [8,14] and for the free energy of directed polymers [4] under the same limit.

1.5 Outline of the Proof

Recall that given an integral kernel $\mathsf{K} : \mathbb{C}^2 \to \mathbb{C}$, its Fredholm determinant is defined as

$$\det(1 + \mathsf{K})_{L^2(\mathcal{C})} := \frac{1}{2\pi\mathbf{i}} \sum_{n=0}^{\infty} \frac{1}{n!} \int_{\mathcal{C}^n} \det[\mathsf{K}(x_i, x_j)]_{i,j=1}^n dx_1 ... dx_n.$$

To prove Theorem 2 we begin with the following Fredholm determinant formula for $\mathbb{P}(H_t(n) < m)$, and perform a saddle point analysis.

Theorem 3. **(**[13, Theorem 1.18]**).**

$$\mathbb{P}(H_t(n) < m) = \det(I - \mathsf{K}_n)_{\mathbb{L}^2(\mathcal{C}_0)},$$

where $\mathcal{C}_0$ is a small positively oriented circle containing 0 but not $-a - b$, and $\mathsf{K}_n : \mathbb{L}^2(\mathcal{C}_0) \to \mathbb{L}^2(\mathcal{C}_0)$ is defined by its integral kernel

$$\mathsf{K}_n(u, u') = \frac{1}{2\pi\mathbf{i}} \int_{1/2-i\infty}^{1/2+i\infty} \frac{e^{ts}}{s} \frac{g(u)}{g(s+u)} \frac{ds}{s+u-u'}, \quad \text{where} \tag{4}$$

$$g(u) = \left(\frac{a+u}{u}\right)^n \left(\frac{a+u}{a+b+u}\right)^m \frac{1}{u}. \tag{5}$$

Remark 2. Note that [13, Theorem 1.18] actually states $\mathbb{P}(H_t(n) < m) = \det(I + \mathsf{K}_n)_{\mathbb{L}^2(\mathcal{C}_0)}$, instead of $\det(I - \mathsf{K}_{t,n})_{\mathbb{L}^2(\mathcal{C}_0)}$ due to a sign mistake.

This result was proved in [13] by taking a zero-temperature limit of a similar formula for the Beta RWRE obtained using the Bethe ansatz solvability of q-Hahn TASEP and techniques from [16,22]. The integral (4) above is oscillatory and does not converge absolutely, but we may deform the contour so that it does. We will justify this deformation in Sect. 2.2.

Theorem 2 is proven in Sect. 2 by applying steep descent analysis to $\det(1 - \mathsf{K}_n)$, however the proofs of several key lemmas are deferred to later sections. The main challenge in proving Theorem 2 comes from the fact that, after a necessary change of variables $\omega = n^{-1/3}u$, the contours of the Fredholm determinant are being pinched between poles of the kernel K_n at $\omega = 0$ and $\omega = \frac{-a-b}{n^{1/3}}$ as $n \to \infty$. In order to show that the integral over the contour near 0 does not affect the asymptotics, we prove bounds for K_n near 0, and carefully choose a family of contours $\mathcal{C}_n$ on which we can control the kernel. This quite technical step is the main goal of Sect. 3. Section 4 is devoted to bounding the Fredholm determinant expansion of $\det(1-\mathsf{K}_n)_{L^2(\mathcal{C}_n)}$, in order to justify the use of dominated convergence in Sect. 2.

1.6 Other Interpretations of the Model

There are several equivalent interpretations of Bernoulli-exponential first passage percolation. We will present the most interesting here.

A Particle System on the Integer Line. The height function of the percolation cluster $H_t(n)$ is equivalent to the height function of an interacting particle system we call geometric jump pushTASEP, which generalizes pushTASEP (the $R = 0$ limit of PushASEP introduced in [23]) by allowing jumps of length greater than 1. This model is similar to Hall-Littlewood pushTASEP introduced in [36], but has a slightly different particle interaction rule.

Definition 2 (Geometric jump pushTASEP). *Let* $\mathrm{Geom}(q)$ *denote a geometric random variable with* $\mathbb{P}(\mathrm{Geom}(q) = k) = q^k(1-q)$. *Let* $1 \leq p_1(t) < p_2(t) < \ldots < p_i(t) < \ldots$ *be the positions of ordered particles in* $\mathbb{Z}_{\geq 1}$. *At time* $t = 0$ *the position* $n \in \mathbb{Z}_{\geq 0}$ *is occupied with probability* $b/(a+b)$. *Each particle has an independent exponential clock with parameter* a, *and when the clock corresponding to the particle at position* p_i *rings, we update each particle position* p_j *in increasing order of* j *with the following procedure. (*$p_i(t-)$ *denotes the position of particle* i *infinitesimally before time* t.*)*

- *If* $j < i$, *then* p_j *does not change.*
- p_i *jumps to the right so that the difference* $p_i(t) - p_i(t-)$ *is distributed as* $1 + \mathrm{Geom}(a/(a+b))$
- *If* $j > i$, *then*
  - *If the update for* $p_{j-1}(t)$ *causes* $p_{j-1}(t) \geq p_j(t-)$, *then* $p_j(t)$ *jumps right so that* $p_j(t) - p_{j-1}(t)$ *is distributed as* $1 + \mathrm{Geom}(a/(a+b))$.

- *Otherwise p_j does not change.*
- *All the geometric random variables in the update procedure are independent.*

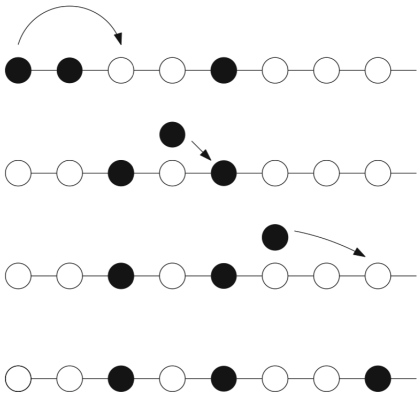


Fig. 3. This figure illustrates a single update for geometric jump pushTASEP. The clock corresponding to the leftmost particle rings, activating the particle. The first particle jumps 2 steps pushing the next particle and activating it. This particle jumps 1 step pushing the rightmost particle and activating it. The rightmost particle jumps 3 steps, and all particles are now in their original order, so the update is complete.

Another way to state the update rule is that each particle jumps with exponential rate a, and the jump distance is distributed as $1 + \mathrm{Geom}(a/(a+b))$. When a jumping particle passes another particle, the passed particle is pushed a distance $1 + \mathrm{Geom}(a/(a+b))$ past the jumping particle's ending location (see Fig. 3).

The height function $\overline{H}_t(n)$ at position n and time t is the number of unoccupied sites weakly to the left of n. If we begin with the distribution of $(n, H_t(n))$ in our percolation model, and rotate the first quadrant clockwise 45°, the resulting distribution is that of $(n, \overline{H}_t(n))$. The horizontal segments in the upper border of the percolation cluster correspond to the particle positions, thus

$$H_t(n) = p_t(n) - n = \sup\{k : \overline{H}_t(n+k) \geq k\}.$$

A direct translation of Theorem 2 gives:

Corollary 1. *Fix parameters $a, b > 0$. For any $t > 0$ and $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{p_t(n) - \left(\frac{a+b}{b}\right) n - dn^{2/3}}{\sigma n^{4/9}} \leq x \right) = F_{\mathrm{GUE}}(x),$$

where $F_{\mathrm{GUE}}(x)$ is the Tracy-Widom GUE distribution.

To the authors knowledge Corollary 1 is the first result in interacting particle systems showing Tracy-Widom fluctuations for the position of a particle at finite time.

Degenerations. If we set $b = 1, t' = t/a$, and $a \to 0$, then in the new time variable t' each particle performs a jump with rate 1 and with probability going to 1, each jump is distance 1, and each push is distance 1. This limit is pushTASEP on $\mathbb{Z}_{\geq 0}$ where every site is occupied by a particle at time 0. Recall that in pushTASEP, the dynamics of a particle are only affected by the (finitely many) particles to its left, so this initial data makes sense.

We can also take a continuous space degeneration. Let x be the spatial coordinate of geometric jump pushTASEP, and let $\exp(\lambda)$ denote an exponential random variable with rate λ. Choose a rate $\lambda > 0$, and set $b = \frac{\lambda}{n}, x' = x/n, a = \frac{n-\lambda}{n}$, and let $n \to \infty$. Then our particles have jump rate $\frac{n-\lambda}{n} \to 1$, jump distance $\frac{\mathrm{Geom}(1-\lambda/n)}{n} \to \exp(\lambda)$, and push distance $\frac{\mathrm{Geom}(1-\lambda/n)}{n} \to \exp(\lambda)$. This is a continuous space version of pushTASEP on $\mathbb{R}_{\geq 0}$ with random initial conditions such that the distance between each particle position p_i and its rightward neighbor p_{i+1} is an independent exponential random variable of rate λ. Each particle has an exponential clock, and when the clock corresponding to the particle at position p_i rings, an update occurs which is identical to the update for geometric jump pushTASEP except that each occurrence of the random variable $1 + \mathrm{Geom}(a/(a+b))$ is replaced by the random variable $\exp(\lambda)$.

A Benchmark Model for Travel Times in a Square Grid City. The first passage times of Bernoulli-exponential FPP can also be interpreted as the minimum amount of time a walker must wait at streetlights while navigating a city [29]. Consider a city, whose streets form a grid, and whose stoplights have i.i.d exponential clocks. The first passage time of a point (n, m) in our model has the same distribution as the minimum amount of time a walker in the city has to wait at stoplights while walking n streets east and m streets north. Indeed at each intersection the walker encounters one green stoplight with zero passage time and one red stoplight at which they must wait for an exponential time. Note that while the first passage time is equal to the waiting time at stoplights along the best path, the joint distribution of waiting times of walkers along several paths is different from the joint passage times along several paths in Bernoulli-exponential FPP.

1.7 Further Directions

Bernoulli-exponential FPP has several features that merit further investigation. From the perspective of percolation theory, it would be interesting to study how long it takes for the percolation cluster to contain all vertices in a given region, or how geodesics from the origin coalesce as two points move together.

From the perspective of KPZ universality, it is natural to ask: what is the correlation length of the upper border of the percolation kernel, and what is the joint law of the topmost few paths.

Under diffusive scaling limit, the set of coalescing simple directed random walks originating from every point of $\mathbb{Z}^2$ converges to the Brownian web [34,35]. Hence the set of all possible tributaries in our model converges to the Brownian web.

One may define a more involved set of coalescing and branching random walks which converges to a continuous object called the Brownian net ([41], [47], see also the review [46]). Thus, it is plausible that there exist a continuous limit of Bernoulli-Exponential FPP where tributaries follow Brownian web paths and branch at a certain rate at special points of the Brownian web used in the construction of the Brownian net.

After seeing Tracy-Widom fluctuations for the edge statistics it is natural to ask whether the density of vertices inside the river along a cross section is also connected to random matrix eigenvalues and whether a statistic of this model converges to the positions of the second, third, etc. eigenvalues of the Airy point process.

1.8 Notation and Conventions

We will use the following notation and conventions.

- $B_\varepsilon(x)$ will denote the open ball of radius $\varepsilon > 0$ around the point x.
- $\mathfrak{Re}[x]$ will denote the real part of a complex number x, and $\mathfrak{Im}[x]$ denotes the imaginary part.
- $\mathcal{C}$ and γ with any upper or lower indices will always denote an integration contour in the complex plane. K with any upper or lower indices will always represent an integral kernel. A lower index like γ_r, $\mathcal{C}_n$, or K_n will usually index a family of contours or kernels. An upper index such as γ^ε, $\mathcal{C}^\varepsilon$, or K^ε will indicate that we are intersecting our contour with a ball of radius ε, or that the integral defining the kernel is being restricted to a ball of radius ε.

2 Asymptotics

2.1 Setup

The steep descent method is a method for finding the asymptotics of an integral of the form

$$I_M = \int_{\mathcal{C}} e^{Mf(z)} dz,$$

as $M \to \infty$, where f is a holomorphic function and $\mathcal{C}$ is an integration contour in the complex plane. The technique is to find a critical point z_0 of f, deform the contour $\mathcal{C}$ so that it passes through z_0 and $\mathfrak{Re}[f(z)]$ decays quickly as z moves along the contour $\mathcal{C}$ away from z_0. In this situation $e^{Mf(z_0)}/e^{Mf(z)}$ has exponential decay in M. We use this along with specific information about our f and $\mathcal{C}$, to argue that the integral can be localized at z_0, i.e. the asymptotics of $\int_{\mathcal{C}\cap B_\varepsilon(z_0)} e^{Mf(z)} dz$ are the same as those of I_M. Then we Taylor expand f near z_0 and show that sufficiently high order terms do not contribute to the asymptotics. This converts the first term of the asymptotics of I_M into a simpler integral that we can often evaluate.

In Sect. 2.1 we will manipulate our formula for $\mathbb{P}(h(n) < m)$, and find a function f_1 so that the kernel K_n can be approximated by an integral of the form

$\int_{\lambda+\mathbf{i}\mathbb{R}} e^{n^{1/3}[f_1(z)-f_1(\omega)]}dz$. Approximating K_n in this way will allow us to apply the steep descent method to both the integral defining K_n and the integrals over $\mathcal{C}_0$ in the Fredholm determinant expansion.

For the remainder of the paper we fix a time $t > 0$, and parameters $a, b > 0$. All constants arising in the analysis below depend on those parameters t, a, b, though we will not recall this dependency explicitly for simplicity of notation.

We also fix henceforth

$$m = \left\lfloor \frac{a}{b}n + dn^{2/3} + n^{4/9}\sigma x \right\rfloor. \tag{6}$$

We consider K_n and change variables setting $\tilde{z} = s + u$, $d\tilde{z} = ds$ to obtain

$$\tilde{\mathsf{K}}_n(u, u') = \frac{1}{2\pi\mathbf{i}} \int_{1/2+u-\mathbf{i}\infty}^{1/2+u+\mathbf{i}\infty} \frac{e^{t(\tilde{z}-u)}}{(\tilde{z}-u)(\tilde{z}-u')} \frac{g(u)}{g(\tilde{z})} d\tilde{z}.$$

In the following lemma, we change our contour of integration in the $\tilde{z}$ variable so that it does not depend on u.

Lemma 1. *For every fixed n,*

$$\tilde{\mathsf{K}}_n(u, u') = \frac{1}{2\pi\mathbf{i}} \int_{n^{1/3}\lambda+\mathbf{i}\mathbb{R}} \frac{e^{t(\tilde{z}-u)}}{(\tilde{z}-u)(\tilde{z}-u')} \frac{g(u)}{g(\tilde{z})} d\tilde{z}.$$

Proof. Choose the contour $\mathcal{C}_0$ to have radius $0 < r < \min[1/4, \lambda]$. This choice of r means that we do not cross $\mathcal{C}_0$ when deforming the contour $1/2 + u + \mathbf{i}\mathbb{R}$ to $\lambda + \mathbf{i}\mathbb{R}$. In this region K is a holomorphic function, so this deformation does not change the integral provided that for M real,

$$\frac{1}{2\pi\mathbf{i}} \int_{1/2+u+\mathbf{i}M}^{n^{1/3}\lambda+\mathbf{i}M} \frac{e^{t(\tilde{z}-u)}}{(\tilde{z}-u)(\tilde{z}-u')} \frac{g(u)}{g(\tilde{z})} d\tilde{z} \xrightarrow[M\to\pm\infty]{} 0.$$

This integral converges to 0 because for all $\tilde{z} \in [n^{1/3}\lambda - \mathbf{i}M, 1/2 + u - \mathbf{i}M] \cup [n^{1/3}\lambda + \mathbf{i}M, 1/2 + u + \mathbf{i}M]$ we have

$$\left| \frac{1}{(\tilde{z}-u)(\tilde{z}-u')g(\tilde{z})} \right| \sim \frac{1}{M},$$

as $M \to \infty$.

Set

$$\tilde{h}_n(z) = -n \log\left(\frac{a+z}{z}\right) - m \log\left(\frac{a+z}{a+b+z}\right), \qquad \text{so that} \qquad e^{\tilde{h}_n(z)} = \frac{z}{g(z)}.$$

Then

$$\mathsf{K}_n(u, u') = \frac{1}{2\pi\mathbf{i}} \int_{n^{1/3}\lambda+\mathbf{i}\mathbb{R}} \frac{e^{t\tilde{z}+\tilde{h}_n(\tilde{z})}}{e^{tu+\tilde{h}_n(u)}} \frac{\tilde{z}}{u} \frac{d\tilde{z}}{(\tilde{z}-u)(\tilde{z}-u')}.$$

Now perform the change of variables

$$z = n^{-1/3}\tilde{z}, \omega = n^{-1/3}u, \omega' = n^{-1/3}u'.$$

If we view our change of variables as occuring in the Fredholm determinant expansion, then due to the $d\omega_i$s, we see that scaling all variables by the same constant does not change the Fredholm determinant $\det(1-\mathsf{K}_n)_{L^2(\mathcal{C})}$. Thus our change of variables gives

$$\mathsf{K}_n(\omega,\omega') = \frac{1}{2\pi\mathbf{i}}\int_{\lambda+\mathbf{i}\mathbb{R}} \frac{e^{n^{1/3}t(z-\omega)}}{(z-\omega)(z-\omega')} e^{h_n(z)-h_n(\omega)}\frac{z}{\omega}dz$$

where

$$h_n(z) = \tilde{h}_n(n^{1/3}z) = -n\log\left(\frac{a+n^{1/3}z}{n^{1/3}z}\right) - m\log\left(\frac{a+n^{1/3}z}{a+b+n^{1/3}z}\right).$$

Remark 3. The contour for ω, ω' becomes $n^{-1/3}\mathcal{C}_0$ after the change of variables, but $\mathsf{K}_n(\omega,\omega')$ is holomorphic in most of the complex plane. Examining of the poles of the integrand for $\mathsf{K}_n(\omega,\omega')$, we see that we can deform the contour for ω,ω' in any way that does not cross the line $\lambda+\mathbf{i}\mathbb{R}$, the pole at $-(a+b)/n^{1/3}$, or the pole at 0, without changing the Fredholm determinant $\det(I-\mathsf{K}_n)_{L^2(n^{-1/3}\mathcal{C}_0)}$.

Taylor expanding the logarithm in the variable n gives

$$h_n(z) = -n^{1/3}\left(\frac{a(a+b)}{2z^2} - \frac{bd}{z}\right) - n^{1/9}\left(\frac{-b\sigma x}{z}\right) + r_n(z).$$

Here $r_n(z) = \mathcal{O}(1)$ in a sense that we make precise in Lemma 3. The kernel can be rewritten as

$$\mathsf{K}_n(\omega,\omega') = \frac{1}{2\pi\mathbf{i}}\int_{\lambda+\mathbf{i}\mathbb{R}} \frac{\exp(n^{1/3}(f_1(z)-f_1(w)) + n^{1/9}(f_2(z)-f_2(\omega)) + (r_n(z)-r_n(\omega)))}{(z-\omega)(z-\omega')}\frac{z}{\omega}dz$$

where

$$f_1(z) = tz - \frac{a(a+b)}{2z^2} + \frac{bd}{z}, \qquad f_2(z) = \frac{b\sigma x}{z}. \tag{7}$$

We have approximated the kernel as an integral of the form $\int e^{n^{1/3}[f_1(z)-f_1(\omega)]}dz$. To apply the steep-descent method, we want to understand the critical points of the function f_1. We have

$$f_1'(z) = t + \frac{a(a+b)}{z^3} - \frac{db}{z^2}, \qquad f_1''(z) = -\frac{3a(a+b)}{z^4} + \frac{2bd}{z^3}, \qquad f_1'''(z) = \frac{12a(a+b)}{z^5} - \frac{6bd}{z^4}. \tag{8}$$

Where a, b are the parameters associated to the model. Let the constant λ be as defined in (3), then $0 = f_1'(\lambda) = f_1''(\lambda) = 0$, and

$$f_1'''(\lambda) = \frac{3a(a+b)}{\lambda^5} = 2\left(\frac{b\sigma}{\lambda^2}\right)^3 = 2\left(\frac{-f_2'(\lambda)}{x}\right)^3,$$

is a positive real number. σ is defined in Eq. (3).

Recall the definition of the Tracy-Widom GUE distribution, which governs the largest eigenvalue of a gaussian hermitian random matrix.

Definition 3. *The Tracy-Widom distribution's distribution function is defined as* $F_{GUE}(x) = \det(1 - \mathsf{K}_{\mathrm{Ai}})_{L^2(x,\infty)}$, *where* K_{Ai} *is the Airy kernel,*

$$\mathsf{K}_{\mathrm{Ai}}(s, s') = \frac{1}{2\pi\mathbf{i}} \int_{e^{-2\pi i/3}\infty}^{e^{2\pi\mathbf{i}/3}\infty} d\omega \frac{1}{2\pi\mathbf{i}} \int_{e^{-\pi\mathbf{i}/3}\infty}^{e^{\pi\mathbf{i}/3}\infty} dz \frac{e^{z^3/3 - zs}}{e^{\omega^3/3 - \omega s'}} \frac{1}{(z-\omega)}.$$

In the above integral the two contours do not intersect. We can think of the inner integral following the contour $(e^{-\pi\mathbf{i}/3}\infty, 1] \cup (1, e^{\pi\mathbf{i}/3}\infty)$, and the outer integral following the contour $(e^{-2\pi\mathbf{i}/3}\infty, 0] \cup (0, e^{2\pi\mathbf{i}/3}\infty)$. Our goal through the rest of the paper is to show that the Fredholm determinant $\det(I - \mathsf{K}_n)$ converges to the Tracy-Widom distribution as $n \to \infty$.

2.2 Steep Descent Contours

Definition 4. *We say that a path* $\gamma : [a, b] \to \mathbb{C}$ *is steep descent with respect to the function* f *at the point* $x = \gamma(0)$ *if* $\frac{d}{dt}\mathfrak{Re}[f(\gamma(t))] > 0$ *when* $t > 0$, *and* $\frac{d}{dt}\mathfrak{Re}[f(\gamma(t))] < 0$ *when* $t < 0$.

We say that a contour $\mathcal{C}$ is steep descent with respect to a function f at a point x, if the contour can be parametrized as a path satisfy the above definition. Intuitively this statement means that as we move along the contour $\mathcal{C}$ away from the point x, the function f is strictly decreasing.

In this section we will find a family of contours γ_r for the variable z and so that γ_r is steep descent with respect to $\mathfrak{Re}[f_1(z)]$ at the point λ, and study the behavior of $\mathfrak{Re}[f_1]$. The contours $\mathcal{C}_n$ for ω are constructed in Sect. 3.

Lemma 2. *The contour* $\lambda + \mathbf{i}\mathbb{R}$ *is steep descent with respect to the function* $\mathfrak{Re}[f_1]$ *at the point* λ.

Proof. We have that

$$\frac{d}{dy}\mathfrak{Re}[f_1(\lambda + \mathbf{i}y)] = -\mathfrak{Im}[f_1'(\lambda + \mathbf{i}y)] = -\mathfrak{Im}\left[t + \frac{a(a+b)}{(\lambda + \mathbf{i}y)^3} - \frac{bd}{\lambda + \mathbf{i}y}\right].$$

Now using the relation $2bd\lambda = 3a(a+b)$ and computing gives

$$\frac{d}{dy}\mathfrak{Re}[f_1(\lambda + \mathbf{i}y)] = \frac{-4a(a+b)y^3}{(\lambda^2 + y^2)^3}.$$

This derivative is negative when $y > 0$ and positive when $y < 0$.

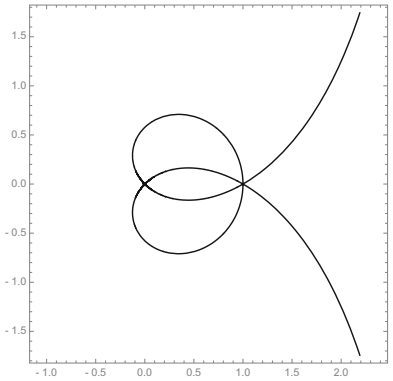


Fig. 4. The level lines of the function $\mathfrak{Re}[f_1(z)]$ at value $\mathfrak{Re}[f_1(\lambda)]$. In this image we take $a = b = t = 1$.

Now we describe the contour lines of $\mathfrak{Re}[f_1(z)]$ seen in Fig. 4. $\mathfrak{Re}[f_1]$ is the real part of a holomorphic function, so its level lines are constrained by its singularities, and because the singularities are not too complicated, we can describe its level lines. The contour lines of the real part of a holomorphic function intersect only at critical points and poles and the number of contour lines that intersect will be equal to the degree of the critical point or pole. We can see from the Taylor expansion of f_1 at λ, that there will be 3 level lines intersecting at λ with angles $\pi/6, \pi/2$, and $5\pi/6$. From the form of f_1, we see that there will be 2 level lines intersecting at 0 at angles $\pi/4$ and $3\pi/4$, and that a pair of contour lines will approach $\mathbf{i}\infty$ and $-\mathbf{i}\infty$ respectively with $\mathfrak{Re}[z]$ approaching $f_1(\lambda)/t$. This shows that, up to a noncrossing continuous deformation of paths, the lines in Fig. 4 are the contour lines $\mathfrak{Re}[f_1(z)] = f_1(\lambda)$. We can also see that on the right side of the figure, tz will be the largest term of $\mathfrak{Re}[f_1(z)]$, so our function will be positive. This determines the sign of $\mathfrak{Re}[f_1(z)]$ in the other regions.

Our contour $\lambda + \mathbf{i}\mathbb{R}$ is already steep descent, but we will deform the tails, so that we can use dominated convergence in the next section.

Definition 5. *For any $r > 0$, define the contour $\gamma_r = (e^{-2\pi\mathbf{i}/3}\infty, \lambda - r\mathbf{i}) \cup [\lambda - r\mathbf{i}, \lambda + r\mathbf{i}] \cup (\lambda + r\mathbf{i}, e^{2\pi\mathbf{i}/3}\infty)$ and $\gamma_r^\varepsilon = \gamma_r \cap B_\varepsilon(\lambda)$. These contours appear in Fig. 5.*

Because for any fixed n, we have $e^{h_n(z)} \to 1$ as $|z| \to \infty$, $\frac{z}{\omega(z-\omega)(z-\omega')}$ has linear decay in z, and $e^{n^{1/3}t(z-\omega)}$ has exponential decay in z, we can deform the vertical contour $\lambda + \mathbf{i}\mathbb{R}$ to the contour γ_r. Thus

$$\mathsf{K}_n(\omega, \omega') = \int_{\gamma_r} \frac{e^{n^{1/3}t(z-\omega)}}{(z-\omega)(z-\omega')} e^{h_n(z)-h_n(\omega)} \frac{z}{\omega} dz.$$

The function $\mathfrak{Re}[f_1]$ is still steep descent on the contour γ_r with respect to the point λ. Lemma 2 shows that $\mathfrak{Re}[f_1]$ is steep descent on the segment $[\lambda - r\mathbf{i}, \lambda + r\mathbf{i}]$,

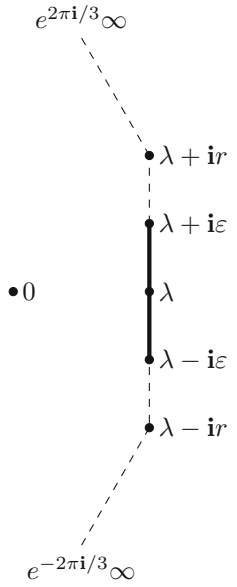


Fig. 5. The contour γ_r is the infinite piecewise linear curve formed by the union of the vertical segment and the two semi infinite rays, oriented from bottom to top. The bold portion of this contour near λ is γ_r^{ε}.

and on $(e^{-2\pi \mathbf{i}/3}\infty, \lambda - r\mathbf{i}) \cup (\lambda + r\mathbf{i}, e^{2\pi \mathbf{i}/3}\infty)$ we inspect $f_1'(z)$ and note that for z sufficiently large, the constant term t dominates the other terms. Because our paths are moving in a direction with negative real component the contour γ_r is steep descent.

Up to this point we have been concerned with contours being steep descent with respect to $\mathfrak{Re}[f_1]$, but the true function in our kernel is $\exp(n^{1/3}t(z-\omega) + h_n(z) - h_n(\omega))$. To show that γ_r is steep descent with respect to this function, we will need to control the error term $n^{1/3}tz + h_n(z) - n^{1/3}f_1(z) = n^{1/9}f_2(z) + r_n(z)$. The following lemma gives bounds on this error term away from $z = 0$.

Lemma 3. *For any $N, \varepsilon > 0$ there is a constant C depending only on ε, N such that*

$$|f_2(\omega)| \leq C \text{ and } |r_n(\omega)| \leq C, \tag{9}$$

for all $n \geq N$, and $\omega \geq \frac{|a+b|+\varepsilon}{N^{1/3}}$.

Similarly for any $\delta > 0$, there exists N_δ and C' depending only on δ, such that

$$|f_2'(\omega)| \leq C' \text{ and } |r_n'(\omega)| \leq C', \tag{10}$$

for all $n \geq N_\delta$, and ω satisfying $|\omega| \geq \delta$.

Lemma 3 is proved in Sect. 3.

At this point we have a contour γ_r for the variable z, which is steep descent with respect to $\mathfrak{Re}[f_1]$. We want to find a suitable contour for ω. The following lemma shows the existence of such a contour $\mathcal{C}_n$, where property (c) below takes the place of being steep descent. This lemma is fairly technical and its proof is the main goal of Sect. 3. To see why observe that the function $n^{1/3}f_1(\omega)$ does not approximate $n^{1/3}t\omega - h_n(\omega)$ well when ω is near 0. The fact that the contribution near 0 is negligible is nontrivial because the function $n^{1/3}t\omega - h_n(\omega)$ has poles at 0 and $\frac{-a-b}{n^{1/3}}$, and our contour $\mathcal{C}_n$ is being pinched between them; we will use Lemma 4 to show that the asymptotics of $\det(1-\mathsf{K}_n)_{L^2(\mathcal{C}_n)}$ are not affected by these poles

Lemma 4. *There exists a sequence of contours $\{\mathcal{C}_n\}_{n\geq N}$ such that:*

(a) For all n, the contour $\mathcal{C}_n$ encircles 0 counterclockwise, but does not encircle $(-a-b)n^{-1/3}$.
(b) $\mathcal{C}_n$ intersects the point λ at angles $-\pi/3$ and $-2\pi/3$.
(c) For all $\varepsilon > 0$, there exists $\eta, N_\varepsilon > 0$ such that for all $n > N_\varepsilon$, $\omega \in \mathcal{C}_n \setminus \mathcal{C}_n^\varepsilon$ and $z \in \gamma_r$, we have

$$\mathfrak{Re}[n^{1/3}t(z-\omega) + h_n(z) - h_n(\omega)] \leq -n^{1/3}\eta,$$

where $\mathcal{C}_n^\varepsilon = \mathcal{C}_n \cap B_\varepsilon(\lambda)$.
(d) There is a constant C such that for all $\omega \in \mathcal{C}_n$,

$$\mathfrak{Re}[n^{1/3}t(\lambda-\omega) + h_n(\lambda) - h_n(\omega)] \leq n^{1/9}C.$$

The next lemma allows us to control $\mathfrak{Re}[n^{1/3}tz + h_n(z)]$ on the contour γ_r.

Lemma 5. *For all $\varepsilon > 0$, and for sufficiently large r, there exists $C, N_\varepsilon > 0$, such that for all $\omega \in \mathcal{C}_n$, and $z \in \gamma_r \setminus \gamma_r^\varepsilon$, then*

$$\mathfrak{Re}[h_n(z) - h_n(\omega) + n^{1/3}t(z-\omega)] \leq -n^{-1/3}C.$$

Proof. We have already shown that γ_r is steep descent with respect to $f_1(z)$.

By Lemma 3, $|r_n| \leq C, |f_2| \leq Cn^{1/9}$ away from 0. We have

$$\begin{aligned} h_n(z) - h_n(\omega) + n^{1/3}t(z-\omega) =& n^{1/3}(f_1(z) - f_1(\omega)) + n^{1/9}(f_2(z) - f_2(\omega)) + (r_n(z) - r_n(\omega)) \\ \leq& n^{1/3}(f_1(z) - f_1(\omega)) + n^{1/9}C + C \leq n^{1/3}(f_1(z) - f_1(\omega) + \delta), \end{aligned}$$

for any sufficiently small $\delta > 0$. Because $f_1(z)$ is decreasing as we move away from λ, we have

$$n^{1/3}tz + h_n(z) < n^{1/3}t\lambda + h_n(\lambda) + Cn^{1/9}.$$

Thus by 3, we have that for all $\varepsilon > 0$ there exists C such that for $z \in \gamma_r \setminus \gamma_r^\varepsilon$,

$$\mathfrak{Re}[h_n(z) - h_n(\lambda) + n^{1/3}t(z-\lambda)] \leq -n^{1/3}C.$$

By Lemma 4(d), we have

$$\mathfrak{Re}[h_n(\lambda) - h_n(\omega) + n^{1/3}t(\lambda-\omega)] \leq n^{1/9}C,$$

for $\omega \in \mathcal{C}_n$. This completes the proof

2.3 Localizing the Integral

In this section we will use Lemmas 4 and 5 to show that the asymptotics of $\det(1-\mathsf{K}_n)_{L^2(\mathcal{C}_n)}$ do not change if we replace $\mathcal{C}_n$ with $\mathcal{C}_n^\varepsilon = \mathcal{C}_n \cap B_\varepsilon(\lambda)$, and replace the contour γ_r defining K_n with the contour $\gamma_r^\varepsilon = \gamma_r \cap B_\varepsilon(0)$.

First we change variables setting $z = \lambda + n^{-1/9}\overline{z}, \omega = \lambda + n^{-1/9}\overline{\omega}$, and $\omega' = \lambda + n^{-1/9}\overline{z}$.

Definition 6. *Define the contours* $\mathcal{D}_0 = [-\mathbf{i}\infty, \mathbf{i}\infty]$, *and* $\mathcal{D}_0^\delta = \mathcal{D}_0 \cap B_\delta(0)$. *(We will often use* $\delta = n^{1/9}\varepsilon$.*)*

Our change of variables applied to the kernel K_n^ε gives

$$\overline{\mathsf{K}}_n^\varepsilon(\overline{\omega},\overline{\omega}') = \frac{1}{2\pi\mathbf{i}}\int_{\mathcal{D}_0^{n^{1/9}\varepsilon}} \frac{1}{(\overline{z}-\overline{\omega})(\overline{z}-\overline{\omega}')}\frac{(\lambda+n^{-1/9}\overline{z})}{(\lambda+n^{-1/9}\overline{\omega})} e^{n^{1/3}f_1(\lambda+n^{-1/9}\overline{z})-f_1(\lambda+n^{-1/9}\overline{\omega})} \times e^{n^{1/9}f_2(\lambda+n^{-1/9}\overline{z})-f_2(\lambda+n^{-1/9}\overline{\omega})} e^{r_n(\lambda+n^{-1/9}\overline{z})-r_n(\lambda+n^{-1/9}\overline{\omega})} d\overline{z}. \tag{11}$$

Definition 7. *The contours* $\mathcal{C}_{-1}$ *and* $\mathcal{C}_{-1}^\varepsilon$ *are defined as* $\mathcal{C}_{-1} = (e^{-2\pi\mathbf{i}/3}\infty, -1) \cup [-1, e^{2\pi\mathbf{i}/3}\infty)$ *and* $\mathcal{C}_{-1}^\varepsilon = \mathcal{C}_{-1} \cap B_{n^{1/9}\varepsilon}(-1)$.

By changing variables, for each m we have

$$\int_{(\mathcal{C}_n^\varepsilon)^m} \det(\mathsf{K}_n^\varepsilon(\omega_i,\omega_j))_{i,j=1}^m d\omega_1 \ldots d\omega_m = \int_{(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} \det(\overline{\mathsf{K}}_n^\varepsilon(\overline{\omega}_i,\overline{\omega}_j))_{i,j=1}^m d\overline{\omega}_1 \ldots d\overline{\omega}_m.$$

This equality follows, because after rescaling the contour C_n^ε, we can deform it to the contour $\mathcal{C}_{-1}^{n^{1/9}\varepsilon}$ without changing its endpoints. The previous equality implies

$$\det(1-\mathsf{K}_n^\varepsilon)_{L^2(\mathcal{C}_\varepsilon^\varepsilon)} = \det(1-\overline{\mathsf{K}}_n^\varepsilon)_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})}.$$

We will make this change of variables often in the following arguments. Given a contour such as $\mathcal{C}_n$ or γ_r, we denote the contour after the change of variables by $\overline{\mathcal{C}}_n$ or $\overline{\gamma}_r$. Now we are ready to localize our integrals.

Proposition 1. *For any sufficiently small* $\varepsilon > 0$,

$$\lim_{n\to\infty} \det(1-\mathsf{K}_n(\omega,\omega'))_{L^2(\mathcal{C})} = \lim_{n\to\infty} \det(1-\mathsf{K}_n^\varepsilon(\omega,\omega'))_{L^2(\mathcal{C}_n^\varepsilon)},$$

where

$$\mathsf{K}_n^\varepsilon = \frac{1}{2\pi\mathbf{i}}\int_{\gamma_r^\varepsilon} \frac{e^{n^{1/3}t(z-\omega)+h_n(z)-h_n(\omega)}}{(z-\omega)(z-\omega')}\frac{z}{w}dz.$$

Proof. The proof will have two steps, and will use several lemmas that are proved in Sect. 4. In the first step we localize the integral in the z variable and show that $\lim_{n\to\infty}\det(1-\mathsf{K}_n)_{L^2(\mathcal{C}^\varepsilon)} = \lim_{n\to\infty}\det(1-\mathsf{K}_n^\varepsilon)_{L^2(\mathcal{C}^\varepsilon)}$ using dominated convergence. In order to prove this, we appeal to Lemmas 12 and 13 to show

that the Fredholm series expansions are indeed dominated. In the second step we localize the integral in the ω, ω' variables by using Lemma 14 to find an upper bound for $\det(1+K_n)_{L^2(\mathcal{C}_n)} - \det(1+K_n)_{L^2(\mathcal{C}_n^\varepsilon)}$. Then we appeal to Lemma 15 to show that this upper bound converges to 0 as $n \to \infty$.

Step 1: By Lemma 5, for any $\varepsilon > 0$, there exists a $C', N > 0$ such that if $\omega \in C_n$ and $z \in \gamma_r \setminus \gamma_r^\varepsilon$, then for all $n > N$,

$$\mathfrak{Re}[h_n(z) - h_n(\omega) + n^{1/3}t(z-\omega)] \leq -n^{1/3}C'.$$

We bound our integrand on $\gamma_r \setminus \gamma_r^\varepsilon$, $\omega, \omega' \in \mathcal{C}_n^\varepsilon$,

$$\left| \frac{e^{h_n(z)-h_n(\omega)+n^{1/3}t(z-\omega)}}{(z-\omega)(z-\omega')} \frac{z}{\omega} \right| \leq \frac{C}{\delta^2} z e^{-n^{1/3}C'} \xrightarrow[n\to\infty]{pointwise} 0.$$

(the δ^2 comes from the fact that $|z-\omega| \geq \delta$). By Lemma 3, there exists a $\eta > 0$ such that for sufficiently large n,

$$\left| \frac{e^{h_n(z)-h_n(\omega)+n^{1/3}t(z-\omega)}}{(z-\omega)(z-\omega')} \frac{z}{\omega} \right| < \left| \frac{e^{n^{1/3}(f_1(z)-f_1(\omega)+\eta)}}{(z-\omega)(z-\omega')} \frac{z}{\omega} \right|.$$

The linear term of $f_1(z)$ in (7) implies

$$\frac{1}{2\pi \mathbf{i}} \int_{\gamma_r} \left| \frac{e^{n^{1/3}(f_1(z)-f_1(\omega)+\eta)}}{(z-\omega)(z-\omega')} \frac{z}{\omega} \right| dz < \infty.$$

In the previous inequality we should write $|dz|$ instead of dz. We will often omit the absolute value in the $d\omega$ portion of the complex integral when the integrand is a positive real valued function.

So for each ω, ω', by dominated convergence

$$\frac{1}{2\pi \mathbf{i}} \int_{\gamma_r \setminus \gamma_r^\varepsilon} \frac{e^{h_n(z)-h_n(\omega)+n^{1/3}t(z-\omega)}}{(z-\omega)(z-\omega')} \frac{z}{\omega} dz \to 0 \quad \text{as} \quad n \to \infty,$$

So $\lim_{n\to\infty} \mathsf{K}_n^\varepsilon(\omega, \omega') = \lim_{n\to\infty} \mathsf{K}_n(\omega, \omega')$.

Now by Lemmas 12, and 13, both Fredholm determinant expansions $\det(1 - \mathsf{K}_n)_{L^2(\mathcal{C}^\varepsilon)}$ and $\det(1 - \mathsf{K}_n^\varepsilon)_{L^2(\mathcal{C}^\varepsilon)}$, are absolutely bounded uniformly in n. Thus we can apply dominated convergence to get

$$\lim_{n\to\infty} \det(1 - \mathsf{K}_n)_{L^2(\mathcal{C}^\varepsilon)} = \lim_{n\to\infty} \det(1 - \mathsf{K}_n^\varepsilon)_{L^2(\mathcal{C}^\varepsilon)}. \tag{12}$$

Step 2: In the expansion

$$\det(1 - \mathsf{K}_n)_{L^2(\mathcal{C}_n)} = \sum_{m=0}^{\infty} \frac{1}{m!} \int_{(\mathcal{C}_n)^m} \det(\mathsf{K}_n(\omega_i, \omega_j'))_{i,j=1}^n d\omega_1, ..., d\omega_m.$$

The mth term can be decomposed as the sum

$$\int_{(\mathcal{C}_n^\varepsilon)^m} \det(\mathsf{K}_n(\omega_i,\omega_j))_{i,j=1}^n d\omega_1 ... d\omega_m + \int_{\mathcal{C}_n^m \setminus (\mathcal{C}_n^\varepsilon)^m} \det(\mathsf{K}_n(\omega_i,\omega_j))_{i,j=1}^n d\omega_1 ... d\omega_m.$$

Lemma 14 along with Hadamard's bound on the determinant of a matrix in terms of it's row norms, implies that when $\omega_1 \in \mathcal{C}_n \setminus \mathcal{C}_n^\varepsilon$ and $\omega_2, ..., \omega_m \in \mathcal{C}^n$,

$$|\det(\overline{\mathsf{K}}_n(\omega_i,\omega_j))_{i,j=1}^m| \leq m^{m/2} M^{m-1/2} L_4 n^{4/9} e^{-n^{1/3}\eta} \to 0 \text{ as } n \to \infty. \tag{13}$$

Now let R be the maximum length of the paths $\mathcal{C}_n$. The rescaled paths $\overline{\mathcal{C}_n}$ will always have length less than $n^{1/9}R$. We have

$$\begin{aligned}
&\int_{\mathcal{C}_n^m \setminus (\mathcal{C}_n^\varepsilon)^m} |\det(\mathsf{K}_n(\omega_i,\omega_j))_{i,j=1}^m| d\omega_1 ... d\omega_m \\
&\qquad \leq m \int_{\mathcal{C}_n \setminus \mathcal{C}_n^\varepsilon} d\omega_1 \int_{\mathcal{C}_n^{m-1}} |\det(\mathsf{K}_n(\omega_i,\omega_j))_{i,j=1}^m| d\omega_2 ... d\omega_m \\
&\qquad \leq m \int_{\overline{\mathcal{C}}_n \setminus \overline{\mathcal{C}}_n^\varepsilon} d\overline{\omega}_1 \int_{\overline{\mathcal{C}}_n^{m-1}} |\det(\overline{\mathsf{K}}_n(\overline{\omega}_i,\overline{\omega}_j))_{i,j=1}^m| d\overline{\omega}_2 ... d\overline{\omega}_m \\
&\qquad \leq \int_{\overline{\mathcal{C}}_n \setminus \overline{\mathcal{C}}_n^\varepsilon} d\overline{\omega}_1 \int_{\overline{\mathcal{C}}_n^{m-1}} m^{m/2} M^{(m-1)/2} L_4 n^{4/9} e^{-n^{1/3}\eta} d\overline{\omega}_2 ... d\overline{\omega}_m \\
&\qquad \leq m (n^{1/9}R)^m m^{m/2} M^{(m-1)/2} L_4 n^{4/9} e^{-n^{1/3}\eta} \\
&\qquad \leq e^{-n^{1/3}\eta} (n^{1/9})^m m^{1+m/2} (MR)^m n^{4/9}.
\end{aligned} \tag{14}$$

The first inequality follows from symmetry of the integrand in the ω_i. In the second inequality, we change variables from ω_i to $\overline{\omega}_i$. In the third inequality we use the first inequality of (13). In the fourth inequality, we use the fact that the total volume of our multiple integral is less than $(n^{1/9}R)^m$. In the fifth inequality we rewrite and use $M^m > M^{(m-1)/2}$.

So we have

$$\begin{aligned}
&\sum_{m=1}^\infty \frac{1}{m!} \int_{\mathcal{C}_n^m \setminus (\mathcal{C}_n^\varepsilon)^m} |\det(\mathsf{K}_n(\omega_i,\omega_j))_{i,j=1}^m| d\omega_1 ... d\omega_m \\
&\qquad \leq \sum_{m=1}^\infty \frac{1}{m!} e^{-n^{1/3}\eta} (n^{1/9})^m m^{1+m/2} (MR)^m n^{4/9} \\
&\qquad = n^{4/9} e^{-n^{1/3}\eta} \sum_{m=1}^\infty \frac{1}{m!} (MRn^{1/9})^m m^{1+m/2}
\end{aligned} \tag{15}$$

Applying Lemma 15 with $C = MRn^{1/9}$ gives.

$$n^{4/9} e^{-n^{1/3}\eta} \sum_{m=1}^\infty \frac{1}{m!} (MRn^{1/9})^m m^{1+m/2} \leq n^{4/9} e^{-n^{1/3}} 16 (MRn^{1/9})^4 e^{2(MR)^2 n^{2/9}} \xrightarrow[n\to\infty]{} 0.$$

Thus

$$\lim_{n\to\infty} \det(1-\mathsf{K}_n)_{L^2(\mathcal{C}_n)} = \lim_{n\to\infty} \det(1-\mathsf{K}_n)_{L^2(\mathcal{C}_n^\varepsilon)}. \tag{16}$$

Combining (12) and (16) concludes the proof of Proposition 1.

2.4 Convergence of the Kernel

In this section we approximate $h_n(z)-h_n(\omega)+n^{1/3}t(z-\omega)$ by its Taylor expansion near λ, and show that this does not change the asymptotics of our Fredholm determinant.

Proposition 2. *For sufficiently small $\varepsilon > 0$,*

$$\lim_{n\to\infty} \det(1-\mathsf{K}_n^\varepsilon)_{L^2(\mathcal{C}_\varepsilon^\varepsilon)} = \lim_{n\to\infty} \det(1-\mathsf{K}_{(x)})_{L^2(\mathcal{C}_{-1})},$$

where

$$\mathsf{K}_{(x)}(\overline{u},\overline{u}') = \frac{1}{2\pi\mathbf{i}} \int_{D'} \frac{e^{s^3/3-xs}}{e^{u^3-xu}} \frac{dz}{(z-u)(z-u')},$$

and

$$D' = (e^{-\pi\mathbf{i}/3}\infty, 0) \cup [0, e^{\pi\mathbf{i}/3}\infty).$$

Proof. Let

$$\mathsf{K}(\overline{\omega},\overline{\omega}') = \frac{1}{2\pi\mathbf{i}} \int_{D'} \frac{d\overline{z}}{(\overline{z}-\overline{\omega})(\overline{z}-\overline{\omega}')} e^{f_1'''(\lambda)(\overline{z}^3-\overline{\omega}^3)/6+f_2'(\lambda)(\overline{z}-\overline{\omega})}, \tag{17}$$

We have seen in Sect. 2.3 that

$$\det(1-\mathsf{K}_n^\varepsilon(\omega,\omega'))_{L^2(\mathcal{C}_\varepsilon^\varepsilon)} = \det(1-\overline{\mathsf{K}}_n^\varepsilon(\overline{\omega},\overline{\omega}'))_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})}.$$

The proof will have two main steps. In the first step we use dominated convergence to show that

$$\lim_{n\to\infty} \det(1-\overline{\mathsf{K}}_n^\varepsilon(\overline{\omega},\overline{\omega}'))_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})} = \lim_{n\to\infty} \det(1-\overline{\mathsf{K}}_{(x)}(\overline{\omega},\overline{\omega}'))_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})}.$$

In the second step we control the tail of the Fredholm determinant expansion to show that

$$\lim_{n\to\infty} \det(1-\overline{\mathsf{K}}_{(x)}(\overline{\omega},\overline{\omega}'))_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})} = \det(1-\overline{\mathsf{K}}_{(x)}(\overline{\omega},\overline{\omega}'))_{L^2(\mathcal{C}_{-1})}.$$

In step 1 we will use Lemma 12 to establish dominated convergence.

Step 1: We have the following pointwise convengences

$$\frac{\lambda+n^{-1/9}\overline{z}}{\lambda+n^{-1/9}\overline{\omega}} \to 1,$$

and for $z = \lambda + n^{-1/9}\overline{z}, \omega = \lambda + n^{-1/9}\overline{\omega}$,

$$n^{1/3}(f_1(z) - f_1(\omega)) + n^{1/9}(f_2(z) - f_2(\omega)) + r_n(z) - r_n(\omega) \to \frac{1}{6}f_1'''(\lambda)(\overline{z}^3 - \overline{\omega}^3) + f_2'(\lambda)(\overline{z} - \overline{\omega}). \tag{18}$$

Because z is purely imaginary, for each $\overline{\omega}, \overline{\omega}'$, the exponentiating the right hand side of (18) gives a bounded function of $\overline{z}$ and $z/\omega \leq \frac{|\lambda+\varepsilon|}{|\lambda-\varepsilon|}$. The left hand side of (18) can be chosen to be within $\delta/n^{1/9}$ of the right hand side by choosing ε small by Taylor's theorem, because all the functions on the left hand side are holomorphic in $B_\varepsilon(\lambda)$. Thanks to the quadratic denominator $\frac{1}{(\overline{z}-\overline{\omega})(\overline{z}-\overline{\omega}')}$, we can apply dominated convergence to get

$$\overline{\mathsf{K}}_n^{\varepsilon}(\overline{\omega}, \overline{\omega}') \xrightarrow[n\to\infty]{pointwise} \frac{1}{2\pi\mathbf{i}} \int_{\mathbf{i}\mathbb{R}} \frac{d\overline{z}}{(\overline{z} - \overline{\omega})(\overline{z} - \overline{\omega}')} e^{f_1'''(\lambda)(\overline{z}^3 - \overline{\omega}^3)/6 + f_2'(\lambda)(\overline{z} - \overline{\omega})}. \tag{19}$$

Because the integrand on the right hand side of (19) has quadratic decay in $\overline{z}$, we can deform the contour from γ_0 to D' without changing the integral, so the right hand side is equal to $\mathsf{K}(\overline{\omega}, \overline{\omega}')$ from 17. Now by Lemma 12 we can apply dominated convergence to the expansion of the Fredholm determinant $\det(1 - \overline{\mathsf{K}}_n^{\varepsilon})_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})}$, to get

$$\lim_{n\to\infty} \det(1 - \overline{\mathsf{K}}_n^{\varepsilon})_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})} = \lim_{n\to\infty} \det(1 - \mathsf{K})_{L^2(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})}.$$

Step 2: Now we make the change of variables $s = -(f_2'(\lambda)/x)\overline{z}$, $u = -(f_2'(\lambda)/x)\overline{\omega}$, and $u' = -(f_2'(\lambda)/x)\overline{\omega}'$. Keeping in mind that $-2(f_2'(\lambda)/x)^3 = f_1'''(\lambda)$, we get

$$\mathsf{K}(\overline{\omega}, \overline{\omega}') = \mathsf{K}_{(x)}(u, u') = \frac{1}{2\pi\mathbf{i}} \int_{D'} \frac{e^{s^3/3 - xs}}{e^{u^3/3 - xu}} \frac{ds}{(s - u)(s - u')}.$$

Recall the expansion:

$$\det(1 - \mathsf{K}_{(x)})_{L^2(\mathcal{C}_{-1}^{\varepsilon})} = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \int_{\mathcal{C}_{-1}^m} \det(\mathsf{K}_{(x)}(\omega_i, \omega_j))_{i,j=1}^m d\omega_1 \ldots d\omega_m,$$

where $\mathcal{C}_{-1} = (e^{-2\pi\mathbf{i}/3}\infty, 1] \cup (1, e^{2\pi\mathbf{i}/3}\infty)$, and $\mathcal{C}_{-1}^m$ is a product of m copies of $\mathcal{C}_{-1}$.

$$|\det(1 - \mathsf{K}_{(x)})_{L^2(\mathcal{C}_{-1})} - \det(1 - \mathsf{K}_{(x)})_{L^2(\mathcal{C}_{-1}^{\varepsilon})}| \leq \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \int_{\mathcal{C}_{-1}^m \setminus (\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} |\det(\mathsf{K}_{(x)}(\omega_i, \omega_j))_{i,j=1}^m| d\omega_1 \ldots d\omega_m,$$

so to conclude the proof of the proposition, we are left with showing that

$$\sum_{m=0}^{\infty} \frac{1}{m!} \int_{\mathcal{C}_{-1}^m \setminus (\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} |\det(\mathsf{K}_{(x)}(\omega_i, \omega_j))_{i,j=1}^m| d\omega_1 \ldots d\omega_m \xrightarrow[n\to\infty]{} 0 \tag{20}$$

Note that

$$\int_{\mathcal{C}_{-1}^m \setminus (\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} |\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m| d\omega_1 ... d\omega_m \leq$$
$$m \int_{\mathcal{C}_{-1}\setminus \mathcal{C}_{-1}^{n^{1/9}\varepsilon}} \int_{\mathcal{C}_{-1}^{m-1}} |\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m| d\omega_1 ... d\omega_m.$$

Set

$$M_1 = \int_{D'} |\overline{z} e^{f_1'''(\lambda)\overline{z}^3/6 + f_2'(\lambda)\overline{z}}| d\overline{z} < \infty.$$

Then $\mathsf{K}_{(x)}(\omega,\omega') \leq M_1 e^{-|\omega|^3 - x|\omega|}$, and Hadamard's bound gives

$$|\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m| \leq m^{m/2} M_1^m \prod_{i=1}^m |e^{-\omega_i^3/3 + x\omega_i}|.$$

We have

$$\int_{\mathcal{C}_{-1}\setminus \mathcal{C}_{-1}^{n^{1/9}\varepsilon}} \int_{\mathcal{C}_{-1}^{m-1}} |\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m| d\omega_1 ... d\omega_m$$
$$\leq M_1 \int_{\mathcal{C}_{-1}\setminus \mathcal{C}_{-1}^{n^{1/9}\varepsilon}} \int_{\mathcal{C}_{-1}^{m-1}} \prod_{i=1}^m |e^{-\omega_i^3/3 + x\omega_i}| d\omega_1 ... d\omega_m$$
$$\leq m^{1+m/2} M_1^m M_2^{m-1} \int_{\mathcal{C}_{-1}\setminus \mathcal{C}_{-1}^{n^{1/9}\varepsilon}} |e^{-\omega_1^3 + x\omega_1}| d\omega_1, \tag{21}$$

where $M_2 = \int_{\mathcal{C}_{-1}} |e^{-\omega^3 - x\omega}| d\omega < \infty$ because $-\omega^3$ lies on the negative real axis. (21) goes to zero because $n^{1/9}\varepsilon \to \infty$. So

$$\int_{\mathcal{C}_{-1}\setminus \mathcal{C}_{-1}^{n^{1/9}\varepsilon}} \int_{\mathcal{C}_{-1}^{m-1}} \left|\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m\right| d\omega_1 ... d\omega_m \xrightarrow[n\to\infty]{} 0.$$

Note also that

$$\int_{\mathcal{C}_{-1}^m \setminus (\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} \left|\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m\right| d\omega_1 ... d\omega_m \leq \int_{\mathcal{C}_{-1}^m} |\det(\mathsf{K}_{(x)}(\omega_i,\omega_j))_{i,j=1}^m| d\omega_1 ... d\omega_m$$
$$\leq m^{1+m/2} M_1 M_2^m.$$

By Stirling's approximation

$$\sum_{m=0}^{\infty} \frac{1}{m!} m^{1+m/2} M_1^m M_2^m < \infty.$$

So by dominated convergence (20) holds which concludes the proof of Proposition 2.

2.5 Reformulation of the Kernel

Now we use the standard $\det(1+AB) = \det(1+BA)$ trick [17, Lemma 8.6] to identify $\det(1-\mathsf{K}_{(x)})_{L^2(\mathcal{C}_{-1})}$ with the Tracy-Widom cumulative distribution function.

Lemma 6. *For $x \in \mathbb{R}$,*

$$\det(1-\mathsf{K}_{(x)})_{L^2(\mathcal{C}_{-1})} = \det(1-\mathsf{K}_{\mathrm{Ai}})_{L^2(x,\infty)}.$$

Proof. First note that because $\mathfrak{Re}[z-\omega] > 0$ along the contours we have chosen, we can write

$$\frac{1}{z-\omega} = \int_{\mathbb{R}_+} e^{-\lambda(z-\omega)} d\lambda.$$

Now let $A : L^2(\mathcal{C}_{-1}) \to L^2(\mathbb{R}_+)$, and $B : L^2(\mathbb{R}_+) \to L^2(\mathcal{C}_{-1})$ be defined by the kernels

$$A(\omega,\lambda) = e^{-\omega^3/3+\omega(x+\lambda)}, \tag{22}$$

$$B(\lambda,\omega') = \int_{e^{-\pi\mathbf{i}/3}\infty}^{e^{\pi\mathbf{i}/3}\infty} \frac{dz}{2\pi\mathbf{i}} \frac{e^{z^3/3-z(x+\lambda)}}{z-\omega'}. \tag{23}$$

We compute

$$\begin{aligned} AB(\omega,\omega') &= \int_{\mathbb{R}_+} e^{-\omega^3/3+\omega(x+\lambda)} \int_{e^{-\pi\mathbf{i}/3}\infty}^{e^{\pi\mathbf{i}/3}\infty} \frac{dz}{2\pi\mathbf{i}} \frac{e^{z^3/3-z(x+\lambda)}}{z-\omega'} \\ &= \frac{1}{2\pi\mathbf{i}} \int_{e^{-\pi\mathbf{i}/3}\infty}^{e^{\pi\mathbf{i}/3}\infty} \frac{e^{z^3/3-zx}}{e^{\omega^3/3-\omega x}} \frac{dz}{(z-\omega)(z-\omega')} \\ &= \mathsf{K}_{(x)}(\omega,\omega'). \end{aligned}$$

Similarly,

$$BA(s,s') = \frac{1}{2\pi\mathbf{i}} \int_{e^{-2\pi\mathbf{i}/3}\infty}^{e^{2\pi\mathbf{i}/3}\infty} d\omega \frac{1}{2\pi\mathbf{i}} \int_{e^{-\pi\mathbf{i}/3}\infty}^{e^{\pi\mathbf{i}/3}\infty} dz \frac{e^{z^3/3-z(x+s)}}{e^{\omega^3/3-\omega(x+s')}} \frac{1}{(z-\omega)} = \mathsf{K}_{\mathrm{Ai}}(x+s,x+s').$$

Because both A and B are Hilbert-Schmidt operators, we have

$$\begin{aligned} \det(1-\mathsf{K}_{(x)})_{L^2(\mathcal{C})} = \det(1-AB)_{L^2(\mathbb{R}_+)} &= \det(1-BA)_{L^2(\mathbb{R}_+)} \\ &= \det(1-\mathsf{K}_{\mathrm{Ai}})_{L^2(x,\infty)} = F_{\mathrm{GUE}}(x). \end{aligned}$$

3 Constructing the Contour $\mathcal{C}_n$

This section is devoted to constructing the contours $\mathcal{C}_n$ and proving Lemma 4. We will prove several estimates for $n^{1/3}\omega + h_n(\omega)$; then we will construct the contour $\mathcal{C}_n$, and prove it satisfies the properties of Lemma 4. We begin by proving that we can approximate $n^{1/3}\omega + h_n(\omega)$ by $n^{1/3}f_1(\omega)$ away from 0.

3.1 Estimates Away from 0: Proof of Lemma 3

Both inequalities for $|f_2| = \frac{b\sigma x}{\omega}$ follow from the fact that f_2 and f_2' are bounded on $\mathbb{C} \setminus B_\varepsilon(0)$. Let $y = 1/\omega$, and let $m = n^{-1/9}$. Define the function $g(y, m) = r_n(\omega)$. First we prove (9). Note that $h_n(\omega)$ is holomorphic in y and m except when $n = \infty$, $n^{1/3}\omega = 0, -a - b$. By Taylor expanding $h_n(\omega)$, we see that $r_n(\omega) = g(y, m)$ is holomorphic in y and m, except at points (y, m) such that $n^{1/3}\omega = 0, -a - b$, in particular there is no longer a pole when $n = \infty$. Thus for any N, $g(y, m)$ is holomorphic with variables y and m, in the region $U = \{(y, m) : n > N, \omega > |a + b|/N^{1/3}\}$, because in this region $n^{1/3}\omega > |a + b|$. The region $U_\varepsilon = \{(y, m) : n > N, \omega \geq \frac{|a+b|+\varepsilon}{N^{1/3}}\}$ is compact in the variables y and m, and because $U_\varepsilon \subset U$, the function $g(y, m)$ is holomorphic in the region U_ε. Thus $g(y, m) = r_n(\omega)$ is bounded by a constant C in the region U_ε.

Now we prove (10). For any δ, pick an arbitrary ε and an N_δ large enough that $\frac{|a+b|+\varepsilon}{N_\delta^{1/3}} \leq \delta$. Because $g(y, m) = r_n(\omega)$ is holomorphic in the variables y and m in the compact set U_ε, the function $\frac{\partial}{\partial y} g(y, m) = -\omega^2 r_n'(\omega)$, is also holomorphic in y, m. So $|\omega^2 r_n'(\omega)| \leq C$ on U_ε. We rewrite as $|r_n'(\omega)| \leq C/|\omega|^2$, and this gives $|r_n'(\omega)| \leq \frac{C}{|\delta|^2} \leq C'$, on the set $U_\varepsilon \cap (\mathbb{N} \times B_\delta(0)^c)$. But by our choice of N_δ, we have $U_\varepsilon \cap (\mathbb{N} \times B_\delta(0)^c)$ is just the set $\{(y, m) : n \geq N_\delta, |\omega| \geq \delta\}$.

3.2 Estimates Near 0

The function $n^{1/3} f_1(\omega)$ only approximates $-n^{1/3}t\omega - h_n(\omega)$ well away from 0. In this section we give two estimates for $-n^{1/3}t\omega - h_n(\omega)$: one in Lemma 7 when ω is of order $n^{-1/3}$ and one in Lemma 8 when ω is of order $n^{\delta-1/3}$ for $\delta \in (0, 1/3)$. Together with Lemma 3 which gives an estimate when ω is of order 1, this will give us the tools we need to control $-n^{1/3}t\omega - h_n(\omega)$ along $\mathcal{C}_n$. First to prove the bound in Lemma 7, we choose a path which crosses the real axis at $-a$, between the poles at 0 and $-a - b$ before rescaling $\tilde{h}_n$ to h_n. We show that after the rescaling, we can bound $\mathfrak{Re}[-n^{-1/3}\omega - h_n(\omega)]$ on this path for small ω.

Lemma 7. *Fix any $c_0 > 1$ and let $s = c_0(a + b)$. For $C = \log\left(\sqrt{s^2 + a^2}\right) - \log(s) > 0$, we have*

$$\limsup_{n\to\infty} \frac{1}{n} \sup_{y\in[-s,s]} \mathfrak{Re}[h_n(\lambda) - h_n(\mathbf{i}n^{-1/3}y - n^{-1/3}a)] < -C.$$

Proof. Let $y \in [-s, s]$ and expand $e^{\mathfrak{Re}[h_n(\lambda)-h_n(iy-an^{-1/3})]}$ to get

$$\left(\frac{y}{\sqrt{y^2 + a^2}}\right)^n \left(\frac{y}{\sqrt{y^2 + b^2}}\right)^m \left(\frac{n^{1/3}\lambda}{n^{1/3}\lambda + a}\right)^n \left(\frac{a + b + n^{1/3}\lambda}{n^{1/3}\lambda + a}\right)^m.$$

The third factor is always less than 1. For sufficiently large n, the second factor times the fourth factor is less than 1, because $|y| \leq |s|$ while $n^{1/3}\lambda \to \infty$. We can bound the first factor by

$$\left| \frac{y}{\sqrt{y^2+a^2}} \right|^n \leq \left(\frac{s}{\sqrt{s^2+a^2}} \right)^n = e^{-nC},$$

with $C = \log\left(\sqrt{(s^2+a^2)}\right) - \log(s)$.

Next we will prove the estimate for ω of order $n^{\delta-1/3}$. In this proof we will consider ω of the form $\omega = -n^{-1/3}a + \mathbf{i}n^{\delta-1/3}c(a+b)$, choose c sufficiently large, then let $n \to \infty$. The largest term in the expansion of $-n^{-1/3}\omega - h_n(\omega)$ will be of order $\frac{n^{1-2\delta}}{c^2}$. We introduce the following definition to let us ignore the terms which are negligible compared to $\frac{n^{1-2\delta}}{c^2}$ uniformly in δ.

Definition 8. *Let A and B be functions depending on n and c, we say $A \sim_\delta B$ or A is δ-equivalent to B, if for sufficiently large c and n,*

$$|A - B| \leq \frac{n^{2/3-2\delta}}{c^2} M_1 + \frac{n^{1-3\delta}}{c^3} M_2 + \frac{n^{4/9-\delta}}{c} M_3.$$

for some constants M_1, M_2, M_3 independent of c and n.

