Exponential Convergence of Projected Langevin Monte Carlo with Non-Convex Potentials

Alireza Daeijavad
Department of Computing and Software
McMaster University
Hamilton, Canada
Email: daeijava@mcmaster.ca

Shahab Asoodeh
Department of Computing and Software
McMaster University
Hamilton, Canada
Email: asoodeh@mcmaster.ca

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. Langevin Monte Carlo (LMC) and its constrained variant, projected LMC (P-LMC), are fundamental algorithms in the sampling literature for generating samples from a target probability distribution $\pi \propto \exp(-u)$ by accessing only the gradient of the potential function u. While tight convergence analyses for these methods have recently been established for convex potentials on bounded domains, their behavior for general potentials remains less understood. This paper extends the analysis to P-LMC for general smooth potentials, both convex and non-convex. We derive exponential convergence rates across multiple distance metrics, including total variation, *I guess squared Hellinger is different from the one we have*squared Hellinger, Rényi divergence, and χ^2-divergence. Our approach leverages techniques from differential privacy, specifically the contractivity of Gaussian kernels over bounded domains.**

I. Introduction

Sampling from a target distribution π using Markov chain Monte Carlo is a fundamental problem in statistics and machine learning [1], and it often amounts to discretizing a diffusion process with π being its stationary measure. When π corresponds to the Gibbs measure $\pi \propto e^{-u}$, where u is the potential function, a popular candidate for such diffusion process is the following stochastic differential equation known as the Langevin dynamics

$$\mathrm{d}X_t = -\nabla u(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{1}$$

where $\{B_t\}_{t\geq 0}$ is Brownian motion in $\mathbb{R}^d$. If $X_t \sim \rho_t$, then ρ_t satisfies the Fokker-Planck equation [2]:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \ \nabla \log \frac{\rho_t}{\pi}\right), \tag{2}$$

where $\nabla\cdot$ represents the divergence. It is evident from (2) that when $\rho_t = \pi$, the term $\frac{\partial \rho_t}{\partial t}$ becomes zero, implying that π is the stationary distribution of the Langevin dynamics.

Discretizing this dynamics, using the Euler–Maruyama method [3], results in the following Markov chain known as Langevin Monte Carlo (LMC) algorithm

$$X_{k+1} = X_k - \eta \nabla u(X_k) + \sqrt{2\eta} Z_k, \tag{3}$$

where $Z_k \sim \mathcal{N}(0, \mathbb{I}_d)$ are independent, and $\eta > 0$ is the discretization parameter. The stationary distribution of LMC algorithm, denotes by π^η, converges to π when η approaches zero, thus we refer to π^η as the *biased* target distribution. We note that the LMC is also referred to as the Unadjusted Langevin Algorithm [2], Langevin MCMC [4], Overdamped Langevin Algorithm [5].

Since the impacts of the discretization bias is rather well-understood in the literature [2, 6–12], tight convergence analysis of (1) typically boils down to finding a tight convergence rate for the LMC. This is the approach adopted in this paper as well.

LMC has been extensively studied in statistical physics [13], statistics[14], and machine learning[1]. However, despite being studied for several decades in multiple communities, the tight convergence rate for a variant of LMC has only recently been determined by Altschuler and Talwar [15]. Specifically, they consider the LMC for target distributions that have finite-sum potentials $u(x) = \sum_{i=1}^n u_i(x)$ and are supported over a compact and convex set $\mathcal{K} \subset \mathbb{R}^d$, leading to the following definition.

Definition 1. *For a compact and convex set $\mathcal{K} \subset \mathbb{R}^d$, potential $u = \sum_{i=1}^n u_i$, batch size $b \leq n$, stepsize $\eta > 0$, and initialization $X_0 \in \mathcal{K}$, the projected Langevin Monte Carlo (P-LMC) is defined as*

$$X_{k+1} = \Pi_{\mathcal{K}}\Big[\psi_{B_k}(X_k) + \sqrt{2\eta} Z_k\Big], \tag{4}$$

where $\Pi_{\mathcal{K}}$ is the Euclidean projection onto $\mathcal{K}$, $\psi_{B_k}(x) \coloneqq x - \frac{1}{b}\sum_{i\in B_k} \eta \nabla u_i(x)$, B_k is a uniform random batch of size b , and $Z_k \sim \mathcal{N}(0, \mathbb{I}^d)$ is an independent noise.

It was shown in [15] that the *mixing time* of P-LMC with convex potentials is $\Theta(\frac{D^2}{\eta}\log\frac{1}{\varepsilon})$, that is the distribution of X_k is within ε total variation (TV) distance of π^η after $k \geq \frac{D^2}{\eta}\log\frac{1}{\varepsilon}$ iterations, where D is the diameter of $\mathcal{K}$. Their proof relies on a novel concept called *shifted divergence* [16–18], which has also been utilized to achieve state-of-the-art privacy analyses for iterative algorithms [19, 20]. This concept, while being powerful, is only applicable to convex potentials.

A. Contribution

In this work, we establish exponential convergence rates for P-LMC with *smooth* potential functions. Compared to existing results (see Table I), our contributions offer two key advantages: (1) the derived bounds apply to a broader class

of potentials, requiring only smoothness, whether convex or non-convex, and (2) the results hold for a wide range of f-divergences, including KL divergence, Rényi divergence, TV distance, and squared Hellinger distance.

Similar to [15], our proof technique builds on a novel privacy analysis framework known as privacy amplification by iteration [21–23]. This framework leverages the contractivity of Markov kernels with respect to a certain f-divergence that underlies differential privacy. Despite the apparent similarity, our proposed approach differs fundamentally from existing techniques in the privacy literature. Privacy analyses typically assess the closeness of two Markov processes initialized identically, whereas convergence analyses, including ours, examine the rate at which a single Markov chain produces indistinguishable outputs when initialized differently. This subtle but important distinction between privacy and sampling necessitates fundamentally different applications of conceptually similar techniques.

B. Related Works

The convergence analysis (or equivalently, the mixing time analysis) of Langevin dynamics (1) and its discretized variants has been extensively studied in the literature under various assumptions. Here, we briefly review those works closely related to P-LMC and defer a more comprehensive literature review to the extended version **[[?]]**.

The study of the mixing time of P-LMC has seen significant progress since the seminal works of Bubeck et al. [24]. Despite this progress, tight mixing bounds (to either π or π^η) remained unresolved until very recently. Altschuler and Talwar [15] provided a complete characterization of the mixing time for P-LMC under convexity and smoothness assumptions. A detailed comparison of our results with theirs is provided in Section IV.

In the non-convex setting, convergence results of LMC (unprojected) have been established for various metrics, including Wasserstein distances [25, 26], KL divergence [2], and Fisher information [27], χ^2-divergence [28], Rényi divergence [28], and f-divergence [29]. Several other convergence results have been established under some assumptions such as Log-Sobolev inequality (LSI) [28], Poincaré inequality (PI) [2], Latała–Oleszkiewicz (LO) inequality [6], Modified Log-Sobolev Inequality (M-LSI) [30], and weak Poincaré inequality (WPI) [31].

