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

Alireza Daeijavad and Shahab Asoodeh
Department of Computing and Software
McMaster University
Hamilton, Canada
Email: {daeijava, 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 of P-LMC for general smooth potentials, both convex and non-convex. We derive exponential convergence rates with respect to multiple distance metrics, including total variation distance, Hellinger divergence, 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 (LD).

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

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]:

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

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)

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

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

Since the impact of the discretization bias is rather well-understood in the literature [2, 6–12], the 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 $u(x) = \sum_{i=1}^n u_i(x)$ which are supported on 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 $b \leq n$, step size $\eta > 0$, 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 $\mathcal{K}$, $\psi_{B_k}(x) := x - \frac{1}{b}\sum_{i\in B_k} \eta \nabla u_i(x)$, 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 the potentials are convex or non-convex, and (2) the results hold for a

wide range of f-divergences, including KL divergence, Rényi divergence, TV distance, and 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 that are 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 the works most closely related to P-LMC and defer a more comprehensive review of the literature to the extended version [24].

The study of the mixing time of P-LMC has seen significant progress since the seminal works of Bubeck et al. [25]. 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 the assumptions of convexity and smoothness. 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 the Wasserstein distances [26, 27], KL divergence [2], and Fisher information [28], χ^2-divergence [29], Rényi divergence [29], and f-divergence [30]. Several other convergence results have been established under assumptions such as Log-Sobolev inequality (LSI) [29], Poincaré inequality [2], Latała–Oleszkiewicz inequality [6], Modified Log-Sobolev inequality [31], and weak Poincaré inequality (WPI) [32].

The closest work to ours is [33] which 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 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 to transition between continuous and discrete time.

Table I summarizes convergence results for various algorithms derived from Langevin dynamics under different assumptions and metrics. The complete table can be found in Appendix D in the longer version [24].

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
[32]	LD	No	WPI, s-Hölder	Rényi	to target
[26]	LMC	No	LSI, M-smooth, dissipative	W_2	to target
[32]	LMC	No	WPI, s-Hölder	Rényi	to target
[30]	LMC	No	M-smooth, f-Sobolev Inequality	f-divergence	to biased
[33]	P-LMC	No	M-smooth, uniform sub-Gaussian gradients	W_1	to target
[25]	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 calligraphic letters to denote sets, except for $\mathcal{N}$, which represents the Gaussian distribution. We define $\mathcal{S}$ to denote $\{x \in \mathbb{R}^d : \|x\|^2 \le dA, A > 0\}$ and $[n]$ to denote $\{1, 2, \dots, n\}$. The set of all distributions on $\mathcal{W}$ is denoted by $\mathcal{P}(\mathcal{W})$. Given $\gamma \ge 1$, the E_γ-divergence between two distributions μ and ν on $\mathcal{X}$ is defined as $\mathsf{E}_\gamma(\mu\|\nu) \coloneqq \sup_{A\in\mathcal{X}}[\mu(A) - \gamma\nu(A)]$. 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 $\{\mathsf{K}(x) \in \mathcal{P}(\mathcal{W}) : x \in \mathcal{K}\}$. 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 $D_f(\mu\|\nu) \coloneqq \int \mathrm{d}\nu f(\mathrm{d}\mu/\mathrm{d}\nu)$. In the sequel, we use KL divergence $\mathsf{KL}(\mu\|\nu)$, χ^2-divergence $\chi^2(\mu\|\nu)$, total variation distance $\mathsf{TV}(\mu, \nu)$, and $\mathcal{H}_\alpha(\mu\|\nu)$ Hellinger divergence of order $\alpha > 1$, that are instances of f-divergence with $f(t) = t\log t, f(t) = (t-1)^2, f(t) = \frac{1}{2}|t-1|$, and $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$, respectively. All f-divergences satisfy the data processing inequality (DPI): $D_f(\mathsf{K}\mu\|\mathsf{K}\nu) \le D_f(\mu\|\nu)$, for any Markov kernel K. This inequality can be improved for some kernels K, that is there may exist $\eta_f \le 1$ such that $D_f(\mathsf{K}\mu\|\mathsf{K}\nu) \le \eta_f D_f(\mu\|\nu)$ for any measures μ and ν. The smallest such η_f is typically referred to as the *contraction coefficient* of K under f-divergence and denoted by $\eta_f(\mathsf{K})$. If $\eta_f(\mathsf{K}) < 1$, we say K satisfies *strong* DPI (SDPI) for f-divergence. For comprehensive exposition of contraction coefficients of Markov kernels, we refer interested readers to [22, 23, 34, 35].

All proofs and more comprehensive literature review are included in the longer version [24].

II. E_γ-MIXING TIME

In this section, we aim to establish an exponential convergence rate for P-LMC under 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.