Now we prove the estimate.

Lemma 8. *For all $\delta \in (0, 1/3)$, setting $\omega = -n^{-1/3}a + \mathbf{i}n^{\delta-1/3}c(a+b)$, gives*

$$\mathfrak{Re}[n^{1/3}t\omega + h_n(\omega)] \sim_\delta \mathfrak{Re}[n^{1/3}f_1(\omega)] \sim_\delta M\frac{n^{1-2\delta}}{c^2},$$

where $\sim_\delta$ is defined in Definition 8.

The proof of this Lemma 8 comes from Taylor expanding h_n and keeping track of the order of different terms with respect to n and c.

Proof. Recall that

$$h_n(\omega) = -n\log\left(1 + \frac{a}{n^{1/3}\omega}\right) + m\log\left(1 + \frac{b}{a + n^{1/3}\omega}\right). \tag{24}$$

For $|n^{1/3}\omega| > a$ and $|a + n^{1/3}\omega| > b$, we can Taylor expand in $n^{1/3}\omega$ to get

$$h_n(\omega) = -n\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{a}{n^{1/3}\omega}\right)^k + m\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{b}{a+n^{1/3}\omega}\right)^k.$$

Let $\omega = -n^{-1/3}a + \mathbf{i}n^{\delta-1/3}c(a+b)$ for $\delta \in (0, 1/3)$, so $|n^{1/3}\omega|, |a + n^{1/3}\omega| > n^\delta c(a+b) > c(a+b)$, for a constant c to be determined later. If $c > 2$, we have

$$\sum_{k=1}^{\infty} \left| \left(\frac{a}{n^{1/3}\omega}\right) \right|^k \leq \sum_{k=1}^{\infty} \left(\frac{b}{n^\delta c(a+b)}\right)^k \leq \frac{a}{n^\delta c(a+b)} \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^k \leq \frac{2a}{n^\delta c(a+b)} = \frac{n^{-\delta}}{c} M, \tag{25}$$

and

$$\sum_{k=1}^{\infty}\left|\left(\frac{b}{a+n^{1/3}\omega}\right)\right|^k \le \sum_{k=1}^{\infty}\left(\frac{a}{n^{\delta}c(a+b)}\right)^k \le \frac{a}{n^{\delta}c(a+b)}\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k = \frac{2a}{n^{\delta}c(a+b)} = \frac{n^{-\delta}}{c}M. \tag{26}$$

In what follows, we will use (25) or (26) when we say that an infinite sum is δ-equivalent to its first term.

We examine the first term in (24).

$$-n\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{a}{n^{1/3}\omega}\right)^k = -\left(\frac{a}{n^{1/3}\omega}\right)+\frac{1}{2}\left(\frac{a}{n^{1/3}\omega}\right)^2 - n\sum_{k=3}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{a}{n^{1/3}\omega}\right)^k,$$
$$\sim_{\delta} -\left(\frac{a}{n^{1/3}\omega}\right)+\frac{1}{2}\left(\frac{a}{n^{1/3}\omega}\right)^2.$$

where the $\delta-$equivalence follows because $\left|n\sum_{k=3}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{a}{n^{1/3}\omega}\right)^k\right| \le \frac{n^{1-3\delta}}{c^3}M$ for some M by (25).

Recall that

$$m\sum_{k=1}^{\infty}\left(\frac{b}{a+n^{1/3}\omega}\right)^k = \left[\left(\frac{a}{b}\right)n + dn^{2/3}+\sigma x n^{4/9}\right]\sum_{k=1}^{\infty}\left(\frac{b}{a+n^{1/3}\omega}\right)^k.$$

We decompose this series as three sums. First the $\left(\frac{a}{b}\right)n$ term gives

$$\frac{a}{b}n\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k =$$
$$n\left(\frac{a}{b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right) - \frac{n}{2}\left(\frac{a}{b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right)^2 + \frac{a}{b}n\sum_{k=3}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k$$
$$\sim_{\delta} n\left(\frac{a}{b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right) - \frac{n}{2}\left(\frac{b}{a+n^{1/3}\omega}\right)^2,$$

because $\left|-\frac{a}{b}n\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k\right| \le Mn^{1-3\delta}/c^3$ for some M. The second term is

$$dn^{2/3}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k = dn^{2/3}\left(\frac{b}{a+n^{1/3}\omega}\right) - dn^{2/3}\sum_{k=2}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k$$
$$\sim_{\delta} dn^{2/3}\left(\frac{b}{a+n^{1/3}\omega}\right)$$

because $\left|dn^{2/3}\sum_{k=2}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k\right| \le Mn^{2/3-2\delta}/c^2$ for some M. The third term is

$$n^{4/9}\sigma x\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\left(\frac{b}{a+n^{1/3}\omega}\right)^k \sim_{\delta} 0,$$

because the full sum $\left| n^{4/9}\sigma x \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{b}{a+n^{1/3}\omega}\right)^k \right| \leq \frac{Mn^{4/9-\delta}}{c}$ for some M. Now we have shown

$$-n\log\left(1+\frac{a}{n^{1/3}\omega}\right) \sim_\delta -n^{2/3}\frac{a}{\omega} + n^{1/3}\frac{a^2}{2\omega^2}, \tag{27}$$

$$m\log\left(1+\frac{b}{a+n^{1/3}\omega}\right) \sim_\delta n\left(\frac{a}{b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right) - n\left(\frac{a}{2b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right)^2 + dn^{2/3}\left(\frac{b}{a+n^{1/3}\omega}\right). \tag{28}$$

Adding (27) and (28) together yields

$$h_n(\omega) \sim_\delta -n^{2/3}\frac{a}{\omega} + n^{1/3}\frac{a^2}{2\omega^2} + n\left(\frac{a}{b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right) - n\left(\frac{a}{2b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right)^2 + dn^{2/3}\left(\frac{b}{a+n^{1/3}\omega}\right). \tag{29}$$

Adding the first and third terms from (29) gives the following cancellation.

$$-n^{2/3}\frac{a}{\omega} + n\left(\frac{a}{b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right) = -n^{2/3}\frac{a}{\omega} + n^{2/3}\frac{a}{\omega}\left[1 - \frac{a}{n^{1/3}\omega} + \sum_{k=2}^{\infty}(-1)^k\left(\frac{a}{n^{1/3}\omega}\right)^k\right] \sim_\delta -n^{1/3}\frac{a^2}{\omega^2},$$

thus

$$h_n(\omega) \sim_\delta -n^{1/3}\left(\frac{a^2}{2\omega^2}\right) - n\left(\frac{a}{2b}\right)\left(\frac{b}{a+n^{1/3}\omega}\right)^2 + dn^{2/3}\left(\frac{b}{a+n^{1/3}\omega}\right).$$

When we expand $\frac{b}{a+n^{1/3}\omega} = \frac{b}{n^{1/3}\omega} + \left(\frac{b}{n^{1/3}\omega}\right)\sum_{k=1}^{\infty}\left(\frac{-a}{n^{1/3}\omega}\right)^k$, we see that because $n^{1/3}\omega \sim_\delta n^\delta \mathbf{i}c(a+b)$, the sum is of order $1/c$ times the first term. So we can take only the first terms in our expansion, just as when we Taylor expand. This approximation leads the $n^{2/3}$ terms to cancel giving

$$h_n(\omega) \sim_\delta -n^{1/3}\left(\frac{a^2+ab}{2\omega^2}\right) + dn^{1/3}\left(\frac{b}{\omega}\right) \sim_\delta n^{1/3}\left(f_1(\omega) - t\omega\right).$$

This implies that $\mathfrak{Re}[n^{1/3}t\omega + h_n(\omega)] \sim_\delta \mathfrak{Re}[n^{1/3}f_1(\omega)]$. Completing the first δ-equivalence in the statement of Lemma 8.

Now observe that in

$$\mathfrak{Re}[n^{1/3}f_1(\omega)] = \mathfrak{Re}\left[n^{1/3}\left(t\omega - \frac{a(a+b)}{2\omega^2} + \frac{bd}{\omega}\right)\right],$$

we can bound the first term $|\mathfrak{Re}[n^{1/3}t\omega]| \leq n^{\delta}M$. We can bound the third term by $\mathfrak{Re}\left[n^{1/3}\frac{bd}{\omega}\right] \leq M\frac{n^{2/3-\delta}}{c}$. For the second term, we have $\left|\frac{a(a+b)}{2\omega^2}\right| \sim_\delta \left(\frac{a(a+b)}{2}\right)\left(\frac{n^{1-2\delta}}{c}\right)$. Thus

$$\mathfrak{Re}[n^{1/3}f_1(\omega)] \sim_\delta \left(\frac{a(a+b)}{2}\right)\left(\frac{n^{1-2\delta}}{c}\right).$$

This gives the second δ-equivalence in the statement of Lemma 8, and completes the proof.

3.3 Construction of the Contour $\mathcal{C}_n$

To construct the contour $\mathcal{C}_n$ we will start with lines departing from λ at angles $e^{\pm 2\pi \mathbf{i}/3}$, and with a vertical line $-n^{1/3}a + \mathbf{i}\mathbb{R}$. We will cut both these infinite contours off at specific values q and p respectively which allow us to use our estimates from the previous section on these contours. We will then connect these contours using the level set $\{z : Re[-f_1(z)] = -f_1(\lambda) - \varepsilon\}$. The rest of this section is devoted to finding the values p and q, showing that our explanation above actually produces a contour, and controlling the derivative of f_1 on the vertical segment near 0.

We note

$$f_1(\lambda) = 3t^{2/3}\left(\frac{a(a+b)}{2}\right)^{1/3} > 0, \tag{30}$$

and let

$$p = \sqrt{\frac{1}{3}\left(\frac{a(a+b)}{2t}\right)^{2/3}} > 0. \tag{31}$$

By simple algebra, we see that $\mathfrak{Re}[-f_1(\pm \mathbf{i}y)] < \mathfrak{Re}[-f_1(\lambda)] < 0$, when $y < p$, with equality at $y = p$.

Lemma 9. *$\frac{d}{dy}\mathfrak{Re}[-f_1(n^{-1/3}a + \mathbf{i}y)]$ is positive for $y \in [n^{-1/3}|a+b|, p]$, and negative for $y \in [-n^{-1/3}|a+b|, -p]$.*

Proof. We compute

$$\frac{d}{dy}\mathfrak{Re}[f_1(n^{-1/3}a + \mathbf{i}y)] = -\mathfrak{Im}(\mathfrak{Re}[f_1(n^{-1/3}a + \mathbf{i}y)]) \tag{32}$$

$$= -\frac{y^3 a(a+b)}{|n^{-1/3}a + \mathbf{i}y|^6} + \frac{a^2(a+b)n^{-2/3}y}{|n^{-1/3}a + \mathbf{i}y|^6} + \frac{3a^2(a+b)bn^{-1/3}y}{2b\lambda|n^{-1/3}a + \mathbf{i}y|^4}. \tag{33}$$

Note that for $y \in [n^{-1/3}|a+b|, p] \cup [-n^{-1/3}|a+b|, -p]$, we have $|n^{-1/3}a + \mathbf{i}y| \sim |y|$, so the first term of (33) is of order y^{-3} and the third term of (33) is of order $y^{-3}n^{-1/3}$. So for large enough n, the third term of (33) is very small compared to the first term. For $y = \pm n^{-1/3}|a+b|$, we have $|n^{-1}a(a+b)^4| = |y^3 a(a+b)| > |a(a+b)n^{-2/3}ay| = |a^2(a+b)^2 n^{-1/3}|$, and the derivative of $y^3 a(a+b)$ is larger

than the derivative of $a(a+b)n^{-2/3}ay$ for $y \in [n^{-1/3}|a+b|, p] \cup [-n^{-1/3}|a+b|, -p]$, so the first term of (33) has larger norm than the second term for $y \in [n^{-1/3}|a+b|, p] \cup [-n^{-1/3}|a+b|, -p]$. Thus the sign $\frac{d}{dy}\mathfrak{Re}[-f_1(n^{-1/3}a + \mathbf{i}y)]$ is determined by the first term of (33) in these intervals.

Now we can define the contour $\mathcal{C}_n$. We will give the definition, and then justify that it gives a well defined contour.

Definition 9. *Let $q > 0$ be a fixed real number such that for $0 < y \leq q$, $\frac{d}{dy}\mathfrak{Re}[-f_1(\lambda \pm ye^{\pm 2\pi\mathbf{i}/3})] < 0$. Let*

$$s = \max\Big\{\mathfrak{Re}[-f_1(\lambda + qe^{-2\pi\mathbf{i}/3})], \mathfrak{Re}[-f_1(\lambda + qe^{2\pi\mathbf{i}/3})], \\ \mathfrak{Re}[-f_1(n^{-1/3}(a - \mathbf{i}|a+b|))], \mathfrak{Re}[-f_1(n^{-1/3}(a + \mathbf{i}|a+b|))]\Big\}. \quad (34)$$

Let α be the contourline $\alpha = \{\omega : \mathfrak{Re}[-f_1(\omega)] = s\}$, and define the set

$$S_n = \{\lambda + ye^{\pm 2\pi\mathbf{i}/3} : 0 \leq y \leq q\} \cup \alpha \cup [-an^{-1/3} - \mathbf{i}p, -an^{-1/3} + \mathbf{i}p].$$

For sufficiently large n, define the path $\mathcal{C}_n$ to begin where α intersects $\{\lambda + ye^{-2\pi\mathbf{i}/3} : 0 \leq y \leq q\}$, follow the path $\{\lambda + ye^{-2\pi\mathbf{i}/3} : 0 \leq y \leq q\}$ toward $y = 0$, then follow the path $\{\lambda + ye^{2\pi\mathbf{i}/3} : 0 \leq y \leq q\}$ until it intersects α. $\mathcal{C}_n$ then follows α in either direction (pick one arbitrarily) until it intersects $[-an^{-1/3} - \mathbf{i}p, -an^{-1/3} + \mathbf{i}p]$ in the upper half plane. $\mathcal{C}_n$ then follows the path $[-an^{-1/3} - \mathbf{i}p, -an^{-1/3} + \mathbf{i}p]$ toward $-an^{-1/3} - \mathbf{i}p$ until it intersects α in the negative half plane. Then $\mathcal{C}_n$ follows α in either direction (pick one arbitrarily) until it reaches its starting point where it intersects $\{\lambda + ye^{-2\pi\mathbf{i}/3} : 0 \leq y \leq q\}$. See Fig. 6

We see that the q in Definition 9 exists by applying Taylor's theorem along with the fact that $f_1'''(\lambda) > 0$, and the $f_1'(\lambda) = f_1''(\lambda) = 0$.

Lemma 10. *The sets $\{\lambda + ye^{2\pi\mathbf{i}/3} : 0 \leq y \leq q\}$ and $\{\lambda + ye^{-2\pi\mathbf{i}/3} : 0 \leq y \leq q\}$ both intersect α at exactly one point. Lemmas 11 and 10 will show that $\mathcal{C}_n$ is a well defined contour.*

This follows from the definition of q and s.

Lemma 11. *There exists $N > 0$ such that for all $n > N$, the sets $[n^{-1/3} + \mathbf{i}n^{-1/3}|a+b|, n^{-1/3}a + p]$ and $[-an^{-1/3} - n^{-1/3}|a+b|, -an^{-1/3} - p]$ both intersect α exactly once.*

Proof. This is true because

$$\mathfrak{Re}[-f_1(-n^{-1/3}(a \pm \mathbf{i}|a+b|))] < \mathfrak{Re}[-f_1(\lambda)]. \quad (35)$$

by the contour lines in Fig. 4. This in addition to Lemma 9, and (30) implies the lemma.

3.4 Properties of the Contour $\mathcal{C}_n$: Proof of Lemma 4

Most of the work is used to prove part (c). The idea of this proof is to patch together the different estimates from the beginning of Sect. 3. Away from 0 we use Lemma 3 and the fact that the contour is steep descent near λ. Very near 0 on the scale $n^{-1/3}$ we use Lemma 7. Moderately near 0 we use Lemma 8, and our control of the derivative of f_1 on the vertical strip of $\mathcal{C}_n$ near 0. This last argument allows us to get bounds uniform in $\delta \in (0, 1/3)$ when ω is on the scale $n^{1/3-\delta}$.

Proof (Proof of Lemma 4). (a) and (b) follow from the definition of $\mathcal{C}_n$. By a slight modification of the proof of Lemma 4, we see that for $z \in \gamma_r$,

$$\mathfrak{Re}[h_n(z) - h_n(\lambda) + n^{1/3}t(z-\lambda)] \leq n^{1/9}C, \tag{36}$$

so to show (c) it suffices to show that for $\omega \in \mathcal{C}_n \setminus \mathcal{C}_n^{\varepsilon}$, we have

$$\mathfrak{Re}[h_n(\lambda) - h_n(\omega) + n^{1/3}t(\lambda-\omega)] \leq -n^{-1/3}\eta. \tag{37}$$

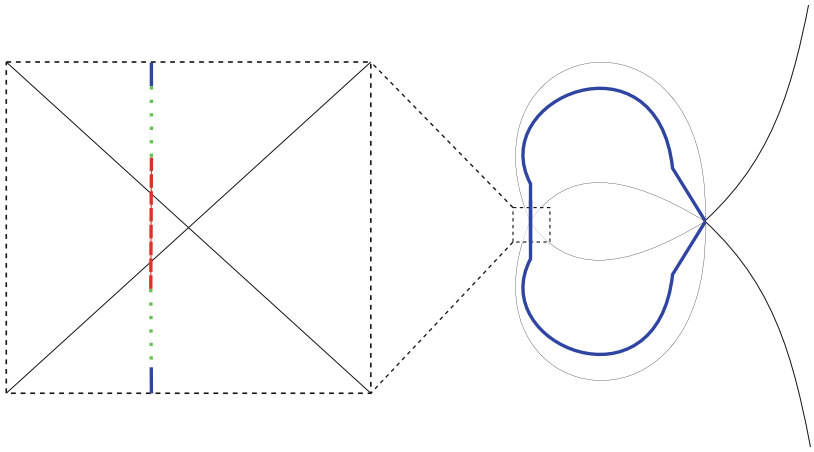


Fig. 6. $\mathcal{C}_n$ is the thick, colored piecewise smooth curve, the contour lines $\{z : \mathfrak{Re}[-f_1(z)] = f_1(\lambda)\}$ are the thin black curves. On the right side of the image we see $\mathcal{C}_n$ as a thick blue curve sandwiched between the contour lines. On the left we zoom in near 0 and see $\mathcal{C}_n$ pass the real axis as a dotted line to the left of zero. The contour lines meet at the point 0 on the left and λ on the right. We will now describe what section of the proof of Lemma 4 bounds $h_n(z) - h_n(\omega) + nt^{1/3}(z-\omega)$ on different portions of $\mathcal{C}_n$. The diagonal segments of $\mathcal{C}_n$ near λ are bounded in (ii). The curved segments in the right image, and the solid dark blue vertical segments at the top and bottom of the left image are bounded in (i). The dark red dashed segment that crosses the real axis in the left image is distance $O(n^{-1/3})$ from 0 and is bounded in (iii). The green dotted segments in the left image are distance $O(n^{\delta-1/3})$ from 0 for $\delta \in (0, 1)$ and are bounded in (iv).

Below we split the contour into 4 pieces and bound each separately. See Fig. 6.

(i) By Lemma 9 and the construction of $\mathcal{C}_n$, we have $\mathfrak{Re}[-f_1(\omega)] \leq s < \mathfrak{Re}[-f_1(\lambda)]$ for $\omega \in \mathcal{C}_n \setminus (\{\lambda + ye^{\pm 2\pi \mathbf{i}/3} : 0 \leq y \leq q\} \cup [n^{-1/3}(-a - \mathbf{i}|a + b|), n^{-1/3}(-a + \mathbf{i}|a + b|)])$. So we can apply Lemma 3 and the fact that f_2 is bounded outside a neighborhood of 0 to show that for any $c_1 < 0$, we have $\mathfrak{Re}[h_n(z) - h_n(\lambda) + n^{1/3}t(z - \lambda)] \leq -n^{-1/3}\eta$ for $\omega \in \mathcal{C}_n \setminus (\{\lambda + ye^{\pm 2\pi \mathbf{i}/3} : 0 \leq y \leq q\} \cup [-n^{-1/3}a - \mathbf{i}c_1|a + b|, -n^{-1/3}a + \mathbf{i}c_1|a + b|])$.

(ii) By the definition of q, The contour $\{\lambda + ye^{\pm 2\pi \mathbf{i}/3} : 0 \leq y \leq q\}$ is steep descent with respect to the function f_1 at the point λ, so we can apply Lemma 3 and the fact that f_2 is bounded outside a neighborhood of 0 to show $\mathfrak{Re}[h_n(z) - h_n(\lambda) + n^{1/3}t(z - \lambda)] \leq -n^{-1/3}\eta$ for $\omega \in \{\lambda + ye^{\pm 2\pi \mathbf{i}/3} : 0 \leq y \leq q\} \setminus B_\varepsilon(\lambda)$.

(iii) By Lemma 7, for any c_0, we have $\mathfrak{Re}[h_n(z) - h_n(\lambda) + n^{1/3}t(z - \lambda)] \leq -n^{-1/3}\eta$ for all $\omega \in [n^{-1/3}(-a - \mathbf{i}c_0|a + b|), n^{-1/3}(-a - \mathbf{i}c_0|a + b|)]$.

(iv) Now we bound the $\mathfrak{Re}[h_n(z) - h_n(\lambda) + n^{1/3}t(z - \lambda)]$ on the last piece of our contour $[n^{-1/3}(-a - \mathbf{i}c_0|a + b|), -n^{-1/3}a + \mathbf{i}c_1|a + b|] \cup [-n^{-1/3}a - \mathbf{i}c_1|a + b|, n^{-1/3}(-a - \mathbf{i}c_0|a + b|)]$. We will do this by fixing a constant $c > c_1$, and bounding the function on $\omega = n^{-1/3}a + \mathbf{i}n^{\delta - 1/3}c(a + b)$ for all pairs $n > N, \delta \in (0, 1/3)$ such that $n^{1/3} \leq c_1/c$.
By Lemma 8, we have that when $\omega = n^{-1/3}a + \mathbf{i}n^{\delta - 1/3}c(a + b)$, there exist constants M_1, M_2, M_3, such that

$$\mathfrak{Re}[n^{1/3}t\omega + h_n(\omega) - n^{1/3}f_1(\omega)] \leq \frac{n^{2/3-2\delta}}{c^2}M_1 + \frac{n^{1-3\delta}}{c^3}M_2 + \frac{n^{4/9-\delta}}{c}M_3,$$

and

$$f_1(\omega) \sim_\delta M\frac{n^{1-2\delta}}{c^2}.$$

First we consider the case when $\delta \in (0, 1/3 - \varepsilon)$. In this case, for any $r > 0$ we can choose c and N_r large enough that for all $n > N_r$,

$$\frac{\frac{n^{2/3-2\delta}}{c^2}M_1 + \frac{n^{1-3\delta}}{c^3}M_2 + \frac{n^{4/9-\delta}}{c}M_3}{\mathfrak{Re}[n^{1/3}f_1(\omega)]} < r/2,$$

uniformly for all $\delta \in (0, 1/3 - \varepsilon)$. In this case we also have that, by Lemma 3,

$$|\mathfrak{Re}[n^{1/3}tz + h_n(z)]| \leq n^{1/3}f_1(\lambda) + n^{1/9}f_2(\lambda) + C.$$

By potentially increasing N_r, we have that for all $n > N_r$

$$\frac{|\mathfrak{Re}[n^{1/3}tz + h_n(z)]|}{\mathfrak{Re}[n^{1/3}f_1(\omega)]} \leq r/2.$$

By Lemma 9 and (35), for all pairs n, δ such that $n^{\delta - 1/3} < c/c_1$, there is an $\eta > 0$ such that

$$\mathfrak{Re}[-f_1(\omega)] \leq \mathfrak{Re}[-f_1(\lambda)] - 2\eta < -2\eta.$$

setting $r = 1/2$ gives

$$\mathfrak{Re}[n^{1/3}t(z-\omega) + h_n(z) - h_n(\omega)] \leq \mathfrak{Re}[-n^{1/3}f_1(\omega)] + \frac{1}{2}\mathfrak{Re}[n^{-1/3}f_1(\omega)] < -\eta n^{1/3}.$$

Now we prove the case $\delta \in (1/3 - \varepsilon, 1/3)$. Note that in the expression

$$\mathfrak{Re}[n^{1/3}t\omega + h_n(\omega) - n^{1/3}f_1(\omega)] \leq \frac{n^{2/3-2\delta}}{c^2}M_1 + \frac{n^{1-3\delta}}{c^3}M_2 + \frac{n^{4/9-\delta}}{c}M_3,$$

when n is sufficiently large, we can bound the right hand side by $(M_1 + M_2)n^{3\varepsilon} \leq (r/2)n^{1/3}$ for any $r > 0$. We also have

$$|\mathfrak{Re}[n^{1/3}t\lambda - h_n(\lambda) - n^{1/3}f_1(\lambda)]| \leq n^{1/9}f_1(\lambda) + C \leq (r/2)n^{1/3}.$$

The first inequality comes from Lemma 3, and the second holds for large enough n. By Lemma 9 and (35), for all pairs n, δ such that $n^{\delta - 1/3} < c/c_1$, there is an $\eta > 0$ such that

$$\mathfrak{Re}[-f_1(\omega)] \leq \mathfrak{Re}[-f_1(\lambda)] - 2\eta < -2\eta.$$

Setting $r = \eta$ gives

$$\mathfrak{Re}[n^{1/3}t(\lambda-\omega) + h_n(\lambda) - h_n(\omega)] \leq n^{1/3}\mathfrak{Re}[f_1(\lambda) - f_1(\omega)] + n^{1/3}\eta \leq -\eta n^{1/3}.$$

The c_1 in part (i) can be chosen as small as desired, the c in part (iv) has already been chosen, and the c_0 in part (iv) can be chosen as large as desired. Choose $c_1 < c < c_0$ to complete the proof of (c).

Given inequalities (36) and (37), part (d) follows if we can show

$$\mathfrak{Re}[n^{1/3}t(\lambda-\omega) + h_n(\lambda) - h_n(\omega)],$$

for $\omega \in \mathcal{C}_n^\varepsilon$. Indeed this follows from Lemma 3 and the fact that the contour $\{\lambda + ye^{\pm 2\pi \mathbf{i}/3} : 0 \leq y \leq q\}$ is steep descent with respect to the function $\mathfrak{Re}[-f_1]$ at the point λ.

4 Dominated Convergence

In this section we carefully prove that the series expansion for $\det(1 - \mathsf{K}_n)_{L^2(\mathcal{C}_n^\varepsilon)}$ gives an absolutely convergent series of integrals bounded uniformly in n. This allows us to use dominated convergence when we localize the integral in Proposition 1, and again when we approximate the kernel by its Taylor expansion in Proposition 2. First we zoom in on a ball of radius epsilon and show that we can absolutely bound $\det(1 - \mathsf{K}_n^\varepsilon)_{L^2(\mathcal{C}_n^\varepsilon)}$ uniformly in n.

Lemma 12. *For any sufficiently small $\varepsilon > 0$, and sufficiently large r, there exists a function $\overline{F}(\overline{\omega}, \overline{\omega}')$, such that for all $\overline{\omega}, \overline{\omega}' \in \mathcal{C}_{-1}^{n^{1/9}\varepsilon}$, $z \in \mathcal{D}_0^{n^{1/9}\varepsilon}$, $n > N$ the integrand of $\overline{\mathsf{K}}_n^\varepsilon(\overline{\omega}, \overline{\omega}')$ in Eq. (11) is absolutely bounded by $\overline{F}(\overline{\omega}, \overline{\omega}', \overline{z})$, and*

$$\sum_{m=0}^{\infty} \int_{(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} \left| \det\left(\int_{\mathcal{D}_0^{n^{1/9}\varepsilon}} \overline{F}(\overline{\omega}_i, \overline{\omega}_j, \overline{z}) d\overline{z} \right)_{i,j=1}^{m} \right| d\overline{\omega}_1 ... d\overline{\omega}_m < \infty. \tag{38}$$

Proof. For $\overline{\omega}, \overline{\omega}' \in \mathcal{C}_{-1}^{\varepsilon}$, and $\overline{z} \in \mathcal{D}_0^{\varepsilon}$, we have

$$\left|\frac{\lambda + n^{-1/9}\overline{z}}{\lambda + n^{-1/9}\overline{\omega}}\right| \leq \left|\frac{\lambda+\varepsilon}{\lambda-\varepsilon}\right|,$$

and by Taylor approximation, we have the additional bounds

$$n^{1/3}(f_1(\lambda + n^{-1/9}\overline{z}) - f_1(\lambda + n^{-1/9}\overline{\omega})) \leq (f_1'''(\lambda) + \delta_1)(\overline{z}^3 - \overline{\omega}^3), \tag{39}$$

$$n^{1/9}(f_2(\lambda + n^{-1/9}\overline{z}) - f_2(\lambda + n^{-1/9}(\overline{\omega}))) \leq (f_2'(\lambda) + \delta_2)(\overline{z} - \overline{\omega}), \tag{40}$$

$$r_n(\lambda + n^{-1/9}\overline{z}) - r_n(\lambda + n^{-1/9}\overline{\omega}) \leq Cn^{-1/9}(\overline{z} - \overline{\omega}) \leq C\varepsilon \leq \delta_3. \tag{41}$$

Note that in these bounds we can make $\delta_1, \delta_2, \delta_3$ as small as desired by choosing ε small. Equations (39) and (40) follow from the fact that f_1, and f_2 are holomorphic in the compact set $\overline{B}_{\varepsilon}(\lambda)$. And Eq. (41) follows from Lemma 3. Note that along $\mathcal{D}_0$, z is purely imaginary, so (39), (40), and (41) show that the full exponential in the integrand in (11) is bounded above by

$$e^{2\delta_3} e^{-(f_1'''(\lambda)-\delta_1)\overline{\omega}^3 - (f_2'(\lambda)-\delta_2)\overline{\omega}}. \tag{42}$$

We choose ε small enough that $\delta_1 < f_1'''(\lambda)$, so that (42) has exponential decay as ω goes to ∞ in directions $e^{\pm 2\pi \mathbf{i}/3}$. Set

$$\overline{F}(\overline{\omega}, \overline{\omega}', \overline{z}) = \left|\left(\frac{\lambda+\varepsilon}{\lambda-\varepsilon}\right) e^{2\delta_3} e^{-(f_1'''(\lambda)-\delta_1)\overline{\omega}^3 - (f_2'(\lambda)-\delta_2)} \frac{1}{(\overline{z}+1)(\overline{z}+1)}\right|.$$

By the sentence preceding (42) $\overline{F}$ absolutely bounds the integrand of $\overline{\mathsf{K}}_n^{\varepsilon}$. Now set $L_1 = \frac{|\lambda+\varepsilon|}{|\lambda-\varepsilon|} e^{2\delta_3} \int_{\mathcal{D}_0} \frac{1}{(\overline{z}+1)(\overline{z}+1)} d\overline{z}$ so that $2e^{2\delta_3} \int_{\mathcal{D}_0} \frac{1}{(\overline{z}-\overline{\omega})(\overline{z}-\overline{\omega}')} d\overline{z} \leq L_1$. Then

$$\int_{\mathcal{D}_0^{\varepsilon}} \overline{F}(\overline{\omega}, \overline{\omega}', \overline{z}) \leq L_1 \left| e^{-(f_1'''(\lambda)-\delta_1)\overline{\omega}^3 - (f_2'(\lambda)-\delta_2)} \right|, \tag{43}$$

By Hadamard's bound

$$\left|\det\left(\int_{\mathcal{D}_0^{n^{1/9}\varepsilon}} \overline{F}(\overline{\omega}_i, \overline{\omega}_j', \overline{z}) d\overline{z}\right)_{i,j=1}^{m}\right| \leq m^{m/2} L_1^m \prod_{i=1}^{m} \left| e^{-(f_1'''(\lambda)-\delta)\overline{\omega}^3 - (f_2'(\lambda)-\delta)\overline{\omega}} \right|.$$

Now because $\delta_1 < f_1'''(\lambda)$, we can set

$$S = \int_{\mathcal{C}_{-1}^{n^{1/9}\varepsilon}} \left| e^{-(f_1'''(\lambda)-\delta)\overline{\omega}^3 - (f_2'(\lambda)-\delta)\overline{\omega}} \right| d\overline{\omega} < \infty.$$

Then we have the bound,

$$\int_{(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} \left|\det\left(\int_{\mathcal{D}_0^{n^{1/9}\varepsilon}} \overline{F}(\overline{\omega}_i, \overline{\omega}_j', \overline{z}) d\overline{z}\right)_{i,j=1}^{m}\right| d\overline{\omega}_1 \ldots d\overline{\omega}_m \leq m^{m/2} (SL_1)^m.$$

So by Stirling's approximation

$$\sum_{m=0}^{\infty} \int_{(\mathcal{C}_{-1}^{n^{1/9}\varepsilon})^m} \left| \det\left(\int_{\mathcal{D}_0^{n^{1/9}\varepsilon}} \overline{F}(\overline{\omega}_i, \overline{\omega}_j, \overline{z}) d\overline{z} \right)_{i,j=1}^{m} \right| d\overline{\omega}_1 ... d\overline{\omega}_m < \infty.$$

The next lemma completes our dominated convergence argument, by controlling the contribution to $\det(I - K_n)_{L^2(\mathcal{C}_n^{\varepsilon})}$ of $z \in \gamma_r \setminus \gamma_r^{\varepsilon}$.

Lemma 13. *For any sufficiently small $\varepsilon > 0$, and sufficiently large r, there is a function $\overline{G}(\overline{\omega}, \overline{\omega}', \overline{z})$, and a natural number N, such that for all $\overline{\omega}, \overline{\omega}' \in \overline{C}_n^{\varepsilon}$ and $\overline{z} \in \overline{\gamma}_r$, $n > N$, the integrand of $\overline{\mathsf{K}}_n(\overline{\omega}, \overline{\omega}')$ is absolutely bounded by $\overline{G}(\overline{\omega}, \overline{\omega}', \overline{z})$, and*

$$\sum_{m=0}^{\infty} \frac{1}{m!} \int_{(\overline{C}^{\varepsilon})^m} \left| \det\left(\int_{\overline{\gamma}_r} \overline{G}(\overline{\omega}_i, \overline{\omega}_j, \overline{z}) dz \right)_{i,j=1}^{m} \right| d\overline{\omega}_i ... d\overline{\omega}_j < \infty, \tag{44}$$

where $\overline{\gamma}_r$ and $\overline{C}_n^{\varepsilon}$ are the rescaled contours of γ_r and C_n^{ε} respectively.

Proof. Let $\overline{G} = \overline{F}$ for $z \in \gamma_r^{\varepsilon}$. We decompose the integral along γ_r in three parts: the integral along γ_r^{ε}, the integral along $(e^{-2\pi \mathbf{i}/3}\infty, -r) \cup (r, e^{2\pi \mathbf{i}/3}\infty)$ and the integral along $[-r, -\varepsilon] \cup [\varepsilon, r]$. For $z \in \gamma_r \setminus \gamma_r^{\varepsilon}$ we have the following bounds

$$\begin{aligned} |e^{n^{1/3}t(z-\omega)+h_n(z)-h_n(\omega)}| &\leq |e^{n^{1/3}(f_1(z)-f_1(\omega))+n^{1/9}C_2+C_3}| \\ &\leq |e^{n^{1/3}(f_1(z)-f_1(\omega)+\delta)}| \\ &\leq |e^{n^{1/3}(f_1(z)-f_1(\lambda)+\delta)}||e^{n^{1/3}(f_1(\lambda)-f_1(\omega))}|. \end{aligned} \tag{45}$$

Where the first inequality follows from Lemma 3. If we choose $\delta < \eta/2$, and recall that if $z \in \gamma_r \setminus \gamma_r^{\varepsilon}$, then $f_1(z) - f_1(\lambda) < -\eta$, so $f_1(z) - f_1(\lambda) + \delta < -\eta/2 < 0$. So if we wish we can bound (45) by either of the following expressions

$$|e^{n^{1/3}(f_1(\lambda)-f_1(\omega))}| \tag{46}$$

$$|e^{n^{1/9}(-tz+t\lambda)}||e^{n^{1/3}(f_1(\lambda)-f_1(\omega))}| \tag{47}$$

The bound (47) follows from the fact that we can choose r large enough so that $|f_1(z) + tz| \leq \delta$ outside $B_r(0)$. Then because the exponent in the first factor of (45) is negative, for large enough n we can remove the constant δ in return for reducing $n^{1/3}$ to $n^{1/9}$.

Now for $z \in [-r, -\varepsilon] \cup [\varepsilon, r]$, we have

$$\left| \frac{z}{\omega} \right| \leq \left| \frac{r+\lambda}{\lambda - \varepsilon} \right|, \qquad \left| \frac{1}{(\overline{z} - \overline{\omega})(\overline{z} - \overline{\omega}')} \right| \leq 1.$$

So for $z \in [-r, -\varepsilon] \cup [\varepsilon, r]$, we set

$$\overline{G}(\overline{\omega}, \overline{\omega}', \overline{z}) = \left| \frac{r+\lambda}{\lambda - \varepsilon} \right| \left| \frac{1}{(\overline{z} - \overline{\omega})(\overline{z} - \overline{\omega}')} \right| \left| e^{n^{1/3}(f_1(\lambda)-f_1(\omega))} \right|.$$

Using the above bounds and (46) we see that the integrand of $\overline{\mathsf{K}}_n$ is absolutely bounded by $\overline{G}$ in this region. Set $L_2 = \int_{i\mathbb{R}} \frac{r+\lambda}{\lambda-\varepsilon} \frac{1}{(\overline{z}+1)(\overline{z}+1)} d\overline{z}$ so that the integral of $\overline{G}$ on the rescaled contour of $[-r,-\varepsilon] \cup [\varepsilon, r]$ is bounded by $L_2 |e^{n^{1/3}(f_1(\lambda)-f_1(\omega))}|$.

For $z \in (e^{-2\pi \mathbf{i}/3}\infty, -r) \cup (r, e^{2\pi \mathbf{i}/3}\infty)$, we have

$$\left| \frac{1}{(\overline{z}-\overline{\omega})(\overline{z}-\overline{\omega}')} \right| \leq 1.$$

So for $z \in (e^{-2\pi \mathbf{i}/3}\infty, -r) \cup (r, e^{2\pi \mathbf{i}/3}\infty)$, we set

$$\overline{G}(\overline{\omega}, \overline{\omega}', \overline{z}) = \left| \frac{z}{\omega} \right| \left| e^{t(\lambda-\overline{z})} \right| \left| e^{(-f_1'''(\lambda)+\delta)\overline{\omega}} \right|.$$

Thus by (47), we can see that the integrand of $\overline{\mathsf{K}_n}$ is absolutely bounded by $\overline{G}$ in this region. Now let $L_3 = \int_{(e^{-2\pi \mathbf{i}/3}\infty, -r] \cup [r, e^{2\pi \mathbf{i}/3}\infty)} \left| \frac{\lambda+\overline{z}}{\lambda-\varepsilon} \right| |e^{t(\lambda-\overline{z})}| d\overline{z}$. For all n, the integral of $\overline{G}$ over the rescaled contour $(e^{-2\pi \mathbf{i}/3}\infty, -r] \cup [r, e^{2\pi \mathbf{i}/3}\infty)$ is bounded above by $L_3 |e^{(-f_1'''(\lambda)+\delta)\overline{\omega}^3}|$.

Let $\overline{\gamma}_r$ be the rescaled contour γ_r in the variable $\overline{z}$

$$\int_{\overline{\gamma}_r} \overline{G} d\overline{z} \leq (L_1 + L_2 + L_3) e^{(-f_1'''(\lambda)+\delta)\overline{\omega}^3} \leq L e^{(-f_1'''(\lambda)+\delta)\overline{\omega}^3}, \tag{48}$$

where the constant L comes from (43). Thus we have bounded $\int_{\overline{\gamma}_r} \overline{G} d\overline{z}$ by a constant times a term which has exponential decay as $\overline{\omega} \to e^{\pm 2\pi \mathbf{i}/3}\infty$. The same argument as in Lemma 12 shows that

$$\sum_{m=0}^{\infty} \frac{1}{m!} \int_{(C^\varepsilon)^m} \left| \det \left(\int_{\gamma_r^\varepsilon} G(\omega_i, \omega_j, z) dz \right)_{i,j=1}^{m} \right| d\omega_i ... d\omega_j < \infty.$$

Lemma 14. *Let $\omega_1 \in \mathcal{C}_n \setminus \mathcal{C}_n^\varepsilon$ and $\omega_2, .., \omega_m \in \mathcal{C}^n$. There exist positive constants $M, L_4, \eta > 0$ so that for sufficiently large n, we have*

$$|\overline{\mathsf{K}}_n(\overline{\omega}_i, \overline{\omega}_j)| \leq M$$

and

$$|\overline{\mathsf{K}}_n(\overline{\omega}_1, \overline{\omega}_i)| \leq L_4 n^{4/9} e^{-n^{1/3}\eta},$$

for all i, j.

Proof. By Lemma 4, for any $\varepsilon > 0$, there exists a $N, C > 0$, such that if $v \in \mathcal{C}_n \setminus \mathcal{C}_n^\varepsilon$, and $z \in \gamma_r$, then for all sufficiently large n, we have

$$\mathfrak{Re}[h_n(z) - h_n(\omega) + n^{1/3} t(z-\omega)] \leq -n^{1/3}\eta.$$

For $z \in \gamma_r$ and $\omega, \omega' \in \mathcal{C}_n \setminus \mathcal{C}_n^\varepsilon$, $n > N$ we have the following bounds:

$$\frac{1}{(z-\omega)(z-\omega')} \leq \left(\frac{2}{\varepsilon} \right)^2, \qquad \frac{1}{\omega} \leq \frac{n^{1/3}}{a},$$

and

$$|e^{n^{1/3}t(z-\omega)+h_n(z)-h_n(\omega)}| \leq |e^{n^{1/3}(f_1(z)-f_1(\omega)+\delta)}| \tag{49}$$

$$\leq |e^{n^{1/3}(f_1(z)-f_1(\lambda)}||e^{n^{1/3}(f_1(\lambda)-f_1(\omega)+\delta)}| \tag{50}$$

where (49) follows from (3) and the fact that f_2 is bounded away from 0. Note that for $z \in \gamma_r$, $|f_1(z)-f_1(\lambda)| \leq 0$, and for $\omega, \omega' \in \mathcal{C}_n \backslash \mathcal{C}_n^{\varepsilon}$, $f_1(\lambda)-f_1(\omega)+\delta < -\eta$, so (50) is bounded above by

$$|e^{(f_1(z)-f_1(\lambda)}||e^{-n^{1/3}\eta}|.$$

Thus if we set $L_4 = \frac{2^2}{a\varepsilon^2}\int_{\gamma_r}|z||e^{f_1(z)-f_1(\lambda)}|dz < \infty$, we get

$$|\mathsf{K}_n(\omega,\omega')| \leq L_4 n^{1/3} e^{-n^{1/3}\eta}.$$

So if we change the variable of integration to $d\overline{z} = n^{1/9}dz$ gives.

$$|\overline{\mathsf{K}}_n(\overline{\omega},\overline{\omega}')| \leq L_4 n^{4/9} e^{-n^{1/3}\eta} \qquad \text{for } \omega, \omega' \in \mathcal{C}_n \setminus \mathcal{C}_n^{\varepsilon} \tag{51}$$

Let $\omega_1 \in \mathcal{C}_n \setminus \mathcal{C}_n^{\varepsilon}$ and $\omega_2, .., \omega_m \in \mathcal{C}^n$, then for $i \neq 1$,

$$|\overline{\mathsf{K}}_n(\overline{\omega}_1,\overline{\omega}_i)| \leq L_4 n^{4/9} e^{-n^{1/3}\eta},$$

$$|\overline{\mathsf{K}}_n(\overline{\omega}_i,\overline{\omega}_j)| \leq \max[Le^{(-f_1'''(\lambda)+\delta)\overline{\omega}^3}, L_4 n^{4/9} e^{-n^{1/3}\eta}] \leq M. \tag{52}$$

The first equality follows from (48) and the second inequality holds for large n, when we set $M = \max[L_4, L]$ because $-f_1'''(\lambda)+\delta < 0$.

The last thing we need to complete the proof of Theorem 2 is to bound (15) from Proposition (2.3). We do so in the following lemma.

Lemma 15. *For any $C > 1$, we have*

$$\sum_{m=1}^{\infty} \frac{1}{m!} C^m m^{1+m/2} \leq 16C^4 e^{2C^2}.$$

Proof. We have

$$\frac{m^{1+m/2}}{m!} \leq \frac{m2^{m/2}}{(\lfloor m/2 \rfloor)!},$$

so that

$$\begin{aligned}\sum_{m=1}^{\infty} \frac{1}{m!} C^m m^{1+m/2} &\leq \sum_{m=1}^{\infty} \frac{m}{(\lfloor m/2 \rfloor)!}(2C^2)^{m/2} \\ &\leq \sum_{k=1}^{\infty} \frac{2k(2C^2)^k}{k!} + \sum_{k=1}^{\infty} \frac{(2k+1)(2C^2)^{k+1}}{k!} \\ &\leq 16C^4 e^{2C^2}.\end{aligned}$$

Acknowledgements. The authors thank Ivan Corwin for many helpful discussions and for useful comments on an earlier draft of the paper. The authors thank an anonymous reviewer for detailed and helpful comments on the manuscript. G. B. was partially supported by the NSF grant DMS:1664650. M. R. was partially supported by the Fernholz Foundation's "Summer Minerva Fellow" program, and also received summer support from Ivan Corwin's NSF grant DMS:1811143.

References

1. Aggarwal, A.: Current fluctuations of the stationary ASEP and six-vertex model. Duke Math. J. **167**(2), 269–384 (2018)
2. Aggarwal, A.: Dynamical stochastic higher spin vertex models. Selecta Math. **24**, 2659–2735 (2018)
3. Aggarwal, A., Borodin, A.: Phase transitions in the ASEP and stochastic six-vertex model. Ann. Probab. **47**(2), 613–689 (2019)
4. Auffinger, A., Baik, J., Corwin, I.: Universality for directed polymers in thin rectangles. arXiv preprint arXiv:1204.4445 (2012)
5. Auffinger, A., Damron, M., Hanson, J.: 50 Years of First-Passage Percolation, University Lecture Series, vol. 68. American Mathematical Society, Providence (2017)
6. Baik, J., Barraquand, G., Corwin, I., Suidan, T.: Facilitated exclusion process. In: Celledoni, E., Di Nunno, G., Ebrahimi-Fard, K., Munthe-Kaas, H.Z. (eds.) Abel Symposium 2016: Computation and Combinatorics in Dynamics, Stochastics and Control, pp. 1–35, Springer, Cham (2018)
7. Baik, J., Deift, P., Johansson, K.: On the distribution of the length of the longest increasing subsequence of random permutations. J. Am. Math. Soc. **12**(4), 1119–1178 (1999)
8. Baik, J., Suidan, T.M.: A GUE central limit theorem and universality of directed first and last passage site percolation. Int. Math. Res. Not. **2005**(6), 325–337 (2005)
9. Balázs, M., Rassoul-Agha, F., Seppäläinen, T.: Large deviations and wandering exponent for random walk in a dynamic beta environment. arXiv preprint arXiv:1801.08070 (2018)
10. Barraquand, G.: A phase transition for q-TASEP with a few slower particles. Stoch. Process. Appl. **125**(7), 2674–2699 (2015)
11. Barraquand, G., Borodin, A., Corwin, I., Wheeler, M.: Stochastic six-vertex model in a half-quadrant and half-line open asymmetric simple exclusion process. Duke Math. J. **167**(13), 2457–2529 (2018)
12. Barraquand, G., Corwin, I.: The q-Hahn asymmetric exclusion process. Ann. Appl. Probab. **26**(4), 2304–2356 (2016)
13. Barraquand, G., Corwin, I.: Random-walk in Beta-distributed random environment. Probab. Theor. Relat. Fields **167**(3–4), 1057–1116 (2017)
14. Bodineau, T., Martin, J.: A universality property for last-passage percolation paths close to the axis. Electron. Commun. Probab. **10**, 105–112 (2005)
15. Borodin, A.: On a family of symmetric rational functions. Adv. Math. **306**, 973–1018 (2017)
16. Borodin, A., Corwin, I.: Macdonald processes. Probab. Theor. Relat. Fields **158**(1–2), 225–400 (2014)
17. Borodin, A., Corwin, I., Ferrari, P.: Free energy fluctuations for directed polymers in random media in 1+1 dimension. Commun. Pure Appl. Math. **67**(7), 1129–1214 (2014)

18. Borodin, A., Corwin, I., Ferrari, P., Vető, B.: Height fluctuations for the stationary KPZ equation. Math. Phys. Anal. Geom. **18**(1), Art. 20, 95 (2015)
19. Borodin, A., Corwin, I., Gorin, V.: Stochastic six-vertex model. Duke Math. J. **165**(3), 563–624 (2016)
20. Borodin, A., Corwin, I., Petrov, L., Sasamoto, T.: Spectral theory for interacting particle systems solvable by coordinate Bethe ansatz. Commun. Math. Phys. **339**(3), 1167–1245 (2015)
21. Borodin, A., Corwin, I., Remenik, D.: Log-gamma polymer free energy fluctuations via a Fredholm determinant identity. Commun. Math. Phys. **324**(1), 215–232 (2013)
22. Borodin, A., Corwin, I., Sasamoto, T.: From duality to determinants for q-TASEP and ASEP. Ann. Probab. **42**(6), 2314–2382 (2014)
23. Borodin, A., Ferrari, P.: Large time asymptotics of growth models on space-like paths I: PushASEP. Electron. J. Probab. **13**, 1380–1418 (2008)
24. Borodin, A., Olshanski, G.: The ASEP and determinantal point processes. Commun. Math. Phys. **353**(2), 853–903 (2017)
25. Borodin, A., Petrov, L.: Higher spin six vertex model and symmetric rational functions. Selecta Math. **24**(2), 751–874 (2018)
26. Chaumont, H., Noack, C.: Fluctuation exponents for stationary exactly solvable lattice polymer models via a Mellin transform framework. ALEA Lat. Am. J. Probab. Math. Stat. **15**(1), 509–547 (2018)
27. Corwin, I.: The Kardar-Parisi-Zhang equation and universality class. Random Matrices Theory Appl. **1**(01), 1130001 (2012)
28. Corwin, I.: The q-Hahn Boson process and q-Hahn TASEP. Int. Math. Res. **2015**(14), 5577–5603 (2014)
29. Corwin, I.: Kardar-Parisi-Zhang universality. Not. AMS **63**, 230–239 (2016)
30. Corwin, I., Gu, Y.: Kardar-Parisi-Zhang equation and large deviations for random walks in weak random environments. J. Stat. Phys. **166**(1), 150–168 (2017)
31. Corwin, I., Petrov, L.: Stochastic higher spin vertex models on the line. Commun. Math. Phys. **343**(2), 651–700 (2016)
32. Corwin, I., Seppäläinen, T., Shen, H.: The strict-weak lattice polymer. J. Stat. Phys. **160**, 1027–1053 (2015)
33. Ferrari, P., Vető, B.: Tracy-Widom asymptotics for q-TASEP, **51**(4), 1465–1485 (2015)
34. Fontes, L., Isopi, M., Newman, C., Ravishankar, K.: The Brownian web. Proc. Nat. Acad. Sci. **99**(25), 15888–15893 (2002)
35. Fontes, L., Isopi, M., Newman, C., Ravishankar, K.: The Brownian web: characterization and convergence. Ann. Probab. **32**(4), 2857–2883 (2004)
36. Ghosal, P.: Hall-Littlewood-pushTASEP and its KPZ limit. arXiv preprint arXiv:1701.07308 (2017)
37. Hammersley, J.M., Welsh, D.J.A.: First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In: Proceedings of an International Research Seminar, Statistical Laboratory, University of California, Berkeley, California, pp. 61–110. Springer (1965)
38. Johansson, K.: Shape fluctuations and random matrices. Commun. Math. Phys. **209**(2), 437–476 (2000)
39. Kardar, M., Parisi, G., Zhang, Y.-C.: Dynamic scaling of growing interfaces. Phys. Rev. Lett. **56**, 889–892 (1986)
40. Krishnan, A., Quastel, J.: Tracy-Widom fluctuations for perturbations of the log-gamma polymer in intermediate disorder. Ann. Appl. Probab. **28**(6), 3736–3764 (2018)

41. Newman, C., Ravishankar, K., Schertzer, E.: Marking (1, 2) points of the Brownian web and applications, **46**(2), 537–574 (2010)
42. O'Connell, N., Ortmann, J.: Tracy-Widom asymptotics for a random polymer model with gamma-distributed weights. Electron. J. Probab. **20**(25), 1–18 (2015)
43. Orr, D., Petrov, L.: Stochastic higher spin six vertex model and q-TASEPs. Adv. Math. **317**, 473–525 (2017)
44. Povolotsky, A.: On the integrability of zero-range chipping models with factorized steady states. J. Phys. A **46**, 465205 (2013)
45. Prähofer, M., Spohn, H.: Universal distributions for growth processes in 1+1 dimensions and random matrices. Phys. Rev. Lett. **84**(21), 4882 (2000)
46. Schertzer, E., Sun, R., Swart, J.: The Brownian web, the Brownian net, and their universality. In: Advances in Disordered Systems, Random Processes and Some Applications, pp. 270–368 (2015)
47. Sun, R., Swart, J.M.: The Brownian net. Ann. Probab. **36**(3), 1153–1208 (2008)
48. Thiery, T.: Stationary measures for two dual families of finite and zero temperature models of directed polymers on the square lattice. J. Stat. Phys. **165**(1), 44–85 (2016)
49. Thiery, T., Le Doussal, P.: On integrable directed polymer models on the square lattice. J. Phys. A **48**(46), 465001 (2015)
50. Thiery, T., Le Doussal, P.: Exact solution for a random walk in a time-dependent 1D random environment: the point-to-point Beta polymer. J. Phys. A **50**(4), 045001 (2016)
51. Tracy, C.A., Widom, H.: A Fredholm determinant representation in ASEP. J. Stat. Phys. **132**(2), 291–300 (2008)
52. Tracy, C.A., Widom, H.: Integral formulas for the asymmetric simple exclusion process. Commun. Math. Phys. **279**(3), 815–844 (2008)
53. Tracy, C.A., Widom, H.: Asymptotics in ASEP with step initial condition. Commun. Math. Phys. **290**(1), 129–154 (2009)
54. Vető, B.: Tracy-Widom limit of q-Hahn TASEP. Electron. J. Probab. **20**, 1–22 (2015)

Hydrodynamics of the N-BBM Process

Anna De Masi[1], Pablo A. Ferrari[2(✉)], Errico Presutti[3], and Nahuel Soprano-Loto[2,3]

[1] Università degli Studi dell'Aquila, 67100 L'Aquila, Italy
demasi@univaq.it
[2] DM-FCEN, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina
{pferrari,nsloto}@dm.uba.ar
[3] Gran Sasso Science Institute, 67100 L'Aquila, Italy
errico.presutti@gssi.infn.it

Abstract. The Branching Brownian Motion (BBM) process consists of particles performing independent Brownian motions in $\mathbb{R}$, and each particle creating a new one at rate 1 at its current position. The newborn particles' increments and branchings are independent of the other particles. The N-BBM process starts with N particles and, at each branching time, the left-most particle is removed so that the total number of particles is N for all times. The N-BBM process has been originally proposed by Maillard, and belongs to a family of processes introduced by Brunet and Derrida. We fix a density ρ with a left boundary $\sup\{r \in \mathbb{R} : \int_r^\infty \rho(x)\mathrm{d}x = 1\} > -\infty$, and let the initial particles' positions be iid continuous random variables with density ρ. We show that the empirical measure associated to the particle positions at a fixed time t converges to an absolutely continuous measure with density $\psi(\cdot, t)$ as $N \to \infty$. The limit ψ is solution of a free boundary problem (FBP). Existence of solutions of this FBP was proved for finite time-intervals by Lee in 2016 and, after submitting this manuscript, Berestycki, Brunet and Penington completed the setting by proving global existence.

Keywords: Hydrodynamic limit · Free boundary problems · Branching Brownian Motion · Brunet-Derrida systems

1 Introduction

Brunet and Derrida [3] proposed a family of one dimensional processes with N branching particles with selection. Start with N particles with positions in $\mathbb{R}$. At each discrete time t, there are two steps. In the first step, each particle creates a number of descendants at positions chosen according to some density as follows: if a particle is located at position x, then its descendants are iid with distribution $Y + x$, where Y is a random variable with a given density. The second step is to keep the N right-most particles, erasing the left-most remaining ones.

The study of N branching Brownian motions has been proposed by Maillard [15,16] as a natural continuous time version of the previous process, also related

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 523–549, 2019.
https://doi.org/10.1007/978-3-030-15096-9_18

with the celebrated BBM process. The particles of the N-BBM process move as independent Brownian motions, and each particle at rate 1 creates a new particle at its current position. When a new particle is created, the left-most particle is removed. The number N of particles is then conserved.

The particles are initially distributed as independent random variables with an absolutely continuous distribution whose density is called ρ. Let $X_t = \{X_t^1, \dots, X_t^N\}$ be the set of positions of the N particles at time t. Here and in the sequel, we will consider multi-sets, allowing repetitions of elements. Denote by $|A|$ the cardinal of a multi-set A, elements being counted several times if repetitions occur. The empirical distribution induced by X_t is defined by

$$\pi_t^N[a,\infty) := \frac{1}{N}|X_t \cap [a,\infty)|,$$

the proportion of particles to the right of a at time t. Our main result is the following hydrodynamic limit.

Theorem 1. *Let ρ be a probability density function satisfying $L^* := \sup_r \left\{ \int_r^{-\infty} \rho(x)\mathrm{d}x = 1 \right\} > -\infty$. Let $X_0^1, \dots, X_0^N$ be independent identically distributed continuous random variables with density ρ. Let X_t be the positions at time t of the N-BBM process starting at $X_0 = \{X_0^1, \dots, X_0^N\}$. For every $t \geq 0$, there exists a density function $\psi(\cdot, t) : \mathbb{R} \to \mathbb{R}^+$ such that, for any $a \in \mathbb{R}$, we have*

$$\lim_{N\to\infty} \int_a^\infty \pi_t^N(\mathrm{d}r) = \int_a^\infty \psi(r,t)\mathrm{d}r \quad \textit{a.s. and in } L^1.$$

In Theorem 2 below, we identify the function $\psi(r,t)$ as the solution $u(r,t)$ of the following free boundary problem.

Free Boundary Problem (FBP). For $T > 0$, find $(u, L) \equiv \big((u(\cdot,t), L_t) : t \in [0,T]\big)$ satisfying the following conditions:

$$u_t = \frac{1}{2}u_{rr} + u \qquad \text{in } D_{L,T} := \{(r,t) : 0 < t < T, L_t < r\}; \tag{1}$$

$$u(r,0) = \rho(r), \qquad r \in \mathbb{R}; \tag{2}$$

$$L_0 = L^*;$$

$$u(L_t, t) = 0, \qquad t \in [0,T]; \tag{3}$$

$$\int_{L_t}^\infty u(r,t)\mathrm{d}r = 1, \quad t \in [0,T]. \tag{4}$$

Berestycki, Brunet and Derrida [1] propose a family of free boundary problems which include this one, and give, under certain conditions, an explicit relation between ρ and L.

Suppose (u, L) is a sufficiently regular solution of the previous free boundary problem. By the use of Leibniz's integral rule, we can take time derivative in condition (4) to obtain

$$\int_{L_t}^\infty u_t(r,t)\mathrm{d}r = 0 \tag{5}$$

for every $t \geq 0$. If in addition $u_r(r,t)$ vanishes as $r \to \infty$, we have

$$\begin{aligned}\frac{1}{2}u_r(L_t,t) &= -\int_{L_t}^{\infty}\frac{1}{2}u_{rr}(r,t)\mathrm{d}r \\ &= -\int_{L_t}^{\infty}\big[u_t(r,t)-u(r,t)\big]\mathrm{d}r \\ &= \int_{L_t}^{\infty}u(r,t)\mathrm{d}r = 1,\end{aligned}$$

so the space derivative at the boundary L is constantly 2. A further use of Leibniz's integral rule, now applied to u_r, gives

$$\begin{aligned}\frac{d}{dt}\Big[\int_{L_t}^{\infty}u_r(r,t)\mathrm{d}r\Big] &= -2\dot{L}_t + \int_{L_t}^{\infty}\partial_t u_r(r,t)\mathrm{d}r \\ &= -2\dot{L}_t + \frac{1}{2}u_{rr}(L_t,t) + \int_{L_t}^{\infty}u_r(r,t)\mathrm{d}r.\end{aligned}$$

This identity together with (5) gives us relation $\dot{L}_t = -\frac{1}{2}u_{rr}(L_t,t)$.

We define a classical solution to the FBP in the interval $[0,T]$ to be a pair (u,L) with $L \in C^1([0,T])$ and $u \in C\big(\overline{D}_{L,T}\big) \cap C^{2,1}(D_{L,T})$ satisfying conditions (1)–(4). For (u,L) a classical solution, the Brownian representation formula

$$\int_a^{\infty}u(r,t)\mathrm{d}r = e^t\int\rho(r)P(B_t^r > a, \tau^{r,L} > t)\mathrm{d}r \tag{6}$$

holds for every $t \in [0,T]$, where $(B_t^r : t \geq 0)$ is the Brownian motion with initial position $B_0^r = r$ and $\tau^{r,L} := \inf\{t \in [0,T] : B_t^r = L_t\}$. This representation formula is the key to prove the following theorem.

Theorem 2. *Suppose (u,L) is a classical solution to the FBP in the interval $[0,T]$. Then the function ψ defined in Theorem 1 coincides with u in that time interval: $\psi(\cdot,t) = u(\cdot,t), t \in [0,T]$.*

We observe that the previous result also gives a uniqueness criteria. About existence, Lee [14] proves that if ρ is such that $\rho \in C_c^2([L^*,\infty))$ and $\rho'(L^*) = 2$, there exists $T > 0$ such that the FBP has a classical solution in $[0,T]$. After submitting this manuscript, Berestycki, Brunet and Penington [2] proved global existence with general hypotheses over the initial condition and, in addition, gave an alternative proof of uniqueness that is independent to the one we give here. Summarizing, Theorem 2 together with the existence result imply that the empirical measure of the N-BBM process starting with iid particles with density ρ converges, in the sense of Theorem 1, to the solution to the FBP.

In the proof of Theorem 1, the presence of the free boundary at the left-most particle spoils usual hydrodynamic proofs. We overcome the difficulty by dominating the process from below and above by auxiliary more tractable processes, a kind of Trotter-Kato approximation. Durrett and Remenik [11] use an upper bound to show the analogous to Theorem 1 for a continuous time

Brunet-Derrida model. The approach with upper and lower bounds is used by three of the authors in [10], and by Carinci, De Masi, Giardina and Presutti in [6] and [5]; see the survey [7]. A further example is [9]. Maillard [16] used upper and lower bounds with a different scaling and scope. In [11], the left-most particle motion is increasing and has natural lower bounds. The lower bounds used in the mentioned papers do not work out-of-the-box here. We introduce labelled versions of the processes and a trajectory-wise coupling to prove the lower bound in Proposition 1 later.

Outline of the Paper. In Sect. 2, we introduce the elements of the proof of hydrodynamics, based on approximating barriers that will dominate the solution from above and below. In Sect. 3, we construct the coupling to show the dominations. In Sect. 4, we show the hydrodynamics for the barriers. Section 5 is devoted to the proof of the existence of the limiting density ψ. In Sect. 6, we prove Theorem 1. In Sect. 7, Theorem 2 is proven. Finally, in Sect. 8, we state a theorem for fixed N establishing the existence of a unique invariant measure for the process as seen from the left-most particle and a description of the traveling wave solutions for the FBP.

2 Domination and Barriers

We define the N-BBM process and the limiting barriers as functions of a ranked version of the BBM process.

Ranked BBM. Denote by $\{B_0^{1,1}, \dots, B_0^{N,1}\}$ the initial positions of N independent BBM processes. The descendants of the same initial particle will be called a family. Let N_t^i be the size of the i-th BBM family (starting at $B_0^{i,1}$) at time t. For $1 \le j \le N_t^i$, let $B_t^{i,j}$ be the position of the j-th member of the i-th family, ordered by birth time. Call (i, j) the *rank* of this particle, and denote

$$B_{[0,t]}^{i,j} := \text{trajectory of the } j\text{-th offspring with initial particle } i \text{ in the interval } [0, t],$$

with the convention that, before its birth time, the trajectory coincides with those of its ancestors. Define the *ranked* BBM as

$$\mathcal{B} := \big(B_{[0,\infty)}^{i,j} : i \in \{1, \dots, N\}, j \in \mathbb{N}\big). \tag{7}$$

We define the BBM process as the positions occupied by the particles at time t:

$$Z_t(\mathcal{B}) = \{B_t^{i,j} : 1 \le i \le N,\ 1 \le j \le N_t^i\}.$$

We drop the dependence on $\mathcal{B}$ in the notation when it is clear from the context.

