Exact targeting of Gibbs distributions using velocity-jump processes

Pierre Monmarché, Mathias Rousset, Pierre-André Zitt

▶ **To cite this version:**

Pierre Monmarché, Mathias Rousset, Pierre-André Zitt. Exact targeting of Gibbs distributions using velocity-jump processes. Stochastics and Partial Differential Equations: Analysis and Computations, 2022, 10.1007/s40072-022-00247-9 . hal-02916073v2

HAL Id: hal-02916073
https://hal.science/hal-02916073v2

Submitted on 13 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Exact targeting of Gibbs distributions using velocity-jump processes.

P. Monmarché[*], M. Rousset[†] and P.A. Zitt[‡]

August 2020

Abstract

This work introduces and studies a new family of velocity jump Markov processes directly amenable to exact simulation with the following two properties: i) trajectories converge in law, when a time-step parameter vanishes, towards a given Langevin or Hamiltonian dynamics; ii) the stationary distribution of the process is always exactly given by the product of a Gaussian (for velocities) by any target log-density. The simulation itself, in addition to the computability of the gradient of the log-density, depends on the knowledge of appropriate explicit upper bounds on lower order derivatives of this log-density. The process does not exhibit any velocity reflections (jumps maximum size can be controlled) and is suitable for the 'factorization method'. We provide rigorous mathematical proofs of the convergence towards Hamiltonian/Langevin dynamics when the time step vanishes, and of the exponentially fast convergence towards the target distribution when a suitable noise on velocities is present. Numerical implementation is detailed and illustrated.

Contents

[*]LJLL & LCT – Laboratoire Jacques-Louis Lions and Laboratoire de Chimie Théorique, Sorbonne Université

[†]Inria & IRMAR – Institut de Recherche en Mathématiques de Rennes, Univ Rennes

[‡]LAMA, Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, F-77454 Marne-la-Vallée, France

1 Introduction

A kinetic process is a Markov process $(X_t, V_t)_{t\geqslant 0}$, where $X_t \in \mathbb{R}^d$ and $V_t \in \mathbb{R}^d$ are respectively called the position and velocity of the process, such that $X_t = X_0 + \int_0^t V_s \mathrm{d}s$ for all $t \geqslant 0$. In addition to modelling a variety of phenomena, these processes can be used as time continuous Markov Chain Monte Carlo algorithms. In this case, given a target probability distribution ν on $\mathbb{R}^d$, the idea is to construct a kinetic process that is ergodic with respect to some probability measure π on $\mathbb{R}^{2d}$ whose first marginal is the target distribution ν. This program generalizes the usual construction of a ν-ergodic process $(X_t)_{t\geqslant 0}$ on $\mathbb{R}^d$. When this ergodicity holds, for observables f that only depend on the position, the empirical estimation $t^{-1}\int_0^t f(X_s)\mathrm{d}s$ still converges in large times towards $\nu(f)$. This idea traces back to the Molecular Dynamics (MD) of Alder and Wainwright [1], based on the Hamiltonian dynamics, introduced shortly after the seminal Metropolis algorithm. Beyond physical applications and motivations — Hamiltonian-based processes simulate the real physical dynamics, an algorithmic motivation is that kinetic processes have a ballistic, rather than diffusive, behaviour: their inertia reduces backtracking, which improves the exploration of the configuration space, by comparison with reversible processes such as Metropolis-Hastings random walk or usual elliptic diffusions.

Langevin diffusion and Hamiltonian Monte-Carlo (HMC) are classical kinetic processes used for sampling purposes. In the last decade, another class of velocity jump samplers has emerged, first obtained as scaling limits of rejection-free lifted Markov chains [20, 4, 16]. In these new samplers, the velocity is piecewise constant and is updated at random times; in particular, the process belongs to the family of piecewise deterministic Markov process (PDMP). The law of these so-called jump (or collision, or event) times is chosen in such a way that the invariant distribution of the process is the target π. An appealing feature of these processes is that they can be implemented in continuous time, since only the value of the process at its jump time is needed, and no supplementary time discretization is required. In particular, the equilibrium of the process effectively implemented is the correct one, which

is usually not the case for discretized diffusions. In HMC-like methods, a Metropolis step is added which corrects for the time discretization; however the introduced rejection requires a velocity reflection which destroys the ballistic dynamics and impairs the efficiency of the algorithm. Another interesting point is that, as detailed in Section 2.1 (see also [17, 18]), different parts of the log-density of ν may be treated at different time scales through a factorization of the target measure, thus reducing the overall computational complexity of the algorithm. This property is somewhat analogous to the deterministic multi-time-step integration methods [22, 14] but, again, without their statistical bias.

If the user is only interested in computing static quantities, that is, integrals of some observables with respect to ν, then any ν-ergodic process, or π-ergodic kinetic process with marginal ν, is theoretically usable, even if some may perform better than others for a finite computational budget. The question is a bit different when the aim is to compute dynamical quantities (diffusion constants, escape rates, quasi-stationnary distributions...) for a given, particular kinetic process, typically the Hamiltonian or Langevin dynamics. Indeed, though they have the same equilibrium, different kinetic processes may have completely different dynamical properties. For instance, bouncy-type samplers, HMC, or other Metropolized schemes based on Langevin diffusions [19] all feature occasional reflections of the velocity; such discontinuities never happen in Hamiltonian or Langevin dynamics.

Errors in the computation of dynamical quantities naturally occur when the computation is done by discretizing in time the continuous time dynamics of interest: a Langevin process discretized with a Verlet-like scheme for example, does not have exactly the same dynamical properties as the reference continuous-time process. In these cases however, there is a parameter, namely the discretization time-step ε, which may be tuned to obtain a trade-off between dynamical precision and cost: smaller ε lead to a better precision on the dynamical properties, at the cost of longer computations —- simulating a trajectory for a given fixed time T typically requires T/ε computations of the gradient of the log-density of ν. Such a precision/computation cost tradeoff does not currently exist for bouncy-type kinetic samplers.

The main contribution of the present work is the design of a new family of velocity jump processes with two interesting properties. Firstly, similarly to discretized Langevin or Hamiltonian schemes, the process does not suffer from regular velocity reflections and moreover converges when a time-step parameter ε vanishes towards a given Langevin of Hamiltonian dynamics. Secondly, similarly to bouncy-type samplers, it is a kinetic MCMC sampler with exact target distribution and suitable for the factorization method.

We provide a rigorous mathematical proof of two related properties. The first one is the convergence in distribution of trajectories of the considered process towards Hamiltonian dynamics, when the time-step parameter ε vanishes. This result relies on classical characterization techniques based on martingale problems. The second property we establish is the exponentially fast convergence of the process time marginal distributions towards the exact target distribution, in an $\mathbb{L}^2$ sense. This result relies on a hypocoercivity analysis based on a Lyapunov function in the form of a well-chosen modified $\mathbb{L}^2$-norm, in the spirit of [10].

The improvement from Hamiltonian integrators and randomized variants is thus that the static properties are unbiased and, maybe more importantly in this context, the factorization

method is still available. The price to pay is the loss of geometric properties as symplecticity. The improvement from bouncy-type samplers is that the proposed method introduces a time-step parameter ε that enables to interpolate the former with Hamiltonian/Langevin dynamics.

Finally, we remark that several recent works [6, 9] in Bayesian statistics have argued that samplers based on Hamiltonian dynamics or Langevin diffusion have good convergence properties, from the fact the continuous-time limit process has dimension-free convergence rate for smooth and concave potentials, and then controlling the distance between this limit and the effective algorithm. In this context, our family of processes may provide a way to keep the dimension-free convergence rate while suppressing the bias (although the dimension should still intervene in the complexity of the algorithm). As said above, we provide explicit $\mathbb{L}^2$ convergence rates in the spirit of [10] and [2] under general assumptions.

The article is organized as follows. The general framework of kinetic samplers and velocity jump processes is introduced in Section 2. Section 3 contains the definition of the new family of processes and the proof of convergence toward the Hamiltonian dynamics (Theorem 3.6). Exponential convergence toward equilibrium with explicit rates is established in Section 4 through Hypocoercivity arguments (Theorem 4.3). The effective simulation of the processes is discussed in Section 5, and numerical experiments are provided in Section 6. Finally, the proof of a general result for the convergence of Markov processes, Theorem 7.1, used in the proof of Theorem 3.6, is postponed to Section 7.

2 Kinetic samplers

2.1 General setting

Let ν and γ be two probability laws on $\mathbb{R}^d$, where ν admits a density with respect to the Lebesgue measure proportional to $\exp(-U)$, for some function $U \in \mathcal{C}^1(\mathbb{R}^d)$ — the log-density. We are interested in kinetic processes for which the Gibbs distribution $\pi = \nu \otimes \gamma$, namely

$$\pi(\mathrm{d}x\mathrm{d}v) \propto \exp\left(-U(x)\right) \mathrm{d}x\gamma(\mathrm{d}v)\,, \tag{2.1}$$

is invariant.

Remark 2.1 (Marginal in the velocities)**.** There are several possible choices for γ. Usual ones are Gaussian distributions and the uniform measure on a sphere or on a discrete set of velocities.

Consider a Markov process on $\mathbb{R}^d \times \mathbb{R}^d$ with — formal — generator $\mathcal{L}$, decomposed as

$$\mathcal{L}\varphi(x,v) = \mathcal{T}\varphi(x,v) + \mathcal{F}\varphi(x,v) + \mathcal{D}\varphi(x,v) \tag{2.2}$$

for smooth, compactly supported test functions $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$, where:

- the *transport* part $\mathcal{T}\varphi(x,v) = v \cdot \nabla_x \varphi(x,v)$ is the free-flight transport operator, and is the only part that acts on the position variable in the sense that $\mathcal{D}\varphi = \mathcal{F}\varphi = 0$ if $\varphi(x,v) = g(x)$ for some function g. In terms of trajectories, this ensures that $X_t = \int_0^t V_s ds$.

- the *dissipative* part $\mathcal{D}$ is a Markov generator that acts on the velocity variables and leaves γ invariant:

$$\forall \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d}), \forall x \in \mathbb{R}^d, \qquad \int_{\mathbb{R}^d} \mathcal{D}\varphi(x,v)\gamma(\mathrm{d}v) = 0\,. \tag{2.3}$$

- the *force* part $\mathcal{F}$ acts on velocity variables and is such that for all $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$,

$$\int_{\mathbb{R}^{2d}} \mathcal{F}\varphi(x,v)\pi(\mathrm{d}x\mathrm{d}v) = -\int_{\mathbb{R}^{2d}} \varphi(x,v)\left(v\cdot\nabla U(x)\right)\pi(\mathrm{d}x\mathrm{d}v)\,. \tag{2.4}$$

Integrating by parts, we see that this last condition means that $\int \mathcal{F}\varphi\pi = -\int \mathcal{T}\varphi\pi$ for all $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$. As a consequence (2.3) together with (2.4) imply that $\pi(\mathcal{L}\varphi) = 0$ for all $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$. If $\mathcal{C}_c^\infty(\mathbb{R}^{2d})$ is a core for $\mathcal{L}$, which is usually true and can be proven through regularization and truncation arguments [12], then this implies that π is invariant for $\mathcal{L}$.

Many operators satisfy the requirements for the dissipative part $\mathcal{D}$; let us mention three usual choices:

- Friction/Dissipation:

$$\mathcal{D}\varphi(x,v) = -v\cdot\nabla_v\varphi(x,v) + \frac{\sigma^2}{2}\Delta_v\varphi(x,v)\,, \tag{2.5}$$

  for some $\sigma > 0$. In this case γ is the centered normal distribution with variance σ^2, and $\mathcal{D}$ is the generator of an Ornstein-Uhlenbeck process acting on velocities.

- Velocity refreshment:

$$\mathcal{D}\varphi(x,v) = \int_{\mathbb{R}^d} \left(\varphi(x,w) - \varphi(x,v)\right)\gamma(\mathrm{d}w)\,. \tag{2.6}$$

  In terms of trajectories this corresponds to resampling the velocity at rate 1, according to the equilibrium measure γ.

- Partial refreshment:

$$\mathcal{D}\varphi(x,v) = \int_{\mathbb{R}^d} \left(\varphi(x, pv + \sqrt{1-p^2}w) - \varphi(x,v)\right)\gamma(\mathrm{d}w)\,, \tag{2.7}$$

  for some $p \in [0,1)$ if γ is a normal distribution. This corresponds to changing the velocity at random times, using the transition kernel of the Ornstein-Uhlenbeck process, and can be seen (up to a rescaling in time) as an interpolation between the previous two exemples.

In general, note that (2.3) implies that for any probability law $\tilde{\nu}$ on $\mathbb{R}^d$, $\tilde{\nu}\otimes\gamma$ is invariant for $\mathcal{D}$. Moreover, if (2.3) holds, then it also holds for the generator $\mathcal{D}_2\varphi(x,v) = \eta(x)\mathcal{D}\varphi(x,v)$ for any positive function η on $\mathbb{R}^d$. For instance, when $\mathcal{D}$ models the interaction of the system with an external heat bath, there may be no coupling with the heat bath in the interior of

some domain, i.e. $\eta(x) = 0$ for x in the domain, and $\eta(x) > 0$ outside. Similarly, if $\mathcal{D}_1$ and $\mathcal{D}_2$ both satisfy (2.3), then $\mathcal{D}_1 + \mathcal{D}_2$ does too.

Let us now discuss in more detail the *force* part $\mathcal{F}$. The most classical choice here is the deterministic drift operator

$$\mathcal{F}\varphi(x,v) = -\sigma^2 \nabla U(x) \cdot \nabla_v \varphi(x,v)$$

which satisfies (2.4) if γ is the centered normal distribution with variance σ^2. With this choice, then $\mathcal{L}$ is the generator of the Hamiltonian dynamics if $\mathcal{D} = 0$, of the Langevin diffusion if $\mathcal{D}$ is given by (2.5), or of the HMC if $\mathcal{D}$ is given by (2.6).

The factorization (or splitting) method relies on the following remark. Suppose that $\nabla U(x) = \sum_{i=1}^N \xi_i(x)$ for some vector fields ξ_i on $\mathbb{R}^d$, $i = 1..N$, and that we have N operators $\mathcal{F}_1, \dots, \mathcal{F}_N$ such that for all $i \in [\![1, N]\!]$,

$$\int_{\mathbb{R}^{2d}} \mathcal{F}_i \varphi(x,v) \pi(\mathrm{d}x\mathrm{d}v) = -\int_{\mathbb{R}^{2d}} \varphi(x,v) \left(v \cdot \xi_i(x)\right) \pi(\mathrm{d}x\mathrm{d}v)\,. \tag{2.8}$$

Then $\mathcal{F} = \sum_{i=1}^N \mathcal{F}_i$ satisfies (2.4). If $\xi_i = \nabla U_i$ for all $i \in [\![1, N]\!]$ for some $U_i \in \mathcal{C}^1(\mathbb{R}^d)$, then the decomposition of $\mathcal{F}$ is based on the factorization

$$\nu(\mathrm{d}x) \propto \prod_{i=1}^N e^{-U_i(x)} \mathrm{d}x\,.$$

Note that in that case it is not necessary that $\exp(-U_i)$ has finite mass. More generally ξ_i is not required to be a gradient. For instance, if $(e_i)_{i \in [\![1,d]\!]}$ is the canonical basis of $\mathbb{R}^d$, then $\xi_i(x) = (\nabla U(x) \cdot e_i) e_i$ gives a decomposition of the forces ∇U as a sum of possibly non-gradient forces.

