SAMPLING WITH THE LANGEVIN DIFFUSION

BENJAMIN BRAIMAN

1. Introduction

The purpose of this expository thesis is to provide a brief introduction to the theory of log-concave sampling. In this setting, one attempts to generate a sample from a unknown density function f where $\log f$ is concave. Currently, the most popular way of obtaining quick convergence results is sampling from the Langevin diffusion

$$dX_t = -\nabla \log f(X_t) + \sqrt{2}\, dB_t.$$

In this idealized continuous-time case, the distribution of X_t converges to f exponentially fast in certain commonly used metrics and divergences. However, in the real-world setting, one is unable to exactly sample directly from the continuous-time Langevin diffusion, unless f is a very nice function (such as a Gaussian) and the Langevin SDE can be explicitly solved. Numerical techniques are required. This has led to a vast wealth of literature, results from which we exhibit in this article (see, for example [8], [28], [13], [26], [30]).

This thesis is organized into three sections. The first section is a review of some preliminary theory; in particular, we highlight a few important results from stochastic analysis, Markov processes, and optimal transport theory which are considered standard in the literature. We then proceed to apply this theory to the Langevin diffusion, performing first a continuous time analysis to illustrate the idealized case. We then exhibit a few numerical techniques from [8] and [7] to illustrate how the Langevin diffusion can be used to generate samples from a stationary distribution in practical applications.

The final section is devoted to an important modification of Langevin Monte-Carlo known as "Stochastic Gradient Langevin Dynamics." In this case, one assumes that the target potential f can be written as a product

$$f = f_1 f_2 \cdots f_m$$

of functions. Rather than spend precious resources computing $\nabla \log f_i$ for each i, this approach tries to choose a random subset of the f_i and compute the gradient over this subset. We present a few results which illustrate the care that needs to be taken when sampling the f_i, and how altering the sampling method can result in good and poor results.

2. Preliminary Theory

In this section, we review some of the technical preliminaries necessary for studying the Langevin diffusion. We assume that the reader has some familiarity with measure-theoretic probability and real analysis at the graduate level. Most of the relevant assumed background can be found in the classical texts [15], [14], and [24]. This section is primarily devoted to reviewing and introducing additional theory that is important to the study of Langevin MCMC.

Stochastic Calculus. Nearly everything explored in this work relies on the theory of stochastic integration and Markov chains. Therefore, it is worthwhile to briefly review the relevant mathematical theory.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. We assume that $(\Omega, \mathcal{F}, \mathbf{P})$ is large enough to support a standard Brownian motion on $\mathbb{R}^d$. That is, a stochastic process $(B_t)_{t\geq 0} \subset \mathbb{R}^d$ such that

(1) $B_0 = 0$.
(2) The mapping $t \mapsto B_t$ is continuous a.s.
(3) If $0 \leq s_1 < s_2 < t_1 < t_2$, then $B_{t_2} - B_{t_1}$ is independent of $B_{s_2} - B_{s_1}$.
(4) $B_t - B_s$ is normally distributed with mean 0 and covariance $(s-t)I_d$, where I_d is the d-dimensional identity matrix.

For $t \geq 0$, we let $\mathcal{F}_t = \sigma(B_s : 0 \leq s \leq t)$ be the σ-algebra generated by $(B_s)_{0\leq s\leq t}$. Recall that a stochastic process $(X_s)_{s\geq 0}$ is *adapted* to $(\mathcal{F}_s)_{s\geq 0}$ if $X_s \in \mathcal{F}_s$ for each $s \geq 0$.

We now introduce one of the most important types of stochastic process.

Definition 2.1. A matrix-valued process $(X_s)_{s\geq 0} \subset \mathbb{R}^m$ adapted to $\mathcal{F}_s$ is said to be a *martingale* if for each $0 \leq s < t$, it holds $\mathbf{E}[X_t \,|\, \mathcal{F}_s] = X_s$ a.s. If $\mathbf{E}|X_t|^2 < \infty$ for each $t > 0$, then we say that X is an L^2 martingale.

It is easy to verify that standard Brownian motion is a martingale according to this definition. It turns out that we can manufacture L^2 martingales in great abundance using Itô integration. To construct the Itô integral, first we declare that an adapted process $X = (X_s)_{s\geq 0} \subset \mathbb{R}^{m\times d}$ is a *simple process* if it is of the form

$$X_t = X_0 \mathbf{1}_{\{0\}}(t) + \sum_{i=1}^{n} H_{t_{i-1}} \mathbf{1}_{(t_{i-1}, t_i]}(t),$$

where $H_{t_{i-1}} \in \mathcal{F}_{t_{i-1}}$ is an L^2 random variable for each $i \geq 1$ Given a simple process X, we will define the Itô $X \cdot B$ of X as the process

$$(X \cdot B)_t = \sum_{i=1}^{n} H_{t_{i-1}} (B_{t_i \wedge t} - B_{t_{i-1} \wedge t}).$$

Proving that $X \cdot B$ is well-defined follows similarly to proving that the Lebesgue-integral is well-defined for simple functions. Clearly, $((X \cdot B)_t)_{t\geq 0} \subset \mathbb{R}^m$ is adapted to $\mathcal{F}_s$ if

B is a d-dimensional standard Brownian motion. If, in addition, it is assumed that $H_{t_{i-1}} \in L^2(\mathbf{P})$, then $(X \cdot B)_t \in L^2(\mathbf{P})$ is a continuous process for each $t \geq 0$. Moreover, $X \cdot B$ satisfies the *Itô isometry*

$$\mathbf{E}[(X \cdot B)_t^2] = \mathbf{E} \int_0^t \|X_s\|^2 \, ds, \tag{2.1}$$

where $\|\cdot\|$ is the Hilbert-Schmidt norm on the space of $\mathbb{R}^{m \times d}$ matrices defined by $\|X\| = \sqrt{\mathrm{Tr}(XX^\top)}$. If $\mathcal{B}([0,t])$ is the Borel σ-algebra on $[0,t]$, then by the density of the simple process in $L^2(\mathcal{B}([0,t]) \otimes \mathcal{F}_t)$, the Itô isometry allows one to uniquely define the Itô integral $X \cdot B$ for any L^2 process X such that $X \cdot B$ is continuous, $\mathcal{F}_s$ adapted, and satisfies the Itô isometry 2.1. Often, the alternative notation

$$\int_0^t X_s \cdot \, dB_s := (X \cdot B)_t$$

is used if it is convenient. For a more detailed treatment of Itô integration, see any one of [10], [18], or [31].

A particular class of stochastic process of interest to our investigation are the Itô diffusions. We say that a process $(X_s) \subset \mathbb{R}^m$ is an *Itô diffusion* if it is of the form

$$X_t = X_0 + \int_0^t b_s \, ds + \int_0^t \sigma_s \cdot \, dB_s \tag{2.2}$$

for $X_0 \in \mathcal{F}_0$ and $\mathcal{F}_s$ adapted processes $(b_s)_{s \geq 0} \subset \mathbb{R}^m$ and $(\sigma_s)_{s \geq 0} \subset \mathbb{R}^{m \times d}$. It is often more convenient to use the shorthand

$$dX_t = b_s \, ds + \sigma_s \cdot dB_s. \tag{2.3}$$

Theorem 2.2. *Let $(X_s)_{s \geq 0} \subset \mathbb{R}^m$ be an Itô diffusion and $f \in C^{1,2}([0,\infty) \times \mathbb{R}^m)$. Then the process $(f(s, X_s))_{s \geq 0}$ is also an Itô diffusion. Namely,*

$$f(t, X_t) = f(0, X_0) + \int_0^t Af(s, X_s) \, ds + \int_0^t Bf(s, X_s) \, dB_s, \tag{2.4}$$

for a one-dimensional Brownian motion B, where

$$Af(s, X_s) = f_t(s, X_s) + \nabla_x f(s, X_s) \cdot b_s + \frac{1}{2} \mathrm{Tr}(H_x f(s, X_s) \sigma_s \sigma_s^\top)$$

and

$$Bf(s, X_s) = \nabla_x f(s, X_s)^\top \sigma_s.$$

As usual, $H_x f(t,x)$ denotes the Hessian matrix of f in the variable x. An important special case arises when $dX_t = dB_t$ (i.e., σ is the identity matrix and $b_s = 0$). In this case, we recover the traditional form of Itô's lemma:

$$f(t, B_t) = f(0, B_0) + \int_0^t \left[f_t(s, B_s) + \frac{1}{2} \triangle_x f(s, B_s) \right] ds + \int_0^t \nabla_x f(s, B_s) \cdot \, dB_s,$$

where $\triangle_x = \sum_{j=1}^m \frac{\partial^2}{\partial x_j^2}$ is the Laplacian. The proof is a long, tedious, but straightforward calculation. The essential idea is captured in the case $m = d = 1$. Given a partition

$0 = t_0 < \cdots < t_n = t$ of $[0,t]$, one expands f to the second order by

$$f(t_i, B_{t_i}) = f(t_{i-1}, B_{t_{i-1}}) + f_t(t_{i-1}, B_{t_{i-1}})\Delta t_i + f_x(t_{i-1}, B_{t_{i-1}})\Delta B_i + \frac{1}{2} f_{xx}(t_{i-1}, B_{t_{i-1}})(\Delta B_i)^2 + O((\Delta t_i)^3 + (\Delta B_i)^3).$$

where $\Delta t_i = (t_i - t_{i-1})$ and $\Delta B_i = B_{t_i} - B_{t_{i-1}}$. This gives

$$f(t, B_t) - f(0, B_0) = \sum_{i=1}^{n} f_t(t_{i-1}, B_{t_{i-1}})\Delta t_i + \sum_{i=1}^{n} f_x(t_{i-1}, B_{t_{i-1}})\Delta B_i + \frac{1}{2}\sum_{i=1}^{n} f_{xx}(t_{i-1}, B_{t_{i-1}})(\Delta B_i)^2 + O\left(\sum_{i=1}^{n}(\Delta t_i)^3 + (\Delta B_i)^3\right)$$

Since $\mathbf{E}[\Delta B_i] = 0$ and $\mathbf{E}[\Delta B_i^2] = t_i - t_{i-1}$, we therefore expect

$$\sum_{i=1}^{n} f_t(t_{i-1}, B_{t_{i-1}})\Delta t_i \to \int_0^t f_t(s, B_s)\, ds, \quad \sum_{i=1}^{n} f_x(t_{i-1}, B_{t_{i-1}})\Delta B_i \to \int_0^t f_x(s, B_s)\, dB_s,$$

$$\sum_{i=1}^{n} f_{xx}(t_{i-1}, B_{t_{i-1}})(\Delta B_i)^2 \to \int_0^t f_{xx}(s, B_s)\, ds,$$

and the error term to vanish as $\sup_{1\le i\le n}(t_i - t_{i-1}) \to 0$ with respect to $L^2(\mathcal{B}([0,t]) \otimes \mathcal{F}_t)$. The formal proof of Itô's lemma just adds all the necessary epsilons to make this informal argument rigorous. We refer the reader to any number of standard texts, such as [31], [18], and [10].

An important subclass of Itô diffusions are known as *stochastic differential equations* (or SDE). Namely, if B is a standard $\mathbb{R}^d$-valued Brownian motion and $b : \mathbb{R}^m \to \mathbb{R}^m$ and $\sigma : \mathbb{R}^m \to \mathbb{R}^{m\times d}$ are sufficiently nice functions, then there exists a continuous process X which solves the stochastic differential equation

$$dX_t = b(X_t)\, dt + \sigma(X_t) \cdot\, dB_t.$$

Solutions to SDE's need not exist nor be unique. We postpone the existence theorem for SDE's until after we have discussed Markov processes.

Markov Processes. It turns out that all Itô diffusions are examples of *Markov processes*. Roughly, a Markov process is a stochastic process X adapted to $\mathcal{F}_s$ such that

$$\mathbf{P}[X_{t+s} \in A \,|\, \mathcal{F}_s] = \mathbf{P}[X_t \in A \,|\, X_s]$$

for each $0 \le s < t$. Informally, this means that X only depends on the most recent information available; it is forgetful of the distant past. However, working with this definition alone is somewhat unsatisfactory, since it fails to encode the time homogeneity we typically expect. To firm up the theory, we need to introduce a definition.

Definition 2.3. A $\mathbb{R}^d$-valued stochastic process X is said to be a Markov process if there exists a family of measures $\{\mathbf{P}_x : x \in \mathbb{R}^d\}$ on Ω such that:

(1) $\mathbf{P}_x[X_0 = x] = 1$.
(2) For each event A, the mapping $x \mapsto \mathbf{P}_x[A]$ is measurable.

(3) For a Borel set $A \subset \mathbb{R}^d$, it holds $\mathbf{P}_x[X_{t+s} \in A \,|\, \mathcal{F}_t] = \mathbf{P}_{X_t}[X_s \in A]$

As usual, $\mathcal{F}_t$ is the filtration generated by the variables $\{X_s : 0 \leq s \leq t\}$.

In words, each measure $\mathbf{P}_x$ describes the dynamics of the chain X given that we start it at $x \in \mathbb{R}^d$. We can naturally expand this definition to include the scenario when $X_0 \sim \mu$ for an arbitrary probability measure μ by defining

$$\mathbf{P}_\mu[A] := \int_{\mathbb{R}^d} \mathbf{P}_x[A]\, d\mu(x).$$

That this is a well-defined measure follows immediately from the standard convergence properties of the Lebesgue integral. Of course, there is a corresponding family of expectations $\mathbf{E}_x$ defined by

$$\mathbf{E}_x[Y] = \int_\Omega Y\, d\mathbf{P}_x.$$

By utilizing the density of the simple functions in $L^1(\mathbf{P})$, one can see that $x \mapsto \mathbf{E}_x Y$ is measurable for each $x \in \mathbb{R}^d$ and that

$$\mathbf{E}_x[f(X_{t+s}) \,|\, \mathcal{F}_t] = \mathbf{E}_{X_t}[f(X_s)]$$

$\mathbf{P}_x$ a.s., a property which gives rise the so-called Markov semi-group associated to X.

Theorem 2.4. *Let X be a Markov process. For each $t \geq 0$, let*

$$P_t f(x) = \mathbf{E}_x[f(X_t)]$$

for a bounded measurable function f. Then $P_t f$ is a bounded measurable function, and the family of operators $(P_s)_{s \geq 0}$ forms an operator semi-group, meaning $P_{t+s} = P_t P_s$ for each $t, s \geq 0$.

Proof. To see that $P_t f$ is a bounded measurable function, observe that f is bounded means

$$|\mathbf{E}_x f(X_t)| \leq \sup |f| < \infty.$$

Measurability follows from the fact $x \mapsto \mathbf{E}_x Y$ is measurable for any $Y \in L^1(\mathbf{P})$. For the final equality,

$$\mathbf{E}_x[f(X_{s+t})] = \mathbf{E}_x[\mathbf{E}_x[f(X_{t+s}) \,|\, \mathcal{F}_t]] = \mathbf{E}_x[\mathbf{E}_{X_t}[f(X_s)]] = \mathbf{E}[P_s f(X_t)] = P_t(P_s f(x)).$$

Thus $P_{s+t} = P_t P_s$, and the equality $P_{s+t} = P_s P_t$ is proved similarly. □

Definition 2.5. Let X be a Markov process. The function $p(s, x, A) := P_s \chi_A(x)$ is called the *transition probability* of the chain X.

Proposition 2.6 (Chapman-Kolmogorov)**.** *Let X be a Markov process with transition probability $p(\cdot, \cdot, \cdot)$. Then for each Borel $A \subset \mathbb{R}^d$ and $t \geq 0$, it holds $x \mapsto p(t, x, A)$ is measurable, and for each fixed $t \geq 0$ and $x \in \mathbb{R}^d$, it holds $A \mapsto p(t, x, A)$ is a Borel probability measure on $\mathbb{R}^d$. Moreover,*

$$p(t + s, x, A) = \int_{\mathbb{R}^d} p(s, y, A) p(t, x, \, dy).$$

Proof. Measurability of $x \mapsto p(t,x,A)$ follows from 2.4. To check that $p(t,x,\cdot)$ is a Borel probability measure, first observe that $p(t,x,\emptyset) = \mathbf{E}_x\chi_\emptyset(X_t) = 0$ and $p(t,x,\mathbb{R}^d) = \mathbf{E}_x\chi_{\mathbb{R}^d}(X_t) = 1$. If $A_1, A_2, \dots$ is a countable disjoint collection of Borel sets whose union is A, then the sequence $\chi_{A_1} + \cdots + \chi_{A_n}$ increases to χ_A as $n \to \infty$, and the monotone convergence theorem gives

$$p\,(t,x,A) = \mathbf{E}_x\chi_A(X_t) = \sum_{j=1}^{\infty} \mathbf{E}_x\chi_{A_j}(X_t) = \sum_{j=1}^{\infty} p(t,x,A_j).$$

Thus $p(t,x,\cdot)$ is a measure.

