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).

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

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

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

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

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 $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} U(\mathbf{x})$ 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 $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

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

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, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:*

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

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

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

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:

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

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:

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

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):

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

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

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

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:

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

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:

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

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:

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

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:*

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

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 $W_2(\bar{p}^*, \bar{p}^*_\mu) + W_2(\bar{p}_K, \bar{p}^*_\mu) \leq \varepsilon$, 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:*

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

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:*

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

where

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

$$\eta \le \frac{\varepsilon^2 \mu^{1-\alpha}\lambda}{1000(L+m)d^{\frac{3-\alpha}{2}}},$$

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

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

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

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:*

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

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

Proof. By the triangle inequality,

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

Applying Lemma A.1 from the appendix,

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

Thus, using Assumption (A3), we get

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

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.