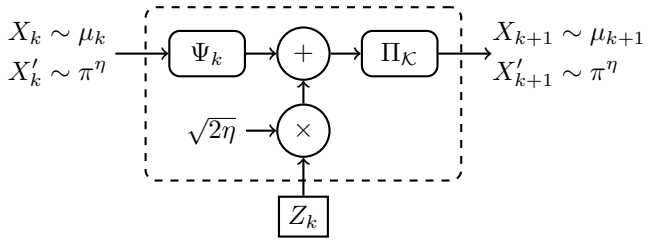

Fig. 1. Visualization of one P-LMC iteration, consisting of three Markov kernels that together satisfy SDPI, ensuring exponential convergence.

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 $\Psi_k \coloneqq \sum_{B\subset[n]} \mathbb{P}(B_k = B)\psi_B$. Here, we use a slight abuse of notation, treating a deterministic function as a Markov kernel.
- $\mathsf{K}_G^{\sqrt{2\eta}} : \mathcal{S} \to \mathcal{P}(\mathbb{R}^d)$ is a $\mathcal{S}$-constrained Gaussian kernel [22], defined as $\mathsf{K}_G^{\sqrt{2\eta}}(y) = \mathcal{N}(y, 2\eta\mathbb{I}_d)$ for all $y \in \mathcal{S}$.
- $\Pi_{\mathcal{K}}(\cdot)$ 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_k with size $|B_k| = b$, we have: $\Psi_k = \sum_{B\subset[n]} \frac{1}{\binom{n}{b}}\psi_B$. Thus, the update rule of P-LMC can be written as:

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

Such representation of each iteration of P-LMC in terms of several Markov kernels enables us to precisely measure the impact of k^{th} iteration on the distance between X_k's distribution and π^η. To formalize this statement, let μ_k denote the distribution of X_k. Note that

$$\mathsf{E}_\gamma(\mu_{k+1}\|\pi^\eta) = \mathsf{E}_\gamma(\mathsf{K}_k\mu_k\|\mathsf{K}_k\pi^\eta).$$

Thus, applying SDPI for E_γ-divergence, we obtain

$$\mathsf{E}_\gamma(\mu_{k+1}\|\pi^\eta) \le \eta_\gamma(\mathsf{K}_k)\mathsf{E}_\gamma(\mu_k\|\pi^\eta), \tag{7}$$

where $\eta_\gamma(\mathsf{K}_k)$ is the contraction coefficient of K_k under E_γ-divergence. The following proposition shows that $\mathcal{S}$-constrained Gaussian kernel, satisfies SDPI.

Proposition 1 ([22, Proposition 1])**.** *Let $\mathcal{S} \subset \mathbb{R}^d$ be a compact set with diameter $\mathsf{dia}(\mathcal{S})$. If K_G^σ is the $\mathcal{S}$-constrained Gaussian kernel, then*

$$\eta_\gamma(\mathsf{K}_G^\sigma) = \theta_\gamma\Big(\frac{\mathsf{dia}(\mathcal{S})}{\sigma}\Big),$$

where

$$\theta_\gamma(r) \coloneqq \mathrm{Q}\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big) - \gamma\mathrm{Q}\Big(\frac{\log\gamma}{r} + \frac{r}{2}\Big), \tag{8}$$

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

We are now in order to state our main result of this section.

Theorem 1. *Let $X_k \sim \mu_k$ denote the output of the k^{th} iteration of P-LMC. Then, we have*

$$\max\{\mathsf{E}_\gamma(\mu_k\|\pi^\eta), \mathsf{E}_\gamma(\pi^\eta\|\mu_k)\} \le \left[\theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\right]^k.$$

Proof sketch: We provide an overview of the proof only for $\mathsf{E}_\gamma(\mu_k\|\pi^\eta)$. The proof for $\mathsf{E}_\gamma(\pi^\eta\|\mu_k)$ is similar.

Consider two initializations for P-LMC: $X_0 \sim \mu_0$, with μ_0 being an arbitrary distribution supported on $\mathcal{K}$, and $X'_0 \sim \pi^\eta$. Let X_k and X'_k be the corresponding outputs of P-LMC after k iterations. Since π^η is the stationary distribution, we have $X'_k \sim \pi^\eta$. Thus, in light of the representation (6) and the convexity of $(\mu,\nu) \mapsto \mathsf{E}_\gamma(\mu\|\nu)$, we have

$$\mathsf{E}_\gamma(\mu_{k+1}\|\pi^\eta) \le \beta \sum_{\substack{B\subset[n]:\\|B|=b}} \mathsf{E}_\gamma((\mathsf{K}_G^{\sqrt{2\eta}}\circ\psi_B)\mu_k\big\|(\mathsf{K}_G^{\sqrt{2\eta}}\circ\psi_B)\pi^\eta),$$

where $\beta \coloneqq \frac{1}{\binom{n}{b}}$.

