Langevin Monte Carlo without smoothness

Niladri S. Chatterji* UC Berkeley
Jelena Diakonikolas* UW-Madison
Michael I. Jordan UC Berkeley
Peter L. Bartlett UC Berkeley

Abstract

Langevin Monte Carlo (LMC) is an iterative algorithm used to generate samples from a distribution that is known only up to a normalizing constant. The nonasymptotic dependence of its mixing time on the dimension and target accuracy is understood mainly in the setting of smooth (gradient-Lipschitz) log-densities, a serious limitation for applications in machine learning. In this paper, we remove this limitation, providing polynomial-time convergence guarantees for a variant of LMC in the setting of nonsmooth log-concave distributions. At a high level, our results follow by leveraging the implicit smoothing of the log-density that comes from a small Gaussian perturbation that we add to the iterates of the algorithm and controlling the bias and variance that are induced by this perturbation.

1 Introduction

The problem of generating a sample from a distribution that is known up to a normalizing constant is a core problem across the computational and inferential sciences (Robert and Casella, 2013; Kaipio and Somersalo, 2006; Cesa-Bianchi and Lugosi, 2006; Rademacher and Vempala, 2008; Vempala, 2005; Chen et al., 2018). A prototypical example involves generating a sample from a log-concave distribution—a probability distribution of the following form:

$$p^*(\mathbf{x}) \propto e^{-U(\mathbf{x})},$$

where the function $U(\mathbf{x})$ is convex and is referred to as the *potential function.* While generating a sample from the exact distribution $p^*(\mathbf{x})$ is often computationally intractable, for most applications it suffices to generate a sample from a distribution $\tilde{p}(\mathbf{x})$ that is close to $p^*(\mathbf{x})$ in some distance (such as, e.g., total variation distance, Wasserstein distance, or Kullback-Leibler divergence).

Proceedings of the 23[rd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

The most commonly used methods for generating a sample from a log-concave distribution are (i) random walks (Dyer et al., 1991; Lovász and Vempala, 2007), (ii) different instantiations of Langevin Monte Carlo (LMC) (Parisi, 1981), and (iii) Hamiltonian Monte Carlo (HMC) (Neal et al., 2011). These methods trade off rate of convergence against per-iteration complexity and applicability: random walks are typically the slowest in terms of the total number of iterations, but each step is fast as it does not require gradients of the log-density and they are broadly applicable, while HMC is the fastest in the number of iterations, but each step is slow as it uses gradients of the log-density and it mainly applies to distributions with smooth log-densities.

LMC occupies a middle ground between random walk and HMC. In its standard form, LMC updates its iterates as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla U(\mathbf{x}_k) + \sqrt{2\eta}\boldsymbol{\xi}_k, \qquad \text{(LMC)}$$

where $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$ are independent Gaussian random vectors. The per-iteration complexity is reduced relative to HMC because it only requires stochastic gradients of the log-density (Welling and Teh, 2011). This also increases its range of applicability relative to HMC. While it is not a reversible Markov chain and classical theory of MCMC does not apply, it is nonetheless amenable to theoretical analysis given that it is obtained via discretization of an underlying stochastic differential equation (SDE). There is, however, a fundamental difficulty in connecting theory to the promised wide range of applications in statistical inference. In particular, the use of techniques from SDEs generally requires $U(\mathbf{x})$ to have Lipschitz-continuous gradients. This assumption excludes many natural applications (Kaipio and Somersalo, 2006; Durmus et al., 2018; Marie-Caroline et al., 2019; Li et al., 2018).

A prototypical example of sampling problems with nonsmooth potentials are different instantiations of sparse Bayesian inference. In this setting, one wants to sample from the posterior distribution of the form:

$$p^*(\mathbf{x}) \propto \exp\big(-f(\mathbf{x}) - \|\Phi\mathbf{x}\|_p^p\big),$$

where $f(\mathbf{x})$ is the log-likelihood function, Φ is a sparsifying dictionary (e.g., a wavelet dictionary), and $p \in [1, 2]$. In the simplest case of Bayesian LASSO (Park and Casella, 2008), $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, $\Phi = \mathbf{I}$, and $p = 1$, where $\mathbf{A}$ is the measurement matrix, $\mathbf{b}$ are the labels, and $\mathbf{I}$ denotes the identity matrix. In general, when Φ is the identity or an orthogonal wavelet transform, proximal maps (i.e., solutions to convex minimization problems of the form $\min_{\mathbf{x}\in\mathbb{R}^d}\{\|\Phi\mathbf{x}\|_p^p + \frac{1}{2\lambda}\|\mathbf{x}-\mathbf{z}\|_2^2\}$, where λ and $\mathbf{z}$ are parameters of the proximal map) are easily computable and proximal LMC methods apply (Cai et al., 2018; Price et al., 2018; Durmus et al., 2019, 2018; Atchadé, 2015). However, in the so-called analysis-based approaches with overcomplete dictionaries, Φ is non-orthogonal and the existence of efficient proximal maps becomes unclear (Elad et al., 2007; Cherkaoui et al., 2018).

In this work, we tackle this problem head-on and pose the following question:

Is it possible to obtain nonasymptotic convergence results for LMC with a nonsmooth potential?

Here, we focus on standard LMC (allowing only minor modifications) and the general case in which proximal maps are not efficiently computable. We answer this question positively through a series of results that involve transformations of the basic stochastic dynamics in (LMC). In contrast to previous work that considered nonsmooth potentials (e.g., Atchadé, 2015; Durmus et al., 2018; Hsieh et al., 2018; Durmus et al., 2019), the transformations we consider are simple (such as perturbing a gradient query point by a Gaussian), they do not require strong assumptions such as the existence of proximal maps, they can apply directly to nonsmooth Lipschitz potentials without any additional structure (such as composite structure in Atchadé (2015); Durmus et al. (2018) or strong convexity in Hsieh et al. (2018)), and the guarantee we provide is on the distribution of the last iterate of LMC as opposed to an average of distributions over a sequence of iterates of LMC in Durmus et al. (2019).

Our main theorem is based on a Gaussian smoothing result summarized in the following theorem.

Main Theorem (Informal)**.** *Let $\bar{p}^*(\mathbf{x}) \propto \exp(-\bar{U}(\mathbf{x}))$ be a probability distribution, where $\bar{U}(\mathbf{x}) = U(\mathbf{x}) + \psi(\mathbf{x})$, $U(\cdot)$ is a convex subdifferentiable function whose subgradients $\nabla U(\cdot)$ satisfy*

$$\|\nabla U(\mathbf{x}) - \nabla U(\mathbf{y})\|_2 \le L\|\mathbf{x}-\mathbf{y}\|_2^\alpha, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,$$

for some $L < \infty$, $\alpha \in [0, 1]$, and $\psi(\cdot)$ is λ-strongly convex and m-smooth. There exists an algorithm—Perturbed Langevin Monte Carlo (P-LMC)—whose iterations have the same computational complexity as (LMC) and that requires no more than $\widetilde{\mathcal{O}}(d^{\frac{5-3\alpha}{2}}/\varepsilon^{\frac{4}{1+\alpha}})$ iterations to generate a sample that is ε-close to $\bar{p}^*$ in 2-Wasserstein distance.

Further, if the goal is to sample from $p^*(\mathbf{x}) \propto \exp(-U(\mathbf{x}))$, a variant of (P-LMC) takes poly(d/ε) iterations to generate a sample from a distribution that is ε-close to p^* in total variation distance.

This informal version of the theorem displays only the dependence on the dimension d and accuracy ε. A detailed statement is provided in Theorems 3.4 and 3.6, and Corollary 4.1.

Our assumption on the subgradients of U from the statement of the Main Theorem is known as Hölder-continuity, or (L, α)-weak smoothness of the function. It interpolates between Lipschitz gradients (smooth functions, when $\alpha = 1$) and bounded gradients (nonsmooth Lipschitz functions, when $\alpha = 0$). In Bayesian inference, the general (L, α)-weakly smooth potentials arise in the Bayesian analog of "bridge regression," which interpolates between LASSO and ridge regression (see, e.g., Park and Casella, 2008) . To the best of our knowledge, our work is the first to consider the convergence of LMC in this general weakly-smooth model of the potentials – previous work only considered its extreme cases obtained for $\alpha = 0$ and $\alpha = 1$.

To understand the behavior of LMC on weakly smooth (including nonsmooth) potentials, we leverage results from the optimization literature. First, by using the fact that a weakly smooth function can be approximated by a smooth function—a result that has been exploited in the optimization literature to obtain methods with optimal convergence rates (Nesterov, 2015; Devolder et al., 2014)—we show that even the basic version of LMC can generate a sample in polynomial time, as long as U is "not too nonsmooth" (namely, as long as $1/\alpha$ can be treated as a constant).

The main impediment to the convergence analysis of LMC when treating a weakly smooth function U as an inexact version of a nearby smooth function is that a constant bias is induced on the gradients, as discussed in Section 3.1. To circumvent this issue, in Section 3.2 we argue that an LMC algorithm can be analyzed as a different LMC run on a Gaussian-smoothed version of the potential using *unbiased stochastic estimates of the gradient.*[1] Building on this reduction, we define a Perturbed Langevin Monte Carlo (P-LMC) algorithm

[1]A similar idea was used in Kleinberg et al. (2018) to view expected iterates of stochastic gradient descent as gradient descent on a smoothed version of the objective. Stochastic smoothing has also been used to lower the parallel complexity of nonsmooth minimization (Duchi et al., 2012).

that reduces the additional variance that arises in the gradients from the reduction.

To obtain our main theorem, we couple a result about convergence of LMC with stochastic gradient estimates in *Wasserstein* distance (Durmus et al., 2019) with carefully combined applications of inequalities relating Kullback-Leibler divergence, Wasserstein distance, and total variation distance. Also useful are structural properties of the weakly smooth potentials and their Gaussian smoothing. As a byproduct of our techniques, we obtain a nonasymptotic result for convergence in *total variation* distance for (standard) LMC with stochastic gradients, which, to the best of our knowledge, was not known prior to our work.

1.1 Related work

Starting with the work of Dalalyan (Dalalyan, 2017), a variety of theoretical results have established mixing time results for LMC (Durmus and Moulines, 2016; Raginsky et al., 2017; Zhang et al., 2017; Cheng and Bartlett, 2018; Cheng et al., 2018b; Dalalyan and Karagulyan, 2019; Xu et al., 2018; Lee et al., 2018) and closely related methods, such as Metropolis-Adjusted LMC (Dwivedi et al., 2018) and HMC (Mangoubi and Smith, 2017; Bou-Rabee et al., 2018; Mangoubi and Vishnoi, 2018; Cheng et al., 2018a). These results apply to sampling from well-behaved distributions whose potential function U is *smooth* (Lipschitz gradients) and (usually) strongly convex. For standard (LMC) with smooth and strongly convex potentials, the tightest upper bounds for the mixing time are $\widetilde{\mathcal{O}}(d/\varepsilon^2)$. They were obtained in Dalalyan (2017); Durmus and Moulines (2016) for convergence in total variation (with a *warm start*; without a warm start the total variation result scales as $\widetilde{\mathcal{O}}(\frac{d^3}{\varepsilon^2})$) and in 2-Wasserstein distance.

When it comes to using (LMC) with nonsmooth potential functions, there are far fewer results. In particular, there are two main approaches: relying on the use of proximal maps (Atchadé, 2015; Durmus et al., 2018, 2019) and relying on averaging of the distributions over iterates of LMC (Durmus et al., 2019, SSGLD). Methods relying on the use of proximal maps require a composite structure of the potential (namely, that the potential is a sum of a smooth and a nonsmooth function) and that the proximal maps can be computed efficiently. Note that this is a very strong assumption. In fact, when the composite structure exists in convex optimization *and* proximal maps are efficiently computable, it is possible to solve nonsmooth optimization problems with the same iteration complexity as if the objective were smooth (see, e.g., Beck and Teboulle, 2009). Thus, while the methods from Durmus et al. (2018, 2019) have a lower iteration complexity than our approach, the use of proximal maps increases their per-iteration complexity (each iteration needs to solve a convex optimization problem). It is also unclear how the performance of the methods degrades when the proximal maps are computed only approximately. Finally, unlike our work, Atchadé (2015); Durmus et al. (2018) and Durmus et al. (2019, SGLD) do not handle potentials that are purely nonsmooth, without a composite structure.

The only method that we are aware of and that is directly applicable to nonsmooth potentials is (Durmus et al., 2019, SSGLD). On a technical level, Durmus et al. (2019) interprets LMC as a gradient flow in the space of measures and leverages techniques from convex optimization to analyze its convergence. The convergence guarantees are obtained for a weighted average of distributions of individual iterates of LMC, which, roughly speaking, maps the standard convergence analysis of the average iterate of projected gradient descent or stochastic gradient descent to the setting of sampling methods. While the iteration complexity for the average distribution (Durmus et al., 2019) is much lower than ours, their bounds for individual iterates of LMC are uninformative. By contrast, our results are for the *last iterate* of perturbed LMC (P-LMC). Note that in the related setting of convex optimization, last-iterate convergence is generally more challenging to analyze and has been the subject of recent research (Shamir and Zhang, 2013; Jain et al., 2019).

It is also worth mentioning that there exist approaches such as the Mirrored Langevin Algorithm (Hsieh et al., 2018) that can be used to efficiently sample from structured nonsmooth distributions such as the Dirichlet posterior. However, this algorithm's applicability to general nonsmooth densities is unclear.

1.2 Outline

Section 2 provides the notation and background. Section 3 provides our main theorems, stated for deterministic and stochastic approximations of the potential (negative log-density) and composite structure of the potential. Section 4 extends the result of Section 3 to non-composite potentials. We conclude in Section 5.

2 Preliminaries

The goal is to generate samples from a distribution $p^* \propto \exp(-U(\mathbf{x}))$, where $\mathbf{x} \in \mathbb{R}^d$. We equip $\mathbb{R}^d$ with the standard Euclidean norm $\|\cdot\| = \|\cdot\|_2$ and use $\langle\cdot,\cdot\rangle$ to denote inner products. We assume the following for the potential (negative log-density) U:

(A1) U is convex and subdifferentiable. Namely, for all

$\mathbf{x} \in \mathbb{R}^d$, there exists a subgradient of U, $\nabla U(\mathbf{x}) \in \partial U(\mathbf{x})$, such that $\forall \mathbf{y} \in \mathbb{R}^d$:

$$U(\mathbf{y}) \geq U(\mathbf{x}) + \langle \nabla U(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle .$$

(A2) There exist $L < \infty$ and $\alpha \in [0, 1]$ such that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$\|\nabla U(\mathbf{x}) - \nabla U(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2^{\alpha}, \tag{2.1}$$

where $\nabla U(\mathbf{x})$ denotes an arbitrary subgradient of U at $\mathbf{x}$.

(A3) The distribution p^* has a finite *fourth moment*:

$$\int_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{x} - \mathbf{x}^*\|_2^4 \cdot p^*(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathcal{M}_4 < \infty,$$

where $\mathbf{x}^* \in \mathrm{argmin}_{\mathbf{x} \in \mathbb{R}^d} U(\mathbf{x})$ is an arbitrary minimizer of U.

Assumption (A2) is known as the (L, α)-weak smoothness or Hölder continuity of the (sub)gradients of U. When $\alpha = 1$, it corresponds to the standard *smoothness* (Lipschitz continuity of the gradients), while at the other extreme, when $\alpha = 0$, U is (possibly) *non-smooth* and *Lipschitz-continuous.*

Properties of weakly smooth functions. A property that follows directly from (2.1) is that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\begin{aligned} U(\mathbf{y}) \leq\, & U(\mathbf{x}) + \langle \nabla U(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ & + \frac{L}{1+\alpha}\|\mathbf{y} - \mathbf{x}\|^{1+\alpha}. \end{aligned} \tag{2.2}$$

One of the most useful properties of weakly smooth functions that has been exploited in optimization is that they can be approximated by smooth functions to an arbitrary accuracy, at the cost of increasing their smoothness parameter Nesterov (2015); Devolder et al. (2014). This was shown in (Nesterov, 2015, Lemma 1) and is summarized in the following lemma for the special case of the unconstrained Euclidean setting.

Lemma 2.1. *Let $U : \mathbb{R}^d \to \mathbb{R}$ be a convex function that satisfies (2.1) for some $L < \infty$ and $\alpha \in [0, 1]$. Then, for any $\delta > 0$ and $M = \left(\frac{1}{\delta}\right)^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}}$, we have that, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:*

$$\begin{aligned} U(\mathbf{y}) \leq\, & U(\mathbf{x}) + \langle \nabla U(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ & + \frac{M}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \frac{\delta}{2}. \end{aligned} \tag{2.3}$$

Furthermore, it is not hard to show that Eq. (2.3) implies (see Devolder et al., 2014, Section 2.2):

$$\|\nabla U(\mathbf{x}) - \nabla U(\mathbf{y})\|_2 \leq M\|\mathbf{x} - \mathbf{y}\|_2 + 2\sqrt{\delta M} \tag{2.4}$$

where $M = \left(\frac{1}{\delta}\right)^{\frac{1-\alpha}{1+\alpha}} \cdot L^{2/(1+\alpha)}$, as in Lemma 2.1.

