Part III Essay 2023-24
Faculty of Mathematics, DPMMS

Sampling from High-Dimensional Distributions

Author:
Ya Shi Zhang
ysz23@cam.ac.uk

Supervisor:
Dr. Randolf Altmeyer

Contents

1 Introduction

In this section, I will first provide one of many formulations of the problem we wish to solve. Afterwards, I will attempt to motivate the reader by first viewing some applications of sampling, then by alluring the reader using the mathematical beauty of sampling and its connections to gradient flows and optimization.

1.1 The Problem

Suppose we have a function $f : \mathbb{R}^p \to \mathbb{R}$, the **sampling problem** is to output samples $X \sim \pi$, where $\pi(x) = (1/Z) \exp(-f(x))$, where $Z := \int_{\mathbb{R}^p} \exp(-f(x))\, dx$ is the normalization constant. In most cases, the normalization constant is computationally intractable, and so we would only have access to π up to proportionality. With the development of computational tools such as AutoGrad [Maclaurin et al., 2015], we are also able to compute the gradient of f and hence the gradient of π via backpropagation.

This, however, is not sufficient for directly sampling form π. Instead, we settle for an algorithm that can produce an approximate (in some sense) sample from π. Given some target accuracy, we wish to minimize the computational resources required to produce an approximate sample satisfying the accuracy requirement. Additionally, p may be extremely large. The main goal for the last 10 years has been to provide samplers that scale well with increasing dimensionality - circumventing the curse of dimensionality for sampling.

Now, given a convex function $f : \mathbb{R}^p \to \mathbb{R}$, the theory of finding $x \in \arg\min_{\mathbb{R}^p} f$ is extremely mature and developed [Fawzi, 2024]. Upper and lower bounds for the speed of convergence of various convex optimization algorithms have been analyzed. This is not the case for sampling. Algorithms for sampling from $\pi \propto \exp(-f)$ are almost always more expensive than algorithms for minimizing f – in fact, most samplers assume that you will initialize your sampler at a minimizer of f, since the computational cost of finding a minima will be dwarfed by that of sampling.

1.2 Why?

One may ask the purpose of posing such a problem. For the computational mathematician, many problems can be solved approximately. For Bayesians, posterior inference is a key step to realizing the models for real-world use cases. Finally, for those only interested in theory, there are deep connections between sampling and convex optimizations, realized through fascinating theories such as optimal transport and differential equations.

A simple and reductive case is estimating the area of the unit circle in $\mathbb{R}^2$. If we somehow forgot the formula for computing this, we can always sample random variables uniform on $[-1, +1]^2$, and compute the ratio of samples that lie within the unit circle - this is known as the Monte Carlo method. Of course, the problem is producing uniform random numbers is highly non-trivial, and this task is delegated to the cryptographers.

A slightly more complex example would be in computing functionals of the target distribution π. Given our framework above, if we only had access to the function f and wanted to compute $\mathbb{E}_{X\sim\pi}[X]$, a simple estimator would be to generate n samples $X_1, \ldots, X_n$ and compute $\bar{X} := n^{-1} \sum_{i=1}^{n} X_i$. We are also allowed to approximate more complex functionals.

An even more complex example would be in generative modelling [Song and Ermon, 2019, Ho et al., 2020, Alamdari et al., 2023]. In this machine learning problem, we are given access to a dataset $\mathcal{D} = \{X_i\}_{i=1}^n \sim \pi$. Our task is to produce a function based on the data $f_{\mathcal{D}}(\cdot)$ such that $f_{\mathcal{D}}(z) \sim \pi$, where $z \sim \mathcal{N}(0, I_p)$ and p is the dimensionality of the data. The popularity of generative models have skyrocketed with commercial applications such as Stable Diffusion Rombach et al. [2021] and DALL-E Ramesh et al. [2022].

Finally, the most common use case of sampling algorithms comes from Bayesian inference. In a general statistical modelling problem, we would first specify a statistical model, parametrized by $\theta \in \mathbb{R}^p$ (note that sampling is also possible in the non-parametric case [Hjort et al., 2010]). Then, the Bayesian approach is to specify a prior density function over these parameters, $p_\theta(\cdot)$, such as a Gaussian or Laplacian. We also specify a likelihood function $\ell(\theta; \mathcal{D}) = p_{\mathcal{D}|\theta}$. Then, after observing the data $\mathcal{D} = \{X_i\}_{i=1}^n$, we can compute, using Bayes' rule, the posterior distribution

$$p_{\theta|\mathcal{D}} = \frac{p_{\mathcal{D}|\theta} p_\theta}{p_{\mathcal{D}}}.$$

Even for 'simple' statistical models such as Bayesian logistic regression, the posterior is often intractable. The problem is only exacerbated in high dimensions, typically when $p \gg n$. Hence, the need for sampling algorithms that can explore the space of the data well is crucial for the purposes of Bayesian inference. Of course, for predictive statistics, the predictive distribution also relies on having a good approximation of the posterior.

Other techniques such as variational inference attempts to directly fit a tractable distribution as an approximation to the posterior [Blei et al., 2017]. In certain problems where knowing the entire distribution is required, variational inference can be better. However, variational inference generally produce samples that have lower fidelity, accompanied with some other issues. While the general consensus is that MCMC methods run slower but produce higher fidelity samples compared to variational inference, developments in computer hardware (GPUs) have enabled massive parallelization of Markov chains.

For most applicable problems, algorithms such as Markov Chain Monte Carlo or Gibbs sampling would be more accurate and even faster [Bishop and Bishop, 2024].

1.3 Outline of Contributions

The main contributions of this paper are to refine and present, in greater detail, the proofs of several recent and/or classical papers that contribute to the theory of sampling, primarily focusing on Langevin-based sampling algorithms. The aim is to provide a clear and accessible introduction to the topic, targeted at early master's level students wising to gain an understanding of the theoretical underpinnings of modern sampling methods.

It should also be noted that the proofs of a select few propositions and lemmas (e.g. Example problems in Chewi [2024]) may have a slightly different approach than in literature (as far as we are concerned). However, the primary focus is still on clarifying and adding details for proofs appearing in literature.

2 Background

This section consists of theory which will be used throughout the rest of this text.

2.1 Notation

Let $I_p \in \mathbb{R}^{p \times p}$ denote the identity matrix and $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product unless otherwise specified. Let $\mathcal{C}^n(\mathbb{R}^p; \mathbb{R}) = \mathcal{C}^n(\mathbb{R}^p)$ be the set of functions $f : \mathbb{R}^p \to \mathbb{R}$ that are n-times continuously differentiable. Smoothness refers to continuously differentiable with Lipschitz gradient. Let $\mathcal{C}_c^n(\mathbb{R}^p)$ be the functions in $\mathcal{C}^n(\mathbb{R}^p)$ with compact support. Let $\mathcal{L}^p(\pi)$ to be (up to $L^p(\pi)$ indistinguishability) the space of functions whose p-th power is π-integrable. Let $\mathcal{L}^p(\mathbb{R}^n)$ for some $n \in \mathbb{N}$ denote the (classical) $\mathcal{L}^p$-space with respect to the Lebesgue measure on $\mathbb{R}^n$. Denote $\mathcal{B}(\mathbb{R}^p)$ the Borel-σ-algebra of $\mathbb{R}^p$. We overload $\sigma(\cdot)$ to denote the diffusion coefficient when the argument is a random variable, and to denote the generated sigma-algebra when the argument is a collection of sets. For two random variables X, Y, we denote independence by $X \perp Y$. For two probability measures $\mathbb{P}, \mathbb{Q}$, we denote equivalence (mutually absolutely continuous) by $\mathbb{P} \sim \mathbb{Q}$, note that $\sim$ is overloaded with 'is distributed as'.

Denote $\mathcal{P}(\mathbb{R}^p)$ to be the set of probability measures over $\mathbb{R}^p$, $\mathcal{P}_2(\mathbb{R}^p)$ the set of probability measures over $\mathbb{R}^p$ with finite second moment (i.e. μ s.t. $\mathbb{E}_{X\sim\mu}[\|X\|_2^2] < \infty$), and $\mathcal{P}_{2,ac}(\mathbb{R}^p)$ those probability measures in $\mathcal{P}_2(\mathbb{R}^p)$ that are absolutely continuous with respect to the Lebesgue measure. For simplicity, we assume that the processes that are solutions to Itô diffusions admit a density with respect to the Lebesgue measure for $t = 0$ (implying the existence of densities for $t > 0$ by applying the Markov semigroup, as will be shown later). Furthermore, if $X \sim \pi$ and π admits density p, we may also write $X \sim p$ instead. When integrating against a probability measure π that admits a density with respect to the Lebesgue measure, we may overload π to denote both the density and measure, i.e. $\int \mu = \int \mu(dx) = \int \mu(x)\, dx$. If the argument of π is a measurable subset $B \in \mathcal{B}(\mathbb{R}^p)$ then π denotes the measure, else if the argument if π is an element $x \in \mathbb{R}^p$ then π denotes the density.

2.2 Basic Markov Semi-group Theory

Before beginning, when referring to some SDE – unless otherwise specified – solutions are assumed to be (1) **strong** and (2) **pathwise unique**. That is, for the Itô diffusion $dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t$ started at $x_0 \in \mathbb{R}^p$, the solution process $(X_t)_{t\geq 0}$ is (1) adapted to $\sigma(W_t : t \geq 0)$ and (2) fixing the ambient space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ and Brownian motion $(W_t)_{t\geq 0}$, for any two solutions $(X_t)_{t\geq 0}, (\widetilde{X}_t)_{t\geq 0}$ started at x_0, we must have $\mathbb{P}(\widetilde{X}_t = X_t\ \forall t \geq 0) = 1$. Hence, we will take the ambient filtration to be the filtration generated by the Brownian motion: $\mathcal{F}_t = \sigma(W_s : s \leq t)$.

One can verify that all of the Itô diffusions indeed have Lipschitz diffusion and drift coefficients, usually due to a $\mathcal{C}^1$ assumption on the drift and constant diffusivity. Hence, by Theorem 7.2 of Miller and Silvestri [2024], the SDE admit – for each starting point $x \in \mathbb{R}^p$ – a pathwise unique strong solution.

Two important things to remark in our setting:

1. By considering only strong and pathwise unique solutions to SDEs, we know that these solutions have the **strong Markov property**. That is, for all bounded and Borel-measurable f on $\mathbb{R}^p$ and τ a.s.-finite stopping times and for all $s \geq 0$, we have

   $$\mathbb{E}^x[f(X_{\tau+s}) \mid \mathcal{F}_\tau] = \mathbb{E}^{X_\tau}[f(X_s)],$$

   where $\mathbb{E}^x[f(X_t)] \coloneqq \mathbb{E}[f(X_t) \mid X_0 = x]$ [Oksendal, 1992].

2. By only considering Itô diffusions drift and diffusion coefficients that do not depend directly on time, solutions to our SDE will be **time-homogeneous** (or simply homogeneous). That is, for all $s, t \geq 0$, $x \in \mathbb{R}^p$, and all $B \in \mathcal{B}(\mathbb{R}^p)$, we have

   $$\mathbb{P}(X_{s+t} \in B \mid X_s = x) = \mathbb{P}(X_t \in B \mid X_0 = x).$$

Now that we know solutions to Itô diffusions are homogeneous strong Markov processes, we can start establishing some theory for Markov processes. Suppose we are given a homogeneous strongly Markovian process $(X_t)_{t\geq 0}$. There are a few properties that we are interested in. We first define the **Markov semigroup** associated with X.

Definition 1 (Markov Semigroups)**.** *Given a homogeneous Markov process $(X_t)_{t\geq 0}$, its associated Markov semigroup (or transition kernel) is the collection of operators $(P^t)_{t\geq 0}$ that acts to the right on bounded Borel-measurable functions f via*

$$(P^t f)(\cdot) = \mathbb{E}[f(X_t) \mid X_0 = \cdot].$$

Notably, if $f = \mathbb{1}_B$ for $B \in \mathcal{B}(\mathbb{R}^p)$, we recover the notion of transition kernels (as seen in the theory of Markov chains):

$$P^t(x, B) \coloneqq (P^t \mathbb{1}_B)(x) = \mathbb{P}(X_t \in B \mid X_0 = x)$$

Furthermore, given a probability measure with density μ with respect to the Lebesgue measure, the operator acts to the left on μ via

$$(\mu P^t)(\cdot) = \int_{\mathbb{R}^p} \mu(x) P^t(x, \cdot)\, dx.$$

The reason why its called the Markov semigroup is because

1. $P^0 = \mathrm{id}$, where id is the identity operator.
2. $P^t(P^s P^r) = (P^t P^s)P^r$ for all $r, s, t \geq 0$.
3. $P^t P^s = P^s P^t = P^{s+t}$ for all $s, t \geq 0$.

The first property is easily verified by definition as $P^0 f(x) = \mathbb{E}^x[f(X_0)] = f(x)$. The second and third properties also directly follow from the definition of time-homogeneity, the strong Markov property, and the tower property Chewi [2024][Lemma 1.2.2].

Now, we very briefly touch on the **infinitesimal generators** of a Markov semigroup. The infinitesimal generator (or generator) associated with a Markov semigroup and hence an Itô diffusion is a second order partial differential operator acting on certain functions (specified in a moment).

Definition 2. *Let $(P^t)_{t\geq 0}$ be a Markov semigroup. Then the infinitesimal generator, or generator, associated with $(P^t)_{t\geq 0}$ acting on f is defined as*

$$(Af)(\cdot) \coloneqq \lim_{t \searrow 0} \frac{(P^t f)(\cdot) - f(\cdot)}{t}$$

whenever the limit is well-defined.

From Oksendal [1992][Theorem 7.3.3], we explicitly have the form of the generator and what functions it can act on.

Fact 3. *Let $(X_t)_{t\geq 0}$ be the solution to the Itô diffusion $dX_t = b(X_t)\, dt + \sigma(X_t)\, dW_t$. If $f \in \mathcal{C}^2(\mathbb{R}^p; \mathbb{R})$ and has compact support, then the limit in Definition 2 exists for f and we have*

$$(Af)(x) = \sum_{i=1}^p b_i(x) \frac{\partial f}{\partial x_i} + \sum_{i,j=1}^p (\frac{\sigma\sigma^\top}{2})_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j} = \langle b(x), \nabla f(x) \rangle + \mathrm{Tr}\left((\nabla^2 f(x)) \frac{\sigma(x)\sigma^\top(x)}{2} \right). \tag{1}$$

Furthermore, we can directly compute the adjoint of A to get

$$(A^* f)(x) = -\sum_{i=1}^p \frac{\partial (b_i f)}{\partial x_i}(x) + \sum_{i,j=1}^p \frac{\partial^2 (\frac{\sigma(x)\sigma^\top(x)}{2} f)}{\partial x_i \partial x_j}(x). \tag{2}$$

Since A is the infinitesimal operator corresponding to the Itô diffusion defined by b, σ, we say that a second-order differential operator of this form is called an A-diffusion, or (b, σ)-diffusion, operator. The operator A and its adjoint A^* are central to the study of both PDEs and SDEs – due to Kolmogorov's backward equation and the forward equation (also known as the Fokker-Planck equation when the SDE is an Itô diffusion). We will state the Fokker-Planck equation.

Fact 4. *For a (b, σ)-Itô diffusion, suppose $X_t \sim \pi_t$ with density $p_t(\cdot)$ for all $t \geq 0$, then*

$$\partial_t p_t = A^* p_t,$$

where A^* is as Equation (2).

Note that it is enough to specify only the initial distribution $X_0 \sim \pi_0$, and then use the adjoints of the Markov semigroup $(P^t)^*$ to evolve the distribution of X_0 towards X_t. More information can be found in Oksendal [1992][8.1-8.3].

Before ending the section, we will also briefly talk about the stationary measure for Markov processes, and how they relate to the generators defined above. A **stationary measure** (or **invariant measure**) of a Markov process is a probability measure π such that if $X_0 \sim \pi$, then $X_t \sim \pi$ for all $t \geq 0$. Using the Fokker-Planck equation, we state some important equivalent conditions for stationarity.

Proposition 5. *Given a Markov process $(X_t)_{t \geq 0}$ with generator A and probability density π over the state space, the following are equivalent:*

1. *If $X_0 \sim \pi$, then for all $t \geq 0$, we have $X_t \sim \pi$.*
2. *$A^* \pi = 0$.*
3. *For all $g \in \mathcal{L}^2(\pi)$, we have $\mathbb{E}_\pi[Ag] = 0$.*

Proof. For (1. $\Rightarrow$ 2.), note that stationarity is equivalent to $\partial_t p_t = 0$ and we are done after viewing Fokker-Planck. This also gives (2. $\Rightarrow$ 1.). For (2. $\Leftrightarrow$ 1.), we note that $(A^*\pi = 0) \Leftrightarrow (\langle g, A^*\pi \rangle_{\mathcal{L}^2(\pi)} = 0$ for all $g \in \mathcal{L}^2(\pi)) \Leftrightarrow (\langle Ag, \pi \rangle_{\mathcal{L}^2(\pi)} = \mathbb{E}_\pi[Ag] = 0$ for all $g \in \mathcal{L}^2(\pi))$ by the definition of the adjoint. ■

To finish off the section, we perform a computation that gives us the stationary distributions of the over-damped Langevin diffusion and underdamped Langevin diffusion.

Proposition 6.

1. *The stationary measure of the Langevin diffusion in Equation (4) is $\pi(x) := e^{-f(x)}/Z$, where $Z := \int_{\mathbb{R}^p} e^{-f(x)}\, dx$ is the normalization constant.*
2. *The stationary measure of the underdamped Langevin diffusion in Equation (8) is $\pi(x, v) := e^{-f(x) - v^2/(2\gamma)}/Z'$, where $Z' := \int_{\mathbb{R}^p \times \mathbb{R}^p} e^{-f(x) - \|v\|_2^2/(2\gamma)}\, dx\, dv$ is the normalization constant. Moreover, the marginalization to x is $\pi(x) \propto e^{-f(x)}$*

Proof of 1. For test functions ϕ, ψ (in this case, $\mathcal{C}^2$ and vanishing at infinities), we have

$$\langle A\phi, \psi \rangle = \int (A\phi)\psi = \int (\Delta\phi - \langle \nabla f, \nabla \phi \rangle)\psi = \int (\Delta\phi)\psi - \int \langle \nabla f, \nabla\phi \rangle \psi.$$

Now, we use integration by parts twice and once on the first and second terms of the right-hand side, respectively, while using the vanishing at infinity property of test functions to get

$$\langle A\phi, \psi \rangle = \int \phi(\underbrace{\Delta\psi + \nabla \cdot (\psi \nabla f)}_{=A^*}) = \int \phi(\nabla \cdot (\psi(\nabla \log \psi + \nabla f)))$$

We could also directly plug in the coefficients of Equation (4) to Fact 3 and get the same result. Regardless, we use condition 2 of Proposition 5 and solve for π in

$$0 = \nabla \cdot (\pi(\nabla \log \pi + \nabla f))$$

to get $\pi = \exp(-f + C)$ for constant C that can be chosen to absorb the normalization constant. ■

Proof of 2. Let $\pi(x, v)$ denote the joint density of the stationary distribution.