Next, we claim that

$$\mathbf{E}_x[f(X_s)] = \int_{\mathbb{R}^d} f(y)p(s,x,\,dy)$$

whenever f is a bounded measurable function. Indeed, if $f = \sum_1^m a_j\chi_{A_j}$ is a simple function, then

$$\mathbf{E}_x[f(X_s)] = \sum_1^m a_j p(s,x,A_j) = \int f(y)\,p(s,x,\,dy),$$

and the claim follows immediately from the density of simple functions in $L^1(p(t,x,\cdot))$. Thus

$$p(t+s,x,A) = P_t(P_s\chi_A(x)) = \int_{\mathbb{R}^d} P_s\chi_A(y)\,p(t,x,\,dy) = \int_{\mathbb{R}^d} p(s,y,A)p(t,x,dy),$$

finishing the proof. □

Remark 2.7. The reader may wonder if it is possible to work backwards, namely if given a family of transition kernels $\{p_t(x,\cdot)\}$, it is possible to find a Markov process which has those kernels. With the axiom of choice, the answer is affirmative, and the way this is usually done is to take $\Omega = (\mathbb{R}^d)^{[0,\infty)}$ and let $\mathcal{F}$ be the Borel σ-algebra generated by the product topology. The measures $\{\mathbf{P}_x\}$ are uniquely determined on cylinder sets by the formula

$$\mathbf{P}_x[(\omega_t)_{t\geq 0} : \omega_{t_1} \in A_1, \dots, \omega_{t_n} \in A_n] =$$
$$\int_{A_1} \cdots \int_{A_n} p_{t_n - t_{n-1}}(x_{n-1}, dx_n) \cdots p_{t_2-t_1}(x_1,\,dx_2)\,p_{t_1}(x, dx_1)$$

for $0 \leq t_1 < \cdots < t_n$, and the associated process is defined by $X_t((\omega_s)_{s\geq 0}) := \omega_t$.

Definition 2.8. We say that a Markov process X is *continuous* if the mapping $t \mapsto X_t$ is continuous for all $t \geq 0$ $\mathbf{P}_x$-a.s. for every $x \in \mathbb{R}^d$. The process is said to be *Feller* if for each $f \in C_0(\mathbb{R}^d)$ it holds $P_s f \in C_0(X)$ for all fixed $s \geq 0$.

The desirability of Feller processes are that the associated semi-group (P_s) maps bounded continuous functions to bounded continuous functions. For a detailed discussion of Feller processes, we refer the reader to the pedagogical treatment given by [19].

At this point it is worthwhile to begin connecting the theory of Markov processes to Brownian motion.

Proposition 2.9. *Let B be a d-dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then B is a continuous Feller process, and the transition probability kernels are given by*

$$p(t, x, A) = \frac{1}{(2\pi t)^{d/2}} \int_A \exp\left(-\frac{1}{2t}|y - x|^2\right) dy$$

Proof. By definition, B is a continuous process. Suppose that B is started at a point $x \in \mathbb{R}^d$. Then we can write $B = x + W$ for a standard Brownian motion W, from which it follows

$$p(t, x, A) = \mathbf{P}_x[B_t \in A] = \frac{1}{(2\pi t)^{d/2}} \int_A \exp\left(-\frac{1}{2t}|y - x|^2\right) dy.$$

This is a continuous function of x by the dominated convergence theorem. We now need to show the Markov property. For fixed $t, s \geq 0$, it holds $X := B_{t+s} - B_t \sim \mathcal{N}(0, sI_d)$ and is independent of $\mathcal{F}_t$. Let V be a Borel subset of $\mathbb{R}^j$ for some $j \geq 1$, and pick points $t_1, \ldots, t_j \in [0, t]$. Let $\Delta B_1 = B_1$, and for $2 \leq i \leq j$, let $\Delta B_i = B_{t_i} - B_{t_{i-1}}$. Let $T : \mathbb{R}^j \to \mathbb{R}^j$ be the invertible linear map defined by $T(x_1, \ldots, x_j) = (x_1, x_2 - x_1, \ldots, x_j - x_{j-1})$. Then $(B_{t_1}, \ldots, B_{t_j}) \in V$ if and only if $T(B_1, \ldots, B_j) = (\Delta B_1, \Delta B_2, \ldots, \Delta B_j) \in T(V)$. Therefore, if μ is the law of $B_{t+s} - B_t$, μ_{j+1} is the law of $B_t - B_{t_j}$, and μ_i is the law of ΔB_i when $1 \leq i \leq j$, it holds

$$\begin{aligned}
&\mathbf{E}_x[\mathbf{1}_V(B_{t_1}, \ldots, B_{t_j}\mathbf{1}_A(B_{t+s})] \\
&= \mathbf{E}_x[\mathbf{1}_{T(V)}(\Delta B_1, \Delta B_2, \ldots, \Delta B_j)\mathbf{1}_A(B_{t+s} - B_t + B_s)] \\
&= \int \mathbf{1}_{T(V)}(x_1, \ldots, x_j)\mathbf{1}_A(y + x_{j+1} + \cdots + x_1)\, d\mu(y)\, \mu_{j+1}(x_{j+1}) \cdots d\mu_1(x_1).
\end{aligned}$$

Now, we observe that

$$\begin{aligned}
\int \mathbf{1}_A(y + x_{j+1} + \cdots + x_1)\, d\mu(y) &= \frac{1}{(2\pi s)^{d/2}} \int_A \exp\left(-\frac{1}{2s}|y - x_{j+1} + \cdots - x_1|^2\right) dy \\
&= \mathbf{P}_{x_1 + \cdots + x_{j+1}}[B_s \in A]
\end{aligned}$$

since $B_{t+s} - B_t$ is normally distributed. Thus

$$\begin{aligned}
&\int \mathbf{1}_{T(V)}(x_1, \ldots, x_j)\mathbf{1}_A(y + x_{j+1} + \cdots + x_1)\, d\mu(y)\, \mu_{j+1}(x_{j+1}) \cdots d\mu_1(x_1) \\
&= \int \mathbf{1}_{T(V)}(x_1, \ldots, x_j)\mathbf{P}_{x_1 + \cdots + x_{j+1}}[B_s \in A]\, d\mu_1(x_1) \cdots d\mu_{j+1}(x_{j+1}) \\
&= \mathbf{E}_x[\mathbf{1}_V(B_{t_1}, \ldots, B_{t_j})\mathbf{P}_{B(t)}[B_s \in A]].
\end{aligned}$$

This proves that

$$\mathbf{E}_x[\mathbf{1}_V(B_{t_1}, \ldots, B_{t_j})\mathbf{1}_A(B_{t+s})] = \mathbf{E}_x[\mathbf{1}_V(B_{t_1}, \ldots, B_{t_j})\mathbf{P}_{B(t)}[B_s \in A]]$$

for all finite sequences of times $t_1, \ldots, t_j \leq t$. It follows from the π-λ theorem that

$$\mathbf{E}_x[\mathbf{1}_U\mathbf{1}_A(B_{t+s})] = \mathbf{E}_x[\mathbf{1}_U\mathbf{P}_{B(t)}[B_s \in A]]$$

for each set $U \in \mathcal{F}_t$, and we deduce $\mathbf{P}_x[B_{t+s} \in A \,|\, \mathcal{F}_t] = \mathbf{P}_{B(t)}[B_s \in A]$ $\mathbf{P}_x$-a.s. This proves that B is a Markov process. To establish that it is Feller, merely observe that

$$\mathbf{E}_x[g(X_t)] = \frac{1}{(2\pi t)^{d/2}} \int_{-\infty}^{\infty} g(y) \exp\left(-\frac{1}{2t}|y - x|^2\right) dy$$

is a C_0 function of x by the dominated convergence theorem whenever g is a C_0 function. □

In light of this theorem, we have the following existence theorem for SDE:

Theorem 2.10. *Let $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be Lipschitz continuous functions satisfying $|b(x)| + \|\sigma(x)\| \leq C(1+|x|)$ for some $C > 0$ that does not depend on x. Then there exists a unique continuous Feller process X adapted to the Brownian filtration satisfying the SDE*

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t.$$

Proof. The existence and uniqueness proof uses the Picard iteration scheme used to prove existence and uniqueness in ODE. See, for example, theorem 5.2 from [31]. □

Definition 2.11. Let X be a continuous Feller process, and let $\mathcal{D}(A)$ be the set of those bounded continuous functions f such that

$$Af := \lim_{t\to 0} \frac{P_t f - f}{t}$$

exists, and so that the convergence is uniform on compact sets. The function A is called the *infinitesimal generator* of X.

Remark 2.12. The set $\mathcal{D}(A)$ is a uniformly dense subspace of the space of bounded continuous functions. For example, fix a bounded continuous function f, and for $\epsilon > 0$, let

$$f_\epsilon(x) = \frac{1}{\epsilon}\int_0^\epsilon P_y f(x)\,dy.$$

A simple calculation shows that $f_\epsilon \in \mathcal{D}(A)$ for each $\epsilon > 0$, and by the fundamental theorem of calculus $f_\epsilon \to f$ uniformly as $\epsilon \to 0$.

Theorem 2.13 (Kolmogorov Backwards Equation)**.** *For each $f \in \mathcal{D}(A)$, it holds Af is continuous, and*

$$\partial_t P_t f = AP_t f = P_t A f.$$

In particular, $t \mapsto P_t f(x)$ is a continuously differentiable function of $t \geq 0$.

Proof. Fix $f \in \mathcal{D}(A)$, and observe that the uniform convergence of $\frac{P_h f - f}{h}$ to Af as $h \to 0$ implies that Af is continuous since $(P_h f - f)/h$ is continuous for all $h > 0$. In addition,

$$\lim_{h\to 0^+} \frac{P_{t+h}f - P_t f}{h} = P_t \lim_{h\to 0^+} \frac{P_h f - f}{h} = P_t A f.$$

The same argument shows that the limit as $t - h \to 0$ exists and equals $P_t Af$. This proves $\partial_t P_t f = AP_t f$. The other equality follows from the fact that we can also write

$$\frac{P_{t+h}f - P_t f}{h} = \frac{P_h(P_t f) - (P_t f)}{h},$$

so taking $h \to 0$ implies also that $\partial_t P_t f = AP_t f$. □

Before exploring the useful properties of the infinitesimal generator, we should compute it for an Itô diffusion. The following lemma is of assistance.

Lemma 2.14 (Dynkin's Formula)**.** *Let X be a continuous Feller process adapted to a filtration $(\mathcal{F}_s)_{s\geq 0}$ and infinitesimal generator A. Fix $f \in \mathcal{D}(A)$. Then Af is the unique continuous function such that*

$$f(X_t) - \int_0^t Af(X_s)\,ds$$

is an $\mathbf{E}_x$-martingale adapted to $\mathcal{F}_s$ for every $x \in \mathbb{R}^d$.

Proof. First, we verify that Af is a function which makes

$$Y_t = f(X_t) - \int_0^t Af(X_y)\,dy$$

into a martingale. Fix $0 \leq s < t$. First, note that $\mathbf{E}_x[f(X_t) \,|\, \mathcal{F}_s] = P_t f(X_s)$. Next, by Fubini's theorem,

$$\begin{aligned}
\mathbf{E}_x\left[\int_s^t Af(X_y)\,dy \,|\, \mathcal{F}_s\right] &= \int_s^t \mathbf{E}_x[Af(X_y) \,|\, \mathcal{F}_s]\,dy \\
&= \int_s^t P_y Af(X_s)\,dy \\
&= \int_s^t \partial_y P_y f(X_s)\,dy \\
&= P_t f(X_s) - P_s f(X_s) = P_t f(X_s) - f(X_s),
\end{aligned}$$

where on the third and fourth lines we used the Kolmogorov backwards equation. Therefore,

$$\begin{aligned}
\mathbf{E}_x\left[Y_t \,|\, \mathcal{F}_s\right] &= \mathbf{E}_x[f(X_t) \,|\, X_s] - \int_0^s Af(X_y)\,dy - \mathbf{E}_x \int_s^t Af(X_y)\,dy \\
&= P_t f(X_s) - \int_0^s Af(X_y)\,dy - (P_t f(X_s) - f(X_s)) \\
&= Y_s.
\end{aligned}$$

On the other hand, suppose that Mf is a continuous function so that $f(X_t) - \int_0^t Mf(X_y)\,dy$ is an $\mathbf{E}_x$-martingale. Then for $t > 0$,

$$P_t f(x) - \int_0^t P_y Mf(x)\,dy = \mathbf{E}_x\left[f(X_t) - \int_0^t Mf(X_y)\,dy\right] = f(x),$$

hence

$$\frac{1}{t}\int_0^t P_y Mf(x)\,dy = \frac{P_t f(x) - f(x)}{t}.$$

Since Mf is continuous and X is Feller, it holds $y \mapsto P_y Mf(x)$ is continuous, so taking $t \to 0$ implies that

$$Mf(x) = Af(x)$$

for every $x \in \mathbb{R}^d$ by the fundamental theorem of calculus. □

Theorem 2.15. *Let $dX_s = b(X_s)\,ds + \sigma(X_s)\cdot dB_s$ be an $\mathbb{R}^d$-valued Itô diffusion for deterministic functions b and σ satisfying the conditions of theorem 2.10. Let A be the infinitesimal generator of X. If $f \in C_c^2(\mathbb{R}^d)$, then $f \in \mathcal{D}(A)$, and*

$$Af(x) = \nabla f(x)\cdot b(x) + \frac{1}{2}\mathrm{Tr}(Hf(x)\sigma(x)\sigma(x)^\top),$$

where Hf is the Hessian of f.

Proof. This follows immediately from Itô's lemma (theorem 2.2), lemma 2.14, and theorem 2.10. □

Remark 2.16. We remark that theorem 2.15 combined with the Kolmogorov backwards equation implies that a solution to the second order linear partial differential equation

$$\partial_t f(t,x) = \nabla_x f(t,x)\cdot b(x) + \frac{1}{2}\mathrm{Tr}(H_x f(t,x)\sigma(x)\sigma(x)^\top)$$

subject to the boundary condition $f(0,x) = g(x)$ is given by $f(t,x) = P_t g(x)$, where P is the semi-group associated to the SDE $dX_s = b(X_s)\,ds + \sigma(X_s)\cdot dB_s$. This is a special case of the Feynman-Kac formula.

Of special interest in the study of Markov processes is the existence of stationary distributions. Namely, whether there exists a probability distribution π such that $X_0 \sim \pi$ implies that $X_t \sim \pi$ for each $t > 0$. In other words,

$$\mathbf{P}_\pi[X_t \in A] = \mathbf{P}_\pi[X_0 \in A] = \pi(A).$$

The infinitesimal generator ends up being of great assistance.

Theorem 2.17. *Let X be a continuous Feller process. Then a probability measure π is stationary for X if and only if*

$$\int Af\,d\pi = 0$$

for all $f \in \mathcal{D}(A)$.

Proof. Suppose that π is stationary. For any $f \in \mathcal{D}(A)$, since $|\partial_t P_t f| = |P_t(Af)| \leq \|Af\|_\infty \in L^1(\mathbb{R}^d)$ by the Kolmogorov backwards equation and the fact $Af \in C_0(\mathbb{R}^d)$, we can pass ∂_t through the integral (see [15], Theorem 2.27) to obtain

$$0 = \partial_t \mathbf{E}_\pi[f(X_t)] = \int \partial_t P_t f\,d\pi = \int P_t Af\,d\pi.$$

Taking $t \to 0$,

$$\int Af\,d\pi = 0.$$

For the converse, if $g \in \mathcal{D}(A)$, then $P_t g \in \mathcal{D}(A)$, and the same argument as before justifies passing ∂_t through the integral to get

$$0 = \int A(P_t g)\,d\pi = \int \partial_t(P_t g)\,d\pi = \partial_t \int P_t g\,d\pi.$$

This shows that $t \mapsto \int P_t g \, d\pi$ is identically the constant $\int g \, d\pi$ for each $g \in \mathcal{D}(A)$. Since $\mathcal{D}(A)$ is uniformly dense in $C_0(\mathbb{R}^d)$ (see remark 2.12), it follows that $\int P_t f \, d\pi = \int f \, d\pi$ for any $f \in C_0(\mathbb{R}^d)$. So

$$\int f \, d\pi = \int P_t f \, d\pi = \iint f(y) p(t, x, \, dy) \, d\pi(x)$$

for all $t \geq 0$, and since f was arbitrary,

$$\pi(E) = \int p(t, x, E) \, d\pi(x) = \mathbf{P}_\pi[X_t \in E]$$

for all Borel sets E and $t \geq 0$. This proves π is stationary. □

Theorem 2.17 provides a convenient way to calculate stationary measures. Indeed, if we assume that a candidate stationary measure π possesses a density, i.e., $d\pi = h \, dx$ for a suitably nice function h, then this reduces the problem to solving the equation $A^* h = 0$, where A^* is the adjoint operator satisfying

$$\int h \, Af \, dx = \int f \, A^* h \, dx$$

for all appropriately defined f and h. If there is a probability density h solving $A^* h = 0$, then Theorem 2.17 implies that $d\pi = h \, dx$ is stationary. It is rather cumbersome to write out the formal definition of the Hermitian adjoint for a general Markov process. We will only require the adjoint for an Itô diffusion, and refer the reader to ([25], Chapter 9) and [19] for the general case.