Gaussian smoothing. Given $\mu \geq 0$, define the Gaussian smoothing U_μ of U as:

$$U_\mu(\mathbf{y}) := \mathbb{E}_{\boldsymbol{\xi}}[U(\mathbf{y} + \mu\boldsymbol{\xi})],$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$. The reason for considering the Gaussian smoothing U_μ instead of U is that it generally enjoys better smoothness properties. In particular, U_μ is smooth even if U is not. Here we review some basic properties of U_μ, most of which can be found in (Nesterov and Spokoiny, 2017, Section 2) for non-smooth Lipschitz functions. We generalize some of these results to weakly smooth functions. While the results can be obtained for arbitrary normed spaces, here we state all the results for the space $(\mathbb{R}^d, \|\cdot\|_2)$, which is the only setting considered in this paper.

The following lemma is a simple extension of the results from (Nesterov and Spokoiny, 2017, Section 2) and it establishes certain regularity conditions for Gaussian smoothing that will be used in our analysis.

Lemma 2.2. *Let $U : \mathbb{R}^d \to \mathbb{R}$ be a convex function that satisfies Eq. (2.1) for some $L < \infty$ and $\alpha \in [0, 1]$. Then:*

(i) For all $\mathbf{x} \in \mathbb{R}^d$:

$$|U_\mu(\mathbf{x}) - U(\mathbf{x})| = U_\mu(\mathbf{x}) - U(\mathbf{x}) \leq \frac{L\mu^{1+\alpha} d^{\frac{1+\alpha}{2}}}{1+\alpha}.$$

(ii) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2 \leq \frac{L d^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}}\|\mathbf{y} - \mathbf{x}\|_2.$$

Additionally, we show that Gaussian smoothing preserves strong convexity, stated in the following (simple) lemma. Recall that a differentiable function ψ is λ-strongly convex if, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \nabla \psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Lemma 2.3. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be λ-strongly convex. Then ψ_μ is also λ-strongly convex.*

Composite potentials and regularization. To prove convergence of the continuous-time process (which requires strong convexity), we work with potentials that have the following composite form:

$$\bar{U}(\mathbf{x}) := U(\mathbf{x}) + \psi(\mathbf{x}), \tag{2.5}$$

where $\psi(\cdot)$ is m-smooth and λ-strongly convex. For obtaining guarantees in terms of convergence to $\bar{p}^* \propto e^{-\bar{U}}$, we do not need Assumption (A3), which bounds the fourth moment of the target distribution—this is only needed in establishing the results for $p^* \propto e^{-U}$.

If the goal is to sample from a distribution $p^*(\mathbf{x}) \propto e^{-U(\mathbf{x})}$ (instead of $\bar{p}^*(\mathbf{x}) \propto e^{-\bar{U}(\mathbf{x})}$), then we need to ensure that the distributions p^* and $\bar{p}^*$ are sufficiently close to each other. This can be achieved by choosing $\psi(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2$, where λ and $\|\mathbf{x}' - \mathbf{x}^*\|_2$ are sufficiently small, for an arbitrary $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} U(\mathbf{x})$ (see Corollary 4.1 for precise details).

Note that by the triangle inequality, we have that:

$$\begin{aligned}\|\nabla \bar{U}(\mathbf{x}) - \nabla \bar{U}(\mathbf{y})\|_2 &\le \|\nabla U(\mathbf{x}) - \nabla U(\mathbf{y})\|_2 + \|\nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{y})\|_2 \\ &\le L\|\mathbf{x}-\mathbf{y}\|_2^\alpha + m\|\mathbf{x}-\mathbf{y}\|_2. \end{aligned} \tag{2.6}$$

Thus, by (2.4), we have the following (deterministic) Lipschitz approximation of the gradients of $\bar{U}$: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, any $\delta > 0$, and $M = M(\delta)$ (as in Lemma 2.1):

$$\begin{aligned}\|\nabla \bar{U}(\mathbf{x}) - \nabla \bar{U}(\mathbf{y})\|_2 \\ \le M\|\mathbf{x}-\mathbf{y}\|_2 + m\|\mathbf{x}-\mathbf{y}\|_2 + 2\sqrt{\delta M}.\end{aligned} \tag{2.7}$$

On the other hand, for Gaussian-smoothed composite potentials, using Lemma 2.2, we have:

$$\begin{aligned}\|\nabla \bar{U}_\mu(\mathbf{x}) - \nabla \bar{U}_\mu(\mathbf{y})\|_2 \\ \le \left(\frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}} + m\right)\|\mathbf{x}-\mathbf{y}\|_2.\end{aligned} \tag{2.8}$$

Distances between probability measures. Given any two probability measures P and Q on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field of $\mathbb{R}^d$, the total variation distance between them is defined as

$$\|P - Q\|_{\mathrm{TV}} := \sup_{A\in\mathcal{B}(\mathbb{R}^d)} |P(A) - Q(A)|.$$

The *Kullback-Leibler* divergence between P and Q is defined as:

$$\mathrm{KL}(P|Q) := \mathbb{E}_P\left[\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right],$$

where $\mathrm{d}P/\mathrm{d}Q$ is the Radon-Nikodym derivative of P with respect to Q.

Define a *transference plan* ζ, a distribution on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that $\zeta(A \times \mathbb{R}^d) = P(A)$ and $\zeta(\mathbb{R}^d \times A) = Q(A)$ for any $A \in \mathcal{B}(\mathbb{R}^d)$. Let $\Gamma(P,Q)$ denote the set of all such transference plans. Then the 2-Wasserstein distance is defined as:

$$W_2(P,Q) := \left(\inf_{\zeta\in\Gamma(P,Q)} \int_{\mathbf{x},\mathbf{y}\in\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|_2^2 \mathrm{d}\zeta(\mathbf{x},\mathbf{y})\right)^{1/2}.$$

3 Sampling for composite potentials

In this section, we consider the setting of composite potentials of the form $\bar{U}(\mathbf{x}) = U(\mathbf{x}) + \psi(\mathbf{x})$, where $U(\cdot)$ is (L,α)-weakly smooth (possibly with $\alpha = 0$, in which case U is nonsmooth and Lipschitz) and $\psi(\cdot)$ is m-smooth and λ-strongly convex. We provide results for mixing times[2] of different variants of overdamped LMC in both 2-Wasserstein and total variation distance.

We first consider the deterministic smooth approximation of U, which follows from Lemma 2.1. This approach does not require making any changes to the standard overdamped LMC. However, it leads to a polynomial dependence of the mixing time on d and $1/\varepsilon$ only when α is bounded away from zero (namely, when $1/\alpha$ can be treated as a constant).

We then consider another approach that relies on a Gaussian smoothing of $\bar{U}$ and that leads to a polynomial dependence of the mixing time on d and $1/\varepsilon$ for all values of α. In particular, the approach leads to the mixing time for 2-Wasserstein distance that matches the best known mixing time of overdamped LMC when U is smooth ($\alpha = 1$) – $\widetilde{\mathcal{O}}(d/\varepsilon^2)$, and preserves polynomial-time dependence on d and $1/\varepsilon$ even if U is nonsmooth ($\alpha = 0$), in which case the mixing time scales as $\widetilde{\mathcal{O}}(d^{\frac{5}{2}}/\varepsilon^4)$. The analysis requires us to consider a minor modification to standard LMC in which we perturb by a Gaussian random variable the points at which $\nabla\bar{U}$ is queried. Note that it is unclear whether it is possible to obtain such bounds for (LMC) without this modification (see Appendix D).

3.1 First attempt: Deterministic approximation by a smooth function

In the optimization literature, deterministic smooth approximations of weakly smooth functions (as in Lemma 2.1) are generally useful for obtaining methods with optimal convergence rates (Nesterov, 2015; Devolder et al., 2014). A natural question is whether the same type of approximation is useful for bounding the mixing times of the Langevin Monte Carlo method invoked for potentials that are weakly smooth.

We note that it is not obvious that such a deterministic approximation would be useful, as the deterministic error introduced by the smooth approximation causes an adversarial bias $2\sqrt{\delta M(\delta)}$ in the Lipschitz approximation of the gradients (see Eq. (2.4)). While this bias can be made arbitrarily small for values of α that are bounded away from zero, when $\alpha = 0$, $M(\delta) = L^2/\delta$, and the induced bias is constant for any value of δ.

We show that it is possible to bound the mixing times of LMC when the potential is "not too nonsmooth". In particular, we show that the upper bound on the

[2] *Mixing time* is defined as the number of iterations needed to reach an ε accuracy in either 2-Wasserstein or total variation distance.

mixing time of LMC when applied to an (L, α)-weakly smooth potential scales with $\mathrm{poly}((\frac{1}{\varepsilon})^{1/\alpha})$ in both the 2-Wasserstein and total variation distance, which is polynomial in $1/\varepsilon$ for α bounded away from zero. Although we do not prove any lower bounds on the mixing time in this case, the obtained result aligns well with our observation that the deterministic bias cannot be controlled for the deterministic smooth approximation of a nonsmooth Lipschitz function, as explained above. Technical details are deferred to Appendix C.

3.2 Gaussian smoothing

The main idea is summarized as follows. Recall that LMC with respect to the potential $\bar{U}$ can be stated as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla\bar{U}(\mathbf{x}_k) + \sqrt{2\eta}\boldsymbol{\xi}_k, \tag{LMC}$$

where $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$ are independent Gaussian random vectors. This method corresponds to the Euler-Mayurama discretization of the Langevin diffusion.

Consider a modification of (LMC) in which we add another Gaussian term:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla\bar{U}(\mathbf{x}_k) + \sqrt{2\eta}\boldsymbol{\xi}_k + \mu\boldsymbol{\omega}_k, \tag{3.1}$$

where $\boldsymbol{\omega}_k \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$ and is independent of $\boldsymbol{\xi}_k$. Observe that (3.1) is simply another (LMC) with a slightly higher level of noise—$\sqrt{2\eta}\boldsymbol{\xi}_k + \mu\boldsymbol{\omega}_k$ instead of $\sqrt{2\eta}\boldsymbol{\xi}_k$. Let $\mathbf{y}_k := \mathbf{x}_k - \mu\boldsymbol{\omega}_{k-1}$. Then:

$$\begin{aligned}\mathbf{y}_{k+1} = \mathbf{y}_k - \eta\Big[\nabla\bar{U}(\mathbf{y}_k + \mu\boldsymbol{\omega}_{k-1}) - \frac{\mu}{\eta}\boldsymbol{\omega}_{k-1}\Big]\\ + \sqrt{2\eta}\boldsymbol{\xi}_k.\end{aligned} \tag{S-LMC}$$

Taking expectations on both sides with respect to $\boldsymbol{\omega}_{k-1}$:

$$\mathbb{E}_{\boldsymbol{\omega}_{k-1}}[\mathbf{y}_{k+1}] = \mathbf{y}_k - \eta\nabla\bar{U}_\mu(\mathbf{y}_k) + \sqrt{2\eta}\boldsymbol{\xi}_k,$$

where $\bar{U}_\mu$ is the Gaussian smoothing of $\bar{U}$, as defined in Section 2. Thus, we can view the sequence $\{\mathbf{y}_k\}$ in Eq. (S-LMC) as obtained by simply transforming the standard LMC chain to another LMC chain using stochastic estimates $\nabla\bar{U}(\mathbf{y}_k + \mu\boldsymbol{\omega}_{k-1}) - \frac{\mu}{\eta}\boldsymbol{\omega}_{k-1}$ of the gradients. However, the variance of this gradient estimate is too high to handle nonsmooth functions, and, as before, our bound on the mixing time of this chain blows up as $\alpha \downarrow 0$ (see Appendix D).

Thus, instead of working with the algorithm defined in (S-LMC), we correct for the extra induced variance and consider the sequence of iterates defined by:

$$\begin{aligned}\mathbf{y}_{k+1} = \mathbf{y}_k - \eta\nabla\bar{U}(\mathbf{y}_k + \mu\boldsymbol{\omega}_{k-1})\\ + \sqrt{2\eta}\boldsymbol{\xi}_k.\end{aligned} \tag{P-LMC}$$

This sequence will have a sufficiently small bound on the variance to obtain the desired results.

Lemma 3.1. *For any $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, let $G(\mathbf{x}, \mathbf{z}) := \nabla\bar{U}(\mathbf{x} + \mu\mathbf{z})$ denote a stochastic gradient of $\bar{U}_\mu$. Then $G(\mathbf{x}, \mathbf{z})$ is an unbiased estimator of $\nabla\bar{U}_\mu$ whose (normalized) variance satisfies:*

$$\begin{aligned}\sigma^2 :=& \frac{\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla\bar{U}_\mu(\mathbf{x}) - G(\mathbf{x}, \mathbf{z})\right\|_2^2\right]}{d}\\ &\le 4d^{\alpha-1}\mu^{2\alpha}L^2 + 4\mu^2 m^2.\end{aligned}$$

Remark 3.2. The variance from Lemma 3.1 can be lowered by using multiple independent samples to estimate $\nabla\bar{U}_\mu$ (instead of a single sample as in (P-LMC)). However, unlike in the case of nonsmooth optimization (Duchi et al., 2012), such a strategy will *not* reduce the mixing times reported here. This is because the variance from Lemma 3.1 is already low enough to not be a limiting factor in the mixing time bounds.

Let the distribution of the k^{th} iterate $\mathbf{y}_k$ be denoted by $\bar{p}_k$, and let $\bar{p}^*_\mu \propto \exp(-\bar{U}_\mu)$ be the distribution with $\bar{U}_\mu$ as the potential. Our overall strategy for proving our main result is as follows. First, we show that the Gaussian smoothing does not change the target distribution significantly with respect to the Wasserstein distance, by bounding $W_2(\bar{p}^*, \bar{p}^*_\mu)$ (Lemma 3.3). Using Lemma 3.1, we then invoke a result on mixing times of Langevin diffusion with stochastic gradients, which allows us to bound $W_2(\bar{p}_k, \bar{p}^*_\mu)$. Finally, using the triangle inequality and choosing a suitable step size η, smoothing radius μ, and number of steps K so that $W_2(\bar{p}^*, \bar{p}^*_\mu) + W_2(\bar{p}_K, \bar{p}^*_\mu) \le \varepsilon$, we establish our final bound on the mixing time of (P-LMC) in Theorem 3.4.

Lemma 3.3. *Let $\bar{p}^*$ and $\bar{p}^*_\mu$ be the distributions corresponding to the potentials $\bar{U}$ and $\bar{U}_\mu$ respectively. Then:*

$$\begin{aligned}W_2(\bar{p}^*, \bar{p}^*_\mu) \le& \frac{8}{\lambda}\Big(\frac{3}{2} + \frac{d}{2}\log\Big(\frac{2(M+m)}{\lambda}\Big)\Big)^{1/2}\\ &\cdot\Big(\beta_\mu + \sqrt{\beta_\mu/2}\Big),\end{aligned}$$

where $\beta_\mu := \beta_\mu(d, L, m, \alpha) = \frac{L\mu^{1+\alpha}d^{\frac{1+\alpha}{2}}}{\sqrt{2}(1+\alpha)} + \frac{m\mu^2 d}{2}$.

Our main result is stated in the following theorem.

Theorem 3.4. *Let the initial iterate $\mathbf{y}_0$ be drawn from a probability distribution $\bar{p}_0$. If the step size η satisfies $\eta < 2/(M + m + \lambda)$, then:*

$$\begin{aligned}W_2(\bar{p}_K, \bar{p}^*) \le& (1 - \lambda\eta)^{K/2}\, W_2(\bar{p}_0, \bar{p}^*_\mu) + W_2(\bar{p}^*, \bar{p}^*_\mu)\\ &+ \Big(\frac{2(M+m)}{\lambda}\eta d\Big)^{1/2} + \sigma\sqrt{\frac{(1+\eta)\eta d}{\lambda}},\end{aligned}$$

where

$$\sigma^2 \le 4d^{\alpha-1}\mu^{2\alpha}L^2 + 4\mu^2 m^2,\ M = \frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}},$$

$$\eta \le \frac{\varepsilon^2 \mu^{1-\alpha}\lambda}{1000(L+m)d^{\frac{3-\alpha}{2}}},$$
$$\mu = \frac{\varepsilon^{\frac{2}{1+\alpha}} \min\{\lambda^{\frac{2}{1+\alpha}}, 1\}/300}{\sqrt{d}\big(\sqrt{m} + L^{\frac{1}{1+\alpha}}\big)\left[10 + d\log\left(\varepsilon^{-2}(m+L)d/\lambda\right)\right]^{\frac{1}{2}}},$$

and $W_2(\bar{p}^*, \bar{p}^*_\mu)$ *is bounded as in Lemma 3.3.*

Further, if, for $\varepsilon \in (0, d^{1/4})$*, we choose*

$$K \ge \frac{1}{\lambda\eta}\log\left(\frac{3W_2(\bar{p}_0, \bar{p}^*_\mu)}{\varepsilon}\right),$$

then $W_2(\bar{p}_K, \bar{p}^*) \le \varepsilon$*.*

Remark 3.5. Treating L, m, λ as constants and using the fact that $W_2(\bar{p}_0, \bar{p}^*_\mu) = \mathcal{O}(\mathrm{poly}(d/\varepsilon))$ (see, Cheng et al., 2018b, Lemma 13, by choosing the initial distribution $\bar{p}_0$ appropriately), we find that Theorem 3.4 yields a bound of $K = \widetilde{\mathcal{O}}\left(d^{\frac{5-3\alpha}{2}}/\varepsilon^{\frac{4}{1+\alpha}}\right)$. When $\alpha = 1$ (the Lipschitz gradient case), we recover the known mixing time of $K = \widetilde{\mathcal{O}}(d/\varepsilon^2)$, while at the other extreme when $\alpha = 0$ (the nonsmooth Lipschitz potential case), we find that $K = \widetilde{\mathcal{O}}(d^{\frac{5}{2}}/\varepsilon^4)$.