N-BBM as Function of the Ranked BBM. Let $(\tau_n)_{n\in\mathbb{N}_0}$ be the branching times of the BBM process (we set $\mathbb{N}_0 = \{0, 1, \ldots\}$). We iteratively define $(\hat{L}_{\tau_n})_{n\in\mathbb{N}_0}$ and the N-BBM process $(X_t)_{t\geq 0}$ at the branching times. Let $X_0 = Z_0$, $\tau_0 = 0$ and $\hat{L}_0 = \min\{X_0^1, \ldots, X_0^N\}$. For $n \geq 1$, let

$$\hat{L}_{\tau_n} := a \in Z_{\tau_n} \text{ such that } \sum_{i=1}^{N}\sum_{j=1}^{N^i_{\tau_n}} \mathbf{1}\{B^{i,j}_{\tau_{n-1}} \in X_{\tau_{n-1}}, B^{i,j}_{\tau_n} \geq a\} = N$$

$$X_{\tau_n} := \{B^{i,j}_{\tau_n} : B^{i,j}_{\tau_{n-1}} \in X_{\tau_{n-1}} \text{ and } B^{i,j}_{\tau_n} \geq \hat{L}_{\tau_n}\},$$

with the convention that if the branching point at time τ_n is at $\hat{L}_{\tau_n}$ and $B^{i,j}_{\tau_n} = B^{i,j'}_{\tau_n}$, $j > j'$, are the two offsprings at that time, only $B^{i,j}_{\tau_n} \geq \hat{L}_{\tau_n}$ while we abuse notation by declaring $B^{i,j'}_{\tau_n} < \hat{L}_{\tau_n}$. Since superposition of particles occur at branching times with zero probability, $\hat{L}_{\tau_n}$ and X_{τ_n} are well defined for every n almost surely. The process

$$X_t(\mathcal{B}) := \{B^{i,j}_t : B^{i,j}_{\tau_n} \geq \hat{L}_{\tau_n} \text{ for all } \tau_n \leq t\}$$

is a version of the N-BBM process described in the introduction.

Stochastic Barriers. For every $\delta > 0$, we define the *stochastic barriers*. These are discrete-time processes denoted by $X^{\delta,-}_{k\delta}$ and $X^{\delta,+}_{k\delta}$, $k \in \mathbb{N}_0$, with initial configurations $X^{\delta,\pm}_0 = Z_0$. Iteratively, assume $X^{\delta,\pm}_{(k-1)\delta} \subset Z_{(k-1)\delta}$ is defined. The barriers at time $k\delta$ are selected points of $Z_{k\delta}$ of cardinal at most N that are defined as follows.

The Upper Barrier. The selected points at time $k\delta$ are the N right-most offsprings of the families of the selected points at time $(k-1)\delta$. The cutting point and the corresponding selected set at time $k\delta$ are

$$L^{N,\delta,+}_{k\delta} := a \in Z_{k\delta} \text{ such that } \sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} \mathbf{1}\{B^{i,j}_{(k-1)\delta} \in X^{\delta,+}_{(k-1)\delta}, B^{i,j}_{k\delta} \geq a\} = N$$

$$X^{\delta,+}_{k\delta} := \{B^{i,j}_{k\delta} : B^{i,j}_{(k-1)\delta} \in X^{\delta,+}_{(k-1)\delta} \text{ and } B^{i,j}_{k\delta} \geq L^{N,\delta,+}_{k\delta}\}. \tag{8}$$

(In Sect. 4, we introduce some deterministic barriers called $L^{\delta,+}_{k\delta}$, without the superscript N; please do not confuse them.) The number of particles in $X^{\delta,+}_{k\delta}$ is exactly N for all k.

The Lower Barrier. The selection is realized at time $(k-1)\delta$. Cut particles from left to right at time $(k-1)\delta$ until the largest possible number non bigger than N of particles is kept at time $k\delta$. While cutting the particles, we also cut all

their ancestors. The cutting point at time $(k-1)\delta$ and the resulting set at time $k\delta$ are given by

$$L^{N,\delta,-}_{(k-1)\delta} := \min\Big\{a \in X^{\delta,-}_{(k-1)\delta} : \sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} \mathbf{1}\big\{B^{i,j}_{(k-1)\delta} \in X^{\delta,-}_{(k-1)\delta} \text{ and } B^{i,j}_{(k-1)\delta} \geq a\big\} \leq N\Big\}$$

$$X^{\delta,-}_{k\delta} := \big\{B^{i,j}_{k\delta} : B^{i,j}_{(k-1)\delta} \in X^{\delta,-}_{(k-1)\delta} \text{ and } B^{i,j}_{(k-1)\delta} \geq L^{N,\delta,-}_{(k-1)\delta}\big\}. \tag{9}$$

Since entire families are cut at time $(k-1)\delta$, it is not always possible to keep exactly N particles at time $k\delta$. Nevertheless, for fixed δ, the number of particles in $X^{\delta,-}_{k\delta}$ is $N - M^{\delta}_{k\delta}$, where $M^{\delta}_{k\delta}/N$ goes to zero almost surely and in L^1. This will be made precise in Lemma 3.

We have the following expression for the barriers as a function of the ranked BBM $\mathcal{B}$:

$$\begin{aligned} X^{\delta,-}_{k\delta}(\mathcal{B}) &= \big\{B^{i,j}_{k\delta} : B^{i,j}_{\ell\delta} \geq L^{N,\delta,-}_{\ell\delta}, 0 \leq \ell \leq k-1\big\} \\ X^{\delta,+}_{k\delta}(\mathcal{B}) &= \big\{B^{i,j}_{k\delta} : B^{i,j}_{\ell\delta} \geq L^{N,\delta,+}_{\ell\delta}, 1 \leq \ell \leq k\big\}. \end{aligned} \tag{10}$$

Partial Order and Domination. Let X and Y be finite particle configurations (multi-sets) and define

$$X \preccurlyeq Y \quad \text{if and only if} \quad |X \cap [a,\infty)| \leq |Y \cap [a,\infty)| \quad \forall a \in \mathbb{R}. \tag{11}$$

In this case, we say that X is dominated by Y. If X and Y are random set of particles, we say that X is stochastically dominated by Y if there exists a random object $(\hat{X}, \hat{Y})$ (a coupling) such that its marginal distributions coincide respectively with the distributions of X and Y, and $\hat{X} \preccurlyeq \hat{Y}$ almost surely.

In Sect. 3, we prove the following dominations.

Proposition 1. *For every $\delta > 0$, there exists a random element $\big\{\big(\hat{X}^{\delta,-}_{k\delta}, \hat{X}_{k\delta}, \hat{X}^{\delta,+}_{k\delta}\big) : k \in \mathbb{N}_0\big\}$ satisfying the following conditions:*

1. *for every k, the marginals of $\big(\hat{X}^{\delta,-}_{k\delta}, \hat{X}_{k\delta}, \hat{X}^{\delta,+}_{k\delta}\big)$ have the same distributions as $X^{\delta,-}_{k\delta}$, $X_{k\delta}$ and $X^{\delta,+}_{k\delta}$;*
2. *for every k,*

$$\hat{X}^{\delta,-}_{k\delta} \preccurlyeq \hat{X}_{k\delta} \preccurlyeq \hat{X}^{\delta,-}_{k\delta} \text{ almost surely.} \tag{12}$$

In other words, $X_{k\delta}$ is stochastically between $X^{\delta,-}_{k\delta}$ and $X^{\delta,+}_{k\delta}$.

Deterministic Barriers. For integrable $u : \mathbb{R} \to \mathbb{R}_+$, the Gaussian kernel G_t is defined by

$$G_t u(a) := \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{-(a-r)^2/2t} u(r)\, \mathrm{d}r,$$

so that $e^t G_t \rho$ is solution of the equation $u_t = \frac{1}{2}u_{rr} + u$ with initial condition ρ. For $m > 0$, the *cut operator* C_m is defined by

$$C_m u(a) := u(a)\mathbf{1}\left\{\int_a^\infty u(r)\mathrm{d}r < m\right\}, \tag{13}$$

so that $C_m u$ has total mass $\left(\int u\right) \wedge m$. For $\delta > 0$ and $k \in \mathbb{N}_0$, define the upper and lower barriers $S_{k\delta}^{\delta,\pm}\rho$ at time δk as follows:

$$S_0^{\delta,\pm}\rho := \rho; \qquad S_{k\delta}^{\delta,-}\rho := \left(e^\delta G_\delta C_{e^{-\delta}}\right)^k \rho; \qquad S_{k\delta}^{\delta,+}\rho := \left(C_1 e^\delta G_\delta\right)^k \rho. \tag{14}$$

To obtain the upper barrier $S_\delta^{\delta,+}\rho$, first diffuse&grow for time δ, and then cut mass from the left to keep mass 1. To get the lower barrier $S_\delta^{\delta,-}\rho$, first cut mass from the left to keep mass $e^{-\delta}$, and then diffuse&grow for time δ (obtaining mass 1 again). Iterate to get the barriers at times $k\delta$. Since $\int e^\delta G_\delta u = e^\delta \int u$, we have $\int S_{k\delta}^{\delta,\pm}\rho = \int \rho = 1$ for all k.

Hydrodynamics of δ-barriers. In Sect. 4, we prove that, for fixed δ, the empirical measures converge as $N \to \infty$ to the macroscopic barriers:

Theorem 3. *Let $\pi_{k\delta}^{N,\delta,\pm}$ be the empirical measures associated to the stochastic barriers $X_{k\delta}^{\delta,\pm}$ with initial configuration X_0. Then, for any $a \in \mathbb{R}$, $\delta > 0$ and $k \in \mathbb{N}_0$,*

$$\lim_{N\to\infty} \int_a^\infty \pi_{k\delta}^{N,\delta,\pm}(\mathrm{d}r) = \int_a^\infty S_{k\delta}^{\delta,\pm}\rho(r)\mathrm{d}r \quad \textit{a.s. and in } L^1.$$

The same is true if we substitute $X_{k\delta}^{\delta,\pm}$ by the coupling marginals $\hat{X}_{k\delta}^{\delta,\pm}$ of Proposition 1.

Convergence of Macroscopic Barriers. For integrable $u, v : \mathbb{R} \to \mathbb{R}^+$, we write

$$u \preccurlyeq v \quad \text{iff} \quad \int_a^\infty u(r)\mathrm{d}r \le \int_a^\infty v(r)\mathrm{d}r \quad \forall a \in \mathbb{R}.$$

In Sect. 5, we fix t and take $\delta = t/2^n$ to prove that, for the order $\preccurlyeq$, the sequence $S_t^{t/2^n,-}\rho$ is increasing, the sequence $S_t^{t/2^n,+}\rho$ is decreasing, and $\|S_t^{t/2^n,+}\rho - S_t^{t/2^n,-}\rho\|_1 \xrightarrow[n\to\infty]{} 0$. As a consequence, we get the following theorem.

Theorem 4. *There exists a continuous function called $\psi(r,t)$ such that, for any $t > 0$,*

$$\lim_{n\to\infty} \|S_t^{t/2^n,\pm}\rho - \psi(\cdot,t)\|_1 = 0.$$

The function ψ is obtained by a limiting procedure. Under certain conditions over ρ and L, we can guarantee ψ is the unique solution of the FBP (Theorem 2), but we do not have a proof of this fact if we assume weaker hypotheses.

Sketch of Proof of Theorems 1 and 2. The coupling of Proposition 1 satisfies $\hat{X}_t^{\delta,-} \preccurlyeq \hat{X}_t \preccurlyeq \hat{X}_t^{\delta,+}$. By Theorem 3, the empirical measures associated to the stochastic barriers $\hat{X}_t^{\delta,\pm}$ converge to the macroscopic barriers $S_t^{\delta,\pm}\rho$. By Theorem 4, the macroscopic barriers converge to a function $\psi(\cdot,t)$ as $\delta \to 0$. Hence the empirical measure of $\hat{X}_t$ must converge to $\psi(\cdot,t)$ as $N \to \infty$. This is enough to get Theorem 1.

In Sect. 6, we show that any solution of the FBP is in between the barriers $S_{k\delta}^{\delta,\pm}\rho$; this is enough to get Theorem 2.

3 Domination Proof of Proposition 1

In this section, we construct versions of $X_{(k+1)\delta}$ and $X_{(k+1)\delta}^{\delta,\pm}$ conditioned to knowing the positions of the particles at time $k\delta$. In order to apply an inductive argument, we suppose the particle configurations at time $k\delta$ upon which we are conditioning are ordered in the sense of the partial order $\preccurlyeq$ defined in (11): $X_{k\delta}^{\delta,-} \preccurlyeq X_{k\delta} \preccurlyeq X_{k\delta}^{\delta,+}$. The property about this domination we will use is the following one: if two particle configurations $y = \{y_1 < \ldots < y_N\}$ and $x = \{x_1 < \ldots < x_N\}$ are such that $y \preccurlyeq x$, then $y_i \le x_i$ for every i. The resulting processes at time $(k+1)\delta$ will again respect the order they had in the previous step. Once these constructions are done, one can easily construct a version of $\{(X_{k\delta}^{\delta,-}, X_{k\delta}, X_{k\delta}^{\delta,+}) : k \in \mathbb{N}_0\}$ that makes the job of Proposition 1.

Stochastic Lower Bound. Fix a particle configuration $x_1 < \ldots < x_N$. For $M \in \mathbb{N}_0$, consider M particles $y_{N-M+1} < y_{N-M+2} < \ldots < y_N$, and suppose that $y_i \le x_i$ for every $i \in \{N-M+1, N-M+2, \ldots, N\}$. Complete the y-particles until getting N of them with particles at $-\infty$: $y_1 = \ldots = y_{N-M} = -\infty$. In such a way, we have $y_i \le x_i$ for every i. As functions of N independent BBM processes $\{\{(B_t^{i,j})_{t\ge 0} : j \in \mathbb{N}\},\ i \in \{1,\ldots,N\}\}$ with initial positions $B_0^{1,1} = \ldots = B_0^{N,1} = 0$, we will construct the following four processes:

1. $(X_t^-)_{t\ge 0}$ with initial configuration $X_0^- = \{y_{N-M+1}, \ldots, y_N\}$;
2. $(Y_t)_{t\ge 0}$ with initial configuration $Y_0 = \{y_1, \ldots, y_N\}$;
3. $\big((\hat{Y}_t, \sigma_t)\big)_{t\ge 0}$ with initial particle-configuration $\hat{Y}_0 = \{y_1, \ldots, y_N\}$;
4. $\big((X_t, \sigma_t)\big)_{t\ge 0}$ with initial particle-configuration $X_0 = \{x_1, \ldots, x_N\}$.

In the last two processes, the second coordinates $(\sigma_t)_{t\ge 0}$ are labels. $(X_t^-)_{t\ge 0}$ at time δ will have the same distribution as the lower barrier $X_{k\delta}^{\delta,-}$ conditioned to taking the value $\{y_{N-M+1}, \ldots, y_N\}$ at time $(k-1)\delta$. The particle-positions X_t of the fourth process will have the same distributions than the N-BBM process with initial positions $\{x_1, \ldots, x_N\}$. $(Y_t)_{t\ge 0}$ and $\big((\hat{Y}_t, \sigma_t)\big)_{t\ge 0}$ will play the role of auxiliary intermediate processes. $(X_t^-)_{t\ge 0}$ will be a subset of $(Y_t)_{t\ge 0}$, the particle-positions $(\hat{Y}_t)_{t\ge 0}$ of the third process will be dominated by $(X_t)_{t\ge 0}$, and $(Y_t)_{t\ge 0}$ will coincide with $(\hat{Y}_t)_{t\ge 0}$; these three facts imply that $(X_t^-)_{t\ge 0}$ is stochastically dominated by $(X_t)_{t\ge 0}$. Once we have this, the first inequality in (12) follows easily from an iterative procedure.

1. As before, set N_t^i to be the cardinal of the i-th BBM process at time $t \geq 0$. The process X_t^- will not be Markovian. For every $t \geq 0$, let $L_t^- := \min\{N - M + 1 \leq \ell \leq N : \sum_{i=\ell}^N N_t^\ell \leq N\}$ (with the convention $\min \emptyset = \infty$), and define
$$X_t^- := \{y_i + B_t^{i,j} : i \geq L_t^-, 1 \leq j \leq N_t^i\}.$$

2. Let $\prec$ be the strict lexicographical order on the set of labels $\{1, \ldots, N\} \times \mathbb{N}$: $(i,j) \prec (i',j')$ if and only if $i < i'$, or $i = i'$ and $j < j'$. For $t \geq 0$, Y_t consists of the N particles with $\prec$-highest labels: $Y_t := \left\{y_i + B_t^{i,j} : 1 \leq i \leq N, 1 \leq j \leq N_t^i, |\{(i',j') : (i,j) \prec (i',j')\}| < N\right\}$. Since X_t^- consists of the descendants at time t of the maximal possible number of right-most particles at time 0 whose total descendance at time t does not exceed N, we have $X_t^- \subset Y_t$, which in turn implies
$$X_t^- \preccurlyeq Y_t.$$

3. We define the fourth process before the third one. Let $\tau_0 = 0$ and $(\tau_n)_{n\in\mathbb{N}}$ be the branching times of the family of BBM processes $\left((B_t^{i,j})_{t\geq 0} : 1 \leq i \leq N, 1 \leq j \leq N_t^i\right)$. Set $X_0^i = x_i$ and $\sigma_0^i = (i,1)$ for every $1 \leq i \leq N$. The label σ_t^i indicates what Brownian motion the i-th particle is following to diffuse at time t (they will have the same role in the process $\left((\hat{Y}_t, \sigma_t)\right)_{t\geq 0}$). For $0 < s < \tau_1$, let $\sigma_s^i := (i,1)$ and $X_s^i := x_i + B_s^{i,1}$ for every $1 \leq i \leq N$. Let $k \geq 1$ and suppose we have defined the process in the time interval $[0, \tau_k)$. Suppose the branching at time τ_k occurs at the particle $B_{\tau_k-}^{n,j}$, in which case the new born particle will have label $(n, N_{\tau_k-}^n + 1)$. If $(n,j) \notin \{\sigma_{\tau_k-}^i : 1 \leq i \leq N\}$, i.e. if the branching particle is not in X_{τ_k-}, the particles continue with their labels and increments: $\sigma_s^i = \sigma_{\tau_k-}^i$ and $X_s^i = X_{\tau_k-}^i + (B_s^{\sigma_s^i} - B_{\tau_1}^{\sigma_s^i})$, $1 \leq i \leq N$, $s \in [\tau_k, \tau_{k+1})$. Suppose $(n,j) \in \{\sigma_{\tau_k-}^i : 1 \leq i \leq N\}$, and let m be the index of the left-most X-particle: $m := \operatorname{argmin}_{1\leq \ell \leq N} X_{\tau_k-}^\ell$. For $s \in [\tau_k, \tau_{k+1})$, set
$$\sigma_s^m := (n, N_{\tau_k-}^n + 1) \text{ and } X_s^m := X_{\tau_k-}^n + (B_s^{\sigma_s^m} - B_{\tau_k}^{\sigma_s^m}),$$
$$\text{and } \sigma_s^\ell := \sigma_{\tau_k-}^\ell \text{ and } X_s^\ell := X_{\tau_k-}^\ell + (B_s^{\sigma_s^\ell} - B_{\tau_k}^{\sigma_s^\ell}) \text{ for every } \ell \neq m.$$
In words, the left-most particle X^m jumps over the branching one X^n, and starts following the increments of the new-born particle on the family of BBM processes.

4. The labels σ_t in the process $\left((\hat{Y}_t, \sigma_t)\right)_{t\geq 0}$ coincide with the labels in the previous one. In other words, the increments of the particle $\hat{Y}^i$ are coupled with the ones of the particle X^i at all times. This information determines the particle-positions for $s \in [0, \tau_1)$: $\hat{Y}_s^i := y_i + B_s^{i,1}$. Suppose the process $\hat{Y}_t$ has been defined in the time-interval $[0, \tau_k)$, and let as before (n,j) be the label at which the branching at time τ_k is carried out. If $(n,j) \notin \{\sigma_{\tau_k-}^i : 1 \leq i \leq N\}$, we change nothing: $\hat{Y}_s^i := \hat{Y}_{\tau_k-}^i + (B_s^{\sigma_s^i} - B_{\tau_k}^{\sigma_s^i})$, $1 \leq i \leq N$, $s \in [\tau_k, \tau_{k+1})$. If $(n,j) \in \{\sigma_{\tau_k-}^i : 1 \leq i \leq N\}$, let m as before be the index of the left-most X-particle (recalling we had $\sigma_s^m = (n, N_{\tau_k-}^n + 1)$, $s \in [\tau_k, \tau_{k+1})$), and h be the index of the lowest label (in the strict lexicographical order): $\sigma_{\tau_k-}^h \prec \sigma_{\tau_k-}^\ell$ for every $\ell \in \{1, \ldots, N\} \setminus \{h\}$. There are two cases (see Figs. 1 and 2):

(1) $n \notin \{m,h\}$. Set $\hat{Y}^m_s := \hat{Y}^n_{\tau_k-} + (B^{\sigma^m_s}_s - B^{\sigma^m_s}_{\tau_k})$ for $s \in [\tau_k, \tau_{k+1})$, and divide in the following two sub-cases:
   (1a) If $h = m$, let the remaining particles keep their positions and ranks: $\hat{Y}^\ell_s := \hat{Y}^\ell_{\tau_k-} + (B^{\sigma^\ell_s}_s - B^{\sigma^\ell_s}_{\tau_k})$, $\ell \neq m$.
   (1b) If $h \neq m$, set $\hat{Y}^h_s := \hat{Y}^m_{\tau_k-} + (B^{\sigma^h_s}_s - B^{\sigma^h_s}_{\tau_k})$ and $\hat{Y}^\ell_s := \hat{Y}^\ell_{\tau_k-} + (B^{\sigma^\ell_s}_s - B^{\sigma^\ell_s}_{\tau_k})$ for $\ell \notin \{m,h\}$. In words, the particle $\hat{Y}^h$ jumps over the position of the particle $\hat{Y}^m$ at time τ_k-, and the rest ones keep their positions.

(2) $n \in \{m,h\}$. We have the following two sub-cases:
   (2a) If $n = h \neq m$, the particle $\hat{Y}^h$ jumps over $\hat{Y}^m$, and the rest of the particles keep their positions: for $s \in [\tau_k, \tau_{k+1})$, $\hat{Y}^h_s := \hat{Y}^m_s + (B^{\sigma^h_s}_s - B^{\sigma^h_s}_{\tau_k})$ and $\hat{Y}^\ell_s := \hat{Y}^\ell_s + (B^{\sigma^\ell_s}_s - B^{\sigma^\ell_s}_{\tau_k})$ for $\ell \neq h$.
   (2b) If $n = m \neq h$ or $n = h = m$, all particles keep their positions: $\hat{Y}^\ell_s := \hat{Y}^\ell_{\tau_k-} + (B^{\sigma^\ell_s}_s - B^{\sigma^\ell_s}_{\tau_k})$ for every ℓ.

Lemma 1. *For every $t \geq 0$, (a) X_t has the distribution of the N-*BBM *process with initial positions $\{x_1, \ldots, x_N\}$, (b) $\hat{Y}_t$ coincide with Y_t, and (c) $\hat{Y}_t \preccurlyeq X_t$.*

Proof. (a) Since the left-most particle jumps over the branching one, and since this branches occurs at rate 1, $(X_t)_{t\geq 0}$ has the distribution of the N-BBM process.

(b) The proof is only the observation that the distribution of both processes $(Y_t)_{t\geq 0}$ and $(\hat{Y}_t)_{t\geq 0}$ can be described as follows: at the beginning, we have N particles with positions $\{y_1, \ldots, y_N\}$; we say that the family name of the particle that starts at y_i is i; each particle diffuses independently; at rate N, one of the N particles is uniformly chosen and a new one is created at that position, inheriting its family name, and the oldest particle among all the ones with lowest family name is killed; the new N particles start to diffuse independently, and the procedure starts again.

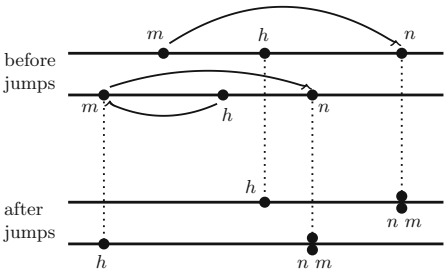


Fig. 1. Relative positions of particles at branching time s for the case (1b). Each X-particle is to the right of the $\hat{Y}$-particle with the same label before and after the branching. This order would be broken if the hth $\hat{Y}$-particle jumped to the nth $\hat{Y}$-particle, in this example.

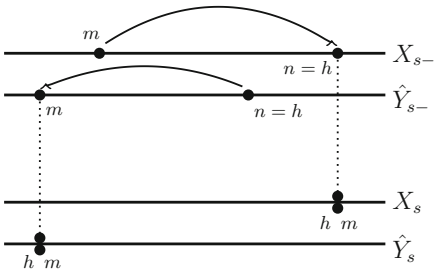


Fig. 2. Relative positions of particles at branching time s for the case (2a).

(c) The domination $\hat{Y}_t^\ell \le X_t^\ell$ holds at time 0 and it is preserved between branching events because the Brownian increments are the same. The domination persists after each branching event obviously in cases (1a) and (2b). In case (1b), we have $\hat{Y}_{\tau_k}^m = \hat{Y}_{\tau_k-}^n \le X_{\tau_k-}^n = X_{\tau_k}^m$ and $\hat{Y}_{\tau_k}^h = \hat{Y}_{\tau_k-}^m \le X_{\tau_k-}^m \le X_{\tau_k-}^h = X_{\tau_k}^h$ because $X_{\tau_k-}^m$ is the minimal X_{τ_k-}-particle and the h-th X particle does not jump at τ_k. In case (2a), $X_{\tau_k}^m = X_{\tau_k}^h = X_{\tau_k-}^h \ge X_{\tau_k-}^m \ge \hat{Y}_{\tau_k-}^m = \hat{Y}_{\tau_k}^h = \hat{Y}_{\tau_k}^m$ (see Fig. 2).

Stochastic Upper Bound. Take again $x_1 < \ldots < x_N$, and let $z_1 < \ldots < z_N$ be another set of N points such that $x_i \le z_i$ for every i. Take the same family of BBM processes $\big\{\{(B_t^{i,j})_{t\ge 0} : j \in \mathbb{N}\},\ i \in \{1,\ldots,N\}\big\}$ with initial positions $B_0^{1,1} = \ldots = B_0^{N,1} = 0$. Let $(X_t)_{t\ge 0}$ be defined as above. We will define the process $(X_t^+)_{t\ge 0}$ that will represent the process $X_{(k+1)\delta}^{\delta,+}$ conditioned to $X_{k\delta}^{\delta,+} = \{z_1,\ldots,z_N\}$. Let $L_t^+ \in \{z_i + B_t^{i,j} : 1 \le i \le N, 1 \le j \le N_t^i\}$ be the unique element satisfying

$$\sum_{i=1}^{N}\sum_{j=1}^{N_t^i} \mathbf{1}\{z_i + B_t^{i,j} \ge L_t^+\} = N,$$

and let $(X_t^+)_{t\ge 0}$ consist of the N right most particles of the BBM processes starting at $\{z_1,\ldots,z_N\}$:

$$X_t^+ := \{z_i + B_t^{i,j} : z_i + B_t^{i,j} \ge L_t^+, 1 \le i \le N, 1 \le j \le N_t^i\}.$$

Since X_t is a subset of $\{x_i + B_t^{i,j} : 1 \le i \le N, 1 \le j \le N_t^i\}$, the former set is dominated by the N right-most particles of the latter one. Also condition $x_i \le z_i$ for every i implies that the set containing the N right-most particles of $\{x_i + B_t^{i,j} : 1 \le i \le N, 1 \le j \le N_t^i\}$ is dominated by X_t^+. The last two facts imply $X_t \preccurlyeq X_t^+$ almost surely.

Proof of Proposition 1. For $k = 0$, the coupling $(\hat{X}_0^{\delta,-}, \hat{X}_0, \hat{X}_0^{\delta,+})$ is simply defined as $\hat{X}_0^{\delta,-} = \hat{X}_0 = \hat{X}_0^{\delta,+} = \{X_0^1,\ldots,X_0^N\}$ (N independent particles with law defined in terms of ρ).

Fix now a realization of the initial configuration $X_0^1,\ldots,X_0^N$, whose points we can suppose to be all different. After reordering their labels, we can identify them with the set $\{x_1,\ldots,x_N\}$ introduced before. In this case, take $M = 0$ and $y_i = x_i = z_i$ for every i. The construction of the pre and post-selection processes gives the coupling $(\hat{X}_\delta^{\delta,-}, \hat{X}_\delta, \hat{X}_\delta^{\delta,+})$ conditioned to having initial configuration $X_0^1,\ldots,X_0^N$ (namely the marginal distributions of $(\hat{X}_\delta^{\delta,-}, \hat{X}_\delta, \hat{X}_\delta^{\delta,+})$ respectively coincide with the conditional distributions of $X_\delta^{\delta,-}$, X_δ and $X_\delta^{\delta,+}$ with initial condition $X_0^1,\ldots,X_0^N$, and $X_\delta^{\delta,-} \preccurlyeq X_\delta \preccurlyeq X_\delta^{\delta,+}$ almost surely). Then make an average of the initial condition weighted with the law of $(\hat{X}_0^{\delta,-}, \hat{X}_0, \hat{X}_0^{\delta,+})$ to get the coupling $(\hat{X}_\delta^{\delta,-}, \hat{X}_\delta, \hat{X}_\delta^{\delta,+})$.

Suppose we have defined the coupling $\{(\hat{X}^{\delta,-}_{\ell\delta}, \hat{X}_{\ell\delta}, \hat{X}^{\delta,+}_{\ell\delta}), 0 \le \ell \le k\}$. Fix a realization of $(\hat{X}^{\delta,-}_{k\delta}, \hat{X}_{k\delta}, \hat{X}^{\delta,+}_{k\delta})$. After reordering, the first, second and third coordinates can be respectively identified with sets $\{y_{N-M+1}, \dots, y_N\}$, $\{x_1, \dots, x_N\}$ and $\{z_1, \dots, z_N\}$, and the coupled conditional distributions of $(\hat{X}^{\delta,-}_{(k+1)\delta}, \hat{X}_{(k+1)\delta}, \hat{X}^{\delta,+}_{(k+1)\delta})$ exists as before. Finally make an average of the conditioning configurations at time $k\delta$ weighted with the law of $(\hat{X}^{\delta,-}_{k\delta}, \hat{X}_{k\delta}, \hat{X}^{\delta,+}_{k\delta})$ to get the coupling $(\hat{X}^{\delta,-}_{(k+1)\delta}, \hat{X}_{(k+1)\delta}, \hat{X}^{\delta,+}_{(k+1)\delta})$.

4 Hydrodynamic Limit for the Barriers

In this section, we prove Theorem 3, namely that the stochastic barriers converge in the macroscopic limit $N \to \infty$ to the deterministic barriers. Recall that ρ is a probability density on $\mathbb{R}$ with a left boundary L^*, and that the N-BBM starts from $X_0 = (X_0^1, \dots, X_0^N)$, iid continuous random variables with density ρ.

It is convenient to have a notation for the cutting points for the macroscopic barriers $S^{\delta,\pm}_{k\delta}$ defined in (14). For $\delta > 0$ and natural number $\ell \le k$, denote

$$L^{\delta,+}_{\ell\delta} := \sup_r \Big\{ \int_r^\infty S^{\delta,+}_{\ell\delta} \rho(r') \mathrm{d}r' = 1 \Big\} \qquad L^{\delta,-}_{\ell\delta} := \sup_r \Big\{ \int_r^\infty S^{\delta,-}_{\ell\delta} \rho(r') \mathrm{d}r' = e^{-\delta} \Big\}. \tag{15}$$

Let B_0 be a continuous random variable with density ρ. Let $B_{[0,t]} = (B_s : s \in [0,t])$ be a Brownian motion starting from B_0, with increments independent of B_0. Let N_t be the random size at time t of a BBM family starting with one member. We have $EN_t = e^t$. Recall $e^t G_t \rho$ is the solution of $u_t = \frac{1}{2} u_{rr} + u$ with $u(\cdot, 0) = \rho$. With this notation, we have the following representation of $e^t G_t \rho$ and the macroscopic barriers as expectation of functions of the Brownian trajectories.

Lemma 2. *For every bounded measurable test function $\varphi : \mathbb{R} \to \mathbb{R}$ and every $t > 0$, we have*

$$\int \varphi(r) e^t G_t \rho(r) \, \mathrm{d}r = e^t E[\varphi(B_t)]. \tag{16}$$

Furthermore

$$\begin{aligned} \int \varphi(r) S^{\delta,+}_{k\delta} \rho(r) \, \mathrm{d}r &= e^{k\delta} E\big[\varphi(B_{k\delta}) \mathbf{1}\big\{B_{\ell\delta} > L^{\delta,+}_{\ell\delta} : 1 \le \ell \le k\big\}\big] \\ \int \varphi(r) S^{\delta,-}_{k\delta} \rho(r) \, \mathrm{d}r &= e^{k\delta} E\big[\varphi(B_{k\delta}) \mathbf{1}\big\{B_{\ell\delta} > L^{\delta,-}_{\ell\delta} : 0 \le \ell \le k-1\big\}\big]. \end{aligned} \tag{17}$$

Proof. Immediate.

Recall the definition of $\mathcal{B}$ in (7). In particular, the trajectory $B^{i,j}_{[0,t]}$ is distributed as $B_{[0,t]}$ for all i, j, and the families $(B^{i,j}_{[0,t]} : j \in \{1, \dots, N^i_t\})$, for $i \in \{1, \dots, N\}$, are iid.

Proposition 2. *Let $t > 0$ and $g : C([0,t],\mathbb{R}) \to \mathbb{R}$ be a bounded measurable function, and define*

$$\mu_t^N g := \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_t^i} g(B^{i,j}_{[0,t]}).$$

Then

$$\lim_{N\to\infty} \mu_t^N g = e^t E g(B_{[0,t]}) \qquad \text{a.s. and in } L^1. \tag{18}$$

Proof. By the many-to-one Lemma (see [18], for instance), we have

$$E\mu_t^N g = E N_t \, E g(B_{[0,t]}) = e^t E g(B_{[0,t]}).$$

This is enough to get the strong law of large numbers (18).

Recall we have defined $M_{k\delta}$ as $N - |X_{k\delta}^{\delta,-}|$.

Lemma 3. *For every $k \in \mathbb{N}_0$, $M_{k\delta}^{\delta}/N \xrightarrow[N\to\infty]{} 0$ almost surely and in L^1.*

Proof. $M_0^\delta = 0$ almost surely. For $k \geq 1$, $M_{k\delta}^\delta$ is non-negative and its law converges as $N \to \infty$ to the law of the Age of a renewal process with inter-renewal intervals distributed as N_δ, the one-particle family size at time δ. The Age law is the size-biased law of N_δ. Since N_δ has all moments finite, we can conclude.

Corollary 1 (Hydrodynamics of the BBM). *For measurable bounded $\varphi : \mathbb{R} \to \mathbb{R}$ and $t > 0$, we have*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_t^i} \varphi(B_t^{i,j}) = e^t E\varphi(B_t) = e^t \int \varphi(r) G_t \rho(r) \mathrm{d}r \qquad \text{a.s. and in } L^1.$$

Proof of Theorem 3. Recalling (10), we have

$$\pi_{k\delta}^{N,\delta,+}\varphi = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_{k\delta}^i} \varphi(B_{k\delta}^{i,j}) \mathbf{1}\{B_{\ell\delta}^{i,j} \geq L_{\ell\delta}^{N,\delta,+} : 1 \leq \ell \leq k\}$$

$$\pi_{k\delta}^{N,\delta,-}\varphi = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_{k\delta}^i} \varphi(B_{k\delta}^{i,j}) \mathbf{1}\{B_{\ell\delta}^{i,j} \geq L_{\ell\delta}^{N,\delta,-} : 0 \leq \ell \leq k-1\}.$$

We want to apply Proposition 2, but do not have an explicit expression for $E\pi_{k\delta}^{N,\delta,\pm}\varphi$ because of the random boundaries in the right hand side. We can use instead the deterministic boundaries $L_{\ell\delta}^{\delta,\pm}$ by defining

$$g_\varphi^+(B_{[0,k\delta]}) := \varphi(B_{k\delta}) \mathbf{1}\{B_{\ell\delta} \geq L_{\ell\delta}^{\delta,+} : 1 \leq \ell \leq k\}$$
$$g_\varphi^-(B_{[0,k\delta]}) := \varphi(B_{k\delta}) \mathbf{1}\{B_{\ell\delta} \geq L_{\ell\delta}^{\delta,-} : 0 \leq \ell \leq k-1\}.$$

By (18) and (17), we have

$$\lim_{N\to\infty} \mu^N_{k\delta} g^{\pm}_{\varphi} = e^{k\delta} E g^{\pm}_{\varphi}(B_{[0,k\delta]}) = \int \varphi(r) S^{\delta,\pm}_{k\delta} \rho(r)\, \mathrm{d}r.$$

To conclude, it suffices to show that $\pi^{N,\delta,\pm}_{k\delta}\varphi - \mu^N_{k\delta} g^{\pm}_{\varphi}$ converges to 0. To get this, observe that, by Proposition 3 below,

$$\begin{aligned}
&\big|\pi^{N,\delta,-}_{k\delta}\varphi - \mu^N_{k\delta} g^{-}_{\varphi}\big| \\
&\quad = \Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} \varphi(B^{i,j}_{k\delta})\Big(\prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{N,\delta,-}_{\ell\delta}\big\} - \prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{\delta,-}_{\ell\delta}\big\}\Big)\Big| \\
&\quad \le \|\varphi\|_\infty \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} \Big|\prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{N,\delta,-}_{\ell\delta}\big\} - \prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{\delta,-}_{\ell\delta}\big\}\Big| \underset{N\to\infty}{\longrightarrow} 0 \\
&\qquad\qquad \text{a.s. and } L^1.
\end{aligned}$$

The same argument works for the upper barrier. □

By Definitions (8) and (9) of the microscopic cutting points $L^{N,\delta,\pm}_{\ell\delta}$ and Lemma 3, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}\prod_{\ell=1}^{k} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{N,\delta,+}_{\ell\delta}\big\} = 1 \\
&\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}\prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{N,\delta,-}_{\ell\delta}\big\} = 1 - O(1/N). \qquad (19)
\end{aligned}$$

Proposition 3. *For $k \in \mathbb{N}$, we have*

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}\Big|\prod_{\ell=1}^{k} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{N,\delta,+}_{\ell\delta}\big\} - \prod_{\ell=1}^{k} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{\delta,+}_{\ell\delta}\big\}\Big| \underset{N\to\infty}{\longrightarrow} 0 \qquad (20)$$

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}\Big|\prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{N,\delta,-}_{\ell\delta}\big\} - \prod_{\ell=0}^{k-1} \mathbf{1}\big\{B^{i,j}_{\ell\delta} \ge L^{\delta,-}_{\ell\delta}\big\}\Big| \underset{N\to\infty}{\longrightarrow} 0 \qquad (21)$$

a.s. and in L^1.

Proof. We first treat the lower barrier, i.e. the "−" quantities. Since at time zero the families have only one element, for the lower barrier at $k = 0$, the left hand side of (21) reads

$$\frac{1}{N}\sum_{i=1}^{N}\Big|\mathbf{1}\big\{B^{i,1}_0 \ge L^{N,\delta,-}_0\big\} - \mathbf{1}\big\{B^{i,1}_0 \ge L^{\delta,-}_0\big\}\Big|.$$

Recalling that all the trajectories $B^{i,j}_{[0,\delta]}$ start at the same point $B^{i,1}_0$, we can bound the above expression by

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_\delta}\Big|\mathbf{1}\{B^{i,j}_0 \geq L^{N,\delta,-}_0\} - \mathbf{1}\{B^{i,j}_0 \geq L^{\delta,-}_0\}\Big| \tag{22}$$

$$= \Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_\delta}\mathbf{1}\{B^{i,j}_0 \geq L^{N,\delta,-}_0\} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_\delta}\mathbf{1}\{B^{i,j}_0 \geq L^{\delta,-}_0\}\Big| \underset{N\to\infty}{\longrightarrow} 0,$$

a.s. and L^1. (23)

The identity holds because the differences of the indicator functions in (22) have the sign of $L^{N,\delta,-}_0 - L^{\delta,-}_0$ for all i, j. The limit (23) holds because (a) by (19) the first term in (23) converges to 1 and (b) the limit of the second term is also 1 by definition (15) of $L^{\delta,-}_0$ and Proposition 2.

For the upper barrier at $k = 1$, put again the sums between the absolute values to get that in this case the left hand side of (20) reads

$$\Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_\delta}\mathbf{1}\{B^{i,j}_\delta \geq L^{N,\delta,+}_\delta\} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_\delta}\mathbf{1}\{B^{i,j}_\delta \geq L^{\delta,+}_\delta\}\Big| \underset{N\to\infty}{\longrightarrow} 0,$$

a.s. and in L^1, (24)

because the first term is 1 by definition of $L^{N,\delta,+}_\delta$ and the second one converges to 1 by the definition of $L^{\delta,+}_\delta$ and Proposition 2.

For the induction step denote

$$A^{N,\delta,+}_{k\delta}(i,j) := \prod_{\ell=1}^{k}\mathbf{1}\{B^{i,j}_{\ell\delta} \geq L^{N,\delta,+}_{\ell\delta}\} \qquad A^{\delta,+}_{k\delta}(i,j) := \prod_{\ell=1}^{k}\mathbf{1}\{B^{i,j}_{\ell\delta} \geq L^{\delta,+}_{\ell\delta}\}$$

$$A^{N,\delta,-}_{k\delta}(i,j) := \prod_{\ell=0}^{k-1}\mathbf{1}\{B^{i,j}_{\ell\delta} \geq L^{N,\delta,-}_{\ell\delta}\} \qquad A^{\delta,-}_{k\delta}(i,j) := \prod_{\ell=0}^{k-1}\mathbf{1}\{B^{i,j}_{\ell\delta} \geq L^{\delta,-}_{\ell\delta}\}.$$

Assume (20) holds for $\ell = k-1$ and write the left hand side of (20) for the upper barrier at $\ell = k$ as

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}\Big|A^{N,\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{N,\delta,+}_{k\delta}\} - A^{\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{\delta,+}_{k\delta}\}\Big|$$

$$\leq \Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}A^{N,\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{N,\delta,+}_{k\delta}\} \tag{25}$$

$$- \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} A^{N,\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{\delta,+}_{k\delta}\}\Big|$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}}\Big|A^{N,\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{\delta,+}_{k\delta}\} - A^{\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{\delta,+}_{k\delta}\}\Big|, \quad (26)$$

where the inequality is obtained by summing and subtracting the same expression and then taking the modulus out of the sums in (25) as all the indicator function differences have the same sign, as before.

By dominated convergence and the induction hypothesis, the expression in (26) converges to zero as $N \to \infty$. In turn, this implies that the second term in (25) has the same limit as the second term in (27) below. Hence we only need to show that expression

$$\Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} A^{N,\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{N,\delta,+}_{k\delta}\}$$

$$- \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} A^{\delta,+}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{k\delta} \geq L^{\delta,+}_{k\delta}\}\Big| \quad (27)$$

vanishes. To see this, since the first term is 1 and the second one converges to 1, we can use the same arguments we used in (24). The same argument shows that the limit (21) for the lower barrier at $\ell = k$ is zero if the expression

$$\Big|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} A^{N,\delta,-}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{\delta} \geq L^{N,\delta,-}_{(k-1)\delta}\}$$

$$- \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N^i_{k\delta}} A^{\delta,-}_{(k-1)\delta}(i,j)\mathbf{1}\{B^{i,j}_{(k-1)\delta} \geq L^{\delta,-}_{k\delta}\}\Big|$$

goes to 0; this follows with the same argument as (23).

5 Existence of the Limit Function ψ

In this Section, we prove Theorem 4, namely the existence of the function ψ at which the N-BBM process converges. We start with the following Proposition whose proof is similar to the one given in Chapters 4, 5 and 6 of [5].

Proposition 4. *The following properties hold for every integrable $u, v : \mathbb{R} \to \mathbb{R}^+$ and $t, m > 0$.*

(a) If $u \preccurlyeq v$, then $C_m u \preccurlyeq v$ for every $m > 0$.
(b) If $u \leq v$ point wise, then $G_t u \leq G_t v$, $G_t u \preccurlyeq G_t v$.
(c) C_m and G_t preserve the order: if $u \preccurlyeq v$, then $C_m u \preccurlyeq C_m v$ and $G_t u \preccurlyeq G_t v$.

(d) $\|C_m u - C_m v\|_1 \le \|u - v\|_1$.
(e) $\|G_t u - G_t v\|_1 \le \|u - v\|_1$.
(f) If $\|u\|_1 = 1$, $\left|\frac{d}{dr} G_t u(r)\right| \le \frac{c}{t}$ *for every* $r \in \mathbb{R}$.

Proof. (Proof of Proposition 4). Items *(a)*, *(b)*, *(e)* and *(f)* are simple and we omit their proofs.
Proof of (c). We start with C_m and assume $m < \|u\|_1 \wedge \|v\|_1$ since the other case is trivial. For $a \in \mathbb{R}$, we have to prove that $\int_a^\infty C_m u \le \int_a^\infty C_m v$. Denote the cutting points by

$$q_m(w) := \sup\Big\{r \in \mathbb{R} : \int_r^\infty w = m\Big\}.\mathbb{R} : \int_{-\infty}^L v > m\Big\}. \tag{28}$$

We suppose $q_m(u) \wedge q_m(v) < a < q_m(u) \vee q_m(v)$ since the other case is trivial. If $q_m(v) < a < q_m(v)$, we have

$$\int_a^\infty C_m u = \int_a^\infty u \le \int_{q_m(v)}^\infty u = m = \int_{q_m(v)}^\infty v = \int_a^\infty C_m v;$$

if $q_m(v) < a < q_m(u)$,

$$\int_a^\infty C_m u = \int_{q_m(u)}^\infty u \le \int_{q_m(u)}^\infty v \le \int_a^\infty v = \int_a^\infty C_m v.$$

Suppose now $\|u\|_1 < \|v\|_1$ and let $m := \|v\|_1 - \|u\|_1$. It is easy to see that $u \preccurlyeq C_m v$. Since $\|u\|_1 = \|C_m v\|_1$, we can apply the previous case to get $G_t u \preccurlyeq G_t C_m v$. We conclude by observing that, because of item (b) and the point-wise dominance $C_m v \le v$, we have $G_t C_m v \preccurlyeq G_t v$.
Proof of (d). We assume $m < \|u\|_1 \wedge \|v\|_1$ since the other case is trivial. Define $q_m(u)$ and $q_m(v)$ as in (28), and suppose $q_m(u) \ge q_m(v)$ without loss of generality. Then

$$\begin{aligned}\|C_m u - C_m v\|_1 &= \left\|u\mathbf{1}_{x \ge q_m(u)} - v\mathbf{1}_{x \ge q_m(v)}\right\|_1 \\ &= \int_{q_m(v)}^{q_m(u)} v - \int_{-\infty}^{q_m(u)} |u - v| + \|u - v\|_1. \end{aligned} \tag{29}$$

Also

$$\int_{q_m(v)}^{q_m(u)} v = \int_{-\infty}^{q_m(v)} v - \int_{-\infty}^{q_m(u)} u + \int_{q_m(v)}^{q_m(u)} v = \int_{-\infty}^{q_m(u)} (v - u) \le \int_{-\infty}^{q_m(u)} |u - v|. \tag{30}$$

Item *(d)* follows from (29) and (30).

Proposition 5. *For every integrable* $u : \mathbb{R} \to \mathbb{R}^+$, $\delta > 0$ *and natural number* $k \ge 0$, *we have*

$$S^{\delta,-}_{k\delta} u \preccurlyeq S^{\delta,+}_{k\delta} u \tag{31}$$

$$S^{\delta,-}_{k\delta} u \preccurlyeq S^{\delta/2,-}_{k\delta} u \tag{32}$$

$$S^{\delta,+}_{k\delta} u \succcurlyeq S^{\delta/2,+}_{k\delta} u. \tag{33}$$

Furthermore, there exists a constant $c > 0$ such that

$$\left\| S^{\delta,+}_{k\delta} u - S^{\delta,-}_{k\delta} u \right\|_1 \leq c\delta. \tag{34}$$

Proof. Inequality (31) is a consequence of Theorem 3 and Proposition 1.

To prove inequalities (32) and (33), we call $H^-_\delta := e^\delta G_\delta C_{e^{-\delta}}$ and $H^+_\delta := C_1 e^\delta G_\delta$, and prove below that

$$H^-_\delta v \preccurlyeq \left(H^-_{\delta/2}\right)^2 v \tag{35}$$

$$H^+_\delta v \succcurlyeq \left(H^+_{\delta/2}\right)^2 v. \tag{36}$$

Now we deduce (32) from (35) by induction. The case $k = 1$ is (35). Assume the assertion holds for $k-1$, and write $\left(H^-_\delta\right)^k u = H^-_\delta \left(H^-_\delta\right)^{k-1} u$. Call $v := \left(H^-_\delta\right)^{k-1} u$ and use the case $k = 1$ to infer $\left(H^-_\delta\right)^k u \preccurlyeq \left(H^-_{\delta/2}\right)^2 \left(H^-_\delta\right)^{k-1} u$. By the inductive hypothesis and the fact that $\left(H^-_{\delta/2}\right)^2$ preserves the order, conclude that $\left(H^-_\delta\right)^k u \preccurlyeq \left(H^-_{\delta/2}\right)^{2k} u$, which is (32). The way to deduce (33) from (36) is similar.

Proof of (35). We first prove that, for $a \in \mathbb{R}$,

$$\int_a^\infty e^{\delta/2} G_{\delta/2} C_{e^{-\delta}} v \leq \int_a^\infty C_{e^{-\delta/2}} H^-_{\delta/2} v. \tag{37}$$

Let $q := q_{e^{-\delta/2}}(H^-_{\delta/2} v)$. If $a \leq q$, identity (37) is satisfied because

$$\int_a^\infty e^{\delta/2} G_{\delta/2} C_{e^{-\delta}} v \leq \int_{-\infty}^\infty e^{\delta/2} G_{\delta/2} C_{e^{-\delta}} v = e^{-\delta/2} = \int_a^\infty C_{e^{-\delta/2}} H^-_{\delta/2} v.$$

If $a > q$, inequality (37) becomes

$$\int_a^\infty e^{\delta/2} G_{\delta/2} C_{e^{-\delta}} v \leq \int_a^\infty H^-_{\delta/2} v,$$

which follows from $C_{e^{-\delta}} v \preccurlyeq C_{e^{-\delta/2}} v$ and from *(c)* of Proposition 4 applied to $e^{\delta/2} G_{\delta/2}$.

From (37), we then have that $e^{\delta/2} G_{\delta/2} C_{e^{-\delta}} v \preccurlyeq C_{e^{-\delta/2}} H^-_{\delta/2} v$. Since $e^\delta G_\delta = e^{\delta/2} G_{\delta/2} e^{\delta/2} G_{\delta/2}$ and $e^{\delta/2} G_{\delta/2}$ preserves the order, we get (35).

Proof of (36). We have to prove that $C_1 C_{e^{\delta/2}} e^{\delta/2} G_{\delta/2} w \succcurlyeq C_1 e^{\delta/2} G_{\delta/2} C_1 w$ for $w := e^{\delta/2} G_{\delta/2} v$. From *(c)* of Proposition 4, this follows if we prove

$$C_{e^{\delta/2}} e^{\delta/2} G_{\delta/2} w \succcurlyeq e^{\delta/2} G_{\delta/2} C_1 w.$$

Denote $q := q_{e^{\delta/2}}\left(e^{\delta/2} G_{\delta/2} w\right)$. If $a \leq q$,

$$\int_a^\infty C_{e^{\delta/2}} e^{\delta/2} G_{\delta/2} w = e^{\delta/2} = \int_{-\infty}^\infty e^{\delta/2} G_{\delta/2} C_1 w \geq \int_a^\infty e^{\delta/2} G_{\delta/2} C_{-1} w.$$

For $a > q$, since $w \le C_1 w$ point-wise, we get

$$\int_a^\infty e^{\delta/2} G_{\delta/2} w \ge \int_a^\infty e^{\delta/2} G_{\delta/2} C_1 w.$$

We now prove (34). Let $k := t/\delta$ and define $u_k := e^{\delta} G_\delta \big(H_\delta^+\big)^{k-1} u$ and $v_k := S_t^{\delta,-} u$. Using that $\|u_k\|_1 = e^\delta$ and assuming δ small enough, we get

$$\begin{aligned} \left\| S_t^{\delta,+} u - S_t^{\delta,-} u \right\|_1 &= \|C_1 u_k - v_k\|_1 \le \|C_1 u_k - u_k\|_1 + \|u_k - v_k\|_1 \\ &= \big(e^\delta - 1\big) + \|u_k - v_k\|_1 \le 2\delta + \|u_k - v_k\|_1. \end{aligned} \tag{38}$$

By items *(d)* and *(e)* of Proposition 4,

$$\begin{aligned} \|u_k - v_k\|_1 &\le e^\delta \|C_1 u_{k-1} - C_{e^{-\delta}} v_{k-1}\|_1 \\ &\le e^\delta \|C_1 u_{k-1} - C_{e^{-\delta}} u_{k-1}\|_1 + e^\delta \|C_{e^{-\delta}} u_{k-1} - C_{e^{-\delta}} v_{k-1}\|_1 \\ &\le e^\delta \big[\big(e^\delta - 1\big) - \big(1 - e^{-\delta}\big)\big] + e^\delta \|u_{k-1} - v_{k-1}\|_1 \\ &\le 3\delta^2 + e^\delta \|u_{k-1} - v_{k-1}\|_1. \end{aligned}$$

Iterating and using that $\|u_1 - v_1\|_1 \le e^\delta \|u - C_{e^{-\delta}} u\|_1 = e^\delta \big(1 - e^{-\delta}\big) \le 2\delta$, we get

$$\begin{aligned} \|v_k - u_k\|_1 &\le 3\delta^2 \sum_{j=0}^{k-2} e^{\delta j} + \|u_1 - v_1\|_1 e^{\delta(k-1)} \\ &\le 3\delta^2 \frac{e^{\delta(k-1)} - 1}{e^\delta - 1} + 2\delta e^t \le 3\delta \big(e^t - 1\big) + 2\delta e^t = c_1 \delta. \end{aligned}$$

We conclude by replacing in (38).

In order to prove Theorem 4, we fix $u : \mathbb{R} \to \mathbb{R}_+$ integrable and $0 < t_0 < T$. Call

$$\mathcal{T}_n := \{k 2^{-n}, k \in \mathbb{N}\},$$

and define the function $\rho_n : \mathbb{R} \times [t_0, T] \to \mathbb{R}_+$ as

$$\rho_n(r,t) := S_t^{2^{-n},-} u \qquad \text{if } r \in \mathbb{R} \text{ and } t \in [t_0, T] \cap \mathcal{T}_n,$$

and by linear interpolation in the rest of the cases.

We will apply Ascoli-Arzelá Theorem, which requires point-wise boundedness and equi-continuity. The first requirement is immediate because, since we are away from $t = 0$, we have a uniform bound:

$$\begin{aligned} &\text{there exists } c = c(u, t_0, T) \text{ such that } |\rho_n(r,t)| \le c \\ &\text{for every } (r,t) \in \mathbb{R} \times [t_0, T] \text{ and every } n. \end{aligned} \tag{39}$$

For the second requirement, we will need to prove space and time equi-continuity separately in Propositions 6 and 7 below.

We will use the following Lemma.

Lemma 4. *Given $\delta > 0$, let $t, s \in \delta\mathbb{N}$ with $s < t$, and let*

$$v^{\delta}_{s,t} := S^{\delta,-}_t u - e^{t-s} G_{t-s} S^{\delta,-}_s u.$$

Then

$$\left\| v^{\delta}_{s,t} \right\|_\infty \le \frac{2e^T \sqrt{t-s}}{\sqrt{2\pi}}. \tag{40}$$

Proof. Under definition $w^{\delta}_{t'} := S^{\delta,-}_{t'} u - C_{e^{-\delta}} S^{\delta,-}_{t'} u$, we have

$$S^{\delta,-}_t u = e^{\delta} G_\delta S^{\delta,-}_{t-\delta} u - e^{\delta} G_\delta w^{\delta}_{t-\delta}. \tag{41}$$

Call m and h the integers such that $s = m\delta$ and $t = h\delta$, and iterate (41) to get

$$S^{\delta,-}_t u = e^{t-s} G_{t-s} S^{\delta,-}_s u - \sum_{k=m}^{h-1} e^{h-k} G_{(h-k)\delta} w^{\delta}_{k\delta}.$$

Since $\|G_{t'} w\|_\infty \le \|w\|_1 / \sqrt{2\pi t'}$ for any $t' > 0$ and any integrable $w : \mathbb{R} \to \mathbb{R}$, using that $\left\| w^{\delta}_{k\delta} \right\|_1 = 1 - e^{-\delta} \le \delta$, we get

$$\begin{aligned} \left\| v^{\delta}_{s,t} \right\|_\infty &\le \sum_{k=m}^{h-1} \left\| e^{h-k} G_{(h-k)\delta} w^{\delta}_{k\delta} \right\|_\infty \le \sum_{k=m}^{h-1} \frac{e^{(h-k)\delta} \left\| w^{\delta}_{k\delta} \right\|_1}{\sqrt{2\pi(h-k)\delta}} \le \frac{e^T \sqrt{\delta}}{\sqrt{2\pi}} \sum_{k=m}^{h-1} \frac{1}{\sqrt{h-k}} \\ &\le \frac{e^T \sqrt{\delta}}{\sqrt{2\pi}} 2\sqrt{h-m} = \frac{2e^T \sqrt{t-s}}{\sqrt{2\pi}}, \end{aligned}$$

that let us conclude.

Proposition 6 (Space equi-continuity). *For any $\varepsilon > 0$, there exist n_0 and $\zeta > 0$ such that*

$$|\rho_n(r,t) - \rho_n(r',t)| < \varepsilon \tag{42}$$

for any $n > n_0$, any $t \in [t_0, T]$, and any $r, r' \in \mathbb{R}$ such that $|r - r'| < \zeta$.

Proof. Fix $\varepsilon > 0$. Choose n_0 such that $\delta_0 := 2^{-n_0} < t_0$ and $\frac{4e^T \sqrt{\delta_0}}{\sqrt{2\pi}} < \varepsilon/2$. Choose $\zeta > 0$ such that $\frac{2c_1 c_2 e^{\delta_0}}{\delta_0} \zeta < \varepsilon/2$, where c_1 is the constant of item *(f)* of Proposition 4, and c_2 is the uniform bound given in (39). We take $\delta = 2^{-n}$ with $n > n_0$, and suppose that $t = h\delta \in [t_0, T]$ observing that it is enough to prove (42) for t of this form since ρ_n is defined by linear interpolation. For $s := t - \delta_0$ and $v^{\delta}_{s,t} := S^{\delta,-}_t u - e^{t-s} G_{t-s} S^{\delta,-}_s u$, we have

$$\begin{aligned} &\left| S^{\delta,-}_t u(r) - S^{\delta,-}_t u(r') \right| \\ &\quad = \left| e^{t-s} G_{t-s} S^{\delta,-}_s u(r) - e^{t-s} G_{t-s} S^{\delta,-}_s u(r') \right| + \left| v^{\delta}_{s,t}(r) - v^{\delta}_{s,t}(r') \right|. \end{aligned}$$

For $r, r' \in \mathbb{R}$ as in (42), and by the choice of ζ, we have

$$\left|e^{t-s}G_{t-s}S_s^{\delta,-}u(r) - e^{t-s}G_{t-s}S_s^{\delta,-}u(r')\right| \leq \frac{2c_1c_2e^{\delta_0}\zeta}{\delta_0} < \varepsilon/2.$$

By the choice of δ_0 and from (40), we get $\left|v_{s,t}^{\delta}(r) - v_{s,t}^{\delta}(r')\right| \leq 2\left\|v_{s,t}^{\delta}\right\|_{\infty} < \varepsilon/2$, which concludes the proof.

Proposition 7 (Time equi-continuity). *For any $\varepsilon > 0$, there exist n_0 and $\zeta > 0$ such that, for any $n > n_0$ and any $r \in \mathbb{R}$,*

$$|\rho_n(r,t) - \rho_n(r,t')| < \varepsilon \qquad \forall t, t' \in [t_0, T] \text{ such that } |t - t'| < \zeta. \tag{43}$$

Proof. Fix $\varepsilon > 0$ and let ζ' and n_0 be the parameters given by Proposition 6 associated to $\varepsilon' := \varepsilon/4$. Take ζ such that $\frac{2e^T}{\sqrt{2\pi}}\sqrt{\zeta} < \varepsilon/4$, $\frac{2\sqrt{\zeta}e^T}{\sqrt{2\pi}\zeta'}e^{-\frac{(\zeta')^2}{2\zeta}} < \varepsilon/4$ and $\left(e^{\zeta} - 1\right)e^T\|u\|_{\infty} < \varepsilon/4$. Let $n > n_0$ and $\delta := 2^{-n}$. We first consider $t, t' \in [t_0, T] \cap \delta\mathbb{N}$ such that $t < t' < t + \zeta$. We have to prove that

$$\left|S_{t'}^{\delta,-}u(r) - S_t^{\delta,-}u(r)\right| = \left|v_{t,t'} + e^{t'-t}G_{t'-t}S_t^{\delta,-}u(r) - S_t^{\delta,-}u(r)\right| < \varepsilon. \tag{44}$$

Using (40), we get

$$\begin{aligned}
&\left|S_{t'}^{\delta,-}u(r) - S_t^{\delta,-}u(r)\right| \\
&\quad\leq \frac{\varepsilon}{4} + \left|t' - tG_{t'-t}S_t^{\delta,-}u(r) - S_t^{\delta,-}u(r)\right| \\
&\quad\leq \frac{\varepsilon}{4} + \left|t' - tG_{t'-t}S_t^{\delta,-}u(r) - G_{t'-t}S_t^{\delta,-}u(r)\right| + \left|G_{t'-t}S_t^{\delta,-}u(r) - S_t^{\delta,-}u(r)\right|.
\end{aligned} \tag{45}$$

We have

$$\begin{aligned}
\left|t' - tG_{t'-t}S_t^{\delta,-}u(r) - G_{t'-t}S_t^{\delta,-}u(r)\right| &\leq (e^{t'-t} - 1)\left\|G_{t'-t}S_t^{\delta,-}u\right\|_{\infty} \\
&\leq (e^{t'-t} - 1)e^T\|u\|_{\infty} < \varepsilon/4.
\end{aligned} \tag{46}$$

We next estimate

$$\begin{aligned}
\int_{-\infty}^{\infty} G_{t'-t}(r,r')\left|S_t^{\delta,-}u(r') - S_t^{\delta,-}u(r)\right|\mathrm{d}r' &\leq \frac{\varepsilon}{4} + e^T\int_{|r'-r|\geq\zeta'} G_{t'-t}(r,r')\mathrm{d}r' \\
&\leq \frac{\varepsilon}{4} + \leq \frac{2\sqrt{\zeta}}{\sqrt{2\pi}\zeta'}e^{-\frac{(\zeta')^2}{2\zeta}}.
\end{aligned} \tag{47}$$

Inserting estimates (46) and (47) in (45), we get (44).

For generic $t, t' \in [t_0, T]$ such that $t < t' < t + \zeta$, we consider $\delta = 2^{-n}$ as before, and $t^-, t^+ \in \delta\mathbb{N}$ such that $t^- \leq t < t^- + \delta$ and $t^+ - \delta < t \leq t^+$. Then

$$|\rho_n(r,t') - \rho_n(r,t)| \leq \max_{t_1,t_2\in[t^-,t^+]\cap\delta\mathbb{N}}\left|S_{t_2}^{\delta,-}u(r) - S_{t_1}^{\delta,-}u(r)\right|.$$

From (44), we get (43).

Proof of Theorem 4. For any integrable function $w : \mathbb{R} \to \mathbb{R}$, we define

$$F(r; w) := \int_r^\infty w(r')\mathrm{d}r'.$$

From Ascoli-Arzelá Theorem, we have convergence by subsequences of $(\rho_n)_{n\geq 1}$. Let ψ be any such a limit point. Observe that, for each $t \in [t_0, T] \cap \mathcal{T}_n$, we have that $\rho_n = S_t^{2^{-n},-}u \in L^1$. Since $F(r; S_t^{2^{-n},-}u)$ is a non increasing function of n —see Proposition 5—, it converges as $n \to \infty$. Then, by dominate convergence, we have that

$$\lim_{n\to\infty} F(r; S_t^{2^{-n},-}u) = F(r; \psi(\cdot, t)) \tag{48}$$

for any $r \in \mathbb{R}$ and $t \in [t_0, T] \cap \mathcal{T}_n$. Thus all limit functions $\psi(r, t)$ agree on $t \in [t_0, T] \cap \mathcal{T}_n$, and hence on the whole $[t_0, T]$ since they are continuous. Then the sequence $\rho_n(r, t)$ converges in sup-norm in compact subsets as $n \to \infty$ to a continuous function $\psi(r, t)$ (and not only by subsequences). Observe that, from (48), we also have $F(r; S_t^{2^{-n},-}u) \leq F(r; \psi(\cdot, t))$ for any n and $t \in \mathcal{T}_n$. □

6 Proof of Theorem 1

Fix $t > 0$, choose $\delta \in \{2^{-n}t, n \in \mathbb{N}\}$ and k such that $k\delta = t$. Take X_0 as in Theorem 1, that is, iid continuous random variables with density ρ. By Proposition 1, there is a coupling between the barriers and N-BBM such that, for increasing and bounded φ,

$$\hat{\pi}_t^{N,\delta,-}\varphi \leq \hat{\pi}_t^N\varphi \leq \hat{\pi}_t^{N,\delta,+}\varphi, \tag{49}$$

where $\hat{\pi}$ are the empiric measures associated to the coupled processes $\hat{X}$ of Proposition 1 with initial condition X_0 in the three coordinates. In Theorem 3, we have proven that, under this initial conditions, $\pi_t^{N,\delta,\pm}\varphi$ converge to $\int \varphi\, S_t^{\delta,\pm}\rho$ almost surely and in L^1. We can conclude that the same convergence holds for the hat-variables.