Definition 2.18. Let $dX_t = b(X_t) \, dt + \sigma(X_t) \cdot dB_t$ be an Itô diffusion. Then the Hermitian adjoint operator A^* is given by the formula

$$A^* f = -\sum_{j=1}^{d} \partial_{x_i}[bf] + \frac{1}{2} \sum_{i,j=1}^{d} \partial^2_{x_i, x_j}[f \Sigma_{ij}]$$

for all suitably chosen f, where Σ_{ij} is the function in the i-th row and j-th column of $\sigma \sigma^\top$.

Checking that the formula for A^* in definition 2.18 is correct just follows from a routine application of integration by parts. In total, these observations yield the following fact, which will be of utility in section 3:

Corollary 2.19. *Let X solve the SDE $dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t$ and A^* be the adjoint of the infinitesimal generator as in definition 2.18. If there is a probability density function h solving $A^* h = 0$, then the measure π with density h is a stationary distribution for X.*

Remark 2.20. The preceding discussion can be used to quickly prove that a standard Brownian motion has no stationary distribution. In this case $A = A^* = \triangle$, and so a stationary distribution π satisfies $\int \triangle f \, d\pi = 0$ for any test function $f \in C_c^\infty(\mathbb{R}^d)$. However, a famous theorem of Weyl [27] implies that π possesses a density u satisfying $\triangle u = 0$. But $u \geq 0$ since it is a probability density, so u is constant, a contradiction.

The Wasserstein Spaces. Given two Borel probability measures μ and ν on $\mathbb{R}^d$, we require a way of measuring the distance between μ and ν. One common family of metrics are the Wasserstein distances W_p for $1 \leq p < \infty$. Recall that a coupling γ of μ and ν is a Borel probability on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\mu(E) = \gamma(E \times \mathbb{R}^d)$ and $\nu(E) = \gamma(\mathbb{R}^d \times E)$ for all Borel $E \subset \mathbb{R}^d$. The set $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν. By an abuse of notation, $\Pi(\mu, \nu)$ is also the set of $\mathbb{R}^d \times \mathbb{R}^d$ random variables (X, Y) such that $X \sim \mu$ and $Y \sim \nu$.

Definition 2.21. Let $\mathcal{P}^p(\mathbb{R}^d)$ be the set of Borel probability measures μ on $\mathbb{R}^d$ with finite p-th moment,

$$\int |x|^p \, d\mu(x) < \infty.$$

For $\mu, \nu \in \mathcal{P}^p$, the p-Wasserstein distance W_p is defined by the formula

$$W_p(\mu, \nu)^p = \inf_{\gamma \in \Pi(\mu,\nu)} \int |x - y|^p \, d\gamma(x, y).$$

Equivalently,

$$W_p(\mu, \nu)^p = \inf_{(X,Y) \in \Pi(\mu,\nu)} \mathbf{E}[|X - Y|^p].$$

By an abuse of notation, $\mathcal{P}^p(\mathbb{R}^d)$ refers to the set of random variables X with $\mathbf{E}[|X|^p] < \infty$. That W_p is well-defined (i.e., $\Pi(\mu, \nu) \neq \emptyset$ and the infimum is finite) follows from the fact that the product measure $\mu \times \nu$ is always an element of $\Pi(\mu, \nu)$, and

$$\int |x - y|^p \, d(\mu \times \nu)(x, y) \leq 2^{p-1} \int |x|^p \, d\mu(x) + 2^{p-1} \int |y|^p \, d\nu(y) < \infty.$$

The last inequality follows from the standard inequality $|x - y|^p \leq 2^{p-1}(|x|^p + |y|^p)$.

We will now proceed to prove a couple of standard facts about the p-Wasserstein distance. The statements of most of these theorems can be found in [8], but we modify some of the arguments or provide them when missing to suit the purposes of this work.

For any locally compact Hausdorff space Ω, recall that a family of Borel probability measures $\mathcal{E}$ on Ω is said to be *tight* if for any $\epsilon > 0$ there exists a compact set $K \subset \Omega$ such that

$$\mu(\Omega \setminus K) < \epsilon$$

for all $\mu \in \mathcal{E}$. Also recall that a sequence of Borel probability measures $(\mu_n)_{n \geq 1}$ is said to converge *vaguely* (or *in distribution*) to a Borel probability measure μ if

$$\lim_{n \to \infty} \int f \, d\mu_n = \int f \, d\mu$$

for each continuous compactly supported function $f : \Omega \to \mathbb{C}$. The topology on the set of Borel probability measures $\mathcal{P}(\Omega)$ generated by this mode of convergence is called the *vague topology*. This is equivalent to the weak-$*$ topology on $C_0(\Omega)$, but saying a sequence converges "vaguely" is more elegant than saying it converges "weakly-$*$". It is well known that vague convergence is metrizable when Ω is separable and metrizable.

There are several metrics which are equivalent to vague convergence, and perhaps the most famous is the Prokhorov distance [4]:

$$\rho(\mu,\nu) := \inf\{\epsilon > 0 : \nu(A^\epsilon) \leq \mu(A) + \epsilon,\ \mu(A^\epsilon) \leq \nu(A) + \epsilon \text{ for all Borel } A\},$$

where $A^\epsilon = \{x : d(x, A) < \epsilon\}$ and $d(x, A)$ is the distance from x to A.

The vague topology and tightness are related by Prokhorov's theorem, which we present in a slightly altered form that suffices for our intentions:

Theorem 2.22 (Prokhorov's Theorem)**.** *Let Ω be a locally compact Hausdorff space. If a family of Borel probability measures $\mathcal{E} \subset \mathcal{P}(\Omega)$ is tight, then there is a sequence in $\mathcal{E}$ which converges vaguely to a regular Borel probability measure μ.*

Proof. Suppose that $\mathcal{E}$ is a tight family of Borel probability measures. To each $\mu \in \mathcal{P}(\Omega)$ we associate a bounded (with respect to the uniform norm on $C_0(\Omega)$) linear operator $T_\mu : C_0(\Omega) \to \mathbb{C}$ by

$$T_\mu f = \int_\Omega f \, d\mu.$$

It follows that $\{T_\mu : \mu \in \mathcal{E}\}$ is contained in the unit ball of the dual space $C_0(\Omega)^*$, and by the Banach-Alaoglu theorem, there is a sequence $\{T_{\mu_k}\}$ and a bounded positive linear operator T such that $T_{\mu_k} f \to Tf$ as $k \to \infty$ for each $f \in C_0(\Omega)$. By the Riesz-Markov theorem (see [24], Theorem 6.2), there is a regular Borel measure μ with $0 \leq \mu \leq 1$ such that $T = T_\mu$. We need to show that μ is a probability measure, namely that $\mu(\Omega) = 1$. By invoking the tightness of $\mathcal{E}$, for an $\epsilon > 0$ pick a compact set $K \subset \Omega$ with $\mu_k(K) \geq 1 - \epsilon$ for each $k \geq 1$. By Urysohn's lemma ([15], Theorem 4.32), we can find a continuous compactly supported $\phi : \Omega \to [0, 1]$ with $\phi|_K = 1$. It follows that

$$\mu(\Omega) \geq \int \phi \, d\mu = \lim_{n\to\infty} \int \phi \, d\mu_n \geq \limsup_{n\to\infty} \mu_n(K) \geq 1 - \epsilon.$$

Since $\epsilon > 0$ was arbitrary, we deduce $\mu(\Omega) = 1$. The condition $T_{\mu_k} f \to T_\mu f$ for each $f \in C_0(\Omega)$ is equivalent to $\mu_k \to \mu$ vaguely, which completes the proof. □

Remark 2.23. A proof of Prokhorov's theorem in a more general setting, including a converse when Ω is a separable metric space, can be found in Billingsley ([4], Theorems 1.5.1 and 1.5.2). While the proof of theorem 2.22 shows that any sequence of Borel probability measures has a vaguely convergent subsequence (this is sometimes called Helly's selection theorem when $\Omega = \mathbb{R}$), it need not hold that the subsequential limit is a probability measure. For example, consider

$$\mu_n(E) = \frac{1}{2n} m(E \cap [-n, n]),$$

where m is the Lebesgue measure. For each $f \in C_0(\mathbb{R})$, one has $\int f \, d\mu_n = \frac{1}{2n} \int_{-n}^{n} f(x) \, dx$, so $\mu_n \to 0$ vaguely. But 0 is not a probability measure.

When Ω is separable and metrizable, the Prokhorov theorem is equivalent to saying that every tight family of Borel probability measures is vaguely pre-compact. An important and frequently used property of W_p can be proven using Prokhorov's theorem:

Theorem 2.24. *For any μ, $\nu \in \mathcal{P}^p$, there is $\gamma \in \Pi(\mu,\nu)$ such that*

$$W_p(\mu,\nu)^p = \int |x-y|^p \, d\gamma(x,y) \tag{2.5}$$

Equivalently, there is a coupling $(X,Y) \in \Pi(\mu,\nu)$ such that $W_p(\mu,\nu)^p = \mathbf{E}[|X-Y|^p]$.

A coupling satisfying equation 2.5 is called an *optimal coupling* of μ and ν.

Lemma 2.25. *The set $\Pi(\mu,\nu)$ is vaguely compact for any $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$.*

Proof. First, we show that $\Pi(\mu,\nu)$ is vaguely closed. Suppose that $(\gamma_n)_{n\geq 1}$ is a sequence of couplings converging vaguely to γ. Let f and g be continuous compactly supported functions on Ω. Then for any continuous compactly supported functions $f : \mathbb{R}^d \to [0,1]$ and $g : \mathbb{R}^d \to [0,1]$,

$$\int f(x)g(y)\, d\gamma(x,y) = \lim_{n\to\infty} \int f(x)g(y)\, d\gamma_n(x,y) = \int f(x)\, d\mu(x) \int g(y)\, d\nu(y).$$

Letting f tend pointwise to the identically 1 function and g tend pointwise to the characteristic function of a compact set K and applying dominated convergence yields

$$\gamma(\mathbb{R}^d \times K) = \nu(K),$$

and using the regularity of Borel probability measures on $\mathbb{R}^d$ implies that $\gamma(\mathbb{R}^d \times E) = \nu(E)$ for all Borel E. Switching the roles of f and g in the above analysis shows $\gamma(E \times \mathbb{R}^d) = \mu(E)$ as well, so that γ is a coupling of (μ,ν). Therefore $\Pi(\mu,\nu)$ is vaguely closed.

Next, we show $\Pi(\mu,\nu)$ is tight. Since μ,ν are probability measures and $\mathbb{R}^d$ is σ-compact, for any $\epsilon > 0$, there exist compact sets $K_1 \subset \mathbb{R}^d$ and $K_2 \subset \mathbb{R}^d$ such that $\mu(\mathbb{R}^d \setminus K_1) < \epsilon/2$ and $\nu(\mathbb{R}^d \setminus K_2) < \epsilon/2$. Then $\mathbb{R}^d \times \mathbb{R}^d \setminus K_1 \times K_2 \subset ((\mathbb{R}^d \setminus K_1) \times \mathbb{R}^d) \cup (\mathbb{R}^d \times (\mathbb{R}^d \setminus K_2))$, so that for any $\gamma \in \Pi(\mu,\nu)$,

$$\begin{aligned}
\gamma(\mathbb{R}^d \times \mathbb{R}^d \setminus K_1 \times K_2) &\leq \gamma((\mathbb{R}^d \setminus K_1) \times \mathbb{R}^d) + \gamma(\mathbb{R}^d \times (\mathbb{R}^d \setminus K_2)) \\
&= \mu(\mathbb{R}^d \setminus K_1) + \nu(\mathbb{R}^d \setminus K_2) \\
&< \epsilon.
\end{aligned}$$

This proves that $\Pi(\mu,\nu)$ is vaguely closed, tight, and therefore vaguely compact by Prokhorov's theorem (2.22). □

We return to proving theorem 2.24.

Proof of Theorem 2.24. In light of lemma 2.25, we need to show that the function

$$\gamma \mapsto \int |x-y|^p \, d\gamma(x,y)$$

is vaguely lower semi-continuous on $\Pi(\mu,\nu)$. If this is shown, then the vague compactness of $\Pi(\mu,\nu)$ guarantees $\gamma \in \Pi(\mu,\nu)$ satisfying 2.5. To this end, suppose $(\gamma_n)_{n\geq 1} \subset$

$\Pi(\mu,\nu)$ vaguely converges to $\gamma \in \Pi(\mu,\nu)$. Fix a continuous compactly supported $\phi : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$. Then

$$\liminf_{n\to\infty} \int |x-y|^p \, d\gamma_n(x,y) \geq \lim_{n\to\infty} \int \phi(x,y)|x-y|^p \, d\gamma_n(x,y) = \int \phi(x,y)|x-y|^p \, d\gamma(x,y).$$

Letting ϕ increase to 1 through a sequence of non-negative compactly supported continuous functions implies that

$$\liminf_{n\to\infty} \int |x-y|^p \, d\gamma_n(x,y) \geq \int |x-y|^p \, d\gamma(x,y)$$

by the monotone convergence theorem. Since $(\gamma_n)_{n\geq 1}$ was an arbitrary vaguely converging sequence, this proves that $\gamma \mapsto \int |x-y|^p \, d\gamma(x,y)$ is vaguely lower semi-continuous. □

Another useful fact about W_p is that it is a *bona fide* metric.

