MARKOV CHAIN MONTE CARLO AND IRREVERSIBILITY

M. OTTOBRE

Abstract. Markov Chain Monte Carlo (MCMC) methods are statistical methods designed to sample from a given measure π by constructing a Markov Chain that has π as invariant measure and that converges to π. Most MCMC algorithms make use of chains that satisfy the detailed balance condition with respect to π; such chains are therefore reversible. On the other hand, recent work [18, 21, 28, 29] has stressed several advantages of using irreversible processes for sampling. Roughly speaking, irreversible diffusions converge to equilibrium faster (and lead to smaller asymptotic variance as well). In this paper we discuss some of the recent progress in the study of non-reversible MCMC methods. In particular: i) we explain some of the difficulties that arise in the analysis of non-reversible processes and we discuss some analytical methods to approach the study of continuous-time irreversible diffusions; ii) most of the rigorous results on irreversible diffusions are available for continuous-time processes; however, for computational purposes one needs to discretize such dynamics. It is well known that the resulting discretized chain will not, in general, retain all the good properties of the process that it is obtained from. In particular, if we want to preserve the invariance of the target measure, the chain might no longer be reversible. Therefore iii) we conclude by presenting an MCMC algorithm, the SOL-HMC algorithm [23], which results from a non-reversible discretization of a non-reversible dynamics.

Keywords. Markov Chain Monte Carlo, non-reversible diffusions, Hypocoercivity, Hamiltonian Monte Carlo.

1. Introduction

The combined use of Bayesian statistics and Markov Chain Monte-Carlo (MCMC) sampling methods has been one of the great successes of applied mathematics and statistics in the last 60 years. While the Bayesian approach constitutes a flexible framework for inference through data assimilation, MCMC turns such a theoretical framework into practice by providing a powerful sampling mechanism to extract information from the posterior measure. For this reason, and because of the wide spectrum of problems that can be recast in Bayesian terms, MCMC has been a revolution in the applied sciences. MCMC is employed in parameter estimation, model validation and, ultimately, in inference. Combined with the Bayesian inference paradigm, MCMC is of current use in finance, biology (population genetics, molecular biology), meteorology, epidemiology, optimization, cryptography, molecular dynamics, computational physics (to gain knowledge about statistical quantities of interest in the study of large particle systems in their equilibrium state), in rare event sampling, in big data analysis and in the field of inverse problems. This list is far from exhaustive.

The increasing popularity of MCMC and the need to tackle problems of growing complexity have brought higher demands on the efficiency of such algorithms, which are often undeniably costly. The answer to such demands has produced both a higher level of sophistication in the design of MCMC algorithms and the introduction of a plethora of different approaches. We would however be very unfair to MCMC if we described it as a mere algorithmic tool: the study of MCMC has in fact opened (or it is related to) a range of beautiful questions in an

unimmaginable wide range of areas of mathematics, from pure probability to analysis, all the way to number theory [11, 36].

The purpose of MCMC is to sample from a given target distribution π or, more commonly, to calculate expectations with respect to π, i.e. integrals of the form

$$\int_{\chi} f(x)d\pi(x)\,, \tag{1.1}$$

when analytic (or deterministic) methods are not feasible. Here π and f are a measure and a function, respectively, both defined on the state space χ. Broadly speaking, the calculation of integrals (1.1) is of interest in the applied sciences for several reasons: i) for different choices of the function f, such integrals represent various statistical properties of a system in equilibrium (with stationary measure π) or properties of the posterior measure, π, in a Bayesian context; ii) if X_t is the solution at time t of a given stochastic differential equation (SDE), then the expectation

$$\mathbb{E}[f(X_t)] \tag{1.2}$$

can be recast in the form (1.1); iii) thanks to the Feynman-Kac formula, integrals of type (1.1) are representations of the solution of a large class of PDEs, as well.

Roughly speaking (we will be more precise in Section 4), the basic prescription behind MCMC can be explained as follows: construct a Markov Chain $\{x_n\}_{n\in\mathbb{N}}$ that converges to our target distribution π. In this way, if we run the chain "long enough", as $n \to \infty$ we will effectively be extracting samples from π. Also, if the chain we constructed enjoys good ergodic properties, the ergodic theorem can be employed, thereby providing an approximation for the quantity (1.1):

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(x_k) = \mathbb{E}_\pi(f) := \int_{\chi} f(x)d\pi(x). \tag{1.3}$$

In order for this process to work efficiently, the constructed chain should: i) converge to equilibirum as fast as possible (all the samples out of equilibrium are not needed); ii) once equilibrium is reached, explore the state space as quickly and thoroughly as possible. This paper intends to comment on some aspects related to point i). Regarding i): the classical MCMC framework - and in particular the popular Metropolis-Hastings (M-H) technique (see Section 4.1) - typically makes use of *reversible chains*, i.e. chains which satisfy the *detailed balance condition* with respect to π. However, it is a well documented principle that, loosely speaking, non-reversible chains might converge to equilibrium faster than reversible ones. We will be more clear on this matter in Section 3. For the moment let us just say that this observation has started to produce a stream of literature aimed at improving the speed of convergence to equilibrium of MCMC methods by designing algorithms that produce non-reversible chains. In this spirit, we will present an algorithm, recently introduced in [23], which does not belong to the M-H framework, as it produces a Markov chain which does not satisfy detailed balance with respect to the target measure. This is the SOL-HMC algorithm (Second Order Langevin- Hybrid Monte Carlo), presented in Section 5. In the present paper we will mostly be concerned with irreversibility and therefore we will only tangentially comment on another important aspect related to the SOL-HMC algorithm: SOL-HMC does not suffer from the so called *curse of dimensionality*. That is, the cost of the algorithm does not increase when the dimension of the space in which it is implemented increases. We will be more precise on this point in Section 4.2.

The remainder of the paper is organized as follows: in Section 2 we recall some basic definitions, mostly with the purpose of fixing the notation for the rest of the paper (references are given for those not familiar with the topic). Section 3 is devoted to the study of exponentially fast convergence to equilibrium for irreversible dynamics. The Markov dynamics presented

here, central to the development of the SOL-HMC algorithm, are hypoelliptic and irreversible; i.e. their generator is non-elliptic and non self-adjoint, so classical techniques do not apply; in order to study these degenerate dynamics the *Hypocoercivity Theory* has been recently introduced in [38]. Section 3 contains a short account of such an approach. Section 4 is devoted to an elementary introduction to MCMC, including the popular Random Walk Metropolis (RWM), Metropolis Adjusted Langevin Algorithm (MALA) and Hybrid (or Hamiltonian) Monte Carlo (HMC). The last section, Section 5, contains an example of an irreversible MCMC algorithm, the so-called SOL-HMC (Second-Order Langevin-Hamiltonian Monte Carlo), introduced in [23]. In this context we will explain how irreversibility can be obtained from the composition of Markov transition probabilities that do satisfy detailed balance.

2. Preliminaries and Notation

In this section we briefly recall some basic facts that will be used in the following. More details about the basic formalism introduced here can be found in [3, 25, 15, 12]. Consider an ordinary stochastic differential equation in $\mathbb{R}^d$ of the form

$$dx(t) = b(x_t)dt + \sigma(x_t)dW_t, \tag{2.1}$$

[1]

where W_t is a d-dimensional standard Brownian motion and the drift and diffusion coefficients ($b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, respectively) are globally Lipshitz. It is a standard fact that under these assumptions there exists a unique strong solution to the SDE (2.1). The solution $x(t)$ is a Markov diffusion process. Because b and σ are time-independent, $x(t)$ is a time-homogeneous Markov process.

To the process x_t we can associate a Markov semigroup as follows. For any function $f : \mathbb{R}^d \to \mathbb{R}$, say $f \in \mathcal{B}_m$, [2] and any point $x \in \mathbb{R}^d$, we can define

$$f(x,t) := \mathbb{E}\left[f(x_t)|x_0 = x\right] ,$$

where $\mathbb{E}$ denotes expected value (with respect to the noise W_t). Notice that the function f is a deterministic function. By using the Itô formula, one can immediately see that $f(x,t)$ solves the Cauchy problem

$$\begin{aligned} \partial_t f(x,t) &= \mathcal{L} f(x,t) \\ f(x,0) &= f(x), \qquad x \in \mathbb{R}^d, \end{aligned} \tag{2.2}$$

where $\mathcal{L}$ is the second order differential operator defined on smooth functions as

$$\mathcal{L} = \sum_{i=1}^{d} b^i(x)\partial_{x_i} + \frac{1}{2}\sum_{i,j=1}^{d} \Sigma_{ij}(x)\partial^2_{x_i x_j}, \qquad \Sigma(x) := \sigma(x)\sigma^T(x),$$

having denoted by σ^T the transpose of the matrix σ. The operator $\mathcal{L}$ is (under the assumptions of the Hille-Yoshida Theorem) the *generator* of the Markov semigroup $\mathcal{P}_t$ associated with the PDE (2.2); i.e., formally:

$$f(x,t) = e^{t\mathcal{L}} f(x) = (\mathcal{P}_t f)(x)\,.$$

With abuse of nomenclature, we will often refer to $\mathcal{L}$ as to the generator of the diffusion process (2.1). The standard example belonging to this framework is the heat semigroup: in this case the process $x(t)$ is simply Brownian motion (i.e. in (2.1) $b = 0$ and σ is the identity matrix) and the generator of the semigroup is the Laplacian operator.

[1] For any time-dependent process or function, we will use the notations h_t and $h(t)$ interchangeably.