The closest work to ours is [32] that analyzed the convergence rate of P-LMC in 1-Wasserstein distance with the non-convex potentials satisfying some mild conditions (such as Lipschitzness and sub-Gaussianity). More specifically, they established convergence to the π by coupling the continuous-time P-LMC with the discrete-time P-LMC. In contrast, our analysis relies exclusively on the discretized version, eliminating the need for transitions between continuous and discrete time.

Table I summarizes convergence results for various LD derived algorithms under different assumptions and metrics. The complete table can be found in Appendix D of the longer version **[CITE longer version]**.

TABLE I
SUMMARY OF CONVERGENCE RESULTS FOR LANGEVIN DYNAMICS AND RELATED ALGORITHMS, WITH 'TYPE' INDICATING CONVERGENCE TO THE TARGET OR BIASED DISTRIBUTION.

Ref	Algo	Convex	Other Assumptions	Metric	Type
[31]	LD	No	Weak PI, s-Hölder	Rényi	to target
[25]	LMC	No	LSI, M-smooth, dissipative	W_2	to target
[31]	LMC	No	Weak PI, s-Hölder	Rényi	to target
[29]	LMC	No	M-smooth, f-Sobolev Inequality	f-divergence	to biased
[32]	P-LMC	No	M-smooth, uniform sub-Gaussian gradients	W_1	to target
[24]	P-LMC	Yes	M-smooth, Lipschitz	TV	to target
[15]	P-LMC	Yes	M-smooth	TV	to biased
Ours	P-LMC	No	M-smooth	f-divergence	to biased

C. Notation and Definitions

Random variables are represented by uppercase letters, such as X. We use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. A mapping $T : \mathcal{X} \to \mathcal{Y}$ between metric spaces $(\mathcal{X}, d)$ and $(\mathcal{Y}, d')$ is a contractive if $d'\big(T(x), T(y)\big) \le \kappa d\big(x, y\big)$ for all $x, y \in \mathcal{X}$, where $0 < \kappa < 1$. Given $\gamma \ge 1$, the E_γ-divergence between two distributions μ and ν on $\mathcal{X}$ is defined as $\mathsf{E}_\gamma(\mu\|\nu) := \sup_{A\in\mathcal{X}}[\mu(A) - \gamma\nu(A)]$. Note that E_γ-divergence reduces to TV distance when $\gamma = 1$. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is M-smooth if ∇f is M-Lipschitz. A Markov kernel $\mathsf{K} : \mathcal{K} \to \mathcal{P}(\mathcal{W})$ is specified by a collection of distributions $\{\mathsf{K}(x) \in \mathcal{P}(\mathcal{K}) : x \in \mathcal{K}\}$. **[Is it how to write Markove kernel? I mean you use $\mathsf{K}(x)$ to denote the conditional distribution $P_{Y|X}(\cdot|x)$.]** Given a Markov kernel $\mathsf{K} : \mathcal{K} \to \mathcal{P}(\mathcal{K})$ and $\mu \in \mathcal{P}(\mathcal{K})$, we denote by $\mathsf{K}\#\mu$ the push-forward of μ under K, i.e.,

$$\mathsf{K}\#\mu = \int_{\mathcal{K}} \mu(\mathrm{d}x)\mathsf{K}(x).$$

Given a convex function f with $f(1) = 0$, we define $D_f(\mu\|\nu) := \int \mathrm{d}\nu f(\mathrm{d}\mu/\mathrm{d}\nu)$. All f-divergences satisfy the data processing inequality (DPI): $D_f(\mathsf{K}\#\mu \parallel \mathsf{K}\#\nu) \le D_f(\mu \parallel \nu)$, for any Markov kernel K.

II. E_γ-MIXING TIME FOR P-LMC

In this section, we aim to establish an exponential convergence rate for P-LMC for for E_γ-divergence for any $\gamma \ge 1$. We will then translate this result to a larger family of f-divergences using properties of E_γ-divergence.

Note that ψ_B the update rule of P-LMC (see Definition 1) can be expressed as a composition of three Markov kernels:

$$\mathsf{K}_k = \Pi_{\mathcal{K}} \circ \mathsf{K}_G^{\sqrt{2\eta}} \circ \Psi_k, \tag{5}$$

where

- $\Psi_k : \mathcal{K} \to \mathcal{P}(\mathcal{K})$, given by $\Psi_k := \sum_{B\subset[n]} \mathbb{P}(B_k = B)\psi_B$. Here, we use a slight abuse of notation, treating a deterministic function as a Markov kernel.
- $\mathsf{K}_G^{\sqrt{2\eta}} : \mathcal{K} \to \mathcal{P}(\mathbb{R}^d)$ is a $\mathcal{K}$-constrained Gaussian kernel $\mathcal{N}(Y_k, 2\eta\mathbb{I}^d)$, where $\mathcal{K} = \{x \in \mathbb{R}^d : \|x\|^2 \le dA, A >$

$0\}$ and Y_k is the output of the previous kernel. **[$\mathcal{K}$-constrained Gaussian kernel is by no means a common term. You need to define better other here or in I.C.]**

- $\Pi_{\mathcal{K}}(\cdot)$ denotes the projection onto the convex set $\mathcal{K}$.

In particular, if we employ sampling without replacement as the method of choosing the batch B_t, we have: $\Psi_t = \sum_{B\subset[n]} \frac{1}{\binom{n}{b}}\psi_B$ **[Don't you need $|B| = b$?]**. Thus, the update rule of P-LMC can be written as:

$$\mathsf{K}_t = \frac{1}{\binom{n}{b}} \sum_{B\subset[n]:|B|=b} \Pi_{\mathcal{K}} \circ \mathsf{K}_G^{\sqrt{2\eta}} \circ \psi_B. \tag{6}$$

Proposition 1 and 2, both play a crucial role in our proof. The proof of Proposition 1 is provided in Appendix A.

Proposition 1. *The diameter of the $\mathcal{S}_B$-constrained Gaussian kernel, where $\mathcal{S}_B \coloneqq \psi_B(\mathcal{K})$, is bounded by $D(\eta M+1)$ where $D = dia(\mathcal{K})$, assuming the potentials are M-smooth.*

Proposition 2 ([22, Proposition 1])**.** *Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact and convex set with diameter D. The contraction coefficient of K_G^σ is equal to $\rho_\gamma(\mathsf{K}_G^\sigma) = \theta_\gamma\left(\frac{D}{\sigma}\right)$, with*

$$\theta_\gamma(r) \coloneqq Q\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big) - \gamma Q\Big(\frac{\log\gamma}{r} + \frac{r}{2}\Big),$$

and $\mathrm{Q}(t) = (2\pi)^{\frac{-1}{2}} \int_t^\infty e^{-u^2/2} du$.

