ON APPROXIMATE SAMPLING FROM NON-LOG-CONCAVE NON-SMOOTH DISTRIBUTIONS VIA A LANGEVIN-TYPE MONTE CARLO ALGORITHM

SHOGO NAKAKITA

Abstract. We propose a Langevin-type Monte Carlo algorithm for approximate sampling from distributions whose potentials are non-convex and non-smooth. We show the sampling complexity of the algorithm to iterate samples whose distributions are close to target distributions in 2-Wasserstein distance. The key tools to analyze the complexity are (i) mollification of the potentials of target distributions, (ii) tractability of random sampling from a distribution with its density equal to a mollifier, and (ii) the Liptser–Shiryeav approach for change of measures.

Keywords: diffusion processes; discrete observations; misspecified models; online gradient descent; simultaneous ergodicity; stochastic differential equations; stochastic mirror descent

1. Introduction

We consider the problem of sampling from a Gibbs distribution $\pi(\mathrm{d}x) \propto \exp(-U(x))\mathrm{d}x$ on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$, where $U : \mathbf{R}^d \to [0, \infty)$ is a non-negative potential function. One of the extensively used types of algorithms for the sampling is the Langevin type motivated by the Langevin dynamics, the solution of the following d-dimensional stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla U\left(X_t\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ X_0 = \xi, \tag{1.1}$$

where $\{B_t\}_{t\geq 0}$ is a d-dimensional Brownian motion and ξ is a d-dimensional random vector with $|\xi| < \infty$ almost surely. Since the 2-Wasserstein or total variation distance between π and the law of X_t is convergent under mild conditions, we expect that the laws of Langevin-type algorithms inspired by X_t should converge to π. However, most of the theoretical guarantees for such algorithms are based on the convexity of U, the twice continuous differentiability of U, or the Lipschitz continuity of the gradient ∇U, which do not hold in some modelling in statistics and machine learning. The main interest of this study is proposal of a Langevin-type algorithm whose convergence can be given under minimal assumptions.

To see what difficulties we need to deal with, we review a typical analysis [23] based on the smoothness of U, that is, the twice continuous differentiability of U and the Lipschitz continuity of ∇U. Firstly, the twice continuous differentiability simplifies discussions or plays significant roles in studies of functional inequalities such as Poincaré inequalities and logarithmic Sobolev inequalities [e.g., 2, 7]. Since the functional inequalities for π are essential in analysis of Langevin algorithms, the assumption that U is of class $\mathcal{C}^2$ frequently appears in previous studies. In the second place, the Lipschitz continuity combined with weak conditions ensures the representation of the likelihood ratio between

Komaba Institute for Science, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
E-mail address: nakakita@g.ecc.u-tokyo.ac.jp.
The author was supported by JSPS KAKENHI Grant Number JP21K20318 and JST CREST Grant Numbers JPMJCR21D2 and JPMJCR2115.

$\{X_t\}$ and $\{Y_t\}$, which is critical when we bound the Kullback–Leibler divergence. Liptser and Shiryaev [17] exhibit much weaker conditions than Novikov's or Kazamaki's condition for the explicit representation if (1.1) has the unique strong solution. Since the Lipschitz continuity of ∇U is sufficient for the existence and the uniqueness of the strong solution of (1.1), the framework of Liptser and Shiryaev [17] is applicable.

Our approaches to overcome the non-smoothness of U are mollification, a classical approach to dealing with non-smoothness in differential equations, and the 'misuse' of moduli of continuity for possibly discontinuous functions. We consider the convolution $\bar{U}_r := U * \rho_r$ on U with a weak gradient, and some sufficiently smooth non-negative function ρ_r with compact support in a ball of centre $\mathbf{0}$ and radius $r \in (0, 1]$. We can let $\bar{U}_r$ be of class $\mathcal{C}^2$ and obtain bounds for the constant of Poincaré inequalities for $\bar{\pi}^r(\mathrm{d}x) \propto \exp(-\bar{U}_r(x))\mathrm{d}x$, which suffice to show the convergence of the law of the mollified dynamics $\{\bar{X}^r_t\}$ defined by the SDE

$$\mathrm{d}\bar{X}^r_t = -\nabla \bar{U}_r\left(\bar{X}^r_t\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ \bar{X}^r_0 = \xi$$

to the corresponding Gibbs distribution $\bar{\pi}^r$ in 2-Wasserstein distance owing to Bakry et al. [2], Liu [18], and Lehec [16]. Since the convolution $\nabla \bar{U}_r$ is Lipschitz continuous if the modulus of continuity of a representative ∇U is finite (the convergence to zero is unnecessary), a concise representation of the likelihood ratios between the mollified dynamics $\{\bar{X}^r_t\}$ and $\{Y_t\}$ is available, and we can evaluate the Kullback–Leibler divergence under weak assumptions.