[2] $\mathcal{B}_m := \{\text{bounded and measurable functions on } \mathbb{R}^d\}$

We recall that a probability measure μ on $\mathbb{R}^d$ is *invariant* for the Markov semigroup $\mathcal{P}_t$ if, for every $h \in \mathcal{B}_m$

$$\int_{\mathbb{R}^d} (\mathcal{P}_t h)(x)\mu(dx) = \int_{\mathbb{R}^d} h(x)\mu(dx)\,.$$

Using the dual semigroup, $\mathcal{P}_t'$, acting on measures, this can also be rewritten as $\mathcal{P}_t'\mu = \mu$ or $\mathcal{L}'\mu = 0$, where $\mathcal{L}'$ denotes the L^2-adjoint of $\mathcal{L}$.[3]

In view of the link between the Markov process x_t solution of the SDE (2.1) and the semigroup $\mathcal{P}_t$, every attribute of the semigroup will also hold for the process and viceversa, unless otherwise stated. So e.g. we say that the measure μ is invariant for the process $x(t)$ if it is invariant for the semigroup associated to $x(t)$. The measure μ is called invariant because if $x(0)$ is distributed according to μ, $x(0) \sim \mu$, then $x(t) \sim \mu$ for every $t \geq 0$. The process x_t is *ergodic* if it admits a unique invariant measure. In this case the only invariant measure is called the ergodic measure of the process and it represents the equilibrium state (in law) of the system.

Central to our discussion will be the definition of reversibility.

Definition 2.1. *A Markov semigroup $\mathcal{P}_t$ is* reversible *with respect to a probability measure μ (or, equivalently, the probability measure μ is reversible for the Markov semigroup $\mathcal{P}_t$) if for any $f, g \in \mathcal{B}_m$*

$$\int (\mathcal{P}_t f) g \, d\mu(x) = \int f (\mathcal{P}_t g) \, d\mu(x). \tag{2.3}$$

In this case it is also customary to say that $\mathcal{P}_t$ satisfies the detailed balance condition *with respect to μ.*

Notice that if μ is reversible then it is invariant as well. If x_t is reversible with respect to μ and $x(0) \sim \mu$ then for any $T > 0$ and any $0 \leq t_1 \leq \ldots \leq t_k < T$, the law of $(x_0, x_{t_1}, \ldots, x_{t_k}, x_T)$ is the same as the law of $(x_T, x_{T-t_1}, \ldots, x_{T-t_k}, x_0)$. In other words, the forward and the time-reversed process have the same law (on this matter see e.g. [25, Section 4.6]). It is easy to show that $\mathcal{P}_t$ is reversible with respect to μ if and only if the generator $\mathcal{L}$ is symmetric in L^2_μ, where

$$L^2_\mu := \left\{ \text{functions } f : \mathbb{R}^d \to \mathbb{C} \text{ such that } \int_{\mathbb{R}^d} f^2 d\mu < \infty \right\}.$$

Because we will be using discrete-time as well as continuous-time Markov processes, we mention here that for a given Markov Chain $x_n, n \in \mathbb{N}$, on a state space S (tipically S will be a finite or countable set, $\mathbb{R}^d$ or a separable Hilbert space $\mathcal{H}$), we will denote by $p(x, A), x \in S, A \subset S$, the transition probabilities of the chain (and by $p^n(x, A)$ the n-step transition probabilities). If S is finite or countable the transition probabilities are specified by $\{p(x, y)\}_{x,y,\in S}$. In this case the detailed balance condition with respect to a measure π on S can be rewritten as follows:

$$\pi(x)p(x, y) = \pi(y)p(y, x), \qquad \forall x, y \in S. \tag{2.4}$$

If the above holds, we say that x_n is reversible with respect to π.

Finally, for a measure μ on $\mathbb{R}^d$, we will use the same Greek letter to denote both the measure and its density (when such a density exists), i.e. we will write $\mu(dx) = \mu(x)dx$; $\mathcal{Z}$ will always denote a generic normalizing constant and for a differential operator A, $\mathcal{D}(A)$ will indicate the domain of A.

[3] $\mathcal{L}$ is the generator of the dynamics and the associated evolution equation, equation (2.2), governs the evolution of the observables. $\mathcal{L}'$ is often refereed to as the *Fokker-Planck* operator; $\mathcal{L}'$ describes the evolution of the law of the process.

3. Irreversibility

In this section we will be concerned with the study of exponentially fast convergence to equilibrium for irreversible Markov dynamics, i.e. dynamics generated by non-symmetric operators. As a term of comparison, let us start from the reversible case.

The theory concerning reversible Markov processes has been much more developed than the theory for non-reversible ones. This is mostly due to the fact that the generator of a reversible Markov process is a symmetric and, under some assumptions, self-adjoint operator; self-adjoint operators enjoy good spectral properties [27], which makes the study of convergence to equilibrium more accessible than in the non self-adjoint, irreversible case.

The study of exponentially fast convergence to equilibrium for reversible processes has been tackled using both probabilistic an analytic techniques. The most comprehensive reference on the analytic approach is [3]. While we do not intend to review the existing methods, we would like to recall some basic results. This is mainly to point out, by comparison, what are some of the difficulties in studying the problem of exponentially fast convergence to equilibrium in the irreversible case. Before stating the next definition we recall the following nomenclature: let $\mathcal{T}$ be the generator of an ergodic Markov semigroup; suppose that the spectrum of $\mathcal{T}$, $\sigma(\mathcal{T})$, is only made of simple isolated eigenvalues, that all such eigenvalues have positive (negative, respectively) real part and assume $0 \in \sigma(\mathcal{T})$. Then the *spectral gap* of $\mathcal{T}$, $\mathfrak{S}(\mathcal{T})$, is the smallest (biggest, respectively) real part of the non-zero eigenvalues of $\mathcal{T}$. Notice that if $\mathcal{T}$ is the generator of a strongly continuous ergodic Markov semigroup then $0 \in \mathfrak{S}(\mathcal{T})$, by the Koopman-Von Neumann Theorem (see [7, Theorem 1.2.1]).

Definition 3.1. *Given a Markov semigroup $\mathcal{P}_t$ with generator $\mathcal{L}$, we say that a measure π which is reversible for $\mathcal{P}_t$ satisfies a* spectral gap inequality *if there exists a constant $\alpha > 0$ such that*

$$\alpha \int_{\mathbb{R}} \left[f - \int_{\mathbb{R}} f d\pi \right]^2 d\pi \leq -\langle \mathcal{L}f, f \rangle_\pi, \quad \text{for every } f \in L^2_\pi \cap \mathcal{D}(\mathcal{L}). \tag{3.1}$$

The largest positive number α such that (3.1) *is satisfied is the* spectral gap *of the self-adjoint operator $\mathcal{L}$.*

The term on the RHS of (3.1) is called the *Dirichlet form* of the operator $\mathcal{L}$.

Remark 3.1. If $\mathcal{L}$ is a self adjoint operator then the form $\langle \mathcal{L}f, f \rangle_\pi$ is real valued. In particular the spectrum of $\mathcal{L}$ is real. If $\mathcal{L}$ is the generator of a strongly continuous Markov semigroup and the semigroup is ergodic then we already know that 0 is a simple eigenvalue of $\mathcal{L}$. If (3.1) holds, then $\langle \mathcal{L}f, f \rangle_\pi \leq 0$ for every f, therefore the self-adjoint operator $-\mathcal{L}$ is *positive* and all the eigenvalues of $-\mathcal{L}$ will be positive. The biggest positive α such that (3.1) holds is the smallest nonzero eigenvalue of $-\mathcal{L}$, i.e. α is the spectral gap.[4] The next proposition clarifies why spectral gap inequalities are so important. Notice however that, at least on a formal level, it makes sense to talk about spectral gap inequalities if one can guarantee that the quantity $\langle \mathcal{L}f, f \rangle_\pi$ is at least real. This can not be guaranteed in general if $\mathcal{L}$ is non self-adjoint. □

Proposition 3.1. *A measure π reversible with respect to the Markov semigroup $\mathcal{P}_t$ satisfies a spectral gap inequality (with constant α) if and only if*

$$\int_{\mathbb{R}} \left(\mathcal{P}_t f - \int_{\mathbb{R}} f d\pi \right)^2 d\pi \leq e^{-2\alpha t} \int_{\mathbb{R}} \left(f - \int_{\mathbb{R}} f d\pi \right)^2 d\pi \,, \tag{3.2}$$

for all $t \geq 0$ and $f \in L^2_\pi$.

[4] This reasoning might appear more transparent if we take mean zero functions, that is f sych that $\int f \, d\pi = 0$.

A proof of the above proposition can be found in [15, Chapter 2]. The spectral gap inequality formalism is one of the most established techniques to study exponential convergence for reversible diffusions. However this cannot be used - at least not as is - in the irreversible case (on this point we also mention the related interesting paper [19]).

If irreversible diffusions are harder to study than reversible ones, it is natural to wonder why one would want to employ them in the study of MCMC. The reason is readily explained: plenty of numerical evidence - although not as many theoretical results - shows that irreversible processes converge to equilibrium faster than reversible dynamics. We illustrate this idea with an example (to the best of our knowledge this is one of the very few examples where rigorous results are available). Consider the Ornstein-Uhlenbeck process (OU)

$$dY_t = -Y_t dt + \sqrt{2} dW_t, \qquad Y_t \in \mathbb{R}^d . \tag{3.3}$$

Y_t is ergodic with unique invariant measure $\pi(y) = e^{-|y|^2/2}/\mathcal{Z}$. Y_t is also reversible with respect to π. Now consider the process Z_t obtained from Y_t by adding a non-reversible perturbation to the drift, i.e. modify the OU process in such a way that the invariant measure of the new process is still π but Z_t is no longer reversible with respect to π:

$$dZ_t = (-Z_t + \gamma(Z_t))dt + \sqrt{2} dW_t, \qquad \text{with } \nabla \cdot (\gamma(z) e^{-V(z)}) = 0 .$$

The condition $\nabla \cdot (\gamma(z) e^{-V(z)}) = 0$ is added in order to preserve the invariance of π. It can be shown (see [21, 18, 24]) that $\mathfrak{S}(Z) \leq \mathfrak{S}(Y)$ and that the process Z_t converges faster than Y_t.

One of the most popular approaches to study exponential convergence to equilibium in the non-reversible case is given by the *hypocoercivity Theory*, which we briefly review below.