The choice of the smoothing radius μ is made such that it is large enough to ensure that the smoothed distribution $\bar{p}_\mu$ is sufficiently smooth, but not too large so as to ensure that the bias, $W_2(\bar{p}^*, \bar{p}_\mu)$, is controlled.

Proof of Theorem 3.4. By the triangle inequality,

$$W_2(\bar{p}_K, \bar{p}^*) \le W_2(\bar{p}_K, \bar{p}^*_\mu) + W_2(\bar{p}^*, \bar{p}^*_\mu). \tag{3.2}$$

To bound the first term, $W_2(\bar{p}_K, \bar{p}^*_\mu)$, we invoke (Durmus et al., 2019, Theorem 21) (see Theorem A.4 in Appendix A). Recall that $\bar{U}_\mu$ is continuously differentiable, $(M+m)$-smooth (with $M = \frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}}$), and λ-strongly convex. Additionally, the sequence of points $\{\mathbf{y}_k\}_{k=1}^K$ can be viewed as a sequence of iterates of overdamped LMC with respect to the potential specified by $\bar{U}_\mu$, where the iterates are updated using unbiased stochastic estimates of $\bar{U}_\mu$. Thus we have:

$$W_2(\bar{p}_K, \bar{p}^*_\mu) \le (1-\lambda\eta)^{K/2} W_2(\bar{p}_0, \bar{p}^*_\mu) + \sqrt{\frac{2(M+m)}{\lambda}\eta d} + \sigma\sqrt{\frac{(1+\eta)\eta d}{\lambda}}, \tag{3.3}$$

and by Lemma 3.1, $\sigma^2 \le 4d^{\alpha-1}\mu^{2\alpha}L^2 + 4\mu^2 m^2$.

The last piece we need is control over the distance between $\bar{p}^*$ and $\bar{p}^*_\mu$. This is established above in Lemma 3.3 . Thus, combining Eqs. (3.2) and (3.3) with Lemma 3.3, the first part of the theorem follows.

It is straightforward to verify that our choice of μ ensures that $W_2(\bar{p}^*, \bar{p}^*_\mu) \le \varepsilon/3$. The choice of η ensures that $(2(M+m)\eta d/\lambda)^{1/2} \le \varepsilon/6$ and the choice of K ensures that the initial error contracts exponentially to $\varepsilon/3$ (see the proof of Theorem 3.6 in Appendix E for a similar calculation). This yields the second claim. □

Further, we show that this result can be generalized to total variation distance.

Theorem 3.6. *Let the initial iterate* $\mathbf{y}_0$ *be drawn from a probability distribution* $\bar{p}_0$*. If we choose the step size such that* $\eta < 2/(M+m+\lambda)$*, then:*

$$\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \le \frac{L\mu^{1+\alpha}d^{(1+\alpha)/2}}{1+\alpha} + \frac{\lambda\mu^2 d}{2} + \sqrt{\mathrm{KL}(\bar{p}_K, \bar{p}^*_\mu)},$$

where $\mathrm{KL}(\bar{p}_K, \bar{p}^*_\mu)$ *is bounded by* $W_2(\bar{p}_K, \bar{p}^*_\mu)$ *in Eq. (3.4), and* $W_2(\bar{p}_K, \bar{p}^*_\mu)$ *is bounded as in Eq. (3.3).*

Further, if, for $\varepsilon \in (0, 1]$*, we choose*

$$\mu = \min\left\{\frac{\varepsilon^{\frac{1}{1+\alpha}}}{4\max\{1, L^{\frac{1}{1+\alpha}}\}d^{1/2}}, \sqrt{\frac{\varepsilon\lambda}{2m^2 d}}\right\},$$
$$\bar{\varepsilon} = \frac{\varepsilon^2}{4\max\{(M+m)(\sqrt{2d/\lambda + 2\|\mathbf{x}^*\|_2^2} + 2\|\mathbf{x}^*\|_2^2), 1\}},$$

then choosing the step size η *and number of steps* K *as*

$$\eta \le \frac{\bar{\varepsilon}^2\lambda}{64d(M+m)} \quad \text{and} \quad K \ge \frac{\log(2W_2(\bar{p}_0, \bar{p}^*_\mu)/\bar{\varepsilon})}{\lambda\eta},$$

we have $\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \le \varepsilon$*.*

Remark 3.7. Treating $L, \mu, \lambda, \|\mathbf{x}^*\|$ as constants and using the fact that $W_2(\bar{p}_0, \bar{p}^*_\mu) = \mathcal{O}(\mathrm{poly}(\frac{d}{\varepsilon}))$ (by Cheng et al., 2018b, Lemma 13, along with an appropriate choice for the initial distribution), Theorem 3.6 gives a bound on the mixing time $K = \widetilde{\mathcal{O}}(d^{5-3\alpha}/\varepsilon^{\frac{7+\alpha}{1+\alpha}})$. When $\alpha = 1$ (Lipschitz gradients), we have $K = \widetilde{\mathcal{O}}(d^2/\varepsilon^4)$, while when $\alpha = 0$ (nonsmooth Lipschitz potential) we have $K = \widetilde{\mathcal{O}}(d^5/\varepsilon^7)$. While the bound for the smooth case (Lipschitz gradients, $\alpha = 1$) is looser than the best known bound for LMC with a warm start (Dalalyan, 2017), we conjecture that it is improvable. The main loss is incurred when relating W_2 to KL distance, using an inequality from Polyanskiy and Wu (2016) (see Appendix A). If tighter inequalities were obtained, either relating W_2 and KL, or W_2 and TV, this result would immediately improve as a consequence. The results for LMC with non-Lipschitz gradients ($\alpha \in [0, 1)$) are novel. Finally, as a byproduct of our approach, we obtain the first bound for stochastic gradient LMC in TV distance (see Remark E.1 in Appendix E).

4 Sampling for regularized potentials

Consider now the case in which we are interested in sampling from a distribution $p^* \propto \exp(-U)$. As mentioned in Section 2, we can use the same analysis as in

$$\mathrm{KL}(\bar{p}_K|\bar{p}_\mu^*) \leq \left(\frac{\bar{M}\sqrt{\frac{2d}{\lambda} + 2\|\mathbf{x}^*\|_2^2}}{2} + \frac{\bar{M}\sqrt{\frac{4d}{\lambda} + 4\|\mathbf{x}^*\|_2^2 + 2W_2^2(\bar{p}_K, \bar{p}_{\mu^*})}}{2} + \bar{M}\|\mathbf{x}^*\|_2\right)W_2(\bar{p}_K, \bar{p}_\mu^*). \tag{3.4}$$

the previous section, by running (P-LMC) with a regularized potential $\bar{U} = U + \lambda\|\mathbf{x} - \mathbf{x}'\|_2^2/2$, where $\mathbf{x}' \in \mathbb{R}^d$. To obtain the desired result, the only missing piece is bounding the distance between $\bar{p}^* \propto \exp(-\bar{U})$ and p^*, leading to the following corollary of Theorem 3.6.

Corollary 4.1. *Let the initial iterate $\mathbf{y}_0$ satisfy $\mathbf{y}_0 \sim \bar{p}_0$, for some distribution $\bar{p}_0$ and let $\bar{p}_K$ denote the distribution of $\mathbf{y}_K$. If we choose the step-size η such that $\eta < 2/(M + 2\lambda)$, then:*

$$\|\bar{p}_K - p^*\|_{\mathrm{TV}} \leq \|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} + \frac{\lambda\sqrt{\mathcal{M}_4}}{2} + \frac{\lambda\|\mathbf{x}' - \mathbf{x}^*\|_2^2}{2},$$

where $\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}}$ is bounded as in Theorem 3.6 and $\mathcal{M}_4$ is the fourth moment of p^*.

Further, if, for $\varepsilon' \in (0, 1]$, we choose $\lambda = \frac{4\varepsilon'}{\sqrt{\mathcal{M}_4} + \|\mathbf{x}' - \mathbf{x}^*\|_2^2}$ and all other parameters as in Theorem 3.6 for $\varepsilon = \varepsilon'/2$, then, we have $\|\bar{p}_K - p^*\|_{\mathrm{TV}} \leq \varepsilon'$.

Proof. By the triangle inequality,

$$\|\bar{p}_K - p^*\|_{\mathrm{TV}} \leq \|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} + \|p^* - \bar{p}^*\|_{\mathrm{TV}}.$$

Applying Lemma A.1 from the appendix,

$$\begin{aligned}\|p^* - \bar{p}^*\|_{\mathrm{TV}} &\leq \frac{1}{2}\left(\int_{\mathbb{R}^d}(U(\mathbf{x}) - \bar{U}(\mathbf{x}))^2 p^*(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{1/2} \\ &= \frac{1}{2}\left(\int_{\mathbb{R}^d}\left(\frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2\right)^2 p^*(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{1/2} \\ &\leq \frac{1}{2}\left(2\int_{\mathbb{R}^d}\left(\frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2\right)^2 p^*(\mathbf{x})\mathrm{d}\mathbf{x}\right. \\ &\qquad \left. + 2\int_{\mathbb{R}^d}\left(\frac{\lambda}{2}\|\mathbf{x}^* - \mathbf{x}'\|_2^2\right)^2 p^*(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{1/2}.\end{aligned}$$

Thus, using Assumption (A3), we get

$$\|p^* - \bar{p}^*\|_{\mathrm{TV}} \leq \frac{\lambda}{2}\sqrt{\mathcal{M}_4} + \frac{\lambda\|\mathbf{x}' - \mathbf{x}^*\|_2^2}{2}.$$

The rest of the proof follows by Theorem 3.6. ☐

Remark 4.2. Treating $L, \|\mathbf{x}^*\|_2, \|\mathbf{x}' - \mathbf{x}^*\|_2$ as constants, the upper bound on the mixing time is $K = \widetilde{\mathcal{O}}(\frac{d^{5-3\alpha}\mathcal{M}_4{}^{3/2}}{\varepsilon^{\frac{10+4\alpha}{1+\alpha}}})$. Thus, when $\alpha = 1$, we have $K = \widetilde{\mathcal{O}}(\frac{d^2\mathcal{M}_4{}^{3/2}}{\varepsilon^7})$, while when $\alpha = 0$, $K = \widetilde{\mathcal{O}}(\frac{d^5\mathcal{M}_4{}^{3/2}}{\varepsilon^{10}})$.

5 Discussion

We obtained polynomial-time theoretical guarantees for a variant of LMC—(P-LMC)—that uses Gaussian smoothing and applies to target distributions with nonsmooth log-densities. The smoothing we apply is tantamount to perturbing the gradient query points in LMC by a Gaussian random variable, which is a minor modification to the standard method.

Beyond its applicability to sampling from more general weakly smooth and nonsmooth target distributions, our work also has some interesting implications. For example, we believe our results can be extended to sampling from structured distributions with nonsmooth and nonconvex negative log-densities, following an argument from, e.g., Cheng et al. (2018a). It should also be possible to work with stochastic gradients instead of exact gradients by coupling our arguments with the bounds in Dalalyan and Karagulyan (2019) or Durmus et al. (2019). Further, it seems plausible that coupling our results with the results for derivative-free LMC (Shen et al., 2019, which only applies to distributions with smooth and strongly convex log-densities) would lead to a more broadly applicable derivative-free LMC.

Several other interesting directions for future research remain. For example, as discussed in Remark 3.7 and Remark E.1 (Appendix E), we conjecture that the asymptotic dependence on d and ε in our bounds on the mixing times for total variation distance (Theorem 3.6) can be improved to match those obtained for the 2-Wasserstein distance (Theorem 3.4). Further, in standard settings of LMC with the exact gradients, Metropolis filter is often used to improve the convergence properties of LMC and it leads to lower mixing times (see, e.g., Dwivedi et al., 2018). However, the performance of Metropolis-adjusted LMC becomes unclear once the gradients are stochastic (as is the case for (P-LMC)). It is an interesting question whether a Metropolis adjustment can speed up (P-LMC).

Acknowledgements

We thank François Lanusse for his useful pointers to the literature on applied Bayesian statistics with nonsmooth posteriors. This research was supported by the NSF grants CCF-1740855 and IIS-1619362, and the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764. Part of this work was done while the authors were visiting Simons Institute for the Theory of Computing.

References

Yves F Atchadé. A Moreau-Yosida approximation scheme for a class of high-dimensional posterior distributions. *arXiv preprint arXiv:1505.07072*, 2015.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

François Bolley and Cédric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse Math.*, 14(3):331–352, 2005.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018.

Xiaohao Cai, Marcelo Pereyra, and Jason D McEwen. Uncertainty quantification for radio interferometric imaging–i. proximal MCMC methods. *Monthly Notices of the Royal Astronomical Society*, 480(3): 4154–4169, 2018.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast MCMC sampling algorithms on polytopes. *J. Mach. Learn. Res.*, 19(1):2146–2231, 2018.

Xiang Cheng and Peter L Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proc. ALT'18*, 2018.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proc. COLT'18*, 2018b.

H Cherkaoui, Loubna El Gueddari, C Lazarus, Antoine Grigis, Fabrice Poupon, Alexandre Vignaud, Sammuel Farrens, J-L Starck, and Philippe Ciuciu. Analysis vs synthesis-based regularization for combined compressed sensing and parallel mri reconstruction at 7 Tesla. In *Proc. IEEE EUSIPCO'18*, 2018.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Series B. Stat. Methodol.*, 79 (3):651–676, 2017.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradients. *Stoch. Process. Their Appl.*, 2019.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2):37–75, 2014.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2): 674–701, 2012.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. *SIAM J. Imaging Sci.*, 11(1):473–506, 2018.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proc. COLT'18*, 2018.

Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM*, 38(1):1–17, 1991.

Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. *Inverse probl.*, 23(3):947, 2007.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. In *Proc. NeurIPS'18*, 2018.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Proc. COLT'19*, 2019.

Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proc. ICML'18*, 2018.

Yin Tat Lee, Zhao Song, and Santosh S Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities. *arXiv preprint arXiv:1812.06243*, 2018.

Yuan Li, Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Graph-based regularization for regression problems with highly-correlated designs. In *Proc. IEEE GlobalSIP'18*, 2018.

László Lovász and Santosh Vempala. The geometry of log-concave functions and sampling algorithms. *Random Struct. Algor.*, 30(3):307–358, 2007.

Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Oren Mangoubi and Nisheeth Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Proc. NeurIPS'18*, 2018.

Corbineau Marie-Caroline, Kouamé Denis, Chouzenoux Emilie, Tourneret Jean-Yves, and Pesquet Jean-Christophe. Preconditioned P-ULA for joint deconvolution-segmentation of ultrasound images. *arXiv preprint arXiv:1903.08111*, 2019.

Radford M Neal et al. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, volume 2, pages 113–162. CRC Press, 2011.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Math. Program.*, 152(1-2): 381–404, 2015.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. of Comput. Math.*, 17(2):527–566, 2017.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

G Parisi. Correlation functions and computer simulations. *Nucl. Phys. B*, 180(3):378–384, 1981.

Trevor Park and George Casella. The Bayesian LASSO. *J. Am. Stat. Assoc.*, 103(482):681–686, 2008.

Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Trans. Inf. Theory*, 62(7):3992–4002, 2016.

Matthew A Price, Xiaohao Cai, Jason D McEwen, Marcelo Pereyra, and Thomas D Kitching. Sparse Bayesian mass-mapping with uncertainties: Local credible intervals. *arXiv preprint arXiv:1812.04017*, 2018.

Luis Rademacher and Santosh Vempala. Dispersion of mass and the complexity of randomized geometric algorithms. *Adv. Math.*, 219(3):1037–1069, 2008.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A non-asymptotic analysis. In *Proc. COLT'17*, 2017.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proc. ICML'13*, 2013.

Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi. Non-asymptotic results for Langevin Monte Carlo: Coordinate-wise and black-box sampling. *arXiv preprint arXiv:1902.01373*, 2019.

Santosh Vempala. Geometric random walks: A survey. *Combinatorial and computational geometry*, 52(2): 573–612, 2005.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. ICML'11*, 2011.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Proc. NeurIPS'18*, 2018.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Proc. COLT'17*, 2017.

Appendix

A Additional background

Here we state the results from related work that are invoked in our analysis.

First, the smooth approximations of the potentials used in this paper are pointwise larger than the original potentials, and have a bounded distance from the original potentials. This allows us to invoke the following lemma from Dalalyan (2017).

Lemma A.1. *(Dalalyan, 2017, Lemma 3) Let U and $\tilde{U}$ be two functions such that $U(\mathbf{x}) \leq \tilde{U}(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$ and both e^{-U} and $e^{-\tilde{U}}$ are integrable. Then the Kullback-Leibler divergence between the distributions defined by densities $p \propto e^{-U}$ and $\tilde{p} \propto e^{-\tilde{U}}$ can be bounded as:*

$$\mathrm{KL}(p|\tilde{p}) \leq \frac{1}{2}\int_{\mathbb{R}^d} (U(\mathbf{x}) - \tilde{U}(\mathbf{x}))^2 p(\mathbf{x})\mathrm{d}\mathbf{x}.$$

As a consequence, $\|p - \tilde{p}\|_{\mathrm{TV}} \leq \frac{1}{2}\|U - \tilde{U}\|_{L^2(p)}$.

The next result that we will be invoking allows us to bound the Wasserstein distance between the target distributions corresponding to the composite potential $\bar{U}$ and its Gaussian smoothing $\bar{U}_\mu$.

Lemma A.2. *(Bolley and Villani, 2005, Corollary 2.3) Let X be a measurable space equipped with a measurable distance ρ, let $p \geq 1$, and let ν be a probability measure on X. Assume that there exist $\mathbf{x}_0 \in X$ and $\gamma > 0$ such that $\int_X e^{\gamma\rho(\mathbf{x}_0,\mathbf{x})^p}\mathrm{d}\nu(\mathbf{x})$ is finite. Then, for any other probability measure μ on X :*

$$W_p(\mu,\nu) \leq C\Big[\mathrm{KL}(\mu|\nu)^{1/p} + \Big(\frac{\mathrm{KL}(\mu|\nu)}{2}\Big)^{1/(2p)}\Big],$$

where

$$C := 2\inf_{\mathbf{x}_0\in X,\gamma>0}\left(\frac{1}{\gamma}\left(\frac{3}{2} + \log\int_X e^{\gamma\rho(\mathbf{x}_0,\mathbf{x})^p}\mathrm{d}\nu(\mathbf{x})\right)\right).$$

Another useful result, due to Polyanskiy and Wu (2016), lets us bound the KL-divergence between two distributions in terms of their 2-Wasserstein distance. This is used to relate the TV distance between distributions to their respective Wasserstein distance in Section 3.2.

Proposition A.3. *(Polyanskiy and Wu, 2016, Proposition 1) Let $Q(\mathbf{x})$ be a density on $\mathbb{R}^d$ such that for all $\mathbf{x} \in \mathbb{R}^d$: $\|\nabla \log Q(\mathbf{x})\|_2 \leq c_1\|\mathbf{x}\|_2 + c_2$ for some $c_1, c_2 \geq 0$. Then,*

$$\mathrm{KL}(P|Q) \leq \left(\frac{c_1}{2}\left[\sqrt{\mathbb{E}_{\mathbf{x}\sim P}\left[\|\mathbf{x}\|_2^2\right]} + \sqrt{\mathbb{E}_{\mathbf{y}\sim Q}\left[\|\mathbf{y}\|_2^2\right]}\right] + c_2\right) W_2(P,Q).$$

In particular, if $Q(\mathbf{x}) \propto e^{-U(\mathbf{x})}$ for some M-smooth function U, then we immediately have:

$$\|\nabla \log Q(\mathbf{x}) - \nabla \log Q(\mathbf{x}^*)\|_2 = \|\nabla \log Q(\mathbf{x})\|_2 \leq M\|\mathbf{x} - \mathbf{x}^*\|_2 \leq M\|\mathbf{x}\|_2 + M\|\mathbf{x}^*\|_2,$$

where $\mathbf{x}^* \in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} U(\mathbf{x})$, and the assumption of the proposition is satisfied with

$$c_1 = M \quad \text{and} \quad c_2 = M\|\mathbf{x}^*\|_2.$$

We will be invoking a result from Dalalyan and Karagulyan (2019) that bounds the Wasserstein distance between the target distribution p^* and the distribution of the K^{th} iterate of LMC with stochastic gradients. The assumptions about the stochastic gradients $G(\mathbf{x}, \mathbf{z})$ is that they are unbiased and their variance is bounded. Namely:

$$\mathbb{E}_{\mathbf{z}_k}[G(\mathbf{x}_k, \mathbf{z}_k) = \nabla U(\mathbf{x}_k),$$

and

$$\mathbb{E}_{\mathbf{z}_k}[\|G(\mathbf{x}_k, \mathbf{z}_k) - \mathbb{E}_{\mathbf{z}'_k}[G(\mathbf{x}_k, \mathbf{z}'_k)]\|_2^2] \leq \sigma^2 d,$$

where the diffusion term $\boldsymbol{\xi}_{k+1}$ is independent of $(\mathbf{z}_1, ..., \mathbf{z}_k)$. The random vectors $(\mathbf{z}_1, ..., \mathbf{z}_k)$ corresponding to the error of the gradient estimate are not assumed to be independent in Dalalyan and Karagulyan (2019); however, in our case it suffices to assume that they are, in fact, independent.

Theorem A.4. *(Durmus et al., 2019, Theorem 21) Let p_K be the distribution of the K^{th} iterate of Langevin Monte Carlo with stochastic gradients, and let $p^* \propto e^{-U}$. If U is M-smooth and λ-strongly convex and the step size η satisfies $\eta \leq \frac{2}{M+\lambda}$, then:*

$$W_2(p_K, p^*) \leq (1-\lambda\eta)^{K/2} W_2(p_0, p^*) + \left(\frac{2M\eta d}{\lambda}\right)^{1/2} + \sigma\sqrt{\frac{(1+\eta)\eta d}{\lambda}}.$$

Next we state the results from Durmus and Moulines (2016) that we use multiple times in our proofs to establish contraction of the solution of the Langevin continuous-time stochastic differential equation:

$$\mathrm{d}\mathbf{y}_t = -\nabla U(\mathbf{y}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \tag{A.1}$$

where $\mathbf{y}_0 \sim q_0$. Let the distribution of $\mathbf{y}_t$ be denoted by q_t.

Theorem A.5. *(Durmus and Moulines, 2016, Proposition 1) Let the function U be L-smooth and λ-strongly convex, let $q_0 \sim \delta_{\mathbf{x}}$ (the Dirac-delta distribution at $\mathbf{x}$), and let $\mathbf{x}^*$ be the minimizer of U. Then:*

1. *For all $t \geq 0$ and $\mathbf{x} \in \mathbb{R}^d$,*
$$\int_{\mathbb{R}^d} \|\mathbf{z}-\mathbf{x}^*\|_2^2 \, q_t(\mathbf{z})\mathrm{d}\mathbf{z} \leq \|\mathbf{x}-\mathbf{x}^*\|_2^2 \, e^{-2\lambda t} + \frac{d}{\lambda}(1-e^{-2\lambda t}).$$
2. *The stationary distribution $p^* \propto \exp(-U)$ satisfies $\int_{\mathbf{y}\in\mathbb{R}^d} \|\mathbf{y}-\mathbf{x}^*\|_2^2 \, p^*(\mathbf{y})\mathrm{d}\mathbf{y} \leq d/\lambda$.*
3. *For any $\mathbf{x} \in \mathbb{R}^d$ and $t > 0$, $W_2(q_t, p^*) \leq e^{-\lambda t}\left\{\|\mathbf{x}-\mathbf{x}^*\|_2 + (d/\lambda)^{1/2}\right\}$.*

Finally, we provide a slight modification of (Dalalyan, 2017, Lemma 5) that we use in the proof of Theorem C.5.

Lemma A.6. *Let the function $\bar{U} = U + \psi$, where U is (L,α)-weakly smooth and ψ is m-smooth and λ-strongly convex. If the initial iterate is chosen as $\mathbf{y}_0 \sim q_0 = \mathcal{N}(\mathbf{x}^*, (M+m)^{-1} I_{d\times d})$ and $\bar{p}^* \propto \exp(-\bar{U})$, then:*

$$\|q_t - \bar{p}^*\|_{\mathrm{TV}} \leq \exp\left\{\frac{d}{4}\log\left(\frac{M+m}{\lambda}\right) + \frac{\delta}{4} - \frac{t\lambda}{2}\right\},$$

where q_t is the distribution of $\mathbf{y}_t$ that evolves according to (A.1).

Proof. Using the definition of p^*,

$$\begin{aligned}
\bar{p}^*(\mathbf{y})^{-1} &= e^{\bar{U}(\mathbf{y})} \int_{\mathbb{R}^d} e^{-\bar{U}(\mathbf{z})}\mathrm{d}\mathbf{z} = e^{\bar{U}(\mathbf{y})-\bar{U}(\mathbf{x}^*)} \int_{\mathbb{R}^d} e^{-\bar{U}(\mathbf{z})+\bar{U}(\mathbf{x}^*)}\mathrm{d}\mathbf{z} \\
&\leq \exp\left\{\langle \nabla\bar{U}(\mathbf{x}^*), \mathbf{y}-\mathbf{x}^*\rangle + \frac{M+m}{2}\|\mathbf{y}-\mathbf{x}^*\|_2^2 + \frac{\delta}{2}\right\} \\
&\quad \cdot \int_{\mathbb{R}^d} \exp\left\{-\langle \nabla\bar{U}(\mathbf{x}^*), \mathbf{z}-\mathbf{x}^*\rangle - \frac{\lambda}{2}\|\mathbf{z}-\mathbf{x}^*\|_2^2\right\}\mathrm{d}\mathbf{z} \\
&\leq \left(\frac{2\pi}{\lambda}\right)^{d/2} \exp\left\{\frac{M+m}{2}\|\mathbf{y}-\mathbf{x}^*\|_2^2 + \frac{\delta}{2}\right\}.
\end{aligned}$$

Thus, we have that the χ^2-divergence between q_0 and p^* is bounded by

$$\begin{aligned}
\chi^2\left(q_0|\bar{p}^*\right) &= \mathbb{E}_{\mathbf{y}\sim\bar{p}^*}\left[\left(\frac{q_0(\mathbf{y})}{\bar{p}^*(\mathbf{y})}\right)^2\right] \\
&= \left(\frac{2\pi}{M+m}\right)^{-d} \int_{\mathbb{R}^d} \exp\left\{-(M+m)\|\mathbf{y}-\mathbf{x}^*\|_2^2\right\} p^*(\mathbf{y})^{-1}\mathrm{d}\mathbf{y} \\
&\leq \exp(\delta/2)\left(\frac{2\pi}{M+m}\right)^{-d}\left(\frac{2\pi}{\lambda}\right)^{d/2} \int_{\mathbb{R}^d} \exp\left\{-\frac{M+m}{2}\|\mathbf{y}-\mathbf{x}^*\|_2^2\right\}\mathrm{d}\mathbf{y} \\
&\leq \exp(\delta/2)\left(\frac{2\pi}{M+m}\right)^{-d}\left(\frac{2\pi}{\lambda}\right)^{d/2}\left(\frac{2\pi}{M+m}\right)^{d/2} \\
&\leq \exp(\delta/2)\left(\frac{M+m}{\lambda}\right)^{d/2}.
\end{aligned}$$

By (Dalalyan, 2017, Lemma 1) (which only relies on the strong convexity of $\bar{U}$), we know that:

$$\|q_t - \bar{p}^*\|_{\mathrm{TV}} \le \frac{\exp(-t\lambda/2)}{2}\chi^2(q_0|\bar{p}^*)^{1/2}, \qquad \forall t \ge 0.$$

Combining this with the upper bound on the initial χ^2 divergence completes the proof. □

B Proofs for Gaussian smoothing

Lemma 2.2. *Let $U : \mathbb{R}^d \to \mathbb{R}$ be a convex function that satisfies Eq. (2.1) for some $L < \infty$ and $\alpha \in [0, 1]$. Then:*

(i) For all $\mathbf{x} \in \mathbb{R}^d$:

$|U_\mu(\mathbf{x}) - U(\mathbf{x})| = U_\mu(\mathbf{x}) - U(\mathbf{x}) \le \frac{L\mu^{1+\alpha} d^{\frac{1+\alpha}{2}}}{1+\alpha}$.

(ii) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2 \le \frac{L d^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}}\|\mathbf{y} - \mathbf{x}\|_2$.

Proof.
Proof of Part (i). First, it is not hard to show that whenever U is convex and $\mu > 0$, $U_\mu(\mathbf{x}) \ge U(\mathbf{x})$, $\forall \mathbf{x}$. By the definition of U_μ and using that $\boldsymbol{\xi}$ is centered, we have:

$$U_\mu(\mathbf{x}) - U(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left[U(\mathbf{x} + \mu\boldsymbol{\xi}) - U(\mathbf{x}) - \mu\left\langle \nabla U(\mathbf{x}), \boldsymbol{\xi}\right\rangle\right] e^{-\|\boldsymbol{\xi}\|_2^2/2} \mathrm{d}\boldsymbol{\xi}.$$

Applying Eq. (2.2):

$$|U_\mu(\mathbf{x}) - U(\mathbf{x})| \le \frac{L}{1+\alpha}\mu^{1+\alpha}\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} \|\boldsymbol{\xi}\|_2^{1+\alpha} e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi}.$$

Finally, using (Nesterov and Spokoiny, 2017, Lemma 1), $\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}\|^{1+\alpha} e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi} \le d^{(1+\alpha)/2}$.

Proof of Part (ii). First, observe that, by Jensen's inequality and Eq. (2.1):

$$\begin{aligned}\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2 &\le \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\|\nabla U(\mathbf{y}+\mu\boldsymbol{\xi}) - \nabla U(\mathbf{x}+\mu\boldsymbol{\xi})\|_2 e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi} \\ &\le L\|\mathbf{y}-\mathbf{x}\|_2^\alpha.\end{aligned} \tag{B.1}$$

Further, by (Nesterov and Spokoiny, 2017, Eq. (21)), the gradient of U_μ can be expressed as:

$$\nabla U_\mu(\mathbf{x}) = \frac{1}{\mu(2\pi)^{d/2}}\int_{\mathbb{R}^d} U(\mathbf{x}+\mu\boldsymbol{\xi})\boldsymbol{\xi} e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi}.$$

Thus, applying Jensen's inequality, we also have:

$$\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2 \le \frac{1}{\mu(2\pi)^{d/2}}\int_{\mathbb{R}^d}|U(\mathbf{x}+\mu\boldsymbol{\xi}) - U(\mathbf{y}+\mu\boldsymbol{\xi})| \cdot \|\boldsymbol{\xi}\|_2 e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi}. \tag{B.2}$$

Using Eq. (2.2), we have that:

$$\begin{aligned}|U(\mathbf{x}+\mu\boldsymbol{\xi}) - U(\mathbf{y}+\mu\boldsymbol{\xi})| \le& \min\Big\{ \left\langle\nabla U(\mathbf{y}+\mu\boldsymbol{\xi}), \mathbf{x}-\mathbf{y}\right\rangle + \frac{L}{1+\alpha}\|\mathbf{y}-\mathbf{x}\|_2^{1+\alpha}, \\ &\qquad \left\langle\nabla U(\mathbf{x}+\mu\boldsymbol{\xi}), \mathbf{y}-\mathbf{x}\right\rangle + \frac{L}{1+\alpha}\|\mathbf{y}-\mathbf{x}\|_2^{1+\alpha}\Big\} \\ \le& \frac{1}{2}\left\langle\nabla U(\mathbf{y}+\mu\boldsymbol{\xi}) - \nabla U(\mathbf{x}+\mu\boldsymbol{\xi}), \mathbf{x}-\mathbf{y}\right\rangle + \frac{L}{1+\alpha}\|\mathbf{y}-\mathbf{x}\|_2^{1+\alpha} \\ \le& \frac{L}{1+\alpha}\|\mathbf{y}-\mathbf{x}\|_2^{1+\alpha},\end{aligned}$$

where the second inequality comes from the minimum being smaller than the mean, and the last inequality is by convexity of U (which implies $\langle \nabla U(\mathbf{x}) - \nabla U(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq 0,\ \forall \mathbf{x},\ \mathbf{y}$). Thus, combining with Eq. (B.2), we have:

$$\begin{aligned}\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2 &\leq \frac{L}{\mu(1+\alpha)}\|\mathbf{y}-\mathbf{x}\|_2^{1+\alpha}\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\|\boldsymbol{\xi}\|_2 e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi} \\ &= \frac{L}{\mu(1+\alpha)}\|\mathbf{y}-\mathbf{x}\|_2^{1+\alpha}d^{1/2}.\end{aligned} \tag{B.3}$$

Finally, combining Eqs. (B.1) and (B.3):

$$\begin{aligned}\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2 &= \|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2^{\alpha}\cdot\|\nabla U_\mu(\mathbf{y}) - \nabla U_\mu(\mathbf{x})\|_2^{1-\alpha} \\ &\leq L^\alpha\Big(\frac{Ld^{1/2}}{\mu(1+\alpha)}\Big)^{1-\alpha}\|\mathbf{y}-\mathbf{x}\|_2 \\ &= \frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}}\|\mathbf{y}-\mathbf{x}\|_2,\end{aligned}$$

as claimed. □

Lemma 2.3. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be λ-strongly convex. Then ψ_μ is also λ-strongly convex.*

Proof. By the definition of a Gaussian smoothing, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\begin{aligned}\psi_\mu(\mathbf{y}) - &\psi_\mu(\mathbf{x}) - \langle\nabla\psi_\mu(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \\ &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\Big(\psi(\mathbf{y}+\mu\boldsymbol{\xi}) - \psi(\mathbf{x}+\mu\boldsymbol{\xi}) - \langle\nabla\psi(\mathbf{x}+\mu\boldsymbol{\xi}), \mathbf{y}-\mathbf{x}\rangle\Big)e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi} \\ &\geq \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\frac{\lambda}{2}\|\mathbf{y}-\mathbf{x}\|^2 e^{-\|\boldsymbol{\xi}\|_2^2/2}\mathrm{d}\boldsymbol{\xi} \\ &= \frac{\lambda}{2}\|\mathbf{y}-\mathbf{x}\|^2,\end{aligned}$$

where we have used λ-strong convexity of ψ. □

C Mixing times for deterministic approximations of negative log-density

In this section, we analyze the convergence of Langevin diffusion in the *2-Wasserstein* distance and *total variation* distance for target distributions of the form $\bar{p}^* \propto e^{-\bar{U}(\mathbf{x})}$, where $\bar{U}(\cdot) = U(\cdot) + \psi(\cdot)$, $U(\cdot)$ is (L, α)-weakly-smooth, and $\psi(\cdot)$ is m-smooth and λ-strongly convex. The techniques we use here are an extension of similar techniques used previously by Dalalyan (2017); Durmus and Moulines (2016).