As our analysis relies on mollification, the bias–variance decomposition in estimation of $\nabla \bar{U}_r$ rather than ∇U is crucial. This decomposition enables us to propose new algorithms for U without continuous differentiability. Concretely speaking, we propose a new algorithm named the spherically smoothed Langevin Monte Carlo (SS-LMC) algorithm, whose errors can be arbitrarily small under the dissipativity of U and the boundedness of the modulus of continuity of weak gradients. In addition, we argue zeroth-order versions of these algorithms which are naturally obtained via integration by parts.

1.1. **Related works.** Non-asymptotic analysis of Langevin-based algorithms under convex potentials has been one of the subjects of much attention and intense research [10, 11, 12], and one without convexity has also gathered keen interest [23, 25, 14]. Whilst most previous studies are based on the Lipschitz continuity of the gradients of potentials, several studies extend the settings to those without global Lipschitz continuity. We can classify the settings of potentials in those studies into three types: (1) potentials with convexity but without smoothness [22, 8, 16]; (2) potentials with Hölder continuous gradients and degenerate convexity at infinity or outside a ball [13, 21, 9]; and (3) potentials with local Lipschitz gradients [6, 26]. We review the results (1) and (2) as our study gives the error estimate of a Langevin-type algorithm with gradients whose discontinuity is uniformly bounded.

Pereyra [22], Chatterji et al. [8], and Lehec [16] study Langevin-type algorithms under the convexity and the non-smoothness of potentials. Pereyra [22] presents proximal Langevin-type algorithms for potentials with convexity but without smoothness, which use the Moreau approximations and proximity mappings instead of the gradients. The algorithms are stable in the sense that they have exponential ergodicity for arbitrary step sizes. Chatterji et al. [8] propose the perturbed Langevin Monte Carlo algorithm for non-smooth potential functions and show its performance to approximate Gibbs distributions. The difference between perturbed LMC and ordinary LMC is the inputs of the gradients; we need to add Gaussian noises not only to the gradients but also to their inputs. The main idea of the algorithm is to use Gaussian smoothing of potential functions studied

in Nesterov and Spokoiny [20]; the expectation of non-smooth convex potentials with inputs perturbed by Gaussian random vectors is smoother than the potentials themselves. Lehec [16] investigates the projected LMC for potentials with convexity, global Lipschitz continuity and discontinuous bounded gradients. The analysis is based on convexity and estimate for local times of diffusion processes with reflecting boundaries. The study also generalizes the result to potentials with local Lipschitz by considering a ball as the support of the algorithm and letting the radius diverge.

Erdogdu and Hosseinzadeh [13], Chewi et al. [9], and Nguyen [21] estimate the error of LMC under non-convex potentials with degenerate convexity, weak smoothness, and weak-dissipativity. Erdogdu and Hosseinzadeh [13] show convergence guarantees of LMC under the degenerate convexity at infinity and weak dissipativity of potentials with Hölder continuous gradients, which are the assumptions for modified logarithmic Sobolev inequalities. Nguyen [21] relaxes the condition of Erdogdu and Hosseinzadeh [13] by considering the degenerate convexity outside a large ball and the mixture weak smoothness of potential functions. Chewi et al. [9] analyse the convergence with respect to the Rényi divergence under either Latała–Oleszkiewics inequalities or modified logarithmic Sobolev inequalities.

Note that our proof of the results uses approaches similar to the smoothing of Chatterji et al. [8] and the control of the radius of Lehec [16], whilst our motivations and settings are close to those of the studies under non-convexity.

1.2. **Contributions.** Theorem 4.1, the main theoretical result of this paper, gives an upper bound for the 2-Wasserstein distance between the law of an algorithm we propose and the target distribution π under weak conditions without the convexity, continuous differentiability, or bounded gradients of U. The proposed algorithms are useful for sampling from posterior distributions for some modelling in statistics and machine learning whose potentials are dissipative and weakly differentiable but neither convex nor continuously differentiable (e.g., some losses with elastic net regularization in nonlinear regression and robust regression). Furthermore, we can use the zeroth-order versions of them inspired by the recent study of Roy et al. [24] for black-box sampling with guaranteed accuracy from distributions whose potentials are not convex or smooth.

2. Notations and assumptions

We give some notation and assumptions before the main result.

