On the Convergence of Langevin Monte Carlo: The Interplay between Tail Growth and Smoothness

Murat A. Erdogdu ERDOGDU@CS.TORONTO.EDU
Department of Computer Science and Department of Statistical Sciences at the University of Toronto, and Vector Institute.

Rasa Hosseinzadeh RASA@CS.TORONTO.EDU
Department of Computer Science at the University of Toronto, and Vector Institute.

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study sampling from a target distribution $\nu_* = e^{-f}$ using the unadjusted Langevin Monte Carlo (LMC) algorithm. For any potential function f whose tails behave like $\|x\|^\alpha$ for $\alpha \in [1, 2]$, and has β-Hölder continuous gradient, we prove that $\widetilde{\mathcal{O}}\big(d^{\frac{1}{\beta}+\frac{1+\beta}{\beta}(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq 1\}})}\epsilon^{-\frac{1}{\beta}}\big)$ steps are sufficient to reach the ϵ-neighborhood of a d-dimensional target distribution ν_* in KL-divergence. This bound, in terms of ϵ dependency, is not directly influenced by the tail growth rate α of the potential function as long as its growth is at least linear, and it only relies on the order of smoothness β. One notable consequence of this result is that for potentials with Lipschitz gradient, i.e. $\beta = 1$, the above rate recovers the best known rate $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ which was established for strongly convex potentials in terms of ϵ dependency, but we show that the same rate is achievable for a wider class of potentials that are degenerately convex at infinity. The growth rate α affects the rate estimate in high dimensions where d is large; furthermore, it recovers the best-known dimension dependency when the tail growth of the potential is quadratic, i.e. $\alpha = 2$, in the current setup.

Keywords: Unadjusted Langevin Algorithm, Rate of Convergence, Markov Chain Monte Carlo

1. Introduction

Sampling from a target distribution using Markov chain Monte Carlo (MCMC) is a fundamental problem in statistics, and it often amounts to discretizing a diffusion process with invariant measure as the target. When the target corresponds to the Gibbs measure $\nu_* = e^{-f}$ where $f : \mathbb{R}^d \to \mathbb{R}$ is the potential function satisfying $\int e^{-f(x)}dx = 1$, a popular candidate diffusion is the overdamped Langevin diffusion, which is the solution of the following stochastic differential equation (SDE),

$$dZ_t = -\nabla f(Z_t)dt + \sqrt{2}dB_t, \tag{1.1}$$

where B_t is a d-dimensional Brownian motion. Langevin diffusion (1.1) admits the target Gibbs measure ν_* as its invariant distribution (Risken, 1996). In general, simulating a continuous-time diffusion such as (1.1) is impractical; thus, a numerical integration scheme is needed to approximate it. In this work, we focus on the unadjusted Langevin Monte Carlo algorithm (LMC) which is the Euler discretization of the overdamped Langevin diffusion (1.1), defined by the update rule

$$x_{k+1} = x_k - \eta\nabla f(x_k) + \sqrt{2\eta}W_k, \tag{1.2}$$

where $W_k \in \mathbb{R}^d$ is an isotropic Gaussian vector independent from W_m and x_m for any $m < k$, and η is a sufficiently small step size.

Convergence rates of LMC have been established under structural assumptions on the potential function, and they quantify the number of iterations sufficient to reach the ϵ-neighborhood of a d-dimensional target distribution $\nu_* = e^{-f}$ under a particular distance measure – our focus is on KL-divergence. Earlier attempts established convergence rate estimates under the global curvature assumptions on the potential function. For example, for strongly convex and smooth potentials, the convergence rate bound of $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ has been shown (Dalalyan, 2017b) (here $\widetilde{\mathcal{O}}$ hides $\log$ factors as well as other constants), whereby a smooth function is a function with Lipschitz continuous gradient. We note that higher-order smoothness on the potential function may improve the convergence rate estimate (Mou et al., 2019b); however, we consider only the first-order smoothness in the current paper. For convex and smooth potentials with growth rate α, a convergence rate estimate of $\widetilde{\mathcal{O}}(d^{1+4/\alpha}\epsilon^{-3})$ is known to hold for LMC (Cheng and Bartlett, 2018), with the caveat that a surrogate strongly convex potential is used instead of the original potential. More recently, however, it has been observed that tail growth structure is the determinant factor in sampling (Cheng et al., 2018a; Eberle, 2016; Erdogdu et al., 2018; Eberle et al., 2019; Majka et al., 2020), rather than the global curvature structure such as (strong) convexity, where in this context, a strongly convex potential is understood to exhibit quadratic growth. Growth-based structural conditions has the additional benefit of allowing for perturbations, which in turn allows for sampling from non-convex potentials. A condition on the target distribution ν_* that fits in this framework is the *log-Sobolev inequality* (LSI) (Bakry and Émery, 1985), given as

$$\forall \rho, \;\; \mathrm{KL}\big(\rho|\nu_*\big) \leq \lambda \mathrm{I}\big(\rho|\nu_*\big), \tag{1.3}$$

where $\mathrm{KL}\big(\rho|\nu_*\big)$ denotes the KL-divergence (relative entropy) and $\mathrm{I}\big(\rho|\nu_*\big)$ denotes the relative Fisher information between ρ and ν_* (defined in (1.7)), and $\lambda > 0$ is the log-Sobolev constant. The LSI condition (1.3) can be verified for potentials with certain growth structure. Indeed, it is known to hold for strongly convex potentials (Bakry and Émery, 1985), and it allows for finite perturbations due to Holley and Stroock (1987) perturbation lemma; thus, potentials that have quadratic growth satisfy LSI. Denoting the distribution of Langevin diffusion (1.1) at time t with ρ_t, by using Fokker-Planck equation it can be shown that $\frac{d}{dt}\mathrm{KL}\big(\rho_t|\nu_*\big) = -\mathrm{I}\big(\rho_t|\nu_*\big)$ which, combined with the LSI condition (1.3) entails a differential inequality of the form $\frac{d}{dt}\log \mathrm{KL}\big(\rho_t|\nu_*\big) \leq -\frac{1}{\lambda}$, which in turn yields an exponential decay in KL-divergence, i.e., $\mathrm{KL}\big(\rho_t|\nu_*\big) \leq e^{-t/\lambda}\mathrm{KL}\big(\rho_0|\nu_*\big)$. LSI coupled with the smoothness condition on the potential function is sufficient to obtain the fast convergence rate estimate $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ for LMC (Vempala and Wibisono, 2019), which is the best known estimate in this framework. The significance of this result is in that, it relaxes the strong convexity assumption which is a global curvature condition on f to the LSI condition (1.3), which can be regarded as a tail growth condition on f, allowing for finite perturbations.

The fundamental idea in the current paper is that the fast convergence of LMC does not require an exponential convergence of the Langevin diffusion, which is essentially obtained under strong tail growth conditions on the potential. A representative convergence analysis of LMC under some distance measure D (our main focus is KL-divergence) establishes a single step bound,

$$\forall k \in \mathbb{N}, \;\; \mathrm{D}(\rho_{k+1}|\nu_*) \leq r(\eta)\, \mathrm{D}(\rho_k|\nu_*) + C\eta^{\theta}, \tag{1.4}$$

where $r : [0, \infty) \to (0, 1]$ is a monotone decreasing function, inherited from the fast decaying diffusion counterpart. The discretization error $C\eta^{\theta}$ can be made small with smaller step size η, and the exponent θ is intrinsic to the numerical scheme as well as the smoothness of f. Elementary algebra

reveals that, one can iterate the inequality (1.4) and achieve convergence as long as $r(\eta) < 1$ and $\lim_{\eta\to 0} \frac{\eta^\theta}{1-r(\eta)} = 0$. Recent literature focused on exponential decays $r_{\exp}(t) = e^{-ct}$ which are usually established under conditions like LSI (1.3), that correspond to potentials exhibiting quadratic growth (see, for example Dalalyan (2017b); Vempala and Wibisono (2019)). Nevertheless, the inequality (1.4) by no means benefits from the exponential decay, as $r(t)$ is only evaluated at short time horizons $t = \eta$. For example, consider the algebraic rate $r_{\text{alg}}(t) = 1/(1 + ct)$ which is much slower than the exponential rate, but it provides the same level of decay in small time horizons, i.e. evaluated at the step size η, one has

$$r_{\text{alg}}(\eta) \approx r_{\exp}(\eta) \approx 1 - c\eta \;\text{ when } \eta \text{ is small.} \tag{1.5}$$

The conditions required for exponential decay depend on the metric, for example LSI coupled with KL, and Poincaré inequality with Chi-square divergence imply an exponential decay. However, for a given metric, algebraic rates can be obtained under weaker conditions on the potential function f.

Modified versions of the LSI condition (1.3) or weak Poincaré inequalities are commonly employed in the analysis of diffusion processes (Gentil et al., 2005; Bakry et al., 2013), and can be used to explain their convergence behavior. For example in Toscani and Villani (2000), a modified log-Sobolev inequality is used to establish a convergence rate of $\mathcal{O}(t^{-\kappa})$ for all $\kappa > 0$ for the Langevin diffusion (1.1) ($\mathcal{O}(t^{-\infty})$ in their notation). Our results build on a similar construction. For a class of potentials that are convex degenerate at infinity, with tails growing like $\|x\|^\alpha$ for $\alpha \in [1, 2]$, we establish the following modified log-Sobolev inequality (mLSI)

$$\forall \rho,\;\; \mathrm{KL}\big(\rho|\nu_*\big) \le \lambda \mathrm{I}\big(\rho|\nu_*\big)^{1-\delta} \mathrm{M}_s(\rho + \nu_*)^\delta \quad \text{with} \quad \delta \in [0, 1/2], \tag{1.6}$$

where $\mathrm{M}_s(h) = \int (1 + \|x\|^2)^{s/2} h(x)dx$ is the s-th moment of any function h. This inequality entails a decay with the desired properties (1.5) under mild conditions on the potential. Our focus is on the algorithmic implications of (1.6), and our contributions can be summarized as follows.

- For a potential f whose tails behave like $\|x\|^\alpha$ for some $\alpha \in [1, 2]$, and has β-Hölder continuous gradient for some $\beta \in (0, 1]$, we prove that $\widetilde{\mathcal{O}}\big(d^{\frac{1}{\beta} + \frac{1+\beta}{\beta}(\frac{2}{\alpha} - \mathbb{1}_{\{\alpha\neq 1\}})} \epsilon^{-\frac{1}{\beta}}\big)$ steps are sufficient for LMC to reach ϵ accuracy in KL-divergence. In moderate dimensions when $d \ll \epsilon^{-1}$, the tail growth rate α does not impact the rate estimate, whereas in high dimensions where $d = \Omega(\epsilon^{-1})$, tail growth enters the rate estimate through dimension dependency.
- As a key step in deriving the above convergence rate, we establish a mLSI (1.6) with an explicit constant λ, and a target dependent moment function $\mathrm{M}_s(\rho + \nu_*)$ for any order $s \ge 2$. Both of these are crucial in deriving a rate estimate with a right dependence on the dimension d as well as the accuracy ϵ. The final result is obtained by employing the mLSI condition (1.6) for the optimal moment order $s = \mathcal{O}(\log(d/\epsilon))$.
- In order to use the condition mLSI (1.6), we establish linearly diverging moment estimates for the LMC iterates under weak dissipativity. Somewhat surprisingly, this is sufficient to establish a convergence rate bound for LMC in KL-divergence. To our knowledge, this is the first convergence result for the LMC for weakly smooth potentials that exhibit subquadratic growth, without relying on regularization and/or smoothing techniques.
- Finally, using Csiszár-Kullback-Pinsker inequalities, the above estimates can be translated to total variation and L_α-Wasserstein metrics with respective rates $\widetilde{\mathcal{O}}\big(d^{\frac{1}{\beta} + \frac{1+\beta}{\beta}(\frac{2}{\alpha} - \mathbb{1}_{\{\alpha\neq 1\}})} \epsilon^{-\frac{2}{\beta}}\big)$ and $\widetilde{\mathcal{O}}\big(d^{\frac{3}{\beta} + \frac{1+\beta}{\beta}(\frac{2}{\alpha} - \mathbb{1}_{\{\alpha\neq 1\}})} \epsilon^{-\frac{2\alpha}{\beta}}\big)$.

Rest of the paper is organized as follows. We compare our results to those of existing works next, and briefly review notation in the remainder of this section. In Section 2, we establish the main technical results on the convergence of LMC for potentials with certain growth and smoothness properties. Section 3 discusses further implications of the tools developed in Section 2. We give concrete examples in Section 4, by applying these tools to non-convex sampling problems that are also weakly smooth. Finally, we conclude in Section 5 with brief remarks on future work. For a detailed survey of additional related work, we refer to Section G. Proofs of the main theorems and corollaries are deferred to appendix and are provided in Sections A, B, C, D, E and F.

Related work and comparisons. In Table 1, we compare the assumptions and results of this paper to those of existing works that only make the first order smoothness assumption. Among these, Dalalyan (2017b); Durmus and Moulines (2017); Cheng and Bartlett (2018); Chatterji et al. (2020); Vempala and Wibisono (2019); Durmus et al. (2019a) are in the quadratic growth regime, and achieve the best rates known to authors. Our results recover the rates of Vempala and Wibisono (2019) for smooth potentials ($\beta = 1$) satisfying the LSI condition ($\alpha = 2$). Cheng and Bartlett (2018); Dalalyan (2017b) establish guarantees for convex and smooth potentials; however, these results rely on surrogate potentials that are strongly convex which causes significant drops in rate estimates, and cannot tolerate perturbations on the potential. In contrast to these results, our analysis provides a continuous interpolation in both the growth rate $\alpha \in (1, 2]$, and the order of smoothness $\beta \in (0, 1]$. In case of linear growth when $\alpha = 1$, there is no convexity in the tails which is why the rate loses an additional factor in dimension dependency (see section E). The results of Chatterji et al. (2020) on the vanilla LMC require the potential to have a composite structure, namely, $f(x) = U(x) + \psi(x)$ where $\psi(x)$ is a strongly convex and smooth function, and $U(x)$ is a convex function with β-Hölder continuous gradient. It is worth emphasizing that the actual rate obtained in Cheng and Bartlett (2018) is $\widetilde{\mathcal{O}}\big(d\epsilon^{-3} \times \mathcal{W}_2^4(\rho_0, \nu_*)\big)$, and depends polynomially on the L_2-Wasserstein distance between the initial distribution and the target, whereas other works depend logarithmically on this difference in terms of KL-divergence. For a potential growing with rate α, one may show $\mathcal{W}_2^2(\rho_0, \nu_*) \lesssim d^{2/\alpha}$ justifying the reported rate in Table 1. For details of the initializations when $\alpha = 2$, we refer to Cheng et al. (2018b). For additional related work, we refer reader to Section G.

Notation. For a real number $x \in \mathbb{R}$, we denote its absolute value with $|x|$. We denote the p-norm of a vector $x \in \mathbb{R}^d$ with $\|x\|_p$ and whenever $p = 2$, we omit the subscript and simply write $\|x\| \triangleq \|x\|_2$ to ease the notation. We use I_d to denote the identity matrix in d-dimensions. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we define its infinity norm as $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$. $\mathrm{M}_s(f)$ is used to denote the modified s-th moment of the function f (which is not necessarily a distribution), defined as $\mathrm{M}_s(f) = \int f(x)(1 + \|x\|^2)^{s/2} dx$.