On the other hand, by (49),

$$\left|\pi_t^N\varphi - \hat{\pi}_t^{N,\delta,\pm}\varphi\right| \leq \left|\pi_t^{N,\delta,+}\varphi - \hat{\pi}_t^{N,\delta,-}\varphi\right| \leq \|\varphi\|\, c\delta,$$

by (34). We can conclude using Theorem 3 that

$$\lim_{N\to\infty} \left|\pi_t^N\varphi - \int \varphi\, S_t^{\delta,\pm}\rho\right| \leq \|\varphi\|\, c\delta, \quad \text{a.s. and in } L^1.$$

Taking $\delta \to 0$ along dyadics, we get a function $\int \varphi\,\psi := \lim_{\delta\to 0} \int \varphi\, S_t^{\delta,\pm}\rho$ in L^1 and

$$\lim_{N\to\infty} \left|\pi_t^N\varphi - \int \varphi\,\psi\right| \quad \text{a.s. and in } L^1. \qquad \square$$

7 Proof of Theorem 2

Fix a density ρ and assume there is a continuous curve $L = (L_t : t \geq 0)$ and density functions $u = (u(r,t) : r \in \mathbb{R}, t \geq 0)$ such that (u, L) solves the FBP. Actually, the only thing we use about being a solution is the representation formula (6). It is convenient to stress the semigroup property of the solution so we call the solution $S_t\rho := u(\cdot, t)$ and notice that the operator S_t is a semigroup. The following theorem shows that the solution is in between the barriers.

Theorem 5. *Let (u, L) be a solution of the FBP in $(0, T]$. Let $t \in (0, T]$ and $\delta \in \{2^{-n}t : n \in \mathbb{N}\}$. Then*

$$S_t^{\delta,-}\rho \preccurlyeq S_t\rho \preccurlyeq S_t^{\delta,+}\rho, \qquad t = k\delta. \tag{50}$$

We show (50) first for time $\delta = 2^{-n}t$ and then use induction to extend to times $k\delta$.

Proposition 8. *For all $r \in \mathbb{R}$, we have*

$$F(r; S_\delta\rho) \leq F(r; S_\delta^{\delta,+}\rho) \tag{51}$$

$$F(r; S_\delta\rho) \geq F(r; S_\delta^{\delta,-}\rho). \tag{52}$$

Proof. If $r \leq L_\delta^{\delta,+}$, by definition of $L_\delta^{\delta,+}$, we have $F(r; S_\delta^{\delta,+}\rho) = F(L_\delta^{\delta,+}; \rho) = 1 \geq F(r; S_\delta\rho)$. If $r > L_\delta^{\delta,+}$, using the Brownian motion representation (6) of $S_t\rho$ with $\tau^L = \inf\{t > 0 : B_t \leq L_t\}$, we get

$$\begin{aligned} F(r; S_\delta^{\delta,+}\rho) = e^\delta \int \rho(x) P_x\big(B_\delta \geq r\big)\mathrm{d}x &\geq e^\delta \int \rho(x) P_x\big(B_\delta \geq r; \tau^L > \delta\big)\mathrm{d}x \\ &= F(r; S_t\rho). \end{aligned}$$

This shows (51).

To show (52), recall the cut operator (13), and denote $\rho_0 := C_{e^{-\delta}}\rho$ and $\rho_1 := \rho - \rho_0$. We then have $\int \rho_0(r)\mathrm{d}r = e^{-\delta}$ and

$$F(r; S_\delta^{\delta,-}\rho) = e^\delta \int \rho_0(x) P_x\big(B_\delta \geq r\big)\mathrm{d}x. \tag{53}$$

We have

$$\begin{aligned} F(r; S_t\rho) &= e^\delta \int \rho(x) P_x\big(B_\delta \geq r; \tau^L > \delta\big)\mathrm{d}x \\ &= e^\delta \int \rho_0(x) P_x\big(B_\delta \geq r\big)\mathrm{d}x - e^\delta \int \rho_0(x) P_x\big(B_\delta \geq r; \tau^L \leq \delta\big)\mathrm{d}x \\ &\quad + e^\delta \int \rho_1(x) P_x\big(B_\delta \geq r; \tau^L > \delta\big)\mathrm{d}x. \end{aligned}$$

Thus, recalling (53), it suffices to show

$$e^{\delta}\int \rho_0(x)P_x\big(B_\delta \geq r;\tau^L \leq \delta\big)\mathrm{d}x \leq e^{\delta}\int \rho_1(x)P_x\big(B_\delta \geq r;\tau^L > \delta\big)\mathrm{d}x. \tag{54}$$

We have that

$$e^{\delta}\int \rho_0(x)P_x(\tau^L \leq \delta)\mathrm{d}x = e^{\delta}\int \rho_1(x)P_x(\tau^L > \delta)\mathrm{d}x, \tag{55}$$

where the last identity follows from subtracting identities

$$e^{\delta}\int (\rho_0(x)+\rho_1(x))P_x(\tau^L > \delta)\mathrm{d}x = \int S_\delta\rho(x)\mathrm{d}x = 1$$
$$e^{\delta}\int \rho_0(x)\mathrm{d}x = e^{\delta}\int C_{e^{-\delta}}\rho(x)\mathrm{d}x = 1.$$

We rewrite (54) as

$$\int \rho_0(x)\int_0^\delta h_x^L(ds)P_{L_s,s}\big(B_\delta \geq r\big)\mathrm{d}x \leq \int \rho_1(x)P_x(\tau^L \geq \delta)P_x\big(B_\delta \geq r|\tau^L > \delta\big)\mathrm{d}x, \tag{56}$$

where $P_{y,s}$ denotes the law of a Brownian motion starting from y at time s, and h_x^L denotes the cumulative distribution function of τ^L under $P_{x,0}$. In Section10.3.2 of [7], it has been proved that if L is a continuous curve then, for every r,

$$P_{L_t;t}\big(B_\delta \geq r\big) \leq P_x\big(B_\delta \geq r \,\big|\, \tau^L > \delta\big), \quad x > L_0,\ t \in [0,\delta).$$

From this and from (55), inequality (56) easily follows.

Remark. Dividing (54) by (55), we have proven the inequality

$$P_{\rho_1}(B_\delta \geq r|\tau^L > \delta) - P_{\rho_0}(B_\delta \geq r|\tau^L \leq \delta) \geq 0,$$

where P_{ρ_i} is the law of Brownian motion with initial distribution

$$P_{\rho_i}(B_0 \in A) = \|\rho_i\|_1^{-1}\int_A \rho_i(x)\mathrm{d}x.$$

Proof (Proof of Theorem 5). Recalling the definitions of C_m and G_t, Proposition 8 shows the following inequalities for $n = 1$:

$$\big(e^{\delta}G_\delta C_{e^{-\delta}}\big)^n\rho \preccurlyeq S_{n\delta}\rho \preccurlyeq \big(C_1e^{\delta}G_\delta\big)^n\rho. \tag{57}$$

Apply (57) with $n = 1$ to $S_{n\delta}\rho$ to get

$$(e^{\delta}G_\delta C_{1-e^{-\delta}})S_{n\delta}\rho \preccurlyeq S_\delta S_{n\delta}\rho \preccurlyeq (C_{e^{\delta}-1}e^{\delta}G_\delta)S_{n\delta}\rho. \tag{58}$$

Apply each inequality in (57) to the corresponding side in (58) to obtain

$$(e^{\delta}G_\delta C_{e^{1-\delta}})^{n+1}\rho \preccurlyeq (e^{\delta}G_\delta C_{e^{-\delta}})S_{n\delta}\rho \preccurlyeq S_{(n+1)\delta}\rho \preccurlyeq (C_1e^{\delta}G_\delta)S_{n\delta}\rho$$
$$\preccurlyeq (C_1e^{\delta}G_\delta)^{n+1}\rho,$$

where we have used that both G_δ and C_m are monotone, by Proposition 4.

8 Traveling Waves

Traveling waves. Fix N and let X_t be N-BBM. Let $X'_t := \{x - \min X_t : x \in X_t\}$ be the process as seen from the left-most particle. In this process there is always a particle at the origin. The following theorem has been proven by Durrett and Remenik [11] for a related Brunet-Derrida process. The proof in this case is very similar so we skip it.

Theorem 6. N-BBM *as seen from the left-most particle is Harris recurrent. Denote* ν_N *its unique invariant measure. Under* ν_N *the process has an asymptotic speed* α_N *given by*

$$\alpha_N = (N-1)\,\nu_N\big[\min(X \setminus \{0\})\big],$$

that is the rate of branching of the $N-1$ *right-most particles times the expected distance between the left-most particle an the second left-most particle.*

N-BBM *starting with an arbitrary configuration converges in distribution to* ν_N *and*

$$\lim_{t\to\infty} \frac{\min X_t}{t} = \alpha_N.$$

Furthermore, α_N *converges to the asymptotic speed of the first particle in* BBM *with a finite initial configuration:*

$$\lim_{N\to\infty} \alpha_N = \sqrt{2}. \tag{59}$$

The analogous to limit (59) was proven by Berard and Gouéré [4] and Durrett and Mayberry [8] for Brunet-Derrida systems.

The traveling wave solutions of the FBP (2) (3) (4) are of the form $u(r,t) = w(r-\alpha t)$, where w must satisfy

$$\frac{1}{2}w'' + \alpha w' + w = 0, \quad w(0) = 0, \quad \int_0^\infty w(r)\mathrm{d}r = 1.$$

Groisman and Jonckheere [12,13] observed that for each speed $\alpha \geq \alpha_c = \sqrt{2}$ there is a solution w_α given by

$$w_\alpha(x) = \begin{cases} M_\alpha\, x e^{-\alpha x} & \text{if } \alpha = \sqrt{2} \\ M_\alpha\, e^{-\alpha x} \sinh\left(x\sqrt{\alpha^2 - 2}\right) & \text{if } \alpha > \sqrt{2} \end{cases}$$

where M_α is a normalization constant such that $\int w_\alpha = 1$. In fact w_α is the unique quasi stationary distribution for Brownian motion with drift $-\alpha$ and absorption rate $w'(0) = 1$; see Proposition 1 of Martínez and San Martín [17]. More precisely, calling $\mathcal{L}_\alpha w = \frac{1}{2}w'' + \alpha w'$, we have that w_α is the unique eigenvector for $\mathcal{L}_\alpha$ with eigenvalue -1. See [12] for the relation between quasi stationary distributions for absorbed Brownian motion and traveling wave solutions for the FBP.

Let X_t be the N-BBM process with initial configuration sampled from the stationary measure ν^N. Show that the empirical distribution of X_t converges to a measure with density $w_{\sqrt{2}}(\cdot - t\sqrt{2})$, as $N \to \infty$. This would be a *strong selection principle* for N-BBM [12,16]; the weak selection principle is already contained in (59), the stationary speed for the finite system converges to the minimal speed in the macroscopic system. A way to show this limit would be to control the particle-particle correlations in the ν_N distributed initial configuration. If instead we start with independent particles with distribution $w_{\sqrt{2}}$, then we can use Theorem 1 and the fact that $w_{\sqrt{2}}(r - t\sqrt{2})$ is a strong solution of the FBP to prove converge of the empirical measure to this solution.

Acknowledgments. PAF thanks Pablo Groisman, Matthieu Jonckheere and Julio Rossi for illuminating discussions on quasi stationary distributions and free boundary problems. PAF thanks Gran Sasso Science Institute and University of Paris Diderot for warm hospitality. We thank kind hospitality at Institut Henri Poincaré during the trimester *Stochastic dynamics out of equilibrium*, where part of this work was performed. We finally thank the referee for his careful reading and helpful comments.

References

1. Berestycki, J., Brunet, É., Derrida, B.: Exact solution and precise asymptotics of a Fisher–KPP type front. J. Phys. A Math. Theor. **51**(3), 035204 (2017)
2. Berestycki, J., Brunet, É., Penington, S.: Global existence for a free boundary problem of Fisher-KPP type. arXiv:1805.03702 (2018)
3. Brunet, É., Derrida, B.: Shift in the velocity of a front due to a cutoff. Phys. Rev. E **56**, 2597–2604 (1997)
4. Bérard, J., Gouéré, J.-B.: Brunet-Derrida behavior of branching-selection particle systems on the line. Comm. Math. Phys. **298**(2), 323–342 (2010)
5. Carinci, G., De Masi, A., Giardinà, C., Presutti, E.: Hydrodynamic limit in a particle system with topological interactions. Arab. J. Math. **3**(4), 381–417 (2014)
6. Carinci, G., De Masi, A., Giardinà, C., Presutti, E.: Super-hydrodynamic limit in interacting particle systems. J. Stat. Phys. **155**(5), 867–887 (2014)
7. Carinci, G., De Masi, A., Giardinà, C., Presutti, E.: Free boundary problems in PDEs and particle systems. In: SpringerBriefs in Mathematical Physics, vol. 12, Springer, Cham (2016)
8. Durrett, R., Mayberry, J.: Evolution in predator-prey systems. Stochast. Process. Appl. **120**(7), 1364–1392 (2010)
9. De Masi, A., Ferrari, P.A.: Separation versus diffusion in a two species system. Braz. J. Probab. Stat. **29**(2), 387–412 (2015)
10. De Masi, A., Ferrari, P.A., Presutti, E.: Symmetric simple exclusion process with free boundaries. Probab. Theor. Relat. Fields **161**(1–2), 155–193 (2015)
11. Durrett, R., Remenik, D.: Brunet-Derrida particle systems, free boundary problems and Wiener-Hopf equations. Ann. Probab. **39**(6), 2043–2078 (2011)
12. Groisman, P., Jonckheere, M.: Front propagation and quasi-stationary distributions: the same selection principle? arXiv:1304.4847 (2013)
13. Groisman, P., Jonckheere, M.: Front propagation and quasi-stationary distributions for one-dimensional Lévy processes. Electron. Commun. Probab. **23**, 11 (2018)

14. Lee, J.: A free boundary problem with non local interaction. Math. Phys. Anal. Geom. **21**(3), 24 (2018)
15. Maillard, P.: The number of absorbed individuals in branching Brownian motion with a barrier. Ann. Inst. Henri Poincaré Probab. Stat. **49**(2), 428–455 (2013)
16. Maillard, P.: Speed and fluctuations of N-particle branching Brownian motion with spatial selection. Probab. Theor. Relat. Fields **166**(3–4), 1061–1173 (2016)
17. Martínez, S., San Martín, J.: Quasi-stationary distributions for a Brownian motion with drift and associated limit laws. J. Appl. Probab. **31**(4), 911–920 (1994)
18. Shi, Z.: Branching Random Walks. Lecture notes from the 42nd Probability Summer School held in Saint Flour, 2012. Lecture Notes in Mathematics, vol. 2151. Springer, Cham (2015)

1D Mott Variable-Range Hopping with External Field

Alessandra Faggionato(✉)

University La Sapienza, P.le Aldo Moro 2, Rome, Italy
faggiona@mat.uniroma1.it

Abstract. Mott variable-range hopping is a fundamental mechanism for electron transport in disordered solids in the regime of strong Anderson localization. We give a brief description of this mechanism, recall some results concerning the behavior of the conductivity at low temperature and describe in more detail recent results (obtained in collaboration with N. Gantert and M. Salvi) concerning the one-dimensional Mott variable-range hopping under an external field.

Keywords: Random walk in random environment · Mott variable-range hopping · Linear response · Einstein relation

1 Mott Variable-Range Hopping

Mott variable range hopping is a mechanism of phonon-assisted electron transport taking place in amorphous solids (as doped semiconductors) in the regime of strong Anderson localization. It has been introduced by N.F. Mott in order to explain the anomalous non-Arrhenius decay of the conductivity at low temperature [20–23,25].

Let us consider a doped semiconductor, which is given by a semiconductor with randomly located foreign atoms (called *impurities*). We write $\xi := \{x_i\}$ for the set of impurity sites. For simplicity we treat spinless electrons. Then, due to Anderson localization, a generic conduction electron is described by a quantum wavefunction localized around some impurity site x_i, whose energy is denoted by E_i. This allows, at a first approximation, to think of the conduction electrons as classical particles which can lie only on the impurity sites, subject to the constraint of site exclusion (due to Pauli's exclusion principle). As a consequence, a microscopic configuration is described by an element $\eta \in \{0,1\}^{\xi}$, where $\eta_{x_i} = 1$ if and only if an electron is localized around the impurity site x_i. The dynamics is then described by an exclusion process, where the probability rate for a jump from x_i to x_j is given by (cf. [1])

$$\mathbb{1}(\eta_{x_i} = 1,\ \eta_{x_j} = 0) \exp\{-2\zeta|x_i - x_j| - \beta\{E_j - E_i\}_+\}\,. \tag{1}$$

Above $\beta = 1/kT$ (k being the Boltzmann's constant and T the absolute temperature), while $1/\zeta$ is the localization length. A physical analysis suggests that the energy marks $\{E_i\}$ can be modeled by i.i.d. random variables. In inorganic doped semiconductors, when the Fermi energy is set equal to zero, the common

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 550–559, 2019.
https://doi.org/10.1007/978-3-030-15096-9_19

distribution ν of E_i is of the form $c\,|E|^\alpha dE$ on some interval $[-A, A]$, where c is the normalization constant and α is a nonnegative exponent.

At low temperature (i.e. large β) the form of the jump rates (1) suggests that long jumps can be facilitated when the energetic cost $\{E_j - E_i\}_+$ is small. This facilitation leads to an anomalous conductivity behavior for $d \geq 2$. Indeed, according to Mott's law, in an isotropic medium the conductivity matrix $\sigma(\beta)$ can be approximated (at logarithmic scale) by $\exp\big(-c\,\beta^{\frac{\alpha+1}{\alpha+1+d}}\big)\mathbb{1}$, where $\mathbb{1}$ denotes the identity matrix and c is a suitable positive constant c with negligible temperature-dependence.

The mathematical analysis of the above exclusion process presents several technical challenges and has been performed only when $\xi \equiv \mathbb{Z}^d$ (absence of geometric disorder) and with jumps restricted to nearest-neighbors (cf. [7,24]). This last assumption does not fit with the low temperature regime, where anomalous conductivity takes place.

To investigate Mott variable range hopping at low temperature, in the regime of low impurity density some effective models have been proposed. One is given by the random Miller-Abrahams resistor network [1,19]. Another effective model is the following (cf. [9]): one approximates the localized electrons by classical non–interacting (independent) particles moving according to random walks with jump probability rate given by the transition rate (1) multiplied by a suitable factor which keeps trace of the exclusion principle. To be more precise, given $\gamma \in \mathbb{R}$, we call μ_γ the product probability measure on $\{0,1\}^\xi$ with $\mu_\gamma(\eta_{x_i}) = \frac{e^{-\beta(E_i-\gamma)}}{1+e^{-\beta(E_i-\gamma)}}$. Then it is simple to check that μ_γ is a reversible distribution for the exclusion process. In the independent particles approximation, the probability rate for a jump from x_i to x_j is given by

$$\mu_\gamma(\eta_{x_i} = 1,\ \eta_{x_j} = 0)\exp\{-2\zeta|x_i - x_j| - \beta\{E_j - E_i\}_+\}\,. \tag{2}$$

It is simple to check that at low temperature, i.e. large β, (2) is well approximated by (cf. [1])

$$\exp\{-2\zeta|x_i - x_j| - \frac{\beta}{2}(|E_i - \gamma| + |E_j - \gamma| + |E_i - E_j|)\}\,. \tag{3}$$

In what follows, without loss of generality, we shift the energy so that the Fermi energy γ equals zero, we take $2\zeta = 1$ and we replace $\beta/2$ by β. Since the above random walks are independent, we can restrict to the analysis of a single random walk, which we call *Mott random walk*.

2 Mott Random Walk and Bounds on the Diffusion Matrix

We give the formal definition of Mott random walk as random walk in a random environment. The environment is given by a marked simple point process $\omega = \{(x_i, E_i)\}$ where $\{x_i\} \subset \mathbb{R}^d$ and the (energy) marks are i.i.d. random variables with common distribution ν having support on some finite interval $[-A, A]$.

Given a realization of the environment ω, Mott random walk is the continuous-time random walk X_t^ω with state space $\{x_i\}$ and probability rate for a jump from x_i to $x_j \neq x_i$ given by

$$r_{x_i,x_j}(\omega) = \exp\{-|x_i - x_j| - \beta(|E_i| + |E_j| + |E_i - E_j|)\} \ . \tag{4}$$

As already mentioned, one expects for $d \geq 2$ that the contribution to the transport of long jumps dominates as $\beta \to \infty$. In [3] a quenched invariance principle for Mott random walk has been proved. Calling $D(\beta)$ the diffusion matrix of the limiting Brownian motion, in [8,9] bounds in agreement with Mott's law have been obtained for the diffusion matrix. More precisely, under very general conditions on the isotropic environment and taking ν of the form $c\,|E|^\alpha dE$ for some $\alpha \geq 0$ and on some interval $[-A, A]$, it has been proved that for suitable β-independent positive constant $c_1, c_2, \kappa_1, \kappa_2$ it holds

$$c_1 \exp\left\{-\kappa_1\, \beta^{\frac{\alpha+1}{\alpha+1+d}}\right\} \mathbb{1} \leq D(\beta) \leq c_2 \exp\left\{-\kappa_2\, \beta^{\frac{\alpha+1}{\alpha+1+d}}\right\} \mathbb{1} \, . \tag{5}$$

We point out that for a genuinely nearest–neighbor random walk $D(\beta)$ would have an Arrenhius decay, i.e. $D(\beta) \approx e^{-c\beta}\mathbb{1}$, thus implying that (5) is determined by long jumps.

In dimension $d = 1$ long jumps do not dominate. Indeed, the following has been derived for $d = 1$ in [2] when ω is a renewal marked simple point process. We label the points in increasing order, i.e. $x_i < x_{i+1}$, take $x_0 = 0$ and assume that $\mathbb{E}[x_1^2] < \infty$. Then the following holds [2]:

(i) If $\mathbb{E}[e^{x_1}] < \infty$, then a quenched invariance principle holds and the diffusion coefficient satisfies

$$c_1 \exp\{-\kappa_1\, \beta\} \leq D(\beta) \leq c_2 \exp\{-\kappa_2\, \beta\} \ , \tag{6}$$

for β–independent positive constants $c_1, c_2, \kappa_1, \kappa_2$;

(ii) If $\mathbb{E}[e^{x_1}] = \infty$, then the random walk is subdiffusive. More precisely, an annealed invariance principle holds with zero diffusion coefficient.

We point out that the above 1d results hold also for a larger class of jump rates and energy distributions ν [2]. Moreover, we stress that the above bounds (5) and (6) refer to the diffusion matrix $D(\beta)$ and not to the conductivity matrix $\sigma(\beta)$. On the other hand, believing in the Einstein relation (which states that $\sigma(\beta) = \beta D(\beta)$), the above bounds would extend to $\sigma(\beta)$, hence one would recover lower and upper bounds on the conductivity matrix in agreement with the physical Mott law for $d \geq 2$ and with the Arrenhius–type decay for $d = 1$. The rigorous derivation of the Einstein relation for Markov processes in random environment is in general a difficult task and has been the object of much investigation also in the last years (cf. e.g. [10,11,13–18]). In what follows, we will concentrate on the effect of perturbing 1d Mott random walk by an external field and on the validity of the Einstein relation. Hopefully, progresses on the Einstein relation for Mott random walk in higher dimension will be obtained in the future.

3 Biased 1D Mott Random Walk

We take $d = 1$ and label points $\{x_i\}$ in increasing order with the convention that $x_0 := 0$ (in particular, we assume the origin to be an impurity site). It is convenient to define Z_i as the interpoint distance $Z_i := x_{i+1} - x_i$ (Fig. 1).

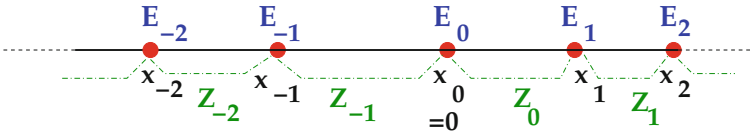


Fig. 1. Points x_i, energy marks E_i and interpoint distances Z_i.

We make the following assumptions:

(A1) The random sequence $(Z_k, E_k)_{k\in\mathbb{Z}}$ is stationary and ergodic w.r.t. shifts;
(A2) $\mathbb{E}[Z_0]$ is finite;
(A3) $\mathbb{P}(\omega = \tau_\ell \omega)$ is zero for all $\ell \in \mathbb{Z} \setminus \{0\}$;
(A4) There exists some constant $d > 0$ satisfying $\mathbb{P}(Z_0 \geq d) = 1$.

Note that we do not restrict to the physically relevant energy mark distributions ν which are of the form $c|E|^\alpha dE$ on some interval $[-A, A]$.

Given $\lambda \in [0, 1)$ we consider the biased generalized Mott random walk $X_t^{\omega,\lambda}$ on $\{x_i\}$ with jump probability rates given by

$$r^\lambda_{x_i,x_j}(\omega) = \exp\{-|x_i - x_j| + \lambda(x_j - x_i) - u(E_i, E_j)\}\,, \qquad x_i \neq x_j\,, \tag{7}$$

and starting at the origin. Above, u is a given bounded and symmetric function. Note that we do not restrict to (4). The special form (4) is relevant when studying the regime $\beta \to \infty$, on the other hand here we are interested in the system at a fixed temperature (which is included in the function u) under the effect of an external field.

In the rest, it is convenient to set $r^\lambda_{x,x}(\omega) \equiv 0$. We also point out that one can easily prove that the random walk $X_t^{\omega,\lambda}$ is well defined since $\lambda \in [0, 1)$.

The following result, obtained in [5], concerns the ballistic/sub–ballistic regime:

Proposition 1 [5]. *For $\mathbb{P}$–a.a. ω the random walk $X_t^{\omega,\lambda}$ is transient to the right, i.e. $\lim_{t\to\infty} X_t^{\omega,\lambda} = +\infty$ a.s.*

Theorem 1 [5]. *Fix $\lambda \in (0, 1)$.*

(i) If $\mathbb{E}\big[e^{(1-\lambda)Z_0}\big] < \infty$ and u is continuous, then for $\mathbb{P}$–a.a. ω the following limit exists

$$v_X(\lambda) := \lim_{t\to\infty} \frac{X_t^{\omega,\lambda}}{t} \qquad a.s.$$

and moreover it is deterministic, finite and strictly positive.

(ii) If $\mathbb{E}\big[e^{-(1+\lambda)Z_{-1}+(1-\lambda)Z_0}\big] = \infty$*, then for* $\mathbb{P}$*–a.a.* ω *it holds*

$$v_X(\lambda) := \lim_{t\to\infty} \frac{X_t^{\xi,\lambda}}{t} = 0 \qquad a.s.$$

As discussed in [5], the condition $\mathbb{E}\big[e^{(1-\lambda)Z_0}\big] = \infty$ does not imply that $v_X(\lambda) = 0$. On the other hand, if $(Z_k)_{k\in\mathbb{Z}}$ are i.i.d. (or even if Z_k, Z_{k+1} are independent for every k) and u is continuous, then the above two cases (i) and (ii) in Theorem 1 are exhaustive and one concludes that $\mathbb{E}\big[e^{(1-\lambda)Z_0}\big] < \infty$ if and only if $v_X(\lambda) > 0$, otherwise $v_X(\lambda) = 0$.

The above Theorem 1 extends also to the jump process associated to $X_t^{\omega,\lambda}$, i.e. to the discrete time random walk $Y_n^{\omega,\lambda}$ with probability $p^\lambda_{x_i,x_k}(\omega)$ of a jump from x_i to $x_k \neq x_i$ given by

$$p^\lambda_{x_i,x_k}(\omega) = \frac{r^\lambda_{x_i,x_j}(\omega)}{\sum_k r^\lambda_{x_i,x_k}(\omega)}\,.$$

In particular, if $\mathbb{E}\big[e^{(1-\lambda)Z_0}\big] < \infty$ and u is continuous then the random walk $Y_n^{\omega,\lambda}$ is ballistic ($v_Y(\lambda) > 0$), while if $\mathbb{E}\big[e^{-(1+\lambda)Z_{-1}+(1-\lambda)Z_0}\big] = \infty$ then the random walk $Y_n^{\omega,\lambda}$ is sub–ballistic (i.e. $v_Y(\lambda) = 0$). We point out that indeed the result has been proved in [5] first for the random walk $Y_n^{\omega,\lambda}$ and then extended to the continuous time case by a random time change argument.

We write $\tau_x\omega$ for the environment ω translated by $x \in \mathbb{R}$, more precisely we set $\tau_x\omega := \{(x_j - x, E_j)\}$ if $\omega = \{(x_j, E_j)\}$. We recall that the environment viewed from the walker $Y_n^{\omega,\lambda}$ is given by the discrete time Markov chain $(\tau_{Y_n^{\omega,\lambda}}\omega)_{n\geq 0}$. This is the crucial object to analyze in the ballistic regime. The following result concerning the environment viewed from the walker $Y_n^{\omega,\lambda}$ is indeed at the basis of the derivation of Theorem 1–(i) as well as the starting point for the analysis of the Einstein relation.

Theorem 2 [5,6]**.** *Fix* $\lambda \in (0,1)$*. Suppose that* $\mathbb{E}\big[e^{(1-\lambda)Z_0}\big] < \infty$ *and that* u *is continuous. Then the environment viewed from the walker* $Y_n^{\omega,\lambda}$ *admits an invariant and ergodic distribution* $\mathbb{Q}_\lambda$ *mutually absolutely continuous w.r.t.* $\mathbb{P}$*. Moreover, it holds*

$$v_Y(\lambda) = \mathbb{Q}_\lambda\big[\varphi_\lambda\big] \quad \textit{and} \quad v_X(\lambda) = \frac{v_Y(\lambda)}{\mathbb{Q}_\lambda\Big[1/\big(\sum_k r^\lambda_{0,x_k}\big)\Big]}\,,$$

where φ_λ *denotes the local drift, i.e.* $\varphi_\lambda(\omega) := \sum_i x_i p^\lambda_{0,x_i}(\omega)$*.*

We point that for $\lambda = 0$ the probability distribution $\mathbb{Q}_0$ defined as

$$d\mathbb{Q}_0 = \frac{\sum_k r^{\lambda=0}_{0,x_k}}{\mathbb{E}[\sum_k r^{\lambda=0}_{0,x_k}]} d\mathbb{P}$$

is indeed reversible for the environment viewed from the walker $Y_n^{\omega,\lambda=0}$, and that $v_Y(0) = v_X(0) = 0$. In what follows, when $\lambda = 0$ we will often drop λ from the notation (in particular, we will write simply $r_{x_i,x_j}(\omega)$, $p_{x_i,x_k}(\omega)$, X_t^ω, Y_n^ω).

The proof of Theorem 2 takes inspiration from the paper [4]. The main technical difficulty comes from the presence of arbitrarily long jumps, which does not allow to use standard techniques based on regeneration times. Following the method developed in [4] we have considered, for each positive integer ρ, the random walk obtained from $Y_n^{\omega,\lambda}$ by suppressing jumps between sites x_i, x_j with $|i-j| > \rho$ [5]. For this ρ-indexed random walk, under the same hypothesis of Theorem 2, we have proved that there exists a distribution $\mathbb{Q}_\lambda^{(\rho)}$ which is invariant and ergodic for the associated environment viewed from the walker, and that $\mathbb{Q}_\lambda^{(\rho)}$ is mutually absolutely continuous w.r.t. $\mathbb{P}$. In particular, the methods developed in [4] provide a probabilistic representation of the Radon–Nykodim $\frac{d\mathbb{Q}_\lambda^{(\rho)}}{d\mathbb{P}}$, which (together with a suitable analysis based on potential theory) allows to prove that $\mathbb{Q}_\lambda^{(\rho)}$ weakly converges to $\mathbb{Q}_\lambda$, and that $\mathbb{Q}_\lambda$ has indeed the nice properties stated in Theorem 2.

4 Linear Response and Einstein Relation for the Biased 1D Mott Random Walk

Let us assume again (A1), (A2), (A3), (A4) as in the previous section. The probabilistic representation of the Radon–Nykodim derivative $\frac{d\mathbb{Q}_\lambda^{(\rho)}}{d\mathbb{P}}$ mentioned in the above section is the starting point for the derivation of estimates on the Radon–Nykodim derivative $\frac{d\mathbb{Q}_\lambda}{d\mathbb{Q}_0}$:

Proposition 2 [6]. *Suppose that for some $p \geq 2$ it holds $\mathbb{E}\left[e^{pZ_0}\right] < +\infty$. Fix $\lambda_0 \in (0,1)$. Then*

$$\sup_{\lambda\in[0,\lambda_0]} \left\| \frac{d\mathbb{Q}_\lambda}{d\mathbb{Q}_0} \right\|_{L^p(\mathbb{Q}_0)} < \infty\,.$$

The above proposition allows to prove the continuity of the expected value $\mathbb{Q}_\lambda(f)$ of suitable functions f:

Theorem 3 [6]. *Suppose that $\mathbb{E}[\mathrm{e}^{pZ_0}] < \infty$ for some $p \geq 2$ and let q be the conjugate exponent of p, i.e. q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. Then, for any $f \in L^q(\mathbb{Q}_0)$ and $\lambda \in [0,1)$, it holds that $f \in L^1(\mathbb{Q}_\lambda)$ and the map*

$$[0,1) \ni \lambda \mapsto \mathbb{Q}_\lambda(f) \in \mathbb{R} \tag{8}$$

is continuous.

Without entering into the details of the proof (which can be found in [6]) we give some comments on the derivation of Theorem 3 from Proposition 2. To this aim, we take for simplicity $p = 2$. Hence we are supposing that $\mathbb{E}(\mathrm{e}^{2Z_0}) < \infty$ and that $f \in L^2(\mathbb{Q}_0)$, and we want to prove that $f \in L^1(\mathbb{Q}_\lambda)$ and that the function in (8) is continuous. For simplicity, let us restrict to its continuity at $\lambda = 0$. The fact that $f \in L^1(\mathbb{Q}_\lambda)$ follows by writing $\mathbb{Q}_\lambda(f) = \mathbb{Q}_0(\frac{d\mathbb{Q}_\lambda}{d\mathbb{Q}_0} f)$ and then by applying Schwarz inequality and Proposition 2. The proof of the continuity at $\lambda = 0$

is more involved. We recall that by Kakutani's theorem balls are compact for the $L^2(\mathbb{Q}_0)$–weak topology. Hence, due to Proposition 2, the family of Radon–Nykodim derivatives $\frac{d\mathbb{Q}_\lambda}{d\mathbb{Q}_0}$, $\lambda \in [0, \lambda_0]$, is relatively compact for the $L^2(\mathbb{Q}_0)$–weak topology. In [6] we then prove that any limit point of this family is given by $\mathbb{1}$. As a byproduct of the representation $\mathbb{Q}_\lambda(f) = \mathbb{Q}_0(\frac{d\mathbb{Q}_\lambda}{d\mathbb{Q}_0} f)$ and of the weak convergence $\frac{d\mathbb{Q}_\lambda}{d\mathbb{Q}_0} \rightharpoonup \mathbb{1}$, we get the continuity of (8) at $\lambda = 0$.

We now move to the study of $\partial_{\lambda=0}\mathbb{Q}_\lambda(f)$. To this aim we introduce the operator $\mathbb{L}_0 : L^2(\mathbb{Q}_0) \to L^2(\mathbb{Q}_0)$ as

$$\mathbb{L}_0 f(\omega) = \sum_k p_{0,x_k}(\omega)[f(\tau_{x_k}\omega) - f(\omega)]\,, \qquad f \in L^2(\mathbb{Q}_0)\,.$$

We recall that a function f belongs to $L^2(\mathbb{Q}_0) \cap H_{-1}$ if there exists $C > 0$ such that

$$|\langle f, g\rangle| \le C\langle g, -\mathbb{L}_0 g\rangle^{1/2} \qquad \forall g \in L^2(\mathbb{Q}_0)\,.$$

Above $\langle\cdot,\cdot\rangle$ is the scalar product in $L^2(\mathbb{Q}_0)$. Due to the theory developed by Kipnis and Varadhan [12], for any $f \in L^2(\mathbb{Q}_0) \cap H_{-1}$, we have the weak convergence

$$\frac{1}{\sqrt{n}}\Big(\sum_{j=0}^{n-1} f(\omega_j), \sum_{j=0}^{n-1} \varphi(\omega_j)\Big) \overset{n\to\infty}{\to} (N^f, N^\varphi)$$

for a suitable 2d gaussian vector (N^f, N^φ). Above φ denotes the local drift φ_λ with $\lambda = 0$ (cf. Theorem 2) and $\omega_j = \tau_{Y_j^\omega}\omega$, i.e. $(\omega_n)_n$ represents the environment viewed from the walker Y_n^ω.

Finally, we need another ingredient coming from the theory of square integrable forms in order to present our next theorem. We consider the space $\Omega \times \mathbb{Z}$ endowed with the measure M defined by

$$M(v) = \mathbb{Q}_0\Big[\sum_{k\in\mathbb{Z}} p_{0,x_k} v(\cdot, k)\Big]\,, \qquad \forall v : \Omega \times \mathbb{Z} \to \mathbb{R} \qquad \text{Borel, bounded}\,.$$

A generic Borel function $v : \Omega \times \mathbb{Z} \to \mathbb{R}$ will be called a *form*. $L^2(M)$ is known as the space of *square integrable forms*. Given a function $g = g(\omega)$ we define

$$\nabla g(\omega, k) := g(\tau_{x_k}\omega) - g(\omega)\,. \tag{9}$$

If $g \in L^2(\mathbb{Q}_0)$, then $\nabla g \in L^2(M)$. The closure in M of the subspace $\{\nabla g : g \in L^2(\mathbb{Q}_0)\}$ is the set of the so called *potential forms* (its orthogonal subspace is given by the so called *solenoidal forms*).

Take again $f \in H_{-1} \cap L^2(\mathbb{Q}_0)$ and, given $\varepsilon > 0$, define $g_\varepsilon^f \in L^2(\mathbb{Q}_0)$ as the unique solution of the equation

$$(\varepsilon - \mathbb{L}_0) g_\varepsilon^f = f\,. \tag{10}$$

As discussed with more details in [6] as ε goes to zero the family of potential forms ∇g_ε^f converges in $L^2(M)$ to a potential form h^f:

$$h^f = \lim_{\varepsilon\downarrow 0} \nabla g_\varepsilon^f \qquad \text{in } L^2(M)\,. \tag{11}$$

Theorem 4 [6]. *Suppose* $\mathbb{E}(e^{pZ_0}) < \infty$ *for some* $p > 2$. *Then, for any* $f \in H_{-1} \cap L^2(\mathbb{Q}_0)$, $\partial_{\lambda=0}\mathbb{Q}_\lambda(f)$ *exists. Moreover the following two probabilistic representations hold with* $h = h^f$ *(see* (11)*):*

$$\partial_{\lambda=0}\mathbb{Q}_\lambda(f) = \begin{cases} \mathbb{Q}_0\big[\sum_{k\in\mathbb{Z}} p_{0,x_k}(x_k - \varphi)h(\cdot,k)\big] \\ -\mathrm{Cov}(N^f, N^\varphi) \end{cases}. \tag{12}$$

We point out that a covariance representation of $\partial_{\lambda=0}\mathbb{Q}_\lambda(f)$ as the second one in (12) appears also in [11] and [18].

We give some comments on the derivation of Theorem 4 up to the first representation in (12). Fix $f \in H_{-1} \cap L^2(\mathbb{Q}_0)$, thus implying that $\mathbb{Q}_0(f) = 0$. Given $\varepsilon > 0$ take $g_\varepsilon \in L^2(\mathbb{Q}_0)$ as the unique solution of the equation $(\varepsilon - \mathbb{L}_0)g_\varepsilon = f$, i.e. $g_\varepsilon = g^f_\varepsilon$ with g^f_ε as in (10). Then we can write

$$\frac{\mathbb{Q}_\lambda(f) - \mathbb{Q}_0(f)}{\lambda} = \frac{\mathbb{Q}_\lambda(f)}{\lambda} = \frac{\varepsilon\mathbb{Q}_\lambda(g_\varepsilon)}{\lambda} - \frac{\mathbb{Q}_\lambda(\mathbb{L}_0 g_\varepsilon)}{\lambda}. \tag{13}$$

The idea is to take first the limit $\varepsilon \to 0$, afterwards the limit $\lambda \to 0$. By the results of Kipnis and Varadhan [12] we have that $\varepsilon\mathbb{Q}_\lambda(g_\varepsilon)$ is negligible as $\varepsilon \to 0$. Hence we have $\partial_{\lambda=0}\mathbb{Q}_\lambda(f) = -\lim_{\lambda\to 0} \frac{\mathbb{Q}_\lambda(\mathbb{L}_0 g_\varepsilon)}{\lambda}$ if the latter exists. On the other hand we have the following identity and approximations for ε, λ small:

$$\begin{aligned} -\frac{\mathbb{Q}_\lambda[\mathbb{L}_0 g_\varepsilon]}{\lambda} &= \mathbb{Q}_\lambda\Big[\frac{(\mathbb{L}_\lambda - \mathbb{L}_0)g_\varepsilon}{\lambda}\Big] = \mathbb{Q}_\lambda\Big[\sum_{k\in\mathbb{Z}} \frac{p^\lambda_{0,x_k} - p_{0,x_k}}{\lambda}(g_\varepsilon(\tau_{x_k}\cdot) - g_\varepsilon)\Big] \\ &\approx \mathbb{Q}_\lambda\Big[\sum_{k\in\mathbb{Z}} \partial_{\lambda=0}p^\lambda_{0,x_k}\, h(\cdot,k)\Big] \approx \mathbb{Q}_0\Big[\sum_{k\in\mathbb{Z}} \partial_{\lambda=0}p^\lambda_{0,x_k}\, h(\cdot,k)\Big] \\ &= \mathbb{Q}_0\big[\sum_{k\in\mathbb{Z}} p_{0,x_k}(x_k - \varphi)h(\cdot,k)\big], \end{aligned} \tag{14}$$

where $\mathbb{L}_\lambda f(\omega) = \sum_k p^\lambda_{0,x_k}(\omega)[f(\tau_{x_k}\omega) - f(\omega)]$ and $h = h^f$ (see (11)). Roughly, the first identity in (14) follows from the stationary of $\mathbb{Q}_\lambda$ for the environment viewed from $Y^{\omega,\lambda}_n$, the first approximation in the second line follows from (11), the second approximation in the second line follows from Theorem 3, the identity in the third line follows from the equality $\partial_{\lambda=0}p^\lambda_{0,k} = p_{0,x_k}(x_k - \varphi)$.

The above steps are indeed rigorously proved in [6]. Since, as already observed, $\partial_{\lambda=0}\mathbb{Q}_\lambda(f) = -\lim_{\lambda\to 0} \frac{\mathbb{Q}_\lambda(\mathbb{L}_0 g_\varepsilon)}{\lambda}$, the content of Theorem 4 up to the first representation in (12) follows from (14).

We conclude this section with the Einstein relation. To this aim we denote by D_X the diffusion coefficient associated to X^ω_t and by D_Y the diffusion coefficient associated to Y^ω_n. D_X is the variance of the Brownian motion to which X^ω_t converges under diffusive rescaling, and a similar definition holds for D_Y.

Theorem 5 [6]. *The following holds:*

(i) If $\mathbb{E}[e^{2Z_0}] < \infty$, *then* $v_Y(\lambda)$ *and* $v_X(\lambda)$ *are continuous functions of* λ;
(ii) If $\mathbb{E}[e^{pZ_0}] < \infty$ *for some* $p > 2$, *then the Einstein relation is fulfilled, i.e.*

$$\partial_{\lambda=0}v_Y(\lambda) = D_Y \qquad \text{and} \qquad \partial_{\lambda=0}v_X(\lambda) = D_X\,. \tag{15}$$

If we make explicit the temperature dependence in the jump rates (7) we would have

$$r^{\lambda}_{x_i,x_j}(\omega) = \exp\{-|x_i - x_j| + \lambda\beta(x_j - x_i) - \beta u(E_i, E_j)\}\ ,$$

where λ is the strength of the external field. Then Einstein relation (15) takes the more familiar (from a physical viewpoint) form

$$\partial_{\lambda=0} v_Y(\lambda,\beta) = \beta D_Y(\beta) \qquad \text{and} \qquad \partial_{\lambda=0} v_X(\lambda,\beta) = \beta D_X(\beta)\,.$$

We conclude by giving some ideas behind the proof of the Einstein relation for $v_Y(\lambda)$ in Theorem 5. We do not fix the details, but only the main arguments. By applying Theorem 4 with $f := \varphi$ and using the first representation in the r.h.s. of (12), one can express $\partial_{\lambda=0}\mathbb{Q}_\lambda[\varphi]$ as a function of $h = h^\varphi$:

$$\partial_{\lambda=0}\mathbb{Q}_\lambda[\varphi] = \mathbb{Q}_0\Big[\sum_{k\in\mathbb{Z}} p_{0,x_k}(x_k - \varphi)h(\cdot,k)\Big]\,.$$

Recall that $v_Y(\lambda) = \mathbb{Q}_\lambda\big[\varphi_\lambda\big]$ (cf. Theorem 2) and that $v_Y(0) = 0$. In [6] we have proved the following approximations and identities (for the last identity see the conclusion of Sect. 9.1 in [6]):

$$\begin{aligned}
\frac{v_Y(\lambda) - v_Y(0)}{\lambda} = \frac{v_Y(\lambda)}{\lambda} = \frac{\mathbb{Q}_\lambda[\varphi_\lambda]}{\lambda} &= \mathbb{Q}_\lambda\big[\frac{\varphi_\lambda - \varphi}{\lambda}\big] + \frac{\mathbb{Q}_\lambda[\varphi] - \mathbb{Q}_0[\varphi]}{\lambda} \\
&\approx \mathbb{Q}_0\big[\partial_{\lambda=0}\varphi_\lambda\big] + \partial_{\lambda=0}\mathbb{Q}_\lambda[\varphi] \\
&= \mathbb{Q}_0\big[\partial_{\lambda=0}\varphi_\lambda\big] + \mathbb{Q}_0\Big[\sum_{k\in\mathbb{Z}} p_{0,x_k}(x_k - \varphi)h(\cdot,k)\Big] \\
&= \mathbb{Q}_0\Big[\sum_{k\in\mathbb{Z}} p_{0,x_k}(x_k - \varphi)(x_k + h(\cdot,k))\Big] = D_Y\,.
\end{aligned}$$

Acknowledgements. It is a pleasure to thank the Institut Henri Poincaré and the Centre Emile Borel for the kind hospitality and support during the trimester "Stochastic Dynamics Out of Equilibrium", as well as the organizers of this very stimulating trimester.
If you want to cite this proceeding and in particular some result contained in it, please cite also the article where the result appeared, so that the contribution of my coworkers can be recognised.

References

1. Ambegoakar, V., Halperin, B.I., Langer, J.S.: Hopping conductivity in disordered systems. Phys. Rev. B **4**, 2612–2620 (1971)
2. Caputo, P., Faggionato, F.: Diffusivity in one-dimensional generalized Mott variable-range hopping models. Ann. Appl. Probab. **19**, 1459–1494 (2009)
3. Caputo, P., Faggionato, A., Prescott, T.: Invariance principle for Mott variable range hopping and other walks on point processes. Ann. Inst. H. Poincaré Probab. Stat. **49**, 654–697 (2013)

4. Comets, F., Popov, S.: Ballistic regime for random walks in random environment with unbounded jumps and Knudsen billiards. Ann. Inst. H. Poincaré Probab. Stat. **48**, 721–744 (2012)
5. Faggionato, A., Gantert, N., Salvi, M.: The velocity of 1D Mott variable range hopping with external field. Ann. Inst. H. Poincaré Probab. Statist. **54**, 1165–1203 (2018)
6. Faggionato, A., Gantert, N., Salvi, M.: Einstein relation and linear response in one–dimensional Mott variable–range hopping. Preprint arXiv:1708.09610 (2017)
7. Faggionato, A., Martinelli, F.: Hydrodynamic limit of a disordered lattice gas. Probab. Theory Relat. Fields **127**, 535–608 (2003)
8. Faggionato, A., Mathieu, P.: Mott law as upper bound for a random walk in a random environment. Commun. Math. Phys. **281**, 263–286 (2008)
9. Faggionato, A., Schulz-Baldes, H., Spehner, D.: Mott law as lower bound for a random walk in a random environment. Commun. Math. Phys. **263**, 21–64 (2006)
10. Gantert, N., Mathieu, P., Piatnitski, A.: Einstein relation for reversible diffusions in a random environment. Commun. Pure Appl. Math. **65**, 187–228 (2012)
11. Gantert, N., Guo, X., Nagel, J.: Einstein relation and steady states for the random conductance model. Ann. Probab. **45**, 2533–2567 (2017)
12. Kipnis, C., Varadhan, S.R.S.: Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusion. Commun. Math. Phys. **104**, 1–19 (1986)
13. Komorowski, T., Olla, S.: Einstein relation for random walks in random environments. Stoch. Process. Appl. **115**, 1279–1301 (2005)
14. Komorowski, T., Olla, S.: On mobility and Einstein relation for tracers in time-mixing random environments. J. Stat. Phys. **118**, 407–435 (2005)
15. Lebowitz, J.L., Rost, H.: The Einstein relation for the displacement of a test particle in a random environment. Stoch. Process. Appl. **54**, 183–196 (1994)
16. Loulakis, M.: Einstein relation for a tagged particle in simple exclusion processes. Commun. Math. Phys. **229**, 347–367 (2005)
17. Loulakis, M.: Mobility and Einstein relation for a tagged particle in asymmetric mean zero random walk with simple exclusion. Ann. Inst. H. Poincaré Probab. Stat. **41**, 237–254 (2005)
18. Mathieu, P., Piatnitski, A.: Steady states, fluctuation-dissipation theorems and homogenization for reversible diffusions in a random environment. Arch. Ration. Mech. Anal. **230**, 277–320 (2018)
19. Miller, A., Abrahams, E.: Impurity conduction at low concentrations. Phys. Rev. **120**, 745–755 (1960)
20. Mott, N.F.: On the transition to metallic conduction in semiconductors. Can. J. Phys. **34**, 1356–1368 (1956)
21. Mott, N.F.: Conduction in non-crystalline materials III. Localized states in a pseudogap and near extremities of conduction and valence bands. Philos. Mag. **19**, 835–852 (1969)
22. Mott, N.F., Davis, E.A.: Electronic Processes in Non-Crystaline Materials. Oxford University Press, New York (1979)
23. Pollak, M., Ortuño, M., Frydman, A.: The Electron Glass. Cambridge University Press, Cambridge (2013)
24. Quastel, J.: Bulk diffusion in a system with site disorder. Ann. Probab. **34**, 1990–2036 (2006)
25. Shklovskii, B., Efros, A.L.: Electronic Properties of Doped Semiconductors. Springer, Berlin (1984)

Invariant Measures in Coupled KPZ Equations

Tadahisa Funaki[1,2(✉)]

[1] Waseda University, Tokyo, Japan
[2] University of Tokyo, Tokyo, Japan
funaki@ms.u-tokyo.ac.jp

Abstract. We discuss coupled KPZ (Kardar-Parisi-Zhang) equations. The motivation comes from the study of nonlinear fluctuating hydrodynamics, cf. [11,12]. We first give a quick overview of results of Funaki and Hoshino [6], in particular, two approximating equations, trilinear condition (T) for coupling constants Γ, invariant measures and global-in-time existence of solutions. Then, we study at heuristic level the role of the trilinear condition (T) in view of invariant measures and renormalizations for 4th order terms. Ertaş and Kardar [2] gave an example which does not satisfy (T) but has an invariant measure. We finally discuss the cross-diffusion case.

Keywords: Coupled KPZ equation · Invariant measure · Renormalization · Trilinear condition

1 Multi-component Coupled KPZ Equation

We consider an $\mathbb{R}^d$-valued KPZ equation for $h(t,x) = (h^\alpha(t,x))_{\alpha=1}^d$ defined on a one-dimensional torus $\mathbb{T} = [0,1)$:

$$\partial_t h^\alpha = \tfrac{1}{2}\partial_x^2 h^\alpha + \tfrac{1}{2}\Gamma^\alpha_{\beta\gamma}\partial_x h^\beta \partial_x h^\gamma + \sigma^\alpha_\beta \xi^\beta. \qquad (\sigma,\Gamma)_{\mathrm{KPZ}}$$

Here, we use Einstein's convention and $\xi(t,x) = (\xi^\alpha(t,x))_{\alpha=1}^d$ (sometimes written as $\dot{W}(t,x)$) is an $\mathbb{R}^d$-valued space-time Gaussian white noise with covariance structure:

$$E[\xi^\alpha(t,x)\xi^\beta(s,y)] = \delta^{\alpha\beta}\delta(x-y)\delta(t-s).$$

The coupled KPZ equation is ill-posed, since the noise is irregular and doesn't match with the nonlinear term. Note that $h \in C^{\frac{1}{4}-,\frac{1}{2}-}([0,\infty)\times\mathbb{T}) \equiv \cap_{\delta>0} C^{\frac{1}{4}-\delta,\frac{1}{2}-\delta}([0,\infty)\times\mathbb{T})$ a.s. when $\Gamma = 0$. Therefore, we need to introduce approximations with smooth noises and renormalizations for the equation $(\sigma,\Gamma)_{\mathrm{KPZ}}$. Indeed, one can introduce two types of approximations: one is simple and commonly used, the other is suitable to study the invariant measures. When $d = 1$, the second type of approximation was introduced by Funaki and Quastel [7].

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 560–568, 2019.
https://doi.org/10.1007/978-3-030-15096-9_20

The coupling constants $\Gamma^\alpha_{\beta\gamma}$ of the nonlinear term satisfy, by the form of the equation, the bilinear condition:

$$\Gamma^\alpha_{\beta\gamma} = \Gamma^\alpha_{\gamma\beta} \qquad \text{for all } \alpha, \beta, \gamma,$$

and sometimes additionally the trilinear condition:

$$\Gamma^\alpha_{\beta\gamma} = \Gamma^\alpha_{\gamma\beta} = \Gamma^\gamma_{\beta\alpha} \qquad \text{for all } \alpha, \beta, \gamma, \tag{T}$$

cf. Ferrari, Sasamoto and Spohn [3], Kupiainen and Marcozz [10]. The noise strength matrix $\sigma = (\sigma^\alpha_\beta)$ is always invertible.

Since σ is invertible, $\hat{h} = \sigma^{-1}h$ transforms the equation $(\sigma, \Gamma)_{\mathrm{KPZ}}$ to another equation $(I, \hat{\Gamma} = \sigma \circ \Gamma)_{\mathrm{KPZ}}$, where $\sigma \circ \Gamma$ is defined by

$$(\sigma \circ \Gamma)^\alpha_{\beta\gamma} := (\sigma^{-1})^\alpha_{\alpha'} \Gamma^{\alpha'}_{\beta'\gamma'} \sigma^{\beta'}_\beta \sigma^{\gamma'}_\gamma.$$

In this way, the KPZ equation with $\sigma = I$ can be considered as a canonical form.

Note that the operation (coordinate change) $\Gamma \mapsto \sigma \circ \Gamma$ keeps the bilinearity, but not the trilinearity. We should say (σ, Γ) satisfies the trilinear condition, if and only if $\hat{\Gamma} := \sigma \circ \Gamma$ satisfies the condition (T). In the following, we assume $\sigma = I$. See [5] for related random interfaces.

2 Two Coupled KPZ Approximating Equations

Two approximations are discussed in [7] when $d = 1$, and extended to the coupled equations in [6]. Let $\eta \in C_0^\infty(\mathbb{R})$ be a usual convolution kernel such that $\eta(x) \geq 0, \eta(-x) = \eta(x)$ and $\int_{\mathbb{R}} \eta(x)dx = 1$. We set $\eta^\varepsilon(x) := \frac{1}{\varepsilon}\eta(\frac{x}{\varepsilon}), 0 < \varepsilon < 1$, and replace the noise by smooth one.

Approximating equation-1 (usual approximation): Let $h^\alpha = h^{\varepsilon,\alpha}$ be the solution of the equation

$$\partial_t h^\alpha = \tfrac{1}{2}\partial_x^2 h^\alpha + \tfrac{1}{2}\Gamma^\alpha_{\beta\gamma}(\partial_x h^\beta \partial_x h^\gamma - c^\varepsilon \delta^{\beta\gamma} - B^{\varepsilon,\beta\gamma}) + \xi^\alpha * \eta^\varepsilon, \tag{1}$$

where $c^\varepsilon = \frac{1}{\varepsilon}\|\eta\|^2_{L^2(\mathbb{R})} (= O(\frac{1}{\varepsilon}))$ and $B^{\varepsilon,\beta\gamma}$ $(= O(\log \frac{1}{\varepsilon})$ in general) is another renormalization factor.

Approximating equation-2 (suitable to study invariant measures): Let $\tilde{h}^\alpha = \tilde{h}^{\varepsilon,\alpha}$ be the solution of the equation

$$\partial_t \tilde{h}^\alpha = \tfrac{1}{2}\partial_x^2 \tilde{h}^\alpha + \tfrac{1}{2}\Gamma^\alpha_{\beta\gamma}(\partial_x \tilde{h}^\beta \partial_x \tilde{h}^\gamma - c^\varepsilon \delta^{\beta\gamma} - \tilde{B}^{\varepsilon,\beta\gamma}) * \eta_2^\varepsilon + \xi^\alpha * \eta^\varepsilon, \tag{2}$$

with a renormalization factor $\tilde{B}^{\varepsilon,\beta\gamma}$, where $\eta_2^\varepsilon = \eta^\varepsilon * \eta^\varepsilon$.

The idea behind (2) is the fluctuation-dissipation relation. The renormalization factor $c^\varepsilon (\equiv c_\epsilon^{\bigvee}) = O(\frac{1}{\varepsilon})$ comes from the second order terms in the Wiener chaos expansion, while the renormalization factors $B^{\varepsilon,\beta\gamma}$ and $\tilde{B}^{\varepsilon,\beta\gamma} = O(\log \frac{1}{\varepsilon})$ are from the fourth order terms involving $C^\varepsilon (= c_\epsilon^{\bigvee\bigvee})$ and $D^\varepsilon (= c_\epsilon^{\bigvee\!\!\bigvee})$; see Sect. 4.2.

3 Quick Overview of Results on Coupled KPZ Equation

In this section, we summarize the results of Funaki and Hoshino [6].

- The convergence of h^ε and $\tilde{h}^\varepsilon$ as $\varepsilon \downarrow 0$ and the local well-posedness of the coupled KPZ equation $(\sigma, \Gamma)_{\mathrm{KPZ}}$ were shown by applying the paracontrolled calculus due to Gubinelli, Imkeller and Perkowski [8]; see Theorem 1 in Sect. 3.1 below. Note that the Cole-Hopf transform does not work for the coupled KPZ equation in general. In scalar-valued case [7], we used it and showed the Boltzmann-Gibbs principle; see Sect. 3.2.
- The second approximation (2) fits to identify the invariant measure under the trilinear condition (T); see Theorem 2-(2) in Sect. 3.2.
- The global solvability for a.s.-initial data under an invariant measure is shown under the condition (T); see Theorem 3 in Sect. 3.3.
- The global well-posedness (existence, uniqueness) for all initial values are established under the condition (T) by combining the strong Feller property shown by Hairer and Mattingly [9] and the global solvability for a.s.-initial values; see Sect. 3.3. The ergodicity and the uniqueness of invariant measure also follow.
- A priori estimate for the first approximation (1), which is available under the condition (T), plays a role.

3.1 Convergence of h^ε and $\tilde{h}^\varepsilon$ and Local Well-Posedness

We state more precisely the results on the convergence of h^ε and $\tilde{h}^\varepsilon$ and local well-posedness of the coupled KPZ equation $(I, \Gamma)_{\mathrm{KPZ}}$. We do not need the condition (T). Recall that we take $\sigma = I$. Let $\mathcal{C}^\kappa = (\mathcal{B}^\kappa_{\infty,\infty}(\mathbb{T}))^d$, $\kappa \in \mathbb{R}$ denote $\mathbb{R}^d$-valued Besov space on $\mathbb{T}$; cf. [6,8].

Theorem 1. (1) *Assume $h_0 \in \cup_{\delta>0}\mathcal{C}^\delta$, then a unique solution h^ε of the Eq. (1) exists up to some $T^\varepsilon \in (0,\infty]$ and $\bar{T} = \liminf_{\varepsilon\downarrow 0} T^\varepsilon > 0$ holds. With a proper choice of $B^{\varepsilon,\beta\gamma}$, h^ε converges in probability to some h in $C([0,T],\mathcal{C}^{\frac{1}{2}-\delta})$ for every $\delta > 0$ and $0 < T \le \bar{T}$.*
(2) *Similar result holds for the solution $\tilde{h}^\varepsilon$ of the Eq. (2) with some limit $\tilde{h}$. Under proper choices of $B^{\varepsilon,\beta\gamma}$ and $\tilde{B}^{\varepsilon,\beta\gamma}$, we can actually make $h = \tilde{h}$.*

3.2 Results Under the Trilinear Condition (T)

Under the condition (T), we have cancellation in logarithmic renormalization factors and invariance of the Wiener measure under the time evolution, and one can compute explicitly the difference of two limits.

Theorem 2. *Assume the trilinear condition* (T).
(1) *Then, $B^{\varepsilon,\beta\gamma}, \tilde{B}^{\varepsilon,\beta\gamma} = O(1)$ so that the solutions of the Eq. (1) with $B = 0$ and the Eq. (2) with $\tilde{B} = 0$ converge. In the limit, we have*

$$\tilde{h}^\alpha(t,x) = h^\alpha(t,x) + c^\alpha t, \quad 1 \le \alpha \le d,$$

where

$$c^{\alpha} = \frac{1}{24} \sum_{\gamma,\gamma'} \Gamma^{\alpha}_{\alpha'\alpha''} \Gamma^{\alpha'}_{\gamma\gamma'} \Gamma^{\alpha''}_{\gamma\gamma'}.$$

(2) *Moreover, the distribution of* $(\partial_x B)_{x\in\mathbb{T}}$*, where* B *is a periodic* d*-dimensional Brownian motion, is invariant under the tilt process* $u = \partial_x h$*. Or, one can say that the periodic Wiener measure on the quotient space* $\mathcal{C}^{\frac{1}{2}-\delta}/\sim$*, where* $h \sim h + c$ *for all constants* $c \in \mathbb{R}$*, is invariant for the height process* h *considered under such identification.*

When $d = 1$ (i.e., for the scalar-valued equation), the condition (T) is automatically satisfied. [7] showed that stationary solutions of two approximating equations without logarithmic renormalization factors satisfy

$$\lim_{\varepsilon\downarrow 0} \tilde{h}^{\varepsilon} = \lim_{\varepsilon\downarrow 0} h^{\varepsilon} + \frac{t}{24} \left(= h_{CH} + \frac{t}{24} \right).$$

Note that $h_{CH} := \lim_{\varepsilon\downarrow 0} h^{\varepsilon}$ is called the Cole-Hopf solution of (scalar-valued) KPZ equation. Theorem 2 extends this result for $d \geq 1$ and in non-stationary setting.

3.3 Global Existence for a.s.-Initial Values Under Stationary Measure

We assume the trilinear condition (T) and that the initial value $h(0)$ is given by $h(0,0) = 0$ and $u(0) := \partial_x h(0) \overset{\text{law}}{=} (\partial_x B)_{x\in\mathbb{T}}$, where B is a periodic d-dimensional Brownian motion. Then, by a similar method to Da Prato and Debussche [1] for two-dimensional stochastic Navier-Stokes equation, one can show the following theorem for $u = \partial_x h$:

Theorem 3. *Assume the condition* (T)*. Then, for every* $T > 0, p \geq 1, \kappa > 0$*, we have*

$$E\left[\sup_{t\in[0,T]} \|u(t; u_0)\|^p_{-\frac{1}{2}-\kappa} \right] < \infty.$$

In particular, $T_{survival}(u(0)) = \infty$ *(i.e., no explosion occurs for the solution) for a.a.-*$u(0)$*.*

In the scalar-valued case, the global existence of solutions for all given $u(0)$ is immediate, since the limit is the Cole-Hopf solution. Hairer and Mattingly [9] proved the global well-posedness for the coupled equation by showing the strong Feller property on the space $\mathcal{C}^{\alpha-1}, \alpha \in (0, \frac{1}{2})$.

4 Role of the Trilinear Condition (T)

4.1 Ertaş and Kardar's example

We give an example that the cancellation of logarithmic renormalization factors and the existence of an invariant measure hold without the condition (T).

Ertaş and Kardar [2] considered the following coupled KPZ equation with $d = 2$:

$$\begin{aligned} \partial_t h^1 &= \tfrac{1}{2}\partial_x^2 h^1 + \tfrac{1}{2}\{\lambda_1(\partial_x h^1)^2 + \lambda_2(\partial_x h^2)^2\} + \xi^1, \\ \partial_t h^2 &= \tfrac{1}{2}\partial_x^2 h^2 + \lambda_1 \partial_x h^1 \partial_x h^2 + \xi^2, \end{aligned} \tag{EK}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}$. The coupling constant Γ satisfies the trilinear condition (T) only when $\lambda_1 = \lambda_2$.