2.1. **Compact polynomial mollifier.** We consider a compact polynomial mollifier [1] $\rho : \mathbf{R}^d \to [0, \infty)$ of class $\mathcal{C}^1$ as follows:

$$\rho(x) = \begin{cases} \left(\frac{\pi^{d/2}(d/2,3)}{\Gamma(d/2)} \right)^{-1} (1 - |x|^2)^2 & \text{if } |x| \le 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $(\cdot, \cdot)$ is the beta function and $\Gamma(\cdot)$ is the gamma function. Note that $\nabla\rho$ has an explicit L^1-bound, which is the reason to adopt ρ as the mollifier in our analysis. Let $\rho_r(x) = r^{-d}\rho(x/r)$ with $r > 0$.

2.2. **Assumptions on potentials.** Let us set the following assumptions on U.

(C1) $U \in W^{1,\infty}_{\mathrm{loc}}(\mathbf{R}^d)$, that is, U is a locally Lipschitz continuous function.

(C2) $|\nabla U(\mathbf{0})| < \infty$ and the modulus of continuity of ∇U is bounded, that is,

$$\omega_{\nabla U}(r) := \sup_{x,y \in \mathbf{R}^d : |x-y| \le r} |U(x) - U(y)| < \infty$$

for some $r \in (0, 1]$.

(C3) There exist $m, b > 0$ such that for all $x \in \mathbf{R}^d$,

$$\langle x, \nabla U(x) \rangle \geq m |x|^2 - b.$$

2.2.1. *Concrete class of potentials functions with the regularity conditions.* We show a simple class of potential functions satisfying (C1)–(C3) and some examples in Bayesian inference; assume = 1 for simplicity of interpretation. Let us consider a possibly non-convex loss with elastic net regularization such that

$$U(x) = L(x) + \frac{\lambda_1}{\sqrt{d}} R_1(x) + \lambda_2 R_2(x),$$

where $L : \mathbf{R}^d \to [0, \infty)$ is in $W^{1,\infty}_{\mathrm{loc}}(\mathbf{R}^d)$ with a weak gradient ∇L satisfying $\|\nabla L\|_\infty < \infty$, $\lambda_1 \geq 0$, $\lambda_2 > 0$, $R_1(x) = \sum_{i=1}^d |x^{(i)}|$ with $x^{(i)}$ indicating the i-th component of x, and $R_2(x) = |x|^2$. Fix a weak gradient of R_1 as $\nabla R_1(x) = (\mathrm{sgn}(x^{(1)}), \ldots, \mathrm{sgn}(x^{(d)}))$; then $\omega_{\nabla U}(1) \leq 2(\|\nabla L\|_\infty + \lambda_1 + \lambda_2) < \infty$ and $\langle x, \nabla U(x) \rangle \geq \lambda_2 |x|^2 - \|\nabla L\|^2_\infty / 4\lambda_2$ since $\langle x, \nabla R_1(x) \rangle \geq 0$ for all $x \in \mathbf{R}^d$. Note that regularization corresponds to the potentials of prior distributions in Bayesian inference; for instance, letting $\lambda_1 = 0$ is equivalent to choosing a Gaussian prior $N(\mathbf{0}, (2\lambda_2)^{-1} I_d)$ on x.

Non-convex losses with bounded weak gradients often appear in nonlinear and robust regression. We first examine a squared loss for nonlinear regression (or equivalently nonlinear regression with Gaussian errors) such that

$$L_{\mathrm{NLR}}(x) = \frac{1}{2\sigma^2} \sum_{\ell=1}^N (y_\ell - \phi_\ell(x))^2,$$

where $N \in \mathbf{N}$, $\sigma > 0$ is fixed, $y_\ell \in \mathbf{R}$, and $\phi_\ell \in W^{1,\infty}_{\mathrm{loc}}(\mathbf{R}^d)$ with $\|\phi_\ell\|_\infty + \|\nabla \phi_\ell\|_\infty < \infty$ for some $\nabla \phi_\ell$; for example, a two-layer neural network with clipped ReLU activation such that

$$\phi_\ell(x) = \frac{1}{W} \sum_{w=1}^W a_w \varphi_{[0,c]}(\langle x_w, f_\ell \rangle),$$

where $\varphi_{[0,c]}(t) = (0 \vee t) \wedge c$ with $t \in \mathbf{R}$, $a_w \in \{-1, 1\}$ and $c > 0$ are fixed, $f_\ell \in \mathbf{R}^F$, $x = (x_1, \ldots, x_W) \in \mathbf{R}^{FW}$, $F, W \in \mathbf{N}$, and $d = FW$. This L_{NLR} indeed satisfies

$$\|\nabla L_{\mathrm{NLR}}\|_\infty \leq \frac{1}{\sigma^2} \sum_{\ell=1}^N (|y_\ell| + \|\phi_\ell\|_\infty) \|\nabla \phi_\ell\|_\infty < \infty.$$

