Degenerated Markov processes for sampling, optimization and modelling.

Pierre Monmarché

Faculté des Sciences — Sorbonne Université

Mémoire d'Habilitation à Diriger des Recherches

Contents

Introduction

The goal of this memoir is to give a comprehensive (yet compact) presentation of my research activity since I defended my thesis. It covers 5 years of research, from 2015 to 2020. During this period, I have collaborated with Michel Benaïm, Patrick Cattiaux, Charles-Edouard Bréhier, Alain Durmus, Virginie Ehrlacher, Nicolas Fournier, Carl-Eric Gauthier, Arnaud Guillin, Lucas Journal, Louis Lagardère, Tony Lelièvre, Eva Löcherbach, Jean-Philip Piquemal, Camille Tardif, Jeremy Weisman and Chaoen Zhang, and interacted with many other particles.

One of my objectives here is to make clear (in particular to myself), *a posteriori*, the general structure and consistency underlying the various themes that I have studied (even though, as one might suspect, the structure owes a great deal to randomness). To do so, it is necessary to put my own works in the larger perspective of the research fields and communities to which I belong (or I am tangent). A very detailed presentation of entire research domains with exhaustive list of up-to-date references exceeding by far the scope of the current memoir, I decided to give in Chapter 1 an overview of some basic topics that are at the center of my work, like metastability, stochastic algorithms, molecular dynamics, etc. While the experienced reader may not learn anything new here, I see the following advantages to this : 1) this memoir could be used as a general self-contained reference for potential future junior colleagues (e.g. PhD students) to get a general glimpse of my research field; 2) this will set up a general picture (highlighting some important points that would otherwise be lost in specific technical details), in which my own work hopefully makes sense 3) a good way to understand my research activity is to see how I see my own research environment (in particular what are the general important questions), which is reflected in the choice of the topics in Chapter 1 and the way they are covered; 4) finally, more prosaically, this will be the occasion to introduce some notions and notations that will be useful in the rest of the memoir.

Chapter 2 is the main part of the memoir, devoted to my own results. In an informal overview, Section 2.1 is an attempt to make some sense of the choices of my specific research topics. In the rest of the chapter, my works are presented in a synthetic way, sorted according to some main themes : hypocoercivity, velocity jump processes, interacting jump processes, simulated annealing algorithms and adaptive algorithms.

Chapter 1

Scientifical context: the basics

1.1 Markov processes and PDE

Though the relation between Markov dynamics and linear PDE is a very basic topic in many master courses, I often discuss with brilliant people from the PDE community – in kinetic theory for instance – who are at most vaguely aware that such a link exists but are not really familiar with this idea, working sometimes on stochastic processes *à la Jourdain*[1]. My research lies precisely at this interface between Markov processes and linear PDE and thus, even for the skilled reader, in order to get a clear view of my work, it is relevant to recall and highlight some very simple notions on that matter.

We will essentially consider the case of a (continuous-time) Markov process $(X_t)_{t\geqslant 0}$ on a finite set, whose law at time t is simply a finite-dimensional vector that solves a linear ordinary differential equation (rather than a density that solves a PDE in an infinite dimensional space). This very elementary context is sufficient to introduce all the notions and notations that will be of interest, without requiring any fancy theoretical tools like stochastic calculus or operator theory.

1.1.1 *On a finite set*

Let E be a finite set, and $(\Omega, \mathbb{P})$ be a probability space. An E-valued stochastic process $(X_t)_{t\geqslant 0}$ on Ω is a measurable function from Ω to the set of càdlàg functions from $\mathbb{R}_+$ to E, endowed with the Borel $\sigma-$algebra associated to the Skorohod topology. In other words it is a random càdlàg function from $\mathbb{R}_+$ to E, $\omega \in \Omega \mapsto (t \geqslant 0 \mapsto X_t(\omega) \in E)$. It is said to be a Markov process if for all function $f : E \mapsto \mathbb{R}$ and all $0 \leqslant s \leqslant t$,

$$\mathbb{E}\left(f(X_t) \mid \mathcal{F}_s\right) \;=\; \mathbb{E}\left(f(X_t) \mid X_s\right),$$

with $(\mathcal{F}_t)_{t\geqslant 0}$ the filtration associated to $(X_t)_{t\geqslant 0}$. In other words, given that I am at some position y at some time s, then (the statistics of) my future trajectory does depend on my current position but not on the past path that brought me here. This is the stochastic extension of the property of solutions of ordinary differential equation, whose trajectory after any time s is completely determined by their position at time s (in particular, solutions of ODEs are Markov processes). This lack of memory is the very reason Markov

[1] Here I am referring to Molière's play *Le Bourgeois gentilhomme* where Mr Jourdain discovers that he had always been speaking prose without even knowing the word *prose*. I am not refering to the Jourdain of [JLR10] who definitely knows what he is doing.

processes are ubiquitous in stochastic algorithms, and paradoxically is also a crucial limitation for such applications. Indeed, Markov processes are easy to sample numerically with an iterative procedure, keeping in memory only the current position. There are many ways to design Markov processes with given invariant measures. On the other hand, stochastic algorithms are mostly based on a random exploration of a large space of parameters, and an amnesic explorer is quite inefficient as its vagrancy brings it back many times to places it has already visited without learning anything new. This general observation is the basis of a significant part of my work, trying to put some inertia or memory in the dynamics to enhance the sampling.

Markov processes on a finite set E are easily described. Define the jump rates $\lambda : E \times E \times \mathbb{R}_+ \to \mathbb{R}$ by

$$\mathbb{P}\left(X_t = y \mid X_s = x\right) \;=\; \mathbb{1}_{\{x=y\}} + (t-s)\lambda_s(x,y) + \underset{t\to s}{o}(t-s)\,.$$

Remark that $\mathbb{1}_{\{x=y\}} = \mathbb{P}\left(X_s = y \mid X_s = x\right)$ and that, probabilities being positive and summing to 1, necessarily $\lambda_s(x,y) \geqslant 0$ for $y \neq x$ and $\lambda_s(x,x) = -\sum_{y\neq x}\lambda_s(x,y)$. Due to the memoryless property of the process, these rates completely characterize the dynamics. A Markov process is said to be time homogeneous if

$$\mathbb{P}(X_t = y \mid X_s = x) \;=\; \mathbb{P}(X_{t-s} = y \mid X_0 = x)$$

for all $0 \leqslant s \leqslant t$ and $x, y \in E$, which is equivalent to say that the rates do not depend on time.

A trajectory $(X_t)_{t\geqslant 0}$ is constructed as follows. Suppose X_s is already constructed for some $s \geqslant 0$. Draw independent standard (i.e. mean 1) exponential variables $(S_y)_{y\in E}$ and for all $y \in E$ let

$$T_y \;=\; \inf\left\{ t \geqslant s,\; S_y \leqslant \int_s^t \lambda_u(X_s, y)\mathrm{d}u \right\}, \qquad T \;=\; \inf_{y\in E} T_y\,.$$

Then T is the next jump time of the process, in other words we set $X_u = X_s$ for all $u \in [s, T[$ and then $X_T = y_*$ where y_* is such that $T_{y_*} = T$ (y_* is almost surely unique). Then X_u is defined for all times u up to T, and the construction can go on by induction. If the jump rates are locally bounded, there is almost surely a finite number of jumps in a finite time interval, so that the trajectory is almost surely defined for all times.

This is a probabilistic description of a process, with the construction of a single random trajectory. Another possibility is to consider the evolution of the time marginals of its law, i.e. the evolution the law at time t of X_t, for all $t \geqslant 0$ (this corresponds to look at the statistics of a large number of particles). Note that, although the collection of marginal laws (for a fixed initial condition) does not characterize the dynamics, the knowledge of all these marginal laws *for all possible initial distributions of* X_s does. By definition, probability measures are given by their action on observables (also called test functions), i.e. bounded measurable functions (of course, here, on a finite set, all real functions are measurable and bounded). With a deterministic initial condition $X_s = x$, we are then lead to consider, for all function $f : E \to \mathbb{R}$,

$$P_{s,t}f(x) \;:=\; \mathbb{E}\left(f(X_t) \mid X_s = x\right)\,. \tag{1.1}$$

Then $f \mapsto P_{s,t}f$ is a linear operator on $\mathcal{L}^\infty(E) = \mathbb{R}^E$, and the Markov property ensure the semi-group property

$$P_{s,u}P_{u,t} = P_{s,t} \qquad \forall 0 \leqslant s \leqslant u \leqslant t\,.$$

Remark that, for homogeneous processes, we simply denote $P_t = P_{0,t}$, so that $P_{s,t} = P_{t-s}$, and the semi-group property reads $P_sP_t = P_{s+t}$.

By definition of the jump rates (and the expression of $\lambda_s(x,x)$), for all $f \in \mathcal{L}^\infty(E)$,

$$\begin{aligned}\mathbb{E}\left(f(X_t) \mid X_s = x\right) &= \sum_{y\in E} f(y)\mathbb{P}(X_t = y \mid X_s)\\ &= f(x) + \sum_{y\in E}\lambda_s(x,y)\left(f(y) - f(x)\right) + \underset{t\to s}{o}(t-s)\\ &:= f(x) + L_sf(x) + \underset{t\to s}{o}(t-s)\,.\end{aligned}$$

We call L_s the generator of the process (simply denoted L for an homogeneous process). This asymptotic expansion, together with the semi-group property, ensures that $(P_{s,t}f)_{t\geqslant s\geqslant 0}$ solves the linear ODE (recall a function on E is equivalently a vector of $\mathbb{R}^E$)

$$\partial_t P_{s,t}f \;=\; P_{s,t}L_tf\,, \qquad P_{s,s}f = f\,.$$

This is the so-called Kolmogorov backward equation associated with the process. In the homogeneous case, this is simply solved by $P_t = e^{tL}$.

Now, although some operators and ODE have appeared, and as far as such distinctions are relevant[2], this is still more of a probabilistic than PDEist point of view. The frontier is rather thin: we just have to write all this again but on the other side of the probability measures/test functions duality. Namely, consider the transition kernel

$$p_{s,t}(x,y) \;=\; \mathbb{P}\left(X_t = y \mid X_0 = x\right)\,,$$

so that for all $f \in \mathbb{R}^E$,

$$P_{s,t}f(x) \;=\; \sum_{y\in E} f(y)p_{s,t}(x,y)\,.$$

For a general initial distribution π_s of X_s, conditionning by the value of X_s, the law π_t of X_t is then

$$\pi_t(x) \;=\; \sum_{y\in E}\pi_0(y)p_{s,t}(y,x) \;=\; \pi_sP_{s,t}$$

(understood as the image of a probability measure by a linear transformation on the test funtions; or here, in finite dimension, by the left multiplication of a line vector by a square matrix, while the action on $P_{s,t}$ on test function is the right multiplication of a column vector by a square matrix). By definition of the jump rates it is straightforward to derive an ODE for π_t or $p_{s,t}$ (which is π_t in the particular cases $\pi_s = \delta_x$ for $x \in E$), but let us rather obtain it by duality. Indeed, for all $f \in \mathbb{R}^E$,

$$\partial_t\left(\pi_tf\right) \;=\; \partial_t\left(\pi_sP_{s,t}f\right) \;=\; \pi_sP_{s,t}L_tf \;=\; \pi_tL_tf\,.$$

[2] so, not much.

In other words, $\partial_t \pi_t = \pi_t L_t$, that is

$$\partial_t \pi_t(x) \;=\; \sum_{y \in E} \lambda_t(y,x)\,(\pi_t(y) - \pi_t(x))\;.$$

This is the so-called Kolmogorov forward equation, or Fokker-Planck equation, associated with the process.

Although this may be a bit pedantic as far as this finite dimensional linear algebra framework is concerned, we can say that $P_t f(x)$ describes *where the particles that started at x are at time t* – this is a Lagrangian description that follows tagged particles – while $\pi_t(x)$ describes *how many particles there are at x at time t* – this is the Eulerian point of view of a stationary observer who measures the flow of particles going through his fixed position. Of course they are perfectly equivalent, and obtained one from the other by duality.

Finally, let us discuss the question of reversibility (in the probabilistic sense, not in the physicist one) in the time homogeneous case. Suppose that the process admits an invariant measure (or equilibrium) μ, so that $\pi_t = \mu$ is a constant solution of the forward equation, i.e.

$$\forall f \in \mathbb{R}^E\,, \quad \int_E Lf \mathrm{d}\mu \;(= \mu L f) \;= 0 \qquad \text{or equivalently} \qquad \mu L = 0\,.$$

Then a natural object to consider is

$$h_t(x) \;=\; \frac{\pi_t(x)}{\mu(x)}$$

the relative density of π_t with respect to the equilibrium (say the support of μ is E). The convergence of π_t to μ is equivalent to the convergence of h_t to $1 = \int h_t \mathrm{d}\mu$. The evolution of h_t is again obtained by duality by

$$\forall f \in \mathbb{R}^E\,, \qquad \partial_t \int_E f h_t \mathrm{d}\mu \;=\; \partial_t \int_E f \mathrm{d}\pi_t \;=\; \partial_t \int_E P_t f \mathrm{d}\pi_0 \;=\; \int_E P_t L f \mathrm{d}\pi_t \;=\; \int_E L f h_t \mathrm{d}\mu\,,$$

in other words $\partial h_t = L^* h_t$ where L^* is the dual of L but in the $L^2(\mu)$ sense (and no more in the measures/functions duality). Equivalently, $h_t = (P_t)^* h_0 = e^{tL^*} h_0$. Now, by definition, the process is said to be reversible with respect to the measure μ if $L^* = L$ (which implies μ is invariant). This means that, for a reversible process, the backward equation is the forward equation *written at the level of the density with respect to the equilibrium*. This is not the case for a non-reversible process but in those cases it is still true that the backward equation has a form that is similar to the equation satisfied by the density of the law with respect to the equilibrium. Also, the operator norm of $P_t - \mu$ and $P_t^* - \mu$ in $L^2(\mu)$ is the same, and

$$\|h_t - 1\|^2_{L^2(\mu)} \;=\; \int_E (h_t - 1)^2 \,\mathrm{d}\mu \;=\; \int_E \frac{(\pi_t - \mu)^2}{\mu} \mathrm{d}x \;=\; \|\pi_t - \mu\|^2_{L^2(1/\mu)}\,,$$

which explains the different (but equivalent) points of view that can be found in the literature.

1.1.2 *The infinite dimensional case: informal discussion*

For a Markov process $(X_t)_{t\geqslant 0}$ with values on $\mathbb{R}^d$, $d \in \mathbb{N}_*$, everything works exactly as in the finite case except that all the definitions have to be proven to make sense in some way. For instance, stochastic calculus gives sense to stochastic differential equations

$$\mathrm{d}X_t \;=\; a_t(X_t)\mathrm{d}t + \sigma_t(X_t)\mathrm{d}W_t \tag{1.2}$$

where a and σ are a vector and matrix fields on $\mathbb{R}^d$ with some regularity conditions (depending on how sophisticated the theory is) and $(W_t)_{t\geqslant 0}$ is a Brownian motion on $\mathbb{R}^d$, or possibly another Lévy process. Assuming non-explosion (i.e. if the solutions of the SDE are defined for all times) this gives the trajectorial definition of a Markov process, so that the semi-group $P_{s,t}$ given by (1.1) is well-defined, at least on $\mathcal{L}^\infty(\mathbb{R}^d)$, possibly under a larger space (like $L^2(\mu)$ where μ is an invariant measure of the process). The definition of the generator

$$L_t f \;=\; \lim_{s\to 0} \frac{P_{t,t+s}f - f}{s}$$

has to be specified: in which topology has this limit to be understood ? In most cases the limit doesn't exist for all functions f, only on some domains, and since domains are usually not explicit it may be simpler in practice to identify so-called cores (like compactly supported $\mathcal{C}^\infty$ functions). For instance, the generator associated with (1.2) is

$$L_t f(x) \;=\; a_t(x)\cdot\nabla_x f(x) + \sum_{i,j=1}^{d} \left(\sigma_t(x)\sigma_t(x)^T\right)_{i,j} \partial_{x_i}\partial_{x_j} f(x)\,, \tag{1.3}$$

which has a clear meaning only for $f \in \mathcal{C}^2(\mathbb{R}^d)$. Now the Kolmogorov backward equation

$$\partial_t P_{s,t} f \;=\; P_{s,t} L_t f\,, \qquad P_{s,s} f = f$$

is no more an ODE but a "genuine" infinite-dimensional (linear) PDE that has to be given some possibly weak sense. The law π_t of X_t have a nice smooth density in the case of hypoelliptic diffusions, but it is not necessarily the case for other Markov processes like Piecewise Deterministic Markov Processes (PDMP). In any cases, denoting L_t' the dual of L_t (in the sense of measure/functions duality) the Kolmogorov forward equation

$$\partial_t \pi_t \;=\; L_t'\pi_t\,,$$

can still be obtained by duality from the backward one. On the contrary, from the PDE viewpoint, the forward equation is generally the starting point, obtained by Eulerian considerations of mass balance, and then the backward equation is possibly used to defined weak solutions of the forward one.

There are several ways to answer the technical questions raised in this informal discussion and to design correct theoretical frameworks in which the objects introduced above can be safely considered. This do not enter the scope of the current memoir, and we refer to classical reference books like [Kal02, Dav93, EK86] on this matter.

1.2 Little zoo of Markovian dynamics

1.2.1 *Diffusion processes*

Diffusions are the Markov processes whose trajectory are continuous, i.e. $t \mapsto X_t(\omega)$ is continuous for $\mathbb{P}$-almost all $\omega \in \Omega$. In sufficiently regular cases (which is always the case for the processes I have been studying) they are solutions of SDEs of the form (1.2), and their generator is of the form (1.3). In other words, from a PDE point of view, diffusion processes correspond to (second-order) differential operators, namely to local operators. During my work, I have mainly encountered three diffusion processes.

The overdamped Langevin diffusion (of Fokker-Planck process) is the process $(X_t)_{t\geqslant 0}$ on $\mathbb{R}^d$ that solves

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t \tag{1.4}$$

where $(W_t)_{t\geqslant 0}$ is a standard Brownian motion on $\mathbb{R}^d$ and $\beta > 0$ and $U \in \mathcal{C}^1(\mathbb{R}^d)$ are respectively called the inverse temperature and potential (or energy) of the dynamics. Its generator is

$$Lf(x) = -\nabla U(x)\cdot\nabla f(x) + \frac{1}{\beta}\Delta f(x)\,,$$

and its unique invariant measure is the Gibbs measure with energy U and inverse temperature β, namely is the probability measure on $\mathbb{R}^d$ with Lebesgue density

$$\mu(x) = \frac{1}{\mathcal{Z}}e^{-\beta U(x)}\,, \qquad \mathcal{Z} = \int_{\mathbb{R}^d} e^{-\beta U(z)}\mathrm{d}z\,, \tag{1.5}$$

provided $\mathcal{Z} < \infty$. The overdamped Langevin diffusion is very nice since it is elliptic, coercive, reversible and the Wasserstein gradient flow of the entropy [OV00], so that many theoretical tools (see e.g. [BGL14]) are available to establish, under various conditions on V, smoothness and integrability of the law of X_t, long-time convergence of the latter toward μ (with possibly explicit speed in $L^2(\mu)$ or Wasserstein metrics or in relative entropy), ergodicity in the sense that

$$\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s \underset{t\to\infty}{\longrightarrow} \int f\mathrm{d}\mu$$

for f in some classes of functions, Large Deviations and Eyring-Kramers formula (i.e. estimates on escape times from metastable states at small temperature: $\beta \to +\infty$, see [BEGK04]), etc.

In many places in this memoir the overdamped Langevin will be used as a basic benchmark to introduce various questions and notions.

The (underdamped, kinetic) Langevin diffusion (or kinetic Fokker-Planck process) is the process $(X_t, V_t)_{t\geqslant 0}$ on $\mathbb{R}^d \times \mathbb{R}^d$ that solves

$$\left\{\begin{array}{rcl} \mathrm{d}X_t &=& V_t\mathrm{d}t \\ \mathrm{d}V_t &=& -\nabla U(X_t)\mathrm{d}t - \gamma V_t + \sqrt{2\gamma\beta^{-1}}\mathrm{d}W_t \end{array}\right. \tag{1.6}$$

with $\gamma > 0$. It can be seen as the Newton's law applied to a particle (here with a mass normalized to 1) that undergoes forces from a potential energy, friction and dissipation (as the resulting forces of microscopic shocks). Alternatively, it is the combination of the Hamiltonian dynamics (case $\gamma = 0$) with an Ornstein-Uhlenbeck process on the velocities (called a Langevin thermostat in molecular dynamics). The generator is $L = L_H + \gamma L_{OU}$ with the Hamiltonian and thermostat parts given by

$$L_H f \;=\; v \cdot \nabla_x f - \nabla U(x) \cdot \nabla_v f\,, \qquad L_{OU} f \;=\; -v \cdot \nabla_v f + \frac{1}{\beta} \Delta_v f\,.$$

The unique invariant measure is the Gibbs measure with inverse temperature β and Hamiltonian $H(x,v) = U(x) + |v|^2/2$. This Hamiltonian is separable, namely, at equilibrium, position and velocity are independent, respectively distributed according to the Gibbs law with energy U and to a Gaussian measure. The Hamiltonian part is skew-symmetric ($L_H^* = -L_H$) while the Ornstein-Uhlenbeck part is self-adjoint. In particular, the Langevin diffusion is not reversible, nor is it elliptic or coercive. Under some conditions on U, it is still hypoelliptic, hypocoercive and ergodic, which means that the law of the process is still smooth and converges exponentially fast toward equilibrium, but many theoretical tools that are available in the overdamped case do not apply in the kinetic case. The overdamped process is the limit process as $\gamma \to +\infty$ of position of the Langevin diffusion (when time is accelerated by a factor γ).

