Convergence of Langevin MCMC in KL-divergence

Xiang Cheng   X.CHENG@BERKELEY.EDU and **Peter Bartlett**   PETER@BERKELEY.EDU

Editor:

Abstract

Langevin diffusion is a commonly used tool for sampling from a given distribution. In this work, we establish that when the target density $\mathbf{p}^*$ is such that $\log \mathbf{p}^*$ is L smooth and m strongly convex, discrete Langevin diffusion produces a distribution $\mathbf{p}$ with $\mathcal{KL}(\mathbf{p} \| \mathbf{p}^*) \le \epsilon$ in $\tilde{O}(\frac{d}{\epsilon})$ steps, where d is the dimension of the sample space. We also study the convergence rate when the strong-convexity assumption is absent. By considering the Langevin diffusion as a gradient flow in the space of probability distributions, we obtain an elegant analysis that applies to the stronger property of convergence in KL-divergence and gives a conceptually simpler proof of the best-known convergence results in weaker metrics.

1. Introduction

Suppose that we would like to sample from a density

$$\mathbf{p}^*(x) = e^{-U(x)+C}$$

where C is the normalizing constant. We know $U(x)$, but we do not know the normalizing constant. This comes up, for example, in variational inference, when the normalization constant is computationally intractable.

One way to sample from $\mathbf{p}^*$ is to consider the Langevin diffusion:

$$\bar{x}_0 \sim \bar{\mathbf{p}}_0$$
$$d\bar{x}_t = -\nabla U(\bar{x}_t)dt + \sqrt{2}dB_t \tag{1}$$

Where $\bar{\mathbf{p}}_0$ is some initial distribution and B_t is Brownian motion (see Section 4). The stationary distribution of the above SDE is $\mathbf{p}^*$.

The Langevin MCMC algorithm, given in two equivalent forms in (3) and (4), is an algorithm based on discretizing (1).

Previous works have shown the convergence of (4) in both total variation distance ([3], [4]) and 2-Wasserstein distance ([5]). The approach in these papers relies on first showing the convergence of (1), and then bounding the discretization error between (4) and (2).

In this paper, our main goal is to establish the convergence of $\mathbf{p}_t$ in (4) in $\mathcal{KL}(\mathbf{p}_t \| \mathbf{p}^*)$. KL-divergence is perhaps the most natural notion of distance between probability distributions in this context, because of its close relationship to maximum likelihood estimation, its interpretation as information gain in Bayesian statistics, and its central role in information

theory. Convergence in KL-divergence implies convergence in total variation and 2-Wasserstein distance, thus we are able to obtain convergence rates in total variation and 2-Wasserstein that are comparable to the results shown in ([3], [4], [5]).

2. Related Work

The first non-asymptotic analysis of the discrete Langevin diffusion (4) was due to Dalalyan in [3]. This was soon followed by the work by Durmus and Moulines in [4], which improved upon the results in [3]. Subsequently, Durmus and Moulines also established convergence of (4) for the 2-Wasserstein distance in [5]. We remark that the proofs of Lemma 7, 11 and 13 are essentially taken from [5].

In a slightly different direction from the goals of this paper, Bubeck et al [6] and Durmus et al [9] studied variants of (4) which work when $-\log \mathbf{p}^*$ is not smooth. This is important, for example, when we want to sample from the uniform distribution over some convex set, so $-\log \mathbf{p}^*$ is the indicator function.

Very recently, Dalalyan et al [13] proved the convergence of Langevin Monte Carlo when only stochastic gradients are available.

Our work also borrows heavily from the theory established in the book of Ambrosio, Gigli and Savare [1], which studies the underlying probability distribution $\bar{\mathbf{p}}_t$ induced by (1) as a gradient flow in probability space. This allows us to view (4) as a *deterministic* convex optimization procedure over the probability space, with KL-divergence as the objective. This beautiful line of work relating SDEs with gradient flows in probability space was begun by Jordan, Kinderlehrer and Otto [2]. We refer any interested reader to an excellent survey by Santambrogio in [10].

Finally, we remark that the theory in [1] has some very interesting connections with the study of normalization flows in [7] and [8]. For example, the tangent velocity of (2), given by $v_t = \nabla \log \mathbf{p}^* - \nabla \log \mathbf{p}_t$, can be thought of as a deterministic transformation that induces a normalizing flow.

3. Our Contribution

In this section, we compare the results we obtain with those in [3], [4] and [5].

Our main contribution is establishing the first nonasymptotic convergence Kullback-Leibler divergence for (4) when $U(x)$ is m strongly convex and L smooth. (see Theorem 1). As a consequence, we also unify the proof of convergence in total variation and W_2 as simple corollaries to the convergence in KL.

The following table compares the number of iterations of (3) required to achieve ϵ error in each of the three quantities according to the analysis of various papers.

Table 1: Comparison of iteration complexity

	TV	W_2	KL
[3], [4]	$\tilde{O}(\frac{d}{\epsilon^2})$	-	-
[5]	$\tilde{O}(\frac{d}{\epsilon^2})$	$\tilde{O}(\frac{d}{\epsilon^2})$	-
this paper	$\tilde{O}(\frac{d}{\epsilon^2})$	$\tilde{O}(\frac{d}{\epsilon^2})$	$\tilde{O}(\frac{d}{\epsilon})$

In Section 7, we also state a convergence result for when U is not strongly convex. The corollary for convergence in total variation has a better dependence on the dimension than the corresponding result in [3], but a worse dependence on ϵ.

4. Definitions

We denote by $\mathscr{P}(\mathbb{R}^d)$ the space of all probability distributions over $\mathbb{R}^d$. In the rest of this paper, only distributions with densities wrt the Lebesgue measure will appear (see Lemma 16), both in the algorithm and in the analysis. With abuse of notation, we use the same symbol (e.g. $\mathbf{p}$) to denote both the probability distribution and its density wrt the Lebesgue measure.

We let B_t be the d-dimensional Brownian motion.

Let $\mathbf{p}^*$ be the target distribution such that $U(x) = -\log \mathbf{p}^*(x) + C$ has L Lipschitz continuous gradients and m strong convexity, i.e. for all x:

$$mI \preceq \nabla^2 U(x) \preceq LI$$

For a given initial distribution $\bar{\mathbf{p}}_0$, the **Exact Langevin Diffusion** is given by the following stochastic differential equation (recall $U(x) - \log \mathbf{p}^*(x)$):

$$\bar{x}_0 \sim \bar{\mathbf{p}}_0$$
$$d\bar{x}_t = -\nabla U(\bar{x}_t)dt + \sqrt{2}dB_t \tag{2}$$

(This is identical to (1), restated here for ease of reference.) For a given initial distribution $\mathbf{p}_0$, and for a given stepsize h, the **Langevin MCMC Algorithm** is given by the following:

$$u^0 \sim \mathbf{p}_0$$
$$u^{i+1} = u^i - h \cdot \nabla U(u^i) + \sqrt{2h}\xi^i \tag{3}$$

Where $\xi^i \overset{iid}{\sim} N(0,1)$.

For a given initial distribution $\mathbf{p}_0$ and stepsize h, the **Discretized Langevin Diffusion** is given by the following SDE:

$$x_0 \sim \mathbf{p}_0 \tag{4}$$
$$dx_t = -\nabla U(x_{\tau(t)})dt + \sqrt{2}dB_t$$

Let $\mathbf{p}_t$ denote the distribution of x_t

3

Where $\tau(t) \triangleq \lfloor \frac{t}{h} \rfloor \cdot h$ (note that $\tau(t)$ is parametrized by h). It is easily verified that for any i, x_{ih} from (4) is equivalent to u^i in (3). Note that the difference between (2) and (4) is in the drift term: one is $\nabla U(\bar{x}_t)$, the other is $\nabla U(x_{\tau(t)})$

For the rest of this paper, we will use $\mathbf{p}_t$ to exclusively denote the distribution of x_t in (4).

We assume without loss of generality that

$$\arg \min_x U(x) = 0$$

, and that

$$U(0) = 0$$

. (We can always shift the space to achieve this, and the minimizer of U is easy to find using, say, gradient descent.)

For the rest of this paper, we will let

$$F(\boldsymbol{\mu}) = \begin{cases} \int \boldsymbol{\mu}(x) \log \left(\frac{\boldsymbol{\mu}(x)}{\mathbf{p}^*(x)} \right) dx, & \text{if } \boldsymbol{\mu} \text{ has density wrt} \\ & \qquad \text{Lebesgue measure} \\ \infty & \text{else} \end{cases}$$

be the KL-divergence between $\boldsymbol{\mu}$ and $\mathbf{p}^*$. It is well known that F is minimized by $\mathbf{p}^*$, and $F(\mathbf{p}^*) = 0$.

Finally, given a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, we define the $L^2(\boldsymbol{\mu})$-norm of v as

$$\|v\|_{L^2(\boldsymbol{\mu})} \triangleq \sqrt{\mathbb{E}_{\boldsymbol{\mu}}[\|v(x)\|_2^2]}$$

4.1 Background on Wasserstein distance and curves in $\mathscr{P}(\mathbb{R}^d)$

Given two distributions $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathbb{R}^d)$, let $\Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$ be the set of all joint distributions over the product space $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals equal $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ respectively. (Γ is the set of all couplings)

The **Wasserstein distance** is defined as

$$W_2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sqrt{\inf_{\gamma \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})} \int (\|x - y\|_2^2) d\gamma(x, y)}$$

Let $(X_1, \mathscr{B}(X_1))$ and $(X_2, \mathscr{B}(X_2))$ be two measurable spaces, $\boldsymbol{\mu}$ be a measure, and $r : X_1 \to X_2$ be a measurable map. The **push-forward measure** of $\boldsymbol{\mu}$ through r is defined as

$$r_{\#}\boldsymbol{\mu}(B) = \boldsymbol{\mu}(r^{-1}(B)) \quad \forall B \in \mathscr{B}(X_2)$$

Intuitively, for any f, $\mathbb{E}_{r_{\#}\boldsymbol{\mu}}[f(x)] = \mathbb{E}_{\boldsymbol{\mu}}[f(r(x))]$.