Through such a decomposition, different forces may be treated with different dynamics. For instance, as we will see in Section 5, jump mechanisms are easily simulated if ∇U is bounded, or Lipschitz, with a known bound, which is not always the case. On the other hand, drift mechanisms suffer the problem of discretization, and a possibly higher computational cost since the forces have to be computed at each time-step. If ∇U can be decomposed in long-range forces which are expansive to compute but easily bounded, and short-range forces which are possibly singular but cheap to compute, then it is natural to treat the first ones with jump processes and the second ones with drift processes [18]. Similarly, if different forces have different time-scales, then instead of using different time-steps in a numerical integration of a drift mechanism, it is possible to use different jump mechanisms as detailed in Section 3.3.2.

In the rest of the paper, unless otherwise specified, we will only consider the non-factorized condition (2.4). Indeed, from an operator $\mathcal{F}$ that satisfies (2.4) (or more precisely (2.14) below) and whose definition only involves U through ∇U, it is then easy to obtain an operator $\mathcal{F}_i$ that satisfies (2.8) by replacing ∇U by ξ_i everywhere in the definition of $\mathcal{F}$ (see Section 3.3).

2.2 Velocity jumps

Let $\lambda(x,v)$ be a non-negative function, and for each x, let $k(x,v;\mathrm{d}v')$ be a Markov kernel. We denote by q the non-normalized kernel $q(x,v;\mathrm{d}v') = \lambda(x,v)k(x,v;\mathrm{d}v')$. From now on, we

consider the case where the jumps on the velocity are given by such a kernel:

$$\mathcal{F}\varphi(x,v) = \int_{v'\in\mathbb{R}^d} (\varphi(x,v') - \varphi(x,v))\, q(x,v;\mathrm{d}v'). \tag{2.9}$$

In this case, the dynamics of a Markov process with generator $\mathcal{T}+\mathcal{F}$ is the following: the x variable evolves deterministically at velocity v; the velocity is piecewise constant, and jumps at a rate $\lambda(x,v)$ to a new velocity v' sampled according to $k(x,v;\mathrm{d}v')$. The number of jumps may go to infinity at finite time, unless for instance λ is bounded. This kind of process is known as a *velocity jump process*.

In a way that is similar to the classical Metropolis algorithm, the jump mechanism q will be constructed by choosing a nice *proposal kernel* q_0, and then *modifying it* to take the log-density U into account, yielding a *corrected kernel* q. We start by stating two conditions that our proposal kernel should satisfy.

Definition 2.2 (Conditions for the proposal kernel)**.** *A non-negative kernel $q_0(x,v;\mathrm{d}v')$ is* reversible *with respect to γ if*

$$q_0(x,v;\mathrm{d}v')\gamma(\mathrm{d}v) = q_0(x,v';\mathrm{d}v)\gamma(\mathrm{d}v') \qquad \forall x\in\mathbb{R}^d. \tag{R}$$

It satisfies the average condition (A) *if moreover $\int_{v'} 1+|v'|\, q_0(x,v;\mathrm{d}v') < +\infty$ for all $(x,v)\in\mathbb{R}^{2d}$ and*

$$\nabla U(x)\cdot\int_{v'\in\mathbb{R}^d}\frac{1}{2}(v-v')q_0(x,v;\mathrm{d}v') = \nabla U(x)\cdot v, \qquad \mathrm{d}x\gamma(\mathrm{d}v)\text{-}a.e.. \tag{A}$$

Note that (A) may be rewritten in terms of the intensity $\lambda_0(x,v) = \int q_0(x,v,\mathrm{d}v')$ and the normalized kernel $k_0 = q_0/\lambda_0$ as

$$\left(\int v' k_0(x,v;\mathrm{d}v')\right)\cdot\nabla U(x) = \left(1-\frac{2}{\lambda_0(x,v)}\right) v\cdot\nabla U(x). \tag{2.10}$$

Let $\psi:\mathbb{R}\to\mathbb{R}_+$ be a measurable function such that

$$\psi(s)-\psi(-s) = s, \quad \forall s\in\mathbb{R}. \tag{2.11}$$

The basic choice for ψ is $\psi(s)=(s)_+$, but as remarked in [3] there are other possibilites, like $\psi(s) = a\ln(e^{s/a}+1)$ for $a>0$. For any proposal kernel q_0, let us define a corrected kernel by:

$$q(x,v,\mathrm{d}v') = \psi\left(\frac{1}{2}\nabla U(x)\cdot(v-v')\right) q_0(x,v;\mathrm{d}v'). \tag{2.12}$$

Our work is based on the following remark.

Lemma 2.3. *Assume that $q_0(x,v;\mathrm{d}v')$ is reversible with respect to γ, in the sense of condition* (R)*. Let q be the corrected non-normalized kernel defined by* (2.12)*, where the function ψ satisfies* (2.11)*. The corresponding operator $\mathcal{F}$ given by* (2.9) *satisfies the condition* (2.4) *if and only if the average condition* (A) *holds true; if this holds then the measure π is invariant for the process.*

Proof. Let $a(x,v,v') = \frac{1}{2}(\nabla U \cdot v - v')$. For any φ,

$$\int \mathcal{F}\varphi(x,v)\mathrm{d}\pi(\mathrm{d}x\mathrm{d}v) = \int \varphi(x,v')\psi(a(x,v,v'))q_0(x,v;\mathrm{d}v')\pi(\mathrm{d}x\mathrm{d}v) - \int \varphi(x,v)\psi(a(x,v,v'))q_0(x,v;\mathrm{d}v')\pi(\mathrm{d}x\mathrm{d}v)\,.$$

In the first integral, use the reversibility assumption and interchange the variables v and v' to get:

$$\begin{aligned}\int \mathcal{F}\varphi(x,v)\mathrm{d}\pi(\mathrm{d}x\mathrm{d}v) &= \int \varphi(x,v)\psi(a(x,v',v))q_0(x,v;\mathrm{d}v')\pi(\mathrm{d}x\mathrm{d}v) \\ &\quad - \int \varphi(x,v)\psi(a(x,v,v'))q_0(x,v;\mathrm{d}v')\pi(\mathrm{d}x\mathrm{d}v) \\ &= \int \left[\int \left(\psi(a(x,v',v)) - \psi(a(x,v,v'))\right) q_0(x,v;\mathrm{d}v')\right] \varphi(x,v)\pi(\mathrm{d}x\mathrm{d}v) \\ &= -\int \left[\int a(x,v,v')q_0(x,v;\mathrm{d}v')\right] \varphi(x,v)\pi(\mathrm{d}x\mathrm{d}v).\end{aligned}$$

As a consequence, the condition (2.4) is met if and only if the term between brackets is almost everywhere equal to $v \cdot \nabla U(x)$, which is exactly the averaging condition (A). □

Remark 2.4. In the case where the corrected kernel is constructed with the function $\psi(s) = (s)_+$, at each jump, the scalar product of the velocity with $-\nabla U$ increases almost surely. In that sense, there is "minimal noise" in the tangential part ∇U. The condition can be relaxed by setting:

$$q(x,v;\mathrm{d}v') = \left[\psi\left(\frac{1}{2}\nabla U(x)\cdot(v-v')\right) + g(x)\right] q_0(x,v;\mathrm{d}v') \tag{2.13}$$

for any non-negative function g on $\mathbb{R}^d$. One checks easily that the averaging condition (A) is unchanged. The process then performs jumps more often, but they are less constrained to be aligned with $-\nabla U$. In fact, if q_0 is reversible for γ, then the kernel $\tilde{q}(x,v;\mathrm{d}v') := g(x)q_0(x,v;\mathrm{d}v')$ leaves invariant $\tilde{\nu}\otimes\gamma$ for all law $\tilde{\nu}$ on $\mathbb{R}^d$, and thus $\tilde{q}$ can be incorporated in the dissipative part $\mathcal{D}$ of the generator. For this reason, in the rest of the paper we only consider the case $g = 0$.

Remark 2.5. In the proof of Lemma 2.3 the integration with respect to the variable x plays no role, so that in fact if q_0 satisfies the conditions (R) and (A) then for all $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$,

$$\int_{\mathbb{R}^d} \mathcal{F}\varphi(x,v)\gamma(\mathrm{d}v) = -\int_{\mathbb{R}^d} \varphi(x,v)\,(v\cdot\nabla U(x))\,\gamma(\mathrm{d}v) \qquad \mathrm{d}x\text{-a.e..} \tag{2.14}$$

2.3 Particular known cases

We now show how special forms of q_0 lead to various known sampling algorithms.

Theorem 2.6 (Zig Zag process)**.** *Let γ be the uniform measure on the finite set $\{-1,1\}^d$. For $v \in \{-1,1\}^d$, let $q_0(x,v;\mathrm{d}v') = \sum_{w:w\sim v}\delta_w(\mathrm{d}v')$, where $w \sim v$ means that v and w are neighbours on the discrete cube, that is, they differ by one coordinate.*

Then q_0 is reversible with respect to γ and satisfies the average condition (A); the corresponding process is the zig-zag process*.*

Proof. The reversibility is clear. To check the average condition, remark that if w and v differ only by the i^{th} coordinate, then $(v-w)/2 = v_i e_i$ where e_i is the i^{th} basis vector. Therefore

$$\nabla U(x) \cdot \int_{v' \in \mathbb{R}^d} \frac{1}{2}(v - v') q_0(x, v; \mathrm{d}v') = \sum_{i=1}^{d} \nabla U(x) \cdot v_i e_i = \nabla U(x) \cdot v.$$

With $\psi(s) = (s)_+$, the corrected kernel is given by

$$q(x, v, \mathrm{d}v') = \frac{1}{2} \left(\nabla U(x) \cdot (v - v')\right)_+ q_0(x, v; \mathrm{d}v') = \sum_{i=1}^{d} \left(\nabla U(x) \cdot v_i e_i\right)_+ \delta_{v - 2v_i e_i}$$

which is exactly the zig-zag jump kernel. □

Theorem 2.7 (Bouncy particle)**.** *Let γ be the uniform measure on a sphere. For v on the sphere, let q be the degenerate kernel $q(x, v; \mathrm{d}v') = \delta_{R(x)v}(\mathrm{d}v')$ where $R(x)$ is the symmetry with respect to the orthogonal of ∇U, that is,*

$$R(x)v = v - 2\mathbb{1}_{\{\nabla U \neq 0\}} \frac{v \cdot \nabla U(x)}{|\nabla U(x)|^2} \nabla U(x).$$

Then q is reversible with respect to γ and satisfies the average condition (A); the corresponding process is the bouncy particle sampler*.*

Proof. Once more, the reversibility is clear. The interesting thing to notice here is that

$$\frac{1}{2} \nabla U(x) \cdot (v - R(x)v) = \nabla U(x) \cdot v.$$

Therefore

$$\nabla U(x) \cdot \int_{v' \in \mathbb{R}^d} \frac{1}{2}(v - v') q_0(x, v; \mathrm{d}v') = \nabla U(x) \cdot v.$$

The corrected kernel with $\psi(s) = (s)_+$ is given by $q(x, v, \mathrm{d}v') = \left(\nabla U(x) \cdot v\right)_+ \delta_{R(x)v}(\mathrm{d}v')$, and we recover the bouncy particle sampler. □

3 The Gaussian case

3.1 The process

In this section we consider the particular case where the normalized proposal kernel $k_0(x, v; \mathrm{d}v')$ and the velocity distribution γ are Gaussian. In fact, up to a change of variables, we assume without loss of generality that γ is the standard Gaussian distribution with mean 0 and variance Id.

Given the particular role of the direction ∇U, it is natural to use the following orthogonal decomposition of the (tangent) space at x. Denote

$$T(x) = \nabla U(x)/|\nabla U(x)|$$

if $\nabla U(x) \neq 0$ and $T(x) = 0$ otherwise. For any $w \in \mathbb{R}^d$, we write $w = w_T + w_O$ where $w_T = (w \cdot T(x))T(x)$ is the projection of w on $\mathrm{Vect}(T(x))$ and w_O is orthogonal to $T(x)$. With this notation, let $k_0(x, v; \cdot)$ be the distribution of the Gaussian random variable V', defined by its decomposition $V' = V'_T + V'_O$:

$$\begin{aligned} V'_T &= \rho_T(x) v_T + \sqrt{1 - \rho_T^2(x)} G_T \\ V'_O &= \rho_O(x) v_O + \sqrt{1 - \rho_O^2(x)} G_O \end{aligned} \tag{3.1}$$

where $\rho_T(x)$, $\rho_O(x)$ are scalars in $[-1, 1]$, and $G = G_T + G_O$ is a d-dimensional standard unit Gaussian. Recall that $k_0(x, v; \cdot) = \mathrm{Law}(V')$ is (up to an intensity λ_0, see below) the proposal kernel for jumps in the velocity. One may therefore interpret the parameters ρ_T and ρ_O as follows:

- the sign of ρ_T encodes whether or not there is a "bounce", that is, a reflection of the component of the velocity that is tangent to the gradient of the log-density;
- $|\rho_T|$ encodes the strength of the memory for this tangential component: if $|\rho_T| = 1$, the memory is perfect (the new tangential component being either equal to the old one or to its opposite); on the contrary, $\rho_T(x) = 0$ means a full resampling without memory (which is called forward event-chain algorithm in [15]);
- similarly $|\rho_O|$ and the sign of ρ_O encodes respectively the balance between full memory and full resampling, and whether or not the orthogonal component of the velocity "bounces".

It is then easy to remark:

Lemma 3.1. *Let $q_0(x, v; \mathrm{d}v') = \lambda_0(x, v) k_0(x, v; \mathrm{d}v')$ where k_0 is defined above by (3.1). For any $x, v \in \mathbb{R}^d$, the average condition (A) (equivalently (2.10)) holds if*

$$\lambda_0(x, v) = \frac{2}{1 - \rho_T(x)};$$

and this latter condition is necessary when $v \cdot \nabla U(x) \neq 0$. Moreover, if λ_0 does not depend on v, then q_0 is reversible with respect to γ (condition (R)).

Proof. To check the second form (2.10) of the average condition, we compute for $x, v \in \mathbb{R}^d$

$$\begin{aligned} \nabla U(x) \cdot \left(\int v' k_0(x, v; \mathrm{d}v') \right) &= \nabla U(x) \cdot (\rho_T(x) v_T + \rho_O(x) v_0) \\ &= \rho_T(x) \nabla U(x) \cdot v. \end{aligned}$$

Then, if $v \cdot \nabla U(x) \neq 0$, (2.10) holds iff $1 - 2/\lambda_0(x, v) = \rho_T(x)$.

Now, if λ_0 does not depend on v, the reversibility of q_0 is a consequence of the reversibility of k_0. Remark that the (density of the) kernel k_0 admits a decomposition $k_0(x,v;v') = k_0^T(x,v_T;v'_T)k_0^O(x,v_O;v'_O)$, and similarly $\gamma(v) = \gamma(v_T)\gamma(v_O)$, with

$$\begin{aligned} k_0^T(x,v_T;v'_T)\gamma(v_T) &= \frac{1}{2\pi\sqrt{1-\rho_T^2(x)}}\exp\left(-\frac{|v'_T-\rho(x)v_T|^2}{2(1-\rho_T^2(x))}-\frac{|v_T|^2}{2}\right) \\ &= k_0^T(x,v'_T;v_T)\gamma(v'_T) \end{aligned}$$

and similarly for the orthogonal part, which concludes the proof of the reversibility. □

Remark 3.2 (Return of the bouncy sampler)**.** The degenerate, deterministic case $\rho_T = -1$, $\rho_O = 1$ gives $\lambda_0 = 1$ and we get back the bouncy sampler.