General Ornstein-Uhlenbeck (OU) processes are Markov processes that are also Gaussian processes, namely are such that $(X_{t_1}, \dots, X_{t_N})$ are Gaussian vectors for all $N \in \mathbb{N}_*$ and $(t_1, \dots, t_N) \in \mathbb{R}_+^N$. They solve SDEs of the form

$$\mathrm{d}X_t \;=\; AX_t \mathrm{d}t + \sigma \mathrm{d}W_t \tag{1.7}$$

where A and σ are constant matrices. Denoting $\Sigma = \sigma\sigma^T$, the generator is simply

$$Lf \;=\; (Ax) \cdot \nabla f + \nabla \cdot (\Sigma \nabla f)\,.$$

For instance, the overdamped and kinetic Langevin are particular cases of OU processes in the case of an harmonic potential $U(x) = x^T R x$ for some matrix R. Many explicit formulas and properties boil down to linear algebra questions for OU processes. For instance, the process is stable if and only if the spectrum of A lies in $\{z \in \mathbb{C},\, \mathfrak{Re}(z) < 0\}$, and it is hypoelliptic if and only if there exist $M \in \mathbb{N}_*$ such that $\sum_{i=1}^N A^k \Sigma (A^T)^k$ is positive definite.

When it is both stable and hypoelliptic, it admits a unique invariant measure which is the Gaussian measure with mean 0 and covariance K that solves the equation $AK + (AK)^T = -2D$.

1.2.2 *Piecewise Deterministic Markov processes (PDMP)*

A PDMP on some space E is defined by three elements:

1. a deterministic flow on E, in the simplest case given by an ODE $\dot{x} = F(x)$.
2. a jump rate $\lambda : E \to \mathbb{R}_+$.
3. a jump kernel $Q : E \to \mathcal{P}(E)$.

The process $(X_t)_{t\geqslant 0}$ is then defined by induction as follows. Let $(S_i)_{i\in\mathbb{N}}$ be a sequence of independent standard exponential variables and let $T_0 = 0$. Suppose the process has been defined up to the n^{th} jump time T_n for some n and let $(x_s)_{s\geqslant 0}$ be the deterministic flow given by $x_0 = X_{T_n}$ and $\dot{x}_s = F(x_s)$. Then the next jump time T_{n+1} is given by

$$T_{n+1} \;=\; T_n + R_n\,, \qquad \text{with} \qquad R_n = \inf\left\{ t\geqslant 0,\; S_n \leqslant \int_0^t \lambda(x_u)\mathrm{d}u \right\}\,,$$

we set $X_{T_n+t} = x_t$ for all $t\in[0,R_n[$ and draw $X_{T_{n+1}}$ according to the distribution $Q(x_{R_n})$ (independently from the past conditionally to x_{R_n}). Under suitable conditions (for instance if the jump rate is bounded) there cannot be an infinite number of jumps in a finite time interval so that $(X_t)_{t\geqslant 0}$ is defined for all times. The associated generator is

$$Lf(x) \;=\; F(x)\cdot\nabla f(x) + \lambda(x)\left(Qf(x) - f(x)\right)$$

where we associate to the Markov kernel $Q : E \to \mathcal{P}(E)$ the operator on $\mathcal{L}^\infty(E)$ given by $Qf(x) = \mathbb{E}(f(Z))$ with $Z\sim Q(x)$. Note that Q is not a local operator.

PDMP are popular in reliability problems or similar modelling contexts, the systems undergoing a deterministic evolution (aging) while random failures occur at some rate depending on the current state. Here are some examples of PDMP (see also [Dav93, ABG+14]).

TCP-like processes. The Transmission Control Protocol (TCP) is a protocol that defines how streams of data are transmitted between devices : as long as no error is observed, the flux is increased (say linearly), which increases the rates of error occurrence, and when an error is observed the flux is divided by some factor (say 2). A very simple model is the PDMP with generator on $\mathbb{R}_+$

$$Lf(x) \;=\; f'(x) + \lambda(x)\left(f\left(\frac{x}{2}\right) - f(x)\right)\,,$$

where $\lambda(x)$ is the failure rate when the flux is x. A similar process that models the concentration of chemicals in a body, ingested at random times and eliminated at constant rate $\alpha > 0$, is given by the generator

$$Lf(x) \;=\; -\alpha f'(x) + \lambda\left(\int_{\mathbb{R}_+} f(x+h)h(y)\mathrm{d}y - f(x)\right)\,,$$

with $\lambda > 0$ a constant rate and h some probability distribution over $\mathbb{R}_+$. See [BCG+13] and references within.

Switched flows constitutes a class of PDMP described by a position X_t in some space E (say $\mathbb{R}^d$) and an index I_t typically in a finite set, say $[\![1,N]\!]$. For each $i\in[\![1,N]\!]$, we are given a vector field F_i such that the flow associated to $\dot{x} = F_i(x)$ is well-defined. We also consider jump rates $\lambda_x(i,j)$ for $x\in E$ and $i,j\in[\![1,N]\!]$. Then $(X_t,I_t)_{t\geqslant 0}$ is the Markov process with generator

$$Lf(x,i) \;=\; F_i(x)\cdot\nabla_x f(x,i) + \sum_{j=1}^N \lambda_x(i,j)\left(f(x,j) - f(x,i)\right)\,.$$

In other words, I_t jumps at rate $\lambda_{X_t}(i,j)$ from i to j while X_t follows the flow $\dot{x} = F_{I_t}(x)$. When the jump rates do not depend on x, then $(I_t)_{t\geqslant 0}$ is a Markov chain on $[\![1,N]\!]$ by itself, and $(X_t)_{t\geqslant 0}$ is sometimes said to be a Markov-modulated ODE. Like the similar Hidden Markov Chain models, they are a simple way to describe non-Markovian processes $(X_t)_{t\geqslant 0}$ with only a few more parameters. See [BLMZ12, BLMZ15] and references within.

Velocity jump processes are kinetic PDMP, namely the process is $(X_t, V_t)_{t\geqslant 0}$ with $\mathrm{d}X_t = V_t \mathrm{d}t$ (so that X_t is the position and V_t the velocity) and V_t is piecewise constant. Then, at a rate $\lambda(x,v)$, the velocity jumps from v to $Q(x,v)$ for some kernel Q. In particular, the trajectory $t \mapsto X_t$ is continuous (but not Markovian, since the velocity is required). Since velocity jump processes are an important and recurrent topic of my work, a more detailed presentation is postponed to Section 2.3.

1.3 Sampling and optimization with stochastic algorithms

The sampling problem is the following: given a target probability measure μ, how can we draw in practice a random variable with law μ ? And how can we estimates expectations with respect to μ ? In many (high-dimensional) applications, μ is only known up to a multiplying factor, typically μ is a probability measure on $\mathbb{R}^d$ with Lebesgue density proportional to $\exp(-U)$ for some energy (or log-likelihood) U but the normalization constant $\mathcal{Z} = \int_{\mathbb{R}^d} \exp(-U(x))\mathrm{d}x$ is intractable. In this context, Markov Chain Monte Carlo (MCMC) algorithms are based on the construction of a Markov process $(X_t)_{t\geqslant 0}$ ergodic with respect to μ so that theoretical expectations $\int f \mathrm{d}\mu$ can be approximated by an empirical average $1/t \int_0^t f(X_s)\mathrm{d}s$ over a long trajectory ($t \to \infty$). Such a Markov process can always be designed by the general Metropolis-Hastings algorithm. Alternatively, the overdamped and kinetic Langevin processes introduced in Section 1.2 are also both suitable for this task, since the knowledge of $\mathcal{Z}$ is not required to sample them. Remark that, in the case of the kinetic Langevin process, if the observable f only depends on the position variable, by ergodicity (take $\beta = 1$)

$$\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s \underset{t\to\infty}{\longrightarrow} \frac{\int_{\mathbb{R}^d\times\mathbb{R}^d} f(x)e^{-H(x,v)}\mathrm{d}x\mathrm{d}v}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-H(x,v)}\mathrm{d}x\mathrm{d}v} = \frac{\int_{\mathbb{R}^d} f(x)e^{-U(x)}\mathrm{d}x}{\int_{\mathbb{R}^d} e^{-U(x)}\mathrm{d}x} = \int_{\mathbb{R}^d} f\mathrm{d}\mu\,,$$

in other words from an MCMC point of view the velocity $(V_t)_{t\geqslant 0}$ is just an auxiliary variable required to sample $(X_t)_{t\geqslant 0}$. In the following, we treat the general case, denoting $(Z_t)_{t\geqslant 0}$ the Markov process (for instance $(X_t, V_t)_{t\geqslant 0}$ in the Langevin case), μ its equilibrium, P_t the associated semi-group and L its generator.

Starting with a distribution π_0, the systematic bias of the scheme is given by

$$\mathrm{bias}(t) = \mathbb{E}\left(\frac{1}{t}\int_0^t f(Z_s)\mathrm{d}s - \int_{\mathbb{R}^d} f\mathrm{d}\mu\right) = \frac{1}{t}\int_0^t\int_{\mathbb{R}^d} (P_s - \mu)f(x)\pi_0(\mathrm{d}x)\mathrm{d}s\,.$$

By the Jensen and Cauchy-Schwarz inequalities, thus,

$$|\mathrm{bias}(t)|^2 \;\leqslant\; \frac{1}{t}\int_0^t \|P_s-\mu\|_{L^2(\mu)}\mathrm{d}s\|f\|_{L^2(\mu)}\left\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu}\right\|_{L^2(\mu)} \;\leqslant\; \frac{C}{\rho t}\|f\|_{L^2(\mu)}\left\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu}\right\|_{L^2(\mu)}$$

if $\|P_t-\mu\|_{L^2(\mu)} \leqslant C\exp(-\rho t)$ for some $C,\rho>0$. Alternatively, if we run in parallel $N\in\mathbb{N}_*$ independent replicas of the Markov process and estimates $\int f\mathrm{d}\mu$ by the empirical average over the replicas at a large time t, the bias is

$$\mathrm{bias}(t) \;=\; \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^N f(Z_t^i)-\int f\mathrm{d}\mu\right) \;=\; \int_{\mathbb{R}^d}(P_t-\mu)f(x)\pi_0(\mathrm{d}x)\,.$$

so that, provided exponential decay,

$$|\mathrm{bias}(t)|^2 \;\leqslant\; Ce^{-\rho t}\|f\|_{L^2(\mu)}\left\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu}\right\|_{L^2(\mu)}\,.$$

In any cases, this bias is small when the simulation time is large with respect to the mixing rate ρ. Concerning the variance part of the quadratic error, a Markovian CLT

$$\sqrt{t}\left(\frac{1}{t}\int_0^t f(Z_s)\mathrm{d}s-\int f\mathrm{d}\mu\right)\underset{t\to\infty}{\longrightarrow}\mathcal{N}(0,\sigma_f^2)$$

usually holds in practical cases, with an asymptotic variance

$$\sigma_f^2 \;=\; -2\int_{\mathbb{R}^d} fL^{-1}f\mathrm{d}\mu \qquad \forall f\in L^2(\mu)\text{ with }\int f\mathrm{d}\mu=0\,.$$

Indeed, provided exponential decay, the kernel of L is reduced to constant functions and, on the orthogonal of this kernel (i.e. if $\int f\mathrm{d}\mu=0$), L^{-1} is bounded :

$$L^{-1}f(x)=\int_0^\infty P_t f\mathrm{d}t \qquad\Rightarrow\qquad \|L^{-1}f\|_{L^2(\mu)}\leqslant\frac{C}{\rho}\|f\|_{L^2(\mu)} \qquad\Rightarrow\qquad \sigma_f^2\leqslant\frac{2C}{\rho}\|f\|_{L^2(\mu)}^2\,.$$

The variance of the empirical average thus scales at $1/(\rho t)$, which is consistent with the discussion on the bias (ignoring again the constant C, which is not necessarily reasonable in all cases). Consequently, in general (i.e. without considering particular observables f), in order to minimize the quadratic error, a reasonable aim in practice is to design a process whose law converges as fast as possible to μ (although there are simple examples where the empirical averages converge very fast to their limit while the law does not converge to μ, for instance $\dot{x}=1$ on the torus).

As a conclusion, in order to get explicit error bounds on the quadratic error of the MCMC estimator (and thus be able to compare the numerical efficiency of different processes, or tune parameters like the simulation time), quantitative estimates of long-time relaxation toward equilibrium are necessary, with in particular explicit dependency of the key quantities of the problem (like the dimension). We have seen above computations in $L^2(\mu)$ but instead of the Cauchy-Schwarz inequality, if f is bounded, we can use

the Pinsker inequality to get

$$|\pi_t f - \mu f| \;\leqslant\; \|f\|_\infty \|\pi_t - \mu\|_{TV} \;\leqslant \|f\|_\infty \sqrt{\int_{\mathbb{R}^d} \ln\left(\frac{\mathrm{d}\pi_t}{\mathrm{d}\mu}\right) \mathrm{d}\pi_t}\,.$$

Similarly, estimates in Wasserstein-1 distance yield controls for Lipschitz functions, and estimates in $\mathcal{V}$-norms where $\mathcal{V} \geqslant 1$ is some Lyapunov function control errors for functions f with $|f| \leqslant \mathcal{V}$.

1.3.1 *Metastability*

Designing a Markov process whose law converges fast to a given target distribution μ is easier said than done. Indeed, convergence can only be achieved in a time where the process has had a reasonable probability to have visited all the regions of the space that have a significant mass with respect to μ. To fix ideas, say that $\mu \propto \exp(-\beta U)$ for some potential U and inverse temperature $\beta > 0$. At small temperature (large β) these regions are the neighborhoods of minima of U. It means that, in order to achieve convergence, the process should have the time to find at least the lowest minima of U. In large dimension with a non-convex U, the location of these minima is unknown. Moreover, the low-energy regions $\{x \in \mathbb{R}^d,\ U(x) \leqslant \min U + \delta\}$ for some $\delta > 0$ typically have a very small Lebesgue measure. It means that, for instance in a Metropolis-Hastings scheme, points sampled according to some isotrope distribution (like uniform in some hypercube, or isotrope Gaussian) have virtually no chance to hit these regions. It's like trying to solve a 5000-piece puzzle by trying uniformly distributed configurations of pieces. Thus, one need to use some information on U to explore the space. However, typically, we don't have any global information about U. For this reason, we are essentially restricted to a local exploration, with a process either with continuous trajectory or moving through very small jumps.

This is for instance the case of the overdamped Langevin process (1.4). Making use of a local information on the landscape ($\nabla U(x)$) it spends most of its time in low-energy regions (at $\beta = +\infty$, i.e. zero temperature, it is a gradient descent).

Now, the problem with a local exploration is the following: when U is not convex, low energy regions may be separated by high energy regions. To fix ideas, suppose there are two minima x_1 and x_2 separated by a region $\mathcal{D} \subset \{U \geqslant U(x_1) \wedge U(x_2) + \Delta\}$ for some $\Delta > 0$. By ergodicity, the ratio of the time spent in the neighborhood of x_1 and x_2 and in $\mathcal{D}$ should converge in large time to the ratio of the invariant probabilities of these regions, namely should be of order larger than $\exp(\beta\Delta)$ for large β. On the other hand, since the exploration is local, each time the process crosses $\mathcal{D}$ to go from x_1 to x_2 or back, it spends a time of order independent from β. This necessarily implies that, for at least one of the two minima, starting from there, the time before the first crossing of $\mathcal{D}$ to reach the other minimum is of order at least $\exp(\beta\Delta)$. This is the so-called Arrhenius Law. As a consequence, convergence cannot be achieved in a time less than $\exp(\beta\Delta)$, which is pretty bad (independently from the Markov sampler as long as it moves locally, according to the previous argument – although we won't try to state a rigorous theorem here).

This is the so-called *metastability* phenomenon. In fact it appears in other contexts than MCMC algorithms at small temperature. In general, a process is called metastable if

it stays for very long time (depending on some parameter of the problem) close to some states before short and unpredictable/memory-less (i.e. happening in a time that follows approximately an exponential distribution) transitions to another part of the space. This happens in particular when the process converges in some regime to a deterministic ODE that has an attracting point, in which case the stochastic process stays close to this attracting point up to a large deviation occurs in its random fluctuations that brings it out of the basin of attraction of the deterministic equilibrium. For the overdamped Langevin, the deterministic ODE at $\beta = +\infty$ is the gradient descent flow.

The situation is similar for instance in some population models. Take the following very simple example : the number of individuals $N(t)$ is modelled by a birth-and-death random process, where each individual dies at rate $r_d > 0$ and gives birth to a new individual at rate $r_b(1 - N(t)/N_0)$, where N_0 is the saturation size of the environment. If $r_b > r_d$ then the two following statements hold true simultaneously: 1) in a finite time interval, by a Law of Large Numbers, $N(t)/N_0$ converges as $N_0 \to \infty$ toward the solution of the deterministic logistic ODE $\dot{x} = x(r_b(1-x) - r_d))$ that admits a globally attractive stationary solution $x_0 > 0$, and 2) extinction is almost sure (i.e. with probability 1, $N(t)$ reaches the absorbing state 0 at some time), since for all time intervals $[t, t+s]$ there is a probability larger than $[\exp(-r_b s)(1 - \exp(-r_d s))]^{N_0} > 0$ that during $[t, t+s]$ no birth occurs and all individuals die. In this case, the system will typically stay close to $N_0 x_0$ (with fluctuations of order $\sqrt{N_0}$) for a time that is exponentially large with N_0 before an abrupt extinction.

There are several ways to make rigorous statements about metastability:

1. Establish that the expectation of the exit time τ from some domain is exponentially large with respect to the parameter β or N, i.e. prove the Arrhenius law, or an Eyring-Kramer formula, that is a refinement of the former where the prefactor in front of the exponential term is specified.

2. Establish the asymptotic exponentiality of τ, namely that $\mathbb{P}(\tau > t\mathbb{E}(\tau))$ converges uniformly in $t > 0$ toward e^{-t} as β or N goes to infinity.

3. Establish that the convergence rate of the law of the process (in the entropy sense, or L^2, Wasserstein, etc.) is exponentially small as a function of β or N.

In population dynamics, metastability means survival. In stochastic algorithms, metastability means poor convergence properties and need for very long simulations, and should be fought. In any case, from a theoretical point of view, this is a nice phenomenon to study.

To go further on the topic of metastability, see for instance [Lel13, BdH15] or [15, 11].

1.3.2 *Simulated annealing*

The issue raised above about the difficulty to explore a high-dimensional non-convex landscapes still holds in the context of optimization (rather than sampling). In particular, the gradient descent flow $\dot{x} = -\nabla U(x)$ locally converge to critical points of U, generically a local minimum, which has no reason to be a global one. A classical algorithm to solve this is to consider the overdamped Langevin diffusion, or more precisely an annealed version :

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2\beta_t^{-1}}\mathrm{d}W_t \tag{1.8}$$

where β_t increases with time and goes to $+\infty$ (in other words, the system is cooled down along time, down to zero temperature). Indeed, for a fixed β, the overdamped Langevin converges in law in large times toward the Gibbs measure $\propto \exp(-\beta U)$, whose mass concentrates as $\beta \to +\infty$ on the global minima of U. As a consequence, the law of the process (1.8) may hopefully concentrates as $t \to +\infty$ on the global minima of U. Nevertheless, for the previous heuristic to hold, we have to give enough time for the law of the process X_t to approach its current target $\exp(-\beta_t U)$ before changing this target. In other words, the temperature β_t^{-1} should decrease sufficiently slowly.

This limitation is easily understood in a toy problem with 3 states $\{0,1,2\}$, with the overdamped Langevin diffusion replaced by a discrete-time Metropolis-Hastings with nearest neighbours proposal (which has the same property that, at a fixed β, the equilibrium is $\propto \exp(-\beta U)$). Suppose that $U(1) > U(0) > U(2)$ and that we start at 0. The probability to move from 0 to 1 in step $t \in \mathbb{N}_*$ is $e^{-\beta_t \Delta}$ with $\Delta = U(1) - U(0)$. From the Borel-Cantelli theorem, if

$$\sum_{k \in \mathbb{N}_*} e^{-\beta_t \Delta} \; < \; +\infty$$

then there is some positive probability that the process never get to 1 (and thus to 2), which means that $\limsup \mathbb{P}(U(X_t) = \min U) < 1$. On the other hand, if the the sum is infinite then almost surely the process will reach the state 2 at some point, and then it is not hard to see that $\mathbb{P}(U(X_t) = \min U) \to 1$ as $t \to +\infty$. This means that the cooling schedule $(\beta_t)_{t \in \mathbb{N}_*}$ should grow at most logarithmically in t. More precisely, if $\beta = 1 + c \ln t$, then there is a phase transition at $c\Delta = 1$: if $c\Delta < 1$, the process converges in probability to the global minimum of U, and if $c\Delta \geqslant 1$ the algorithm has a positive probability to fail.