From Fact 4, we know that the stationary distribution π must satisfy:

$$A^*\pi = 0,$$

where A^* is the adjoint of the generator of the diffusion process.

Using Equation (2) with drift coefficient $b(x, v) = (v, -\nabla f(x) - \gamma v)$ and constant diffusion coefficient

$$\sigma(x, v) = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2\gamma} I_p \end{bmatrix} \in \mathbb{R}^{2p \times 2p},$$

we must have

$$A^*\pi = -\sum_{i=1}^{p} \frac{\partial(\pi v_i)}{\partial x_i} + \sum_{i=1}^{p} \frac{\partial(\pi(\frac{\partial f}{\partial x_i} + \gamma v_i))}{\partial v_i} + \gamma \sum_{i=1}^{p} \frac{\partial^2 \pi}{\partial v_i^2} = 0$$

in order to satisfy the stationarity condition. Now, we ansatz $\pi(x, v) = \pi_X(x)\pi_V(v) = C \exp(-f(x) - \|v\|_2^2/(2\gamma))$. Plugging this in, we can verify that the stationarity condition indeed holds. Now choose C to normalize our density. ■

2.3 Basic Optimal Transport Theory

Here, we introduce the basics of optimal transport theory for probability measures over Euclidean space. In fact, we will only need the definition of the Wasserstein metric. We first define a **coupling** between two measures over $\mathbb{R}^p$.

Definition 7. *Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^p)$, the set of couplings between μ and ν is defined as*

$$\Pi(\mu, \nu) \coloneqq \{\gamma \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p) \mid \gamma(B \times \mathbb{R}^p) = \mu(B),\, \gamma(\mathbb{R}^p \times B) = \nu(B)\ \forall B \in \mathcal{B}(\mathbb{R}^p)\},$$

that is, the set of measures on $\mathbb{R}^p \times \mathbb{R}^p$ such that the first marginal agrees with μ and the second marginal agrees with ν.

Now, for some cost function $c : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty)$, the c-optimal transport cost between μ and ν is $\inf_{\gamma \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}[c(X, Y)]$. However, for our purposes, we are only concerned with the case when $c(x, y) = \|x - y\|_2^2$. This is because – as will be seen in the next section – the optimal transport cost paired with the space of distributions with finite second moment forms a nice space with 'good-enough' geometry that we can start solving an ODE on it.

Definition 8. *For $\mu, \nu \in \mathcal{P}(\mathbb{R}^p)$, the **2-Wasserstein metric** between μ and ν is*

$$W_2(\mu, \nu) \coloneqq \inf_{\gamma \in \Pi(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\gamma} \|X - Y\|_2^2\right)^{\frac{1}{2}}.$$

Now, we need to show that there actually exists a coupling such that the infimum in the above definition is attained. Recall that if a function $f : \mathcal{T} \to \bar{\mathbb{R}}$, where $\mathcal{T}$ is any topological space, is lower semicontinuous and $K \subseteq \mathcal{T}$ compact, then $\inf_{x \in K} f(x)$ exists and is attained by some $x \in K$. We refer the reader to Villani et al. [2009][Theorem 4.1, Lemma 4.3, Lemma 4.4] or Thorpe [2019][Proposition 1.5] for the full proof. This is to avoid having to talk about weak-* topologies.

Fact 9. *$W_2(\mu, \nu)$ always exists for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^p)$ and is attained by some $\gamma \in \Pi(\mu, \nu)$.*

There is one more thing to say, which is that the original formulation of optimal transport, known as Monge's problem, had a tighter restriction (our relaxed notion of optimal transport is called the Kantorovich problem or the Kantorovich relaxation). Instead of optimizing over all couplings, we instead need to optimize over all transport plans $T : \mathbb{R}^p \to \mathbb{R}^p$ such that $(X, T(X)) \sim \mu \otimes \nu$. The celebrated Brenier's theorem, in part, provides conditions on when Monge's problem coincides with the Kantorovich problem. This is also what motivates us to switch from using $\mathcal{P}_2(\mathbb{R}^p)$ to $\mathcal{P}_{2,ac}(\mathbb{R}^p)$ in the next section.

Theorem 10 (Brenier [1991])**.** *Suppose $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$, then the optimal transport plans $T_{\mu,\nu}$ $T_{\nu,\mu}$, satisfying the transport from μ to ν and ν to μ, respectively, exists and is unique.*

Before ending the section, it should be noted that this is not even scratching the surface of optimal transport theory: basic notions such as the duality formula are fundamental but will not be covered.

2.4 Langevin Diffusion in $\mathcal{P}_{2,ac}(\mathbb{R}^p)$: The JKO Scheme

In this section, we develop the JKO scheme for the Langevin diffusion [Jordan et al., 1998]. The JKO scheme allows us to easily couple the analysis between the discrete time Langevin Monte Carlo with the continuous time Langevin diffusion. Instead of analyzing the random dynamic (X_t) over $\mathbb{R}^p$ for minimizing f, we can analyze the deterministic gradient flow of $\text{law}(X_t)$ over $(\mathcal{P}_{2,ac}(\mathbb{R}^p), W_2)$ for minimizing $D_{KL}(\cdot \| e^{-f}/Z)$.

We first show that the Wasserstein space is a metric space. In order to do so, we require a technical lemma from Villani et al. [2009], itself requiring another theorem (disintegration of measures or simply disintegration theorem). We will simply state it.

Fact 11 (Gluing Lemma)**.** *If $\gamma_1, \gamma_2 \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p)$ are such that γ_1 has the same marginal distribution in its second argument as γ_2 in its first argument, then there exists $\gamma \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p)$ such that the first two marginals of γ is γ_1 and the last two marginals of γ is γ_2.*

Proposition 12. *The space $(\mathcal{P}_{2,ac}(\mathbb{R}^p), W_2)$ is a metric space.*

Proof. Symmetry is simple, and non-negativity follows directly from the non-negativity of the 2-norm. To show $W_2(\mu, \nu) = 0 \iff \mu = \nu$, first note that if $\mu = \nu$, then choosing $\gamma(x, y) = \delta_x(y)\mu(x)$ upper bounds $W_2^2(\mu, \nu)$ by 0. On the other hand, if $W_2^2(\mu, \nu) = 0$, then we must have $X = Y$ γ-a.e., where γ is the optimal coupling. So for all test functions $f : \mathbb{R}^p \to \mathbb{R}$, we have

$$\int_{\mathbb{R}^p} f \, d\mu = \int_{\mathbb{R}^p \times \mathbb{R}^p} f(x) \, d\gamma(x, y) = \int_{\mathbb{R}^p \times \mathbb{R}^p} f(y) \, d\gamma(x, y) = \int_{\mathbb{R}^p} f \, d\nu,$$

and so $\mu = \nu$.

For the triangle inequality, we make use of the gluing lemma. First let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R}^p)$ and suppose $\gamma_1 \in \Pi(\mu_1, \mu_2), \gamma_2 \in \Pi(\mu_2, \mu_3)$ be the optimal couplings. Now, by the gluing lemma, we can find $\gamma \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p)$ such that $T^1_\# \gamma = \gamma_1$ and $T^2_\# \gamma = \gamma_2$, where $T^1 : (x, y, z) \mapsto (x, y), T^2 : (x, y, z) \mapsto (y, z)$

are projection mappings from $(\mathbb{R}^p)^3$ to $(\mathbb{R}^p)^2$. Also define $T : (x, y, z) \mapsto (x, z)$ and $\pi \coloneqq T_{\#}\gamma$. Now, for $B \in \mathcal{B}(\mathbb{R}^p)$, we have

$$\begin{aligned}\pi(B \times \mathbb{R}^p) = T_{\#}\gamma(B \times \mathbb{R}^p) &= \gamma\{(x, y, z) \in (\mathbb{R}^p)^3 \mid T(x, y, z) \in B \times \mathbb{R}^p\} \\ &= \gamma\{(x, y, z) \in (\mathbb{R}^p)^3 \mid x \in B\} = \gamma(B \times \mathbb{R}^p \times \mathbb{R}^p) = \mu_1(B)\end{aligned}$$

, which also gives $\pi(\mathbb{R}^p \times B) = \mu_3(B)$. Hence $\pi \in \Pi(\mu_1, \mu_3)$. Now, we can easily use π as an upper bound on the infimum part of $W_2(\mu_1, \mu_3)$ as well as the triangle inequality (followed by Jensen's) of $\| \cdot \|_2$:

$$\begin{aligned}W_2(\mu_1, \mu_3) &\leq \left(\int_{\mathbb{R}^p \times \mathbb{R}^p} \|x - z\|_2^2 \, d\pi(x, z)\right)^{\frac{1}{2}} \\ &= \left(\int_{(\mathbb{R}^p)^3} \|x - z\|_2^2 \, d\gamma(x, y, z)\right)^{\frac{1}{2}} \\ &\leq \left(\int_{\mathbb{R}^p \times \mathbb{R}^p} \|x - y\|_2^2 \, d\gamma_1(x, y)\right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^p \times \mathbb{R}^p} \|y - z\|_2^2 \, d\gamma_2(y, z)\right)^{\frac{1}{2}} = W_2(\mu_1, \mu_2) + W_2(\mu_2, \mu_3).\end{aligned}$$

■

Not only do we have a metric space here, it is also complete and separable. The 2-Wasserstein distance also simultaneously metrizes weak convergence and $\mathcal{L}^2$-convergence. That is, for $(\mu_k)_{k\in\mathbb{N}} \subseteq \mathcal{P}_{2,ac}(\mathbb{R}^p)$, $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$, $W_2(\mu_k, \mu) \overset{k\to\infty}{\to} 0$ if and only if $\int_{\mathbb{R}^p} \| \cdot \|_2^2 \, d\mu_k \overset{k\to\infty}{\to} \int_{\mathbb{R}^p} \| \cdot \|_2^2 \, d\mu$ ***and*** $\mu_k \overset{\text{weak}}{\to} \mu$ [Thorpe, 2019][Theorem 5.8].

Remark. *Unfortunately, we must wave our hands vigorously throughout the rest of the section to rid the background regarding the theory of Riemmanian and smooth manifolds. For great coverage of this theory, please refer to Dafermos [2012]. For a rigorous coverage on the geometry of Wasserstein space, please refer to Ambrosio et al. [2005]. I apologize profusely, let us proceed.*

Furthermore, we can do basic geometry in Wasserstein space such as talk about gradient flows by defining a **Riemannian metric**: at every 'point' $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$ we can define a tangent space $T_\mu\mathcal{P}_{2,ac}(\mathbb{R}^p)$, itself equipped with an inner product $\langle \cdot, \cdot \rangle_\mu$ such that it is 'smooth when varying μ'. The Riemannian metric then induces a metric on the original space which surprisingly matches the 2-Wasserstein distance. The implication and takeaway should be that, for a 'smooth' curve in Wasserstein space, i.e. $t \mapsto \mu_t \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$, at each time t we can define a 'tangent vector' $v_t \in T_{\mu_t}\mathcal{P}_{2,ac}(\mathbb{R}^p)$.

So, given a functional $F : \mathcal{P}_{2,ac}(\mathbb{R}^p) \to [0, \infty]$ that we wish to minimize, a good start is to compute the 'gradient' of F at some point μ, denoted $\nabla_{W_2}F(\mu)$. It is an element of the tangent space at μ, $T_\mu\mathcal{P}_{2,ac}(\mathbb{R}^p)$ such that for every curve $t \mapsto \mu_t$ with $\mu_0 = \mu$, the following is satisfied:

$$\partial_t F(\mu_t)|_{t=0} = \langle \nabla_{W_2}F(\mu), v_0 \rangle_\mu,$$

where v_0 is the tangent vector to the curve at time 0. This leads us to define the **first variation** of a functional at some point.

Definition 13. *Given a functional $F : \mathcal{P}_{2,ac}(\mathbb{R}^p) \to [0, \infty]$ and some point $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$, the first variation of F at μ is the function $\delta F(\mu) : \mathbb{R}^p \to \mathbb{R}$, up to an additive constant, such that*

$$\partial_t F(\mu_t)|_{t=0} = \int_{\mathbb{R}^p} \delta F(\mu)(x) \, d(\partial_t \mu_t|_{t=0})(x)$$

for all smooth curves (μ_t) started at μ.

Before we state the main theorem of the section, we need one more result from Chewi [2024][Theorem 1.3.17].

Fact 14. *Let $t \mapsto v_t$ be a family of vector fields $(v_t : \mathbb{R}^p \to \mathbb{R}^p)_{t\geq 0}$. Suppose a random process satisfies the 'SDE' $dX_t = v_t(X_t)\, dt$. Let μ_t be the probability density of X_t at each time t. Then μ_t follows the* ***continuity equation****, written as*

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0$$

Finally, this allows us to characterize the Wasserstein gradient of a functional at μ.

Lemma 15. *For a functional $F : \mathcal{P}_{2,ac}(\mathbb{R}^p) \to [0, \infty]$ and a point $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$, we have*

$$\nabla_{W_2} F(\mu) = \nabla(\delta F(\mu)).$$

Proof. Let $t \mapsto \mu_t$ be a smooth curve of measures in $\mathcal{P}_{2,ac}(\mathbb{R}^p)$ started at μ. If $v_t \in T_{\mu_t}\mathcal{P}_{2,ac}(\mathbb{R}^p)$ and the continuity equation holds, then v_t is the tangent vector to the curve $t \mapsto \mu_t$ at time t. Therefore,

$$\begin{aligned}\partial_t F(\mu_t)|_{t=0} &= \int_{\mathbb{R}^p} \delta F(\mu)(x)\, d(\partial_t \mu_t|_{t=0})(x) \\ &= -\int_{\mathbb{R}^p} \delta F(\mu)(x) \nabla \cdot ((v_0\mu)(x)) = \int_{\mathbb{R}^p} \langle \nabla \delta F(\mu)(x), v_0(x)\rangle\, d\mu(x).\end{aligned}$$

Now, we need to show $\nabla \delta F(\mu) \in T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^p)$. Although it was not defined in the hand-waving section, to complete the proof it suffices to know that

$$T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^p) = \mathrm{close}_{\mathcal{L}^2(\mu)}\{\nabla \phi \mid \phi \in \mathcal{C}_c^\infty(\mathbb{R}^p)\},$$

where $\mathrm{close}_{\mathcal{L}^2(\mu)}$ is the closure of the set with respect to μ and the squared norm $\int (\cdot)^2\, d\mu$, recall the construction of the classical Lebesgue space $\mathcal{L}^2(dx)$. So by definition we arrive at the result. ■

Finally, we have a notion of the **Wasserstein gradient flow** for a functional F. By intuition, it should be the smooth curve of measures $t \mapsto \mu_t$ such that at each time t, its tangent vector is $v_t = -\nabla_{W_2} F(\mu_t)$. We get a PDE over $\mathbb{R}^p$ by using to continuity equation in Fact 14 then substituting in Lemma 15:

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla \delta F(\mu_t)). \tag{3}$$

Here is the most beautiful part, where we see that the Fokker-Planck equation for Langevin diffusions matches the continuity equation for the Wasserstein gradient flow of the KL divergence against the stationary measure $\pi = e^{-f}$.

Corollary 16. *Consider a Langevin diffusion with potential $f : \mathbb{R}^p \to \mathbb{R}$ and the functional $F(\cdot) = D_{KL}(\cdot\|\pi)$, where $\pi \propto \exp(-f)$. The law of the Langevin diffusion $t \mapsto \mu_t$ is the Wasserstein gradient flow of $D_{KL}(\cdot\|\pi)$.*

Proof. Notice that

$$F(\cdot) = \int \mu \log \frac{\mu}{\pi} = \int \mu \log \mu - \int \mu \log(\exp(-f)) = \int f\, d\mu + \int \log \mu\, d\mu.$$

Now, we wish to compute the first variation at some measure $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$. Let $t \mapsto \mu_t$ be a sufficiently smooth curve, we have

$$\begin{aligned}\partial_t F(\mu_t)|_{t=0} &= \partial_t \left(\int f(x)\mu_t(dx) + \int \log(\mu_t(x))\mu_t(dx) \right)\Bigg|_{t=0} \\ &= \int f(\partial_t \mu_t|_{t=0}) + \int (\partial_t \mu_t)|_{t=0} \log \mu + \underbrace{\int \mu_t \left(\frac{1}{\mu_t} (\partial_t \mu_t|_{t=0}) \right)}_{=0} \\ &= \int (f(x) + \log \mu(x))(\partial_t \mu_t|_{t=0})(dx).\end{aligned}$$

Remark. *Note that this calculation also shows why first variations are defined up to an additive constant as the rightmost term on the right-hand side goes to 0. We could technically multiply it by any constant and absorb it into the integral.*

So we have computed that $\delta F(\mu) = f + \log \mu$, therefore we have

$$\begin{aligned}\nabla_{W_2} F(\mu) &= \nabla \delta F(\mu) = \nabla f + \nabla \log \mu \\ &= \nabla(-\log \pi) + \nabla \log \mu = \nabla \log \frac{\mu}{\pi}\end{aligned}$$

by our characterization of the Wasserstein gradient in Lemma 15. Plugging into the continuity equation shows that the Wasserstein gradient flow of $D_{KL}(\cdot \| \pi)$ satisfies

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla \log \frac{\mu_t}{\pi}) = A^* \mu_t,$$

where A^* is the adjoint of the Langevin diffusion operator. ■

This is a beautiful result that requires its own dedicated analysis. Here, however, we hope to have provided a small peek into what Jordan et al. [1998] have kick-started. Before we end off, we need one more lemma, to be used later, that transfers the α-strong convexity of the potential function to the α-strong convexity of $D_{KL}(\cdot \| \pi)$ in a geodesic sense.