It is a well known result that for any two distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ which have density wrt the Lebesgue measure, the optimal coupling is induced by a map $T_{opt} : \mathbb{R}^d \to \mathbb{R}^d$, i.e. $W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \int (\|x - y\|_2^2) d\gamma^*(x, y)$ for

$$\gamma^* = (Id, T_{opt})_{\#}\boldsymbol{\mu}$$

Where Id is the identity map, and T_{opt} satisfies $T_{opt\#}\boldsymbol{\mu} = \boldsymbol{\nu}$, so by definition, $\gamma^* \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$. We call T_{opt} **the optimal transport map**, and $T_{opt} - Id$ the **optimal displacement map**.

Given two points $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ in $\mathscr{P}(\mathbb{R}^d)$, a curve $\boldsymbol{\mu}_t : [0, 1] \to \mathscr{P}(\mathbb{R}^d)$ is a **constant-speed-geodesic** between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ if $\boldsymbol{\mu}_0 = \boldsymbol{\nu}$, $\boldsymbol{\mu}_1 = \boldsymbol{\pi}$ and $W_2(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t) = (t - s)W_2(\boldsymbol{\nu}, \boldsymbol{\pi})$ for all $0 \leq s \leq t \leq 1$. If $v_{\boldsymbol{\nu}}^{\boldsymbol{\pi}}$ is the optimal displacement map between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$, then the constant-speed-geodesic $\boldsymbol{\mu}_t$ is nicely characterized by

$$\boldsymbol{\mu}_t = (Id + tv_{\boldsymbol{\nu}}^{\boldsymbol{\pi}})_{\#}\boldsymbol{\nu} \tag{5}$$

Given a curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$, we define its **metric derivative** as

$$|\boldsymbol{\mu}_t'| \triangleq \limsup_{s \to t} \frac{W_2(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t)}{|s - t|} \tag{6}$$

. Intuitively, this is the speed of the curve in 2-Wasserstein distance. We say that a curve $\boldsymbol{\mu}_t$ is **absolutely continuous** if $\int_a^b |\boldsymbol{\mu}_t'|^2 < \infty$ for all $a, b \in \mathbb{R}$.

Given a curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$ and a sequence of velocity fields $v_t : \mathbb{R}^+ \to (\mathbb{R}^d \to \mathbb{R}^d)$, we say that $\boldsymbol{\mu}_t$ and v_t satisfy the **continuity equation** at t if

$$\frac{d}{dt}\boldsymbol{\mu}_t(x) + \nabla \cdot (\boldsymbol{\mu}_t(x) \cdot v_t(x)) = 0 \tag{7}$$

(We assume that $\boldsymbol{\mu}_t$ has density wrt Lebesgue measure for all t)

Remark 1 *If $\boldsymbol{\mu}_t$ is a constant-speed-geodesic between $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$, then $v_{\boldsymbol{\nu}}^{\boldsymbol{\pi}}$ satisfies (7) at $t = 0$, by the characterization in (5).*

We say that v_t is tangent to $\boldsymbol{\mu}_t$ at t if the continuity equation holds and $\|v_t + w\|_{L^2(\boldsymbol{\mu}_t)} \leq \|v_t\|_{L^2(\boldsymbol{\mu}_t)}$ for all w such that $\nabla \cdot (\boldsymbol{\mu}_t \cdot w) = 0$. Intuitively, v_t is tangent to $\boldsymbol{\mu}_t$ if it minimizes $\|v_t\|_{L^2(\boldsymbol{\mu}_t)}$ among all velocity fields v that satisfy the continuity equation.

5. Preliminary Lemmas

This section presents some basic results needed for our main theorem.

5.1 Calculus over $\mathscr{P}(\mathbb{R}^d)$

In this section, we present some crucial Lemmas which allow us to study the evolution of $F(\boldsymbol{\mu}_t)$ along a curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$. These results are all immediate consequences of results proven in [1].

Lemma 1 *For any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, let $\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu}) : \mathbb{R}^d \to \mathbb{R}$ be the first variation of F at $\boldsymbol{\mu}$ defined as $\left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu})\right)(x) \triangleq \log\left(\frac{\boldsymbol{\mu}(x)}{\mathbf{p}^*(x)}\right) + 1$. Let the subdifferential of F at $\boldsymbol{\mu}$ be given by*

$$w_{\boldsymbol{\mu}} \triangleq \nabla\left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu})\right) : \mathbb{R}^d \to \mathbb{R}^d$$

. For any curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$, and for any v_t that satisfies the continuity equation for $\boldsymbol{\mu}_t$ (see equation (7)), the following holds:

$$\frac{d}{dt}F(\boldsymbol{\mu}_t) = \mathbb{E}_{\boldsymbol{\mu}_t}\left[\langle w_{\boldsymbol{\mu}_t}(x), v_t(x)\rangle\right]$$

Based on Lemma 1, we define (for any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$) the operator

$$\mathcal{D}_{\boldsymbol{\mu}}(v) \triangleq \mathbb{E}_{\boldsymbol{\mu}}\left[\langle w_{\boldsymbol{\mu}}(x), v(x)\rangle\right] : (\mathbb{R}^d \to \mathbb{R}^d) \to \mathbb{R} \tag{8}$$

$\mathcal{D}_{\boldsymbol{\mu}}(v)$ **is linear in** v.

Lemma 2 *Let $\boldsymbol{\mu}_t$ be an absolutely continuous curve in $\mathscr{P}(\mathbb{R}^d)$ with tangent velocity field v_t. Let $|\boldsymbol{\mu}'_t|$ be the metric derivative of $\boldsymbol{\mu}_t$.*

Then

$$\|v_t\|_{L^2(\boldsymbol{\mu}_t)} = |\boldsymbol{\mu}'_t|$$

Lemma 3 *For any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, let $\|\mathcal{D}_{\boldsymbol{\mu}}\|_* \triangleq \sup_{\|v\|_{L^2(\boldsymbol{\mu})} \leq 1} \mathcal{D}_{\boldsymbol{\mu}}(v)$, then*

$$\|\mathcal{D}_{\boldsymbol{\mu}}\|_* = \sqrt{\int \left\|\nabla\left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu})\right)(x)\right\|_2^2 \boldsymbol{\mu}(x)dx}$$

Furthermore, for any absolutely continuous curve $\boldsymbol{\mu}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$ with tangent velocity v_t, we have

$$\left|\frac{d}{dt}F(\boldsymbol{\mu}_t)\right| \leq \|\mathcal{D}_{\boldsymbol{\mu}_t}\|_* \|v_t\|_{L^2(\boldsymbol{\mu}_t)}$$

As a Corollary of Lemma 2 and Lemma 3, we have the following result:

Corollary 4 *Let $\boldsymbol{\mu}_t$ be an absolutely continuous curve with tangent velocity field v_t. Then*

$$\frac{d}{dt}F(\boldsymbol{\mu}_t) \leq \|\mathcal{D}_{\boldsymbol{\mu}_t}\|_* \cdot |\boldsymbol{\mu}'_t|$$

5.2 Exact and Discrete Gradient Flow for $F(\mathbf{p})$

In this section, we will study the curve $\mathbf{p}_t : \mathbb{R}^+ \to \mathscr{P}(\mathbb{R}^d)$ defined in (4). Unless otherwise specified, we will assume that $\mathbf{p}_0$ is an arbitrary distribution.

Let x_t be as defined in (4).

For any given t and for all s, we define a stochastic process y_s^t as

$$y_s^t = x_s \qquad\qquad\qquad\qquad \text{for } s \leq t$$
$$dy_s^t = -\nabla U(y_s^t)ds + \sqrt{2}dB_s \qquad\qquad\qquad \text{for } s \geq t \qquad (9)$$
$$\text{let } \mathbf{q}_s^t \text{ denote the distribution for } y_s^t$$

From $s = t$ onwards, this is the exact Langevin diffusion with $\mathbf{p}_t$ as the initial distribution (compare with expression (2)).

Finally, for each t, we define a sequence z_s^t by

$$z_s^t = x_s \qquad\qquad\qquad\qquad \text{for } s \leq t$$
$$dz_s^t = (-\nabla U(z_{\tau(t)}^t) + \nabla U(z_s^t))ds, \qquad\qquad\qquad \text{for } s \geq t \qquad (10)$$
$$\text{let } \mathbf{g}_s^t \text{ denote the distribution for } z_s^t$$

z_s^t represents the discretization error of $\mathbf{p}_s$ through the divergence between $\mathbf{q}_s^t$ and $\mathbf{p}_s$ (formally stated in Lemma 5). Note that $z_{\tau(t)}^t = x_{\tau(t)}^t$ because $\tau(t) \leq t$.

Remark 2 *The the B_s in (4) and (9) are the same. Thus, x_s (from (4)), y_s^t (from (9)) and z_s^t (from (10)) define a coupling between the the curves $\mathbf{p}_s$, $\mathbf{q}_s^t$ and $\mathbf{g}_s^t$.*

Our proof strategy is as follows:

1. In Lemma 5, we demonstrate that the divergence between $\mathbf{p}_s$ (discretized Langevin) and $\mathbf{q}_s^t$ (exact Langevin) can be represented as a curve $\mathbf{g}_s^t$.

2. In Lemma 6, we demonstrate that the "decrease in $F(\mathbf{p}_t)$ due to exact Langevin" given by $\frac{d}{ds}F(\mathbf{q}_s^t)\big|_{s=t}$ is sufficiently negative.

3. In Lemma 7, we show that the "discretization error" given by $\frac{d}{ds}(F(\mathbf{p}_s) - F(\mathbf{q}_s^t))\big|_{s=t}$ is small.