To analyze the convergence, in both cases we will use a coupling argument that bounds the discretization error after Euler-Mayurama discretization is applied to the Langevin diffusion. Consider the first process which describes the exact continuous time process:

$$\mathrm{d}\mathbf{x}_t = -\nabla\bar{U}(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{C.1}$$

with initial condition $\mathbf{x}_0 \sim p_0 \equiv q_0$. Let the distribution of $\mathbf{x}_t$ be denoted by q_t. Let $p_0\mathbb{P}_t$ denote the distribution of the entire path $\{\mathbf{x}_s\}_{s=0}^t$. Consider a second process that describes the Euler-Mayurama discretization of (C.1),

$$\mathrm{d}\tilde{\mathbf{x}}_t = -\mathbf{b}_t(\tilde{\mathbf{x}}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{C.2}$$

with the same initial condition $\tilde{\mathbf{x}}_0 \sim p_0 \equiv q_0$, $\mathbf{b}_t(\tilde{\mathbf{x}}_t) = \sum_{k=0}^\infty \nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\cdot\mathbb{I}\left[t\in[k\eta,(k+1)\eta)\right]$, and the same Brownian motion (synchronous coupling). Let the distribution of $\tilde{\mathbf{x}}_t$ be denoted by $\tilde{p}_t$.

We will analyze the following (Langevin) iterative algorithm for which the initial point $\mathbf{x}_0$ satisfies $\mathbf{x}_0 \sim p_0 \equiv q_0$ and the k^{th} iterate is given by:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta\nabla\bar{U}(\mathbf{x}_{k-1}) + \sqrt{2\eta}\boldsymbol{\xi}_k, \tag{C.3}$$

where $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$. Observe that this algorithm corresponds to the discretized process (C.2) with a fixed step size η, and thus we will use $\bar{p}_k$ to denote the distribution of $\mathbf{x}_k$.

C.1 Guarantees for Wasserstein distance

Define the difference process between $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$ as the process $\mathbf{z}_t$ which evolves according to:

$$\mathrm{d}\mathbf{z}_t = -\left(\nabla\bar{U}(\mathbf{x}_t) - \nabla\bar{U}(\mathbf{x}_0)\right)\mathrm{d}t.$$

To bound the discretization error, here we want to bound the 2-Wasserstein distance between the distributions $q_\eta, \tilde{p}_\eta$ after one step of size η. This is established in the following lemma.

Lemma C.1 (Discretization Error)**.** *Let $\mathbf{x}_0 \sim p_0$ for some probability distribution p_0 and $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$. Let q_η denote the distribution of point $\mathbf{x}_t$ defined by the process* (C.1) *and $\tilde{p}_t$ denote the distribution of point $\tilde{\mathbf{x}}_t$ defined by the process* (C.2)*, as described above. If $\eta < 1/(2\lambda)$ then:*

$$W_2^2(q_\eta, \tilde{p}_\eta) \leq \frac{8d(M+m)^4}{\lambda}\eta^4 + 8(M+m)^4\eta^4 W_2^2(p_0, \bar{p}^*) + 32\delta(M+m)^2 M\eta^4 + 4d(M+m)^2\eta^3 + 8\delta M\eta^2.$$

Proof. By the definition of Wasserstein distance (as the infimum over all couplings) we have,

$$\begin{aligned} W_2^2(q_\eta, \tilde{p}_\eta) \leq \mathbb{E}\left[\|\tilde{\mathbf{x}}_\eta - \mathbf{x}_\eta\|_2^2\right] = \mathbb{E}\left[\|\mathbf{z}_\eta\|_2^2\right] &= \mathbb{E}\left[\left\|-\int_0^\eta \left(\nabla\bar{U}(\mathbf{x}_t) - \nabla\bar{U}(\mathbf{x}_0)\right)\mathrm{d}t\right\|_2^2\right] \\ &\leq \eta\int_0^\eta \mathbb{E}\left[\left\|\left(\nabla\bar{U}(\mathbf{x}_t) - \nabla\bar{U}(\mathbf{x}_0)\right)\right\|_2^2\right]\mathrm{d}t, \end{aligned}$$

where we have used Jensen's inequality. Continuing by using the smoothness of $\bar{U}$ (Eq. (2.7)) and applying Young's inequality ($(a+b)^2 \leq 2a^2 + 2b^2$), we get,

$$\begin{aligned} W_2^2(q_\eta, \tilde{p}_\eta) &\leq \eta\int_0^\eta \mathbb{E}\left[2(M+m)^2\|\mathbf{x}_t - \mathbf{x}_0\|_2^2 + 8\delta M\right]\mathrm{d}t \\ &= 2(M+m)^2\eta\int_0^\eta \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_0\|_2^2\right]\mathrm{d}t + 8\delta M\eta^2 \\ &= 2(M+m)^2\eta\int_0^\eta \mathbb{E}\left[\left\|\int_0^t\left(-\nabla\bar{U}(\mathbf{x}_s)\mathrm{d}s + \sqrt{2}\mathrm{d}\mathbf{B}_s\right)\right\|_2^2\right]\mathrm{d}t + 8\delta M\eta^2, \end{aligned}$$

where the last equality is by the definition of the continuous process (C.1).

By another application of Young's inequality:

$$\begin{aligned} &W_2^2(q_\eta, \tilde{p}_\eta) - 8\delta M\eta^2 \\ &\quad\leq 4(M+m)^2\eta\int_0^\eta \mathbb{E}\left[\left\|\int_0^t \nabla\bar{U}(\mathbf{x}_s)\mathrm{d}s\right\|_2^2\right]\mathrm{d}t + 8(M+m)^2\eta\int_0^\eta \underbrace{\mathbb{E}\left[\|\mathbf{B}_t\|_2^2\right]}_{=d\cdot t}\mathrm{d}t \\ &\quad\overset{(i)}{\leq} 4(M+m)^2\eta\int_0^\eta t\int_0^t \mathbb{E}\left[\underbrace{\left\|\nabla\bar{U}(\mathbf{x}_s) - \nabla\bar{U}(\mathbf{x}^*)\right\|_2^2}_{\leq 2(M+m)^2\|\mathbf{x}_s - \mathbf{x}^*\|_2^2 + 16\delta M}\right]\mathrm{d}s\mathrm{d}t + 4d(M+m)^2\eta^3 \\ &\quad\leq 16(M+m)^4\eta\int_0^\eta t\int_0^t \mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}^*\|_2^2\right]\mathrm{d}s\mathrm{d}t + 32\delta(M+m)^2M\eta^4 + 4d(M+m)^2\eta^3, \end{aligned} \tag{C.4}$$

where (i) is again by Jensen's inequality. We now look to bound the term $\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}^*\|_2^2\right]$. By (Durmus and Moulines, 2016, Proposition 1) (see the first part of Theorem A.5) we have

$$\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}^*\|_2^2\right] \leq \frac{d}{\lambda} + e^{-2\lambda s}\mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\right] = \frac{d}{\lambda} + e^{-2\lambda s}\mathbb{E}_{\mathbf{y}\sim p_0}\left[\|\mathbf{y} - \mathbf{x}^*\|_2^2\right].$$

Let $\mathbf{z} \sim \bar{p}^*$ and assume it is optimally coupled to $\mathbf{y}$, so that $\mathbb{E}_{\mathbf{y}\sim p_0,\mathbf{z}\sim\bar{p}^*}\left[\|\mathbf{y}-\mathbf{z}\|_2^2\right] = W_2^2(p_0,\bar{p^*})$. Then, combining with the inequality above we get:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_s-\mathbf{x}^*\|_2^2\right] &\leq \frac{d}{\lambda} + e^{-2\lambda s}\mathbb{E}_{\mathbf{y}\sim p_0}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right] \\ &\leq \frac{d}{\lambda} + e^{-2\lambda s}\mathbb{E}_{\mathbf{y}\sim p_0,\mathbf{z}\sim\bar{p}^*}\left[\|\mathbf{y}-\mathbf{z}+\mathbf{z}-\mathbf{x}^*\|_2^2\right] \\ &\leq \frac{d}{\lambda} + 2e^{-2\lambda s}\mathbb{E}_{\mathbf{y}\sim p_0,\mathbf{z}\sim\bar{p}^*}\left[\|\mathbf{y}-\mathbf{z}\|_2^2\right] + 2\mathbb{E}_{\mathbf{z}\sim\bar{p}^*}\left[\|\mathbf{z}-\mathbf{x}^*\|_2^2\right] \\ &\leq \frac{3d}{\lambda} + 2e^{-2\lambda s}W_2^2(p_0,\bar{p}^*),\end{aligned}$$

where in the last inequality the bound on $\mathbb{E}_{\mathbf{z}\sim\bar{p}^*}\left[\|\mathbf{z}-\mathbf{x}^*\|_2^2\right] \leq d/\lambda$ follows from (Durmus and Moulines, 2016, Proposition 1) (see second part of Theorem A.5 in Appendix A). Plugging this into (C.4) we get

$$\begin{aligned}W_2^2(q_\eta,\tilde{p}_\eta) &\leq \frac{48d(M+m)^4}{\lambda}\eta\int_0^\eta t\int_0^t \mathrm{d}s\mathrm{d}t + 16(M+m)^4\eta\cdot W_2^2(p_0,\bar{p}^*)\int_0^\eta t(1-e^{-2\lambda t})\mathrm{d}t \\ &\quad + 32\delta(M+m)^2M\eta^4 + 4d(M+m)^2\eta^3 + 8\delta M\eta^2 \\ &\overset{(i)}{\leq} \frac{48d(M+m)^4}{\lambda}\eta\int_0^\eta t\int_0^t \mathrm{d}s\mathrm{d}t + 16(M+m)^4\eta\cdot W_2^2(p_0,\bar{p}^*)\int_0^\eta t^2\mathrm{d}t \\ &\quad + 32\delta(M+m)^2M\eta^4 + 4d(M+m)^2\eta^3 + 8\delta M\eta^2 \\ &\leq \frac{8d(M+m)^4}{\lambda}\eta^4 + 8(M+m)^4\eta^4W_2^2(p_0,\bar{p}^*) \\ &\qquad + 32\delta(M+m)^2M\eta^4 + 4d(M+m)^2\eta^3 + 8\delta M\eta^2.\end{aligned}$$

where (i) follows as $\eta < 1/(2\lambda)$ which implies $1-e^{-2\lambda t} \leq 2\lambda t$. ☐

Theorem C.2. *Let $\bar{p}_k$ denote the distribution of the k^{th} iterate $\mathbf{x}_k$ of the discrete Langevin Monte Carlo given by Eq. (C.3), where $\mathbf{x}_0 \sim \bar{p}_0$. If the step size η satisfies:*

$$\eta \leq \min\left\{\frac{\min\{\lambda,\lambda^2\}\varepsilon^2}{90000d}\cdot\frac{1}{\left((\frac{24}{\lambda\varepsilon})^{\frac{1-\alpha}{\alpha}}L^{\frac{1}{\alpha}}+m\right)}, \frac{1}{2\lambda}, \frac{\lambda}{36(M+m)}\right\},$$

then, for any

$$k \geq \frac{720000d}{\min\{1,\lambda\}\varepsilon^2\lambda^2}\left(\left(\frac{24}{\lambda\varepsilon}\right)^{\frac{1-\alpha}{\alpha}}L^{\frac{1}{\alpha}}+m\right)\log\left(\frac{W_2(\bar{p}_0,\bar{p}^*)}{\varepsilon}\right),$$

we have that $W_2(\bar{p}_k,\bar{p}^*) \leq \varepsilon$, where $\bar{p}^* \propto e^{-\bar{U}}$ is the target distribution.

Proof. By the triangle inequality, we have that at any step k:

$$\begin{aligned}W_2(\bar{p}_k,\bar{p}^*) \leq &\underbrace{e^{-\lambda\eta}W_2(\bar{p}_{k-1},\bar{p}^*)}_{\text{Continuous Process Contraction}} \\ &+ \underbrace{\frac{9\sqrt{d}(M+m)^2}{\sqrt{\lambda}}\eta^2 + 9(M+m)\eta^2W_2(\bar{p}_{k-1},\bar{p}^*) + 6\sqrt{\delta}(M+m)\sqrt{M}\eta^2 + 2\sqrt{d}(M+m)\eta^{3/2} + 4\sqrt{\delta M}\eta}_{\text{Discretization Error}},\end{aligned}$$

where the continuous process contraction follows from (Durmus and Moulines, 2016, Proposition 1) (see third point in Theorem A.5 in Appendix A), while the discretization error is due to the Lemma C.1. First we club together the two terms that contain $W_2(\bar{p}_{k-1}, \bar{p}^*)$ and observe that:

$$\begin{aligned} e^{-\lambda\eta} W_2(\bar{p}_{k-1}, \bar{p}^*) + 9(M+m)\eta^2 W_2(\bar{p}_{k-1}, \bar{p}^*) &\leq \left(1 - \frac{\lambda\eta}{2} + 9(M+m)\eta^2\right) W_2(\bar{p}_{k-1}, \bar{p}^*) \\ &\overset{(i)}{\leq} \left(1 - \frac{\lambda\eta}{2} + \frac{\lambda\eta}{4}\right) W_2(\bar{p}_{k-1}, \bar{p}^*) \\ &= \left(1 - \frac{\lambda\eta}{4}\right) W_2(\bar{p}_{k-1}, \bar{p}^*) \\ &\leq e^{-\lambda\eta/8} W_2(\bar{p}_{k-1}, \bar{p}^*), \end{aligned}$$

where (i) follows as $\eta < \lambda/(36(M+m))$.