We now present our main Theorem. The sole assumption here is that the potentials are M-smooth. Moreover, using this upper bound, we obtain mixing time of the P-LMC:

Theorem 1. *Consider any batch size $b \in [n]$, discretization parameter η, and any M-smooth potentials $u : \mathcal{K} \to \mathbb{R}$ with $D = dia(\mathcal{K})$. Let $X_k \sim \mu_k$ evolves following P-LMC (Definition 1). The E_γ-divergence between μ_T and π^η is upper bounded by:*

$$\mathsf{E}_\gamma(\mu_T \parallel \pi^\eta) \le \left[\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\right]^T, \tag{7}$$

Moreover, P-LMC mixes to π^η in time:

$$T_{mix,\mathsf{E}_\gamma}(\varepsilon) \le \frac{\log\varepsilon}{\log\left(\theta_\gamma\left(\frac{D(\eta M+1)}{\sqrt{2\eta}}\right)\right)}. \tag{8}$$

Proof sketch: Consider a coupling of two random variables X_0 and X_0', representing two different initializations for P-LMC. The first initialization is chosen randomly, such that $X_0 \sim \mu_0$, while the second initialization is sampled from the stationary distribution of P-LMC, i.e., $X_0' \sim \pi^\eta$. Our goal is to analyze the E_γ-divergence between the laws of these two random variables after applying the update rule of P-LMC for T iterations. Specifically, we aim to compute $\mathsf{E}_\gamma(\mu_T \parallel \pi^\eta)$, where $X_T \sim \mu_T$, and π^η is the distribution of output X_T' after T iterations starting from X_0' which is rooted in the definition of the stationary distribution of a Markov chain.

The update rule of P-LMC consists of three consecutive Markov kernels, as defined in (6) and shown in Fig. 1. The goal is to prove its contractivity: $\mathsf{E}_\gamma(\mu_{k+1} \parallel \pi^\eta) \le c \cdot \mathsf{E}_\gamma(\mu_k \parallel \pi^\eta)$ with $c < 1$. Here, the left-hand side can be expressed as

$$\mathsf{E}_\gamma\Big((\Pi_{\mathcal{K}} \circ \mathsf{K}_G^{\sqrt{2\eta}} \circ \Psi_k)\mu_k \parallel (\Pi_{\mathcal{K}} \circ \mathsf{K}_G^{\sqrt{2\eta}} \circ \Psi_k)\pi^\eta\Big).$$

Since the projection kernel is a form of post-processing, we can upper bound the previous equation by

$$\mathsf{E}_\gamma\Big((\mathsf{K}_G^{\sqrt{2\eta}} \circ \Psi_k)\mu_k \parallel (\mathsf{K}_G^{\sqrt{2\eta}} \circ \Psi_k)\pi^\eta\Big).$$

Moreover, the Gaussian noise addition kernel is a contraction mapping, as established using Proposition 2. Taking this into account and considering joint convexity of E_γ-divergence, the preceding equation is upper-bounded as follows.

$$\frac{1}{\binom{n}{b}} \sum_{B\subset[n]:|B|=b} \theta_\gamma\Big(\frac{dia(\mathcal{S}_B)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\psi_B(\mu_k) \parallel \psi_B(\pi^\eta)\Big)$$

Due to DPI and after some simplifications, the prior equation admits an upper bound given by the next equation.

$$\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_k \parallel \pi^\eta\Big)$$

Consequently, each iteration reduces the divergence between X_k and X_k', ensuring that these variables become progressively closer at each step. Thus, running the process for T iterations retains the contraction property.

Finally, we leverage the fact that the E_γ-divergence between any two distributions, including μ_0 and π^η, is bounded above by one. This completes the proof.

The mixing time for E_γ-divergence is defined as

$$T_{mix,\mathsf{E}_\gamma}(\varepsilon) = \min\{t \in \mathbb{N} \mid \mathsf{E}_\gamma(\mu_t \parallel \pi^\eta) \le \varepsilon\}$$

Based on our upper bound, if we find T such that

$$\left[\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\right]^T \le \varepsilon,$$

we derive an upper bound for $T_{mix,\mathsf{E}_\gamma}(\varepsilon)$, which can be easily obtained through simple calculations.

A detailed mathematical proof is provided in Appendix A. ∎

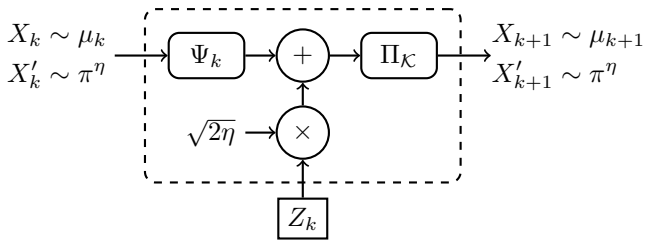

Fig. 1. Visualization of each iteration of P-LMC. Each iteration is composed of three Markov kernels that collectively form a contraction mapping, ensuring exponential convergence.

III. From E_γ-Mixing Time to f-divergence-Mixing Time

The following corollary directly follows for TV distance:

Corollary 1. *Consider the same setting as in Theorem 1. We have the following upper bound for TV distance:*

$$TV(\mu_T, \pi^\eta) \leq \left[1 - 2Q\Big(\frac{D(\eta M+1)}{2\sqrt{2\eta}}\Big)\right]^T, \tag{9}$$

Moreover, P-LMC mixes to π^η in time:

$$T_{mix,TV}(\varepsilon) \leq \frac{\log \varepsilon}{\log\Big(1 - 2Q\big(\frac{D(\eta M+1)}{2\sqrt{2\eta}}\big)\Big)}. \tag{10}$$

A. Mixing results in other notions of distance

In the previous subsection, we derived an upper bound for the E_γ-divergence and TV distance in the non-convex setting. With appropriate modifications, similar results hold for other f-divergences. In particular, every f-divergence can be expressed as an integral of the E_γ-divergence over γ. Consequently, we extend our results to a broad range of f-divergences, with particular emphasis on the KL-divergence, the χ^2 divergence, Hellinger divergence, and Rényi divergence.

Theorem 2. *Consider the same setting as in Theorem 1. Let $f : (0, \infty) \to \mathbb{R}$ be a twice-differentiable convex function with $f(1) = 0$ and $f''(t)$ continuous, satisfying:*

1) *For $t \geq s : t^{-2} f''(t^{-1}) \leq L$.*
2) *Minimum $\delta \in \mathbb{N}$ such that for $t \geq s : t^{1-\delta} f''(t) \leq N$.*

Then, for any f-divergence D_f, the following upper bound holds for $T > \delta$:

$$\begin{aligned} D_f(\mu_T \parallel \pi^\eta) \leq & \frac{r\big(L + Ne^{\delta r^2}\big)}{T-1}(2\pi)^{\frac{-T}{2}} \\ & + \left[f'(s) - \frac{f'(s^{-1})}{s} + f(s^{-1})\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^T \end{aligned}$$

where $r = \frac{D(\eta M+1)}{\sqrt{2\eta}}$ and $s = e^{\frac{r^2}{2}+r}$.