4. Added together, they imply that $\frac{d}{ds}F(\mathbf{p}_s)\big|_{s=t}$ is sufficiently negative.

Lemma 5 *For all $x \in \mathbb{R}^d$ and $t \in \mathbb{R}^+$*

$$\frac{d}{ds}\mathbf{g}_s^t(x)\bigg|_{s=t} = \left(\frac{d}{ds}\mathbf{p}_s(x) - \frac{d}{ds}\mathbf{q}_s^t(x)\right)\bigg|_{s=t}$$

Lemma 6 *For all $s, t \in \mathbb{R}^+$*

$$\frac{d}{ds}F(\mathbf{q}_s^t) = -\|\mathcal{D}_{\mathbf{q}_s^t}\|_*^2$$

Lemma 7 *For all $t \in \mathbb{R}^+$*

$$\frac{d}{ds}\left(F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\right)\bigg|_{s=t} \leq \left(2L^2h\sqrt{\mathbb{E}_{\mathbf{p}_{\tau(t)}}\left[\|x\|_2^2\right]} + 2L\sqrt{hd}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_*$$

6. Strong Convexity Result

In this section, we study the consequence of assuming m strong convexity and L smoothness of $U(x)$.

6.1 Theorem statement and discussion

Theorem 1 *Let x_t and $\mathbf{p}_t$ be as defined in (4) with $\mathbf{p}_0 = N(0, \frac{1}{m})$.*

If

$$h = \frac{m\epsilon}{16dL^2}$$

and

$$k = 16\frac{L^2}{m^2}\frac{d\log\frac{dL}{m\epsilon}}{\epsilon}$$

Then $\mathcal{KL}\left(\mathbf{p}_{kh}\|\mathbf{p}^*\right) \le \epsilon$

The above theorem immediately allows us to obtain the convergence rate of $\mathbf{p}_{kh}$ in both total variation and 2-Wasserstein distance.

Corollary 8 *Using the choice of k and h in Theorem 1, we get*

1. $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \le \sqrt{\epsilon}$

2. $W_2(\mathbf{p}_{kh}, \mathbf{p}^*) \le \sqrt{\frac{2\epsilon}{m}}$

The first item follows from Pinsker's inequality. The second item follows from (12), where we take $\boldsymbol{\mu}_0$ to be $\mathbf{p}^*$ and $\boldsymbol{\mu}_1$ to be $\mathbf{p}_{kh}$, and noting that $\mathcal{D}_{\mathbf{p}^*} = 0$. To achieve δ accuracy in Total Variation or W_2, we apply Theorem 1 with $\epsilon = \delta^2$ and $\epsilon = m\delta^2$ respectively.

Remark 3 *The* log *term in Theorem 1 is not crucial. One can run (3) a few times, each time aiming to only halve the objective $F(\mathbf{p}_t) - F(\mathbf{p}^*)$ (thus the stepsize starts out large and is also halved each subsequent run). The proof is quite simple and will be omitted.*

6.2 Proof of Theorem 1

We now state the Lemmas needed to prove Theorem 1. We first establish a notion of strong convexity of $F(\boldsymbol{\mu})$ with respect to W_2 metric.

Lemma 9 *If $\log \mathbf{p}^*(x)$ is m strongly convex, then*

$$F(\boldsymbol{\mu}_t) \le (1-t)F(\boldsymbol{\mu}_0) + tF(\boldsymbol{\mu}_1) - \frac{m}{2}t(1-t)W_2^2(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) \tag{11}$$

for all $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathscr{P}(\mathbb{R}^d)$ and $t \in [0,1]$, let $\boldsymbol{\mu}_t : [0,1] \to \mathscr{P}(\mathbb{R}^d)$ be the constant-speed geodesic between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. (recall from (5) that If $v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1}$ is the optimal displacement map from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_1$, then $\boldsymbol{\mu}_t = (Id + t \cdot v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1})_{\#}\boldsymbol{\mu}_0$.)

Equivalently,

$$F(\boldsymbol{\mu}_1) \ge F(\boldsymbol{\mu}_0) + \mathcal{D}_{\boldsymbol{\mu}_0}(v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1}) + \frac{m}{2}W_2^2(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) \tag{12}$$

We call this the m-strong-geodesic-convexity of F wrt the W_2 distance.

The above Lemma is essentially implied by well known results on strong-geodesic-convexity, see for example [1].

Next, we use the m strong geodesic convexity of F to upper bound $F(\boldsymbol{\mu}) - F(\mathbf{p}^*)$ by $\frac{1}{2m}\|\mathcal{D}_{\boldsymbol{\mu}}\|_*^2$ (for any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$). This is analogous to how $f(x) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x)\|_2^2$ for standard m-strongly-convex functions in R^d.

Lemma 10 *Under our assumption that* $-\log \mathbf{p}^*(x)$ *is* m *strongly convex, we have that for all* $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$,

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq \frac{1}{2m}\|\mathcal{D}_{\boldsymbol{\mu}}\|_*^2$$

Now, recall $\mathbf{p}_t$ from (4). We use strong convexity to obtain a bound on $\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right]$ for all t. This will be important for bounding the discretization error in conjunction with Lemma 7

Lemma 11 *Let* $\mathbf{p}_t$ *be as defined in* (4). *If* $\mathbf{p}_0$ *is such that* $\mathbb{E}_{\mathbf{p}_0}\left[\|x\|_2^2\right] \leq \frac{4d}{m}$, *and* $h \leq \frac{1}{L}$ *in the definition of* (4), *then for all* $t \in \mathbb{R}^+$,

$$E_{\mathbf{p}_t}\|x\|^2 \leq \frac{4d}{m}$$

Finally, we put everything together to prove Theorem 1.

Proof of Theorem 1

We first note that $h = \frac{m\epsilon}{16L^2} \leq \frac{1}{L}$.

By Lemma 11, for all t, $\mathbb{E}_{\mathbf{p}_t}\left[\|x\|_2^2\right] \leq \frac{4d}{m}$. Combined with Lemma 7, we get that for all $t \in \mathbb{R}^+$

$$\frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\bigg|_{s=t} \leq \left(4L^2h\sqrt{\frac{d}{m}} + 2L\sqrt{hd}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_*$$

Suppose that $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq \epsilon$, and let

$$h = \frac{m\epsilon}{16dL^2} \leq \frac{1}{16}\min\left\{\frac{m}{L^2}\sqrt{\frac{\epsilon}{d}}, \frac{m\epsilon}{L^2d}\right\}$$

then $\forall t$

$$\frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\bigg|_{s=t} \leq \left(4L^2h\sqrt{\frac{d}{m}} + 2L\sqrt{hd}\right)$$
$$\leq \frac{1}{2}\sqrt{m\epsilon}\|\mathcal{D}_{\mathbf{p}_t}\|_* \leq \frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2$$

Where the last inequality is because Lemma 10 and the assumption that $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq \epsilon$ together imply that $\|\mathcal{D}_{\mathbf{p}_t}\|_* \geq \sqrt{2m\epsilon}$.

9

So combining Lemma 6 and Lemma 5, we have

$$\frac{d}{dt}F(\mathbf{p}_t) = \left.\frac{d}{ds}F(\mathbf{q}_s^t)\right|_{s=t} + \left.\frac{d}{ds}F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\right|_{s=t}$$

$$\leq -\|\mathcal{D}_{\mathbf{p}_t}\|_*^2 + \frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2$$

$$= -\frac{1}{2}\|\mathcal{D}_{\mathbf{p}_t}\|_*^2$$

$$\leq -m(F(\mathbf{p}_t) - F(\mathbf{p}^*)) \tag{13}$$

Where the last line once again follows from Lemma 10.

To handle the case when $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq \epsilon$, we use the following argument:

1. We can conclude that $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq \epsilon$ implies $\frac{d}{dt}F(\mathbf{p}_t) \leq 0$.

2. By the results of Lemma 16 and Lemma 17, for all t, $|\mathbf{p}_t'|$ is finite and $\|\mathcal{D}_{\mathbf{p}_t}\|$ is finite, so $\frac{d}{dt}F(\mathbf{p}_t)$ is finite and $F(\mathbf{p}_t)$ is continuous in t.

3. Thus, if $F(\mathbf{p}_t) \leq \epsilon$ for some $t \leq kh$, then $F(\mathbf{p}_s) \leq \epsilon$ for all $s \geq t$ as $F(\mathbf{p}_t) \geq \epsilon$ implies $\frac{d}{dt}F(\mathbf{p}_t) \leq 0$ and $F(\mathbf{p}_t)$ is continuous in t. Thus $F(\mathbf{p}_{kh}) - F(\mathbf{p}^*) \leq \epsilon$.

Thus, we need only consider the case that $F(\mathbf{p}_t) \geq \epsilon$ for all $t \leq kh$. This means that (13) holds for all $t \leq kh$.

By Gronwall's inequality, we get

$$F(\mathbf{p}_{kh}) - F(\mathbf{p}^*) \leq (F(\mathbf{p}_0) - F(\mathbf{p}^*))\exp(-mkh)$$

We thus need to pick

$$k = \frac{\frac{1}{m}\log\frac{F(\mathbf{p}_0)-F(\mathbf{p}^*)}{\epsilon}}{h} = 16\frac{L^2}{m^2}\frac{d\log\frac{F(\mathbf{p}_0)-F(\mathbf{p}^*)}{\epsilon}}{\epsilon}$$

Using the fact that $\mathbf{p}_0 = N(0, \frac{1}{m})$. Using L-smoothness and m-strong convexity, we can show that

$$-\log\mathbf{p}^*(x) \leq \frac{L}{2}\|x\|_2^2 + \frac{d}{2}\log(\frac{2\pi}{m})$$

, and

$$\log\mathbf{p}_0(x) = -\frac{m}{2}\|x\|_2^2 - \frac{d}{2}\log(\frac{2\pi}{m})$$

. We thus get that $F(\mathbf{p}_0) - F(\mathbf{p}^*) = \mathcal{KL}(\mathbf{p}_0\|\mathbf{p}^*) \leq \frac{dL}{m}$, so

$$k = 16\frac{L^2}{m^2}\frac{d\log\frac{dL}{m\epsilon}}{\epsilon}$$

■

10

7. Weak convexity result

In this section, we study the case when $\log \mathbf{p}^*$ is not m strongly convex (but still convex and L smooth). Let $\boldsymbol{\pi}_h$ be the stationary distribution of (4) with stepsize h.