Assume that:

$$2\sqrt{d}(M+m)\eta^{3/2} \geq \max\left\{6\sqrt{\delta}(M+m)\sqrt{M}\eta^2, \frac{9\sqrt{d}(M+m)^2}{\sqrt{\lambda}}\eta^2\right\}.$$

(It is not hard to check that this assumption holds for the choice of the step size η and for δ specified below.) Unrolling the recursive inequality for $W_2(\bar{p}_k, \bar{p}^*)$ over k steps, we get:

$$\begin{aligned} W_2(\bar{p}_k, \bar{p}^*) &\leq e^{-\lambda k\eta/8} W_2(\bar{p}_0, \bar{p}^*) + \left(4\sqrt{\delta M} + 6\sqrt{d}(M+m)\eta^{3/2}\right)\eta \sum_{s=0}^{\infty} e^{-\lambda s\eta/8} \\ &\leq e^{-\lambda k\eta/8} W_2(\bar{p}_0, \bar{p}^*) + \frac{4\sqrt{\delta M}\eta}{1 - e^{-\lambda\eta/8}} + \frac{6\sqrt{d}(M+m)\eta^{3/2}}{1 - e^{-\lambda\eta/8}}. \end{aligned}$$

Recalling that $M = (\frac{1}{\delta})^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}}$, we further have:

$$\begin{aligned} W_2(\bar{p}_k, \bar{p}^*) &\leq e^{-\lambda k\eta/8} W_2(\bar{p}_0, \bar{p}^*) + \frac{4\sqrt{\delta \cdot (\frac{1}{\delta})^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}}}\eta}{1 - e^{-\lambda\eta/8}} + \frac{6\sqrt{d}((\frac{1}{\delta})^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} + m)\eta^{3/2}}{1 - e^{-\lambda\eta/8}} \\ &\overset{(i)}{\leq} e^{-\lambda k\eta/8} W_2(\bar{p}_0, \bar{p}^*) + \frac{64 L^{\frac{1}{1+\alpha}} \delta^{\frac{\alpha}{1+\alpha}} \eta}{\lambda\eta} + \frac{96\sqrt{d}((\frac{1}{\delta})^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} + m)\eta^{3/2}}{\lambda\eta} \\ &= e^{-\lambda k\eta/8} W_2(\bar{p}_0, \bar{p}^*) + \frac{64 L^{\frac{1}{1+\alpha}} \delta^{\frac{\alpha}{1+\alpha}}}{\lambda} + \frac{96\sqrt{d}((\frac{1}{\delta})^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} + m)\eta^{1/2}}{\lambda}, \end{aligned}$$

where (i) uses $1 - e^{-a} \geq \frac{a}{2}$, which holds for any $a \in [0, 1]$.

Choosing $\delta = \left(\frac{\lambda\varepsilon}{24 L^{\frac{1}{1+\alpha}}}\right)^{\frac{1+\alpha}{\alpha}}$ and recalling that the step-size is:

$$\eta \leq \frac{\lambda^2\varepsilon^2}{3^2 \cdot (96)^2 d((\frac{1}{\delta})^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} + m)^2} = \frac{\lambda^2\varepsilon^2}{90000d} \cdot \frac{1}{\left(\left(\frac{24}{\lambda\varepsilon}\right)^{\frac{1-\alpha}{\alpha}} L^{\frac{1}{\alpha}} + m\right)},$$

we have:

$$W_2(\bar{p}_k, \bar{p}^*) \leq e^{-\lambda k\eta} W_2(\bar{p}_0, \bar{p}^*) + \frac{2\varepsilon}{3}.$$

Thus, for any:

$$k \geq \frac{720000d}{\min\{1, \lambda\}\varepsilon^2\lambda^2} \left(\left(\frac{24}{\lambda\varepsilon}\right)^{\frac{1-\alpha}{\alpha}} L^{\frac{1}{\alpha}} + m\right) \log\left(\frac{W_2(\bar{p}_0, \bar{p}^*)}{\varepsilon}\right),$$

we have $W_2(\bar{p}_k, \bar{p}^*) \leq \varepsilon$, as claimed. □

C.2 Guarantees for total variation distance

Let $p_0\tilde{\mathbb{P}}_t$ denote the distribution of the entire stochastic process $\{\tilde{\mathbf{x}}_s\}_{s\in[0,t]}$ described by (C.2). Similar to Dalalyan (2017), we use Girsanov's formula (Øksendal, 2003, Chapter 8) to control the Kullback-Leibler divergence between the distributions $p_0\mathbb{P}_t$ and $p_0\tilde{\mathbb{P}}_t$.

$$\mathrm{KL}\left(p_0\mathbb{P}_t|p_0\tilde{\mathbb{P}}_t\right) = \frac{1}{4}\int_0^t \mathbb{E}\left[\|\nabla\bar{U}(\tilde{\mathbf{x}}_s)+\mathbf{b}(\tilde{\mathbf{x}}_s)\|_2^2\right]\mathrm{d}s. \tag{C.5}$$

The application of this identity allows us to bound the discretization error, as in the following lemma.

Lemma C.3 (Discretization Error Bound)**.** *Let $\mathbf{x}^*$ be a point such that $\nabla\bar{U}(\mathbf{x}^*)=0$. Then, for any integer $K\geq 1$ we have:*

$$\begin{aligned}\mathrm{KL}\left(p_0\mathbb{P}_{K\eta}|p_0\tilde{\mathbb{P}}_{K\eta}\right) \leq \frac{(M+m)^3\eta^2}{9}\mathbb{E}_{\mathbf{y}\sim p_0}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right] + \frac{(M+m)^2Kd\eta^3}{9}\\ + \frac{(M+m)^2(K+1)\delta\eta^2}{9} + \frac{(M+m)^2Kd\eta^2}{2} + \delta\eta KM.\end{aligned}$$

Proof. By Girsanov's formula Eq. (C.5) and the definition of $\mathbf{b}(\tilde{\mathbf{x}})$ we have,

$$\begin{aligned}\mathrm{KL}\left(p_0\mathbb{P}_{K\eta}|p_0\tilde{\mathbb{P}}_{K\eta}\right) &= \frac{1}{4}\int_0^{K\eta}\mathbb{E}\left[\|\nabla\bar{U}(\tilde{\mathbf{x}}_s)+\mathbf{b}(\tilde{\mathbf{x}}_s)\|_2^2\right]\mathrm{d}s\\ &= \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla\bar{U}(\tilde{\mathbf{x}}_s)-\nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right]\mathrm{d}s\end{aligned}$$

Using the smoothness property of $\bar{U}$, Eq. (2.7), and Young's inequality, we get:

$$\begin{aligned}\mathrm{KL}\left(p_0\mathbb{P}_{K\eta}|p_0\tilde{\mathbb{P}}_{K\eta}\right) &\leq \frac{(M+m)^2}{2}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\tilde{\mathbf{x}}_s-\tilde{\mathbf{x}}_{k\eta}\|_2^2\right]\mathrm{d}s + \sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\delta M\mathrm{d}s\\ &= \frac{(M+m)^2}{2}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\tilde{\mathbf{x}}_s-\tilde{\mathbf{x}}_{k\eta}\|_2^2\right]\mathrm{d}s + \delta\eta KM.\end{aligned} \tag{C.6}$$

Let us unpack and bound the first term on the right hand side. By the definition of $\tilde{\mathbf{x}}_s$, we have that for each $k\in\{0,\ldots,K-1\}$:

$$\begin{aligned}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\tilde{\mathbf{x}}_s-\tilde{\mathbf{x}}_{k\eta}\|_2^2\right]\mathrm{d}s &= \int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\left\|\int_{k\eta}^{s}\left(\nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})+\sqrt{2}\mathrm{d}\mathbf{B}_r\right)\mathrm{d}r\right\|_2^2\right]\mathrm{d}s\\ &= \int_{k\eta}^{(k+1)\eta}\left(\mathbb{E}\left[\left\|\nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\right\|_2^2(s-k\eta)^2\right]+2d(s-k\eta)\right)\mathrm{d}s.\end{aligned}$$

Plugging this back into Eq. (C.6), we get:

$$\mathrm{KL}\left(p_0\mathbb{P}_{K\eta}|p_0\tilde{\mathbb{P}}_{K\eta}\right) \leq \frac{(M+m)^2\eta^3}{6}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\right\|_2^2\right] + \frac{dK(M+m)^2\eta^2}{2} + \delta\eta KM.$$

By invoking Lemma C.4, we get the desired result. □

Lemma C.4. *Let $\eta\leq 1/(2(M+m))$ and let $K\geq 1$ be an integer. Then:*

$$\eta\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right] \leq \frac{2(M+m)}{3}\mathbb{E}\left[\|\tilde{\mathbf{x}}_0-\mathbf{x}^*\|_2^2\right] + \frac{4(M+m)K\eta d}{3} + \frac{2(K+1)\delta}{3}.$$

Proof. Let $\bar{U}^{(k)} := \bar{U}(\tilde{\mathbf{x}}_{k\eta})$. Then, by the smoothness of $\bar{U}$ (Eq. (2.3)), we have:

$$\begin{aligned}\bar{U}^{(k+1)} &\le \bar{U}^{(k)} + \langle \nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta}), \tilde{\mathbf{x}}_{(k+1)\eta} - \tilde{\mathbf{x}}_{k\eta}\rangle + \frac{M+m}{2}\left\|\tilde{\mathbf{x}}_{(k+1)\eta} - \tilde{\mathbf{x}}_{k\eta}\right\|_2^2 + \frac{\delta}{2} \\ &= \bar{U}^{(k)} - \eta\|\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2 + \sqrt{2\eta}\langle \nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta}), \boldsymbol{\xi}_k\rangle + \frac{M+m}{2}\left\|\eta\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta}) - \sqrt{2\eta}\boldsymbol{\xi}_k\right\|_2^2 + \frac{\delta}{2},\end{aligned}$$

where $\boldsymbol{\xi}_k = \int_{s=k\eta}^{(k+1)\eta} \mathrm{d}\mathbf{B}_s$ is independent Gaussian noise. Taking expectations on both sides:

$$\begin{aligned}\mathbb{E}\left[\bar{U}^{(k+1)}\right] &\le \mathbb{E}\left[\bar{U}^{(k)}\right] - \eta\mathbb{E}\left[\|\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right] + \frac{M+m}{2}\eta^2\mathbb{E}\left[\|\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right] + (M+m)\eta d + \frac{\delta}{2} \\ &= \mathbb{E}\left[\bar{U}^{(k)}\right] - \eta\left(1 - \frac{(M+m)\eta}{2}\right)\mathbb{E}\left[\|\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right] + (M+m)\eta d + \frac{\delta}{2}.\end{aligned}$$

Rearranging the above inequality and summing from $k = 0$ to $K-1$, we get that:

$$\eta\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right] \le \frac{4}{3}\mathbb{E}\left[\bar{U}^{(0)} - \bar{U}^{(K)}\right] + \frac{4(M+m)K\eta d}{3} + \frac{2K\delta}{3}.$$

Let $\bar{U}^* = \inf_{\mathbf{x}\in\mathbb{R}^d}\bar{U}(\mathbf{x})$. Therefore, we have $\bar{U}^K \ge \bar{U}^*$ and $\bar{U}^{(0)} - \bar{U}^* \le (M+m)\|\tilde{\mathbf{x}}_0 - \mathbf{x}^*\|_2^2/2 + \delta/2$. Combining this with the inequality above, we finally have:

$$\eta\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla \bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2\right] \le \frac{2(M+m)}{3}\mathbb{E}\left[\|\tilde{\mathbf{x}}_0 - \mathbf{x}^*\|_2^2\right] + \frac{4(M+m)K\eta d}{3} + \frac{2(K+1)\delta}{3},$$

as claimed. □

Theorem C.5. *Let the initial point be drawn from a Normal distribution $\tilde{\mathbf{x}}_0 \sim p_0 \equiv \mathcal{N}(\mathbf{x}^*, (M+m)^{-1}I_{d\times d})$, where $\mathbf{x}^*$ is a point such that $\nabla\bar{U}(\mathbf{x}^*) = 0$. Let the step size satisfy $\eta < 1/(2(M+m))$. Then, for any integer $K \ge 1$, we have:*

$$\begin{aligned}\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \le& \frac{1}{2}\exp\left(\frac{d}{4}\log\left(\frac{M+m}{\lambda}\right) + \frac{\delta}{4} - \frac{K\eta\lambda}{2}\right) + (M+m)\eta\sqrt{\frac{d}{18}} + (M+m)\sqrt{\frac{Kd\eta^3}{18}} \\ &+ (M+m)\eta\sqrt{\frac{(K+1)\delta}{18}} + (M+m)\eta\sqrt{\frac{Kd}{2}} + \sqrt{\frac{\delta\eta KM}{2}}.\end{aligned}$$

Proof. By applying the triangle inequality to total variation distance, we have:

$$\|\tilde{p}_{K\eta} - \bar{p}^*\|_{\mathrm{TV}} \le \|q_{K\eta} - \bar{p}^*\|_{\mathrm{TV}} + \|\tilde{p}_{K\eta} - q_{K\eta}\|_{\mathrm{TV}}.$$

Recall that, by definition, $\bar{p}_K$ (distribution of the K^{th} iterate) is the same as $\tilde{p}_{K\eta}$, and $q_{K\eta}$ denotes the distribution of the solution to continuous process defined by (C.1) at time $K\eta$. We start off by choosing the initial distribution to be a Gaussian $\tilde{\mathbf{x}}_0 \sim \mathcal{N}(\mathbf{x}^*, (M+m)^{-1}I_{d\times d})$. Therefore, by Lemma A.6, we have that $\|q_{K\eta} - \bar{p}^*\|_{\mathrm{TV}}$ can be bounded as:

$$\|q_{K\eta} - \bar{p}^*\|_{\mathrm{TV}} \le \frac{1}{2}\exp\left(\frac{d}{4}\log\left(\frac{M+m}{\lambda}\right) + \frac{\delta}{2} - \frac{K\eta\lambda}{2}\right).$$

While by Lemma C.3, which holds for a fixed initial point $\mathbf{x}$, combined with the convexity of KL-divergence, we get that,

$$\begin{aligned}\|\tilde{p}_{K\eta} - q_{K\eta}\|_{\mathrm{TV}} \le& \|p_0\mathbb{P}_{K\eta} - p_0\tilde{\mathbb{P}}_{K\eta}\|_{\mathrm{TV}} \le \left(\frac{1}{2}\mathrm{KL}\left(p_0\mathbb{P}_{K\eta}|p_0\tilde{\mathbb{P}}_{K\eta}\right)\right)^{1/2} \\ \le& \sqrt{\frac{(M+m)^3\eta^2}{18}\mathbb{E}_{\mathbf{y}\sim p_0}\left[\|\mathbf{y} - \mathbf{x}^*\|_2^2\right]} + \sqrt{\frac{(M+m)^2Kd\eta^3}{18}} \\ &+ \sqrt{\frac{(M+m)^2(K+1)\delta\eta^2}{18}} + \sqrt{\frac{(M+m)^2Kd\eta^2}{2}} + \sqrt{\frac{\delta\eta KM}{2}},\end{aligned}$$

where the first inequality is by the data-processing inequality and the second is by Pinsker's inequality. It is a simple calculation to show that $\mathbb{E}_{\mathbf{y}\sim p_0}\left[\|\mathbf{y}-\mathbf{x}^*\|_2^2\right] = d/(M+m)$. Combining this with the inequality above yields the desired claim. □

Corollary C.6. *In the setting of the theorem above, if we choose*

$$K \geq \max\left\{\beta, \frac{d}{4\eta\lambda}\log\left(\frac{M+m}{\lambda}\right) + \frac{\delta}{4\eta\lambda}\right\}, \qquad \delta = \min\left\{\left[\frac{\lambda\varepsilon^2}{8d\log(\frac{M+m}{\lambda})L^{\frac{2}{1+\alpha}}}\right]^{\frac{1+\alpha}{2\alpha}}, 1\right\},$$

$$\text{and } \eta \leq \min\left\{1, \frac{1}{2\beta(M+m)}, \frac{\lambda\varepsilon^2}{32d^2(M+m)^2\log\left(\frac{M+m}{\lambda}\right)}\right\}$$

for some $\beta \geq 1$, *where* $M = M(\delta) = \left(\frac{1}{\delta}\right)^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}}$, *then,* $\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \leq \min\{\varepsilon, 1\}$.

Proof. The proof follows by invoking the theorem above and by elementary algebra. □

Remark C.7. In the corollary above, if we treat L, β, and λ as constants, then we find that the mixing time K scales as $\widetilde{\mathcal{O}}(d^{\frac{1+2\alpha}{\alpha}}/\varepsilon^{\frac{2}{\alpha}})$. This recovers the rate obtained in Dalalyan (2017) when no warm start is used of $K = \widetilde{\mathcal{O}}(d^3/\varepsilon^2)$ in the smooth case, $\alpha = 1$. However, as the potential U gets nonsmooth, that is, $\alpha \to 0$, the mixing time blows up.

In this section we have established results in the setting when we sample from distributions with composite potential functions $\bar{U} = U + \psi$, where U is (L,α)-weakly smooth and ψ is m-smooth and λ-strongly convex. If however, we are interested in sampling from a distribution with potential U that is (L,α)-weakly smooth, then we can add a small regularization to the potential exactly as we do in Section 4 to obtain results similar to Corollary 4.1. Again these bounds on the mixing time would blow up as $\alpha \to 0$, but would be polynomial in d and ε when α is sufficiently far from 0.