From now on we assume that there is no noise on the orthogonal part, that is, $\rho_O(x) = 1$, and that $\lambda_0(x,v) = \lambda_0(x) = 2/(1-\rho_T(x))$ for all $x, v \in \mathbb{R}^d$ and $\psi(s) = (s)_+$ for all $s \in \mathbb{R}$. Introducing the notation

$$\varepsilon(x) := \frac{1-\rho_T(x)}{1+\rho_T(x)} \in [0,+\infty] \tag{3.2}$$

we can express (dropping the x dependence notation in the following for simplicity)

$$\rho_T = \frac{1-\varepsilon^2}{1+\varepsilon^2}\,, \qquad \sqrt{1-\rho_T^2} = \frac{2\varepsilon}{1+\varepsilon^2}\,, \qquad \lambda_0 = \frac{1+\varepsilon^2}{\varepsilon^2}.$$

The consequences of the previous discussion are gathered in the following result.

Lemma 3.3 (Velocity-jump sampler)**.** *Let $\varepsilon = \varepsilon(x) > 0$ denote a strictly positive function on $\mathbb{R}^d$. Let q_0 be the proposal kernel given by*

$$\int \varphi(v')q_0(x,v,dv') = \frac{1+\varepsilon^2}{\varepsilon^2}\mathbb{E}\left(\varphi(V')\right),$$

where the random variable V' is constructed from the tangent vector $T = T(x) = \nabla U(x)/|\nabla U(x)|$ (if $\nabla U(x) \neq 0$, and 0 otherwise), and a standard one-dimensional Gaussian G by the formula:

$$V' = v - \frac{2\varepsilon}{1+\varepsilon^2}\left(\varepsilon v\cdot T + G\right)T. \tag{3.3}$$

Consider the PDMP generator $\mathcal{L} = \mathcal{T} + \mathcal{F}$ where $\mathcal{T} = v\cdot\nabla_x$ and $\mathcal{F}$ is the velocity jump operator given by correcting q_0:

$$\mathcal{F}(\varphi)(x,v) = \int \left(\varphi(v')-\varphi(v)\right)q(x,v,dv') = \frac{1}{2}\int \left(\varphi(v')-\varphi(v)\right)\left(\nabla_x U\cdot(v-v')\right)_+ q_0(x,v,dv').$$

We have the following properties:

1. *the proposal kernel q_0 satisfies the average condition* (A) *and is reversible with respect to the unit Gaussian distribution in velocity variables.*

2. *Consequently, the process with generator $\mathcal{L}$ leaves the target distribution π invariant.*

In particular, $\varepsilon = +\infty$ is the full bouncy particle, $\varepsilon = 1$ the full resampling, $\varepsilon < 1$ a partial memory, and $\varepsilon \to 0$ corresponds to small changes at an increasing jump rate.

For theoretical and practical reasons, it is interesting to derive a more explicit formula for the corrected kernel.

Theorem 3.4 (Corrected jump rate)**.** *The corrected kernel associated with the velocity jumps process of Lemma 3.3 is given by*

$$\int \varphi(v')q(x,v,\mathrm{d}v') = \frac{|\nabla U|}{\varepsilon}\mathbb{E}\left[\varphi\left(x, v - \frac{2\varepsilon}{1+\varepsilon^2}\left(\varepsilon v\cdot T + G\right)T\right)\left(\varepsilon v\cdot T + G\right)_+\right].$$

As a consequence, the corrected jump rate λ is given by

$$\lambda(x,v) = \frac{|\nabla U(x)|}{\varepsilon(x)}\mathbb{E}\left[\left(v\cdot T(x)\varepsilon(x) + G\right)_+\right] = \frac{|\nabla U(x)|}{\varepsilon(x)}\Theta(\varepsilon(x)v\cdot T(x))$$

where

$$\Theta(u) := \mathbb{E}\left[\left(u+G\right)_+\right] = u\mathbb{P}(G > -u) + \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-u^2/2}.$$

Proof. By definition,

$$\int \varphi(v')q(x,v,\mathrm{d}v') = \frac{1}{2}\lambda_0(x)\mathbb{E}\left[\varphi\left(x,V'\right)\left(\nabla U\cdot(v-V')\right)_+\right].$$

Replacing V' by its expression given by (3.3) concludes. □

3.2 Convergence toward the Hamiltonian dynamics

From now on, we denote by $\mathcal{L}_\varepsilon$ the generator of the velocity jump process with kernel $q = q_\varepsilon$ given by Theorem 3.4 for some positive function ε on $\mathbb{R}^d$, i.e.

$$\mathcal{L}_\varepsilon\varphi(x,v) = v\cdot\nabla_x\varphi(x,v) + \int_{v'\in\mathbb{R}^d}\left(\varphi(x,v') - \varphi(x,v)\right)q_\varepsilon(x,v;\mathrm{d}v'). \tag{3.4}$$

It can then be formally expanded using Taylor's formula as

$$\mathcal{L}_\varepsilon(\varphi) = v\cdot\nabla_x\varphi + \sum_{n=1}^{+\infty}\frac{|\nabla U|}{n!}D_v^n\varphi\left(T,\dots,T\right)\varepsilon^{n-1}\left(\frac{-2}{1+\varepsilon^2}\right)^n\mathbb{E}\left[\left(\varepsilon v\cdot T+G\right)_+^{n+1}\right]. \tag{3.5}$$

For $n \geqslant 1$,

$$\mathbb{E}\left[\left(u+G\right)_+^{n+1}\right] = \mathbb{E}\left[G_+^{n+1}\right] + (n+1)u\mathbb{E}\left[G_+^n\right] + \underset{u\to 0}{\mathcal{O}}(u^2),$$

with

$$\mathbb{E}(G_+) = 1/\sqrt{2\pi}, \qquad \mathbb{E}(G_+^2) = 1/2, \qquad \mathbb{E}(G_+^3) = \sqrt{2/\pi}.$$

As a consequence, as ε vanishes, we formally get back the Hamiltonian dynamics

$$\mathcal{L}_\varepsilon(\varphi) = v\cdot\nabla_x\varphi - \nabla U\cdot\nabla_y\varphi + \mathcal{O}(\varepsilon),$$

and at first order in ε, we obtain a (degenerate) Langevin diffusion

$$\begin{aligned}\mathcal{L}_\varepsilon(\varphi) &= v\cdot\nabla_x\varphi - \nabla U\cdot\nabla_y\varphi + \varepsilon\frac{4|\nabla U|}{\sqrt{2\pi}}\left[-(v\cdot T)T\cdot\nabla_v\varphi + D_v^2\varphi(T,T)\right] + \mathcal{O}(\varepsilon^2)\\ &= v\cdot\nabla_x\varphi - \nabla U\cdot\nabla_y\varphi + \varepsilon\frac{4|\nabla U|}{\sqrt{2\pi}}\left[e^{\frac{|v|^2}{2}}T\cdot\nabla_v\left(e^{-\frac{|v|^2}{2}}T\cdot\nabla_v\varphi\right)\right] + \mathcal{O}(\varepsilon^2)\,,\end{aligned}$$

which can be interpreted as the Langevin process that is degenerate along the force direction; is reversible (up to velocity reversal) with respect to the target distribution π, and has a typical relaxation time of order $1/(|\nabla_x U|\,\varepsilon)$.

We now give conditions under which the convergence of the velocity jump process towards an Hamiltonian dynamics can be proven rigorously. It is remarkable that the limit can be identified as soon as the martingale problem for the *deterministic* Hamiltonian dynamics is well-posed. If ∇U is Lipschitz, this is a consequence of the standard Cauchy-Lipschitz theory; the minimal conditions on ∇U being still an open problem. We define first martingale problems in $\mathbb{R}^d$.

Definition 3.5. *A càdlàg random process $(Z_t)_{t\geq 0}$ in $\mathbb{R}^d$ with initial distribution μ is solution to the martingale problem associated with $(\mu, L, C_c^\infty(\mathbb{R}^d))$, where L is a Markov generator, if for any $\varphi \in C_c^\infty(\mathbb{R}^d)$ the process*

$$t \mapsto \varphi(Z_t) - \int_0^t L\varphi(Z_s)ds$$

is a martingale with respect to the natural filtration of Z. We say that uniqueness holds if all solutions have the same probability distribution on the usual Polish space of càdlàg trajectories.

Theorem 3.6. *Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence of strictly positive measurable functions on $\mathbb{R}^d$ that vanishes uniformly on all compact sets as $n \to +\infty$. Denote $\mathcal{L}_{\varepsilon_n}$ the associated PDMP generator and denote*

$$\mathcal{L}_0 \overset{\text{def}}{=} v\cdot\nabla_x - \nabla U\cdot\nabla_v\,,$$

and consider $\mu \in \mathcal{P}(\mathbb{R}^{2d})$ an initial distribution. Assume that

- *For each n, the velocity jump process associated to $\mathcal{L}_{\varepsilon_n}$ is defined for all time (the sequence of jump times converges to $+\infty$).*
- *∇U is continuous and the martingale problem associated with $(\mu, \mathcal{L}_0, C_c^\infty(\mathbb{R}^{2d}))$ is well-posed on $\mathbb{R}^{2d}$.*

Then, as $n \to +\infty$, the velocity jump process associated to $\mathcal{L}_{\varepsilon_n}$ converges in distribution in the space of càdlàg trajectories endowed with the Skorohod topology towards the unique martingale solution of the Hamiltonian dynamics $\mathcal{L}_0$.

Proof. The proof follows from a general result, Theorem 7.1, postponed to an Appendix section. Indeed, according to Theorem 7.1, it is sufficient to check that for any $\varphi \in C_c^\infty(\mathbb{R}^{2d})$ and any compact $K \subset \mathbb{R}^{2d}$

$$\lim_{n\to+\infty}\sup_K |\mathcal{L}_{\varepsilon_n}\varphi - \mathcal{L}_0\varphi| = 0.$$

Using the definition of $\mathcal{L}_{\varepsilon_n}$ from Equation (3.4), and denoting by $u_n = u_n(x,v) = \varepsilon_n(x) v \cdot T(x) + G$, (where $T(x) = \nabla U(x)/|\nabla U(x)|$), the difference $(\mathcal{L}_{\varepsilon_n}\varphi - \mathcal{L}_0\varphi)(x,v)$ may be rewritten as:

$$\mathbb{E}\left[\frac{|\nabla U(x)|\,(u_n)_+}{\varepsilon_n(x)}\left(\varphi\left(v - \frac{2\varepsilon_n u_n(x)}{1+\varepsilon_n(x)^2}T(x)\right) - \varphi(v)\right) + \nabla U(x)\cdot\nabla_v\varphi(v)\right].$$

Omitting the dependency in x in the notations for legibility, we may apply Taylor's theorem at the first order on the difference $(\varphi\left(v - \frac{2\varepsilon_n u_n}{1+\varepsilon_n^2}T\right) - \varphi(v))$ to get:

$$|(\mathcal{L}_\varepsilon\varphi - \mathcal{L}_0\varphi)(x,v)| \leqslant \left|1 - \frac{2\mathbb{E}\left[(u_n)_+^2\right]}{1+\varepsilon_n^2}\right| |\nabla U\cdot\nabla_v\varphi(v)| + |\nabla U|\|\nabla^2\varphi\|_\infty \varepsilon_n \mathbb{E}\left[(u_n)_+^3\right].$$

Since $\mathbb{E}\left[(u_n)_+^2\right]$ converges to $1/2$ uniformly on all compact sets and $\mathbb{E}\left[(u_n)_+^3\right]$ is uniformly bounded in n on all compact sets, the right hand side vanishes uniformly on all compact sets of $\mathbb{R}^{2d}$ as $n \to +\infty$. □

Remark 3.7. More generally, considering a limit generator $\mathcal{L}_0 + \mathcal{D}_0$ for some dissipative $\mathcal{D}_0$, the proof of Theorem 3.6 is straightforwardly adapted to get the convergence of the processes associated to generators $\mathcal{L}_{\varepsilon_n} + \mathcal{D}_{\varepsilon_n}$ with $\mathcal{D}_{\varepsilon_n}\varphi \to \mathcal{D}_0\varphi$ for all $\varphi \in C_c^\infty(\mathbb{R}^{2d})$. For instance, that way we can design velocity jump processes that converge toward the Langevin diffusion or the HMC process.

3.3 Drift limit and factorization

As discussed in Section 2.1, if the forces are decomposed as $\nabla U(x) = \sum_{i=1}^N \xi_i(x)$ for some vector fields ξ_i, then we can consider the operators given by

$$\mathcal{F}_i\varphi(x,v) = \int_{v'\in\mathbb{R}^d} (\varphi(x,v') - \varphi(x,v))\, q_i(x,v;\mathrm{d}v'),$$

with

$$\int \varphi(v') q_i(x,v,\mathrm{d}v') = \frac{|\xi_i|}{\varepsilon_i}\mathbb{E}\left[\varphi\left(v - \frac{2\varepsilon_i}{1+\varepsilon_i^2}\left(\varepsilon_i v\cdot T_i + G\right)T_i\right)\left(\varepsilon_i v\cdot T_i + G\right)_+\right]$$

where G is a one-dimensional standard Gaussian variable, $x \mapsto \varepsilon_i(x)$ is a positive function and $T_i(x) = \xi_i(x)/|\xi_i(x)|$ if $\xi_i(x) \neq 0$ and $T_i(x) = 0$ otherwise. In other words, the process with generator $\mathcal{T} + \mathcal{F}_i$ is exactly the velocity jump process introduced in Section 3.1, except that ∇U is replaced everywhere by ξ_i. In particular, the previous results are straightforwardly extended: from Lemma 3.1, the generators $\mathcal{F}_i$ satisfy (2.8) (and more precisely (2.14) with ∇U replaced by ξ_i), so that $\mathcal{L} = \mathcal{T} + \sum_{i=1}^N \mathcal{F}_i$ satisfies $\int \mathcal{L}\varphi d\pi = 0$ for all $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$. Similarly, from the computations of Section 3.2,

$$\mathcal{F}_i(\varphi) = -\xi_i\cdot\nabla_y\varphi + \mathcal{O}(\varepsilon_i),$$

and thus we still get the convergence toward the Hamiltonian dynamics, since

$$\mathcal{L}(\varphi) = v\cdot\nabla_x\varphi - \nabla U\cdot\nabla_y\varphi + \mathcal{O}(\max_{1\leqslant i\leqslant N}\varepsilon_i).$$

Let us give two examples of such a factorization.

3.3.1 Gibbs velocity jump processes

For $i \in [\![1, d]\!]$, set $\xi_i(x) = \partial_{x_i} U(x) e_i$, where e_i is the i^{th} vector of the canonical basis and

$$\mathcal{L} = \mathcal{T} + \sum_{i=1}^{d} \mathcal{F}_i = \sum_{i=1}^{d} (v_i \partial_{x_i} + \mathcal{F}_i)$$

where $\mathcal{F}_i$ is defined as above, with some ε_i. The corresponding process can be seen as a (kinetic) Gibbs sampler: indeed, each generator $v_i \partial_{x_i} + \mathcal{F}_i$ leaves invariant the conditional law $(x_i, v_i) \mapsto \pi(dxdv)$. When $\varepsilon_i(x) = +\infty$ for all i, we recover the zig-zag process, which may thus be seen as a Gibbs version of the bouncy sampler (remark that, when $\varepsilon = +\infty$, the norm of the velocity is unchanged at jump times, so that although $\pi = \nu \otimes \gamma$ is indeed invariant for $\mathcal{L}$ with a a Gaussian distribution γ, it won't be ergodic).