It can be verified that $\mathsf{K}_G^{\sqrt{2\eta}} \circ \psi_B$ is an $\mathcal{S}_B$-constrained Gaussian kernel with $\mathcal{S}_B \coloneqq \psi_B(\mathcal{K})$. It is straightforward to show that

$$\mathsf{dia}(\mathcal{S}_B) \le D(\eta M+1), \tag{9}$$

(see Proposition 2 in the longer version [24] for a proof.) Thus, invoking Proposition 1 and monotonicity of $r \mapsto \theta_\gamma(r)$ for $\gamma \ge 1$, we arrive at

$$\mathsf{E}_\gamma(\mu_{k+1}\|\pi^\eta) \le \theta_\gamma\Big(\frac{D(\eta M+1)}{\sigma}\Big)\mathsf{E}_\gamma(\mu_k\|\pi^\eta).$$

Applying this argument for k times yields the result. ∎

This theorem establishes that the E_γ-divergence between μ_k and π^η decays exponentially, even for non-convex potentials, provided they satisfy a smoothness condition. Notably, this smoothness assumption is considerably less restrictive than those required by existing results on the convergence analysis of P-LMC in the non-convex setting (see Table I for a summary of such assumptions).

This theorem can naturally be translated into an argument about mixing time. Borrowing the recently defined notion of E_γ-*mixing-time* [36], we define

$$T_{mix,\mathsf{E}_\gamma}(\varepsilon) = \min\{k \in \mathbb{N} : \mathsf{E}_\gamma(\mu_k\|\pi^\eta) \le \varepsilon\}, \tag{10}$$

for $\gamma \ge 1$ and $\varepsilon > 0$. The following is an immediate corollary of Theorem 1 that gives an upper bound for E_γ-mixing-time.

Corollary 1. *For any $\gamma \ge 1$ and $0 < \varepsilon < 1$, the E_γ-mixing-time of P-LMC satisfies*

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

where θ_γ was defined in (8)*.*

It is worth making a remark on the E_γ-mixing-time. Zamanlooy et al. [36] established that $T_{mix,\mathsf{E}_\gamma}(0) < \infty$ for finite-state Markov chains, deviating from traditional notions of mixing

time, i.e., those based on TV distance, KL divergence, Rényi divergence, or χ^2-divergence. However, Corollary 1, which provides the first analysis of the E_γ-mixing-time for Markov chains with a continuous state space, does not exhibit this property. Whether their result extends to the continuous-state setting remains an open question.

We conclude this section with another immediate corollary of Theorem 1, which establishes the convergence rate and mixing time in terms of TV distance. Note that the TV-mixing time is defined analogously to (10), with the E_γ-divergence replaced by the TV distance (i.e., setting $\gamma = 1$).

Corollary 2. *Let $X_k \sim \mu_k$ denote the output of the k^{th} iteration of P-LMC. Then, we have*

$$\mathsf{TV}(\mu_k, \pi^\eta) \leq \left[1 - 2Q\Big(\frac{D(\eta M + 1)}{2\sqrt{2\eta}}\Big)\right]^k. \tag{11}$$

Moreover, for $0 < \varepsilon < 1$, the TV-mixing time satisfies

$$T_{mix,\mathsf{TV}}(\varepsilon) \leq \frac{\log \varepsilon}{\log\Big(1 - 2Q\big(\frac{D(\eta M+1)}{2\sqrt{2\eta}}\big)\Big)}. \tag{12}$$

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

In the previous section, we studied a convergence analysis for P-LMC, in terms of how fast the E_γ-divergence between μ_k and π^η decays and provided the corresponding mixing time.

It is well established that a broad class of f-divergences can be expressed in terms of E_γ-divergence. Specifically, for twice-differentiable f, we have [37, Corollary 3.7]

$$D_f(\mu\|\nu) = \int_1^\infty \big[f''(\gamma)\mathsf{E}_\gamma(\mu\|\nu) + \frac{1}{\gamma^3} f''(1/\gamma)\mathsf{E}_\gamma(\nu\|\mu)\big]\mathrm{d}\gamma. \tag{13}$$

By combining Theorem 1 with this elegant representation, we directly obtain the convergence rate of P-LMC for an extensive family of divergences. This universality significantly enhances the applicability of our results, making them a powerful tool for analyzing a wide range of divergences.