Under a few conditions on U, this is exactly what happens for the process (1.8): there exist a constant Δ, called the critical depth of the potential U (explicit in theory, unknown in practice) such that if $\beta_t = c \ln t$ with $c < \Delta$ then $\limsup \mathbb{P}(U(X_t) = \min U) < 1$ (for all initial condition) while for $c > \Delta$, $\mathbb{P}(U(X_t) = \min U) \to 1$. A failure may happen if the process has a non-zero probability to stay stuck forever in a domain that does not contain any global minimum of U.

As we see in this short presentation of the problem, the question of convergence of the simulated annealing is related both to the escape time from metastable domains, and to the scaling in β of the long-time convergence rate of the process (at low but fixed temperature) to its equilibrium.

Note that the theoretical criteria on the cooling schedule to ensure convergence in probability toward the global minima is far from the end of the story for the study and use of simulated annealing algorithms. Indeed, in practice, logarithmic schemes yield a very slow convergence, and thus faster schemes can be used, with the objective to find a sufficiently low (although possibly not global) minimum with good probability in a reasonable time. Nevertheless, my own work on simulated annealing is concerned with the theoretical criteria to ensure convergence in probability.

To go further, see e.g. [HKS89, Mic92] or [15, 19].

1.3.3 *Non-reversibility*

For either sampling (MCMC) or optimization (simulated annealing) purpose, we have seen that the objective is to explore the space in the more efficient way. However, Markov processes are lousy explorers as they don't learn from what they have already seen. Metastability around a local minimum is an example where the process repeatedly "rediscover" the same local minimum over and over again for a very long time. In fact, the worst way to explore is to stay at the same position, which is what a Metropolis-Hastings acceptance/rejection step does when there is rejection. This explains that the Metropolis-Hastings suffers in many contexts from a high variance, so that for the overdamped Langevin for instance it may be more efficient in order to reduce the quadratic error to have a systematic discretization bias with no rejection step than to add a rejection step [DM17].

When the process doesn't stay exactly at the same place, then the second worst way to explore is to go back to the place we were in the previous step. Yet, for a reversible process, the detailed balance condition $\mu(x)p_t(x,y) = \mu(y)p_t(y,x)$ precisely prescribes that if there is a probability to go from x to y then there is necessarily a probability to go back to x in the next step. It means that, to reduce backtracking, we should leave the reversible realm.

Consider generators L_s and L_a respectively symmetric and antisymmetric with respect to some measure μ (which implies that μ is invariant for both L_s and L_a). Suppose that L_s has a spectral gap $\rho > 0$, which means that $\langle f, L_s f\rangle \leqslant -\rho\|f\|^2$ where $\langle\cdot\rangle$ and $\|\cdot\|$ stands for the scalar product and norm of $L^2(\mu)$. Considering as in Section 1.1 h_t the solution of $\partial_t h_t = (L_s + L_a)h_t$ with $\int h_0 d\mu = 0$, we get that

$$\partial_t\|h_t\|^2 \;=\; 2\langle h_t, (L_s+L_a)h_t\rangle \;=\; 2\langle h_t, L_s h_t\rangle \;\leqslant -2\rho\|h_t\|^2 \qquad \Rightarrow \qquad \|h_t\| \;\leqslant\; e^{-\rho t}\|h_0\|\,.$$

In other words, $\|e^{t(L_s+L_a)}\| \leqslant e^{-\rho t} = \|e^{tL_s}\|$. Adding a non-reversible part to L_s can only improve the convergence rate in $L^2(\mu)$.

It is also possible that the reversible part by itself has no spectral gap, while the non-reversible generator does. This is typically the case for kinetic processes, or more generally lifted processes. The general idea of lifted Markov chain is the following : if we want to sample some distribution μ on a space E, then we can consider a process (X,Y) on $E \times F$ for some space F with equilibrium $\mu \otimes \nu$ for some auxiliary distribution ν on F. Such construction often makes the design of non-reversible processes easier. A major example is the case where $Y = V = dX/dt$ is the velocity of the process X. In discrete time, it means $V_n = X_n - X_{n-1}$, in other words instead of constructing a Markov chain $(X_n)_{n\in\mathbb{N}}$ we construct a chain $(X_n)_{n\in\mathbb{N}}$ which is *not* Markovian but such that $(X_n, V_n)_{n\in\mathbb{N}}$, or alternatively $(X_n, X_{n-1})_{n\in\mathbb{N}}$, is a Markov chain. This is a simple way to go beyond the Markovian constraint (although an enlarged process is still Markovian), with a short-term memory. Also, note that adding the velocity to the process allows to reduce backtracking. Indeed, giving some persistence to the velocity – some inertia – forces the process to go roughly in the same direction for several steps, which can improve the exploration. Indeed, in that case, the behaviour of the process is ballistic, covering a distance of order t in a time t, while a reversible walker has a diffusive behaviour, covering a distance of order $\sqrt{t}$ in a time t (due to backtracking).

To go further on the topic of non-reversible sampling, see for instance [LNP13, Nea04] and references within.

1.4 Molecular dynamics

In some sense, in theory, the motion of atoms has been "solved" in 1925 with the Schrödinger equation (as said Dirac in 1929 : *The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble*). Since then, all the work for simulations has been devoted to design approximations of this (numerically completely intractable) ideal model. Following the Born–Oppenheimer approximation, in many cases we can decouple the motion of electrons and of atomic nuclei, and consider only electrons as quantum particles, while the nuclei are treated like classical Newtonian pointwise particles. Denoting $(q,p) = (q_i,p_i)_{i\in[\![1,N]\!]} \in \mathbb{R}^{3N}$ the positions and momenta of N nuclei, they are thus supposed to follow the classical Hamiltonian dynamics

$$\dot{q} \;=\; M^{-1}p \;=\; \nabla_p H(q,p) \qquad\qquad \dot{p} \;=\; -\nabla E(q) \;=\; -\nabla_q H(q,p) \tag{1.9}$$

where $M = diag(m_1 I_3, \dots, m_N I_3)$ is the mass matrix and $H(q,p) = E(q) + p^T M^{-1} p/2$ is the Hamiltonian of the system, sum of the potential and kinetic energy. The potential energy E may contain a contribution from the quantum density of electrons (determined for a fixed q) or be constituted of sums of empirical potential, that are supposed to approximate the quantum reality. The simplest empirical potentials, like the Coulomb and Van der Waals one, involve interaction between pairs of particles :

$$E_{Coulomb}(q) \;=\; \sum_{i}^{N}\sum_{j\neq i}\frac{\varepsilon_{i,j}}{|q_i-q_j|}, \qquad E_{VdW}(q) \;=\; \sum_{i=1}^{N}\sum_{j\neq i} E_{i,j}\left(\left(\frac{\sigma_{i,j}}{|q_i-q_j|}\right)^{12} - \left(\frac{\sigma_{i,j}}{|q_i-q_j|}\right)^{6}\right)$$

where the parameters $\varepsilon_{i,j}, E_{i,j}$ and $\sigma_{i,j}$ depends on the nature of the nuclei i and j. Designing, parametrizing and computing efficiently such empirical potentials is one of the many issues of the field. In the following, let us assume we are given some energy function E. Remark on the previous examples that the energy in MD is a singular function (the energy blows up when two nuclei get close).

The Hamiltonian is conserved along the deterministic dynamics (1.9). In particular, any probability density that is a function of H is left invariant by the process. Along all these distributions, the Gibbs laws $\mu_\beta \propto \exp(-\beta H)$ plays a particular role : they are the minimiser of the Boltzmann entropy $\int(\ln\mu)\mu$ among probability densities that have a given average energy $\{\nu,\ \int H d\nu = C\}$ (to a value of C corresponds a value of β, which is obtained as a Lagrange mulitplier when minimising over $f = \ln\mu$ [LRS10]). They are thus the less informative distributions with a given average energy. The reason to consider an average energy rather than a fixed one (as would be the case along the Hamiltonian trajectory) is that (1.9) describes the evolution of an isolated system, while simulations are often concerned with systems that are immersed in some non-simulated environment it exchanges energy with. Besides, alternatively, μ_β is in some sense the limit distribution of a small Hamiltonian system in contact with a much larger one (called a heat bath) at a given temperature, see details in [Tuc10, Section 4.3]. For these reasons, macroscopic quantities in statistical physics are computed as averages with respect to the Gibbs law μ_β where $\beta = 1/(k_B T)$ with T the average temperature (i.e. the expected kinetic energy up to a normalization factor) and k_B the Boltzmann's constant (at least in the NVT-ensemble, namely when the number N of particles, the volume V

and the average temperature T are fixed, which is the only case I will consider). This is thus a (high-dimensional non-convex) sampling problem.

The Langevin diffusion (1.6) is widely used in this context for sampling the Gibbs measure. Indeed, it is nothing different from a Newtonian motion where a friction and a dissipation forces, due to the interaction with the heat bath, are added to the potential force. These forces would be justified for a mesoscopic particle undergoing many collisions with small particles constituting the heat bath (like a pollen grain in water in the initial experiment of Brown), nevertheless this doesn't make sense at the level of atoms. As far as computing static averages (expectations with respect to μ_β), this is not a problem since any Markov process with the correct target distribution would be suitable, so that the Langevin thermostat (i.e. the friction/dissipation part in (1.6)) can be seen as a simple artificial way to ensure the ergodicity that lacks (1.9) and select the Gibbs law among all invariant measures of the latter. However, in many situations of chemistry or material science, dynamical properties are also of interest, like transition times between metastable domains, exit locations from such domains, reaction path (typical paths from a domain to another), mean square displacement, etc., which are not averages with respect to μ_β but averages over trajectories of the process. Of course, different Markov processes with the same statistical equilibrium may have completely different dynamical properties. Even in that context, the Langevin diffusion is still one of the most widely used in MD.

We finish this brief discussion on MD with a few order of magnitudes. Due to the very high frequency oscillations of the light hydrogen atoms, for the discretization scheme of the process, the timestep is constrained to be of the order of the picosecond (10^{-12}s). On the other hand, phenomena of interest (transitions) may happen at the scale of the second. This means that very long trajectories are needed. Moreover, in modern simulations, systems may contain millions of atoms, possibly all interacting with the others, which makes the computation of forces (required at each timestep to propagate the system) computationally expensive. It can be estimated that 40% of supercomputing power in the world nowadays is devoted to MD (for chemistry or material science)[3]

For more details on Molecular Dynamics, see [Tuc10, LRS10] and references within.

1.5 Two approaches to quantitative long-time relaxation

1.5.1 *Couplings*

Given a distance ρ on $\mathbb{R}^d$ and $p \in \mathbb{N}_*$, define the associated $\mathcal{W}_p$ Wasserstein distance over $\mathcal{P}_p(\mathbb{R}^d) := \{\nu \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} \rho^p(x,0)\nu(\mathrm{d}x) < \infty\}$ as

$$\mathcal{W}_p(\nu,\mu) \;=\; \inf_{\pi\in\Gamma(\nu,\mu)} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \rho^p(x,y)\pi(\mathrm{d}x,\mathrm{d}y) \right)^{\frac{1}{p}}$$

where $\Gamma(\mu,\nu)$ is the set of transference plans of ν and μ, namely the set of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with d-marginals μ and ν. If the distribution of a random variable (X,Y) is in $\Gamma(\mu,\nu)$ we say that (X,Y) is a coupling of μ and ν. Remark that the

[3] see for instance page 9 of the CSCS Annual Report 2016, https://www.cscs.ch/fileadmin/user_upload/contents_publications/annual_reports/Annual_Report_2016_print.pdf

total variation distance is a particular case of Wasserstein distances with $\rho(x,y) = \mathbb{1}_{x\neq y}$ the discrete distance over $\mathbb{R}^d$.

A nice feature of Wasserstein distances is that, since they are defined as an infimum (although dual representations of $\mathcal{W}_p$ as supremum over some set of functions are also possible), an upper bound is provided by any particular coupling.

Couplings are classically constructed under a Lyapunov + local Doeblin condition, with a distance $\rho(x,y) = \mathbb{1}_{x\neq y}(1+\mathcal{V}(x)+\mathcal{V}(y))$ where $\mathcal{V}$ is a Lyapunov function, see e.g. [HM11]. Another usual case is the parallel coupling for SDE, which consists in considering two solutions of the same SDE driven by the same Brownian noise but with different initial conditions. For instance, for the overdamped Langevin process, if

$$\mathrm{d}X_t \;=\; -\nabla U(X_t)+\sqrt{2}\mathrm{d}W_t \qquad \text{and} \qquad \mathrm{d}Y_t \;=\; -\nabla U(Y_t)+\sqrt{2}\mathrm{d}W_t$$

with a λ-convex U for some $\lambda > 0$ then almost surely

$$\mathrm{d}|X_t-Y_t|^2 \;\leqslant\; -2\lambda|X_t-Y_t|^2\mathrm{d}t\,, \qquad \text{and thus} \qquad |X_t-Y_t|^2 \;\leqslant\; e^{-2\lambda t}|X_0-Y_0|^2\,.$$

Taking the expectation and choosing (X_0,Y_0) according to an optimal coupling for the $\mathcal{W}_2$ distance with the Euclidian norm immediatly yields

$$\mathcal{W}_2(\nu P_t,\mu P_t) \;\leqslant\; e^{-\lambda t}\mathcal{W}_2(\nu,\mu)$$

for all initial conditions $\nu,\mu\in\mathcal{P}_2$.

For general considerations on Wasserstein distances, couplings and optimal transport, see [Vil09b].

1.5.2 *Functional inequalities*

Consider the example of the overdamped Langevin equation (1.4). Then the evolution of the relative entropy of the law at time t of the process with respect to the equilibrium μ_β is given by the Fischer Information :

$$\partial_t \int_{\mathbb{R}^d} h_t \ln(h_t)\mathrm{d}\mu_\beta \;=\; -\int_{\mathbb{R}^d}\frac{|\nabla h_t|^2}{h_t}\mathrm{d}\mu_\beta\,.$$

As a consequence, if we are able to prove an inequality between the entropy and its dissipation, which means here that μ_β satisfies a so-called log-Sobolev inequality

$$\forall h>0\,,\ \int_{\mathbb{R}^d} h = 1\,, \qquad \int_{\mathbb{R}^d} h\ln(h)\mathrm{d}\mu_\beta \;\leqslant\; \frac{1}{\rho_\beta}\int_{\mathbb{R}^d}\frac{|\nabla h|^2}{h}\mathrm{d}\mu_\beta$$

for some $\rho_\beta>0$, then we immediatly get an exponential decay of the relative entropy at rate ρ_β. Reciprocally, if

$$\int_{\mathbb{R}^d} h_t\ln(h_t)\mathrm{d}\mu_\beta \;\leqslant\; e^{-\rho_\beta t}\int_{\mathbb{R}^d} h_0\ln(h_0)\mathrm{d}\mu_\beta$$

for all $t\geqslant 0$ and all h_0, then the log-Sobolev is obtained by letting $t\to 0$, so that it is equivalent (for the overdamped Langevin process) to the exponential decay of the

entropy. We have thus replaced the dynamical question of estimating the convergence rate by a static question of proving a functional inequality for μ_β. The same arguments work with the L^2-norm (with the Poincaré inequality) or other entropies.

There are then many tools to establish such inequalities. Let us simply give an illustration. It is clear that if a probability distribution μ satisfies a log-Sobolev inequality with some constant $\rho > 0$ then any law $\mu_h = h\mu$ with a positive density h lower and upper bounded satisfies a log-Sobolev inequality with constant $\rho/(\|h\|_\infty \|1/h\|_\infty)$. On the other hand, through the Bakry-Emery calculus (more details in Section 2.2) is is known that if U is ρ-convex then $\mu \propto \exp(-U)$ satisfies a log-Sobolev inequality with constant $\rho > 0$. As a consequece, if U is the sum of a ρ-convex and of a bounded potentials, we immediatly get that the overdamped Langevin (1.4) has a convergence rate larger than $\rho\beta e^{-2\beta(\max U - \min U)}$. The exponential scaling in β is to be expected, as we saw in Section 1.3.1, although we don't get the exact factor of β with this rough argument. A refined analysis (still based on the previous perturbation argument) allows to obtain the correct exponential term, $e^{-\beta\Delta}$ with Δ the critical depth of the potential, and even the leading term of the sub-expoential prefactor (see [Mic92, HKS89, MS14]).

General references on functional inequalities are [ABC+00, BGL14].

Hypocoercivity

If we try to treat similarly the Langevin process (1.6), we see that

$$\partial_t \int_{\mathbb{R}^d} h_t \ln(h_t) \mathrm{d}\mu_\beta \;=\; -\int_{\mathbb{R}^d} \frac{|\nabla_y h_t|^2}{h_t} \mathrm{d}\mu_\beta \,,$$

due to the fact that there is only noise in the velocity variable. In particular, if h_0 only depends on x (but is not equal to 1) then the derivative of the entropy is 0 at time 0 so there is no way for an exponential decay to occur, i.e.

$$\int_{\mathbb{R}^d} h_t \ln(h_t) \mathrm{d}\mu_\beta \;\leqslant\; e^{-\rho t} \int_{\mathbb{R}^d} h_0 \ln(h_0) \mathrm{d}\mu_\beta$$

only holds with $\rho = 0$. Nevertheless an exponential decay in the sense that

$$\int_{\mathbb{R}^d} h_t \ln(h_t) \mathrm{d}\mu_\beta \;\leqslant\; Ce^{-\rho t} \int_{\mathbb{R}^d} h_0 \ln(h_0) \mathrm{d}\mu_\beta$$

for some $\rho > 0$ may hold for some $C > 1$. This is called an hypocoercive convergence toward equilibrium. Two main differences with the previous (coercive) case is that an hypocoercive convergence is not equivalent to a functional inequality (as letting $t \to 0$ in the inequality doesn't give anything), and that hypocoercivity is invariant if the entropy is replaced by an equivalent quantity. Indeed, a way to establish hypocoercivity is to work with another equivalent quantity in which the process is coercive. This is basically an extension of the case of linear ODE in finite dimension $\dot{x} = Ax$ with A that is not symmetric but whose eigenvalues all have negative part, so that a norm equivalent to the L^2-norm decays at constant rate.

Hypocoercivity will be more discussed in Section 2.2, see also [Vil09a, DMS15].

1.6 Interacting processes

Apart from lifted processes, another way to leave the Markovian realm (and possibly, from a stochastic algorithm viewpoint, accelerate long-time convergence) is to consider interacting processes. For instance, the process may interact with its own law (mean-field interaction) or with its past trajectory (self-interaction). Let us briefly discuss examples based on the overamped Langevin process.

1.6.1 *Mean-field interaction*

Consider a system of N particles $X = (X^1, \dots, X^N)$ on $\mathbb{R}^{dN}$ solving the SDE

$$\forall i \in [\![1, N]\!] \qquad \mathrm{d}X_t^i \;=\; -\nabla U(X_t^i)\mathrm{d}t - \frac{1}{N}\sum_{j=1}^{N} \nabla W(X_i - X_j)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i$$

where U and W (assumed symmetric) are respectively called the confining (or exterior) and interaction potentials, and $B = (B^1, \dots, B^N)$ is a Brownian motion on $\mathbb{R}^{dN}$. In other words, X is an overdamped Langevin process solving

$$\mathrm{d}X_t \;=\; -\nabla U_N(X)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad \text{with} \qquad U_N(x) \;=\; \sum_{i=1}^{N} U(x_i) + \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} W(x_i - x_j)\,.$$

Remark that

$$\frac{1}{N}\sum_{j=1}^{N} \nabla W(X_i - X_j) \;=\; \int_{\mathbb{R}^d} \nabla W(X_i - z)\pi_t^N(\mathrm{d}z) \qquad \text{with} \qquad \pi_t^N \;=\; \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}$$

is the empirical distribution of the system. If the X_i's were independent (which they are not, at least as soon as $t > 0$) and identically distributed, then by the Law of Large Number π_t^N would converge to their common law. In fact, as $N \to +\infty$, two given particles become more and more independent and a *propagation of chaos* phenomenon can be established, according to which $\pi_t^N \to m_t$ as $N \to +\infty$ where m_t is the law of the process $\bar{X}$ on $\mathbb{R}^d$ solving

$$\mathrm{d}\bar{X}_t \;=\; -\nabla U(\bar{X}_t)\mathrm{d}t - \int_{\mathbb{R}^d} \nabla W(\bar{X} - z)m_t(\mathrm{d}z) + \sqrt{2}\mathrm{d}B_t\,. \tag{1.10}$$

In other words, m_t solves the non-linear PDE

$$\partial_t m_t \;=\; \nabla \cdot (\nabla m_t + (\nabla U + \nabla W * m_t)m_t)$$

(where $*$ stands for the convolution operator), called the granular media equation.