Theorem 2.26. *W_p is a metric on $\mathcal{P}^p(\mathbb{R}^d)$.*

To prove the triangle inequality, we require a slightly complicated technical lemma:

Lemma 2.27. *Suppose that $\gamma_{1,2} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ and $\gamma_{2,3} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ are Borel probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\gamma_{1,2}(\mathbb{R}^d \times E) = \gamma_{2,3}(E \times \mathbb{R}^d)$ for all Borel sets E. Then there is a Borel probability measure $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ where $\gamma(A \times \mathbb{R}^d) = \gamma_{1,2}(A)$ and $\gamma(\mathbb{R}^d \times A) = \gamma_{2,3}(A)$ for all Borel $A \subset \mathbb{R}^d \times \mathbb{R}^d$.*

Proof. Let $\pi_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ be the projection $\pi_1(x,y) = x$ and $\pi_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ be the projection $\pi_2(x,y) = y$. Let $\mathcal{F}_1 = \sigma(\pi_1)$ and $\mathcal{F}_2 = \sigma(\pi_2)$ be the sub-Borel σ-algebras generated by π_1 and π_2, respectively. Observe that for a Borel measurable $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, it holds $\mathbf{E}_{\gamma_{1,2}}[f \,|\, \mathcal{F}_2](x,y)$ is a function only of y, and $\mathbf{E}_{\gamma_{2,3}}[f \,|\, \mathcal{F}_1](x,y)$ is a function solely of x. Therefore, for each $x \in \mathbb{R}^d$, we can define measures on $\mathbb{R}^d \times \mathbb{R}^d$ by

$$\rho_1(x,A) := \mathbf{E}_{\gamma_{1,2}}[\mathbf{1}_A \,|\, \mathcal{F}_2](0,x), \quad \rho_2(x,A) := \mathbf{E}_{\gamma_{2,3}}[\mathbf{1}_A \,|\, \mathcal{F}_1](x,0).$$

Let μ be the common middle marginal of $\gamma_{1,2}$ and $\gamma_{2,3}$. By the Riesz-Markov theorem, there exists a Borel probability measure γ on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ such that

$$\int f \, d\gamma = \iiint f(x,y,z)\rho_1(w, d(x,y))\rho_2(r, d(w,z)) \, d\mu(r)$$

for each $f \in C_0(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$. To verify that γ is the desired measure, fix a bounded measurable function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and let

$$g(w) = \int f(x,y)\rho_1(w, \, d(x,y)).$$

Then

$$\int f(x,y) \, d\gamma(x,y,z) = \iint g(w)\rho_2(r, \, d(z,w)) \, d\mu(r).$$

The function $(w,z) \mapsto g(w)$ is $\mathcal{F}_1$-measurable, and so

$$\int g(w)\rho_2(r,\, d(w,z)) = g(r)$$

Thus

$$\begin{aligned}\int f(x,y)\, d\gamma(x,y,z) &= \int g(r)\, d\mu(r) \\ &= \iint f(x,y)\, \rho_1(w, d(x,y))\, d\mu(y) \\ &= \int \mathbf{E}_{\gamma_{1,2}}[f \mid \mathcal{F}_2]\, d\gamma_{1,2} \\ &= \int f\, d\gamma_{1,2}.\end{aligned}$$

To see the third equality, note that the conditional expectation is a $\mathcal{F}_2$-measurable function and $\int \cdot\, d\mu = \int \cdot\, d\gamma_{1,2}$ for all $\mathcal{F}_2$-measurable integrands. This proves $\gamma(A\times\mathbb{R}^d) = \gamma_{1,2}(A)$. Similarly, observe that

$$f(w,z) = \int f(y,z)\rho_1(w,\, d(x,y))$$

since the map $(x,y) \mapsto f(y,z)$ is $\mathcal{F}_2$-measurable. Therefore,

$$\begin{aligned}\int f(y,z)\, d\gamma(x,y,z) &= \iint f(w,z)\rho_2(r, d(w,z))\, d\mu(r) \\ &= \int \mathbf{E}_{\gamma_{2,3}}[f \mid \mathcal{F}_1]\, d\gamma_{2,3} \\ &= \int f\, d\gamma_{2,3}.\end{aligned}$$

This proves $\gamma(\mathbb{R}^d \times A) = \gamma_{2,3}(A)$, as required. □

Proof of Theorem 2.26. That $W_p(\mu,\nu) = W_p(\nu,\mu)$ is obvious, and $\mu = \nu$ clearly implies $W_p(\mu,\nu) = 0$. If $W_p(\mu,\nu) = 0$, then by theorem 2.24 there is an optimal coupling $(X,Y) \in \Pi(\mu,\nu)$ (where X and Y are random variables) satisfying $0 = W_p(\mu,\nu)^p = \mathbf{E}[|X-Y|^p]$. It follows $X = Y$ a.s., hence X and Y are identically distributed. That is, $\mu = \nu$ since $X \sim \mu$ and $Y \sim \nu$.

For the triangle inequality, let $\mu,\nu,\sigma \in \mathcal{P}^p(\mathbb{R}^d)$, and let $\gamma_{1,2} \in \Pi(\mu,\sigma)$ and $\gamma_{2,3} \in \Pi(\sigma,\nu)$ be optimal couplings. Since $\gamma_{1,2}(\mathbb{R}^d \times E) = \gamma_{2,3}(E \times \mathbb{R}^d) = \sigma(E)$, choose $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ as in lemma 2.27. Let X, Y, Z be $\mathbb{R}^d$-valued random variables with $(X,Y,Z) \sim \gamma$. Then Minkowski's inequality implies that

$$\mathbf{E}[|X-Z|^p]^{1/p} \le \mathbf{E}[|X-Y|^p]^{1/p} + \mathbf{E}[|Y-Z|^p]^{1/p} = W_p(\mu,\sigma) + W_p(\sigma,\nu)$$

since $(X,Y) \sim \gamma_{1,2}$ and $(Y,Z) \sim \gamma_{2,3}$. Since $X \sim \mu$ and $Z \sim \nu$, we have $(X,Z) \in \Pi(\mu,\nu)$ so that

$$W_p(\mu,\nu) \le \mathbf{E}[|X-Z|^p]^{1/p} \le W_p(\mu,\sigma) + W_p(\sigma,\nu),$$

which completes the proof. □

Theorem 2.28. *The p-Wasserstein distance gives $\mathcal{P}^p(\mathbb{R}^d)$ the topological structure of a complete separable metric space such that $\mu_n \to \mu$ as $n \to \infty$ if and only if $\mu_n \to \mu$ vaguely and*

$$\lim_{n\to\infty} \int |x|^p \, d\mu_n = \int |x|^p \, d\mu.$$

Proof. Suppose that $\mu_n \to \mu$ in the W_p distance. Let γ_n be a sequence of optimal couplings. Then

$$\left| \left(\int |x|^p \, d\mu_n \right)^{1/p} - \left(\int |x|^p \, d\mu \right)^{1/p} \right| \leq \left(\int |x-y|^p \, d\gamma_n(x,y) \right)^{1/p} = W_p(\mu_n, \nu)$$

which proves that $\int |x|^p \, d\mu_n \to \int |x|^p \, d\mu$. Next, suppose that f is a compactly supported Lipschitz function, with Lipschitz constant M. Then

$$\int |f(x) - f(y)|^p \, d\gamma_n(x,y) \leq M^p \int |x-y|^p \, d\gamma_n(x,y) = M^p W_p(\mu_n, \nu)^p.$$

Therefore

$$|\|f\|_{L^p(\mu_n)} - \|f\|_{L^p(\mu)}| \leq M W_p(\mu, \nu),$$

hence $\int |f|^p \, d\mu_n \to \int |f|^p \, d\mu$. Now, if f is an arbitrary non-negative function in $C_c(\mathbb{R}^d)$ and $\epsilon > 0$, then the uniform density of Lipschitz functions in $C_c(\mathbb{R}^d)$ furnishes a Lipschitz function g with $\|f^p - |g|^p\| < \epsilon$. It follows

$$\begin{aligned}
&\left| \int f^p \, d\mu_n - \int f^p \, d\mu \right| \\
&\leq \int ||f|^p - |g|^p| \, d\mu_n + \int ||f|^p - |g|^p| \, d\mu + \left| \int |g|^p \, d\mu_n - \int |g|^p \, d\mu \right| \\
&\leq 2\epsilon + \left| \int |g|^p \, d\mu_n - \int |g|^p \, d\mu \right|
\end{aligned}$$

This yields

$$\limsup_{n\to\infty} \left| \int |f|^p \, d\mu_n - \int |f|^p \, d\mu \right| \leq 2\epsilon,$$

and since $\epsilon > 0$ was arbitrary we recover

$$\lim_{n\to\infty} \left| \int |f|^p \, d\mu_n - \int |f|^p \, d\mu \right| = 0.$$

By replacing f with $f^{1/p}$ we conclude that $\int f \, d\mu_n \to \int f \, d\mu$ when f is non-negative. Finally, if f is an arbitrary function in $C_c(\mathbb{R}^d)$ (not strictly non-negative), then we can write $f = f^+ - f^-$ for non-negative functions f^+ and f^- in $C_c(\mathbb{R}^d)$. Since we just proved that $\int f^\pm \, d\mu_n \to \int f^\pm \, d\mu$, the linearity of the integral shows $\int f \, d\mu_n \to \int f \, d\mu$ for any $f \in C_c(\mathbb{R}^d)$, as required.

For the converse, since $\mu_n \to \mu$ vaguely we can find a sequence of random variables $\{X_n\}$ such that $X_n \to X$ a.s. and $X_n \sim \mu_n$ and $X \sim \mu$. Combining this with the fact $\|X_n\|_p \to \|X\|_p$, a routine application of dominated convergence implies that $\|X_n - X\|_p \to 0$. But,

$$W_p(\mu_n, \mu) \leq \|X_n - X\|_p,$$

from which it follows $\mu_n \to \mu$ in the W_p sense.

Completeness and separability follows from the fact that the vague topology is also complete and separable. □

At this point we specialize to the case W_2. The L^2 norm provides additional geometric information which endows $\mathcal{P}^2$ with an especially nice structure. We let $\mathcal{P}^2_{ac}$ be the subset of (Lebesgue) absolutely continuous measures in $\mathcal{P}^2$. For a measurable function f and a measure μ, we write $f_\sharp\mu$ to denote the pushforward measure $E \mapsto \mu(\{x : f(x) \in E\})$. Moreover, id denotes the identity map.

Theorem 2.29 (Brenier-McCann Theorem)**.** *Suppose that $\mu, \nu \in \mathcal{P}^2(\mathbb{R}^d)$ and that $\mu \in \mathcal{P}^2_{ac}(\mathbb{R}^d)$. Then there exists a convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that $(\nabla\phi)_\sharp\mu = \nu$ and the measure $\gamma := (\mathrm{id} \times \nabla\phi)_\sharp\mu$ is the unique optimal coupling of μ and ν.*

In the language of probability, if X is an absolutely continuous L^2 random variable and Y is another L^2 random variable, then there is a unique convex function ϕ such that $\nabla\phi(X)$ and Y are identically distributed and $(X, \nabla\phi(X))$ is the optimal coupling the respective distributions.

We remark that the convex function guaranteed in Brenier's theorem need not be differentiable everywhere. This is the primary reason that we require $\mu \in \mathcal{P}^2_{ac}$. Convex functions are locally Lipschitz ([17], Lemma 1.1.6) hence absolutely continuous. In particular, they are differentiable Lebesgue almost everywhere. The requirement $\mu \in \mathcal{P}^2_{ac}$ therefore ensures that ϕ is differentiable μ -a.e., hence the pushforward measure $(\nabla\phi)_\sharp\mu$ is well-defined.

Proof. For a detailed proof, see ([20], Chapter 6) or [16]. □

One of the key consequences of Brenier's theorem is the geometry on $\mathcal{P}^2_{ac}$ it implies. Before we explore this construction in detail, let us motivate why a geometry is desirable property.

Suppose that we are given a continuous-time Markov process X with a stationary distribution π_0. A property that many Markov processes enjoy, including the ones investigated in this work, is a form of ergodicity. Namely, if $\pi_0 \in \mathcal{P}_{2,ac}$ is a probability distribution and π_t is the distribution of X_t when $X_0 \sim \pi_0$, then in many cases we can show that

$$W_2(\pi_t, \pi) \to 0$$

as $t \to \infty$. But proving that this is the case, in particular establishing the rate of convergence (i.e., the mixing time of the chain), is a delicate matter. A nice geometry on $\mathcal{P}_{2,ac}$ would be beneficial, as it would introduce useful tools from calculus, such as gradients.

With this motivation in mind, we begin the process of formally constructing the Wasserstein geometry. Most of this work was pioneered in the early 2000s, with notable contributors including L. Ambrosio, R. McCann, C. Villani, and A. Figalli, and Y. Brenier. We defer to the works of Villani [29] and Ambrosio [1] for a comprehensive treatment of Wasserstein geometry. Many of the formal proofs required in this construction rely on concepts from Riemannian geometry which would take us too far afield for the purposes of this work. Therefore, we follow the more informal exposition in [8], and refer the reader to the texts [20], [1], and [29] for needed rigor.

Definition 2.30. A curve μ in $\mathcal{P}^2$ is a function $\mu_t : [0,\infty) \to \mathcal{P}^2(\mathbb{R}^d)$ such that for every test function $\phi \in C_c^\infty(\mathbb{R}^d)$, the map

$$t \in [0,1] \mapsto \int \phi(x)\, d\mu_t(x)$$

is Borel measurable. Given a vector field $u : \mathbb{R}^d \times [0,\infty) \to \mathbb{R}^d$ we say that μ is transported by u, or that u is an Eulerian velocity of μ, and symbolically write

$$\partial_t \mu_t + \operatorname{div}(u\mu_t) = 0,$$

if the following two conditions hold:

(1) *Finiteness of the action*: $u(\cdot,t) \in L^2(\mu_t)$ for almost every $t \in (0,1)$, and has finite action:
$$\int_0^\infty \int_{\mathbb{R}^n} |u(x,t)|^2\, d\mu_t(x)\, dt < \infty.$$
(2) The distributional *continuity equation* is satisfied: If $\psi \in C_c^\infty(\mathbb{R}^d \times (0,\infty))$, it holds
$$\int_0^\infty \int_{\mathbb{R}^n} (\partial_t \psi(x,t) + \nabla\psi(x,t)\cdot u(x,t))\, d\mu_t(x)\, dt = 0.$$

Remark 2.31. Definition 2.30 might seem slightly obscure without proper motivation. To motivate it, suppose that $\mu_0 \in \mathcal{P}^2(\mathbb{R}^d)$, and suppose that $X_0 \sim \mu_0$. Suppose further that there is a solution X_t to the flow $\partial_t X_t = u(X_t,t)$. Then the distribution μ_t of X_t satisfies the distributional continuity equation. Indeed, if $\psi \in C_c^\infty(\mathbb{R}^d \times (0,\infty))$, then

$$\partial_t[\psi(X_t,t)] = u(X_t,t)\cdot\nabla\psi(X_t,t) + (\partial_t\psi)(X_t,t),$$

from which it follows

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^n} (\partial_t\psi(x,t) + \nabla\psi(x,t)\cdot u(x,t))\, d\mu_t(x)\, dt \\
&= \mathbf{E}\left[\int_0^\infty (\partial_t\psi)(X_t,t) + \nabla\psi(X_t,t)\cdot u(X_t,t)\, dt\right] \\
&= \mathbf{E}\left[\int_0^\infty \partial_t[\psi(X_t,t)]\, dt\right] \\
&= 0
\end{aligned}$$

by the fundamental theorem of calculus and since the compact support of ψ implies $\psi(\cdot,0) = \psi(\cdot,\infty) = 0$.

Another useful concept is that of the metric derivative. Namely, given a metric space (X, d) and a continuous map $\gamma : [0, 1] \to X$, one can define the *metric derivative* by

$$\dot{\gamma}(t) := \lim_{s \to t} \frac{d(\gamma(s), \gamma(t))}{|s - t|},$$

wherever the limit exists. If there is a non-negative function g such that

$$d(\gamma(s), \gamma(t)) \leq \int_s^t g(x)\, dx < \infty$$

for all $0 \leq s < t \leq 1$, then we say that γ is *absolutely continuous*.

Lemma 2.32 ([1], Theorem 9.2)**.** *If γ is absolutely continuous, then $\dot{\gamma}$ exists for almost every $t \in [0, 1]$.*

Proof. Since the image of γ is compact, we might as well assume that X is compact. Pick a countable dense subset $\{p_i\}_{i \geq 1}$ of X, and define functions $\{f_i\}_{i \geq 1}$ by

$$f_i(t) = d(\gamma(t), p_i).$$

Since

$$|f_i(t) - f_i(s)| \leq d(\gamma(s), \gamma(t)) \leq \int_s^t g(x)\, dx \tag{2.6}$$

each f_i is absolutely continuous and $|f_i'| \leq g$ for all $i \geq 1$. Let

$$m(t) = \sup_{i \geq 1} |f_i'(t)|,$$

where m is defined for almost every t by absolute continuity. Then by equation 2.6,

$$\liminf_{s \to t} \frac{d(\gamma(s), \gamma(t))}{|s - t|} \geq m(t).$$

Since the p_i are dense, we have

$$d(\gamma(t), \gamma(s)) = \sup_{i \geq 1} |f_i(t) - f_i(s)| \leq \sup_{i \geq 1} \int_s^t |f_i'(x)|\, dx \leq \int_s^t m(x)\, dx,$$

from which it follows

$$\limsup_{s \to t} \frac{d(\gamma(t), \gamma(s))}{|s - t|} \leq \lim_{s \to t} \frac{1}{|s - t|} \int_s^t m(x)\, dx = m(t)$$

for almost every $t \in [0, 1]$ by the Lebesgue differentiation theorem. This proves

$$\dot{\gamma}(t) = m(t)$$

a.e., finishing the proof. □

Given an absolutely continuous curve $\gamma : [0, 1] \to X$, where X is a metric space, we can define its *length* by the formula

$$\ell(\gamma) := \int_0^1 |\dot{\gamma}(t)|\, dt.$$

As the proof of lemma 2.32 shows, we have

$$\ell(\gamma) \geq d(\gamma(0), \gamma(1)).$$

In the special case $\ell(\gamma) = d(\gamma(0), \gamma(1))$, we say that γ is a *geodesic*. If $\dot{\gamma}$ is a.e. a constant, then γ is said to be of *constant speed*. Denote by $\mathrm{Geo}(X, d)$ the set of constant speed geodesics on the metric space (X, d).