3.1. **Hypocoercivity theory and Second Order Langevin Equation.** Let us start by introducing the Second Order Langevin (SOL) equation, which is possibly the simplest example of dynamics that retains all the properties that we are interested in. Also, it is the dynamics that we will use to construct the SOL-HMC algorithm in Section 5. By SOL we will mean the following SDE (or slight variations):

$$\begin{aligned} dq &= p dt \\ dp &= -\partial_q V(q) dt - p dt + \sqrt{2} dW_t, \end{aligned} \tag{3.4}$$

where, $(q, p) \in \mathbb{R}^2$, $V(q) \in \mathcal{C}^\infty$ is a confinig potential (i.e. $V(q) \to \infty$ as $|q| \to \infty$ and $V(q)$ grows at least quadratically at infinity [5]) and W_t is a one dimensional standard Brownian motion. The generator of (3.4) is

$$\mathcal{L} = p\partial_q - \partial_q V(q)\partial_p - p\partial_p + \partial_p^2 \tag{3.5}$$

and the corresponding Fokker-Planck operator is

$$\mathcal{L}' = -p\partial_q + \partial_q V(q)\partial_p + \partial_p(p\cdot) + \partial_p^2 . \tag{3.6}$$

Notice that $\mathcal{L}'$ is non-uniformly elliptic. In particular, it is *hypoelliptic*. We will not linger on this fact here and refer the reader to [39] for a coincise and clear introduction to the hypoellipticity theory. We just observe that the fact that $\partial_t - \mathcal{L}'$ is hypoelliptic on $\mathbb{R}_+ \times \mathbb{R}^2$ implies that the law of the process (3.5) has a density for every $t > 0$. The dynamics generated by the operator (3.5) is ergodic as well and the density of the unique invariant measure of such a dynamics is

$$\rho(q, p) = \frac{e^{-(V(q)+p^2/2)}}{\mathcal{Z}} . \tag{3.7}$$

[5] Under this assumption strong uniqueness and non-explosivity are guaranteed, see e.g. [35, Chapter 10]

The dynamics described by (3.4) can be thought of as split into a Hamiltonian component,

$$\begin{aligned} \dot q &= p \\ \dot p &= -\partial_q V(q) \end{aligned} \tag{3.8}$$

plus a OU process (in the p variable, see (3.3)):

$$\begin{aligned} dq &= p dt \\ dp &= -\partial_q V(q) dt \underbrace{-p dt + \sqrt{2} dW_t}_{\text{OU process}}. \end{aligned}$$

Indeed the equations (3.8) are the equations of motion of an Hamiltonian system with Hamiltonian

$$H(q,p) = V(q) + \frac{p^2}{2}.$$

At the level of the generator this is all very clear: we can write the operator $\mathcal{L}$ as

$$\mathcal{L} = \mathcal{L}_H + \mathcal{L}_{OU},$$

where

$$\mathcal{L}_H := p\partial_q - \partial_q V(q)\partial_p \tag{3.9}$$

is the Liouville operator of classical Hamiltonian mechanics and

$$\mathcal{L}_{OU} := -p\partial_p + \partial_p^2$$

is the generator of a OU process in the p variable. By the point of view of our formalism, the Hamiltonian dynamics (3.8) admits infinitely many invariant measures, indeed

$$-\mathcal{L}_H' f(H(q,p)) = \mathcal{L}_H f(H(q,p)) = 0 \quad \text{for every } f \text{ (smooth enough)}.$$

So any integrable and normalized function of the Hamiltonian is an invariant probability measure for (3.8). Adding the OU process amounts to selecting one equilibrum.

To distinguish between the flat L^2 adjoint of an operator T and the adjoint in the weighted L^2_ρ, we shall denote the first by T' and the latter by T^*. The scalar product and norm of L^2_ρ will be denoted by $\langle\cdot,\cdot\rangle_\rho$ and $\|\cdot\|_\rho$, respectively. Notice now that the generator $\mathcal{L}_H$ of the Hamiltonian part of the Langevin equation is antisymmetric both in L^2 and in L^2_ρ. It is indeed straightforward to see that

$$\mathcal{L}_H = -\mathcal{L}_H'.$$

Also, $\langle \mathcal{L}_H f, g\rangle_\rho = -\langle f, \mathcal{L}_H g\rangle_\rho$ for every f, g say in $L^2_\rho \cap \mathcal{D}(\mathcal{L}_H)$:

$$\begin{aligned} \langle \mathcal{L}_H f, g\rangle_\rho &= \int_{\mathbb{R}}\int_{\mathbb{R}} (p\partial_q f - q\partial_p f)\, g\rho \, dpdq \\ &= -\int_{\mathbb{R}}\int_{\mathbb{R}} fp\partial_q(g\rho)\, dpdq + \int_{\mathbb{R}}\int_{\mathbb{R}} fq\partial_p(g\rho)\, dpdq \\ &= -\int_{\mathbb{R}}\int_{\mathbb{R}} fp(\partial_q g)\rho + \int_{\mathbb{R}}\int_{\mathbb{R}} qf(\partial_p g)\rho = -\langle f, \mathcal{L}_H g\rangle_\rho. \end{aligned}$$

The generator of the OU process is instead symmetric in $\mathcal{L}^2_\rho$ and in particular

$$\mathcal{L}_{OU} = -T^*T,$$

where

$$T = \partial_p, \quad \text{so that} \quad T^* = -\partial_p + p.$$

In conclusion, the generator of the Langevin equation decomposes into a symmetric and antisymmetric part. Moreover, the antisymmetric part comes from the Hamiltonian deterministic component of the dynamics, the symmetric part comes from the stochastic component.

Using Stone's Theorem (see e.g. [27]) we also know that the semigroup generated by $\mathcal{L}_H$ is norm-preserving, while it is easy to see that the semigroup generated by $\mathcal{L}_{OU}$ is dissipative, indeed

$$\begin{aligned}\frac{d}{dt}\|e^{t\mathcal{L}_{OU}}h\|_\rho^2 &= 2\langle \mathcal{L}_{OU}e^{t\mathcal{L}_{OU}}h, e^{t\mathcal{L}_{OU}}h\rangle_\rho \\ &= -2\langle T^*Th_t, h_t\rangle_\rho = -2\|Th_t\|_\rho^2 < 0\,,\end{aligned}$$

where we used the notation $h_t(x) = e^{t\mathcal{L}_{OU}}h(x)$. In conclusion, so far we have the following picture:

$$\begin{array}{ccccc} \mathcal{L} = & \underbrace{\mathcal{L}_H} & - & \underbrace{T^*T} & \\ & \text{skew symmetric} & & \text{symmetric} & \\ & \downarrow & & \downarrow & \\ & \text{deterministic} & & \text{stochastic} & \\ & \text{conservative} & & \text{dissipative} & \end{array}$$

This is precisely the setting of the hypocoercivity theory. The hypocoercivity theory, subject of [38], is concerned with the problem of exponential convergence to equilibrium for evolution equations of the form

$$\partial_t h + (A^*A - B)\, h = 0,^{6} \tag{3.10}$$

where B is an antisymmetric operator [7]. We shall briefly present some of the basic elements of such a theory and then see what are the outcomes of such a technique when we apply it to the Langevin equation (3.4).

We first introduce the necessary notation. Let $\mathcal{H}$ be a Hilbert space, *real* and separable, $\|\cdot\|$ and $(\cdot,\cdot)$ the norm and scalar product of $\mathcal{H}$, respectively. Let A and B be unbounded operators with domains $\mathcal{D}(A)$ and $\mathcal{D}(B)$ respectively, and assume that B is antisymmetric, i.e. $B^* = -B$, where * denotes adjoint in $\mathcal{H}$. We shall also assume that there exists a vector space $\mathcal{S} \subset \mathcal{H}$, dense in $\mathcal{H}$, where all the operations that we will perform involving A and B are well defined.

Writing the involved operator in the form $\mathcal{T} = A^*A - B$ has several advantages. Some of them are purely computational. For example, for operators of this form checking the contractivity of the semigroup associated with the dynamics (3.10) becomes trivial. Indeed, the antisymmetry of B implies

$$(Bx, x) = -(x, Bx) \quad \Longrightarrow \quad (Bx, x) = 0. \tag{3.11}$$

This fact, together with $(A^*Ax, x) = \|Ax\|^2 \geq 0$, immediately gives

$$\frac{1}{2}\frac{d}{dt}\|e^{-t\mathcal{T}}h\|^2 \overset{(3.11)}{=} -\|Ah_t\|^2 \leq 0\,.$$

On the other hand, conceptually, the decomposition $A^*A - B$ is physically meaningful as the symmetric part of the operator, A^*A, corresponds to the stochastic (dissipative) part of the dynamics, whereas the antisymmetric part corresponds to the deterministic (conservative) component.

[6] Generalizations to the form $\partial_t h + \left(\sum_{i=1}^m A_i{}^*A_i - B\right) h = 0$ as well as further generalizations are presented in [38]. We refer the reader to such a monograph for these cases.

[7] Notice that, for less than regularity issues, any second order differential operator $\mathcal{L}$ can be written in the form $A^*A - B$.