Among other issues, there are two intertwinned questions: the behaviour as $N \to +\infty$, and as $t \to +\infty$. This is also related to mestability questions. Indeed, in cases where U has several local minima and $W(x) = x^2$ (which means particles are attracted one to the others), if all particles start close to the same local minimum, it is possible that the first time a large proportion of the system (say $N/2$ particles) is not in the basin of attraction of this initial minimum becomes arbitrarily large as $N \to +\infty$, which means that this event never happens in the limit $N = +\infty$. In other word, if m_0 is concentrated close

to a given local minimum, half of the mass will remain forever around this minimum. Now, the same situation may hold for another solution that starts in another well, which means that there is no chance for both solutions to converge to the same equilibrium. On the contrary, a finite system of N particles being a classical overdamped Langevin process, it is ergodic with respect to the Gibbs measure $\exp(-U_N)$, in particular, whatever the initial condition, the law of X_t^1 will converge in large time toward the first d-marginal of the Gibbs measure. This two facts also means that the convergence of the law of X_t^1 to the law of $\bar{X}_1$ cannot be uniform in time, in this case. Alternatively, there are cases where the two limits commutes, see [Mal01] and Sections 2.2.2 and 2.4.

1.6.2 *Self-interaction*

Considering as previously confining and interaction potentials U and W, one can consider a process solving

$$\mathrm{d}X_t \;=\; -\nabla U(X_t)\mathrm{d}t - \int_0^t K(t,s)\nabla W\left(X_t - X_s\right)\mathrm{d}s + \sqrt{2}\mathrm{d}B_t$$

for some weight K. In the following, consider the case $K(t,s) = 1/t$, so that the process is similar to (1.10) except that the law of the process m_t has been replaced by its occupation measure $\nu_t = 1/t \int_0^t \delta_{X_s}\mathrm{d}s$, which is a random probability measure. Due to the $1/t$ factor, for large times the evolution of the occupation measure is slow with respect to the motion of X_t and with the mixing time of the overdamped Langevin process with potential $U + W * \nu_t$ (with a fixed t). In other words, in some sense,

$$\nu_{t+T} \;=\; \frac{t}{t+T}\nu_t + \frac{T}{t+T}\times\frac{1}{T}\int_0^T \delta_{X_{t+s}}\mathrm{d}s \;\underset{t\gg T\gg 1}{\simeq}\; \left(1-\frac{T}{t}\right)\nu_t + \frac{T}{t}\frac{e^{-U+W*\nu_t}}{\int_{\mathbb{R}^d} e^{-U(x)+W*\nu_t(x)}\mathrm{d}x}\,,$$

or, after a time rescaling, ν_{e^t} approximately follows the deterministic flow on probability measures given by

$$\partial_t \mu_t \;=\; \frac{e^{-U+W*\mu_t}}{\int_{\mathbb{R}^d} e^{-U(x)+W*\mu_t(x)}\mathrm{d}x} - \mu_t\,.$$

This is the so-called ODE method. The objective is then, first, to make rigorous the link between ν_t and this deterministic flow and, second, to study the behaviour of the deterministic flow (does it have a unique global attractor ? Does it have periodic solutions ?) to deduce properties on ν_t.

An equilibrium of the deterministic flow is a probability measure μ that satisfies

$$\mu \;=\; \frac{e^{-U+W*\mu}}{\int_{\mathbb{R}^d} e^{-U(x)+W*\mu(x)}\mathrm{d}x}\,,$$

which exactly means that it is an equilibrium of the mean-field process (1.10).

For more details about self-interacting diffusions, see [BLR02].

Chapter 2

Organization of my work

Let us start with an informal general discussion to outline how my works more or less logically ensue one from the others, with the role of some particular encounters. All the articles mentioned in Section 2.1 are then detailed in the rest of the chapter.

2.1 Overview of a research trajectory

The title of my PhD thesis was *Hypocoercivity : alternative approaches and application to stochastic algorithms*. The main objective was to prove the convergence of the simulated annealing algorithm based on the (kinetic) Langevin process (see Section 2.2.2 below), and one of the idea my advisor Laurent Miclo wanted me to try in order to solve this question was to develop an approach to hypocoercivity that differs from Villani's method [Vil09a], namely : instead of considering a modified H^1 norm, one should keep the usual L^2 norm but, instead of differentiating it once along time, it should be differentiated three times. This is the method I implemented in [13] to prove a long-time exponential convergence for two kinetic models, the Langevin and run-and-tumble processes. It turned out the method and results were not so different from the H^1 framework (in fact the L^2 norm and its three derivatives alone cannot give a closed differential inequality, an H^1 term is added in an auxiliary steps and then disappears at the end). This work awakened my interest on the question of hypocoercivity methods, and the fact that arguments that were meant to be general, at the end, were always applied to the same benchmark (the Langevin diffusion) and failed in most other contexts (since then, the Dolbeault-Mouhot-Schmeiser approach [DMS15] turned out to be quite robust, but it was not so clear at that time, at least to me). In particular, I have to admit that, to some extent, rather than looking for solutions to relevant problems, for some time I had been looking for problems to my solutions, namely for benchmarks to test the limits of the domain of applications of hypocoercivity methods. This is how I ended up applying the H^1 hypocoercivity approach to some PDMP in [14]. A new difficulty from previous works, among others, was that the equilibria of the processes did not have an explicit expression, which compelled me to put a first finger in Bakry-Emery calculus and gradient/semigroup sub-commutations. The work [20] (shortly after my PhD) then naturally followed (still in this "problems for my solutions" posture), where I gave systematic adaptations of classical Bakry-Emery arguments to typical hypocoercivity cases. At the end of this work, I had a clearer view on hypocoercivity computations, and in particular on the explicit estimates one could get. This directly lead to three other works

(again, shortly after the PhD) : first, using explicit convergence rates for the Langevin diffusion in term of the inverse temperature β, I was able (finally) to solve the initial question of my PhD, that is the convergence of the Langevin simulated annealing algorithms [19]. Second, ("problems for my solutions", again) using explicit convergence rates for the Langevin diffusion in term of the number of particles, I studied the Vlasov-Fokker-Planck equation and the associated system of interacting particles in the convex potential/convex interaction case [17] (this was my first contact with the propagation of chaos phenomenon). Third, since I had obtained in [20] that the convergence rate for Ornstein-Uhlenbeck (OU) processes is completely determined by the spectral gap of the drift matrix, after a discussion with Arnaud Guillin we decided to solve the question of finding the optimal OU process (possibly hypoelliptic) in term of convergence rate for a given target gaussian equilibrium [8].

In the meanwhile, in parallel to this direction that followed the work [20] about hypocoercivity, another axe had emerged in my research from my first works [12, 13, 14], namely the use of PDMPs for sampling target measures. This led to [15], which contains two parts. The first part is an elementary study of the one-dimensional run-and-tumble process at low-temperature and in an annealing version. The second part is concerned with multi-dimensional target measures, in particular a velocity jump sampler for arbitrary target in any dimension is introduced and studied. I was later informed that in fact a similar process had already been introduced in the statistical physics community [PdW12]. Later, independently, another multi-dimensional velocity jump sampler was proposed in [BFR19]. The link between these three works has later been made by the computational statistics community [BVD18, VBDD17], in which these piecewise deterministic samplers have since then gained a high interest (and initiated a full range of works). Thus, by chance, I happened to be already working on this topic when it became somehow trendy.

Around that time, I left Toulouse and went for a post-doc position at Neuchâtel in the team of Michel Benaïm. One of my objective was to learn the techniques of Michel for studying self-interacting processes. Indeed, during my PhD, I had been deeply impressed by a talk of Tony Lelièvre about adaptive algorithms, based on interacting processes [JLR10, LRS08]. While I was at that time realising that, at low temperature, all Markov processes (the kinetic Langevin diffusion, the velocity jump processes, etc.) were essentially doomed to have the same exponentially bad rate and thus the same condition for convergence of the simulated annealing processes (see the informal argument in Section 1.3.1), the adaptive algorithms were successfully reducing the critical depth of the potential. I had the impression that with these algorithms, the metastability issue was definitively solved (several years later, now that I have worked for some times on these algorithms and in contact with real applications in molecular dynamics, I can safely say that this impression is long gone). As a first training on the ODE method, I studied the self-interacting Adaptive Biasing Force (ABF) algorithm, but written with velocity jump processes [24]. With Carl-Eric Gauthier, who was a student of Michel at that time, we also performed an elementary study of a toy model for metadynamics (another adaptive algorithm), which is a stongly self-interacting process [7] with some similarities with the simulated annealing algorithm. After a year, I left Neuchâtel and, quite naturally, went to work as a post-doc with Virginie Ehrlacher and Tony at École des Ponts Paristech. A specific subject was attached to the post-doc position, based on a idea of Tony and Virginie : the aim was to combine the ABF algorithm with tensor ap-

proximation schemes to allow for a (moderately) large number of reaction coordinates. Based on my training in Neuchâtel, I quickly tackled the theoretical questions and then, taking advantage of Tony's and Virginie's patience, I took ages to finish the numerical part of the project [5]. In the meanwhile, as a natural continuation of our respective previous works, with Charles-Edouard Bréhier and Michel, we started to work on the convergence of the self-interacting ABF algorithm but in the case where the occupation measure is not reweighted (since this is the practical case, while all previous works considered for theoretical simplicity the reweighted case), which gave [1].

These works on adaptive algorithms and a year in the CERMICS team had brought me closer to the molecular dynamics realm, precisely the year where a math/chemistry position opened in Université Pierre et Marie Curie (soon to be renamed Sorbonne Université). It was a challenge for me to go in that direction – as I was still mostly concerned with the theoretical study of stochastic algorithms – but the math/chemistry interactions at UPMC were already vivid (in particular around Jean-Philip Piquemal and Yvon Maday) and my integration has been smooth. Notably, during my year at CERMICS, I had become convinced that programming was 1) subtle, 2) crucial and 3) not for me but rather for more qualified persons. I was clear on that topic during the audition for the position at UPMC, and I was answered that the team was lacking maybe theoretical skills in stochastic algorithms, but certainly not programming skills, and indeed since then I have had the pleasure to work with very able people in that domain (in particular Louis Lagardère). When I arrived in the team, Jean-Philip and Louis were working (among a full range of other things) on multi-time step methods. After a few discussion on that topic, I realised that velocity jump processes could be a non-biased alternative to these schemes. This lead to [22, 23] with Jean-Philip, Louis and Jeremy Weisman. It was my first experience on designing an original algorithm that is then implemented in a real applied software and indeed gives competitive results. This is not a particularly unpleasant experience.

In the meanwhile, a new direction had branched from my work on velocity jump processes, about couplings. I first met coupling methods in [17] for propagation of chaos, but it went on another level when, with Alain Durmus and Arnaud Guillin, we decided to extend my arguments of [15] for the ergodicity of the Bouncy Particle Sampler to the non-compact space case [4] (in fact we took too much time and another group released a paper on the same topic before us [DBD19]. This is the drawback of working on trendy subjects...). In this work, contrary to my initial work [15] and the similar [DBD19], instead of proving a Doeblin condition based on controllability considerations, we constructed an explicit coupling of refreshed velocity jump processes. Moreover, in the course of the study, we had to prove a few technical results on PDMPs, that we ended up gathering in a companion paper [3]. In the latter, in particular for some smooth approximation arguments, developing an idea of [24], we used synchronous couplings of jump mechanisms to obtain bounds on transition kernels of PDMPs with different jump rate and kernel. Synchronous coupling for jump mechanisms is somehow an analogous of parallel coupling for diffusions, and I realised it could be used for systems of interacting jump processes (where interaction plays at the level of the jump mechanism). This lead to a first general work [18], followed later by a work specific to the Flemming-Viot algorithm [10] with Lucas Journel during his master thesis, and more recently by a work with Eva Löcherbach on systems of interacting neurons [11]. The latter brought me back to questions of metastability, and we gave a general result on exponentiality of escape

times for Markov processes (extending previous works restricted to diffusions).

Despite these new directions, five years after the end of my Phd thesis, hypocoercivity remains a topic in my research. In the continuation of my work [20] on Γ calculus and my first work with Arnaud Guillin [8], we started with Patrick Cattiaux, Chaoen Zhang (at that time a PhD student of Arnaud) and Arnaud to work on the problem of establishing hypocoercivity estimates in the entropy sense when the Hessian of the potential is not bounded. This lead us to a generalisation of the multipliers method of Villani [Vil09a], working with a weighted modified Fischer Information [2]. Later, Arnaud and Chaoen with Wei Liu and Liming Wu established in [GLWZ19] log-Sobolev inequalities uniform in N for the McKean-Vlasov process, from which with Arnaud we treated the kinetic case [9], extending the arguments and results of my previous work [17] to possibly non-convex cases. While working on [2] and [4] I had a close and prolonged contact with (probabilistic) Lyapunov arguments, which led me to study the relation between Lyapunov arguments and hypocoercivity methods [21], extending the classical works of Patrick, Arnaud and co-authors in the reversible case [CGZ13, CG14, BCG08].

Finally, let me mention the recent work [6]. Nicolas Fournier and Camille Tardif, who had been interested for some time already on processes with heavy tails, got interested in the study of the simulated annealing under minimal growth assumption of the potential. They came to discuss with me in my position of "expert in simulated annealing". We established that, for potentials of the form $U(x) = f(|x|)$ with an increasing f, then a transition occurs for $f(t) = c \ln \ln t$ for some c, and that more generally one can essentially find potentials with an arbitrarily slow growth and with an unbounded set of local minima for which the simulated annealing converges.

2.2 Hypocoercivity

2.2.1 *General results*

In [20] I adapted many classical arguments of Bakry-Emery Γ calculus to more general (possibly hypoelliptic, hypocoercive, non-local) contexts. This direction had already been pioneered by Baudoin in [Bau17], centered on the study of the Langevin process. One of the main object of the classical Γ-calculus is the carré-du-champ operator Γ given by $\Gamma(f) = 1/2Lf^2 - fLf$ where L is the generator of a Markov semi-group $(P_t)_{t\geqslant 0}$. This quantity naturally appears when one considers the interpolation between $P_t f^2$ and $(P_t f)^2$, given by $s \mapsto P_s(P_{t-s}f)^2$ for $s \in [0,t]$. This first computation leads to the question of interpolating between $P_t(\Gamma f)$ and $\Gamma(P_t f)$, which defines similarly the Γ_2 operator. The computations are exactly the same as

$$\partial_t \int_{\mathbb{R}^d} (P_t f)^2 \mathrm{d}\mu = -\int_{\mathbb{R}^d} \Gamma(P_t f)\mathrm{d}\mu\,, \qquad \partial_t \int_{\mathbb{R}^d} \Gamma(P_t f)\mathrm{d}\mu = -\int_{\mathbb{R}^d} \Gamma_2(P_t f)\mathrm{d}\mu\,,$$

simply integrating with respect to an invariant measure μ. In particular the same problems as mentioned in Section 1.5.2 arise when considering for instance the Langevin process, for which $\Gamma(f) = |\nabla_y f|^2$, so that most classical arguments fail. Following the integrated computations of Villani [Vil09a], the idea of Baudoin was to add to Γ another operator Γ^Z. In [20] however I completely set aside Γ and just considered any operator. In other words, I considered the general question of interpolating quantities of the form $\Phi(P_t f)$ and $P_t(\Phi f)$ with operators Φ that have possibly nothing to do with the dynamics

of the process. This was in particular motivated by my previous work [14] where I established estimates of the form $|\nabla P_t f|^2 \leqslant e^{-\rho t} P_t |\nabla f|^2$ for PDMPs whose non-local carré du champ had nothing to do with the gradient. It turned out to be important also to obtain the optimal convergence rate for Ornstein-Uhlenbeck processes, based on a estimate of the form $|R\nabla P_t f|^2 \leqslant e^{-\rho t} P_t |R\nabla f|^2$ where the matrix R only depends on the drift matrix of the process, and not at all on the diffusion part (that gives the carré-du-champ).

So take any operator Φ and consider for some f the interpolation

$$\psi(s) \;=\; P_s\left(\Phi(P_{t-s}f)\right), \qquad s \in [0,t].$$

Then we see that

$$\psi'(s) \;=\; P_s\left(L\Phi - (D\Phi)L\right)(P_{t-s}f) \;:=\; 2P_s\Gamma_\Phi(P_{t-s}f)$$

has the same structure as ψ. From this remark, I was able to extend many classical arguments, like gradient/semigroup subcommutation, proof of functional inequalities, longtime convergence, small time regularization. The two main general results concern diffusion processes :

Theorem 1 (from Theorems 9 and 10 of [20]). Consider on $\mathbb{R}^d$ a diffusion operator in Hörmander form $L = B_0 + \sum_{i=1}^d B_i^2$ where the B_j's are smooth vector fields. Suppose there exist $N_c \in \mathbb{N}$ and $\lambda, \Lambda, m > 0$ such that for $i \in [\![0, N_c+1]\!]$ there exist derivation operators C_i and R_i and a matrix field Z_i satisfying:

1. $C_{N_c+1} = 0$, and $[B_0, C_i] = Z_{i+1}C_{i+1} + R_{i+1}$ for all $i \in [\![0, N_c]\!]$.
2. $[B_j, C_i] = 0$ for all $i \in [\![0, N_c]\!]$, $j \in [\![1, d]\!]$,
3. $\lambda \leq \frac{Z_i + Z_i^T}{2} \leq \Lambda$ for all $i \in [\![0, N_c]\!]$,
4. $C_0^T C_0 \leq m \sum_{j\geq 1} B_j^T B_j$ and $R_i^T R_i \leq m \sum_{j<i} C_j^T C_j$ for all $i \in [\![0, N_c+1]\!]$.

Then there exists $c > 0$ such that for all $f > 0$, $x \in \mathbb{R}^d$, $t > 0$ and $i \in [\![0, N_c]\!]$

$$\frac{|C_i \nabla P_t f|^2(x)}{P_t f(x)} \;\leq\; c(1-e^{-t})^{-2i-1}\left(P_t(f\ln f)(x) - P_t f(x)\ln\left(P_t f(x)\right)\right) \tag{2.1}$$

If, moreover, there exist $\rho, K > 0$ and a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ such that

1. $\sum_{i\geq 0} C_i^T C_i \geq \rho$,
2. μ satisfies a log-Sobolev inequality with constant K,

then there exist $\kappa > 0$ such that for all $t > 0$ and for all $f > 0$ with $\int f \mathrm{d}\mu = 1$,

$$\int_{\mathbb{R}^d} P_t f \ln(P_t f)\mathrm{d}\mu \;\leq\; e^{-\kappa t(1-e^{-t})^{2N_c}} \int_{\mathbb{R}^d} f\ln f \mathrm{d}\mu. \tag{2.2}$$

This theorem is similar to previous results of Villani [Vil09a, Theorems 28 and A.15], but a crucial difference is that there is no need to compute adjoint of operators in $L^2(\mu)$,

which would require to have an explicit expression of μ, which is not necessary in my case. Moreover, an explicit expression is given for c and κ in (2.1) and (2.2) :

$$\begin{aligned} c &= \left(\frac{100}{\lambda}\left(N_c^2+\frac{\Lambda^2}{\lambda}+m\right)\right)^{20N_c^2} \\ \kappa &= \frac{\rho}{cK}. \end{aligned} \tag{2.3}$$

This is very rough but in many cases N_c, Λ, λ, m and ρ are just of order 1 in a parameter of interest of the problem so that only the constant of the log-Sobolev inequality matters, exactly as in the reversible case.

After this general study, the rest of [20] is dedicated to two particular processes: first, Ornstein-Uhlenbeck processes (1.7), for which I obtained the optimal long-time rate of convergence (even when the drift matrix admits defective eigenvalues) and the short-time decay of the entropy (as t^{2M+1} where M is the number of Hörmander brackets needed to span the whole space), thus extending the previous results of [AE14]. Second, systems of interacting particles

$$\forall i \in G, \quad \mathrm{d}X_i(t) = -\sum_{i\sim j}\nabla W_{i,j}\left(X_i(t)-X_j(t)\right)\mathrm{d}t + \sum_{j=1}^{\infty}\sigma_{i,j}\mathrm{d}B_j(t)$$

where G is a finite graph of interactions and for all edge (i,j), $W_{i,j}$ is a convex potential, for which I established that the convergence rate could be bounded by the spectral gap of a simple Markov chain on G. In the particular case where G is $[\![1,N]\!]$ with the nearest neighbour relation, it gives a convergence of the chain at a rate of order N^{-2}.

More recently, the note [21] also tackles a general question about hypocoercivity, which is the link between the Dolbeault-Mouhot-Schmeiser (DMS) approach to L^2 hypocoercivity [DMS15] and the existence of a Lyapunov function, namely of a positive function $\mathcal{V}$ that goes to infinity at infinity and such that

$$\mathcal{L}\mathcal{V} \leqslant -\rho\mathcal{V}+C$$

for some $\rho, C > 0$. The existence of such a Lyapunov function classically implies that the hitting time of the compact set $\{\mathcal{V} \leqslant 2C/\rho\}$ has an exponential moment. Conversely, the exponential integrability of hitting times implies the existence of a Lyapunov function, at least under some regularity assumptions or in a weaker sense [BCG08].