Another example is a Cauchy loss for robust linear regression (or equivalently linear regression with Cauchy errors) such that

$$L_{\mathrm{RLR}}(x) = \sum_{\ell=1}^N \log(1 + |y_\ell - \langle f_\ell, x \rangle|^2 / \sigma^2)$$

, where $N \in \mathbf{N}$, $\sigma > 0$ is fixed, $y_\ell \in \mathbf{R}$, and $f_\ell \in \mathbf{R}^d$. The fact $|\frac{\mathrm{d}}{\mathrm{d}t} \log(1 + t^2/\sigma^2)| = |2t/(t^2 + \sigma^2)| \leq 1/\sigma$ for all $t \in \mathbf{R}$ yields $\|\nabla L_{\mathrm{RLR}}\|_\infty \leq \sum_{\ell=1}^N |f_\ell| / \sigma < \infty$.

3. Spherically smoothed Langevin Monte Carlo (SS-LMC) algorithm

We propose an algorithm named spherically smoothed Langevin Monte Carlo.

3.1. **Basic idea.** The core idea of the algorithm is approximate sampling from distributions with smoother potentials than the target distribution π rather than direct approximate sampling from π. Concretely speaking, we approximate $\pi(\mathrm{d}x) \propto \exp(-U(x))\mathrm{d}x$ with $\bar{\pi}^r(\mathrm{d}x) \propto \exp(-\bar{U}_r(x))\mathrm{d}x$, where $\bar{U}_r(x) := (\rho_r * U)(x)$. This $\bar{\pi}^r$ has the properties such that $\pi \approx \bar{\pi}^r$ for sufficiently small r and $\bar{\pi}^r$ has a smooth potential $\bar{U}_r \in \mathcal{C}^2$ for all $r > 0$ if $U \in W^{1,\infty}_{\mathrm{loc}}$. Therefore, we expect that sampling from $\bar{\pi}^r$ must be easier than sampling from π itself and it approximates π if $r \ll 1$.

Whilst the standard LMC algorithm aims at the Euler–Maruyama discretization of the solution (Langevin dynamics) of the SDE such as

$$\mathrm{d}X_t = -\nabla U\left(X_t\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

we propose an algorithm being a discretization of the solution of the following mollified SDE:

$$\mathrm{d}\bar{X}^r_t = -\nabla \bar{U}_r\left(\bar{X}^r_t\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

Under mild conditions, $D(\mathcal{L}(\bar{X}^r_t)\|\bar{\pi}^r) \to 0$ as $t \to \infty$.

If we knew $\nabla\bar{U}_r$, the following Langevin-type Monte Carlo algorithm would work:

$$\hat{y}^r_{i+1} = \hat{y}^r_i - \eta\nabla\bar{U}_r(\hat{y}^r_i) + \sqrt{2\eta}z_i,\ \hat{y}^r_0 = \xi,\ i = 0, \ldots, k-1,$$

where $k \in \mathbf{N}$ is the number of iteration, $\eta \in (0, 1]$ is the stepsize, $r \in (0, 1]$ is the radius of mollification, and $z_i \sim^{\text{i.i.d.}} N(\mathbf{0}, I_d)$. However, the computation of $\nabla\bar{U}_r$ is another integration problem and we do not know its explicit representation except for some special cases such as $U(x) = |x|^2$.

Alternatively, we consider a Monte Carlo approximation of the integral $\nabla\bar{U}_r$. Note that the mollifier ρ is also the probability density function of a random variable $\zeta = \tau_1\sqrt{\tau_2}$, where $\tau_1 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\tau_2 \sim \mathrm{Beta}(d/2, 3)$ are independent random variables. Therefore, we can consider spherical smoothing with the random variables whose density is ρ_r as an analogue to Gaussian smoothing of Chatterji et al. [8]. A Monte Carlo approximation of $\nabla\bar{U}_r$ can be given as

$$G(x, \{\zeta_j\}) := \frac{1}{N_B}\sum_{j=1}^{N_B}\nabla U\left(x + r\zeta_j\right), \tag{3.1}$$

where $N_B \in \mathbf{N}$ is the minibatch size of the Monte Carlo approximation and $\zeta_j \sim^{\text{i.i.d.}} \rho$. Under Assumptions (C1) and (C2), it holds that for all $x \in \mathbf{R}^d$,

$$\mathbf{E}\left[G(x, \{\zeta_j\})\right] = \nabla\bar{U}_r(x),\ \mathbf{E}\left[\left|G\left(x, \{\zeta_j\}\right) - \nabla\bar{U}_r(x)\right|^2\right] \le \frac{\left(2\omega_{\nabla U}(r)\right)^2}{N_B}. \tag{3.2}$$

Hence we can approximate $\nabla\bar{U}_r$ if N_B is sufficiently large.