For a general choice of ε_i, this factorization ensures the following property: in the case where the target law is a tensor product of one-dimensional laws, i.e. if $U(x) = \sum_{i=1}^{d} U_i(x_i)$ for some one-dimensional potentials U_i, then the coordinates of the corresponding kinetic Gibbs process are independent one-dimensional processes.

Note that

$$\mathcal{L}_\varepsilon(\varphi) \;=\; v \cdot \nabla_x \varphi - \nabla U \cdot \nabla_y \varphi + \sum_{i=1}^{d} \varepsilon_i \frac{4|\partial_{x_i} U|}{\sqrt{2\pi}} \left[-v_i \partial_{v_i} \varphi + \partial^2_{v_i} \varphi \right] + \mathcal{O}(\max_{1 \leqslant i \leqslant N} \varepsilon_i^2)\,.$$

The fact that, in that case, the order one term is a non-degenerate Langevin diffusion is reminiscent of the fact the Zig-Zag process is irreducible in cases where the bouncy sampler is not, see [5].

3.3.2 Multi-time-stepping

Suppose that $\nabla U = \xi_1 + \xi_2$ where ξ_1 is large and numerically cheap to compute by comparison with ξ_2, smaller but numerically more intensive. To fix ideas, suppose that $\|\xi_i\|_\infty \leqslant L_i$ for $i = 1, 2$ with known constants $L_1 \gg L_2$. For $i = 1, 2$, take $\varepsilon_i(x) = \varepsilon_0$ for some $\varepsilon_0 > 0$. Then, in order to sample a trajectory of the process corresponding to the splitting $\nabla U = \xi_1 + \xi_2$, as detailed in Section 5, ξ_i will be computed at a rate L_i/ε_0. Hence, the splitting reduces the number of computations of ξ_2. This extends the strategy of [18] where $\varepsilon_1 = 0$ and $\varepsilon_2 = +\infty$ (bounce/drift process).

3.4 Non-irreducibility

The bouncy particle sampler and the Hamiltonian dynamics are well-known to be both non-irreducible in general. There are in fact non-irreducible counterexamples for all the processes with generator $\mathcal{L}_\varepsilon = \mathcal{T} + \mathcal{F}_\varepsilon$, $\varepsilon > 0$ in the case with no additional noise ($\mathcal{D} = 0$). For instance, for a symmetric Gaussian target (or more generally any target with radial potential, i.e. that is invariant by isometries preserving the origin) in dimension larger than one, $\nabla U(X_t)$ being collinear to X_t, note that $X_t, V_t \in \mathrm{span}(X_0, V_0)$ for all $t \geqslant 0$. Moreover, assuming that X_0 and V_0 are not collinear, even within this two-dimensional plane, the process is not irreducible. Indeed, in the following, still for a symmetric Gaussian target, suppose that $d = 2$ and

$(x_0, v_0) \in \mathbb{R}^2 \times \mathbb{R}^2$ with $\mathrm{span}(x_0, v_0) = \mathbb{R}^2$. Remark that $X_t \wedge V_t := X_t^1 V_t^2 - X_t^2 V_t^1$ is unchanged by the free transport and by the jumps, hence is constant along time. In particular, starting from a deterministic condition (x_0, v_0) the law of the process will never converge to the Gaussian target measure. More precisely, we expect the law of the process to converge to the law of a standard Gaussian variable (X, V) on $\mathbb{R}^4$ conditioned to $X \wedge V = x_0 \wedge v_0$ (since the standard Gaussian on $\mathbb{R}^4$ is invariant for the process, so is this conditional law). Even if we are only concerned with the law of X, this induces a bias (see the numerical section).

4 Hypocoercivity

The question of long-time convergence and ergodicity for velocity jump samplers have been addressed in various cases in [5, 8, 11] with a Meyn-Tweedie approach and in [2] with the L^2 hypocoercivity method of Dolbeault-Mouhot-Schmeiser [10]. Our approach will be similar to the latter. Since the process is not irreducible in general, a dissipative part is added for the velocities. In all this section, the target measure π is given by (2.1) with γ the standard (mean 0, variance Id) Gaussian distribution on $\mathbb{R}^d$ and we consider a kinetic process with generator $\mathcal{L} = \mathcal{T} + \mathcal{F} + \mathcal{D}$ as in Section 2.1 and $\mathcal{F}$ is the operator defined in Lemma 3.3 for some non-negative function ε on $\mathbb{R}^d$.

We would like to emphasize that we will only conduct a formal study, disregarding in particular the question of domains and extensions of the operators involved. The technical arguments to make the proofs valid would be exactly those of [2], and thus we omit them for the sake of clarity and in order to focus on the (formal) computations.

Assumption 4.1. *The dissipative part $\mathcal{D}$ may be written as $\mathcal{D} = \eta(x)\mathcal{D}_0$, where $\eta : \mathbb{R}^d \to \mathbb{R}_+$ is such that*

$$\forall x \in \mathbb{R}^d\,, \qquad 0 < \underline{\eta} < \eta(x) < \overline{\eta}\,(1 + |\nabla U(x)|)\,,$$

for some $\overline{\eta} \geqslant \underline{\eta} > 0$, and $\mathcal{D}_0$ is a self-adjoint operator on $L^2(\gamma)$ such that $\mathcal{D}_0(v) = -v$ and with a spectral gap of 1, in the sense that, for all nice $g \in L^2(\gamma)$,

$$\langle \mathcal{D}_0 g, g \rangle_{L^2(\gamma)} \leqslant -\|g - \int g \mathrm{d}\gamma\|^2_{L^2(\gamma)}\,.$$

Moreover, $U \in \mathcal{C}^2(\mathbb{R}^d)$ and there exists $C_1 \geqslant 0$ such that

$$\nabla^2 U(x) \succeq -C_1 I \tag{C_1}$$

(in the sense of positive symmetric matrices) for all $x \in \mathbb{R}^d$ and

$$\liminf_{|x| \to \infty} \left(\frac{1}{2} |\nabla U|^2 - \Delta U \right) > 0\,. \tag{4.1}$$

Finally, $\mathcal{T} = v \cdot \nabla_x$ and $\mathcal{F}$ belongs to the class of operators defined in Lemma 3.3.

Remark 4.2. The three classical dissipative operators $\mathcal{D}_0$ given by (2.5), (2.6) and (2.7) are all self-adjoint in $L^2(\gamma)$ with a spectral gap of 1 and with $\mathcal{D}_0(v) = -v$.

The condition (4.1) classically implies that the measure ν satisfies a Poincaré inequality with some constant $c_P > 0$: for all $f \in H^1(\nu)$,

$$\|f - \nu f\|^2_{L^2(\nu)} \;\leqslant\; \frac{1}{c_P} \|\nabla_x f\|^2_{L^2(\nu)} \,. \tag{c_P}$$

It also implies that there exist $C_2 > 0$ such that

$$\forall x \in \mathbb{R}^d \,, \qquad \Delta U(x) \leqslant C_2 + |\nabla U(x)|^2/2 \,. \tag{C_2}$$

In the following, $\|\cdot\|$ and $\langle \cdot \rangle$ stands respectively for the norm and scalar product in $L^2(\pi)$. We denote by m_2 (respectively m_4) the second (respectively fourth) moment of γ:

$$m_2 = \int |v|^2 \, d\gamma(v) = d, \qquad\qquad m_4 = \int |v|^4 \, d\gamma(v) = d(d+2).$$

Let $(P_t)_{t \geqslant 0}$ be the Markov semi-group with generator $\mathcal{L}$.

Theorem 4.3 (Exponential convergence in $\mathbb{L}^2$)**.** *Under Assumption 4.1, for all $f \in L^2(\pi)$ and all $t \geqslant 0$,*

$$\|(P_t - \pi) f\|^2 \;\leqslant\; \frac{4}{3} e^{-\kappa t} \|(I - \pi) f\|^2 \,,$$

where κ is given by:

$$\frac{1}{\kappa} \;=\; \frac{6}{\underline{\eta}} \left(1 + \frac{1}{c_P^2} \left(1 + \frac{C_1}{2 c_P} \right) \left(1 + 4 C_2 + 16 c_P^2 \right) \left(\overline{\eta}/\sqrt{d} + 5 \sqrt{1 + 2/d^2} + 4/d \right)^2 \right) .$$

Remark 4.4. The main point here is that κ does not depend on ε. Also note that, as a function of η, the convergence rate scales for large η as $\underline{\eta}/\max(1, \overline{\eta}^2)$, which is well-known for the Langevin dynamics with a constant η and suggests that the constant remains finite in the overdamped regime under proper rescaling (albeit with a sub-optimal constant of order c_P^3 instead of c_P). For $c_P \ll 1$ and $d \gg 1$ we obtain

$$\frac{1}{\kappa} \;\sim\; \frac{3}{c_P^3} \left(2 c_P + C_1 \right) \left(1 + 4 C_2 \right) \frac{1}{\underline{\eta}} \left(\overline{\eta}/\sqrt{d} + 5 \right)^2 .$$

Alternatively, if U is ρ-convex for some $\rho > 0$ independent from the dimension (so that $C_1 = 0$ and $c_P = \rho$), choosing a constant $\eta = \sqrt{d}$, we get $\kappa = \mathcal{O}(C_2/\sqrt{d})$. For instance, for a standard d-dimensional Gaussian target, $C_2 = d$.

Denote $\mathcal{M}^*$ the dual of an operator $\mathcal{M}$ in $L^2(\pi)$, $\mathcal{S} = (\mathcal{L} + \mathcal{L}^*)/2$ and $\mathcal{A} = (\mathcal{L} - \mathcal{L}^*)/2$ the symmetric and skew symmetric parts of $\mathcal{L}$ and

$$\Pi_v f(x, v) \;=\; \int f(x, v') \gamma(\mathrm{d}v') \,.$$

The Dolbeault-Mouhot-Schmeiser method [10] relies on the modified norm

$$\mathbf{H}(f) = \frac{1}{2} \|f\|^2 + \delta \, \langle \mathcal{B} f, f \rangle \;,$$

where $\mathcal{B}$ is defined by

$$\mathcal{B} \;=\; -\left(mI + (\mathcal{A}\Pi_v)^* \mathcal{A}\Pi_v\right)^{-1} (\mathcal{A}\Pi_v)^* \;,$$

for some scalar parameters $\delta, m > 0$ to be chosen later on. From [2, Proposition 26-(d)] (applied to the operator $-\mathcal{A}\Pi_v/\sqrt{m}$) , $\|\mathcal{B}\| \leqslant 1/\sqrt{m}$ so that $\mathbf{H}$ is equivalent to the $L^2(\pi)$ norm for $\delta < \sqrt{m}/2$. The aim is thus to prove that $\mathbf{H}$ decays exponentially fast along the semi-group $(P_t)_{t\geqslant 0}$, which proves the hypocoercive decay in $L^2(\pi)$ (in the sense of [23], that is: exponential decay up to a constant factor $C > 1$). Formally, the general result is the following:

Theorem 4.5. *Assume that*

$$\mathcal{S}\Pi_v = 0 \qquad \Pi_v \mathcal{A}\Pi_v = 0$$

and that there exist $c_v, R(m) = R > 0$ and $c_x(m) = c_x \in (0,1]$ such that, for all nice $f \in L^2(\pi)$ with $\pi f = 0$, it holds:

$$\textit{(microscopic coercivity)} \qquad \langle \mathcal{S}f, f\rangle \leqslant -c_v \|(I-\Pi_v)f\|^2 \tag{4.2}$$

$$\textit{(macroscopic coercivity)} \qquad \langle \mathcal{B}\mathcal{A}\Pi_v f, f\rangle \leqslant -c_x \|\Pi_v f\|^2 \tag{4.3}$$

$$\textit{(auxiliary bound)} \qquad \langle \mathcal{B}\mathcal{L}(1-\Pi_v)f, f\rangle \leqslant R\|\Pi_v f\|\|(I-\Pi_v)f\| \,. \tag{4.4}$$

Then, for all $f \in L^2(\pi)$ and all $t \geqslant 0$,

$$\|(P_t - \pi)f\|^2 \;\leqslant\; \frac{4}{3} e^{-\kappa t} \|(I-\pi)f\|^2 \,,$$

where

$$\kappa \;=\; c_x \inf_{m>0} \min\left(\frac{\sqrt{m}}{6}, \frac{2c_v}{6 + 3R^2/c_x}\right) .$$

Remark 4.6. Typically, the macroscopic coercivity amounts to a spectral gap of the operator $(\mathcal{A}\Pi_v)^* \mathcal{A}\Pi_v$ restricted to functions of space variables. In that case (which indeed occurs for our PDMP), one has

$$c_x = \frac{c}{m+c},$$

where c is the spectral gap of $(\mathcal{A}\Pi_v)^* \mathcal{A}\Pi_v$. Then one can choose $m = c$ to get

$$\kappa \;=\; \min\left(\frac{\sqrt{c}}{12}, \frac{c_v}{6 + 6R^2}\right).$$

Proof. We only recall the main steps and refer to [10, 2] for details. Denoting $f_t = P_t f - \int f \mathrm{d}\pi$, from $\partial_t f_t = \mathcal{L} f_t$ we get that

$$\partial_t \mathbf{H}(f_t) = \langle f_t, \mathcal{L} f_t\rangle + \delta \langle \mathcal{B} f_t, \mathcal{L} f_t\rangle + \delta \langle \mathcal{B}\mathcal{L} f_t, f_t\rangle \;.$$

The microscopic coercivity condition (4.2) intervenes in the first term

$$\langle f, \mathcal{L} f\rangle \;=\; \langle f, \mathcal{S} f\rangle \;\leqslant\; -c_v \|(I-\Pi_v)f\|^2 \,.$$

Under the condition $\Pi_v \mathcal{A} \Pi_v = 0$, the second term is bounded as

$$\langle \mathcal{B}f, \mathcal{L}f \rangle \;\leqslant\; \|(I - \Pi_v)f\|^2 \,,$$

see [2, Lemma 5]. From the macroscopic coercivity and auxiliary bounds conditions (4.3) and (4.4), the third term gives

$$\langle \mathcal{B}\mathcal{L}f, f \rangle \;=\; \langle \mathcal{B}\mathcal{L}\Pi_v f, f \rangle + \langle \mathcal{B}\mathcal{L}(1-\Pi_v)f, f \rangle \;\leqslant\; -c_x \|\Pi_v f\|^2 + R\|(I-\Pi_v)f\|\|\Pi_v f\| \,,$$

where we used that $\mathcal{S}\Pi_v = 0$. Denoting $\alpha = \|\Pi_v f_t\|^2 / \|f_t\|^2 \in [0,1]$, we have thus obtained

$$\begin{aligned} \frac{\partial_t \mathbf{H}(f_t)}{\|f_t\|^2} &\leqslant (\delta - c_v)(1-\alpha) - c_x \delta \alpha + \delta R \sqrt{\alpha(1-\alpha)} \\ &\leqslant \left(\delta \left(1 + \frac{R^2}{2c_x} \right) - c_v \right) (1-\alpha) - \frac{1}{2} c_x \delta \alpha \,. \end{aligned}$$

In particular, if $\delta \leqslant c_v / (2 + R^2/c_x)$, we get that

$$\frac{\partial_t \mathbf{H}(f_t)}{\|f_t\|^2} \;\leqslant\; -\frac{1}{2} c_v (1-\alpha) - \frac{1}{2} c_x \delta \alpha \;\leqslant\; -\frac{1}{2} c_x \delta$$

for all $\alpha \in [0,1]$, where we used that $c_x \leqslant 1$ and $\delta \leqslant c_v$. If moreover $\delta \leqslant \sqrt{m}/4$ we get that $\|f\|^2 \leqslant 4\mathbf{H}(f) \leqslant 3\|f\|^2$ and

$$\partial_t \mathbf{H}(f_t) \;\leqslant\; -\frac{1}{2} c_x \delta \|f_t\|^2 \;\leqslant\; -\frac{2}{3} c_x \delta \mathbf{H}(f_t) \,.$$