Fact 17 (Chewi [2024] Definition 1.3.25, Theorem 1.4.4, Theorem 1.4.5)**.** *If $\pi \propto \exp(-f)$ is α-strongly log-concave, then for all $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$, we have that*

$$D_{KL}(\nu \| \pi) \geq D_{KL}(\mu \| \pi) + \langle \nabla_{W_2} F(\mu), T_{\mu,\nu} - \mathrm{id} \rangle_\mu + \frac{\alpha}{2} W_2^2(\mu, \nu),$$

where $F(\cdot) = D_{KL}(\cdot \| \pi)$, $\langle \cdot, \cdot \rangle_\mu$ is the Riemannian metric for $(\mathcal{P}_{2,ac}(\mathbb{R}^p), W_2)$ at μ, and $T_{\mu,\nu}$ is the optimal transport plan from μ to ν.

3 Langevin-type Sampling

Now, we begin our analysis of a small subset of sampling algorithms: Langevin-type samplers.

3.1 Langevin Monte Carlo

The principal Langevin-type sampling algorithm is **Langevin Monte Carlo** (LMC): an Euler discretization of the Langevin diffusion:

$$d\theta_t = -\nabla f(\theta_t)\, dt + \sqrt{2}\, dW_t, \tag{4}$$

where $f \in \mathcal{C}^1(\mathbb{R}^p; \mathbb{R})$ is the 'potential' function, and $W : [0, \infty) \times \Omega \to \mathbb{R}^p$ is the standard p-dimensional Brownian motion. That is, for an initial value $\vartheta_0 \in \mathbb{R}^p$ and 'step size' $h > 0$, LMC has updates given by

$$\vartheta_{h(k+1)} \coloneqq \vartheta_{hk} - h \nabla f(\vartheta_{hk}) + \sqrt{2} \xi_{hk} \tag{5}$$

for each $k \in \mathbb{N}$, where $(\xi_{hk})_{k \geq 0} \overset{i.i.d.}{\sim} \mathcal{N}(0, hI_p)$.

Note that the law of the iterates match the law of the solution of the following SDE at times kh $\forall k$:

$$d\vartheta_t = b_t(\vartheta)\, dt + \sqrt{2}\, dW_t, \tag{6}$$

where $b_\cdot(\cdot) : (\theta, t) \mapsto -\sum_{k=0}^{\infty} \nabla f(\theta_{kh}) \mathbb{1}_{[kh,(k+1)h}(t)$. Also notice that both SDEs are driven by the same Brownian motion. From here on out, we take $(\theta_t^x)_{t\geq 0}$ and $(\vartheta_t^x)_{t\geq 0}$ to be the (pathwise unique, strong) solutions to Equation (4) and Equation (6) started at $x \in \mathbb{R}^p$, respectively (refer to first paragraph of Section 2.2 for why).

We analyze the performance of Equation (5) when the potential is strongly convex and smooth. That is, we further assume that $f \in \mathcal{C}^2(\mathbb{R}^p; \mathbb{R})$ and $\exists \alpha, \beta \in (0, \infty)$ such that

$$0 \prec \alpha I_p \preceq \nabla^2 f(\theta) \preceq \beta I_p. \tag{7}$$

Now, we move towards a non-asymptotic bound for convergence of the law of the LMC iterates against the invariant measure in the total variation norm, as seen in Theorem 2 of Dalalyan [2017].

First, we state – without proof – two useful facts about Langevin diffusions.

Fact 18 (Roberts and Tweedie [1996], Theorem 2.1). *If $f \in \mathcal{C}^1(\mathbb{R}^p)$, then solutions to Equation (4) are non-explosive, i.e. $\sup_{0\leq t\leq T} \|\theta_t\|_2 < \infty$ a.s. for all $T < \infty$.*

Since we always assume f is at least once continuously differentiable to satisfy the smoothness condition, non-explosivity is assumed henceforth. Now, we briefly define **reversibility** of stochastic processes, which holds for Langevin diffusions and makes the succeeding proofs easier to work with.

Definition 19 (Reversible Stochastic Process). *A stochastic process $(X_t)_{t\geq 0}$ is reversible (or time-reversible) if for all sets of finite time increments $0 \leq t_1 < \ldots < t_n < \infty$, $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}^p)$, $n \in \mathbb{N}$, and for all $\tau \in (t_n, \infty)$, we have*

$$\mathbb{P}(X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n) = \mathbb{P}(X_{\tau - t_1} \in B_1, \ldots, X_{\tau - t_n} \in B_n).$$

Fact 20 (Kolmogorov's Characterization of Reversibility). *If b is Lipschitz, and $\theta_0 \in \mathcal{L}^2(\mathbb{P})$. Then the solution of $dX_t = b(X_t)\,dt + dW_t$ is reversible.*

Note that not all of our sampling algorithms will be reversible. In fact, non-reversible SDEs, such as the Underdamped Langevin dynamic in Section 3.2, when discretized and used properly, can often lead to improvements in sampling complexity [Wu and Robert, 2019].

Proposition 21. *Assume Equation (7) holds, then*

$$\mathbb{E}[f(\vartheta_{(k+1)h})] \leq \mathbb{E}[f(\vartheta_{kh})] - \frac{1}{2}h(2 - \beta h)\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] + \beta h p.$$

Proof. By the Descent Lemma [Fawzi, 2024], then substituting the definition of $\vartheta_{(k+1)h}$ from Equation (5), we have

$$\begin{aligned} f(\vartheta_{(k+1)h}) &\leq f(\vartheta_{kh}) + \langle \nabla f(\vartheta_{kh}), \vartheta_{(k+1)h} - \vartheta_{kh}\rangle + \frac{\beta}{2}\|\vartheta_{(k+1)h} - \vartheta_{kh}\|_2^2 \\ &= f(\vartheta_{kh}) + \langle \nabla f(\vartheta_{kh}), -h\nabla f(\vartheta_{kh}) + \sqrt{2}\xi_{kh}\rangle + \frac{\beta}{2}\| - h\nabla f(\vartheta_{kh}) + \sqrt{2}\xi_{kh}\|_2^2 \\ &= f(\vartheta_{kh}) - h\|\nabla f(\vartheta_{kh})\|_2^2 + \sqrt{2}\langle \nabla f(\vartheta_{kh}), \xi_{kh}\rangle + \frac{\beta}{2}\|h\nabla f(\vartheta_{kh}) - \sqrt{2}\xi_{kh}\|_2^2 \end{aligned}$$

Taking expectations, expanding the rightmost term, and recalling $\mathbb{E}[\xi_{kh}] = 0$, $\mathbb{E}[\|\xi_{kh}\|_2^2] = ph$, and $\xi_{kh} \perp \nabla f(\vartheta_{kh})$ yields the result:

$$\begin{aligned} \mathbb{E}[f(\vartheta_{(k+1)h})] &\leq \mathbb{E}[f(\vartheta_{kh})] - h\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] + \frac{\beta}{2}\mathbb{E}[h^2\|\nabla f(\vartheta_{kh})\|_2^2 + 2\|\xi_{kh}\|_2^2 - 2\sqrt{2}h\langle \nabla f(\vartheta_{kh}), \xi_{kh}\rangle] \\ &= \mathbb{E}[f(\vartheta_{kh})] + (\frac{\beta h^2}{2} - h)\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] + \beta p h \end{aligned}$$

■

A corollary of the above proposition will be used to prove our first lemma.

Corollary 22. *Assume Equation* (7). *Let* $\theta^* = \arg\min_{x\in\mathbb{R}^p} f(x)$*,* $f^* = f(\theta^*)$*,* $\vartheta_0 = \theta_0$*,* $h \le (C\beta)^{-1}$ *with* $C \ge 1$*, and let* $K \ge 1$*. Then*

$$h\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] \le \frac{C}{2C-1}\beta\mathbb{E}[\|\theta_0-\theta^*\|_2^2] + \frac{2C}{2C-1}\beta Khp.$$

Proof. First note $h \le 1/(C\beta)$ if and only if $2-\beta h \ge (2C-1)/C$. Now, by Proposition 21, we have

$$\frac{h(2C-1)}{2C}\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] \le \mathbb{E}[f(\vartheta_{kh}) - f(\vartheta_{(k+1)h})] + \beta hp,\ \forall k \in \mathbb{N}.$$

Summing both sides from $k = 0,\dots,K-1$, the first term on the right hand side telescopes, we get

$$\begin{aligned}\frac{h(2C-1)}{2C}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] &\le \mathbb{E}[f(\vartheta_0) - f(\vartheta_{Kh})] + K\beta hp\\ &\le \mathbb{E}[f(\vartheta_0) - f^*] + K\beta hp,\end{aligned}$$

as $f^* \le f(\theta)\ \forall\theta \in \mathbb{R}^p$. Rearranging:

$$h\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] \le \frac{2C}{2C-1}\mathbb{E}[f(\vartheta_0) - f^*] + \frac{2C}{2C-1}K\beta hp.$$

As f is β-smooth, the descent lemma [Fawzi, 2024] applied to (ϑ_0, θ^*) gives

$$f(\vartheta_0) - f^* \le \langle\nabla f(\theta^*), \vartheta_0 - \theta^*\rangle + \frac{\beta}{2}\|\vartheta_0 - \theta^*\|_2^2,$$

as $\nabla f(\theta^*) = 0$. Taking expectations gives $2\mathbb{E}[f(\vartheta_0) - f^*] \le \beta\mathbb{E}[\|\vartheta_0 - \theta^*\|_2^2]$. Since $\vartheta_0 = \theta_0$, we are done:

$$h\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2] \le \frac{\beta C}{2C-1}\mathbb{E}[\|\theta_0-\theta^*\|_2^2] + \frac{2C}{2C-1}K\beta hp.$$

■

Now, due to Dalalyan and Tsybakov [2012], we show an explicit formula for computing the KL divergence between the continuous and 'discretized' Langevin diffusions, assuming an affine condition on the drift coefficient.

Proposition 23. *Suppose, for some* $B > 0$*, we have* $\|b_t(X)\|_2 \le B(1 + \|X\|_\infty)$ *for every* $t \in [0, Kh]$ *and every continuous process* X*. Assume* f *is* β*-smooth and let* θ^* *be* ***any*** *stationary point, i.e.* $\nabla f(\theta^*) = 0$*. Then* $\mathbb{P}_{x,Kh} \sim \widetilde{\mathbb{P}}_{x,Kh}$ *and*

$$D_{KL}(\mathbb{P}_{x,Kh}\|\widetilde{\mathbb{P}}_{x,Kh}) := \int_{\mathbb{R}^p}\log\left(\frac{d\mathbb{P}_{x,Kh}}{d\widetilde{\mathbb{P}}_{x,Kh}}\right)d\mathbb{P}_{x,Kh} = \frac{1}{4}\int_0^{Kh}\mathbb{E}[\|\nabla f(\vartheta_t) + b_t(\vartheta)\|_2^2]\,dt,$$

where $\mathbb{P}_{x,Kh}, \widetilde{\mathbb{P}}_{x,Kh}$ *is the law of* $(\theta_t)_{0\le t\le Kh}, (\vartheta_t)_{0\le t\le Kh}$ *started at* x*, respectively.*

Proof. First let $T \coloneqq Kh$, now recall one form of Girsanov's theorem [Girsanov, 1960][Theorem 1]: For a $\mathbb{P}$-Brownian motion $(W_t)_{0\leq t\leq T}$, $T > 0$, and previsible process γ_t satisfying Novikov's condition

$$\mathbb{E}_{\mathbb{P}}\big[\exp(\frac{1}{2}\int_0^T \gamma_t^2\, dt)\big] < \infty,$$

there exists a measure $\mathbb{Q} \sim \mathbb{P}$ such that $(\widetilde{W}_t)_{0\leq t\leq T}$, defined as $\widetilde{W}_t \coloneqq W_t + \int_0^t \gamma_s\, ds$, is a $\mathbb{Q}$-Brownian motion. Further, the Radon-Nikodym derivative on the measurable space of continuous functions $\mathcal{C}([0,T];\mathbb{R})$ [Üstünel, 2010][Chapter 1.1] is given by

$$\frac{d\mathbb{Q}}{d\mathbb{P}}(\gamma) = \exp\big(-\int_0^T \gamma_s\, dW_s - \frac{1}{2}\int_0^T \gamma_s^2\, ds\big).$$

Choosing $\gamma_s \coloneqq \frac{1}{\sqrt{2}}(-b_s(\vartheta) - \nabla f(\vartheta_s))$, we easily have previsibility. Now we quickly verify Novikov's condition:

$$\begin{aligned}\mathbb{E}\big[\exp(\frac{1}{4}\int_0^{Kh} \|-b_s(\vartheta)-\nabla f(\vartheta_s)\|_2^2\, ds)\big] &= \mathbb{E}\big[\exp(\frac{1}{4}\int_0^{Kh} (\|b_s(\vartheta)\|_2^2 + \|\nabla f(\vartheta_s)\|_2^2 + 2\langle b_s(\vartheta), \nabla f(\vartheta_s)\rangle)\, ds)\big] \\ &\leq \mathbb{E}\big[\exp(\frac{1}{4}\int_0^{Kh} (B^2(1+\|\vartheta\|_\infty)^2 \\ &+ \beta^2\|\vartheta_s - \theta^*\|_2^2 + B\beta\|\vartheta_s - \theta^*\|_2(1+\|\vartheta\|_\infty))\, ds)\big] < \infty,\end{aligned}$$

where the first inequality is due to $\|b_s(\vartheta)\|_2^2 \leq B^2(1+\|\vartheta\|_\infty)^2$, $\|\nabla f(\vartheta_s)\|_2^2 = \|\nabla f(\vartheta_s) - \nabla f(\theta^*)\|_2^2 \leq \beta^2\|\vartheta_s - \theta^*\|_2^2$, and $\langle b_s(\vartheta), \nabla f(\vartheta_{kh})\rangle \leq B\beta\|\vartheta_s - \theta^*\|_2(1+\|\vartheta\|_\infty)$ by Cauchy-Schwarz. The second inequality is due to (ϑ_s) being non-explosive and continuous. Therefore there exists a probability measure $\mathbb{Q} \sim \mathbb{P}_{x,Kh}$ such that $W_t + \int_0^{Kh} \frac{1}{\sqrt{2}}(-b_s(\vartheta) - \nabla f(\vartheta_s))\, ds$ is a $\mathbb{Q}$-Brownian motion. By rearranging and putting into differential form, we have

$$-\nabla f(\vartheta_s)\, dt + \sqrt{2}\, dW_t = b_t(\vartheta)\, dt + \sqrt{2}\, d\widetilde{W}_t,$$

i.e. $\mathbb{Q} = \widetilde{\mathbb{P}}_{x,Kh}$. So we have

$$\begin{aligned}-\log\big(\frac{d\widetilde{\mathbb{P}}_{x,Kh}}{d\mathbb{P}_{x,Kh}}(\vartheta)\big) &= -\frac{1}{\sqrt{2}}\int_0^{Kh} (-b_s(\vartheta) - \nabla f(\vartheta_s))\, dW_s + \frac{1}{4}\int_0^{Kh} \|-b_s(\vartheta) - \nabla f(\vartheta_s)\|_2^2\, ds \\ \implies D_{KL}(\mathbb{P}_{x,Kh}\|\widetilde{\mathbb{P}}_{x,Kh}) &= \mathbb{E}_{\vartheta\sim\mathbb{P}_{x,Kh}}\big[-\log\big(\frac{d\widetilde{\mathbb{P}}_{x,Kh}}{d\mathbb{P}_{x,Kh}}(\vartheta)\big)\big] \\ &= \frac{1}{4}\int_0^{Kh} \mathbb{E}[\|b_s(\vartheta) + \nabla f(\vartheta_s)\|_2^2]\, ds.\end{aligned}$$

■

Now, we are ready to prove the first useful lemma from Dalalyan [2017].

Lemma 24. *Let f be β-smooth and $\theta^* \in \mathbb{R}^p : \nabla f(\theta^*) = 0$ (not necessarily a global minimum). Fixing* law(θ_0) *such that $\mathbb{E}[\theta_0] = x$ and number of iterations $K \in \mathbb{N}$, if $h \leq 1/(C\beta)$ with $C \geq 1$, then we have*

$$D_{KL}(\mathbb{P}_{x,Kh}\|\widetilde{\mathbb{P}}_{x,Kh}) \leq \frac{\beta^3 h^2 C}{12(2C-1)}(\|x - \theta^*\|_2^2 + 2Khp) + \frac{1}{4}pK\beta^2h^2$$

Proof. By Proposition 23, we have

$$\begin{aligned}
D_{KL}(\mathbb{P}_{x,Kh}\|\widetilde{\mathbb{P}}_{x,Kh}) &= \frac{1}{4}\int_0^{Kh}\mathbb{E}[\|\nabla f(\vartheta_t)+b_t(\vartheta)\|_2^2]\,dt \\
&= \frac{1}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\mathbb{E}[\|\nabla f(\vartheta_t)-\nabla f(\vartheta_{kh})\|_2^2]\,dt \\
&\le \frac{\beta^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\underbrace{\mathbb{E}[\|\vartheta_t-\vartheta_{kh}\|_2^2]}_{(a)}\,dt,
\end{aligned}$$

where the inequality comes from the β-smoothness of f. Now, to compute (a), we have the following equality (in distribution):

$$\vartheta_t-\vartheta_{kh}\overset{d}{=}b_t(\vartheta)(t-kh)+\sqrt{2}\xi, \xi\sim\mathcal{N}(0,(t-kh)I_p),\, t\ge kh.$$

So we have

$$\begin{aligned}
\mathbb{E}[\|\vartheta_t-\vartheta_{kh}\|_2^2] &= \mathbb{E}[\|b_t(\vartheta)\|_2^2(t-kh)^2+2\xi^2+2\sqrt{2}\xi b_t(\vartheta)(t-kh)] \\
&= \mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2](t-kh)^2+2p(t-kh).
\end{aligned}$$

Substituting back in, we have

$$\begin{aligned}
D_{KL}(\mathbb{P}_{x,Kh}\|\widetilde{\mathbb{P}}_{x,Kh}) &\le \frac{\beta^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}(\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2](t-kh)^2+2p(t-kh))\,dt \\
&= \frac{\beta^2}{4}\sum_{k=0}^{K-1}(\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2]\frac{h^3}{3}+ph^2) \\
&= \frac{\beta^2h^3}{12}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_{kh})\|_2^2]+\frac{\beta^2Kph^2}{4} \\
&\le \frac{\beta^2h^2}{12}\big(\frac{2C}{2C-1}\beta\mathbb{E}[\|x-\theta^*\|_2^2]+\frac{2C}{2C-1}\beta Khp\big)+\frac{\beta^2Kph^2}{4},
\end{aligned}$$

where the last inequality is due to Corollary 22. ■

Now we prove the second important lemma.