We will assume that we can choose an initial distribution $\mathbf{p}_0$ which satisfies

$$W_2(\mathbf{p}_0, \mathbf{p}^*) = C_1 \tag{14}$$

and

$$\sqrt{E_{\mathbf{p}^*} \|x\|_2^2} = C_2 \tag{15}$$

. Let h' be the largest stepsize such that

$$W_2(\boldsymbol{\pi}_h, p^*) \le C_1 \quad , \forall h \le h' \tag{16}$$

7.1 Theorem statement and discussion

Theorem 2 *Let C_1, C_2 and h' be defined as in the beginning of this section.*

Let x_t and $\mathbf{p}_t$ be as defined in (4) with $\mathbf{p}_0$ satisfying (14). If

$$h = \frac{1}{48} \min\left\{ \frac{\epsilon}{C_1(C_1 + C_2)L^2}, \frac{\epsilon^2}{C_1^2 dL^2}, h' \right\}$$

$$= \frac{1}{48} \min\left\{ \frac{\epsilon}{C_1 C_2 L^2}, \frac{\epsilon^2}{C_1^2 dL^2}, h' \right\}$$

and

$$k = \frac{2C_1^2}{\epsilon h} + \frac{2C_1^2 \log(F(\mathbf{r}^0) - F(\mathbf{p}^*))}{h}$$

Then $\mathcal{KL}\left(\mathbf{p}_{kh} \| \mathbf{p}^*\right) \le \epsilon$

Once again, applying Pinsker's inequality, we get that the above choice of k and t yields $d_{TV}(\mathbf{r}^k, \mathbf{p}^*) \le \sqrt{\epsilon}$. Without strong convexity, we cannot get a bound on W_2 from bounding $F(\mathbf{r}^k) - F(\mathbf{p}^*)$ like we did in corollary 8.

In [3], a proof in the non-strongly-convex case was obtained by running Langevin MCMC on

$$\tilde{\mathbf{p}}^* \propto \mathbf{p}^* \cdot \exp(-\frac{\delta}{d} \|x\|_2^2)$$

$\log \tilde{\mathbf{p}}^*$ is thus strongly convex with $m = \frac{\delta}{d}$, and $d_{TV}(\mathbf{p}^*, \tilde{\mathbf{p}}^*) \le \delta$. By the results of [3], or [4], or Theorem 1, we need

$$k = \tilde{O}(\frac{d^3}{\delta^4}) \tag{17}$$

iterations to get $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \le \delta$.

11

On the other hand, if we assume $\log(F(\mathbf{p}_0) - F(\mathbf{p}^*)) \leq \frac{1}{\epsilon}$ and $h' \geq \frac{1}{10} \min\left\{\frac{\epsilon}{C_1 C_2 L^2}, \frac{\epsilon^2}{C_1^2 d L^2}\right\}$ the results of Theorem 2 implies that

$$h = \frac{\epsilon}{L^2 C_1} \min\left\{\frac{1}{C_2}, \frac{\epsilon}{dC_1}\right\}$$

To get $d_{TV}(\mathbf{p}_{kh}, \mathbf{p}^*) \leq \delta$, we need

$$k = \frac{L^2 C_1^3}{\delta^4} \max\left\{C_2, \frac{dC_1}{\delta^2}\right\}$$

Even if we ignore C_1 and C_2, our result is not strictly better than (17) as we have a worse dependence on δ. However, we do have a better dependence on d.

The proof of Theorem 2 is quite similar to that of Theorem 1, so we defer it to the appendix.

Acknowledgments

References

[1] Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré. Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2008.

[2] Jordan, Richard, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation." SIAM journal on mathematical analysis 29.1 (1998): 1-17.

[3] Dalalyan, Arnak S. "Theoretical guarantees for approximate sampling from smooth and log-concave densities." Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2016).

[4] Durmus, Alain, and Eric Moulines. "Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm." arXiv preprint arXiv:1507.05021 (2015).

[5] Durmus, Alain, and Eric Moulines. "High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm." (2016).

[6] Bubeck, Sébastien, Ronen Eldan, and Joseph Lehec. "Sampling from a log-concave distribution with Projected Langevin Monte Carlo." Advances in Neural Information Processing Systems 28 (2015).

[7] Danilo Jimenez Rezende and Shakir Mohamed. "Variational inference with normalizing flows" Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (2015)

[8] Liu, Qiang, and Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm." Advances In Neural Information Processing Systems. 2016.

[9] Durmus, Alain, Eric Moulines, and Marcelo Pereyra. "Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau." arXiv preprint arXiv:1612.07471 (2016).

[10] Santambrogio, Filippo. "Euclidean, metric, and Wasserstein gradient flows: an overview." Bulletin of Mathematical Sciences 7.1 (2017): 87-154.

[11] Santambrogio, Filippo. "Optimal transport for applied mathematicians." Birkäuser, NY (2015).

[12] Eberle, Andreas. "Reflection couplings and contraction rates for diffusions." Probability theory and related fields 166.3-4 (2016): 851-886.

[13] Dalalyan, Arnak S., and Avetik G. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." arXiv preprint arXiv:1710.00095 (2017).

8. Supplementary Materials

Proof of Lemma 1 The proof is directly from results in [1]. See Theorem 10.4.9, with $\mathcal{F}(\mu|\gamma) = \mathcal{KL}(\mu\|\gamma)$, with $\mu = \boldsymbol{\mu}$, $\gamma = \mathbf{p}^*$, $\sigma = \frac{\boldsymbol{\mu}}{\mathbf{p}^*}$, $F(\rho) = \rho\log\rho$, $L_F(\sigma) = \sigma$, and $w_{\boldsymbol{\mu}} = \frac{\nabla L_F(\sigma)}{\sigma} = \nabla\log\frac{\boldsymbol{\mu}}{\mathbf{p}^*}$. The expression for $\frac{d}{dt}F(\mathbf{p}_t)$ comes from expression 10.1.16 (section E of chapter 10.1.2, page 233). See also expressions 10.4.67 and 10.4.68.

(One can also refer to Theorem 10.4.13 and Theorem 10.4.17 for proofs of $w_{\boldsymbol{\mu}}$ for the KL-divergence functional in more general settings.) By Lemma 16, $w_{\mathbf{p}_t}$ is well defined for all t. ∎

Proof of Lemma 2 Theorem 8.3.1 of [1].

∎

Proof of Lemma 3 By definition of $\mathcal{D}_{\boldsymbol{\mu}_t}(v)$ in (8) and Lemma 1 and Cauchy Schwarz. ∎

Proof of Lemma 5 In this proof, we treat t as a fixed but arbitrary number, and prove the Lemma for all $t \in \mathbb{R}^+$. We will use x_s, y_s^t, z_s^t, $\mathbf{p}_s$, $\mathbf{q}_s^t$ and $\mathbf{g}_s^t$ as defined in (4), (9) and (10).

First, consider the case when $t = \tau(t)$. By definition, $x_t = y_t^t = z_t^t$, and $\mathbf{p}_t = \mathbf{q}_t^t = \mathbf{g}_t^t$. By Fokker Planck,

$$
\begin{aligned}
\frac{d}{ds}\mathbf{p}_s(x)\Big|_{s=t} &= -\nabla U(x_t) + \operatorname{tr}(\nabla^2\mathbf{p}_t) \\
&= -\nabla U(y_t^t) + \operatorname{tr}(\nabla^2\mathbf{q}_t^t) \\
&= \frac{d}{ds}\mathbf{q}_s^t(x)\Big|_{s=t}
\end{aligned}
$$

On the other hand

$$
dz_s^t\big|_{s=t} = -\nabla U(z_{\tau(t)}^t) + \nabla U(z_t^t) = -\nabla U(x_t) + \nabla U(x_t) = 0
$$

Thus $\frac{d}{ds}\mathbf{g}_s^t\big|_{s=t} = 0$ So Lemma 5 holds.

In the remainder of this proof, we assume that $t \neq \tau(t)$.

For a given $\Theta \in \mathbb{R}^{2d}$, we let $\Pi_1(\Theta)$ denote the projection of Θ onto its first d coordinates, and $\Pi_2(\Theta)$ denote the projection of Θ onto its last d coordinates. With abuse of notation, for $\mathbf{P} \in \mathscr{P}(\mathbb{R}^{2d})$, we let $\Pi_1(\mathbf{P})$ and $\Pi_2(\mathbf{P})$ denote the corresponding marginal densities.

We will consider three stochastic processes: $\Theta_s, \Lambda_s^t, \Psi_s^t$ over $\mathbb{R}^{2d}$ for $s \in [\tau(t), \tau(t)+h)$.