We may then apply Gronwall's Lemma to conclude: for all $\delta \leqslant \min(\sqrt{m}/4, c_v/(2 + R^2/c_x))$,

$$\|f_t\|^2 \;\leqslant\; 4\mathbf{H}(f_t) \;\leqslant\; 4e^{-2c_x \delta t/3} \mathbf{H}(f_0) \;\leqslant\; \frac{4}{3} e^{-2c_x \delta t/3} |f_0\|^2 \,. \qquad \square$$

We now have to check that the conditions of Theorem 4.5 are met under Assumption 4.1. This is usually done by computing explicitly $\mathcal{S}$ and $\mathcal{A}$ for particular processes. In fact we will only need the following information, which is obtained from the condition (2.14), satisfied by all usual kinetic samplers:

Lemma 4.7. *Under Assumption 4.1,*

$$\mathcal{S}\Pi_v = 0 \,, \qquad \mathcal{A}\Pi_v \;=\; \mathcal{T}\Pi_v \,, \qquad \Pi_v \mathcal{A} \Pi_v = 0 \,. \tag{4.5}$$

Proof. Since $\mathcal{F}$ and $\mathcal{D}$ only act on the v variable and $\Pi_v f$ only depends on x for all f, $\mathcal{D}\Pi_v = \mathcal{F}\Pi_v = 0$. Moreover, from condition (2.14), for all $f, g \in \mathcal{C}_c^\infty(\mathbb{R}^{2d})$,

$$\begin{aligned} \langle \mathcal{F}^* \Pi_v f, g \rangle &= \int_{\mathbb{R}^{2d}} \mathcal{F} g(x,v) \Pi_v f(x,v) \pi(\mathrm{d}x \mathrm{d}v) \\ &= \int_{\mathbb{R}^d} \int_{R^d} f(x,w) \gamma(\mathrm{d}w) \int_{\mathbb{R}^d} \mathcal{F} g(x,v) \gamma(\mathrm{d}v) \nu(\mathrm{d}x) \\ &= -\int_{\mathbb{R}^{2d}} (v \cdot U(x)) g(x,v) \Pi_v f(x,v) \pi(\mathrm{d}x \mathrm{d}v) \,. \end{aligned}$$

In other words, $\mathcal{F}^*\Pi_v f(x,v) = -v\cdot\nabla U(x)\Pi_v f(x,v)$. Besides, $\mathcal{D}^* = \mathcal{D}$ by assumption and, integrating by parts, $\mathcal{T}^* f(x,v) = -\mathcal{T} f(x,v) + v\cdot\nabla U(x) f(x,v)$. As a consequence,

$$2\mathcal{S}\Pi_v \;=\; (\mathcal{T}+\mathcal{T}^*+\mathcal{F}+\mathcal{F}^*+\mathcal{D}+\mathcal{D}^*)\Pi_v \;=\; 0$$

and

$$2\mathcal{A}\Pi_v \;=\; (\mathcal{T}-\mathcal{T}^*+\mathcal{F}-\mathcal{F}^*+\mathcal{D}-\mathcal{D}^*)\Pi_v \;=\; 2\mathcal{T}\Pi_V\,.$$

Finally, for all $f\in\mathcal{C}_c^\infty(\mathbb{R}^{2d})$,

$$\Pi_v\mathcal{T}\Pi_v f(x,v) \;=\; \int_{\mathbb{R}^d} w\gamma(\mathrm{d}w)\cdot\nabla_x\int_{\mathbb{R}^d} f(x,w)\gamma(\mathrm{d}w) \;=\; 0\,. \qquad \square$$

In particular, the operator $\mathcal{B}$ being defined from the operator $\mathcal{A}\Pi_v = \mathcal{T}\Pi_v$, it is the same in our case and in [2] (up to the choice of the parameter m, which is $m = m_2$ in [2]). From [2, Lemma 9], $(\mathcal{A}\Pi_v)^*\,\mathcal{A}\Pi_v f = m_2\nabla_x^*\nabla_x\Pi_v f$ and thus

$$\mathcal{B}^* f \;=\; -\mathcal{T}u\,, \qquad \text{where} \qquad u \;=\; (mI + m_2\nabla_x^*\nabla_x)^{-1}\,\Pi_v f\,. \tag{4.6}$$

Remark that u is a function of x alone.

Lemma 4.8. *Under Assumption 4.1, the microscopic and macroscopic coercivity conditions* (4.2) *and* (4.3) *respectively hold with* $c_v = \underline{\eta}$ *and* $c_x = m_2c_P/(m+m_2c_P)$.

Proof. To get the microscopic coercivity estimate, we remark that $\mathcal{T}+\mathcal{F}$ is the generator of a Markov semigroup that fixes π, so that

$$0 \;\geqslant\; \int f(\mathcal{T}+\mathcal{F})f\pi \;=\; \frac{1}{2}\int f(\mathcal{T}+\mathcal{F}+\mathcal{T}^*+\mathcal{F}^*f)\pi\,,$$

and thus

$$\begin{aligned}\langle\mathcal{S}f,f\rangle \;\leqslant\; \langle\mathcal{D}f,f\rangle \;&=\; \int\eta(x)\int f(x,v)\mathcal{D}_0 f(x,v)\gamma(\mathrm{d}v)\nu(\mathrm{d}x)\\ &\leqslant\; -\int\eta(x)\int\left(f(x,v)-\Pi_v f(x,v)\right)^2\gamma(\mathrm{d}v)\nu(\mathrm{d}x) \leqslant\; -\underline{\eta}\|f-\Pi_v f\|^2\,.\end{aligned}$$

For the macroscopic condition, remark that

$$\mathcal{B}\mathcal{A}\Pi_v f = -\Phi\left((\mathcal{A}\Pi_v)^*\,\mathcal{A}\Pi_v\right) f$$

with $\Phi(z) = z/(m+z)$, which is a non-decreasing function from $\mathbb{R}_+$ to $[0,1]$. Moreover, $(\mathcal{A}\Pi_v)^*\,\mathcal{A}\Pi_v$ is self-adjoint and for all $f\in L^2(\pi)$ such that $\pi f = 0$ (so that $\nu\Pi_v f = 0$),

$$\langle(\mathcal{A}\Pi_v)^*\,\mathcal{A}\Pi_v f,f\rangle \;=\; m_2\,\langle\nabla_x^*\nabla_x\Pi_v f,\Pi_v f\rangle \;=\; m_2\|\nabla_x\Pi_v f\|^2 \;\geqslant\; m_2c_P\|\Pi_v f\|^2\,.$$

From the spectral mapping theorem [7, Theorem 2.5.1, Corollary 2.5.4], $\Phi\left((\mathcal{A}\Pi_v)^*\,\mathcal{A}\Pi_v\right)$ is self-adjoint with a spectral gap bounded by $\Phi(m_2c_P)$, which concludes. $\square$

The previous results have been established using only the general condition (2.14). By contrast, the proof of the auxiliary bound (4.4) is based on the particular form of $\mathcal{L}$.

Lemma 4.9. *Under Assumption 4.1,*

$$\langle \mathcal{B}\mathcal{L}(1-\Pi_v)f, f\rangle \;\leqslant\; R\|\Pi_v f\|\|f-\Pi_v f\|\,,$$

where R is given by

$$R^2 = \frac{1}{m_2}\left(\frac{1}{m_2}+\frac{C_1}{2m}\right)\frac{1+4C_2+16c_P^2}{c_P^2}\left(\overline{\eta}\sqrt{m_2}+5\sqrt{m_4}+4\right)^2\,.$$

Proof. First, we bound

$$\langle \mathcal{B}\mathcal{L}(1-\Pi_v)f, f\rangle \;=\; \langle (1-\Pi_v)f, \mathcal{L}^*\mathcal{B}^* f\rangle \;\leqslant\; \|(I-\Pi_v)f\|\|\mathcal{L}^*\mathcal{B}^* f\|\,.$$

Let u be defined by (4.6). Using the process definition in Lemma 3.3, we first remark that since i) $\mathcal{T}+\mathcal{F}$ conserves the target distribution π and ii) q_0 is reversible, one has:

$$(\mathcal{T}+\mathcal{F})^*\,\varphi(x,v) = -v\cdot\nabla_x\varphi + \int \left(\varphi(x,v')-\varphi(x,v)\right)\left(\nabla U\cdot(v'-v)\right)_+ q_0(x,v,dv').$$

Using that $\mathcal{D}_0(v)=-v$ and that $(\nabla U\cdot T)(\nabla_x u\cdot T)=\nabla U\cdot\nabla_x u$,

$$\begin{aligned}
\mathcal{L}^*\mathcal{B}^* f &= -\mathcal{T}^2 u - \nabla_x u\cdot\mathcal{D}(v) + \frac{1}{2}\nabla_x u\cdot\int\left(\nabla U\cdot(v'-v)\right)_+(v'-v)q_0(x,v;\mathrm{d}v')\\
&= -v\cdot\nabla_x^2 u\, v + \eta v\nabla_x u - \frac{2}{1+\varepsilon^2}(\nabla U\cdot\nabla_x u)\int(\varepsilon v\cdot T+w)_-^2\, e^{-w^2/2}\frac{\mathrm{d}w}{\sqrt{2\pi}}\\
&=: -v\cdot\nabla_x^2 u\, v + \eta v\nabla_x u - (\nabla U\cdot\nabla_x u)H(v,x)
\end{aligned}$$

(Recall that ε, hence H, can depend on x). We bound

$$H(v) \;\leqslant\; \frac{2}{1+\varepsilon^2}\int(\varepsilon v\cdot T+w)_-^2\, e^{-w^2/2}\frac{\mathrm{d}w}{\sqrt{2\pi}} \;\leqslant\; 4|v|^2+4\,.$$

As a consequence,

$$|\mathcal{L}^*\mathcal{B}^* f| \;\leqslant\; |v|^2|\nabla_x^2 u| + \overline{\eta}|v||\nabla_x u|\sqrt{1+|\nabla U|^2} + \left(4|v|^2+4\right)|\nabla U||\nabla_x u|\,,$$

and

$$\|\mathcal{L}^*\mathcal{B}^* f\| \;\leqslant\; \sqrt{m_4}\|\nabla_x^2 u\| + (\overline{\eta}\sqrt{m_2}+4\sqrt{m_4}+4)\,\|\sqrt{1+|\nabla U|^2}\nabla_x u\|\,.$$

Finally, the following elliptic regularity estimates are proven in [2, Corollary 35 and Proposition 33]:

$$\begin{aligned}
\|\nabla_x^2 u\|^2 &\leqslant \left(\frac{1}{m_2^2}+\frac{C_1}{2mm_2}\right)\|\Pi_v f\|^2\\
\|\sqrt{1+|\nabla U|^2}\nabla_x u\|^2 &\leqslant \left(\frac{1}{m_2^2}+\frac{C_1}{2mm_2}\right)\frac{1+4C_2+16c_P^2}{c_P^2}\|\Pi_v f\|^2\,,
\end{aligned}$$

which concludes using $\frac{1+4C_2+16c_P^2}{c_P^2}\geq 1$. □

We may now conclude the proof.

Proof of Theorem 4.3. By (4.5) and Lemmas 4.8 and 4.9, Theorem 4.5 applies with any choice of $m > 0$. We take $m = m_2 c_P$, so that $c_x = 1/2$ in Lemma 4.8, R^2 given in Lemma 4.9 is

$$R^2 = \frac{1}{m_2^2 c_P^3}\left(c_P + \frac{C_1}{2}\right)\left(1 + 4C_2 + 16c_P^2\right)\left(\overline{\eta}\sqrt{m_2} + 5\sqrt{m_4} + 4\right)^2$$

while one has from Theorem 4.5

$$\kappa \;=\; \min\left(\frac{\sqrt{m_2 c_P}}{12}, \frac{\underline{\eta}}{6(1+R^2)}\right).$$

Recall $m_2 = d$ and $m_4 = d(d+2)$. Let us show that the minimum is always given by the second term. Using that $C_1, C_2 \geqslant 0$, we simply bound

$$\frac{6(1+R^2)}{\underline{\eta}} \;\geqslant\; \frac{6R^2}{\overline{\eta}} \;\geqslant\; \frac{6\left(1+16c_P^2\right)\left(\overline{\eta}/\sqrt{d}+5\right)^2}{c_P^2\overline{\eta}}.$$

Optimizing with respect to $\overline{\eta}$ we remark that $(\overline{\eta}/\sqrt{d}+5)^2/\overline{\eta} \geqslant 20/\sqrt{d}$. Moreover, we always have $(1+16c_P^2)/c_P^2 \geqslant 1/\sqrt{c_P}$, and thus $6(1+R^2)/\underline{\eta} \geqslant 120/\sqrt{dc_P}$. As a conclusion, $\kappa = \underline{\eta}/(6+6R^2)$. □

5 Simulation of velocity-jump processes

5.1 General strategy

The practical implementation of our velocity jumps processes rely on two assumptions:

i) the gradient $\nabla U(x)$ of the log-density can be computed numerically,

ii) some prior estimates on ∇U are given, typically its uniform norm or global Lipshitz constant.

For the sake of simplicity we only consider the case $\psi(s) = (s)_+$, although the extension to other cases is straightforward.

In order to simulate exactly a velocity jump-process we need some *a priori* information on the jump rate evolution.

Definition 5.1. *Let $\lambda(x,v) = \int_{v'} q(x,v;\mathrm{d}v')$ be the total jump rate of a velocity jump process. A function $\bar{\lambda} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^+$ is called a* prior rate upper bound *if*

$$\lambda(x+tv, v) \leq \bar{\lambda}(x,v,t) \qquad \forall x, v \in \mathbb{R}^d,\, t \in \mathbb{R}^+.$$

The simulation of the process is based on increasing the number of jumps at the price of adding uneffective (also called ghost) jumps. The jump times and velocities at those jump times, which determine the whole trajectory, are defined by induction. The simulation of the jumps then follows the algorithmic rules:

(i) At time t, compute $t + S$ the next jump time so that

$$\int_0^S \bar{\lambda}(X_t, V_t, s)\mathrm{d}s = E,$$

where E is independent unit exponentially distributed. The expression of the prior rate bound $\bar{\lambda}$ shall be sufficiently simple to compute S cheaply and exactly (up to round-off).

(ii) Note that $X_{(t+S)-} = X_t + SV_t$ and $V_{(t+S)-} = V_t$. With probability

$$\frac{\lambda(X_t + SV_t, V_t)}{\bar{\lambda}(X_t, V_t, S)}$$

sample a new velocity V_{t+S} according to the probability kernel $k(X_t + SV_t, V_t; \mathrm{d}v')$; else do not change velocity.

In the rest of this section, we present how a suitable prior rate upper bound can be established and how to sample according to k in the case of the Gaussian velocity jump samplers introduced in Theorem 3.4.