Theorem 2. *Let $f : (0, \infty) \to \mathbb{R}$ be a twice-differentiable convex function with continuous second derivative f'' and $f(1) = 0$. Take $r = \frac{D(\eta M+1)}{\sqrt{2\eta}}$ and $s = e^{\frac{r^2}{2}+r}$. Let $X_k \sim \mu_k$ denote the output of the k^{th} iteration of P-LMC. If there exist constants L and N such that for all $t \geq s$:*

$$t^{-2} f''(t^{-1}) \leq L, \tag{14}$$

and

$$t^{1-K} f''(t) \leq N, \quad \text{for some } K \in \mathbb{N}, \tag{15}$$

then, for any $k \geq K$, we have

$$D_f(\mu_k\|\pi^\eta) \leq \frac{r\big(L + Ne^{Kr^2}\big)}{k-1}(2\pi)^{\frac{-k}{2}} + \left[f'(s) - \frac{f'(s^{-1})}{s} + f(s^{-1})\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^k.$$

Proof sketch: Note that the integral representation of f-divergence in (13) yields

$$D_f(\mu_k\|\pi^\eta) = \int_1^\infty \big[f''(\gamma)\mathsf{E}_\gamma(\mu_k\|\pi^\eta) + \frac{1}{\gamma^3} f''(1/\gamma)\mathsf{E}_\gamma(\pi^\eta\|\mu_k)\big]\mathrm{d}\gamma.$$

Applying Theorem 1, we can write

$$D_f(\mu_k\|\pi^\eta) \leq \underbrace{\int_1^s \big[f''(\gamma) + \gamma^{-3} f''(\gamma^{-1})\big]\Big[\theta_\gamma(r)\Big]^k \mathrm{d}\gamma}_{A} + \underbrace{\int_s^\infty f''(\gamma)\Big[\theta_\gamma(r)\Big]^k \mathrm{d}\gamma}_{B} + \underbrace{\int_s^\infty \gamma^{-3} f''(\gamma^{-1})\Big[\theta_\gamma(r)\Big]^k \mathrm{d}\gamma}_{C}$$

We now provide upper bounds for each of these integrals.

We begin by A. First, it holds that $\theta_\gamma(r) \leq Q\left(\frac{\log\gamma}{r} - \frac{r}{2}\right)$. The monotonicity of $\gamma \mapsto Q\big(\frac{\log\gamma}{r} - \frac{r}{2}\big)$ for all $\gamma \geq 1$ implies that

$$A \leq \left[Q\left(\frac{-r}{2}\right)\right]^k \int_1^s \big[f''(\gamma) + \gamma^{-3} f''(\gamma^{-1})\big]\mathrm{d}\gamma. \tag{16}$$

An application of integration by parts yields

$$A \leq \left[Q\left(\frac{-r}{2}\right)\right]^k \big[f'(s) - s^{-1} f'(s^{-1}) + f(s^{-1})\big].$$

Next, we bound B and C. Notice that $Q(x) < \frac{p(x)}{x}$ for $x > 0$, where $p(x)$ denotes the probability density function of the normal distribution. By leveraging this inequality and applying the trivial bound mentioned for $\theta_\gamma(r)$, we obtain

$$B \leq \int_s^\infty \gamma^{K-1}\gamma^{1-K} f''(\gamma)\left[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\right]^k \mathrm{d}\gamma,$$

and

$$C \leq \int_s^\infty \gamma^{-3} f''(\gamma^{-1})\left[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\right]^k \mathrm{d}\gamma.$$

Following (14) and (15), we have

$$B \leq N\int_s^\infty \gamma^{K-1}\left[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\right]^k \mathrm{d}\gamma,$$

and

$$C \leq L\int_s^\infty \gamma^{-1}\left[\frac{p\Big(\frac{\log\gamma}{r} - \frac{r}{2}\Big)}{\frac{\log\gamma}{r} - \frac{r}{2}}\right]^k \mathrm{d}\gamma,$$

from which, and a simple application of algebraic simplification, we arrive at

$$B \leq Nre^{Kr^2}(2\pi)^{\frac{-k}{2}}\left(\frac{1}{k-1}\right), \tag{17}$$

and

$$C \leq Lr(2\pi)^{\frac{-k}{2}}\left(\frac{1}{k-1}\right). \tag{18}$$

Combining (16), (17), and (18) gives the desired result. ∎

This theorem establishes an exponential convergence rate for P-LMC across a broad range of instances of f-divergences, provided that f'' satisfies the growth conditions outlined in (14) and (15). Notably, these assumptions hold for many commonly used f-divergences. For example, it can be verified with $N = L = 1$ for KL divergence, $N = L = 2$ for χ^2-divergence, and with $N = L = \alpha$ for the Hellinger divergence of order α. The following corollary, derived from Theorem 2, explicitly identifies the exponential convergence rate for these specific metrics. Furthermore, the one-to-one relationship between Rényi divergence and Hellinger divergence allows us to directly extend the convergence rate derived for Hellinger divergence to Rényi divergence.