14

First, we introduce the stochastic process Θ_s for $s \in [\tau(t), \tau(t) + h)$

$$\Theta_{\tau(t)} = \begin{bmatrix} x_{\tau(t)} \\ -\nabla U(x_{\tau(t)}) \end{bmatrix}$$

$$d\Theta_s = \begin{bmatrix} \Pi_2(\Theta_s) \\ 0 \end{bmatrix} dt + \begin{bmatrix} \sqrt{2}dB_t \\ 0 \end{bmatrix} \quad \text{for } s \in [\tau(t), \tau(t) + h)$$

We let $\mathbf{P}_s$ denote the density for Θ_s. Intuitively, $\mathbf{P}_s$ is the joint density between x_s and $-\nabla U(x_{\tau(t)})$. One can verify that $\Pi_1(\Theta_s) = x_s$ and $\Pi_1(\mathbf{P}_s) = \mathbf{p}_s$. By Fokker-Planck, we have $\forall \Theta \in \mathbb{R}^{2d}$

$$\frac{d}{ds}\mathbf{P}_s(\Theta)\Big|_{s=t} = -\nabla \cdot \left(\mathbf{P}_t(\Theta) \cdot \begin{bmatrix} \Pi_2(\Theta) \\ 0 \end{bmatrix} \right)$$

$$+ \sum_{i=1}^d \frac{\partial^2}{\partial \Theta_i^2} \mathbf{P}_t(\Theta) \tag{18}$$

Next, for any given t, we introduce the stochastic process Λ_s^t for $s \in [\tau(t), \tau(t) + h)$.

$$\Lambda_s^t = \Theta_s \qquad\qquad\qquad \text{for } s \leq t$$

$$d\Lambda_s^t = \begin{bmatrix} -\nabla U(\Pi_1(\Lambda_s^t)) \\ 0 \end{bmatrix} ds + \begin{bmatrix} \sqrt{2}dB_s \\ 0 \end{bmatrix} \qquad\qquad \text{for } s \geq t$$

Let $\mathbf{Q}_s^t$ denote the density for Λ_s^t. One can verify that $\Pi_1(\Lambda_s^t) = y_s^t$ and $\Pi_1(\mathbf{Q}_s^t) = \mathbf{q}_s^t$. By Fokker-Planck, we have $\forall \Theta \in \mathbb{R}^{2d}$

$$\frac{d}{ds}\mathbf{Q}_s^t(\Theta)\Big|_{s=t} = -\nabla \cdot \left(\mathbf{Q}_t^t(\Theta) \cdot \begin{bmatrix} -\nabla U(\Pi_1(\Theta)) \\ 0 \end{bmatrix} \right)$$

$$+ \sum_{i=1}^d \frac{\partial^2}{\partial \Theta_i^2} \mathbf{Q}_t^t(\Theta) \tag{19}$$

Finally, define

$$\Psi_s^t = \Theta_s \qquad\qquad\qquad \text{for } s \leq t$$

$$d\Psi_s^t = \begin{bmatrix} \Pi_2(\Psi_s^t) + \nabla U(\Pi_1(\Psi_s^t)) \\ 0 \end{bmatrix} ds$$

$$+ \begin{bmatrix} \sqrt{2}dB_s \\ 0 \end{bmatrix} \qquad\qquad \text{for } s \geq t$$

Let $\mathbf{G}_s^t$ denote the density for Ψ_s^t. One can verify that $\Pi_1(\Psi_s^t) = z_s^t$ and $\Pi_1(\mathbf{G}_s^t) = \mathbf{g}_s^t$. By Fokker-Planck, we have $\forall \Theta \in \mathbb{R}^{2d}$

$$\frac{d}{ds}\mathbf{G}_s^t(\Theta)\Big|_{s=t} = -\nabla \cdot \left(\mathbf{G}_t^t(\Theta) \cdot \begin{bmatrix} \Pi_2(\Theta) + \nabla U(\Pi_1(\Theta)) \\ 0 \end{bmatrix} \right) \tag{20}$$

15

By definition, $\Theta_t = \Lambda_t^t = \Psi_t^t$ almost surely, and $\mathbf{P}_t = \mathbf{Q}_t^t = \mathbf{G}_t^t$. Taking the difference between (18), (19) thus gives

$$\frac{d}{ds}\mathbf{P}_s(\Theta) - \mathbf{Q}_s^t(\Theta)\Big|_{s=t} = -\nabla \cdot \left(\mathbf{P}_t(\Theta) \cdot \begin{bmatrix} \Pi_2(\Theta) + \nabla U(\Pi_1(\Theta)) \\ 0 \end{bmatrix} \right)$$

$$= \frac{d}{ds}\mathbf{G}_s^t(\Theta)\Big|_{s=t}$$

Finally, marginalizing out the last d coordinates on both sides, and recalling that $\Pi_1(\mathbf{P}_s) = \mathbf{p}_s$, $\Pi_1(\mathbf{Q}_s^t) = \mathbf{q}_s^t$ and $\Pi_1(\mathbf{G}_s^t) = \mathbf{g}_s^t$, we prove the Lemma. ∎

Proof of Lemma 6 The fact that $\mathbf{q}_s^t$ is the steepest descent follows from the fact that Fokker-Planck equation for Langevin diffusion yields, for all $x \in \mathbb{R}^d$

$$\frac{d}{ds}\mathbf{q}_s^t(x) = \nabla \cdot (\mathbf{q}_s^t(x)\nabla \log \mathbf{p}^*(x)) + \mathrm{tr}(\nabla^2 \mathbf{q}_s^t(x))$$

$$= \nabla \cdot \left(\mathbf{q}_s^t(x) \left(\nabla \log \frac{\mathbf{p}^*(x)}{\mathbf{q}_s^t(x)} \right) \right)$$

By definition of (7), we get that

$$v_s = \nabla \log \frac{\mathbf{p}^*(x)}{\mathbf{q}_s^t(x)} \tag{21}$$

satisfies the continuity equation for $\mathbf{q}_s^t$. By Lemma 1,

$$w_{\mathbf{q}_s^t} = \nabla \log \left(\frac{\mathbf{q}_s^t}{\mathbf{p}^*} \right)$$

Thus

$$\frac{d}{ds}F(\mathbf{q}_s^t) = \mathcal{D}_{\mathbf{q}_s^t}(v_s) = -\mathbb{E}_{\mathbf{q}_s^t}\left[\|w_{\mathbf{q}_s^t}\|_2^2 \right] = -\|\mathcal{D}_{\mathbf{q}_s^t}\|_*^2$$

The last equality follows from the first statement of Lemma 3.

∎

Proof of Lemma 7 Consider z_s^t and $\mathbf{g}_s^t$ as defined in (10). By Lemma 5, $\frac{d}{ds}\mathbf{g}_s^t\big|_{s=t} = (\frac{d}{ds}\mathbf{P}_s - \frac{d}{ds}\mathbf{q}_s^t)\big|_{s=t}$. The first variation of F, defined by

$$\lim_{\epsilon \to 0} \frac{F(\boldsymbol{\mu} + \epsilon\Delta) - F(\boldsymbol{\mu})}{\epsilon} = \int \left(\frac{\delta F}{\delta \boldsymbol{\mu}}(\boldsymbol{\mu}) \right)(x) \cdot \Delta(x)dx$$

is linear (see Chapter 7.2 of [11]). (In the above, $\Delta : \mathbb{R}^d \to \mathbb{R}$ is an arbitrary 0-mean perturbation). In addition, because $\mathbf{p}_t = \mathbf{q}_t^t = \mathbf{g}_t^t$, we have $\frac{\delta F}{\delta \boldsymbol{\mu}}(\mathbf{p}_t) = \frac{\delta F}{\delta \boldsymbol{\mu}}(\mathbf{q}_t^t) = \frac{\delta F}{\delta \boldsymbol{\mu}}(\mathbf{g}_t^t)$, we get that

$$\frac{d}{ds}F(\mathbf{g}_s^t)\Big|_{s=t} = \left(\frac{d}{ds}F(\mathbf{p}_s) - \frac{d}{ds}F(\mathbf{q}_s^t) \right)\Big|_{s=t}$$

16

We will upper bound $|\mathbf{g}_s^{t\prime}|\big|_{s=t}$, then apply Corollary 4.

$$|\mathbf{g}_s^{t\prime}|\big|_{s=t}$$

$$=\lim_{\epsilon\to 0}\frac{1}{\epsilon}W_2(\mathbf{g}_{t+\epsilon}^t,\mathbf{g}_t^t)$$

$$\leq\lim_{\epsilon\to 0}\frac{1}{\epsilon}\sqrt{\mathbb{E}\left[\left\|\epsilon(\nabla U(x_t)-\nabla U(x_{\tau(t)}))\right\|_2^2\right]}$$

$$=\sqrt{\mathbb{E}\left[\left\|\nabla U(x_t)-\nabla U(x_{\tau(t)})\right\|_2^2\right]}$$

$$\leq\sqrt{\mathbb{E}\left[L^2\|x_t-x_{\tau(t)}\|_2^2\right]}$$

$$=L\sqrt{\mathbb{E}\left[\left\|(t-\tau(t))\nabla U(x_{\tau(t)})+\sqrt{2}(B_t-B_{\tau(t)})\right\|_2^2\right]}$$

$$\leq 2L(t-\tau(t))\sqrt{\mathbb{E}\left[\|\nabla U(x_{\tau(t)})\|_2^2\right]}+2L\sqrt{(t-\tau(t))d}$$

$$\leq 2L(t-\tau(t))\sqrt{L^2\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]}+2L\sqrt{(t-\tau(t))d}$$

Where the first line is by definition of metric derivative, second line is by the coupling between $\mathbf{g}_t^t$ and $\mathbf{g}_{t+\epsilon}^t$) induced by the joint distribution $(z_t^t, z_{t+\epsilon}^t)$ and the fact that $z_{\tau(t)}^t=x_t$. The fourth line is by Lipschitz-gradient of $U(x)$, fifth line is by definition of x_t, sixth line is by variance of B_t-B_0, seventh line is once again by Lipschitz-gradient of $U(x)$.

Thus, we upper bound $|\mathbf{g}_s^{t\prime}|\big|_{s=t}$ by $2L^2(t-\tau(t))\sqrt{\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]}+2L\sqrt{(t-\tau(t))d}$. Applying Corollary 4, and using the fact that for all t, $t-\tau(t)\leq h$, we get

$$\frac{d}{ds}F(\mathbf{g}_s^t)\bigg|_{s=t}$$

$$\leq\left(2L^2h\sqrt{\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]}+2L\sqrt{hd}\right)\|\mathcal{D}_{\mathbf{g}_t^t}\|_*$$

$$\leq\left(2L^2h\sqrt{\mathbb{E}\left[\|x_{\tau(t)}\|_2^2\right]}+2L\sqrt{hd}\right)\|\mathcal{D}_{\mathbf{p}_t}\|_*$$

The last line is because $\mathbf{g}_t^t=\mathbf{p}_t$ by definition.