Based on these analyses, we propose the spherically smoothed Langevin Monte Carlo algorithm (Algorithm 1) using G as an approximation of $\nabla\bar{U}_r$.

4. Convergence analysis of SS-LMC

We give a convergence analysis of the proposed algorithm. We first assume the warm start of the algorithm as follows.

(A0) The initial value ξ has the law $\mu_0(\mathrm{d}x) = (\int_{\mathbf{R}^d}\exp(-\Psi(x))\mathrm{d}x)^{-1}\exp(-\Psi(x))\mathrm{d}x$ with $\Psi : \mathbf{R}^d \to [0, \infty)$ and $\psi_0, \psi_2 > 0$ such that $(2 \vee (|\nabla U(\mathbf{0})| + \omega_{\nabla U}(1)))|x|^2 - \psi_0 \le \Psi(x) \le \psi_2|x|^2 + \psi_0$ for all $x \in \mathbf{R}^d$.

The following theorem gives an error estimate of the SS-LMC algorithm. Let $\mu_{i\eta}$ denote the law of y_i.

Algorithm 1 SS-LMC algorithm [19]

Input: $k, n \in \mathbf{N}, \eta, r > 0, y_0 = \xi$
$i \leftarrow 0$
while $i < k$ **do**
  $j \leftarrow 0$
  $G \leftarrow \mathbf{0}$
  while $j < n$ **do**
    $\zeta_j \sim \rho$
    $G \leftarrow G + n^{-1}\nabla U(y_i + r\zeta_j)$
    $j \leftarrow j + 1$
  end while
  $z_i \sim N(\mathbf{0}, I_d)$
  $y_{i+1} \leftarrow y_i - \eta G + \sqrt{2\eta} z_i$
  $i \leftarrow i + 1$
end while
Output: $\{y_i : i = 1, \dots, k\}$

Theorem 4.1 (error estimate of SS-LMC, [19])**.** *Under (C1)–(C3) and (A0), there exists a constant $C \geq 1$ independent of $N_B, r, k, \eta, d, c_{\mathrm{P}}$ such that for all $k \in \mathbf{N}$, $\eta \in (0, 1 \wedge (m/(4(\omega_{\nabla U}(1))^2))]$, $r \in (0, 1]$, and $N_B \in \mathbf{N}$ with $(d^2(\omega_{\nabla U}(r)/r)\eta + (\omega_{\nabla U}(r))^2/N_B)k\eta + r\omega_{\nabla U}(r) \leq 1$,*

$$\mathcal{W}_2\left(\mu_{k\eta}, \pi\right) \leq C\sqrt{d}\sqrt[4]{\left(d^2\frac{\omega_{\nabla U}(r)}{r}\eta + \frac{(\omega_{\nabla U}(r))^2}{N_B}\right)k\eta + r\omega_{\nabla U}(r)} + e^{Cd}\exp\left(-\frac{k\eta}{Cc_{\mathrm{P}}}\right),$$

where c_{P} is the Poincaré constant of π.

4.1. **The sampling complexity of SS-LMC.** We analyse the behaviour of SS-LMC; to see that the convergence $\omega_{\nabla U}(r) \downarrow 0$ is unnecessary, we consider a rough version of the upper bound by replacing $\omega_{\nabla U}(r)$ with the constant $\omega_{\nabla U}(1)$.

Corollary 4.2. *Under (C1)–(C3) and (A0), there exists a constant $C \geq 1$ independent of $N_B, r, k, \eta, d, c_{\mathrm{P}}$ such that for all $N_B \in \mathbf{N}$, $k \in \mathbf{N}$, $\eta \in (0, 1 \wedge (m/(4(\omega_{\nabla U}(1))^2))]$, and $r \in (0, 1]$ with $\left(d^2 r^{-1}\eta + N_B^{-1}\right)k\eta + r \leq 1$,*

$$\mathcal{W}_2\left(\mu_{k\eta}, \pi\right) \leq C\sqrt{d}\sqrt[4]{\left(d^2 r^{-1}\eta + N_B^{-1}\right)k\eta + r} + e^{Cd}\exp\left(-\frac{k\eta}{Cc_{\mathrm{P}}}\right).$$

We yield the following estimate of the sampling complexity.