Lemma 25. *Assume Equation (7), let $\mu_{h,x}$ be the probability density for ϑ_h started at x, i.e. $\mathcal{N}(x-h\nabla f(x),2hI_p)$. If $h\le 1/(2\beta)$, then*

$$\mathbb{E}_\pi\left[\frac{\mu_{h,x}(\vartheta)^2}{\pi(\vartheta)^2}\right]\le\exp\left(\frac{1}{2\alpha}\|\nabla f(x)\|_2^2-\frac{p}{2}\log(2h\alpha)\right).$$

Proof. Notice that

$$\begin{aligned}
\pi(\vartheta)^{-1} &= e^{f(\vartheta)}\int_{\mathbb{R}^p}e^{-f(z)}\,dz \\
&= e^{f(\vartheta)-f(x)}\int_{\mathbb{R}^p}e^{-(f(z)-f(x))}\,dz \\
&\le \exp\left(\nabla f(x)^\top(\vartheta-x)+\frac{\beta}{2}\|\vartheta-x\|_2^2\right)\int_{\mathbb{R}^p}\exp\left(-\nabla f(x)^\top(z-x)-\frac{\alpha}{2}\|z-x\|_2^2\right)\,dz \\
&= \exp\left(\nabla f(x)^\top(\vartheta-x)+\frac{\beta}{2}\|\vartheta-x\|_2^2\right)\underbrace{\int_{\mathbb{R}^p}\exp\left(-\nabla f(x)^\top z-\frac{\alpha}{2}\|z\|_2^2\right)\,dz}_{=:(a)}.
\end{aligned}$$

Let us focus on the integral term, write

$$\begin{aligned}(a) &= \int_{\mathbb{R}^p} \exp\left(-\frac{\alpha}{2}\left(\|z + \frac{1}{\alpha}\nabla f(x)\|_2^2 - \frac{1}{\alpha^2}\|\nabla f(x)\|_2^2\right)\right) dz \\ &= \exp\left(\frac{1}{2\alpha}\|\nabla f(x)\|_2^2\right)\int_{\mathbb{R}^p} \exp\left(-\frac{\alpha}{2}\|z\|_2^2\right) dz \\ &= \left(\frac{2\pi}{\alpha}\right)^{p/2} \exp\left(\frac{1}{2\alpha}\|\nabla f(x)\|_2^2\right).\end{aligned}$$

Therefore, we have

$$\pi(\vartheta)^{-1} = \left(\frac{2\pi}{\alpha}\right)^{p/2} \exp\left(\nabla f(x)^\top(\vartheta - x) + \frac{\beta}{2}\|\vartheta - x\|_2^2 + \frac{1}{2\alpha}\|\nabla f(x)\|_2^2\right).$$

Now, we have

$$\begin{aligned}&\mathbb{E}_\pi\left[\frac{\mu_{h,x}(\vartheta)^2}{\pi(\vartheta)^2}\right] = \frac{1}{(4\pi h)^p}\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2h}\|\vartheta - x + h\nabla f(x)\|_2^2\right)\frac{\pi(\vartheta)}{\pi^2(\vartheta)}\, d\vartheta \\ &\le \frac{1}{(4\pi h)^p}\left(\frac{2\pi}{\alpha}\right)^{p/2} \exp\left(\frac{1}{2\alpha}\|\nabla f(x)\|_2^2\right) \\ &\times \int_{\mathbb{R}^p} \underbrace{\exp\left(-\frac{1}{2h}\|\vartheta - x + h\nabla f(x)\|_2^2 + \nabla f(x)^\top(\vartheta - x) + \frac{\beta}{2}\|\vartheta - x\|_2^2\right)}_{=:(b)} d\vartheta.\end{aligned}$$

Where integrand can be simplified:

$$(b) = \exp\left(\left(\frac{\beta}{2} - \frac{1}{2h}\|\vartheta - x\|_2^2\right) - \frac{h}{2}\|\nabla f(x)\|_2^2\right) = \exp\left(-\frac{1-\beta h}{2h}\|\vartheta - x\|_2^2\right)\exp\left(-\frac{h}{2}\|\nabla f(x)\|_2^2\right).$$

So

$$\begin{aligned}\mathbb{E}_\pi\left[\frac{\mu_{h,x}(\vartheta)^2}{\pi(\vartheta)^2}\right] &\le \frac{1}{(4\pi h)^p}\left(\frac{2\pi}{\alpha}\right)^{p/2} \exp\left(\frac{1-h\alpha}{2\alpha}\|\nabla f(x)\|_2^2\right)\int_{\mathbb{R}^p} \exp\left(-\frac{1-\beta h}{2h}\|\vartheta - x\|_2^2\right) d\vartheta \\ &= \frac{1}{(4\pi h)^p}\left(\frac{2\pi}{\alpha}\right)^{p/2}\left(\frac{2\pi h}{1-\beta h}\right)^{p/2} \exp\left(\frac{1-h\alpha}{2\alpha}\|\nabla f(x)\|_2^2\right) \\ &= \exp\left(-\frac{p}{2}\log\left(4h\alpha(1-\beta h)\right)\right)\exp\left(\frac{1-h\alpha}{2\alpha}\|\nabla f(x)\|_2^2\right) \\ &\le \exp\left(\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 - \frac{p}{2}\log(2h\alpha) - \frac{h}{2}\|\nabla f(x)\|_2^2\right) \\ &\le \exp\left(\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 - \frac{p}{2}\log(2h\alpha)\right),\end{aligned}$$

where the second inequality comes from $h \le 1/(2\beta) \implies (1-\beta h) \le 1/2$, and the last inequality comes from $(h/2)\|\nabla f(x)\|_2^2 \ge 0$. ■

Now, for the last lemma, we require the notion of geometric (or exponential) ergodicity for Markov processes. Here, we only need to define a slightly weaker notion: $\mathcal{L}^2$-geometric ergodicity.

Definition 26. *A Markov process with transition kernels $(P^t)_{t\ge 0}$ and invariant measure π is* ***geometrically ergodic with respect to*** $\mathcal{L}^2(\pi)$*, or* $\mathcal{L}^2(\pi)$***-geometrically ergodic****, if there is $\gamma > 0$ such that for each $\varphi \in \mathcal{L}^2(\pi)$, we have*

$$\|P^t\varphi - \pi(\varphi)\|_{\mathcal{L}^2(\pi)}^2 \le \pi(\varphi^2)e^{-\gamma t}$$

Recalling Markov chains over discrete space, we say a Markov chain is **irreducible** if any state can be reached from any other state in finite time with positive probability. Here, we extend this notion, which is useful for showing when $\mathcal{L}^2(\pi)$-geometric ergodicity is equivalent to the (stronger, usual) notion of π-a.e. geometric ergodicity.

Definition 27. *A Markov process with transition kernels $(P^t)_{t\geq 0}$ and invariant measure π is **ϕ-irreducible** (or **irreducible**) if for the (or there exists a) positive measure ϕ over $\mathcal{B}(\mathbb{R}^p)$ such that for every $x \in \mathbb{R}^p$ and $B \in \mathcal{B}(\mathbb{R}^p)$ with $\phi(B) > 0$, there exists $t > 0$ such that*

$$P^t(x, B) > 0.$$

This leads us to a strong result by Roberts and Tweedie [1996, Theorem 2.1] which gives conditions on when the Langevin diffusion is dx-irredible, where dx is the Lebesgue measure on $\mathbb{R}^p$.

Fact 28 (Roberts and Tweedie [1996] Theorem 2.1)**.** *Suppose $f \in \mathcal{C}^2(\mathbb{R}^p)$ and, for some $N, a, b < \infty$, we satisfy*

$$\langle \nabla f(x), x \rangle \leq a\|x\|_2^2 + b$$

whenever $\|x\|_2 > N$. Then the Langevin diffusion with drift coefficient $-\nabla f$ and constant diffusion coefficient is dx-irreducible.

We also define the spectral gap with respect to diffusions, it will be useful very shortly.

Definition 29. *For a (b, σ)-diffusion operator on $\mathbb{R}^p$ acting on $f \in \mathcal{C}_c^2(\mathbb{R}^p)$, if we have non-explosivity and other mild conditions, then the spectral gap of L is defined as*

$$\operatorname{gap}(L) := \inf \left\{ \int_{\mathbb{R}^p} \langle a(x)\nabla f(x), \nabla f(x)\rangle \, \pi(dx) : f \in \mathcal{D}, \int f\pi = 0, \int f^2\pi = 1 \right\},$$

where $\mathcal{D} = \{f + c \mid f \in \mathcal{C}_c^\infty(\mathbb{R}^p),\ c \in \mathbb{R}\}$ is the space of test functions (up to constants as the expression only involves gradients of f). We say that L has a spectral gap if $\operatorname{gap}(L) \geq C > 0$ for some $C > 0$.

An important result by Roberts and Tweedie [2001] allows us to reduce the task of showing $\mathcal{L}^2(\pi)$-geometric ergodicity to the existence of a spectral gap. This will be stated as fact.

Fact 30 (Roberts and Tweedie [2001] Eqn 4. Remark + Theorem 2)**.** *If a Markov process is reversible and ϕ-irreducible, then it is $\mathcal{L}^2(\pi)$-geometrically ergodic iff there exists a spectral gap.*

Now, thanks to Chen and Wang [1997], we have the following explicit lower bound for $\operatorname{gap}(L)$.

Fact 31 (Chen and Wang [1997] Remark 4.14)**.** *For the SDE $dX_t = b(X_t)\, dt + \sigma(X_t)\, dW_t$ on $\mathbb{R}^p$, suppose there exists $\bar{a} > 0$ such that $\langle a(x)u, u\rangle \leq \bar{a}\|u\|_2^2$ for all $x, u \in \mathbb{R}^p$, where $a(\cdot) := \sigma(\cdot)\sigma(\cdot)^\top$. Also let*

$$K := \sup_{x \neq y} \frac{\|\sigma(x) - \sigma(y)\|_{op}^2 + \langle b(x) - b(y), x - y\rangle}{\|x - y\|_2^2}.$$

Then the corresponding L-diffusion satisfies

$$gap(L) \geq -K\bar{a}^{-1} \inf_x \lambda_{\min}(a(x)).$$

Furthermore, if the right-hand side is strictly greater than 0, then the Markov process associated with L is also $\mathcal{L}^2(\pi)$-geometrically ergodic with exponential parameter equal to the right-hand side.

This sequence of facts tells us that, given the drift and diffusion coefficients, we can directly compute the lower bound for the spectral gap

Now, we prove our final lemma for this section.

Lemma 32. *Let* $\mathbb{P}_\theta^t(x, B) \coloneqq \mathbb{P}(\theta_t \in B \mid \theta_0 = x),\ B \in \mathcal{B}(\mathbb{R}^p)$ *be the transition kernel of the Markov process that is the (pathwise unique, strong) solution of Equation* (4). *Assuming Equation* (7) *holds, then for any density* μ*, we have*

$$\|\mu\mathbb{P}_\theta^t - \pi\|_{TV} \le \frac{1}{2} D_{\chi^2}(\mu\|\pi)^{1/2} e^{-t\alpha/2}$$

Proof. We have

$$\begin{aligned}
\|\mu\mathbb{P}_\theta^t - \pi\|_{TV} &= \sup_{B\in\mathcal{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} \mathbb{P}_\theta^t(x, B)\mu(x)\, dx - \pi(B) \right| \\
&= \sup_{B\in\mathcal{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} (\mathbb{P}_\theta^t(x, B) - \pi(B))\mu(x)\, dx \right| \\
&= \sup_{B\in\mathcal{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} (\mathbb{P}_\theta^t(x, B) - \pi(B))(\mu(x) - \pi(x))\, dx \right| \\
&= \sup_{B\in\mathcal{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} (\mathbb{P}_\theta^t(x, B) - \pi(B)) \left(\frac{\mu(x)}{\pi(x)} - 1 \right) \pi(x)\, dx \right| \\
&\le \sup_{B\in\mathcal{B}(\mathbb{R}^p)} \int_{\mathbb{R}^p} \left| \mathbb{P}_\theta^t(x, B) - \pi(B) \right| \cdot \left| \frac{\mu(x)}{\pi(x)} - 1 \right| \pi(x)\, dx \\
&\le \sup_{B\in\mathcal{B}(\mathbb{R}^p)} \left(\int_{\mathbb{R}^p} \left| \mathbb{P}_\theta^t(x, B) - \pi(B) \right|^2 \pi(dx) \right)^{1/2} \left(D_{\chi^2}(\mu\|\pi) \right)^{1/2},
\end{aligned}$$

where the first inequality is due to Jensen's, and the second inequality is Cauchy-Schwarz on $\mathcal{L}^2(\pi)$. Now, we want to show that Equation (4) is $\mathcal{L}^2(\pi)$-geometrically ergodic. To use Fact 31, we must find $\bar{a}$ and K. In our case, $b(\theta_t) = -\nabla f(\theta_t)$ and $\sigma(\theta_t) = \sqrt{2} I_p$, so we can choose $\bar{a} = \sqrt{2}$. We also lower bound K:

$$\begin{aligned}
K &= \sup_{x\neq y} \frac{\|\sigma(x) - \sigma(y)\|_{op}^2 + \langle b(x) - b(y), x - y\rangle}{\|x - y\|_2^2} \\
&= \sup_{x\neq y} \frac{\langle \nabla f(y) - \nabla f(x), x - y\rangle}{\|x - y\|_2^2} \\
&\le \sup_{x\neq y} \frac{-\frac{\alpha}{2}\|x - y\|_2^2}{\|x - y\|_2^2} = -\frac{\alpha}{2}.
\end{aligned}$$

So $\text{gap}(L) \ge \frac{\alpha}{2}\frac{1}{\sqrt{2}} \inf_x \lambda_{\min}(a(x)) = \frac{\alpha}{2}$. Reversibility directly follows from Fact 20 with $\theta_0 \sim \delta_x \in \mathcal{L}^2(\mathbb{P})$ and our β-smooth assumption on the potential f. ϕ-irreducibility follows using Fact 28 with since our β-smoothness condition gives us

$$\langle \nabla f(x), x\rangle = \langle \nabla f(x) - \nabla f(x^*), x\rangle \le \beta\|x - x^*\|_2\|x\|_2 \le \beta(\|x - x^*\|_2^2 \vee \|x\|_2^2)$$

where $x^* = \arg\min_{x\in\mathbb{R}^p} f(x)$ and using Cauchy-Schwarz. WLOG we can assume that f is minimized at 0, and so Fact 28 holds with $a = \beta$, $b = 0$, and $N = 0$. Now that we have $\mathcal{L}^2(\pi)$-geometric ergodicity by Fact 30, for any $B \in \mathcal{B}(\mathbb{R}^p)$, choose $\varphi(\cdot) = \mathbb{1}_B(\cdot) - \pi(B) \in \mathcal{L}^2(\pi)$ so that

$$\begin{aligned}
(\mathbb{E}[\varphi(\theta_t) \mid \theta_0 = x] - \pi(\varphi))^2 &= (\mathbb{E}[\mathbb{1}_B(\theta_t) - \pi(B) \mid \theta_0 = x] - \mathbb{E}_{Y\sim\pi}[\mathbb{1}_B(Y) - \pi(B)])^2 \\
&= (\mathbb{P}_\theta^t(x, B) - \pi(B))^2.
\end{aligned}$$

Using this identity for $\varphi(\cdot) = \mathbb{1}_B(\cdot) - \pi(\cdot)$ with the definition of $\mathcal{L}^2(\pi)$-geometric ergodicity, we have

$$\begin{aligned}\|P^t\varphi - \pi(\varphi)\|^2_{\mathcal{L}^2(\pi)} &= \int_{\mathbb{R}^p} (\mathbb{P}^t_\theta(x, B) - \pi(B))^2\pi(x)\, dx \le e^{-\alpha t}\mathbb{E}_{Y\sim\pi}[(\mathbb{1}_B(Y) - \pi(B))^2] \\ &= e^{-\alpha t}\pi(B)(1 - \pi(B)) \le \frac{1}{4}e^{-\alpha t}.\end{aligned}$$

Taking square roots gives us the result. ∎

Finally, we are ready to state the result, due to Dalalyan [2017], on non-asymptotic bounds of Langevin Monte Carlo iterates in the total variation norm.