∎

Proof of Lemma 9 By Theorem 9.4.11 and Proposition 9.3.2 of [1], m-strong-convexity of $\log \mathbf{p}^*$ implies geodesic convexity. Expression (11) then follows from the definition of geodesic convexity in definition 9.1.1 of [1].

Rearrranging terms, dividing by t and taking limit as $t\to 0$, we get

$$F(\boldsymbol{\mu}_1)\geq F(\boldsymbol{\mu}_0)+\lim_{t\to 0}\frac{F(\boldsymbol{\mu}_t)-F(\boldsymbol{\mu}_0)}{t}+\frac{m}{2}W_2^2(\boldsymbol{\mu}_0,\boldsymbol{\mu}_1)$$

$$=F(\boldsymbol{\mu}_0)+\mathcal{D}_{\boldsymbol{\mu}_0}(v_{\boldsymbol{\mu}_0}^{\boldsymbol{\mu}_1})+\frac{m}{2}W_2^2(\boldsymbol{\mu}_0,\boldsymbol{\mu}_1)$$

The last equality follows by Lemma 1 and by the remark immediately following (7).

We remark that the proof of (12) is completely analogous to the proof of first-order characterization of strongly convex functions over $\mathbb{R}^d$.

∎

Proof of Lemma 10 We consider (12), and use two facts

1. For any $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$, $\mathcal{D}_{\boldsymbol{\mu}}(v)$ is linear in v. (see (8))

2. For any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathbb{R}^d)$, $W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbb{E}_{\boldsymbol{\mu}} \|v_{\boldsymbol{\mu}}^{\boldsymbol{\nu}}(x)\|_2^2$, by definition of W_2 and $v_{\boldsymbol{\mu}}^{\boldsymbol{\nu}}$ as the optimal displacement map.

We apply Lemma 10 with $\boldsymbol{\mu}_0 = \mathbf{p}^*$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$. Let $v_{\boldsymbol{\mu}}^{\mathbf{p}^*}$ be the optimal displacement map from $\boldsymbol{\mu}$ to $\mathbf{p}^*$, so (12) gives

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq -\mathcal{D}_{\boldsymbol{\mu}}(v_{\boldsymbol{\mu}}^{\mathbf{p}^*}) - \frac{m}{2} W_2^2(\boldsymbol{\mu}, \mathbf{p}^*)$$

$$= -\mathcal{D}_{\boldsymbol{\mu}}(v_{\boldsymbol{\mu}}^{\mathbf{p}^*}) - \frac{m}{2} \mathbb{E}_{\boldsymbol{\mu}} \|v_{\boldsymbol{\mu}}^{\mathbf{p}^*}(x)\|_2^2$$

Let $v^* \triangleq \arg\max_{\|v\|_{L^2(\boldsymbol{\mu})} \leq 1} -\mathcal{D}_{\boldsymbol{\mu}}(v)$, so $\mathcal{D}_{\boldsymbol{\mu}}(v^*) = -\|\mathcal{D}_{\boldsymbol{\mu}}\|_*$ by linearity. We know that the maximizer of

$$\arg\max_v -\mathcal{D}_{\boldsymbol{\mu}}(v) - \frac{m}{2} \mathbb{E}_{\boldsymbol{\mu}} \|v_{\boldsymbol{\mu}}^{\mathbf{p}^*}(x)\|_2^2 = c \cdot v^*$$

for some real number c. Taking derivatives wrt c gives $c = \frac{1}{m} \|\mathcal{D}_{\boldsymbol{\mu}}\|_*$. Thus we get

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq \frac{m}{2} \|\mathcal{D}_{\boldsymbol{\mu}}\|_*^2$$

∎

Proof of Lemma 11 We prove this by induction on k. First, by definition of $\mathbf{p}_0 = N(0, \frac{1}{m})$, we get that

$$\mathbb{E}_{\mathbf{p}_t} \left[\|x\|_2^2 \right] = \frac{d}{m} \leq \frac{4d}{m} \quad, \forall t \leq 0h$$

Next, we assume that for some k, and for all $t \leq kh$, $\mathbb{E}_{\mathbf{p}_t} \left[\|x\|_2^2 \right] \leq \frac{4d}{m}$.

For the inductive step, we consider $t \in (kh, (k+1)h]$

From (4),

$$x_t = x_{kh} - (t - kh)\nabla U(x_{kh}) + \sqrt{2}(B_t - B_{kh})$$

By smoothness and strong convexity and the assumption that $\arg\min_x U(x) = 0$, we get that for all x and for all t:

$$\|(x - (t - kh)\nabla U(x)) - 0\|_2 \leq (1 - mt)\|x - 0\|_2$$

18

(note that $h \leq \frac{1}{L}$ implies that $t - kh \leq \frac{1}{L}$.) So for all t

$$\mathbb{E}_{x \sim \mathbf{p}_t} \|x\|_2^2$$
$$=\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x - (t-kh)\nabla U(x) + \sqrt{2}(B_t - B_{kh})\|_2^2$$
$$=\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x - (t-kh)\nabla U(x)\|_2^2 + \mathbb{E}\|\sqrt{2}(B_t - B_{kh})\|_2^2$$
$$\leq (1-mt)\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x\|_2^2 + 2dt$$
$$=\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x\|_2^2 + (2dt - mt\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x\|_2^2)$$

By inductive hypothesis, we have $\mathbb{E}_{x \sim \mathbf{p}_t} \|x\|_2^2 \leq \frac{4d}{m}$ for all $t \leq kh$

If $\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x\|_2^2 \geq \frac{2d}{m}$, then $\mathbb{E}_{x \sim \mathbf{p}_t} \|x\|_2^2 \leq \mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x\|_2^2 \leq \frac{4d}{m}$.

If $\mathbb{E}_{x \sim \mathbf{p}_{kh}} \|x\|_2^2 \leq \frac{2d}{m}$, then $\mathbb{E}_{x \sim \mathbf{p}_t} \|x\|_2^2 \leq \frac{2d}{m} + \frac{2d}{L} \leq \frac{4d}{m}$ (by $t - kh \leq \frac{1}{L}$ and by $L \geq m$).

Thus if $\mathbf{p}_{kh}$ is such that $E_{\mathbf{p}_{kh}} \|x\|_2^2 \leq \frac{4d}{m}$, then it must be that $E_{\mathbf{p}_t} \|x\|^2 \leq \frac{4d}{m}$ for all $t \in (kh, (k+1)h]$, thus proving the inductive step. ∎

8.1 Proof of Theorem 2

First, we present a Lemma for upper bounding $F(\boldsymbol{\mu}) - F(\mathbf{p}^*)$ for $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$ in the absence of strong convexity. The following Lemma plays an analogous role to Lemma 10.

Lemma 12 *Let F be convex in W_2, then for all $\boldsymbol{\mu} \in \mathscr{P}(\mathbb{R}^d)$,*

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq \|\mathcal{D}_{\boldsymbol{\mu}}\|_* W_2(\boldsymbol{\mu}, \mathbf{p}^*)$$

Proof of Lemma 12 Similar to the proof of Lemma 10, we consider (12), but with $m = 0$, (and once again $v_{\boldsymbol{\mu}}^{\mathbf{p}^*}$ denotes the optimal displacement map from $\boldsymbol{\mu}$ to $\mathbf{p}^*$):

$$F(\boldsymbol{\mu}) - F(\mathbf{p}^*) \leq -\mathcal{D}_{\mathbf{p}}(v_{\boldsymbol{\mu}}^{\mathbf{p}^*})$$
$$\leq \|\mathcal{D}_{\boldsymbol{\mu}}\|_* \cdot \|v_{\boldsymbol{\mu}}^{\mathbf{p}^*}\|_{L^2(\boldsymbol{\mu})}$$
$$\leq \|\mathcal{D}_{\boldsymbol{\mu}}\|_* \cdot W_2(\boldsymbol{\mu}, \mathbf{p}^*)$$

Where first inequality is from (12), second line is by definition of $\|\mathcal{D}_{\boldsymbol{\mu}}\|_*$, third line is by defintion of Wasserstein distance and the fact that $v_{\boldsymbol{\mu}}^{\mathbf{p}^*}$ is the optimal transport map.

∎

Next, we establish that for a fixed stepsize h, $W_2(\mathbf{p}_t, \boldsymbol{\pi}_h)$ is nonincreasing, using a synchronous coupling technique taken from [5].

Lemma 13 *Let $\mathbf{p}_t$ be defined as in the statement of Theorem (2). Let h be a fixed stepsize satisfying $h \leq \min\{\frac{1}{L}, h'\}$. Then for all k,*

$$W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) \leq W_2(\mathbf{p}_0, \boldsymbol{\pi}_h)$$

Proof of Lemma 13

First, we demonstrate that (4) is contractive in W_2.

We will prove this by induction.

Base case: trivially true.

Inductive Hypothesis: $W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) \leq W_2(\mathbf{p}_0, \boldsymbol{\pi}_h)$ for some k.

Inductive Step: Let T be the optimal transport map from $\mathbf{p}_{kh}$ to $\boldsymbol{\pi}_h$. We will demonstrate a coupling between $\mathbf{p}_{(k+1)h}$ and $\boldsymbol{\pi}_h$ with cost less than $W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h)$. The Lemma then follows from induction.

Since $x_{kh} \sim \mathbf{p}_{kh}$ (see (4)), the optimal coupling between $\mathbf{p}_{kh}$ and $\boldsymbol{\pi}_h$ is given by the pair of random variables $(x_{kh}, T(x_{kh}))$. For $t \in [kh, (k+1)h]$,

$$x_{(k+1)h} = x_{kh} - h\nabla U(x_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}))$$

. Consider the coupling γ between $\mathbf{p}_{kh}$ and $\boldsymbol{\pi}_h$ defined by the following pair of random variables

$$\Big(x_{kh} - h\nabla U(x_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}),$$
$$T(x_{kh}) - h\nabla U(T(x_{kh})) + \sqrt{2}(B_{(k+1)h} - B_{kh}) \Big)$$

(Note that $\boldsymbol{\pi}_h$ is stationary under the discrete Langevin diffusion with stepsize h, so γ does have the right marginals).