Proposition 4.3. *Assume (C1)–(C3) and (A0) and fix $\epsilon \in (0, 1]$. If $r = \epsilon^4/48C^4d^2$, $N_B \geq 48C^4d^2(Cc_{\mathrm{P}}(\log(2/\epsilon) + Cd) + 1)/\epsilon^4$, and η satisfies*

$$\eta \leq 1 \wedge \frac{m}{4\left(\omega_{\nabla U}(1)\right)^2} \wedge \frac{r\epsilon^4}{48C^4d^4(Cc_{\mathrm{P}}(\log(2/\epsilon) + Cd) + 1)},$$

then $\mathcal{W}_2\left(\mu_{k\eta}, \pi\right) \leq \epsilon$ for $k = \lceil Cc_{\mathrm{P}}(\log(2/\epsilon) + Cd)/\eta \rceil$.

Since the complexities of N_B and k are given as $N_B = \mathcal{O}(d^2 c_{\mathrm{P}}(\log \epsilon^{-1} + d)/\epsilon^4)$ and $k = \mathcal{O}(d^6 c_{\mathrm{P}}^2(\log \epsilon^{-1} + d)^2/\epsilon^8)$, we obtain the sampling complexity of SS-LMC as $N_B k = \mathcal{O}(d^8 c_{\mathrm{P}}^3(\log \epsilon^{-1} + d)^3/\epsilon^{12})$ or $N_B k = \widetilde{\mathcal{O}}(d^{11} c_{\mathrm{P}}^3/\epsilon^{12})$, where $\widetilde{\mathcal{O}}$ ignores logarithmic factors.

4.2. **Bounds for Poincaré constants.** Whilst we give the sampling complexity in terms of d, ϵ, and c_{P}, it is difficult to obtain dimension-free estimates for c_{P} in general. If we set only Assumptions (C1)–(C3), we obtain $c_{\mathrm{P}} = \mathcal{O}(\exp(\mathcal{O}(d)))$ ([23, 19]) and it is quite loose in the dimension d. We introduce some known assumptions on potentials to give tighter bounds on the constants.

4.2.1. *Bounded perturbation of potentials.* The following result of the perturbation theory [3] is fundamental and essential: if $\mu(\mathrm{d}x) \propto \exp(-V(x))\mathrm{d}x$ has a finite Poincaré constant c_{P}, then distributions $\mu_F(\mathrm{d}x) \propto \exp(-F(x) - V(x))\mathrm{d}x$ with $\|F\|_{L^\infty} < \infty$ have Poincaré constants $c_{\mathrm{P}}(\mu_F)$ such that

$$c_{\mathrm{P}}(\mu_F) \leq \exp(\operatorname{ess\,sup} F - \operatorname{ess\,inf} F) c_{\mathrm{P}}(\mu). \tag{4.1}$$

For instance, we consider $U(x) = \bar{F} \wedge F(x) + |x|^2/2$, where $|x|^2/2$ is the potential function of the d-dimensional standard Gaussian distribution, a nonnegative loss function $F \in W^{1,\infty}_{\mathrm{loc}}$ and a clipping constant $\bar{F} > 0$. Since the Poincaré constant of the$d$-dimensional standard Gaussian distribution is 1, a Poincaré constant c_{P} of the distribution π with this potential U satisfies

$$c_{\mathrm{P}} \leq \exp\left(\bar{F}\right). \tag{4.2}$$

If F satisfies $\sup_{x \in \mathcal{X}} |\nabla F(x)| < \infty$ with the sublevel set $\mathcal{X} := \{x \in \mathbf{R}^d : F(x) \leq \bar{F}\}$, the following estimate holds by regarding $\bar{F}$ as a constant:

$$nk = \tilde{\mathcal{O}}\left(d^{11}\epsilon^{-12}\right). \tag{4.3}$$

4.2.2. *Miclo's trick.* If a distribution satisfies a logarithmic Sobolev inequality with a constant c_{LS}, then the distribution satisfies a Poincaé inequality with a constant $c_{\mathrm{P}} (\leq c_{\mathrm{LS}})$. Using this fact, we give an estimate via the Miclo's trick ([4]). Assume that the potential U has a representation $U = U_c + U_l$, where $U_c \in \mathcal{C}^2(\mathbf{R}^d)$ satisfies $\nabla^2 U_c \geq \lambda I_d, \lambda > 0$ and U_l is M-Lipschitz. Then it holds that

$$c_{\mathrm{P}} \leq c_{\mathrm{LS}} \leq \frac{4}{\lambda} \exp\left(\frac{4M^2\sqrt{2d}}{\lambda\sqrt{\pi}}\right) \tag{4.4}$$