Corollary 3. *Let $X_k \sim \mu_k$ denote the output of the k^{th} iteration of P-LMC. Let $r = \frac{D(\eta M+1)}{\sqrt{2\eta}}$ and $s = e^{\frac{r^2}{2}+r}$. We have the following upper bounds:*

- *KL divergence: For $k \geq 2$, we have*

$$\mathsf{KL}(\mu_k\|\pi^\eta) \leq \left[\frac{r^2}{2}+r+1-\frac{1}{s}\right]\left[Q\big(\frac{-r}{2}\big)\right]^k + \frac{r\big(1+e^{r^2}\big)}{k-1}\big(\frac{1}{2\pi}\big)^{\frac{k}{2}},$$

- *χ^2-divergence: For $k \geq 2$, we have*

$$\chi^2(\mu_k\|\pi^\eta) \leq \left[2s-1-\frac{1}{s^2}\right]\left[Q\big(\frac{-r}{2}\big)\right]^k + \frac{2r\big(1+e^{r^2}\big)}{k-1}\big(\frac{1}{2\pi}\big)^{\frac{k}{2}},$$

- *Hellinger divergence: For $k \geq \lceil\alpha\rceil$ and $\alpha \in (1,\infty)$, we have*

$$\mathcal{H}_\alpha(\mu_k\|\pi^\eta) \leq \left[\frac{\alpha s^{\alpha-1}-1}{\alpha-1}-\frac{1}{s^\alpha}\right]\left[Q\big(\frac{-r}{2}\big)\right]^k + \frac{\big(1+e^{\lceil\alpha-1\rceil r^2}\big)}{(k-1)\big(\alpha r\big)^{-1}}\big(\frac{1}{2\pi}\big)^{\frac{k}{2}},$$

- *Rényi divergence: For $k \geq \lceil\alpha\rceil$ and $\alpha \in (1,\infty)$, we have*

$$D_\alpha(\mu_k\|\pi^\eta) \leq \frac{1}{\alpha-1}\log\left[\frac{\alpha r\big(1+e^{\lceil\alpha-1\rceil r^2}\big)}{(k-1)(\alpha-1)^{-1}}\big(\frac{1}{2\pi}\big)^{\frac{k}{2}} + \left[\alpha s^{\alpha-1}-1-\frac{\alpha-1}{s^\alpha}\right]\left[Q\big(\frac{-r}{2}\big)\right]^k+1\right].$$

IV. Mixing time for convex functions

In this section, we demonstrate that the convergence rate established in Theorem 1 improves when the potential is convex. This improvement follows from a well-known result in convex optimization (see, e.g., [38]): If g is M-smooth and convex, then $w \mapsto w-\eta\nabla g(w)$ is contractive for $\eta \leq 2/M$. In our context, this implies that the update rule ψ_B is contractive for smooth and convex potentials (refer to the longer version for a formal proof). This contractivity further ensures that $\mathsf{dia}(\mathcal{S}_B) \leq D$, where $\mathcal{S}_B \coloneqq \psi_B(\mathcal{K})$, which provides a tighter bound compared to (9).

The following theorem establishes an improved convergence rate specifically tailored for convex potentials. For ease of comparison with [15], we present this result in terms of total variation distance.

Theorem 3. *Consider the potentials are convex and $\eta \leq \frac{2}{M}$. Let $X_k \sim \mu_k$ denote the output of the k^{th} iteration of P-LMC. Then, we have*

$$\mathsf{TV}(\mu_k,\pi^\eta) \leq \left[1-2Q\left(\frac{D}{2\sqrt{2\eta}}\right)\right]^k. \tag{19}$$

Moreover, for $0<\varepsilon<1$, the TV-mixing time satisfies

$$T_{mix,\mathsf{TV}}(\varepsilon) \leq \frac{\log\varepsilon}{\log\left[1-2Q\big(\frac{D}{2\sqrt{2\eta}}\big)\right]}. \tag{20}$$

Proof sketch: It can be verified that $\mathsf{dia}(\mathcal{S}_B) \leq D$. Following the steps as those in the proof of Theorem 2, with the new estimate for the diameter of $\mathcal{S}_B$, we obtain

$$\mathsf{TV}(\mu_k,\pi^\eta) \leq \left[Q\left(\frac{-D}{2\sqrt{2\eta}}\right)-Q\left(\frac{D}{2\sqrt{2\eta}}\right)\right]^k$$

Since $Q(x) = 1-Q(-x)$, we derive the final result. ∎

It is worth noting that a tight convergence bound for the mixing time in the convex setting, recently established by Altschuler and Talwar [15], is of order $\frac{D^2}{\eta}$. While our result does not achieve this bound, it represents an improvement over the existing convergence analyses in non-convex settings.

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.