D Shifted Langevin Monte Carlo

Here, we focus on bounding the mixing time of the sequence defined in Eq. (S-LMC), to which we refer as the Shifted Langevin Monte Carlo. Recall that this sequence is given by:

$$\begin{aligned}\mathbf{y}_{k+1} &= \mathbf{y}_k + \mu\boldsymbol{\omega}_{k-1} - \eta\nabla\bar{U}(\mathbf{y}_k + \mu\boldsymbol{\omega}_{k-1}) + \sqrt{2\eta}\boldsymbol{\xi}_k \\ &= \mathbf{y}_k - \eta\left[\nabla\bar{U}(\mathbf{y}_k + \mu\boldsymbol{\omega}_{k-1}) - \frac{\mu}{\eta}\boldsymbol{\omega}_{k-1}\right] + \sqrt{2\eta}\boldsymbol{\xi}_k.\end{aligned}$$

The only difference compared to the Perturbed Langevin method analyzed in Section 3.2 is in the bound on the variance, established in the following lemma.

Lemma D.1. *For any* $\mathbf{x} \in \mathbb{R}^d$, *and* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, *let* $G(\mathbf{x},\mathbf{z}) := \nabla\bar{U}(\mathbf{x}+\mu\mathbf{z}) - \frac{\mu}{\eta}\mathbf{z}$ *denote a stochastic gradient of* $\bar{U}_\mu$. *Then* $G(\mathbf{x},\mathbf{z})$ *is an unbiased estimator of* $\nabla\bar{U}_\mu$ *whose (normalized) variance can be bounded as:*

$$\sigma^2 := \frac{\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla\bar{U}_\mu(\mathbf{x}) - G(\mathbf{x},\mathbf{z})\right\|_2^2\right]}{d} \leq 8d^{\alpha-1}\mu^{2\alpha}L^2 + 8\mu^2 m^2 + \frac{2\mu^2}{\eta^2}.$$

Proof. Recall that by definition of $\bar{U}_\mu$, we have $\nabla\bar{U}_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{w}}\left[\bar{U}(\mathbf{x}+\mu\mathbf{w})\right]$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, and is independent of $\mathbf{z}$. Clearly, $\mathbb{E}_{\mathbf{z}}[G(\mathbf{x},\mathbf{z})] = \nabla\bar{U}_\mu(\mathbf{x})$.

We now proceed to bound the variance of $G(\mathbf{x},\mathbf{z})$. First, using Young's inequality (which implies $(a+b)^2 \leq 2(a^2+b^2)$, $\forall a, b$) and that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, we have:

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla\bar{U}_\mu(\mathbf{x}) - G(\mathbf{x},\mathbf{z})\right\|_2^2\right] &\leq 2\mathbb{E}_{\mathbf{z}}\left[\|\mathbb{E}_{\mathbf{w}}\left[\bar{U}(\mathbf{x}+\mu\mathbf{w})\right] - \nabla\bar{U}(\mathbf{x}+\mu\mathbf{z})\|_2^2\right] + \frac{2\mu^2\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{z}\|_2^2\right]}{\eta^2} \\ &= 2\mathbb{E}_{\mathbf{z}}\left[\|\mathbb{E}_{\mathbf{w}}\left[\bar{U}(\mathbf{x}+\mu\mathbf{w})\right] - \nabla\bar{U}(\mathbf{x}+\mu\mathbf{z})\|_2^2\right] + \frac{2\mu^2 d}{\eta^2}.\end{aligned}$$

The rest of the proof follows the same argument as the proof of Lemma 3.1 and is omitted. □

We can now establish the following theorem.

Theorem D.2. *Let the initial iterate satisfy* $\mathbf{y}_0 \sim \bar{p}_0$. *If we choose the step-size such that we have* $\eta < 2/(M+m+\lambda)$, *then:*

$$W_2(\bar{p}_K, \bar{p}^*) \leq (1-\lambda\eta)^{K/2} W_2(\bar{p}_0, \bar{p}^*_\mu) + \left(\frac{2(M+m)}{\lambda}\eta d\right)^{1/2} + \sigma\sqrt{\frac{\eta d}{\lambda}}$$
$$+ \frac{8}{\lambda}\left(\frac{3}{2} + \frac{d}{2}\log\left(\frac{2(M+m)}{\lambda}\right)\right)^{1/2}\left(\beta_\mu + \sqrt{\beta_\mu/2}\right),$$

where $\sigma^2 \leq 8d^{\alpha-1}\mu^{2\alpha}L^2 + 8\mu^2 m^2 + \frac{2\mu^2}{\eta^2}$, $M = \frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}}$ *and* $\beta_\mu = \frac{L\mu^{1+\alpha}d^{\frac{1+\alpha}{2}}}{\sqrt{2}(1+\alpha)} + \frac{m\mu^2 d}{2}$.

Proof. By a triangle inequality, we can bound above the Wasserstein distance between p_K and $\bar{p}^*$ by:

$$W_2(\bar{p}_K, \bar{p}^*) \leq W_2(\bar{p}_K, \bar{p}^*_\mu) + W_2(\bar{p}^*, \bar{p}^*_\mu). \tag{D.1}$$

To bound the first term—$W_2(\bar{p}_K, \bar{p}^*_\mu)$—we invoke (Durmus et al., 2019, Theorem 21) (see Theorem A.4 in Appendix A). Recall that $\bar{U}_\mu$ is continuously differentiable, $(M+m)$-smooth (with $M = \frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}}$) and λ-strongly convex. Additionally, $\{\mathbf{y}_k\}_{k=1}^K$ can be viewed as iterates of overdamped Langevin MCMC with respect to the potential specified by $\bar{U}_\mu$ and is updated using unbiased noisy gradients of $\bar{U}_\mu$. Thus, we get:

$$W_2(\bar{p}_K, \bar{p}^*_\mu) \leq (1-\lambda\eta)^{K/2} W_2(\bar{p}_0, \bar{p}^*_\mu) + \left(\frac{2(M+m)}{\lambda}\eta d\right)^{1/2} + \sigma\sqrt{\frac{\eta d}{\lambda}}. \tag{D.2}$$

As was shown in Lemma D.1,

$$\sigma^2 \leq 8d^{\alpha-1}\mu^{2\alpha}L^2 + 8\mu^2 m^2 + \frac{2\mu^2}{\eta^2}. \tag{D.3}$$

The last piece we need is control over the distance between the distributions $\bar{p}^*$ and $\bar{p}^*_\mu$. Notice that by Lemma 2.2 it is possible to control the point-wise distance between $\bar{U}$ and $\bar{U}_\mu$, and hence the likelihood ratio and the KL-divergence between $\bar{p}$ and $\bar{p}_\mu$. We can then use Lemma A.2 to upper bound the Wasserstein distance between these distribution by the KL-divergence. These calculations are worked out in detail in Lemma 3.3 to get:

$$W_2(\bar{p}^*, \bar{p}^*_\mu) \leq \frac{8}{\lambda}\left(\frac{3}{2} + \frac{d}{2}\log\left(\frac{2(M+m)}{\lambda}\right)\right)^{1/2}\left(\beta_\mu + \sqrt{\beta_\mu/2}\right), \tag{D.4}$$

where β_μ is as defined above. By combining Eqs. (D.1)-(E.1) we get a bound on $W_2(p_K, \bar{p}^*)$ in terms of the relevant problem parameters. □

Consider the following choice of μ, η and K:

$$K \geq \frac{1}{\lambda\eta}\log\left(\frac{10W_2(p_0, \bar{p}^*_\mu)}{\varepsilon}\right), \qquad \mu = \left[\frac{\eta L d^{\frac{1-\alpha}{2}}}{2\sqrt{\lambda}}\right]^{\frac{1}{2-\alpha}} \text{ and,} \tag{D.5}$$
$$\eta = \min\left\{\left(\frac{\varepsilon}{1000}\right)^{\frac{2(2-\alpha)}{\alpha}} \frac{\lambda^{\frac{4(2-\alpha)^2}{\alpha(3-\alpha)}}}{L^{2/\alpha}d^{\frac{3-2\alpha}{\alpha}}}, \frac{\lambda^{\frac{3-\alpha}{2(1+\alpha)}}\varepsilon^{\frac{2(1-\alpha)}{1+\alpha}}}{5000C_3^{\frac{2(1-\alpha)}{1+\alpha}}L^{\frac{1-\alpha}{1+\alpha}}d^{\frac{(3+\alpha)(1-\alpha)}{2(1+\alpha)}}}\right\}.$$

and consider a regime of *target accuracy* $C_1 < \varepsilon < C_2$, for two positive constants C_1, C_2, such that the following holds:

1. $M > m$.
2. $\frac{\mu^2}{\eta^2} > 4\max\{d^{\alpha-1}\mu^{2\alpha}L^2, \mu^2 m^2\}$.
3. $d\log\left(\frac{2(M+m)}{\lambda}\right) > 3$.

4. $\frac{L\mu^{1+\alpha}d^{\frac{1+\alpha}{2}}}{\sqrt{2}(1+\alpha)} > \frac{m\mu^2 d}{2}$ and $\beta_\mu < 1$.

5. $\log\left(\frac{2(M+m)}{\lambda}\right) \leq C_3$.

Observe that K blows up as $\alpha \downarrow 0$, since η scales with $(\frac{1000}{\varepsilon})^{2(2-\alpha)/\alpha}$, which tends to zero as $\alpha \downarrow 0$, for any $\varepsilon < 1000$.

A constant $C_2(d, \alpha, L, m, \lambda)$ will exist as our parameters μ and η are monotonically increasing functions of ε and hence $M \propto 1/\mu^{1-\alpha}$ is a monotonically decreasing function of ε. We choose a lower bound on $C_1 \leq \varepsilon$ to simplify the presentation of our corollary that follows to ensure that Condition 5 specified above holds; it is possible to get rid of this condition and it would only change the results by poly-logarithmic factors.

Corollary D.3. *Under the conditions of Theorem D.2, there exist positive constants C_1, C_2, and C_3 such that if $C_1 < \varepsilon < C_2$, then under the choice of η, μ, and K in Eq. (D.5), we have:*

$$W_2(\bar{p}_K, \bar{p}^*) \leq \varepsilon.$$

Proof. The proof follows by invoking Theorem D.2 and using elementary algebra. □

E Omitted results and proofs from Section 3

Lemma 3.1. *For any $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, let $G(\mathbf{x}, \mathbf{z}) := \nabla \bar{U}(\mathbf{x} + \mu\mathbf{z})$ denote a stochastic gradient of $\bar{U}_\mu$. Then $G(\mathbf{x}, \mathbf{z})$ is an unbiased estimator of $\nabla \bar{U}_\mu$ whose (normalized) variance satisfies:*

$$\begin{aligned}\sigma^2 :=& \frac{\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla \bar{U}_\mu(\mathbf{x}) - G(\mathbf{x}, \mathbf{z})\right\|_2^2\right]}{d} \\ \leq& 4d^{\alpha-1}\mu^{2\alpha}L^2 + 4\mu^2 m^2.\end{aligned}$$

Proof. Recall that by definition of $\bar{U}_\mu$, we have $\nabla \bar{U}_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{w}}\left[\bar{U}(\mathbf{x} + \mu\mathbf{w})\right]$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, and is independent of $\mathbf{z}$. Clearly, $\mathbb{E}_{\mathbf{z}}[G(\mathbf{x}, \mathbf{z})] = \nabla \bar{U}_\mu(\mathbf{x})$.

We now proceed to bound the variance of $G(\mathbf{x}, \mathbf{z})$. First, by the definition of $G(\mathbf{x}, \mathbf{z})$:

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla \bar{U}_\mu(\mathbf{x}) - G(\mathbf{x}, \mathbf{z})\right\|_2^2\right] = \mathbb{E}_{\mathbf{z}}\left[\|\mathbb{E}_{\mathbf{w}}\left[\nabla \bar{U}(\mathbf{x} + \mu\mathbf{w})\right] - \nabla \bar{U}(\mathbf{x} + \mu\mathbf{z})\|_2^2\right].$$

By Jensen's inequality,

$$\mathbb{E}_{\mathbf{z}}\left[\|\mathbb{E}_{\mathbf{w}}\left[\nabla \bar{U}(\mathbf{x} + \mu\mathbf{w})\right] - \nabla \bar{U}(\mathbf{x} + \mu\mathbf{z})\|_2^2\right] \leq \mathbb{E}_{\mathbf{z},\mathbf{w}}\left[\|\nabla \bar{U}(\mathbf{x} + \mu\mathbf{w}) - \nabla \bar{U}(\mathbf{x} + \mu\mathbf{z})\|_2^2\right].$$

Hence, using (2.6) and applying Young's inequality ($(a+b)^2 \leq 2(a^2 + b^2)$, $\forall a, b$), we further have:

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla \bar{U}_\mu(\mathbf{x}) - G(\mathbf{x}, \mathbf{z})\right\|_2^2\right] \leq& \mathbb{E}_{\mathbf{z},\mathbf{w}}\left[\left(L\|\mu(\mathbf{w} - \mathbf{z})\|_2^\alpha + m\|\mu(\mathbf{w} - \mathbf{z})\|_2\right)^2\right] \\ \leq& 2L^2\mu^{2\alpha}\mathbb{E}_{\mathbf{z},\mathbf{w}}\left[\|\mathbf{w} - \mathbf{z}\|_2^{2\alpha}\right] + 2m^2\mu^2\mathbb{E}_{\mathbf{z},\mathbf{w}}[\|\mathbf{w} - \mathbf{z}\|_2^2].\end{aligned}$$

Observe that $f(y) = y^\alpha$ is a concave function, $\forall \alpha \in [0, 1]$. Hence, we have that $\mathbb{E}_{\mathbf{z},\mathbf{w}}\left[\|\mathbf{w} - \mathbf{z}\|_2^{2\alpha}\right] \leq \left(\mathbb{E}_{\mathbf{z},\mathbf{w}}\left[\|\mathbf{w} - \mathbf{z}\|_2^2\right]\right)^\alpha$. As $\mathbf{w}$ and $\mathbf{z}$ are independent, $\mathbf{w} - \mathbf{z} \sim \mathcal{N}(\mathbf{0}, 2I_{d\times d})$. Thus, we finally have:

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\nabla \bar{U}_\mu(\mathbf{x}) - G(\mathbf{x}, \mathbf{z})\right\|_2^2\right] \leq 4d^\alpha\mu^{2\alpha}L^2 + 4d\mu^2 m^2,$$

as claimed. □

Lemma 3.3. *Let $\bar{p}^*$ and $\bar{p}^*_\mu$ be the distributions corresponding to the potentials $\bar{U}$ and $\bar{U}_\mu$ respectively. Then:*

$$\begin{aligned}W_2(\bar{p}^*, \bar{p}^*_\mu) \leq& \frac{8}{\lambda}\left(\frac{3}{2} + \frac{d}{2}\log\left(\frac{2(M+m)}{\lambda}\right)\right)^{1/2} \\ &\cdot \left(\beta_\mu + \sqrt{\beta_\mu/2}\right),\end{aligned}$$

where $\beta_\mu := \beta_\mu(d, L, m, \alpha) = \frac{L\mu^{1+\alpha}d^{\frac{1+\alpha}{2}}}{\sqrt{2}(1+\alpha)} + \frac{m\mu^2 d}{2}$.