Theorem 33. *Under the assumption of Equation* (7)*, let $h \le (C\beta)^{-1}$ and $K \ge C$ for some $C \ge 1$. Also let $\theta_0 \sim \mu = \mathcal{N}(\theta^*, \beta^{-1}I_p)$, where $\theta^* = \arg\min_{\theta\in\mathbb{R}^p} f(\theta)$ (uniqueness of minimizer is due to strong convexity). Then $\forall K \in \mathbb{N}$, we have*

$$\|\mu\mathbb{P}^{Kh}_\vartheta - \pi\|_{TV} \le \frac{1}{2}\exp\left(\frac{p}{4}\log\left(\frac{\beta}{\alpha}\right) - \frac{Kh\alpha}{2}\right) + \sqrt{\frac{p\beta^2Kh^2C}{4(2C-1)}},$$

where $\mu\mathbb{P}^{Kh}_\vartheta$ is the law of the K-th iterate of Equation (5) *started at $\theta_0 \sim \mu$.*

Proof. By the triangle inequality over the total variation norm, we have

$$\|\mu\mathbb{P}^{Kh}_\vartheta - \pi\|_{TV} \le \underbrace{\|\mu\mathbb{P}^{Kh}_\theta - \pi\|_{TV}}_{(a)} + \underbrace{\|\mu\mathbb{P}^{Kh}_\vartheta - \mu\mathbb{P}^{Kh}_\theta\|_{TV}}_{(b)},$$

where $\mathbb{P}^{Kh}_\theta$ is the law of Equation (4) at time Kh. For (a), we first apply Lemma 32 to get

$$(a) \le \frac{1}{2}D_{\chi^2}(\mu\|\pi)^{1/2}e^{-(Kh\alpha)/2}.$$

Evaluating the χ^2 divergence:

$$\begin{aligned}D_{\chi^2}(\mu\|\pi) &= \int_{\mathbb{R}^p}\left(\frac{\mu(x)}{\pi(x)} - 1\right)^2\pi(x)\, dx \\ &= \int_{\mathbb{R}^p}\left(\frac{\mu^2(x)}{\pi^2(x)} - 2\frac{\mu(x)}{\pi(x)} + 1\right)\pi(x)\, dx \\ &\le \exp\left(\frac{1}{2\alpha}\|\nabla f(\theta^*)\|_2^2 - \frac{p}{2}\log(2Kh\alpha)\right) - 1 \\ &\le \exp\left(-\frac{p}{2}\log(2Kh\alpha)\right) \\ &\le \exp\left(\frac{p}{2}\log\left(\frac{\beta}{2\alpha}\right)\right) \\ &\le \exp\left(\frac{p}{2}\log\left(\frac{\beta}{\alpha}\right)\right)\end{aligned}$$

where the first inequality comes from Lemma 25, the second inequality comes from removing the -1 and $\nabla f(\theta^*) = 0$, and the third inequality comes from $h \le 1/(C\beta)$ and $K \ge C$. So we have

$$(a) \le \frac{1}{2}\exp\left(\frac{p}{4}\log\left(\frac{\beta}{\alpha}\right) - \frac{1}{2}Kh\alpha\right).$$

For (b), we use Pinsker's inequality to get

$$(b) = \|\mu\mathbb{P}_{\vartheta}^{Kh} - \mu\mathbb{P}_{\theta}^{Kh}\|_{TV} \leq \|\mu\widetilde{\mathbb{P}}_{\theta^*,Kh} - \mu\mathbb{P}_{\theta^*,Kh}\|_{TV} \leq \left(\frac{1}{2}D_{KL}(\mu\mathbb{P}_{\theta^*,Kh}\|\mu\widetilde{\mathbb{P}}_{\theta^*,Kh})\right)^{1/2}.$$

We can use Lemma 24 to get

$$\begin{aligned}
D_{KL}(\mu\mathbb{P}_{\theta^*,Kh}\|\mu\widetilde{\mathbb{P}}_{\theta^*,Kh}) &\leq \frac{\beta^3h^2C}{12(2C-1)}(\underbrace{\mathbb{E}_{\theta_0\sim\mu}[\|\theta_0-\theta^*\|_2^2]}_{=p/\beta}+2Khp) + \frac{pK\beta^2h^2}{4} \\
&= \frac{\beta^2h^2Cp}{12(2C-1)} + \frac{\beta^3h^3KCp}{6(2C-1)} + \frac{pK\beta^2h^2}{4} \\
&= \frac{pK\beta^2h^2}{4}\left(\frac{C}{3K(2C-1)} + \frac{2\beta hC}{3(2C-1)} + 1\right) \\
&\leq \frac{p\beta^2Kh^2C}{2(2C-1)},
\end{aligned}$$

where the last inequality is due to $K \geq C$ and $h \leq 1/(C\beta)$. So we have

$$(b) \leq \sqrt{\frac{p\beta^2Kh^2C}{4(2C-1)}}.$$

Combining both bounds gives us the desired result. ■

By this theorem, if we wish to be sampling ϵ-close to our target distribution (in the TV distance), it is sufficient to perform $\mathcal{O}(\epsilon^{-2}p^3)$ iterations of Equation (5).

3.2 Under-damped Langevin Monte Carlo

Now, we consider an 'under-damped' version of Equation (4), where – in addition to the position vector – we include a momentum vector. Intuitively, we should observe an 'acceleration' effect (although this is difficult to explicitly show). One can intuit that this is to sampling what Nesterov gradient methods (see Fawzi [2024]) are to optimization, or 'simply' read Ma et al. [2021]. Below is the continuous underdamped Langevin diffusion, a system of SDEs:

$$\begin{cases} dX_t &= V_t\,dt \\ dV_t &= -\nabla f(X_t)\,dt - \gamma V_t\,dt + \sqrt{2\gamma}\,dW_t, \end{cases} \tag{8}$$

where $\gamma > 0$ is the momentum coefficient.

Here is a quick proposition about the generator of both the Langevin and underdamped Langevin diffusions, which will be useful later.

Proposition 34. *The generator of Langevin diffusion is $L = \Delta - \langle\nabla f, \nabla(\cdot)\rangle$. That is, $\mathbb{E}_\pi[Lg] = 0\ \forall g \in \mathcal{L}^2(\pi)$.*

Proof. Plug the diffusion and drift coefficients directly into Fact 3. ■

This immediately gives us a useful lemma from Chewi [2024].

Lemma 35 (Chewi [2024] Lemma 4.0.1)**.** *For $f \in \mathcal{C}^2(\mathbb{R}^p)$, we have the following:*

1. *If f is $\alpha-$strongly-convex, and x^* minimizes f, then $\mathbb{E}_\pi\| \cdot - x^*\|_2^2 \leq p/\alpha$.*
2. *If f is $\beta-$smooth, then $\mathbb{E}_\pi\|\nabla f\|_2^2 \leq \beta p$.*

Proof. Using Proposition 34, $Ag = \Delta g - \langle \nabla f, \nabla g\rangle$, and setting $g = (1/2)\| \cdot - x^*\|_2^2$ and $g = f$ gives 1. and 2., respectively. For 1., we have $\nabla g = \cdot - x^*$, hence

$$0 = \mathbb{E}_\pi[\Delta g - \langle \nabla f, \nabla g\rangle] = p - \mathbb{E}_{X\sim\pi}\langle \nabla f, X - x^*\rangle \leq d - \alpha \mathbb{E}_{X\sim\pi}[\|X - x^*\|_2^2],$$

then rearrange.

For 2. we have $\nabla^2 f \preceq \beta I_p$, hence $\Delta f = \nabla \cdot (\nabla f) \leq \beta p$

$$0 = \mathbb{E}_\pi[\Delta f - \|\nabla f\|_2^2] \leq \beta p - \mathbb{E}_\pi\|\nabla f\|_2^2,$$

then rearrange. ■

Here is a solution to the underdamped Langevin diffusion.

Proposition 36 (Cheng et al. [2018] Lemma 10)**.** *The following process solves Equation* (8)*:*

$$\begin{cases} V_t = V_0 e^{-\gamma t} - \int_0^t e^{-\gamma(t-s)} \nabla f(X_s)\, ds + \sqrt{2\gamma}\int_0^t e^{-\gamma(t-s)}\, dW_s \\ X_t = X_0 + \int_0^t V_s\, ds. \end{cases} \tag{9}$$

Proof. Take derivatives and use Itô's lemma. ■

This yields a discretization scheme, which will form the iterates of ULMC.

Fact 37 (Zhang et al. [2023] Appendix A)**.** *The conditional law of $(X_{(k+1)}h, V_{(k+1)}h)$ on (X_{kh}, V_{kh}) is $\mathcal{N}(M_{(k+1)h}, \Sigma)$, $M_{(k+1)h} \in \mathbb{R}^{2p}$, $\Sigma \in \mathbb{R}^{2p\times 2p}$, where*

$$M_{(k+1)h} = \begin{bmatrix} X_{kh} + \frac{1}{\gamma}(1 - \exp(-\gamma h))V_{kh} - \frac{1}{\gamma}(h - \frac{1}{\gamma}(1 - \exp(-\gamma h)))\nabla f(X_{kh} \\ V_{kh}\exp(-\gamma h) - \frac{1}{\gamma}(1 - \exp(-\gamma h))\nabla f(X_{kh}) \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \frac{2}{\gamma}(h - \frac{2}{\gamma}(1 - \exp(-\gamma h)) + \frac{1}{2\gamma}(1 - \exp(-2\gamma h)))I_p & \frac{1}{\gamma}(1 - 2\exp(-\gamma h) + \exp(-2\gamma h))I_p \\ \frac{1}{\gamma}(1 - 2\exp(-\gamma h) + \exp(-2\gamma h))I_p & (1 - \exp(-2\gamma h))I_p \end{bmatrix}.$$

Following Cheng et al. [2018], under a strongly convex and smooth potential, we hope to derive non-asymptotic convergence guarantees. We intuitively expect faster rates due to the (unproven) acceleration effect of ULMC. Our first lemma (Theorem 5 of Cheng et al. [2018]) provides a contraction bound under a certain norm (which is particularly useful for analyzing underdamped Langevin diffusions).

Lemma 38 (Cheng et al. [2018] Theorem 5)**.** *Let $(X_t^0, V_t^0), (X_t^1, V_t^1)$ be two underdamped Langevin diffusions driven by the same Brownian motion, with strongly convex and smooth potential as in Equation* (7)*. Defining the norm*

$$|||(x, v)|||^2 := \|x + \sqrt{\frac{2}{\beta}}v\|_2^2 + \|x\|_2^2$$

and setting $\gamma = \sqrt{2\beta}$, we have

$$\left|\left|\left|(X_t^0, V_t^0) - (X_t^1, V_t^1)\right|\right|\right| \leq \exp\left(-\frac{\alpha t}{\sqrt{2\beta}}\right)\left|\left|\left|(X_0^0, V_0^0) - (X_0^1, V_0^1)\right|\right|\right|.$$

Proof. Let $\delta X_t \coloneqq X_t^1 - X_t^0$, $\delta V_t \coloneqq V_t^1 - V_t^0$. By Itô's formula, we have

$$\begin{aligned}
d(\delta X_t + \eta \delta V_t) &= (\delta V_t - \eta(\nabla f(X_t^1) - \nabla f(X_t^0)) - \gamma\eta\delta V_t)\,dt \\
&= \left(-(\gamma\eta - 1)\delta V_t - \eta(\underbrace{\int_0^t \nabla^2 f((1-s)X_t^0 + sX_t^1)\,ds}_{=:H_t})\delta X_t \right) dt \\
&= \left(-(\gamma - \frac{1}{\eta})(\delta X_t + \eta\delta V_t) + (\gamma - \frac{1}{\eta} - \eta H_t)\delta X_t \right) dt
\end{aligned}$$

where the second step is using Taylor's theorem. Note that $H_t \in \mathbb{R}^{p\times p}$. We use Itô's formula again to obtain

$$d(\delta X_t) = \delta V_t\,dt = \left(\frac{1}{\eta}(\delta X_t - \eta\delta V_t) - \frac{1}{\eta}\delta X_t \right) dt.$$

Therefore, we have

$$\frac{1}{2}\frac{d}{dt}\left(\|\delta X_t + \eta\delta V_t\|_2^2 + \|\delta V_t\|_2^2 \right) = -\begin{bmatrix} \delta X_t + \eta\delta V_t \\ X_t \end{bmatrix}^\top \underbrace{\begin{bmatrix} (\gamma - \frac{1}{\eta})I_p & \frac{1}{2}(\eta H_t - \gamma I_p) \\ \frac{1}{2}(\eta H_t - \gamma I_p) & \frac{1}{\eta}I_p \end{bmatrix}}_{=:S_t} \begin{bmatrix} \delta X_t + \eta\delta V_t \\ X_t \end{bmatrix}.$$

For convenience, $S_t \in \mathbb{R}^{2p\times 2p}$ is written in $p \times p$ blocks. Since we have the strong convexity and smoothness assumption, the eigenvalues of H_t, denoted by $(\Lambda_i)_{i=1}^p$, are bounded below and above by α, β, respectively. Now, we substitute in $\eta = \sqrt{2/\beta}$ and $\gamma = 2/\eta = \sqrt{2\beta}$ and explicitly compute the eigenvalues of S_t. That is, we solve

$$\begin{aligned}
&\det\left(\begin{bmatrix} (\gamma - \frac{1}{\eta} - \lambda)I_p \frac{1}{2}(\eta H_t - \gamma I_p) & \\ \frac{1}{2}(\eta H_t - \gamma I_p) & (\frac{1}{\eta} - \lambda)I_p \end{bmatrix} \right) = 0 \\
&\implies (\frac{1}{\eta} - \lambda) = -\frac{1}{2}(\eta\Lambda_i - \frac{2}{\eta}),\ i = 1, \ldots, p \\
&\implies \lambda = \frac{\Lambda_i}{\sqrt{2\beta}} \geq \frac{\alpha}{\sqrt{2\beta}}.
\end{aligned}$$

Therefore, we have

$$\frac{1}{2}\frac{d}{dt}\|\!|\!|(\delta X_t, \delta V_t)|\!|\!| \leq \frac{\alpha}{\sqrt{2\beta}}\|\!|\!|\delta X_t, \delta V_t)|\!|\!|.$$

We can now use the differential form of Grönwall's lemma [Miller and Silvestri, 2024] to obtain the contraction bound. ■

Now, we note that our norm with three lines is equivalent to the Euclidean norm, i.e. $(1/3)(\|x\|_2^2 + (2/\beta)\|v\|_2^2) \leq |\!|\!|(x,v)|\!|\!| \leq 3(\|x\|_2^2) + (2/\beta)\|v\|_2^2$. This follows immediately from

$$\begin{aligned}
|\!|\!|(x,v)|\!|\!|^2 = \|x + \sqrt{\frac{2}{\beta}}v\|_2^2 + \|x\|_2^2 \leq 2(\|x\|_2^2 + \frac{2}{\beta}\|v\|_2^2) + \|x\|_2^2 \\
\frac{2}{\beta}\|v\|_2^2 \leq 2\|x + \sqrt{\frac{2}{\beta}}v\|_2^2 + 2\|x\|_2^2.
\end{aligned}$$

The takeaway is that bounding terms in the $|\!|\!|(\cdot,\cdot)|\!|\!|$ norm is equivalent to bounding both arguments in the Euclidean norm. Our next lemma is important: bounding the expected movement of our ULD particles in the Euclidean norm.

Lemma 39 (Chewi [2024] Exercise 5.7)**.** *Let (X_t, V_t) be a underdamped Langevin diffusion with β-smooth potential. If $t \le \beta^{-1/2} \wedge \gamma^{-1}$, then*

$$\mathbb{E}\|X_t - X_0\|_2^2 \lesssim t^2 \mathbb{E}\|V_0\|_2^2 + \gamma p t^3 + t^4 \mathbb{E}\|\nabla f(X_0)\|_2^2.$$

Proof.

$$\|X_t - X_0\|_2 = \| \int_0^t V_s \, ds \|_2 \le t\|V_0\|_2 + \| \int_0^t (V_s - V_0) \, ds \|_2,$$

by subtracting and adding V_0 then using triangle inequality. Since $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\|X_t - X_0\|_2^2 \le 2t^2\|V_0\|_2^2 + 2\| \int_0^t (V_s - V_0) \, ds \|_2^2.$$

Focusing on the second term on the right (without the square), we use the definition of ULD followed by the triangle inequality to write

$$\left\| \int_0^t (V_s - V_0) \, ds \right\|_2 \le \underbrace{\gamma \left\| \int_0^t \int_0^s V_r \, dr \, ds \right\|_2}_{(a)} + \underbrace{\left\| \int_0^t \int_0^s \nabla f(X_r) \, dr \, ds \right\|_2}_{(b)} + \underbrace{\sqrt{2\gamma} \left\| \int_0^t W_s \, ds \right\|_2}_{(c)}.$$

An important inequality is $\int_0^t g(s) \, ds \le t \sup_{s \in [0,t]} g(s)$. The non-decreasing property of our terms – e.g. $\int_0^s \|V_r\|_2 \, dr$ – along with the compactness of $[0,t]$ allows us to get rid of the supremum. We also use Jensen's inequality to throw the norm inside of the integrals. Now, for each of the terms, we have

$$\begin{aligned}
(a) &\le \gamma \int_0^t \int_0^s \|V_r\|_2 \, dr \, ds \\
&\le \gamma t \sup_{s \in [0,t]} \int_0^s \|V_r\|_2 \, dr \\
&= \gamma t \int_0^t \|V_s\|_2 \, ds \\
&\le \gamma t \int_0^t \|V_s - V_0\| \, ds + \gamma t^2 \|V_0\|_2 \\
(b) &\le \int_0^t \int_0^s \|\nabla f(X_r)\|_2 \, dr \, ds \\
&= \int_0^t \int_0^s \|\nabla f(X_r) - \nabla f(X_0) + \nabla f(X_0)\|_2 \, dr \, ds \\
&\le \beta \int_0^t \int_0^s \|X_r - X_0\|_2 \, dr \, ds + \frac{t^2}{2} \|\nabla f(X_0)\|_2 \\
&\le \beta t \int_0^t \|X_s - X_0\|_2 \, ds + \frac{t^2}{2} \|\nabla f(X_0)\|_2 \\
(c) &\le \sqrt{2\gamma} t \sup_{s \in [0,t]} \|W_s\|_2.
\end{aligned}$$

Now, we use $(\sum_{i=1}^n a_i)^2 \le n \sum_{i=1}^n a_i^2$ with $n = 5$ to get

$$\begin{aligned}
\left\| \int_0^t (V_s - V_0) \, ds \right\|_2^2 \le 5\gamma^2 t^2 \int_0^2 \|V_s - V_0\|_2^2 \, ds + 5\gamma^2 t^4 \|V_0\|_2^2 + \frac{5}{4} t^4 \|\nabla f(X_0)\|_2^2 + 10\gamma t^2 \sup_{s \in [0,t]} \|W_s\|_2^2 \\
+ 10\beta^2 t^2 \int_0^t \|X_s - X_0\|_2^2 \, ds.
\end{aligned}$$

We can plug this expression back into our inequality for $\|X_t - X_0\|_2^2$ and take expectations to get

$$\mathbb{E}\|X_t - X_0\|_2^2 \leq t^2(2 + 10\gamma^2 t^2)\mathbb{E}\|V_0\|_2^2 + 10\gamma^2 t^2 \int_0^t \mathbb{E}\|V_s - V_0\|_2^2 \, ds + \frac{5}{2}t^4 \mathbb{E}\|\nabla f(X_0)\|_2^2$$
$$+ 20\gamma t^2 \mathbb{E}[\sup_{s\in[0,t]} \|W_s\|_2^2] + 10\beta^2 t^2 \int_0^t \mathbb{E}\|X_s - X_0\|_2^2 \, ds.$$

For the $\mathbb{E}[\sup_s \|W_s\|_2^2]$ term, since the p-dimensional Brownian motion is the concatenation of p many independent 1-dimensional standard Brownian motions $(B^i)_{i=1}^p$ [Sousi, 2023], we have

$$\begin{aligned}\mathbb{E}[\sup_{s\in[0,t]} \|W_s\|_2^2] &= \sum_{i=1}^p \mathbb{E}[\sup_{s\in[0,t]} \|B_s^i\|_2^2] \\ &= p\mathbb{E}[\sup_{s\in[0,t]} |B_s^i|^2] \leq 4pt,\end{aligned}$$

where the first equality is by independence, second equality is by identical distributions, and the inequality is by Doob's $\mathcal{L}^p$-inequality for $p = 2$ [Sousi, 2023].