Under the transform $\hat{h} = sh$ with $s = \begin{pmatrix} \lambda_1 & (\lambda_1\lambda_2)^{1/2} \\ \lambda_1 & -(\lambda_1\lambda_2)^{1/2} \end{pmatrix}$, the equation (EK) is transformed into

$$\partial_t \hat{h}^\alpha = \tfrac{1}{2}\partial_x^2 \hat{h}^\alpha + \tfrac{1}{2}(\partial_x \hat{h}^\alpha)^2 + s^\alpha_\beta \xi^\beta. \tag{EK$_T$}$$

Namely, $\hat{\Gamma} = s \circ \Gamma$ in (EK$_T$) is given by $\hat{\Gamma}^\alpha_{\alpha\alpha} = 1, = 0$ otherwise, so that $\hat{\Gamma}$ satisfies the condition (T). But, this has no special meaning, since (EK) is the canonical form (with $\sigma = I$) and not the equation (EK$_T$).

The equation (EK) does not satisfy the trilinear condition (T) if $\lambda_1 \neq \lambda_2$. However, since the nonlinear term is decoupled in (EK$_T$), the Cole-Hopf transform $Z^\alpha := \exp \hat{h}^\alpha$ works for each component $\hat{h}^\alpha$ so that the global well-posedness holds even for the coupled equation (EK$_T$) and therefore (EK). This also implies that the logarithmic renormalization factors are unnecessary for this equation. Moreover, the equation (EK$_T$) has an invariant measure whose marginals are Wiener measures, but the joint distribution of such invariant measure is unclear (presumably non-Gaussian). Indeed, one can easily check the tightness on the space $\mathcal{C}_0^{\delta-1}/\sim$ of the Cesàro mean $\mu_T = \frac{1}{T}\int_0^T \mu(t)dt$ of the distributions $\mu(t)$ of $\partial_x \hat{h}(t)$ having an initial distribution $\otimes_\alpha \mu_\alpha$. In fact, this is seen from the tightness of two marginals of μ_T, so that the limit of μ_T as $T \to \infty$ exists and is an invariant measure of the equation (EK$_T$). Thus, the equation (EK) also has an invariant measure.

4.2 Cancellation of Logarithmic Renormalization Factors

We first give concrete formulas of renormalization factors $B^{\varepsilon,\beta\gamma}, \tilde{B}^{\varepsilon,\beta\gamma}$ in the Eqs. (1) and (2):

$$B^{\varepsilon,\beta\gamma} = F^{\beta\gamma} C^\varepsilon + 2G^{\beta\gamma} D^\varepsilon, \ \tilde{B}^{\varepsilon,\beta\gamma} = F^{\beta\gamma} \tilde{C}^\varepsilon + 2G^{\beta\gamma} \tilde{D}^\varepsilon,$$

where

$$\begin{aligned} F^{\beta\gamma} = \Gamma^\beta_{\gamma_1\gamma_2}\Gamma^\gamma_{\gamma_1\gamma_2}, \ G^{\beta\gamma} = \Gamma^\beta_{\gamma_1\gamma_2}\Gamma^{\gamma_1}_{\gamma\gamma_2}, \\ C^\varepsilon + 2D^\varepsilon = -\tfrac{1}{12} + O(\varepsilon), \quad \tilde{C}^\varepsilon + 2\tilde{D}^\varepsilon = 0. \end{aligned}$$

Indeed, $C^\varepsilon, D^\varepsilon$ appearing above and c^ε appearing in the Eqs. (1) and (2) are sometimes denoted as

$$c^\varepsilon = c_\epsilon^{\blacktriangledown}, C^\varepsilon = c_\epsilon^{\blacktriangledown\blacktriangledown}, D^\varepsilon = c_\epsilon^{\blacktriangledown\!\!\blacktriangledown}.$$

One can show that the trilinear condition (T) is equivalent to "$F = G$", which is further equivalent to $B, \tilde{B} = O(1)$. But, for cancellation of logarithmic renormalization factors, what we really need is: "$\Gamma B, \Gamma\tilde{B} = O(1)$". This holds if $\Gamma F = \Gamma G$, which is weaker than the trilinear condition (T).

The condition "$\Gamma F = \Gamma G$" holds if and only if Γ satisfies the condition

$$\Gamma^\alpha_{\beta\gamma}\Gamma^\beta_{\gamma_1\gamma_2}\Gamma^\gamma_{\gamma_1\gamma_2} = \Gamma^\alpha_{\beta\gamma}\Gamma^\beta_{\gamma_1\gamma_2}\Gamma^{\gamma_1}_{\gamma\gamma_2},$$

for all α. This holds under (T) and also for Ertaş and Kardar's example.

We can summarize the above arguments as follows:

$$\begin{aligned}(\mathrm{T}) &\iff \text{"}F = G\text{"}\\ &\implies \text{"}\Gamma F = \Gamma G\text{"}\\ &\iff \text{Cancellation of log-renormalization factors}\end{aligned}$$

4.3 Infinitesimal Invariance

In order to explain the role of the trilinear condition (T) from a viewpoint of the invariant measure, let us discuss the infinitesimal invariance; cf. [4]. The arguments are rather heuristic, for instance, derivatives of $h^\alpha(x)$ appear in the computations though they do not really exist. To be precise, we need to discuss at the level of the discrete approximation (cf. [7]) or at the level of the approximating equation-2 (cf. [4]).

Let $\mathcal{L} = \mathcal{L}_0 + \mathcal{A}$ be the generator of the coupled KPZ equation $(I, \Gamma)_{\mathrm{KPZ}}$ with $\sigma = I$. Here, $\mathcal{L}_0$ is the generator of the Ornstein-Uhlenbeck part, while $\mathcal{A}$ is that of the nonlinear part (as we mentioned above, we ignore the renormalization factors):

$$\mathcal{L}_0\Phi = \frac{1}{2}\sum_\alpha \left\{ \int_{\mathbb{T}} D^2_{h^\alpha(x)}\Phi\, dx + \int_{\mathbb{T}} \ddot{h}^\alpha(x) D_{h^\alpha(x)}\Phi\, dx \right\}, \tag{3}$$

$$\mathcal{A}\Phi = \frac{1}{2}\sum_{\alpha,\beta,\gamma} \Gamma^\alpha_{\beta\gamma} \int_{\mathbb{T}} \dot{h}^\beta(x)\dot{h}^\gamma(x) D_{h^\alpha(x)}\Phi\, dx, \tag{4}$$

for $\Phi = \Phi(h)$, where $D_{h^\alpha(x)}, D^2_{h^\alpha(x)}$ are functional derivatives and $\dot{h}^\beta(x) := \partial_x h^\beta(x), \ddot{h}^\alpha(x) := \partial_x^2 h^\alpha(x)$.

We say that the infinitesimal invariance $(ST)_{\mathcal{L}}$ holds for ν if

$$\int \mathcal{L}\Phi d\nu = 0,$$

holds for all Φ (though this statement is already heuristic).

If the invariant measure ν is Gaussian, $(ST)_{\mathcal{L}_0}$ is the condition for the second order Wiener chaos of Φ, while $(ST)_{\mathcal{A}}$ is that for the third order Wiener chaos of Φ. Therefore, the condition $(ST)_{\mathcal{L}}$ is separated into two conditions:

$$(ST)_{\mathcal{L}} \Longleftrightarrow (ST)_{\mathcal{L}_0} + (ST)_{\mathcal{A}}. \tag{5}$$

Since $\mathcal{L}_0$ is an Ornstein-Uhlenbeck operator, from the condition $(ST)_{\mathcal{L}_0}$, ν must be the Wiener measure.

4.4 (T) is Necessary and Sufficient for Wiener Measure ν to Satisfy $(ST)_{\mathcal{A}}$

We have the integration-by-parts formula for the Wiener measure ν (actually we need to discuss at approximating level as we mentioned above):

$$\int \mathcal{A}\Phi d\nu = -\frac{1}{2}\Gamma^{\alpha}_{\beta\gamma}c^{\beta\gamma}_{\alpha},$$

where

$$c^{\beta\gamma}_{\alpha} \equiv c^{\beta\gamma}_{\alpha}(\Phi) := E^{\nu}\left[\Phi \int_{\mathbb{T}} \dot{h}^{\beta}(x)\dot{h}^{\gamma}(x)\ddot{h}^{\alpha}(x)dx\right].$$

One easily see that the constants $c^{\beta\gamma}_{\alpha}$ satisfy the following two conditions:
(1) (bilinearity) $c^{\beta\gamma}_{\alpha} = c^{\gamma\beta}_{\alpha}$,
(2) (integration by parts on $\mathbb{T}$) $c^{\beta\gamma}_{\alpha} + c^{\gamma\alpha}_{\beta} + c^{\alpha\beta}_{\gamma} = 0$.
In particular, we have $c^{\alpha\alpha}_{\alpha} = 0$ for all α. When $d = 1$, this implies $(ST)_{\mathcal{A}}$: $\int \mathcal{A}\Phi d\nu = 0$ for all Φ.

If Γ satisfies the trilinear condition (T), by the above condition (2) for $c^{\beta\gamma}_{\alpha}$, we have

$$\Gamma^{\alpha}_{\beta\gamma}c^{\beta\gamma}_{\alpha} = \frac{1}{3}\Gamma^{\alpha}_{\beta\gamma}(c^{\beta\gamma}_{\alpha} + c^{\gamma\alpha}_{\beta} + c^{\alpha\beta}_{\gamma}) = 0.$$

Therefore, the condition (T) implies $(ST)_{\mathcal{A}}$. This was shown in [4] (at approximating level).

Conversely, $(ST)_{\mathcal{A}}$ implies (T). In fact, by the condition (2) for $c^{\beta\gamma}_{\alpha}$, one can rewrite as

$$\begin{aligned} -2\int \mathcal{A}\Phi d\nu = &\sum_{\alpha\neq\beta}(\Gamma^{\alpha}_{\beta\beta} - \Gamma^{\beta}_{\alpha\beta})c^{\beta\beta}_{\alpha} \\ &+ 2\sum_{\alpha>\beta>\gamma}(\Gamma^{\alpha}_{\beta\gamma} - \Gamma^{\gamma}_{\alpha\beta})c^{\beta\gamma}_{\alpha} + 2\sum_{\beta>\alpha>\gamma}(\Gamma^{\alpha}_{\beta\gamma} - \Gamma^{\gamma}_{\alpha\beta})c^{\beta\gamma}_{\alpha}. \end{aligned}$$

If the left hand side is 0, noting that $c^{\beta\beta}_{\alpha}, c^{\beta\gamma}_{\alpha}(\alpha > \beta > \gamma, \beta > \alpha > \gamma)$ move freely, we obtain the trilinear condition (T).

Ertaş and Kardar's example does not satisfy (T), but has an invariant measure. This should be "non-separating class" (i.e., (5) does not hold) and the invariant measure is presumably non-Gaussian (but has Gaussian marginals).

5 Cross-diffusion Case

Let us consider the following coupled KPZ equation on $\mathbb{T}$ as a generalization of the equation $(\sigma, \Gamma)_{\mathrm{KPZ}}$:

$$\partial_t h^\alpha = \tfrac{1}{2} d^\alpha_\beta \partial_x^2 h^\beta + \tfrac{1}{2} \Gamma^\alpha_{\beta\gamma} \partial_x h^\beta \partial_x h^\gamma + \sigma^\alpha_\beta \xi^\beta, \qquad (D, \sigma, \Gamma)_{\mathrm{KPZ}}$$

with a cross-diffusion matrix $D = (d^\alpha_\beta)$. We assume $\sigma = I$ and D is symmetric and positive definite. Note that the cross-diffusion system

$$\partial_t h^\alpha = \tfrac{1}{2} d^\alpha_\beta \partial_x^2 h^\beta,$$

is well-posed under these conditions on D, since D is diagonalizable: $PDP^{-1} = \mathrm{diag}(d_1, \dots, d_d)$ with $d_\alpha > 0$ and $\hat{h} := Ph$ is decoupled.

5.1 Invariant Measure of Ornstein-Uhlenbeck Part

Dropping the nonlinear term and taking $\sigma = I$, we have a linear stochastic partial differential equation:

$$\partial_t h^\alpha = \tfrac{1}{2} d^\alpha_\beta \partial_x^2 h^\beta + \xi^\alpha, \tag{6}$$

and its generator is given as a modification of (3)

$$\mathcal{L}_0 \Phi = \frac{1}{2} \sum_\alpha \left\{ \int_{\mathbb{T}} D^2_{h^\alpha(x)} \Phi \, dx + \sum_\beta d^\alpha_\beta \int_{\mathbb{T}} \ddot{h}^\beta(x) D_{h^\alpha(x)} \Phi \, dx \right\}.$$

Then, for $V = V(h)$,

$$\begin{aligned} e^{-V} \mathcal{L}_0^* e^V =& \frac{1}{2} \sum_\alpha \int_{\mathbb{T}} \left\{ D^2_{h^\alpha(x)} V + (D_{h^\alpha(x)} V)^2 \right\} dx \\ & - \frac{1}{2} \int_{\mathbb{T}} \left\{ \sum_\alpha d^\alpha_\alpha \delta''_x(x) + \sum_{\alpha,\beta} d^\alpha_\beta \ddot{h}^\beta(x) D_{h^\alpha(x)} V \right\} dx. \end{aligned}$$

If we choose

$$V(h) = -\frac{1}{2} v_{\beta\beta'} \int_{\mathbb{T}} \dot{h}^\beta(y) \dot{h}^{\beta'}(y) dy$$

with a symmetric matrix $(v_{\beta\beta'}) = (d^\beta_{\beta'})$, one can show that $e^{-V} \mathcal{L}_0^* e^V = 0$. This determines the invariant measure $\nu \equiv \nu_D = e^V dh$ of the Ornstein-Uhlenbeck part. Note that ν_D is considered as the distribution of $\sqrt{D}^{-1} B$ with a periodic d-dimensional Brownian motion B in the sense of Theorem 2-(2). This fact is easily seen also by diagonalizing the Eq. (6) by the orthogonal matrix P.

5.2 Trilinear Condition for Cross-diffusion Coupled KPZ Equation

The generator $\mathcal{A}$ of the nonlinear part is the same as (4), so that $(ST)_{\mathcal{A}}$ holds for $\nu = \nu_D (= e^V dh)$ if and only if

$$0 = e^{-V}\mathcal{A}^* e^V = -\frac{1}{2}\sum_{\alpha,\beta,\gamma} \Gamma^\alpha_{\beta\gamma} \int_{\mathbb{T}} \dot{h}^\beta(x)\dot{h}^\gamma(x) D_{h^\alpha(x)} V \, dx$$
$$= -\frac{1}{2}\sum_{\alpha,\beta,\gamma,\beta'} \Gamma^\alpha_{\beta\gamma} d^{\beta'}_\alpha \int_{\mathbb{T}} \dot{h}^\beta(x)\dot{h}^\gamma(x)\ddot{h}^{\beta'}(x)\, dx,$$

since $D_{h^\alpha(x)} V = v_{\alpha\beta'}\ddot{h}^{\beta'}(x) = \sum_{\beta'} d^{\beta'}_\alpha \ddot{h}^{\beta'}(x)$. Thus, the condition for ν_D to satisfy $(ST)_{\mathcal{A}}$ is that

$$\tilde{\Gamma}^\alpha_{\beta\gamma} := d^\alpha_{\alpha'} \Gamma^{\alpha'}_{\beta\gamma}$$

satisfies the trilinear condition (T).

References

1. Da Prato, G., Debussche, A.: Two-dimensional Navier-Stokes equations driven by a space-time white noise. J. Funct. Anal. **196**, 180–210 (2002)
2. Ertaş, D., Kardar, M.: Dynamic roughening of directed lines. Phys. Rev. Lett. **69**, 929–932 (1992)
3. Ferrari, P.L., Sasamoto, T., Spohn, H.: Coupled Kardar-Parisi-Zhang equations in one dimension. J. Stat. Phys. **153**, 377–399 (2013)
4. Funaki, T.: Infinitesimal invariance for the coupled KPZ equations. In: Memoriam Marc Yor – Séminaire de Probabilités XLVII. Lecture Notes in Mathematics, vol. 2137, pp. 37–47. Springer (2015)
5. Funaki, T.: Lectures on Random Interfaces. SpringerBriefs in Probability and Mathematical Statistics. Springer, Singapore (2016)
6. Funaki, T., Hoshino, M.: A coupled KPZ equation, its two types of approximations and existence of global solutions. J. Funct. Anal. **273**, 1165–1204 (2017)
7. Funaki, T., Quastel, J.: KPZ equation, its renormalization and invariant measures. Stoch. PDE Anal. Comp. **3**, 159–220 (2015)
8. Gubinelli, M., Imkeller, P., Perkowski, N.: Paracontrolled distributions and singular PDEs. Forum Math. Pi. **3**, 1–75 (2015)
9. Hairer, M., Mattingly, J.: The strong Feller property for singular stochastic PDEs. Ann. Inst. Henri Poincaré Probab. Stat. **54**, 1314–1340 (2018)
10. Kupiainen, A., Marcozzi, M.: Renormalization of generalized KPZ equation. J. Stat. Phys. **166**, 876–902 (2017)
11. Spohn, H.: Nonlinear fluctuating hydrodynamics for anharmonic chains. J. Stat. Phys. **154**, 1191–1227 (2014)
12. Spohn, H., Stoltz, G.: Nonlinear fluctuating hydrodynamics in one dimension: the case of two conserved fields. J. Stat. Phys. **160**, 861–884 (2015)

Reversible Viscosity and Navier–Stokes Fluids

Giovanni Gallavotti[1,2(✉)]

[1] Università di Roma "La Sapienza", Rome, Italy
giovanni.gallavotti@roma1.infn.it
[2] INFN, Rome, Italy

Abstract. Exploring the possibility of describing a fluid flow via a time-reversible equation and its relevance for the fluctuations statistics in stationary turbulent (or laminar) incompressible Navier-Stokes flows.

1 Introduction

Studies on non equilibrium statistical mechanics progressed after the introduction of thermostats, [1]. Finite thermostats have not only permitted a new series of simulations of many particle systems, but have been essential to clarify that *irreversibility* and dissipation *should not* be identified.

Adopting the terminology of [2] it is convenient to distinguish the finite system of interest, *i.e.* particles forming the *test system* in a container $\mathcal{C}_0$, from the thermostats. The thermostats $\mathbf{T}_1, \mathbf{T}_2, \ldots$ are also particle systems, forming the *interaction systems*, acting on the test systems: they are in infinite containers and, *asymptotically at infinity*, are always supposed in equilibrium states with given densities $\rho_1, \rho_2, \ldots$ and temperatures $T_1, T_2, \ldots$.

The thermostats particles in each thermostat may interact with each other and with the particles of the test system *but not directly* with the particles of the other thermostats. The test system and the interaction systems, together, form a Hamiltonian system (classical or quantum) that can be symbolically illustrated as in Fig. 1: Finite thermostats have been introduced recently and fulfill the main function of replacing, [1], the above test systems and "perfect thermostats", consisting of infinite systems of particles in a state in a well defined equilibrium state at infinity, with finite systems suitable for simulations.

The perfect thermostats, being infinite, are not suited in simulations, while the finite ones have the drawback that their equations of motion contain "unphysical forces".

The basic idea is that, asymptotically *e.g.* for large number of particles ("thermodynamic limit"), most statistical properties of the "test" system do not depend on the particular thermostat model but only on its equilibrium parameters defined at infinity.

Several finite thermostats employed in simulations are governed by reversible equations of motion: denoting $u \to S_t u, t \in R$ the time evolution of a point u in phase space $\mathcal{F}$, this means that the map $u \to Iu$ in which all velocities in u are

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 569–580, 2019.
https://doi.org/10.1007/978-3-030-15096-9_21

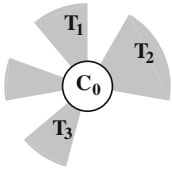


Fig. 1. The "*test*" system are particles enclosed in $\mathcal{C}_0$ while the external $\mathcal{C}_j$ systems are thermostats at respective temperatures T_j (marked in the figure) or, following the terminology of Feynman–Vernon, [2], *"interaction"* systems.

reversed is such that $S_t I = I S_{-t}$, so that if $u(t), t \in R$ is a possible solution of the equations of motion also $I\,u(-t), t \in R$ is a possible one.

If u describes the state of a system in which dissipation occurs, *i.e.* in which external forces perform work on the test subsystem, it might be thought that, unless the interaction systems are infinite, the motion is not reversible: this has been clearly shown to be not true by the many simulations performed since the early '80s, reviewed in [1]. And the simulations have added evidence that the same physical phenomenon occurring in the test system is largely independent of several (appropriate) realizations of thermostat models (reversible or not).

A remarkable instance is an example of a system of particles interacting with a single thermostat at temperature $\beta^{-1} = T$ which has a stationary state described by a probability distribution $\mu(du)$ which is different from the canonical distribution (say) but which is nevertheless equivalent to it in the sense, [3], of the theory of ensembles, *i.e.* in the thermodynamic limit, see [1].

In the different context of turbulence theory a similar example can be found in the simulation in [4]: where viscosity is set = 0 but "unphysical forces" are introduced to constrain the energy value on each "energy shell" to fulfill the OK "$\frac{5}{3}$ law". The stationary distribution of the velocity field for many observables, *e.g.* the large scale velocity components, remains the same as in the viscous unconstrained system and in the reversible new one, at very large Reynolds number.

Then one is led to think that the root of the equivalence between very different equations of motion for the same physical system lies in the fundamental microscopic reversibility of the equations of motion, [5,6], and to a precise formulation of the "conjecture" that "*in microscopically reversible (chaotic) systems time reversal symmetry cannot be spontaneously broken, but only phenomenologically so*" and a program to test it, was proposed, [7]. The program has been followed so far in a few works, [8,9], with results apparently not always satisfactory [10].

Here, after a general discussion of the conjecture and its precise formulation, several tests will be proposed, on the statistical properties of the stationary states of the $2D$ incompressible Navier-Stokes equation, and performed with results described in some detail.

2 Irreversible and Reversible ODE's

More generally an ODE $\dot{x} = h(x)$ on the "phase space" R^N has a *time reversal symmetry* I if the solution operator $x \to S_t x, x \in R^N$, and the map I are such that $I^2 = 1$, $S_t I = I S_{-t}$.

Non trivial examples are provided, as mentioned, by many Hamiltonian equations, but there are also interesting examples not immediately related to Hamiltonian systems, as the equations of the form $\dot{x}_j = f_j(x)$, $\nu > 0$, $j = 1, \ldots, N$ with $f_j(x) = f_j(-x)$, like the Lorenz96 model at $\nu = 0$:

$$\dot{x}_j = x_{j-1}(x_{j+1} - x_{j-2}) + F - \nu x_j, \tag{2.1}$$

with $F = const$ and periodic b.c. $x_0 = x_N$.

Another example is provided by the GOY *shell model*, [11,12], given by:

$$\begin{aligned} \dot{u}_n = \; & -\nu k_n^2 u_n + g\delta_{n,4} \\ & + i k_n \big(-\frac{1}{4}\overline{u}_{n-1}\overline{u}_{n+1} + \overline{u}_{n+1}\overline{u}_{n+2} - \frac{1}{8}\overline{u}_{n-1}\overline{u}_{n-2} \big) \end{aligned}$$

where $k_n = 2^n$, $u_n = u_{n,1} + i u_{n,2}$, with $u_n = 0$ for $n = -1, 0$ or $n > N$, if $\nu = 0$.

A reversible equation often evolves initial data x into functions $x(t)$ which are unbounded as $t \to \infty$. The case of Hamiltonian systems with bounded energy surfaces are an important exception. Therefore, particularly in problems dealing with stationary states in chaotic systems, the equations contain additional terms which arise by taking into account that the systems under study are also subject to stabilizing mechanisms forcing motions to be confined to some sphere in phase space.

A typical additional term is $-\nu x_j$ or $-\nu(Lx)_j$ with $\nu > 0$ and L a positive defined matrix: such extra terms are often introduced empirically. This is the case in the above two examples. And they can be thought as empirical realizations of the action of thermostats acting on the systems.

At this point it is necessary to distinguish the models in which

(1) the equations $\dot{x} = h(x) - \nu L x$ arise, possibly in some limit case, from a system of particles, as the one of the Feynman-Vernon system in Sect. 1, Fig. 1, or

(2) the equations are not directly related to a fundamental microscopic description of the system.

The above Lorenz96 and GOY models are examples of the second case, while the Navier-Stokes equations, since the beginning, were considered macroscopic manifestations of particles interacting via Newtonian forces, [13, Eq. (128)].

The success of the simulations using artificial thermostat forces with finite thermostats and the independence of the results from the particular choice of the thermostats used to contain energy growth in nonequilibrium, [1], induces to think that there might be alternative ways to describe the same systems via equations that maintain the time reversal symmetry shown by the non thermostatted equations. A first proposal that seems natural is the following.

Consider an equation

$$\dot{x} = h(x) - \nu Lx, \text{ with } h(x) = h(-x) \tag{2.2}$$

time reversible if $\nu = 0$, for the time reversal $Ix = -x$; suppose that $|x \cdot h(x)| \le G|x|$. Then the motions will be asymptotically confined, if $\nu > 0$, to the ellipsoid $(x \cdot Lx) \le \frac{G}{\nu}$ and the system will be able to reach a stationary state, *i.e.* an invariant probability distribution $\mu^C_{\frac{1}{\nu}}$ of the phase space points. Frequently, if ν is small enough, the motions will be chaotic and there will be a unique stationary distribution, the "SRB distribution", [14].

The family of stationary distributions forms what will be called the "*viscosity ensemble*" $\mathcal{F}^C$ whose elements are parameterized by ν (and possibly by an index distinguishing the extremal distributions which can be reached as stationary states, for the same ν, from different initial data); then consider the new equation

$$\dot{x} = h(x) - \alpha(x)Lx, \qquad \alpha(x) = \frac{(Lx \cdot h(x))}{(Lx \cdot Lx)} \tag{2.3}$$

where α has been determined so that the observable $\mathcal{D}(x) \overset{def}{=} (x \cdot Lx)$ is an exact constant of motion. For each choice of the parameter $\mathcal{E}$ the evolution will determine a family $\mu^M_{\mathcal{E}}$ of stationary probability distributions parameterized by the value $\mathcal{E}$ that $\mathcal{D}$ takes on the initial x generating the distribution. The collection $\mathcal{F}^M$ of such distributions will be called "*reversible viscosity ensemble*" because the distributions are stationary states for Eq. (2.3) which is reversible (for $Ix = -x$).

Also in this case if $\mathcal{E}$ is large the evolution Eq. (2.3) is likely to be chaotic and for each such $\mathcal{E}$ the distribution $\mu^M_{\mathcal{E}}$ is unique: *if not* extra parameter needs to be introduced the identify each of the extremal ones.

Suppose for simplicity that $\frac{1}{\nu}, \mathcal{E}$ are large enough and the stationary states $\mu^C_{\frac{1}{\nu}}, \mu^M_{\mathcal{E}}$ are unique. Then say that $\mu^C_{\frac{1}{\nu}}$ and $\mu^M_{\mathcal{E}}$ are *correspondent* if

$$\mu^M_{\mathcal{E}}(\alpha) = \frac{1}{\nu}, \qquad \text{or if} \qquad \mu^C_{\frac{1}{\nu}}(\mathcal{D}) = \mathcal{E} \tag{2.4}$$

Then the following proposal appears in [5,6] about the properties of the fluctuations of "K-local observables", *i.e.* of observables $F(x)$ depending only on the coordinates x_i with $i < K$

If $\frac{1}{\nu}$ and $\mathcal{E}$ *are large enough so that the motions generated by the equations Eqs. (2.2) and (2.3) are chaotic, e.g. satisfy the "Chaotic hypothesis", [15,16], then corresponding distributions* $\mu^C_{\frac{1}{\nu}}, \mu^M_{\mathcal{E}}$ *give the same distribution to the fluctuations of a given K-local observable* F *in the sense that*

$$\mu^M_{\mathcal{E}}(F) = \mu^C_{\frac{1}{\nu}}(F)(1 + o(F, \nu)) \tag{2.5}$$

with $o(F, \nu) \xrightarrow[\frac{1}{\nu} \to \infty]{} 0$.

There have been a few attempts to check this idea, [8,9] and more recently in [17].

3 Reversible Viscosity

The ideas of the preceding section will next be studied in the case of the Navier-Stokes equation. This is particularly interesting because the equation can be formally derived as an equation describing the macroscopic evolution of microscopic Newtonian particles (*i.e.* point masses interacting via a short range force), [13]. Hence the equation belongs to the rather special case (1) in Sect. 2.

The incompressible Navier-stokes equations with viscosity ν for a velocity field $\mathbf{v}(\mathbf{x},t)$ in a periodic container of size L and with a forcing $\mathbf{F}=F\mathbf{g}$ acting on large scale, *i.e.* with Fourier components $\mathbf{F}_{\mathbf{k}}\neq 0$ only for a few $|\mathbf{k}|$. *To fix the ideas* in 2 dimensions choose $F_{\mathbf{k}}\neq 0$ only for the single mode $\mathbf{k}=\pm\frac{2\pi}{L}(2,-1)$ with $||\mathbf{F}||_2=F$ (*i.e.* $\mathbf{g}_{\pm\mathbf{k}}=\frac{e^{\pm i\vartheta}}{\sqrt{2}}$ for some phase ϑ).

The equations can be written in dimensionless form: introduce rescaling parameters V,T for velocity and time, and write $\underset{\sim}{\mathbf{v}}(\mathbf{x},\tau)=V\underset{\sim}{\mathbf{u}}(\mathbf{x}/L,\tau/T)$. Define $V=(FL)^{\frac{1}{2}}$, $T=(\frac{L}{F})^{\frac{1}{2}}$ and fix $\frac{TV}{L}=1$ and $\frac{FT}{V}=1$; then the equation for $\mathbf{u}(\mathbf{x},t)$ can be written as, "I-NS":

$$\dot{\underset{\sim}{\mathbf{u}}}+(\mathbf{u}\cdot\partial)\mathbf{u}=\frac{1}{R}\Delta\underset{\sim}{\mathbf{u}}+\mathbf{g}-\boldsymbol{\partial}p,\quad \boldsymbol{\partial}\cdot\mathbf{u}=0 \tag{3.6}$$

where $R\equiv\frac{LV}{\nu}\equiv(\frac{FL^3}{\nu^2})^{\frac{1}{2}}$ and p is the pressure. In this way the inverse of the viscosity can be identified with the dimensionless parameter R, here called "Reynolds number" (often called, in this specific case, "Grashof number").

The units for L,F will be fixed so that $F=1$ and $L=2\pi$: hence the modes $\mathbf{k}$ will be pairs of integers $\mathbf{k}=(k_1,k_2)$. The reality conditions $\mathbf{u}_{\mathbf{k}}=\overline{\mathbf{u}}_{-\mathbf{k}}, F_{\mathbf{k}}=\overline{F}_{-\mathbf{k}}$ implies that only the components with

$$\mathbf{k}=(k_1,k_2)\in I^+\overset{def}{=}\{k_1>0 \text{ or } k_1=0,k_2\geq 0\} \tag{3.7}$$

are independent components (and it is assumed that $\mathbf{u_0}=0$).

We shall consider the case of 2 dimensional incompressible fluids to avoid the problem that the 3 dimensional equations have not yet been proved to admit a (classical or even just constructive) solution. In spite of this, below, the 3 dimensional case, and the vortex stretching present only in 3D, will also be commented and essentially everything that will be presented in the 2 dimensional case *turns out also relevant in* 3 *dimensions*.

Proceeding as in Sect. 2, define the family $\mathcal{F}^C$ of stationary probability distribution $\mu_R^{\mathbf{C}}(d\mathbf{u})$ on the fields $\mathbf{u}$ corresponding to the *stationary state* for the Eq. (3.6).

Consider, *alternatively*, the equation (reversible for the symmetry $I\mathbf{u}=-\mathbf{u}$), "R-NS":

$$\dot{\underset{\sim}{\mathbf{u}}}+(\mathbf{u}\cdot\boldsymbol{\partial})\underset{\sim}{\mathbf{u}}=\alpha(\underset{\sim}{\mathbf{u}})\Delta\underset{\sim}{\mathbf{u}}+\underset{\sim}{\mathbf{F}}-\underset{\sim}{\boldsymbol{\partial}}p\ ,\quad \underset{\sim}{\boldsymbol{\partial}}\cdot\underset{\sim}{\mathbf{u}}=0 \tag{3.8}$$

in which the viscosity $\nu = \frac{1}{R}$, *c.f.r.* Eq. (3.6), is replaced by the multiplier $\alpha(\mathbf{u})$ which is fixed so that

$$\mathcal{D}(\mathbf{u}) = \int |\underset{\sim}{\partial}\mathbf{u}(\mathbf{x})|^2 d\mathbf{x} = \text{exact const. of motion} \tag{3.9}$$

Therefore, if the space dimension is 2, the multiplier $\alpha(\mathbf{u})$ will be expressed, in terms of the Fourier transform $\mathbf{u_k}$ (defined via $\mathbf{u}(\mathbf{x}) = \sum_{\mathbf{k}} e^{2\pi i \mathbf{k}\cdot\mathbf{x}}\mathbf{u_k}$) as:

$$\alpha(\mathbf{u}) = \frac{\sum_{\mathbf{k}} \mathbf{k}^2 \overline{\mathbf{g}}_{\mathbf{k}} \cdot \mathbf{u_k}}{\sum_{\mathbf{k}} \mathbf{k}^4 |\mathbf{u_k}|^2} \equiv \frac{\sum_{\mathbf{k}\in I^+} \mathbf{k}^2 (g^r_{\mathbf{k}} \mathbf{u}^r_{\mathbf{k}} + g^i_{\mathbf{k}} \mathbf{u}^i_{\mathbf{k}})}{2\sum_{\mathbf{k}\in I^+} \mathbf{k}^4 |\mathbf{u_k}|^2} \tag{3.10}$$

and the stationary distribution for Eq. (3.10) with the value of $\mathcal{D}(\mathbf{u})$ fixed to $\mathcal{E}$, will be denoted $\mu^{\mathbf{M}}_{\mathcal{E}}(d\mathbf{u})$. The expression for α is slightly more involved (see [6, Eq. (1.11)]).

The collection of all stationary distributions $\mu^{\mathbf{C}}_R$ as R varies and of all stationary distributions $\mu^{\mathbf{M}}_{\mathcal{E}}$ as $\mathcal{E}$ varies will be denoted $\mathcal{F}^C$ and $\mathcal{F}^M$ and called *viscosity ensemble*, as in Sect. 2, and, respectively, *enstrophy ensemble*.

Call *K-local* an observable $f(\mathbf{u})$ which depends on the finite number of components $\mathbf{u_k}$ with $|\mathbf{k}| < K$, of the velocity field; then in the above cases the *conjecture* proposed in [5,6] becomes

In the limit of large Reynolds number the distribution μ^C_R attributes to any given K-local observable $f(\mathbf{u})$ the same average, in the sense of Eq. (2.5) with $R \equiv \frac{1}{\nu}$, as the distribution $\mu^{\mathbf{M}}_{\mathcal{E}}$ if

$$\mathcal{E} = \int \mu^{\mathbf{C}}_R(d\mathbf{u})\mathcal{D}(\mathbf{u}) \tag{3.11}$$

Remarks: (1) The size of R might (see however Sect. 4) depend on the observable f, *i.e.* on how many Fourier modes are needed to define f.

(2) Therefore locality in Fourier space is here analogous to locality in space in the equivalence between equilibrium ensembles.

(3) The notations $\mu^{\mathbf{C}}_R, \mu^{\mathbf{M}}_{\mathcal{E}}$ have been used to evoke the analogy of the equivalence between canonical and microcanonical ensembles in equilibrium statistical mechanics: the viscosity ensemble can be likened to the canonical ensemble, with the viscosity $\nu = \frac{1}{R}$ corresponding to β, and the enstrophy ensemble to the microcanonical one, with the enstrophy corresponding to the total energy.

(4) The equivalence has roots in the *chaotic hypothesis*, [16]: if the motion is sufficiently chaotic, as expected if R or $\mathcal{E}$ are large, [14,18], the multiplier $\alpha(\mathbf{u})$ fluctuates in time and the conjecture is based on a possible "*self-averaging*" of α implying homogenization of $\alpha(\mathbf{u})$ in Eq.(3.8) to a constant value, namely $\nu = \frac{1}{R}$.

(5) the latter remark, if μ^C_R is equivalent to $\mu^M_{\mathcal{E}}$ (*e.g.* if $\mu^C_R(\mathcal{D}) = \mathcal{E}$, see Eqs. (2.4) and (3.11)), leads to expect a relation like:

$$\mu^{\mathbf{M}}_R(\alpha) = \frac{1}{R}(1 + o(\frac{1}{R})), \tag{3.12}$$

(6) The property $\alpha(\mathbf{u}) = -\alpha(-\mathbf{u})$ implies that the evolution defined by Eq. (3.8) is *time reversible*, so that $\alpha(\mathbf{u})$ can be called "*reversible viscosity*".

4 Regularization

In Eqs. (2.5) and (3.12) the question on how large should R be for equivalence is implicitly raised. An answer, which may become relevant in simulations, that it would be interesting to investigate, is that the equivalence might hold much more generally, at least in the cases (1) in Sect. 2 above: therefore for the Navier Stokes equations in dimension 2 (and 3, see below).

The Navier-Stokes equation in $2D$ is known to admit unique evolution of smooth initial data, [3]. The same question has not yet been studied for the reversible viscosity case. In *both cases*, however, simulations impose that the field $\mathbf{u}$ must be represented by a finite number of data, *i.e.* it must be "regularized", to use the language of field theory, [19].

Here the regularization will simply be enforced by considering Eqs. (3.6) and (3.8) with fields with $\mathbf{u_k} \neq 0$ only if $\mathbf{k} \in I_N \stackrel{def}{=} \{|k_j| \leq N\}$. Consequently all statements will depend on the cut-off value N. In particular the conjecture of equivalence will have to be studied also as a function of N and for a fixed local observable.

Pursuing the analogy with equilibrium statistical mechanics, SM, of a system with energy E, temperature β^{-1} and observables localized in a volume V_0, mentioned above, consider.

(a) *the cut-off N as analogous to the total volume in SM,*
(b) *K–local observables (defined before Eq. (2.4)) as analogous to the observables localized in a volume $V_0 = K$ in SM*
(c) *the enstrophy $\mathcal{D}(\mathbf{u})$ as analogous to the energy in SM*

Furthermore the incompressible Navier Stokes equations (as well as the Euler equations or the more general transport equations) can be regarded, if $N = \infty$, as macroscopic versions of the atomic motion: the latter is certainly reversible (if appropriately described together with the external interactions) and essentially always strongly chaotic.

Therefore, for $N = \infty$ and at least for 2 dimensions, no matter whether R is small or large, the equivalence should not only remain valid but could hold in stronger form. Let $\mu^M_{\mathcal{E},N}, \mu^C_{R,N}$ be the stationary distributions for the regularized Navier-Stokes equations, then

Fixed K let F be a K-local observable; suppose that the equivalence condition $\mu^C_{R,N}(\mathcal{D}) = \mathcal{E}$ (or $\mu^M_{\mathcal{E},N}(\alpha) = \frac{1}{R}$) holds, then:

$$\begin{aligned} &(a)\ \mu^C_R = \lim_{N\to\infty} \mu^C_{R,N},\ \ \mu^M_{\mathcal{E}} = \lim_{N\to\infty} \mu^M_{\mathcal{E},N} \text{ exist} \\ &(b) \qquad \mu^C_R(F) = \mu^M_{\mathcal{E}}(F), \qquad\qquad \text{for all} \quad R, \mathcal{E} \end{aligned} \tag{4.13}$$

Remarks: (1) The statement is much closer in spirit to the familiar thermodynamic limit equivalence between canonical and microcanonical ensembles.
(2) Since the basis is that the microscopic motions that generate the Navier-Stokes equations are chaotic and reversible the limit $N \to \infty$ is essential.

(3) The truncated or full Navier stokes equations at *low Reynolds number* admit, for the same R, fixed point solutions, periodic solutions or even coexisting chaotic solutions, [20–23], and the condition of equivalence must be interpreted as meaning that when there are several coexisting stationary *ergodic* distributions then there is a one-to-one correspondence between the ones that in the two ensembles $\mathcal{F}^C, \mathcal{F}^M$ obey the equivalence condition and the averages of local observables obey Eq. (4.14).
(4) The possibility of coexisting stationary distributions actually observed in truncated NS equations in [20–22], (see also [3, (4.4.8)]), or predicted and observed at high turbulence, [24,25], is analogous to the phase coexistence in equilibrium statistical mechanics (and in that case too the equivalence can hold only in the thermodynamic limit).
(5) It is remarkable that above conjecture *really deals only with the regularized equations*: therefore it makes sense irrespective of whether the non regularized equations dimensionality is 2 or 3.
Of course in the 3–dimensional equation the $\alpha(\mathbf{u})$ has a somewhat different form, due to the presence of vortex stretching [3,26]; furthermore in the developed turbulence regimes, *in dimension* 3, the picture may become simpler: this is so because of the *natural cut-off due to the OK41* $\frac{5}{3}$*-law*: namely $|k_j| \le N = R^{\frac{3}{4}\varepsilon}$, $\varepsilon > 0$, [3].
(6) The equivalence also suggests that there might be even some relation between the "*T-local Lyapunov exponents*" of pairs of equivalent distributions. Here T–local exponents are defined via the Jacobian matrix $M_T(\mathbf{u}) = \partial S_T(\mathbf{u})$ and its RU-decomposition: they are the averages of the diagonal elements $l_j(\mathbf{u})$ of the R-matrix over T time steps of integration, [27]. Although the "local exponents" cannot be considered to be among the K-local observables it is certainly worth to compare the two spectra.
(7) A suggestion emerges that it would be interesting to study the R-NS equations with $\alpha(\mathbf{u})$ *replaced by a stochastic process* like a white noise centered at $\frac{1}{R}$ with the reversibility taken into account by imposing the width of the fluctuations to be also $\frac{1}{R}$, as required by the fluctuation relation, see below. As R varies stationary states describe a new ensemble which could be equivalent to $\mathcal{E}^C$ in the sense of the conjecture.
(8) A heuristic comment: if the *Chaotic hypothesis*, [16], is assumed for the evolution in the regularized equations the *fluctuation relation*, see below, should also hold, thus yielding a prediction on the large fluctuations of the observable "divergence of the equations of motion" $\sigma_N(\mathbf{u})$ in the distributions $\mu^{\mathbf{M}}_{\mathcal{E},N}$ which, in the 2-dimensional case, is:

$$\sigma_N(\mathbf{u}) = -\frac{\sum_{\mathbf{h}\in I_N}(\mathbf{h}^2)^2 \mathrm{Re}(\overline{\mathbf{g}}_{\mathbf{h}}\cdot\mathbf{u}_{\mathbf{h}}) - 2\alpha E_6}{E_4} - \alpha\sum_{\mathbf{h}\in I_N}\mathbf{h}^2 \qquad (4.14)$$

where $I_N \stackrel{def}{=} \{|k_j|\} \le N$, $E_{2m} = \sum_{\mathbf{h}\in I_N}(\mathbf{h}^2)^m\,|\mathbf{u}_{\mathbf{h}}|^2$ which follows, if $g_{\mathbf{h}} \stackrel{def}{=} g^r_{\mathbf{h}} + i\,g^i_{\mathbf{h}}$, from

$$\frac{\partial\alpha}{\partial\mathbf{u}^b_{\mathbf{h}}} = \frac{\mathbf{h}^2 g^b_{\mathbf{h}}}{E_4} - 2\alpha\frac{\mathbf{h}^4\mathbf{u}^b_{\mathbf{h}}}{E_4}, \quad b = r, i \tag{4.15}$$

Notice that the cut-off N is essential to define $\sigma_N(\mathbf{u})$ as the last (and main) term in Eq. (4.14) would be, otherwise, infinite.

If $\sigma_{N,+}$ is the infinite time average of $\sigma_N(S_t\mathbf{u})$, *i.e.* $\sigma_{+,N} \equiv \int \mu^M_R(d\mathbf{u})\,\sigma_N(\mathbf{u})$ and if $p_\tau(\mathbf{u}) = \frac{1}{\tau}\int_0^\tau \frac{\sigma_N(S_t\mathbf{u})}{\sigma_{N,+}}dt$ then the variable $p_\tau(\mathbf{u})$ satisfies the fluctuation relation in $\mu^M_{En,N}$ if, asymptotically as $\tau \to \infty$,

$$\frac{\mu^{\mathbf{M}}_{En,N}(p_\tau(\mathbf{u}) \sim p)}{\mu^{\mathbf{M}}_{En,N}(p_\tau(\mathbf{u}) \sim -p)} = e^{p\sigma_{N,+}\tau + o(\tau)} \tag{4.16}$$

The average $\sigma_{N,+}$ becomes infinite in the limit $N \to \infty$: which implies that the probability of $|p-1| > \varepsilon$ tends to 0 (exponentially in N^4, *i.e.* proportionally to $\varepsilon^2 \sum_{|\mathbf{k}|<N} \mathbf{k}^2$) so that the reversible viscosity (proportional to $\alpha \sim \frac{\sigma}{\sigma_+}$) will have probability tending to 0 as $N \to \infty$ (if the large deviation function has a quadratic maximum at $p = 1$ or faster if the maximum is steeper). Large fluctuations of the reversible viscosity away from $\frac{1}{R}$ are still possible if $N < \infty$ but not observable, [28, Eq. (5.6.3)].

Some of the questions raised in the remarks in the above sections will now be analyzed in a series of simulations in the Appendix. They are very preliminary tests and are meant just to propose tests to realize in the future to test validity, dependence/stability of the results as N, R vary. Source-codes (in progress) available on request

Acknowledgements. This is an extended version of part of my talk at and includes only the material prepared to propose simulations to the attending postdocs during my stay at the Institut Henri Poincaré - Centre Emile Borel during the trimester Stochastic Dynamics Out of Equilibrium. I am grateful to the organizers for the support and hospitality and also for the possibility of starting and performing the presented simulations on the IHP computer cluster; I thank also L. Biferale for providing computer facilities to improve the graphs. I am grateful for long critical discussions with L. Biferale, M. Cencini, M. De Pietro, A. Giuliani and V. Lucarini: they provided hints and stimulated the ideas here.

A Appendix: Reversible Viscosity and Reynolds Number

We first analyze the evolution and distribution of the reversible viscosity $\alpha(\mathbf{u})$ defined in Eq. (3.10) considered as an observable for the evolution Eq. (3.6), *i.e.* for the *irreversible NS2D evolution*.

Consider the NS2D with regularization $(2N+1) \times (2N+1)$. For $N = 3$ a simulation gives the running average of the value of $R\alpha(\mathbf{u})$ (drawn every 5 data to avoid a too dense a figure), the actual fluctuating values of $R\alpha(\mathbf{u})$ and the straight line at quota 1. It shows that $R\alpha(\mathbf{u})$ fluctuates strongly, yet $R\alpha(\mathbf{u})$ averages to a value close ($\sim$2%) to 1, *i.e.* $\alpha(\mathbf{u})$ averages to the viscosity value:

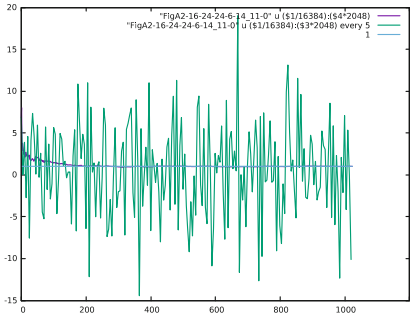


Fig. 2. The modes are in the 7×7 box centered at the origin, corresponding to a cut-off $N = 3$; the Reynolds number is $R = 2^{11}$; the time step is 2^{-14} and the time axis is in units of 2^{14} (*i.e.* the evolution history is obtained via 2^{24} time steps).

the analogy, mentioned earlier in Eq. (3.12), with equilibrium thermodynamics would suggest checking that at large R, $\mu^C_{R,N}(\alpha) = \frac{1}{R}(1+o_{R,N})$ with $o_{R,N}$ small. A check is also necessary because $\alpha(\mathbf{u})$ is not a K-local observable.

The same data considered in Fig. 1 for $R = 2014$ and 2^{26} integration steps of size 2^{-15} drawn every $10 \cdot 2^{15}$ yield:

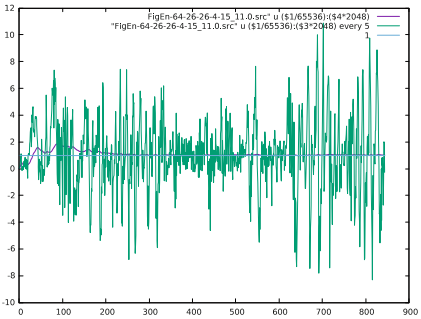


Fig. 3. At 960 modes and $R = 2048$: the evolution of the observable "*reversible viscosity*", *i.e.* $\alpha(\mathbf{u})$ in Eq. (3.10) in the I-NS: the time average of α should be $\frac{1}{R}(1 + o(\frac{1}{R}))$. Represents the fluctuating values of α every $5 \cdot 2^{16}$ integration steps; the middle line is the running average of α and it *is close* to $\frac{1}{R}$ (horiz. line).

It would be interesting to present a few more recent results on the closeness of the Lyapunov spectra of the R-NS and I-NS but the analysis requires further work.

References

1. Evans, D.J., Morriss, G.P.: Statistical Mechanics of Nonequilibrium Fluids. Academic Press, New York (1990)

2. Feynman, R.P., Vernon, F.L.: The theory of a general quantum system interacting with a linear dissipative system. Ann. Phys. **24**, 118–173 (1963)
3. Gallavotti, G.: Foundations of Fluid Dynamics, 2nd edn. Springer, Berlin (2005)
4. She, Z.S., Jackson, E.: Constrained Euler system for Navier-Stokes turbulence. Phys. Rev. Lett. **70**, 1255–1258 (1993)
5. Gallavotti, G.: Equivalence of dynamical ensembles and Navier Stokes equations. Phys. Lett. A **223**, 91–95 (1996)
6. Gallavotti, G.: Dynamical ensembles equivalence in fluid mechanics. Physica D **105**, 163–184 (1997)
7. Gallavotti, G.: Breakdown and regeneration of time reversal symmetry in nonequilibrium statistical mechanics. Physica D **112**, 250–257 (1998)
8. Gallavotti, G., Rondoni, L., Segre, E.: Lyapunov spectra and nonequilibrium ensembles equivalence in 2d fluid. Physica D **187**, 358–369 (2004)
9. Gallavotti, G., Lucarini, V.: Equivalence of non-equilibrium ensembles and representation of friction in turbulent flows: the Lorenz 96 model. J. Stat. Phys. **156**, 1027–1065 (2014)
10. Rondoni, L., Mejia-Monasterio, C.: Fluctuations in nonequilibrium statistical mechanics: models, mathematical theory, physical mechanisms. Nonlinearity **20**, R1–R37 (2007)
11. Benzi, R., Paladin, G., Parisi, G., Vulpiani, A.: Multifractal and intermittency in turbulence. In: Benzi, R., Basdevant, C., Ciliberto, S. (eds.) Nova Science Publishers, Commack (1993)
12. Biferale, L.: Shell models of energy cascade in turbulence. Ann. Rev. Fluid Mech. **35**, 441–468 (2003)
13. Maxwell, J.C.: On the dynamical theory of gases. In: Niven, W.D. (ed.) The Scientific Papers of J.C. Maxwell, vol. 2, pp. 26–78. Cambridge University Press, Cambridge (1986)
14. Ruelle, D.: Turbulence, Strange Attractors and Chaos. World Scientific, New York (1995)
15. Gallavotti, G., Cohen, D.: Dynamical ensembles in nonequilibrium statistical mechanics. Phys. Rev. Lett. **74**, 2694–2697 (1995)
16. Gallavotti, G., Cohen, D.: Dynamical ensembles in stationary states. J. Stat. Phys. **80**, 931–970 (1995)
17. De Pietro, M.: Nonlinear helical interactions in Navier-Stokes and shell models for turbulence. Ph.D. thesis, Università Tor Vergata, Roma, pp. 1–102 (2017)
18. Ruelle, D., Takens, F.: On the nature of turbulence. Commun. Math. Phys. **20**, 167–192 (1971)
19. Gallavotti, G.: Renormalization theory and ultraviolet stability for scalar fields via renormalization group methods. Rev. Mod. Phys. **57**, 471–562 (1985)
20. Franceschini, V., Tebaldi, C.: Sequences of infinite bifurcations and turbulence in a five-mode truncation of the Navier-Stokes equations. J. Stat. Phys. **21**, 707–726 (1979)
21. Franceschini, V., Tebaldi, C.: Truncations to 12, 14 and 18 modes of the Navier-Stokes equations on a two-dimensional torus. Meccanica **20**, 207–230 (1985)
22. Franceschini, V., Tebaldi, C., Zironi, F.: Fixed point limit behavior of N-mode truncated Navier-Stokes equations as N increases. J. Stat. Phys. **35**, 387–397 (1984)
23. Marchioro, C.: An example of absence of turbulence for any Reynolds number. Commun. Math. Phys. **105**, 99–106 (1986)

24. Frisch, U.: Fully developed turbulence: where do we stand? In: Diner, S., Fargue, D., Lochak, G. (eds.) Dynamical Systems. A Renewal of Mechanism: Centennial of G.D. Birkhoff. World Scientific (1986)
25. Huisman, S.G., Van der Veen, R.C., Sun, C., Lohse, D.: Multiple states in highly turbulent Taylor-Couette flow. Nat. Commun. **5**(3820), 1–5 (2014)
26. Gallavotti, G.: New methods in nonequilibrium gases and fluids. Open Syst. Inf. Dyn. **6**, 101–136 (1999)
27. Benettin, G., Galgani, L., Giorgilli, A., Strelcyn, J.: Lyapunov charateristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part I, theory. Meccanica **15**, 9–20 (1980)
28. Gallavotti, G.: Nonequilibrium and Irreversibility. Theoretical and Mathematical Physics. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06758-2

On the Nonequilibrium Entropy of Large and Small Systems

Sheldon Goldstein[1], David A. Huse[2], Joel L. Lebowitz[3(✉)], and Pablo Sartori[4]

[1] Department of Mathematics, Rutgers University, Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA
oldstein@math.rutgers.edu

[2] Department of Physics, Princeton University, Jadwin Hall, Washington Road, Princeton, NJ 08544-0708, USA
huse@princeton.edu

[3] Departments of Mathematics and Physics, Rutgers University, Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA
lebowitz@math.rutgers.edu

[4] The Center for Studies in Physics and Biology and Laboratory of Living Matter, Rockefeller University, New York, NY 10065, USA
psartori@rockefeller.edu

Abstract. Thermodynamics makes definite predictions about the thermal behavior of macroscopic systems in and out of equilibrium. Statistical mechanics aims to derive this behavior from the dynamics and statistics of the atoms and molecules making up these systems. A key element in this derivation is the large number of microscopic degrees of freedom of macroscopic systems. Therefore, the extension of thermodynamic concepts, such as entropy, to small (nano) systems raises many questions. Here we shall reexamine various definitions of entropy for nonequilibrium systems, large and small. These include thermodynamic (hydrodynamic), Boltzmann, and Gibbs-Shannon entropies. We shall argue that, despite its common use, the last is not an appropriate physical entropy for such systems, either isolated or in contact with thermal reservoirs: physical entropies should depend on the microstate of the system, not on a subjective probability distribution. To square this point of view with experimental results of Bechhoefer we shall argue that the Gibbs-Shannon entropy of a nano particle in a thermal fluid should be interpreted as the Boltzmann entropy of a dilute gas of Brownian particles in the fluid.

Keywords: Nonequilibrium thermodynamics · Statistical mechanics

1 Introduction

The role of probability in the statistical mechanical analysis of the thermal behavior of individual physical systems is subtle. Indeed, it has frequently been a source of confusion and controversy: note e.g. the conflict between Boltzmann

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 581–596, 2019.
https://doi.org/10.1007/978-3-030-15096-9_22

and Zermelo about the H-theorem [1]. A crucial ingredient in the statistical mechanical analysis of this problem by Maxwell, Thomson, Boltzmann, Gibbs and Einstein is the "law of large numbers", which permits "almost sure" predictions, i.e. with probability approaching 1, when the number of quasi-independent entities in the system become very large. This is clearly the case for macroscopic systems (MS), which contain a large number of atoms or molecules, to which statistical mechanics was historically restricted. Thus the microcanonical ensemble and the other equilibrium Gibbs ensembles make definite predictions for equilibrium MS. We can therefore speak of the "typical" behavior of such a system[1]. This restriction to MS was historically natural, since the notions of heat, entropy and the second law were all developed in the nineteenth century for such systems. The subsequent development of statistical mechanics had as its aim to describe and explain microscopically the observed thermal phenomena in such MS. It therefore also considered only systems consisting of very large numbers of particles.

In going beyond equilibrium, where the theory is fundamentally complete, this disparity in sizes between microscopic and macroscopic also plays a critical role. It forms the basis of the explanation by Boltzmann, Maxwell and Thomson of how time-asymmetric behavior, as expressed for example by the second law of thermodynamics, can originate from time-symmetric microscopic laws [2,3]. In particular, time-asymmetric macroscopic equations like the heat equation or the Navier-Stokes equations, as well as the mesoscopic Boltzmann equation (to which Zermelo objected), can be seen as being expressions of the law of large numbers, valid in the limit of particle number $N \to \infty$ [4]. Unfortunately, a rigorous mathematical derivation of such equations from time-symmetric microscopic dynamics is still beyond our reach for realistic systems. In fact, the only cases for which hydrodynamic-type equations have been derived rigorously are systems with bulk stochastic interactions, like lattice gases [5–7]. Therefore, there are still many open problems for nonequilibrium MS.

The reliance on the law of large numbers raises the issue of understanding the thermal behavior of nanosystems (NS), in which there is currently much interest. This interest is fueled by technological advances that make such systems experimentally accessible. Nanosystems can be well isolated from their environment, or can be in contact with reservoirs. Here we will focus on the latter case, an example being the recent work of Bechhoefer et al. [8], of a nanoparticle immersed in a fluid (a talk by Bechhoefer triggered this work). Recent work on such NS goes under the name of "stochastic thermodynamics", see [9] and other articles in that issue. Stochastic thermodynamics, as the name indicates, takes explicit account of the stochastic modeling of the effective interactions between the small system, such as a nanoparticle or a polymer, and the equilibrium thermal reservoir that it is in contact with, usually a macroscopic fluid. There is also much current interest in isolated quantum systems having only a few degrees of freedom [10], but we shall not consider these here.

[1] We are taking for granted here an assumed underlying (approximately) equal a priori probability of different microstates for a specified macrostate.

The consideration of thermal properties of NS raises the question of whether and how the thermodynamic and statistical mechanics formalism can be extended to systems with a small number of degrees of freedom. In the stochastic thermodynamic extensions the Gibbs-Shannon (GS) entropy (defined in (9)) plays a central role. This raises the questions: does the GS entropy of a probability measure μ, when μ is not a Gibbs measure for an equilibrium macroscopic system, have physical meaning? And when it does, what is that meaning? This entropy has some very nice mathematical properties and it is very alluring to consider it, as is generally done in the stochastic thermodynamic literature, as the "proper entropy" of a nonequilibrium system. We shall argue against this for MS. On the other hand the experiments of Bechhoefer, mentioned earlier, actually have measured this quantity, more or less directly, for a nanoparticle immersed in a liquid. We shall discuss our interpretation of these experiments at the end of this note. Let us however start from the beginning and first consider the statistical mechanical entropy for isolated MS. We shall then consider both MS and NS in contact with thermal baths.

2 Thermodynamic and Boltzmann Entropy of a MS

The discovery by Clausius of the existence of an entropy function $S(E, V, N)$ for equilibrium macroscopic systems (with energy E, particle number N and volume V), and its central role in the time asymmetric evolution of the world, as expressed by the second law, is one of the key events of nineteenth century science, c.f. [11] and [12]. This discovery raised immediately the question of how to define S as a function of the microstate $X = (\mathbf{r}_1, \mathbf{p}_1, \ldots, \mathbf{r}_N, \mathbf{p}_N)$ of the particles composing the system, where $\mathbf{r}_i \in V$ is the position and $\mathbf{p}_i \in \mathbb{R}^3$ the momentum of the ith particle. The problem was compounded by the fact that the time evolution $X(t)$ in the phase space Γ, as given by the Hamiltonian $H(X)$, is time symmetric, c.f. [3].

The answer arrived at by Boltzmann was to identify S for a MS with

$$S_{\mathrm{B}}(X) = \log |\Gamma(M(X))|. \tag{1}$$

Here X is a phase point (microstate) in the energy shell, $E \leq H(X) \leq E + \Delta E$ and $M(X)$ is the macrostate of the system. This macrostate is defined, e.g., by dividing V into $\mathcal{N}$ cells ω_α, $\alpha = 1, \ldots \mathcal{N}$, with $1 \ll \mathcal{N} \ll N$, and then specifying the number, energy, and total momentum of the particles in each ω_α with a certain tolerance. $|\Gamma(M)|$ is the Liouville (Lebesgue) volume of the phase space region $\Gamma(M)$ containing all microstates X belonging to the macrostate M, c.f. [2,3] and also Sect. 7 of this work. (For classical systems there is an arbitrary overall additive constant in the entropy coming from the unit of phase space volume, but this has no impact on what we discuss here, so will be ignored.) For a macroscopic system there is a special macrostate M_{eq}, corresponding to equilibrium, such that $\Gamma(M_{\mathrm{eq}})$ covers almost the whole surface of energy E. That is, $|\Gamma(M_{\mathrm{eq}})| \sim |\Gamma_E|$, the volume of the energy shell between E and $E + \Delta E$.

This definition of $S_{\mathrm{B}}(X)$ assigns an entropy even to microstates X which do not behave at all as expected from the second law, as discussed below after (2).

The Boltzmann entropy $S_{\mathrm{B}}(M_{\mathrm{eq}})$ agrees, to leading order in the size of the system, with the experimentally determined equilibrium Clausius entropy S. This was shown for a dilute gas by Boltzmann and for general systems by Gibbs. To calculate this entropy Gibbs (and Boltzmann) introduced the microcanonical ensemble μ_{m}, as the uniform probability density in the energy shell $E \leq H(X) \leq E + \Delta E$ for describing the almost sure properties of equilibrium MS with energy E. They naturally equated the precisely defined logarithm of $|\Gamma_E|$, the volume of the energy shell which for MS is very close to $|\Gamma(M_{\mathrm{eq}})|$, with the entropy $S(E)$.

Time-Evolution: The time evolution of the microstate $X(t)$ will induce a time evolution of the macrostate M. Boltzmann then argued that $S_{\mathrm{B}}(X(t))$ will, for a typical X in $\Gamma(M)$, evolve in time according to the second law, i.e.

$$\frac{\mathrm{d}S_{\mathrm{B}}(X(t))}{\mathrm{d}t} \geq 0, \quad t > 0 \tag{2}$$

see [4] and references there. This can be proven when one assumes that $M(t)$ evolves under an autonomous macroscopic equation, e.g. the Navier-Stokes or diffusion equation, but the rigorous derivations of these equations from the microscopic dynamics is not available at the present time.

Nota Bene: Equation (2) can only be true for typical microstates X, i.e. for the overwhelming majority of the X's with respect to Liouville measure in $\Gamma(M)$: there are special microstates for which it is definitely false. An example of such a special state can be obtained by starting in a typical low-entropy state at some time in the past, and then evolving that state in time to a higher-entropy state in the present, followed by exactly reversing all velocities.

Hydrodynamic Time-Evolution: We now describe a class of nonequilibrium systems for which the time evolution of the Boltzmann entropy is given by hydrodynamic equations and satisfies the second law. Consider a system in a macrostate $M \neq M_{\mathrm{eq}}$ for which one can define a "smooth" energy and mass density profile $e(\mathbf{r})$ and $n(\mathbf{r})$, where $\mathbf{r} \in V$ denotes different spatial points of the system. For such systems $S_{\mathrm{B}}(X)$ coincides to leading order with $S_{\mathrm{h}}(\{e(\mathbf{r}), n(\mathbf{r})\})$, the hydrodynamic entropy of systems in local thermal equilibrium (LTE) given by [13]

$$S_{\mathrm{h}}(\{e(\mathbf{r}), n(\mathbf{r})\}) = \int_V s(e(\mathbf{r}), n(\mathbf{r}))\mathrm{d}\mathbf{r}, \tag{3}$$

where $s(e, n)$ is the equilibrium entropy per unit volume, $s = S/|V|$ (in the thermodynamic limit). [We have assumed for simplicity that the local velocity $u(\mathbf{r})$ is zero, otherwise $e(\mathbf{r}) \to [e(\mathbf{r}) - \frac{1}{2}n(\mathbf{r})|u(\mathbf{r})|^2]$. Note that S_{h} coincides with the equilibrium entropy $S(E, V, N)$ when e and n are independent of $\mathbf{r}$.

As an example of the hydrodynamic time evolution of S_{h}, consider a system in LTE with a temperature profile $T(\mathbf{r}, t)$. Starting then with the general equation

for the time evolution of the entropy density of a system in LTE

$$\frac{\partial s(\mathbf{r},t)}{\partial t} = -\frac{\text{div } j(\mathbf{r},t)}{T(\mathbf{r},t)} = -\text{div } (j/T) + j \cdot \nabla \left(\frac{1}{T}\right), \tag{4}$$

where $j(\mathbf{r},t)$ is the energy flux vector, we get

$$\frac{\mathrm{d}S_\mathrm{h}}{\mathrm{d}t} = -\int_Q \frac{j(\mathbf{q},t)}{T(\mathbf{q})} \cdot \mathrm{d}\mathbf{q} + \int_V j(\mathbf{r},t) \cdot \nabla \left(\frac{1}{T(\mathbf{r},t)}\right) \mathrm{d}\mathbf{r}. \tag{5}$$

In (5) Q is the surface of V and $\mathrm{d}\mathbf{q}$ is the (outward directed) surface area element. The flux integrand vanishes on the parts of the surface which are insulated, the whole surface if the system is isolated. There will, however, be a contribution from the parts of the surface which are held at specified temperatures $T(\mathbf{q})$ by external reservoirs. The integral over Q can be identified with the entropy production in the thermal reservoirs, $đS_\mathrm{r}/\mathrm{d}t$, which maintain the temperature $T(\mathbf{q})$. (Since we have idealized the reservoirs as infinite systems with fixed temperatures, their entropy is formally infinite, but the rate of change in their entropies is finite.) The second term in (5) corresponds to the hydrodynamic or Boltzmann entropy change in the bulk of the MS,

$$\sigma_\mathrm{B}(t) = \int_V j(\mathbf{r},t) \cdot \nabla \left(\frac{1}{T(\mathbf{r},t)}\right) \mathrm{d}\mathbf{r} \geq 0, \tag{6}$$

due to local "dissipation". The integrand is in fact everywhere non-negative, an expression of the second law: the component of the energy flux parallel to the temperature gradient cannot be directed from 'cold' towards 'hot'.