Definition 3.2. *We say that an unbounded linear operator $\mathcal{T}$ on $\mathcal{H}$ is relatively bounded with respect to the linear operators $T_1, ..., T_n$ if the domain of $\mathcal{T}$, $\mathcal{D}(\mathcal{T})$, is contained in the intersection $\cap\mathcal{D}(T_j)$ and there exists a constant $\alpha > 0$ s.t.*

$$\forall h \in \mathcal{D}(\mathcal{T}), \qquad \|\mathcal{T}h\| \leq \alpha(\|T_1 h\| + ... + \|T_n h\|).$$

Definition 3.3 (Coercivity)**.** *Let $\mathcal{T}$ be an unbounded operator on a Hilbert space $\mathcal{H}$, denote its kernel by $\mathcal{K}$ and assume there exists another Hilbert space $\tilde{\mathcal{H}}$ continuously and densely embedded in $\mathcal{K}^\perp$. If $\|\cdot\|_{\tilde{\mathcal{H}}}$ and $(\cdot,\cdot)_{\tilde{\mathcal{H}}}$ are the norm and scalar product on $\tilde{\mathcal{H}}$, respectively, then the operator $\mathcal{T}$ is said to be λ-coercive on $\tilde{\mathcal{H}}$ if*

$$(\mathcal{T}h, h)_{\tilde{\mathcal{H}}} \geq \lambda \|h\|^2_{\tilde{\mathcal{H}}}, \quad \forall h \in \mathcal{K}^\perp \cap D(\mathcal{T}),$$

where $D(\mathcal{T})$ is the domain of $\mathcal{T}$ in $\tilde{\mathcal{H}}$.

Notice the parallel with (3.1). Notice also that, from the above discussion, for every $h \in \mathcal{D}(\mathcal{T})$, the number $(\mathcal{T}h, h)$ is always real. Not surprisingly, the following proposition gives an equivalent definition of coercivity (cfr Proposition 3.1).

Proposition 3.2. *With the same notation as in Definition 3.3, $\mathcal{T}$ is λ-coercive on $\tilde{\mathcal{H}}$ iff*

$$\| e^{-\mathcal{T}t} h \|_{\tilde{\mathcal{H}}} \leq e^{-\lambda t} \| h \|_{\tilde{\mathcal{H}}} \quad \forall h \in \tilde{\mathcal{H}} \text{ and } t \geq 0.$$

Definition 3.4 (Hypocoercivity)**.** *With the same notation of Definition 3.3, assume $\mathcal{T}$ generates a continuous semigroup. Then $\mathcal{T}$ is said to be λ-hypocoercive on $\tilde{\mathcal{H}}$ if there exists a constant $\kappa > 0$ such that*

$$\| e^{-\mathcal{T}t} h \|_{\tilde{\mathcal{H}}} \leq \kappa e^{-\lambda t} \| h \|_{\tilde{\mathcal{H}}}, \quad \forall h \in \tilde{\mathcal{H}} \text{ and } t \geq 0. \tag{3.12}$$

Remark 3.2. We remark that the only difference between Definition 3.3 and Definition 3.4 is in the constant κ on the right hand side of (3.12), when $\kappa > 1$. Thanks to this constant, the notion of hypocoercivity is invariant under a change of equivalent norm, as opposed to the definition of coercivity which relies on the choice of the Hilbert norm. Hence the basic idea employed in the proof of exponentially fast convergence to equilibrium for degenerate diffusions generated by operators in the form (3.10), is to appropriately construct a norm on $\tilde{\mathcal{H}}$, equivalent to the existing one, and such that in this norm the operator is coercive. □

We will state in the following the basic theorem in the theory of hypocoercivity. Generalizations can be found in [38].

Theorem 3.1. *With the notation introduced so far, let $\mathcal{T}$ be an operator of the form $\mathcal{T} = A^*A - B$, with $B^* = -B$. Let $\mathcal{K} = Ker\mathcal{T}$, define $C := [A, B]$, [8] and consider the norm*

$$\|h\|^2_{\mathcal{H}^1} := \|h\|^2 + \|Ah\|^2 + \|Ch\|^2.$$

on $\mathcal{K}^\perp$. [9] Suppose the following holds:

1. *A and A^* commute with C;*
2. *$[A, A^*]$ is relatively bounded with respect to I and A;*
3. *$[B, C]$ is relatively bounded with respect to A, A^2, C and AC,*

[8] Given two differential operators X and Y we denote by $[X, Y] = XY - YX$ the commutator between X and Y.

[9] One can prove that space $\mathcal{K}^\perp$ is the same irrespective of whether we consider the scalar product $\langle\cdot,\cdot\rangle$ of $\mathcal{H}$ or the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{H}^1}$ associated with the norm $\|\cdot\|_{\mathcal{H}^1}$.

then there exists a scalar product $((\cdot,\cdot))$ *on* $\mathcal{H}^1/\mathcal{K}$ *defining a norm equivalent to the* $\mathcal{H}^1$ *norm such that*

$$((h,\mathcal{T}h)) \geq k(\|Ah\|^2 + \|Ch\|^2), \quad \forall h \in \mathcal{H}^1/\mathcal{K}, \tag{3.13}$$

for some constant $k > 0$*. If, in addition to the above assumptions, we have*

$$A^*A + C^*C \text{ is } \kappa - \text{coercive for some } \kappa > 0, \tag{3.14}$$

then $\mathcal{T}$ *is hypocoercive in* $\mathcal{H}^1/\mathcal{K}$*: there exist constants* c, $\lambda > 0$ *such that*

$$\|e^{-t\mathcal{L}}\|_{\mathcal{H}^1/\mathcal{K}\to\mathcal{H}^1/\mathcal{K}} \leq ce^{-\lambda t}.$$

Remark 3.3. Let $\mathcal{K}$ be the kernel of $\mathcal{T}$ and notice that $Ker(A^*A) = Ker(A)$ and $\mathcal{K} = Ker(A)\cap Ker(B)$. Suppose $KerA \subset KerB$; then $Ker\mathcal{T} = KerA$. In this case the coercivity of $\mathcal{T}$ is equivalent to the coercivity of A^*A. So the case we are interested in is the case in which A^*A is coercive and $\mathcal{T}$ is not. In order for this to happen A^*A and B cannot commute; if they did, then $e^{-t\mathcal{T}} = e^{-tA^*A}e^{tB}$. Therefore, since e^{tB} is norm preserving, we would have $\|e^{-t\mathcal{T}}\| = \|e^{-tA^*A}\|$. This is the intuitive reason why commutators (especially of the form $[A,B]$) appear in Theorem 3.1. □

Comment.[On the Proof of Theorem 3.1] We will not write a proof of this theorem but I will explain how it works. The idea is the same that we have explained in Remark 3.2. Consider the norm

$$((h,h)) := \|h\|^2 + a\|Ah\|^2 + c\|Ch\|^2 + 2b(Ah,Ch),$$

where a, b and c are three strictly positive constants to be chosen. Assumptions (1), (2) and (3) are needed to ensure that this norm is equivalent to the $\mathcal{H}^1$ norm, i.e. that there exist constant $c_1, c_2 > 0$ such that

$$c_1\|h\|_{\mathcal{H}^1} \leq ((h,h)) \leq c_2\|h\|_{\mathcal{H}^1}.$$

If we can prove that $\mathcal{T}$ is coercive in this norm, then by Proposition 3.2 and Remark 3.2 we have also shown exponential convergence to equilibrium in the $\mathcal{H}^1$ norm i.e. hypocoercivity. So the whole point is proving that

$$((\mathcal{T}h,h)) \geq K((h,h)),$$

for some $K > 0$. If (1), (2) and (3) of Theorem 3.1 hold, then (with a few lengthy but surprisingly not at all complicated calculations) (3.13) follows. From now on $K > 0$ will denote a generic constant which might not be the same from line to line. The coercivity of $A^*A + C^*C$ means that we can write

$$\begin{aligned}\|Ah\|^2 + \|Ch\|^2 &= \frac{1}{2}(\|Ah\|^2 + \|Ch\|^2) + \frac{1}{2}(\|Ah\|^2 + \|Ch\|^2)\\ &\geq \frac{1}{2}(\|Ah\|^2 + \|Ch\|^2) + \frac{K}{2}\|h\|^2\\ &\geq K\|h\|_{\mathcal{H}^1}.\end{aligned}$$

Combining this with (3.13), we obtain

$$((h,\mathcal{T}h)) \geq k(\|Ah\|^2 + \|Ch\|^2) \geq K\|h\|_{\mathcal{H}^1} \geq K((h,h)).$$

This concludes the sketch of the proof. Another important observation is that, in practice, the coercivity of $A^*A + C^*C$ boils down to a Poincaré inequality. This will be clear when we apply this machinery to the Langevin equation, see proof of Theorem 3.2.

□

We now use Theorem 3.1 to prove exponentially fast convergence to equilibrium for the Langevin dynamics. We shall apply such a theorem to the operator $\mathcal{L}$ defined in (3.5) on the

space $\mathcal{H} = L^2_\rho$, where ρ is the equilibrium distribution (3.5). (The space $\mathcal{S}$ can be taken to be the space of Schwartz functions.) The operators A and B are then

$$A = \partial_p \quad \text{and} \quad B = p\partial_q - \partial_q V \partial_p,$$

so that

$$C := [A, B] = AB - BA = \partial_q \,.$$

The kernel $\mathcal{K}$ of the operator $\mathcal{L}$ is made of constants and in this case the norm $\mathcal{H}^1$ will be the Sobolev norm of the weighted $H^1(\rho)$:

$$\|f\|^2_{H^1_\rho} := \|f\|^2_\rho + \|\partial_q f\|^2_\rho + \|\partial_p f\|^2_\rho \,.$$

Let us first calculate the commutators needed to check the assumptions of Theorem 3.1.

$$[A, C] = [A^*, C] = 0, \quad [A, A^*] = Id \tag{3.15}$$

and

$$[B, C] = -\partial_q^2 V(q) \partial_p \,. \tag{3.16}$$

Theorem 3.2. *Let $V(q)$ be a smooth potential such that*

$$|\partial_q^2 V| \le \alpha\big(1 + |\partial_q V|\big), \qquad \textit{for some constant } \alpha > 0. \tag{3.17}$$