5.2 Bounds on the corrected rate

Consider the jump rate λ defined in Theorem 3.4. The bound $\mathbb{E}[(a + bG)_+] \leq (a)_+ + b/\sqrt{2\pi}$ yields

$$\lambda(x, v) \leqslant (v \cdot \nabla U(x))_+ + \frac{|\nabla U(x)|}{\sqrt{2\pi}\varepsilon(x)}\,.$$

Natural choices for $\varepsilon(x)$ are $\varepsilon(x) = \varepsilon_0|\nabla U(x)|$ (as this gives a uniform bound on the second part of the jump rate), $\varepsilon(x) = \varepsilon_0$ and $\varepsilon(x) = \varepsilon_0/(1 + |\nabla U(x)|)$ (for which, according to the discussion in Section 3.2, the degenerate Langevin term that appears as the first order error with respect to the Hamiltonian dynamics as $\varepsilon \to 0$ is then uniformly bounded in x). In any of those cases, a prior rate upper bound can be obtained from bounds on $v \cdot \nabla U(x + tv)$ and $|\nabla U(x + t)|$. Such bounds are easily obtained if ∇U is uniformly bounded by some known constant L, or if the the Hessian H of U is globally bounded in the Euclidean matrix norm, i.e. $M := \sup_{x\in\mathbb{R}^d} \|H(x)\|_2 < \infty$, in which case

$$v \cdot \nabla U(x + vt) \leqslant v \cdot \nabla U(x) + M|v|^2 t\,, \qquad |\nabla U(x + t)| \leqslant |\nabla U(x)| + M|v|t\,.$$

Each of the three choices of ε above yields a bound of the form

$$\lambda(x + tv, v) \leqslant \bar{\lambda}(x, v, t) := M|v|^2(t - t_0(x, v))_+ + a(x, v) + b(x, v)t^k$$

for some $k \in \{1, 2\}$ and $a, b, t_0 \geqslant 0$. Remark that, from the properties of the exponential law, then

$$S := \inf\left\{s > 0, \int_0^s \bar{\lambda}(X_t, V_t, s)\mathrm{d}s > E\right\}$$

has the same law as $S_1 \wedge S_2 \wedge S_3$ where, denoting $\bar{\lambda}_1 = M|v|^2(t - t_0)_+$, $\bar{\lambda}_2 = a$ and $\bar{\lambda}_3 = bt^k$,

$$S_i := \inf\left\{s > 0, \int_0^s \bar{\lambda}_i(X_t, V_t, s)\mathrm{d}s > E_i\right\}\,,$$

for $i = 1, 2, 3$, where E_1, E_2, E_3 are independent with unit exponential distribution. Here,

$$S_1 = t_0 + \sqrt{\frac{2E_1}{M|v|^2}}\,, \qquad S_2 = \frac{E_2}{a}\,, \qquad S_3 = \left(\frac{(k+1)E_3}{b}\right)^{\frac{1}{k+1}}\,.$$

5.3 Sampling according to the corrected kernel

5.3.1 General strategy

Consider the process defined in Theorem 3.4. Then, omitting in the notation the dependency of ε and T on x, the velocity after jump is $v - 2\varepsilon(\varepsilon v \cdot T + \tilde{G})T/(1+\varepsilon^2)$ where $\tilde{G}$ is a one-dimensional random variable with density

$$f_m(y) = \frac{1}{\Theta(m)\sqrt{2\pi}}(m+y)_+ \exp\left(-y^2/2\right),$$

where $m = \varepsilon v \cdot T$. We sample $\tilde{G}$ using rejection sampling, with various proposal distributions, depending on the value of the parameter m. In order to fix notations, we briefly recall the procedure. We look for a function g_m satisfying the two requirements:

1. g_m is a probability density from which we know how to sample;
2. there exists $C_m > 0$ such that for all x, $f_m(x) \le C_m g_m(x)$ and the ratio $f_m(x)/(C_m g_m(x))$ is computable.

The rejection sampling then consists in drawing Y according to g_m, and accepting it with probability $f_m(Y)/(C_m g_m(Y))$, and repeating until a proposal is accepted. It is well-known that this leads to a sample distributed according to f_m, and that the number of proposals needed is geometrically distributed with mean C_m.

5.3.2 Proposal distributions

We now list various choices for the proposal distribution with the corresponding computations; these choices are compared in terms of the expected number of trials and the CPU time in our implementation below.

Gamma proposal For $m < 0$, one can choose a $\Gamma(2, -m)$ proposal, shifted by $(-m)$:

$$g_m(y) = (y+m)_+(-m)^2 \exp\left(-(-m)(y+m)\right),$$

which is the distribution of $(-m)+(E_1+E_2)/(-m)$, where E_1 and E_2 are standard exponential random variables. This choice yields

$$\begin{aligned}\frac{f_m(y)}{g_m(y)} &= \frac{1}{\sqrt{2\pi}m^2\Theta(m)} \exp\left(-y^2/2 - m(y+m)\right) \\ &= \frac{1}{\sqrt{2\pi}m^2\Theta(m)} \exp\left(-(y+m)^2/2 - m^2/2\right),\end{aligned}$$

which is less than $C_m = \exp(-m^2/2)/(\sqrt{2\pi}m^2\Theta(m))$. A proposed value y is accepted with probability $\exp(-(y+m)^2/2)$, and the expected number of trials $C_m \to 1$ for $m \to -\infty$.

Exponential proposal Still for $m < 0$, we can use an exponentially distributed proposal, shifted by $(-m)$:

$$g_m(y) = \lambda \exp\left(-\lambda(y+m)\right) \mathbb{1}_{y>-m}.$$

The choice $\lambda = -m$ leads to simple bounds:

$$\frac{f_m(y)}{g_m(y)} = \frac{1}{(-m)\Theta(m)\sqrt{2\pi}}(m+y)_+ \exp\left(-y^2/2 - m(y+m)\right)$$

is maximized for $y = (-m) + 1$, so $f_m(y) \leq C_m g(y)$ where

$$C_m = \frac{1}{(-m)\Theta(m)\sqrt{2\pi}} \exp\left(-1/2 - m^2/2\right).$$

The acceptance probability in y is

$$\begin{aligned}\frac{f(y)}{C_m g(y)} &= (-m)(m+y)\exp\left(-y^2/2 - my - m^2 + 1/2 + m^2/2\right) \\ &= (-m)(m+y)\exp(1/2)\exp\left(-(y+m)^2/2\right).\end{aligned}$$

The constant $C_m \sim \exp(-1/2)(-m)$ is unbounded for $m \to -\infty$. However it behaves better than the Gamma proposal for small values of $|m|$.

Shifted Rayleigh proposal Consider once more the case $m < 0$. In the density f_m, $(m+y)_+$ is then bounded above by $y\mathbb{1}_{y>-m}$, leading to the bound

$$f_m(y) \leq \frac{1}{\Theta(m)\sqrt{2\pi}} \mathbb{1}_{y>-m} y \exp\left(-y^2/2\right) = C_m g_m(y),$$

where

$$C_m = \frac{\exp(-m^2/2)}{\sqrt{2\pi}\Theta(m)}, \qquad\qquad g_m(y) = \mathbb{1}_{y>-m} y \exp\left(m^2/2 - y^2/2\right)$$

It is easily checked that g_m is the distribution of $\sqrt{m^2 + 2E}$ for E an exponentially distributed random variable.

From the expansion $\mathbb{P}(G \geqslant x) \simeq \exp(-x^2/2)(1/x - 1/x^3)/\sqrt{2\pi}$ as $x \to \infty$, we get the asymptotic behaviour

$$C_m = m^2 + \underset{m\to-\infty}{o}(m^2)$$

implying that this choice is bad when $|m|$ is large. On the contrary, C_m converges to the optimal value 1 when m goes to 0_-.

Mixture between Rayleigh and Gaussian distribution We now turn to the case $m > 0$ and bound $(m+y)_+$ from above by $m + y\mathbb{1}_{y>0}$.

$$f_m(y) \leq \frac{1}{\Theta(m)\sqrt{2\pi}}(m + y\mathbb{1}_{y>0}) \exp\left(-y^2/2\right) := C_m g_m(y)$$

where $C_m = (m + 1/\sqrt{2\pi})/\Theta(m)$ and g_m is a probability density. One easily checks that g_m is the density of the mixture

$$\tilde{Y} = G\, \mathbb{1}_{U \leqslant m/(m+1/\sqrt{2\pi})} + \sqrt{2E}\, \mathbb{1}_{U > m/(m+1/\sqrt{2\pi})}$$

where G, E and U are independent and respectively distributed according to the standard Gaussian law, the standard exponential distribution and the uniform law over $[0, 1]$; it is therefore easy to sample. The proposal is accepted with probability $(m + Y)_+/(m + Y\mathbb{1}_{Y \geq 0})$.

The bound

$$\Theta(m) \geq \mathbb{E}((m + G)\mathbb{1}_{G \geq 0}) = \frac{m}{2} + \frac{1}{\sqrt{2\pi}}$$

shows that C_m is always less than 2 and converges to 1 when m vanishes. For $m \to \infty$, $\Theta(m) \sim m$ and $C_m \to 1$.

Gaussian proposal If $m > 0$, the mode of f_m is $\alpha = (\sqrt{m^2 + 4} - m)/2$. Let g_m be the density of the Gaussian random variable $\mathcal{N}(\alpha, 1)$. Then

$$\begin{aligned} \frac{f(x)}{g(x)} &= \frac{1}{\Theta(m)}(m + y)_+ \exp\left(-y^2/2 + (y - \alpha)^2/2\right) \\ &= \frac{1}{\Theta(m)}(m + y)_+ \exp\left(-\alpha y + \alpha^2/2\right). \end{aligned}$$

This is maximized for $y + m = 1/\alpha$, leading to the bound

$$\frac{f_m(x)}{g_m(x)} \leq C_m = \frac{1}{\Theta(m)\alpha} \exp\left(-\alpha^2/2\right).$$

The algorithm then consists in sampling from g_m and accepting with probability

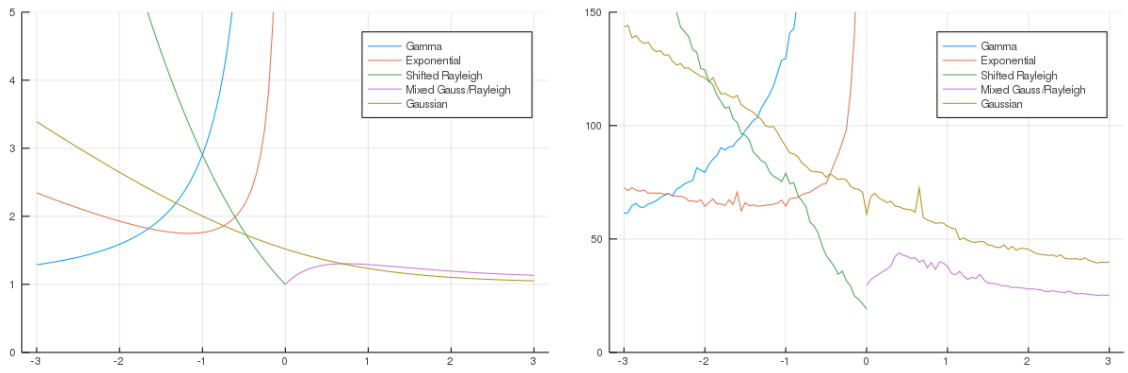
$$\frac{f(x)}{C_m g_m(x)} = \alpha(m + y)_+ \exp\left(-\alpha y + \alpha^2\right).$$

If m goes to infinity, $\alpha \sim 1/m$, so $C_m \sim m/\Theta(m) \to 1$. If m goes to zero, α goes to 1, and C_m to $\exp(-1/2)\frac{1}{\Theta(0)} = \sqrt{2\pi}\exp(-1/2) \approx 1.52$.

5.3.3 Choice of the proposal

We compare in Figure 1 the various choices for the proposal distributions, both theoretically and empirically. The best method depending on m will of course depend on implementation details; the important point is that by choosing an appropriate proposal we are able to keep the expected number of samples before acceptance C_m bounded. For our implementation we are led to the following choices.

m	Best proposal
$m \lesssim 2.5$	Gamma
$-2.5 \lesssim m \lesssim -1$	Exponential
$-1 \lesssim m \lesssim 0$	Rayleigh
$0 \lesssim m$	Mixed Rayleigh/Gaussian



On the left, we plot the value of C_m, the expected number of samples before acceptance, as a function of m, for the five proposal distributions discussed above. On the right we plot the empirical time (in nanoseconds) used by our implementation of the various methods. Note that the Gaussian proposal is in practice, for our implementation, a little slower than its competitors. From both point of views, the minimum of the curves stays uniformly bounded.

Figure 1: Comparison of proposal distributions

6 Numerical experiments

We provide in this section a numerical illustration for the very simple case of the two dimensional unit Gaussian distribution. We choose the precision parameter to be constant $\varepsilon(x) = \varepsilon$, and the simulated process is the velocity-jump process described in Lemma 3.3, without any additional noise on velocity.

Motivation Although this example may seem *a priori* naïve, it is motivated by the practical problem of sampling according to distributions with "multiscale" densities in Euclidean space. Indeed, near a local minimum, the potential (log-density) is approximately quadratic, which justifies the choice of the potential. Moreover, the few fastest time scales of the process – corresponding to stiffest directions of the local minimum – typically cannot be identified, and may be considered decoupled from: i) other degrees of freedom, and ii) additional noise on velocity which is usually restricted to the slowest time-scale. Those fastest degrees of freedom are the ones we arguably emulate here.

Simulation parameters Simulations are carried out with the following parameters:

- An initial condition $(x_0, v_0) \in \mathbb{R}^4$.
- A number of force evaluations $n \geq 1$.
- A quadratic potential of the form:

$$V(x) = x_1^2/2 + \lambda x_2^2/2$$

  with asymmetry parameter $\lambda \geq 1$. $\lambda = 1$ corresponds to the potential with (vectorial) isometry symmetry.

- A constant dynamical precision parameter $\varepsilon(x) = \varepsilon$ (see Lemma 3.3).

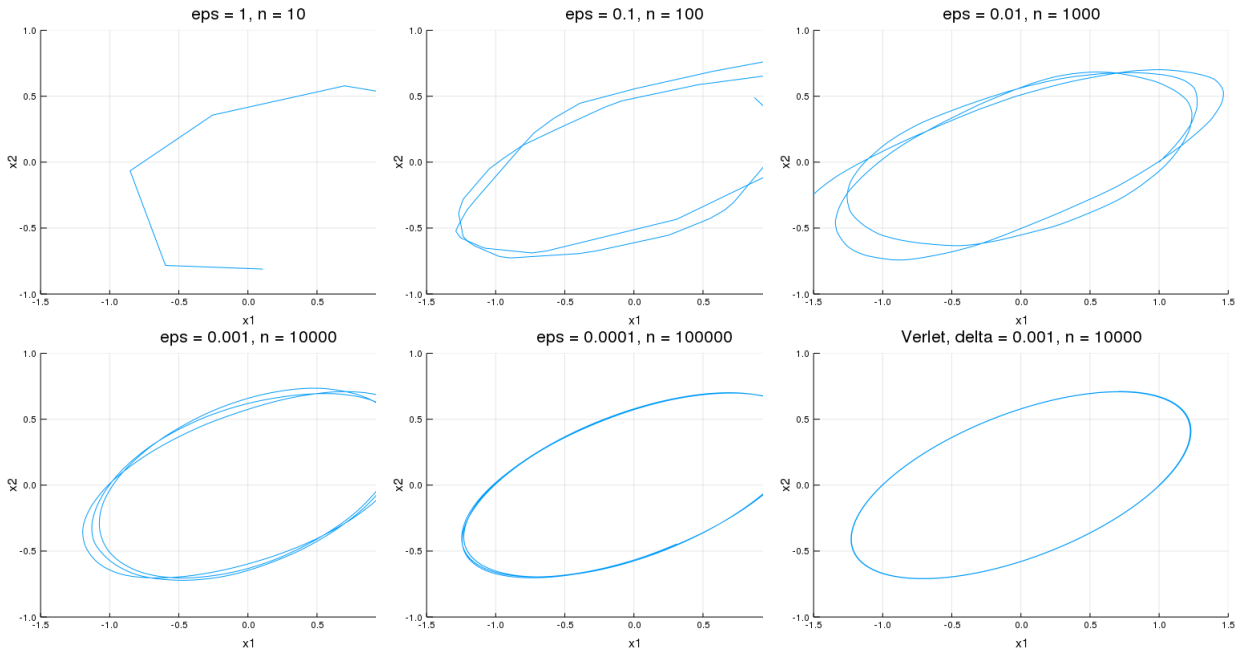


Figure 2: Examples of trajectories with the same (approximate) time length. The velocity-jump process is compared to the Hamiltonian limit computed with a Verlet scheme. Various ε are compared.