Proof: The proof of this theorem relies on substituting the E_γ-divergence with the integral

$$\begin{aligned} & D_f(P \parallel Q) \\ & \quad = \int_1^\infty \big[f''(\gamma)\mathsf{E}_\gamma(P \parallel Q) + \frac{1}{\gamma^3} f''(\frac{1}{\gamma})\mathsf{E}_\gamma(Q \parallel P)\big]\mathrm{d}\gamma \end{aligned} \tag{11}$$

(See [33, Proposition 3]). The integral can be partitioned into these three integrals.

$$\begin{aligned} & = \underbrace{\int_1^s \left[f''(\gamma) + \gamma^{-3} f''(\gamma^{-1})\right]\left[\theta_\gamma\Big(r\Big)\right]^T \mathrm{d}\gamma}_{A} \\ & \quad + \underbrace{\int_s^\infty f''(\gamma)\left[\theta_\gamma\Big(r\Big)\right]^T \mathrm{d}\gamma}_{B} + \underbrace{\int_s^\infty \gamma^{-3} f''(\gamma^{-1})\left[\theta_\gamma\Big(r\Big)\right]^T \mathrm{d}\gamma}_{C} \end{aligned}$$

To find an upper bound, we ignore the second Q function in $\theta_\gamma(r)$ for all terms A,B, and C. Due to the monotonicity of $Q\big(\frac{\log\gamma}{r} - \frac{r}{2}\big)$ with respect to γ, the term A is bounded by

$$\int_1^s \left[f''(\gamma) + \gamma^{-3} f''(\gamma^{-1})\right]\left[Q\Big(\frac{-r}{2}\Big)\right]^T \mathrm{d}\gamma$$

By leveraging standard integration techniques, we arrive at the desired result for A

$$\left[Q\big(\frac{-r}{2}\big)\right]^T \left[f'(s) - s^{-1} f'(s^{-1}) + f(s^{-1})\right]$$

For terms B and C, we use the inequality $Q(x) < \frac{p(x)}{x}$ for $x > 0$, where $p(x)$ is the probability density function of the normal distribution. Thus we have

$$\begin{aligned} B & \leq \int_s^\infty \gamma^{\delta-1}\gamma^{1-\delta} f''(\gamma)\left[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\right]^T \mathrm{d}\gamma \\ C & \leq \int_s^\infty \gamma^{-3} f''(\gamma^{-1})\left[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\right]^T \mathrm{d}\gamma \end{aligned}$$

Which again Using standard integration techniques and the given assumptions, we derive:

$$\begin{aligned} B & \leq Nre^{\delta r^2}(2\pi)^{\frac{-T}{2}}\left(\frac{1}{T-1}\right) \\ C & \leq Lr(2\pi)^{\frac{-T}{2}}\left(\frac{1}{T-1}\right)\mathrm{d}x \end{aligned}$$

Adding these bounds for A,B, and C finishes the proof. A detailed proof is on Appendix B. ∎

Note that $Q\left(\frac{-r}{2}\right) < 1$, so both terms decay exponentially fast as T increases. The following corollary holds for several popular divergences.

Corollary 2. *Consider the same setting as in Theorem 1. We have the following upper bounds:*

- *For KL-divergence:*

$$\begin{aligned} & D_{KL}(\mu_T \parallel \pi^\eta) \\ & \quad \leq \left[\frac{r^2}{2} + r + 1 - \frac{1}{s}\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^T + \frac{r\big(1+e^{r^2}\big)}{T-1}\big(\frac{1}{2\pi}\big)^{\frac{T}{2}}, \end{aligned}$$

- *For χ^2-divergence with $T > \delta = 1$:*

$$\begin{aligned} & D_{\chi^2}(\mu_T \parallel \pi^\eta) \\ & \quad \leq \left[2s - 1 - \frac{1}{s^2}\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^T + \frac{2r\big(1+e^{r^2}\big)}{T-1}\big(\frac{1}{2\pi}\big)^{\frac{T}{2}}, \end{aligned}$$

- *For Hellinger divergence with $T > \delta = \lceil \alpha - 1 \rceil$:*

$$\begin{aligned} D_{\mathcal{H}_\alpha}(\mu_T \parallel \pi^\eta) \leq & \left[\frac{\alpha s^{\alpha-1} - 1}{\alpha - 1} - \frac{1}{s^\alpha}\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^T \\ & + \frac{\big(1 + e^{\lceil \alpha-1 \rceil r^2}\big)}{(T-1)\big(\alpha r\big)^{-1}}\big(\frac{1}{2\pi}\big)^{\frac{T}{2}}, \end{aligned}$$

- *And for Rényi divergence of order $\alpha \in (1, \infty)$:*

$$D_\alpha(\mu_T \parallel \pi^\eta) \leq \frac{1}{\alpha-1} \log \left[\frac{\alpha r\big(1+e^{\lceil \alpha-1 \rceil r^2}\big)}{(T-1)(\alpha-1)^{-1}} (\frac{1}{2\pi})^{\frac{T}{2}} + \left[\alpha s^{\alpha-1} - 1 - \frac{\alpha-1}{s^\alpha} \right] \left[Q\big(\frac{-r}{2}\big) \right]^T + 1 \right],$$

where $r = \frac{D(\eta M+1)}{\sqrt{2\eta}}$ and $s = e^{\frac{r^2}{2}+r}$.

Proof: The first three inequalities follow directly by substituting $f(t) = t \ln t$ for KL-divergence, $f(t) = (t-1)^2$ for Chi-squared divergence, and $f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}$ for Hellinger divergence. The upper bound for Rényi divergence is derived using the inequality [33, Definition 4]:

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha-1} \log\left(1 + (\alpha-1) D_{\mathcal{H}_\alpha}(P \parallel Q)\right).$$

■

IV. Mixing time for convex functions

Here, we incorporate both the convexity and M-smoothness assumptions for the potentials. This addition highlights that our method for deriving bounds is comprehensive and applicable to both convex and non-convex settings. For comparison with [15], we present our result in terms of TV distance.

Theorem 3. *Consider the assumptions in Theorem 1. Also, assume that the potentials are convex and $\eta \leq \frac{2}{M}$. The TV distance between μ_T and π^η is bounded by:*

$$TV(\mu_T, \pi^\eta) \leq \left[1 - 2Q\Big(\frac{D}{2\sqrt{2\eta}}\Big) \right]^T \tag{12}$$

Furthermore, P-LMC mixes to π^η in time:

$$T_{mix,TV}\Big(\frac{1}{4}\Big) \leq \frac{-2}{\log\left[1 - 2Q\big(\frac{D}{2\sqrt{2\eta}}\big)\right]} \tag{13}$$

Proof: Appendix C ■

The current state-of-the-art upper and lower bounds for the mixing time of P-LMC in convex settings, as established by Altschuler and Talwar [15], are $O\big(\frac{D^2}{\eta}\big)$. While our result does not achieve the optimal performance reported by Altschuler and Talwar [15] in convex regimes, it represents a significant improvement in the upper bounds for non-convex settings. Notably, our approach demonstrates the flexibility to address both non-convex and convex cases with minimal modifications, underscoring its versatility.

V. Discussion

This work establishes exponential convergence guarantees for P-LMC, a constrained version of LMC, across several f-divergences. The sole assumption is that the potentials are smooth, allowing them to be either convex or non-convex. While our results do not match state-of-the-art bounds for convex potentials, our method can be seamlessly applied to the convex setting, demonstrating its versatility.