For probability densities p,q on $\mathbb{R}^d$, we use $\mathrm{KL}\big(p|q\big)$ and $\mathrm{I}\big(p|q\big)$ to denote their KL-divergence (or relative entropy) and relative Fisher information, respectively, which are defined as

$$\mathrm{KL}\big(p|q\big) = \int p(x) \log \frac{p(x)}{q(x)} dx, \quad \text{and} \quad \mathrm{I}\big(p|q\big) = \int p(x) \Big\| \nabla \log \frac{p(x)}{q(x)} \Big\|^2 dx. \tag{1.7}$$

We denote the entropy of p with $\mathrm{H}\,(p) = -\int p(x) \log p(x) dx$. Denoting the Borel σ-field of $\mathbb{R}^d$ with $\mathcal{B}(\mathbb{R}^d)$, L_α-Wasserstein for $\alpha > 0$ and total variation metrics are defined as

$$\mathcal{W}_\alpha(p, q) = \inf_\nu \left(\int \|x - y\|^\alpha d\nu(p, q) \right)^{1/\alpha}, \text{ and } \mathrm{TV}\,(p, q) = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left| \int_A (p(x) - q(x)) dx \right|,$$

(a) KL-divergence

WORK	CONVERGENCE RATE	SMOOTHNESS	CURVATURE	PERTURBATION
Cheng and Bartlett (2018) Durmus et al. (2019a)	$\widetilde{\mathcal{O}}\left(d\epsilon^{-1}\right)$	Lipschitz gradient	Strongly Convex	None
Vempala and Wibisono (2019)	$\widetilde{\mathcal{O}}\left(d\epsilon^{-1}\right)$	Lipschitz gradient	Strongly Convex	Bounded difference
Cheng and Bartlett (2018)	$\widetilde{\mathcal{O}}\left(d^{1+\frac{4}{\alpha}}\epsilon^{-3}\right)$	Lipschitz gradient	Convex Growth rate α	None
This work	$\widetilde{\mathcal{O}}\left(d^{\frac{1}{\beta}+\frac{1+\beta}{\beta}\left(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq 1\}}\right)}\epsilon^{-\frac{1}{\beta}}\right)$	β-Hölder gradient	Tail growth $\sim \|x\|^{\alpha}$	Bounded difference

(b) Total Variation

WORK	CONVERGENCE RATE	SMOOTHNESS	CURVATURE	PERTURBATION
Dalalyan (2017b) Durmus and Moulines (2017)	$\widetilde{\mathcal{O}}\left(d\epsilon^{-2}\right)$	Lipschitz gradient	Strongly convex	None
Dalalyan (2017b)	$\widetilde{\mathcal{O}}\left(d^{3}\epsilon^{-4}\right)$	Lipschitz gradient	Convex	None
Chatterji et al. (2020)	$\widetilde{\mathcal{O}}\left(d^{2+\frac{1}{\beta}}\epsilon^{-\frac{2}{\beta}}\right)$	Lipschitz+β-Hölder gradient	Strongly Convex	None
This work	$\widetilde{\mathcal{O}}\left(d^{\frac{1}{\beta}+\frac{1+\beta}{\beta}\left(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq 1\}}\right)}\epsilon^{-\frac{2}{\beta}}\right)$	β-Hölder gradient	Tail growth $\sim \|x\|^{\alpha}$	Bounded difference

Table 1: Convergence rate estimates in (a) KL-divergence and (b) TV distance for the LMC algorithm in various papers and their accompanying assumptions. Comparison is made with results relying only on first order smoothness. The perturbation column indicates whether the results are robust against adding a bounded perturbation to the potential (see Lemma 7).

where in the first formula, infimum runs over the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ that has marginals with corresponding densities p and q.

2. Main Results

We develop our explicit bounds on the convergence rate of the LMC algorithm in three key steps. First, in Theorem 1, we prove a modified log-Sobolev inequality (mLSI) for a class of asymptotically convex degenerate potentials described in Assumption 1, which can accommodate for sub-quadratic tail growth. The condition mLSI relies on the moments of the Markov chain defined by the iterates of LMC; thus, in Proposition 2, we prove that any order moments of the LMC iterates grow at most linearly in the number of iterations, an estimate that is diverging in the limit. Finally in Theorem 3, we invoke these two results for an arbitrary moment order and establish a general convergence result, which in turn yields the main result of this paper after tuning the moment order in Corollary 4. We focus on the following class of potentials functions.

Assumption 1 (Degenerate convexity at ∞) *The potential function $f(x)$ is degenerately convex at infinity in the sense that there exist a function $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ such that for a constant $\xi \geq 0$*

$$\left\|f - \tilde{f}\right\|_{\infty} \leq \xi, \quad \textit{where } \tilde{f} \textit{ satisfies } \quad \nabla^2 \tilde{f}(x) \succeq \frac{\mu}{\left(1+\frac{1}{4}\|x\|^2\right)^{\theta/2}} I_d, \tag{2.1}$$

for some $\mu > 0$ and $\theta \geq 0$.

The parameter θ is intimately related to the tails of f. Consider, for example, the following potential function $f(x) = \|x\|^\alpha$ for $\alpha \in [1, 2]$. The case $\alpha = 2$ corresponds to quadratic growth with $\theta = 0$, and it is easy to see that for a superlinear tail $\alpha \in (1, 2]$, one has $\theta = 2 - \alpha$. However, when the tail is exactly linear with $\alpha = 1$, the assumption can be shown to hold for any $\theta > 2$. In section E, we show that for functions with linear growth, this assumption does not hold for any $\theta \leq 2$.

It is known that the LSI condition (1.3) is not satisfied when $\alpha < 2$, for example for the potential $f(x) = |x|^\alpha + c$ (see e.g. Bobkov and Götze (1999)); therefore, for the above class of potentials, we state the following log-Sobolev-type inequality.

Theorem 1 (mLSI) *If the potential $f = -\log \nu_*$ satisfies Assumption 1, then the following inequality holds for all $s \geq 2$,*

$$\forall \rho, \;\; \mathrm{KL}\big(\rho|\nu_*\big) \leq \lambda \mathrm{I}\big(\rho|\nu_*\big)^{1-\delta} \mathrm{M}_s(\rho + \nu_*)^\delta, \tag{2.2}$$

where δ and λ are constants that depend on s, and are defined as

$$\delta \triangleq \frac{\theta}{s - 2 + 2\theta} \in [0, 1/2], \;\; \textit{and} \;\; \lambda \triangleq 4e^{2\xi} \mu^{-\frac{s-2}{s-2+2\theta}}.$$

The constants λ and δ are explicit, and the above inequality reduces exactly to the LSI condition (1.3) up to the absolute constant 4 when $\theta = 0$ and $\xi = 0$, in which case the potential function f is strongly convex. Modified LSI-type inequalities such as (2.2) as well as weak Poincaré inequalities appear in the analysis of diffusion operators (Bakry et al., 2013). The mLSI condition (2.2) is similar in nature to the modified LSI of Toscani and Villani (2000); yet, the latter was established for the purpose of proving the rate $\mathcal{O}(t^{-\infty})$ for the diffusion process (1.1), and will yield a bound on the convergence rate that is worse than what will be established below in Corollary 4. It also cannot recover the existing rates (e.g. Vempala and Wibisono (2019)) in the limit case $\alpha \to 2$. Our proof builds on the construction made in Toscani and Villani (2000) and uses the results of Bakry and Émery (1985); Holley and Stroock (1987), which we defer to Section A.

The gradient of the potential function is employed as the drift of Langevin diffusion (1.1), and it also governs its discretization, the LMC algorithm (1.2). The growth behavior of this term is regulated in the following assumption which is a relaxation of the standard 2-dissipativity condition, $\langle \nabla f(x), x \rangle \geq a\|x\|^2 - b$ for some $a, b > 0$ (Mattingly et al., 2002; Meyn and Tweedie, 2012), also commonly used in non-convex optimization (Raginsky et al., 2017; Yu et al., 2020).

Assumption 2 (α-dissipativity & ζ-growth of gradient) *For $\alpha \in [1, 2]$ and $a, b > 0$, we have*

$$\langle \nabla f(x), x \rangle \geq a\|x\|^\alpha - b \;\; \textit{for all} \;\; x \in \mathbb{R}^d. \tag{2.3}$$

Moreover, for a positive constant $\zeta \leq \alpha/2$, the gradient satisfies the following growth condition,

$$\|\nabla f(x)\| \leq M(1 + \|x\|^\zeta) \;\; \textit{for all} \;\; x \in \mathbb{R}^d. \tag{2.4}$$

Note that when the tail growth is superlinear $\alpha \in (1, 2]$, the parameter θ in Assumption 1 satisfies $\theta = 2 - \alpha$ where α is as in Assumption 2. The key difference between the cases $\alpha = 2$ and $\alpha < 2$ is that the former implies that the LMC iterates have uniformly bounded moments of all orders (Erdogdu et al., 2018), whereas in the latter case, we are not aware of any methods to establish such uniform bounds on all moments. This poses significant challenges in the proof. That is, we establish that the moments of LMC can diverge at most linearly, and even though it is not immediately

clear that LMC even converges in this setup, we are able to show that this estimate is sufficient to establish a non-asymptotic bound on the convergence rate of the algorithm. It is also worth noting that under an additional condition on the gradient perturbation, i.e. $\|\nabla f - \nabla \tilde{f}\|_\infty \leq \xi$, it can be shown that (2.1) implies (2.3) in Assumption 2 (see Lemma 24); however, the above setting is more general and covers a wider range of potentials, justifying the current presentation.

In a representative analysis of LMC, one considers a sequence of interpolation diffusion processes $\{\widetilde{x}_{k,t}\}_{k\in\mathbb{N},t\geq 0}$ where each iteration x_{k+1} of the LMC (1.2) can be written as $\widetilde{x}_{k,\eta}$, where

$$d\widetilde{x}_{k,t} = -\nabla f(x_k)dt + \sqrt{2}dB_t \quad \text{with} \quad \widetilde{x}_{k,0} = x_k, \tag{2.5}$$

for an appropriate Brownian motion B_t. Denoting the distribution of $\widetilde{x}_{k,t}$ with $\widetilde{\rho}_{k,t}$, it can be shown that the time derivative of the KL-divergence between $\widetilde{\rho}_{k,t}$ and the target, $d\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big)/dt$, reduces to the negative relative Fisher information $-\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big)$ up to an additive error term that depends on the difference between the LMC iterate x_k and the its interpolating diffusion $\widetilde{x}_{k,t}$ (see for example Vempala and Wibisono (2019, Proof of Lemma 3)), which yields the inequality

$$\forall k \in \mathbb{N}, \forall t \geq 0, \quad \tfrac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) \leq -\tfrac{3}{4}\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\right].$$

Combining this with mLSI (2.2) for $\rho = \widetilde{\rho}_{k,t}$, one obtains the following differential inequality for the interpolating diffusion process (2.5),

$$\tfrac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) \leq -\tfrac{3}{4}\left(\tfrac{1}{\lambda}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big)\right)^{\frac{1}{1-\delta}} \mathrm{M}_s(\widetilde{\rho}_{k,t}+\nu_*)^{-\frac{\delta}{1-\delta}} + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\right]. \tag{2.6}$$

The convergence rate of LMC can be derived by analyzing the differential inequality (2.6), which requires appropriate estimates on the moments $\mathrm{M}_s(\widetilde{\rho}_{k,\eta}+\nu_*) = \mathrm{M}_s(\rho_{k+1}+\nu_*)$.

Proposition 2 *If the potential $f = -\log \nu_*$ satisfies Assumption 2, then denoting the distribution of the k-th iterate of LMC with ρ_k, for a step size satisfying $\eta \leq \frac{1}{2}\big(1 \wedge \frac{a}{2M^2}\big)$, we have*

$$\mathrm{M}_s(\rho_k+\nu_*) \leq \mathrm{M}_s(\rho_0+\nu_*) + C_s k\eta, \quad \textit{for even integer } \ s \geq 2, \tag{2.7}$$

$$\textit{where} \qquad C_s \triangleq \left(\tfrac{3a+2b+3}{1\wedge a}\right)^{\frac{s-2}{\alpha}+1} s^s d^{\frac{s-2}{\alpha}+1}. \tag{2.8}$$

Although the bound (2.7) grows linearly with the number of iterations and diverges in the limit $k \to \infty$, this estimate is sufficient to establish a global convergence guarantee for the LMC algorithm. The leading coefficient in the bound C_s (2.8) is of order $\mathcal{O}(d^{\frac{s-2}{\alpha}+1})$ which is the same order as in the continuous-time case (cf. Lemma 10).

In what follows, we make an assumption on the order of smoothness of the potential function f in order to obtain an estimate for the additive error term in the differential inequality (2.6). In this context, order of smoothness refers to the Hölder exponent of the gradient of the potential.

Assumption 3 (Order of smoothness) *The potential function f is differentiable with β-Hölder continuous gradient with constant L, i.e.*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|^\beta \quad \textit{for all} \quad x, y \in \mathbb{R}^d, \tag{2.9}$$

where the order of smoothness β satisfies $\zeta \leq \beta \leq 1$ for the constant ζ in (2.4).

Potentials with order of smoothness $\beta = 1$ are termed as *smooth* and those with $\beta < 1$ are often referred to as *weakly smooth* in the literature (Chatterji et al., 2020; Nesterov, 2015), a term that is borrowed from optimization theory. Our results cover potentials satisfying (2.9) for any $\beta \in (0, 1]$.

β-Hölder continuity already imposes a growth condition on the gradient (2.4) with $\zeta = \beta$. However, we state these separately as the order of smoothness β and the growth rate ζ need not be the same; a smaller growth rate on the gradient improves certain estimates in the main result, which in turn allows us to cover a wider class of potentials. For example, the function $f(x) = \sqrt{1 + \|x\|^2}$ is smooth with Lipschitz gradient, but its gradient is also bounded implying $\zeta = 0$. One cannot simply use $\zeta = 1$ since the condition $\zeta \leq \alpha/2$ in Assumption 2 implies that $\alpha \geq 2$ which is clearly not true for this potential. Hence, keeping these parameters separate allows us to cover a wider range of potentials. The relationship among these parameters can be summarized as

$$2\zeta \leq \alpha \leq \zeta + 1 \leq \beta + 1.$$

If one requires quadratic growth on the potential, i.e. $\alpha = 2$, this immediately implies that the smoothness order is at least 1, i.e $\beta \geq 1$, in which case only smooth potentials become feasible.

Before we present the main technical result of this paper, we note that when $\alpha > 1$, all assumptions are satisfied for potentials of the form

$$f(x) = \|x\|^\alpha + 10\cos(\|x\|) + c.$$

This potential is non-convex and it does not have a Lipschitz gradient, and serves as a canonical example that demonstrates the wide applicability of the following result. For additional (non-trival) examples, we refer to Section 4 (specifically 4.2).