Also, assume that $V(q)$ is such that the measure $e^{-V(q)}$ satisfies a Poincaré inequality. [10] *Then, there exist constants C, $\lambda > 0$ such that for all $h_0 \in H^1(\rho)$,*

$$\left\| e^{-t\,\mathcal{L}} h_0 - \int h_0 \, d\rho \right\|_{H^1(\rho)} \le C e^{-\lambda t} \|h_0\|_{H^1(\rho)}, \tag{3.18}$$

where we recall that here $\mathcal{L}$ is the operator (3.5) *.*

Proof. We will use Theorem 3.1. Conditions (1) and (2) are satisfied, due to (3.15). In [38, page 56 and Lemma A.19] it is shown that condition (3) holds under the assumption (3.17) on the potential V. Now we turn to condition (3.14). Let us first write the operator $\widehat{\mathcal{L}} = A^*A + C^*C$ (notice that $\widehat{\mathcal{L}}$ is elliptic):

$$\widehat{\mathcal{L}} = p\partial_p - \partial_p^2 + \partial_q V \partial_q - \partial_q^2.$$

The operator $\widehat{\mathcal{L}}$ is coercive if

$$\int \Big(|\partial_q h|^2 + |\partial_p h|^2 \Big) \, d\rho \ge \kappa \|h\|^2_\rho.$$

The above is a Poincaré inequality for the measure ρ (as we have already observed, the kernel of $\mathcal{T}$ is the set of constant functions, so it suffices to write the Poincaré inequality for mean zero functions, as we have done in the above). Therefore, in order for $\widehat{\mathcal{L}}$ to be coercive, it is sufficient for the measure $\rho = e^{-V(q)} e^{-p^2/2}$ to satisfy a Poincaré inequality. This probability measure is the product of a Gaussian measure (in p) which satisfies a Poincaré inequality, and of the probability measure $e^{-V(q)}$. In order to conclude the proof it is sufficient, therefore, to use the assumption that $e^{-V(q)}$ satisfies a Poincaré inequality. □

More details about the above proof can also be found in [25].

We mention that while the hypocoercivity theory has rapidly become one of the most popular techniques to study return to equilibrium for hypoelliptic-irreversible processes, other avenues have recently been opened [24], based on spectral theory and semiclassical analysis (in this context, we would also point out the paper [13]). While the first approach mostly provides qualitative results, the latter allows a more quantitative study. In other words, through the

[10] Theorem A.1 in [38] gives some sufficient conditions in order for e^{-V} to satisfy such an inequality.

hypocoercivity techniques we only know that some $\lambda > 0$ exists, such that (3.18) holds; the spectral approach [24] gives instead the exact rate of exponential convergence, i.e. it determines λ. However, in comparison to the hypocoercivity framework, spectral techniques only apply to a more restricted class of hypoelliptic diffusions. Quantitative information for the Ornstein-Uhlenbeck process has been obtained also by using hypocoercivity-type techniques [1].

4. Markov Chain Monte Carlo

A standard and practical reference on MCMC is the book [30]. A rigorous approach to the theory of Markov chains and some theoretical results about MCMC are contained in [22]. The case for using MCMC is passionately argued in [11].

As we have already mentioned in the Introduction, MCMC algorithms can be employed for two purposes: i) sampling from a given target distribution $\pi(x)$ which is known only up to its normalizing constant or ii) approximate statistical quantities of π, that is, calculate integrals of the form (1.1). In order to achieve either i) or ii), the MCMC approach consists in building a Markov Chain x_n that has π as (unique) invariant measure. Then, for example under an assumption of positive recurrence, the ergodic theorem holds (e.g. for all $f \in L^1_\pi$), and the average on the left hand side of (1.3) is, for n large enough, a good approximation of the integral on the right hand side. We will not discuss here the very important practical issue of how big n should be and other related issues.

In algorithmical practice, it is a standard procedure to start by building a chain which admits the target measure π as unique invariant measure. This obviously does not ensure that the chain will converge to π (in whichever sense, see Example 4.1 below) and therefore a significant amount of literature has been devoted to the study of convergence criteria applicable to MCMC chains. Reviewing these criteria is beyond the scope of the present paper and we refer the reader to [14, 22, 34] and references therein. However, for Markov Chains as well as for continuous time Markov processes, it is still the case that the great majority of the convergence results concern reversible processes. This is mostly due to the popularity of the Metropolis-Hastings algorithm, which we introduce in Section 4.1. Before presenting the general algorithm, we start with a simple example (see [2]).

Example 4.1. Suppose we want to sample from a measure π defined on a finite state space S. In order to do so, we shall construct a Markov Chain x_n that converges to π, in the sense that if $p(x, y)$ are the transition probabilities of the Markov chain x_n then we want

$$\lim_{n\to\infty} p^n(x, y) = \pi(y)\,. \tag{4.1}$$

With the intent of constructing x_n (or, equivalently, $p(x, y)$) we can proceed as follows. Let $q(x, y)$ be a an arbitrary transition probability on S. Suppose the transition matrix $Q = (q(x, y))_{(x,y)\in S}$ is symmetric and irreducible. Given such a Q (usually called *proposal* transition matrix) and a probability distribution $\pi(x)$ on S such that $\pi(x) > 0$ for all $x \in S$, let us now construct a new transition matrix $P = (p(x, y))$ as follows :

$$p(x,y) = \begin{cases} q(x,y) & \text{if } \pi(y) \geq \pi(x) \text{ and } x \neq y \\ q(x,y)\frac{\pi(y)}{\pi(x)} & \text{if } \pi(y) < \pi(x) \text{ and } x \neq y \\ 1 - \sum_{x\neq y} p(x,y) & \text{otherwise.} \end{cases} \tag{4.2}$$

It is easy to check that the matrix $P = (p(x, y))$ constructed in this way is an irreducible transition matrix. [11] Being the state space finite, this also implies that P is recurrent and that there exists a unique stationary distribution. We can easily show that such an invariant

[11] Meaning that the whole state space is irreducible under P; this implies that the state space is also closed under P (here we mean *closed* in the sense of Markov Chains; that is, we say that a set A of the state space is

distribution is exactly π as P is reversible with respect to π in the sense (2.4). (2.4) is obviously true when $x = y$. So suppose $x \neq y$ and $\pi(y) \geq \pi(x)$. Then, by construction, $\pi(x)p(x,y) = \pi(x)q(x,y)$ but also $\pi(y)p(y,x) = q(y,x)[\pi(x)/\pi(y)]\pi(y)$ so that using the symmetry of q we get $\pi(y)p(y,x) = q(x,y)\pi(x)$ and we are done. If $\pi(y) < \pi(x)$ we can repeat the above with roles of x and y reversed. We are left with proving that the chain x_n with transition matrix P converges to π. We show in Appendix that convergence (in the sense (4.1)) happens for any proposal Q unless π is the uniform distribution on S (see Lemma A.1). This is just to highlight, on a simple example where calculations can be easily made by hand, that the convergence of the scheme can depend on the target measure and not only on Q. More complex (and meaningful) examples on this point can be found in [33]. □

The procedure (4.2) can be expressed as follows: given $X_n = x_n$,

(1) generate $y_{n+1} \sim q(x_n, \cdot)$;
(2) calculate
$$\alpha(x_n, y_{n+1}) := \min\left\{1, \frac{\pi(y_{n+1})}{\pi(x_n)}\right\} \tag{4.3}$$
(3) set $X_{n+1} = \begin{cases} y_{n+1} & \text{with probability } \alpha(x_n, y_{n+1}) \\ x_n & \text{otherwise.} \end{cases}$

In practice, if $\mathcal{U}[0,1]$ is the uniform distribution on $[0,1]$, the algorithm that realizes the above is

Algorithm 4.1. Given $X_n = x_n$,

(1) generate $y_{n+1} \sim q(x_n, \cdot)$;
(2) generate $u \sim \mathcal{U}[0,1]$;
(3) if $u < \pi(y_{n+1})/\pi(x_n)$ then $X_{n+1} = y_{n+1}$; otherwise $X_{n+1} = x_n$.

In words, given the state of the chain at time n, we pick the *proposal* $y_{n+1} \sim q(x_n, \cdot)$. Then the proposed move is accepted with probability α (4.3). If it is rejected, the chain remains where it was. For this reason $\alpha(x,y)$ is called the *acceptance probability*.

Algorithm 4.1 is a first example of a *Metropolis-Hastings algorithm*. Intuitively, it is clear why we always accept moves towards points with higher probability. We anyway make the obvious remark that if we want to construct an ergodic chain (in the sense (1.3)) with invariant probability π then the time spent by the chain in each point y of S needs to equal, in the long run, the probability assigned by π to y, i.e. $\pi(y)$. So we have to accept more frequently points with higher probability.

4.1. **Metropolis-Hastings algorithm.** Throughout this section our state space is $\mathbb{R}^N$. For simplicity we will assume that all the measures we use have a density with respect to the Lebesgue measure, so $\pi(x)$ will be the density of π and e.g. $q(x,y)$ will denote the density of the proposal $q(x,\cdot)$. A very nice presentation of the theory underlying the M-H algorithm in general state space can be found in [37].