Theorem 2.33 ([1], Corollary 10.10)**.** *Let μ_0 and μ_1 be absolutely continuous Borel probability measures on $\mathbb{R}^d$. Then the function $\mu : [0,1] \to \mathcal{P}^2_{ac}(\mathbb{R}^d)$ given by*

$$\mu(t) = ((1-t)\mathrm{id} + t\nabla\phi)_\sharp \mu_0$$

is the unique element of $\mathrm{Geo}(\mathcal{P}^2_{ac}(\mathbb{R}^d), W_2)$ connecting μ_0 and μ_1, where ϕ is the convex function given by the Brenier-McCann theorem 2.29.

The function μ in Theorem 2.33 is called the *displacement interpolation*.

Proof. As in Theorem 2.29, the proof is long and requires the introduction of many concepts and technical lemmas which are not of much use elsewhere in this work. We refer the interested reader to chapters 9 and 10 of [1], chapter 15 of [20], or chapter 8 of [29]. □

We next take a page out of Riemannian geometry and define the tangent space of a measure $\mu \in \mathcal{P}^2_{ac}$ by

$$T_\mu \mathcal{P}^2_{ac}(\mathbb{R}^d) = \overline{\{\nabla\phi : \phi \in C_c^\infty(\mathbb{R}^d)\}},$$

where the closure is understood in the $L^2(\mu)$ sense. Note that $\nabla\phi$, by an abuse of notation, represents the *measure* associated to $\nabla\phi$ defined by $\nabla\phi\, dx$. We equip $T_\mu \mathcal{P}^2_{ac}(\mathbb{R}^d)$ with the $L^2(\mu)$ norm to endow it with Hilbert space structure. Taking a cue from Riemannian geometry, we can define an associated length on $\mathcal{P}^2_{ac}(\mathbb{R}^d)$ by

$$d(\mu,\nu) = \inf\left\{ \int \|v_t\|_{L^2(\mu_t)}\, dt : \mu \text{ is absolutely continuous, } \mu(0) = \mu,\ \mu(1) = \nu \right\}.$$

Here, v is the "tangent vector" associated with the curve μ. We will not specify precisely what this means, but we will remark that in the case μ has an Eulerian velocity u satisfying the continuity equation 2.30, it holds $v = u$. In particular,

$$d(\mu,\nu) = \inf\left\{ \int_0^1 \|u(t,\cdot)\|_{L^2(\mu_t)}\, dt : \mu_0 = \mu,\ \mu_1 = \nu,\ \partial_t \mu_t + \mathrm{div}(u\mu_t) = 0 \right\}.$$

There are two things to note. The first is that we are implicitly assuming that absolutely continuous curves admit Eulerian velocities. This is true in general, see ([20], Theorem 15.5). The second is that there are two quantities being optimized over in the above equation: flows of measures and Eulerian velocities on the flow of measures. This is because an Eulerian velocity need not be unique. For example, suppose that the flow μ satisfies $d\mu_t = \rho_t\, dx$. If v is a divergence-free vector field and u is an Eulerian velocity for μ, then $u + \epsilon(v/\rho)$ is also an Eulerian velocity for μ. The great insight of the Benemou-Brenier theorem is that the metric d is actually equivalent to W_2.

Theorem 2.34 (Benemou-Brenier)**.** *For any $\mu_0, \mu_1 \in \mathcal{P}^2_{ac}(\mathbb{R}^d)$, it holds*

$$W_2(\mu_0,\mu_1)^2 = \inf_\mu \mathbb{A}(\mu),$$

where the infimum is taken over all absolutely continuous flows of measures μ with $\mu(0) = \mu_0$ and $\mu(1) = \mu_1$, and

$$\mathbb{A}(\mu) = \inf \left\{ \int_0^1 \int_{\mathbb{R}^d} |u(x,t)|^2 \, d\mu_t(x) \, dt : \partial_t \mu_t + \operatorname{div}(u\mu_t) = 0 \right\}.$$

Proof. See [20], chapter 15, or [29], chapter 8. □

In other words, the Theorems 2.34 and 2.33 essentially tell us that the space $\mathcal{P}^2_{ac}(\mathbb{R}^d)$ can be endowed with a geometry where the geodesics are given by displacement interpolations. An important consequence of this geometry is the associated calculus it induces. This calculus was first proposed by Felix Otto in the context of the porous medium equation in PDE (see [22]). One example of the utility of having a geometry is that it supplies a gradient. In ([8], Chapter 1.4.1), the 2-Wasserstein gradient of a functional $\mathcal{F} : \mathcal{P}^2_{ac} \to \mathbb{R}$ is defined as the function $\nabla_{W_2}\mathcal{F}$ such that $\nabla_{W_2}\mathcal{F}(\mu) \in T_\mu \mathcal{P}^2_{ac}$ and $\partial_t|_{t=0}\mathcal{F}(\mu_t) = (\nabla_{W_2}\mathcal{F}(\mu), v_0)_{L^2(\mu)}$ for each absolutely continuous flow of measures μ with $\mu_0 = \mu$ and ν is an Eulerian velocity at time $t = 0$ for μ, i.e., $\partial_t \mu_t + \operatorname{div}(\nu_t \mu_t) = 0$. Recall that such Eulerian velocities can be thought of as "tangent" vectors in the Wasserstein space. On the other hand, in the case where the first variation of $\mathcal{F}$ exists, i.e., there is a function $\delta F : \mathcal{P}^2_{ac} \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = \int \delta\mathcal{F}(\mu)\partial_t\mu_t \, dx.$$

Using the continuity equation, we have the following symbolic calculation:

$$\partial_t|_{t=0}\mathcal{F}(\mu_t) = -\int \delta\mathcal{F}(\mu)\operatorname{div}(\mu_0\nu_0) \, dx = \int (\nabla\delta\mathcal{F}(\mu), \nu_0) \, d\mu.$$

While this calculation should not be taken too literally, note that $\nabla\delta\mathcal{F}(\mu) \in T_\mu\mathcal{P}^2_{ac}$ by the definition of the Wasserstein tangent space. So, this informal calculation should motivate the following definition:

Definition 2.35 ([8], Theorem 1.4.1)**.** Let $\mathcal{F} : \mathcal{P}^2_{ac} \to \mathbb{R}$ be a functional such that $\nabla\delta\mathcal{F}$ is well-defined. Then the Wasserstein gradient of $\mathcal{F}$ is defined by $\nabla_{W_2}\mathcal{F} := \nabla\delta\mathcal{F}$.

An important example is when $\mathcal{F}$ is the KL-divergence with respect to the probability density $\exp(-V)$, where $V : \mathbb{R}^d \to \mathbb{R}$ is a function so that $\pi := \exp(-V) \, dx \in \mathcal{P}^2_{ac}$. I.e.,

$$\mathcal{F}(\rho \, dx) = \operatorname{KL}(\rho \, dx \,||\, \pi) = \int \rho \log \rho/\pi \, dx = \int \rho \log \rho \, dx - \int V\rho \, dx.$$

This implies

$$\nabla\delta\mathcal{F}(\rho \, dx) = \nabla V + \nabla \log \rho = \nabla \log \frac{\rho}{\pi} \tag{2.7}$$

when ρ is suitably nice.

3. The Langevin Diffusion

This section will analyze the convergence of the Langevin stochastic differentiation equation

$$dX_t = -\nabla V(X_t)\,dt + \sqrt{2}\,dB_t.$$

The Langevin SDE can be roughly viewed as a noisy version of a gradient flow $\dot{x}_t = -\nabla V(x_t)$. As has been well studied in optimization theory, for all suitably nice V, the gradient flow converges exponentially fast to the unique minimizer of V, and a similar convergence phenomena occurs for Langevin SDE. This will be the principle motivation for the analysis in continuous time. Instead, we shall see that the *distribution* of X_t converges exponentially fast to the stationary distribution under the assumption V is strongly convex. Under some additional assumptions, such as a Poincaré inequality and a log-Sobolev Inequality (LSI), L^2 estimates and an estimate in the KL-divergence can also be established.

Of course, in the practical setting we cannot directly compute the solution to a continuous Langevin SDE. Just as one approximates the solution to the gradient flow $\dot{x}_t = -\nabla V(x_t)$ by the gradient descent $x_{k+1} = x_k - \gamma \nabla V(x_k)$ for a suitable $\gamma > 0$, we can approximate the solution to the Langevin SDE by an analogous discretization scheme, which requires a separate analysis to ensure the dynamics of the chain behave as in the continuous time case.

Analysis in Continuous Time. For the remainder of this section, $V : \mathbb{R}^d \to \mathbb{R}$ denotes a C^1 function on $\mathbb{R}^d$. The *Langevin SDE* (or *Langevin diffusion*) (with respect to the potential V) is a continuous time stochastic process $(X_t)_{t\geq 0}$ satisfying the SDE

$$dX_t = -\nabla V(X_t)\,dt + \sqrt{2}\,dB_t.$$

What is important is that a Langevin diffusion exists. In general, this can only be guaranteed through certain requirements on the potential function V.

Theorem 3.1. *Suppose that $V \in C^1$ is chosen so that ∇V is Lipschitz. Then there is a unique Langevin diffusion with potential V.*

Henceforth, we assume that V is chosen such that the diffusion exists. Using 2.15 the infinitesimal generator of the Langevin SDE is given by the formula

$$Af(x) = -\nabla f(x) \cdot \nabla V(x) + \triangle f(x),$$

which has Hermitian adjoint

$$A^* f(x) = \operatorname{div}(\nabla V(x) f(x)) + \triangle f(x).$$

Writing $\triangle = \operatorname{div}\nabla$, this is equivalent to

$$A^* f = \operatorname{div}(f\,\nabla(V + \log f))$$

if $f > 0$. Thus $A^* f = 0$ when $f = Ce^{-V}$ for $C > 0$. Therefore, if $e^{-V} \in L^1(\mathbb{R}^n)$, corollary 2.19 implies that a stationary distribution of the Langevin diffusion (in fact,

we will soon prove that it is *the* stationary distribution) is the so-called *Gibbs distribution* with density

$$\pi(x) = Ce^{-V(x)}$$

where C is a normalizing constant. As shorthand, we write $\pi \propto e^{-V}$.

Continuous Time Analysis. In this section, we will prove that the exact solution to the Langevin SDE exhibits many desirable convergence properties. Many of these desirable convergence properties can be efficiently proven using a simplified version of Gronwall's inequality

Lemma 3.2. *Let $u : [a, b] \to \mathbb{R}$ be a differentiable function satisfying*

$$u'(x) \leq Cu(x)$$

for some constant C. Then $u(x) \leq u(a)e^{C(x-a)}$ for all $x \in [a, b]$.

Proof. The constraints imply that the function $x \mapsto u(x)e^{-Cx}$ decreases, hence $u(x)e^{-Cx} \leq u(a)e^{-Ca}$. □

This yields a first estimate on the mixing time of a Langevin diffusion:

Theorem 3.3. *Suppose that $V \in C^2(\mathbb{R}^d)$ is α-strongly convex, i.e., $\nabla^2 V \succeq \alpha I_d$ for some $\alpha > 0$, where I_d is the identity matrix. Let μ_t be the law at time $t \geq 0$ of a Langevin diffusion with potential V, and let π be the Gibbs measure of V. Then for all $t \geq 0$,*

$$W_2(\mu_t, \pi) \leq \exp(-\alpha t) W_2(\mu_0, \pi).$$

Here, ∇^2 denotes the Hessian operator, and $A \preceq B$ means $B - A$ is a positive semi-definite matrix.

Proof. Suppose that (X_0, Y_0) is an optimal coupling of μ_0 and π. Let X_t be a Langevin diffusion started at X_0 and Y_t be a Langevin diffusion started from Y_0, and suppose that X and Y are driven by the same Brownian motion. That is,

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2}\, dB_t, \quad dY_t = -\nabla V(Y_t)\, dt + \sqrt{2}\, dB_t.$$

Observe that $X_t - Y_t$ is then a C^1 function of t, and in particular,

$$\partial_t |X_t - Y_t|^2 = -2\left(\nabla V(X_t) - \nabla V(Y_t)\right) \cdot (X_t - Y_t) \leq -2\alpha |X_t - Y_t|^2.$$

By lemma 4.3, it holds

$$|X_t - Y_t|^2 \leq \exp(-2\alpha t)|X_0 - Y_0|^2,$$

whence

$$W_2(\mu_t, \pi)^2 \leq \mathbf{E}[|X_t - Y_t|^2] \leq \exp(-2\alpha t)\mathbf{E}[|X_0 - Y_0|^2] = \exp(-2\alpha t) W_2(\mu_0, \pi)^2.$$

By taking square roots, we recover the claim. □

In the practical application of Monte-Carlo methods, one usually tries to estimate $\int f\,d\pi$ by calculating $P_t f(x) = \mathbf{E}_x[f(X_t)]$ for a large t. The next inequality is actually equivalent to the exponential convergence of $P_t f$ to $\int f\,d\pi$ in the L^2 norm:

Definition 3.4 (Poincaré Inequality)**.** Let X be a continuous Feller process with generator A and stationary distribution π. We say that X satisfies a Poincaré inequality if, for all functions $f \in \mathcal{D}(A)$,

$$-\int f(x)Af(x)\,d\pi(x) \geq C^{-1}\mathbf{Var}(f(X_0)).$$

We remark that this is claiming that A is a negative-definite operator with strictly negative eigenvalues. Indeed, if f is a solution to the eigenvalue problem $Af = \lambda f$, then one has $(f, -Af)_{L^2(\pi)} = \lambda\|f\|^2_{L^2(\pi)} \geq C^{-1}\|f\|_{L^2(\pi)}$, which means $-\lambda \geq C^{-1} > 0$. Combining the Poincaré inequality with Gronwall's lemma yields a good convergence result for the semi-group.

Lemma 3.5 ([8], Chapter 1.5)**.** *Suppose that X is a continuous Feller process with generator A and stationary distribution π which satisfies a Poincaré inequality with constant C. Then for any sufficiently nice $f \in L^2(\pi)$, it holds*

$$\|P_t f\|_{L^2(\pi)} \leq \exp(-C^{-1}t)\|f\|_{L^2(\pi)}.$$

The "sufficiently nice" condition conceals some obvious conditions that f needs to satisfy for the statement of the theorem to even make sense. For example, $P_t f \in L^2(\pi)$ for each $t > 0$.

Proof. Fix $f \in \mathcal{D}(A)$. Then the Kolmogorov Backwards equation yields

$$\partial_t \int |P_t f|^2\,d\pi = 2\int P_t f A P_t f \leq -2C^{-1}\|P_t f\|_{L^2(\pi)}.$$

Applying Gronwall's inequality to the function $u(t) = \|P_t f\|^2_{L^2(\pi)}$ gives the proof for $f \in \mathcal{D}(A)$. Since $\mathcal{D}(A)$ is dense in $L^2(\pi)$, we have the conclusion. □

In the special case where X is reversible (i.e., P_t is L^2 self-adjoint) Langevin diffusion, the Poincare inequality is a gradient domination condition. Indeed, since π is a stationary distribution, one has $\int A(fg)\,d\pi = 0$ for all suitably nice f and g, and the self-adjoint condition of the semi-group means A is self-adjoint. Therefore,

$$\int f(-Ag)\,d\pi = \int g(-Af)\,d\pi = \frac{1}{2}\int A(fg) - gAf - fAg\,d\pi.$$

Using $Af = \triangle f - \nabla V \cdot \nabla f$, one can calculate $A(fg) - gAf - fAg = |\nabla f|^2$, and so a Poincaré inequality reads

$$\|\nabla f\|^2_{L^2(\pi)} = \int f(-Af)\,d\pi \geq C^{-1}\|f\|^2_{L^2(\pi)}.$$

The converse of 3.5 also holds, but the statement of that lemma shows that proving a Poincaré inequality is enough for exponential $L^2(\pi)$ convergence of $P_t f$ to $\int f\,d\pi$.

Another type of convergence often employed in applications with respect to the KL-divergence,

$$\mathrm{KL}\left(\mu \,||\, \pi\right) := \int \frac{d\mu}{d\pi} \log \frac{d\mu}{d\pi}\, dx, \quad \mu \ll \pi.$$

A *log-Sobolev* inequality is equivalent to this type of convergence:

Definition 3.6. A continuous Feller process is said to satisfy a log-Sobolev inequality with constant C if for all density functions $\rho \in \mathcal{D}(A^*)$,

$$\mathrm{Ent}_\pi \rho := \int \rho \log \rho \, d\pi \leq \frac{C}{2} \int (A^* \rho) \log \rho \, d\pi.$$

Just as in the case of a Poincaré inequality, it turns out that a log-Sobolev inequality yields exponential convergence by utilizing Gronwall's inequality.