Theorem 3 *Suppose the potential $f = -\log \nu_*$ satisfies Assumptions 1, 2, 3, and denote the distribution of the k-th iterate of LMC with ρ_k. Then, for a sufficiently small ϵ satisfying $\epsilon \leq \psi$ where ψ is defined in (C.10), and for some $\Delta_0 > 0$ upper bounding the error at initialization, i.e. $\mathrm{KL}(\rho_0|\nu_*) \leq \Delta_0$, if the step size satisfies*

$$\eta = (\sigma c_\gamma)^{-\frac{1}{1+\beta}} d^{-\frac{\alpha+\theta}{\alpha\beta} - \frac{\gamma}{\beta+1}} \left(1 + (1-\alpha/2)\log(d)\right)^{-\frac{1}{\beta}} \log\left(\frac{\Delta_0}{\epsilon}\right)^{-\frac{\gamma}{1+\beta}} \left(\frac{2}{\epsilon}\right)^{-\frac{1}{\beta} - \frac{\gamma}{1+\beta}}, \quad (2.10)$$

then the LMC iterates reach ϵ-accuracy of the target, i.e. $\mathrm{KL}(\rho_N|\nu_*) \leq \epsilon$, after N steps for

$$N = c_\gamma d^{\frac{\alpha+\theta+\beta\theta}{\alpha\beta}+\gamma} \left(1 + (1-\alpha/2)\log(d)\right)^{\frac{1}{\beta}} \log\left(\frac{2\Delta_0}{\epsilon}\right)^{1+\gamma} \left(\frac{2}{\epsilon}\right)^{\frac{1}{\beta}+\gamma},$$

where γ is given by

$$\gamma \triangleq \gamma(s) = \tfrac{(1+\beta)\theta}{\beta(s-2)} \quad \textit{for any even integer } \ s \geq 4, \quad (2.11)$$

and σ and c_γ are constants given as

$$\sigma = 4L^2\left(1 + 2a^\beta\left[1 + \tfrac{2\alpha}{a}\left(\log\left(16\pi/a\right) + M\left(2 + 2b/a\right)^2 + b + |f(0)|\right)\right]\right),$$
$$c_\gamma = \sigma^{\frac{1}{\beta}}(16\lambda)^{1+\frac{1}{\beta}+2\gamma}\left(\tfrac{\mathrm{M}_s(\rho_0+\nu_*)}{16 d^{\frac{s-2}{\alpha}+1}} \vee \tfrac{s^s}{16}\left(\tfrac{3a+2b+3}{1\wedge a}\right)^{\frac{s-2}{\alpha}+1}\right)^\gamma.$$

The above theorem, proved in Section C, implies that for smooth potentials that satisfy LSI, i.e. $\alpha = 2$ and $\beta = 1$, we have $\gamma = \theta = 0$; thus, LMC achieves the rate of $\widetilde{\mathcal{O}}\left(d\epsilon^{-1}\right)$, recovering the rate established by Vempala and Wibisono (2019). In the general case, Theorem 3 implies the convergence rate bound $\widetilde{\mathcal{O}}\big(\gamma^{-1}d^{\frac{\alpha+\theta+\beta\theta}{\alpha\beta}+\gamma}\epsilon^{-\frac{1}{\beta}-\gamma}\big)$ where $\gamma > 0$ is given in (2.11) and can be arbitrarily small. However, one cannot simply let $\gamma \to 0$ by taking the limit $s \to \infty$; for any other potential function with subquadratic tail growth $\alpha < 2$, the parameter γ requires tuning.

The upper bound ψ on accuracy, as stated in (C.10), is $\mathcal{O}(1)$, depending only on the fixed problem parameters and the initial KL-divergence Δ_0. When initialized with a Gaussian, Δ_0 can also be characterized with the fixed problem parameters (see Lemma 26). More importantly, the upper bound on ϵ, as stated in (C.10), does not depend on the moment order s, which enables us to choose $s = \mathcal{O}\left(\log\left(d\epsilon^{-1}\right)\right)$ and accordingly $\gamma = \mathcal{O}\big(1/\log\left(d\epsilon^{-1}\right)\big)$, which in turn yields the best known bound on the convergence rate that can be achieved by our method. This is formalized in the next corollary which is the main result of this paper.

Corollary 4 *Suppose the potential $f = -\log \nu_*$ satisfies Assumptions 1, 2, 3, and denote by ρ_k, the distribution of the k-th iterate of LMC initialized with $x_0 \sim \mathcal{N}(x, I_d)$ for any $x \in \mathbb{R}^d$ and Δ_0 upper bounding the error at initialization (see Lemma 26). Then, for a sufficiently small ϵ satisfying $\epsilon \le \psi$ where ψ is defined in (C.10), if the step size satisfies (2.10) for $s = 2 + 2\lceil\log(\frac{6d}{\epsilon})\rceil$, the iterates of LMC reaches ϵ-accuracy of the target, i.e. $\mathrm{KL}\big(\rho_N|\nu_*\big) \le \epsilon$, after N steps satisfying*

$$N \le cd^{\frac{\alpha+\theta+\beta\theta}{\alpha\beta}}\left(1+(1-\alpha/2)\log(d)\right)^{\frac{1}{\beta}}\log\left(\tfrac{2\Delta_0}{\epsilon}\right)^{1+\frac{(1+\beta)\theta}{2\beta}}\left(2+2\left\lceil\log\left(\tfrac{6d}{\epsilon}\right)\right\rceil\right)^{\frac{2(1+\beta)\theta}{\beta}}\left(\tfrac{2}{\epsilon}\right)^{\frac{1}{\beta}},$$

where c is a constant independent of d and ϵ, and given as

$$c = e^{\frac{(1+\alpha)(1+\beta)\theta}{\alpha\beta}}\sigma^{\frac{1}{\beta}}\left(\frac{64e^{2\xi}}{1\wedge\mu}\right)^{1+\frac{1+\theta+\beta\theta}{\beta}}\left(\tfrac{3a+2b+3}{1\wedge a}\right)^{\frac{2(1+\beta)\theta}{\alpha\beta}}.$$

The above corollary, proved in Section F, implies that the LMC algorithm achieves ϵ accuracy of the target in KL-divergence in $\widetilde{\mathcal{O}}\big(d^{\frac{\alpha+(1+\beta)\theta}{\alpha\beta}}\epsilon^{-\frac{1}{\beta}}\big)$ steps. Whenever the tail growth of the potential is superlinear and behaves like $\|x\|^\alpha$ for $\alpha \in (1,2]$, Assumption 1 holds for $\theta = 2-\alpha$; thus, Corollary 4 can be invoked for this choice of θ, yielding the convergence rate estimate $\widetilde{\mathcal{O}}\big(d^{\frac{2}{\alpha}(1+\frac{1}{\beta})-1}\epsilon^{-\frac{1}{\beta}}\big)$. On the other hand, when the potential function has linear tail growth (i.e. $f(x) \sim \|x\|$), then by setting $\tilde{f} = \big(1+\|x\|^{1+\tau}\big)^{1/(1+\tau)}$ where $\tau \in (0,1)$, one can verify that Assumption 1 holds for $\theta = 2+\tau$. By tuning this parameter with $\tau = 1/\log(6d)$, we obtain an estimate of $\widetilde{\mathcal{O}}\big(d^{2+\frac{3}{\beta}}\epsilon^{-\frac{1}{\beta}}\big)$. Putting this all together, one can simply use $\theta = 2-\alpha\mathbb{1}_{\{\alpha\neq 1\}}$, which yields the advertised estimate $\widetilde{\mathcal{O}}\big(d^{\frac{1}{\beta}+(1+\frac{1}{\beta})\left(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq1\}}\right)}\epsilon^{-\frac{1}{\beta}}\big)$. We emphasize that, in moderate dimensions where $d \ll \epsilon^{-1}$, the estimate only depends on the order of smoothness β, whereas in high dimensions where $d = \Omega(\epsilon^{-1})$, the tail growth rate α enters the bound through dimension dependency.

The obtained rate is continuous in the domain $\alpha \in (1,2]$ and $\beta \in (0,1]$; however, there is a discontinuous jump at $\alpha = 1$ due to lack of convexity, this is explored further in Section E, where we prove that linear potentials cannot be convex degenerate with $\theta \le 2$. One can verify that $\theta = 1$ implies a tail growth of $\|x\|\log(1+\|x\|)$ which is superlinear, but in terms of Assumption 2, we still have $\alpha = 1$. In this case, the tail growth cannot be explained with a polynomial in $\|x\|$; therefore, $\theta = 1 \neq 2-\alpha\mathbb{1}_{\{\alpha\neq1\}}$ because of the additional logarithmic factor. One should also note that $\alpha = 1$ is the exception in the sense that introducing additional $\log$ factors when $\alpha > 1$ does not change θ, and ultimately the bound on the convergence rate stays the same.

In the next corollary, we translate the result in KL-divergence to other measures of distance between probability distributions such as total variation (TV) and L_α-Wasserstein metrics. The proof is straightforward, and postponed to Section F. In order to reach the same level of accuracy in different probability metrics, one needs to adapt the step size accordingly. This requires a different upper bound on the accuracy ϵ in each metric.

Corollary 5 (Other Measures of Distance) *Instantiate the assumptions and notation of Theorem 3. If LMC (1.2) is initialized with $x_0 \sim \mathcal{N}(x, I_d)$ for any $x \in \mathbb{R}^d$, then, the following table summarizes the number of steps that are sufficient for obtaining an ϵ-accurate sample in various distance measures.*

METRIC	NUMBER OF STEPS	BOUND ON ϵ
TV	$\widetilde{\mathcal{O}}\big(d^{\frac{\alpha+\theta+\beta\theta}{\alpha\beta}}\epsilon^{-\frac{2}{\beta}}\big)$	$\epsilon \le \sqrt{\psi/2}$ (see (C.10))
$\mathcal{W}_\alpha$	$\widetilde{\mathcal{O}}\big(d^{\frac{3\alpha+\theta+\beta\theta}{\alpha\beta}}\epsilon^{-\frac{2\alpha}{\beta}}\big)$	(F.1)
$\mathcal{W}_2$ ($\alpha=2$, $\beta=1$ and $\theta=0$)	$\widetilde{\mathcal{O}}\big(d\epsilon^{-2}\big)$	(F.2)

Table 2: Convergence rate estimates in various metrics.

As before, one can simply use $\theta = 2 - \alpha\mathbb{1}_{\{\alpha\neq 1\}}$. In the case of strongly convex and smooth potentials, i.e. $\alpha = 2$ and $\beta = 1$, the corollary recovers the rate $\widetilde{\mathcal{O}}\left(d\epsilon^{-2}\right)$ in TV distance, which was established in Durmus and Moulines (2017). For functions that have quadratic growth, the convergence rate in $\mathcal{W}_\alpha$ can be made better, because the result relies on the CKP inequality (see Lemma 23) which does not recover Talagrand's inequality when $\alpha = 2$ (Bolley and Villani, 2005). Therefore, the case $\alpha = 2$ is handled separately, where Talagrand's inequality is available. We emphasize that $\theta = 0$ corresponds to potentials with tail growth rate $\alpha = 2$. Since in this case $\gamma = 0$, there is no need to tune s to a specific moment. The quadratic growth setting only covers smooth potentials, because Assumption 3 implies that the gradient of the potential has a tail growth rate upper bounded by β, which in turn upper bounds the tail growth of f with $\beta + 1$. Thus, the only feasible value for β is 1 in this case.

3. Further Implications

Convex potentials. If a potential function has tail growth rate α, then there exist $a, b > 0$ such that

$$f(x) \ge a\|x\|^\alpha - b, \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{3.1}$$

The next proposition (cf. Bakry et al. (2008, Lemma 2.2)) shows that convex potentials exhibit at least linear tail growth, i.e. (3.1) holds for some $\alpha \ge 1$. Furthermore, Assumption 2 is also satisfied for the same value of α in (3.1). The proof is deferred to Section D.

Proposition 6 *For any differentiable convex potential $f : \mathbb{R}^d \to \mathbb{R}$ (e.g. $\int e^{-f} < \infty$), there exist constants $a, b > 0$ such that (3.1) is satisfied for $\alpha = 1$. If further (3.1) is satisfied for some $\alpha \ge 1$, then α-dissipativity condition (2.3) of Assumption 2 holds for the same value of α.*

In light of Proposition 6, one can argue that for convex potentials, the limiting factor is the smoothness of the potential function f. For example, consider $f_1(x) = \sqrt{1 + \|x\|^2}$ and $f_2(x) = \|x\|$. These are both convex functions with linear growth and they satisfy Assumptions 1 and 2 with the same θ and α. While f_1 is smooth, f_2 does not satisfy Assumption 3 for any β.

Non-convex potentials. Assumptions 1, 2 and 3 are robust to bounded perturbations. In other words, if these assumptions are satisfied for a potential, then they also hold for its finite perturbation. The following lemma formalizes this statement and its proof is deferred to Section F.

Lemma 7 *Let f be a potential satisfying Assumptions 1, 2 and 3 for $\alpha > 1$. Then, for any bounded function ϕ with β-Hölder continuous and bounded gradient, $f + \phi$ can be normalized to a potential, also satisfying Assumptions 1, 2 and 3. Further, if we additionally have $\sup_{x \in \mathbb{R}^d} \|\nabla \phi(x)\| < a$ for the constant a as in Assumption 2, the above result also holds for $\alpha = 1$.*

The previous lemma shows that Corollary 4 is robust to finite perturbations. Moreover, investigating the proof reveals that the growth rate α and the order of smoothness β do not change (along with θ), which means that our estimate of the convergence rate of LMC for the perturbed potential is the same as that of the original potential.

4. Applications

In this section, we apply the results of Sections 2 and 3 to various illustrative potential functions. We begin with a few basic examples in order to demonstrate the effect of tail growth and the order of smoothness on the convergence of LMC.

4.1. Pedagogical examples

Example 1 (Weakly smooth potential with subquadratic tails) Consider the potential $f(x) = \|x\|^\alpha$ for $\alpha \in (1, 2)$. This potential is not smooth with an unbounded Hessian near the origin, and its tails are subquadratic which means the tails of the target $\nu_* \propto e^{-f}$ are heavier than those of the Gaussian distribution. It is straightforward to verify our assumptions for this potential (see e.g. Lemma 35 for Assumption 3). Therefore, Corollary 4 implies that we can reach ϵ accuracy in KL-divergence after taking $\widetilde{\mathcal{O}}\big(d^{\frac{3-\alpha}{\alpha-1}} \epsilon^{-\frac{1}{\alpha-1}}\big)$ steps. To highlight the impact of the order of smoothness, consider the smooth potential $f(x) = (1 + \|x\|^2)^{\alpha/2}$ which has the same tail growth as $\|x\|^\alpha$, for which our rate estimate improves to $\widetilde{\mathcal{O}}\big(d^{\frac{4-\alpha}{\alpha}} \epsilon^{-1}\big)$.

Our results allow for finite perturbations, for example, consider the function $\phi(x) = \cos(\|x\|)$ which is bounded with bounded first derivative. Its gradient is given by $\nabla \phi(x) = -\frac{x}{\|x\|} \sin(\|x\|)$ which is Lipschitz continuous; hence, Lemma 34 implies that it is also β-Hölder continuous. By Lemma 7, the rate obtained from Corollary 4 is applicable to the potential $g(x) = \|x\|^\alpha + 10 \cos(\|x\|) + \xi$, and the convergence rate estimate $\widetilde{\mathcal{O}}\big(d^{\frac{3-\alpha}{\alpha-1}} \epsilon^{-\frac{1}{\alpha-1}}\big)$ still holds.

Example 2 (Smooth potential with linear tails): Since $\|x\|$ has discontinuous gradient at the origin, we consider $f(x) = \sqrt{1 + \|x\|^2}$ as an example of a smooth potential with linear growth. It is straightforward to verify our assumptions with the parameter values $\xi = 0$, $\alpha = 1$, $\beta = 1$ and $\theta = 3$ (by setting $\tilde{f} = f$). Plugging these parameters in Corollary 4, we obtain the convergence rate bound $\widetilde{\mathcal{O}}\left(d^7 \epsilon^{-1}\right)$ for the LMC algorithm in KL-divergence.

The dimension dependency in the previous convergence rate bound can be improved by changing $\tilde{f}$ to a function that is different from f. Observe that the difference between $\sqrt{1 + \|x\|^2}$ and

$\left(1+\|x\|^{1+\tau}\right)^{1/(1+\tau)}$ is bounded for any $\tau \in (0,1)$. Thus, if we set $\tilde{f}(x) = \left(1+\|x\|^{1+\tau}\right)^{1/(1+\tau)}$, Assumption 1 is satisfied with $\theta = 2+\tau$ and $\mu = \mathcal{O}(\tau)$. Setting $\tau = \mathcal{O}(\log(6d)^{-1})$ and invoking Corollary 4 implies an estimate of $\widetilde{\mathcal{O}}(d^5\epsilon^{-1})$, for a potential like $f(x) = \sqrt{1+\|x\|^2}+0.5\cos(\|x\|)$. We note that in this case, the norm of the perturbation needs to be strictly smaller than 1, otherwise Assumption 2 is no longer satisfied.

4.2. Bayesian regression

In this section, the fixed problem parameters such as M and L depend on the data, and the rates are obtained assuming these constants are $\mathcal{O}(1)$. Depending on the data scaling, these parameters may depend on the dimension as well as the number of samples n, in which case the rates can still be obtained by using the explicit formulas presented in Theorem 3 and Corollary 4.