The DMS approach relies on a modified norm

$$\mathbf{H}(f) = \int_{R^d}\left(f^2+fAf\right)\mathrm{d}\mu$$

where A is a bounded operator on $L^2(\mu)$ (with $\|A\| \leqslant 1/2$ so that $\mathbf{H}$ is equivalent to the squared L^2 norm) defined from the symmetric and antisymmetric parts of L. Then in [DMS15] are established some conditions under which

$$\partial_t\mathbf{H}(P_tf) \leqslant -\rho\mathbf{H}(P_tf)$$

for some $\rho > 0$. This method (with a systematic definition of $\mathbf{H}$) proved to work

for many degenerated processes : Langevin process and linear Boltzmann equation [DMS15], velocity jump processes [ADNR18], hypoelliptic diffusions related to strongly self-interacting processes [BG17], and so on. On the contrary, the construction of a Lyapunov function usually requires to have some understanding of the particular process in study. The link between (coercive) L^2 convergence and the existence of Lyapunov function had been studied in a series of paper or Cattiaux, Guillin and co-authors [CGZ13, CG14, BCG08] in the reversible case. I showed that an exponential contraction in a modified L^2-norm (as what is provided by the DMS method) is in fact sufficient to have the exponential integrability of hitting times of compact sets, and thus the existence of a Lyapunov function.

2.2.2 *Variations on the Langevin diffusion*

Let us now detail the works [17, 19, 2, 9], which are all concerned with the Langevin process (1.6) (so is [13] but I don't detail it here as it was already in the PdD thesis).

First, [19] is concerned with the study of the kinetic simulated annealing process

$$\left\{ \begin{array}{rcl} \mathrm{d}X_t &=& V_t \mathrm{d}t \\ \mathrm{d}V_t &=& -\nabla U(X_t)\mathrm{d}t - V_t \mathrm{d}t + \sqrt{2\beta_t^{-1}}\mathrm{d}B_t \end{array} \right.$$

for some positive decreasing cooling schedule $t \mapsto \beta_t^{-1}$. Let E_* be the critical depth of U, namely

$$E_* = \sup_{x \in m_U} \inf_{y \in M_U} \inf \left\{ \max_{s\in[0,1]} U(\gamma(s)) - U(x)\,, \gamma \in \mathcal{C}^1([0,1], \mathbb{R}^d),\ \gamma(0) = x,\ \gamma(1) = y \right\}$$

where m_U and M_U are respectively the sets of local and global minima of U. The main result of [19] is the following.

Theorem 2. Assume U is a Morse function with a finite number of critical points, quadratic at infinity with bounded Hessian, and $\partial_t \beta_t \leqslant 1/(Et)$ with $E > E_*$. Then for all $\delta > 0$,

$$\mathbb{P}\left(U(X_t) \leqslant \min U + \delta\right) \underset{t\to+\infty}{\longrightarrow} 1\,.$$

This is not an surprising result, but it was yet to be established. It is essentially based on an adaptation of the argument of Miclo [Mic92] in the overdamped Langevin case, and requires an explicit estimates of the convergence rate of the entropy for the Langevin process at fixed but low temperature $\beta^{-1} > 0$, which is easily obtained by the explicit estimate (2.3) and the results on log-Sobolev inequalities for reversible diffusions like [Mic92, MS14]. The Morse assumption is only made for simplicity.

Similarly, [17, 9] are concerned with mean-field interacting Langevin diffusions,

$$\forall i \in [\![1, N]\!]\,, \qquad \left\{ \begin{array}{rcl} \mathrm{d}X_t^i &=& V_t^i \mathrm{d}t \\ \mathrm{d}V_t^i &=& -\nabla U(X_t^i)\mathrm{d}t - \frac{1}{N}\sum_{j=1}^N \nabla_x W(X_t^i, X_t^j) - V_t^i \mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i\,. \end{array} \right.$$

In fact, $X = (X^1, \dots, X^N)$ and $V = (V^1, \dots, V^N)$ form a standard Langevin diffusion on

$\mathbb{R}^{dN}$, since

$$\left\{ \begin{array}{rcl} \mathrm{d}X_t &=& V_t \mathrm{d}t \\ \mathrm{d}V_t &=& -\nabla U_N(X_t^i)\mathrm{d}t - V_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i \end{array} \right.$$

with

$$U_N(x) \;=\; \sum_{i=1}^N U(x_i) + \frac{1}{2N}\sum_{i,j=1}^N W(x_i,x_j)\,.$$

Assuming that U and W have bounded Hessians, in the explicit estimate (2.3) applied to (X,V), only the log-Sobolev constant of the equilibrium may depend on N. In [17], I used that this log-Sobolev constant is uniform in N when U is convex and W is not too concave (so that U_N is ρ-convex for some ρ uniformly in N). In [9] with Arnaud Guillin we used the result of [GLWZ19] where a log-Sobolev inequality with a constant uniform in N is proven for the Gibbs measure $\exp(-U_N)$ under some assumptions that allows non-convex confining potential (in particular at high temperature). In both cases, this gives convergence rate for (X,V) that is uniform in N.

From this, we obtain long-time convergence for the Vlasov-Fokker-Planck equation

$$\partial_t m_t + v\cdot\nabla_x m_t \;= \nabla_v\cdot\left(\nabla_v m_t + \left(\int \nabla_x U\,(x,u)\,m_t(u,w)\mathrm{d}w\mathrm{d}w + v\right) m_t\right)$$

which is the mean-field limit as $N\to+\infty$ of the law of X_t^1. The long-time convergence (uniform in N) also allows to establish uniform in time propagation of chaos : denoting $m_t^{(n,N)}$ the law of $(X^1,\dots,X^n)$ for $n\in[\![1,N]\!]$, we obtain under the previous assumptions on U and W estimates of the form

$$\begin{array}{rcl} \mathcal{W}_2\left(m_t^{(n,N)}, m_t^{\otimes n}\right) &\leq& \dfrac{K\sqrt{n}}{N^\kappa} \\ \|m_t^{(n,N)} - m_t^{\otimes n}\|_{TV} &\leqslant& \dfrac{K\sqrt{n}}{N^\kappa}\,. \end{array}$$

for some $K,\kappa>0$ that do not depend on t. We also consider the empirical distribution $\pi_t^N = 1/N\sum_{i=1}^N \delta_{X_i,V_i}$ and prove that

$$\mathbb{E}\left(\mathcal{W}_2^2\left(\pi_t^N, m_\infty\right)\right) \;\leqslant\; K\left(e^{-\chi t} + a(N)\right)$$

for some $K,\chi>0$ where

$$a(N) \;=\; \left\{ \begin{array}{ll} N^{-1/2} & \text{if } d=1 \\ \ln(1+N)N^{-1/2} & \text{if } d=2 \\ N^{-2/d} & \text{if } d\geqslant 3\,. \end{array} \right.$$

This is optimal even for non-interacting reversible diffusions. This last estimate is relevant from a numerical point of view where the equilibrium of the Vlasov-Fokker-Planck is approximated by the empirical distribution of a system of particles at a large t.

In [17, 19, 9], the Hessian of the potential is assumed to be bounded. Indeed, at that time the existing results of long-time convergence in the relative sense (either mine in

[20] or Villani's in [Vil09a]) required this assumption. This is different from the L^2 case, where a condition of the form $|\nabla^2 U| \leqslant C(1+|\nabla U|^2)$ is sufficient. With Patrick Cattiaux, Arnaud Guillin and Chaoen Zhang, we relaxed this assumption in [2], obtaining long-time hypocoercive convergence in relative entropy for the Langevin diffusion (1.6) in particular when U behaves at infinity like $|x|^\alpha$ for some $\alpha \geqslant 2$. This is based on an extension of the multipliers method of Villani [Vil09a] in L^2, namely we use a modified entropy of the form

$$\int_{\mathbb{R}^d} h \ln h \mathrm{d}\mu + \int_{\mathbb{R}^d} \frac{|R\nabla h|^2}{h} \mathrm{d}\mu$$

where $R = R(t,x,v)$ is a non-constant matrix. The log-Sobolev inequality has to be replaced by a weighted inequality of the form

$$\int_{\mathbb{R}^d} h \ln h \mathrm{d}\mu \leqslant C \int_{\mathbb{R}^d} \frac{|\tilde{R}\nabla h|^2}{h} \mathrm{d}\mu$$

for some C with some non-constant $\tilde{R}(x,v)$. Thanks to Lyapunov arguments, we establish simple criteria on U to ensure the validity of such inequalities.

2.2.3 *Optimal Gaussian diffusions*

For a given symmetric positive definite matrix S on $\mathbb{R}^d$, consider the Gaussian measure with density

$$\mu_S(x) = \frac{(\det S)^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left(\frac{-x^T S x}{2}\right)$$

and, for two matrices A and D on $\mathbb{R}^d$ with D symmetric positive semi-definite, the Markov generator $L_{A,D}$ given by

$$L_{A,D} f(x) := (Ax)^T \nabla f(x) + \nabla \cdot (D \nabla f)(x).$$

Then μ_S is invariant by $L_{A,D}$ if, by definition, $L'_{A,D}\mu_S = 0$ where $L'_{A,D}$ is the adjoint of $L_{A,D}$ in $L^2(\mathrm{d}x)$. Besides, from my results on Ornstein-Uhlenbeck processes in [20] the convergence rate of $L_{A,D}$ is the spectral gap of A. Thus, from an MCMC point of view, in order to sample according to μ_S in the more efficient way, the question is to find among all matrices A and D such that $L'_{A,D}\mu_S = 0$ the maximal $\rho(A) = \inf\{-\Re(\lambda),\ \lambda \in \sigma(A)\}$, where σ stands for the spectrum. First, remark that this is a toy problem in the sense that Gaussian distributions are much simpler to sample than most of the distributions met in MCMC problems (although this may give some indication to sample a target measure $\propto \exp(-U)$ close to a local minimum of U). Second, note that the problem is ill-posed since, if μ_S is invariant for $L_{A,D}$ then it is invariant for $\lambda L_{A,D}$ for any $\lambda > 0$ and taking λ arbitrarily large yields arbitrarily large convergence rates. This corresponds to an artificial change of time, an it is not interesting in practice. Following [GM13], with Arnaud Guillin [8] we thus added the constraint that the trace of D is bounded by the trace of the identity of $\mathbb{R}^d$, which means that the amount of randomness injected in the system at all time corresponds to a standard Brownian motion. The optimisation

problem is then set on

$$\mathcal{I}(S) := \left\{ (A,D) \in \mathcal{M}_{d\times d}(\mathbb{R}) \times \mathcal{S}_d^{\geqslant 0},\ \mathrm{Tr}D \leqslant d,\ L'_{A,D}\mu_S = 0 \right\}.$$

Remark that the classical choice of the overdamped Langevin process (1.4) corresponds to $D = I_d$ and $A = -S$, which yields a convergence rate of $\rho(-S) = \min\sigma(S)$.

Our main result in [8] is that

$$\max\left\{\rho(A),\ (A,D) \in \mathcal{I}(S)\right\} \quad = \quad \max\sigma(S).$$

The optimisation problem had previously been considered in [HHMS93, LNP13] in the case where $D = I_d$, and it was proven that

$$\max\left\{\rho(A),\ A \text{ s.t. } (A,I_d) \in \mathcal{I}(S)\right\} \quad = \quad \frac{\mathrm{Tr}S}{d},$$

which is the arithmetic mean of all eigenvalues of S. On the other hand, for $(A,D) \in \mathcal{I}(S)$, the process is reversible if and only if $A = -(2DS + A)$, namely $A = -DS$, and for $D = \frac{d}{\mathrm{Tr}S^{-1}}S^{-1}$ this gives

$$\max\left\{\rho(A),\ (A,D) \in \mathcal{I}(S),\ L^*_{A,D} = L_{A,D}\right\} \quad = \quad \frac{d}{\mathrm{Tr}S^{-1}},$$

which is the harmonic mean of the eigenvalues of S. Note that

$$\min\sigma(S) \quad \leqslant \quad \frac{d}{\mathrm{Tr}S^{-1}} \quad \leqslant \quad \frac{\mathrm{Tr}S}{d} \quad \leqslant \quad \max\sigma(S)$$

and that the equalities hold only when S is a homogeneous dilation, in which case no non-reversible dynamics can yield any improvement of the rate of convergence to equilibrium.

The optimal rare over $\mathcal{I}(S)$ is obtained with a very degenerate diffusion. Indeed, only one coordinate receives all the noise. As can be seen for instance in Theorem 1, for an hypoelliptic diffusion, the entropy initially decays as t^{2N_c+1} where N_c is the number of Hörmander Bracket necessary to span the whole space, which for Ornstein-Uhlenbeck processes corresponds to the smallest N such that $\sum_{i=0}^{N} A^k D (A^T)^k$ is definite positive. Hence, when the rank of D is one, $N_c = d$. In particular it is clear that, for small times, an elliptic diffusion beats a non-elliptic one, in term of entropy decay. This means that, although a very hypoelliptic (A,D) gives asymptotically the best convergence rate, it is still possible that, due to a very high prefactor (geometric in d) in front of the exponential decay, the hypoelliptic diffusion only becomes better than the basic overdamped Langevin diffusion after a very long time. Moreover, another issue is that the optimal A may have a very high norm so that an Euler scheme for the hypoelliptic diffusion would require a very small timestep, which would impair the numerical efficiency. To tackle these concerns, we established a result at finite time:

Theorem 3 (From Theorem 2.2 of [8]). It is possible to construct $(A,D) \in \mathcal{I}(S)$ with $\|A\|_F \leq 4d^2\sqrt{\frac{(\max\sigma(S))^3}{\min\sigma(S)}}$ (where $\|A\|_F = \sqrt{\mathrm{Tr}\left(A^TA\right)}$ is the Frobenius norm) such that

for all $h > 0$, with finite entropy, denoting for all $t \geqslant t_0$

$$h_t \;=\; e^{(t-t_0)L^*_{A,D}} e^{t_0 L_{-S,I_N}} h\,,$$

then

$$\mathrm{Ent}_{\mu_S}\left(h_t\right) \;\leqslant\; \frac{\max\sigma(S)}{t_0\left(\min\sigma(S)\right)^2} e^{-(\max\sigma(S))(t-t_0)} \mathrm{Ent}_{\mu_S}\left(h\right).$$

In other words, in practice the small-time regularisation is not a real issue as it is sufficient to run a basic reversible overdamped Langevin for some time t_0 and then switch on the optimal hypoelliptic diffusion for the rest of the simulation.

2.2.4 *Non-local operators*

Let me briefly mention two of my works concerned with hypocoercivity for jump processes, [14, 16]. First, H^1 hypocoercivity for a class of PDMP is established in [14] (already presented in my PhD thesis). Second, [16] establishes the hypoercive decay in entropy for the linear Boltzmann equation when the potential is a small perturbation of an harmonic one. It is largely inspired by the work [Eva17] of Evans that tackles the same problem but in the periodic torus with no potential. The difference is thus the addition of a (close to) linear drift, as I had already studied in the case of Gaussian diffusions. The main difficulty (solved by Evans) is that a non-local generator does not satisfy the chain rule, which doesn't change anything for L^2/H^1 computations but is crucial when considering relative entropy/Fischer informations, due to the non-linear character of these quantities.

2.2.5 *Some on-going projects and perspectives*

I am currently interested in the following questions :

- Establish functional inequalities (log-Sobolev or Poincaré) for the invariant measures of non-equilibrium (possibly hypoelliptic non-elliptic) diffusions, which are not explicit. In some cases this can be done via coupling or generalized Γ calculus arguments. Long-time convergence will thus follow from Theorem 1.

- Establish long-time convergence in relative entropy for singular potentials for the Langevin diffusion, through weighted inequalities, in the spirit of [BGH19] or [2]. Related to this, study the spectral gap of Coulomb gas [BCF18] as a function of the number of particles in dimension larger than 1.

- Hypocoercivity and Γ calculus (in particular in entropy) with non-local operators (for instance Levy noise).

2.3 Velocity jump processes

My works [12, 13, 15, 24, 3, 4, 22, 23] all revolve around velocity jump processes for sampling. During my master thesis with Laurent Miclo, in [12], I studied the continuous-time limit of the persistent walk of [DHN00], that is the Markov process on $\mathbb{T} \times \{-1, 1\}$

with generator

$$Lf(x,v) = v\partial_x f(x,v) + a\left(f(x,-v) - f(x,v)\right) \tag{2.4}$$

for some $a > 0$. Depending on the community, nowaday such a process can be called a BGK, integrated telegraph, run-and-tumble, Zig-Zag or bouncy particle process. In [13, 15] (presented in my PhD thesis) I studied the long-time convergence and metastability of the extension of this process, still in dimension 1 (with velocities ± 1), designed to target any distribution $\mu(x,v) \propto \exp(-U(x))$ on $\mathbb{R} \times \{-1,1\}$. In [15] I also defined the velocity jump process on $\mathbb{R}^d \times \mathbb{S}^{d-1}$ for $d \in \mathbb{N}_*$ with generator

$$Lf(x,v) = \tag{2.5}$$
$$v \cdot \nabla_x f(x,v) + \lambda(x,v)\left(f(x,R(x,v)) - f(x,v)\right) + a\left(\int_{\mathbb{S}^{d-1}} f(x,w)\mathrm{d}w - f(x,v)\right),$$

where $a > 0$ is a parameter (called the refreshment rate), $\mathrm{d}w$ denotes the uniform measure on $\mathbb{S}^{d-1}$ and

$$\lambda(x,v) = (v \cdot \nabla_x U(x))_+ , \qquad R(x,v) = v - 2\left(v \cdot \frac{\nabla_x U(x)}{|\nabla_x U(x)|}\right)\frac{\nabla_x U(x)}{|\nabla_x U(x)|}.$$

The process is designed so that $\exp(-U(x))\mathrm{d}x\mathrm{d}v$ on $\mathbb{R}^d \times \mathbb{S}^{d-1}$ is left invariant. In fact, as already mentioned above, a similar process (except that velocities where in $\mathbb{R}^d$, refreshed according to a Gaussian distribution) had been introduced in [PdW12], that also contained a formal proof that the Gibbs measure $\propto \exp(-H(x,v))$ with $H(x,v) = U(x) + |v|^2/2$ is invariant. Since [BVD18] this process is known as the the bouncy particle sampler in the MCMC community. Interesting features of velocity jump processes is that they have a ballistic behaviour and that they can be sampled exactly through a thinning algorithm (thus targeting the exact distribution without discretization bias, without any accept/reject step).

In [15], through controllability arguments, I proved the geometric ergodicity in the case of a position in the periodic torus and velocities on the unit sphere, and thus it was natural to expend this result in a non-compact state space. To do so, it remained to construct a Lyapunov function for the generator. This is what we did with Alain Durmus and Arnaud Guillin in [4] where we established geometric ergodicity under some conditions on U. These conditions are a bit technical in their general form but, for instance, they hold if U behaves at infinity like $|x|^\alpha$, $\alpha \geqslant 1$. During this work, several technical questions arised, in particular: is it sufficient to check that $\int_{\mathbb{R}^d \times \mathbb{R}^d} Lf\mathrm{d}\mu = 0$ for all smooth and compactly supported test function f to conclude that μ is invariant for L ? This is true if the set of $\mathcal{C}^\infty$ functions with compact support is a core for L. It would be the case if $P_t f$ is smooth with compact support for all $t \geqslant 0$ if this is the case of f, however this is false. Thus we had to introduce an approximation step where the process is approximated through a truncation/smoothing step so that compactly-supported smooth functions are fixed for all times by the approximate semi-group. This required to construct rigorously the synchronous coupling of two PDMPs, and other results concerning non-explosion and error bounds on the invariant measures of different PDMPs. All these general considerations on PDMPs have been gathered in [3].

In fact I had already used the synchronous coupling of PDMPs in [24] to prove the convergence of the ABF algorithm (see Section 2.6 below) based on velocity jump processes similar to the bouncy particle sampler. The coupling played a similar role as, in the diffusion case, regularization results (which do not hold in the PDMP case) necessary for the proof.

Finally, [22, 23] are two companion papers, a theoretic one and an applied one. More precisely, there are two rather distinct parts in [22], the link being that there are both concerned with discrete-time kinetic chains used for sampling, and thus share some common features like thinning, factorization and continuous-time limits. The first part is concerned with a multi-dimensional generalization of the persistent walk to sample any target distribution $\propto \exp(-U)$ on $\mathbb{Z}^d$ (exactly as (2.5) is the generalization of (2.4), which is the continuous-time limit of the persistent walk). Through an elementary study, in dimension 1, the Law of Large Number, Central Limit Theorem, metastability (Eyring-Kramers formulae and asymptotic exponentiallity of the exit times) and continuous-time scaling limit towards the Zig-Zag process are established under very general conditions. In larger dimension, the continuous-time scaling limit towards the Zig-Zag process is obtained through classical general arguments on the convergence of the generator, and the long-time convergence (and thus a LLN and a CLT) is obtained through the construction of a Lyapunov function and the identification of the irreducibility classes of the process under a growth condition on U. Indeed, the process is neither irreducible nor aperiodic, because the signature $\sigma(x,v) = ((-1)^{x_i} v_i)_{i \in [\![1,d]\!]}$ is multiplied by one at each step of the chain. Nevertheless, the x-marginal of the equilibrium is $\exp(-U)$ in each of the irreducibility classes, which means that for observables that depend only on the position x then we still get the convergence of ergodic means toward the target mean.