Proof. By (Bolley and Villani, 2005, Corollary 2.3) (see Lemma A.2 in Appendix A), we have:

$$W_2(\bar{p}^*, \bar{p}^*_\mu) \leq C_{\bar{U}_\mu} \cdot \left(\sqrt{\mathrm{KL}(\bar{p}^* | \bar{p}^*_\mu)} + \left(\frac{\mathrm{KL}(\bar{p}^* | \bar{p}^*_\mu)}{2} \right)^{1/4} \right), \tag{E.1}$$

where

$$C_{\bar{U}_\mu} := 2 \inf_{\mathbf{y} \in \mathbb{R}^d, \gamma > 0} \left(\frac{1}{\gamma} \left(\frac{3}{2} + \log \left(\int_{\mathbb{R}_d} e^{\gamma \|\mathbf{y} - \mathbf{x}\|_2^2} \bar{p}^*_\mu(\mathbf{x}) \mathrm{d}\mathbf{x} \right) \right) \right)^{1/2}.$$

First, let us control the constant $C_{\bar{U}_\mu}$. Without loss of generality, let $\mathbf{0}$ be the global minimizer of $\bar{U}_\mu$ (the minimizer is unique, as the function is strongly convex) and let $\bar{U}_\mu(\mathbf{0}) = 0$ (as the target distribution is invariant to constant shift in the potential). Choose $\mathbf{y} = \mathbf{0}$ and $\gamma = \lambda/4$. By smoothness and strong convexity of $\bar{U}_\mu$ (see Section 2):

$$\lambda \|\mathbf{x}\|_2^2/2 \leq \bar{U}_\mu(\mathbf{x}) \leq (M+m)\|\mathbf{x}\|_2^2/2.$$

Therefore:

$$\begin{aligned} C_{\bar{U}_\mu} &\leq \frac{8}{\lambda}\Big(\frac{3}{2} + \log\Big(\int_{\mathbb{R}_d} e^{\lambda\|\mathbf{x}\|_2^2/4} \bar{p}^*_\mu(\mathbf{x})\mathrm{d}\mathbf{x}\Big)\Big)^{1/2} \leq \frac{8}{\lambda}\Big(\frac{3}{2} + \log\Big(\frac{\int_{\mathbb{R}_d} e^{\lambda\|\mathbf{x}\|_2^2/4} e^{-\lambda\|\mathbf{x}\|_2^2/2}\mathrm{d}\mathbf{x}}{\int_{\mathbb{R}_d} e^{-(M+m)\|\mathbf{x}\|_2^2/2}\mathrm{d}\mathbf{x}}\Big)\Big)^{1/2} \\ &\leq \frac{8}{\lambda}\Big(\frac{3}{2} + \log\Big(\frac{(4\pi/\lambda)^{d/2}}{(2\pi/(M+m))^{d/2}}\Big)\Big)^{1/2} \leq \frac{8}{\lambda}\Big(\frac{3}{2} + \frac{d}{2}\log\Big(\frac{2(M+m)}{\lambda}\Big)\Big)^{1/2}. \end{aligned} \tag{E.2}$$

Next, we can control the Kullback-Leibler divergence between the distributions by using (Dalalyan, 2017, Lemma 3) (see Lemma A.1 in Appendix A). Using Lemma 2.2, we have $0 \leq \bar{U}_\mu - \bar{U} \leq \frac{L\mu^{1+\alpha} d^{(1+\alpha)/2}}{1+\alpha} + \frac{m\mu^2 d}{2}$. Therefore:

$$\begin{aligned} \mathrm{KL}(\bar{p}^* | \bar{p}^*_\mu) &\leq \frac{1}{2} \int \left(\bar{U}(\mathbf{x}) - \bar{U}_\mu(\mathbf{x}) \right)^2 \bar{p}^*(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &\leq \Big(\frac{L\mu^{1+\alpha} d^{(1+\alpha)/2}}{1+\alpha} + \frac{m\mu^2 d}{2} \Big)^2 = {\beta_\mu}^2. \end{aligned} \tag{E.3}$$

Combining Eqs. (E.1)-(E.3), we get the claimed result. □

Theorem 3.6. *Let the initial iterate $\mathbf{y}_0$ be drawn from a probability distribution $\bar{p}_0$. If we choose the step size such that $\eta < 2/(M + m + \lambda)$, then:*

$$\begin{aligned} \|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \leq & \frac{L\mu^{1+\alpha} d^{(1+\alpha)/2}}{1+\alpha} + \frac{\lambda\mu^2 d}{2} \\ & + \sqrt{\mathrm{KL}(\bar{p}_K, \bar{p}^*_\mu)}, \end{aligned}$$

where $\mathrm{KL}(\bar{p}_K, \bar{p}^*_\mu)$ is bounded by $W_2(\bar{p}_K, \bar{p}^*_\mu)$ in Eq. (3.4), and $W_2(\bar{p}_K, \bar{p}^*_\mu)$ is bounded as in Eq. (3.3).

Further, if, for $\varepsilon \in (0, 1]$, we choose

$$\mu = \min\left\{ \frac{\varepsilon^{\frac{1}{1+\alpha}}}{4\max\{1, L^{\frac{1}{1+\alpha}}\} d^{1/2}}, \sqrt{\frac{\varepsilon\lambda}{2m^2 d}} \right\},$$

$$\bar{\varepsilon} = \frac{\varepsilon^2}{4\max\{(M+m)(\sqrt{2d/\lambda + 2\|\mathbf{x}^*\|_2^2} + 2\|\mathbf{x}^*\|_2^2), 1\}},$$

then choosing the step size η and number of steps K as

$$\eta \leq \frac{\bar{\varepsilon}^2 \lambda}{64 d(M+m)} \quad \textit{and} \quad K \geq \frac{\log(2W_2(\bar{p_0}, \bar{p}^*_\mu)/\bar{\varepsilon})}{\lambda\eta},$$

we have $\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \leq \varepsilon$.

Proof. By a triangle inequality, we can upper bound the total variation distance between $\bar{p}_K$ and $\bar{p}^*$:

$$\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \le \|\bar{p}_K - \bar{p}^*_\mu\|_{\mathrm{TV}} + \|\bar{p}^* - \bar{p}^*_\mu\|_{\mathrm{TV}}. \tag{E.4}$$

Same as in the proof of Theorem 3.4, to bound the Wasserstein distance between $\bar{p}_K$ and $\bar{p}^*_\mu$, we invoke (Durmus et al., 2019, Theorem 21) (see Theorem A.4 in Appendix A), which leads to:

$$W_2(\bar{p}_K, \bar{p}^*_\mu) \le (1-\lambda\eta)^{K/2} W_2(\bar{p}_0, \bar{p}^*_\mu) + \left(\frac{2(M+m)}{\lambda}\eta d\right)^{1/2} + \sigma\sqrt{\frac{(1+\eta)\eta d}{\lambda}}, \tag{E.5}$$

Our next step is to relate this bound on the $W_2(\bar{p}_K, \bar{p}^*_\mu)$ to the total variation distance between these distributions. Let $\bar{M} := M + m$, then by the smoothness of $\bar{U}_\mu$ (Lemma 2.2):

$$\|\nabla\bar{U}_\mu(\mathbf{x}) - \bar{U}_\mu(\bar{\mathbf{x}}^*_\mu)\|_2 = \|\nabla\bar{U}_\mu(\mathbf{x})\|_2 \le \bar{M}\|\mathbf{x} - \bar{\mathbf{x}}^*_\mu\|_2 \le \bar{M}\|\mathbf{x}\|_2 + \bar{M}\|\bar{\mathbf{x}}^*_\mu\|_2,$$

where $\bar{\mathbf{x}}^*_\mu$ is the (unique) minimizer of $\bar{U}_\mu$. Further, by (Durmus and Moulines, 2016, Proposition 1) (see Theorem A.5 in Appendix A) we have that $\mathbb{E}_{x\sim\bar{p}^*_\mu}\left[\|\mathbf{x}\|_2^2\right] \le \frac{2d}{\lambda} + 2\|\mathbf{x}^*\|_2^2$. Using these facts it is also possible to bound the second moment of $\bar{p}_K$. Consider random variables $\mathbf{y} \sim \bar{p}_K$ and $\mathbf{x} \sim \bar{p}^*_\mu$, such that $\mathbf{x}$ and $\mathbf{y}$ are optimally coupled; that is, $\mathbb{E}\left[\|\mathbf{x}-\mathbf{y}\|_2^2\right] = W_2^2(\bar{p}_K, \bar{p}^*_\mu)$. Then, using Young's inequality, we have,

$$\mathbb{E}\left[\|\mathbf{y}\|_2^2\right] = \mathbb{E}\left[\|\mathbf{y}-\mathbf{x}+\mathbf{x}\|_2^2\right] \le 2\mathbb{E}\left[\|\mathbf{y}-\mathbf{x}\|_2^2\right] + 2\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] = \frac{4d}{\lambda} + 4\|\mathbf{x}^*\|_2^2 + 2W_2^2(\bar{p}_K, \bar{p}_{\mu^*}).$$

Thus, by invoking (Polyanskiy and Wu, 2016, Proposition 1) (see Proposition A.3 in Appendix A), we get:

$$\begin{aligned}&\mathrm{KL}(\bar{p}_K|\bar{p}^*_\mu)\\ &\quad\le \left(\frac{\bar{M}\sqrt{\frac{2d}{\lambda} + 2\|\mathbf{x}^*\|_2^2}}{2} + \frac{\bar{M}\sqrt{\frac{4d}{\lambda} + 4\|\mathbf{x}^*\|_2^2 + 2W_2^2(\bar{p}_K, \bar{p}_{\mu^*})}}{2} + \bar{M}\|\mathbf{x}^*\|_2\right) W_2(\bar{p}_K, \bar{p}^*_\mu).\end{aligned} \tag{E.6}$$

Finally, by Pinsker's inequality, we have:

$$\begin{aligned}&\|\bar{p}_K - \bar{p}^*_\mu\|_{\mathrm{TV}}\\ &\quad\le \sqrt{\left(\frac{\bar{M}\sqrt{\frac{2d}{\lambda} + 2\|\mathbf{x}^*\|_2^2}}{4} + \frac{\bar{M}\sqrt{\frac{4d}{\lambda} + 4\|\mathbf{x}^*\|_2^2 + 2W_2^2(\bar{p}_K, \bar{p}_{\mu^*})}}{4} + \frac{\bar{M}\|\mathbf{x}^*\|_2}{2}\right) W_2(\bar{p}_K, \bar{p}^*_\mu)},\end{aligned} \tag{E.7}$$

where $W_2(\bar{p}_K, \bar{p}^*_\mu)$ was bounded above in Eq. (E.5). By Lemma E.2 (see Appendix E), we have that:

$$\|\bar{p}^* - \bar{p}^*_\mu\|_{\mathrm{TV}} \le \frac{L\mu^{1+\alpha}d^{(1+\alpha)/2}}{1+\alpha} + \frac{\lambda\mu^2 d}{2}. \tag{E.8}$$

Combining Eqs. (E.4), (E.5), (E.7) and (E.8) yields the first claim.

For the remaining claim, we first choose μ so that $\|\bar{p}^* - \bar{p}^*_\mu\|_{\mathrm{TV}} \le \frac{\varepsilon}{2}$. It is not hard to verify that the following choice suffices, as $m \ge \lambda$ (smoothness is always at least as high as the strong convexity of a function, and m and λ parameters come from the same function ψ):

$$\mu = \min\left\{\frac{\varepsilon^{\frac{1}{1+\alpha}}}{4\max\{1, L^{\frac{1}{1+\alpha}}\}d^{1/2}}, \sqrt{\frac{\varepsilon\lambda}{2m^2 d}}\right\}.$$

It remains to bound $\|\bar{p}_K - \bar{p}^*_\mu\|_{\mathrm{TV}}$ by $\varepsilon/2$. To do so, we first bound $W_2(\bar{p}_K, \bar{p}^*_\mu)$.

To simplify the upper bound on $W_2(\bar{p}_K, \bar{p}^*_\mu)$ from (E.5), we first show that under our choice of μ, $\sigma \le (2(M+m))^{1/2}$. Indeed, as σ, M and m are all non-negative, we have that it suffices to show that $\sigma^2 \le 2(M+m)$. Recall from Lemma 3.1 that:

$$\sigma^2 \le \frac{4\mu^{2\alpha}L^2}{d^{1-\alpha}} + 4\mu^2 m^2.$$

By the choice of μ, as $\varepsilon \leq 1$ and $d \geq 1$, we have that $4\mu^2 m^2 \leq 2m$. Hence, to prove the claim, it remains to show that $\frac{4\mu^{2\alpha}L^2}{d^{1-\alpha}} \leq 2M$. Recalling that $M = \frac{Ld^{\frac{1-\alpha}{2}}}{\mu^{1-\alpha}(1+\alpha)^{1-\alpha}} \geq \frac{Ld^{\frac{1-\alpha}{2}}}{\sqrt{2}\mu^{1-\alpha}}$, and using elementary algebra and the choice of μ, the claim follows.

Hence, we have $W_2(\bar{p}_K, \bar{p}_\mu^*) \leq (1-\lambda\eta)^{K/2} W_2(\bar{p}_0, \bar{p}_\mu^*) + 2\left(\frac{2(M+m)}{\lambda}\eta d\right)^{1/2}$. Choosing:

$$\eta \leq \frac{\bar{\varepsilon}^2 \lambda}{64d(M+m)} \quad \text{and} \quad K \geq \frac{\log(2W_2(\bar{p}_0, \bar{p}_\mu^*)/\bar{\varepsilon})}{\lambda\eta},$$

for some $\bar{\varepsilon}$, ensures $W_2(\bar{p}_K, \bar{p}_\mu^*) \leq \bar{\varepsilon}$. It only remains to choose $\bar{\varepsilon}$ so that $\|\bar{p}_K - \bar{p}_\mu^*\|_{\mathrm{TV}}$, which was bounded in Eq. (E.7), is at most $\varepsilon/2$. Choosing:

$$\bar{\varepsilon} = \frac{\varepsilon^2}{4\max\{(M+m)(\sqrt{2d/\lambda + 2\|\mathbf{x}^*\|_2^2} + 2\|\mathbf{x}^*\|_2^2), 1\}}$$

suffices, which gives the choice of parameters from the statement of the theorem, completing the proof. □

Remark E.1. An interesting byproduct of the sequence of inequalities used in the proof of Theorem 3.6 is that they lead to bounds in total variation distance for (LMC) with stochastic gradients. This follows by combining the result from (Durmus et al., 2019, Theorem 21) (see Theorem A.4 in Appendix A) with the inequality from Eq. (E.7). Combining the two, we have that under the assumptions of Theorem A.4:

$$\begin{aligned}
&\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \\
&\leq \sqrt{\left(\frac{M\sqrt{\frac{2d}{\lambda} + 2\|\mathbf{x}^*\|_2^2}}{4} + \frac{M\sqrt{\frac{4d}{\lambda} + 4\|\mathbf{x}^*\|_2^2 + 2W_2^2(\bar{p}_K, \bar{p}^*)}}{4} + \frac{M\|\mathbf{x}^*\|_2}{2}\right) W_2(\bar{p}_K, \bar{p}^*)},
\end{aligned} \tag{E.9}$$

where $W_2(\bar{p}_K, \bar{p}^*)$ is bounded as in Theorem A.4. Thus, treating M, λ, and $\|\mathbf{x}^*\|_2$ as constants, (LMC) with stochastic gradients takes at most as many iterations to converge to $\|\bar{p}_K - \bar{p}^*\|_{\mathrm{TV}} \leq \varepsilon$ as it takes to converge to $W_2(\bar{p}_K, \bar{p}^*) \leq \varepsilon^2$.

Note that the inequality from Eq. (E.7) is also precisely the reason the mixing time we obtain for (P-LMC) with $\alpha = 1$ (smooth potential) in total variation distance is quadratically higher than for the 2-Wasserstein distance. If this inequality improved, our bound would improve as well.

Finally, we note that an obvious obstacle to carrying out the analysis directly in the total variation distance using the coupling technique (see Appendix C) is the application of Girsanov's formula (see the proof of Lemma C.3). Specifically, when applying Girsanov's formula, we need to bound

$$\mathbb{E}\|\nabla\bar{U}_\mu(\tilde{\mathbf{x}}_s) - \nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2 = \mathbb{E}\|\nabla\bar{U}_\mu(\tilde{\mathbf{x}}_s) - \nabla\bar{U}_\mu(\tilde{\mathbf{x}}_{k\eta}) + \nabla\bar{U}_\mu(\tilde{\mathbf{x}}_{k\eta}) - \nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2.$$

Applying Young's inequality, we need bounds on $\mathbb{E}\|\nabla\bar{U}_\mu(\tilde{\mathbf{x}}_s) - \nabla\bar{U}_\mu(\tilde{\mathbf{x}}_{k\eta})\|_2^2$ and $\mathbb{E}\|\nabla\bar{U}_\mu(\tilde{\mathbf{x}}_{k\eta}) - \nabla\bar{U}(\tilde{\mathbf{x}}_{k\eta})\|_2^2$. While bounding the former is not an issue, the latter can only be bounded using the variance bound from Lemma 3.1. Unfortunately, when $\alpha = 0$, the bound on the variance is at least $\frac{4L^2}{d}$ (a constant independent of the step size), which leads to the similar blow up in the bound on the mixing time as in Corollary C.6.

Lemma E.2. *Let $\bar{p}^*$, and $\bar{p}_\mu^*$ be the distributions corresponding to the potentials $\bar{U}$, and $\bar{U}_\mu$, respectively. Then, we have:*

$$\|\bar{p}^* - \bar{p}_\mu^*\|_{\mathrm{TV}} \leq \frac{L\mu^{1+\alpha}d^{(1+\alpha)/2}}{1+\alpha} + \frac{\lambda\mu^2 d}{2}.$$

Proof. We can control the Kullback-Leibler divergence between the distributions by using (Dalalyan, 2017, Lemma 3) (see Lemma A.1 in Appendix A). Using Lemma 2.2, we have $0 \leq \bar{U}_\mu - \bar{U} \leq \frac{L\mu^{1+\alpha}d^{(1+\alpha)/2}}{1+\alpha} + \frac{\lambda\mu^2 d}{2}$. Therefore:

$$\begin{aligned}
\mathrm{KL}(\bar{p}^*|\bar{p}_\mu^*) &\leq \frac{1}{2}\int \left(\bar{U}(\mathbf{x}) - \bar{U}_\mu(\mathbf{x})\right)^2 \bar{p}^*(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\leq \left(\frac{L\mu^{1+\alpha}d^{(1+\alpha)/2}}{1+\alpha} + \frac{\lambda\mu^2 d}{2}\right)^2
\end{aligned} \tag{E.10}$$

Invoking Pinsker's inequality: $\|\bar{p}^* - \bar{p}_\mu^*\|_{\mathrm{TV}} \leq \sqrt{\mathrm{KL}(p^*|\bar{p}_\mu^*)/2}$ yields the claim. □