Example 3 (Bridge regression): Our analysis shows that the LMC algorithm can handle potentials that are weakly smooth, which comes up frequently in Bayesian statistics. For example, denoting the matrix of covariates with $V = \{v_i\}_{i=1}^n \in \mathbb{R}^{n\times d}$, and the response vector with $Y = \{y_i\}_{i=1}^n \in \mathbb{R}^n$, in Bridge linear regression (Fu, 1998; Frank and Friedman, 1993), one assumes a linear model $Y = Vx + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I_n)$, and a prior proportional to $\exp(-\sum_{i=1}^d |x_i|^q)$. Therefore, sampling from the resulting posterior is equivalent to sampling from the potential function $f(x) = \|Y - Vx\|^2 + \sum_{i=1}^d |x_i|^q$. Assume that we have $V^\top V \succ 0$ and $q \in (1,2)$. Then, the above potential has quadratic growth which, in our framework, translates to setting $\theta = 0$ in Assumption 1, and choosing $\alpha = 2$ and $\zeta = 1$ in Assumption 2. This potential lacks smoothness; yet in the close neighborhood of the origin, Assumption 3 holds with $\beta = q-1$. On the other hand, ∇f has linear growth and when $\|x-y\| \geq 1$, Assumption 3 holds with $\beta = 1$. Initially, it might seem that our results are not applicable to this potential but by adapting Assumption 3 to this setting as

$$\|\nabla f(x) - \nabla f(y)\| \leq L\left(\|x-y\|^{\beta_1} + \|x-y\|^{\beta_2}\right) \quad \text{for all} \quad x, y \in \mathbb{R}^d,$$

where $\beta_1 < \beta_2$, and some minor changes to Lemma 15, our convergence results also cover this potential. In this example, we need to set $\beta_1 = q-1$, $\beta_2 = 1$, and $\zeta \leq \beta_2$ which yields the convergence estimate $\widetilde{\mathcal{O}}\big(d^{\frac{1}{q-1}}\epsilon^{-\frac{1}{q-1}}\big)$ in KL-divergence.

Example 4 (Bayesian logistic regression): In Bayesian logistic regression, we are given n samples according to the logistic regression model

$$V = \{v_i\}_{i=1}^n \in \mathbb{R}^{n\times d}, Y = \{y_i\}_{i=1}^n \in \{0,1\}^n \quad \text{and} \quad \mathbb{P}(y_i = 1|v_i) = 1/(1+\exp(-\langle x, v_i\rangle))$$

for some parameter $x \in \mathbb{R}^d$. It is common to use LMC to generate samples from the posterior distribution $p(x|V,Y)$ with the following potential function

$$f(x) = -\log p(x) - \langle Y, Vx\rangle + \textstyle\sum_{i=1}^n \log(1+\exp(\langle x, v_i\rangle)),$$

where $p(x)$ is a prior on x. In practice, the prior distribution can be arbitrary whereas most theoretical results require the prior to be the Gaussian distribution in order to ensure that the posterior is smooth and has quadratic growth (Dalalyan, 2017b; Li et al., 2019). The framework in this paper allows for priors that have heavier tails than a Gaussian and/or have potentials that do not have Lipschitz gradients. For example, consider a pseudo-Huber prior $p(x) \propto \exp\left(-\sqrt{1+\|x\|^2}\right)$ (Gorham et al., 2019; Hartley and Zisserman, 2003), which results in a similar setting as in Example 2, in the sense that a careful choice of $\tilde{f}$ yields a convergence rate bound of $\widetilde{\mathcal{O}}\left(d^5\epsilon^{-1}\right)$.

Next, consider the prior $p(x) \propto \exp\big(\sum_{i=1}^{d} |x_i|^q\big)$ for $q \in (1,2)$ which is similar to the Bridge linear regression setting. The resulting potential function is not smooth in this case, and the potential lacks quadratic growth. Our analysis can be used to show that the LMC algorithm reaches ϵ-accuracy in KL-divergence after $\widetilde{\mathcal{O}}\big(d^{\frac{3-q}{q-1}}\epsilon^{-\frac{1}{q-1}}\big)$ steps.

5. Conclusion

In this paper, we analyzed the convergence of unadjusted LMC algorithm for a class of potentials whose tails behave like $\|x\|^\alpha$ for $\alpha \in [1,2]$, and have β-Hölder continuous gradients. This covers potentials that are weakly smooth, and can be written as finite perturbations of a function which is convex degenerate at ∞. To establish this, we proved a moment dependent modified log-Sobolev inequality for any order moment of the LMC. Further establishing a diverging moment estimate on the LMC iterates under α-dissipativity, we obtained a differential inequality which can be iterated to obtain our main convergence result after tuning the moment order. To demonstrate the applicability of our results, we showed that any convex potential have at least linear growth, and further we verified our main assumptions on a variety of sampling problems. The established bound on the convergence rate of LMC can be described as a function of the smoothness order and the tail growth rate in high dimensions. Our results hold only for the last iterate of the LMC algorithm; investigating the behavior of the subsequent iterates is an interesting direction left for another study.

Acknowledgments

This research is partially funded by NSERC Grant [2019-06167], Connaught New Researcher Award, and CIFAR AI Chairs program at the Vector Institute.

References

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

Dominique Bakry and Michel Émery. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985.

Dominique Bakry, Franck Barthe, Patrick Cattiaux, Arnaud Guillin, et al. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. 2013.

Lorenzo Bertini and Boguslaw Zegarlinski. Coercive inequalities for Gibbs measures. *Journal of functional analysis*, 162(2):257–286, 1999.

Kush Bhatia, Yi-An Ma, Anca D Dragan, Peter L Bartlett, and Michael I Jordan. Bayesian robustness: A nonasymptotic viewpoint. *arXiv preprint arXiv:1907.11826*, 2019.

Adrien Blanchet and Jérôme Bolte. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *Journal of Functional Analysis*, 275(7):1650–1673, 2018.

Sergey Bobkov and Mokshay Madiman. The entropy per coordinate of a random vector is highly constrained under convexity conditions. *IEEE Transactions on Information Theory*, 57(8):4940–4954, 2011.

S.G Bobkov and F Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

François Bolley and Cédric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pages 331–352, 2005.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Conference on Learning Theory*, pages 319–342. PMLR, 2017.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

EA Carlen and Avraham Soffer. Entropy production by block variable summation and central limit theorems. *Communications in mathematical physics*, 140(2):339–371, 1991.

Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

Niladri Chatterji, Jelena Diakonikolas, Michael I Jordan, and Peter Bartlett. Langevin Monte Carlo without smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 1716–1726. PMLR, 2020.

Xiang Cheng and Peter L Bartlett. Convergence of Langevin MCMC in KL-divergence. *PMLR 83*, (83): 186–211, 2018.

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

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018b.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a.

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

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

Arnak S Dalalyan, Lionel Riou-Durand, and Avetik Karagulyan. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *arXiv preprint arXiv:1906.08530*, 2019.

Anh Duc Doan, Xin Dang, and Dao Nguyen. Black-box sampling for weakly smooth Langevin Monte Carlo using p-generalized gaussian smoothing. *arXiv preprint arXiv:2002.10071*, 2020.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 06 2017.

Alain Durmus, Gareth O Roberts, Gilles Vilmart, Konstantinos C Zygalakis, et al. Fast Langevin based algorithm for MCMC in high dimensions. *The Annals of Applied Probability*, 27(4):2195–2237, 2017.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

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

Alain Durmus, Eric Moulines, et al. High-dimensional bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019b.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3-4):851–886, 2016.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018.

LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering Langevin Monte Carlo ii: An improved proof using soft Markov chain decomposition. *arXiv preprint arXiv:1812.00793*, 2018.

Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Ivan Gentil, Arnaud Guillin, and Laurent Miclo. Modified logarithmic Sobolev inequalities and transportation inequalities. *Probability theory and related fields*, 133(3):409–436, 2005.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, Lester Mackey, et al. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. 2003.

Ye He, Krishnakumar Balasubramanian, and Murat A Erdogdu. On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method. *Advances in Neural Information Processing Systems*, 33, 2020.

Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, 1987.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887, 2018.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Petar M. Vasić Jovan D. Kečkić. Some inequalities for the gamma function. *Publications de l'Institut Mathématique*, 11(25)(31):107–114, 1971.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic runge-kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems 32*, pages 7748–7760. 2019.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Tung Duy Luu, Jalal Fadili, and Christophe Chesneau. Sampling from non-smooth distributions through Langevin diffusion. *Methodology and Computing in Applied Probability*, 2020.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A complete recipe for stochastic gradient MCMC. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, page 2917–2925, 2015.

Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019a.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019b.

Mateusz B Majka, Aleksandar Mijatović, Łukasz Szpruch, et al. Nonasymptotic bounds for sampling algorithms without log-concavity. *Annals of Applied Probability*, 30(4):1534–1581, 2020.

P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*, pages 1–29, 1999.

Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2): 185–232, 2002.

Alex McNabb. Comparison theorems for differential equations. *Journal of mathematical analysis and applications*, 119(1-2):417–428, 1986.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. 2012.

Grigori Noah Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics*. 2013.

Grigorii Noikhovich Milstein. *Numerical integration of stochastic differential equations*, volume 313. 1994.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. An efficient sampling algorithm for non-smooth composite potentials. *arXiv preprint arXiv:1910.00551*, 2019a.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *arXiv preprint arXiv:1907.11331*, 2019b.

Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *arXiv preprint arXiv:1908.10859*, 2019c.

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

F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361 – 400, 2000.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703, 2017.

Hannes Risken. Fokker-Planck equation. In *The Fokker-Planck Equation*, pages 63–95. Springer, 1996.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2098–2109, 2019.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

Michel Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

G. Toscani and C. Villani. On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds. *Journal of Statistical Physics*, 98(5):1279–1309, 2000.

Giuseppe Toscani. Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation. *Quarterly of Applied Mathematics*, 57(3):521–541, 1999.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8092–8104, 2019.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference On Learning Theory*, pages 2093–3027, 2018.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. *arXiv preprint arXiv:2006.07904*, 2020.

Boguslaw Zegarlinski. Entropy bounds for Gibbs measures with non-Gaussian tails. *Journal of Functional Analysis*, 187(2):368–395, 2001.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022. PMLR, 2017.

Appendix A. Proof of Modified Log-Sobolev Inequality

We start with a lemma that allows us to construct a finite perturbation of the potential function that has polynomially decaying Hessian which is unbounded at 0. This will allow us to optimize the final bound.

Lemma 8 *Suppose Assumption 1 holds. Then for sufficiently small $\varepsilon > 0$, there exist a function $\tilde{f}_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\big\|f - \tilde{f}_\varepsilon\big\|_\infty \leq \xi + \varepsilon/2,$$

where $\tilde{f}_\varepsilon(x)$ satisfies,

$$\nabla^2 \tilde{f}_\varepsilon(x) \succeq m(\|x\|) I_d,$$

where $m : \mathbb{R}_+ \to \mathbb{R}_+$ is a monotonically decreasing and onto function satisfying

$$m(r) \geq \frac{\mu - \alpha_\theta \varepsilon}{(1 + r^2/4)^{\theta/2}},$$

where $\alpha_\theta < \infty$ is a constant depending only on θ.

Proof of Lemma 8. Let $\tilde{f}_\varepsilon(x) = \tilde{f}(x) + \varepsilon \|x\|^{3/2} e^{-\|x\|^2}$, and notice that f satisfies

$$\|f - \tilde{f}_\varepsilon\|_\infty \leq \xi + \varepsilon/2.$$

For its Hessian, we write

$$\begin{aligned}\nabla^2 \tilde{f}_\varepsilon(x) = \nabla^2 \tilde{f}(x) + \varepsilon e^{-\|x\|^2} \big\{ &\big(\tfrac{3}{2}\|x\|^{-1/2} - 2\|x\|^{3/2}\big) I_d \\ &- \big(6\|x\|^{-1/2} - 4\|x\|^{3/2} + \tfrac{3}{4}\|x\|^{-5/2}\big) xx^\top \big\},\end{aligned}$$

and by choosing $\alpha_\theta = 8 \sup_{r \geq 0} r^{1.5}(1 + r^2/4)^{\theta/2} e^{-r^2}$, we observe that

$$\nabla^2 \tilde{f}_\varepsilon(x) \succeq \frac{\mu - \alpha_\theta \varepsilon}{(1 + \|x\|^2/4)^{\theta/2}} I_d.$$

Also, $\nabla^2 \tilde{f}_\varepsilon(x) > \varepsilon \|x\|^{-1/2}/2$ when $x \leq 0.1$. Now by selecting

$$m(r) = (\mu - \alpha_\theta \varepsilon)(1 + r^2/4)^{-\theta/2} \vee \frac{\varepsilon}{2} r^{-1/2} \mathbb{1}_{\{r \leq 0.1\}},$$

the lemma follows. Note that $m : \mathbb{R}_+ \to \mathbb{R}$ is both monotone and onto for $\varepsilon < 1/(\alpha_\theta + 2)$. ■

Proof of Theorem 1. We follow a similar construction developed in Toscani and Villani (2000), and define the functions h and $\tilde{h}_\varepsilon$ as

$$\begin{aligned} h(x) &= f(x) + m(2r)\big(\|x\| - r\big)^2 \mathbb{1}_{\{\|x\| \geq r\}} + C_r \quad \text{and} \\ \tilde{h}_\varepsilon(x) &= \tilde{f}_\varepsilon(x) + m(2r)\big(\|x\| - r\big)^2 \mathbb{1}_{\{\|x\| \geq r\}} + C_r, \end{aligned} \tag{A.1}$$

where C_r is the normalizing constant for the unnormalized potential h satisfying

$$e^{C_r} = \int_{\|x\| < r} e^{-f(x)} dx + \int_{\|x\| \geq r} e^{-f(x)} e^{-m(2r)(\|x\| - r)^2} dx. \tag{A.2}$$

Using Assumption 1 and Lemma 8, we notice that h and $\tilde{h}_\varepsilon$ satisfy $\big|h(x) - \tilde{h}_\varepsilon(x)\big| \leq \xi + \varepsilon/2$, and also the growth of the function $\tilde{h}_\varepsilon$ can be characterized in the following three regions.

- For $\|x\| < r$, we have

$$\begin{aligned}\nabla^2 \tilde{h}_\varepsilon(x) = \nabla^2 \tilde{f}_\varepsilon(x) \succeq & m(\|x\|) I_d \\ \succeq & m(2r) I_d,\end{aligned}$$

  where in the last step we used the monotonicity of m.

- For $r \le \|x\| < 2r$, we have

$$\begin{aligned}\nabla^2 \tilde{h}_\varepsilon(x) =& \nabla^2 \tilde{f}_\varepsilon(x) + m(2r) \left\{ 2I_d - \frac{2r}{\|x\|} I_d + 2r \frac{xx^\top}{\|x\|^3} \right\} \\ \succeq & m(\|x\|) I_d + m(2r) \left\{ 2I_d - 2I_d + 0 \right\} \\ \succeq & m(2r) I_d,\end{aligned}$$

  where again the last step follows from the monotonicity of m.