For instance, by setting $U_c = |x|^2/2$ as the potential of a prior distribution and a Lipschitz continuous function as the potential of a likelihood function, we obtain (C1)–(C3) since Lipschitz continuous functions have bounded weak gradient [15]. Therefore, under this setting, we obtain the following complexity estimate:

$$nk = \tilde{\mathcal{O}}\left(\exp\left(\mathcal{O}\left(\sqrt{d}\right)\right)\epsilon^{-12}\right) \tag{4.5}$$

Whilst it diverges faster than any polynomial functions of d, it improves the order in comparison to that given only by Assumptions (C1)–(C3).

5. Remark on a zeroth-order version of the algorithm

First-order algorithms are sometimes prohibitive when gradients are not available or their derivation is computationally expensive; hence it motivates us to consider zeroth-order (or gradient-free) sampling algorithms. Let us consider a zeroth-order version of SS-LMC as an analogue to Roy et al. [24] with the following G_0, an estimator of $\nabla \bar{U}_r$, under (C1)–(C3) and the assumption $|U(x)| < \infty$ for all $x \in \mathbf{R}^d$:

$$G_0(x, \{\zeta_j\}) := \frac{1}{N_B} \sum_{j=1}^{N_B} \frac{U\left(x + r\zeta_j\right) - U\left(x\right)}{r} \frac{4\zeta_j}{(1 - |\zeta_j|^2)},$$

where $N_B \in \mathbf{N}$, $r \in (0,1]$, and $\{\zeta_j\}$ is an i.i.d. sequence of random variables with the density ρ. The fact that

$$\frac{U(x+r\zeta_j)-U(x)}{r}\frac{4\zeta_j}{(1-|\zeta_j|^2)} = \frac{U(x+r\zeta_j)-U(x)}{r}\frac{-\nabla\rho(\zeta_j)}{\rho(\zeta_j)},$$

the symmetricity of ρ, and approximation of $\rho \in \mathcal{C}_0^1(\mathbf{R}^d) \cap W^{1,\infty}(\mathbf{R}^d)$ yield that for all $x \in \mathbf{R}^d$,

$$\begin{aligned}
\mathbf{E}\left[G_0(x,\{\zeta_j\})\right] &= \int_{\mathbf{R}^d} \frac{U(x+rz)-U(x)}{r}\frac{-\nabla\rho(z)}{\rho(z)}\rho(z)\,\mathrm{d}z \\
&= -\int_{\mathbf{R}^d} \frac{U(x+rz)-U(x)}{r}\nabla\rho(z)\,\mathrm{d}z \\
&= -\int_{\mathbf{R}^d} (U(x+y)-U(x))\left(\frac{1}{(r)^{d+1}}\nabla\rho\left(\frac{y}{r}\right)\right)\mathrm{d}y \\
&= \int_{\mathbf{R}^d} \nabla U(x+y)\,\rho_r(y)\,\mathrm{d}y \\
&= \nabla\bar{U}_r(x).
\end{aligned}$$

Therefore, $G_0(x,\{\zeta_j\})$ also gives an unbiased estimation of $\nabla\bar{U}_r(x)$. By evaluating the variance of $G_0(x,\{\zeta_j\})$, we can obtain the result that $\mathcal{W}_2(\mu_{k\eta},\pi) \le \epsilon$ with arbitrary $\epsilon > 0$ for the zeroth-order algorithm adopting this G_0 as the estimator of $\nabla\bar{U}_r$. Note that the complexity deteriorates by a factor of $\mathcal{O}(d^3)$ in comparison to SS-LMC; this is a worse cost than $\mathcal{O}(d)$ which is the deterioration of the zeroth-order algorithm proposed by Roy et al. [24] with respect to the Langevin Monte Carlo algorithm.