A Metropolis-Hastings (M-H) algorithm is a method of constructing a time-homogeneous Markov chain or, equivalently, a transition kernel $p(x,y)$, that is reversible with respect to a given target distribution $\pi(x)$. To construct the π-invariant chain X_n we make use of a proposal kernel $q(x,y)$ which we know how to sample from and of an accept/reject mechanism with acceptance probability
$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\}. \tag{4.4}$$

closed if whenever $x \in A$ and y is accessible from x then also y belongs to A. For a precise Definition see [10, page 246])

For simplicity we require that $\pi(y)q(y,x) > 0$ and $\pi(x)q(x,y) > 0$. The M-H algorithm consists of two steps:

Algorithm 4.2 (Metropolis-Hastings algorithm)**.** Given $X_n = x_n$,

(1) generate $y_{n+1} \sim q(x_n, \cdot)$;
(2) calculate $\alpha(x_n, y_{n+1})$ according to the prescription (4.4)
(3) set $X_{n+1} = \begin{cases} y_{n+1} & \text{with probability } \alpha(x_n, y_{n+1}) \\ x_n & \text{otherwise.} \end{cases}$

Lemma 4.1. *If α is the acceptance probability* (4.4)*, (and assuming $\pi(y)q(y,x) > 0$ and $\pi(x)q(x,y) > 0$) the Metropolis-Hastings algorithm, Algorithm 4.2, produces a π- invariant time-homogeneous Markov chain.* [12]

Proof. A proof of this fact can be found in [37]. □

Remark 4.1. In order to implement Algorithm 4.2 we don't need to know the normalizing constant for π, as it gets canceled in the ratio (4.4). However we do need to know the normalizing constant for q: q is a transition probability so by definition for every fixed x the function $y \to q(x,y)$ is a probability density i.e. it integrates to one. However the normalizing constant of $q(x,\cdot)$ can, and in general will, depend on x. In other words, $q(x,y)$ will in general be of the form $q(x,y) = \mathcal{Z}_x^{-1}\tilde{q}(x,y)$, with $\int dy\, \tilde{q}(x,y) = \mathcal{Z}_x$ so that the ratio in the acceptance probability (4.4) can be more explicitly written as

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)\mathcal{Z}_x\tilde{q}(y,x)}{\pi(x)\mathcal{Z}_y\tilde{q}(x,y)}\right\}.$$

Clearly, if the proposal kernel is symmetric, $q(x,y) = q(y,x)$, then there is no need to know the normalizing constant for q, as the above expression for α reduces to (4.3). This is a big appeal of algorithms with symmetric proposals, such as Random Walk Metropolis, which we introduce below. □

Remark 4.2. Let us repeat that M-H is a method to generate a π-reversible time-homogeneous Markov chain. As we have already noticed, the fact that the chain is π-reversible does not imply that π is the only invariant distribution for the chain or even less that the chain converges to π. The matter of convergence of the chain constructed via M-H is probably better studied case by case (i.e. depending on the proposal we decide to use and on the target measure that we are trying to sample from). Some results concerning convergence of the chain can be found in [22, Chapter 20] and references therein or in [32, 33]. □

4.1.1. *Random Walk Metropolis (RWM).* A very popular M-H method is the so called *Random Walk Metropolis*, where the proposal y_{n+1} is of the form

$$y_{n+1} = x_n + \sigma\xi_{n+1}, \qquad \sigma > 0;$$

for the algorithm that is most commonly referred to as RWM, the noise ξ is Gaussian, i.e. $\xi \sim \mathcal{N}(0,\sigma^2)$ so that $q(x,y) \sim \mathcal{N}(x,\sigma^2)$.[13] Therefore the acceptance probability reduces to $\alpha = \min\{1, \pi(y)/\pi(x)\}$. The case in which the noise ξ is Gaussian has been extensively studied in the literature, for target measures defined on $\mathbb{R}^N$. We stress that the variables $\xi_1, \dots, \xi_n, \dots$ are i.i.d. random variables, independent of the current state of the chain x_n. Therefore the proposal move doesn't take into account any information about the current state of the chain or about the target measure. This is in contrast with the MALA algorithm, Section 4.1.2 below, where

[12] Lemma 4.1 can be made a bit more general, see [37].

[13] In principle ξ could be chosen to be any noise with density $g(x)$ symmetric with respect to the origin, $g(x) = g(|x|)$.

the proposal move incorporates information about the target. This makes RMW a more naive algorithm than MALA.

Moreover, RWM is not immune to the *curse of dimensionality*: the cost of the algorithm increases with the dimension N of the state space in which it is implemented. Simply put: sampling from a measure that is defined on $\mathbb{R}^N$ is more expensive than sampling from a measure defined on $\mathbb{R}^{N-1}$. Here by cost of the algorithm we mean the number of MCMC steps needed in order to explore the state space in stationarity. In order to ameliorate this problem, it is crucial to choose the proposal variance appropriately. In $\mathbb{R}^N$ it is customary to consider $\sigma^2 = cN^{-\gamma}$ where $c, \gamma > 0$ are two parameters to be appropriately tuned, the most interesting of the two being γ. If γ is too large then σ^2 is too small, so the proposed moves tend to stay close to the current value of the chain and the state space is explored very slowly. If instead γ is too small, more precisely smaller than a critical value γ_c, the average acceptance rate decreases very rapidly to zero as N tends to infinity. This means that the algorithm will reject more and more as N increases. It was shown in the seminal paper [31] that the choice $\gamma = 1$ is the one that optimally compromises between the need of moving far enough away from the current position and the need of accepting frequently enough.

4.1.2. *Metropolis Adjusted Langevin Algorithm (MALA).* Consider the first order Langevin equation

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}}dW_t \tag{4.5}$$

where $X_t \in \mathbb{R}^d$, $V(x)$ is a confining potential and W_t is a d-dimensional standard Brownian motion. $\beta > 0$ is a parameter (typically β^{-1} is the temperature) which from now on we fix to be equal to one, $\beta = 1$. This dynamics is ergodic; the (unique) invariant measure has a density, $\rho(x)$, explicitly given by

$$\rho(x) = \frac{e^{-V(x)}}{\mathcal{Z}}, \tag{4.6}$$

where $\mathcal{Z}$ is the normalizing constant. Moreover, under the stated assumptions on the potential, X_t converges exponentially fast to the equilibrium ρ. If we want to sample from measures of the form (4.6), it is a natural idea to construct a Markov Chain that converges to ρ by discretizing the continuous-time dynamics (4.5). Unfortunately one can readily see that naive discretizations can completely destroy the good properties of the dynamics (4.5). Indeed, as pointed out in [32], suppose we discretize (4.5) by using the Euler scheme with step h; that is, suppose we create a chain according to

$$X_{n+1} \sim \mathcal{N}(X_n - h\nabla V(X_n), 2hI_d), \qquad I_d = d\text{-dimensional identity matrix.}$$

Suppose your target distribution is Gaussian with zero mean and unit variance (corresponding to $V(x) = |x|^2/2$) and choose $h = 1$. Then $X_n \sim \mathcal{N}(0, 2)$ for every n. So clearly the chain converges immediately, but to the wrong invariant measure. This is the most drastic example of what can go wrong. In general when discretizing, the invariance of the target measure is only approximately preserved. To correct for the bias introduced by the discretization one can make use of the M-H accept-reject mechanism, which guarantees that the resulting chain will be reversible with respect to the target measure; in this way we can guarantee that, if the chain converges, it can only converge to the correct measure. To summarize, the MALA algorithm is as follows: suppose at step n we are in X_n. From X_n we propose to move to Y_{n+1}

$$Y_{n+1} := X_n - h\nabla V(X_n) + \sqrt{2h}\,\xi_{n+1}, \qquad \xi_{n+1} \sim \mathcal{N}(0, 1)\,.$$

Using (4.4)[14] we then accept or reject the move to Y_{n+1}. If Y_{n+1} is accepted we set $X_{n+1} = Y_{n+1}$, otherwise $X_{n+1} = X_n$.

[14] In this case $q(x, \cdot) \sim \mathcal{N}(x - h\nabla V(x), 2hI_d)$

We stress again that in the context of the MALA algorithm the accept-reject mechanism can be seen as a way of properly discretizing the first order Langevin dynamics. The resulting chain is reversible with respect to the target distribution. Finally, also the MALA algorithm sufferes from the curse of dimensionality.

4.2. **Sampling measures defined on infinite dimensional spaces.** As in Section 3.1, let $\mathcal{H}$ be a separable Hilbert space. Throughout the remainder of the paper we assume that C is a bounded, positive and symmetric operator on $\mathcal{H}$ with associated eigenvalues $\{\lambda_j^2\}_{j\in\mathbb{N}}$ and orthonormal eigenvectors $\{\varphi_j\}_{j\in\mathbb{N}}$, that is

$$C\varphi_j = \lambda_j^2\varphi_j.$$

We will also assume that C is trace class [15] and that for some $\kappa > 1/2$ we have

$$\lambda_j \asymp j^{-\kappa}.$$ [16]

The next two algorithms that we present are aimed at sampling from measures on the space $\mathcal{H}$, in particular from measures of the form

$$d\pi(q) \propto e^{-\Phi(q)}d\pi_0(q),\ ^{17} \qquad \pi_0 \sim \mathcal{N}(0,C),\ q \in \mathcal{H}. \tag{4.7}$$

That is, the measure π that we want to sample from is a change of measure from the underlying Gaussian π_0. By the Bayesian point of view, (4.7) can be interpreted to be a posterior measure, given prior π_0 and likelyhood Φ. More details on the functional setting and in general on the material of this section can be found e.g. in [4, 36]. For background reading on Gaussian measures on infinite dimensional spaces see [7]. It is natural to wonder why we would want to sample from a measure that is defined on an infinite dimensional space. We explain this fact with an example.

Example 4.2 (Conditioned Diffusions)**.** Consider the Langevin equation (4.5) in a double well potential. That is, $V(x)$ is confining and has two minima, say x^- and x^+. Suppose we are interested only in the paths X_t that satisfy (4.5), together with $X(0) = x^-$ and $X(1) = x^+$. It is well known that, at least for low temperatures, if we start the path in x^-, the jump to the other potential well is a rare event, so just simulating (4.5) subject to the initial condition $X(0) = x^-$ does not sound like a good idea. The approach that we want to present here is the following: one can prove that the measure on path space (i.e. on $L^2[0,1]$) induced by the diffusion (4.5), with $X(0) = x^-$ and $X(1) = x^+$, is indeed of the form (4.7) [36, Section 3.8 and references therein]. Sampling from such a measure means extracting information from the desired paths. □

If we want to sample from π by using the MCMC approach, then we need to construct a chain x_n, defined on $\mathcal{H}$, $\{x_n\} \subset \mathcal{H}$, such that π is the only invariant measure of x_n and x_n converges to π as well. In other words, we need to construct an algorithm that is well defined on the infinite dimensional space $\mathcal{H}$. Assume we have been able to find such an algorithm. It is clear that in computational practice we cannot use the infinite dimensional algorithm directly. So instead of using the chain x_n, we will use the chain x_n^N, which is obtained by projecting each element of x_n on the space $\mathcal{H}^N$. Therefore $\{x_n^N\} \subset \mathbb{R}^N$. One can prove that the chain obtained in this way, as projection of an infinite dimensional algorithm, *does not* suffer from the curse of dimensionality. For example, the RWM algorithm suffers from the curse of dimensionality (and it is in fact not

[15] We recall that a bounded, positive and symmetric operator on a Hilbert space is *trace class* if $\sum_{k=1}^{\infty}\langle C\varphi_k,\varphi_k\rangle < \infty$.