Lemma 3.7 ([8], Chapter 1.5)**.** *Suppose that X is a continuous Feller process with stationary distribution π and $X_0 \sim \pi_0$. If X satisfies a log-Sobolev inequality with constant $C > 0$, then*

$$\mathrm{KL}\left(\pi_t \,||\, \pi\right) \leq \exp(-2C^{-1}t)\mathrm{KL}\left(\pi_0 \,||\, \pi\right).$$

Proof. This is just applying the same Gronwall's theorem argument used in lemma 3.5, instead using the Kolmogorov forward equation instead of the backwards equation. □

We remark that a log-Sobolev inequality is generally much stronger than a Poincare inequality. The first indication of this fact is that 3.7 implies that a process satisfying an LSI inequality always converges in the KL-divergence to its stationary distribution.

Theorem 3.8. *Suppose that the Langevin diffusion with potential V satisfies a log-Sobolev inequality, and let μ_t be the density of the Langevin diffusion with potential V at time t. If $e^{-V} \in L^1(\mathbb{R}^d)$, then*

$$\mathrm{KL}\left(\mu_t \,||\, \pi\right) \leq \exp(-2\alpha t)\mathrm{KL}\left(\mu_0 \,||\, \pi\right).$$

That is, μ_t vaguely converges to the stationary distribution exponentially fast as $t \to \infty$.

This stationary distribution $\pi \propto e^{-V}$ is what makes the Langevin diffusion interesting from a MCMC perspective. The fact that certain estimates, such as the ones explored above, imply the chain is "ergodic" (in the sense that there is a unique stationary distribution for the chain, and μ_t always converges vaguely to this distribution) makes Langevin MCMC algorithms a fruitful area of study. Poincaré and log-Sobolev inequalities are but one of several different types of estimates that are fruitful when studying Markov processes. For a host of other useful inequalities, we refer the reader to Bakry et al.'s encyclopedic text on the convergence of Markov processes [2].

One property which we have been taking for granted this entire time is the fact that we can explicitly compute the solution to the Langevin SDE at all time points $t > 0$. In practical application, this is almost certainly never the case. As will be shown in the next section, we can only use numerical schemes to approximate the solution to the Langevin SDE. However, these approximate solutions will not identically preserve

the distribution of the continuous time process, and as such more clever tricks and additional assumptions are needed to make sure that the approximate chain is ergodic and converges to the intended stationary distribution.

Euler Discretization. In practice, we can almost never exactly compute the solution to the Langevin SDE. In practice, certain numerical schemes are used to generate approximate solutions. We fix $h > 0$ and some X_0. Then for each $k \geq 0$ we calculate $X_{(k+1)h}$ through the update

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}).$$

Sampling the random variable $B_{(k+1)h} - B_{kh}$ is quite straightforward, since this is Gaussian distributed. This is the so-called *Euler-Mayurama* discretized Langevin diffusion. Just as in ordinary gradient descent, it turns out that the discretized Langevin diffusion converges fairly rapidly in a certain sense. There are multiple interpretations of this statement, and we will present a few, starting with a proof via Wasserstein coupling.

The following lemma is considered standard (see, for example, [8], Lemma 4.0.1).

Lemma 3.9. *Let $\pi \propto \exp(-V) \in L^1(\mathbb{R}^d)$ have a finite second moment. If $\nabla^2 V \succeq \alpha I_d \succ 0$, and V is minimized at x_*, then if $X_0 \sim \pi$,*

$$\mathbf{E}[\|X_0 - x_*\|^2] \leq d\alpha^{-1},$$

and if $\beta I_d \succeq \nabla^2 V$, then

$$\mathbf{E}[\|\nabla V(X_0)\|^2] \leq \beta d.$$

Proof. Let $f = \frac{1}{2}\|\cdot - x_*\|^2$ and let A be the generator of the Langevin diffusion with potential V. Fix a non-negative and $[0,1]$-valued $\phi \in C_c^2(\mathbb{R}^d)$. Then $f\phi \in \mathcal{D}(A)$. By applying the product rule for the laplacian and gradient,

$$\begin{aligned} A(f\phi)(x) = d\phi(x) - 2\nabla\phi(x)\cdot(x - x_*) + f(x)\nabla\phi(x) \\ - \nabla V(x)\cdot(x - x_0)\phi(x) - (\nabla V(x)\cdot\nabla\phi(x))f(x). \end{aligned}$$

Let $R(x) = -2\nabla\phi(x)\cdot(x - x_*) + f(x)\nabla\phi(x) - (\nabla V(x)\cdot\nabla\phi(x))f(x)$, so that

$$\begin{aligned} 0 = \mathbf{E}[A(f\phi)(X_0)] = d\mathbf{E}[\phi(X_0)] - \mathbf{E}[\nabla V(X_0)\cdot(X_0 - x_*)] \\ + \mathbf{E}[R(X_0)] \leq d - \alpha\mathbf{E}_\pi[\|X_0 - x_*\|^2] + \mathbf{E}[R(X_0)]. \end{aligned}$$

As we let ϕ increase to the identically 1 function, we use the moment assumption on π to apply dominated convergence and get $\mathbf{E}[R(X_0)] \to 0$. Rearranging gives the desired inequality.

For the final statement, we can just take $f = V$ and run through the same set of steps to get

$$0 = \mathbf{E}[AV(X_0)] \leq d\beta - \mathbf{E}[\nabla V(X_0)\cdot(X_0 - x_*)] \leq d\beta - \mathbf{E}[\|V(X_0)\|^2].$$

□

We remark that the Langevin Monte-Carlo algorithm is essentially a noisy gradient descent. Therefore, many of the proofs of convergence of the Langevin diffusion will utilize the same essential ingredients as the proof of ordinary gradient descent, with various bells and whistles designed to regulate the stochastic elements. Therefore, we briefly review the convergence theorem for ordinary (non-stochastic) gradient descent.

Theorem 3.10 (Ordinary Gradient Descent Convergence.)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a C^2 convex function with a finite minimum $f(x_*)$ such that $0 \prec \alpha I_d \preceq \nabla^2 f \preceq \beta I_d$. Then the sequence of points $\{x_k\}_{k\geq 0}$ defined by the gradient descent $x_{k+1} = x_k - \gamma \nabla f(x_k)$ converges to x_* when $\gamma < 2\beta^{-1}$ for any choice of x_0.*

Proof. We observe that the following inequality holds:

$$\gamma\left(\frac{\alpha\gamma}{2} - 1\right)|\nabla f(x_{n-1})|^2 \leq f(x_n) - f(x_{n-1}) \leq \gamma\left(\frac{\beta\gamma}{2} - 1\right)|\nabla f(x_{n-1})|^2.$$

Since $0 < \gamma < 2\beta^{-1}$ it follows that the sequence $\{f(x_n)\}$ decreases to a limit, which must be finite since $f(x_n) \geq f(x_*) > -\infty$ for all n. Since $|\nabla f(x_{n-1})| \geq \alpha|x_{n-1} - x_*|$, we get

$$f(x_n) - f(x_{n-1}) \leq \gamma\left(\frac{\beta\gamma}{2} - 1\right)\alpha^2|x_{n-1} - x_*|^2,$$

whence

$$|x_{n-1} - x_*|^2 \leq \frac{f(x_n) - f(x_{n-1})}{\gamma\left(\frac{\beta\gamma}{2} - 1\right)\alpha^2},$$

and taking $n \to \infty$ implies that $x_n \to x_*$ as $n \to \infty$. □

A similar result holds for the discretized Langevin diffusion, but of course the diffusion term adds additional complexity to the proof.

Theorem 3.11 ([8], Theorem 4.12)**.** *For $h > 0$, define a sequence of random variables $\{X_{kh}\}$ by*

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}).$$

If $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, and μ_{kh} is the law of X_{kh}, then we have

$$W_2(\mu_{kh}, \pi) \leq \exp\left(-\frac{\alpha k h}{2}\right) W_2(\mu_0, \pi) + O\left(\frac{\beta d^{1/2} h^{1/2}}{\alpha}\right),$$

if let $h \sim \frac{\epsilon^2}{\beta d}$ then we obtain $\sqrt{\alpha} W_2(\mu_{Nh}, \pi) \leq \epsilon$ in

$$N = O\left(\frac{d}{\epsilon^2}\log\left(\frac{\sqrt{\alpha} W_2(\mu_0, \pi)}{\epsilon}\right)\right)$$

iterations.

See figures 1 and 2 for experimental evidence of theorem 3.11.

We need a quick technical lemma which relies on Gronwall's inequality.

Lemma 3.12. *Let $\{Z_t\}$ denote the Langevin diffusion and π_t denote its law. If ∇V is β-Lipschitz, then for $t \leq \frac{1}{3\beta}$ it holds*

$$\mathbf{E}[|Z_t - Z_0|^2] \leq 8t^2\mathbf{E}[|\nabla V(Z_0)|^2] + 8dt.$$

Proof. By definition of the Langevin diffusion and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\mathbf{E}\left[|Z_t - Z_0|^2\right] &= \mathbf{E}\left[\left|-\int_0^t \nabla V(Z_s)\,ds + \sqrt{2}B_t\right|\right] \\ &\leq 2t\int_0^t |\nabla V(Z_s)|^2\,ds + 4td.\end{aligned}$$

Since ∇V is β-Lipschitz, it holds $|\nabla V(Z_s)|^2 \leq 2|\nabla V(Z_0)|^2 + 2\beta^2|Z_t - Z_0|^2$, whence

$$\mathbf{E}\left[|Z_t - Z_0|^2\right] \leq 4\beta^2 t\int_0^t \mathbf{E}[|Z_s - Z_0|^2]\,ds + 4t^2\mathbf{E}[|\nabla V(Z_0)|^2] + 4td.$$

Gronwall's inequality implies that

$$\mathbf{E}[|Z_t - Z_0|^2] \leq (4t^2\mathbf{E}[|\nabla V(Z_0)|^2] + 4td)e^{4\beta^2t^2},$$

and the estimate $t \leq 3\beta^{-1}$ means $e^{4\beta^2t^2} \leq 2$, and the rest follows. □

We now begin the proof of Theorem 3.11.

Proof. The proof is broken into two steps. First, we prove an estimate for the single step case, and then use that to prove the estimate for the general n-step case.

In the single-step case, we have $X_h = X_0 - h\nabla V(Z_0) + \sqrt{2}B_h$. We couple this with the Langevin diffusion powered by the same Brownian motion:

$$Z_h = X_0 - \int_0^h \nabla V(Z_t)\,ds + \sqrt{2}B_h.$$

Then

$$\begin{aligned}W_2(\mu_h, \pi_h)^2 &= \mathbf{E}\left[\left|-\int_0^h \nabla V(Z_t)\,ds + h\nabla V(Z_0)\right|^2\right] \\ &\leq h\int_0^h \mathbf{E}[|\nabla V(Z_t) - \nabla V(Z_0)|^2]\,dt \\ &\leq h\beta^2\int_0^h \mathbf{E}[|Z_t - Z_0|^2]\,dt.\end{aligned}$$

Using Lemma 3.12 when $h \leq \frac{1}{3\beta}$, it yields

$$\begin{aligned} W_2(\mu_h, \pi_h)^2 &\leq h\beta^2 \int_0^h 8t^2 \mathbf{E}[|\nabla V(Z_0)|^2] + 8td\, dt \\ &= \frac{8}{3} h^4 \beta^2 + 4h^3 d\beta^2 \\ &\leq 3h^4\beta^2 \mathbf{E}[|\nabla V(Z_0)|^2] + 4h^3 d\beta^2. \end{aligned}$$

To deduce the multistep bound, let $X_{kh} \sim \mu_{kh}$ and $Z_{kh} \sim \pi$ be the optimal coupling of (μ_{kh}, π). Then let

$$\begin{aligned} X_{(k+1)h} &= X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}) \\ Z_t &= Z_{kh} - \int_0^t \nabla V(Z_s)\, ds + \sqrt{2}(B_t - B_{kh}) \quad t \in [kh, (k+1)h). \end{aligned}$$

By definition it holds $X_{(k+1)h} \sim \mu_{(k+1)h}$ and $Z_{(k+1)h} \sim \pi$, since π is the stationary distribution of the Langevin diffusion. Furthermore, let $\{\overline{X}_t : kh \leq t \leq (k+1)h\}$ denote the Langevin diffusion started at X_{kh}. Then we have

$$W_2(\mu_{(k+1)h}, \pi) \leq \|\overline{X}_{(k+1)h} - Z_{(k+1)h}\|_2 + \|\overline{X}_{(k+1)h} - X_{(k+1)h}\|_2.$$

Since $\overline{X}$ and Z are both Langevin diffusions run by the same Brownian motions, we have

$$\|\overline{X}_{(k+1)h} - Z_{(k+1)h}\|_2 \leq \exp(-\alpha h)\|X_{kh} - Z_{kh}\|_2 = \exp(-\alpha x) W_2(\mu_{kh}, \pi).$$

By our one-step bound,

$$\begin{aligned} \|\overline{X}_{(k+1)h} - X_{(k+1)h}\|_2 &\leq \sqrt{3h^4\beta^2 \mathbf{E}[|\nabla V(X_{kh})|^2] + 4h^3 d\beta^2} \\ &\leq \sqrt{6h^4\beta^2(\beta^2 \mathbf{E}[|X_{kh} - Z_{kh}|^2] + \mathbf{E}[|\nabla V(Z_{kh})|^2]) + 4h^3 d\beta^2} \\ &\leq \sqrt{6h^4\beta^2(\beta^2 W_2(\mu_{kh}, \pi)^2 + \beta d) + 4h^3 d\beta^2} \\ &\leq \sqrt{6}h^2\beta^2 W_2(\mu_{kh}, \pi) + \sqrt{6h^4\beta^3 d + 4h^3 d\beta^2}. \end{aligned}$$

So, in total we get

$$W_2(\mu_{(k+1)h}, \pi) \leq \exp(-\alpha h) W_2(\mu_{kh}, \pi) + O(\beta^2 h^2 W_2(\mu_{kh}, \pi) + \beta\sqrt{dh^3}).$$

If $h \approx \alpha\beta^{-2}$, then $\exp(-\alpha h) + O(\beta^2 h^2) \leq \exp(-\alpha h/2)$, so we will get

$$W_2(\mu_{(k+1)h}, \pi) \leq \exp(-\alpha h/2) W_2(\mu_{kh}, \pi) + O(\beta\sqrt{dh^3}).$$

Iterating the recursion yields the conclusion of the theorem. □

Essentially, Theorem 3.11 states that the convergence of the discretized chain in the Wasserstein sense exhibits an exponential contraction as in the continuous time case, plus an extra $O(h^{1/2})$ term to account for the discretization error. Observe that if we take $k = \lfloor t/h \rfloor$, as $h \to 0$, we get the estimate

$$\limsup_{h \to 0} W_2(\mu_{kh}, \pi) \leq \exp(-\alpha t/2) W_2(\mu_0, \pi),$$

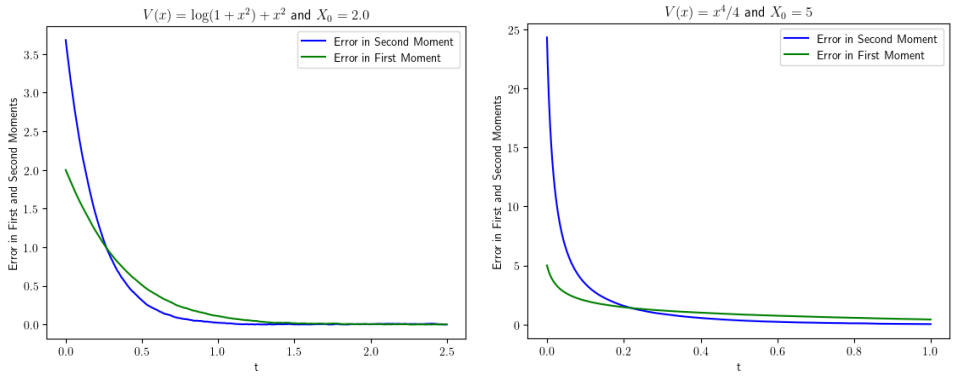

FIGURE 1. $\left|\int |x|^2 \, d\mu_t(x) - \int |x|^2 \, d\pi(x)\right|$ and $\left|\int |x| \, d\mu_t(x) - \int |x| \, d\pi(x)\right|$ for an Euler-Mayurama discretized Langevin diffusion when $V(x) = \log(1 + x^2) + x^2$ and $V(x) = x^4/4$ and $h = 0.001$ using 10000 samples. Both decay exponentially as both theorems 3.11 and 3.3 assert.