Finally, using Grönwall's inequality [Miller and Silvestri, 2024] on $\mathbb{E}\|X_t - X_0\|_2^2$ along with our assumption that $t \leq \beta^{-1/2} \wedge \gamma^{-1}$ gives us our desired movement bound. ■

With this, we can now state the main theorem.

Theorem 40 (Chewi [2024] Theorem 5.3.8)**.** *Assuming Equation* (7)*. Let* law$(X_0, V_0) = \mu_0 \otimes \mathcal{N}(0, I_p)$*,* $h = \varepsilon/\sqrt{\beta^2 p/\alpha}$*,* $\gamma = \sqrt{2\beta}$*, then* $\sqrt{\alpha} W_2(\mu_{Kh}, \pi) \leq \varepsilon$ *after*

$$N = \mathcal{O}\left(\frac{(\frac{\beta}{\alpha})^{3/2} p^{1/2}}{\varepsilon} \log \frac{\sqrt{\alpha} W_2(\mu_0, \pi)}{\varepsilon}\right)$$

iterations of ULMC.

Proof. Let (X_t, V_t) be the discrete ULD iterates that agrees with ULMC on each $t \in [kh, (k+1)h)$ and $(\bar{X}_t, \bar{V}_t)$ be the ULD. We first show an inequality for one step of ULMC.

$$\mathbb{E}\|X_h - \bar{X}_h\|_2^2 = \mathbb{E}\|\int_0^h (V_t - \bar{V}_t \, dt)\|_2^2 \leq h \int_0^h \mathbb{E}\|V_t - \bar{V}_t\|_2^2 \, dt,$$

where the last inequality is due to the fact that

$$\|\int_0^t g(s) \, ds\|_2^2 = \|\frac{1}{t}\int_0^t \langle t, g(s)\rangle \, ds\|_2^2 \leq t \int_0^t \|g(s)\|_2^2 \, ds$$

and the linearity of expectation. Of course, we can rewrite the $(V_s - V_0)$ term inside the norm using the definition of ULD and the fact $(a + b)^2 \leq 2(a^2 + b^2)$ to get

$$\mathbb{E}\|V_t - V_0\|_2^2 \leq 2h \int_0^t (\mathbb{E}\|\nabla f(\bar{X}_s) - \nabla f(X_s)\|_2^2 + \gamma^2 \mathbb{E}\|V_s - \bar{V}_s\|_2^2) \, ds.$$

We can use Grönwall's inequality [Miller and Silvestri, 2024] with the assumption that $h \leq 1/\gamma$ to get

$$\begin{aligned}\mathbb{E}\|V_t - \bar{V}_t\|_2^2 &\leq 2h \int_0^t \mathbb{E}\|\nabla f(\bar{X}_s) - \nabla f(X_s)\|_2^2 \, ds \exp(2h^2\gamma^2) \\ &\leq 2h \int_0^h \mathbb{E}\|\nabla f(\bar{X}_s) - \nabla f(X_s)\|_2^2 \, ds \\ &\leq 2\beta^2 h \int_0^t \mathbb{E}\|\bar{X}_s - X_s\|_2^2 \, ds.\end{aligned}$$

Now, we get to use our movement bound Lemma 39 to bound the triple norm of between one step of our discrete and continuous underdamped diffusion.

$$\begin{aligned}\mathbb{E}|||(X_h, V_h) - (\bar{X}_h, \bar{V}_h)||| &\lesssim \mathbb{E}\|X_h - \bar{X}_h\|_2^2 + \frac{1}{\beta}\mathbb{E}\|V_h - \bar{V}_h\|_2^2 \\ &\lesssim \beta h^4 \mathbb{E}\|V_0\|_2^2 + \beta^{3/2} p h^5 + \beta h^6 \mathbb{E}\|\nabla f(X_0)\|_2^2.\end{aligned}$$

With the single step bound finished, we are almost there. Let

$$\mathcal{C}^2(\mu, \nu) \coloneqq \inf_{\varpi \in \Pi(\mu,\nu)} \mathbb{E}_{(X_0, X_1)\sim\varpi,\, (V_0, V_1)\sim\mathcal{N}(0, I_p)\otimes\mathcal{N}(0,I_p)} \left|\left|\left|(X^0, V^0) - (X^1, V^1)\right|\right|\right|^2,$$

This is reminiscent of an optimal transport cost under the triple norm. Since the triple norm is equivalent to the Euclidean norm, we can use the triangle inequality followed by the contraction bound in Lemma 38 to get

$$\begin{aligned}\mathcal{C}(\mu_{(k+1)h}, \pi) &\leq \mathcal{C}(\mu_{(k+1)h, \bar{\mu}_{(k+1)h}}) + \mathcal{C}(\bar{\mu}_{(k+1)h}, \pi) \\ &\leq \exp\left(-\frac{\alpha h}{\sqrt{2\beta}}\right)\mathcal{C}(\mu_{kh}, \pi) + \sqrt{\mathcal{O}(\beta h \mathbb{E}\|V_{kh}\|_2^2 + \beta^{3/2} p h^5 + \beta h^6 \mathbb{E}\|\nabla f(X_{kh})\|_2^2}.\end{aligned}$$

Now, from Lemma 35, we have $\mathbb{E}\|\bar{V}_\infty\|_2^2 = p$ and $\mathbb{E}\|\nabla f(\bar{X}_\infty)\|_2^2 \leq \beta p$, where $(\bar{X}_\infty, \bar{V}_\infty) \sim \pi$. It is possible to apply this lemma to get $\mathbb{E}\|\nabla f(\bar{X}_\infty)\|_2^2 \leq \beta p$ as the marginal of $\bar{\pi}$ to x is a stationary measure of the (overdamped) Langevin diffusion (c.f. Equation (4)), and the marginal of $\bar{\pi}$ to v is a p-dimensional standard normal. This gives us

$$\begin{aligned}\mathbb{E}\|V_{kh}\|_2^2 + h^2\mathbb{E}\|\nabla f(X_{kh})\|_2^2 &\lesssim p + \beta p h^2 + \beta\mathcal{C}(\mu_{kh}, \pi) \\ &\lesssim p + \beta\mathcal{C}(\mu_{kh}, \pi).\end{aligned}$$

So

$$\begin{aligned}\mathcal{C}(\mu_{(k+1)h}, \pi) &\leq \exp(-\frac{\alpha h}{\sqrt{2\beta}})\mathcal{C}(\mu_{kh}, \pi) + \mathcal{O}(\beta^{1/2} p^{1/2} h^2 + \beta h^2 \mathcal{C}(\mu_{kh}, \pi) \\ &\leq \exp(-\frac{\alpha h}{2\sqrt{\beta}})\mathcal{C}(\mu_{kh}, \pi) + \mathcal{O}(\beta^{1/2} p^{1/2} h^2),\end{aligned}$$

where the second inequality comes from $h \lesssim \alpha/\beta^{3/2}$ and absorbing the $\beta h^2 \mathcal{C}(\mu_{kh}, \pi)$ term into the exponential term. Finally, we can iterate the above inequality K times to get

$$\mathcal{C}(\mu_{Kh}, \pi) \leq \exp(-\frac{\alpha K h}{2\sqrt{\beta}})\mathcal{C}(\mu_0, \pi) + \mathcal{O}(\frac{\beta p^{1/2} h}{\alpha}).$$

Solving backwards for ε gives us the desired result. ■

3.3 Metropolis-Adjusted Langevin Algorithm

One of the original remarks made by Roberts and Tweedie [1996] was to avoid using Langevin Monte Carlo due to the sensitivity of the Markov chain with respect to the choice of step size h. Another issue is that LMC is biased: looking at Theorem 33, we see that the second term provides an upper bound on the bias term $\|\mu\mathbb{P}_\vartheta^\infty - \pi\|_{TV}$. This is due to the Euler-Maruyama discretization of the continuous Langevin diffusion. There have been attempts throughout the years to resolve this issue, such as the recent Random Midpoint Method by Shen and Lee [2019]. The main idea is to use an extra source of randomness in the length of the step size to mitigate bias.

In this section, we will focus on the Metropolis adjustment. Similar to the Random Midpoint Method, we add an extra source of randomness. However, instead of randomly choosing the length of the step size and performing two iterations, we **propose** a new point based on our current point and the dynamics, then **accept or reject** the new point based on a randomized rule. To be more specific, let $Q(\cdot,\cdot)$ be a proposal kernel (or a Markov transition kernel), i.e. for each $x \in \mathbb{R}^p$, $Q(x,\cdot)$ is a probability density on $\mathbb{R}^p$. Supposing we are at a point $x_k \in \mathbb{R}^p$ during the k-th iteration, we then sample $y \sim Q(x_k,\cdot)$. Now, we will set $x_{k+1} := y$ with probability $A(x_k, y)$, or set $x_{k+1} := x$ with probability $1 - A(x_k, y)$, where $A(x,y) \in [0,1]$ outputs an acceptance probability based on the original and proposed location.

Algorithm 1 Generic Metropolis Algorithm

```
Input: Initial state $x_0$, number of iterations $N$, proposal kernel $Q(\cdot,\cdot)$, acceptance probability $A(\cdot,\cdot)$, target distribution $\pi(x)$
Output: (Approximate) samples from the target distribution $\pi(x)$
  for $k = 1$ to $N$ do
      Sample $y \sim Q(x_{k-1},\cdot)$
      Compute :
      $A(x_{k-1}, y) = 1 \wedge \frac{\pi(y)Q(y,x_{k-1})}{\pi(x_{k-1})Q(x_{k-1},y)}$
      Sample $u \sim \mathrm{Uniform}(0,1)$
      if $u \leq A(x_{k-1}, y)$ then
          Accept the proposed step: $x_k = y$
      else
          Reject the proposed step: $x_k = x_{k-1}$
      end if
  end for
  return $\{x_1, x_2, \ldots, x_N\}$
```

In the context of the Metropolis adjustment, we explicitly define the acceptance probability to be

$$A(x,y) := 1 \wedge \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}. \tag{10}$$

Now, our choice of $Q(\cdot,\cdot)$ will completely determine the sampling algorithm. For example, if we choose Q such that $Q(x,\cdot) = \mathcal{N}(x, hI_p)$, where $h > 0$ is the step size, then we obtain the Metropolized random walk. This is a simple but rather effective algorithm that does not require the evaluation of the gradient of the potential. We are only required to know the density of our invariant distribution π up to the normalization constant.

In the context of Langevin samplers, the obvious thing to do is to choose

$$Q(x,\cdot) = \mathcal{N}(x - h\nabla f(x), 2hI_d), \tag{11}$$

which is one iterate of Equation (5). This will be the main focus of this section.

The Metropolis-adjusted samplers, being the gold standard for empirical samplers, is notoriously difficult to analyze. This section's aim is to survey – without proofs – some important convergence results under different settings. At the end, we will see how ULMC can assist in speeding up the convergence of MALA.

3.3.1 Feasible Starts

Feasible start refers to when we set our initial distribution to be $\mathcal{N}(x^*, \beta^{-1}I_p)$, where $x^* = \arg\min_x f(x)$ – assuming that the potential is strongly convex and $\beta-$smooth. We have the following results from Chen et al. [2020].

Fact 41 (MRW, Feasible Start)**.** *Assuming $\pi \propto e^{-f}$ is α-strongly-log-concave and β-smooth. Then, using the Metropolis Random Walk algorithm, under a feasible start with step size $h = \mathcal{O}(\alpha/(\beta^2 p))$, we achieve*

$$\sqrt{D_{\chi^2}(\mu_{Nh}\|\pi)} \leq \varepsilon$$

after

$$N = \mathcal{O}\left(p\frac{\beta^2}{\alpha^2}\log\left(\frac{p}{\varepsilon}\right)\right)$$

many iterations, where μ_{Nh} is the law of the N-th MRW iterate.

We provide a similar statement for MALA.

Fact 42 (MALA, Feasible Start)**.** *Assuming $\pi \propto e^{-f}$ is α-strongly-log-concave and β-smooth. Then, using the Metropolis-adjusted Langevin Algorithm, under a feasible start with step size $h = \mathcal{O}(1/\beta p(1 \vee \sqrt{\beta/\alpha p}))$, we achieve*

$$\sqrt{D_{\chi^2}(\mu_{Nh}\|\pi)} \leq \varepsilon$$

after

$$N = \mathcal{O}\left(p\frac{\beta}{\alpha}\left(1 \vee \sqrt{\frac{\beta}{\alpha p}}\right)\log\left(\frac{p}{\varepsilon}\right)\right)$$

many iterations, where μ_{Nh} is the law of the N-th MALA iterate.

3.3.2 Warm Starts

A warm start initialization refers to any initial distribution μ_0 such that

$$D_{\chi^2}(\mu_0\|\pi) = \mathcal{O}(1).$$

Here is a more precise definition.

Definition 43. *A probability measure μ is said to be M-warm with respect to another probability measure π if*

$$\mu(B) \leq M\pi(B)\ \forall B \in \mathcal{B}(\mathbb{R}^p).$$

We easily see that this definition implies the imprecise definition since

$$\int (\frac{d\mu}{d\pi} - 1)^2 d\pi \leq M - 1.$$

Warm start mixing times, as well as how to obtain warm starts, have been a recent hot topic. This is due to practitioners wanting to circumvent the curse of dimensionality: the χ^2 divergence of the initial distribution from the target distribution grows **at least** exponentially with p [Lee et al., 2021a][Theorem 3]. We first showcase some mixing times, then explain how ULMC can be used to obtain warm starts for MALA.

Fact 44 (Wu et al. [2021], Chewi et al. [2020])**.** *Assuming α-strong log-concavity and β-smooth potential. For any accuracy $\varepsilon > 0$, the Metropolis-adjusted Langevin algorithm with an M-warm start and step size*

$$h = \mathcal{O}\left(\frac{\alpha 1/2}{\beta^{4/3}p^{1/2}\log(d\beta M/(\alpha\varepsilon))}\right)$$

will achieve

$$\sqrt{D_{\chi^2}(\mu_{Nh}\|\pi)} \leq \varepsilon$$

after

$$N = \mathcal{O}\left(\frac{\beta^{4/3}p^{1/2}}{\alpha^{3/2}} \log\left(\frac{M}{\varepsilon}\right) \log\left(p\frac{\beta}{\alpha} + \frac{M}{\varepsilon}\right)\right)$$

many iterations of MALA.

If we absorb the poly-logarithmic terms, for the purposes of sampling from high-dimensional distributions, under a warm start, we only require $N = \widetilde{\mathcal{O}}(\sqrt{p})$ many iterations. Naturally, we ask how one can obtain a warm initialization.

3.3.3 ULMC for Warm Starts

The main issue with obtaining warm starts is that the algorithm used to obtain the warm measure must have complexity less than or equal to MALA (or any metropolized sampler). Most importantly, in the case of sampling from high-dimensional distributions, is the dependence on dimension p.

Surprisingly, we are able to obtain a warm-start distribution via the underdamped Langevin Monte Carlo sampler discussed in Section 3.2. The extremely recent result Altschuler and Chewi [2024] proves this. We first define the Rényi divergence and discuss its relation to the χ^2-divergence.

Definition 45. *For $q > 1$, the q-Rényi divergence between two probability measures μ, ν is*

$$D_{R_q}(\mu\|\nu) = \frac{1}{q-1} \log\left(\int_{\mathbb{R}^p} (\frac{d\mu}{d\nu})^q \, d\nu\right).$$

By direct manipulation of the definition, we have that

$$\log(1 + D_{\chi^2}(\mu\|\nu)) = D_{R_2}(\mu\|\nu).$$

Therefore, the bound that we will state in Rényi divergence will be directly applicable to the above mixing time theorems. Now, we are ready to state the theorem.

Fact 46 (Altschuler and Chewi [2024])**.** *Assume that the potential is α-strongly convex and β-smooth. Let μ_{kh} denote the law of the k-th iterate of ULMC with $\gamma = \sqrt{2\beta}$ and $\mu_0 = \mathcal{N}(x^*, \beta^{-1}I_p)$, where $x^* = \arg\min_x f(x)$. Let $\varepsilon > 0$ be the target accuracy. Then, for step size*

$$h = \mathcal{O}(\frac{\varepsilon}{\sqrt{\beta^2 p/\alpha}}),$$

we will have

$$\sqrt{D_{R_2}(\mu_{Nh}\|\pi)} \leq \varepsilon$$

after

$$N = \mathcal{O}\left(\frac{\beta^{3/2}p^{1/2}}{\alpha^{3/2}\varepsilon}\right) = \widetilde{\mathcal{O}}(\sqrt{p})$$

many ULMC iterations.

3.4 Proximal Sampling

Proximal sampling is an new class of samplers and an active area of research [Lee et al., 2021b, Chen et al., 2022]. The theory of proximal sampling is extremely useful and is often applied to analyze related sampling algorithms. Additionally, proximal sampling also has connections to Wasserstein gradient flows, an important by-product of optimal transport theory [Villani et al., 2009].