6. Sketch of the proof of Theorem 4.1

Let us review the proof of Theorem 4.1 briefly. We decompose the 2-Wasserstein distance as follows:

$$\mathcal{W}_2(\mu_{k\eta},\pi) \le \mathcal{W}_2(\mu_{k\eta},\bar{\nu}^r_{k\eta}) + \mathcal{W}_2(\bar{\nu}^r_{k\eta},\bar{\pi}^r) + \mathcal{W}_2(\bar{\pi}^r,\pi),$$

where $\bar{\nu}^r$ is the law of $\bar{X}^r_t$. If $k\eta$ is sufficiently large, then the second term on the right-hand side converges to zero by the exponential decay of entropy and the Talagrand's inequality. The third term also converges to zero as $r \to 0$ due to the convergence of the Kullback–Leibler divergence. Hence the key term in the analysis is the first term on the right-hand side.

We can reduce the problem to an estimate of the Kullback–Leibler divergence of $\mu_{k\eta}$ from $\bar{\nu}^r_{k\eta}$ owing to Bolley and Villani [5]. The Liptser–Shiryaev approach [17] plays a significant role in the estimate of the divergence. It gives an explicit representation of the likelihood ratio of the solutions of SDEs given that one of the SDEs has a unique strong solution and the other one has a unique weak solution; for details, see Chapter 7 of [17]. Roughly speaking, they use a Brownian motion of the weak solution as the driving process of the SDE with a unique strong solution after a change of measures; it is possible because the unique strong solution can exist for any Brownian motion.

We obtain the following estimate via a slight change of the discussion of [17].

Lemma 6.1 ([19]). *For some $C \ge 1$, for any $k \in \mathbf{N}$ and $\eta \in (0, 1/C]$, it holds true that*

$$D(\mu_{k\eta}\|\bar{\nu}^r_{k\eta}) \le C\left(d^2\frac{\omega_{\nabla U}(r)}{r}\eta + \frac{\omega_{\nabla U}(r)^2}{N_B}\right)k\eta.$$

Using this estimate, we can bound the 2-Wasserstein distance.

References

[1] Anderson, C. R. (2014). Compact polynomial mollifiers for Poisson's equation. Technical report, Department of Mathematics, UCLA, Los Angeles, California.

[2] Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. (2008). A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66.

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

[4] Bardet, J.-B., Gozlan, N., Malrieu, F., and Zitt, P.-A. (2018). Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence. *Bernoulli*, 24(1):333–353.

[5] Bolley, F. and Villani, C. (2005). Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 14(3):331–352.

[6] Brosse, N., Durmus, A., Moulines, É., and Sabanis, S. (2019). The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663.

[7] Cattiaux, P., Guillin, A., and Wu, L.-M. (2010). A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. *Probability Theory and Related Fields*, 148:285–304.

[8] Chatterji, N., Diakonikolas, J., Jordan, M. I., and Bartlett, P. (2020). Langevin Monte Carlo without smoothness. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1716–1726.

[9] Chewi, S., Erdogdu, M. A., Li, M., Shen, R., and Zhang, S. (2022). Analysis of Langevin Monte Carlo from Poincare to Log-Sobolev. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 1–2.

[10] Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676.

[11] Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587.

[12] Durmus, A. and Moulines, E. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882.

[13] Erdogdu, M. A. and Hosseinzadeh, R. (2021). On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 1776–1822.

[14] Erdogdu, M. A., Mackey, L., and Shamir, O. (2018). Global non-convex optimization with discretized diffusions. *32nd Conference on Neural Information Processing Systems*.

[15] Evans, L. C. (2010). *Partial Differential Equations*, volume 19. American Mathematical Society.

[16] Lehec, J. (2021). The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *To appear in the Annals of Applied Probability*.

[17] Liptser, R. S. and Shiryaev, A. N. (2001). *Statistics of Random Processes: I. General theory*. Springer, 2nd edition.

[18] Liu, Y. (2020). The poincaré inequality and quadratic transportation-variance inequalities. *Electronic Journal of Probability*, 25(1):1–16.

[19] Nakakita, S. (2023). Non-asymptotic analysis of Langevin-type Monte Carlo algorithms. `arXiv:2303.12407`.

[20] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566.

[21] Nguyen, D. (2022). Unadjusted Langevin algorithm for sampling a mixture of weakly smooth potentials. *Brazilian Journal of Probability and Statistics*, 36(3):504–539.
[22] Pereyra, M. (2016). Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26:745–760.
[23] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1674–1703.
[24] Roy, A., Shen, L., Balasubramanian, K., and Ghadimi, S. (2022). Stochastic zeroth-order discretizations of langevin diffusions for bayesian inference. *Bernoulli*, 28(3):1810–1834.
[25] Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *32nd Conference on Neural Information Processing Systems*.
[26] Zhang, Y., Akyildiz, Ö. D., Damoulas, T., and Sabanis, S. (2023). Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25.