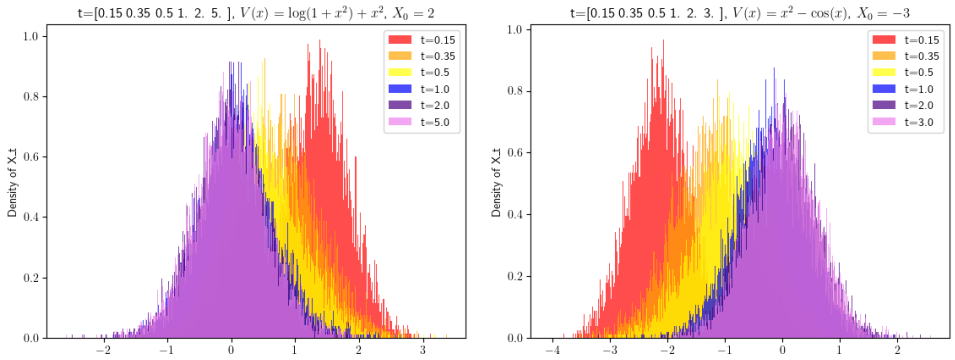

FIGURE 2. Empirical density function of X_t plotted as a function of t. By time $t = 2$, the histograms have essentially converged.

which is just a slightly worse version of the exponential mixing bound in Theorem 3.3. This is the desired behavior, since in the limit we should have μ_{kh} tends to the law at time t of the continuous Langevin SDE.

We remark that this is bound is essentially state of the art for the Wasserstein distance. It is possible to get similar estimates with a different choice of "distance", namely the Kullback-Leibler divergence. The current best estimate for the KL divergence is given as follows:

Theorem 3.13. *If* $0 \prec \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, *and* μ_{kh} *is as in theorem 3.11 and* $\overline{\mu}_{Nh} := \frac{1}{N}\sum_{k=1}^{N} \mu_{jh}$, *then* $\sqrt{\mathrm{KL}\left(\overline{\mu}_{Nh} \,||\, \pi\right)} \leq \epsilon$ *when*

$$N = O\left(\frac{\beta d W_2^2(\mu_0, \pi)}{\epsilon^4}\right).$$

Proof. See [13] or [8] Theorem 4.3.6. □

Metropolised Langevin Diffusion. As we have seen, one of the drawbacks of using a discretization approach to sampling is the error induced by discretization itself. In the ideal case, the Langevin SDE converges to stationary in exponential time, but the discretization picks up an additional $O(h^{1/2})$ term. Avoiding this additional error would be desirable. Therefore, we conclude this section with an alternative approach which manages to avoid a discretization error. It utilizes one of the most famous and ancient techniques in MCMC: the Metropolis-Hastings filter [21].

A slight change of notation is warranted for these discrete time Markov processes. Let $\{p(k, x, \cdot) : k \in \mathbb{N}, x \in \mathbb{R}^n\}$ be the transition kernels of a discrete time process X. By the Chapman-Kolmogorov equation, the kernels $p(k, \cdot, \cdot)$ are given recursively by the formula

$$p(k, x, A) = \int p(1, y, A)p(k-1, x, \, dy)$$

when $k \geq 2$, and in particular all of the kernels are uniquely determined by $p(1, \cdot, \cdot)$. It makes sense to call $P(x, \cdot) := p(1, x, \cdot)$ *the* transition kernel of X.

For the Metropolis-Hastings algorithm, we assume that the transition kernel P has a strictly positive density for each x, namely $P(x, dy) = Q(x, y)\, dy$. For a given strictly positive probability density π, we define an "acceptance" probability using the formula

$$A(x, y) = 1 \wedge \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}.$$

We then generate a sequence of points $\{X_k\}$ by the following algorithm: Fix some $X_0 \in \mathbb{R}^d$, and for each $k \geq 1$, iterate the following sequence of steps:

(1) Sample a point Y_k with distribution $P(X_{k-1}, \cdot)$
(2) With probability $A(X_{k-1}, Y_k)$, let $X_k = Y_k$. Otherwise, $X_k = X_{k-1}$.

The sequence of random variables $\{X_k\}$ form a Markov process with transition kernel

$$M(x, \, dy) = A(x, y)P(x, dy) + \left(1 - \int A(x, s)Q(x, ds)\right) \delta_x(dy).$$

We remark that π is a stationary distribution for this Markov process. This comes from the fact that M is reversible with respect to π, i.e., $\pi(x)M(x, y) = \pi(y)M(y, x)$, the usual definition from elementary Markov chain theory. The utility of Metropolis-Hastings is that the stationary distribution π needs only to be known up to a scaling constant, since only the quotient of π appears in the calculation of A. This spares the user of having to compute $\int \pi$, one of the most attractive features of this algorithm.

The relationship between the Metropolis-Hastings filter and the Langevin Monte-Carlo is the choice of the proposal. In the *Metropolis-Adjusted Langevin Algorithm* (MALA), we choose our proposal density Q to be a d-dimensional Gaussian with standard deviation $2h$ and mean $x - h\nabla V(x)$. The following result is the state of the art for MALA:

Theorem 3.14 ([7])**.** *Suppose that the potential V of a Langevin diffusion satisfies $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$ and $\nabla V(0) = 0$. Let X_0 be Gaussian distributed with mean $\beta^{-1} I_d$,*

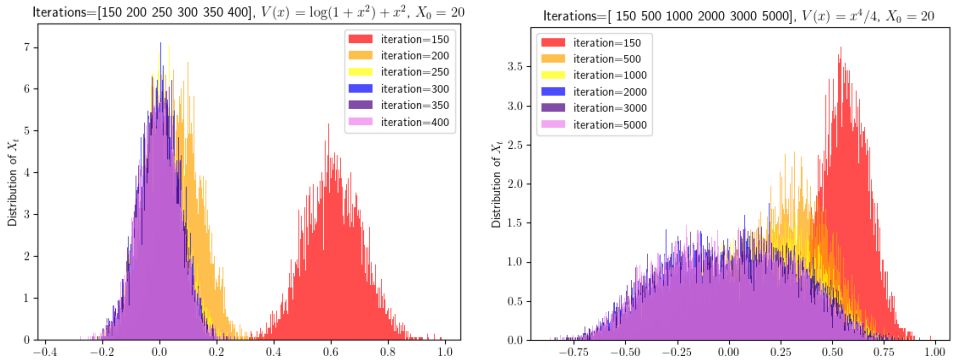

FIGURE 3. Probability distributions at a given iteration of the Metropolis-Adjusted Langevin algorithm.

and suppose that μ_N is the law of the N-iterate of the MALA algorithm. Then

$$\sqrt{\chi^2(\mu_n \,||\, \pi)} \leq \epsilon$$

when $N = O\left(\frac{\beta}{\alpha} d \operatorname{polylog} \epsilon^{-1}\right)$.

Here,

$$\chi^2(\mu \,||\, \pi) := \int \left(\frac{d\mu}{d\pi} - 1\right)^2 d\pi.$$

For empirical evidence of χ^2-convergence, see figure 3.

Proof. The proof of this theorem would take us too far afield for the purposes of this work. The proof involves deep insights into the geometric structure of the Metropolis-Hastings algorithm, and in particular relating the spectral gap of the semi-group to the Cheeger constant and conductance of the chain. We refer the reader to [7] and ([8], Chapter 7). □

The utility of using the Metropolis-Hastings filter is its simplicity to implement and its widespread adoption in practical applications and its ability to avoid picking up a discretization error. Indeed, the polylog in ϵ^{-1} convergence is almost identical to the exponentially fast convergence that we expect from solving the continuous time Langevin SDE.

4. Stochastic Gradient Langevin

In many important applications, the target potential V can be written in the form

$$V = \frac{1}{m} \sum_{i=1}^{m} V_i$$

for some smooth V_i When one runs an Euler discretized Langevin diffusion with potential V, that is,

$$X_{(k+1)h} = X_{kh} - \nabla V(X_{kh})h + \sqrt{2}(B_{(k+1)h} - B_{kh}),$$

it is necessary to compute the gradient for each of the V_i. In many cases, this does not pose a problem, especially when m is relatively small and the V_i are elementary functions (such as exp, log, polynomial, etc.). However, if m is too large or the functions V_i are particularly complicated, computing V_i for every i on each iteration can be prohibitively expensive. One way to overcome this difficulty is to, on each iteration k of the Euler discretization, instead of computing the full gradient

$$\nabla V = \sum_{i=1}^{m} \nabla V_i,$$

one selects a random subset S_k of $\{1, \dots, m\}$ independent of the Brownian motion (called a *mini-batch*, and $\#S_k$ is usually much smaller than m) such that

$$\mathbf{E}\left[\sum_{i \in S_k} \nabla V_i(X_{kh}) \,|\, X_{kh}\right] = \nabla V(X_{kh}).$$

For example, if we declare $\#S_k = 1$, then at each step k, we can select an index $i_k \in \{1, \dots, m\}$ uniformly at random. This provides an unbiased estimate of $\nabla V(X_{kh})$ since

$$\mathbf{E}\left[\nabla V_{i_k}(X_{kh}) \,|\, X_{kh}\right] = \frac{1}{m}\sum_{j=1}^{m} \nabla V_i(X_{kh}) = \nabla V(X_{kh}).$$

At each time step k, we instead update X_{kh} according to the rule

$$X_{(k+1)h} = X_{kh} - h\sum_{i \in S_k} \nabla V_i(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}). \tag{4.1}$$

This modified version of the Langevin algorithm motivates the *stochastic gradient langevin diffusion* (SGLD), which appears to have been first proposed by Welling and Teh [30]. Of course, sampling from only a subset of the gradient changes the dynamics of the system, and so new analysis is required. A few results in this direction can be found in ([6], [26], [3]). We will exhibit a few results from these works.

The first results we present in this section are meant to illustrate the pitfalls of SGLD, following the paper of Brosse, Durmus, and Moulines [6]. There, they make the following set of assumptions:

(1) V is α-strongly convex. That is, for all $x, y \in \mathbb{R}^d$, it holds

$$(\nabla V(x) - \nabla V(y)) \cdot (x - y) \geq \alpha |x - y|^2$$

for an $\alpha > 0$.

(2) V_i is convex for each i.

They also assume that each V_i is a C^4 function, but it is not necessary for the proof of the specific theorem we analyze. In this theorem, let $R^k(\cdot, \cdot)$ be the transition probability kernel at the k-th iteration.

Proposition 4.1 ([6], Lemma 1)**.** *Suppose at each iteration k, the mini-batches are sampled IID from $\{1,\dots,n\}$ and $\#S_k = p$. Then, for a step size $\gamma \in (0, 2/L] \cap (0, 1/(2\alpha)]$, there exists a invariant measure $\tilde{\pi} \in \mathcal{P}^2(\mathbb{R}^d)$ for the process $\{X_{k\gamma}\}$ such that*

$$W_2^2(R^k(\theta,\cdot),\tilde{\pi}) \le (1-\alpha h)^k \int |\theta-\vartheta|^2\, d\tilde{\pi}(\vartheta)$$

for any $\theta \in \mathbb{R}^d$.

Proof. Fix two initial distributions λ_1 and λ_2 in $\mathcal{P}^2$, to be determined later. Let (θ_0, θ_0') be an optimal coupling of (λ_1, λ_2), and let θ_k and θ_k' be SGLD's starting from θ_0 and θ_0', respectively. That is,

$$\theta_{k+1} = \theta_k + \gamma \left[\nabla V_0(\theta_k) + \sum_{i\in S_{k+1}} \nabla V_i(\theta_k)\right] + \sqrt{2\gamma} Z_k$$

$$\theta'_{k+1} = \theta'_k + \gamma \left[\nabla V_0(\theta'_k) + \sum_{i\in S_{k+1}} \nabla V_i(\theta'_k)\right] + \sqrt{2\gamma} Z_k,$$

where the Z_k are IID standard normals. For ease of notation, let

$$U_k = \nabla V_0 + \sum_{i\in S_{k+1}} \nabla V_i$$

Each U_k is strongly convex, and it follows that

$$\begin{aligned}
\|\theta_{k+1} - \theta'_{k+1}\|^2 &= \|\theta_k - \theta'_k\|^2 + \gamma^2 \|U_k(\theta_k) - U_k(\theta'_k)\|^2 - 2\gamma (x-y)\cdot(U_k(\theta_k) - U_k(\theta'_k)) \\
&\le \|\theta_k - \theta'_k\|^2 + (\gamma^2 L - 2\gamma)(x-y)\cdot(U_k(\theta_k) - U_k(\theta'_k)) \\
&= \|\theta_k - \theta'_k\|^2 - 2\gamma\left(1 - \frac{L\gamma}{2}\right)(x-y)\cdot(U_k(\theta_k) - U_k(\theta'_k)).
\end{aligned}$$

Conditioning on $\mathcal{F}_k$ and using the fact the S_k are independent of $\mathcal{F}_k$ and provide an unbiased estimate of ∇V, we get

$$\mathbf{E}[|\theta_{k+1} - \theta'_{k+1}|^2 \,|\, \mathcal{F}_k] \le \|\theta_k - \theta'_k\|^2 - 2\gamma\left(1 - \frac{L\gamma}{2}\right)(x-y)\cdot(\nabla V(\theta_k) - \nabla V(\theta'_k))$$

Since $1 - 2\frac{\gamma}{L} > 0$, the α-strong convexity of ∇V yields:

$$\mathbf{E}[|\theta_{k+1} - \theta'_{k+1}|^2 \,|\, \mathcal{F}_k] \le \left(1 - 2\alpha\gamma\left(1 - \frac{L\gamma}{2}\right)\right)|\theta_k - \theta'_k|^2.$$

Undoing the conditioning and iterating we get

$$W_2(\lambda_1 R^{k+1}, \lambda_2 R^{k+1}) \le \left(1 - 2\alpha\gamma\left(1 - \frac{L\gamma}{2}\right)\right)^k W_2(\lambda_1,\lambda_2)^2.$$

and since $2\alpha\gamma < 1$ it holds $0 \le 1 - 2\alpha\gamma\left(1 - \frac{L\gamma}{2}\right) < 1$. It follows that

$$\sum_{k\ge 0} W_2(\lambda_1 R^k, \lambda_2 R^k) < \infty.$$

We now take $\lambda_2 = \lambda_1 R$, where $R = R^1$, and it yields

$$\sum_{k\geq 0} W_2(\lambda_1 R^k, \lambda_1 R^{k+1}) < \infty,$$

so that $\{\lambda R^k\}$ is a Cauchy sequence in W_2. Since W_2 is complete with its metric, $\lambda_1 R^k \to \tilde{\pi}$ for a measure $\tilde{\pi}$.

We need to show that $\tilde{\pi}$ is the unique invariant measure of the SGLD, and that it is invariant under the choice of λ_1.

To see that it is invariant under the choice of λ_1, we observe that if λ_2 is another distribution and $\tilde{\mu}$ is the W_2 limit of $\{\lambda_2 R^k\}$, then we have

$$W_2(\tilde{\pi}, \tilde{\mu}) = \lim_{k\to\infty} W_2(\lambda_1 R^k, \lambda_2 R^k) = 0,$$

since $\sum_k W_2(\lambda_1 R^k, \lambda_2 R^k) < \infty$. To see that it is the stationary measure of the SGLD, observe that by Chapman-Kolmogorov we have $\tilde{\pi} R^{k+1} = (\tilde{\pi} R^k) R$, whence

$$W_2(\tilde{\pi}, \tilde{\pi} R) = \lim_{k\to\infty} W_2(\tilde{\pi} R^k, \tilde{\pi} R^{k+1}) = 0.$$

Thus $\tilde{\pi} R = \tilde{\pi}$, as required. □

Thus the semigroup of the SGLD converges to a stationary measure, but this proof does not promise that it is the stationary measure of the Langevin diffusion.

Random Reshuffling. Thus far we have put little burden on how the gradient is sampled. As shown in proposition 4.1, under the assumption of an IID sampling algorithm we can achieve exponentially fast convergence of the SGLD algorithm to *some* stationary distribution, but we lose the ability to fully describe the stationary distribution. We can partially remedy this situation through a better choice of random sampling. In the context of stochastic gradient descent algorithms, a method known as *random reshuffling* has been observed have better convergence rates in certain circumstances (see [5] and [12]). Naturally, the Langevin diffusion is analogous to a gradient descent, and so a recent paper due to Shaw and Welley [26] investigates SGLD with this random reshuffling scheme. With random reshuffling, one can actually guarantee convergence to the stationary distribution of the Langevin diffusion, a result which has not been promised thus far.