Irreducibility issues Without additional noise (which provides not only ergodicity but also exponentially fast mixing, see Section 4), the simulated velocity-jump process may not be irreducible with respect to the normal distribution (see Section 3.4). In the present section, we will observe the following two cases.

- The invariant distribution is the unit normal distribution, hence it is invariant by origin preserving isometries. In that case, the process is not irreductible, and it is easy to check that $t \mapsto X_t \wedge V_t$ is constant through time ($x \wedge v = x_1 v_2 - x_2 v_1$ in an orthonormal basis so that $x \wedge v = 0$ if and only if x and v are collinear). The process seems to be irreducible with respect to the unit normal (X, V) conditioned by $X \wedge V = x_0 \wedge v_0$ and $X, V \in \mathrm{Vect}(x_0, v_0)$ where (x_0, v_0) are the initial conditions of the process.
- The invariant distribution is an asymmetric normal distribution, and the process seems to be irreducible in dimension 2 in that case.

Rigorous analysis of irreducibility issues without additional noise is left for future work.

Results — short trajectories In Fig.2 and 3 we plot short/medium time trajectories for $\lambda = 1$ (the unit, symmetric quadratic potential $|x|^2$) and initial condition is $x_0 = (1, 0)$, $v_0 = (1, 1)$. Total physical time is (roughly) constant, so that the number of force evaluations increases with the precision parameter ε. We observe that when $\varepsilon \to 0$, trajectories indeed converge to the expected Hamiltonian dynamics of a two dimensional harmonic oscillator (integrated with a Verlet scheme here).

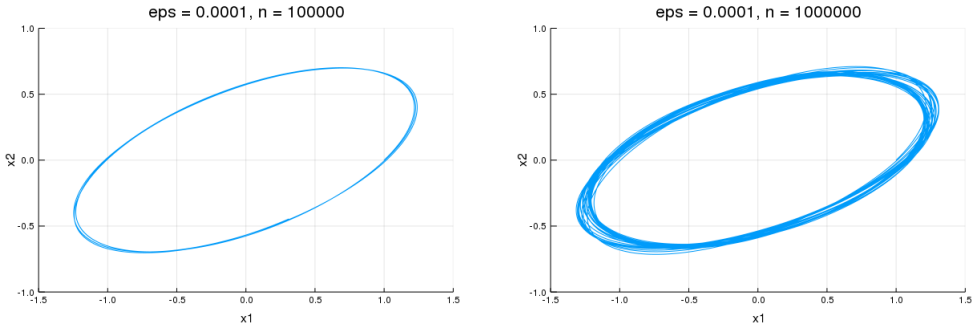


Figure 3: Same as 2 but for a longer trajectory

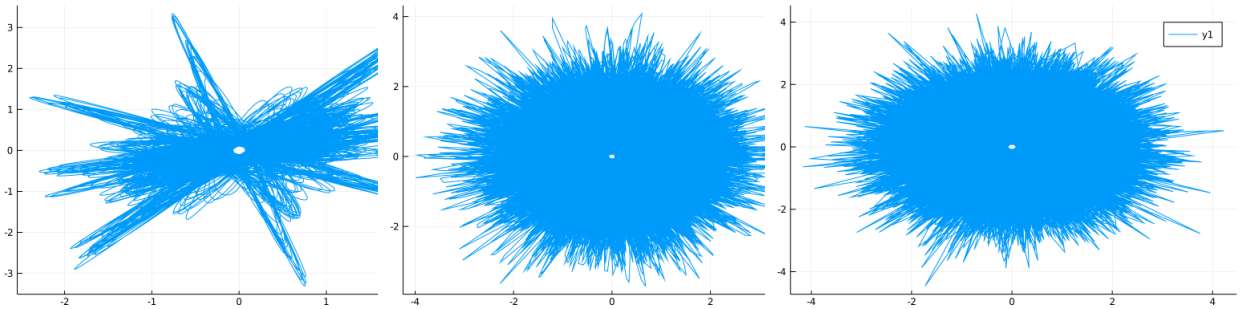


Figure 4: Examples of long trajectories for the (non-irreducible) unit Gaussian for, from left to right, $\varepsilon \in \{.01, 1, 100\}$.

Results — long non-ergodic trajectories In Fig.4 we plot long time trajectories for $\lambda = 1$ (the unit, symmetric quadratic potential $|x|^2$), initial condition $x_0 = (0, 0.5)$, $v_0 = (0.5, 0)$, and total number of force evaluations $n = 10^5$. The expected non-ergodicity is observed.

Results – mixing In Fig.5, we fix the initial condition $x_0 = (0, 0.5)$, $v_0 = (0.5, 0)$, and the number of force evaluations $n = 10^5$. We consider the position observable given by the time average of the square distance to the origin

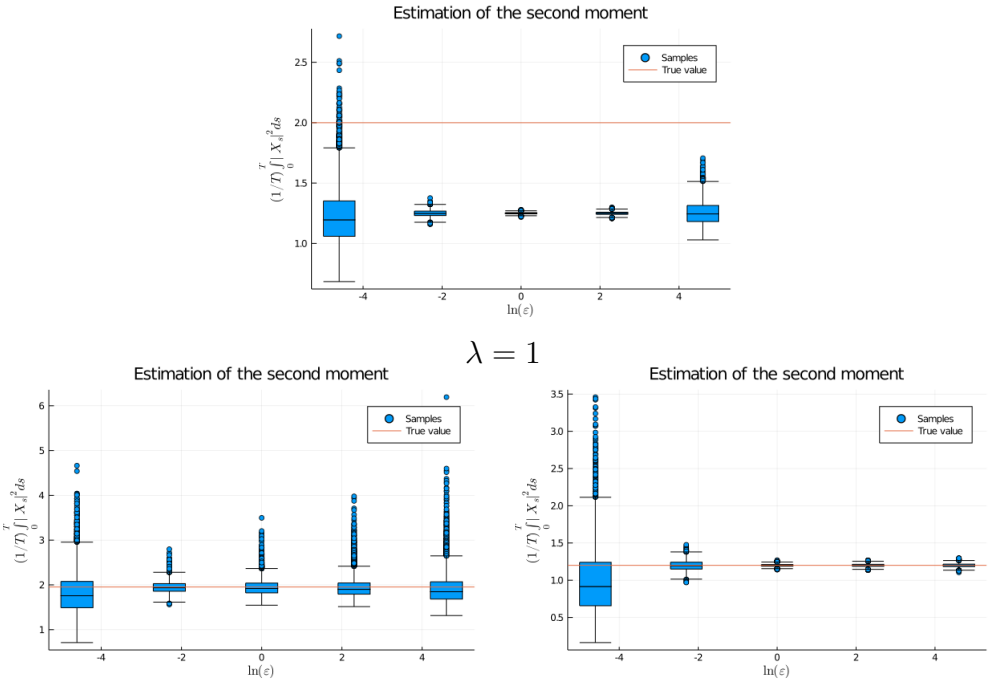
$$\frac{1}{T}\int_0^T |X_t|^2 \, dt.$$

For this observable, we compare the mixing efficiency for various $\varepsilon \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ and $\lambda \in \{1, 1.05, 5\}$ using various independent samples obtained by simulating the velocity-jump process. Let us recall that $\varepsilon = 0$ corresponds to the Hamiltonian dynamics, while $\varepsilon = +\infty$ is exactly the bouncy sampler. The figure consists of three (left, right, bottom) groups of box plots of those samples (each corresponding to a value of λ), the horizontal axis being $\ln \varepsilon$.

Several remark and results:

- As expected, for $\lambda = 1$ (and in this case only), the process is not irreducible and the sample is biased. A quick calculation shows that if $(X, V) \in \mathbb{R}^4$ is unit Gaussian then

$$\mathbb{E}(|X|^2 | X \wedge V = c) = 1 + cK_1(c)/K_0(c)$$



$\lambda = 1$

$\lambda = 1.05$ $\lambda = 5$

Figure 5: Box plots of samples obtained with fixed number of force evaluation $n = 10^5$. Comparison between: $\varepsilon \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ on the horizontal axis, as well as symmetric versus asymmetric potentials — eigenvalues ratio 1 (up chart), 1.05 (left chart) and 5 (right chart). Observe: i) the bias due to lack of ergodicity in the symmetric case, iii) a decrease of variance in the $\lambda = 5$ very asymmetric case, and iii) an efficiency which seems optimal for various non-extremal values of ε.

where K denotes the modified Bessel special function of the second kind. With our choice of initial conditions, $c = 1/4$ and the above quantity is roughly 1.224 which is consistent with the observed bias.

- For $\lambda = 1.05$, the proximity to $\lambda = 1$ where the breakdown of irreducibility (conservation law) occurs, seems to result in a larger variance than the other cases.
- For very small value of ε, the process: i) is simulated on comparatively shorter timescale due to the required precision, ii) is closed to an Hamiltonian dynamics which possesses additional conserved quantities (in particular energy). This translates into a poor mixing and thus a larger variance.
- We observe that the optimal sample quality is obtained for various intermediate values of ε ($\varepsilon = 1$ or even lower, that is full tangential resampling or closer to the Hamiltonian limit). These intermediate cases seems to consistently outperform the bouncy sampler ($\varepsilon = +\infty$).

Conclusion The process exhibits irreducibility issues in the presence of radial symmetries that are similar to the ones of the bouncy sampler. A moderate addition of velocity noise is thus recommended in general. The optimal sampling efficiency seems to be obtained for intermediate values of ε, around 1 or a bit lower (closer to the Hamiltonian limit than the full resampling, but not too much). However, this particular optimal value seems to vary with the target model and requires further and more exhaustive analysis.

7 Supplementary material

In this section, we establish a general result on the convergence of a family of Markov processes, Theorem 7.1, which is used in the proof of Theorem 3.6.

Consider a family $(L_\varepsilon)_{\varepsilon>0}$ of Markov generators on $\mathbb{R}^d$, and Markov processes (X_t^ε) associated to these generators by a martingale problem. There is a large literature (a reference monograph we will abundantly use here is [13]) linking convergence properties of (L_ϵ) with a convergence at the level of stochastic processes. Our purpose is to provide a simple generic setting in which checking the convergence of L_ϵ to a limiting generator L *locally* is enough to imply weak convergence at the process level. The classical 'weak' (convergence in distribution) approach of [13] relies on characterization of Markov processes by their generator through martingale problems (see below). Applying the convergence of generators at the level of the martingale problem typically enables to obtain tightness of the process distributions, extract a limit from them, and identify it.

In order to state the result, let us briefly recall that if E is a Polish state space, the set of càdlàg trajectories indexed by $\mathbb{R}_+$ may be equipped with the Skorokhod topology, forming a Polish space denoted by $\mathbb{D}_E$ (Section 5 and 6, Chapter 3 of [13]). We also recall that a sequence of càdlàg trajectories x^n converges to x in $\mathbb{D}_E$ if, on any finite time interval, it converges uniformly up to a uniformly small time change.

Theorem 7.1. *Let $\left((X_t^\varepsilon)_{t\geq 0}\right)_{\varepsilon>0}$ denotes a family of càdlàg processes in $\mathbb{R}^d$ with initial distribution μ. Assume the following:*

- *For each ε, $(X_t^\varepsilon)_{t\geq 0}$ solves the martingale problem associated with $\left(\mu, L_\varepsilon, C_c^\infty(\mathbb{R}^d)\right)$.*
- *For all $\varphi \in C_c^\infty(\mathbb{R}^d)$, $L_\varepsilon\varphi$ converges to $L\varphi$ uniformly on compacts.*
- *$L\varphi$ is continuous and the martingale problem associated with $\left(\mu, L, C_c^\infty(\mathbb{R}^d)\right)$ is well-posed in $\mathbb{R}^d$ (in particular the solution exists for all time) for any initial probability distribution μ.*

Then X^ε converges in distribution in the Skorohod space towards the unique solution of the limiting martingale problem.

Remark 7.2. The case of $\mathbb{R}^d$ could be easily generalized to any locally compact Polish space.

Proof. The proof uses heavily the classical technical apparatus developed in [13]. Let us first give an outline of the strategy before going into details.

The key point in order to use the "local" convergence of L_ϵ to L is to stop the processes when they leave large compact sets of $\mathbb{R}^d$, say balls defined by

$$B_r \stackrel{\text{def}}{=} \{x,\, |x| \leq r\}\,,$$

and to remark that the family of stopped processes is tight with respect to the Skorohod topology. Using the convergence of L_ε to L, any limit of extracted ε-sequences is then shown to coincide, when stopped, with the unique solution of a stopped martingale problem associated with L. In the last step, stopping the processes outside an appropriate ball, the global convergence is established.

Let us now give details on these three steps.

Tightness for stopped processes. If $F \subset E$ is closed, and $x \in \mathbb{D}_E$ we consider the hitting time

$$\tau(F) = \tau(F, x) \stackrel{\text{def}}{=} \inf\{t \geq 0,\, |x_t| \in F \,\text{or}\, x_{t^-} \in F\} \in [0, +\infty],$$

which is a stopping time for the canonical natural filtration of the Borel sets of $\mathbb{D}_E$. We fix an $r > 0$, let $F = \{x : |x| \geq r\}$ and consider $X^{\varepsilon,F}$ the stopped process

$$X^{\varepsilon,F}(t) = X^\varepsilon(t \wedge \tau(F)).$$

The goal of this first step is to prove that $(X^{\varepsilon,F})_{\varepsilon>0}$, whose trajectories stay in the bounded set $\{x : |x| \leq r\}$, is tight. The proof follows a very classical pattern; we sketch it using [13] as reference for the sake of completeness. Details can be found in [21], Section 3.2.

Using [13, Theorem 9.1, Chapter 3, p.142], tightness is equivalent to the tightness in $\mathbb{D}_\mathbb{R}$ of $\left(\varphi(X^{\varepsilon,F})\right)_{\varepsilon>0}$ for each $\varphi \in C_c^\infty(\mathbb{R}^d)$. Fix $\varphi \in C_c^\infty(\mathbb{R}^d)$. Classically: i) expand squares of the form $\left(\varphi(X_{t+h}^{\varepsilon,F}) - \varphi(X_t^{\varepsilon,F})\right)^2$; ii) consider the two (stopped) martingales associated with $\varphi(X_t^{\varepsilon,F})$ and $\varphi^2(X_t^{\varepsilon,F})$; and iii) use the uniform boundedness on compacts of $L_\varepsilon\varphi$ and $L_\varepsilon\varphi^2$ (which follows from the convergence assumption). Standard tightness criteria like [13, Theorem 8.6, Chapter 4, p.137] enables to conclude.

Identification of the limit through a stopped martingale problem. Let $X^n = X^{\epsilon_n,F}$ be an arbitrary convergent subsequence of $X^{\epsilon,F}$, and call its limit Y. Call X a solution of the limit martingale problem; recall that we assume well-posedness so X is unique in distribution.