- For $2r \le \|x\|$, we have

$$\begin{aligned}\nabla^2 \tilde{h}_\varepsilon(x) =& \nabla^2 \tilde{f}_\varepsilon(x) + m(2r) \left\{ 2I_d - \frac{2r}{\|x\|} I_d + 2r \frac{xx^\top}{\|x\|^3} \right\} \\ \succeq & 0 + m(2r) \left\{ 2I_d - I_d + 0 \right\} \\ \succeq & m(2r) I_d.\end{aligned}$$

In all three cases, we obtain that the function $\tilde{h}_\varepsilon$ has a positive definite Hessian which is lower bounded by $m(2r)$ which implies, by the Bakry-Emery's LSI result on strongly convex potentials (Bakry and Émery, 1985) that the distribution $e^{-\tilde{h}_\varepsilon}$ satisfies (1.3). Combining this with the Holley-Stroock perturbation lemma (Holley and Stroock, 1987), we obtain

$$\forall \rho, \;\; \mathrm{KL}\big(\rho | e^{-h}\big) \le \frac{e^{2\xi+\varepsilon}}{2m(2r)} \mathrm{I}\big(\rho | e^{-h}\big). \tag{A.3}$$

We will convert the above inequality on the perturbed potential h to an inequality on the potential function f. Using the definition in (A.1), we can obtain an upper bound on the relative entropy for all $r > 0$,

$$\begin{aligned}\mathrm{KL}\big(\rho | e^{-f}\big) =& \mathrm{KL}\big(\rho | e^{-h}\big) + \int \rho(x)(f(x) - h(x)) dx \\ =& \mathrm{KL}\big(\rho | e^{-h}\big) - m(2r) \int_{\|x\| \ge r} \rho(x) (\|x\| - r)^2 dx - C_r.\end{aligned} \tag{A.4}$$

For the normalizing constant C_r explicitly given in (A.2), one can obtain a lower bound using the Jensen's inequality,

$$\begin{aligned}C_r =& \log \int e^{-f(x)} e^{-m(2r)(\|x\|-r)^2 \mathbb{1}_{\{\|x\| \ge r\}}} dx \\ \ge & - m(2r) \int_{\|x\| \ge r} e^{-f(x)} (\|x\| - r)^2 dx.\end{aligned}$$

Combining this with (A.4), we obtain

$$\begin{aligned}\mathrm{KL}\big(\rho|e^{-f}\big) \leq& \mathrm{KL}\big(\rho|e^{-h}\big) + m(2r)\int_{\|x\|\geq r}\big(e^{-f(x)}-\rho(x)\big)(\|x\|-r)^2dx\\ \leq& \mathrm{KL}\big(\rho|e^{-h}\big) + m(2r)\int_{\|x\|\geq r} e^{-f(x)}\|x\|^2dx\\ \leq& \mathrm{KL}\big(\rho|e^{-h}\big) + m(2r)\tfrac{\mathrm{M}_s(\nu_*)}{(1+r^2)^{s/2-1}} \end{aligned} \tag{A.5}$$

where the second step follows since $\rho \geq 0$, and $\|x\|^2 \geq (\|x\|-r)^2$ in the domain of integration, and the last step follows from Lemma 9 below.

Lemma 9 *For a given distribution ρ and for a constant $r > 0$, we have*

$$\int_{\|x\|\geq r}\rho(x)\|x\|^2dx \leq \tfrac{\mathrm{M}_s(\rho)}{(1+r^2)^{s/2-1}}.$$

Proof of Lemma 9. For positive constants $p, q, s > 0$ satisfying $1/p + 1/q = 1$, we apply the Hölder's inequality and get

$$\begin{aligned}\int_{\|x\|\geq r}\rho(x)\|x\|^2dx =& \int \rho(x)\|x\|^2\mathbb{1}_{\{\|x\|\geq r\}}dx\\ \leq& \big(\int\rho(x)\|x\|^{2p}dx\big)^{1/p}\,\mathbb{P}\left(\big(1+\|x\|^2\big)^{s/2}\geq(1+r^2)^{s/2}\right)^{1/q}\\ \leq& \tfrac{\mathrm{M}_{2p}(\rho)^{1/p}\mathrm{M}_s(\rho)^{1/q}}{(1+r^2)^{s/2q}},\end{aligned}$$

where the last step follows from Markov's inequality. Final result follows by choosing $p = s/2$. ■

Similarly for the Fisher information, we write

$$\mathrm{I}\big(\rho|e^{-h}\big) \leq 2\mathrm{I}\big(\rho|e^{-f}\big) + 2\int\rho(x)\|\nabla h(x)-\nabla f(x)\|^2dx. \tag{A.6}$$

For the second term on the right hand side, we write

$$\begin{aligned}\int\rho(x)\|\nabla h(x)-\nabla f(x)\|^2dx =& 4m(2r)^2\int\rho(x)(\|x\|-r)^2\mathbb{1}_{\{\|x\|\geq r\}}dx\\ \leq& 4m(2r)^2\int\rho(x)\|x\|^2\mathbb{1}_{\{\|x\|\geq r\}}dx\\ \leq& \frac{4m(2r)^2}{(1+r^2)^{s/2-1}}\mathrm{M}_s(\rho),\end{aligned}$$

where in the last step we applied Lemma 9. Plugging this back in (A.6), we get

$$\mathrm{I}\big(\rho|e^{-h}\big) \leq 2\mathrm{I}\big(\rho|e^{-f}\big) + \tfrac{8m(2r)^2}{(1+r^2)^{s/2-1}}\mathrm{M}_s(\rho). \tag{A.7}$$

Combining the inequalities (A.3), (A.5), and (A.7), we obtain

$$\begin{aligned}\forall\rho,\ \ \mathrm{KL}\big(\rho|e^{-f}\big) \leq& \mathrm{KL}\big(\rho|e^{-h}\big) + \tfrac{m(2r)\mathrm{M}_s(\nu_*)}{(1+r^2)^{s/2-1}}\\ \leq& \tfrac{e^{2\xi+\varepsilon}}{2m(2r)}\mathrm{I}\big(\rho|e^{-h}\big) + \tfrac{m(2r)\mathrm{M}_s(\nu_*)}{(1+r^2)^{s/2-1}}\\ \leq& \tfrac{e^{2\xi+\varepsilon}}{m(2r)}\mathrm{I}\big(\rho|e^{-f}\big) + \tfrac{m(2r)}{(1+r^2)^{s/2-1}}\big(4e^{2\xi+\varepsilon}\mathrm{M}_s(\rho)+\mathrm{M}_s(\nu_*)\big)\\ \leq& \tfrac{e^{2\xi+\varepsilon}}{m(2r)}\mathrm{I}\big(\rho|e^{-f}\big) + 4e^{2\xi+\varepsilon}\tfrac{m(2r)}{(1+r^2)^{s/2-1}}\mathrm{M}_s(\rho+\nu_*).\end{aligned}$$

Finally, using the Lemma 28 and optimizing over $m(2r)$, we get

$$\forall \rho, \;\; \mathrm{KL}\big(\rho|e^{-f}\big) \leq \lambda_\varepsilon \mathrm{I}\big(\rho|e^{-f}\big)^{\frac{s-2+\theta}{s-2+2\theta}} \mathrm{M}_s(\rho+\nu_*)^{\frac{\theta}{s-2+2\theta}},$$

where

$$\lambda_\varepsilon = \frac{4e^{2\xi+\varepsilon}}{(\mu-\alpha_\theta\varepsilon)^{\frac{s-2}{s-2+2\theta}}},$$

for all sufficiently small $\varepsilon > 0$. Taking the limit of $\varepsilon \downarrow 0$ concludes the proof. ∎

Appendix B. Moment Bounds on the LMC Iterates

Proof of Proposition 2. Similar to the continuous-time case, it suffices to prove

$$\mathrm{M}_s(\rho_k) \leq \mathrm{M}_s(\rho_0) + C_s k\eta.$$

Part 1. We prove a linear bound on the second moment of $\widetilde{x}_{k,t}$ conditioned on x_k. Consider the distribution $\rho(\widetilde{x}_{k,t}|x_k)$ which is the distribution of $\widetilde{x}_{k,t}$ given x_k.

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{x}_{k,t}\|^2|x_k\right] &= \|x_k\|^2 - 2t\left\langle \nabla f(x_k), x_k\right\rangle + t^2\|\nabla f(x_k)\|^2 + 2dt \\
&\overset{1}{\leq} \|x_k\|^2 - 2t(a\|x_k\|^\alpha - b) + 2t^2M^2(1+\|x_k\|^{2\zeta}) + 2dt \\
&= \|x_k\|^2 + 2t\left(-a\left(1+\|x_k\|^\alpha\right) + \eta M^2\|x_k\|^{2\zeta} + a + b + d + \eta M^2\right) \\
&\overset{2}{\leq} \|x_k\|^2 + 2\left(a+b+d+\eta M^2\right)t \\
&\leq \|x_k\|^2 + C_2 t,
\end{aligned}$$

for any C_2 satisfying

$$C_2 \geq 3a + 2b + 2d. \tag{B.1}$$

Step 1 is obtained using Assumptions 2, and step 2 is because $4\eta M^2 \leq a$. Adding one to both sides, we get the following equation

$$\mathrm{M}_2(\widetilde{\rho}_{k,t}|x_k) \leq g_2(x_k) + C_2 t,$$

where $g_s(x) = \left(1+x^2\right)^{s/2}$ and $\mathrm{M}_s(\widetilde{\rho}_{k,t}|x_k)$ denotes the s-moment of $\widetilde{x}_{k,t}$ conditioned on x_k.

Part 2. We upper bound a term which will become useful in the proof of the induction step. (In below, Z denotes a standard Gaussian vector that is independent of x_k.)

$$\begin{aligned}
&\mathbb{E}[-\left\langle \nabla f(x_k), Z\right\rangle g_2(\widetilde{x}_{k,t})|x_k] \\
&\qquad \overset{1}{=} 2\sqrt{2t}\mathbb{E}\left[-\left\langle \nabla f(x_k), Z\right\rangle \left\langle x_k, Z\right\rangle + t\left\langle \nabla f(x_k), Z\right\rangle \left\langle \nabla f(x_k), Z\right\rangle |x_k\right] \\
&\qquad = 2\sqrt{2t}\left(-\left\langle \nabla f(x_k), x_k\right\rangle + t\|\nabla f(x_k)\|^2\right) \\
&\qquad \leq 2\sqrt{2\eta}\left(-a\|x_k\|^\alpha + b + 2\eta M^2(1+\|x_k\|^{2\zeta})\right) \\
&\qquad \leq 2\sqrt{2\eta}\left(-a\left(\|x_k\|^\alpha + 1\right) + 2\eta M^2\|x_k\|^{2\zeta} + a + b + 2\eta M^2\right) \\
&\qquad \overset{2}{\leq} 2\sqrt{2\eta}\left(a+b+2\eta M^2\right) \leq N_2,
\end{aligned}$$

where

$$N_2 \triangleq 2\sqrt{2\eta}\,(1.5a+b)\,. \tag{B.2}$$

Step 1 follows from odd Gaussian moments being zero, and step 2 uses $4\eta M^2 < a$. Note that Z is independent of x_k, with zero mean.

Part 3. Now we use induction to prove the linear bound for even moments of the conditional distribution. The base case ($s=2$) is already proved. Hence, we can assume $s \geq 4$ which implies $(s-4)$ is an even non-negative integer. For the proof to work, we need to strengthen the induction hypothesis for which part 2 in the proof will be useful. We will prove by induction that for all even s, we have the following

1. $\mathrm{M}_s(\widetilde{\rho}_{k,t}|x_k) \leq g_s(x_k) + C_s t$.
2. $\mathbb{E}\left[-\left\langle \nabla f(x_k), Z\right\rangle g_s(\widetilde{x}_{k,t})|x_k\right] \leq N_s$.

For the first inequality above, we will bound the time derivative of $\mathrm{M}_s(\widetilde{\rho}_{k,t}|x_k)$ as follows.

$$\begin{aligned}
&\frac{\partial}{\partial t}\mathrm{M}_s(\widetilde{\rho}_{k,t}|x_k)\\
&\quad= \mathbb{E}\left[-s\left\langle \nabla f(x_k), \widetilde{x}_{k,t}\right\rangle g_{s-2}(\widetilde{x}_{k,t}) + s(d+s-2)g_{s-2}(\widetilde{x}_{k,t}) - s(s-2)g_{s-4}(\widetilde{x}_{k,t})|x_k\right]\\
&\quad\leq s\mathbb{E}\left[\left(-\left\langle \nabla f(x_k), x_k - t\nabla f(x_k) + \sqrt{2t}Z\right\rangle + (d+s-2)\right) g_{s-2}(\widetilde{x}_{k,t})|x_k\right]\\
&\quad\leq s\left(-\left\langle \nabla f(x_k), x_k\right\rangle + t\|\nabla f(x_k)\|^2 + (d+s-2)\right)\mathbb{E}\left[g_{s-2}(\widetilde{x}_{k,t})|x_k\right]\\
&\qquad + s\sqrt{2t}\mathbb{E}\left[\left\langle -\nabla f(x_k), Z\right\rangle g_{s-2}(\widetilde{x}_{k,t})|x_k\right]\\
&\quad\leq s\left[-a\|x_k\|^\alpha + b + 2\eta M^2(1+\|x_k\|^{2\zeta}) + (d+s-2)\right]_+ (g_{s-2}(x_k) + C_{s-2}t)\\
&\qquad + s\sqrt{2\eta}N_{s-2}\\
&\quad\overset{1}{\leq} \max_{u\geq 1} s\left(-au^\alpha + 2\eta M^2 u^{2\zeta} + (2\eta M^2 + a + b + d + s - 2)\right)(u^{s-2} + C_{s-2}t)\\
&\qquad + s\sqrt{2\eta}N_{s-2}\\
&\quad\leq s\max_{u\geq 1}\left(-\frac{a}{2}u^{\alpha+s-2} + 2\eta M^2 u^{2\zeta+s-2}\right)\\
&\qquad + s\max_{u\geq 1}\left(-\frac{a}{2}u^{\alpha+s-2} + (2\eta M^2 + a + b + d + s - 2)u^{s-2}\right)\\
&\qquad + sC_{s-2}\eta\max_{u\geq 1}\left(-au^\alpha + 2\eta M^2 u^{2\zeta} + (2\eta M^2 + a + b + d + s - 2)\right) + s\sqrt{2\eta}N_{s-2}\\
&\quad\overset{2}{\leq} s\Bigg[(2\eta M^2 + a + b + d + s - 2)\left(\frac{2(2\eta M^2 + a + b + d + s - 2)(s-2)}{a(\alpha+s-2)}\right)^{\frac{s-2}{\alpha}}\\
&\qquad + C_{s-2}\eta(2\eta M^2 + a + b + d + s - 2) + \sqrt{2\eta}N_{s-2}\Bigg],
\end{aligned}$$

in which substitution $u = \sqrt{1+\|x_k\|^2}$ is used in step 1 and Lemma 30 is used in step 2. The above inequality shows $\mathrm{M}_s(\widetilde{\rho}_{k,t}|x_k) \leq \mathrm{M}_s(x_k) + C_s t$ for any C_s satisfying

$$\begin{aligned}
\frac{C_s}{s} \geq\ & (2\eta M^2 + a + b + d + s - 2)\left(\frac{2(2\eta M^2 + a + b + d + s - 2)}{a}\right)^{\frac{s-2}{\alpha}}\\
&+ C_{s-2}\eta(2\eta M^2 + a + b + d + s - 2) + \sqrt{2\eta}N_{s-2}.
\end{aligned} \tag{B.3}$$

For proving the second part of the induction step, we use Stein's lemma (Stein, 1981) (the version we use is stated in Lemma 31) in the first equality below.

$$
\begin{aligned}
&\mathbb{E}[-\langle \nabla f(x_k), Z\rangle g_s(\widetilde{x}_{k,t})|x_k] \\
&\quad= \mathbb{E}\left[-s\sqrt{2t}\left\langle \nabla f(x_k), g_{s-2}(\widetilde{x}_{k,t})\left(x_k - t\nabla f(x_k) + \sqrt{2t}Z\right)\right\rangle |x_k\right] \\
&\quad\le s\sqrt{2t}\left(-\langle \nabla f(x_k), x_k\rangle + t\|\nabla f(x_k)\|^2\right)\mathbf{M}_{s-2}(\widetilde{\rho}_{k,t}|x_k) + 2stN_{s-2} \\
&\quad\le s\sqrt{2\eta}\left[-a\|x_k\|^\alpha + b + 2\eta M^2(1+\|x_k\|^{2\zeta})\right]_+ (g_{s-2}(x_k) + C_{s-2}\eta) + 2s\eta N_{s-2} \\
&\quad\le s\sqrt{2\eta}\max_{u\ge 1}\left(-au^\alpha + 2\eta M^2 u^{2\zeta} + (b+a+2\eta M^2)\right)(u^{s-2} + C_{s-2}\eta) + 2s\eta N_{s-2} \\
&\quad\le s\sqrt{2\eta}\max_{u\ge 1}\left(-\frac{a}{2}u^{\alpha+s-2} + 2\eta M^2 u^{2\zeta+s-2}\right) \\
&\qquad + s\sqrt{2\eta}\max_{u\ge 1}\left(-\frac{a}{2}u^{\alpha+s-2} + (b+a+2\eta M^2)u^{s-2}\right) \\
&\qquad + s\sqrt{2\eta}C_{s-2}\eta\max_{u}\left(-au^\alpha + 2\eta M^2 u^{2\zeta} + (b+a+2\eta M^2)\right) + 2s\eta N_{s-2} \\
&\quad\le s\Bigg[(b+a+2\eta M^2)\sqrt{2\eta}\left(\frac{2(b+a+2\eta M^2)(s-2)}{a(\alpha+s-2)}\right)^{\frac{s-2}{\alpha}} \\
&\qquad + C_{s-2}\eta\sqrt{2\eta}(b+a+2\eta M^2) + 2\eta N_{s-2}\Bigg] \le N_s,
\end{aligned}
$$

where

$$
\frac{N_s}{s} = (b+a+2\eta M^2)\sqrt{2\eta}\left(\frac{2(b+a+2\eta M^2)}{a}\right)^{\frac{s-2}{\alpha}} + C_{s-2}\eta\sqrt{2\eta}(b+a+2\eta M^2) + 2\eta N_{s-2}. \tag{B.4}
$$

Again, the substitution $u = \sqrt{1+\|x\|^2}$ is used here. This completes the induction.