To introduce random reshuffling, as usual we assume the potential V can be written as

$$V = \sum_{i=1}^{N} V_i.$$

One picks integers m, n, which, for simplicity, satisfy $mn = N$. For each $1 \leq i \leq m$ and $1 \leq j \leq n$, we draw a sample ω_{ij} without replacement uniformly at random from the index set $\{1, \ldots, N\}$, and form the $m \times n$ matrix $\Omega = [\omega_{ij}]_{i,j}$ with rows ω_i for $1 \leq i \leq m$. On the i-th iteration for $1 \leq i \leq m$, we let the *mini-batch* S_i be the set of those indices contained in the vector ω_i. If we wish to go beyond m iterations, then on the $(m+1)$-st iteration we resample the matrix Ω in the same fashion and repeat until sufficiently

many iterations have passed. Each time we resample the matrix Ω is called an *epoch*. Since m is fixed throughout, the total number of iterations of SGLD using the random reshuffling method is typically measured in the number epochs R, or equivalently an integral multiple of m.

For analysis of SGLD with random reshuffling (SGLD-RR), we make some minor tweaks to our assumptions on V.

(1) V is continuously differentiable with $|\nabla V(x) - \nabla V(y)| \leq L|x-y|$.
(2) Each V_i is convex and α-strongly convex.
(3) The potential V is twice continuously differentiable and the Hessian has $\|\nabla^2 V(x) - \nabla^2 V(y)\| \leq L_1|x-y|$.
(4)
$$\sigma_*^2 := \mathbf{E}\left|\frac{1}{n}\sum_{i\in\omega} V_i(X) - V(X)\right|^2 < \infty,$$
where $X \sim \pi \propto e^{-V}$ and $\omega \subset \{1,\dots,N\}$ is sampled independently and uniformly at random with $\#\omega = n$.

For ease of notation, we will let $V_\omega := \frac{1}{\#\omega}\sum_{i\in\omega} V_i$ for a nonempty subset $\omega \subset \{1,\dots,N\}$.

Theorem 4.2 ([26], Theorem 3.9)**.** *Let $\pi \propto \exp(-V)$ and let $\{x_{kh}\}$ be the iterates of SGLD-RR with $x_0 \sim \pi_0$ for some distribution π_0. If $0 < h < 1/L$, then supposing $x_{kh} \sim \pi_{kh}$, we have*
$$W_2(\pi_{kh}, \pi) \leq (1-h\alpha)^k W_2(\pi_0, \pi) + O\left(\frac{L^2 h^{1/2} L_1 \sigma_* m}{\alpha} d\right).$$

To prove this theorem, we require some lemmas. The first one should be familiar at this point:

Lemma 4.3 ([26] Proposition A.1)**.** *Let $\{u_i\}_{i\geq 0}$ and $\{v_i\}_{i\geq 0}$ be two realizations of SLGD-RR driven by the same noise. Then for $h < 2/L$, it holds*
$$\mathbf{E}[|u_{r+k} - v_{r+k}|^2] \leq (1-h\alpha)^k \mathbf{E}[|u_r - v_r|^2].$$

Proof. This is just the same contractive property that we have used in Proposition 4.1 and in the discretized Euler-Mayurama scheme. □

The next lemma is a simple discrete analogue of Gronwall's inequality.

Lemma 4.4. *Suppose that $\{a_k\}$ is a non-decreasing sequence, $C > 0$ is a constant, and $\{x_k\}$ is a sequence such that $x_0 \leq a_0$ and*
$$x_k \leq a_k + C\sum_{i=0}^{k-1} x_i$$
for all $k \geq 0$. Then $x_k \leq a_k(1+C)^k$ for all $k \geq 0$.

Proof. If the claim is true for $0 \leq i \leq k$, then

$$x_{k+1} \leq a_{k+1}\left(1 + C\sum_{i=0}^{k}(1+C)^i\right) = a_{k+1}\left(1 + C\frac{(1+C)^{k+1}-1}{C}\right) = a_{k+1}(1+C)^{k+1},$$

which proves the claim by induction since $x_0 \leq a_0$. □

We abbreviate some steps of this lemma, since they are mostly tedious and mechanical to verify:

Lemma 4.5 ([26] Proposition A.2)**.** *Under the assumptions of theorem 4.2, and X_r is a continuous time Langevin diffusion with potential $V = \sum V_i$ started from π, and x_r is the k-th iterate of SLGD-RR,*

$$\|x_{r+k} - X_{(r+k)h}\|_2 \leq$$
$$4e^{khL}(h\sqrt{R}\sigma_* + h^{3/2}L\sqrt{k}(R\sqrt{hLd} + \sqrt{2Rd}) + \frac{1}{2}kh^2(L\sqrt{Ld} + L_1 d) + \frac{1}{3}h^{3/2}L\sqrt{kd}).$$

Proof. Let ω_k be the vector of indices at iteration k used by SGLD-RR, and let $\Delta_k = x_k - X_{kh}$. Then it holds

$$\begin{aligned}\Delta_k &= \Delta_{k-1} - \int_{(k-1)h}^{kh} \nabla V_{\omega_{k-1}}(x_{k-1}) - \nabla V(X_t)\,dt \\ &= \Delta_{k-1} - hQ_{k-1} + hD_{k-1} + E_{k-1},\end{aligned}$$

where

$$\begin{aligned}Q_{k-1} &:= \nabla V_{\omega_{k-1}}(x_{k-1}) - \nabla V_{\omega_{k-1}}(X_{kh}) \\ D_{k-1} &:= \nabla V(X_{(k-1)h}) - \nabla V_{\omega_{k-1}}(X_{(k-1)h}) \\ E_{k-1} &:= \int_{(k-1)h}^{kh} \nabla V(X_t) - \nabla V(X_{(k-1)h})\,dt.\end{aligned}$$

Observe that when $\Delta_r = 0$, we have

$$\Delta_{r+k} = \sum_{i=r+1}^{r+k} \Delta_i - \Delta_{i-1} = \sum_{i=r}^{r+k-1} hQ_i + \sum_{i=r}^{r+k-1}(hD_i + E_i).$$

If R is the epoch count, then since $\|Q_k\|_2 \leq hL\|\Delta_k\|$, we have

$$\begin{aligned}\|\Delta_{r+k}\|_2 \leq hL\sum_{i=r}^{r+k-1}\|\Delta_i\|_2 + h\left\|\sum_{i=R\lfloor (r+k-1)/R\rfloor}^{r+k-1} D_i\right\| + h\left\|\sum_{i=r}^{R\lfloor r/R\rfloor - 1} D_i\right\| \\ + h\left\|\sum_{i=R\lfloor r/R\rfloor}^{\lfloor (r+k-1)/R\rfloor - 1} D_i\right\| + \left\|\sum_{i=r}^{r+k-1} E_i\right\|_2.\end{aligned}$$

The idea is that the sums containing the D_i are broken up into three terms: the first two are head and tail terms, and the last are those D_i contained in a full epoch of the

RR algorithm. Shaw and Welley then estimate the head/tail terms:

$$h\left\|\sum_{i=R\lfloor (r+k-1)/R\rfloor}^{r+k-1} D_i\right\| + h\left\|\sum_{i=r}^{R\lfloor r/R\rfloor -1} D_i\right\| \leq 2h\sqrt{\frac{7}{2}R}\sigma_*,$$

which utilizes lemma 14 from [23], and bound the terms contained in a full epoch by

$$h\left\|\sum_{i=R\lfloor r/R\rfloor}^{\lfloor (r+k-1)/R\rfloor -1} D_i\right\| \leq 2\sqrt{\frac{7}{2}}h\sqrt{k}L(hR\sqrt{Ld} + \sqrt{2hRd}).$$

Finally, they estimate

$$\left\|\sum_{i=r}^{r+k-1} E_i\right\|_2 \leq 2h^2k(L\sqrt{Ld} + L_1 d) + h^{3/2}L\sqrt{2kd}$$

using ([11], Lemma 2). Adding these all up gives

$$\begin{aligned}\|\Delta_{r+k}\|_2 \leq hL\sum_{i=r}^{r+k-1}\|\Delta_i\|_2 + 2h\sqrt{\frac{7}{2}R}\sigma_* \\ + 2\sqrt{\frac{7}{2}}h\sqrt{k}L(hR\sqrt{Ld} + \sqrt{2hRd}) + 2h^2k(L\sqrt{Ld} + L_1 d) + h^{3/2}L\sqrt{2kd}\end{aligned}$$

The claim then follows by the discrete form of Gronwall's inequality 4.4, using obvious choice for a_k, the fact $1 + hL \leq e^{hL}$, and the estimate

$$\begin{aligned}&2h\sqrt{\frac{7}{2}R}\sigma_* + 2\sqrt{\frac{7}{2}}h\sqrt{k}L(hR\sqrt{Ld} + \sqrt{2hRd}) + 2kh^2(L\sqrt{Ld} + L_1 d) + h^{3/2}L\sqrt{d} \\ &< 4\left(h\sqrt{R}\sigma_* + h\sqrt{k}L(hR\sqrt{Ld} + \sqrt{2hRd}) + \frac{1}{2}kh^2(L\sqrt{Ld} + L_1 d) + \frac{1}{2}h^{3/2}L\sqrt{kd}\right).\end{aligned}$$

□

With these two lemmas, Shaw and Welley deliver the full proof of theorem 4.2.

Proof of Theorem 4.2. Let $(x_0, \tilde{x}_0)$ be an optimal coupling of π_0 and π. Let $\{x_i\}_{i\geq 0}$ be the iterates of the SGLD-RR scheme started from x_0 and $\{\tilde{x}_i\}_{i\geq 0}$ be the iterates of the SGLD-RR scheme started from $\tilde{x}_0$, driven by the same Brownian motions. Let $(X_{ih})_{i=0}^k$ be sampled from a continuous Langevin diffusion driven by the same Brownian motion as $\{\tilde{x}_i\}$ with $X_0 = \tilde{x}_0$.

For a fixed $k \geq 0$ and $0 \leq \ell \leq k$ Shaw and Welley also introduce an interpolated sequence $\{X_i^\ell\}_{i=1}^k$ by:

(1) $X_0^\ell = x_0$
(2) $X_i^\ell = X_{ih}$ if $0 \leq i \leq \ell$
(3) $X_i^\ell = x_i$ when $\ell + 1 \leq i \leq k$.

If $\tilde{k} = \lceil (hL)^{-1} \rceil$, then

$$\|x_k - X_{kh}\|_2 \leq \|X_k^{\lfloor k/\tilde{k} \rfloor} - X_k^k\|_2 + \sum_{j=0}^{\lfloor k/\tilde{k} \rfloor - 1} \|X_k^{j\tilde{k}} - X_k^{(j+1)\tilde{k}}\|_2.$$

By lemmas 4.3 and 4.5 we have

$$\|X_k^{(j\tilde{k})} - X_k^{(j+1)\tilde{k}}\|_2 \leq A(1-h\mu)^{k-(j+1)\tilde{k}}$$

where A is the non-exponential expression depending on $\tilde{k}$ in 4.5. Summing,

$$\|x_k - X_{kh}\|_2 \leq \frac{2A}{1-(1-h\alpha)^{\tilde{k}}} \leq 2A\left(1 + \frac{1}{h\alpha\tilde{k}}\right).$$

Using the fact $\tilde{k} = \lceil \frac{1}{hL} \rceil$, we recover

$$W_2(\tilde{\pi}_k, \pi) \leq \|x_k - X_{kh}\|_2 \leq \frac{2A}{1-(1-h\alpha)^{\tilde{k}}} \leq 4A\frac{L}{\alpha}$$

where $x_k \sim \tilde{\pi}_k$. Letting π_{kh} being as defined in the theorem statement (the law of k-th iteration of SGLD-RR started at a distribution π_0) we have

$$W_2(\pi_{kh}, \pi) \leq W_2(\pi_{kh}, \tilde{\pi}_k) + W_2(\tilde{\pi}_k, \pi) \leq W_2(\pi_{kh}, \tilde{\pi}_k) + 4A\frac{L}{\alpha}.$$

Using lemma 4.3 and using the choice of $(x_0, \tilde{x}_0)$ as an optimal coupling of π_0 and π,

$$W_2(\pi_{kh}, \pi) \leq (1-h\alpha)^k W_2(\pi_0, \pi) + \frac{4AL}{\alpha}.$$

Since A grows like

$$A = O\left(Lh^{1/2}L_1\sigma_* md\right),$$

the proof is complete. □

Theorem 4.2 says that with random reshuffling, convergence of the SGLD-RR algorithm is virtually identical to Theorem 3.11, since it both cases there is an upper bound of $\exp(-\alpha kh)W_2(\pi_0, \pi) + O(h^{1/2})$. There is slightly worse dependency on the condition number L/α in SGLD-RR, but otherwise SGLD-RR nearly matches the state-of-the-art estimate for the Euler discretization for the non-stochastic Langevin SDE. This indicates that SGLD-RR is an optimal way to incorporate stochastic gradient into the discretized Langevin process.

5. Concluding Remarks

As hinted at in the discussion of MALA, sampling from a Langevin diffusion with log-concave potential is not the only approach to log-concave sampling. There are several other avenues which this work did not explore. In fact, the Langevin diffusion is only a specific case of a more general SDE known as the Hamilton-Jacobi equations. Moreover, while a log-concave potential yields exponentially fast convergence to stationary in several different senses (2-Wasserstein, Kullback-Leibler, χ^2, etc.), it is a much more

difficult and open problem to determine optimal convergence rates when the target distribution is not log-concave. For more recent progress in this direction, see ([8], chapter 11) and [3].

Moreover, the astute reader may have noticed that the Wasserstein geometry introduced in the background section had little relevance in the analysis we presented later. While much of the existing literature can be painfully derived using the classical techniques employed in most of the theorems presented here, the Wasserstein geometry (especially the Otto calculus) can be exploited to derive alternative proofs of many theorems (such as Theorem 3.3). However, this would require a much more thorough exposition on the Otto calculus, which requires a background in Riemannian geometry not assumed here. However, the interested reader can view the work of Chewi-Niles-Weed-Rigollet [9] which introduces the Otto calculus with an eye towards probabilistic applications, or the classic work of Villani [29] for a more rigorous exposition for pure mathematicians.

References

[1] L. Ambrosio, E. Brué, and D. Semola. *Lectures on Optimal Transport*, volume 169 of *UNITEXT*. Springer, Cham, Switzerland, 2024.

[2] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

[3] K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and S. Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

[4] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, 2nd edition, 1999.

[5] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pages 2624–2633. Citeseer, 2009.

[6] N. Brosse, A. Durmus, and E. Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Y. Chen and K. Gatmiry. A simple proof of the mixing of metropolis-adjusted langevin algorithm under smoothness and isoperimetry, 2023.

[8] S. Chewi. Log-concave sampling. Available at `https://chewisinho.github.io/main.pdf`, 2025.

[9] S. Chewi, J. Niles-Weed, and P. Rigollet. Statistical optimal transport, 2024.

[10] K. L. Chung. *Foundations of Stochastic Analysis*. Springer, New York, revised edition, 2001.

[11] A. S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent, 2017.

[12] C. M. De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33:5957–5967, 2020.

[13] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of langevin monte carlo via convex optimization, 2018.

[14] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, Cambridge; New York, 5th edition, 2019.

[15] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, New York, 2nd edition, 1999.

[16] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[17] C. E. Gutiérrez. *The Monge-Ampère Equation*, volume 89 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Cham, Switzerland, 2nd edition, 2016.

[18] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, New York, 2nd edition, 1991.
[19] T. M. Liggett. *Continuous Time Markov Processes: An Introduction*, volume 113 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, R.I., 2010.
[20] F. Maggi. *Optimal Mass Transport on Euclidean Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge; New York, 2023.
[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. Technical report, Los Alamos Scientific Lab., Los Alamos, NM (United States); Univ. of Chicago, IL (United States), 03 1953.
[22] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[23] D. Paulin, P. A. Whalley, N. K. Chada, and B. Leimkuhler. Sampling from bayesian neural network posteriors with symmetric minibatch splitting langevin dynamics, 2025.
[24] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 2nd edition, 1974.
[25] W. Rudin. *Functional Analysis*. McGraw-Hill, New York, second edition edition, 1991.
[26] L. Shaw and P. A. Whalley. Random reshuffling for stochastic gradient langevin dynamics, 2025.
[27] D. W. Stroock. Weyl's lemma, one of many. In K. Tent, editor, *Groups and Analysis: The Legacy of Hermann Weyl*, pages 164–173. Cambridge University Press, 2008.
[28] S. S. Vempala and A. Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices, 2022.
[29] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.
[30] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, Bellevue, WA, USA, 2011. ACM.
[31] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin; New York, 6th edition, 2003.