[16] The notation $\asymp$ means : there esist two positive constants $c_1, c_2 > 0$ such that $c_1 j^{-\kappa} \le \lambda_j \le c_2 j^{-\kappa}$.

[17] We use the symbol "$\propto$" to mean "proportional to", i.e. the LHS is equal to the RHS for less than a multiplicative constant.

well defined in infinite dimension). However it can be modified in such a way that the resulting algorithm is well defined in $\mathcal{H}$; such a modification is the *pre-conditioned Crank-Nicolson* (pCN) algorithm (see [36]). It is also possible to prove that while the spectral gap of the RWM chain tends to 0 as $N \to \infty$, the spectral gap of pCN does not, see [16].

4.3. **Hybrid Monte Carlo.** In view of the previous section, we will describe a version of the HMC algorithm which is adapted to sampling from measures of the form (4.7) and is well defined in infinite dimension [5]. A very nice introduction to HMC can be found in [26]. The basic principle behind HMC is relatively simple: in order to sample from the measure π defined on $\mathcal{H}$ we will create a Markov Chain $(q_k, v_k) \in \mathcal{H} \times \mathcal{H}$ that samples from the measure Π, on $\mathcal{H} \times \mathcal{H}$, defined as follows:

$$d\Pi(q, v) \propto d\pi(q) d\pi_0(v), \quad \pi_0 \sim \mathcal{N}(0, C).$$

Notice that the measure Π is the product of our target measure with a Gaussian measure (in the v component). So effectively, in the long run, the only component of the chain that we will be interested in is the first one, which is the one that will be converging to π. The measure Π can be more explicitly written as

$$d\Pi(q, v) \propto e^{-\Phi(q)} d\pi_0(q) d\pi_0(v), \quad \pi_0 \sim \mathcal{N}(0, C).$$

If we introduce the Hamiltonian

$$H(q, v) = \frac{1}{2}\langle v, C^{-1} v\rangle + \frac{1}{2}\langle q, C^{-1} q\rangle + \Phi(q), \tag{4.8}$$

then one has

$$d\Pi(q, v) \propto e^{-H(q,v)}.$$

The Hamiltonian flow associated with the Hamiltonian function (4.8) can be written as

$$\mathcal{F}^t : \begin{cases} \dot{q} = v \\ \dot{v} = -q - C\nabla\Phi(q)\,. \end{cases}$$

The Hamiltonian flow $\mathcal{F}^t$ preserves functions of the Hamiltonian and, at least in finite dimensions, the volume element $dqdv$. It therefore preserves the measure Π. For this reason it is a natural idea to think of using a time-step discretization of the Hamiltonian flow as a proposal move to create the chain (q_k, v_k). However, like in the MALA case, we still need to discretize the flow $\mathcal{F}^t$. We discretize the Hamiltonian flow by "splitting" it into its linear and non-linear part, i.e. by using the Verlet integrator. The Verlet integrator is defined as follows: let R^t and Θ^t be the flows associated with the following ODEs:

$$R^t : \begin{cases} \dot{q} = v \\ \dot{v} = -q \end{cases} \qquad \Theta^t : \begin{cases} \dot{q} = 0 \\ \dot{v} = -C\nabla\Phi(q) \end{cases} \tag{4.9}$$

and let

$$\chi_\tau := \Theta^{\tau/2} \circ R^\tau \circ \Theta^{\tau/2}\,. \tag{4.10}$$

A time step discretization (of size h) of the flow $\mathcal{F}^t$ is then given by

$$\chi_\tau^h = \chi_\tau \circ \cdots \circ \chi_\tau \quad \left[\frac{h}{\tau}\right] \text{ times.} \tag{4.11}$$

We now have all the notation in place to introduce the HMC algorithm. Suppose at time k the first component of the chain is in q_k. Then

(1) pick $v_k \sim \mathcal{N}(0, C)$;
(2) compute
$$(q_{k+1}^*, v_{k+1}^*) = \chi_\tau^t(q_k, v_k),$$
and propose q_{k+1}^* as next move;

(3) calculate the acceptance probability α_k, according to

$$\alpha_k = 1 \wedge e^{-(H(\chi_\tau^t(q_k,v_k))-H(q_k,v_k))}; \tag{4.12}$$

(4) set $q_{k+1} = q_{k+1}^*$ with probability α. Otherwise $q_{k+1} = q_k$.

Remark 4.3. Some comments are in order:

- Notice that at each step the component v_k is sampled independently from q_k. If the velocity variable was not resampled, the algorithm would be stuck in areas with approximately the same probability.
- If $\mathcal{H}$ is infinite dimensional, the Hamiltonian function (4.8) is almost surely infinite. However in order for the algorithm to be well defined, all we need is for the difference $(H(\chi_\tau^t(q_k, v_k)) - H(q_k, v_k)$ appearing in (4.12) to be finite. This is indeed the case (and the choice of integrator was in fact driven by the need to satisfy this requirement [5]).
- The generated chain is reversible with respect to the target density function.
- The above algorithm is well posed in infinite dimension i.e. for $(q, v) \in \mathcal{H} \times \mathcal{H}$.

□

5. An irreversible MCMC algorithm: the SOL-HMC

We now want to construct an MCMC algorithm which results from appropriately discretizing the Second Order Langevin equation. The algorithm that we will present has been introduced in [23] and can be understood as a generalization of [17]. In order to carry out such a discretization we will make use of a modification of the HMC algorithm which we have just presented. Again, we want to sample from a measure π of the form (4.7). First of all, let us rewrite the SOL equation in a way adapted to our context

$$\begin{aligned} dq &= v\, dt \\ dv &= [-q - C\,\nabla\Phi(q)]\, dt - v\, dt + \sqrt{2C} dW_t\,. \end{aligned} \tag{5.1}$$

Equation (5.1)is well posed in an infinite dimensional context [23], it is ergodic and it admits our target π as unique invariant measure. Again, like for the MALA algorithm, if we discretize the equation naively we risk to destroy all the good properties of the dynamics. In particular, if we were to discretize and then use the Metropolis-Hastings accept-reject mechanism, we would end up with a chain that does sample from the correct measure, but such a chain would be reversible. What we want to do here instead is to discretize the irreversible Markov dynamics (5.1) in such a way to produce an irreversible chain. It is clear that in order to do so we will have to leave the comfort of the Metropolis-Hastings setting.

In order to present the SOL-HMC algorithm, we first need to introduce the numerical integrator that we will use. To integrate (5.1) numerically, we construct an integrator which takes advantage of the structure of the equation highlighted in Section 3.1. Namely, we look again at the splitting "Hamiltonian + OU process". Recall the definition of the flows R^t, Θ^t, equation (4.9), and define $\mathcal{O}^t$, to be the map that gives the solution at time t of the system

$$\mathcal{O}^t : \begin{cases} \dot{q} = 0 \\ \dot{v} = -v\, dt + \sqrt{2C} dW_t \end{cases}$$

Let χ_τ and χ_τ^h be defined as in (4.10) and (4.11), respectively. For given positive parameters h and δ (to be appropriately tuned), the proposal move and acceptance probability of the SOL-HMC algorithm are then given by

$$(q^*, v^*) = (\chi_\tau^h \circ \mathcal{O}^\delta)(q, v) \tag{5.2}$$

and

$$\alpha = 1 \wedge e^{-[H(q^*,v^*)-H(\mathcal{O}^\delta(q,v))]}, \tag{5.3}$$

respectively. With this notation in place, the SOL-HMC algorithm proceeds as follows:

(1) given (q_k, v_k), let
$$(q'_k, v'_k) = \mathcal{O}^\delta(q_k, v_k)$$
and propose
$$(q^*_{k+1}, v^*_{k+1}) = (\chi^u_\tau)(q'_k, v'_k);$$
(2) calculate the acceptance probability α_k, according to (5.3);
(3) set
$$(q_{k+1}, v_{k+1}) = \begin{cases} (q^*_{k+1}, v^*_{k+1}) & \text{with probability } \alpha \\ \\ \mathcal{O}^\delta(q_k, -v_k) & \text{with probability } 1-\alpha\,. \end{cases}$$

In words: if at step k we are in (q_k, v_k), we first calculate (q'_k, v'_k) (notice that $q'_k = q_k$). Then we propose a move to (q^*_{k+1}, v^*_{k+1}). If the move is accepted then $(q_{k+1}, v_{k+1}) = (q^*_{k+1}, v^*_{k+1})$. Otherwise we change the sign of the velocity, i.e. we consider $(q_k, -v_k)$ and evolve for time δ according to $\mathcal{O}^\delta$, so that $(q_{k+1}, v_{k+1}) = \mathcal{O}^\delta(q_k, -v_k)$. Notice that in case of rejection of the proposal (q^*_{k+1}, v^*_{k+1}) we do not stay where we started from, i.e. in (q_k, v_k), but we move to $\mathcal{O}^\delta(q_k, -v_k)$.

Remark 5.1. Again, let us make a few observations about the algorithm.