To demonstrate contraction in W_2:

$$W_2^2(\mathbf{p}_{(k+1)h}, \boldsymbol{\pi}_h)$$
$$\leq \mathbb{E}\left[\left\|\left(x_{kh} - h\nabla U(x_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})\right)\right.\right.$$
$$\left.\left. - \left(T(x_{kh}) - h\nabla U(T(x_{kh})) + \sqrt{2}(B_{(k+1)h} - B_{kh})\right)\right\|_2^2\right]$$
$$= \mathbb{E}\left[\|(x_{kh} - h\nabla U(x_{kh})) - (T(x_{kh}) - h\nabla U(T(x_{kh})))\|_2^2\right]$$
$$\leq \mathbb{E}[\|x_{kh} - T(x_{kh})\|_2^2 - 2h\langle\nabla U(x_{kh}) - \nabla U(T(x_{kh})), x_{kh} - T(x_{kh})\rangle$$
$$+ h^2\|\nabla U(x_{kh}) - \nabla U(T(x_{kh}))\|_2^2]$$
$$\leq \mathbb{E}\left[\|x_{kh} - T(x_{kh})\|_2^2\right]$$
$$= W_2^2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h)$$

where the last equality follows by optimality of T, and the last inequality follows because L-smoothness of $U(x)$ implies

$$- 2h\langle\nabla U(x_{kh}) - \nabla U(T(x_{kh})), x_{kh} - T(x_{kh})\rangle$$
$$\leq - \frac{h}{L}\|\nabla U(x_{kh}) - \nabla U(T(x_{kh}))\|_2^2$$
$$\leq - h^2\|\nabla U(x_{kh}) - \nabla U(T(x_{kh}))\|_2^2$$

This completes the inductive step. ■

Corollary 14 *Let* $\mathbf{p}_t$ *be as defined in* (2). *Then for all* t,

$$W_2(\mathbf{p}_t, \mathbf{p}^*) \leq 4C_1$$

Proof of Corollary 14 First, if $t = \tau(t)$, then by Lemma 13 and (16) and triangle inequality, we get our conclusion.

So assume that $t \neq \tau(t)$. Using identical arguments as in Lemma 13, and noting the assumption on h' in (16) and the fact that $h \leq h'$, we can show that

$$W_2(\mathbf{p}_t, \boldsymbol{\pi}_{t-\tau(t)}) \leq W_2(\mathbf{p}_{\tau(t)}, \boldsymbol{\pi}_{t-\tau(t)}) \tag{22}$$

By triangle inequality and the assumption in (16), we have

$$
\begin{aligned}
&W_2(\mathbf{p}_t, \mathbf{p}^*) \\
&\leq W_2(\mathbf{p}_t, \boldsymbol{\pi}_{t-\tau(t)}) + W_2(\boldsymbol{\pi}_{t-\tau(t)}, \mathbf{p}^*) \\
&\leq W_2(\mathbf{p}_{\tau(t)}, \boldsymbol{\pi}_{t-\tau(t)}) + W_2(\boldsymbol{\pi}_{t-\tau(t)}, \mathbf{p}^*) \\
&\leq W_2(\mathbf{p}_{\tau(t)}, \boldsymbol{\pi}_h) + W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) \\
&\quad + W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) + W_2(\boldsymbol{\pi}_{t-\tau(t)}, \mathbf{p}^*) \\
&\leq 4C_1
\end{aligned}
$$

Where the first inequality is by triangle inequality, the second inequality is by (22), third inequality is by triangle inequality, fourth inequality is by assumption (16) and the fact that $t - \tau(t) \leq h \leq h'$.

■

Next, we use Lemma 13, to bound $\mathbb{E}\left[\|x_{kh}\|_2^2\right]$ for all k:

Lemma 15 *Let* h, x_t *and* $\mathbf{p}_t$ *be as defined in the statement of Theorem 2. Then for all* k

$$\mathbb{E}\left[\|x_{kh}\|_2^2\right] \leq 4(C_1^2 + C_2^2)$$

Proof of Lemma 15

Let $\gamma(x, y)$ be the optimal coupling between $\mathbf{p}_{kh}$ and $\boldsymbol{\pi}_h$. Let $\gamma'(x, y)$ be the optimal coupling between $\boldsymbol{\pi}_h$ and $\mathbf{p}^*$. Then

$$\begin{aligned}
\mathbb{E}_{\mathbf{p}_{kh}}\left[\|x\|_2^2\right] &= \mathbb{E}_\gamma\left[\|x\|_2^2\right] \\
&= \mathbb{E}_\gamma\left[\|x-y+y\|_2^2\right] \\
&\leq 2\mathbb{E}_\gamma\left[\|x-y\|_2^2\right] + 2\mathbb{E}_\gamma\left[\|y\|_2^2\right] \\
&= 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 2\mathbb{E}_{\boldsymbol{\pi}_h}\left[\|y\|_2^2\right] \\
&= 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 2\mathbb{E}_{\gamma'}\left[\|x\|_2^2\right] \\
&= 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 2\mathbb{E}_{\gamma'}\left[\|x-y+y\|_2^2\right] \\
&\leq 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 4\mathbb{E}_{\gamma'}\left[\|x-y\|_2^2\right] + 4\mathbb{E}_{\gamma'}\left[\|y\|_2^2\right] \\
&\leq 2W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) + 4W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) + 4\mathbb{E}_{\mathbf{p}^*}\left[\|x\|_2^2\right]
\end{aligned}$$

By definition of C_2 at the start of Section 7, we have

$$\mathbb{E}_{\mathbf{p}^*}\left[\|x\|_2^2\right] \leq C_2^2$$

.

By Lemma 13, we have

$$W_2(\mathbf{p}_{kh}, \boldsymbol{\pi}_h) \leq W_2(\mathbf{p}_0, \boldsymbol{\pi}_h) \leq C_1$$

By definition of h' at the start of Section 7, and h in Theorem 2 (which ensures $h \leq h'$), we have

$$W_2(\boldsymbol{\pi}_h, \mathbf{p}^*) \leq C_1$$

∎

Proof of Theorem 2 First, we bound the discretization error (for an arbitrary t). By Lemma 7:

$$\begin{aligned}
\frac{d}{ds}\left(F(\mathbf{p}_s) - F(\mathbf{q}_s^t)\right)\Big|_{s=t} &\leq \left(2L^2 t\sqrt{\mathbb{E}_{\mathbf{p}_{\tau(t)}}\|x_{\tau(t)}\|_2^2} + 2L\sqrt{td}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_* \\
&\leq \left(2L^2 t\sqrt{\mathbb{E}_{\mathbf{p}_{\tau(t)}}\|x_{\tau(t)}\|_2^2} + 2L\sqrt{td}\right) \cdot \|\mathcal{D}_{\mathbf{p}_t}\|_*
\end{aligned}$$

Given the choice of

$$h = \frac{1}{48}\min\left\{\frac{\epsilon}{C_1(C_1+C_2)L^2}, \frac{\epsilon^2}{L^2 C_1^2 d}, h'\right\}$$

, we can ensure that

$$\begin{aligned}
\left(L^2 h\sqrt{E\|x_{\tau(t)}\|_2^2} + 2L\sqrt{hd}\right) &\leq \frac{1}{4}\left(L^2 h\sqrt{18(C_1^2+C_2^2)} + 2L\sqrt{hd}\right) \\
&\leq \frac{\epsilon}{8C_1}
\end{aligned}$$

where the first inequality comes from Lemma 15.

Assume that $F(\mathbf{p}_s) - F(\mathbf{p}^*) \geq \epsilon$. By Lemma 12 and Corollary 14, we have

$$
\begin{aligned}
\|\mathcal{D}_{\mathbf{p}_s}\|_* &\geq \frac{F(\mathbf{p}_s) - F(\mathbf{p}^*)}{W_2(\mathbf{p}_s, \mathbf{p}^*)} \\
&\geq \frac{\epsilon}{W_2(\mathbf{p}_s, \mathbf{p}^*)} \\
&\geq \frac{\epsilon}{4C_1}
\end{aligned}
\tag{23}
$$

This implies that

$$
\left. \frac{d}{ds} F(\mathbf{p}_s) - F(\mathbf{q}_s^t) \right|_{s=t} \leq \frac{1}{2} \|\mathcal{D}_{\mathbf{p}_t}\|_*^2
$$

The rate of decrease of $F(\mathbf{p}_t)$ thus satisfies

$$
\begin{aligned}
\frac{d}{dt} F(\mathbf{p}_t) - F(\mathbf{p}^*) &= \left. \frac{d}{dt} F(\mathbf{q}_s^t) - F(\mathbf{p}^*) \right|_{s=t} + \left. \frac{d}{dt} (F(\mathbf{p}_s) - F(\mathbf{q}_s^t)) \right|_{s=t} \\
&= -\|\mathcal{D}_{\mathbf{p}_t}\|_*^2 + \frac{1}{2} \|\mathcal{D}_{\mathbf{p}_t}\|_*^2 \\
&\leq -\frac{1}{2} \|\mathcal{D}_{\mathbf{p}_t}\|_*^2 \\
&\leq -\frac{1}{2C_1^2} (F(\mathbf{p}_t) - F(\mathbf{p}^*))^2
\end{aligned}
$$

We now study two regimes. The first regime is when $F(\mathbf{p}_t) - F(\mathbf{p}^*) \geq 1$, $\frac{d}{dt} F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq -\frac{1}{2C_1^2} (F(\mathbf{p}_t) - F(\mathbf{p}^*))$, which implies

$$
F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq (F(\mathbf{p}_0) - F(\mathbf{p}^*)) \exp(-\frac{t}{2C_1^2})
$$

We thus achieve $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq 1$ in

$$
t \geq 2C_1^2 \log(F(\mathbf{p}_0) - F(\mathbf{p}^*))
$$

In the second regime, $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq 1$. By noting that $f_t = \frac{1}{t}$ is the solution to $\frac{d}{dt} f_t = -f_t^2$, and letting $f_t = \frac{1}{2C_1^2} (F(\mathbf{p}_t) - F(\mathbf{p}^*))$, we get $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq \frac{2C_1^2}{t}$. To achieve $F(\mathbf{p}_t) - F(\mathbf{p}^*) \leq \epsilon$, we set $t = \frac{2C_1^2}{\epsilon}$. Overall, we just need to set

$$
t \geq \frac{2C_1^2}{\epsilon} + 2C_1^2 \log(F(\mathbf{p}_0) - F(\mathbf{p}^*))
$$

This, combined with the choice of h earlier, proves the theorem.