The previous induction showed $\mathbf{M}_s(\widetilde{\rho}_{k,t}|x_k) \le g_s(x_k) + C_s t$ when s is a positive even integer. We take expectation with respect to x_k in order to get

$$
\mathbf{M}_s(\widetilde{\rho}_{k,t}) \le \mathbf{M}_s(\rho_k) + C_s t,
$$

setting $t = \eta$ yields

$$
\mathbf{M}_s(\rho_{k+1}) \le \mathbf{M}_s(\rho_k) + C_s\eta,
$$

and finally, induction on k gives

$$
\mathbf{M}_s(\rho_k) \le \mathbf{M}_s(\rho_0) + C_s k\eta.
$$

Part 4. In this part, we establish a non-recursive formula for C_s. Note that the theorem holds for a larger C_s, this helps us to derive a closed-form formula for C_s. We combine (B.1) and (B.2) to get $N_2 \le C_2\sqrt{2\eta}$, then we use (B.3) and (B.4) inductively, to establish $N_s \le C_s\sqrt{2\eta}$. By combining the previous inequality with $4\eta M^2 \le a$, we can strengthen the bound (B.3) to

$$
C_s \ge \left(\frac{3a+2b+2d+2s}{1\wedge a}\right)^{\frac{s-2}{\alpha}+1} s + \frac{3a+2b+2d+2s}{1\wedge a} \times C_{s-2}s\eta.
$$

C_s, as defined in (2.8), satisfies the previous inequality and (B.1), which in turn implies that it also satisfies (B.3) and (B.1). ■

The next proposition is the analog moment bound for the continuous-time process, and is adapted from Toscani and Villani (2000) for the sake of comparison with the bound for the discrete time process.

Lemma 10 *Let f satisfy Assumption 2 and p_t be the distribution of Z_t, then*

$$\mathrm{M}_s(p_t) \leq \mathrm{M}_s(p_0) + K_s t,$$

where $K_s = (b+d+a+s-2)\left(\frac{b+d+a+s-2}{a}\right)^{\frac{s-2}{\alpha}} s$.

Proof If $s < s'$ then $\mathrm{M}_s(p_t) = \int p_t(x)(1+\|x\|^2)^{\frac{s}{2}} \leq \int p_t(x)(1+\|x\|^2)^{\frac{s'}{2}} = \mathrm{M}_{s'}(p_t)$. We differentiate $\mathrm{M}_s(p_t)$ with respect to time.

$$\begin{aligned}
\frac{d}{dt}\mathrm{M}_s(p_t) &= \int p_t(x)\left[\Delta(1+\|x\|^2)^{\frac{s}{2}} - \left\langle \nabla f(x), \nabla(1+\|x\|^2)^{\frac{s}{2}}\right\rangle\right] \\
&= (ds + s(s-2))\,\mathrm{M}_{s-2}(p_t) - s(s-2)\mathrm{M}_{s-4}(p_t) \\
&\quad - s\int p_t(x)\,\langle \nabla f(x), x\rangle\,(1+\|x\|^2)^{\frac{s-2}{2}} \\
&\leq (b+d+s-2)s\mathrm{M}_{s-2}(p_t) - s\int p_t(x) a\|x\|^{\alpha}(1+\|x\|^2)^{\frac{s-2}{2}} \\
&\leq (b+d+a+s-2)s\mathrm{M}_{s-2}(p_t) - \frac{as}{2}\mathrm{M}_{s+\alpha-2}(p_t) \\
&\leq (b+d+a+s-2)s\mathrm{M}_{s+\alpha-2}(p_t)^{\frac{s-2}{s+\alpha-2}} - \frac{as}{2}\mathrm{M}_{s+\alpha-2}(p_t) \\
&\overset{1}{\leq} (b+d+a+s-2)s\left(\frac{2(b+d+a+s-2)(s-2)}{a(s+\alpha-2)}\right)^{\frac{s-2}{\alpha}} \\
&\leq (b+d+a+s-2)\left(\frac{b+d+a+s-2}{a/2}\right)^{\frac{s-2}{\alpha}} s,
\end{aligned}$$

where step 1 follows from Lemma 30. ■

Appendix C. Proof of The Main Theorem

The proof will be done in three parts. In the first part, we bound the α-th moment of a given distribution with its KL-divergence from the ν_*. In the second part, the bound derived in the first part will be used to construct a differential inequality on the interpolation diffusion. Next, using a comparison theorem on the differential inequality, we will derive a single step bound on the LMC iterates. Finally, in the last part, we will iterate the single step bound to obtain a non-asymptotic convergence rate.

C.1. Bounding LMC moments with KL-divergence

The behavior of the discrete-time process is different from that of the continuous-time diffusion in that, a step size dependent bias term appears in the differential inequality that governs its evolution. The results in this section will help us handle the bias term. First, using Assumption 2, we prove that the potential grows at least like $\|x\|^\alpha$ in Lemma 11. Using this growth, we bound the α-th exponential moment of the target ν_* in Lemma 12. Finally, using the exponential moment bound, in Lemma 13, we upper bound the α-th moment of a given distribution with its KL-divergence from ν_*. Although this step can be handled easily by Talagrand's inequality in the case of $\alpha = 2$, it is more challenging for $\alpha \in [1, 2)$.

Lemma 11 *If f satisfies Assumption 2, then*

$$f(x) \geq \frac{a}{2\alpha}\|x\|^\alpha + f(0) - M\left(\frac{2a+2b}{a}\right)^2 - b.$$

Proof For notational ease, let $R = \left(\frac{2b}{a}\right)^{\frac{1}{\alpha}}$. First, using the gradient growth condition in Assumption 2, we upper bound $\|\nabla f(x)\|$ when $x \leq R$.

$$\|\nabla f(x)\| \leq \max_{\|x\|\leq R} M(1+\|x\|^\zeta) \leq M\left(1 + \left(\frac{2b}{a}\right)^{\zeta/\alpha}\right) \leq \frac{M(2a+2b)}{a}.$$

Now using Assumption 2 we lower bound f.

$$\begin{aligned}
f(x) &= f(0) + \int_0^{\frac{R}{\|x\|}} \langle \nabla f(tx), x\rangle \, dt + \int_{\frac{R}{\|x\|}}^1 \langle \nabla f(tx), x\rangle \, dt \\
&\geq f(0) - \int_0^{\frac{R}{\|x\|}} \|\nabla f(tx)\|\|x\| dt + \int_{\frac{R}{\|x\|}}^1 \frac{1}{t}\langle \nabla f(tx), tx\rangle \, dt \\
&\geq f(0) - \left(\frac{M(2a+2b)}{a}\right) R + \int_{\frac{R}{\|x\|}}^1 \frac{1}{t}\left(a\|tx\|^\alpha - b\right) dt \\
&\overset{1}{\geq} f(0) - M\left(\frac{2a+2b}{a}\right)^2 + \frac{a}{2}\|x\|^\alpha \int_{\frac{R}{\|x\|}}^1 t^{\alpha-1} dt \\
&\geq f(0) - M\left(\frac{2a+2b}{a}\right)^2 + \frac{a}{2\alpha}\|x\|^\alpha \left(1 - \frac{R^\alpha}{\|x\|^\alpha}\right) \\
&\geq \frac{a}{2\alpha}\|x\|^\alpha + f(0) - M\left(\frac{2a+2b}{a}\right)^2 - b.
\end{aligned}$$

where step 1 uses the fact that if $t \geq \frac{R}{\|x\|}$ then $a\|tx\|^\alpha - b \geq \frac{a}{2}\|tx\|^\alpha$. ■

We use Lemma 11 to prove that the α-th exponential moment of the target ν_* is bounded.

Lemma 12 *If f satisfies Assumption 2, then*

$$0 < \log\left(\int e^{\frac{a}{4\alpha}\|x\|^\alpha - f(x)}\right) \leq \tilde{d}\tilde{\mu},$$

where,

$$\begin{cases} \tilde{\mu} &= \log\left(\frac{16\pi}{a}\right) + M\left(\frac{2a+2b}{a}\right)^2 + b + |f(0)|\,, \\ \tilde{d} &= d\left(1 + (1-\alpha/2)\log(d)\right). \end{cases} \tag{C.1}$$

Proof Using Lemma 11 we get

$$\begin{aligned} \int e^{\frac{a}{4\alpha}\|x\|^\alpha - f(x)}dx &\leq e^{-f(0)+M\left(\frac{2a+2b}{a}\right)^2+b}\int e^{-\frac{a}{4\alpha}\|x\|^\alpha}dx \\ &= \frac{2\pi^{d/2}}{\alpha}\left(\frac{4\alpha}{a}\right)^{d/\alpha} e^{-f(0)+M\left(\frac{2a+2b}{a}\right)^2+b}\frac{\Gamma(d/\alpha)}{\Gamma(d/2)}. \end{aligned}$$

Next, using an inequality for the ratio of Gamma functions (Jovan D. Kečkić, 1971), we obtain

$$\frac{\Gamma(d/\alpha)}{\Gamma(d/2)} \leq \frac{(d/\alpha)^{\frac{d}{\alpha}-\frac{1}{2}}}{(d/2)^{\frac{d}{2}-\frac{1}{2}}} e^{\frac{d}{2}-\frac{d}{\alpha}}.$$

By plugging this back into the previous bound and taking logs, we obtain

$$\begin{aligned} \log\left(\int e^{\frac{a}{4\alpha}\|x\|^\alpha - f(x)}dx\right) \leq& \tfrac{d}{2}\log(\pi) + \tfrac{d}{\alpha}\log\left(\tfrac{4\alpha}{a}\right) + \left(\tfrac{d}{\alpha}-\tfrac{d}{2}\right)\log\left(\tfrac{d}{2e}\right) \\ &+ \left(\tfrac{d}{\alpha}+\tfrac{1}{2}\right)\log\left(\tfrac{2}{\alpha}\right) + M\left(\tfrac{2a+2b}{a}\right)^2 + b + |f(0)| \\ \leq& \tfrac{d}{\alpha}\left(\log\left(\tfrac{16\pi}{a}\right) + \left(1-\tfrac{\alpha}{2}\right)\log\left(\tfrac{d}{2e}\right)\right) + M\left(\tfrac{2a+2b}{a}\right)^2 + b + |f(0)| \\ \leq& \tilde{d}\tilde{\mu}. \end{aligned}$$

■

Finally, using the previous lemma, we will bound the α-th moment of any distribution ρ using its KL-divergence from the target ν_*.

Lemma 13 *If the potential f satisfies Assumption 2, then for $\nu_* = e^{-f}$ and any distribution ρ, we have*

$$\frac{4\alpha}{a}\left[\mathrm{KL}\big(\rho|\nu_*\big) + \tilde{d}\tilde{\mu}\right] \geq \mathbb{E}_\rho\left[\|x\|^\alpha\right].$$

Proof Let $q(x) = e^{\frac{a}{4\alpha}\|x\|^\alpha - f(x)}$. Let z be number such that $q(x)/z$ be a probability distribution. Lemma 12 implies $\log z \leq \tilde{d}\tilde{\mu}$. Using this bound on z we get

$$\mathrm{KL}\big(\rho|\nu_*\big) = \int \rho\log\frac{\rho}{q/z} + \int \rho\log\frac{q/z}{\nu_*} = \mathrm{KL}\big(\rho|q/z\big) + \mathbb{E}_\rho\left[\log\frac{q/z}{e^{-f}}\right] \geq \frac{a}{4\alpha}\mathbb{E}_\rho\left[\|x\|^\alpha\right] - \tilde{d}\tilde{\mu}.$$

Rearranging this yields the desired inequality. ■

C.2. Single step bound

The proof strategy is to consider the continuous-time interpolation of a single LMC iteration

$$d\widetilde{x}_{k,t} = -\nabla f(x_k)dt + \sqrt{2}dB_t \quad \text{with} \quad \widetilde{x}_{k,0} = x_k, \tag{C.2}$$

where x_k is the k-th iterate of the LMC algorithm (1.2). Denoting the distributions of x_k and $\widetilde{x}_{k,t}$ with ρ_k and $\widetilde{\rho}_{k,t}$, respectively, we notice that $\widetilde{\rho}_{k,0} = \rho_k$ and $\widetilde{x}_{k,\eta} \sim \rho_{k+1}$. In the following, we construct a differential inequality for the KL-divergence between $\widetilde{\rho}_{k,t}$ and the target. This inequality will be used together with the modified log-Sobolev inequality of Theorem 1 and the linear moment bounds of Proposition 2 to obtain a single step bound.

The time derivative of the KL-divergence between $\widetilde{\rho}_{k,t}$ and the target ν_* has an additional bias term compared to the diffusion process (1.1). The next lemma characterizes this bias and is adapted from Vempala and Wibisono (2019).