In our proof, we modeled the update rule of P-LMC as a sequence of three Markov kernels. For two different initializations, each kernel either acts as a contraction or, based on the DPI, does not increase the divergence. Consequently, the combination of these three kernels forms a contraction for each iteration of P-LMC. If one of the initializations is sampled from π^η, this contraction implies that the output of P-LMC progressively approaches the stationary distribution. As a result, we derive an exponentially decaying divergence between the output of P-LMC with a random initialization and π^η. This result is initially formulated in terms of E_γ-divergence and then extended to other notions of distance using tools from Information Theory.

We outline a few natural directions for future work inspired by the results of this paper. First is to extend our proof technique to other sampling algorithms, such as the Metropolis-Adjusted Langevin Algorithm (MALA) [34, 35]. Our approach, which relies on the contractivity of Markov chains, can potentially be adapted to MALA and related methods. This could enhance our proof framework's applicability and robustness in Markov chain-based sampling techniques.

The second direction involves examining the bias of P-LMC. Since our results pertain to the stationary distribution of P-LMC, they must be combined with discretization error bounds to estimate the divergence from the target distribution. While several works [2, 6–12] have addressed this area, it remains unclear whether these bounds are tight.

References

[1] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 2011, pp. 681–688.

[2] S. Vempala and A. Wibisono, "Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] S. Chewi, "Log-concave sampling," *Book draft available at https://chewisinho. github. io*, 2023.

[4] X. Cheng and P. Bartlett, "Convergence of langevin mcmc in kl-divergence," in *Proceedings of Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, F. Janoos, M. Mohri, and K. Sridharan, Eds., vol. 83. PMLR, 07–09 Apr 2018, pp. 186–211.

[5] T. Lelièvre, G. A. Pavliotis, G. Robin, R. Santet, and G. Stoltz, "Optimizing the diffusion of overdamped langevin dynamics," *arXiv preprint arXiv:2404.12087*, 2024.

[6] S. Chewi, M. A. Erdogdu, M. Li, R. Shen, and S. Zhang, "Analysis of langevin monte carlo from poincare to log-sobolev," in *Conference on Learning Theory*. PMLR, 2022, pp. 1–2.

[7] A. Durmus, S. Majewski, and B. Miasojedow, "Analysis of langevin monte carlo via convex optimization," *Journal of Machine Learning Research*, vol. 20, no. 73, pp. 1–46, 2019.

[8] J. M. Altschuler and K. Talwar, "Concentration of the langevin algorithm's stationary distribution," *arXiv preprint arXiv:2212.12629*, 2022.

[9] X. Cheng, D. Yin, P. Bartlett, and M. Jordan, "Stochastic gradient and langevin processes," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1810–1819.

[10] J. Lehec, "The langevin monte carlo algorithm in the non-smooth log-concave case," *The Annals of Applied Probability*, vol. 33, no. 6A, pp. 4858–4874, 2023.

[11] W. Mou, N. Flammarion, M. J. Wainwright, and P. L. Bartlett, "Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity," *Bernoulli*, vol. 28, no. 3, pp. 1577–1601, 2022.

[12] A. Ganesh and K. Talwar, "Faster differentially private samplers via rényi divergence analysis of discretized langevin mcmc," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7222–7233, 2020.

[13] W. Coffey and Y. P. Kalmykov, *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*. World Scientific, 2012, vol. 27.

[14] A. S. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 79, no. 3, pp. 651–676, 2017.

[15] J. Altschuler and K. Talwar, "Resolving the mixing time of the langevin algorithm to its stationary distribution for log-concave sampling," in *Proceedings of Thirty Sixth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, G. Neu and L. Rosasco, Eds., vol. 195. PMLR, 12–15 Jul 2023, pp. 2509–2510.

[16] J. M. Altschuler and S. Chewi, "Shifted composition i: Harnack and reverse transport inequalities," *IEEE Transactions on Information Theory*, 2024.

[17] ——, "Shifted composition ii: shift harnack inequalities and curvature upper bounds," *arXiv preprint arXiv:2401.00071*, 2023.

[18] ——, "Shifted composition iii: Local error framework for kl divergence," *arXiv preprint arXiv:2412.17997*, 2024.

[19] J. Altschuler and K. Talwar, "Privacy of noisy stochastic gradient descent: More iterations without more privacy loss," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3788–3800, 2022.

[20] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2018, pp. 521–532.

[21] S. Asoodeh, M. Diaz, and F. P. Calmon, "Privacy amplification of iterative algorithms via contraction coefficients," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 896–901.

[22] ——, "Contraction of E_γ-divergence and its applications to privacy," *arXiv preprint arXiv:2012.11035*, 2020.

[23] S. Asoodeh and M. Diaz, "Privacy loss of noisy stochastic gradient descent might converge even for non-convex losses," *arXiv preprint arXiv:2305.09903*, 2023.

[24] S. Bubeck, R. Eldan, and J. Lehec, "Sampling from a log-concave distribution with projected langevin monte carlo," *Discrete & Computational Geometry*, vol. 59, pp. 757–783, 2018.

[25] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis," in *Conference on Learning Theory*. PMLR, 2017, pp. 1674–1703.

[26] N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang, "On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 3, pp. 959–986, 2021.

[27] K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and S. Zhang, "Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo," in *Conference on Learning Theory*. PMLR, 2022, pp. 2896–2923.

[28] M. A. Erdogdu, R. Hosseinzadeh, and S. Zhang, "Convergence of langevin monte carlo in chi-squared and rényi divergence," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8151–8175.

[29] S. Mitra and A. Wibisono, "Fast convergence of ϕ-divergence along the unadjusted langevin algorithm and proximal sampler," in *36th International Conference on Algorithmic Learning Theory*, 2025.

[30] M. A. Erdogdu and R. Hosseinzadeh, "On the convergence of langevin monte carlo: The interplay between tail growth and smoothness," in *Conference on Learning Theory*. PMLR, 2021, pp. 1776–1822.

[31] A. Mousavi-Hosseini, T. K. Farghly, Y. He, K. Balasubramanian, and M. A. Erdogdu, "Towards a complete analysis of langevin monte carlo: Beyond poincaré inequality," in *Proceedings of Thirty Sixth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 195. PMLR, 12–15 Jul 2023, pp. 1–35.

[32] A. Lamperski, "Projected stochastic gradient langevin algorithms for constrained sampling and non-convex learning," in *Conference on Learning Theory*. PMLR, 2021, pp. 2891–2937.

[33] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[34] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to langevin diffusions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 1, pp. 255–268, 1998.

[35] J. M. Altschuler and S. Chewi, "Faster high-accuracy log-concave sampling via algorithmic warm starts," *Journal of the ACM*, vol. 71, no. 3, pp. 1–55, 2024.

[36] D. Bakry, I. Gentil, M. Ledoux *et al.*, *Analysis and geometry of Markov diffusion operators*. Springer, 2014, vol. 103.

[37] A. S. Dalalyan and A. Karagulyan, "User-friendly guarantees for the langevin monte carlo with inaccurate gradient," *Stochastic Processes and their Applications*, vol. 129, no. 12, pp. 5278–5311, 2019.