3 The Gibbs-Shannon Entropy

The entropy of the micro-canonical ensemble, $S(E) = \log |\Gamma_E|$, can also be written as

$$S(E) = S_\mathrm{G}(\mu_\mathrm{m}) = -\int \mu_\mathrm{m}(X) \log \mu_\mathrm{m}(X) \mathrm{d}X. \tag{7}$$

Using Legendre transforms, Gibbs showed that if one considers the canonical ensemble with probability density μ_β given by

$$\mu_\beta = Z^{-1} \exp\left[-\beta H(X)\right], \tag{8}$$

with $\beta = 1/T$, then $S_\mathrm{G}(\mu_\beta)$ also gives the equilibrium entropy of a MS as a function of temperature. The same is true for the grand-canonical ensemble and other equilibrium ensembles. They all agree to leading order in the size of the system.

It is a natural step to extend this notion of entropy to general probability measures with densities $\mu(X,t)$ which depend on X and t as

$$S_\mathrm{G}(\mu) = -\int \mu(X,t) \log \mu(X,t) \mathrm{d}X. \tag{9}$$

The quantity $S_{\mathrm{G}}(\mu)$ is the Shannon entropy of an arbitrary measure μ on a space Ω (relative to the measure $\mathrm{d}X$). It plays a central role in information theory as developed by Shannon [14]. However, as is well known, for an isolated physical system evolving under Hamiltonian dynamics, μ changes in time according to the Liouville equation, $\partial\mu/\partial t = -\{\mu, H\}$, and the GS entropy $S_{\mathrm{G}}(\mu(t))$ does not change at all. Thus $S_{\mathrm{G}}(\mu(t))$ cannot be identified with the thermodynamic or hydrodynamic entropy S_{h} of an isolated macroscopic system which is not in global thermal equilibrium, even if it is in local thermal equilibrium, a situation in which S_{h} is unambiguous. (This was already noted by Gibbs and discussed by P. and T. Ehrenfest in their 1916 article [15]). This raises the question of what is the physical meaning of $S_{\mathrm{G}}(\mu)$ for any system for which μ is not an equilibrium Gibbs measure of a MS.

The behavior of the GS entropy associated with a measure μ is very different when the system is in contact with stochastic reservoirs. As will be seen below, the rate of change of S_{G} is no longer zero and is related to the thermodynamic entropy change in the reservoirs. We shall discuss the physical significance, if any, of this later, after we introduce the mathematical formalism to describe such systems.

4 Model of System in Contact with Thermal Reservoirs

The formalism we shall use was developed by Bergmann and Lebowitz [16,17] who studied the dynamics of a system evolving under the combined action of its own Hamiltonian $H(X)$ and of n thermal reservoirs at different temperatures (and chemical potentials). These reservoirs were thought of as being infinite and acting at the boundaries of the MS. To simplify matters the interaction with the reservoirs was idealized as being of the collision type: when a collision occurs the phase point of the system, X, jumps to X', while the reservoir particle goes off to infinity, never to be seen again. The system thus sees an ever fresh stream of reservoir particles with a Maxwellian distribution, at the temperature $T_\alpha = \beta_\alpha^{-1}$ of that reservoir, $\alpha = 1, \cdots, n$. The time evolution of the system will thus be given by a continuous time Markov process.

Denoting $K_\alpha(X, X')\mathrm{d}X$ the transition rate from the phase point X' to the phase space volume $\mathrm{d}X$ around X due to collisions with reservoir α, yields the following stochastic Liouville master equation for the probability density $\mu(X, t)$,

$$\frac{\partial\mu(X,t)}{\partial t} + \{\mu, H\} = \sum_{\alpha=1}^{n} \int \left[K_\alpha(X, X')\mu(X', t) - K_\alpha(X', X)\mu(X, t)\right] \mathrm{d}X', \tag{10}$$

where $\{\mu, H\}$ is the usual Poisson bracket describing the deterministic Hamiltonian evolution of the isolated system.

Using the time reversibility of the collision dynamics yields a condition for each α,

$$K_\alpha(X, X') = e^{\beta_\alpha H(X')} L_\alpha(X, X') \tag{11}$$

with $L_\alpha(X, X') = L_\alpha(\overline{X'}, \overline{X})$, where $\overline{X}$ corresponds to reversal of the velocity coordinates of X. Some further simplifications give $L_\alpha(\overline{X'}, \overline{X}) = L_\alpha(X', X)$, so that $L_\alpha(X, X') = L_\alpha(X', X)$, corresponding to "detailed balance" for each reservoir, i.e.

$$K_\alpha(X, X')/K_\alpha(X', X) = \exp[-\beta_\alpha(H(X) - H(X'))]. \tag{12}$$

It was proven in [16], under quite general conditions on $L_\alpha(X, X')$ that, as $t \to \infty$, a system started in some arbitrary initial $\mu(X, 0)$ will approach a stationary state

$$\lim_{t\to\infty} \mu(X, t) = \mu_{\mathrm{s}}(X). \tag{13}$$

This state is unique and is absolutely continuous with respect to Liouville measure. When there is only one reservoir at reciprocal temperature, β_α, then clearly

$$\mu_{\mathrm{s}}(X) = \mu_\alpha(X) \equiv Z^{-1} \exp\left[-\beta_\alpha H(X)\right] \tag{14}$$

is the unique stationary state. When the temperatures β_α^{-1} are different μ_{s} will be a nonequilibrium stationary state (NESS), for which the dynamics do not satisfy detailed balance. It was further shown that this NESS will satisfy the Onsager reciprocal relations when all β_α are close to some $\overline{\beta}$, as well as a generalized Kubo relation in the presence of an external field.

5 Time Evolution of Gibbs Entropy for a System in Contact with Thermal Reservoirs

For a closed system, given the phase point $X(t_0)$, $X(t)$ is determined for all t. The only randomness expressed in $\mu(X, t)$ for a closed system is that introduced initially, which could be due to ignorance. There is therefore no intrinsic physical significance to $\mu(X, t)$ for an isolated system. On the other hand when the system is in contact with a stochastic reservoir then $X(t)$ is no longer determined by $X(t_0)$, and $\mu(X, t)$ acquires some "objective" meaning. The GS entropy now evolves in time in a non trivial way, which can be calculated from (10). It consists of two contributions,

$$\frac{\mathrm{d}}{\mathrm{d}t} S_{\mathrm{G}}(\mu) = \sum_{\alpha=1}^{n} J_\alpha(t)/T_\alpha + \sigma_{\mathrm{G}}(t). \tag{15}$$

In the first contribution J_α is the average energy flux *from* the αth reservoir into the system, that is

$$J_\alpha(t) = \int \mu(X, t) \int K_\alpha(X', X) \left[H(X') - H(X)\right] \mathrm{d}X' \mathrm{d}X, \tag{16}$$

with $\sum_{\alpha=1}^{n} J_\alpha(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int H(X)\mu(X,t)\mathrm{d}X$. The second contribution in (15) is

$$\sigma_{\mathrm{G}}(t) = \frac{1}{2}\sum_{\alpha=1}^{n}\int\int L_\alpha(X,X')\left[\nu_\alpha(X,t)-\nu_\alpha(X',t)\right]\log\left[\frac{\nu_\alpha(X,t)}{\nu_\alpha(X',t)}\right]\mathrm{d}X\mathrm{d}X' \geq 0, \tag{17}$$

where

$$\nu_\alpha(X,t) = \mu(X,t)\exp\left[\beta_\alpha H(X)\right]. \tag{18}$$

Equation (15) can be rewritten in the suggestive form

$$\sigma_{\mathrm{G}}(t) = \frac{đS_{\mathrm{r}}}{\mathrm{d}t} + \frac{\mathrm{d}S_{\mathrm{G}}}{\mathrm{d}t} \geq 0, \tag{19}$$

where we have written $đS_{\mathrm{r}}/\mathrm{d}t = \sum_\alpha đS_\alpha/\mathrm{d}t$ and

$$\frac{đS_\alpha}{\mathrm{d}t} = -J_\alpha/T_\alpha \tag{20}$$

is the rate of change of the entropy of the αth reservoir caused by the energy (heat) flow $-J_\alpha$ *into* that reservoir. Equation (19) is reminiscent of the second law, and has therefore prompted the interpretation of $S_{\mathrm{G}}(\mu)$ for systems in contact with thermal reservoirs as a physical entropy, despite the fact that it is not so for an isolated system and is not specified by the microstate of the system, c.f. Sect. 9.

We want to argue however that (19) does not justify the interpretation of S_{G} as the physical entropy of an open nonequilibrium system unless it agrees, at least to leading order, with S_{B}. In fact for a MS in contact with reservoirs at its surface all the entropy production σ_{G} is caused, as can be seen from (17), by the stochastic interactions at its surface. This is in contrast to the entropy production σ_{B}, given in (6), which is due to the chaotic microscopic dynamics in the bulk of the system, as it should be from a physical point of view.

We note further that, as is well known, for general Markov processes the GS entropy relative to the stationary measure,

$$S_{\mathrm{G}}(\mu|\mu_{\mathrm{s}}) = -\int_\Gamma \mu(X,t)\log\left(\frac{\mu(X,t)}{\mu_{\mathrm{s}}(X)}\right)\mathrm{d}X = S_{\mathrm{G}}(\mu) + \int \mu\log\mu_{\mathrm{s}}\mathrm{d}X, \tag{21}$$

is monotone non decreasing [18]. We thus always have

$$\frac{\mathrm{d}}{\mathrm{d}t}S_{\mathrm{G}}(\mu) + \frac{\mathrm{d}}{\mathrm{d}t}\int \mu\log\mu_{\mathrm{s}}\mathrm{d}X \geq 0, \tag{22}$$

irrespective of whether the stochasticity comes from thermal reservoirs or not. This time derivative coincides in our case with σ_{G} when the system is in contact with only one reservoir and $\mu_{\mathrm{s}} \sim \exp[-\beta_\alpha H]$. When the system is in contact with several reservoirs then in addition to (21) we also have

$\frac{\mathrm{d}}{\mathrm{d}t}S_{\mathrm{G}}(\mu) + \frac{\mathrm{d}}{\mathrm{d}t}\sum_\alpha \int \mu \log \mu_\alpha \mathrm{d}X \geq 0$. The positivity of σ_{G} is thus simply a consequence of (22) and the detailed balance condition for each reservoir which gives (12). Equation (22) would thus hold whatever the stationary states, μ_α, of the system in contact with only one reservoir.

The relationship between S_{G} and S_{B} for open systems is an interesting question. It may be considered in the following example in which a macroscopic system is in contact with a single reservoir, e.g. a metal ball of radius 10 cm immersed in a large tub of water. Consider the case when at $t = 0$ we have $\mu(X, 0) = \mu_{\beta_0}(X)$, i.e. the system is in equilibrium with a reservoir at the temperature T_0. At $t = 0$ the system is suddenly coupled to a thermal reservoir at temperature T_f, with which it comes to equilibrium as $t \to \infty$. We thus have $S_{\mathrm{G}}(0) = S_{\mathrm{B}}(0)$ and $S_{\mathrm{G}}(\infty) = S_{\mathrm{B}}(\infty)$, but what about the times in between? Under the very reasonable assumption of LTE, say $T_0 = 50\,^\circ\mathrm{C}$ and $T_f = 30\,^\circ\mathrm{C}$, $S_{\mathrm{B}}(t)$ can be computed for all $t > t_0$ from the heat equation, but what about $S_{\mathrm{G}}(t)$? Does it agree with $S_{\mathrm{B}}(t)$ to leading order? Or do we have $S_{\mathrm{G}}(t) < S_{\mathrm{B}}(t)$ to leading order for some values of t? We do not know.

A similar question can be asked when the system is in contact with two (or more) reservoirs at different temperatures on its surface. We expect that if the system is macroscopic and chaotic, i.e. it satisfies Fourier's law, then the energy and density profile in the stationary state computed as an average over μ_{s} will be that corresponding to LTE. The quantity σ_{G} will then be given by

$$\sigma_{\mathrm{G}} = -\sum_\alpha J_\alpha / T_\alpha, \tag{23}$$

since $\mathrm{d}S_{\mathrm{G}}(\mu_{\mathrm{s}})/\mathrm{d}t = 0$, in (15). As long as the first integral in (5) can be identified with (23), there will be a similar expression for the hydrodynamic entropy production σ_{B} in (5) and (6). This raises the following question: to what extent does the stationary measure μ_{s} for a "chaotic" system correspond to a LTE state when the only stochasticity is the one at the surface induced by the reservoirs. In other words, does $S_{\mathrm{G}}(\mu_{\mathrm{s}}) = S_{\mathrm{B}}$ in this scenario? For the particular case where there are also bulk stochastic interactions which satisfy detailed balance this has been proven [19–21]. However, for the more general case in which the bulk dynamics is Hamiltonian this remains an open question.

In fact, it is not true that $S_{\mathrm{G}}(\mu_{\mathrm{s}}) = S_{\mathrm{B}}$ when the reservoirs at the surface are dissipative but deterministic, see [22]. There, it is considered a NESS produced by driving the system via deterministic non Hamiltonian forces of the type used in Gaussian thermostats at the surface. These yield a NESS, μ_{s}, which is singular with respect to Liouville measure. Its Gibbs entropy, $S_{\mathrm{G}}(\mu_{\mathrm{s}})$, is thus equal to $-\infty$. On the other hand molecular dynamics simulations of this model show that it is in LTE, corresponding to shear flow, as far as thermodynamic quantities are concerned. Whether such a situation can also occur when the NESS is produced by stochastic thermal reservoirs is an open question. This problem is also discussed in Sect. 6 of [23] for different kinds of reservoirs.

6 Small System in Contact with a Thermal Reservoir

Isolated small classical systems, such as a few particles in a box, and systems with only a few relevant degrees of freedom, such as the center of mass motion of a massive pendulum or the moon, are not thought to have any thermodynamic functions, such as entropy, associated with them. For this reason the second law is, as noted by Maxwell, constantly being violated in small systems [24]. It is certainly no great surprise if an isolated box containing 10 Argon atoms is frequently seen to have 8 or more particles in the right half of the box. Such a percentage of particles on one side of an isolated system would certainly be a violation of the second law if the system consisted of 10^{20} or more particles. Just how large does the system have to be to rule out "ever" seeing such a violation in an isolated system of N particles during a period of 100 years depends (strongly) on the nature of the interaction between the particles, shape of the container, initial state, and on whether we are considering classical or quantum dynamics. Leaving quantum systems for a separate consideration, we will analyze now what happens to a small classical system in contact with a thermal reservoir.

The Hamiltonian of such a small system in contact with a thermal reservoir is given by

$$H_{\mathrm{tot}} = H_{\mathrm{sys}}(X) + H_{\mathrm{r}}(Y) + V(X,Y), \tag{24}$$

where X describes the relevant part of the microstate of the small system with Hamiltonian H_{sys}, Y that of the reservoir with Hamiltonian H_{r}, and V is the interaction between system and reservoir. If the total system is in equilibrium and is described by a microcanonical or canonical ensemble at a temperature β^{-1}, this induces a probability density for the system $\tilde{\mu}(X) = \int \mu_\beta(X,Y)\mathrm{d}Y \sim \exp[-\beta(H_{\mathrm{sys}}(X) + \tilde{V}(X,\beta))]$. Note that $\tilde{V}$ will generally be determined by both, H_{r} and V, and can depend on β. This has to be taken into account when one considers, for example, the collapse transition of a polymer in a solvent [25,26].

We note here that $\tilde{\mu}(X)$ no longer gives almost sure predictions about the properties of the small system. We will presumably however get the same $\tilde{\mu}(X)$ when the size of the reservoir is very large for all different Gibbs ensembles describing the total system. Nevertheless, it is not clear how meaningful it is to assign thermodynamic functions to the small system based on $\tilde{\mu}(X)$: see discussion in [25]. We shall focus here on cases where the interaction $V(X,Y)$ can be taken to be of the impulsive type, as in Sect. 3, where $\tilde{V}(X)$ can be taken to be essentially independent of X and set equal to zero. The paradigm of such a system is a Brownian particle [BP] immersed in an equilibrium fluid at temperature T. The only relevant degree of freedom for such a (spherical) particle is the location of its center of mass. The phase space Γ of the system is thus six dimensional, $X = (\mathbf{r}, \mathbf{v})$, where we have set the mass of the BP equal to unity so that $\mathbf{p} = \mathbf{v}$. Treating the fluid (approximately) as an infinite thermal reservoir one obtains a stochastic Liouville equation of the form of Eq. (10) for $\mu(\mathbf{r}, \mathbf{v}, t)$ with $H(\mathbf{r}, \mathbf{v}) = \frac{1}{2}|\mathbf{v}|^2 + U(\mathbf{r})$, where $U(\mathbf{r})$ is an external potential which varies slowly on the microscopic spatial scale.

For a sufficiently idealized fluid in thermal equilibrium at temperature β^{-1} one can obtain, in an appropriate limit, a Fokker-Planck equation for the time evolution of the probability density of the BP $\mu(\mathbf{r}, \mathbf{v}, t)$

$$\frac{\partial \mu}{\partial t} + \mathbf{v} \cdot \frac{\partial \mu}{\partial \mathbf{r}} - \frac{\partial U}{\partial \mathbf{r}} \cdot \frac{\partial \mu}{\partial \mathbf{v}} = \xi \frac{\partial}{\partial \mathbf{v}} \cdot \left[\mu_\beta \frac{\partial}{\partial \mathbf{v}} (\mu/\mu_\beta) \right], \tag{25}$$

where $\mu_\beta = Z^{-1} \exp\left\{-\beta \left[\frac{1}{2}|\mathbf{v}|^2 + U(\mathbf{r})\right]\right\}$, c.f. [27]. Thus within the approximate, but physically appropriate, scheme Eq. (25) treats the fluid as an infinite thermal reservoir which exerts a stochastically stationary, delta-time correlated, Gaussian force on the particle. The particle distribution then evolves towards its stationary value μ_β on a time scale T/ξ.

For the Fokker-Planck Eq. (25), one can formally follow the approach of Sect. 5 [16,17] to calculate the corresponding change in the Gibbs-Shannon entropy production. The result is given by

$$\sigma_{\mathrm{G}} = \xi \int \mu(\mathbf{r}, \mathbf{v}, t) \left| \frac{\partial}{\partial \mathbf{v}} \log \nu \right|^2 \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} = \frac{\mathrm{d}S_{\mathrm{G}}}{\mathrm{d}t} - J/T \geq 0, \tag{26}$$

where $\nu = \mu/\mu_\beta$ as in Eq. (18) and J is the average energy flux from the fluid to the Brownian particle $J = \frac{\mathrm{d}}{\mathrm{d}t} \int (\frac{1}{2}|\mathbf{v}|^2 + U)\mu \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v}$. One can take further limits when the Fokker-Planck equation becomes a Langevin equation but we shall not go into that here [30].

Equations (25) and (26) and their analogues play a central role in "stochastic thermodynamics" where $S_{\mathrm{G}}(\mu)$ is generally taken for granted to represent a thermodynamic entropy and thus (26) is considered to be an expression of the second law [28,29]. There are in fact, as already noted, recent experiments which give some support to this interpretation [8]. The question therefore naturally arises of why this should be true for small systems in contact with thermal reservoirs when, as argued above, this is not the case for isolated systems and may not be true for MS in contact with reservoirs at their surfaces.

7 The Brownian Gas

We shall now attempt to justify the identification of $S_{\mathrm{G}}(\mu)$ of a nano-particle in contact with a thermal reservoir, such as a BP in a fluid, with a thermodynamic entropy. Consider a dilute gas of N such BP, $N \gg 1$, and call it a Brownian gas [BG]. The gas is so dilute that interactions between the BP are negligible. This BG is a macroscopic system in contact with a thermal reservoir not just at its boundaries, but "everywhere". Let γ be the 6 dimensional phase space of the Brownian particle (in the older literature γ is called the $\mu-$space, where μ stands for molecule). Then, the phase space Γ of the BG will have $6N$ dimensions.

The ("meso") macrostate of the Brownian gas is given by specifying the number of Brownian particles in each region $\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v}$ of the 6 dimensional γ space, to be

$Nf(\mathbf{r},\mathbf{v})\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v}$. This corresponds to a region Γ_f in the $6N$ dimensional phase space Γ. The log of the Liouville volume $|\Gamma_f|$ is given, up to constants, by [2]

$$\log|\Gamma_f| = -N\int f(\mathbf{r},\mathbf{v})\log f(\mathbf{r},\mathbf{v})\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} + N. \tag{27}$$

The Boltzmann entropy, $S_{\mathrm{B}}(f)$, of this meso state is then given by (27), whose right hand side coincides up to a constant term with $NS_{\mathrm{G}}(\mu)$. This is so even though the physical interpretations of $NS_{\mathrm{G}}(\mu)$ and $S_{\mathrm{B}}(f)$ are quite different.

The entropy production in a Brownian gas plus fluid is given by

$$\sigma_{\mathrm{B}} = -N\frac{\mathrm{d}}{\mathrm{d}t}\int f\log f\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} - J_N/T \geq 0, \tag{28}$$

where J_N is the flux of energy from the fluid to the Brownian gas, and $-J_N/T$ is the rate of entropy change of the fluid at temperature T. The right side of Eq. (28) is just N times the right hand side of Eq. (26) if one identifies $J_N = NJ$.

We remark here that Boltzmann's famous $\mathcal{H}-$theorem shows the monotone increase of $-\int f\log f\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v}$ for an isolated dilute gas evolving in time according to the Boltzmann equation. Boltzmann interpreted this as a microscopic derivation of the second law for $S_{\mathrm{B}}(\{f\})$ and says [31]: "we have thus succeeded in defining entropy for a system not in equilibrium".

An important observation now is that, unlike an isolated gas, where some interaction between the particles is essential to make the system satisfy the second law (rather than behaving like an ideal gas), the Brownian gas gets thermalized via its interaction with the fluid. Hence the behavior of a single Brownian particle averaged over many trials will be the same as that of a Brownian gas. It is therefore meaningful to consider the Gibbs-Shannon entropy of a single Brownian particle as having a thermodynamic meaning, i.e. being equal to that of a Brownian gas divided by the number of particles. This should be true both when the Brownian gas is in global equilibrium, or in a meso (macro) state described by $f(\mathbf{r},\mathbf{v})$.

The above considerations will hold also in the case when the Brownian particle is acted on by a time dependent external potential $U(\mathbf{r},t)$. The Brownian gas will behave like a MS on which work is being done. In particular when $U(\mathbf{r},t)$ varies sufficiently slowly in time compared to the time it takes the Brownian particle to relax to equilibrium with $U(\mathbf{r},t)$, then the entropy will change adiabatically and the right hand side of (26) will be an equality. The behavior of a single Brownian particle will then be similar to that of the Brownian gas, with vanishing fluctuations. This is what is observed in the experiments in [8] which we discuss next.

[2] The derivation of Eq. (27), due to Boltzmann, is straightforward. Divide the $\gamma-$space into regions Δ_α, with $\alpha = 1,\ldots,M$, and let N_α be the number of particles in Δ_α. Then, one has that $|\Gamma_f| \sim \prod \frac{|\Delta_\alpha|^{N_\alpha}}{N_\alpha!}$. Using Stirling's formula, one obtains Eq. (27), see [4] for details.

8 Experiments on a Brownian Particle

An idealized version of the experiment in [8] is as follows: The thermal reservoir fluid occupies a volume V which is divided into regions V_1 and V_2. At $t = 0$ the BP is in equilibrium with the fluid in V_1 and a confining (infinite) external potential $U_0(\mathbf{r})$ which excludes it from V_2. At $t = 0$, $U_0(\mathbf{r})$ is changed to $U_1(\mathbf{r})$ without any work being done, e.g. one suddenly removes the infinite potential confining the particle to V_1. One then waits until time t_1 for the particle to come to equilibrium with the fluid at the new potential $U_1(\mathbf{r})$, e.g. no confining potential. One then changes U_1 to U_0 by gradually raising the height of the potential in V_2 over a time interval τ. During this time one does work $W(\tau)$ on the particle.

We make the time variation of $U(\mathbf{r}, t)$ during τ very slow compared to the relaxation time of the particle to its equilibrium distribution. Hence during the time interval τ of a given realization the probability of the particle being in position $\mathbf{r}$ with velocity $\mathbf{v}$ varies in a quasistatic way. From a thermodynamic point of view the macro (meso) state of the corresponding Brownian gas is given up to a factor N by

$$\mu_\beta(\mathbf{r}, \mathbf{v}, t) = \frac{1}{Z(t)} \exp\left(-\beta\left[\frac{1}{2}|\mathbf{v}|^2 + U(\mathbf{r}, t)\right]\right) \tag{29}$$

with $U(\mathbf{r}, t_1) = U_1(\mathbf{r})$ and $U(\mathbf{r}, t_1 + \tau) = U(\mathbf{r}, 0) = U_0(\mathbf{r})$.

We can now use standard thermodynamics to calculate the work done by an external agent that slowly manipulates the potential $U(\mathbf{r}, t)$. The work done per unit time is $\dot{w}(t) = -\mathbf{v} \cdot \partial U/\partial \mathbf{r}|_{\mathbf{r}(t)}$, where $\mathbf{r}(t)$ is the position of the BP at time t. The total work done over the duration of the period $(0, t_1 + \tau)$ is then different from zero only during the interval $(t_1, t_1 + \tau)$, and is given by

$$W(\tau) = \int_{t_1}^{t_1+\tau} \dot{w}(t)\mathrm{d}t = \int_{t_1}^{t_1+\tau} \frac{\partial U(\mathbf{r}, t)}{\partial t}\mathrm{d}t, \tag{30}$$

where it is important for the equality that the potential is the same at the beginning and at the end of the protocol [32,33]. We can now relate the average total work $\langle W(\tau)\rangle$, where the average is taken with respect to $\mu_\beta(\mathbf{r}, \mathbf{v}, t)$, with the change of $\log Z$ during the period τ. More precisely

$$\begin{aligned}\langle W(\tau)\rangle &= \int \left(\int_{t_1}^{t_1+\tau} \frac{\partial U(\mathbf{r}, t)}{\partial t}\mu_\beta(\mathbf{r}, \mathbf{v}, t)\mathrm{d}t\right)\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} \\ &= -\int \left(\int_{t_1}^{t_1+\tau} \frac{1}{\beta Z(t)}\frac{\partial}{\partial t}\left[\mathrm{e}^{-\beta(\frac{1}{2}|\mathbf{v}|^2 + U(\mathbf{r},t))}\right]\right)\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} \\ &= T\log[Z(t_1)/Z(t_1 + \tau)].\end{aligned} \tag{31}$$

To interpret this work as a change in Gibbs-Shannon entropy we can integrate by parts in the definition of $\langle W(\tau)\rangle$. This gives

$$\langle W(\tau)\rangle = E(t_1 + \tau) - E(t_1) - T[S_{\mathrm{G}}(t_1 + \tau) - S_{\mathrm{G}}(t_1)], \tag{32}$$

where we have defined the average energy as $E(t) = \langle \frac{1}{2}|\mathbf{v}|^2 + U(\mathbf{r}, t)\rangle$. In the actual experiment $E(t_1) = E(0) = E(t_1 + \tau)$. Hence measuring $\langle W(\tau)\rangle$, which was shown experimentally to have very little variance for large τ, lets one measure $S_{\mathrm{G}}(\mu_\beta(t_1)) - S_{\mathrm{G}}(\mu_\beta(t_1 + \tau))$ for different external potentials $U(\mathbf{r}, t)$, e.g. for different confining volumes of $U_0(\mathbf{r})$.

Using the BG interpretation, the quantities in Eq. (32) can all be interpreted as macroscopic quantities divided by the number of particles. Within this interpretation S_{G} coincides with the hydrodynamic entropy which changes in an irreversible way. The entropy of the heat bath (here the fluid) has increased during the cycle from 0 to $t_1 + \tau$ by $\langle W(\tau)\rangle/T$. When τ is not so large so that one can not assume instantaneous equilibrium of the BG there will be extra entropy production in the BG during this period. This work, obtained by averaging $W(\tau)$ over repetitions of the experiment, was indeed found to be greater than the right hand side of (32). In the analysis of the experiment and in stochastic thermodynamics one goes beyond the simple equality or inequality of (32). One actually computes the distribution of $W(\tau)$. We shall not go into that here. We thus conclude that the interpretation of $S_{\mathrm{G}}(\mu)$ as the thermodynamic Boltzmann entropy per particle of a Brownian gas is consistent.

9 Concluding Remarks

The point of view taken in this note is that the entropy of a physical system should be a property of the state of the *individual* system and thus it should be possible to define the entropy of a system without referring to any ensembles, see [3,4]. For a classical system the most detailed description of the physical state of the system is that given by its microstate $X \in \Gamma$. Thus any physical entropy S is a function of X. This is the case for the Boltzmann entropy $S_{\mathrm{B}}(X)$ of a macroscopic system in a well defined macrostate M (for which S_{B} is in fact the same for all $X \in \Gamma(M)$). Going beyond the hydrodynamic entropy (3), appropriate only for systems in local thermal equilibrium, S_{B} can be extended to dilute gases not in local thermal equilibrium. For these the macro (meso) state M is specified by the empirical distribution $f(\mathbf{r}, \mathbf{v}, t)$, which is of course determined by X, see [4]. In contrast the Gibbs-Shannon entropy (of a measure) not only fails to be determined by the microstate of the system – it also fails to change in time for an isolated system, large or small, even for a large isolated system that is undergoing (internally) dissipative relaxation and thus producing thermodynamic entropy. Of course the Gibbs-Shannon entropy of the microcanonical ensemble μ_{m} is meaningful for an isolated macroscopic system in global thermal equilibrium, where it coincides with S_{B} to leading order in the size of the system.

The existence of a useful general notion of entropy for an isolated nanosystem is not so clear. Such systems have been studied theoretically and experimentally for quantum systems [10,34], which we have not discussed here. In fact our main concern here has been with the significance of S_{G} for systems large and small in contact with heat baths. We have not resolved this issue for macroscopic systems,

but have given a possible, to us plausible, answer for the case of a nanosystem studied experimentally in [8].

Acknowledgements. We thank John Bechhoefer, Rafaël Chetrite, Stanislas Leibler, Eugene Speer and Bingkan Xue for fruitful discussions. The work of JLL was supported by an AFOSR grant FA9550-16-1-0037. The work of PS has been partly supported by grants from the Simons Foundation to Stanislas Leibler through The Rockefeller University (Grant 345430) and the Institute for Advanced Study (Grant 345801). DAH, JLL, and PS thank the Institute for Advanced Study for its hospitality during the elaboration of this work.

References

1. Klein, M.J.: The development of Boltzmann's statistical ideas. In: Cohen, E.G.D., Thirring, W. (eds.) The Boltzmann Equation. Theory and Application, pp. 53–106. Springer, Berlin (1973)
2. Penrose, O.: Foundations of Statistical Mechanics: A Deductive Treatment. Courier Corporation, North Chelmsford (2005)
3. Lebowitz, J.L.: From time-symmetric microscopic dynamics to time-asymmetric macroscopic behavior: an overview. In: Boltzmann's Legacy, pp. 63–88 (2007)
4. Goldstein, S., Lebowitz, J.L.: On the (Boltzmann) entropy of non-equilibrium systems. Physica D **193**(1), 53–66 (2004)
5. Kipnis, C., Landim, C.: Scaling Limits of Interacting Particle Systems, vol. 320. Springer, Heidelberg (2013)
6. Giacomin, G., Lebowitz, J.L., Presutti, E.: Deterministic and stochastic hydrodynamic equations arising from simple microscopic model systems. Math. Surv. Monogr. **64**, 107–152 (1998)
7. Lebowitz, J.L., Presutti, E., Spohn, H.: Microscopic models of hydrodynamic behavior. J. Stat. Phys. **51**(5), 841–862 (1988)
8. Gavrilov, M., Chétrite, R., Bechhoefer, J.: Direct measurement of weakly nonequilibrium system entropy is consistent with Gibbs–Shannon form. Proc. Nat. Acad. Sci. **114**(42), 11097–11102 (2017)
9. Ciliberto, S.: Experiments in stochastic thermodynamics: short history and perspectives. Phys. Rev. X **7**(2), 021051 (2017)
10. Kaufman, A.M., Tai, M.E., Lukin, A., Rispoli, M., Schittko, R., Preiss, P.M., Greiner, M.: Quantum thermalization through entanglement in an isolated many-body system. Science **353**(6301), 794–800 (2016)
11. Brush, S.G.: Science and Culture in the Nineteenth Century: Thermodynamics and History. University of Texas, Texas (1967)
12. Callen, H.B.: Thermodynamics and an Introduction to Thermostatistics. Wiley, New York (1998)
13. De Groot, S.R., Mazur, P.: Non-Equilibrium Thermodynamics. Courier Corporation, North Chelmsford (2013)
14. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
15. Ehrenfest, P., Ehrenfest, T.: The Conceptual Foundations of the Statistical Approach in Mechanics. Courier Corporation, North Chelmsford (2002)
16. Bergmann, P.G., Lebowitz, J.L.: New approach to nonequilibrium processes. Phys. Rev. **99**(2), 578 (1955)

17. Lebowitz, J.L., Bergmann, P.G.: Irreversible Gibbsian ensembles. Ann. Phys. **1**(1), 1–23 (1957)
18. Doob, J.L.: Stochastic Processes. Wiley, New York (1953)
19. Derrida, B., Lebowitz, J., Speer, E.: Entropy of open lattice systems. J. Stat. Phys. **126**(4–5), 1083–1108 (2007)
20. Bonetto, F., Lebowitz, J.L., Lukkarinen, J.: Fourier's law for a harmonic crystal with self-consistent stochastic reservoirs. J. Stat. Phys. **116**(1), 783–813 (2004)
21. Kosygina, E.: The behavior of the specific entropy in the hydrodynamic scaling limit. Ann. Probab. **29**(3), 1086–1110 (2001)
22. Chernov, N., Lebowitz, J.L.: Stationary nonequilibrium states in boundary-driven Hamiltonian systems: shear flow. J. Stat. Phys. **86**(5), 953–990 (1997)
23. Lebowitz, J.L., Spohn, H.: A Gallavotti–Cohen-type symmetry in the large deviation functional for stochastic dynamics. J. Stat. Phys. **95**(1), 333–365 (1999)
24. Maxwell, J.C.: Theory of Heat, p. 308: Tait's Thermodynamics. Nature 17, 257 (1878). Quoted in M. J. Klein "The development of Boltzmann's statistical ideas". See ref. [1]
25. Jarzynski, C.: Stochastic and macroscopic thermodynamics of strongly coupled systems. Phys. Rev. X **7**(1), 011008 (2017)
26. Lebowitz, J.L., Pastur, L.: On the equilibrium state of a small system with random matrix coupling to its environment. J. Phys. A Math. Theor. **48**(26), 265201 (2015)
27. Dürr, D., Goldstein, S., Lebowitz, J.: A mechanical model of Brownian motion. Commun. Math. Phys. **78**(4), 507–530 (1981)
28. Seifert, U.: Stochastic thermodynamics, fluctuation theorems and molecular machines. Rep. Prog. Phys. **75**(12), 126001 (2012)
29. Baiesi, M., Maes, C., Wynants, B.: Fluctuations and response of nonequilibrium states. Phys. Rev. Lett. **103**(1), 010602 (2009)
30. Risken, H.: Fokker-Planck equation. In: The Fokker-Planck Equation, pp. 63–95. Springer (1996)
31. Boltzmann, L.: Vorlesungen über Gastheorie: 2. Teil, Leipzig: Barth, 1896, 1898. This book has been translated into English by Brush, S.G. Lectures on Gas Theory, Cambridge University Press, London (1964)
32. Vilar, J.M., Rubi, J.M.: Failure of the work-Hamiltonian connection for free-energy calculations. Phys. Rev. Lett. **100**(2), 020601 (2008)
33. Peliti, L.: On the work-Hamiltonian connection in manipulated systems. J. Stat. Mech. Theory Exp. **2008**(05), P05002 (2008)
34. Goldstein, S., Huse, D.A., Lebowitz, J.L., Tumulka, R.: Thermal equilibrium of a macroscopic quantum system in a pure state. Phys. Rev. Lett. **115**(10), 100402 (2015)

Marginal Relevance for the γ-Stable Pinning Model

Hubert Lacoin(✉)

IMPA, Institudo de Matemática Pura e Aplicada,
Estrada Dona Castorina 110, Rio de Janeiro 22460-320, Brazil
lacoin@impa.br

Abstract. We investigate disorder relevance for the pinning of a renewal when the law of the random environment is in the domain of attraction of a stable law with parameter $\gamma \in (1,2)$. Assuming that the renewal jumps have power-law decay, we determine under which condition the critical point of the system is altered by the introduction of a small quantity of disorder. In an earlier study of the problem [20] we have shown that the answer depends on the value of the tail exponent α associated to the distribution of renewal jumps: when $\alpha > 1-\gamma^{-1}$ a small amount of disorder shifts the critical point whereas it does not when $\alpha < 1-\gamma^{-1}$. The present paper is focused on the boundary case $\alpha = 1-\gamma^{-1}$. We show that in this case, the critical point is shifted, and obtain an estimate for the intensity of this shift.

Keywords: Pinning model · Disorder relevance · Stable laws · Harris criterion

1 Introduction

The renewal pinning model has been developed as a toy model to understand phenomena like wetting in two dimensions [1] and pinning of a polymer to a defect line [11]. Due to its simplicity and the fact that the critical exponent associated to the localization transition can be tuned to any value just by modifying one parameter (the tail exponent of the renewal process in (2.1)), it has also been employed as benchmark to test prediction concerning the effect of disorder obtained by non-rigorous renormalization group arguments. We refer to the monographs [13,14] for a complete introduction to the subject.

More precisely a rich literature has been developed (see [2,3,5–7,17,19,21] and references therein) to establish rigorously that the sensibility of the system to disorder is determined by the sign of the critical exponent associated to the specific heat as predicted by Harris [18]. More precisely it was shown that when the specific-heat exponent is positive (which corresponds to $\alpha > 1/2$ for the exponent in (2.1)) disorder even of small intensity shifts the critical point and modifies the critical exponent, while when it is negative ($\alpha < 1/2$) the critical point and the critical exponent of the localization transition are conserved.

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 597–616, 2019.
https://doi.org/10.1007/978-3-030-15096-9_23

The criterion developed by Harris does not yield any prediction when the specific heat exponent vanishes: this corresponds to a tail exponent $\alpha = 1/2$ for the renewal process. This case is of special importance in the case of pinning as it corresponds to the original random walk pinning model (see e.g. [10]). A more detailed renormalization group analysis in [7] yielded that in this so-called marginal case, disorder should also be relevant (a prediction conflicting with others made in the literature e.g. [12], see the introduction of [15] for a more detailed account on the controversy). This conjecture was proved in [15] (see also [4,16]).

As most heuristics concerning disorder relevance rely on second moment expansion, a natural question is:

"Is Harris criterion valid when the disorder has infinite variance?"

The issue was raised for pinning model in [20] and it was shown that when the disorder is in the domain of attraction of a γ-stable law with $\gamma \in (1,2)$, Harris criterion is not satisfied. More precisely we showed that the critical point is shifted when $\alpha > 1 - \gamma^{-1}$ and that critical points and exponents are not perturbed by a small amount of disorder when $\alpha < 1 - \gamma^{-1}$.

In the present work we investigate the marginal case $\alpha = 1 - \gamma^{-1}$ for which we prove disorder relevance. It presents strong analogies with the Random Walk pinning model treated in [7,15]. While the methods used to resolve it are clearly inspired by those used in the marginal case with second moment [4,15,16], they also incorporate new ingredients which are necessary to deal with heavier-tail disorder.

2 Model and Results

2.1 Disordered Pinning and Phase Transition

Consider $\tau = (\tau_n)_{n \geqslant 0}$ a recurrent integer valued renewal process, that is a random sequence starting from $\tau_0 = 0$ whose increments $(\tau_{n+1} - \tau_n)$ are independent, identically distributed (IID) positive integers. We let $\mathbf{P}$ denote the associated probability distribution and assume that the inter-arrival distribution has power-law decay or more precisely

$$K(n) := \mathbf{P}[\tau_1 = n] \overset{n\to\infty}{\sim} C_K n^{-(1+\alpha)}, \quad \alpha \in (0,1), \tag{2.1}$$

where $C_K > 0$ is an arbitrary constant. Note that τ can alternatively be considered as an infinite subset of $\mathbb{N}$ and in our notation $\{n \in \tau\}$ is equivalent to $\{\exists k \in \mathbb{N},\ \tau_k = n\}$.

We consider a sequence of IID random variables $(\omega_n)_{n\geq 0}$, with law denoted by $\mathbb{P}$, which satisfies $\mathbb{E}[\omega_1] = 0$ and for some $a \in (0,1)$

$$\mathbb{P}[\omega_1 \geq -a] = 1. \tag{2.2}$$

We work under the assumption that ω is in the domain of attraction of a γ-stable law, or more precisely we assume that for some $C_{\mathbb{P}} > 0$ we have

$$\mathbb{P}[\omega_n \geq x] \overset{x\to\infty}{\sim} C_{\mathbb{P}} x^{-\gamma}, \quad \gamma \in (1,2). \tag{2.3}$$

Given $\beta \in [0,1]$, $h \in \mathbb{R}$, and $N \in \mathbb{N}$, we define a modified renewal measure $\mathbf{P}_{N,h}^{\beta,\omega}$ whose Radon-Nikodym derivative with respect to $\mathbf{P}$ is given by

$$\frac{\mathrm{d}\mathbf{P}_{N,h}^{\beta,\omega}}{\mathrm{d}\mathbf{P}}(\tau) = \frac{1}{Z_{N,h}^{\beta,\omega}} \left(\prod_{n\in[1,N]\cap\tau} e^h(\beta\omega_n + 1) \right) \mathbf{1}_{\{N\in\tau\}}, \tag{2.4}$$

where

$$Z_{N,h}^{\beta,\omega} = \mathbf{E}\left[\left(\prod_{n\in[1,N]\cap\tau} e^h(\beta\omega_n + 1) \right) \mathbf{1}_{\{N\in\tau\}} \right]. \tag{2.5}$$

In the case $\beta = 0$, we retrieve the homogeneous pinning model which, setting $\delta_n := \mathbf{1}_{\{n\in\tau\}}$, is defined by

$$\frac{\mathrm{d}\mathbf{P}_{N,h}}{\mathrm{d}\mathbf{P}}(\tau) := \frac{1}{Z_{N,h}} e^{h\sum_{n=1}^N \delta_n} \delta_N \quad \text{and} \quad Z_{N,h} := \mathbf{E}\left[e^{h\sum_{n=1}^N \delta_n} \delta_N \right]. \tag{2.6}$$

We investigate the behavior of τ under $\mathbf{P}_{N,h}^{\beta,\omega}$ using the notion of *free energy per monomer*, which is defined as the asymptotic growth rate of the partition function

$$\mathrm{F}(\beta,h) := \lim_{N\to\infty} \frac{1}{N} \log Z_{N,h}^{\beta,\omega} \overset{\mathbb{P}-a.s.}{=} \lim_{N\to\infty} \tfrac{1}{N}\mathbb{E}\left[\log Z_{N,h}^{\beta,\omega}\right] < \infty. \tag{2.7}$$

We refer to [13, Theorem 4.1] for a proof of existence of $\mathrm{F}(\beta,h)$. Note that $\mathrm{F}(\beta,h)$ is non-negative, and that $h \mapsto \mathrm{F}(\beta,h)$ is non-decreasing and convex (as a limit of non-decreasing convex functions). By exchanging limit and derivative, as allowed by convexity, we obtain that the derivative of F w.r.t. h corresponds to the asymptotic contact fraction

$$\partial_h \mathrm{F}(\beta,h) := \lim_{N\to\infty} \frac{1}{N} \mathbf{E}_{N,h}^{\beta,\omega}\left[|\tau \cap [1,N]|\right]. \tag{2.8}$$

In particular, if one sets

$$h_c(\beta) := \inf\{h \in \mathbb{R} \ : \ \mathrm{F}(\beta,h) > 0\}, \tag{2.9}$$

we have

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \mathbf{E}_{N,h}^{\beta,\omega}\left[|\tau \cap [1,N]|\right] &= 0 \quad \text{if } h < h_c(\beta), \\ \liminf_{N\to\infty} \frac{1}{N} \mathbf{E}_{N,h}^{\beta,\omega}\left[|\tau \cap [1,N]|\right] &> 0 \quad \text{if } h > h_c(\beta). \end{aligned} \tag{2.10}$$

We say in the first case that τ is delocalized and in the second one that it is localized. It can be proved using simple inequalities (see below or [13, Proposition 5.1]), that $h_c(\beta) \notin \{-\infty, \infty\}$ meaning that this phase transition really occurs.

2.2 Annealed Comparison and Disorder Relevance

Using Jensen's inequality and the assumption that the ωs have zero mean we have

$$\mathbb{E}\left[\log Z_{N,h}^{\beta,\omega}\right] \le \log \mathbb{E}\left[Z_{N,h}^{\beta,\omega}\right] = \log Z_{N,h}. \tag{2.11}$$

Hence

$$\forall \beta \in (0,1], \quad \mathrm{F}(\beta,h) \le \mathrm{F}(h), \tag{2.12}$$

Our assumption (2.2) also implies that $\mathrm{F}(\beta,h) \ge \mathrm{F}(h+\log(1-a\beta))$.

The localization transition is easier to analyze when $\beta = 0$, and this makes the inequality (2.12) more interesting: $\mathrm{F}(h)$ is the solution of an explicit inverse problem

$$\mathrm{F}(h) = \begin{cases} 0 & \text{if } h \le 0, \\ g^{-1}(h) & \text{if } h > 0, \end{cases} \tag{2.13}$$

where g is defined on $\mathbb{R}_+$ by

$$g(x) := -\log\left(\sum_{n=1}^{\infty} e^{-nx} K(n)\right).$$

In particular we have $h_c(0) = 0$ and from a closer analysis of g (see [13, Theorem 2.1]) we obtain

$$\mathrm{F}(h) \overset{h\to 0+}{\sim} \left(\frac{\alpha h}{C_K \Gamma(1-\alpha)}\right)^{\frac{1}{\alpha}}.$$

A natural question is to ask whether the annealed comparison (2.12) is sharp, in the following sense:

(A) Is the critical value of h preserved when disorder is introduced:
Do we have $h_c(\beta) = 0$?
(B) Is the critical exponent for the phase transition preserved:
Do we have $\mathrm{F}(h,\beta) \approx h^{1/\alpha}$ in some sense?

If these two property hold, it means that the introduction of disorder in the system does not change its property and this situation is referred to as *irrelevant disorder*. In the case where the critical properties of the system are changed disorder is said to be *relevant*.

2.3 Harris Criterion and Former Results

Harris [18] developed a criterion in order to predict disorder relevance. For one dimensional systems such as the one studied in the present paper, it can be interpreted as follows: If the critical exponent for free-energy of the pure (i.e. $\beta = 0$) model is larger than 2 then disorder is irrelevant for small values of β, whereas disorder is always relevant in the case when the exponent is larger than 2. In the case of pinning model, this means that disorder is irrelevant for $\alpha < 1/2$ and relevant for $\alpha > 1/2$.

The validity of the Harris criterion has been confirmed in various cases for the pinning model, in the case where the environment has finite second moment $\mathbb{E}[\omega_1^2] < \infty$ (see [2,19,21] for the irrelevant disorder case, and [3,6,17] in the relevant case). This assumption is far from being only technical as Harris heuristics is based on a second moment expansion at the vicinity of the critical point in order to test stability.

For this reason we suspected that with an environment with an heavier tail distribution, Harris criterion may not be valid. This has been confirmed in [20] where we have shown that disorder is irrelevant when $\alpha < 1 - \gamma^{-1}$ and relevant for $\alpha > 1 - \gamma^{-1}$.

Theorem A (From Theorems 2.3 and 2.4 in [20]).

(A) If $\alpha < 1 - \gamma^{-1}$, then there exists β_0 such that for all $\beta \in (0, \beta_0]$ we have $h_c(\beta) = 0$ and furthermore

$$\lim_{h \to 0+} \frac{\log \mathrm{F}(\beta, h)}{\log(h)} = \frac{1}{\alpha}. \tag{2.14}$$

(B) If $\alpha > 1 - \gamma^{-1}$, then for all β we have $h_c(\beta) > 0$ and

$$\lim_{\beta \to 0+} \frac{\log h_c(\beta)}{\log \beta} = \frac{\alpha\gamma}{1 - \gamma(1-\alpha)}. \tag{2.15}$$

These results indicate that Harris criterion has to be reinterpreted in the case where the environment is heavy-tailed. A question which has been left open in [20] is the case $\alpha = 1 - \gamma^{-1}$ which we refer to as the *marginal case*.

2.4 Main Result

The main achievement of this paper is to prove that disorder shifts the critical point for all values of β also in the marginal case $\alpha = 1 - \gamma^{-1}$. The result bears some similarity with the one proved in [15], when it is shown that under finite second moment assumption for ω ([15] actually only treats the case of Gaussian environment but the generalization can be found in [16]), disorder is relevant when the renewal exponent satisfies $\alpha = 1/2$.

Theorem 1. *Assume that (2.1) and (2.3) are satisfied for $\alpha = 1 - \gamma^{-1}$. Then, for any $\beta \in [0, 1]$, $h_c(\beta) > 0$ and furthermore, there exists a constant $A > 0$ such that*

$$\forall \beta \in (0, 1], \quad h_c(\beta) \geq \exp(-A\beta^{-2\gamma}). \tag{2.16}$$

Remark 2. *We are discussing in this paper only the case where the inter-arrival distribution $K(\cdot)$ has a pure power-law behavior, cf. (2.1). When a slowly varying function is introduced instead of the constant C_K, the picture gets slightly more complicated and a necessary and sufficient condition for disorder relevance was proved in [4] under the finite second moment assumption. For γ-stable environment we refer to [20, Section 2.5.1] for a conjecture.*

Remark 3. *We do not believe that this lower bound on $h_c(\beta)$ is optimal but we know from [20, Proposition 6.1] that $h_c(\beta)$ is smaller than any power of β at the vicinity of 0. This contrasts with the case $\alpha > 1 - \gamma^{-1}$, cf. (2.15). It seems plausible that improving the technique presented in the present paper in the same spirit as what is done in [4], we can bring the exponent in the exponential in (2.16) from 2γ down to γ, which could be the optimal answer. We would not know however how to obtain a matching upper bound.*

2.5 Organization of the Paper

The proof of our main statement is divided into three main steps: In Sect. 3 we present a sequence of inequalities which combines coarse graining ideas (in a very similar spirit with what has been done e.g. in [4,15]) and a change of measure which penalizes environment which displays atypical "dual peaks". This reduces the problem to estimating the coarse-grained partition function under the penalized measure (Proposition 4), provided we control the "cost" of the penalization procedure (Proposition 3). In Sect. 4, Proposition 3 is proved while Proposition 4 is reduced to a one block estimate (Proposition 5), which is itself proved in Sect. 5.

3 Fractional Moments, Coarse Graining and Change of Measure

3.1 Fractional Moments

Let us consider $\theta \in (0,1)$. A more efficient bound than (2.11) can be achieved on the free-energy by applying Jensen's inequality in a different manner.

$$\mathbb{E}\left[\log Z_{N,h}^{\beta,\omega}\right] = \frac{1}{\theta}\mathbb{E}\left[\log\left(Z_{N,h}^{\beta,\omega}\right)^{\theta}\right] \le \frac{1}{\theta}\log\mathbb{E}\left[\left(Z_{N,h}^{\beta,\omega}\right)^{\theta}\right]. \tag{3.1}$$

In particular we can prove that $\mathrm{F}(\beta,h) = 0$ if we have

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{E}\left[\left(Z_{N,h}^{\beta,\omega}\right)^{\theta}\right] = 0. \tag{3.2}$$

We set

$$h_\beta := \exp\left(-A\beta^{-2\gamma}\right), \tag{3.3}$$

where $A > 0$ is a sufficiently large constant, and consider a special length

$$\ell_\beta := h_\beta^{-1}. \tag{3.4}$$

In the remainder of the paper, we often drop the superscript β when referring to ℓ for readability's sake. We consider a system whose size $N = m\ell$ is an integer multiple of ℓ. In view of (3.2) the proof of Theorem 1 can be reduced to that of the following statement.

Proposition 1. *Given $\theta \in \left(\frac{\gamma}{2\gamma-1}, 1\right)$, if A is chosen sufficiently large, we have for all $\beta \in (0,1]$*

$$\limsup_{m\to\infty} \mathbb{E}\left[\left(Z^{\beta,\omega}_{m\ell_\beta,h_\beta}\right)^\theta\right] < \infty. \tag{3.5}$$

In particular we have

$$h_c(\beta) \geq h_\beta.$$

The proof of the proposition goes in two steps: Firstly, we use a kind of bootstrapping argument in order to reduce the problem to estimates of partition functions of systems of size smaller or equal to ℓ. Secondly, to control the partition function of these smaller system we introduce a change of measure procedure which has the effect of penalizing some atypical environments whose contribution to the annealed partition function is significant.

The approach adopted in [20] to prove disorder relevance when $\alpha > 1 - \gamma^{-1}$ used a finite volume criterion from [6], and penalized environments for which uncommonly large values of ω appeared. This approach fails to give any result in the present case and we need to perform a finer analysis to catch the critical point shift.

We introduce two improvements with respect to the method used in [20]: The first is to replace [6, Proposition 2.5] by a finer coarse graining. This is not a new idea and is very similar to the method applied e.g. in [22]. The second improvement is the main novelty of this paper and concerns the type of penalization considered in the change of measure procedure: we design a new form of penalization which involves considering pairs of site where ω displays high values.

This approach contrasts with what has been done in the marginal case under finite second moment assumption: In the case of Gaussian environment, a penalization that would induce a change of the covariance structure was considered [15], and more generally for an environment with finite second moment a tilting by a quadratic form, or a multi-linear form of higher order [4,16] was used in order to prove marginal disorder relevance. Under assumption (2.3) quadratic forms in ω seems trickier to analyze and we need to select another function of ω which is easier to manipulate. We decide to look only for extremal values in ω and to penalize environments which present two high-peaks close to each other. The exact threshold that we use is determined by a function of the distance between the two sites.

3.2 The Coarse Graining Procedure

For the sake of completeness let us repeat in full details the coarse graining procedure from [4]. We split the system into blocks of size ℓ, we define for $i \in [\![1, m]\!]$

$$B_i := [\![\ell(i-1)+1, \ell i]\!]. \tag{3.6}$$

Given $\mathcal{I} = \{i_1, \ldots, i_{|\mathcal{I}|}\} \subset [\![1, m]\!]$ we define the event

$$E_{\mathcal{I}} := \left\{ \left\{ i \in [\![1,m]\!] \ : \ \tau \cap B_i \neq \emptyset \right\} = \mathcal{I} \right\}, \tag{3.7}$$

and set $Z^{\mathcal{I}}$ to be the contribution to the partition function of the event $E_{\mathcal{I}}$,

$$Z^{\mathcal{I}} := Z^{\beta,\omega}_{N,h}(E_{\mathcal{I}}) = Z^{\beta,\omega}_{N,h}\mathbf{E}^{\beta,\omega}_{N,h}[E_{\mathcal{I}}]. \tag{3.8}$$

Note that $Z^{\mathcal{I}} > 0$ if and only if $m \in \mathcal{I}$. When $\tau \in E_{\mathcal{I}}$, the set $\mathcal{I}$ is called the coarse-grained trajectory of τ. As the $E_{\mathcal{I}}$ are mutually disjoint events, $Z^{\beta,\omega}_{N,h} = \sum_{\mathcal{I}\subset[\![1,m]\!]} Z^{\mathcal{I}}$ and thus using the inequality $(\sum a_i)^\theta \leqslant \sum a_i^\theta$ for non-negative a_i's, we obtain

$$\mathbb{E}\left[\left(Z^{\beta,\omega}_{N,h}\right)^\theta\right] \leq \sum_{\mathcal{I}\subset\{1,\dots,m\}} \mathbb{E}\left[\left(Z^{\mathcal{I}}\right)^\theta\right]. \tag{3.9}$$

We therefore reduced the proof to that of an upper bound on $\mathbb{E}\left[\left(Z^{\mathcal{I}}\right)^\theta\right]$, which can be interpreted as the contribution of the coarse grained trajectory $\mathcal{I}$ to the fractional moment of the partition function.

Proposition 2. *Given $\eta > 0$, and $\theta \in (0,1)$, if A is sufficiently large then for all $\beta \in (0,1]$ there exists a constant C_β such that for all $m \geq 1$ and $\mathcal{I} \subset [\![1,m]\!]$*

$$\mathbb{E}\left[\left(Z^{\mathcal{I}}\right)^\theta\right] \leq C_\ell \prod_{k=1}^{|\mathcal{I}|} \frac{\eta}{(i_k - i_{k-1})^{(1+\alpha)\theta}}, \tag{3.10}$$

where by convention we have set $i_0 := 0$.

Proof (Proof of Proposition 1 from Proposition 2). Note that by Jensen's inequality we just have to prove the statement for θ close to one. We choose $\theta < 1$ which satisfies

$$(1+\alpha)\theta > 1.$$

Using (3.9) and Proposition 2 for all $m > 0$ we have

$$\mathbb{E}\left[\left(Z^{\beta,\omega}_{N,h}\right)^\theta\right] \leq C_\beta \sum_{\substack{\mathcal{I}\subset[\![1,m]\!]\\ m\in\mathcal{I}}} \prod_{k=1}^{|\mathcal{I}|} \frac{\eta}{(i_k - i_{k-1})^{(1+\alpha)\theta}}. \tag{3.11}$$

By considering the sum over all finite subsets of $\mathbb{N}$ with cardinal at most m instead of subsets of $[\![1,m]\!]$ and reorganizing the sum we obtain that

$$\mathbb{E}\left[\left(Z^{\beta,\omega}_{N,h}\right)^\theta\right] \leq C_\beta \sum_{j=1}^{m} \left(\eta \sum_{n\geq 1} n^{-(1+\alpha)\theta}\right)^j. \tag{3.12}$$

We can check that choosing

$$\eta = \left(2\sum_{n=1}^{\infty} n^{-(1+\alpha)\theta}\right)^{-1}, \tag{3.13}$$

the l.h.s. of (3.12) is smaller C_ℓ.

3.3 Penalization of Favorable Environments

Let us now introduce the notion of penalization of the environment in a cell, which is the main tool to prove Proposition 2.
Given $G_{\mathcal{I}}(\omega)$ a positive function of $(\omega_n)_{n\in\bigcup_{i\in\mathcal{I}} B_i}$, using Hölder's inequality, we have

$$\mathbb{E}\left[\left(Z^{\mathcal{I}}\right)^{\theta}\right] \leqslant \left(\mathbb{E}\left[G_{\mathcal{I}}(\omega)^{-\frac{\theta}{1-\theta}}\right]\right)^{1-\theta} \left(\mathbb{E}\left[G_{\mathcal{I}}(\omega)Z^{\mathcal{I}}\right]\right)^{\theta}. \tag{3.14}$$

We decide to apply this inequality with $G_{\mathcal{I}}$, which is a product of functions of $(\omega_n)_{n\in B_i}$, for $i \in \mathcal{I}$. More precisely, given $g : \mathbb{R}^{\ell} \to \mathbb{R}^d$ we set

$$G_{\mathcal{I}}(\omega) := \prod_{i\in\mathcal{I}} g(\omega_{i(\ell-1)+1}, \ldots, \omega_{i\ell}) =: \prod_{i\in\mathcal{I}} g_i(\omega). \tag{3.15}$$

We decide to use (3.14) for some g that takes values in $[0, 1]$, which should be equal to 1 for "typical environments" but close to zero for environments that gives too much contribution to $Z^{\mathcal{I}}$. The difficulty lies in finding a function g such that the cost for introducing the penalization $\mathbb{E}\left[G_{\mathcal{I}}(\omega)^{-\frac{\theta}{1-\theta}}\right]$ is not too big, and such that $\mathbb{E}\left[G_{\mathcal{I}}(\omega)Z^{\mathcal{I}}\right]$ is much smaller than $\mathbb{E}\left[Z^{\mathcal{I}}\right]$ (so that we get a large benefit out of it).

Let us now introduce our choice for the function g. Instead of giving a fixed penalty (i.e multiplication by some factor smaller than 1) for each ω_n above a certain threshold (something of order $\ell^{\gamma^{-1}}$) like in [20], which would not give any conclusive result in the case presently studied, we decide to introduce a g that penalizes the presence of "dual peaks in the environment" $(\omega_n)_{n\in[\![1,\ell]\!]}$. Given M a large constant, we set

$$g(\omega_1, \ldots, \omega_\ell) := \exp\left(-M\mathbf{1}_{\mathcal{A}_\ell}\right), \tag{3.16}$$

where

$$\mathcal{A}_\ell := \left\{\exists i, j \in [\![1, \ell]\!],\ i \neq j,\ \min(\omega_i, \omega_j) \geq V(M, \ell, i - j)\right\}, \tag{3.17}$$

and

$$V(M, \ell, n) = V(n) := e^{M^2}(\ell(\log \ell)n)^{\frac{1}{2\gamma}}. \tag{3.18}$$

We are going to prove that with this choice for g, the benefits of the penalization overcome the cost. This is the object of the two following results, whose proofs are postponed to the next section.

Proposition 3. *Given θ, if $M > M_0(\theta)$ is sufficiently large then we have (for all ℓ, m and $\mathcal{I}$)*

$$\mathbb{E}\left[G_{\mathcal{I}}(\omega)^{-\frac{\theta}{1-\theta}}\right] \leq 2^{|\mathcal{I}|}. \tag{3.19}$$

Proposition 4. *Given $\eta > 0$, and M, there exists A such that for all $\beta \in (0, 1]$ we have (for all ℓ, m and $\mathcal{I}$)*

$$\mathbb{E}\left[G_{\mathcal{I}}(\omega)Z^{\mathcal{I}}\right] \leq \frac{C_\beta \eta^{|\mathcal{I}|}}{\prod_{j=1}^{|\mathcal{I}|} |i_j - i_{j-1}|^{1+\alpha}}. \tag{3.20}$$

It is quite straightforward using (3.14) to check that Proposition 2 is a consequence of Propositions 3 and 4 with adequate changes for the value of η and C_ℓ.

4 The Costs and Benefits of the Penalization Procedure

4.1 The Proof of Proposition 3

Using the fact that the environment is IID and the block structure of $G_{\mathcal{I}}$, it is sufficient to show that

$$\mathbb{E}\left[g(\omega_1,\dots,\omega_\ell)^{-\frac{\theta}{1-\theta}}\right] \le 2. \tag{4.1}$$

We have

$$\mathbb{E}\left[g(\omega_1,\dots,\omega_\ell)^{-\frac{\theta}{1-\theta}}\right] \le 1 + e^{\frac{M\theta}{1-\theta}}\mathbb{P}[\mathcal{A}_\ell]. \tag{4.2}$$

Using a union bound and the tails distribution of the ω (2.3) (recall also (3.18)) the probability above can be bounded as follows

$$\begin{aligned}\mathbb{P}\left[\mathcal{A}_\ell\right] &\le \sum_{1\le i<j\le \ell} \mathbb{P}\left[\min(\omega_i,\omega_j)\ge V(j-i)\right] \\ &\le C\frac{e^{-2\gamma M^2}}{\ell(\log \ell)} \sum_{1\le i<j\le \ell} \frac{1}{(j-i)} \le C' e^{-2\gamma M^2}.\end{aligned} \tag{4.3}$$

Hence if M is sufficiently large, the second term in (4.2) is sufficiently small and we can conclude.

□

4.2 The Proof of Proposition 4

For any couple of integers $a < b$ we define

$$Z^h_{[a,b]} := \mathbf{E}\left[\prod_{n\in\tau\cap[a,b]} e^h(1+\beta\omega_n) \;\middle|\; a,b\in\tau\right]. \tag{4.4}$$

We have

$$Z^h_{[a,b]} = \frac{\mathbf{E}\left[\prod_{n\in\tau\cap[0,b-a]} e^h(1+\beta\omega_{a+n})\mathbf{1}_{\{b-a\in\tau\}}\right]}{u(b-a)}, \tag{4.5}$$

where $u(n) := \mathbf{P}[n\in\tau]$. Let us mention an asymptotic equivalent of $u(n)$ [8, Theorem 1], which holds under assumption 2.1 and is used in the rest of the proof

$$u(n) \overset{n\to\infty}{\sim} \frac{\alpha\sin(\pi\alpha)}{\pi C_K}(n+1)^{\alpha-1}. \tag{4.6}$$

The main tool to prove Proposition 4 is the following result which quantifies how the multiplication by g affects the expected value of the partition functions in a single block. Its proof is detailed in the next section.