Lemma 14 (Vempala and Wibisono (2019)) *Suppose $\widetilde{x}_{k,t}$ is the interpolation of the discretized process (C.2). Let $\widetilde{\rho}_{k,t}$ denote its distribution. Then*

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) &= -\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\left\langle \nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k), \nabla \log\left(\frac{\widetilde{\rho}_{k,t}(\widetilde{x}_{k,t})}{\nu_*(\widetilde{x}_{k,t})}\right)\right\rangle\right] \\ &\le -\frac{3}{4}\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\right].\end{aligned} \tag{C.3}$$

Proof The following proof is included for reader's convenience. For further notational convenience, we denote with $\widetilde{\rho}_{k|t}$ and $\widetilde{\rho}_{t|k}$, distributions of x_k conditioned on $\widetilde{x}_{k,t}$, and $\widetilde{x}_{k,t}$ conditioned on x_k, respectively. The distribution $\widetilde{\rho}_{t|k}(x)$ evolves by the following Fokker-Planck equation.

$$\frac{\partial \widetilde{\rho}_{t|k}(x)}{\partial t} = \nabla \cdot \big(\widetilde{\rho}_{t|k}(x)\nabla f(x_k)\big) + \Delta \widetilde{\rho}_{t|k}(x).$$

Taking expectation with respect to x_k yields

$$\begin{aligned}\frac{\partial \widetilde{\rho}_{k,t}(x)}{\partial t} &= \nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\int \widetilde{\rho}_{k|t}(x_k)\nabla f(x_k)dx_k\right) + \Delta\widetilde{\rho}_{k,t}(x) \\ &= \nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\mathbb{E}\left[\nabla f(x_k)|\widetilde{x}_{k,t} = x\right]\right) + \Delta\widetilde{\rho}_{k,t}(x).\end{aligned}$$

This equality is combined with the time derivative of KL-divergence to prove the claim.

$$
\begin{aligned}
\frac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) &= \int \frac{\partial \widetilde{\rho}_{k,t}}{\partial t}(x) \log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right) dx \\
&= \int \left(\nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\mathbb{E}\left[\nabla f(x_k)|\widetilde{x}_{k,t}=x\right]\right) + \Delta \widetilde{\rho}_{k,t}(x)\right) \log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right) dx \\
&\overset{1}{=} \int \nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\left(\mathbb{E}\left[\nabla f(x_k) - \nabla f(x)|\widetilde{x}_{k,t}=x\right] + \nabla \log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)\right)\right) \log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right) dx \\
&\overset{2}{=} -\int \widetilde{\rho}_{k,t}(x)\left\langle \mathbb{E}\left[\nabla f(x_k) - \nabla f(x)|\widetilde{x}_{k,t}=x\right] + \nabla \log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right), \nabla \log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)\right\rangle dx \\
&= -\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\left\langle \nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k), \nabla \log\left(\frac{\widetilde{\rho}_{k,t}(\widetilde{x}_{k,t})}{\nu_*(\widetilde{x}_{k,t})}\right)\right\rangle\right] \\
&\overset{3}{\leq} -\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - f(x_k)\|^2\right] + \frac{1}{4}\mathbb{E}\left[\|\nabla \log\left(\frac{\widetilde{\rho}_{k,t}(\widetilde{x}_{k,t})}{\nu_*(\widetilde{x}_{k,t})}\right)\|^2\right] \\
&= -\frac{3}{4}\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - f(x_k)\|^2\right],
\end{aligned}
$$

in which equality 1 follows from $\Delta \widetilde{\rho}_{k,t} = \nabla \cdot (\nabla \widetilde{\rho}_{k,t})$, equality 2 follows from the divergence theorem and inequality 3 follows from $\langle u, v\rangle \leq \|u\|^2 + \frac{1}{4}\|v\|^2$. ■

Next, using Lemma 14, we bound the time derivative of the KL-divergence $\frac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big)$, and obtain a useful differential inequality.

Lemma 15 *If the potential f satisfies Assumptions 1, 2 and 3, then*

$$
\begin{aligned}
\frac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) \leq -\frac{3}{4}\lambda^{-\frac{1}{1-\delta}}\left(\mathrm{M}_s(\rho_0+\nu_*) + C_s(k+1)\eta\right)^{-\frac{\delta}{1-\delta}}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big)^{\frac{1}{1-\delta}} \\
+ \frac{16\alpha L^2 M^{2\beta}}{a}\mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta} + 4L^2\left(1 + M^{2\beta}\left(1 + \frac{2\alpha\tilde{\mu}}{a}\right)\right)\tilde{d}\eta^{\beta},
\end{aligned} \tag{C.4}
$$

when $t \leq \eta \leq \frac{1}{2}\left(1 \wedge \frac{a}{2M^2}\right)$. The constants $\tilde{d}$ and $\tilde{\mu}$ are defined in (C.1).

Proof We bound $\mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\right]$ using Assumption 3

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\right] \\
&\qquad \leq L^2 \mathbb{E}\left[\|-t\nabla f(x_k) + \sqrt{2t}Z\|^{2\beta}\right] \\
&\qquad \overset{1}{\leq} 2L^2 t^{2\beta} \mathbb{E}\left[\|\nabla f(x_k)\|^{2\beta}\right] + 4L^2 t^{\beta} \mathbb{E}\left[\|Z\|^{2\beta}\right] \\
&\qquad \overset{2}{\leq} 2L^2 t^{2\beta} \mathbb{E}\left[\left(2M^2(1 + \|x_k\|^{2\zeta})\right)^{\beta}\right] + 4L^2 t^{\beta} \mathbb{E}\left[\|Z\|^2\right]^{\beta} \\
&\qquad \leq 4t^{2\beta} L^2 M^{2\beta} \mathbb{E}\left[1 + \|x_k\|^{2\beta\zeta}\right] + 4L^2 d^{\beta} t^{\beta} \\
&\qquad \overset{3}{\leq} 4t^{2\beta} L^2 M^{2\beta} \mathbb{E}\left[2 + \|x_k\|^{\alpha}\right] + 4L^2 d^{\beta} t^{\beta} \\
&\qquad \overset{4}{\leq} \frac{16\alpha L^2 M^{2\beta}}{a} \mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta} + 4\eta^{\beta} L^2 \left(d^{\beta} + 2\left(\eta M^2\right)^{\beta}\left(1 + \frac{2\alpha\tilde{\mu}\tilde{d}}{a}\right)\right) \\
&\qquad \leq \frac{16\alpha L^2 M^{2\beta}}{a} \mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta} + 4\tilde{d}L^2 \left(1 + 2a^{\beta}\left(1 + \frac{2\alpha\tilde{\mu}}{a}\right)\right)\eta^{\beta},
\end{aligned}$$

where step 1 follows from Lemma 29, step 2 from Assumption 2, step 3 from the fact that $2\zeta\beta \leq \alpha$, and step 4 from Lemma 13 and $\eta < 1$. Plugging the above inequality back in (C.3) and using Theorem 1 and Proposition 2 results in (C.4). ■

Finally, using a differential comparison argument, a single step bound is obtained on the KL-divergence of steps of LMC (1.2) from the target.

Lemma 16 *Suppose f satisfies Assumptions 1, 2 and 3, then*

$$\begin{aligned}
\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq &\mathrm{KL}\big(\rho_k|\nu_*\big)\left[1 - \frac{3\eta}{8\lambda^{\frac{1}{1-\delta}}}\left(\frac{\mathrm{KL}\big(\rho_k|\nu_*\big)}{\mathrm{M}_s(\rho_0+\nu_*) + C_s(k+1)}\right)^{\frac{\delta}{1-\delta}} + \frac{16\alpha L^2 M^{2\beta}\eta^{2\beta+1}}{a}\right] \\
&+ \sigma\tilde{d}\eta^{\beta+1},
\end{aligned} \tag{C.5}$$

where $\sigma = 4L^2\left(1 + 2a^{\beta}\left(1 + \frac{2\alpha\tilde{\mu}}{a}\right)\right)$. The step size needs to be sufficiently small, satisfying

$$\eta \leq \frac{1}{2}\left(1 \wedge \frac{a}{2M^2}\right) \wedge \left(\frac{4\lambda^{\frac{1}{1-\delta}}}{3}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*) + C_s(k+1)\eta}{\mathrm{KL}\big(\rho_k|\nu_*\big)}\right)^{\frac{\delta}{1-\delta}}\right).$$

Proof Let

$$\begin{aligned}
\kappa_1 &= \frac{3}{4}\lambda^{-\frac{1}{1-\delta}}\left(\mathrm{M}_s(\rho_0+\nu_*) + C_s(k+1)\eta\right)^{-\frac{\delta}{1-\delta}}, \\
\kappa_2 &= \frac{16\alpha L^2 M^{2\beta}}{a}\mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta} + \sigma\tilde{d}\eta^{\beta}, \\
\psi(t,x) &= -\kappa_1 x^{\frac{1}{1-\delta}} + \kappa_2,
\end{aligned}$$

where κ_1 and κ_2 are constants independent of t. We can rewrite (C.4) as

$$\frac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) \leq \psi\left(t, \mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big)\right).$$

For positive and sufficiently small $\tilde{\varepsilon}$ (less than $\mathrm{KL}\big(\rho_k|\nu_*\big)^{-\frac{\delta}{1-\delta}}$), consider the function

$$h_{\tilde{\varepsilon}}(t) = \left(\mathrm{KL}\big(\rho_k|\nu_*\big)^{-\frac{\delta}{1-\delta}} + \kappa_1\frac{\delta}{1-\delta}t - \tilde{\varepsilon}\right)^{-\frac{1-\delta}{\delta}} + \kappa_2 t.$$

We will use the following basic comparison lemma for differential inequalities; see, for example McNabb (1986) for a simple proof.

Lemma 17 *Suppose $u(t)$ and $v(t)$ are continuous on interval $[a,b]$ and differentiable on $(a,b]$, $f:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is a continuous mapping and*

$$u(a) < v(a) \qquad \textit{and} \qquad \frac{du}{dt} - f(t,u) < \frac{dv}{dt} - f(t,v), \qquad \textit{on} \quad (a,b\,].$$

Then $u<v$ on $[a,b]$.

For positive t, we have

$$\frac{d}{dt}h_{\tilde{\varepsilon}}(t) - \psi(t, h_{\tilde{\varepsilon}}(t)) > 0 \geq \frac{d}{dt}\mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big) - \psi\left(t, \mathrm{KL}\big(\widetilde{\rho}_{k,t}|\nu_*\big)\right).$$

Since $h_{\tilde{\varepsilon}}(0) > \mathrm{KL}\big(\widetilde{\rho}_{k,0}|\nu_*\big)$, the previous comparison lemma implies

$$h_{\tilde{\varepsilon}}(\eta) > \mathrm{KL}\big(\widetilde{\rho}_{k,\eta}|\nu_*\big) = \mathrm{KL}\big(\rho_{k+1}|\nu_*\big).$$

Taking the limit of $\tilde{\varepsilon}\downarrow 0$ gives

$$\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \left(\mathrm{KL}\big(\rho_k|\nu_*\big)^{-\frac{\delta}{1-\delta}} + \kappa_1\frac{\delta}{1-\delta}\eta\right)^{-\frac{1-\delta}{\delta}} + \kappa_2\eta.$$

Plugging the values for κ_1 and κ_2 back in the previous inequality reads

$$\begin{aligned}\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \left(\mathrm{KL}\big(\rho_k|\nu_*\big)^{-\frac{\delta}{1-\delta}} + \frac{3\lambda^{-\frac{1}{1-\delta}}\delta}{4(1-\delta)}\left(\mathrm{M}_s(\rho_0+\nu_*) + C_s(k+1)\eta\right)^{-\frac{\delta}{1-\delta}}\eta\right)^{-\frac{1-\delta}{\delta}} \\ + \frac{16\alpha L^2M^{2\beta}}{a}\mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta+1} + \sigma\tilde{d}\eta^{\beta+1}.\end{aligned}$$

We rewrite the previous inequality to get

$$\begin{aligned}\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \frac{16\alpha L^2M^{2\beta}}{a}\mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta+1} + \sigma\tilde{d}\eta^{\beta+1} \\ + \frac{\mathrm{KL}\big(\rho_k|\nu_*\big)}{\left(1+\frac{3\lambda^{-\frac{1}{1-\delta}}\delta}{4(1-\delta)}\left(\frac{\mathrm{KL}\left(\rho_k|\nu_*\right)}{\mathrm{M}_s(\rho_0+\nu_*)+C_s(k+1)\eta}\right)^{\frac{\delta}{1-\delta}}\eta\right)^{\frac{1-\delta}{\delta}}}.\end{aligned}$$

Using the fact that $(1+x)^{\frac{1-\delta}{\delta}} \geq 1 + \frac{1-\delta}{\delta}x$, in the denominator, yields

$$\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \frac{\mathrm{KL}\big(\rho_k|\nu_*\big)}{1+\frac{3}{4\lambda^{\frac{1}{1-\delta}}}\left(\frac{\mathrm{KL}\big(\rho_k|\nu_*\big)}{\mathrm{M}_s(\rho_0+\nu_*)+C_s(k+1)\eta}\right)^{\frac{\delta}{1-\delta}}\eta} + \frac{16\alpha L^2 M^{2\beta}}{a}\mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta+1} + \sigma\tilde{d}\eta^{\beta+1}.$$

Since $\frac{1}{1+x} < 1 - \frac{x}{2}$, when $x \leq 1$, and $\frac{3}{4\lambda^{\frac{1}{1-\delta}}}\left(\frac{\mathrm{KL}\big(\rho_k|\nu_*\big)}{\mathrm{M}_s(\rho_0+\nu_*)+C_s(k+1)\eta}\right)^{\frac{\delta}{1-\delta}}\eta < 1$, we have

$$\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \mathrm{KL}\big(\rho_k|\nu_*\big)\left(1 - \frac{3}{8\lambda^{\frac{1}{1-\delta}}}\left(\frac{\mathrm{KL}\big(\rho_k|\nu_*\big)}{\mathrm{M}_s(\rho_0+\nu_*)+C_s(k+1)\eta}\right)^{\frac{\delta}{1-\delta}}\eta\right) + \frac{16\alpha L^2 M^{2\beta}}{a}\mathrm{KL}\big(\rho_k|\nu_*\big)\eta^{2\beta+1} + \sigma\tilde{d}\eta^{\beta+1}.$$

Rearranging the above inequality yields the desired result. ■

C.3. Proof of the main theorem

In this section, we prove the convergence of the LMC algorithm by iterating the single step bound, obtained in the previous section. More specifically, we establish that the algorithm reaches the desired accuracy ϵ after N steps, for which our argument relies on two steps. In the first step, we prove that if an iterate of LMC reaches the desired accuracy before N steps, then it will remain below that accuracy level until the step N. In the second step, we show that if LMC does not reach ϵ accuracy before N steps, it is guaranteed to reach that accuracy at the step N. Since the single step bound obtained in Lemma 16 is quite convoluted, we first simplify it to a manageable recursive formula, and iterate the resulting expression. Special care is taken to determine the upper bound on the accuracy for the aforementioned claims to hold. The bound on ϵ is independent of the moment order s, which is crucial for tuning this parameter to obtain the final bound on the convergence rate leading to the main corollary.

Proof of Theorem 3. We simplify the recurrence relation for the single step bound in (C.5). For notational convenience, let $A = \frac{\lambda^{-1/(1-\delta)}}{16}\left(\frac{\sigma\tilde{d}}{\mathrm{M}_s(\rho_0+\nu_*)\vee C_s}\right)^{\delta/(1-\delta)}$. We remind that $\tilde{d}$ is defined as $\tilde{d} = d\,(1 + (1-\alpha/2)\log(d))$. We will show that under the conditions and notations of Lemma 16, if $k < N$ and $\mathrm{KL}\big(\rho_k|\nu_*\big) \geq \epsilon/2$, then

$$\mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \left(1 - \frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}\right)\mathrm{KL}\big(\rho_k|\nu_*\big) + \sigma\tilde{d}\eta^{\beta+1}. \tag{C.6}$$

The above expression depends on the choice of step size η and number of steps N; thus, given (C.5), we verify the inequality (C.6) for

$$\eta^{-1} = (\sigma\tilde{d})^{\frac{1}{\beta}}(16\lambda)^{\frac{1}{\beta(1-2\delta)}}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*)\vee C_s}{16}\right)^{\frac{\delta}{\beta(1-2\delta)}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\frac{\delta}{\beta(1-2\delta)}}\left(\frac{2}{\epsilon}\right)^{\frac{1-\delta}{\beta(1-2\delta)}},$$
$$N = (\sigma\tilde{d})^{\frac{1}{\beta}}(16\lambda)^{\frac{1+\beta}{\beta(1-2\delta)}}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*)\vee C_s}{16}\right)^{\frac{(1+\beta)\delta}{\beta(1-2\delta)}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{1+\frac{(\beta+1)\delta}{\beta(1-2\delta)}}\left(\frac{2}{\epsilon}\right)^{\frac{1-\delta(1-\beta)}{\beta(1-2\delta)}}. \tag{C.7}$$

For the above choices of η and N, using (C.5) together with the fact that $k < N$ and $\mathrm{KL}\big(\rho_k|\nu_*\big) \geq \frac{\epsilon}{2}$, in order for (C.6) to hold, it suffices to prove the following inequality

$$\frac{3\lambda^{-1/(1-\delta)}}{8}\left(\frac{\epsilon/2}{\mathrm{M}_s(\rho_0+\nu_*)+C_s(N+1)\eta}\right)^{\delta/(1-\delta)}\eta - \frac{16\alpha L^2M^{2\beta}}{a}\eta^{2\beta+1} \geq \frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}.$$

We will prove this inequality by showing that the following two inequalities hold,

$$\begin{cases} \frac{3\lambda^{-1/(1-\delta)}}{8}\left(\frac{\epsilon/2}{\mathrm{M}_s(\rho_0+\nu_*)+C_s(N+1)\eta}\right)^{\delta/(1-\delta)} \geq \frac{2A\eta^{\delta\beta/(1-\delta)}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}, \\ \frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}} \geq \frac{16\alpha L^2M^{2\beta}}{a}\eta^{2\beta+1}. \end{cases} \tag{C.8}$$

For the second inequality, we simply plug in the values for η and A. Then, by using $\epsilon < 2\Delta_0/e$ and $\mathrm{M}_s(\rho_0+\nu_*) \geq 1$, this inequality holds if the following is satisfied,

$$\left(\frac{2}{\epsilon}\right)^{\frac{2-3\delta}{1-2\delta}} \geq \frac{16\alpha L^2M^{2\beta}}{a}\left(\frac{1}{16\lambda^{\frac{1}{1-\delta}}}\right)^{\frac{1-\delta}{1-2\delta}}(\sigma\tilde{d})^{-2}.$$

This yields an upper bound on the accuracy. In order to simplify this bound and make it independent of s, we define $\tilde{\lambda} = \frac{4e^{2\xi}}{1\vee\mu} \leq \lambda$. Also using $4L^2 < \sigma$ and $d \leq \tilde{d}$, the bound can be simplified to

$$\epsilon \leq 2\left(\tilde{\lambda}^{0.5}\wedge\tilde{\lambda}^2\right)\left(1\wedge\frac{2a\sigma d^2}{M^{2\beta}}\right)^{0.5},$$

under which the second inequality in (C.8) holds.