Our main results are based on the observation that the update rule of P-LMC can be modeled as a composition of three Markov kernels. For two different initializations of P-LMC, 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 f-divergence between the output distribution of P-LMC and π^η decays exponentially.

This observation may hold for other sampling algorithms. Thus, as a natural future direction, we aim to extend our result to other such sampling algorithms, e.g., the Metropolis-Adjusted Langevin Algorithm (MALA) [39, 40].

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] A. Daeijavad and S. Asoodeh, "Supplementary material for exponential convergence of projected langevin monte carlo with non-convex potentials," https://daeijavad.github.io/docs/SupplementaryMaterial.pdf, 2025.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] A. Makur and L. Zheng, "Comparison of contraction coefficients for f-divergences," *Probl. Inf. Transm.*, vol. 56, no. 2, p. 103–156, Apr. 2020.

[35] M. Raginsky, "Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[36] B. Zamanlooy, S. Asoodeh, M. Díaz, and F. du Pin Calmon, "E_γ-mixing time," *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 3474–3479, 2024.

[37] J. Cohen, J. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices, with Applications in Information Theory, Statistics, Economics, and Population Sciences*. Birkhäuser, 1998.

[38] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. Springer Publishing Company, Incorporated, 2014.

[39] 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.

[40] 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.

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

[42] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, pp. 341–363, 1996.

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

[44] 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.

[45] 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.

[46] 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.

[47] M. B. Majka, A. Mijatović, and Ł. Szpruch, "Nonasymptotic bounds for sampling algorithms without log-concavity," *The Annals of Applied Probability*, vol. 30, no. 4, pp. 1534–1581, 2020.

[48] 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.

[49] Y. Zheng and A. Lamperski, "Constrained langevin algorithms with l-mixing external random variables," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, 2022, pp. 20 511–20 521.

[50] J. Liang, S. Mitra, and A. Wibisono, "On independent samples along the langevin diffusion and the unadjusted langevin algorithm," *arXiv preprint arXiv:2402.17067*, 2024.

[51] 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.

[52] X. Cheng, B. Wang, J. Zhang, and Y. Zhu, "Fast conditional mixing of mcmc algorithms for non-log-concave distributions," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

APPENDIX A
PROOF OF THEOREM 1

We begin by stating the following proposition which will be required subsequently.

Proposition 2. *Let $\mathcal{S}_B \coloneqq \psi_B(\mathcal{K})$ for M-smooth potential functions. Then, we have*

$$\mathsf{dia}(\mathcal{S}_B) \leq D(\eta M + 1),$$

where $D = \mathsf{dia}(\mathcal{K})$.

Proof of Proposition 2: We have

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

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

Using Proposition 2, we compute $\mathsf{E}_\gamma(\mu_{k+1}\|\pi^\eta)$ after $k+1$ iterations, where the initial inputs are sampled from π^η and μ_0:

$$\begin{aligned}
&\mathsf{E}_\gamma(\mu_{k+1}\|\pi^\eta) \\
&= \mathsf{E}_\gamma\Big((\Pi_\mathcal{K}\circ \mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_k)\mu_k \| (\Pi_\mathcal{K}\circ \mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_k)\pi^\eta\Big) \\
&\leq \mathsf{E}_\gamma\Big((\mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_k)\mu_k \| (\mathsf{K}_G^{\sqrt{2\eta}}\circ\Psi_k)\pi^\eta\Big) \\
&\leq \frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\ |B|=b}} \mathsf{E}_\gamma\Big((\mathsf{K}_G^{\sqrt{2\eta}}\circ\psi_B)\mu_k \| (\mathsf{K}_G^{\sqrt{2\eta}}\circ\psi_B)\pi^\eta\Big) \\
&\leq \frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\ |B|=b}} \theta_\gamma\Big(\frac{\mathsf{dia}(\mathcal{S}_B)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\psi_B(\mu_k)\|\psi_B(\pi^\eta)\Big) \\
&\leq \frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\ |B|=b}} \theta_\gamma\Big(\frac{\mathsf{dia}(\mathcal{S}_B)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_k\|\pi^\eta\Big) \qquad (21) \\
&\leq \frac{1}{\binom{n}{b}}\sum_{\substack{B\subset[n]:\\ |B|=b}} \theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_k\|\pi^\eta\Big) \\
&= \theta_\gamma\Big(\frac{D(\eta M+1)}{\sqrt{2\eta}}\Big)\mathsf{E}_\gamma\Big(\mu_k\|\pi^\eta\Big)
\end{aligned}$$

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 $(P,Q)\mapsto \mathsf{E}_\gamma(P\|Q)$. The next step leverages Proposition 1. Again, we apply DPI and Proposition 2. The last step holds because the terms are identical, so we can multiply them by their count.