[38] A. Durmus and E. Moulines, "High-dimensional bayesian inference via the unadjusted langevin algorithm," *arXiv preprint arXiv:1605.01559*, 2016.

[39] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan, "Sharp convergence rates for langevin dynamics in the nonconvex setting," *arXiv preprint arXiv:1805.01648*, 2018.

[40] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan, "Sampling can be faster than optimization," *Proceedings of the National Academy of Sciences*, vol. 116, no. 42, pp. 20 881–20 885, 2019.

[41] A. S. Dalalyan, A. Karagulyan, and L. Riou-Durand, "Bounding the error of discretized langevin algorithms for non-strongly log-concave targets," *Journal of Machine Learning Research*, vol. 23, no. 235, pp. 1–38, 2022.

[42] D. Nguyen, X. Dang, and Y. Chen, "Unadjusted langevin algorithm for non-convex weakly smooth potentials," *Communications in Mathematics and Statistics*, pp. 1–58, 2023.

APPENDIX A
PROOF OF THEOREM 1

We begin by proving Proposition 1, which bounds the diameter of a $\mathcal{S}$-constrained Gaussian kernel:

$$
\begin{aligned}
&dia(\mathcal{S}_B)\\
&= \sup_{w_1,w_2\in\mathcal{K}} ||\psi_B(w_2)-\psi_B(w_1)||\\
&= \sup_{w_1,w_2\in\mathcal{K}} \Big\|(w_2-w_1)+\frac{\eta}{b}\sum_{i\in B}\big(\nabla u_i(w_1)-\nabla u_i(w_2)\big)\Big\|\\
&\le \sup_{\substack{w_1\in\mathcal{K}\\ w_2\in\mathcal{K}}} \Big\|(w_2-w_1)\Big\| + \frac{\eta}{b}\sum_{i\in B}\sup_{\substack{w_1\in\mathcal{K}\\ w_2\in\mathcal{K}}} \Big\|\nabla u_i(w_1)-\nabla u_i(w_2)\Big\|\\
&\le D+\frac{\eta}{b}\sum_{i\in B}\sup_{w_1,w_2\in\mathcal{K}}\Big\|M\times(w_1-w_2)\Big\|\\
&= D(\eta M+1).
\end{aligned}
$$

The first step follows from substituting the definition of the function ψ_B, the second step uses the triangle inequality, and the next step relies on the M-smoothness of the potentials.

Using Proposition 1, we now analyze the E_γ-divergence between the outputs of the Markov chain after $T+1$ iterations, where the initial inputs are sampled from π^η and μ_0:

$$
\begin{aligned}
&\mathsf{E}_\gamma(\mu_{T+1}\parallel\pi^\eta)\\
&\quad=\mathsf{E}_\gamma\Big((\Pi_\mathcal{K}\circ\mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_T)\mu_T\parallel(\Pi_\mathcal{K}\circ\mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_T)\pi^\eta\Big)\\
&\quad\le\mathsf{E}_\gamma\Big((\mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_T)\mu_T\parallel(\mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_T)\pi^\eta\Big)\\
&\quad\le\frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\|B|=b}}\mathsf{E}_\gamma\Big((\mathsf{K}_G^{\sqrt{2\eta}}\circ\psi_B)\mu_T\parallel(\mathsf{K}_G^{\sqrt{2\eta}}\circ\psi_B)\pi^\eta\Big)\\
&\quad\le\frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\|B|=b}}\theta_\gamma\Big(\frac{dia(\mathcal{S}_B)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\psi_B(\mu_T)\parallel\psi_B(\pi^\eta)\Big)\\
&\quad\le\frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\|B|=b}}\theta_\gamma\Big(\frac{dia(\mathcal{S}_B)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_T\parallel\pi^\eta\Big) \qquad (14)\\
&\quad\le\frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\|B|=b}}\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_T\parallel\pi^\eta\Big)\\
&\quad=\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_T\parallel\pi^\eta\Big)
\end{aligned}
$$

The first step follows directly from the definition of the P-LMC Markov kernel in (6) and the fact that π^η is the stationary distribution of this Markov kernel. Next, we apply DPI, followed by utilizing the convexity of $(P,Q)\mapsto\mathsf{E}_\gamma(P\parallel Q)$. The next step leverages Proposition 2. Again, we apply DPI and Proposition 1. The last step holds because the terms are identical, so we can multiply them by their count.

By applying the same operations on $\mathsf{E}_\gamma(\mu_T\parallel\pi^\eta)$ over T iterations, we obtain the following result:

$$
\mathsf{E}_\gamma(\mu_{T+1}\parallel\pi^\eta)\le\left[\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\right]^{(T+1)}\mathsf{E}_\gamma\Big(\mu_0\parallel\pi^\eta\Big).
$$

Finally, since E_γ-divergence is trivially bounded by 1, we obtain the desired result.

We now turn to proving the second part of the Theorem which is to find an upper bound for the mixing time $T_{mix,\mathsf{E}_\gamma}(\varepsilon)$ under E_γ-divergence ($\varepsilon<1$). Specifically, we aim to determine T such that $\mathsf{E}_\gamma(\mu_T\parallel\pi^\eta)\le\varepsilon$, which holds when

$$
\left[\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\right]^T\le\varepsilon
$$

Taking the logarithm of both sides, we have

$$
T\ge\frac{\log\varepsilon}{\log\left(\theta_\gamma\left(\frac{D(\eta M+1)}{\sqrt{2\eta}}\right)\right)}
$$

As a result

$$
T_{mix,\mathsf{E}_\gamma}(\varepsilon)\le\frac{\log\varepsilon}{\log\left(\theta_\gamma\left(\frac{D(\eta M+1)}{\sqrt{2\eta}}\right)\right)}
$$

APPENDIX B
PROOF OF THEOREM 2

We set $r=\frac{D(\eta M+1)}{\sqrt{2\eta}}$ and $s=e^{\frac{r^2}{2}+r}$. By substituting our upper bound from Theorem 1 into (11), we obtain:

$$
D_f(\mu_T\parallel\pi^\eta)\le\int_1^\infty\Big(f''(\gamma)+\gamma^{-3}f''(\gamma^{-1})\Big)\left[\theta_\gamma\Big(r\Big)\right]^T\mathrm{d}\gamma
$$

To simplify the analysis and computation, the previous integral is split as follows:

$$
\begin{aligned}
&=\underbrace{\int_1^s\left[f''(\gamma)+\gamma^{-3}f''(\gamma^{-1})\right]\left[\theta_\gamma\Big(r\Big)\right]^T\mathrm{d}\gamma}_{A}\\
&\quad+\underbrace{\int_s^\infty f''(\gamma)\left[\theta_\gamma\Big(r\Big)\right]^T\mathrm{d}\gamma}_{B}+\underbrace{\int_s^\infty\gamma^{-3}f''(\gamma^{-1})\left[\theta_\gamma\Big(r\Big)\right]^T\mathrm{d}\gamma}_{C}
\end{aligned}
$$