- The relevant energy difference here is $H(q', v') - H(q^*, v^*)$ (rather than $H(q, v) - H(q^*, v^*)$); indeed the first step in the definition of the proposal (q^*, v^*), namely the OU process $\mathcal{O}^\delta(q, v)$, is based on an exact integration and preserves the desired invariant measure. Therefore the accept-reject mechanism (which is here only to account for the numerical error made by the integrator χ^h_τ) doesn't need to include also the energy difference $H(q, v) - H(q', v')$.
- The flip of the sign of the velocity in case of rejection of (q^*, v^*) is there to guarantee that the overall proposal moves are symmetric. This is done in order to ensure that the acceptance probability can be defined only in terms of the ratio $\Pi(q^*, v^*)/\Pi(q', v')$, i.e. in terms of the energy difference $H(q', v') - H(q^*, v^*)$. An interesting discussion on the matter can be found in [20, Chapter 2].
- The algorithm is well posed in finite as well as in infinite dimension.
- Most importantly, the algorithm produces an irreversible chain. How did we lose reversibility? The important observation that this algorithm is based on is the following [26]: detailed balance is not preserved under composition. That is, if we consider a Markov transition kernel, say r, resulting from the composition of transition kernels, each of them satisfying detailed balance, r does not, in general, satisfy detailed balance as well. In the same way, *each step of the SOL-HMC algorithm satisfies detailed balance; however their composition does not.* □

Beyond [17, 23] the only other MCMC irreversible algorithms that we know of are [6, 8] (see also references therein). The advantages of irreversibility by the point of view of asymptotic variance have also been investigated in [9, 28, 29].

Acknowledgments. The author is grateful to the anonymous referee for very useful comments that helped improving the paper.

Appendix A.

Lemma A.1. *With the setting and assumptions of Example 4.1, if π is not the uniform distribution then the chain x_n with transition probabilities $p(x, y)$ defined in* (4.2) *converges in the*

sense (4.1) *to the target distribution* π *for any choice of the (irreversible and symmetric) proposal matrix* Q*. If* π *is the uniform distribution then convergence may happen or not, depending on* Q*.*

Proof. (See [2] for more details on this proof) The proof is quite simple so we only sketch it. A time-homogeneous Markov Chain (MC) on a finite state space S is said to be *regular* if there exists a positive integer $k > 0$ such that $p^k(x, y) > 0$ for all $x, y \in S$. Clearly a regular MC is irreducible. It is easy to prove the following: if for any x and y in S there exists an integer $n > 0$ such that $p^n(x, y) > 0$ and there exists $a \in S$ such that $p(a, a) > 0$ then the chain is regular. (Notice that k is independent of x and y whereas $n = n(x, y)$ i.e. it depends on the choice of x and y.) A standard result in the basic theory of MCs states that if x_n is a regular chain on a finite state space then the chain has exactly one stationary distribution, π, and

$$\lim_{n\to\infty} p^n(x, y) = \pi(y), \quad \text{for all } x \text{ and } y \in S. \tag{A.1}$$

With these premises, and assuming that π is not the uniform distribution on S, we want to show that the chain with transition matrix P is regular. Recall that Q is irreducible hence P is irreducible as well, therefore it is true that for all x, y there exists $n = n(x, y) > 0$ such that $p^{n(x,y)}(x, y) > 0$. Therefore we only need to find a state $a \in S$ such that $p(a, a) > 0$. Let M be the set $M = \{x \in S : \pi(x) = \max_{y\in S} \pi(y)\}$. Because Q is irreducible there exist $a \in M$ and $b \in M^c$ such that $q(a, b) > 0$ and clearly by construction $\pi(a) > \pi(b)$. Notice also that from the definition of P, $p(x, y) \le q(x, y)$ for all $x \neq y$. Then

$$\begin{aligned}
p(a, a) &= 1 - \sum_{x\neq a} p(a, x) = 1 - \sum_{x\neq a,b} p(a, x) - p(a, b) \\
&\ge 1 - \sum_{x\neq a,b} q(a, x) - q(a, b)\pi(b)/\pi(a) \\
&= 1 - \sum_{x\neq a} q(a, x) + q(a, b)\left[1 - \pi(b)/\pi(a)\right] \\
&= q(a, a) + q(a, b)\left[1 - \pi(b)/\pi(a)\right] \ge q(a, b)\left[1 - \pi(b)/\pi(a)\right] > 0.
\end{aligned}$$

On the other hand if $\pi(x)$ is the uniform distribution on S then $p(x, y) = q(x, y)$ so, because $q(x, y)$ is symmetric, detailed balance is still satisfied so π is still invariant [18]. However if $q(x, y)$ is periodic then convergence in the sense (A.1) does not take place. (However ergodic averages will still converge). □

References

[1] A. Arnold and J. Erb. *Sharp entropy decay for hypocoercive and non–symmetric Fokker–Planck equations with linear drift.* arXiv:1409.5425

[2] P. Baldi, Calcolo delle Probabilità e Statistica. Second Edition. Mc Graw-Hill, Milano 1998.

[3] D. Bakry, I. Gentil and M. Ledoux. Analysis and geometry of Markov Diffusion operators. Springer, 2014.

[4] A. Beskos and A. M. Stuart. MCMC methods for sampling function space. In *ICIAM Invited Lecture 2007.* European Mathematical Society, Zürich.

[5] A. Beskos, F. Pinski, J.M. Sanz-Serna, and A.M. Stuart. Hybrid Monte-Carlo on Hilbert spaces. *Stochastic Processes and Applications*, 121:2201–2230, 2011.

[6] J. Bierkens. *Non-Reversible Metropolis Hastings.* arXiv:1401.8087

[7] G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions*, Cambridge University Press, Cambridge, 1992.

[8] P.Diaconis, S. Holmes and R. M. Neal. *Analysis of non-reversible markov Chain Sampler.* Annals of Applied Probability, Vol. 10, No. 3, 726–752, 2000.

[18] Notice that, in the setting of this lemma, if $q(x, y)$ is not symmetric then the uniform distribution might not be an invariant measure

[9] A. Duncan, T. Lelievre, G.A. Pavliotis. *Variance reduction using non-reversible Langevin Samplers*. arXiv:1506.04934
[10] R. Durret. *Probability: theory and examples*, 2010.
[11] P. Diaconis. *The Markov chain Monte Carlo revolution*. Bull. Amer. Math. Soc. 46, 179–205, 2009.
[12] R. Durrett. *Probability: theory and examples*. Fourth edition. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.
[13] S.Gadat and L. Miclo. *Spectral decompositions and L2-operator norms of toy hypocoercive semi-groups*. Kinet. Relat. Models 6, no. 2, 317372, 2013.
[14] W.R. Gilks, S. Richardson and D.J. Spiegelhalter. Markov Chain Monte Carlo in practice. Chapman and Hall, London, 1996.
[15] A. Guionnet and B. Zegarlinski. *Lectures on Logarithmic Sobolev inequalities*. Lecture Notes.
[16] M. Hairer, A. M. Stuart and S. Vollmer. *Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions*. arxiv 1202.0709
[17] A. M. Horowitz. A generalized guided Monte Carlo algorithm. Physics Letters B, 268(2):247 – 252, 1991
[18] C. Hwang, S. Hwang-Ma and S. Sheu. *Accelerating diffusions*. The Annals of Applied Probability, 2005.
[19] I. Kontoyiannis and S. P. Meyn. *Geometric ergodicity and the spectral gap of non-reversible Markov chains*. Probab. Theory Relat. Fields 154:327 – 339, 2012.
[20] T. Lelièvre, M. Rousset and G. Stoltz. *Free energy computations: A mathematical perspective*. Imperial College Press, 2010
[21] T. Lelièvre, F. Nier, G. A. Pavliotis. *Optimal Non-reversible Linear Drift for the Convergence to Equilibrium of a Diffusion*, J. Stat. Phys. **152** (2013), 237–274.
[22] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.
[23] M.Ottobre, N. Pillai, F. Pinski and A.M.Stuart. *A Function Space HMC Algorithm With Second Order Langevin Diffusion Limit*. To appear in Bernoulli.
[24] M. Ottobre, G.A. Pavliotis, K. Pravda-Starov. *Exponential return to equilibrium for hypoelliptic quadratic systems*, J. Funct. Anal. **262** (2012), 4000–4039.
[25] G. Pavliotis. *Stochastic Process and Applications*. Springer, 2015.
[26] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.
[27] M. Reed and B. Simon. *Methods of Modern Mathematical Physics*. Academic Press, New York, 1980.
[28] L. Rey-Bellet and K. Spiliopoulos. *Irreversible Langevin samplers and variance reduction: a large deviation approach*. Nonlinearity, Vol. 28, pp. 2081-2103, 2015.
[29] L. Rey-Bellet and K. Spiliopoulos. *Variance reduction for irreversible Langevin samplers and diffusion on graphs*. Electronic Communications in Probability, Vol. 20, no. 15, pp. 116, 2015.
[30] C.P. Robert and G. Casella. Introducing Monte Carlo methods with R. Use R!. *Springer, New York*, 2010.
[31] G. O. Roberts, A. Gelman and W. R. Gilks. *Weak convergence and optimal scaling of random walk Metropolis algorithms*. Ann. Appl. Probab., 1997.
[32] G.Roberts and R. Tweedie. *Exponential convergence of Langevin distributions and their discrete approximations*. Bernoulli, 341–363, 1996.
[33] G.Roberts and R. Tweedie. *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms*. Biometrika, 95–110, 1996.
[34] C. Sherlock, P. Fearnhead and G.O. Roberts. The random walk Metropolis: linking theory and practice thorugh a case study.
[35] D.W. Stroock and S.R.S. Varadhan.*Multidimensional diffusion processes*. Springer, Berlin, 1979.
[36] A. M. Stuart. *Inverse Problems: a Bayesian Perspective*. Acta Numerica, **19**, 451–559, 2010.
[37] L. Tierney. *A note on Metropolis-Hastings kernels for general state spaces*. Ann. Appl. Probab. 8 (1998), no. 1, 1–9.
[38] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950), 2009.
[39] D.Williams. *To begin at the beginning:...* Proc. Sympos., Univ. Durham, Durham, 1980. pp. 1-55, Lecture Notes in Math., 851, Springer, Berlin-New York, 1981.

Michela Ottobre, Department of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, UK

E-mail address: m.ottobre@hw.ac.uk