The second part of [22] is concerned with discretization schemes for hybrid (drift/jump) kinetic processes with generators

$$L \;=\; A_1 + A_2 + \sum_{i=1}^{M} A_{3,i} + \gamma A_4 + \lambda A_5 \,,$$

where, given a potential U and vector fields $F_0, \ldots, F_M$ on $\mathbb{R}^d$ with $\sum_{i=0}^{M} F_i = \nabla U$,

$$\begin{aligned}
A_1 f(x,v) &= v \cdot \nabla_x f(x,v) \\
A_2 f(x,v) &= -F_0(x) \cdot \nabla_v f(x,v) \\
A_{3,i} f(x,v) &= (v \cdot F_i(x))_+ \left(f\left(x, R_i(x,v)\right) - f(x,v) \right) \qquad \forall i \in [\![1,M]\!] \\
A_4 f(x,v) &= v \cdot \nabla_v f(x,v) + \Delta_v f(x,v) \\
A_5 f(x,v) &= \int_{\mathbb{R}^d} \left(f(x,w) - f(x,v) \right) \nu_d(\mathrm{d}w)
\end{aligned}$$

and where $\gamma, \lambda \geqslant 0$, ν_d is the standard d-dimensional Gaussian distribution and

$$R_i(x,v) \;=\; v - 2 \left(\frac{F_i(x) \cdot v}{|F_i(x)|^2} F_i(x) \right) \mathbb{1}_{F_i(x) \neq 0}$$

is the orthogonal reflection of v with respect to $F_i(x)$. For instance the Hamiltonian, Hybrid Monte-Carlo, Langevin, Zig-Zag or bouncy particle processes all have a generator

of this form. As soon as $F_0 \neq 0$, apart from some very particular cases (like when F_0 is linear) the equation $\ddot{x} = -F_0(x)$ cannot be solved exactly (and similarly if $\gamma \neq 0$ the diffusion with generator $A_1 + A_2 + \gamma A_4$ cannot be sampled exactly), so that we lose one of the interest of velocity jump samplers, namely exact simulation and unbiasedness of the equilibrium. Nevertheless, here I add in view applications in Molecular Dynamics, where systems present some high frequency oscillations, constraining the timesteps to be small, so that the discretization error is small with respect to the variance of the estimators (which is often high due to metastability). Moreover, recall that, in MD, ∇U is singular, and thus thinning methods (based on bounds on ∇U or higher derivatives) to sample the jump times would not be very efficient (many jumps would be proposed and rejected). Finally, in this context, the numerical cost of the simulation is mainly due to the computations of ∇U, that involves in particular a loop over all pairs of atoms of the system to compute interactions (in particular a bad thinning method, leading to many computations of ∇U, would be disastrous). With hybrid processes, the objective is precisely to reduce this numerical cost by evaluating the most costly part of ∇U less often.

Indeed, the fundamental remark is the following: the high-frequency and singular parts of ∇U are in fact quite cheap to compute since they only involve atoms that are close one to the others. On the other hands, long-range forces, that are costly since basically any atom interacts with all the others, are low-frequency and bounded. For instance, with the Van der Waals potential $W(r) = r^{-12} - r^{-6}$, the forces decay very fast with r the distance between the nuclei. Thus we can decompose

$$W'(r) \;=\; W'(r)\chi(r) + W'(r)\,(1-\chi(r))$$

with χ a cut-off function (equal to 1 in $[0, r_0]$ and to 0 in $[r_1, +\infty)$ for some $r_1 > r_0 > 0$). Summing over all the interactions the part $W'\chi$ in F_0 and using jump mechanisms $F_1, \dots, F_M$ to deal with the parts $W'(1-\chi)$, we end up with with a hybrid drift/jump process that cannot be sampled exactly.

A second-order discretization scheme is obtained through a so-called Trotter-Strang splitting, based on the approximation

$$e^{t(L_1+L_2+L_3)} \;=\; e^{tL_1/2}e^{t(L_2+L_3)}e^{tL_1/2} + \underset{t\to 0}{o}(t^2) \;=\; e^{tL_1/2}e^{tL_2/2}e^{tL_3}e^{tL_2/2}e^{tL_1/2} + \underset{t\to 0}{o}(t^2)\,.$$

In [22] I considered the case

$$L_1 = A_1\,, \qquad L_2 = A_2 + \gamma A_4\,, \qquad L_3 = \sum_{i=1}^{M} A_{3,i} + \lambda A_5\,.$$

Other choices are possible: the important thing is that we should be able to sample exactly each step. The novelty here with respect to classical schemes for the Langevin diffusion comes from the generator L_3, which corresponds to velocity jumps. This is the only place where long-range forces have to be computed. This is done efficiently through thinning. Indeed, to sample the jumps of the velocity of an atom following L_3 for a time Δt, instead of computing $N-1$ terms where N is the number of atoms in the system, we computed S such terms where S is a random variable distributed according to a Poisson law with parameter $(N-1)Kc\Delta t$ where c is some normalization constant and K is a bound on $\|F_i\|_\infty$, $i \in [\![1, M]\!]$. Since Δt is chosen depending on the high forces

F_0, we typically have $Kc\Delta t \ll 1$. Indeed, with Jeremy Weisman, Louis Lagardère and Jean-Philip Piquemal we implemented in [23] this scheme in the MD software Tinker [LJL+17]. In cases where all the forces are pair interactions (Coulomb and Van der Waals forces), the simulation times where divided by 4. In cases with many-body interactions, we still obtained a speed-up with respect to state-of-the-art integrators.

2.3.1 *Some on-going projects and perspectives*

1. With Mathias Rousset and Pierre-André Zitt we are currently finishing a work where we introduce a jump mechanism (i.e. jump rate and kernel) depending on some parameter $\varepsilon \in (0, +\infty]$ so that : 1) for all ε the target $\exp(-H)$ with $H(x,v) = U(x) + |v|^2/2$ is invariant for the corresponding velocity jump process, 2) $\varepsilon = +\infty$ corresponds to the jump mechanism of the bouncy particle sampler, and 3) as $\varepsilon \to 0$ the jump mechanism converges to the continuous drift $-\nabla U(x)\nabla_v$ (so, depending on whether there is some dissipative mechanism, the corresponding velocity jump process converges for instance to the Hamiltonian dynamics, the Langevin diffusion or the HMC process). In other words, this gives an interpolation between bounces and continuous drift ; between the bouncy particle sampler and the Hamiltonian dynamics (see Figure 2.1). The parameter ε then plays a role similar to the timestep of a classical integrator : the smaller it is, the more numerically intensive is the simulation, but the closer the process is to the Hamiltonian dynamics. This allows to compute dynamical properties (like diffusion coefficients, transition times and paths...), while conserving unbiased estimation for statistical properties, and the possibility to use the forces splitting method.

2. With Boris Nectoux we started to work on the metastable behaviour of velocity jump processes at small temperature, more precisely on Eyring-Kramer formulas for the exit time of potential wells.

3. The work [23] is a proof of concept, and there is now much work to do in order to derive full benefit from the method. In particular, up to now we have only treated pairwise interactions, since it was the simpler case and thus we knew how to split the forces. But most of the numerical cost comes from forces whose computation involve all the atoms. This is the case of the reciprocal Coulomb forces (see [ELB+95]), that involve some Fourier computations in the whole space, and of polarization forces (see [LJL+17]) that require to solve some linear system involving all particles. Reducing the number of computations of such forces per simulation time would have an impact even larger than pairwise interactions. However, obtaining explicit (and efficient) bounds on the jump rate is then impossible. A possible strategy is to use an empirical (constant) bound. This introduces an additional numerical error, but which is small according to [3, Proposition 26].

2.4 Uniform estimates for mean-field interacting jump processes

While [17, 9] were concerned with mean-field kinetic diffusions (in particular with uniform in N longtime convergence rates), the works [18, 10, 11] are concerned with simi-

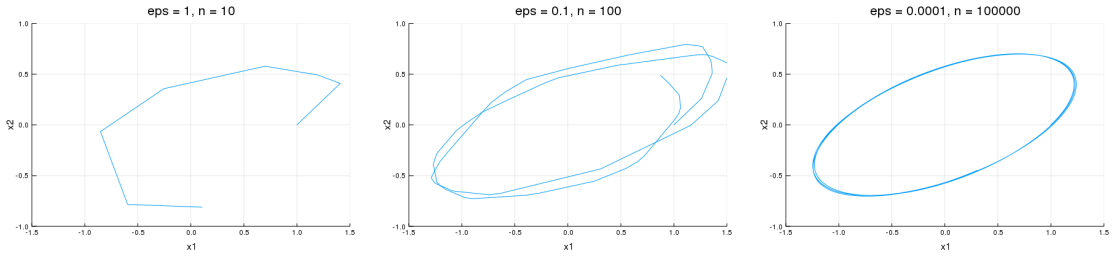

Figure 2.1: Trajectory of the process in a quadratic potential in dimension 2 for $\varepsilon = 1$, 10^{-1} and 10^{-4}, respectively after $n = 10$, 10^2 and 10^5 jumps.

lar questions but for processes where the interaction intervenes in the jump mechanism rather than in the drift of the dynamics.

More precisely, [18] is interested in the general question of the long-time behaviour of mean-field semi-linear integro-differential equations of the form

$$\partial_t m_t(x) \quad = \quad L' m_t(x) + Q'_{m_t}\left(\lambda_{m_t} m_t\right)(x) - \lambda_{m_t}(x) m_t(x)\,,$$

where $m_t(x)$ is the density of particles at time t at point x, L is the generator of a Markov process, A' denotes the dual on measures of an operator A on functions, and for all probability distribution ν, $\lambda_\nu > 0$ is a jump rate and Q_ν is a jump kernel. In other words, the stochastic process whose law m_t solves this equation follows the dynamics of L and, at a rate λ_{m_t}, jumps to a new position drawn according to the kernel Q_{m_t}.

The main argument is simple and based on coupling methods. In the classical linear case, a contraction in (weighted) total variation distance can be obtained by constructing simultaneously two processes starting at different positions in such a way that they have merged at some positive time with some positive probability. Once they have met, then they remain equal forever. However, in the non-linear case, even if we manage to merge two processes that started at different positions, they still don't have exactly the same law, hence the same dynamics, and thus they may have to split after some time. Nevertheless, as they stay together, their laws get closer one to the other, which makes a splitting less and less likely. Provided the non-linearity is small enough, the balance between this two merge/split mechanisms is on the side of contraction.

It is not possible to do this when the non-linearity lies at the level of the drift of the process since, even if the two processes are equal at some time, then instantaneously they split due to the difference of their drift (although if there is some diffusion, it is possible to consider the sticky coupling of [EZ19] to merge them again instantaneously). This is why, for the Vlasov-Fokker-Planck equation [17, 9] for instance, it is more natural to work with Wasserstein distances rather than the total variation one since we can still ensure that the processes stay close (although they have splitted). The story is different when the non-linearity lies in the jump mechanism: indeed, considering the synchronous coupling of the two processes, then they stay together up to the first asynchronous jump. The main assumption in [18] is that there exists $\theta > 0$ such that for all ν, μ probability distributions

$$\sup_{\|f\|_\infty \leqslant 1} \|\lambda_\nu(Q_\nu f - f) - \lambda_\mu(Q_\mu f - f)\|_\infty \;\leqslant\; \theta \|\nu - \mu\|_{TV}\,. \tag{2.6}$$

Indeed, the left hand side is exactly the bound on the asynchronous jump rate we ob-

tained in [3]. The assumption then means that the closer the laws μ and ν, the longest it is possible to keep together processes with jumps given respecitvely by λ_μ, Q_μ and λ_ν, Q_ν. In fact the condition (2.6) is sufficient to prove the existence of a weak solution m_t to the non-linear equation and of an associated inhomogeneous Markov process. If, additionally, the jump rate is uniformly bounded, i.e. $\lambda_\nu(x) \leqslant \lambda_*$ for some $\lambda_* > 0$ for all ν and x, and if the Markov semigroup $(P_t)_{t\geqslant 0}$ associated to the generator L satisfies a Doeblin condition

$$\forall x, y, \qquad \|\delta_x P_{t_0} - \delta_y P_{t_0}\|_{TV} \leqslant 2(1-\alpha)$$

for some $t_0, \alpha > 0$, there the previous merge/split coupling argument yields

$$\|m_t - h_t\|_{TV} \leqslant e^{\theta t_0}\left(e^{\theta t_0} - \alpha e^{-\lambda_* t_0}\right)^{\lfloor t/t_0 \rfloor} \|m_0 - h_0\|_{TV}\,.$$

for all $t \geqslant 0$ if $t \mapsto m_t, h_t$ are two different solutions of the non-linear equation. In particular, if θ is small enough (i.e. if the non-linearity is small enough), this gives a contraction. This first result is then extended to the case where the Doeblin condition for P_{t_0} is only local and there exist a Lyapunov function.

Still in [18], the same argument is applied to the system of interacting particles associated to the non-linear equation, namely $X_1, \dots, X_N$ evolve independently according to the generator L between jumps that occurs for X_i at rate $\lambda_{\pi_t^N}(X_i)$ where $\pi_t^N = 1/N \sum_{j=1}^N \delta_{X_j}$ is the empirical distribution of the system. Then a crucial point is that we should *not* use the total variation distance (i.e. the Wasserstein distance associated to the distance $(x,y) \mapsto \mathbb{1}_{(x_1,\dots,x_N)\neq(y_1,\dots,y_N)}$), but rather the Wasserstein distance associated to the distance $(x,y) \mapsto \sum_{i=1}^N \mathbb{1}_{x_i \neq y_i}$ (which up to a factor $1/N$ is in fact nothing different from the total variation distance of the empirical distributions of the two systems, at least when all particles are distinct). Indeed, with the total variation distance, one need to merge in a single step all the particles, which happens with a probability α^N (which yields a convergence rate geometrically small in N), while with the other distance it is possible to couple only some particles and then to try and keep them together as long as possible, waiting for other particles to merge. Under the condition (2.6), of course, the more pairs of particles (x_i, y_i) have already merged, the smaller is the total variation distance between the empirical distributions of the two systems, and thus the easier it is to keep together pairs that have already merged. It is then not surprising that we end up with the same rate as for the non-linear equation, namely

$$\|m_0 R_t - h_0 R_t\|_{TV} \leqslant N e^{\theta t_0}\left(e^{\theta t_0} - \alpha e^{-\lambda_* t_0}\right)^{\lfloor t/t_0 \rfloor} \|m_0 - h_0\|_{TV}\,,$$

where R_t is the Markov semi-group of the system and m_0, h_0 are any initial conditions. The factor N comes from the bound $\sum_{i=1}^N \mathbb{1}_{x_i\neq y_i} \leqslant N \mathbb{1}_{(x_1,\dots,x_N)\neq(y_1,\dots,y_N)}$. Moreover, denoting m_t' and h_t' the respective laws of $X_{I,t}$ and $Y_{I,t}$ where $X_t \sim m_0 R_t$, $Y_t \sim h_0 R_t$ and I is uniformly distributed over $[\![1, N]\!]$ and independent from X_t and Y_t, then

$$\|m_t' - h_t'\|_{TV} \leqslant e^{\theta t_0}\left(e^{\theta t_0} - \alpha e^{-\lambda_* t_0}\right)^{\lfloor t/t_0 \rfloor} \|m_0' - h_0'\|_{TV}\,.$$

As in the non-linear case, a similar result is also stated and proved with a Lyapunov/local Doeblin condition.

These general results are then illustrated on several examples: mean-field interacting run-and-tumble processes, MCMC for granular media equilibrium, the Zig-Zag process or Hybrid drift/bounce kinetic samplers with a close to tensor target, Selection/Mutation algorithms and mean-field TCP processes.

The work [10] with Lucas Journel follows the same strategy. It is devoted to the proof of the Flemming-Viot algorithm for diffusions on the torus with smooth killing. Consider a diffusion

$$\mathrm{d}X_t \;=\; b(X_t)\mathrm{d}t + \mathrm{d}B_t$$

for some continuous b on the torus $\mathbb{T}^d$, and a continuous killing rate $\lambda : \mathbb{T}^d \to \mathbb{R}_+$. The aim of the Flemming-Viot algorithm is to approximate the Quasi-Stationary Distribution (QSD) of the diffusion killed at rate λ, which means here the limit as $t \to +\infty$ of the law of X_t conditionally to the event $\{T > t\}$ where, given E a standard exponential random variable independent from X, the death time T is given by

$$T \;=\; \inf\left\{ t \geqslant 0\,, E < \int_0^t \lambda(X_s)\mathrm{d}s \right\}\,.$$

The algorithm is based on a system of interacting particles $(X_1, \dots, X_N)$ that are independent diffusions up to the death of one of the particle which is then resurrect on X_J with J uniformly distributed over $[\![1, N]\!]$. Note that, as in the previous framework of [18], if we consider two systems $(X_1, \dots, X_N)$ and $(Y_1, \dots, Y_N)$, then if a pair (X_i, Y_i) has merged we can keep them together up to their death (which can be chosen to be synchronous) and then we can resurrect them on (X_J, Y_J) (with the same random index J) so that the probability to resurrect them at the same place is exactly given by the proportion of pairs that have already merged. As a consequence, a similar result is established, so that, provided λ is small enough with respect to the mixing properties of the diffusion, the convergence rate of the Markov system $(X_1, \dots, X_N)$ is independent from N. In fact in [10] we decided to work with the usual $\mathcal{W}_1$ distance rather than the total variation distance, so that the smallnest assumption does not concern λ but the Lipschitz norm of λ. The reason is that we wanted to treat both the limits $t \to +\infty$ and $N \to +\infty$ at the same time. For the limit $N \to +\infty$, one needs to couple an empirical distribution with the limit law that has a density with respect to the Lebesgue measure, so that the total variation distance is not suitable (the distance would always be 2). Besides we went one step closer to the real implementation by considering, instead of the continuous-time diffusion, its discrete-time Euler scheme. Then there are three sources of error between the empirical distribution π_t of the system (which is what is available in practice) and the theoretical QSD (the target): N, t, and the timestep $\gamma > 0$. We proved that, provided the Lipschitz norm of λ is small enough, then the three convergences $(N, t \to +\infty, \gamma \to 0)$ are uniform with respect to the two other parameters. In particular this gives an error bound between the result of the algorithm and its target:

Theorem 4. There exists $c_0, \gamma_0, C, \kappa > 0$ that depends only on the drift b and the dimension d such that, if $\gamma \in (0, \gamma_0]$ and if λ is Lipschitz with a constant L_λ such that

$$L_\lambda e^{\gamma \|\lambda\|_\infty} \;<\; c_0\,,$$

then for all $N \in \mathbb{N}$, $t \geqslant 0$ and initial distribution,

$$\mathbb{E}\left[\mathcal{W}_1\left(\pi_t, \nu_*\right)\right] \leqslant C\left(\sqrt{\gamma} + \alpha(N) + e^{-\kappa t}\right),$$

where

$$\alpha(N) = \begin{cases} N^{-1/2} & \text{if } d = 1, \\ N^{-1/2}\ln(1+N) & \text{if } d = 2, \\ N^{-1/d} & \text{if } d > 2. \end{cases}$$

Finally, with Eva Löcherbach, we studied in [11] a model of mean-field interacting neurons. Each neuron emits action potentials (spikes) at a rate $\lambda(u)$ depending on its membrane potential value u. At the spiking time, the neuron's potential is reset to 0. At the same time all other neurons receive an additional amount of potential h/N, where $h > 0$ is the synaptic weight and N the size of the system. Finally, in between successive jumps, each neuron's potential undergoes some leak effect and looses potential at exponential rate $\alpha > 0$. The associated generator is given by

$$Lf(u) = \sum_{i=1}^{N} \lambda(u_i)\left[f(u + \Delta_i(u)) - f(u)\right] - \alpha x \cdot \nabla f(x),$$

where

$$(\Delta_i(u))_j = \left\{ \begin{array}{ll} \frac{h}{N} & j \neq i \\ -u_i & j = i \end{array} \right\},$$

As $N \to +\infty$, spikes are more and more frequent, but have less and less effect on other particles, so that in the mean-field limit, the interaction intervenes in a deterministic drift rather than a jump mechanism. More precisely, the limit process solves

$$d\bar{U}(t) = -\alpha\bar{U}(t)dt + h\mathbb{E}(\lambda(\bar{U}(t))dt - \bar{U}(t-)\int_{\mathbb{R}_+} \mathbb{1}_{\{z \leq \lambda(\bar{U}(t-))\}}\pi(dt, dz),$$

where $\pi(dt, dz)$ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}_+$ having intensity $dtdz$.