Proposition 5. *Given $\eta \in (0,1)$, if M and A are chosen sufficiently large then for any $\beta \in (0,1]$, and $(d,f) \in [\![1,\ell]\!]^2$ satisfying $(f-d) \geq \eta\ell$, we have*

$$\mathbb{E}\left[g(\omega_1,\dots,\omega_\ell)Z^0_{[d,f]}\right] \leq \eta. \tag{4.7}$$

Proof (Proof of Proposition 4 from Proposition 5). We decompose $Z^{\mathcal{I}}$ according to the first and last contact points in each block $(B_i)_{i\in\mathcal{I}}$ where $\mathcal{I} := \{i_1,\dots,i_l\}$. We have

$$Z^{\mathcal{I}} := \sum_{\substack{d_1,f_1\in B_{i_1}\\ d_1\leqslant f_1}} \cdots \sum_{\substack{d_l\in B_{i_l}\\ f_l=N}} K(d_1)u(f_1-d_1)Z^h_{[d_1,f_1]}K(d_2-f_1) \cdots K(d_l-f_{l-1})u(N-d_l)Z^h_{[d_l,N]}. \tag{4.8}$$

Then we use the fact that, due to our choice for the value of h, we have $Z^h_{d,f} \leq e^{\ell h}Z^0_{d,f} = eZ^0_{d,f}$, for any d and f such that $(f-d)\leq\ell$. We obtain thus using the product structure of $G_{\mathcal{I}}$ (3.15)

$$\begin{aligned}
\mathbb{E}\left[G_{\mathcal{I}}(\omega)Z^{\mathcal{I}}\right] &\leq e^{|\mathcal{I}|}\sum_{\substack{d_1,f_1\in B_{i_1}\\ d_1\leqslant f_1}} \cdots \sum_{\substack{d_l\in B_{i_l}\\ f_l=N}} K(d_1)u(f_1-d_1)\mathbb{E}\left[g_{i_1}(\omega)Z^0_{[d_1,f_1]}\right]K(d_2-f_1)\\
&\qquad \cdots K(d_l-f_{l-1})u(N-d_l)\mathbb{E}\left[g_{i_l}(\omega)Z^0_{[d_l,N]}\right]\\
&\leq e^{|\mathcal{I}|}\sum_{\substack{d_1,f_1\in B_{i_1}\\ d_1\leqslant f_1}} \cdots \sum_{\substack{d_l\in B_{i_l}\\ f_l=N}} K(d_1)u(f_1-d_1)\left[\eta+(1-\eta)\mathbf{1}_{\{(f_1-d_1)\leq\eta\ell\}}\right]K(d_2-f_1)\\
&\qquad \cdots K(d_l-f_{l-1})u(N-d_l)\left[\eta+(1-\eta)\mathbf{1}_{\{(f_l-d_l)\leq\eta\ell\}}\right],
\end{aligned} \tag{4.9}$$

where in the last line we used Proposition 5, and when $(f_j-d_j)\leq\eta\ell$, the fact that

$$\mathbb{E}\left[g_{i_j}(\omega)Z^0_{[d_j,f_j]}\right] \leq \mathbb{E}\left[Z^0_{[d_j,f_j]}\right] = 1. \tag{4.10}$$

Now we only need to obtain a bound on the r.h.s of (4.9). Using (2.1) and (4.6), we can replace $K(n)$ and $u(n)$ by $n^{-(\alpha+1)}$ and $(n+1)^{1-\alpha}$ (the quantity $n+1$ is present instead of n because we also consider $u(0)$) at the cost of losing a constant factor per cell. Thus we need to prove that given $\delta>0$, if η is sufficiently small, we have for some constant C_ℓ

$$\begin{aligned}
&\sum_{\substack{d_1,f_1\in B_{i_1}\\ d_1\leqslant f_1}} \cdots \sum_{\substack{d_l\in B_{i_l}\\ f_l=N}} (d_1)^{-(1+\alpha)}(f_1-d_1+1)^{\alpha-1}[\eta+(1-\eta)\mathbf{1}_{\{(f_1-d_1)\leq\eta\ell\}}](d_2-f_1)^{-(1+\alpha)}\\
&\qquad \cdots (d_l-f_{l-1})^{-(1+\alpha)}(N-d_l+1)^{\alpha-1}[\eta+(1-\eta)\mathbf{1}_{\{(f_l-d_l)\leq\eta\ell\}}]\\
&\qquad\qquad \leq C_\ell\delta^{|\mathcal{I}|}\prod_{j=1}^{l}|i_j-i_{j-1}|^{-\frac{1+\alpha}{2}}.
\end{aligned} \tag{4.11}$$

For the remainder of the proof, $\delta > 0$ is fixed and η is chosen sufficiently small in a way that depends on η. To obtain the bound, we proceed as in the computation [4, Equations (4.25) to (4.37)]. We observe that the l.h.s. in (4.11) is bounded above by

$$\sum_{\substack{d_1, f_1 \in B_{i_1} \\ d_1 \leqslant f_1}} \cdots \sum_{\substack{d_l \in B_{i_l} \\ f_l = N}} (d_1)^{-\frac{(1+\alpha)}{2}} \times \prod_{j=1}^{l-1} \left((d_j - f_{j-1}^{\max})^{-\frac{(1+\alpha)}{2}} (f_j - d_j + 1)^{\alpha-1} [\eta + (1-\eta)\mathbf{1}_{\{(f_i - d_i) \leq \eta \ell\}}] (d_{j+1}^{\min} - f_j)^{-\frac{(1+\alpha)}{2}} \right) \times (d_l - f_{l-1}^{\max})^{-\frac{(1+\alpha)}{2}}, \quad (4.12)$$

where $f_{j-1}^{\max}$ is the maximal element of $B_{i_{j-1}}$ ($f_0^{\max} = 0$ by convention) and $d_{j+1}^{\min}$ is the minimal element of $B_{i_{j+1}}$ and we decide to bound each term in the product separately. We are going to prove that for all $j \in \{1, \dots, l-1\}$ we have, provided that η is sufficiently small

$$\sum_{\substack{d_j, f_j \in B_{i_j} \\ d_j \leq f_j}} (d_j - f_{j-1}^{\max})^{-\frac{(1+\alpha)}{2}} (f_j - d_j + 1)^{\alpha-1} [\eta + (1-\eta)\mathbf{1}_{\{(f_i - d_i) \leq \eta \ell\}}] (d_{j+1}^{\min} - f_j)^{-\frac{(1+\alpha)}{2}} \leq \delta \left[(i_j - i_{j-1})(i_{j+1} - i_j) \right]^{-(1+\alpha)}. \quad (4.13)$$

Additionally we need two additional inequalities to bound the contribution of the first and last jump respectively. The reader will readily check that

$$\begin{aligned} (d_1)^{-\frac{(1+\alpha)}{2}} &\leq i_1^{-\frac{(1+\alpha)}{2}}, \\ \sum_{d_l \in B_m} (d_l - f_{l-1}^{\max})^{-\frac{(1+\alpha)}{2}} &\leq \ell (m - i_{l-1})^{-\frac{(1+\alpha)}{2}}. \end{aligned} \quad (4.14)$$

Equation (4.11) follows by multiplying the three inequalities given in (4.13) and (4.14).

Let us now prove (4.13): we split the set of indices $\{d_j, f_j \in B_{i_j} \ : \ d_j \leq f_j\}$ in the r.h.s. of (4.13) into three subsets by adding an extra condition:

(i) $\{d_j, f_j \in B_{i_j} \ : \ (i_j - 3/4)\ell \leq d_j \leq f_j\}$,
(ii) $\{d_j, f_j \in B_{i_j} \ : \ d_j \leq f_j \leq (i_j - 1/4)\ell\}$,
(iii) $\{d_j, f_j \in B_{i_j} \ : \ f_j \geq d_j + \ell/2\}$.

It is easy to check that the union of these (non disjoint) subsets give us back the original set. We estimate the contribution of each set separately, the idea being that each condition in $(i) - (iii)$ allows to replace one of the variable factors by

an asymptotic equivalent which does not depend on d_i nor f_i. This makes the computation easier. First we can bound the contribution (i) as follows

$$\sum_{\substack{d_j, f_j \in B_{i_j} \\ (i_j - 3/4)\ell \le d_j \le f_j}} \cdots$$
$$\le [\ell(i_j - i_{j-1})/4]^{-\frac{(1+\alpha)}{2}} \sum_{\substack{d_j, f_j \in B_{i_j} \\ d_j \le f_j}} (f_j - d_j + 1)^{\alpha-1} [\eta + (1-\eta)\mathbf{1}_{\{(f_j - d_j) \le \eta\ell\}}] (d_{j+1}^{\min} - f_j)^{-\frac{(1+\alpha)}{2}} \tag{4.15}$$

Then considering the sum over d_j separately, we obtain that the remaining double sum is smaller than

$$\left(\sum_{a=1}^{\eta\ell} a^{\alpha-1} + \eta \sum_{a=\eta\ell}^{\ell} a^{\alpha-1} \right) \quad \left(\sum_{f_j \in B_{i_j}} (d_{j+1} - f_j)^{-\frac{(1+\alpha)}{2}} \right). \tag{4.16}$$

The first factor is smaller than $\varepsilon \ell^\alpha$ where ε can be made arbitrarily small by considering small η, and the second factor is of order $|i_{j+1} - i_j|^{-\frac{(1+\alpha)}{2}} \ell^{\frac{1-\alpha}{2}}$. All the powers of ℓ cancel out and we obtain (provided ε is sufficiently small) that

$$\sum_{\substack{d_j, f_j \in B_{i_j} \\ (i_j - 3/4)\ell \le d_j \le f_j}} \cdots \le \frac{\delta}{3} \left[(i_j - i_{j-1}) (i_{j+1} - i_j) \right]^{-\frac{(1+\alpha)}{2}}, \tag{4.17}$$

where the term in the sum is the same as in (4.13). We obtain similarly by symmetry

$$\sum_{\substack{d_j, f_j \in B_{i_j} \\ d_j \le f_j \le (i_j - 1/4)\ell}} \cdots \le \frac{\delta}{3} \left[(i_j - i_{j-1}) (i_{j+1} - i_j) \right]^{-\frac{(1+\alpha)}{2}}. \tag{4.18}$$

Finally in the case $f_j - d_j \ge \ell/2$ we have, provided $\eta < 1/2$,

$$[\eta + (1-\eta)\mathbf{1}_{\{(f_j - d_j) \le \eta\ell\}}](f_j - d_j + 1)^{\alpha-1} \le \eta(\ell/2 + 1)^{\alpha-1}$$

and thus

$$\sum_{\substack{d_j, f_j \in B_{i_j} \\ f_i \ge d_j + \ell/2}} \cdots \le \delta(\ell/2+1)^{\alpha-1} \sum_{d_j, f_j \in B_{i_j}} (d_j - f_{j-1}^{\max})^{-\frac{(1+\alpha)}{2}} (d_{j+1}^{\min} - f_j)^{-\frac{(1+\alpha)}{2}}. \tag{4.19}$$

The double sum factorizes and can be shown to be of order

$$\ell^{1-\alpha} \left[(i_j - i_{j-1}) (i_{j+1} - i_j) \right]^{-\frac{(1+\alpha)}{2}}.$$

This yields (when η is sufficiently small)

$$\sum_{\substack{d_j, f_j \in B_{i_j} \\ f_j \ge d_j + \ell/2}} \cdots \le \frac{\delta}{3} \left[(i_j - i_{j-1}) (i_{j+1} - i_j) \right]^{-\frac{(1+\alpha)}{2}}. \tag{4.20}$$

This concludes the proof of (4.13) and thus of Proposition 4.

5 Proof of Proposition 5

5.1 Reduction to a Simpler Statement

The aim of this section is to reduce the proof of Proposition 5 to the estimation of the probability of some nice event for the environment ω. As $\mathbb{E}\left[Z^0_{[d,f]}\right] = 1$, $Z^0_{[d,f]}$ can be considered as a probability density. To prove (4.7) we must thus show that the probability of $\mathcal{A}_\ell$ under the probability $Z^0_{[d,f]}(\omega)\mathbb{P}[\,\mathrm{d}\omega]$ is close to one, whenever $(f-d) \geq \eta\ell$.

More precisely, given a fixed realization of τ, with $a, b \in \tau$ we define $\mathbb{P}^{a,b}_\tau$

$$\frac{\mathrm{d}\mathbb{P}^{d,f}_\tau}{\mathrm{d}\mathbb{P}}(\omega) := \prod_{n\in\tau\cap[\![d,f]\!]} (1+\beta\omega_n). \tag{5.1}$$

We have

$$\mathbb{E}\left[g(\omega_1,\dots,\omega_\ell)Z^h_{d,f}\right] = e^{-M} + (1-e^{-M})\mathbf{E}\left[\mathbb{P}^{d,f}_\tau(\mathcal{A}^{\complement}_\ell) \mid d, f \in \tau\right]. \tag{5.2}$$

We notice that under $\mathbb{P}^{d,f}_\tau$, the ω_ns are still independent, but they are not identically distributed anymore, as for $n \in [d,f]\cap\tau$, the distribution of ω_n has been tilted and thus peaks are more likely to appear on those sites. In order to bound the probability of $\mathcal{A}_\ell$ we are going to check only sites with tilted environment. Let us consider the alternative event (recall (3.18))

$$\mathcal{A}(d,f,\tau) := \{\exists i,j \in \tau\cap[d,f],\, i\neq j,\, \min(\omega_i,\omega_j) \geq V(j-i)\} \tag{5.3}$$

As $\mathcal{A}(d,f,\tau)$ is clearly included in $\mathcal{A}_\ell$, it is sufficient for us to obtain a bound on $\mathbb{P}^{d,f}_\tau(\mathcal{A}(d,f,\tau)^{\complement})$. If we let $\widetilde{\mathbb{P}}$ denote the probability obtained by tilting all the variables: the ω_ns are IID and with distribution

$$\widetilde{\mathbb{P}}[\omega_1 \in \mathrm{d}x] = (1+\beta x)\mathbb{P}[\omega_1 \in \mathrm{d}x], \tag{5.4}$$

then we have

$$\mathbb{P}^{d,f}_\tau(\mathcal{A}(d,f,\tau)) = \widetilde{\mathbb{P}}\left(\mathcal{A}(d,f,\tau)\right).$$

Hence we can prove Proposition 5 provided we show

$$\mathbf{E}\left[\widetilde{\mathbb{P}}(\mathcal{A}(d,f,\tau)^{\complement}) \mid d, f \in \tau\right] \leq \varepsilon, \tag{5.5}$$

for an arbitrary ε. Without loss of generality, let us assume that $d = 0$. We set for notational simplicity $r := \eta\ell/4$ and we define a new event $\mathcal{B}(r,\tau)$ satisfying $\mathcal{B}(r,\tau) \subset \mathcal{A}(0,f,\tau)$ for all $f \geq \eta$,

$$\mathcal{B}(r,\tau) := \left\{\exists (i,j) \in [\![1,r]\!]\times[\![1,r^{\alpha/4}]\!],\, i\in\tau,\, i+j\in\tau,\, \min(\omega_i,\omega_{i+j}) \geq V(j)\right\}. \tag{5.6}$$

Furthermore it is measurable with respect to $\tau\cap\sigma([0,f/2])$. We want to use this assumption to drop the conditioning in τ present in (5.5). The reason to consider only dual peaks with relatively small distance ($\leq r^{\alpha/4}$) is not of crucial importance but it notably simplifies the computation (cf. (5.29)).

Lemma 1. *There exists a constant such that for all $N > 0$ for any function F measurable with respect to $\sigma(\tau \cap [0, N/2])$ we have*

$$\mathbf{E}[F(\tau) \mid N \in \tau] \leq C\mathbf{E}[F(\tau)] \tag{5.7}$$

Proof. If we let $X_N := \max\{\tau \cap [0, N/2]\}$, the left-hand side can be rewritten as

$$\sum_{i=0}^{N/2} \mathbf{E}[F(\tau) \mid X_N = i, N \in \tau]\mathbf{P}[X_N = i \mid N \in \tau] = \sum_{i=0}^{N/2} \mathbf{E}[F(\tau) \mid X_N = i]\mathbf{P}[X_N = i \mid N \in \tau], \tag{5.8}$$

where the equality comes from the Markov property for the renewal τ. With this formulation, (5.7) is simply a consequence of [6, Equation (A.15)].

As a consequence of the lemma we have

$$\mathbf{E}\left[\widetilde{\mathbb{P}}(\mathcal{A}(0, f, \tau)^{\complement}) \mid f \in \tau\right] \leq \mathbf{E}\left[\widetilde{\mathbb{P}}(\mathcal{B}(r, \tau)^{\complement}) \mid f \in \tau\right] \leq C\mathbf{E}\left[\widetilde{\mathbb{P}}(\mathcal{B}(r, \tau)^{\complement})\right] \tag{5.9}$$

Hence to conclude the proof we only need to show the following which we do in the next section.

Lemma 2. *Recall that $r = \eta\ell/4$. If A is chosen sufficiently large (depending on η, M and ε), we have for all $\beta \in (0, 1]$*

$$\mathbf{E}\left[\widetilde{\mathbb{P}}\left[\mathcal{B}(r, \tau)\right]\right] \geq 1 - \varepsilon. \tag{5.10}$$

5.2 Proving Lemma 2

For i, j, k in $\mathbb{N}$ let δ_i, $\delta_{i,j}$ and $\delta_{i,j,k}$ be the indicator function of the respective events $\{i \in \tau\}$, $\{i, i+j \in \tau\}$ and $\{i, i+j, i+j+k \in \tau\}$. Using the independence of renewal jumps we have (recall the definition of $u(n)$ above (4.6))

$$\mathbf{E}[\delta_i] = u(i), \quad \mathbf{E}[\delta_{i,j}] = u(i)u(j), \quad \text{and} \quad \mathbf{E}[\delta_{i,j,k}] = u(i)u(j)u(k). \tag{5.11}$$

Let us also set (recall (3.18))

$$W(i, j) := \mathbf{1}_{\{\min(\omega_i, \omega_{i+j}) \geq V(j)\}} \tag{5.12}$$

We define

$$Y(\omega, \tau) := \sum_{i=1}^{r} \sum_{j=1}^{r^{\alpha/4}} W(i, j)\delta_{i,j}. \tag{5.13}$$

With this notation we have $\mathcal{B}(r,\tau) = \{Y(\omega,\tau) \geq 1\}$. We are going to prove the lemma by controlling the two first moments of Y w.r.t measure $\widetilde{\mathbb{P}}$.

We are going to use repeatedly the following estimates which can be deduced from the assumption (2.3), the definition of the size biased measure and the value chosen for ℓ: There exists a constant (depending on M) such that for every value of i, j and $\beta \in (0,1]$ chosen we have (recall $\alpha = 1 - \gamma^{-1}$)

$$(C_M)^{-1}\beta\left(\ell(\log \ell)n\right)^{-\frac{\alpha}{2}} \leq \widetilde{\mathbb{P}}[\omega_1 \geq V(n)] \leq C_M\beta\left(\ell(\log \ell)n\right)^{-\frac{\alpha}{2}} \tag{5.14}$$

From now on, all the constants displayed in the equation might depend on M and η but not on other parameters. Using (5.14) we have for some $c > 0$

$$\widetilde{\mathbb{E}}[Y] \geq c\beta^2(\ell \log \ell)^{-\alpha} \sum_{i=1}^{r} \sum_{j=1}^{r^{\alpha/4}} j^{-\alpha}\delta_{i,j}. \tag{5.15}$$

To compute the variance, we ignore after developing $Y^2 = \sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} \sum_{i'=1}^{r}\sum_{j'=1}^{r^{\alpha/4}} \cdots$ all the terms which have covariance zero. We are left with the diagonal terms but also terms for which $|\{i, i+j\} \cap \{i', i'+j'\}| = 1$ (three cases must be considered). Reordering the sum this gives the following estimate

$$\begin{aligned}
\mathrm{Var}_{\widetilde{\mathbb{P}}}[Y(\tau,\omega)] \leq\ & \widetilde{\mathbb{E}}\left[\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} W(i,j)\delta_{i,j}\right] \\
&+ 2\widetilde{\mathbb{E}}\left[\sum_{i=1}^{r}\sum_{1\leq j<k\leq r^{\alpha/4}} \left[W(i,j)W(i,k) + W(i,k)W(i+j,k-j)\right]\delta_{i,j,k-j}\right] \\
&+ 2\widetilde{\mathbb{E}}\left[\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}}\sum_{k=1}^{r^{\alpha/4}} W(i,j)W(i+j,k)\delta_{i,j,k}\right]
\end{aligned} \tag{5.16}$$

Using (5.14) to control all the expectation we obtain

$$\begin{aligned}
\mathrm{Var}_{\widetilde{\mathbb{P}}}[Y(\tau,\omega)] \leq\ & C\beta^2(\ell \log \ell)^{-\alpha} \sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} j^{-\alpha}\delta_{i,j} \\
&+ C\beta^3(\ell \log \ell)^{-3\alpha/2}\Bigg(\sum_{i=1}^{r}\sum_{1\leq j<k\leq r^{\alpha/4}} \left(j^{-\alpha/2}k^{-\alpha} + k^{-\alpha/2}(k-j)^{-\alpha}\right)\delta_{i,j,k-j} \\
&\qquad + \sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}}\sum_{k=1}^{r^{\alpha/4}} j^{-\alpha/2}k^{-\alpha/2}\max(j,k)^{-\alpha/2}\delta_{i,j,k}\Bigg).
\end{aligned} \tag{5.17}$$

To conclude the proof of Lemma 2 we use the following estimates proved in the next section.

Proposition 6. *The following estimates hold for some universal constant C*

(i) $\mathbf{E}[\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} j^{-\alpha}\delta_{i,j}] \le Cr^{\alpha}(\log r)$.
(ii) There exists $\varepsilon > 0$ such that for all r sufficiently large

$$\mathbf{P}\left[r^{-\alpha}(\log r)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} j^{-\alpha}\delta_{i,j} \ge \varepsilon\right] \ge 1-\varepsilon,$$

(iii)

$$\mathbf{E}\left[\sum_{i=1}^{r}\sum_{1\le j<k\le r^{\alpha/4}}\left(j^{-\alpha/2}k^{-\alpha}+k^{-\alpha/2}(k-j)^{-\alpha}\right)\delta_{i,j,k-j}\right] \le Cr^{3\alpha/2}(\log r),$$
$$\mathbf{E}\left[\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}}\sum_{k=1}^{r^{\alpha/4}} j^{-\alpha/2}k^{-\alpha/2}\max(j,k)^{-\alpha/2}\delta_{i,j,k}\right] \le Cr^{3\alpha/2}(\log r). \tag{5.18}$$

From (ii) and (5.15) we obtain directly that provided ε is sufficiently small (how small can depend on η and M) we have

$$\mathbf{P}\left[\widetilde{\mathbb{E}}[Y(\tau,\omega)] \ge \varepsilon\beta^2(\log\ell)^{\gamma^{-1}}\right] \ge 1-\varepsilon \tag{5.19}$$

From (i) and (iii) and (5.16), we obtain a bound on $\mathbf{E}[\mathrm{Var}_{\widetilde{\mathbb{P}}}[Y(\omega,\tau)]]$. Then applying Markov inequality we obtain that with $\mathbf{P}$-probability larger than $1-\varepsilon$ we have

$$\mathrm{Var}_{\widetilde{\mathbb{P}}}[Y(\omega,\tau)] \le C(\eta,M)\varepsilon^{-1}\left[\beta^2(\log\ell)^{\gamma^{-1}}+\beta^3(\log\ell)^{\gamma^{-1}-\alpha/2}\right]. \tag{5.20}$$

With our choice $\ell = e^{A\beta^{-2\gamma}}$ with probability larger than $(1-2\varepsilon)$ we have

$$\begin{aligned}\widetilde{\mathbb{E}}[Y(\tau,\omega)] &\ge \varepsilon A^{\gamma^{-1}},\\ \mathrm{Var}_{\widetilde{\mathbb{P}}}[Y(\tau,\omega)] &\le C(\eta,M)\varepsilon^{-1}A^{\gamma^{-1}}(1+O(\beta^{1+\alpha\gamma})).\end{aligned} \tag{5.21}$$

Thus by choosing A sufficiently large (depending on η, M and ε), using Chebychev inequality we conclude that

$$\mathbf{E}\widetilde{\mathbb{P}}\left[Y(\tau,\omega) \ge 1\right] \ge 1-3\varepsilon.$$

□

5.3 Proof of Proposition 6

We start with point (i) and (iii) which are simpler to prove. Using (5.11) to rewrite the sum in (i) and (4.6) to estimate it, we obtain

$$\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} u(i)u(j)j^{-\alpha} \le C\sum_{i=1}^{r} i^{\alpha-1}\sum_{j=1}^{r^{\alpha/4}} j^{-1}, \tag{5.22}$$

For (iii) let us perform the computation only for the first of the three sum we have to control since the two other cases similar. Using (5.11) and (4.6) we have

$$\sum_{i=1}^{r} \sum_{1\le j<k\le r^{\alpha/4}} u(i)u(j)u(k-j)j^{-\alpha/2}k^{-\alpha} \le C\sum_{i=1}^{r} i^{\alpha-1}\sum_{j=1}^{r^{\alpha/4}} j^{-\alpha/2-1}\sum_{k=j+1}^{r^{\alpha/4}}(k-j)^{-1} \le C' r^{3\alpha/2}\log r. \tag{5.23}$$

Let us now consider the more delicate point (ii). We set

$$\begin{aligned} X_r^1 &:= r^{-\alpha}\sum_{j=1}^{r}\delta_i, \\ X_r^2 &:= \left(\sum_{k=1}^{r^{\alpha/4}} k^{-\alpha}u(k)\right)^{-1} r^{-\alpha}\sum_{i=1}^{r}\sum_{j=1}^{r^{\alpha/4}} j^{-\alpha}\delta_{i,j}. \end{aligned} \tag{5.24}$$

Note that as $\left(\sum_{k=1}^{r^{\alpha/4}} k^{-\alpha}u(k)\right)^{-1}$ is of order $\log r$, X_r^2 is asymptotically equivalent to the expression appearing in (ii). Hence it is sufficient to prove that

$$\lim_{r\to\infty} \mathbf{P}\left[X_r^2 \ge \varepsilon\right] \ge 1-\varepsilon. \tag{5.25}$$

We are going to show that X_r^2 converges in law and that the limit distribution does not give any mass to zero. First we notice that as $n^{-1/\alpha}\tau_{\lceil n\rceil}$ converges to an α-stable subordinator (see e.g. [9, Chap. 16]) X_r^1 converges to the first hitting time of $[1,+\infty)$ for this limiting process. This hitting time is strictly positive with probability 1. Hence we conclude the proof using the following technical lemma, which readily implies that X_r^2 converges in distribution to the same random variable.

Lemma 3. *We have*

$$\lim_{r\to\infty} \mathbf{E}\left[(X_r^1 - X_r^2)^2\right] = 0. \tag{5.26}$$

Proof. We have

$$r^{\alpha}\left(\sum_{k=1}^{r^{\alpha/4}} k^{-\alpha}u(k)\right)(X_n^2 - X_n^1) = \sum_{i=1}^{r}\delta_i\left(\sum_{j=1}^{r^{\alpha/4}} j^{-\alpha}\left(\delta_{i+j} - u(j)\right)\right) =: \sum_{i=1}^{r} U_i. \tag{5.27}$$

Hence we have

$$\mathbf{E}\left[(X_n^2 - X_n^1)^2\right] \le Cr^{-2\alpha}(\log r)^{-2}\sum_{i_1,i_2=1}^{r}\mathbf{E}[U_{i_1}U_{i_2}]. \tag{5.28}$$

We are going to show that we have

$$\mathbf{E}[U_i^2] \le C r^{\alpha/2} i^{\alpha-1}. \tag{5.29}$$

and

$$|i_1 - i_2| \ge r^{\alpha/4} \quad \Rightarrow \quad \mathbf{E}[U_{i_1} U_{i_2}] = 0, \tag{5.30}$$

Using these estimates we obtain that

$$\begin{aligned}\sum_{i_1,i_2=1}^{r} \mathbf{E}[U_{i_1} U_{i_2}] &\le \sum_{i=1}^{r} \mathbf{E}[U_i^2] + 2 \sum_{i_1=1}^{r} \sum_{i_2=(i_1+1)}^{\max(r, i_1+r^{\alpha/4})} \mathbf{E}[U_{i_1} U_{i_2}] \\ &\le (1 + 2r^{\alpha/4}) \sum_{i=1}^{r} \mathbf{E}[U_i^2] \le C r^{3\alpha/4}) \sum_{i=1}^{r} i^{\alpha-1} \le C r^{7\alpha/4},\end{aligned} \tag{5.31}$$

which in regards of (5.28) allows to conclude.

The inequality (5.29) is simple to obtain. We have

$$\left| \sum_{j=1}^{r^{\alpha/4}} j^{-\alpha} \left(\delta_{i+j} - u(j) \right) \right| \le r^{\alpha/4},$$

and hence

$$\mathbf{E}[U_i^2] \le r^{\alpha/2} \mathbf{E}[\delta_i] \le C r^{\alpha/2} i^{\alpha-1}.$$

For (5.30), we assume that i_1 is the smallest index. Note that with the assumption $i_2 - i_1 \ge r^{\alpha/4}$, U_{i_1} is measurable w.r.t. $\sigma(\tau \cap [0, i_2])$. Hence we have

$$\begin{aligned}\mathbf{E}[U_{i_1} U_{i_2} \mid \tau \cap [0, i_2]] &= U_{i_1} \delta_{i_2} \sum_{j=1}^{r^{\alpha/4}} j^{-\alpha} \mathbf{E}\left[\delta_{i_2+j} - u(j) \mid \tau \cap [0, i_2] \right] \\ &= U_{i_1} \delta_{i_2} \sum_{j=1}^{r^{\alpha/4}} j^{-\alpha} \mathbf{E}\left[\delta_{i_2+j} - u(j) \mid i_2 \in \tau \right] = 0.\end{aligned} \tag{5.32}$$

To obtain the second equality, we observe that both terms are equal to zero if $i_2 \notin \tau$ and that conditionally to $i_2 \in \tau$, $\tau \cap [0, i_2]$ and $\tau \cap [i_2, \infty)$ are independent.

References

1. Abraham, D.B.: Surface structures and phase transitions, exact results. In: Domb, C., Lebowitz, J.L. (eds.) Phase Transitions and Critical Phenomena, vol. 10, pp. 1–74. Academic Press, London (1986)
2. Alexander, K.S.: The effect of disorder on polymer depinning transitions. Commun. Math. Phys. **279**, 117–146 (2008)

3. Alexander, K.S., Zygouras, N.: Quenched and annealed critical points in polymer pinning models. Comm. Math. Phys. **291**, 659–689 (2009)
4. Berger, Q., Lacoin, H.: Pinning on a defect line: characterization of marginal disorder relevance and sharp asymptotics for the critical point shift. J. Inst. Math. Jussieu **17**, 305–346 (2018)
5. Caravenna, F., Den Hollander, F.: A general smoothing inequality for disordered polymers. Elec. Comm. Probab. **18**, 1–15 (2013)
6. Derrida, B., Giacomin, G., Lacoin, H., Toninelli, F.L.: Fractional moment bounds and disorder relevance for pinning models. Comm. Math. Phys. **287**, 867–887 (2009)
7. Derrida, B., Hakim, V., Vannimenus, J.: Effect of disorder on two-dimensional wetting. J. Statist. Phys. **66**, 1189–1213 (1992)
8. Doney, R.A.: One-sided local large deviation and renewal theorems in the case of infinite mean. Probab. Theory Relat. Fields **107**, 451–465 (1997)
9. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. II. Wiley, New York-London-Sydney (1966)
10. Fisher, M.E.: Walks, walls, wetting, and melting. J. Stat. Phys. **34**, 667–729 (1984)
11. Forgacs, G., Lipowsky, R., Nieuwenhuizen, Th.M.: The behavior of interfaces in ordered and disordered systems. In: Domb, C., Lebowitz, J.L. (eds.) Phase Transitions and Critical Phenomena, vol. 14, pp. 135–363. Academic Press, London (1991)
12. Forgacs, G., Luck, J.M., Nieuwenhuizen, Th.M., Orland, H.: Wetting of a disordered substrate: exact critical behavior in two dimensions. Phys. Rev. Lett. **57**, 2184–2187 (1986)
13. Giacomin, G.: Random Polymer Models. Imperial College Press, World Scientific, London (2007)
14. Giacomin, G.: Disorder and Critical Phenomena Through Basic Probability Models: École d'Été de Probabilités de Saint-Flour XL 2010. Lecture Notes in Mathematics, vol. 2025. Springer, Heidelberg (2011)
15. Giacomin, G., Lacoin, H., Toninelli, F.L.: Marginal relevance of disorder for pinning models. Commun. Pure Appl. Math. **63**, 233–265 (2010)
16. Giacomin, G., Lacoin, H., Toninelli, F.L.: Disorder relevance at marginality and critical point shift. Ann. Inst. H. Poincaré **47**, 148–175 (2011)
17. Giacomin, G., Toninelli, F.L.: Smoothing effect of quenched disorder on polymer depinning transitions. Commun. Math. Phys. **266**, 1–16 (2006)
18. Harris, A.B.: Effect of random defects on the critical behaviour of Ising models. J. Phys. C **7**, 1671–1692 (1974)
19. Lacoin, H.: The martingale approach to disorder irrelevance for pinning models. Elec. Comm. Probab. **15**, 418–427 (2010)
20. Lacoin, H., Sohier, J.: Disorder relevance without Harris Criterion: the case of pinning model with γ-stable environment. Electron. J. Probab. **22**, 1–26 (2017)
21. Toninelli, F.L.: A replica-coupling approach to disordered pinning models. Commun. Math. Phys. **280**, 389–401 (2008)
22. Toninelli, F.L.: Coarse graining, fractional moments and the critical slope of random copolymers. Electron. J. Probab. **14**, 531–547 (2009)

A Rate of Convergence Result for the Frederickson-Andersen Model

Thomas Mountford[1(✉)] and Glauco Valle[2]

[1] Département de Mathématiques, École Polytechnique Fédérale, 1015 Lausanne, Switzerland
thomas.mountford@epfl.ch

[2] Instituto de Matemática, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, Rio de Janeiro 21945-970, Brazil
glauco.valle@im.ufrj.br

Our paper [2] considers the Frederickson-Andersen model (FA model) on a general class of graphs of bounded degree. Given a graph $G = (V, E)$ we write $x \sim y$ for $x, y \in V$ if $\{x, y\}$ is in E. The FA model with parameter $q \in (0, 1)$ is a spin system on $\{0, 1\}^V$ with flip rates at a site x given by

$$c(x, \eta) = \begin{cases} q & \text{if } \eta(x) = 0, \sum_{y \sim x} \eta(y) \neq 0\,, \\ 1 - q & \text{if } \eta(x) = 1, \sum_{y \sim x} \eta(y) \neq 0\,, \end{cases}$$

and is otherwise zero.

It can be thought of equivalently as a process where if, for site x, the condition $\sum_{y \sim x} \eta(y) \neq 0$ is satisfied, then at rate 1 the value $\eta(x)$ is replaced by Bernoulli(q) random variables independent of the process up to this time.

We note that the condition for a strictly positive rate at x depends only on neighbouring spins but not on $\eta(x)$ itself. Thus while it is immediate that $\delta_{\overline{0}}$, i.e. the point mass at the identically zero configuration $\overline{0}$, is invariant for the process, it is also true that if $\eta_0 \neq \overline{0}$ then $\overline{0}$ can never be attained by the process $(\eta_t : t \geq 0)$ corresponding to these flip rates.

Another consequence is that, by detailed balance, the measure γ_q on $\{0, 1\}^V$ for which all variables $(\eta(x))_{x \in V}$ are iid Bernoulli(q) is an equilibrium.

The starting point for our work was the article of Blondel, Cancrini, Martinelli, Roberto and Toninelli [1] concerning the speed convergence to γ_q for initial distributions satisfying for some $c > 0$

$$\sup_x E[e^{cd(x, \eta_0)}] < \infty\,, \tag{1}$$

where $d(x, \eta_0) = \inf\{d(x, y) : \eta_0(y) = 1\}$. The parameter q was required to be strictly greater than $\frac{1}{2}$. The graphs allowed were connected but completely general except that a growth condition was required:

$$M_d : \exists C < \infty \quad \text{so that} \quad \forall \quad x \in V \quad |B(x, r)| \leq C r^d$$

where as usual $B(x, r) = \{y \in V : d(x, y) < r\}$. It is natural to view the inf of the $d's$ for which this condition holds as the "dimension" of the graph.

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 617–620, 2019.
https://doi.org/10.1007/978-3-030-15096-9_24

Under these minimalistic conditions it was shown that for all cylinder function f there exists a constant C_f, depending upon f, and the "dimension" d so that

$$|Ef(\eta_t) - \nu_q(f)| \leq \begin{cases} C_f e^{-t/C_f}, & d = 1, \\ C_f e^{-(\frac{t}{\log t})^{\frac{1}{d}}/C_f}, & d > 1. \end{cases}$$

So only in the special case $d = 1$ the convergence rate is exponential. The techniques used were sophisticated log-Sobolev estimates and exploitation of spectral gap estimates for clever choices of auxiliary finite Markov chains.

Our result, obtained using contact processes and oriented percolation techniques is the following.

Theorem 1. *Let $G = (V, E)$ be a countable connected graph of bounded degree satisfying the growth condition $\exists M < \infty$ and $\epsilon > 0$ so that for every $x \in V, r \in \mathbb{Z}_+ \quad |B(x, r)| \leq Me^{r^{1-\epsilon}}$. For q sufficiently close to one, $\exists c = c(q, \epsilon, M) > 0$ such that for any non null η_0 and cylinder function f there exists $C = C(f, c, \eta_0)$ so that*

$$\left|\mathrm{E}^{\eta_0}\left[f(\eta_t)\right] - \gamma_q(f)\right| \leq Ce^{-ct}.$$

Our result took a fixed $\eta_0 \not\equiv \bar{0}$ and a function f rather than the condition (1). This is apparently more general but this is illusory. In order to replicate the results of [1], i.e. obtain the constant C in the statement depending only on the size of the support of the cylinder function of f and not on its location, we would need to pass to their condition. It is also worth noting (see discussion of parameter q below) that the condition (1) is quite reasonable given our process. So for this part of the hypotheses there is no difference. On the other hand our formulation makes clear that under our conditions $\delta_{\bar{0}}$ and γ_q are the only extremal equilibria, which we do not see as obvious or simple.

The condition of a finite dimensional graph was replaced by the sub geometric growth condition (on a connected graph)

$$\exists\ C < \infty,\ t > 0 \quad \text{so that} \quad \forall \quad x \quad |B(x, r)| \leq Ce^{r^{1-\epsilon}}$$

This condition represents a substantial loosening of conditions. Not only does it include graphs of infinite "dimension" but also it treats all graphs of finite dimension equally rather than providing a bound which is progressively worse as the dimension increases.

Our condition on parameter q is radically inferior to [1]. The bounds on how close q is required to be to 1 can be calculated in terms of M and ϵ if need be but will typically be very close to 1. In addition the condition $q > \frac{1}{2}$ used by [1] seems more likely to be loosened than the constraints we demand.

On the other hand the bound we provide is qualitatively the correct one we believe. (We do not establish lower bounds.)

We do believe (but cannot show) that there is no medium regime where exponential convergence does not take place.

Our approach started from a common idea: we compared our process η_t (starting from a given η_0) with a process η_t^q starting from γ_q but generated by

the same Harris system. It would be sufficient to show that for x fixed in V

$$P(\eta_t(x) = \eta_t^q(x)) \geq 1 - Ce^{-ct}$$

for fixed $C,\ c$. To do this we considered dual paths associated with (x,t). A key observation was that if $\eta_t(x) \neq \eta_t^q(x)$ then there must be a dual path down from (x,t) to $V \times \{0\}$ on which $\eta_s(y) \neq \eta_s^q(y)$, $0 \leq s \leq t$. Our argument was based on showing that this was not possible.

Our proof used the contact process (or discrete contact-process-like processes) as tools: the first step consisted of using the observation of [1] that for $q > \frac{1}{2}$ and any $x_0 \in V$ the process $d(x_0, \eta_t)$ was stochastically dominated by a reflecting nearest neighbour random walk on $\mathbb{Z}_+$ which jumped up at rate $1-q$ and down (unless at 0) at rate ϵ. This enabled us to say that outside exponentially small probability the process η at time $\frac{t}{4}$, say, had produced many $1's$.

The second step consisted in showing that for every space time point (y,s) with $t/2 \leq s \leq t$ we could define an oriented percolation structure (with time flowing in reverse direction to the flow of the original process) so that if it survived for order t amount of time then outside exponentially in t small probability

(I) it would survive for ever,
(II) at time of order t it would have of order t sites belonging to it.

It would also have the property that if at dual time one of the sites in this percolation structure was occupied then necessarily $\eta_s(y) = 1$. It is important to realize that this dual is entirely a function of the generating Harris system over the relevant time interval; the dual is the same for processes generated by the given Harris system, though of course a site in the dual may have value 1 for one process and 0 for another. Given the result of the first step this meant that survival would (outside exponential probability) guarantee the existence of a site for which $\eta_s(y) = \eta_s^q(y) = 1$. Of course any oriented percolation process may die out no matter how high the (nontrivial) connection probabilities but high connection probabilities ensured that there were many "percolation points" (y,s). This, it is argued, leads to the conclusion that any dual path (in the sense of step one) cannot avoid one of the "percolating points". This means that there cannot be a dual path of any length on which η and η^q are different.

Thirdly we showed that any dual path must hit percolation points (outside exponentially small probability). This argument used a coarse graining and it was there that our below exponential growth condition was used.

We have shown exponentially fast convergence to equilibrium for q close to 1 (for graphs with below exponential growth) but in fact we believe that the convergence should be exponentially fast for every non zero q (for this class of graphs).

A reason for believing this is that if we consider q very small then we expect particles to be fairly isolated. As such at small rate q birth is given to a particle at a neighbouring site. Then at rate 1 (outside probability q) one of these two particles will die. If it is the new particle then essentially nothing has happened but if the original dies then effectively our particle has performed the step of a

random walk. Thus we can view the process as a branching random walk where particles move at rate q and give birth at rate q^2 to particles separated by 1.

As such it is believable that on a graph with polynomial growth we can have an embedded process on V, $(X_t : t \geq 0)$ so that $\eta_t(X_t) = 1 \ \forall \ t$ and so that X_t is attracted towards a given site. This is suggestive of exponential convergence but by no means a compelling argument. But in particular we believe there are only two extremal equilibria for every q for graphs of below exponential growth.

It is easy to convince oneself that for regular graphs of high degree if q is small then there need not be convergence, let alone more exponential convergence for all non zero η_0, and perhaps here the passage from below to above exponential growth is critical.

It also seems to be the case that for these high degree graphs there exist equilibria that are not combinations of $\delta_{\overline{0}}$ and γ_q.

However it still seems reasonable to guess that for a graph of exponential growth it is still the case that if q is sufficiently large then we have exponential convergence but our methods do not extend to this case.

References

1. Blondel, O., Cancrini, N., Martinelli, F., Roberto, C., Toninelli, C.: Fredrickson-Andersen one spin facilitaded model out of equilibrium. Markov Process Relat. Fields **19**, 383–406 (2013)
2. Mountford, T., Valle, G.: Exponential convergence for the Fredrikson-Andersen one spin facilitated model. arXiv:1609.01364

Stochastic Duality and Eigenfunctions

Frank Redig(✉) and Federico Sau

Delft Institute of Applied Mathematics, Delft University of Technology,
van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
f.h.j.redig@tudelft.nl, F.Sau@tudelft.nl

Abstract. We start from the observation that, anytime two Markov generators share an eigenvalue, the function constructed from the product of the two eigenfunctions associated to this common eigenvalue is a duality function. We push further this observation and provide a full characterization of duality relations in terms of spectral decompositions of the generators for finite state space Markov processes. Moreover, we study and revisit some well-known instances of duality, such as Siegmund duality, and extract spectral information from it. Next, we use the same formalism to construct all duality functions for some solvable examples, i.e., processes for which the eigenfunctions of the generator are explicitly known.

1 Introduction

Stochastic duality is a technique to connect two Markov processes via a so-called *duality function*. This connection, interesting in its own right, turns out to be extremely useful when the *dual* process is more tractable than the original process.

Several applications of stochastic duality may be found in the context of interacting particle systems [28] as, for instance, in the study of hydrodynamic limits and fluctuations [9,10,23], characterization of extremal measures [28,33], derivation of Fourier law of transport [3,24] and correlation inequalities [17]. Other fields rich of applications are population genetics, where the coalescent process arises as a natural dual process (see [11] and references therein) and branching-coalescing processes [13]. Duality and related notions have already been used in the study of spectral gaps and convergence to stationarity by several authors, see e.g. [6,12,14,29,32].

Part of the research about stochastic duality deals with the problem of *finding* and *characterizing* duality functions relating two given Markov processes. This means that, for a given pair of Markov generators, one wants to find all duality functions or, alternatively, a basis of the linear space of duality functions. See, for instance, in this direction [30] in the context of population genetics, while for particle systems the works [1,2,15,33] for symmetric and [4,5,34] for asymmetric processes. For Markov processes, algebraic constructions of duality relations for specific classes of models have also been provided (see e.g. [1,4,16,19,26]).

G. Giacomin et al. (Eds.): *Stochastic Dynamics Out of Equilibrium*, PROMS 282, pp. 621–649, 2019.
https://doi.org/10.1007/978-3-030-15096-9_25

In this paper we first show that, viewing a duality relation as a *spectral* relation among the associated Markov generators, duality functions can be obtained from linear combinations of products of eigenfunctions associated to a common eigenvalue. Secondly, we establish this connection with the general aim of characterizing all possible dualities in terms of eigenfunctions and generalized eigenfunctions of the generators involved. To this purpose, our discussion mainly focuses on continuous-time finite-state Markov chains for which no reversibility is assumed but canonical eigendecompositions of Jordan-type of the generators are available.

We emphasize that this connection between duality and eigenfunctions goes both ways: not only eigenfunctions of a shared spectrum give rise to duality functions, but also the existence of duality relations carries information about the spectrum of the generators. Here we can already see a clear distinction between the notion of *self-duality* and *integrability*: knowing certain linear combinations of products of eigenfunctions (self-duality) rather than knowing the eigenfunctions themselves (integrability).

The rest of the paper is organized as follows. In Sect. 2 we provide all preliminary notions of stochastic duality for continuous-time Markov chains. After an introductory study of self-duality and duality in the reversible setting in Sects. 3 and 4, in Sect. 5, via Jordan canonical decompositions, we make precise to which extent spectrum and eigenstructure of generators in duality are shared. In fact, the assumed orthonormality of the eigenfunctions in Sects. 3 and 4 has the only role of simplifying the exposition at a first reading. There, products of orthonormal eigenfunctions are a natural tensor basis w.r.t. which express duality functions; this fact allows a direct description of the linear subspace of duality functions in terms of this tensor basis. In Sect. 5, we show how, by dropping reversibility of the generators and thus orthonormality of the associated eigenfunctions, a tensor basis in terms of product of generalized eigenfunctions is always possible.

We further investigate the connection between eigenfunctions and particular instances of dualities that typically appear in the context of interacting particle systems, see e.g. [15,33], in Sects. 3 and 4. In Sect. 5.4 we revisit the notion of *intertwining* (see e.g. [21]) in this setting and provide an application to the symmetric exclusion process in Sect. 5.5. In Sect. 6 we provide an alternative way of proving and characterizing Siegmund duality [21,36] in the finite context.

2 Setting and Notation

Let Ω be a finite state space with cardinality $|\Omega| = n$. We consider an *irreducible* continuous-time Markov process $\{X_t,\ t \geq 0\}$ on Ω, with generator L given by

$$Lf(x) = \sum_{y \in \Omega} \ell(x, y)(f(y) - f(x)),$$

where $f : \Omega \to \mathbb{R}$ is a real-valued function and $\ell : \Omega \times \Omega \to [0, +\infty)$ gives the transition rates. For $x \in \Omega$, we define the exit rate from $x \in \Omega$ as

$$\ell(x) = \sum_{y \in \Omega \setminus \{x\}} \ell(x, y).$$

In the finite context we can identify L with the matrix, still denoted by L, given by

$$L(x, y) \;=\; \ell(x, y) \;\text{ for }\; x \neq y, \quad L(x, x) \;=\; -\ell(x).$$

Given two state spaces Ω, $\widehat{\Omega}$ of cardinalities $|\Omega| = n$, $|\widehat{\Omega}| = \widehat{n}$, and two *Markov processes* with generators L, $\widehat{L}$, we say that they are *dual* with *duality function* $D : \widehat{\Omega} \times \Omega \to \mathbb{R}$ if, for all $x \in \Omega$ and $\widehat{x} \in \widehat{\Omega}$, we have

$$\widehat{L}_{\text{left}} D(\widehat{x}, x) = L_{\text{right}} D(\widehat{x}, x), \tag{1}$$

where "left", resp. "right", refers to action on the left, resp. right, variable. If the laws of the two processes coincide, we speak about *self-duality*. The same notion in terms of matrix multiplication, where D also denotes the matrix with entries $\{D(\widehat{x}, x),\ \widehat{x} \in \widehat{\Omega},\ x \in \Omega\}$, is expressed as

$$\sum_{\widehat{y} \in \widehat{\Omega}} \widehat{L}(\widehat{x}, \widehat{y}) D(\widehat{y}, x) = \sum_{y \in \Omega} L(x, y) D(\widehat{x}, y),$$

or, shortly, as

$$\widehat{L} D = D L^{\mathsf{T}}, \tag{2}$$

where the symbol $^{\mathsf{T}}$ denotes *matrix transposition*, i.e., for a matrix A,

$$(A^{\mathsf{T}})(x, y) = A(y, x), \quad x, y \in \Omega.$$

More generally, we define two *operators* $\widehat{L}$ and L *dual* with duality function D if relation (1), or equivalently (2) in matrix notation, holds.

3 Self-duality from Eigenfunctions: Reversible Case

As in Sect. 2, let Ω be a finite set of cardinality $|\Omega| = n$, and let L be a generator of an irreducible *reversible* Markov process on Ω w.r.t. the positive measure μ. This measure then satisfies the detailed balance condition

$$\mu(x) L(x, y) = \mu(y) L(y, x), \tag{3}$$

for all $x, y \in \Omega$. This relation can be rewritten as a self-duality with self-duality function the so-called *cheap self-duality function*:

$$D_{\text{cheap}}(x, y) = \frac{\delta_{x,y}}{\mu(y)}. \tag{4}$$

The reversibility of μ implies that L is self-adjoint in $L^2(\mu)$ and, as a consequence, there exists a basis $\{u_1, \ldots, u_n\}$ of eigenfunctions of L with $u_1(x) = 1/\sqrt{n}$ corresponding to eigenvalue zero and $\{u_1, \ldots, u_n\}$ orthonormal, i.e., $\langle u_i, u_j \rangle_\mu = \delta_{i,j}$ where $\langle \cdot, \cdot \rangle_\mu$ denotes inner product in $L^2(\mu)$. We denote by $\{\lambda_1, \ldots, \lambda_n\}$ the corresponding real eigenvalues with

$$0 \;=\; \lambda_1 \;> \lambda_2 \;\geq\; \ldots \;\geq\; \lambda_n.$$

The following proposition then shows how to obtain and characterize self-duality functions in terms of this orthonormal system. The last statement recovers an earlier result from [16].

Proposition 1. *(i) For $a_1, a_2 \ldots, a_n \in \mathbb{R}$, the function*

$$D(x,y) = \sum_{i=1}^{n} a_i u_i(x) u_i(y) \tag{5}$$

is a self-duality function.

(ii) Every self-duality function has a unique decomposition of the form

$$D(x,y) = \sum_{i,j:\lambda_i = \lambda_j} a_{ij} u_i(x) u_j(y). \tag{6}$$

(iii) If a function of the form $D(x,y) = f(x)g(y)$ is a non-zero self-duality function, then f and g are eigenfunctions corresponding to the same eigenvalue.

(iv) The $L^2(\mu)$ inner product of self-duality functions produces self-duality functions, i.e., if D and D' are self-duality functions, then

$$\langle D(x,\cdot), D'(x',\cdot) \rangle_\mu = D''(x,x') \tag{7}$$

defines a self-duality function D''.

Proof. For *(i)*, by definition of eigenfunction $Lu_i = \lambda_i u_i$ with $\lambda_i \in \mathbb{R}$, we obtain

$$\begin{aligned} L_{\text{left}} D(x,y) &= \sum_{i=1}^{n} a_i L u_i(x) u_i(y) = \sum_{i=1}^{n} a_i \lambda_i u_i(x) u_i(y) \\ &= \sum_{i=1}^{n} a_i u_i(x) \lambda_i u_i(y) = \sum_{i=1}^{n} a_i u_i(x) L u_i(y) = \; L_{\text{right}} D(x,y), \end{aligned}$$

hence (1).

For *(ii)*, start by noticing that every function $D : \Omega \times \Omega \to \mathbb{R}$ can be written in a unique way as

$$D(x,y) = \sum_{i,j=1}^{n} a_{i,j} u_i(x) u_j(y),$$

Now using the duality relation (1), it follows that

$$\sum_{i,j} a_{i,j} \lambda_i u_i(x) u_j(y) = \sum_{i,j} a_{i,j} \lambda_j u_i(x) u_j(y),$$

which implies that, for all $i, j = 1, \ldots, n$,

$$a_{i,j}\lambda_i \;=\; a_{i,j}\lambda_j.$$

For item *(iii)*, first write

$$f(x)g(y) = \sum_{i,j=1}^{n} a_{ij}u_i(x)u_j(y).$$

Then we find $a_{ij} = \langle f, u_i\rangle_\mu\langle g, u_j\rangle_\mu =: \alpha_i\beta_j$. From self-duality we conclude, for all $i, j = 1, \ldots, n$,

$$\alpha_i\beta_j(\lambda_i - \lambda_j) = 0.$$

Now use that $f(x)g(y)$ is not identically zero to conclude that there exists i with $\alpha_i \neq 0$. Then if $\lambda_j \neq \lambda_i$ we conclude $\beta_j = 0$, which implies that g is an eigenfunction with eigenvalue λ_i. Because g is not identically zero, we can reverse the argument and conclude.

For *(iv)*, by exchanging the order of summations and using $\langle u_j, u_l\rangle_\mu = \delta_{j,l}$, the l.h.s. of (7) reads

$$\begin{aligned}
&\sum_{y\in\Omega} D(x,y)D(x',y)\mu(y)\\
&= \sum_{y\in\Omega}\left(\sum_{i,j:\lambda_i=\lambda_j} a_{i,j}u_i(x)u_j(y)\right)\left(\sum_{k,l:\lambda_k=\lambda_l} a_{k,l}u_k(x')u_l(y)\right)\mu(y)\\
&= \sum_{j=1}^{n}\left(\sum_{i:\lambda_i=\lambda_j} a_{i,j}u_i(x)\right)\left(\sum_{k:\lambda_k=\lambda_j} a_{k,j}u_k(x')\right).
\end{aligned}$$

By noting that, for all $j = 1, \ldots, n$, the function $u'_j = \sum_{i:\lambda_i=\lambda_j} a_{i,j}u_i$ is either vanishing or is an eigenfunction of L associated to λ_j, the proof is concluded. □

In the next propositions we study particular instances of self-duality functions. More precisely, by using Proposition 1, we recover the cheap self-duality function in (4), while in Proposition 3 we characterize *orthogonal* self-duality functions (cf. (11)–(12) below).

Proposition 2 (Cheap self-duality)

(i) For the choice $a_1 = a_2 = \ldots = a_n = 1$ in (5), we obtain the cheap self-duality function, i.e.,

$$D_{cheap}(x,y) \;=\; \frac{\delta_{x,y}}{\mu(y)} \;=\; \sum_{i=1}^{n} u_i(x)u_i(y). \tag{8}$$

(ii) Conversely, if $\{v_1, \dots, v_n\}$ *is a basis of* $L^2(\mu)$ *and satisfies*

$$\sum_{i=1}^{n} v_i(x) v_i(y) = \frac{\delta_{x,y}}{\mu(y)} \tag{9}$$

for all $x, y \in \Omega$, *then* $\{v_1, \dots, v_n\}$ *is an orthonormal basis of* $L^2(\mu)$.

Proof. To show (8), by the positivity of μ, we need to show that, for all $f : \Omega \to \mathbb{R}$ and $x \in \Omega$,

$$\sum_{y \in \Omega} \sum_{i=1}^{n} u_i(x) u_i(y) \mu(y) f(y) \;=\; f(x).$$

Now note, by interchanging the sum over i with the sum over y, that the l.h.s. equals

$$\sum_{i=1}^{n} u_i(x) \langle u_i, f \rangle_\mu \;=\; f(x),$$

and hence we obtain *(i)*.

For *(ii)* we need to show that for all $f : \Omega \to \mathbb{R}$ and $x \in \Omega$

$$f(x) \;=\; \sum_{i=1}^{n} v_i(x) \langle v_i, f \rangle_\mu \;=\; \sum_{i=1}^{n} \sum_{y \in \Omega} v_i(x) v_i(y) f(y) \mu(y). \tag{10}$$

We conclude by interchanging the order of the two summations in the r.h.s. above and using (9), we indeed obtain (10). □

Remark that the cheap self-duality function is the only, up to multiplicative constants, *diagonal* self-duality, and that it is *orthogonal* in the sense that, for all $x, x' \in \Omega$,

$$\langle D_{\text{cheap}}(x, \cdot), D_{\text{cheap}}(x', \cdot) \rangle_\mu = \delta_{x,x'} \, \langle D_{\text{cheap}}(x, \cdot), D_{\text{cheap}}(x, \cdot) \rangle_\mu, \tag{11}$$

and similarly, for all $y, y' \in \Omega$,

$$\langle D_{\text{cheap}}(\cdot, y), D_{\text{cheap}}(\cdot, y') \rangle_\mu = \delta_{y,y'} \, \langle D_{\text{cheap}}(\cdot, y), D_{\text{cheap}}(\cdot, y) \rangle_\mu. \tag{12}$$

The next proposition shows how to find *all* orthogonal self-duality functions.

Proposition 3 (Orthogonal self-duality)

(i) If $\{\tilde{u}_1, \dots, \tilde{u}_n\}$ *is an orthonormal system in* $L^2(\mu)$ *of eigenfunctions of* L, *corresponding to the same eigenvalues* $\{\lambda_1, \dots, \lambda_n\}$, *then*

$$D(x, y) = \sum_{i=1}^{n} \tilde{u}_i(x) u_i(y) \tag{13}$$

is an orthogonal self-duality function. More precisely, for all $x, x' \in \Omega$,

$$\langle D(x, \cdot), D(x', \cdot) \rangle_\mu = \frac{\delta_{x,x'}}{\mu(x')}. \tag{14}$$

(ii) The self-duality functions of the form (13) are the only, up to a multiplicative factor, orthogonal self-duality functions.

Proof. For *(i)*, we compute, for all $k = 1, \dots, n$ and $x \in \Omega$, the following quantity

$$\sum_{x' \in \Omega} \langle D(x, \cdot), D(x', \cdot) \rangle_\mu \tilde{u}_k(x') \mu(x').$$

By $\langle u_i, u_j \rangle_\mu = \langle \tilde{u}_i, \tilde{u}_j \rangle_\mu = \delta_{i,j}$, the line above rewrites as follows:

$$\begin{aligned}
&\sum_{x' \in \Omega} \sum_{y \in \Omega} \left(\sum_{i=1}^n \tilde{u}_i(x) u_i(y) \right) \left(\sum_{j=1}^n \tilde{u}_j(x') u_j(y) \right) \mu(y) \tilde{u}_k(x') \mu(x') \\
&= \sum_{i=1}^n \sum_{j=1}^n \tilde{u}_i(x) \left(\sum_{y \in \Omega} u_i(y) u_j(y) \mu(y) \right) \left(\sum_{x' \in \Omega} \tilde{u}_j(x') \tilde{u}_k(x') \mu(x') \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n \tilde{u}_i(x) \delta_{i,j} \delta_{j,k} \;=\; \tilde{u}_k(x).
\end{aligned}$$

This together with Proposition 2 concludes the proof of part *(i)*.

For *(ii)*, by starting from a general self-duality function

$$D(x, y) = \sum_{i,j : \lambda_i = \lambda_j} a_{i,j} u_i(x) u_j(y),$$

the l.h.s. of (14) rewrites as

$$\sum_{j=1}^n u'_j(x) u'_j(x'),$$

where $\{u'_1, \dots, u'_n\}$ is defined as

$$u'_j(x) = \sum_{i : \lambda_i = \lambda_j} a_{i,j} u_i(x).$$

By remarking that either $u'_j = 0$ or u'_j is an eigenfunction of L associated to λ_j and applying Proposition 2, we have that

$$\langle u'_i, u'_j \rangle_\mu = \delta_{i,j},$$

and that the self-duality function D has the form (13) with $\tilde{u}_i = u'_i$. □

4 Duality from Eigenfunctions: Reversible Case

Now we consider two generators L, $\widehat{L}$ on the same finite state space Ω with reversible measures μ, $\widehat{\mu}$ respectively, and orthonormal systems of eigenfunctions $\{u_1, \dots, u_n\}$, $\{\widehat{u}_1, \dots, \widehat{u}_n\}$ corresponding to the *same* real eigenvalues

$\{\lambda_1, \ldots, \lambda_n\}$, i.e., we assume that L and $\widehat{L}$ are self-adjoint in $L^2(\mu)$, resp. in $L^2(\widehat{\mu})$, and that they are iso-spectral.

In what follows we state - without proofs - analogous relations between duality functions and orthonormal systems of eigenfunctions of L and $\widehat{L}$.

Proposition 4. *(i) For $a_1, \ldots, a_n \in \mathbb{R}$ the function*

$$D(\widehat{x}, x) = \sum_{i=1}^{n} a_i \widehat{u}_i(\widehat{x}) u_i(x)$$

is a duality function for duality between $\hat{L}$ and L.

(ii) Every duality function has a unique decomposition of the form

$$D(\widehat{x}, x) = \sum_{i,j:\lambda_i=\lambda_j} a_{ij} \widehat{u}_i(\widehat{x}) u_j(x).$$

(iii) If a function of the form $D(\widehat{x}, x) = f(\widehat{x})g(x)$ is a non-zero duality function, then f and g are eigenfunctions of $\widehat{L}$, resp. L, corresponding to the same eigenvalue.

(iv) The $L^2(\mu)$ and $L^2(\widehat{\mu})$ inner products of duality functions produce self-duality functions, i.e., if D and D' are duality functions, then

$$\langle D(\widehat{x}, \cdot), D'(\widehat{x}', \cdot)\rangle_\mu = \widehat{D}(\widehat{x}, \widehat{x}')$$

defines a self-duality function $\widehat{D}$ for $\widehat{L}$, and similarly

$$\langle D(\cdot, x), D'(\cdot, x')\rangle_{\widehat{\mu}} = \widetilde{D}(x, x')$$

determines a self-duality function $\widetilde{D}$ for L.

Proposition 5 (Orthogonal duality)

(i) If $\{\tilde{u}_1, \ldots, \tilde{u}_n\}$ is an orthonormal system in $L^2(\widehat{\mu})$ of eigenfunctions of $\widehat{L}$ corresponding to the same eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, then

$$D(\widehat{x}, x) = \sum_{i=1}^{n} \tilde{u}_i(\widehat{x}) u_i(x) \tag{15}$$

is an orthogonal duality function, i.e.,

$$\langle D(\widehat{x}, \cdot), D(\widehat{x}', \cdot)\rangle_\mu = \frac{\delta_{\widehat{x},\widehat{x}'}}{\widehat{\mu}(\widehat{x}')}$$

and

$$\langle D(\cdot, x), D(\cdot, x')\rangle_{\widehat{\mu}} = \frac{\delta_{x,x'}}{\mu(x')}.$$

(ii) These are the only, up to multiplicative constants, orthogonal dualities between $\widehat{L}$ and L.

5 Duality from Eigenfunctions: Non-reversible Case

Working in the *non-reversible* context, i.e., whenever there *does not* exist a probability measure μ on Ω for which the generator L is self-adjoint in $L^2(\mu)$, a spectral decomposition of the generator in terms of real non-positive eigenvalues and orthonormal real eigenfunctions is typically lost. In recent years, the study of the eigendecomposition of non-reversible generators has received an increasing attention [6–8,32,37] and duality-related notions have been introduced to relate spectral information of one process, typically a reversible one, to another, typically non-reversible [14,29].

However, regardless of the spectral eigendecomposition of the generators, in principle interesting dualities can still be constructed from eigenfunctions, either real or complex, and generalized eigenfunctions of the generators involved. The key on which this relation builds up, in the finite context, is the *Jordan canonical decomposition* of the generators. A relation between duality and the Jordan canonical decomposition has already been used in the context of models of population dynamics in [30].

Below, before studying the most general result that exploits the Jordan form of the generators, we treat some special cases reminiscent of the previous sections. In the sequel, for a function $u : \Omega \to \mathbb{C}$, we denote by $u^* : \Omega \to \mathbb{C}$ its complex conjugate.

5.1 Duality from Complex Eigenfunctions

A first feature that typically drops as soon as one moves to the non-reversible situation is the appearance of only real eigenvalues. Indeed, given a non-reversible generator L of an irreducible Markov process on Ω, pairs of complex conjugates eigenvalues $\{\lambda, \lambda^*\}$ and eigenfunctions $\{u, u^*\}$ may arise as in the following example.