Recall from convex optimization that the proximal operator is defined on the Euclidean space as

$$\mathrm{prox}_{h,f}(\cdot) \coloneqq \operatorname*{arg\,min}_{x\in\mathbb{R}^p} \left\{ f(x) + \frac{1}{2h}\|x - \cdot\|_2^2 \right\}.$$

In the context of minimizing convex composite functions $f + g$, the proximal gradient method is useful when one of the functions is not differentiable, but whose proximal operator is easily computable and can be viewed as an **oracle**. Intuition is that the proximal operator regularizes the problem. This motivates the definition of a proximal sampler.

Recalling that $\pi \propto e^{-f}$ is our target distribution. Fixing $h > 0$, we can augment the target with an additional variable of the same dimension to have the following joint distribution

$$\widetilde{\pi}(x, y) \propto \exp\left(-f(x) - \frac{\|y - x\|_2^2}{2h} \right)$$

on $\mathbb{R}^p \times \mathbb{R}^p$.

Viewing this augmented joint distribution, we can check four things. Note that $\propto_x$ denotes proportionality up to a constant dependent not on x. In this case, the constant depends only on y.

- The marginal of $\widetilde{\pi}$ with respect to X is $\pi^X \propto \pi$.
- The conditional distribution of Y given X is $\pi^{Y|X}(\cdot|x) = \pi^{Y|X=x}(\cdot|X = x) = \mathcal{N}(x, hI_p)$.
- The marginal of $\widetilde{\pi}$ with respect to Y is $\pi^Y = \pi^X * \mathcal{N}(x, hI_p)$, where $*$ denotes the convolution operator.
- The conditional distribution of X given Y is $\pi^{X|Y}(x|y) \propto_x \exp\left(-f(x) - \frac{1}{2h}\|x - y\|_2^2\right)$.

The idea of the proximal sampler is to sample from both conditional distributions (Gibb's sampling on $\widetilde{\pi}$). Given $x_0 \in \mathbb{R}^p$, for each iterate $k = 0, 1, \ldots$, we perform:

1. Sample $y_k \sim \pi^{Y|X}(\cdot|x_k) = \mathcal{N}(x_k, hI_p)$.
2. Sample $x_{k+1} \sim \pi^{X|Y}(\cdot|y_k)$

Since we are essentially performing Gibbs sampling, we borrow their theory and state that the proximal sampler is unbiased [Douc et al., 2018][Chapter 2.3.3].

Sampling from a normal distribution is easy enough, but we are not quite sure how to perform the second step. If we had access to, say, an **oracle** sampler that, when queried with $y \in \mathbb{R}^p$, returns a random variable with distribution $\pi^{X|Y}(\cdot|y)$, things would be nice. This object in proximal sampling is called the **restricted Gaussian oracle** (RGO).

Of course, in practical applications, we do not magically have access to oracles. Instead, we sample from the RGO using other sampling methods – such as rejection sampling, or the warm-started Metropolis-adjusted Langevin Algorithm. This also means the complexity of the proximal sampler is equal to the complexity of the RGO sampler multipied by the number of iterations.

We will prove convergence under strong log-concavity. The analysis of proximal samplers is fascinating – we use tools developed in Jordan et al. [1998] – as will be seen shortly. In fact, the RGO is a proximal operator (in the classical optimization sense) in the Wasserstein space $(\mathcal{P}_{2,ac}(\mathbb{R}^p), W_2)$. That is, given a functional $F : \mathcal{P}_{2,ac}(\mathbb{R}^p) \to \bar{\mathbb{R}}$, the proximal operator on Wasserstein space is defined as

$$\operatorname{prox}_F(\mu) := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)} \left\{ F(\nu) + \frac{1}{2} W_2^2(\mu, \nu) \right\}. \tag{12}$$

This follows from Lemma 15 and Equation (3) via a discretization, which Jordan et al. [1998] computes in greater detail. Now, we prove a connection between the Wasserstein proximal operator and the RGO. This requires a result on product measures.

Proposition 47. *Let $x \in \mathbb{R}^p$, then for any $\mu \in \mathcal{P}(\mathbb{R}^p)$ we have $\Gamma(\mu, \delta_x) = \{\mu \otimes \delta_x\}$ (i.e. the product measure is the unique coupling).*

Proof. Let $\gamma \in \Gamma(\mu, \delta_x)$, we wish to show $\gamma(A \times B) = \mu(A)\delta_x(B)$ for all $\mathbb{R}^p-$Borel sets A, B. Obviously, if $x \notin B$ then $\mu(A)\delta_x(B) = 0$ and $\gamma(A \times B) \le \gamma(\mathbb{R}^p \times B) = \delta_x(B) = 0$. If $x \in B$, then $\mu(A)\delta_x(B) = \mu(A)$ and $\gamma(A \times (\mathbb{R}^p \setminus B)) = 0$. Hence

$$\begin{aligned} \mu(A)\delta_x(B) = \mu(A) &= \gamma((A \times B) \cup (A \times (\mathbb{R}^p \setminus B))) \\ &= \gamma(A \times B) + \gamma(A \times (\mathbb{R}^p \setminus B)) = \gamma(A \times B) \end{aligned}$$

■

Lemma 48 (Chewi [2024] Exercise 8.1)**.** *For all $y \in \mathbb{R}^p$,*

$$\pi^{X|Y}(\cdot|y) = \operatorname{prox}_{hD_{KL}(\cdot\|\pi^X)}(\delta_y),$$

where δ_y is the Dirac measure at y.

Proof. We borrow a fact from Ambrosio et al. [2005][Remark 9.4.2]: For a Borel measure μ on $\mathbb{R}^p$ and a Borel map $V : \mathbb{R}^p \to (-\infty, +\infty]$ such that

- $V(x) \vee 0 \le A + B\|\bar{x} - x\|_2^2$ for all $x \in \mathbb{R}^p$ and some $A, B \ge 0$ and some $\bar{x} \in \mathbb{R}^p$,
- $\bar{\mu} := e^{-V}\mu$ is a probability measure,

we have, for all $\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$,

$$D_{KL}(\nu\|\mu) = D_{KL}(\nu\|\bar{\mu}) - \int_{\mathbb{R}^p} V(x)\nu(dx).$$

Choosing $\mu = \pi^X$, $\bar{\mu} = \pi^{X|Y=y}$, and $V = (1/2h)\|\cdot - y\|_2^2 + C(y)$ – where $C(y)$ is a constant depending only on y which appears after absorbing the y terms in $\pi^{X|Y=y} \propto_x \exp(-(1/2h)\|x - y\|_2^2)\pi^X$ into the exponential – we have

$$D_{KL}(\mu\|\pi^X) = D_{KL}(\mu\|\pi^{X|Y=y}) - \frac{1}{2h}\int_{\mathbb{R}^p} \|x - y\|_2^2 \mu(dx) + C(y).$$

Rearranging, we have

$$D_{KL}(\mu\|\pi^{X|Y=y}) = D_{KL}(\mu\|\pi^X) + \frac{1}{2h}\int_{\mathbb{R}^p} \|x - y\|_2^2 \mu(dx) - C(y).$$

Obviously, the functional $D_{KL}(\cdot\|\pi^{X|Y=y})$ is minimized over $\mathcal{P}_{2,ac}(\mathbb{R}^p)$ at $\pi^{X|Y=y}$. So by putting arg min over $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^p)$ on both sides of the equation, we have

$$\begin{aligned}\pi^{X|Y=y} &= \operatorname*{arg\,min}_{\mu\in\mathcal{P}_{2,ac}(\mathbb{R}^p)} \left\{ D_{KL}(\mu\|\pi^X) + \frac{1}{2h}\int_{\mathbb{R}^p} \|x-y\|_2^2 \mu(dx) \right\} \\ &= \operatorname*{arg\,min}_{\mu\in\mathcal{P}_{2,ac}(\mathbb{R}^p)} \left\{ D_{KL}(\mu\|\pi^X) + \frac{1}{2h} W_2^2(\mu,\delta_y) \right\} \\ &= \operatorname{prox}_{hD_{KL}(\cdot\|\pi^X)}(\delta_y),\end{aligned}$$

where the second equality comes from Proposition 47, and the last is by definition. Note that we indeed have uniqueness of arg min due to the strong geodesic convexity of $D_{KL}(\cdot\|\pi)$ as seen in Fact 17. ■

Now we can present the result from Chen et al. [2022].

Theorem 49 (Chewi [2024] Theorem 8.2.1)**.** *Assuming π^X is α-strongly log-concave, let $(\mu_k^X), (\bar{\mu}_k^X), k \geq 0$ be two copies of the iterates of the proximal sampler. Then*

$$W_2(\mu_k^X, \bar{\mu}_k^X) \leq \frac{1}{(1+\alpha h)^k} W_2(\mu_0^X, \bar{\mu}_0^X).$$

Proof. Note that the proof utilizes sub-differential calculus from convex optimization [Fawzi, 2024]. From Lemma 48, we have

$$\pi^{X|Y=y} = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_{2,ac}(\mathbb{R}^p)} \left\{ D_{KL}(\mu\|\pi^X) + \frac{1}{2h} W_2^2(\mu,\delta_y) \right\}.$$

Using another fact from Ambrosio et al. [2005][Lemma 10.1.2], we know that optimality conditions are the same for Wasserstein as it is on Euclidean space. That is, our minimizer $\pi^{X|Y=y}$ must satisfy

$$\frac{1}{h}(y - \mathrm{id}) \in \partial D_{KL}(\pi^{X|Y=y}\|\pi^X),$$

where $\partial D_{KL}(\cdot\|\pi^X)$ is the sub-differential. Therefore, for two $y, \bar{y} \in \mathbb{R}^p$, we have

$$\begin{aligned}&\mathrm{id} \in y - h\,\partial D_{KL}(\pi^{X|Y=y}\|\pi^X),\ \pi^{X|Y=y}\text{-a.s.}\\ &\mathrm{id} \in \bar{y} - h\,\partial D_{KL}(\pi^{X|Y=\bar{y}}\|\pi^X),\ \pi^{X|Y=\bar{y}}\text{-a.s.}\end{aligned}$$

Let T be the optimal transport map from $\pi^{X|Y=y}$ to $\pi^{X|Y=\bar{y}}$. Using Ambrosio et al. [2005][Theorem 6.2.4], we know T is the unique transport map. We apply T on both sides to get

$$T \in \bar{y} - h\,\partial D_{KL}(\pi^{X|Y=y}|\pi^X) \circ T,\ \pi^{X|Y=y}\text{-a.s.}$$

Now we can write, $\pi^{X|Y=y}$-a.s., that

$$T - \mathrm{id} \in (\bar{y} - y) - h\,\partial\left(D_{KL}(\pi^{X|Y=\bar{y}}\|\pi^X)\circ T - D_{KL}(\pi^{X|Y=y}\|\pi^X)\right).$$

Let $\delta(\pi^{X|Y=y}) \in \partial D_{KL}(\pi^{X|Y=y}\|\pi^X)$ and $\delta(\pi^{X|Y=\bar{y}}) \in \partial D_{KL}(\pi^{X|Y=\bar{y}}\|\pi^X)$. Then we can write

$$\begin{aligned}\|T-\mathrm{id}\|^2 &= \|\bar{y}-y\|^2 - 2h\langle \delta(\pi^{X|Y=\bar{y}})\circ T - \delta(\pi^{X|Y=y}), T - \mathrm{id}\rangle - h^2\|\delta(\pi^{X|Y=\bar{y}})\circ T - \delta(\pi^{X|Y=y})\|^2 \\ &\leq \|\bar{y}-y\|^2 - 2h\alpha\|T-\mathrm{id}\|^2 - h^2\alpha^2\|T-\mathrm{id}\|^2,\end{aligned}$$

where $\|\cdot\|$ is an abbreviation for $\|\cdot\|_{\mathcal{L}^2(\pi^{X|Y=y})}$, and the inequality comes from the geodesic α-strong convexity of $D_{KL}(\cdot\|\pi^X)$ induced by the α-strong log-concavity of π^X in Fact 17. Now we integrate with respect to $\pi^{X|Y=y}$, obtaining

$$W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) \leq \|\bar{y}-y\|_2^2 - 2\alpha h W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}) - \alpha^2 h^2 W_2^2(\pi^{X|Y=y}, \pi^{X|Y=\bar{y}}).$$

We do a little rearranging and get the result. ■

We note the elegance of the proof based on only using calculus in the Wasserstein space. Before ending the section, it is crucial to note that there are important techniques for analyzing proximal samplers that are not considered here. Notably, we have simultaneous flow – based on the observation that applying the heat equation (diffusion with no drift coefficient and constant diffusion coefficient) to π^X and μ_k^X transforms the measures into π^Y and μ_k^Y, respectively. This, combined with the fact that we can reverse the heat flow with a specified final condition (in the form of a probability measure) [Anderson, 1982], provides us with a way to move both forwards and backwards in time, as well as between the X and Y marginals of $\widetilde{\pi}$.

3.5 Hamiltonian Monte Carlo

There are two more important algorithms that are very popular and should be covered, albeit very briefly: Hamiltonian Monte Carlo methods and its Metropolized version. They can be viewed as a generalization of Langevin Monte Carlo methods. We only provide basic definitions and how they work, as they are still an important class of samplers. A good introduction would be Betancourt [2017], while those wishing to dive into greater theory would enjoy Arnold [1978], Hairer et al. [2006].

Similar to the underdamped Langevin diffusion, we add a momentum variable v. This leads to an augmentation of the target density to be

$$\widetilde{\pi} \propto \exp\left(-f(x) - \frac{1}{2}\|v\|_2^2\right).$$

The term is the exponential is also known as the **Hamiltonian**, denoted by $H(x, v) \coloneqq f(x) + \frac{1}{2}\|v\|_2^2$. The pair (x, v) are then governed by **Hamilton's equations**:

$$\begin{cases} \frac{d}{dt}x_t = \nabla_v H(x_t, v_t) = v_t \\ \frac{d}{dt}v_t = -\nabla_x H(x_t, v_t) = -\nabla f(x_t). \end{cases}$$

This system can also be written as

$$\begin{bmatrix} \frac{d}{dt}x_t \\ \frac{d}{dt}v_t \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & I_p \\ -I_p & 0 \end{bmatrix}}_{=:J} \nabla H(x_t, v_t). \tag{13}$$

For initial conditions $(x_0, v_0) \in \mathbb{R}^p \times \mathbb{R}^p$, let $F : \mathbb{R}^+ \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p \times \mathbb{R}^p$ be the solution of Equation (13) started at (x_0, v_0) at time t. That is, for every $t \geq 0$ and initial conditions $(x_0, v_0) \in \mathbb{R}^p \times \mathbb{R}^p$, $F_t(x_0, v_0) = (x_t, v_t)$.

We note that, for fixed t, our augmented distribution is invariant under F_t. That is, $(F_t)_{\#}\widetilde{\pi} = \widetilde{\pi}$, where $\#$ denotes the pushforward action of maps on measures. However, simply having $(x_0, v_0) \sim \mu_0$ and running $t \to \infty$ does not necessarily lead to convergence. For example, $\lim_{t\to\infty} D_{KL}((F_t)_{\#}\mu_0\|\widetilde{\pi}) \neq 0$ when $f(x) = (1/2)\|x\|_2^2$ and μ_0 is chosen such that $D_{KL}(\mu_0\|\widetilde{\pi}) \neq 0$.

This motivates the **refreshing** of momentum. We fix an **integration time** $T > 0$, then after every T units of time, we sample a new momentum vector to replace the current momentum vector. Intuitively, this means every T seconds, a random Gaussian force acts on our particle. Thus, in the ideal world, we would have the following algorithm.

Definition 50 (Ideal HMC)**.** *Given $T > 0$ and $\mu_0 \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p)$, sample $(x_0, v_t) \sim \mu_0$. Now, for each $k \in \mathbb{N}$:*

1. *Sample $v'_{kT} \sim \mathcal{N}(0, I_p)$.*

2. Set $(x_{(k+1)T}, v_{(k+1)T} \coloneqq F_T(x_{kT}, v'_{kT})$.

Afterwards, output $(x_{kT})_{k=0}^K$ for some $K \in \mathbb{N}$. These are the samples from the Ideal HMC.

As such, we have an optimal result for the convergence of Ideal HMC from Chen and Vempala [2022].

Fact 51. *Assume that the target distribution $\pi \propto \exp(-f)$ satisfies β-smoothness and α-strong log-concavity. For each $k \in \mathbb{N}$, let π_{kT}^X denote the law of x_{kT} defined from the Ideal HMC iterates with integration time $T > 0$. Specifically, if we choose $T = 1/(2\sqrt{\beta})$, then we have*

$$W_2^2(\pi_{NT}^X, \pi) \leq \exp\left(-\frac{N\alpha}{16\beta}\right) W_2^2(\mu_0^X, \pi).$$

However, it is intractable to integrate Hamilton's equations exactly, and we must resort to a numerical ODE solver. The most popular solver in the context of HMC is the leapfrop integrator [Neal, 2012]. Fixing $N \in \mathbb{N}$ and integration time $T > 0$, the leapfrog integrator with $K \in \mathbb{N}$ iterations and step size $h > 0$ has the following iterations for $k = 0, 1, \ldots, K-1$:

1. Set $v_{(k+1/2)h} \coloneqq p_{kh} - (h/2)\nabla f(x_{kh})$.
2. Set $x_{(k+1)h} \coloneqq x_{kh} + hv_{(k+1/2)h}$.
3. Set $v_{(k+1)h} \coloneqq v_{(k+1/2)h} - (h/2)\nabla f(x_{(k+1)h})$.

After the iterations, we have approximately integrated Hamilton's equation up until TN. Now, we refresh our momentum vector and repeat the above procedure again.

3.5.1 Metropolized Hamiltonian Monte Carlo

The obvious thing to do now is to add a Metropolis-adjustment to our above algorithm. Namely, we have, for initial position vector $x_0 \sim \pi_0^X$. For $k = 0, 1, \ldots$:

1. Sample $v_k \sim \mathcal{N}(0, I_p)$.
2. Propose $(x', v') = F_{l,N,h}(x_k, v_k)$.
3. With probability $1 \wedge \exp(H(x_k, v_k) - H(x', v'))$, set $x_{k+1} \coloneqq x'$. Otherwise, set $x_{k+1} \coloneqq x_k$.

Notice that when $N = 1$, i.e. we use a 1-step leapfrog integrator, the above algorithm reduces to MALA.