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

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

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 (Assuming $\varepsilon < 1$. If $\varepsilon \geq 1$ we get the trivial bound of $T_{mix} \geq 0$). Specifically, we aim to determine k such that $\mathsf{E}_\gamma(\mu_k\|\pi^\eta) \leq \varepsilon$, which holds when

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

Taking the logarithm of both sides, we have

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

As a result

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

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 (13), we obtain:

$$D_f(\mu_k\|\pi^\eta) \leq \int_1^\infty \Big(f''(\gamma) + \gamma^{-3}f''(\gamma^{-1})\Big)\left[\theta_\gamma\Big(r\Big)\right]^k \mathrm{d}\gamma$$

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

$$\begin{aligned}
&= \underbrace{\int_1^s \left[f''(\gamma) + \gamma^{-3}f''(\gamma^{-1})\right]\left[\theta_\gamma\Big(r\Big)\right]^k \mathrm{d}\gamma}_{A} \\
&\quad + \underbrace{\int_s^\infty f''(\gamma)\left[\theta_\gamma\Big(r\Big)\right]^k \mathrm{d}\gamma}_{B} + \underbrace{\int_s^\infty \gamma^{-3}f''(\gamma^{-1})\left[\theta_\gamma\Big(r\Big)\right]^k \mathrm{d}\gamma}_{C}
\end{aligned}$$

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 $1 \leq \gamma \leq s$, it attains its maximum at $\gamma^* = 1$. Consequently, we have:

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

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:

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

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 $\forall x \geq s : x^{1-K} f''(x) \leq N$ 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., $\forall x : e^{-x^2} \leq 1$. This sequence of steps leads to the final result.

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

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

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 $\forall x \geq s : x^{-2} f''(x^{-1}) \leq L$. 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:

$$\begin{aligned}
D_f(\mu_k \| \pi^\eta) \leq & \frac{r\big(L + Ne^{Kr^2}\big)}{k-1}(2\pi)^{\frac{-k}{2}} \\
& + \left[f'(s) - \frac{f'(s^{-1})}{s} + f(s^{-1})\right]\Big[Q\big(\frac{-r}{2}\big)\Big]^k
\end{aligned}$$

APPENDIX C
PROOF OF THEOREM 3

We start by modifying Proposition 2 for the convex case:

Proposition 3. *Let $\mathcal{S}_B \coloneqq \psi_B(\mathcal{K})$ for M-smooth and convex potential functions. Then, for $\eta \leq \frac{2}{M}$ we have*

$$\mathsf{dia}(\mathcal{S}_B) \leq D,$$

where $D = \mathsf{dia}(\mathcal{K})$.

Proof of Proposition 3: First, we begin by showing that when $\eta \leq \frac{2}{M}$, the convexity and smoothness assumptions on the potential imply the 1-Lipschitzness of $\psi_B(w)$:

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

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:

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

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:

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

■

Having Proposition 3, we revise the upper bound for TV distance and mixing time for P-LMC. A straightforward manipulation of (21) leads to following bound for TV distance:

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

This yields the following upper bound for mixing time:

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

APPENDIX D
OVERVIEW TABLE OF CONVERGENCE RESULTS

Theoretical analyses of algorithms related to LD can be categorized into several subsections. In the continuous-time setting, Bakry et al. [41] demonstrated that the Log-Sobolev inequality (LSI) and the Poincaré inequality (PI) imply exponential convergence in KL divergence and χ^2 divergence, respectively. The convergence of LD under other metrics and assumptions has been explored in works such as [2, 6, 32].

In the discrete-time setting, LMC and P-LMC serve as cornerstones of sampling methods due to their simplicity and scalability. The first attempts at providing a theoretical analysis for LMC were made by Roberts and Tweedie [42]. Recent work on LMC focuses on improving theoretical understanding by imposing structural assumptions on the potentials. Convexity, as one of the earliest and most influential assumptions, provides a clear framework for analyzing its convergence.

a) Convex LMC: Under smoothness and strong convexity of the potential, Dalalyan [14] showed that LMC converges to an ε-neighborhood of π in TV distance within $\tilde{O}(d^3\varepsilon^{-2})$ iterations, later improved in [43, 44]. It was broaden to KL divergence in [4]. Having smoothness, strong convexity relaxed to convexity in [7, 45]. The analysis were extended to cases where the potential is non-convex within a ball and strongly convex outside it by [9, 46–49]. On a different approach, Liang et al. [50] showed that the initial and current output distributions become independent exponentially fast when potential is convex and smooth.

b) Non-convex LMC: In the unconstrained non-convex setting, the literature has expanded both the scope of metrics, covering 1-Wasserstein distance [26], 2-Wasserstein distance [27], KL divergence [2], and Fisher information [28], as well as relaxed assumptions on potential functions, including α-mixture weak smoothness [51], LSI [29], PI [2], the Latała–Oleszkiewicz (LO) inequality [6], Modified Log-Sobolev Inequality (M-LSI) [31], and weak Poincaré inequality (WPI) [32], enabling new results under less restrictive assumptions on potentials.