Example 1. The continuous-time Markov chain on the state space $\Omega = \{1, 2, 3\}$ and described by the generator L, which, viewed as a matrix, reads

$$L = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix},$$

represents a basic example of this situation. Indeed, the Markov chain is irreducible, the eigenvalues $\{\lambda_1, \lambda_2, \lambda_3\}$ are

$$\lambda_1 = 0, \quad \lambda_2 = \lambda_3^* = -\frac{3}{2} + i\frac{\sqrt{3}}{2},$$

while the associated eigenfunctions $\{u_1, u_2, u_3\}$ are, for $x \in \{1, 2, 3\}$,

$$u_1(x) = \frac{1}{\sqrt{3}}, \quad u_2(x) = u_3^*(x) = e^{(i\frac{2}{3}\pi)x}.$$

□

Let us, thus, consider two irreducible non-reversible generators L, $\widehat{L}$ on the same state space Ω. We investigate the situation in which there exist $\lambda \in \mathbb{C}\backslash\mathbb{R}$ and functions $u, \widehat{u} : \Omega \to \mathbb{C}$ such that

$$Lu \;=\; \lambda u, \qquad \widehat{L}\widehat{u} \;=\; \lambda \widehat{u}. \tag{16}$$

Remark that, as L, $\widehat{L}$ are real operators, this implies that

$$Lu^* \;=\; \lambda^* u^*, \qquad \widehat{L}\widehat{u}^* \;=\; \lambda^* \widehat{u}^*. \tag{17}$$

A real duality function arising from a shared pair of complex eigenvalues is obtained in the following proposition.

Proposition 6. *For $a \in \mathbb{R}$, the function*

$$D(\widehat{x}, x) = a\widehat{u}(\widehat{x})u(x) + a\widehat{u}^*(\widehat{x})u^*(x)$$

takes values in $\mathbb{R}$ and is a duality function between $\widehat{L}$ and L.

Proof. It is clear that $D(\widehat{x}, x)$ is in $\mathbb{R}$. Then, by using (16) and (17), we obtain

$$\begin{aligned}
\widehat{L}_{\text{left}} D(\widehat{x}, x) \;&=\; a(\widehat{L}\widehat{u})(\widehat{x})u(x) + a(\widehat{L}\widehat{u}^*)(\widehat{x})u^*(x) \\
&=\; a\lambda\widehat{u}(\widehat{x})u(x) + a\lambda^*\widehat{u}^*(\widehat{x})u^*(x) \;=\; a\widehat{u}(\widehat{x})\lambda u(x) + a\widehat{u}^*(\widehat{x})\lambda^* u^*(x) \\
&=\; a\widehat{u}(\widehat{x})(Lu)(x) + a\widehat{u}^*(\widehat{x})(Lu^*)(x) \;=\; L_{\text{right}} D(\widehat{x}, x).
\end{aligned}$$

□

5.2 Duality from Generalized Eigenfunctions

A second feature that may be lacking is the existence of a linear independent system of eigenfunctions. However, if L is an irreducible non-reversible generator on the state space Ω with real non-negative eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, there always exists a linearly independent system of so-called *generalized eigenfunctions*, i.e., for each eigenvalue λ_i, there exists a set of linearly independent functions $\{u_i^{(1)}, \ldots, u_i^{(m_i)}\}$ such that $m_i \leq n$,

$$Lu_i^{(1)} = \lambda_i u_i^{(1)}$$

and, for $1 < k \leq m_i$,

$$Lu_i^{(k)} = \lambda_i u_i^{(k)} + u_i^{(k-1)}.$$

We refer to $u_i^{(k)}$ as the *k-th order generalized eigenfunction* associated to λ_i. Moreover, if $\lambda_i \neq \lambda_j$, then the set $\{u_i^{(1)}, \ldots, u_i^{(m_i)}, u_j^{(1)}, \ldots, u_j^{(m_j)}\}$ is linearly independent and any arbitrary function $f : \Omega \to \mathbb{R}$ can be written as linear combination of functions in $\{u_i^{(k)},\ i = 1, \ldots, n;\ k = 1, \ldots, m_i\}$.

Example 2. The irreducible generator L on the state space $\Omega = \{1, 2, 3, 4\}$ given by

$$L = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & -1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & -1 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & -1 \end{pmatrix},$$

represents a basic example of this situation. Indeed, the eigenvalue $\lambda = -1$ has $u^{(1)}$ given by

$$u^{(1)}(x) \;=\; \frac{(-1)^x}{2}, \quad x \in \{1, 2, 3, 4\},$$

as eigenfunction and

$$u^{(2)}(x) \;=\; \cos\left(\frac{\pi}{2}(x+1)\right), \quad x \in \{1, 2, 3, 4\},$$

as a second order generalized eigenfunction, i.e.,

$$Lu^{(2)} \;=\; -u^{(2)} + u^{(1)}.$$

□

In this situation, in case of two generators L, $\widehat{L}$ sharing a real eigenvalue λ with associated generalized eigenfunctions $\{u^{(1)}, \ldots, u^{(m)}\}$, $\{\widehat{u}^{(1)}, \ldots, \widehat{u}^{(m)}\}$, the main idea is that a duality function is readily constructed from sums of products of generalized eigenfunctions whose order is, nevertheless, reversed. This connection is the content of the following proposition.

Proposition 7. *The function*

$$D(\widehat{x}, x) = \sum_{k=1}^{m} \widehat{u}^{(k)}(\widehat{x}) u^{(m+1-k)}(x)$$

is a duality function between $\widehat{L}$ *and* L.

Proof. By using the definition of k-th order generalized eigenfunction, we obtain

$$\begin{aligned} \widehat{L}_{\text{left}} D(\widehat{x}, x) \;&=\; \sum_{k=1}^{m} (\widehat{L}\widehat{u}^{(k)})(\widehat{x}) u^{(m+1-k)}(x) \\ &= \sum_{k=1}^{m} \lambda \widehat{u}^{(k)}(\widehat{x}) u^{(m+1-k)} + \sum_{k=2}^{m} \widehat{u}^{(k-1)}(\widehat{x}) u^{(m+1-k)}(x) \\ &= \sum_{k=1}^{m} \lambda \widehat{u}^{(k)}(\widehat{x}) u^{(m+1-k)} + \sum_{k=1}^{m-1} \widehat{u}^{(k)}(\widehat{x}) u^{(m-k)}(x) \\ &= \sum_{k=1}^{m} \widehat{u}^{(k)}(\widehat{x}) (L u^{(m+1-k)})(x) = L_{\text{right}} D(\widehat{x}, x). \end{aligned}$$

□

5.3 Duality and the Jordan Canonical Decomposition: General Case

In this section we provide a general framework that allows us to cover all instances of duality encountered so far in the finite setting. The standard strategy of decomposing generators - viewed as matrices - into their Jordan canonical form builds a bridge between dualities and spectral information of the generators involved. In particular, this linear algebraic approach is useful for the problem of *existence* and *characterization* of duality functions: on one side, the existence of a Jordan canonical decomposition for any generator leads, for instance, to the existence of self-dualities; on the other side, dualities between generators carry information about a common, at least partially, spectral structure of the generators.

Before stating the main result, we introduce some notation. Given a generator L on the state space Ω with cardinality $|\Omega| = n$, L is in *Jordan canonical form* if it can be written as

$$L = UJU^{-1},$$

where $J \in \mathbb{C}^{n\times n}$ is the *unique*, up to permutations, *Jordan matrix* [20, Definition 3.1.1] associated to L and $U \in \mathbb{C}^{n\times n}$ is an invertible matrix. Recall that columns $\{u_1, \dots, u_n\}$ of U consists of (possibly generalized) eigenfunctions of L, while the rows $\{w_1, \dots, w_n\}$ of U^{-1} the (possibly generalized) eigenfunctions of L^{T}, chosen in such a way that

$$\langle w_i, u_j\rangle = \sum_{x\in\Omega} w_i(x)u_i^*(x) = \delta_{i,j}.$$

For all Jordan matrices $J \in \mathbb{C}^{n\times n}$ of the form

$$J = \begin{pmatrix} J_{m_1}(\lambda_1) & & \cdots & 0 \\ & J_{m_2}(\lambda_n) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & J_{m_k}(\lambda_k) \end{pmatrix},$$

with $m_1+\ldots+m_k = n$ and *Jordan blocks* $J_m(\lambda)$ of size m associated to eigenvalue $\lambda \in \mathbb{C}$, we define the matrix $B_J \in \mathbb{R}^{n\times n}$ as follows

$$B_J = \begin{pmatrix} H_{m_1} & & \cdots & 0 \\ & H_{m_2} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & H_{m_k} \end{pmatrix},$$

where, for all $m \in \mathbb{N}$, the matrix $H_m \in \mathbb{R}^{m\times m}$ is defined as

$$H_m = \begin{pmatrix} 0 & \cdots & & 1 \\ \vdots & & \iddots & \\ & \iddots & & \vdots \\ 1 & & \cdots & 0 \end{pmatrix},$$

i.e., in such a way that $B_J^{\mathsf{T}} = B_J^{-1} = B_J$ and $JB_J = B_J J^{\mathsf{T}}$. Moreover, we say that two matrices $L \in \mathbb{R}^{n\times n}$, $\widehat{L} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ are *r-similar* for some $r = 1, \dots, \min\{n, \widehat{n}\}$ if there exist Jordan canonical forms

$$L \;=\; UJU^{-1}, \qquad \widehat{L} \;=\; \widehat{U}\widehat{J}\widehat{U}^{-1}, \tag{18}$$

matrices $S_r \in \mathbb{R}^{\widehat{n}\times n}$ and $I_r \in \mathbb{R}^{r\times r}$ of the form

$$S_r \;=\; \begin{pmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad I_r \;=\; \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix},$$

and permutation matrices $\widehat{P} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ and $P \in \mathbb{R}^{n\times n}$ such that

$$T_r = \widehat{P} S_r P$$

and

$$\widehat{J}T_r = T_r J. \tag{19}$$

Of course, if two matrices are r-similar, then they are necessarily r'-similar, for all $r' = 1, \dots, r$ and if $r = n = \widehat{n}$ then we simply say that they are *similar*.

In the following theorem we establish a general connection between duality relations and Jordan canonical forms for generators L, $\widehat{L}$.

Theorem 1. *The following statements are equivalent:*

(i) There exists a duality function $D(\widehat{x}, x)$ of rank r between $\widehat{L}$ and L.
(ii) L and $\widehat{L}$ are r-similar*.*

If either condition holds, any duality function is of the form

$$D = \widehat{U} T_r B_J U^{\mathsf{T}}. \tag{20}$$

In particular if $L = \widehat{L}$, for any $r = 1, \dots, n$, there always exists a self-duality function D of rank r and it must be of the form (20).

Proof. We start with proving that *(ii)* implies *(i)*. By using the property of r-similarity (19) with Jordan decompositions as in (18), with the choice (20) of the candidate duality function D, we obtain

$$\widehat{L}\widehat{U}T_r B_J U^{\mathsf{T}} \;=\; \widehat{U}\widehat{J}T_r B_J U^{\mathsf{T}} \;=\; \widehat{U}T_r J B_J U^{\mathsf{T}} \;=\; \widehat{U}T_r B_J J^{\mathsf{T}} U^{\mathsf{T}} \;=\; \widehat{U}T_r B_J U^{\mathsf{T}} L^{\mathsf{T}},$$

i.e., the duality relation (2) in matrix form.

For the other implication, as the matrices U, $\widehat{U}$ in (18) and B_J are invertible, the following chains of identities are equivalent:

$$\begin{aligned} \widehat{L}D \;=\; DL^{\mathsf{T}} &\iff \widehat{U}\widehat{J}\widehat{U}^{-1}D \;=\; D(U^{-1})^{\mathsf{T}} J^{\mathsf{T}} U^{\mathsf{T}} \\ &\iff \widehat{J}\widehat{U}^{-1}D(U^{-1})^{\mathsf{T}} \;=\; \widehat{U}^{-1}D(U^{-1})^{\mathsf{T}} J^{\mathsf{T}} \\ &\iff \widehat{J}\widehat{U}^{-1}D(U^{-1})^{\mathsf{T}} B_J \;=\; \widehat{U}^{-1}D(U^{-1})^{\mathsf{T}} B_J J. \end{aligned}$$

Moreover, if D has rank r, then $\widehat{U}^{-1}D(U^{-1})^{\mathsf{T}}B_J$ must have rank r as well. The last relation is of the form

$$\widehat{J}A = AJ,$$

where $A = \widehat{U}^{-1}D(U^{-1})^{\mathsf{T}}B_J$ is a matrix of rank r. Therefore, we conclude that there exists a permutation matrix $P \in \mathbb{R}^{n\times n}$ such that

$$\widehat{J}S_r = S_r P J P^{-1},$$

i.e., L and $\widehat{L}$ are r-similar according to the Jordan canonical decompositions

$$L = \widetilde{U}\widetilde{J}\widetilde{U}^{-1}, \qquad \widehat{L} = \widehat{U}\widehat{J}\widehat{U}^{-1},$$

with $\widetilde{U} = UP^{-1}$ and $\widetilde{J} = PJP^{-1}$. □

Remark 1. (a) In words, the theorem above states that there exists a rank-r duality matrix if and only if the generators $\widehat{L}$ and L have r eigenvalues (with multiplicities) in common with "compatible" structure of eigenspaces. Additionally, Eq. (20) provides the most general form of the duality function D in terms of matrices U, $\widehat{U}$. In particular, if J is *diagonal* (i.e., B_J is the identity matrix) *all* duality functions $D(\widehat{x}, x)$ of rank r read as

$$D(\widehat{x}, x) = \sum_{i=1}^{r} a_i \widehat{u}_i(\widehat{x}) u_i(x),$$

for $a_1, \ldots, a_n \in \mathbb{R} \setminus \{0\}$, given $\{u_1, \ldots, u_n\}$, $\{\widehat{u}_1, \ldots, \widehat{u}_{\widehat{n}}\}$ are the columns of U, $\widehat{U}$, invertible matrices in the Jordan decompositions (18) satisfying (19) with $T_r = S_r$. Note the analogy with the duality function described in Propositions 1, 4 and 6. If J is *non-diagonal*, all duality functions D have a similar form up to some index permutations as in Proposition 7.

(b) We note that the *constant duality function* is always a trivial duality function between any two generators L, $\widehat{L}$ on Ω, $\widehat{\Omega}$. Indeed, $\lambda = 0$ is always an eigenvalue for both L and $\widehat{L}$ with associated constant eigenfunctions $u : \Omega \to \mathbb{R}$, $\widehat{u} : \widehat{\Omega} \to \mathbb{R}$, i.e., for all $x \in \Omega$ and $\widehat{x} \in \widehat{\Omega}$,

$$u(x) = 1, \quad \widehat{u}(\widehat{x}) = 1,$$

are eigenfunctions for L, $\widehat{L}$ associated to $\lambda = 0$.

(c) Another consequence, as already mentioned in [18], is that in the finite context *self-duality* functions *always* exist. In fact, a generator L, viewed as a matrix, is always *similar* to itself. Hence, viewing duality relations between generators as similarity relations among matrices allows one to transfer statements about *existence* of Jordan canonical decompositions to statements regarding the *existence* of duality relations, even when neither any explicit formula of the duality functions nor reversible measures for the processes are known. However, Theorem 1 above provides information on how to construct any self-duality matrix. Indeed, given any two Jordan decompositions of L, say

$$LU = UJ, \qquad L\widetilde{U} = \widetilde{U}J,$$

the matrix D constructed from $U, \widetilde{U}$ and J as in (20), namely

$$D \;=\; \widetilde{U} B_J U^{\mathsf{T}}, \tag{21}$$

turns out to be a self-duality function for L and, viceversa, any self-duality matrix D for L is of the form (21).

Typically, to find the eigenvalues and eigenfunctions of the generator associated to a Markov chain is a much more challenging task than establishing duality relations. However, we have seen that the knowledge of the eigenfunctions leads to a full characterization of duality and/or self-duality functions. This is, indeed, the case of the example below, in which we exploit the knowledge of eigenfunctions of two generators to characterize the family of self-duality and duality functions.

Example 3 (One-dimensional symmetric random walks on a finite grid). Let us introduce the symmetric random walk on $\Omega = \{1, \dots, n\}$ *reflected on the left* and *absorbed on the right*. We describe the action of the generator L on functions $f : \Omega \to \mathbb{R}$ as

$$Lf(x) = (f(x+1) - f(x)) + (f(x-1) - f(x)), \quad x \in \Omega \setminus \{1, n\},$$

while for $x \in \{1, n\}$ we have

$$Lf(1) \;=\; 2(f(2) - f(1)), \qquad Lf(n) \;=\; 0.$$

Similarly, we denote by $\widehat{L}$ the generator of the symmetric random walk on Ω *reflected on the right* and *absorbed on the left*. Namely,

$$\widehat{L}f(x) = (f(x+1) - f(x)) + (f(x-1) - f(x)), \quad x \in \Omega \setminus \{1, n\},$$

and

$$\widehat{L}f(1) \;=\; 0, \qquad \widehat{L}f(n) \;=\; 2(f(n-1) - f(n)).$$

As an application of Theorem 1, we prove the following dualities: *self-duality* of L, *self-duality* of $\widehat{L}$ and *duality* between L and $\widehat{L}$. The key is to explicitly find eigenvalues and eigenfunctions of the generators. Indeed, the eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ of L and $\widehat{L}$ read as follows:

$$\lambda_1 \;=\; 0, \quad \lambda_i = 2(\cos(\theta_i) - 1), \quad \theta_i \;=\; \frac{i - \frac{1}{2}}{n-1}\pi, \quad i = 2, \dots, n. \tag{22}$$

The eigenfunctions $\{u_1, \dots, u_n\}$ of L are, for $x \in \Omega$,

$$u_1(x) \;=\; \frac{1}{\sqrt{n}}, \quad u_i(x) \;=\; \frac{1}{\sqrt{n}} \cos(\theta_i(x-1)), \quad i = 2, \dots, n,$$

while the eigenfunctions $\{\widehat{u}_1, \dots, \widehat{u}_n\}$ of $\widehat{L}$ are, for $x \in \Omega$,

$$\widehat{u}_1(\widehat{x}) \;=\; \frac{1}{\sqrt{n}}, \quad \widehat{u}_i(\widehat{x}) \;=\; \frac{1}{\sqrt{n}} \sin(\theta_i(\widehat{x}-1)), \quad i = 2, \dots, n.$$

Hence, we conclude the following:

(a) *Self-duality functions for L.* For all values $a_1, \dots, a_n \in \mathbb{R}$, the function

$$D(x,y) \;=\; \sum_{i=1}^{n} a_i u_i(x) u_i(y) \;=\; \frac{a_1}{n} + \sum_{i=2}^{n} \frac{a_i}{n} \cos(\theta_i(x-1)) \cos(\theta_i(y-1)) \quad (23)$$

is a self-duality function for L and all self-duality functions are of this form.

(b) *Self-duality functions for $\widehat{L}$.* For all $a_1, \dots, a_n \in \mathbb{R}$,

$$\widehat{D}(\widehat{x},\widehat{y}) \;=\; \sum_{i=1}^{n} a_i \widehat{u}_i(\widehat{x}) \widehat{u}_i(\widehat{y}) \;=\; \frac{1}{n} + \sum_{i=2}^{n} \frac{a_i}{n} \sin(\theta_i(\widehat{x}-1)) \sin(\theta_i(\widehat{y}-1)) \quad (24)$$

is a self-duality function for $\widehat{L}$ and all self-duality functions are of this form.

(c) *Duality functions between L and $\widehat{L}$.* For all $a_1, \dots, a_n \in \mathbb{R}$,

$$D'(\widehat{x},x) \;=\; \frac{a_1}{n} + \sum_{i=2}^{n} \frac{a_i}{n} \sin(\theta_i(\widehat{x}-1)) \cos(\theta_i(x-1)), \quad (25)$$

is a duality function between L and $\widehat{L}$ and all duality functions are of this form. □

We can now provide an analogue of Proposition 2 beyond the reversible context. To fix notation, let L be a generator on Ω, with $|\Omega| = n$. Lacking reversibility, we have seen that complex eigenvalues and generalized eigenfunctions of the generator may arise. However, in the irreducible case, i.e., in case there exists a unique stationary measure $\mu > 0$ for which the adjoint of L in $L^2(\mu)$, say $L^\dagger$, is itself a generator, a trivial duality relation between L and $L^\dagger$ is available. Indeed, from the adjoint relation

$$\langle L^\dagger f, g\rangle_{L^2(\mu)} = \langle f, Lg\rangle_{L^2(\mu)}, \quad f, g : \Omega \to \mathbb{R},$$

it follows that the diagonal function $D : \Omega \times \Omega \to \mathbb{R}$ given by

$$D(x,y) = \frac{\delta_{x,y}}{\mu(y)}, \quad x, y \in \Omega, \quad (26)$$

is a duality function for $L^\dagger$, L. In analogy with (4), we refer to it as *cheap duality function*, also $D = D_{\text{cheap}}$.

From Theorem 1, the above duality tells us that, beside the fact that the generators L and $L^\dagger$ are indeed similar as matrices, the cheap duality function D_{cheap} in (26) should be represented in terms of functions $\{u_1, \dots, u_n\}$ and $\{\widetilde{u}_1, \dots, \widetilde{u}_n\}$, which, up to suitably reordering, are indeed the generalized eigenfunctions of L and $L^\dagger$, respectively.

As a consequence of the following lemma, which we use in the proof of Theorem 5, we obtain that a relation of *bi-orthogonality* w.r.t. μ among the generalized eigenfunctions of L and those of $L^\dagger$ can be derived from the duality w.r.t. D_{cheap}. For the proof, we refer back to the proof of Proposition 2.

Proposition 8. *Let L be a generator, μ a positive measure on Ω (not necessarily stationary for L) and let $L^\dagger$ be the adjoint operator of L in $L^2(\mu)$. Let the spans of the generalized eigenfunctions of L and $L^\dagger$, say $\{u_1, \dots, u_n\}$ and $\{\widetilde{u}_1, \dots, \widetilde{u}_n\}$, both coincide with $L^2(\mu)$. Then the following statements are equivalent:*

(i) Cheap duality from generalized eigenfunctions. *For $x, y \in \Omega$,*

$$\sum_{i=1}^{n} \widetilde{u}_i(x) u_i(y) = \frac{\delta_{x,y}}{\mu(y)}.$$

(ii) Bi-orthogonality of generalized eigenfunctions. *For all $i, j = 1, \dots, n$,*

$$\langle \widetilde{u}_i, u_j^* \rangle_\mu = \sum_{x' \in \Omega} \widetilde{u}_i(x') u_j(x') \mu(x') \;=\; \delta_{i,j}. \tag{27}$$

Two families $\{u_1^*, \dots, u_n^*\}$, $\{\widetilde{u}_1, \dots, \widetilde{u}_n\}$ satisfying condition (27) are also said to be bi-orthogonal *w.r.t. the measure μ.*

5.4 Intertwining Relations, Duality and Generalized Eigenfunctions

Symmetries of the generators or, more generally, *intertwining relations* have proved to be useful in producing new duality relations from existing ones, e.g. cheap dualities [4,33]. Here, we analyze this technique and revisit [33, Theorem 5.1] from the point of view of generalized eigenfunctions.

Theorem 2 (Intertwining relations and duality). *Let L, $\widetilde{L}$ and $\widehat{L}$ be three generators on Ω, $\widetilde{\Omega}$ and $\widehat{\Omega}$ respectively. We assume that L and $\widetilde{L}$ are* intertwined, *i.e., there exists a linear operator $\Lambda : L^2(\Omega) \to L^2(\widetilde{\Omega})$ such that, for all $f \in L^2(\Omega)$, we have*

$$\widetilde{L}\Lambda f = \Lambda L f. \tag{28}$$

Moreover, we assume that L and $\widehat{L}$ are dual with duality function $D : \widehat{\Omega} \times \Omega \to \mathbb{R}$, i.e.,

$$\widehat{L}_{left} D(\widehat{x}, x) = L_{right} D(\widehat{x}, x).$$

Then, the function $\Lambda_{right} D : \widehat{\Omega} \times \widetilde{\Omega} \to \mathbb{R}$ is a duality function for $\widetilde{L}$ and $\widehat{L}$, i.e.,

$$\widehat{L}_{left} \Lambda_{right} D(\widehat{x}, \widetilde{x}) = \widetilde{L}_{right} \Lambda_{right} D(\widehat{x}, \widetilde{x}).$$

Proof. We observe that the *intertwining operator* Λ maps eigenspaces of L to eigenspaces of $\widetilde{L}$. More formally, if there exists a subset $\{u^{(1)}, \dots, u^{(m)}\}$ of $L^2(\Omega)$ such that, for some $\lambda \in \mathbb{C}$,

$$Lu^{(1)} \;=\; \lambda u^{(1)}, \quad Lu^{(k)} \;=\; \lambda u^{(k)} + u^{(k-1)}, \quad k = 2, \dots, m, \tag{29}$$

then, by (28), the subset $\{\Lambda u^{(1)}, \dots, \Lambda u^{(m)}\}$ in $L^2(\widetilde{\Omega})$ satisfy the same identities as in (29) up to replace L by $\widetilde{L}$:

$$\widetilde{L}\Lambda u^{(1)} \;=\; \lambda \Lambda u^{(1)}, \quad \widetilde{L}\Lambda u^{(k)} \;=\; \lambda \Lambda u^{(k)} + \Lambda u^{(k-1)}, \quad k = 2, \dots, m. \tag{30}$$

By Theorem 1, the duality function is given by

$$D(\widehat{x}, x) = \sum_{i=1}^{n} \widehat{u}_i(\widehat{x}) u_i(x),$$

where $\{u_1, \dots, u_n\}$, $\{\widehat{u}_1, \dots, \widehat{u}_n\}$ are sets of (possibly generalized) eigenfunctions of L, $\widehat{L}$. Then, by applying the intertwining operator Λ on the right variables, we obtain

$$\Lambda_{\text{right}} D(\widehat{x}, \widetilde{x}) = \sum_{i=1}^{n} \widehat{u}_i(\widehat{x})(\Lambda u_i)(\widetilde{x}).$$

We conclude from the considerations in (30), (29) and Theorem 1. □

Typical examples of intertwining relations occur when either Λ is a *symmetry* of a generator, i.e., $\widetilde{L} = L$ in (28) (see e.g. [4]) or when Λ is a *positive contractive operator* such that $\Lambda 1 = 1$, i.e., viewed as a matrix, it is a stochastic matrix from the space $\widetilde{\Omega}$ to Ω (see e.g. [21]). A particular instance, which recovers the so-called *lumpability*, of this last situation is when Λ is a "deterministic" stochastic kernel, i.e., induced by a map from $\widetilde{\Omega}$ to Ω.

5.5 Intertwining of Exclusion Processes

In this section we provide an application of Theorem 2 above. Indeed, after finding suitable intertwining relations between a particular instance of the symmetric simple exclusion process and a generalized symmetric exclusion process, we obtain as in Theorem 2 a large class of self-duality functions for the latter process from self-duality functions of the former. In what follows, we fix $\gamma \in \mathbb{N}$, a finite set V of cardinality $|V| = m$ and a function $p : V \times V \to \mathbb{R}_+$ such that $p(x, x) = 0$ for all $x \in V$.

The γ-*ladder*-SEP is the finite-state Markov process on $\widetilde{\Omega} = \{0, 1\}^{V \times \{1, \dots, \gamma\}}$ with generator $\widetilde{L}$ acting on functions $\widetilde{f} : \widetilde{\Omega} \to \mathbb{R}$ as

$$\begin{aligned} \widetilde{L}\widetilde{f}(\widetilde{\eta}) \;=\; & \sum_{x,y \in V} p(x,y) \Bigg[\sum_{a=1}^{\gamma} \sum_{b=1}^{\gamma} \widetilde{\eta}(x,a)(1 - \widetilde{\eta}(y,b)) \left(\widetilde{f}(\widetilde{\eta}^{(x,a),(y,b)}) - \widetilde{f}(\widetilde{\eta}) \right) \\ & + \; \widetilde{\eta}(y,b)(1 - \widetilde{\eta}(x,a)) \left(\widetilde{f}(\widetilde{\eta}^{(y,b),(x,a)}) - \widetilde{f}(\widetilde{\eta}) \right) \Bigg], \quad \widetilde{\eta} \in \widetilde{\Omega}, \end{aligned}$$

where $\widetilde{\eta}^{(x,a),(y,b)}$ denotes the configuration obtained from $\widetilde{\eta}$ by removing a particle at position (x, a) and placing it at (y, b). As already mentioned, this process may be considered as a special case of a simple symmetric exclusion process on the set $\widetilde{V}_\gamma = V \times \{1, \dots, \gamma\}$ where $\widetilde{p} : \widetilde{V} \times \widetilde{V} \to \mathbb{R}_+$ is such that

$$\widetilde{p}((x,a),(y,b)) \;=\; p(x,y), \quad (x,a),(y,b) \in \widetilde{V}_\gamma.$$

The SEP(γ) is the finite-state Markov process on $\Omega = \{0, \dots, \gamma\}^V$ with generator L acting on functions $f : \Omega \to \mathbb{R}$ as

$$Lf(\eta) = \sum_{x,y \in V} p(x,y)\, [\eta(x)(\gamma - \eta(y))\,(f(\eta^{x,y}) - f(\eta)) + \eta(y)(\gamma - \eta(x))\,(f(\eta^{y,x}) - f(\eta))], \quad \eta \in \Omega.$$

It is well known (see e.g. [18]) that L and $\widetilde{L}$ are *intertwined* via a deterministic intertwining operator $\Lambda : L^2(\Omega) \to L^2(\widetilde{\Omega})$. The intertwining operator Λ is defined, given the mapping $\pi : \widetilde{\Omega} \to \Omega$ such that

$$\pi(\widetilde{\eta}) = (|\widetilde{\eta}(1,\cdot)|, \dots, |\widetilde{\eta}(n,\cdot)|) \ \in \ \Omega, \quad |\widetilde{\eta}(x,\cdot)| \ := \ \sum_{a=1}^{\gamma} \widetilde{\eta}(x,a),$$

as acting on functions $f : \Omega \to \mathbb{R}$ as

$$\Lambda f(\widetilde{\eta}) = f(\pi(\widetilde{\eta})), \quad \widetilde{\eta} \in \widetilde{\Omega}.$$

The intertwining relation then reads, for all $f : \Omega \to \mathbb{R}$, as

$$\widetilde{L}\Lambda f(\widetilde{\eta}) = \Lambda L f(\widetilde{\eta}),$$

for $\widetilde{\eta} \in \widetilde{\Omega}$. Given any self-duality for L with self-duality function $D(\xi, \eta)$, we can build a duality function, namely $D'(\xi, \widetilde{\eta}) = \Lambda_{\text{right}} D(\xi, \widetilde{\eta})$ for L and $\widetilde{L}$ and, furthermore, a self-duality function $D''(\widetilde{\xi}, \widetilde{\eta}) = \Lambda_{\text{left}} \Lambda_{\text{right}} D(\widetilde{\xi}, \widetilde{\eta})$ for $\widetilde{L}$.

However, we ask whether there exists an "inverse" intertwining relation, i.e., $\widetilde{\Lambda} : L^2(\widetilde{\Omega}) \to L^2(\Omega)$ such that, for $\widetilde{f} : \widetilde{\Omega} \to \mathbb{R}$,

$$\widetilde{\Lambda}\widetilde{L}\widetilde{f}(\eta) = L\widetilde{\Lambda}\widetilde{f}(\eta), \quad \eta \in \Omega. \tag{31}$$

In what follows, we say that $\widetilde{\eta} \in \widetilde{\Omega}$ is *compatible* with $\eta \in \Omega$ or, shortly, $\widetilde{\eta} \sim \eta$, if $\pi(\widetilde{\eta}) = \eta$.

Proposition 9. *The operator $\widetilde{\Lambda} : L^2(\widetilde{\Omega}) \to L^2(\Omega)$ defined as*

$$\widetilde{\Lambda}\widetilde{f}(\eta) = \left(\prod_{x \in V} \frac{1}{\binom{\gamma}{\eta(x)}} \right) \sum_{\widetilde{\eta} \sim \eta} \widetilde{f}(\widetilde{\eta}), \quad \eta \in \Omega, \tag{32}$$

is the inverse intertwining in (31). Moreover, the intertwining operator above is a stochastic *intertwining.*

Proof. Without loss of generality, we consider $V = \{x, y\}$. By expanding the l.h.s. of (31) with $\widetilde{\Lambda}$ as in (32), we obtain four terms:

$$\ell_1 \ = \ -\frac{1}{\binom{\gamma}{\eta(x)}} \frac{1}{\binom{\gamma}{\eta(y)}} \sum_{\widetilde{\eta} \sim \eta} \sum_{a=1}^{\gamma} \sum_{b=1}^{\gamma} \widetilde{\eta}(x,a)(1 - \widetilde{\eta}(y,b))\widetilde{f}(\widetilde{\eta})$$

$$\ell_2 = \frac{1}{\binom{\gamma}{\eta(x)}}\frac{1}{\binom{\gamma}{\eta(y)}}\sum_{\widetilde{\eta}\sim\eta}\sum_{a=1}^{\gamma}\sum_{b=1}^{\gamma}\widetilde{\eta}(x,a)(1-\widetilde{\eta}(y,b))\widetilde{f}(\widetilde{\eta}^{(x,a),(y,b)})$$

$$\ell_3 = -\frac{1}{\binom{\gamma}{\eta(x)}}\frac{1}{\binom{\gamma}{\eta(y)}}\sum_{\widetilde{\eta}\sim\eta}\sum_{a=1}^{\gamma}\sum_{b=1}^{\gamma}\widetilde{\eta}(y,b)(1-\widetilde{\eta}(x,a))\widetilde{f}(\widetilde{\eta})$$

$$\ell_4 = \frac{1}{\binom{\gamma}{\eta(x)}}\frac{1}{\binom{\gamma}{\eta(y)}}\sum_{\widetilde{\eta}\sim\eta}\sum_{a=1}^{\gamma}\sum_{b=1}^{\gamma}\widetilde{\eta}(y,b)(1-\widetilde{\eta}(x,a))\widetilde{f}(\widetilde{\eta}^{(y,b),(x,a)}).$$

By doing the same thing with the r.h.s., we obtain:

$$r_1 = -\frac{1}{\binom{\gamma}{\eta(x)}}\frac{1}{\binom{\gamma}{\eta(y)}}\eta(x)(\gamma-\eta(y))\sum_{\widetilde{\eta}\sim\eta}\widetilde{f}(\widetilde{\eta})$$

$$r_2 = \frac{1}{\binom{\gamma}{\eta(x)-1}}\frac{1}{\binom{\gamma}{\eta(y)+1}}\eta(x)(\gamma-\eta(y))\sum_{\widetilde{\eta}\sim\eta^{x,y}}\widetilde{f}(\widetilde{\eta})$$

$$r_3 = -\frac{1}{\binom{\gamma}{\eta(x)}}\frac{1}{\binom{\gamma}{\eta(y)}}\eta(y)(\gamma-\eta(x))\sum_{\widetilde{\eta}\sim\eta}\widetilde{f}(\widetilde{\eta})$$

$$r_4 = \frac{1}{\binom{\gamma}{\eta(x)+1}}\frac{1}{\binom{\gamma}{\eta(y)-1}}\eta(y)(\gamma-\eta(x))\sum_{\widetilde{\eta}\sim\eta^{y,x}}\widetilde{f}(\widetilde{\eta}).$$

Note that $\ell_1 = r_1$ because, for all $\widetilde{\eta}\sim\eta$,

$$\sum_{a=1}^{\gamma}\sum_{b=1}^{\gamma}\widetilde{\eta}(x,a)(1-\widetilde{\eta}(y,b)) = \eta(x)(\gamma-\eta(y)),$$

and similarly for $\ell_3 = r_3$. For $\ell_2 = r_2$ it is enough to verify that, for each $\widetilde{\eta}_*\sim\eta^{x,y}$,

$$\sum_{\widetilde{\eta}\sim\eta}\sum_{a=1}^{\gamma}\sum_{b=1}^{\gamma}\widetilde{\eta}(x,a)(1-\widetilde{\eta}(y,b))\mathbb{1}\{\widetilde{\eta}^{(x,a),(y,b)}=\widetilde{\eta}_*\} = (\eta(y)+1)(\gamma-\eta(x)+1).$$

This last identity indeed holds, as the configurations $\widetilde{\eta}\sim\eta$ can be obtained from $\widetilde{\eta}_*$ by picking one of the $\eta(y)+1$ particles on $y\in V$ and putting it back on one of the $\gamma-\eta(x)+1$ holes of $x\in V$. Analogously for $\ell_4 = r_4$. □

As a consequence of this proposition, by starting from self-duality of the γ-ladder-SEP, we can produce duality functions for $\widetilde{L}$ and L and self-duality functions for L. We use the following result of [35, Theorem 2.8] to obtain a large class of "factorized" self-duality functions for $\widetilde{L}$.

Theorem 3 ([35]). *The* simple symmetric exclusion process $\{\widetilde{\eta}_t,\ t\geq 0\}$ *on the vertex set* $V\times\{1,\dots,\gamma\}$ *is self-dual w.r.t. the duality function*

$$\widetilde{D}(\widetilde{\xi},\widetilde{\eta}) = \prod_{(x,a)\in V\times\{1,\dots,\gamma\}}(\alpha+\beta\widetilde{\eta}(x,a))^{\epsilon+\delta\widetilde{\xi}(x,a)},\quad \widetilde{\xi},\widetilde{\eta}\in\widetilde{\Omega}, \tag{33}$$

for all α,β,ϵ *and* $\delta\in\mathbb{R}$.

Now, we apply the intertwining operator $\widetilde{\Lambda}$ first on the right and then on the left variables of $\widetilde{D}$ above.

Theorem 4. *All self-duality functions for SEP(γ) derived from self-duality functions of γ-ladder-SEP as in (33) are all in factorized form, i.e.,*

$$D(\xi,\eta) = \widetilde{\Lambda}_{left}\widetilde{\Lambda}_{right}\widetilde{D}(\xi,\eta) = \prod_{x\in V} d_x^{\alpha,\beta,\epsilon,\delta}(\xi(x),\eta(x)).$$

Moreover, the single-site self-duality functions $d_x^{\alpha,\beta,\epsilon,\delta}(k,n)$, *for* $k,n \in \{0,\dots,\gamma\}$, *are in one of the following forms: either the* classical *polynomials*

$$d_x^{0,\beta,0,\delta}(k,n) = (\beta^\delta)^k \frac{(\gamma-k)!}{\gamma!}\frac{n!}{(n-k)!}\mathbb{1}\{n\geq k\},$$

the orthogonal *polynomials*

$$d_x^{\alpha,\beta,\epsilon,\delta}(k,n) = (-1)^{\delta k}\alpha^{\epsilon\gamma-\epsilon n+\delta k}(\alpha+\beta)^{\epsilon n}\,{}_2F_1\left[\begin{matrix}-k-n\\ -\gamma\end{matrix};1-\left(1+\tfrac{\beta}{\alpha}\right)^\delta\right],$$

or other degenerate functions:

$$\begin{aligned} d_x^{\alpha,\beta,\epsilon,0}(k,n) &= (\alpha+\beta)^{\epsilon n}\alpha^{\epsilon(\gamma-n)}\\ d_x^{0,\beta,\epsilon,\delta}(k,n) &= \beta^{\epsilon\gamma+\delta k}\mathbb{1}\{n=\gamma\}\\ d_x^{\alpha,0,\epsilon,\delta}(k,n) &= \alpha^{\epsilon\gamma+\delta k}\\ d_x^{\alpha,-\alpha,\epsilon,\delta}(k,n) &= \alpha^{\epsilon\gamma+\delta k}\mathbb{1}\{n=0\}. \end{aligned}$$

Proof. First thing to note is that the factorized structure of D is preserved under $\widetilde{\Lambda}$. Indeed, if we use the notation

$$\mathtt{d}(k,n) = (\alpha+\beta n)^{\epsilon+\delta k},$$

then

$$\widetilde{\Lambda}_{\text{right}}D(\widetilde{\xi},\eta) = \prod_{x\in V}\left(\frac{1}{\binom{\gamma}{\eta(x)}}\sum_{\widetilde{\eta}(x,\cdot)\sim\eta(x)}\prod_{a=1}^{\gamma}\mathtt{d}(\widetilde{\xi}(x,a),\widetilde{\eta}(x,a))\right).$$

Hence we compute only what is inside the parenthesis (which will see does depend on $\widetilde{\xi}(x,\cdot)$ only through $|\widetilde{\xi}(x,\cdot)|$):

$$\begin{aligned} &d_x^{\alpha,\beta,\epsilon,\delta}(\xi(x),\eta(x))\\ &\quad= (\alpha+\beta)^{\epsilon\eta(x)}\alpha^{\epsilon(\gamma-\eta(x))}\frac{1}{\binom{\gamma}{\eta(x)}}\sum_{\widetilde{\eta}(x,\cdot)\sim\eta(x)}\prod_{a=1}^{\gamma}(\alpha+\beta\widetilde{\eta}(x,a))^{\delta\widetilde{\xi}(x,a)}. \end{aligned} \tag{34}$$

The last summation

$$\frac{1}{\binom{\gamma}{\eta(x)}} \sum_{\widetilde{\eta}(x,\cdot)\sim\eta(x)} \prod_{a=1}^{\gamma} (\alpha+\beta\widetilde{\eta}(x,a))^{\delta\widetilde{\xi}(x,a)}$$

clearly does not depend on $\widetilde{\xi}(x,\cdot)$ but only on $\xi(x) = |\widetilde{\xi}(x,\cdot)|$ and equals

$$\frac{1}{\binom{\gamma}{\eta(x)}} \sum_{\ell=0}^{\xi(x)} \binom{\xi(x)}{\xi(x)-\ell} \binom{\gamma-\xi(x)}{\eta(x)-(\xi(x)-\ell)} (\alpha+\beta)^{\delta(\xi(x)-\ell)} \alpha^{\delta\ell}. \tag{35}$$

If $\delta = 0$, this last expression in (35) by Chu-Vandermonde identity equals 1, hence

$$d_x^{\alpha,\beta,\epsilon,0}(\xi(x),\eta(x)) = (\alpha+\beta)^{\epsilon\eta(x)} \alpha^{\epsilon(\gamma-\eta(x))}.$$

If $\delta \neq 0$ and $\alpha = 0$, expression (35) rewrites as

$$\begin{aligned} &\frac{1}{\binom{\gamma}{\eta(x)}} \binom{\gamma-\xi(x)}{\eta(x)-\xi(x)} \beta^{\delta\xi(x)} \mathbb{1}\{\eta(x) \geq \xi(x)\} \\ &\qquad = (\beta^{\delta})^{\xi(x)} \frac{(\gamma-\xi(x))!}{\gamma!} \frac{\eta(x)!}{(\eta(x)-\xi(x))!} \mathbb{1}\{\eta(x) \geq \xi(x)\}, \end{aligned}$$

and hence, for $\epsilon = 0$, (34) becomes

$$d_x^{0,\beta,0,\delta}(\xi(x),\eta(x)) = (\beta^{\delta})^{\xi(x)} \frac{(\gamma-\xi(x))!}{\gamma!} \frac{\eta(x)!}{(\eta(x)-\xi(x))!} \mathbb{1}\{\eta(x) \geq \xi(x)\},$$

i.e., the classical single-site self-duality functions, while, for $\epsilon \neq 0$,

$$d_x^{0,\beta,\epsilon,\delta}(\xi(x),\eta(x)) = \beta^{\epsilon\gamma+\delta\xi(x)} \mathbb{1}\{\eta(x) = \gamma\}.$$

If $\delta \neq 0$ and $\alpha \neq 0$ and $\beta = 0$, then again we get some trivial:

$$d_x^{\alpha,0,\epsilon,\delta}(\xi(x),\eta(x)) = \alpha^{\epsilon\gamma+\delta\xi(x)}.$$

The most interesting case is when $\delta \neq 0$, $\alpha \neq 0$, $\beta \neq 0$ and $\alpha \neq -\beta$. In this case the quantity in (35) equals

$$(\alpha+\beta)^{\delta\xi(x)} \frac{1}{\binom{\gamma}{\eta(x)}} \sum_{\ell=0}^{\xi(x)} \binom{\xi(x)}{\xi(x)-\ell} \binom{\gamma-\xi(x)}{\eta(x)-(\xi(x)-\ell)} \left(\frac{\alpha}{\alpha+\beta}\right)^{\delta\ell},$$

which rewrites, by using two known relations in [31, p. 51], as

$$(-\alpha)^{\delta\xi(x)} {}_2F_1 \left[\begin{matrix} -\xi(x)-\eta(x) \\ -\gamma \end{matrix} ; 1-\left(1+\tfrac{\beta}{\alpha}\right)^{\delta} \right],$$

leading to

$$\begin{aligned} &d_x^{\alpha,\beta,\epsilon,\delta}(\xi(x),\eta(x)) \\ &\quad = (-1)^{\delta\xi(x)} \alpha^{\epsilon\gamma-\epsilon\eta(x)+\delta\xi(x)} (\alpha+\beta)^{\epsilon\eta(x)} {}_2F_1 \left[\begin{matrix} -\xi(x)-\eta(x) \\ -\gamma \end{matrix} ; 1-\left(1+\tfrac{\beta}{\alpha}\right)^{\delta} \right], \end{aligned}$$

i.e., we recover the *orthogonal polynomial single-site self-duality functions* for the SEP(γ), namely families of Kravchuk polynomials. If $\alpha = -\beta$, then we have

$$d_x^{\alpha,-\alpha,\epsilon,\delta}(\xi(x),\eta(x)) = \alpha^{\epsilon\gamma+\delta\xi(x)}\mathbb{1}\{\eta(x)=0\}.$$

□

6 Siegmund Duality

This connection between duality functions and eigenfunctions enables us to recover another special instance of duality, the so-called *Siegmund duality*. Siegmund duality, which arises in the context of totally ordered state spaces $\Omega = \widehat{\Omega}$, was first established by Siegmund [36] for pairs of absorbed/reflected-at-0 processes on the positive real line and on the positive integers. Further applications and generalizations of Siegmund dualities were studied by many authors, see for instance [25,27,28].

What we focus here on is a finite-context characterization of Siegmund duality already obtained via an intertwining relation in [21]. However, by using a representation of duality in terms of generalized eigenfunctions of the generators, the characterization result of Siegmund duality that we obtain, besides simplifying the proof of an analogous result in [36, Theorem 3], adds spectral information to the proof in [21].

Moreover, as Siegmund duality can be seen as a full-rank duality between two processes, cf. Theorem 1, a spectral approach provides a strategy to find other duality relations in the presence of Siegmund duality.

6.1 Characterization of Siegmund Duality

On the totally ordered state space $\Omega = \{1,\dots,n\}$, two generators L, $\widehat{L}$ are said to be *Siegmund dual* if

$$\widehat{L}_{\text{left}} D_{\text{S}}(x,y) = L_{\text{right}} D_{\text{S}}(x,y), \tag{36}$$

with duality function $D_{\text{S}} : \Omega \times \Omega \to [0,1]$ given by

$$D_{\text{S}}(x,y) = \mathbb{1}\{x \geq y\}. \tag{37}$$

Note that the duality relation (36) with duality function D_{S} (37) reads out

$$\sum_{x'=y}^{n} \widehat{L}(x,x') = \sum_{y'=1}^{x} L(y,y'). \tag{38}$$

From (38), a *necessary* relation between two Siegmund dual generators L and $\widehat{L}$ reads as follows:

$$L(y,x) \;= \sum_{x'=y}^{n} \widehat{L}(x,x') - \widehat{L}(x-1,x'), \quad x,y \in \Omega, \tag{39}$$

with the convention $\widehat{L}(0,\cdot) = 0$. As (39) implies (38), this condition is indeed also *sufficient*.

Remark 2 (Sub-generators and monotonicity). If we require that only $\widehat{L}$ is a generator, the operator L as defined in (39) is not necessarily a generator. However, the following implications hold:

(a) If $\widehat{L}$ is a generator and $L(y,x) \geq 0$ for all $x \neq y$, then L is a *sub-generator* on Ω, i.e.,

$$L(y,x) \;\geq\; 0,\; x \neq y \quad \text{and} \quad \sum_{x=1}^{n} L(y,x) \;\leq\; 0,\; y \in \Omega. \tag{40}$$

The proof goes as follows:

$$\sum_{x=1}^{n} L(y,x) \;=\; \sum_{x'=y}^{n} \sum_{x=1}^{n} \widehat{L}(x,x') - \widehat{L}(x-1,x') \;=\; \sum_{x'=y}^{n} \widehat{L}(n,x') \leq \sum_{x'=1}^{n} \widehat{L}(n,x') \;=\; 0,$$

where we used (39) in the first equality and the last inequality is a consequence of $\widehat{L}$ being a generator.

(b) Note that, by [22, Theorem 2.1],

$$\sum_{x'=y}^{n} \widehat{L}(x,x') - \widehat{L}(x-1,x') \;\geq\; 0, \quad x \neq y, \tag{41}$$

is equivalent to require that the continuous-time Markov chain with generator $\widehat{L}$ is *monotone* (see [28]).

As a consequence, L is a sub-generator if and only if $\widehat{L}$ is associated to a monotone process on Ω.

In the following theorem, we study the relation between eigenfunctions of Siegmund dual (sub-)generators and how the Siegmund duality function D_{S} in (37) is constructed from the eigenfunctions.

Theorem 5. *(i) Let L and $\widehat{L}$ be Siegmund dual (sub-)generators in the sense of (36). If $\widehat{w}$ is a k-th order generalized eigenfunction of $\widehat{L}^{\mathsf{T}}$ associated to eigenvalue λ, then*

$$u(x) \;=\; \sum_{y=x}^{n} \widehat{w}(y), \quad x \in \Omega, \tag{42}$$

is a k-th order generalized eigenfunction of L associated to the eigenvalue λ.

(ii) In the same context as in item (i), given a set $\{\widehat{w}_1, \ldots, \widehat{w}_n\}$ of (generalized) eigenfunctions of $\widehat{L}^{\mathsf{T}}$ whose span coincides with $L^2(\Omega)$, if $\{\widehat{u}_1, \ldots, \widehat{u}_n\}$ are (generalized) eigenfunctions of $\widehat{L}$ such that

$$\langle \widehat{w}_i, \widehat{u}_j^* \rangle = \sum_{x=1}^{n} \widehat{w}_i(x) \widehat{u}_j(x) \;=\; \delta_{i,j}, \tag{43}$$

and $\{u_1, \dots, u_n\}$ *are defined in terms of* $\{\widehat{w}_1, \dots, \widehat{w}_n\}$ *as in* (42)*, then the function*

$$D(x, y) = \sum_{i=1}^{n} \widehat{u}_i(x) u_i(n), \quad x, y \in \Omega,$$

is the Siegmund duality function D_{S}.

(iii) *Let* L *and* $\widehat{L}$ *be (sub-)generators on* Ω*. If for any k-th order generalized eigenfunction* $\widehat{w}$ *of* $\widehat{L}^{\mathsf{T}}$ *associated to eigenvalue* λ*,* u *as defined in* (42) *is a k-th order generalized eigenfunction of* L *associated to the same eigenvalue* λ*, then* L *and* $\widehat{L}$ *are Siegmund dual and* D_{S} *is obtained as in item (ii).*

Proof. Let $\widehat{w}$ and u be as in item (i). Then,

$$\begin{aligned}\sum_{x=1}^{n} L(y, x) u(x) &= \sum_{x=1}^{n} \left(\sum_{x'=y}^{n} \widehat{L}(x, x') - \widehat{L}(x-1, x') \right) u(x) \\ &= \sum_{x'=y}^{n} \sum_{x=1}^{n} \left(\widehat{L}^{\mathsf{T}}(x', x) u(x) - \widehat{L}^{\mathsf{T}}(x', x-1) u(x) \right),\end{aligned}$$

which, by noting that $\widehat{w}(n) = u(n)$, reads as

$$\sum_{x'=y}^{n} \sum_{x=1}^{n} \widehat{L}^{\mathsf{T}}(x', x) \widehat{w}(x) = \sum_{x'=y}^{n} \lambda \widehat{w}(x') \;=\; \lambda \sum_{x'=y}^{n} \widehat{w}(x') \;=\; \lambda u(y),$$

thus, u is eigenfunction with eigenvalue λ. For the generalized eigenfunctions, the proof follows the same line.

For item (ii) and (iii), from the sets $\{\widehat{w}_1, \dots, \widehat{w}_n\}$ and $\{u_1, \dots, u_n\}$ of generalized eigenfunctions of $\widehat{L}^{\mathsf{T}}$ and L related as in (42), by Theorem 1 the function

$$D(x, y) \;=\; \sum_{i=1}^{n} \widehat{u}_i(x) u_i(y) \;=\; \sum_{i=1}^{n} \widehat{u}_i(x) \sum_{x'=y}^{n} \widehat{w}_i(x') \;=\; \sum_{x'=y}^{n} \sum_{i=1}^{n} \widehat{u}_i(x) \widehat{w}_i(x') \tag{44}$$

is a full-rank duality for L and $\widehat{L}$. By Proposition 8 and condition (43), by passing to the conjugates, we obtain

$$\sum_{i=1}^{n} \widehat{u}_i(x) \widehat{w}_i(x') = \delta_{x,x'},$$

and hence the function $D(x, y)$ in (44) writes as

$$D(x, y) = \sum_{x'=y}^{n} \delta_{x,x'} = \mathbb{1}\{x \geq y\} \;=\; D_{\mathrm{S}}(x, y).$$

□

In this final example, by using item (iii) of Theorem 5, we show how to obtain Siegmund duality from the knowledge of eigenvalues and eigenfunctions of (sub-)generators. The example we consider here concerns two symmetric random walks on $\Omega = \{1, \dots, n\}$.

Example 4 (Blocked vs absorbed random walks on a finite grid). The first symmetric nearest-neighbor random walk is *blocked at the boundaries*, namely the generator $\widehat{L}$ is described, for $f : \Omega \to \mathbb{R}$, as

$$\widehat{L}f(x) = (f(x+1) - f(x)) + (f(x-1) - f(x)), \quad x \in \Omega \setminus \{1, n\},$$

and, on the boundaries,

$$\widehat{L}f(1) \;=\; f(2) - f(1), \qquad \widehat{L}f(n) \;=\; f(n-1) - f(n).$$

The second random walk is *absorbed at the boundaries*, i.e., it is a sub-Markov process on $\Omega = \{1, \dots, n\}$ with sub-generator L which acts on functions $f : \Omega \to \mathbb{R}$ as

$$Lf(x) = (f(x+1) - f(x)) + (f(x-1) - f(x)), \quad x \in \Omega \setminus \{1, n\},$$

and

$$Lf(1) \;=\; 0, \qquad Lf(n) \;=\; f(n-1) - 2f(n),$$

i.e. $x = 1$ is an absorbing point, while at $x = n$ the random walk either jumps to the left at rate 1 or "exits the system" at rate 1.

To explicitly obtain eigenfunctions and eigenvalues in this setting we use the following *ansatz*:

$$f_{a,b,c,\theta}(x) = a\cos(\theta x + c) + b\sin(\theta x + c), \quad x \in \Omega,$$

where a, b, c and $\theta \in \mathbb{R}$ are the parameters to be determined. Regarding the eigenvalues $\{\lambda_1, \dots, \lambda_n\}$, in both cases we have

$$\lambda_1 \;=\; 0, \quad \lambda_i \;=\; 2(\cos(\theta_i) - 1), \quad \theta_i \;=\; \frac{i-1}{n}\pi, \quad i = 2, \dots, n.$$

Hence, all eigenvalues are distinct. The eigenfunctions $\{\widehat{u}_1, \dots, \widehat{u}_n\}$ of $\widehat{L}$ are, for $x \in \{1, \dots, n\}$ and $i = 2, \dots, n$,

$$\widehat{u}_1(x) \;=\; \frac{1}{\sqrt{n}},$$

and

$$\widehat{u}_i(x) \;=\; \frac{1}{\sqrt{n(1-\cos(\theta_i))}}(-\sin(\theta_i)\cos(\theta_i(x-1)) + (1-\cos(\theta_i))\sin(\theta_i(x-1))).$$

The eigenfunctions $\{u_1, \dots, u_n\}$ of L are given, for $x \in \{1, \dots, n\}$ and $i = 2, \dots, n$, by

$$u_1(x) \;=\; \frac{n+1-x}{\sqrt{n}}, \quad u_i(x) \;=\; \frac{1}{\sqrt{n(1-\cos(\theta_i))}}\sin(\theta_i(x-1)).$$

Hence, we note that:

(a) By Theorem 1, L and $\widehat{L}$ are dual and any duality function is of the form

$$D(x,y) = \sum_{i=1}^{n} a_i \widehat{u}_i(x) u_i(y), \tag{45}$$

for $a_1, \ldots, a_n \in \mathbb{R}$.

(b) By denoting by ν the counting measure on $\Omega = \{1, \ldots, n\}$, the generator $\widehat{L}$ is self-adjoint in $L^2(\nu)$ and is, as a matrix, symmetric, i.e., $\widehat{L}^{\mathsf{T}} = \widehat{L}$. As a consequence, $\{\widehat{u}_1, \ldots, \widehat{u}_n\}$ are eigenfunctions of both $\widehat{L}$ and $\widehat{L}^{\mathsf{T}}$.

(c) For all $i = 1, \ldots, n$,

$$u_i(x) = \sum_{y=x}^{n} \widehat{u}_i(y), \quad x \in \Omega,$$

i.e., the eigenfunctions $\{u_1, \ldots, u_n\}$ are related to $\{\widehat{u}_1, \ldots, \widehat{u}_n\}$ as in (42).

(d) The eigenfunctions $\widehat{u}_1, \ldots, \widehat{u}_n$ are normalized in $L^2(\nu)$, i.e., for all $i, j = 1, \ldots, n$,

$$\langle \widehat{u}_i, \widehat{u}_j \rangle_{L^2(\nu)} = \delta_{i,j}.$$

As a consequence, by Theorem 5, for the choice $a_1 = \ldots = a_n = 1$, the duality function $D(x,y)$ in (45) is the Siegmund duality function $D_{\mathrm{S}}(x,y)$ in (37), namely, for all $x, y \in \Omega$,

$$\frac{n+1-y}{n} + \sum_{i=2}^{n} \frac{\sin(\theta_i(y-1))}{n(1-\cos(\theta_i))} \left(-\sin(\theta_i)\cos(\theta_i(x-1)) + (1-\cos(\theta_i))\sin(\theta_i(x-1))\right) = \mathbb{1}\{x \geq y\}.$$

As a final remark, we note that, by adding the cemetery state $\Delta = \{n+1\}$ accessible at rate 1 only from the state $\{n\}$, the absorbed sub-Markov random walk associated to L becomes a proper Markov process with $\{1\}$ and $\{n+1\}$ as absorbing states. If we denote by L^{ext} the generator on the extended space $\Omega \cup \Delta$, it follows that the eigenvalues of L^{ext} remain unchanged, while the new eigenfunctions $\{u_1^{\mathrm{ext}}, \ldots, u_n^{\mathrm{ext}}, u_{n+1}^{\mathrm{ext}}\}$ are such that

$$u_{n+1}^{\mathrm{ext}}(x) = 1, \quad x \in \Omega \cup \Delta,$$

and, for all $i = 1, \ldots, n$,

$$u_i^{\mathrm{ext}}(n+1) = 0, \qquad u_i^{\mathrm{ext}}(x) = u_i(x), \quad x \in \Omega.$$

Hence, the function

$$D_{\mathrm{S}}^{\mathrm{ext}}(x,y) = \sum_{i=1}^{n} \widehat{u}_i(x) u_i^{\mathrm{ext}}(y), \quad x \in \Omega, \quad y \in \Omega \cup \Delta,$$

equals $\mathbb{1}\{x \geq y\}$. □

Acknowledgments. The authors thank *Institut Henri Poincaré*, where part of this work was done, for very kind hospitality. F.S. acknowledges NWO for financial support via the TOP1 grant 613.001.552. The same author is indebted to G. Carinci for fruitful discussions.

References

1. Borodin, A., Corwin, I., Gorin, V.: Stochastic six-vertex model. Duke Math. J. **165**, 563–624 (2016)
2. Carinci, G., Giardinà, C., Giberti, C., Redig, F.: Dualities in population genetics: a fresh look with new dualities. Stoch. Process. Appl. **125**, 941–969 (2015)
3. Carinci, G., Giardinà, C., Giberti, C., Redig, F.: Duality for stochastic models of transport. J. Stat. Phys. **152**, 657–697 (2013)
4. Carinci, G., Giardinà, C., Redig, F., Sasamoto, T.: A generalized asymmetric exclusion process with $U_q(\mathfrak{sl}_2)$ stochastic duality. Probab. Theory Relat. Fields **166**, 887–933 (2016)
5. Carinci, G., Giardinà, C., Redig, F., Sasamoto, T.: Asymmetric stochastic transport models with $\mathscr{U}_q(\mathfrak{su}(1,1))$ symmetry. J. Stat. Phys. **163**, 239–279 (2016)
6. Choi, M.C.H., Patie, P.: A sufficient condition for continuous-time finite skip-free Markov chains to have real eigenvalues. In: Bélair, J., et al. (eds.) Mathematical and Computational Approaches in Advancing Modern Science and Engineering, pp. 529–536. Springer, Cham (2016)
7. Choi, M., Patie, P.: Skip-free Markov chains. Preprint, Research gate (2016)
8. Conrad, N.D., Weber, M., Schütte, C.: Finding dominant structures of nonreversible Markov processes. Multiscale Model. Simul. **14**, 1319–1340 (2016)
9. Corwin, I., Shen, H., Tsai, L.-C.: ASEP(q, j) converges to the KPZ equation. Annales de l'Institut Henri Poincaré Probabilités et Statistiques **54**, 995–1012 (2018)
10. De Masi, A., Presutti, E.: Mathematical Methods for Hydrodynamic Limits. Lecture Notes in Mathematics, vol. 1501. Springer, Heidelberg (1991)
11. Depperschmidt, A., Greven, A., Pfaffelhuber, P.: Tree-valued Fleming-Viot dynamics with mutation and selection. Ann. Appl. Probab. **22**, 2560–2615 (2012)
12. Diaconis, P., Fill, J.A.: Strong stationary times via a new form of duality. Ann. Probab. **18**, 1483–1522 (1990)
13. Etheridge, A., Freeman, N., Penington, S.: Branching Brownian motion, mean curvature flow and the motion of hybrid zones. Electron. J. Probab. **22**, 1–40 (2017)
14. Fill, J.A.: On hitting times and fastest strong stationary times for skip-free and more general chains. J. Theor. Probab. **22**, 587–600 (2009)
15. Franceschini, C., Giardinà, C.: Stochastic Duality and Orthogonal Polynomials. Preprint, arXiv:1701.09115 (2016)
16. Franceschini, C., Giardinà, C., Groenevelt, W.: Self-duality of Markov processes and intertwining functions. Math. Phys. Anal. Geom. **21**, 29–49 (2018)
17. Giardinà, C., Kurchan, J., Redig, F.: Duality and exact correlations for a model of heat conduction. J. Math. Phys. **48**, 033301 (2007)
18. Giardinà, C., Kurchan, J., Redig, F., Vafayi, K.: Duality and hidden symmetries in interacting particle systems. J. Stat. Phys. **135**, 25–55 (2009)
19. Groenevelt, W.: Orthogonal stochastic duality functions from Lie algebra representations. J. Stat. Phys. **174**, 97–119 (2019)

20. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)
21. Huillet, T., Martinez, S.: Duality and intertwining for discrete Markov kernels: relations and examples. Adv. Appl. Prob. **43**, 437–460 (2011)
22. Keilson, J., Kester, A.: Monotone matrices and monotone Markov processes. Stoch. Process. Appl. **5**, 231–241 (1977)
23. Kipnis, C., Landim, C.: Scaling Limits of Interacting Particle Systems, vol. 320. Springer, Heidelberg (1999)
24. Kipnis, C., Marchioro, C., Presutti, E.: Heat flow in an exactly solvable model. J. Stat. Phys. **27**, 65–74 (1982)
25. Kolokol'tsov, V.N.: Stochastic monotonicity and duality for one-dimensional Markov processes. Math. Notes **89**, 652–660 (2011)
26. Kuan, J.: An algebraic construction of duality functions for the stochastic $\mathscr{U}_q(A_n^{(1)})$ vertex model and its degenerations. Commun. Math. Phys. **359**, 121–187 (2018)
27. Lee, R.X.: The existence and characterisation of duality of Markov processes in the Euclidean space. Ph.D. thesis, University of Warwick (2013)
28. Liggett, T.M.: Interacting Particle Systems. Springer, Heidelberg (2005)
29. Miclo, L.: On the Markovian similarity. In: Donati-Martin, C., Lejay, A., Rouault, A. (eds.) Séminaire de Probabilités XLIX, pp. 375–403. Springer, Cham (2018)
30. Möhle, M.: The concept of duality and applications to Markov processes arising in neutral population genetics models. Bernoulli **5**, 761–777 (1999)
31. Nikiforov, A.F., Suslov, S.K., Uvarov, V.B.: Classical Orthogonal Polynomials of a Discrete Variable. Springer, Heidelberg (1991)
32. Patie, P., Savov, M., Zhao, Y.: Intertwining, Excursion Theory and Krein Theory of Strings for Non-self-adjoint Markov Semigroups. Preprint, arXiv:1706.08995 (2017)
33. Redig, F., Sau, F.: Factorized duality, stationary product measures and generating functions. J. Stat. Phys. **172**, 980–1008 (2018)
34. Schütz, G.M.: Duality relations for asymmetric exclusion processes. J. Stat. Phys. **86**, 1265–1287 (1997)
35. Schütz, G.M.: Fluctuations in stochastic interacting particle systems. In: Giacomin, G., Olla, S., Saada, E., Spohn, H. (eds.) Stochastic Dynamics out of Equilibrium. PROMS, vol. 282, pp. 67–134. Springer, Cham (2019). https://indico.math.cnrs.fr/event/852/material/1/0.pdf
36. Siegmund, D.: The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. Ann. Probab. **4**(6), 914–924 (1976)
37. Weber, M.: Eigenvalues of non-reversible Markov chains - a case study. Technical report 17-13, ZIB, Takustr. 7, 14195 Berlin (2017)