If λ is differentiable at 0 then extinction is almost sure for a system of N interacting neurons, namely at some point there is a last spike, from which the system simply goes to zero following $\dot{u} = -\alpha u$. Nevertheless, when $h\lambda'(0)$ is sufficiently large with respect to α, we can expect that mutual excitation of the neurons is stronger than the leakage, so that the extinction time should be very long (in N) and the system should be metastable. We prove in [11] several results in this direction, some under general conditions on λ and some in the particular case where $\lambda(u) = \max(ku, \lambda_*)$ for some $k, \lambda_* > 0$. The results in the latter case depends on the parameters $a = \alpha/(kh)$ and $b = \lambda_*/(kh)$, and are summarized in Figure 2.2.

The proofs heavily rely on synchronous coupling of processes. In particular, we couple the average spiking rate $\Lambda_N(t) = 1/N\sum_{i=1}^{N}\lambda(U_i(t))$ (which is not a Markov process) with a simple Markov process $(Z_N(t))_{t\geqslant 0}$ on $\mathbb{R}_+$ in such a way that $Z_N(t) \leqslant \Lambda_N(t)$ for all $t \geqslant 0$ and that all jumps of Z_N are spikes for the neuron system. The idea is that if $\Lambda_N(t)$ is small, then necessarily a large portion of neurons have a potential close to 0, so that at the next spiking time their potential will increase approximately by $h\lambda'(0)$.

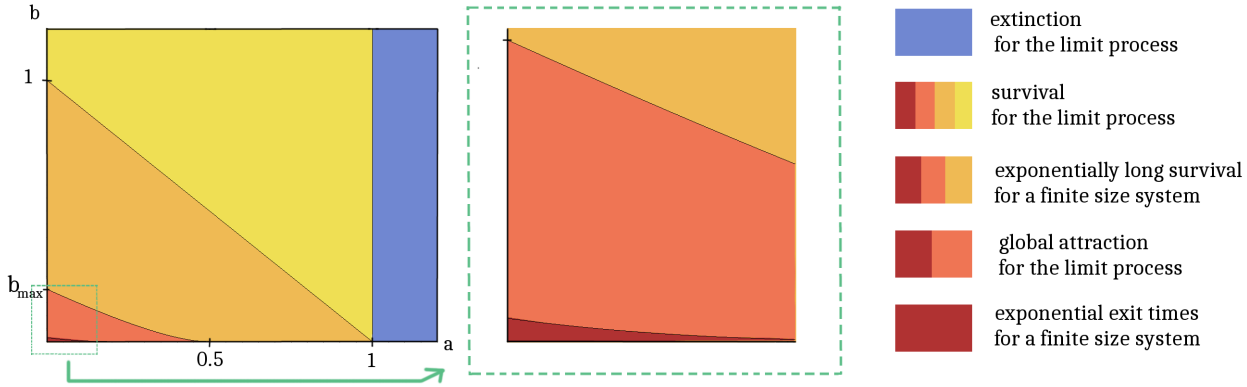

Figure 2.2: If $a > 1$, δ_0 is the unique equilibrium of the limit process, globally attractive and if $a < 1$ it is unstable and there exists at least a positive equilibrium. If $a + b < 1$, the last spike time for a finite system is exponentially large with the size of the system. Under some explicit condition on a and b, the positive equilibrium of the limit process is unique and globally attractive. Under another explicit (stronger) condition on a and b, the exit times of a finite size system from a neighborhood of the limit equilibrium converge to an exponential distribution as $N \to +\infty$.

That way we get a lower bound on $\Lambda_N(t)$ in terms of $\Lambda_N(t-)$ when there is a spike at time t, and we define the jumps of Z_N to be this lower bound. The study of Z_N, which is a one-dimensional jump process, relies then on classical Large Deviation tools. Letting N go to infinity, since $\Lambda_N(t)$ converges to $\mathbb{E}(\lambda(\bar{U}(t)))$, we get that the latter is bounded below by the deterministic ODE that is the limit of Z_N. Provided a and b are sufficiently small, this ODE has a positive equilibrium, sufficiently large so that $\bar{U}$ has a large positive drift and thus $\lambda(\bar{U})$ reaches its saturation value λ_* in a sufficiently small time. This is then crucial to bound two limit processes $\bar{U}$ and $\hat{U}$ with different initial conditions. Indeed, contrary to the works [18, 10], since here the non-linearity is in the drift, we cannot keep two processes equal for some time, they split instantaneously. But since the jump rate saturates, an asynchronous jump can only occur if one of the rate is smaller than λ_*, which is only true for a short time after each spike. So, provided a and b are small enough, the positive drift is sufficiently large so that the probability to see an asynchronous jump is sufficiently small, while a synchronous spikes make the processes equal (although they split immediatly, their distance is then related to the difference of their drift, hence of their law, and this is sufficient to conclude). A similar strategy can then be adapted for a system of N interacting neurons, provided we don't see a large deviation in the jump mechanisms (that happens with exponentially small probability in N). This is an ingredient for proving the asymptotic exponentiality of the exit times. Indeed, adapting classical arguments in the case of diffusion processes to a more general framework, we proved in [11] the following general result.

Theorem 5 (from Theorem 4.3 of [11]). For all $\delta, C > 0$ there exist $M_0 > 0$ such that the following holds. Let $(X_t)_{t \geqslant 0}$ be a time-homogeneous strong Markov process taking values in some Polish space E. Let $\mathcal{D}, \mathcal{K}$ be measurable subsets of E with $\emptyset \neq \mathcal{K} \subset \mathcal{D}$. For $A \subset E$, denote $\tau_A = \inf\{t \geqslant 0 \ : \ X_t \in A\}$. Let $\varepsilon_1, \varepsilon_2, \varepsilon_3 \in [0,1]$ and $s_1 \geqslant s_2 > 0$ be

such that

$$\begin{aligned}\varepsilon_1 &\geqslant \sup_{x\in\mathcal{K}} \mathbb{P}_x\left(\tau_{\mathcal{D}^c} \leqslant s_1\right)\\ \varepsilon_2 &\geqslant \sup_{x\in\mathcal{D}} \mathbb{P}_x\left(\tau_{\mathcal{D}^c} \wedge \tau_{\mathcal{K}} > s_2\right)\end{aligned}$$

and that for all $x, y \in \mathcal{K}$ there exists a coupling $(X_t, Y_t)_{t\geqslant 0}$ of two processes with respective initial condition x and y such that

$$\mathbb{P}\left(X_t = Y_t \ \forall t \geqslant s_1\right) \ \geqslant \ 1-\varepsilon_3\,.$$

Suppose that $\tau_{\mathcal{D}^c}$ is $\mathbb{P}_{x_0}$-almost surely finite for some $x_0 \in \mathcal{K}$. Finally, suppose that

$$\varepsilon_1+\varepsilon_2+\varepsilon_3 \ \leqslant \ Ce^{-\delta M}\,, \qquad \mathbb{P}_{x_0}\left(\tau_{\mathcal{D}^c} \leqslant s_2 e^{\delta M}/C\right) \leqslant \frac{1}{4}$$

for some $M \geqslant M_0$. Then,

$$\sup_{y\in\mathcal{D}} \mathbb{E}_y(\tau_{\mathcal{D}^c}) \ < \ +\infty$$

and for all $x, y \in \mathcal{K}$

$$\sup_{t\geqslant 0} \left|\mathbb{P}_x\left(\tau_{\mathcal{D}^c} > t\mathbb{E}_x\left(\tau_{\mathcal{D}^c}\right)\right) - e^{-t}\right| \ \leqslant \ K' M^3 e^{-\min(\delta/3,1/2)M}$$

and

$$\left|\frac{\mathbb{E}_x\left(\tau_{\mathcal{D}^c}\right)}{\mathbb{E}_y\left(\tau_{\mathcal{D}^c}\right)} - 1\right| \ \leqslant \ K' M^3 e^{-\min(\delta/3,1/2)M}\,.$$

2.4.1 *Some on-going projects and perspectives*

The work [10] is in fact a first step, the aim is to prove a similar result of convergence of the Flemming-Viot algorithm but in the hard killing case, namely when the process is not killed at some continuous rate but when it reaches the boundary of some domain. More precisely, an interesting case would be the low temperature regime for a diffusion process, when the metastability within the domain is smaller than the exit time from the domain. Indeed, in that case, the death time is large with respect to the mixing time, which is exactly the condition required for the strategy of [10] to work. This is also the most interesting case from a practical point of view. The additional difficulty is then that the probability to die in a given time is not bounded uniformly over the whole domain, because of the areas close to the boundary.

2.5 Simulated annealing

My two works [19, 15] concerned with simulated annealing based respectively on the Langevin process and the bouncy particle sampler have already been mentioned above. In the more recent [6] with Nicolas Fournier and Camille Tardif, we considered the

classical simulated annealing based on the overdamped Langevin diffusion, namely

$$\mathrm{d}X_t \;=\; -\nabla U(X_t)\mathrm{d}t + \sqrt{2\beta_t^{-1}}\mathrm{d}B_t\,.$$

The question was to find minimal conditions (slow growth at infinity, possibly unbounded set of local minima) on U to ensure the convergence in probability of X_t toward the global minima of U. Indeed, all classical results are mostly interested in the behaviour of U in some compact set and just assume a convenient confining behaviour at infinity, in particular in order to have a Poincaré or log-Sobolev inequality for $\exp(-\beta U)$ for all $\beta > 0$ as in [Mic92, HKS89]. The only anterior attempt to deal with slowly growing potential was the work [Zito8] of Zitt (still based on functional inequalities argument), under the condition that, outside some compact set, $U(x) \geqslant (\ln|x|)^a - C$ for some $a > 1$ and $C > 0$ and $\Delta U < 0$ (which is a strong assumption).

Assume that U goes to infinity at infinity, that its critical depth E_* is finite and that $\beta_t = \beta_0 + \ln(1+t)/c$ for some $c > E_*$. In [6] we use that, from classical results on the simulated annealing, whenever the process comes back to a given compact set containing the global minima of U, then it has some probability to stay forever in a set $\{U \leqslant C\}$ for some $C > 0$, so that its convergence toward the local minima is thus essentially the result of the convergence results in the compact case (say on the periodic torus). So what has to be checked is whether the process always come back to compact sets. We consider two different conditions (in fact three but the third is a variation of the second):

- First, we consider the case where, outside a compact set, the drift is directed toward the origin, namely where $x \cdot \nabla U(x) \geqslant 0$ for x large enough. In that case there is a transition at $U(x) = a\ln\ln(|x|)$ with $a = c(d-2)/2$. More precisely, if we assume that, outside some compact set, $x \cdot \nabla U(x) \geqslant a/\ln(|x|)$ with $a > c(d-2)/2$, then $U(X_t)$ converges in probability to $\min U$. On the contrary, for $U(x) = a\ln\big(1+\ln(1+|x|^2)\big)$ with $a < c(d-2)/2$ (for which the critical depth is 0) then, for all initial conditions, $\mathbb{P}(\lim_{t\to+\infty} U(X_t) = +\infty) > 0$. These two results are established through explicit comparison with Bessel processes, whose recurrence properties are known.

- Second, we consider cases where there is an unbounded family of ring on which the potential grows sufficiently. More precisely, we assume that there exist $\alpha, \delta_0 > 0$ and three sequences $(a_i)_{i\geqslant 1}$, $(b_i)_{i\geqslant 1}$ and $(\delta_i)_{i\geqslant 1}$ such that $0 \leqslant a_1 < b_1 \leqslant a_2 < b_2 \leqslant \ldots$ and, for all $i \geqslant 1$, $\delta_i \geqslant \delta_0$, $b_i \geqslant a_i + \alpha\delta_i$, and

$$|x| \in [a_i, b_i] \quad\Longrightarrow\quad \frac{x}{|x|}\cdot\nabla U(x) \geq \frac{1}{\delta_i}.$$

  Then, if $\alpha > c$, $U(X_t)$ converges in probability to $\min U$. This assumption allows for unbounded sets of local minima, and arbitrarily slow growth, in the sense that for all $p \geqslant 1$, denoting $\ln^{\circ p} = \ln(\ln(\ldots$ composed p times, there exist potentials such that $\ln^{\circ p}|x| \leqslant U(x) \leqslant 3\ln^{\circ p}|x|$ outside some compact that satisfies this assumption. The result is obtained by considering on each ring $[a_i, b_i]$ a Bessel process initialized at a_i and reflected at a_i, in such a way that this Bessel process is above $|X_t|$ as long as the two processes stay in $[a_i, b_i]$, and such that there is a probability (bounded away from 0 uniformly in i) that the process never reaches

b_i. This implies that there is a random $i \in \mathbb{N}$ such that $|X_t|$ will never hit b_i.

2.5.1 *Some on-going projects and perspectives*

- Pursuing the study started in [6], Nicolas Fournier and Camille Tardif have established in [FT20] the convergence of the simulated annealing if $c > E_*$ as soon as U goes to infinity at infinity and $\exp(-\alpha U)$ has finite mass for some $\alpha > 0$. On the one hand, has we saw in [6], this condition is not necessary. On the other hand, it is a very simple, general and natural condition. The strategy of [FT20] is still to use a localization procedure : basically, all one need to prove is that the process does not go to infinity, and then conclusion follows from the convergence of the algorithm on compact sets. Of course this is not so easy, but the arguments are sufficiently clear to be expected to extend to other processes than the overdamped Langevin diffusion. In particular this will allow to widely extend the results of [19] which required very restrictive behaviour of U at infinity due to entropic hypocoercive computations.

- Let me simply mention here a project of using simulated annealing versions of saddle-search algorithms in the spirit of [LO17]. The basic idea is to replace the gradient descent $\dot{x} = -\nabla U(x)$ by the $\dot{x} = -R(x)\nabla U(x)$ where $R(x) = I - 2v_x v_x^T$ is the orthogonal reflection with respect to v_x a normalized eigenvector of $\nabla^2 U(x)$ associated its the smallest eigenvalue. That way, local minima become unstable but index-1 saddle points become local attractors. Now the deterministic system is no more a gradient flow and, adding some small noise (say at constant temperature), the corresponding diffusion is a non-equilibrium non-reversible process, with non-explicit equilibrium. Compared to the reversible case, small-noise non-equilibrium diffusion processes may have very rich behaviours, depending on the limiting ODE. Finally, in order to lower metastability and find new local minima, a strategy is to switch at random times between a classical overdamped Langevin dynamics and a noisy saddle-point search (and possibly other similar processes). The behaviour of such a process (in particular of its invariant measure and mixing rate at low-temperature) would then depend on the relation between the Brownian noise intensity and the switching rate. As a conclusion, this algorithm opens many practical and theoretical questions.

2.6 Adaptive algorithms

The works [7, 24, 5, 1] are concerned with adaptive bias algorithms, mostly the Adaptive Biasing Force (ABF) algorithm (except for [7] on metadynamics, a similar algorithm). This class of algorithms, meant for sampling metastable high-dimensional Gibbs measures in Molecular Dynamics, are based on so-called reaction coordinates. A reaction coordinate is a function $\xi : \mathbb{R}^d \to \mathbb{R}^m$ where $m \ll d$ (typically $m \leqslant 3$ while $d \geqslant 10^3$) which is supposed to capture the macroscopic, slow variables of a molecular system. While $q \in \mathbb{R}^d$ describes all the positions of the nuclei of the system, $\xi(q)$ can be the torsion angle of a molecule, or a particular distance in the system (for instance between a ligand and its binding site in a protein). Up to now reaction coordinates have been designed by practitioners, based on a priori knowledge or intuitions on the system, and in the last year the question of unsupervised learning of such reaction coordinates have

gained interest, see [GSB+20]. Here we will assume ξ is given. A "good" reaction coordinate is such that it gathers most of the metastability of the system, in the sense that if $\xi(q)$ has visited all the space, then q has visited all the low-energy regions of the space. In the rest of this section, for the sake of simplicity, we will assume periodic boundary condition (i.e. $q \in \mathbb{T}^d$) and that $\xi : \mathbb{T}^d \to \mathbb{T}^m$ is simply given by $\xi(q_1, \dots, q_d) = (q_1, \dots, q_m)$. While this seems a very restrictive case, it is in fact the case for so-called Alchemic transformations [KB96], and in general it is always possible to get back to this case by considering extended variables as in [LRS07], which is also motivated by practical issues. In the following we denote $q = (x, y) \in \mathbb{T}^m \times \mathbb{T}^{d-m}$, so that $\xi(x, y) = x$. We also consider a potential $U \in \mathcal{C}^1(\mathbb{T}^d, \mathbb{R})$.

To a reaction coordinate ξ is associated its free energy, which is by definition the function A such that, if q is a random variable distributed according to the Gibbs law $\propto \exp(-U)$, then $\xi(q)$ follows the distribution $\propto \exp(-A)$ (in particular, the free energy is in fact defined up to an additive constant). With our present choice of ξ, A is well-defined and is simply given by

$$A(x) \;=\; -\ln \int_{\mathbb{T}^{d-m}} e^{-U(x,y)} \mathrm{d}y \,.$$

The fundamental remark on which are based the adaptive bias algorithms is the following: the invariant measure of the diffusion

$$\left\{ \begin{array}{rcl} \mathrm{d}X_t &=& -\nabla_x U(X_t, Y_t)\mathrm{d}t + \nabla_x A(X_t) + \sqrt{2}\mathrm{d}B_t^1 \\ \mathrm{d}Y_t &=& -\nabla_y U(X_t, Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^2 \end{array} \right.$$

with $B = (B^1, B^2)$ a standard Brownian motion, is $\propto \exp(-U(x, y) + A(x))$. In particular, if (X, Y) is at equilibrium, then $\xi(X)$ is distributed according to $\exp(A(x) - A(x)) = 1$, in other words it is uniformly distributed over $\mathbb{T}^m$. This means there is no more any metastability at the level of $\xi(X)$ (which means that, if ξ is a "good" reaction coordinate, there is much less metastability in the whole system), all the values are equally visited. Of course, the process is not targeting the correct Gibbs measure, but then this is simply a classical Importance Sampling scheme, and we can just compute averages with respect to $\exp(-U)$ by a re-weighting step:

$$\frac{\int_0^t \varphi(X_s, Y_s) e^{-A(X_s)} \mathrm{d}s}{\int_0^t e^{-A(X_s)} \mathrm{d}s} \underset{t \to +\infty}{\longrightarrow} \frac{\int_{\mathbb{T}^m} \varphi(x, y) e^{-U(x,y)} \mathrm{d}x\mathrm{d}y}{\int_{\mathbb{T}^m} e^{-U(x,y)} \mathrm{d}x\mathrm{d}y}$$

(or, alternatively, an estimator can be based on the fact that at equilibrium the conditional densities proportional to $y \mapsto \exp(-U + A \circ \xi)$ are correct).

There is a slight issue here: A is unknown. In fact, computing $A(x)$ for a given value of x is as difficult as the initial sampling problem. In many applications in MD, the goal is precisely to compute the free energy. The general idea of adaptive biasing algorithms is to consider a process

$$\left\{ \begin{array}{rcl} \mathrm{d}X_t &=& -\nabla_x U(X_t, Y_t)\mathrm{d}t + \nabla_x A_t(X_t) + \sqrt{2}\mathrm{d}B_t^1 \\ \mathrm{d}Y_t &=& -\nabla_y U(X_t, Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^2 \end{array} \right. \tag{2.7}$$

where A_t is constructed on the fly, based on the trajectory $(X_s, Y_s)_{s\in[0,t)}$, in such a way it converges in the longtime to A. There are several ways to do so, which yields various similar algorithms. In the ABF algorithm, instead of approximating A, the target is the force ∇A. Indeed,

$$\nabla A(x) \;=\; \frac{\int_{\mathbb{T}^{d-m}} \nabla_x U(x,y) e^{-U(x,y)} \mathrm{d}y}{\int_{\mathbb{T}^{d-m}} e^{-U(x,y)} \mathrm{d}y} \;=\; \mathbb{E}\left(\nabla_x U(X,Y) \mid X\right)$$

when $(X,Y) \sim \exp(-U)$. More generally, for any B,

$$\nabla A(x) \;=\; \frac{\int_{\mathbb{T}^{d-m}} \nabla_x U(x,y) e^{-U(x,y)+B(x)} \mathrm{d}y}{\int_{\mathbb{T}^{d-m}} e^{-U(x,y)+B(x)} \mathrm{d}y} \;=\; \mathbb{E}\left(\nabla_x U(X,Y) \mid X\right)$$

when $(X,Y) \sim \exp(-U(x,y)+B(x))$. Another way to see this formula is that A is the minimizer over $\mathcal{H}^1(\mathbb{T}^m,\mathbb{R})$ of

$$\mathcal{E}(f) \;:=\; \int\limits_{\mathbb{T}^m\times\mathbb{T}^{d-m}} |\nabla_x U(x,y) - \nabla_x f(x)|^2 e^{-U(x,y)+B(x)} \mathrm{d}x\mathrm{d}y$$

for any B. The idea is then to replace the Gibbs law $\exp(-U)$ by an approximation obtained from the trajectory $(X_s, Y_s)_{s\in[0,t]}$. Consider the (possibly reweighted) occupation measure

$$\nu_t \;=\; \frac{\int_0^t \omega_s \delta_{(X_s,Y_s)} \mathrm{d}s}{\int_0^t \omega_s \mathrm{d}s}$$

with either $\omega_s = 1$ (non-reweighted case) or $\omega_s = \exp(-A_s(X_s))$ (reweighted case). Assuming that A_s (which is yet to define) converges to A, this occupation measure should converge either to $\exp(-U)$ or to $\exp(-U+A\circ\xi)$, so in both case we would like to define A_t by replacing $\exp(-U+B\circ\xi)$ in the defintion of $\mathcal{E}(f)$ by ν_t. This is not possible since the minimization problem would then be ill-posed (as ν_t is a probability measure supported on the trajectory of the process).