We claim that the stopped process $Y^\tau = Y(t \wedge \tau(F))$ solves a *stopped martingale problem* ([13, Section 6, Chapter 4]): for any φ,

$$\varphi(Y(t \wedge \tau(F))) - \varphi(Y(0)) - \int_0^{t\wedge\tau(F)} L\varphi(Y(s))ds \tag{7.1}$$

is a martingale with respect to the natural filtration of Y. By [13, Theorem 6.1 p. 217 Ch.4], there is a unique solution of this stopped martingale problem, namely the distribution of X stopped at F, so that:

$$Y^\tau \overset{(d)}{=} X^\tau \overset{\text{def}}{=} X(\cdot \wedge \tau(X, F)). \tag{7.2}$$

Let us now justify the claim. By Lemma 7.4 below, there exists a sequence δ_n such that, denoting by F_n the δ_n-neighborhood $F(\delta_n)$ of F, the sequence $(X^n, \tau(F_n, X^n))$ converges in distribution towards $(Y, \tau(F, Y))$. In particular, as can be seen by a Skorohod almost sure representation of the latter convergence, we get the convergence in distribution of $\tilde{X}^n(\cdot) = X^n(\cdot \wedge \tau(F_n, X^n))$:

$$(\tilde{X}^n, \tau(F_n, \tilde{X}_n)) \underset{(d)}{\longrightarrow} (Y^\tau, \tau(F, Y)).$$

Now $\tilde{X}^n$ solves the martingale problem associated to L_{ϵ_n}, stopped at time $\tau(F_n)$: let us briefly see how to send n to infinity and justify the claim.

By [13, Lemma 7.7, Chapter 3, p. 131] there exists a dense subset of times $\mathcal{C} \subset \mathbb{R}$ where the limit Y^τ is continuous with probability one. Let t_1,... t_{K+1} be arbitrary times in $\mathcal{C}$ and φ, φ_1, ... φ_K be bounded test functions. By definition of the stopped martingale problem solved by $\tilde{X}^n$ and the characterization of martingales given in [13, p.174],

$$\mathbb{E}\left[\left(\varphi(\tilde{X}^n(t_{K+1})) - \varphi(\tilde{X}^n(t_K)) - \int_{t_K\wedge\tau(F_n,\tilde{X}^n)}^{t_{K+1}\wedge\tau(F_n,\tilde{X}^n)} L_{\varepsilon_n}\varphi(\tilde{X}^n(s))ds\right)\prod_{k=1}^K \varphi_k(\tilde{X}^n(t_k))\right] = 0. \tag{7.3}$$

The left-hand side may be written as $\mathbb{E}\left[\Phi(\tilde{X}^n, \tau(F_n, \tilde{X}^n))\right]$ for some function Φ. Remarking by dominated convergence that since $L\varphi$ is continuous, integrals of the form $x \mapsto \int_0^t L\varphi(x_s)ds$ are continuous with respect to the Skorokhod topology, and since the t_k are in $\mathcal{C}$, Φ is almost surely continuous at the limit $(Y^\tau, \tau(F, Y))$. This justifies taking the limit in (7.3), which yields

$$\mathbb{E}\left[\left(\varphi(Y^\tau(t_{K+1})) - \varphi(Y^\tau(t_K)) - \int_{t_K\wedge\tau(F,Y)}^{t_{K+1}\wedge\tau(F,Y)} L\varphi(Y^\tau(s))ds\right)\prod_{k=1}^K \varphi_k(Y^\tau(t_k))\right] = 0.$$

Using again the previously mentioned characterization of martingales, this entails that Y^τ indeed satisfies the martingale problem with generator L, stopped at time $\tau = \tau(F, Y)$.

Convergence of the original processes. We fix a bounded continuous observable $\Psi(x) = \Psi(x_s, 0 \le s \le T)$ on $\mathbb{D}_E$ mesurable with respect to paths restricted to a given finite time interval $[0, T]$. Since the limit martingale problem is assumed to be well-posed in $\mathbb{R}^d$, the solution X exists for all time, and thus for each $\eta > 0$ there exists a $r = r(T, \eta)$ such that denoting $F = \{x, |x| \ge r\}$,

$$\mathbb{P}(\tau(F, X) \le 2T) \le \eta.$$

Our goal is to prove that $|\mathbb{E}[\Psi(X^\varepsilon)] - \mathbb{E}[\Psi(X)]| \to 0$, or in other words that

$$D \stackrel{\text{def}}{=} \limsup_{\epsilon \to 0} |\mathbb{E}[\Psi(X^\varepsilon)] - \mathbb{E}[\Psi(X)]|$$

is zero. Let us extract a sequence $X^n = X^{\epsilon_n}$ such that $D = \lim_n |\mathbb{E}[\Psi(X^n)] - \mathbb{E}[\Psi(X)]|$; up to extracting a further subsequence we may assume by tightness that $X^{n,F}$ converges in distribution. By (7.1) and (7.2) from the previous step, there exists a sequence (δ_n) such that, for $F_n = F(\delta_n)$,

$$(\tilde{X}^n, \tau(F_n, \tilde{X}^n)) \xrightarrow{(d)} (X^\tau, \tau(F, X)), \tag{7.4}$$

where $\tilde{X}^n = X^{n,F_n}$ is the process stopped when it reaches F_n. Since $X^n(t) = \tilde{X}^n(t)$ when $t < \tau(F_n, X_n)$,

$$\begin{aligned} |\mathbb{E}[\Psi(X^n)] - \mathbb{E}[\Psi(X)]| \leq & \left| \mathbb{E}\left[\Psi(\tilde{X}^n)\mathbb{1}_{\tau(F_n,\tilde{X}^n)>T}\right] - \mathbb{E}\left[\Psi(X^\tau)\mathbb{1}_{\tau(F,X)>T}\right] \right| \\ & + \|\Psi\|_\infty \mathbb{P}(\tau(F_n, \tilde{X}^n) \leq T) \\ & + \|\Psi\|_\infty \mathbb{P}(\tau(F, X) \leq T) \end{aligned}$$

By (7.4), the first term vanishes in the limit so $D \leq 2\|\Psi\|_\infty \eta$. Since η is arbitrary, D must be zero, concluding the proof of convergence. □

The above proof uses a technical result to handle the fact that, for a given closed set F, the map $x \mapsto \tau(F, x)$ is only lower semicontinuous with respect to the Skorokhod topology. To understand what may go wrong, consider X^n the deterministic motion in $\mathbb{R}$ that goes upwards or downwards at speed one and is reflected on the boundary of $[-1+1/n, 1-1/n]$: for $F = \mathbb{R}\backslash]-1, 1[$, the hitting time of F is infinite for X^n but finite for the limiting process X; in particular, the stopped process $X^n(t \wedge \tau(\{-1, 1\}, X^n)) = X^n(t)$ does *not* converge to $X(t \wedge \tau(\{-1, 1\}, X))$.

We first prove a deterministic result showing that we may almost recover continuity by considering δ-neighborhoods of F, $F(\delta) = \{x : d(x, F) \leq \delta\}$.

Lemma 7.3. *Suppose that $x^n \to x$ in $\mathbb{D}_E$ and let F be a closed set. Then $\delta \mapsto \tau(F(\delta), x)$ is decreasing and*

$$\limsup_n \tau(F(\delta), x^n) \leq \tau(F, x) \leq \liminf_n \tau(F, x^n), \tag{7.5}$$

$$\tau(F(\delta), x) \xrightarrow[\delta \to 0]{} \tau(F, x). \tag{7.6}$$

Consequently for any sequence $\delta_n \to 0$,

$$\liminf_n \tau(F(\delta_n), x^n) \geq \tau(F, x). \tag{7.7}$$

Note that we can only expect to get a statement on the liminf if the sequence δ_n is arbitrary: indeed,in the example detailed above, whether $\limsup_n \tau(F(\delta_n), X^n) \leq \tau(F, X)$ depends on how δ_n compares to $1/n$.

Proof. If $\tau(F,x) > t$ then the compactified trajectory $\Gamma = \overline{x([0,t])}$ is entirely contained in F^c; by compactness $\Gamma(\delta)$ is also contained in F^c for δ small enough. For any $t' < t$, by definition of the Skorokhod topology the compactified trajectory $\Gamma_n = \overline{x^n([0,t'])}$ is included in $\Gamma(\delta)$ for n large enough, so $\tau(F,x^n) \geq t'$ for n large enough, proving the second inequality in (7.5).

Similarly, fixing δ and $t > \tau(F,x)$, we get that $x(\tau(F,x)) \in F$ so that for n large enough, $x^n(t_n) \in F(\delta)$ for some $t_n \leq t$; in other words $\tau(F(\delta),x^n) \leq t$ for n large enough. Therefore $\limsup \tau(F(\delta),x^n) \leq \tau$, completing the proof of (7.5) since $t > \tau(F,x)$ is arbitrary.

We now prove (7.6). Clearly if $F \subset G$ then $\tau(F,x) \geq \tau(G,x)$, so $\delta \mapsto \tau(F(\delta),x)$ decreases. Let δ_n be a sequence decreasing to zero: $\tau_n = \tau(F(\delta_n),x)$ is increasing. Let τ_∞ be its limit; since $\tau_n \leq \tau(F,x)$, $\tau_\infty \leq \tau(F,x)$. If $\tau_\infty = \infty$ then $\tau_\infty = \tau(F,x)$. If τ_∞ is finite, for each n one of $x(\tau_n)$ or $x((\tau_n)_-)$ is in $F(\delta_n)$: call it y_n. By compactness of $\overline{x([0,\tau_\infty])}$ y_n must converge; its limit is in $\cap_n F(\delta_n) = F$, and is either $x(\tau_\infty)$ or $x((\tau_\infty)_-)$, so $\tau(F,x) \leq \tau_\infty$ and once more they are equal.

Suppose δ_n converges to 0. Fix a $\delta > 0$. For n large enough, $\delta_n \leq \delta$ so $\tau(F(\delta_n),x^n) \geq \tau(F(\delta),x^n)$. Taking limits we get

$$\liminf \tau(F(\delta_n),x^n) \geq \liminf \tau(F(\delta),x^n) \geq \tau(F(\delta),x),$$

using (7.5). Taking δ to zero and using (7.6) yields Equation (7.7). □

The following probabilistic corollary shows that δ_n may be chosen to decay slowly enough so that the hitting times converge.

Lemma 7.4. *Suppose that X^n converges in distribution to X. For any closed set F, there exists a sequence of radii $(\delta_n)_{n\geq 0}$ such that*

$$(X^n, \tau(F(\delta_n),X^n)) \xrightarrow[n\to\infty]{(d)} (X,\tau(F,X)).$$

Proof. By the Skorokhod representation theorem we may assume without loss of generality that X^n converges almost surely to X; it is then enough to construct $(\delta_n)_{n\geq 0}$ such that $\tau(F(\delta_n),X^n)$ converges in probability to $\tau(F,X)$. By Lemma 7.3, we have almost surely, for any sequence $(\delta_n)_{n\geq 0}$, $\liminf \tau(F(\delta_n),X^n) \geq \tau(F,X)$. To prove the upper bound, we fix $\varepsilon > 0$, and it remains to show that we can construct a sequence (δ_n) with

$$\lim_n \mathbb{P}\left[\tau(F(\delta_n,X^n)) > \tau(F,X) + \epsilon\right] = 0.$$

Define the events

$$A(n,\delta) \stackrel{\text{def}}{=} \left\{\tau(F(\delta),X^m) \leq \tau(F,X) + \delta, \forall m \geq n\right\},$$

and say that (n,δ) is good if $\mathbb{P}(A(n,\delta)) \geq 1-\delta$. It is easily checked that goodness is doubly monotonous:

$$(n' \geq n, \delta' \geq \delta, (n,\delta) \text{ good }) \implies (n',\delta') \text{ good.}$$

Now, for any fixed $\delta > 0$, the events $\{A_{n,\delta}\}_{n\geq 1}$ form an increasing sequence, and

$$\bigcup_{n\geq 0} A_{n,\delta} = \left\{\limsup_{n\geq 0} \tau(F(\delta),X^n) \leq \tau(F,X) + \delta\right\}$$

has probability one by (7.5). As a consequence, for each $\delta > 0$, there is a finite $n(\delta)$ such that $(n(\delta), \delta)$ is good, for instance, $n(\delta) = \min\{n \geq 1 | (n, \delta) \text{ is good}\}$; and using the monotony of goodness, one can then easily construct a decreasing sequence $(\delta(n))_{n\geq 0}$ that decreases to zero and such that (n, δ_n) is good for each $n \geq 1$.

Finally, on $A(n, \delta_n)$, $\tau(F(\delta_n, X^n)) \leq \tau(F, X) + \delta_n$, so for any $\epsilon > 0$, and for n large enough to ensure $\delta_n \leq \epsilon$,

$$\mathbb{P}\left[\tau(F(\delta_n, X^n)) > \tau(F, X) + \epsilon\right] \leq \mathbb{P}\left[A(n, \delta_n)^c\right] \leq \delta_n \xrightarrow[n\to\infty]{} 0,$$

concluding the proof that $\tau(F(\delta_n), X^n)$ converges to $\tau(F, X)$ in probability. □

References

[1] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. I. General method. *J. Chem. Phys.*, 31:459–466, 1959.

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

[3] C. Andrieu and S. Livingstone. Peskun-tierney ordering for markov chain and process monte carlo: beyond the reversible scenario, 2019.

[4] J. Bierkens and G. Roberts. A piecewise deterministic scaling limit of lifted Metropolis-Hastings in the Curie-Weiss model. *Ann. Appl. Probab.*, 27(2):846–882, 2017.

[5] J. Bierkens, G. Roberts, and P.-A. Zitt. Ergodicity of the zigzag process. *ArXiv e-prints*, December 2017.

[6] X. Cheng, N.S. Chatterji, P.L. Bartlett, and M.I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *COLT*, 2018.

[7] E. Brian Davies. *Spectral Theory and Differential Operators*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

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

[9] G. Deligiannidis, D. Paulin, A. Bouchard-Côté, and A. Doucet. Randomized hamiltonian monte carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates, 2018.

[10] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for kinetic equations with linear relaxation terms. *C. R. Math. Acad. Sci. Paris*, 347(9-10):511–516, 2009.

[11] A. Durmus, A. Guillin, and P. Monmarché. Geometric ergodicity of the bouncy particle sampler. *arXiv e-prints*, page arXiv:1807.05401, Jul 2018.

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

[13] 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.

[14] D.A. Gibson and E.A. Carter. Time-reversible multiple time scale ab initio molecular dynamics. *The Journal of Physical Chemistry*, 97:13429–13434, 1993.

[15] M. Michel, A. Durmus, and S. Sénécal. Forward Event-Chain Monte Carlo: Fast sampling by randomness control in irreversible Markov chains. *arXiv e-prints*, page arXiv:1702.08397, Feb 2017.

[16] L. Miclo and P. Monmarché. Étude spectrale minutieuse de processus moins indécis que les autres. *Lecture Notes in Mathematics*, 2078:459–481, September 2012.

[17] P. Monmarché. Kinetic walks for sampling. *to appear in ALEA*, 2020.

[18] 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. *arXiv e-prints*, page arXiv:2002.07109, Feb 2020.

[19] M. Ottobre, N. S. Pillai, F.J. Pinski, and A. M. Stuart. A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, 22(1):60–106, 2016.

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

[21] Mathias Rousset, Yushun Xu, and Pierre-André Zitt. A weak overdamped limit theorem for langevin processes. *ALEA*, 2019.

[22] M.E. Tuckerman, B.J. Berne, and A. Rossi. Molecular dynamics algorithm for multiple time scales: Systems with disparate masses. *J. Chem. Phys.*, 94, 1991.

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