For term A, after ignoring the second term in $\theta_\gamma(r)$, we observe that $Q\big(\frac{\log\gamma}{r}-\frac{r}{2}\big)$ is a monotonically decreasing function of γ. Therefore, for $1\le\gamma\le s$, it attains its maximum at $\gamma^*=1$. Consequently, we have:

$$
\begin{aligned}
A&=\int_1^s\left[f''(\gamma)+\gamma^{-3}f''(\gamma^{-1})\right]\left[\theta_\gamma\Big(r\Big)\right]^T\mathrm{d}\gamma\\
&\le\int_1^s\left[f''(\gamma)+\gamma^{-3}f''(\gamma^{-1})\right]\left[Q\Big(\frac{\log\gamma}{r}-\frac{r}{2}\Big)\right]^T\mathrm{d}\gamma\\
&\le\int_1^s\left[f''(\gamma)+\gamma^{-3}f''(\gamma^{-1})\right]\left[Q\Big(\frac{-r}{2}\Big)\right]^T\mathrm{d}\gamma\\
&=\left[Q\Big(\frac{-r}{2}\Big)\right]^T\int_1^s\left[f''(\gamma)+\gamma^{-3}f''(\gamma^{-1})\right]\mathrm{d}\gamma\\
&=\left[Q\big(\frac{-r}{2}\big)\right]^T\left[f'(s)-f'(1)+\int_1^s\gamma^{-3}f''(\gamma^{-1})\mathrm{d}\gamma\right]\\
&=\left[Q\big(\frac{-r}{2}\big)\right]^T\left[f'(s)-f'(1)+\int_{s^{-1}}^1 tf''(t)\mathrm{d}t\right]
\end{aligned}
$$

$$= \Big[Q\big(\frac{-r}{2}\big)\Big]^T \Big[f'(s) - f'(1) + tf'(t)\Big|_{\frac{1}{s}}^1 - \int_{s^{-1}}^1 f'(t)\mathrm{d}t\Big]$$
$$= \Big[Q\big(\frac{-r}{2}\big)\Big]^T \big[f'(s) - s^{-1}f'(s^{-1}) - f(1) + f(s^{-1})\big]$$
$$= \Big[Q\big(\frac{-r}{2}\big)\Big]^T \big[f'(s) - s^{-1}f'(s^{-1}) + f(s^{-1})\big]$$

As mentioned earlier, the first step involves ignoring the second term in $\theta_\gamma(r)$, followed by upper bounding the Q-function by its maximum. Assuming that f'' is continuous allows us to compute the integral for the first term. Then, by applying integration by substitution ($t = \gamma^{-1}$) and integration by parts in the next two steps, we derive the bound. The final equality holds because $f(1) = 0$ for f-divergences.

For term B, we apply the following inequality: $Q(x) < \frac{p(x)}{x}$ for $x > 0$, where $p(x)$ is the probability density function of the normal distribution. Therefore, we have:

$$\begin{aligned}
B &= \int_s^\infty f''(\gamma)\Big[\theta_\gamma\Big(r\Big)\Big]^T \mathrm{d}\gamma \\
&\le \int_s^\infty f''(\gamma)\Big[Q\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)\Big]^T \mathrm{d}\gamma \\
&\le \int_s^\infty f''(\gamma)\Big[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\Big]^T \mathrm{d}\gamma \\
&= \int_s^\infty \gamma^{\delta-1}\gamma^{1-\delta} f''(\gamma)\Big[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\Big]^T \mathrm{d}\gamma \\
&\le N\int_s^\infty \gamma^{\delta-1}\Big[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\Big]^T \mathrm{d}\gamma \\
&= N(2\pi)^{\frac{-T}{2}}\int_{\frac{r^2}{2}+r}^\infty e^{\delta t}\left[\frac{\exp\Big(-\frac{\left(\frac{t}{r} - \frac{r}{2}\right)^2}{2}\Big)}{\frac{t}{r} - \frac{r}{2}}\right]^T \mathrm{d}t \\
&= Nr(2\pi)^{\frac{-T}{2}}\int_1^\infty \Big(e^{rx + \frac{r^2}{2}}\Big)^\delta \left[\frac{e^{-\frac{x^2}{2}}}{x}\right]^T \mathrm{d}x \\
&= Nr(2\pi)^{\frac{-T}{2}}\int_1^\infty \left[\frac{e^{\frac{-(x-r)^2}{2} + r^2}}{x}\right]^\delta \left[\frac{e^{-\frac{x^2}{2}}}{x}\right]^{T-\delta} \mathrm{d}x \\
&\le Nre^{\delta r^2}(2\pi)^{\frac{-T}{2}}\int_1^\infty \frac{1}{x^\delta}\left[\frac{1}{x}\right]^{T-\delta} \mathrm{d}x \\
&= Nre^{\delta r^2}(2\pi)^{\frac{-T}{2}}\left(\frac{1}{T-1}\right)
\end{aligned}$$

We began by omitting the second term in $\theta_\gamma(r)$ for simplicity in the initial analysis. Next, we applied the inequality introduced earlier for the Q function. The assumption $\forall x \ge s : x^{1-\delta}f''(x) \le N$ enabled the derivation of the subsequent term. By performing two substitutions during integration ($t = \log\gamma$ and $x = \frac{t}{r} - \frac{r}{2}$), we further simplified the expression. Finally, we upper-bounded all terms of the form e^{-x^2} by their maximum value of one, i.e., $\forall x : e^{-x^2} \le 1$. This sequence of steps leads to the final result.

The first steps for C are similar to those for B. We have:

$$\begin{aligned}
C &= \int_s^\infty \gamma^{-3}f''(\gamma^{-1})\Big[\theta_\gamma\Big(r\Big)\Big]^T \mathrm{d}\gamma \\
&\le \int_s^\infty \gamma^{-3}f''(\gamma^{-1})\Big[Q\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)\Big]^T \mathrm{d}\gamma \\
&\le \int_s^\infty \gamma^{-3}f''(\gamma^{-1})\Big[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\Big]^T \mathrm{d}\gamma \\
&\le L\int_s^\infty \gamma^{-1}\Big[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\Big]^T \mathrm{d}\gamma \\
&= L(2\pi)^{\frac{-T}{2}}\int_{\frac{r^2}{2}+r}^\infty e^t e^{-t}\left[\frac{exp\Big(-\frac{\left(\frac{t}{r} - \frac{r}{2}\right)^2}{2}\Big)}{\frac{t}{r} - \frac{r}{2}}\right]^T \mathrm{d}t \\
&= Lr(2\pi)^{\frac{-T}{2}}\int_1^\infty \left[\frac{e^{-\frac{x^2}{2}}}{x}\right]^T \mathrm{d}x \\
&\le Lr(2\pi)^{\frac{-T}{2}}\int_1^\infty \left[\frac{1}{x}\right]^T \mathrm{d}x \\
&= Lr(2\pi)^{\frac{-T}{2}}\left(\frac{1}{T-1}\right)
\end{aligned}$$

Here, we once again ignored the last term and apply the inequality for the Q function ($\forall x > 0 : Q(x) < \frac{p(x)}{x}$). Next, we used the assumption $\forall x \ge s : x^{-2}f''(x^{-1}) \le L$. By performing two integrations by substitution ($t = \log\gamma$ and $x = \frac{t}{r} - \frac{r}{2}$) and subsequently upper-bounding e^{-2rx} by one for $x > 1$, we derived the upper bound for the term C.