To circumvent this difficulty, we consider a regularization parameter $\lambda \geqslant 0$ and a smooth symmetric positive density kernel $K \in \mathcal{C}^\infty(\mathbb{T}^m\times\mathbb{T}^m,\mathbb{R}_+)$ with

$$\int\limits_{\mathbb{T}^d} K(x,z)\mathrm{d}z = 1 \qquad \text{and} \qquad K(x,z) = K(z,x) \qquad \forall x,z\in\mathbb{T}^m\,.$$

In practice, $K(x,\cdot)$ should be close to a Dirac mass at x. For instance, a possible choice for K would be the so-called von-Mises kernel for a given small parameter $\varepsilon > 0$, i.e.

$$K(x,z) \;\propto\; \prod_{i=1}^d \exp\left(-\frac{1}{\varepsilon^2/2}\sin^2\left(\frac{z_i-x_i}{2}\right)\right)\,.$$

Now, for $f\in\mathcal{H}^1(\mathbb{T}^m,\mathbb{R})$, we define

$$\mathcal{J}_t(f) \;:=\; \int\limits_{\mathbb{T}^{d-m}\times\mathbb{T}^m\times\mathbb{T}^m} |\nabla_x U(x,y) - \nabla_x f(z)|^2 K(x,z)\mathrm{d}z\mathrm{d}\nu_t(x,y) + \lambda \int\limits_{\mathbb{T}^m} |\nabla_z f(z)|^2 \mathrm{d}z\,,$$

Note that, as $K(x,\cdot)$ converges toward the Dirac mass at x, λ goes to 0, and ν_t goes to $\exp(-U+B\circ\xi)$ for some B then for all $f\in H$, $\mathcal{J}_t(f)$ converges towards $\mathcal{E}(f)$. On the other hand if we assume either that $\lambda>0$ or that $K>0$, then the problem of minimizing $\mathcal{J}_t$ is well-posed over $\mathcal{H}^1(\mathbb{T}^m,\mathbb{R})$ (with uniqueness of the minimizer up to an additive constant, so we can just chose the normalization $\int_{\mathbb{T}^m}A_t(x)\mathrm{d}x=0$). In fact, it is equivalent to the problem of minimizing

$$\hat{\mathcal{J}}_t(f) \;=\; \int_{\mathbb{T}^m}|F_t(z)-\nabla f(z)|^2\theta_t(z)\mathrm{d}z$$

with

$$\begin{aligned}\theta_t(z) &:= \frac{1}{\lambda+1}\left(\lambda+\int_{\mathbb{T}^d}K(x,z)\mathrm{d}\nu_t(x,y)\right)\\ F_t(z) &:= \frac{1}{(\lambda+1)\theta_\nu(z)}\int_{\mathbb{T}^m\times\mathbb{T}^m}\nabla_x U(x,y)K(x,z)\mathrm{d}\nu_t(x,y)\,.\end{aligned}$$

In other words, F_t is a regularized estimator of ∇A and A_t is an Helmholtz projection of F_t (which has no reason to be a gradient) in a weighted H^1 space (with the weight θ_t which is a smooth probability density).

To sum up, finally, the ABF algorithm is based on the process $(X_t,Y_t)_{t\geqslant 0}$ solving (2.7) where A_t is the unique minimizer in $\mathcal{H}^1(\mathbb{T}^m,\mathbb{R})$ with integral zero of $\mathcal{J}_t$. In particular, it is a self-interacting process since its drift depends on its occupation measure. There also exist mean-field versions [JLR10], with a particle system $(X_1,Y_1,\dots,X_N,Y_N)$ with a similar dynamics except that in $\mathcal{J}_t$ the occupation measure is replaced by the empirical measure of the system at time t. The first works on the longtime convergence of the ABF algorithm [LRS08] tackled the non-linear limit $N=+\infty$.

The works [24, 5, 1] are all concerned with the longtime convergence of the self-interacting ABF algorithm (i.e. the convergence of A_t to A and of ν_t). The specificities are the following:

In [24] I considered the algorithm based upon velocity jump processes instead of the overdamped Langevin diffusion. The definition of the adaptive bias is the same, based on the reweighted occupation measure, and then the force used in the jump mechanism is $\nabla(U-A_t\circ\xi)$. Other self-interacting velocity jump processes are studied.

In [5] with Virginie Ehrlacher and Tony Lelièvre, the process is the overdamped Langevin diffusion and the bias is defined from the reweighted occupation measure, but the minimization of $\mathcal{J}_t$ is solved through an approximation in tensor form, namely A_t is the sum of terms of the form $r_1(x_1)\times\cdots\times r_m(x_m)$. The motivation is to be able to consider larger values of m. Indeed, in the basic algorithm, the occupation measure and the bias are kept in memory on a discrete grid of dimension m, which prevents m to be larger than, say, 3 or 4. On the contrary, the memory necessary to record a tensor term is linear in m (and not exponential). We proved, at a fixed t, the convergence of the approximation scheme toward the minimizer of $\mathcal{J}_t$, and then the longtime convergence of the self-interacting ABF algorithm.

In [1] with Michel Benaïm and Charles-Edouard Bréhier we considered the case of the overdamped Langevin diffusion but with a bias defined by the *non-reweighted* case. This is more difficult for the following reason. As explained in Section 1.6.2, the study

relies on the ODE method, namely on a deterministic flow on probability measures. When the occupation measure is reweighted, this flow is simply

$$\dot{\nu} \;=\; \mu - \nu$$

with $\mu \propto \exp(-U)$, so it is perfectly clear that μ is the unique global attractor. On the contrary, in the non-reweighted case, the limit flow is

$$\dot{\nu} \;=\; \mu_\nu - \nu$$

with $\mu_\nu \propto \exp(-U + A_\nu)$ where A_ν is the minimizer of $\mathcal{J}_\nu$, defined like $\mathcal{J}_t$ except that ν_t is replaced by ν. We proved that, if the regularization is sufficiently small (i.e. if $K(x, \dot{})$ is sufficiently close in the $\mathcal{W}_2$ sense to δ_x and if λ is sufficiently small) then this flow still admits a unique global attractor. From this, we obtain the convergence of the algorithm.

2.6.1 *Some on-going projects and perspectives*

- With Tony Lelièvre and Lise Maurin we are currently finishing a work about the convergence of the mean-field non-linear ABF algorithm when the forces are not the gradient of some potential (i.e. ∇U is replaced by some vector field $\mathcal{F}$). A motivation comes from ab initio computations, where forces are obtained through some approximate scheme. Proving that, if $\mathcal{F}$ is close to ∇U, then the adaptive bias converges toward something close to the free energy associated to U is not completely obvious because an equilibrium of the non-linear process is given as the solution of some fixed-point problem. Moreover, in a non-gradient case, many explicit computations break down.

- With Lise Maurin, Louis Lagardère, Jérôme Hénin and Jean-Philip Piquemal, we are currently implementing some adaptive algorithms in the molecular dynamics software Tinker [LJL+17].

- In practice, molecular dynamics simulations are carried out with the kinetic Langevin diffusion. Theoretical results on adaptive algorithms are established for the over-damped process for simplicity. Nevertheless, adapting the proof to the kinetic case should not raise any particular difficulty in the case of a self-interacting process, since hypoelliptic and hypocoercive estimates are available for the (linear) Langevin diffusion. The mean-field case may be more troublesome as it would require non-linear hypocoercive entropic computations.

- The combination of MCMC and tensor approximation in [5] is a proof of concept, kept as simple as possible for the sake of clarity. As a consequence it is quite limited in practice. As discussed in this paper, there are many ways to improve its performances, and these variations should be implemented and studied. Moreover, as also discussed in the paper, the method to compute tensor approximation of a free energy from an MCMC sample in moderately high dimension can be used to conduct statistical studies on the reaction coordinates (sensitivity analysis, selection...). This opens many interesting perspectives.

Bibliography

Publications and preprints

[1] M. Benaïm, C.-E. Bréhier, and P. Monmarché. Analysis of an Adaptive Biasing Force method based on self-interacting dynamics. *arXiv e-prints*, page arXiv:1910.04428, October 2019.

[2] P. Cattiaux, A. Guillin, P. Monmarché, and P. Zhang. Entropic multipliers method for langevin diffusion and weighted log sobolev inequalities. *Journal of Functional Analysis*, 277(11):108288, 2019.

[3] A. Durmus, A. Guillin, and P. Monmarché. Piecewise Deterministic Markov Processes and their invariant measure. *arXiv e-prints*, page arXiv:1807.05421, Jul 2018.

[4] A. Durmus, A. Guillin, and P. Monmarché. Geometric ergodicity of the bouncy particle sampler. *To appear in Annals of Applied Probability*, 2020.

[5] V. Ehrlacher, T. Lelièvre, and P. Monmarché. Adaptive force biasing algorithms: new convergence results and tensor approximations of the bias. Hal preprint, December 2019.

[6] N. Fournier, P. Monmarché, and C. Tardif. Simulated Annealing In $\mathbf{R}^d$ With Slowly Growing Potentials. *arXiv e-prints*, page arXiv:1909.01570, September 2019.

[7] C.-E. Gauthier and P. Monmarché. Strongly self-interacting processes on the circle. *Stochastics*, 91(8):1249–1271, 2019.

[8] A. Guillin and P. Monmarché. Optimal linear drift for an hypoelliptic diffusion. *Electronic Communication of Probability*, 21, 2016.

[9] A. Guillin and P. Monmarché. Uniform long-time and propagation of chaos estimates for mean field kinetic particles in non-convex landscapes. *arXiv e-prints*, page arXiv:2003.00735, March 2020.

[10] L. Journel and P. Monmarché. Convergence of the Fleming-Viot algorithm: uniform in time estimates in a compact soft case. *arXiv e-prints*, page arXiv:1910.05060, October 2019.

[11] E. Löcherbach and P. Monmarché. Metastability for systems of interacting neurons. *arXiv e-prints*, page arXiv:2004.13353, April 2020.

[12] L. Miclo and P. Monmarché. Étude spectrale minutieuse de processus moins indécis que les autres. In *Séminaire de Probabilités XLV*, volume 2078 of *Lecture Notes in Math.*, pages 459–481. Springer, Cham, 2013.

[13] P. Monmarché. Hypocoercive relaxation to equilibrium for some kinetic models. *Kinet. Relat. Models*, 7(2):341–360, 2014.

[14] P. Monmarché. On $\mathcal{H}^1$ and entropic convergence for contractive PDMP. *Electronic Journal of Probability*, 20, December 2015.

[15] P. Monmarché. Piecewise deterministic simulated annealing. *ALEA Lat. Am. J. Probab. Math. Stat.*, 13(1):357–398, 2016.

[16] P. Monmarché. A note on Fisher Information hypocoercive decay for the linear Boltzmann equation. *arXiv e-prints*, page arXiv:1703.10504, March 2017.

[17] P. Monmarché. Long-time behaviour and propagation of chaos for mean field kinetic particles. *Stochastic Process. Appl.*, 127(6):1721–1737, 2017.

[18] P. Monmarché. Elementary coupling approach for non-linear perturbation of Markov processes with mean-field jump mechanims and related problems. *arXiv e-prints*, page arXiv:1809.10953, Sep 2018.

[19] P. Monmarché. Hypocoercivity in metastable settings and kinetic simulated annealing. *Probability Theory and Related Fields*, Jan 2018.

[20] P. Monmarché. Generalized Γ calculus and application to interacting particles on a graph. *Potential Analysis*, 50:439–466, 2019.

[21] P. Monmarché. Hypocoercivité L^2, inégalité de concentration, temps d'atteinte et fonctions de Lyapunov. *arXiv e-prints*, page arXiv:1911.01748, November 2019.

[22] P. Monmarché. Kinetic walks for sampling. *To appear in ALEA Lat. Am. J. Probab. Math. Stat.*, 2020.

[23] P. Monmarché, J. Weisman, L. Lagardère, and J.-P. Piquemal. Velocity jump processes : an alternative to multi-timestep methods for faster and accurate molecular dynamics simulations. *To appear in Journal of Chemical Physics*, page arXiv:2002.07109, February 2020.

[24] P. Monmarché. Weakly self-interacting piecewise deterministic bacterial chemotaxis. *Markov Process. Related Fields*, 23(4):609–659, 2017.

Exogenous bibliography

[ABC+00] C. Ané, S. Blachère, D. Chafaï, P Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer. *Sur les inégalités de Sobolev logarithmiques*. Panoramas et synthèses. Société mathématique de France, Paris, 2000.

[ABG+14] R. Azaïs, J.-B. Bardet, A. Génadot, N. Krell, and P.-A. Zitt. Piecewise deterministic Markov process—recent results. In *Journées MAS 2012*, volume 44 of *ESAIM Proc.*, pages 276–290. EDP Sci., Les Ulis, 2014.

[ADNR18] C. Andrieu, A. Durmus, N. Nüsken, and J. Roussel. Hypocoercivity of Piecewise Deterministic Markov Process-Monte Carlo. *arXiv e-prints*, page arXiv:1808.08592, August 2018.

[AE14] A. Arnold and J. Erb. Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift. *ArXiv e-prints*, September 2014.

[Bau17] F. Baudoin. Bakry-emery meet villani. *J. Funct. Anal.*, 273(7):2275–2291, 2017.

[BCF18] F. Bolley, D. Chafaï, and J. Fontbona. Dynamics of a planar coulomb gas. *Ann. Appl. Probab.*, 28(5):3152–3183, 10 2018.

[BCG08] D. Bakry, P. Cattiaux, and A. Guillin. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *J. Funct. Anal.*, 254(3):727–759, 2008.

[BCG+13] J.-B. Bardet, A. Christen, A. Guillin, F. Malrieu, and P.-A. Zitt. Total variation estimates for the TCP process. *Electron. J. Probab.*, 18:no. 10, 21, 2013.

[BdH15] A. Bovier and F. den Hollander. *Metastability: A Potential-Theoretic Approach*. Grundlehren dermathematischen Wissenschaften. Springer, Cham, 2015.

[BEGK04] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6(4):399–424, 2004.

[BFR19] J. Bierkens, P. Fearnhead, and G. Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320, 2019.

[BG17] M. Benaïm and C. Gauthier. Self-repelling diffusions on a riemannian manifold. *Probab. Theory Relat. Fields*, 169:63–104, 2017.

[BGH19] F. Baudoin, M. Gordina, and D.P. Herzog. Gamma calculus beyond villani and explicit convergence estimates for langevin dynamics with singular potentials, 2019.

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

[BLMZ12] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt. On the stability of planar randomly switched systems. *ArXiv e-prints, to appear to the Annals of Applied Probability*, April 2012.

[BLMZ15] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt. Qualitative properties of certain piecewise deterministic Markov processes. *Ann. Inst. Henri Poincaré Probab. Stat.*, 51(3):1040–1075, 2015.

[BLR02] M. Benaïm, M. Ledoux, and O. Raimond. Self-interacting diffusions. *Probab. Theory Related Fields*, 122(1):1–41, 2002.

[BVD18] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.*, 113(522):855–867, 2018.

[CG14] P. Cattiaux and A. Guillin. Semi log-concave Markov diffusions. In *Séminaire de Probabilités XLVI*, volume 2123 of *Lecture Notes in Math.*, pages 231–292. Springer, Cham, 2014.

[CGZ13] P. Cattiaux, A. Guillin, and P.-A. Zitt. Poincaré inequalities and hitting times. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(1):95–118, 2013.

[Dav93] M.H.A Davis. *Markov Models and Optimization*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1993.

[DBD19] G. Deligiannidis, A. Bouchard-Côté, and A. Doucet. Exponential ergodicity of the bouncy particle sampler. *Ann. Statist.*, 47(3):1268–1287, 2019.

[DHN00] P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.*, 10(3):726–752, 2000.

[DM17] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017.

[DMS15] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Trans. Amer. Math. Soc.*, 367(6):3807–3828, 2015.

[EK86] S. N. Ethier and T. G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. Characterization and convergence.

[ELB+95] U. Essmann, Perera L., M.L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys*, 103(19):8577–8593, 1995.

[Eva17] J. Evans. Hypocoercivity in Phi-entropy for the Linear Boltzmann Equation on the Torus. *ArXiv e-prints*, February 2017.

[EZ19] A. Eberle and R. Zimmer. Sticky couplings of multidimensional diffusions with different drifts. *Ann. Inst. H. Poincaré Probab. Statist.*, 55(4):2370–2394, 11 2019.

[FT20] N. Fournier and C. Tardif. On The Simulated Annealing In $\mathbf{R}^d$. *arXiv e-prints*, page arXiv:2003.06360, March 2020.

[GLWZ19] A. Guillin, W. Liu, L. Wu, and C. Zhang. Uniform Poincar{é} and logarithmic Sobolev inequalities for mean field particles systems. *arXiv e-prints*, page arXiv:1909.07051, September 2019.

[GM13] S. Gadat and L. Miclo. Spectral decompositions and $\mathbb{L}^2$-operator norms of toy hypocoercive semi-groups. *Kinet. Relat. Models*, 6(2):317–372, 2013.

[GSB+20] P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. Chodera, A. R. Dinner, A. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora, and T. Leliévre. Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *arXiv e-prints*, page arXiv:2004.06950, April 2020.

[HHMS93] C.R. Hwang, S.Y. Hwang-Ma, and S.J. Sheu. Accelerating gaussian diffusions. *Ann. Appl. Probab.*, 3:897–913, 1993.

[HKS89] R. A. Holley, S. Kusuoka, and D. W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.*, 83(2):333–347, 1989.

[HM11] M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63 of *Progr. Probab.*, pages 109–117. Birkhäuser/Springer Basel AG, Basel, 2011.

[JLR10] B. Jourdain, T. Lelièvre, and R. Roux. Existence, uniqueness and convergence of a particle approximation for the adaptive biasing force process. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(5):831–865, 2010.

[Kal02] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[KB96] X. Kong and C. L. Brooks. $\lambda-$ dynamics: A new approach to free energy calculations. *The Journal of Chemical Physics*, 105(6):2414–2423, 1996.

[Lel13] T. Lelièvre. Two mathematical tools to analyze metastable stochastic processes. In *Numerical mathematics and advanced applications 2011*, pages 791–810. Springer, Heidelberg, 2013.

[LJL+17] L. Lagardère, L.-H. Jolly, F. Lipparini, F. Aviat, B. Stamm, Z. F Jing, M. Harger, H. Torabifard, G. A. Cisneros, M. Schnieders, N. Gresh, Y. Maday, P. Ren, J. W Ponder, and J.-P. Piquemal. Tinker-HP: a Massively Parallel Molecular Dynamics Package for Multiscale Simulations of Large Complex Systems with Advanced Point Dipole Polarizable Force Fields. *Chemical Science* , November 2017.

[LNP13] T. Lelièvre, F. Nier, and G. A. Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *J. Stat. Phys.*, 152(2):237–274, 2013.

[LO17] A. Levitt and C. Ortner. Convergence and cycling in Walker-type saddle search algorithms. *SIAM J. Numer. Anal.*, 55(5):2204–2227, 2017.

[LRS07] T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy profiles with parallel adaptive dynamics. 4 pages, 1 Figure, 2007.

[LRS08] T. Lelièvre, M. Rousset, and G. Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155–1181, 2008.

[LRS10] T. Lelièvre, M. Rousset, and G. Stoltz. *Free energy computation: a mathematical perspective.* Imperial College Press, 2010.

[Mal01] F. Malrieu. Logarithmic Sobolev inequalities for some nonlinear PDE's. *Stochastic Process. Appl.*, 95(1):109–132, 2001.

[Mic92] L. Miclo. Recuit simulé sur R^n. Étude de l'évolution de l'énergie libre. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(2):235–266, 1992.

[MS14] G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 2014.

[Nea04] R. M. Neal. Improving Asymptotic Variance of MCMC Estimators: Non-reversible Chains are Better. *arXiv Mathematics e-prints*, page math/0407281, Jul 2004.

[OV00] F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361 – 400, 2000.

[PdW12] E. A. J. F. Peters and G. de With. Rejection-free monte carlo sampling for general potentials. *Phys. Rev. E 85, 026703*, 2012.

[Tuc10] M.E. Tuckerman. *Statistical mechanics theory and molecular simulation*. Oxford University Press, 2010.

[VBDD17] P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. Piecewise-Deterministic Markov Chain Monte Carlo. *arXiv e-prints*, page arXiv:1707.05296, Jul 2017.

[Vil09a] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950):iv+141, 2009.

[Vil09b] C. Villani. *Optimal transport, old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009.

[Zit08] P-A Zitt. Annealing diffusions in a potential function with a slow growth. *Stochastic Process. Appl.*, 118(1):76–119, 2008.