4 Simulations

In this section, we briefly perform some simple experiments that utilize the samplers developed in Section 3. It should be noted that there is still a large gap between observed empirical phenomena of samplers and our theoretical understanding of their dynamics and mixing times. Therefore, this section serves more as a playground, and we cannot reasonably, even empirically, conclude anything concrete. However, it is still fruitful to see the samplers that we have developed throughout the paper be applied. Also note that many of the parameters and hyper-parameters chosen in the two sections below are mostly arbitrary and do not use any theoretically-backed or empirically-backed guidelines. The (anonymized) code is available at HTTPS://ANONYMOUS.4OPEN.SCIENCE/R/LANGEVINSAMPLING/.

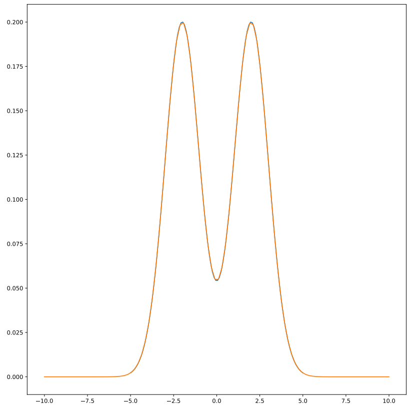


Figure 1: A plot of the probability density function induced by the potential in Equation (14).

4.1 Sampling from a Mixture of Gaussians in 1D

We test our samplers by sampling from the density $\pi \propto \exp(-f)$, $p = 1$, where:

$$f(x) = \log\left(\exp\left(-\frac{1}{2}\|x-2\|_2^2\right) + \exp\left(-\frac{1}{2}\|x+2\|_2^2\right)\right), \tag{14}$$

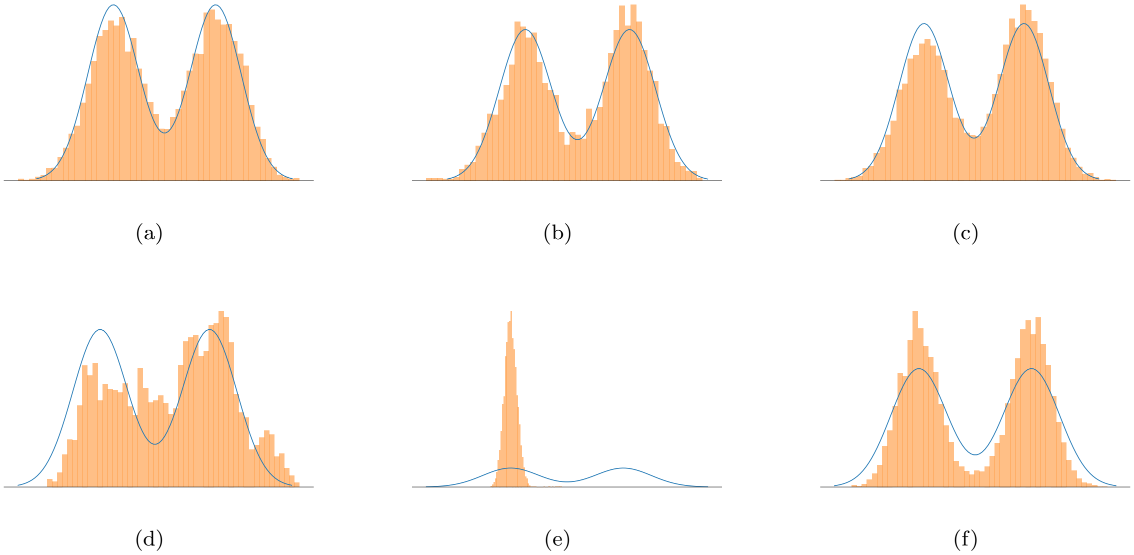
i.e. a sum of two Gaussians (see Figure 1). Since the theory developed in Section 3 does not cover the case when the potential if non-convex, it may be of interest to consider such a potential. For each sampler, we will run $K = 10000$ iterations starting at $x_0 = 0$. All samplers will use $h = 0.1$. For ULMC, we will sample $v_0 \sim \mathcal{N}(0, 1)$ and choose $\gamma = 1$. For the Hamiltonian Monte Carlo and Metropolize Hamiltonian samplers, we use $N = 10$ leapfrog steps. The results are seen in Figure 2.

For LMC, ULMC, HMC, and MHMC, we observe good matching between the empirical density formed from the Markov chain's samples and true density. MRW is also relatively great, although it is slightly biased towards the rightward mode, presumably due to the initialization at 0 followed by a first rightward step (for this random seed). MALA, however, seems to be completely stuck in the leftward mode. Some possible explanations include (1) incompatible step-size, (2) lack of 'burn-in period', which involves throwing away the first, say, 10% of samples generated by the method. It may also be due to (3) an unconditioned proposal distribution, i.e. it may be more fruitful to propose $\mathcal{N}(x - hH(x)\nabla f(x), 2hI_d)$, where H is some chosen matrix-valued function, e.g. the inverse Hessian of f, if it exists. This is known as pre-conditioning.

4.2 Bayesian Logistic Regression

A simple application of these samplers is to sample from the posterior distribution of the parameters governing a logistic regression model without intercept, also known as Bayesian logistic regression [Gelman et al., 1995]. For our toy problem, we will generate a classification dataset with 2 features and 100 data-points, half of which are held-out until the testing/predictive phase. These data-points are standardized to mean 0 and standard deviation 1 before-hand.

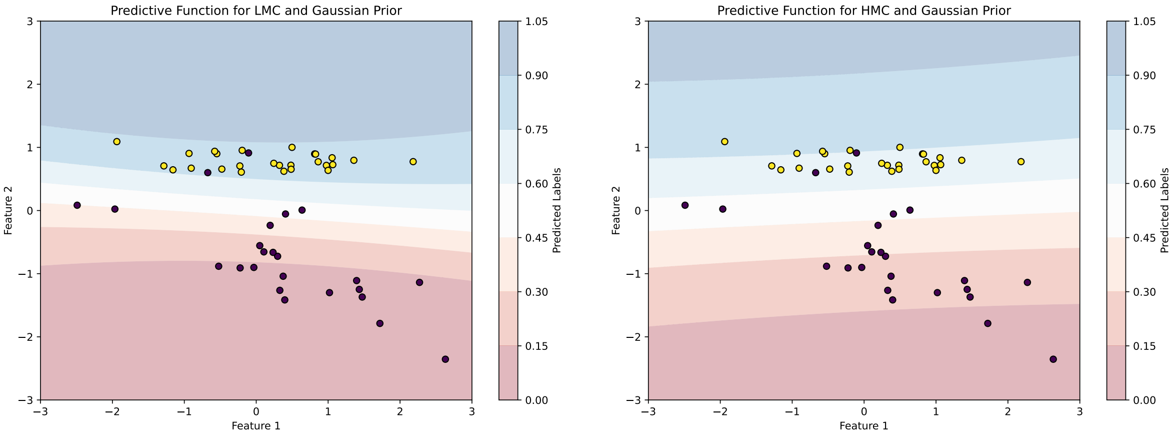
By specifying a multivariate Gaussian (or Laplacian) as the prior, we implicitly define a posterior distribution, which our samplers will attempt to sample from. The covariance matrix for these distributions can be specified to be isotropic (i.e. $\mathcal{N}(0, \lambda I_p)$) with hyper-parameter $\lambda > 0$ to up-weight or down-weight the effect of the prior. We choose $\lambda = 10$, $h = 0.01$, $K = 100$. To form our prediction, we sample 1000 times from the



(a) (b) (c)

(d) (e) (f)

Figure 2: A table that plots the empirical densities in orange – with the target density in blue – for each of the following samplers: (a) Langevin Monte Carlo, (b) Underdamped Langevin Monte Carlo, (c) Hamiltonian Monte Carlo, (d) Metropolized Random Walk, (e) Metropolis-adjusted Langevin Algorithm, (f) Metropolized Hamiltonian Monte Carlo.



(a) (b)

Figure 3: For both plots, the points colored yellow refer to test points under class $Y_i = 1$, the points colored purple refer to test points under class $Y_i = 0$. The abscissa and ordinate refer to the first and second coordinates of the test inputs X_i^1 and X_i^2, respectively. The contour in the background is the posterior predictive functions' output on uniformly spaced points in $[-3, +3]^2$. As seen in the temperature scale on the right sides of both figures, areas where the posterior predictive distribution has high confidence that the input is class 1 exhibit darker blue colors, vice versa for class 0 exhibiting darker red colors. For points where the predictive distribution is more uncertain, i.e. near the decision boundary, the color is lighter and closer to white. (a) refers to using the LMC sampler to sample from the posterior after specifying a isotropic Gaussian prior, (b) refers to using the HMC sampler to sample from the posterior after specifying a isotropic Gaussian prior.

posterior predictive distribution and average the outputs. For Hamiltonian Monte Carlo, we run $N = 10$ leapfrog steps. The predictive landscape is displayed in Figure 3.

5 Discussion

We have surveyed a small (but important) subset of sampling algorithms with a focus on Langevin-type samplers. The theoretical results presented can provide valuable insight for practical applications. However, these convergence properties have only been shown for strong convexity and smoothness of the potential function. In real-world applications, these assumptions seldom hold — in fact, most Bayesian posteriors are not even convex [Altmeyer, 2022]. Other distributions, such as the multi-modal distribution considered in Section 4.1, also do not satisfy the assumptions. In spite of these violations, empirical performance is still good [Dias and Wedel, 2004]. As such, a large portion of current research is bridging this gap. Another research direction is in transporting ideas from convex optimization to log-concave sampling.

Practically, there is also the possibility of exploring diagnostic tools for assessing the quality of the samples generated by the algorithm. Refer to Cowles and Carlin [1996] for details. Using domain-specific knowledge to create ad-hoc sampling algorithms that incorporate structural assumptions about the target distribution is also an active area of work, for example in protein design [Ovchinnikov and Huang, 2021, Hosur et al., 2012]. As the field of machine learning continues to evolve to tackle the increasingly high-dimensional and complex problems, the development of more efficient or effective sampling techniques — along with their analysis — will continue to play a key role.

References

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X. Lu, Nicolo Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. **bioRxiv**, 2023. doi: 10.1101/2023.09.11.556673. URL `https://www.biorxiv.org/content/early/2023/09/12/2023.09.11.556673`.

Randolf Altmeyer. Polynomial time guarantees for sampling based posterior inference in high-dimensional generalised linear models. **arXiv preprint arXiv:2208.13296**, 2022.

Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. **J. ACM**, mar 2024. ISSN 0004-5411. doi: 10.1145/3653446. URL `https://doi.org/10.1145/3653446`. Just Accepted.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. **Gradient flows in metric spaces and in the space of probability measures**. Birkhäuser Verlag, 2005.

Brian D. O. Anderson. Reverse-time diffusion equation models. **Stochastic Processes and their Applications**, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5.

V. I. Arnold. **Mathematical Methods of Classical Mechanics**. Springer New York, 1978. ISBN 9781475716931. doi: 10.1007/978-1-4757-1693-1. URL `http://dx.doi.org/10.1007/978-1-4757-1693-1`.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. **arXiv preprint arXiv:1701.02434**, 2017.

Christopher M. Bishop and Hugh Bishop. **Deep Learning: Foundations and Concepts**. Springer International Publishing, 2024. ISBN 9783031454684. doi: 10.1007/978-3-031-45468-4. URL `http://dx.doi.org/10.1007/978-3-031-45468-4`.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. **Journal of the American statistical Association**, 112(518):859–877, 2017.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. **Communications on Pure and Applied Mathematics**, 44:375–417, 1991. URL `https://api.semanticscholar.org/CorpusID:123428953`.

Mu-Fa Chen and Feng-Yu Wang. Estimation of spectral gap for elliptic operators. **Transactions of the American Mathematical Society**, 349(3):1239–1267, 1997. ISSN 00029947. URL `http://www.jstor.org/stable/2155416`.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In Po-Ling Loh and Maxim Raginsky, editors, **Proceedings of Thirty Fifth Conference on Learning Theory**, volume 178 of **Proceedings of Machine Learning Research**, pages 2984–3014. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/chen22c.html`.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: benefits of multi-step gradients. **J. Mach. Learn. Res.**, 21(1), jan 2020. ISSN 1532-4435.

Zongchen Chen and Santosh S. Vempala. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. **Theory of Computing**, 18(9):1–18, 2022. doi: 10.4086/toc.2022.v018a009. URL `https://theoryofcomputing.org/articles/v018a009`.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, **Proceedings of the 31st Conference On Learning Theory**, volume 75 of **Proceedings of Machine Learning Research**, pages 300–323. PMLR, 06–09 Jul 2018. URL `https://proceedings.mlr.press/v75/cheng18a.html`.

Sinho Chewi. **Log-Concave Sampling**. Draft, 2024. URL `https://chewisinho.github.io/main.pdf`.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In **Annual Conference Computational Learning Theory**, 2020. URL `https://api.semanticscholar.org/CorpusID:229363679`.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. **Journal of the American statistical Association**, 91(434):883–904, 1996.

Mihalis Dafermos. Part III differential geometry lecture notes. In **Part III Differential Geometry Lecture Notes**, 2012. URL `https://api.semanticscholar.org/CorpusID:51762027`.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, 79(3):651–676, 2017. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/44681805`.

A.S. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin montecarlo. **Journal of Computer and System Sciences**, 78(5):1423–1443, 2012. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.12.023. URL `https://www.sciencedirect.com/science/article/pii/S0022000012000220`. JCSS Special Issue: Cloud Computing 2011.

José G Dias and Michel Wedel. An empirical comparison of em, sem and mcmc performance for problematic gaussian mixture likelihoods. **Statistics and Computing**, 14:323–332, 2004.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. **Markov Chains**. Springer International Publishing, 2018. ISBN 9783319977041. doi: 10.1007/978-3-319-97704-1. URL `http://dx.doi.org/10.1007/978-3-319-97704-1`.

Hamza Fawzi. Part III topics in convex optimisation lecture notes. **Lecture Notes**, 2024.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. **Bayesian data analysis**. Chapman and Hall/CRC, 1995.

I. V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. **Theory of Probability & Its Applications**, 5(3):285–301, January 1960. ISSN 1095-7219. doi: 10.1137/1105027. URL `http://dx.doi.org/10.1137/1105027`.

Ernst Hairer, Gerhard Wanner, and Christian Lubich. **Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations**. Springer-Verlag, 2006. ISBN 3540306633. doi: 10.1007/3-540-30666-8. URL `http://dx.doi.org/10.1007/3-540-30666-8`.

Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. **Bayesian nonparametrics**, volume 28. Cambridge University Press, 2010.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. **Advances in neural information processing systems**, 33:6840–6851, 2020.

Raghavendra Hosur, Jian Peng, Arunachalam Vinayagam, Ulrich Stelzl, Jinbo Xu, Norbert Perrimon, Jadwiga Bienkowska, and Bonnie Berger. A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. **Genome biology**, 13:1–14, 2012.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. **SIAM Journal on Mathematical Analysis**, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL `https://doi.org/10.1137/S0036141096303359`.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Lower bounds on metropolized sampling methods for well-conditioned distributions. In **Neural Information Processing Systems**, 2021a. URL `https://api.semanticscholar.org/CorpusID:235390877`.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In Mikhail Belkin and Samory Kpotufe, editors, **Proceedings of Thirty Fourth Conference on Learning Theory**, volume 134 of **Proceedings of Machine Learning Research**, pages 2993–3050. PMLR, 15–19 Aug 2021b. URL `https://proceedings.mlr.press/v134/lee21a.html`.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of nesterov acceleration for gradient-based mcmc? **Bernoulli**, 27(3), May 2021. ISSN 1350-7265. doi: 10.3150/20-bej1297. URL `http://dx.doi.org/10.3150/20-BEJ1297`.

Dougal Maclaurin, David Duvenaud, and Ryan P Adams. Autograd: Effortless gradients in numpy. In **ICML 2015 AutoML workshop**, volume 238-5, 2015.

Jason Miller and Vittoria Silvestri. Part III stochastic calculus lecture notes. **Lecture Notes**, 2024.

Radford Neal. Mcmc using hamiltonian dynamics. **Handbook of Markov Chain Monte Carlo**, 06 2012. doi: 10.1201/b10905-6.

Bernt Oksendal. **Stochastic differential equations (3rd ed.): an introduction with applications**. Springer-Verlag, Berlin, Heidelberg, 1992. ISBN 3387533354.

Sergey Ovchinnikov and Po-Ssu Huang. Structure-based protein design with deep learning. **Current opinion in chemical biology**, 65:136–144, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. **arXiv preprint arXiv:2204.06125**, 1(2):3, 2022.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. **Bernoulli**, 2(4):341 – 363, 1996.

Gareth O. Roberts and Richard L. Tweedie. Geometric l2 and l1 convergence are equivalent for reversible markov chains. **Journal of Applied Probability**, 38:37–41, 2001. ISSN 00219002. URL `http://www.jstor.org/stable/3215866`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/eb86d510361fc23b59f18c1bc9802cc6-Paper.pdf`.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. **Advances in neural information processing systems**, 32, 2019.

Perla Sousi. Part III advanced probability lecture notes. **Lecture Notes**, 2023.

Matthew Thorpe. Introduction to optimal transport. **Lecture Notes**, 3, 2019.

Ali S Üstünel. Analysis on wiener space and applications. **arXiv preprint arXiv:1003.1649**, 2010.

Cédric Villani et al. **Optimal transport: old and new**, volume 338. Springer, 2009.

Changye Wu and Christian P. Robert. Coordinate sampler: a non-reversible gibbs-like mcmc sampler. **Statistics and Computing**, 30(3):721–730, December 2019. ISSN 1573-1375. doi: 10.1007/s11222-019-09913-w. URL `http://dx.doi.org/10.1007/s11222-019-09913-w`.

Keru Wu, Scott C. Schmidler, and Yuansi Chen. Minimax mixing time of the metropolis-adjusted langevin algorithm for log-concave sampling. **J. Mach. Learn. Res.**, 23:270:1–270:63, 2021. URL `https://api.semanticscholar.org/CorpusID:237940357`.

Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A Erdogdu. Improved discretization analysis for underdamped langevin monte carlo. In **The Thirty Sixth Annual Conference on Learning Theory**, pages 36–71. PMLR, 2023.