The final step is to combine the upper bounds for A, B, and C. This gives us:

$$\begin{aligned}
D_f(\mu_T \parallel \pi^\eta) \le &\frac{r\big(L + Ne^{\delta r^2}\big)}{T-1}(2\pi)^{\frac{-T}{2}} \\
&+ \left[f'(s) - \frac{f'(s^{-1})}{s} + f(s^{-1})\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^T
\end{aligned}$$

Appendix C
Proof of Theorem 3

We start by modifying Proposition 1 for the convex case:

Proposition 3. *The diameter of the $\mathcal{S}_B$-constrained Gaussian kernel, where $\mathcal{S}_B := \psi_B(\mathcal{K})$, is less than D, where $D = dia(\mathcal{K})$ for $\eta \le \frac{2}{M}$, assuming the potentials are both M-smooth and convex.*

Proof: First, we begin by showing that when $\eta \le \frac{2}{M}$, the convexity and smoothness assumptions on the potential imply the 1-Lipschitzness of $\psi_B(w)$:

$$\begin{aligned}
&\Big\|\psi_B(w_2) - \psi_B(w_1)\Big\|^2 \\
&= \Big\|w_2 - \frac{1}{b}\sum_{i\in B}\eta\nabla u_i(w_2) - w_1 + \frac{1}{b}\sum_{i\in B}\eta\nabla u_i(w_1)\Big\|^2 \\
&= \Big\|w_2 - w_1\Big\|^2 + \frac{\eta^2}{b^2}\Big\|\sum_{i\in B}\big(\nabla u_i(w_2) - \nabla u_i(w_1)\big)\Big\|^2
\end{aligned}$$

$$-\frac{2\eta}{b}\Big\langle\sum_{i\in B}\big(\nabla u_i(w_2)-\nabla u_i(w_1)\big)\ ,\ w_2-w_1\Big\rangle$$

All the u_i functions are convex and M-smooth, so the function $u_B(w)=\sum_{i\in B}u_i(w)$ is also convex and bM-smooth, where $b=|B|$. Moreover, using the fact that in the convex function u_B, bM-smoothness is equivalent to co-coercivity of ∇u_B, we can write:

$$\begin{aligned}&\Big\|\psi_B(w_2)-\psi_B(w_1)\Big\|^2\\ &\leq\Big\|w_2-w_1\Big\|^2+\frac{\eta^2}{b^2}\Big\|\sum_{i\in B}\big(\nabla u_i(w_2)-\nabla u_i(w_1)\big)\Big\|^2\\ &\quad-\frac{2\eta}{b^2M}\Big\|\sum_{i\in B}\big(\nabla u_i(w_2)-\nabla u_i(w_1)\big)\Big\|^2\\ &\leq\Big\|w_2-w_1\Big\|^2+\frac{\eta}{b^2}(\eta-\frac{2}{M})\Big\|\sum_{i\in B}\nabla u_i(w_2)-\nabla u_i(w_1)\Big\|^2\end{aligned}$$

Now if $\eta\leq\frac{2}{M}$ holds, we have:

$$\Big\|\psi_B(w_2)-\psi_B(w_1)\Big\|^2\leq\Big\|w_2-w_1\Big\|^2$$

And therefore, the $\psi_B(w)$ function is 1-Lipschitz. As a result:

$$dia(\mathcal{S}_B)=\sup_{w_1,w_2\in\mathcal{K}}||\psi_B(w_2)-\psi_B(w_1)||=D$$

■

Having Proposition 3, we revise the upper bound for TV distance and mixing time for P-LMC. A straightforward manipulation of (14) leads to following bound for TV distance:

$$TV(\mu_T,\pi^\eta)\leq\left[1-2Q\Big(\frac{D}{2\sqrt{2\eta}}\Big)\right]^T.$$

This yields the following upper bound for mixing time:

$$T_{mix,TV}\Big(\frac{1}{4}\Big)\leq\frac{-2}{\log\Big[1-2Q\big(\frac{D}{2\sqrt{2\eta}}\big)\Big]}.$$

APPENDIX D
OVERVIEW TABLE OF CONVERGENCE RESULTS

TABLE II
OVERVIEW OF PAPERS PRESENTING CONVERGENCE RESULTS FOR LANGEVIN DYNAMICS AND RELATED ALGORITHMS.

Ref	Algo	Convex	Other Assumptions	Metric	Type
[36]	LD	No	PI	χ^2	to target
[36]	LD	No	LSI	KL	to target
[2]	LD	No	LSI	Rényi	to target
[6]	LD	No	Latała–Oleszkiewicz inequality	Rényi	to target
[6]	LD	No	Modified LSI	Rényi	to target
[31]	LD	No	Weak PI, s-Hölder	Rényi	to target
[14]	LMC	Strong	M-smooth	TV	to target
[37]	LMC	Strong	M-smooth	W_2	to target
[38]	LMC	Strong	M-smooth	W_2	to target
[4]	LMC	Strong	M-smooth	KL	to target
[39]	LMC	Strong outside a ball	M-smooth	W_1	to target
[40]	LMC	Strong outside a ball	M-smooth	TV	to target
[9]	LMC	Strong outside a ball	M-smooth	W_1	to biased
[7]	LMC	Yes	M-smooth	KL	to target
[41]	LMC	Yes	M-smooth	W_q	to target
[25]	LMC	No	LSI, M-smooth, dissipative	W_2	to target
[26]	LMC	No	M-smooth, dissipative	W_1	to target
[2]	LMC	No	LSI, M-smooth	KL	to target
[2]	LMC	No	LSI, M-smooth	Rényi	to biased
[2]	LMC	No	PI, M-smooth	Rényi	to biased
[42]	LMC	No	LSI, α-mix weakly smooth	KL	to target
[28]	LMC	No	LSI, M-smooth, dissipative	KL	to target
[28]	LMC	No	LSI, M-smooth, dissipative	Rényi	to target
[30]	LMC	No	Modified LSI, s-Hölder, dissipative	KL	to target
[6]	LMC	No	Latała–Oleszkiewicz inequality, s-Hölder	Rényi	to target
[6]	LMC	No	Modified LSI, s-Hölder	Rényi	to target
[31]	LMC	No	Weak PI, s-Hölder	Rényi	to target
[29]	LMC	No	M-smooth, f-Sobolev Inequality	f-divergence	to biased
[27]	Average-LMC	No	M-smooth	Fisher information	to target
[32]	P-LMC	No	M-smooth, Uniform sub-Gaussian gradients	W_1	to target
[24]	P-LMC	Yes	M-smooth, Lipschitz	TV	to target
[15]	P-LMC	Yes	M-smooth	TV	to biased
Ours	P-LMC	No	M-smooth	f-divergences	to biased