It has been shown in [30] that, under the assumptions of smoothness, the f-Sobolev inequality, and bounded η, exponential convergence in the corresponding f-divergence is achieved. Additionally, Cheng et al. [52] proposed conditional convergence on a state space subset, showing that with smoothness and local LSI, the probability mass is either small or the LMC output is close π.

c) Convex P-LMC: In the constrained setting, there is a relative scarcity of research. For the convex case, Bubeck et al. [25] made a key contribution, demonstrating that under Lipschitz continuity and smoothness assumptions, the iteration complexity of P-LMC to achieve an ε-neighborhood of π in TV distance is $\tilde{O}(d^{12}\varepsilon^{-12})$. Altschuler and Talwar [15] characterize the mixing time of P-LMC to π^η under convexity and smoothness assumptions, showing that the mixing time in TV distance is $\Theta\big(\frac{D^2}{\eta}\big)$. A detailed comparison is provided in Section IV.

d) Non-convex P-LMC: Lamperski [33] analyzed the convergence rate of the 1-Wasserstein distance between output distribution of P-LMC and π. In their framework, the potential function is assumed to take the form $\nu(X_k, Z_k)$, where the first argument (X_k) corresponds to the same variable as in our setting, and the second argument (Z_k) represents I.I.D. external random variables.

The assumptions in [33] are: (i) The mean potential function, defined as $\bar{\nu}(x) = \mathbb{E}_W[\nu(x, z)]$, is M-smooth; (ii) For each z, the gradient $\nabla_x \nu(x, z)$ is ℓ-Lipschitz; (iii) For each $x \in \mathbb{R}^d$, the deviations $\nabla_x \nu(x, Z) - \nabla_x \bar{\nu}(x)$ are uniformly sub-Gaussian. Specifically, there exists $\sigma > 0$ such that for all $\alpha \in \mathbb{R}^d$, the following bound holds:

$$\mathbb{E}\left[\exp\left(\alpha^\intercal\big(\nabla_x \nu(x, Z) - \nabla_x \bar{\nu}(x)\big)\right)\right] \leq e^{\sigma^2\|\alpha\|^2/2}. \quad (22)$$

Under these assumptions, Lamperski [33] proved that for $\eta \leq \frac{1}{2}$ and constants $\{c_i : i \in [3]\}$, this bound holds:

$$W_1\big(\mathcal{L}(X_T), \pi_{\bar{\nu}}\big) \leq\ c_1(\eta \log T)^{\frac{1}{4}} + c_2 e^{-\eta c_3 T} \quad (23)$$

To compare our results with those in [33], we assume that the second argument in [33] is drawn from a uniform distribution, and we set the batch size in our setting to one. Therefore, $u_i(x)$ corresponds to $\nu(x, z_i)$ in [33] and satisfies all the assumptions stated in [33].

When comparing our findings in Theorem 2 with (23), we observe a key distinction in the rate of convergence. Our result establishes exponential convergence in TV distance, whereas [33] demonstrates logarithmic convergence in terms of iterations under the 1-Wasserstein distance. This difference stems from the methodological approaches employed. Specifically, Lamperski [33] achieved 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, thereby resulting in a sharper convergence bound.

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

Ref	Algo	Convex	Other Assumptions	Metric	Type
[41]	LD	No	PI	χ^2	to target
[41]	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
[32]	LD	No	Weak PI, s-Hölder	Rényi	to target
[14]	LMC	Strong	M-smooth	TV	to target
[44]	LMC	Strong	M-smooth	W_2	to target
[43]	LMC	Strong	M-smooth	W_2	to target
[4]	LMC	Strong	M-smooth	KL	to target
[46]	LMC	Strong outside a ball	M-smooth	W_1	to target
[48]	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
[45]	LMC	Yes	M-smooth	W_q	to target
[26]	LMC	No	LSI, M-smooth, dissipative	W_2	to target
[27]	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
[51]	LMC	No	LSI, α-mix weakly smooth	KL	to target
[29]	LMC	No	LSI, M-smooth, dissipative	KL	to target
[29]	LMC	No	LSI, M-smooth, dissipative	Rényi	to target
[31]	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
[32]	LMC	No	Weak PI, s-Hölder	Rényi	to target
[30]	LMC	No	M-smooth, f-Sobolev Inequality	f-divergence	to biased
[28]	Average-LMC	No	M-smooth	Fisher information	to target
[33]	P-LMC	No	M-smooth, Uniform sub-Gaussian gradients	W_1	to target
[25]	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