■

8.2 Some regularity results

In this subsection, we provide some regularity results needed in various parts of the paper.

Lemma 16 *Let $w_{\boldsymbol{\mu}}$ be as defined in Lemma 1. Let $\mathbf{p}_t$ be as defined in 4. For all t, $w_{\mathbf{p}_t}$ is well defined, and $\mathbb{E}_{\mathbf{p}_t}\left[\|w_{\mathbf{p}_t}\|_2^2\right]$ is finite.*

Proof of Lemma 16 First, we establish the following statement: For any t, there exists a $\delta \in \mathbb{R}$ with $\boldsymbol{\mu}_{\delta,y}(x)$ being the distribution of $N(y,\delta)$ and $\mathbf{p} \in \mathscr{P}(\mathbb{R}^d)$ such that

1. For all $x \in \mathbb{R}^d$, $\mathbf{p}_t(x) = \mathbb{E}_{y \sim \mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right]$

2. $\mathbb{E}_{\mathbf{p}}\left[\|x\|_2^2\right]$ is finite.

If $t = \tau(t)$, then let $\mathbf{p} = (Id(\cdot) - h\nabla U(\cdot))_{\#}\mathbf{p}_{\tau(t)-1}$ and let $\delta = 2h$. Otherwise, if $t \neq \tau(t)$, then let $\mathbf{p} = (Id(\cdot) - (t - \tau(t))\nabla U(\cdot))_{\#}\mathbf{p}_{\tau(t)}$ and $\delta = 2(t - \tau(t))$. Where we used the definition of push-forward distribution from (4). 1. now can be easily verified.

To see 2, let $t' = \tau(t) - 1$ in case 1 and let $t' = \tau(t)$ in case 2.

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{p}}\left[\|x\|_2^2\right] \\
=&\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x - h\nabla U(x)\|_2^2\right] \\
\leq&2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x\|_2^2\right] + 2h^2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|\nabla U(x)\|_2^2\right] \\
\leq&2\mathbb{E}_{\mathbf{p}_{t'}}\left[\|x\|_2^2\right] + 2h^2 L^2 \mathbb{E}_{\mathbf{p}_{t'}}\left[\|x\|_2^2\right] \\
\leq&(2 + 2h^2 L^2)\frac{4d}{m}
\end{aligned}
$$

Where the last inequality follows by Lemma 11.

Since $\boldsymbol{\mu}_{\delta,y}(x)$ for all x, y, $\mathbb{E}_{\mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right]$ is differentiable for all x. This proves the first part of the Lemma.

Next, a nice property of Gaussians is that

$$
\nabla_x \boldsymbol{\mu}_{\delta,y}(x) = -\frac{\boldsymbol{\mu}_{\delta,y}}{\delta}(x - y)
$$

Thus,

$$
\begin{aligned}
&\nabla \log \mathbf{p}_t(x) \\
=&\frac{1}{\delta}\nabla \log \mathbb{E}_{y \sim \mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right] \\
=&\frac{1}{\delta}\frac{1}{\mathbb{E}_{y \sim \mathbf{p}}\left[\boldsymbol{\mu}_{\delta,y}(x)\right]}\mathbb{E}_{y \sim \mathbf{p}}\left[\left((y - x)\boldsymbol{\mu}_{\delta,y}(x)\right)\right] \\
=&\frac{1}{\delta}\mathbb{E}_{y \sim \boldsymbol{\mu}_{\delta}^x}\left[y\right] - x
\end{aligned}
$$

Where $\boldsymbol{\mu}_{\delta}^x$ denotes the conditional distribution of y given x, when $y \sim \mathbf{p}$ and $x \sim \boldsymbol{\mu}_{\delta,y}$.

Thus

$$\mathbb{E}_{x \sim \mathbf{p}_t(x)} \left[\|\nabla \log \mathbf{p}_t(x)\|_2^2 \right]$$
$$\leq \frac{1}{\delta} \mathbb{E}_{x \sim \mathbf{p}_t(x)} \left[2 \mathbb{E}_{y \sim \boldsymbol{\mu}_\delta^x} \left[\|y\|_2^2 \right] + 2 \|x\|_2^2 \right]$$
$$= \frac{2}{\delta} \mathbb{E}_{y \sim \mathbf{p}} \left[\|y\|_2^2 \right] + \frac{2}{\delta} \mathbb{E}_{x \sim \mathbf{p}_t} \left[\|x\|_2^2 \right]$$
$$\leq \infty$$

Where the first inequality is by Jensen's inequality and Young's inequality and the preceding result, the second inequality is by definition of conditional distribution, the third inequality is by the fact that $\delta > 0$ (by definition at the start of the proof), the fact that $\mathbb{E}_{x \sim \mathbf{p}_t} \left[\|x\|_2^2 \right] \leq \frac{4d}{m}$ (by Lemma 11), and by the fact that $\mathbb{E}_{y \sim \mathbf{p}} \left[\|y\|_2^2 \right] < \infty$ (see item 2. at the start of the proof)

Finally, we have that

$$\|w_{\mathbf{p}_t}\|_{L^2(\mathbf{p}_t)}^2$$
$$= \mathbb{E}_{\mathbf{p}_t} \left[\|w_{\mathbf{p}_t}(x)\|_2^2 \right]$$
$$= \mathbb{E}_{\mathbf{p}_t} \left[\|\nabla \log \mathbf{p}_t(x) - \nabla \log \mathbf{p}^*(x)\|_2^2 \right]$$
$$\leq 2 \mathbb{E}_{\mathbf{p}_t} \left[\|\nabla \log \mathbf{p}_t(x)\|_2^2 \right] + 2 \mathbb{E}_{\mathbf{p}_t} \left[\|\nabla \log \mathbf{p}^*(x)\|_2^2 \right]$$
$$< \infty$$

Where the last inequality uses the fact that $\|\nabla \log \mathbf{p}^(x)\|_2 = \|\nabla U(x)\|_2 \leq L\|x\|_2$ and $\mathbb{E}_{\mathbf{p}_t} \left[\|x\|_2^2 \right] \leq \frac{4d}{m}$. ∎

Lemma 17 *Let $\mathbf{p}_t$ be as defined in (4). Then $|\mathbf{p}_t'|$ is finite for all t, where $|\mathbf{p}_t'|$ is the metric derivative of $\mathbf{p}_t$, as defined in (6).*

Proof of Lemma 17 We define the random variable ξ to be distributed as $N(0,1)$. For all t, let $\mathbf{x}_t$ be as defined in (4). One can verify that the random variable $y_t \triangleq x_{\tau(t)} - \nabla(t - \tau(t)) U(x_{\tau(t)}) + \sqrt{2(t - \tau(t))}\xi$ has the same distribution as x_t. Thus y_t and $y_{t+\epsilon}$ define a coupling between $\mathbf{p}_t$ and $\mathbf{p}_{t+\epsilon}$. We thus have

25

Let $h \triangleq t - \tau(t)$

$$|\mathbf{p}_t'|$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} W_2(\mathbf{p}_t, \mathbf{p}_{t+\epsilon})$$

$$\leq \lim_{\epsilon \to 0} \frac{1}{\epsilon} \sqrt{\mathbb{E}\left[\|y_t - y_{t+\epsilon}\|_2^2\right]}$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \sqrt{\mathbb{E}_{x \sim \mathbf{P}_{\tau(t)}}\left[\|\epsilon \nabla U(x) + (\sqrt{2(h+\epsilon)} - \sqrt{2h})\xi\|_2^2\right]}$$

$$= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \sqrt{\mathbb{E}_{x \sim \mathbf{P}_{\tau(t)}}\left[\|\epsilon \nabla U(x)\|_2^2\right] + \mathbb{E}\left[\|(\sqrt{2(h+\epsilon)} - \sqrt{2h})\xi\|_2^2\right]}$$

$$\leq \lim_{\epsilon \to 0} \frac{1}{\epsilon} \sqrt{\mathbb{E}_{x \sim \mathbf{P}_{\tau(t)}}\left[\|\epsilon \nabla U(x)\|_2^2\right]}$$

$$\qquad + \frac{1}{\epsilon} \sqrt{\mathbb{E}\left[\|(\sqrt{2(h+\epsilon)} - \sqrt{2h})\xi\|_2^2\right]}$$

$$= \sqrt{\mathbb{E}_{x \sim \mathbf{P}_{\tau(t)}}\left[\|\nabla U(x)\|_2^2\right]} + \frac{1}{\sqrt{8h}} \sqrt{\mathbb{E}\left[\|\xi\|_2^2\right]}$$

Where the last inequality follows by Taylor expansion of $\sqrt{2h + 2\epsilon}$. We can bound the first term by a finite number using $\|\nabla U(x)\|_2^2 \leq L^2 \|x\|_2^2$, then applying Lemma 11. The second term is finite for $h \neq 0$.

For the case $h = 0$, we know that $w_{\mathbf{p}_t}$ satisfies the continuity equation for $\mathbf{p}_t$ at t, and so $|\mathbf{p}_t'| = \|w_{\mathbf{p}_t}\|_{L^2(\mathbf{p}_t)} < \infty$, by Lemma 2 and Lemma 16.

∎