For the first inequality in (C.8), we consider two cases. In the first case $N\eta \geq 1$, since we have

$$\mathrm{M}_s(\rho_0+\nu_*)+C_s(N+1)\eta \leq 3(\mathrm{M}_s(\rho_0+\nu_*)\vee C_s)N\eta,$$

the following condition implies the desired inequality

$$\frac{3\lambda^{-1/(1-\delta)}}{8}\left(\frac{\epsilon/2}{3(\mathrm{M}_s(\rho_0+\nu_*)\vee C_s)N\eta}\right)^{\delta/(1-\delta)} \geq \frac{2A\eta^{\delta\beta/(1-\delta)}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}.$$

This inequality can be verified by plugging in the values of A, η and N. In the other case $N\eta < 1$, we simply drop $N\eta$ since we have

$$\mathrm{M}_s(\rho_0+\nu_*)+C_s(N+1)\eta \leq 3(\mathrm{M}_s(\rho_0+\nu_*)\vee C_s),$$

hence, the following condition suffices

$$\frac{3\lambda^{-1/(1-\delta)}}{8}\left(\frac{\epsilon/2}{3(\mathrm{M}_s(\rho_0+\nu_*)\vee C_s)}\right)^{\delta/(1-\delta)} \geq \frac{2A\eta^{\delta\beta/(1-\delta)}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}.$$

For this to hold, it is sufficient if $\epsilon < 2$ and

$$\log\left(\frac{2\Delta_0}{\epsilon}\right) \geq \frac{1}{16}\lambda^{-\frac{1}{1-\delta}}.$$

which can be further strengthened to

$$\epsilon \leq 2\Delta_0 e^{\frac{-1}{16(\tilde{\lambda}\wedge\tilde{\lambda}^2)}}.$$

Hence, the simplified single step bound (C.6) holds when KL-divergence is not too small, i.e. when it is greater than $\epsilon/2$. For handling the case where KL-divergence is small, we need to show that once LMC reaches ϵ-accuracy, it remains below that threshold until the last step. In other words

$$\mathrm{KL}\big(\rho_k|\nu_*\big) \leq \epsilon \implies \mathrm{KL}\big(\rho_{k+1}|\nu_*\big) \leq \epsilon, \text{ for } k < N. \tag{C.9}$$

We split this into two cases. First, consider the case $\epsilon/2 \leq \mathrm{KL}\big(\rho_k|\nu_*\big) \leq \epsilon$. In this case, using (C.6) and $\mathrm{KL}\big(\rho_k|\nu_*\big) \leq \epsilon$, it suffices to show

$$\sigma\tilde{d}\eta^{\beta+1} \leq \epsilon\frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}},$$

which can be verified by plugging in the values for A, η and $\tilde{d}$. When $\mathrm{KL}\big(\rho_k|\nu_*\big) \leq \epsilon/2$, using Lemma 16, we need to show

$$\frac{16\alpha L^2M^{2\beta}}{a}\eta^{2\beta+1}\frac{\epsilon}{2} + \sigma\tilde{d}\eta^{\beta+1} \leq \frac{\epsilon}{2}.$$

We bound each of the terms on the left hand side with $\epsilon/4$. By simplifying the expressions and further using $\epsilon < 2\Delta_0/e$ and $\mathrm{M}_s \geq 1$, we obtain the following two conditions on the accuracy ϵ to be combined together later,

$$\epsilon \leq 8(\tilde{\lambda}\wedge\tilde{\lambda}^2)\left(1\wedge\frac{a}{\alpha L^2M^{2\beta}}\right)^{\frac{1}{3}}(1\wedge\sigma d) \leq 2^{5-\frac{5\beta(1-2\delta)}{(1+2\beta)(1-\delta)}}\left(\frac{a}{\alpha L^2M^{2\beta}}\right)^{\frac{\beta(1-2\delta)}{(1+2\beta)(1-\delta)}}\tilde{\lambda}^{\frac{1}{1-\delta}}(\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta}},$$
$$\epsilon \leq 32(\tilde{\lambda}\wedge\tilde{\lambda}^2)(1\wedge\sigma d) \leq 2^{5+\frac{3\beta(1-2\delta)}{1-\delta+\delta\beta}}\lambda^{\frac{1+\beta}{1-\delta+\delta\beta}}(\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta+\delta\beta}}.$$

Next, our analysis continues with considering the following two cases.

1. LMC reaches ϵ accuracy at a step $k < N$.
2. LMC does not reach ϵ accuracy at a step $k < N$.

For the first case above, if at any step $k < N$, we have $\mathrm{KL}\big(\rho_k|\nu_*\big) \leq \epsilon$, then by using (C.9) we conclude $\mathrm{KL}\big(\rho_N|\nu_*\big) \leq \epsilon$. For the second case, we have $\mathrm{KL}\big(\rho_k|\nu_*\big) > \epsilon$ for all $k < N$; therefore, (C.6) combined with Lemma 32 and the fact that $\mathrm{KL}\big(\rho_0|\nu_*\big) \leq \Delta_0$ imply

$$\mathrm{KL}\big(\rho_N|\nu_*\big) \leq \exp\left(\frac{-A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}N\right)\Delta_0 + \frac{\sigma\tilde{d}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}\eta^{\beta(1-2\delta)/(1-\delta)}}{A}.$$

Notice that to reach ϵ accuracy at step N, it is sufficient that each of the above terms on the right hand side is upper bounded by $\epsilon/2$. Simplifying these bounds, we obtain

$$\log\left(\frac{2\Delta_0}{\epsilon}\right) \leq \frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}N,$$

$$\eta \leq A^{\frac{1-\delta}{\beta(1-2\delta)}}(\sigma\tilde{d})^{-\frac{1-\delta}{\beta(1-2\delta)}}\left(\frac{\epsilon}{2}\right)^{\frac{1-\delta}{\beta(1-2\delta)}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{-\frac{\delta}{\beta(1-2\delta)}}.$$

The second inequality holds with the selection of η. Plugging the value for η in the first inequality yields

$$(\sigma\tilde{d})^{\frac{1-\delta+\delta\beta}{\beta(1-2\delta)}}A^{-\frac{(1+\beta)(1-\delta)}{\beta(1-2\delta)}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\frac{\beta(1-\delta)+\delta}{\beta(1-2\delta)}}\left(\frac{2}{\epsilon}\right)^{\frac{1-\delta+\delta\beta}{\beta(1-2\delta)}} \leq N,$$

which is true because of the value of N.

Finally, we translate the bound on the step size in Lemma 16, to a condition on the accuracy ϵ. That is, we have

$$\eta \leq \frac{1}{2}\left(1 \wedge \frac{a}{2M^2}\right) \wedge \frac{4\lambda^{\frac{1}{1-\delta}}}{3}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*) + C_s(k+1)\eta}{\mathrm{KL}\big(\rho_k|\nu_*\big)}\right)^{\frac{\delta}{1-\delta}}.$$

By plugging the value of η, in $\eta \leq \frac{1}{2}\left(1 \wedge \frac{a}{2M^2}\right)$, we get

$$\left(\frac{\epsilon}{2}\right)\log\left(\frac{2\Delta_0}{\epsilon}\right)^{-\frac{\delta}{1-\delta}} \leq 32\left(\frac{1}{2}\left(1 \wedge \frac{a}{2M^2}\right)\right)^{\beta(\frac{1-2\delta}{1-\delta})}\lambda^{\frac{1}{1-\delta}}\left(\mathrm{M}_s(\rho_0+\nu_*) \vee C_s\right)^{\frac{\delta}{1-\delta}}(\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta}},$$

but since $\epsilon < \frac{2\Delta_0}{e}$ and $\mathrm{M}_s \geq 1$ and $\beta(\frac{1-2\delta}{1-\delta}) \leq 1$, it suffices to have

$$\epsilon \leq 16(\tilde{\lambda} \wedge \tilde{\lambda}^2)(1 \wedge \sigma d)\left(1 \wedge \frac{a}{2M^2}\right) \leq 16\left(1 \wedge \frac{a}{2M^2}\right)\lambda^{\frac{1}{1-\delta}}(\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta}}.$$

For the other constraint on η, if we show $\eta \leq \frac{4\lambda^{\frac{1}{1-\delta}}}{3}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*)}{\Delta_0}\right)^{\frac{\delta}{1-\delta}}$, Lemma 16 shows that the first step is decreasing and $\mathrm{KL}\big(\rho_1|\nu_*\big) \leq \Delta_0$. Continuing inductively from there, we get either $\mathrm{KL}\big(\rho_k|\nu_*\big)$ is decreasing or it is less than ϵ, in both of the cases, we have $\mathrm{KL}\big(\rho_k|\nu_*\big) \leq \Delta_0$. This in turn shows that the constraint on η is getting looser, so all we need to consider is

$$\eta \leq \frac{4\lambda^{\frac{1}{1-\delta}}}{3}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*)}{\Delta_0}\right)^{\frac{\delta}{1-\delta}},$$

which holds whenever

$$\epsilon \log\left(\frac{2\Delta_0}{\epsilon}\right)^{-\frac{\delta}{1-\delta}} \leq 32\lambda^{\frac{1-\delta+\beta(1-2\delta)}{(1-\delta)^2}} \left(\mathrm{M}_s(\rho_0+\nu_*)\vee C_s\right)^{\frac{\delta}{1-\delta}} \left(\frac{\mathrm{M}_s(\rho_0+\nu_*)}{\Delta_0}\right)^{\frac{\delta(1-2\delta)\beta}{(1-\delta)^2}} (\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta}}.$$

Once again, since $\epsilon < \frac{2\Delta_0}{e}$ and $\mathrm{M}_s \geq 1$, all we need is

$$\epsilon \leq 32(1\wedge\sigma d)(\tilde{\lambda}\wedge\tilde{\lambda}^3)(1\wedge\Delta_0^{-1})^{\frac{\beta}{4}} \leq 32\lambda^{\frac{1-\delta+\beta(1-2\delta)}{(1-\delta)^2}} \Delta_0^{\frac{-\delta(1-2\delta)\beta}{(1-\delta)^2}} (\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta}}.$$

Collecting all the upper bounds on the accuracy we get

$$\begin{aligned}\psi = \min\Bigg\{ & 2, \frac{2\Delta_0}{e}, 2\Delta_0 e^{\frac{-1}{16(\tilde{\lambda}\wedge\tilde{\lambda}^2)}}, 32(1\wedge\sigma d)(\tilde{\lambda}\wedge\tilde{\lambda}^3)(1\wedge\Delta_0^{-1})^{\frac{\beta}{4}}, \\ & 16(\tilde{\lambda}\wedge\tilde{\lambda}^2)(1\wedge\sigma d)\left(1\wedge\frac{a}{2M^2}\right), 2\left(\tilde{\lambda}^{0.5}\wedge\tilde{\lambda}^2\right)\left(1\wedge\frac{2a\sigma d^2}{M^{2\beta}}\right)^{0.5}, \\ & 8(\tilde{\lambda}^2\wedge\tilde{\lambda})\left(1\wedge\frac{a}{\alpha L^2 M^{2\beta}}\right)^{\frac{1}{3}}(1\wedge\sigma d), 32(\tilde{\lambda}\wedge\tilde{\lambda}^2)(1\wedge\sigma d)\Bigg\},\end{aligned} \tag{C.10}$$

where $\tilde{\lambda}$ is defined as $\tilde{\lambda} = \frac{4e^{2\xi}}{1\vee\mu}$. Note that the upper bound on ϵ is of order $\mathcal{O}(1)$, and it depends on the fixed parameters except for Δ_0 which depends on the initial distribution. In case of starting with a Gaussian random vector, Lemma 26 provides a bound on Δ_0. More importantly, the upper bound on the accuracy does not depend on the moment order s, which enables us to optimize over this parameter which is done in Corollary 4. Finally, we plug in the values for δ, $\tilde{d}$ and C_s back into (C.7) to get

$$\begin{aligned}\eta = {} & \sigma^{-\frac{1}{\beta}}(16\lambda)^{-\frac{s-2+2\theta}{\beta(s-2)}}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*)}{16d^{(s-2+\alpha)/\alpha}}\vee\left(\frac{3a+2b+3}{1\wedge a}\right)^{\frac{s-2+\alpha}{\alpha}}\frac{s^s}{16}\right)^{-\frac{\theta}{\beta(s-2)}} \\ & d^{-\frac{1}{\beta}-\frac{(s-2+\alpha)\theta}{\alpha\beta(s-2)}}(1+(1-\alpha/2)\log(d))^{-\frac{1}{\beta}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{-\frac{\theta}{\beta(s-2)}}\left(\frac{\epsilon}{2}\right)^{\frac{s-2+\theta}{\beta(s-2)}}, \\ N = {} & \sigma^{\frac{1}{\beta}}(16\lambda)^{\frac{(1+\beta)(s-2+2\theta)}{\beta(s-2)}}\left(\frac{\mathrm{M}_s(\rho_0+\nu_*)}{16d^{(s-2+\alpha)/\alpha}}\vee\left(\frac{3a+2b+3}{1\wedge a}\right)^{\frac{s-2+\alpha}{\alpha}}\frac{s^s}{16}\right)^{\frac{(1+\beta)\theta}{\beta(s-2)}} \\ & d^{\frac{1}{\beta}+\frac{(s-2+\alpha)(1+\beta)\theta}{\alpha\beta(s-2)}}(1+(1-\alpha/2)\log(d))^{\frac{1}{\beta}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{1+\frac{(\beta+1)\theta}{\beta(s-2)}}\left(\frac{2}{\epsilon}\right)^{\frac{1}{\beta}+\frac{(1+\beta)\theta}{\beta(s-2)}}.\end{aligned}$$

■

Appendix D. Linear Growth of Convex Potentials

First, we prove a lemma about one dimensional convex potentials, which will be used to prove the unboundedness in the general case. A similar result can be found in (Bobkov and Madiman, 2011, Equation (9)).

Lemma 18 *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\int_{\mathbb{R}} e^{-f(x)}dx < \infty$, then f is lower bounded, i.e. $\inf_{x\in\mathbb{R}} f(x) > -\infty$.*

Proof Shifting f does not affect convexity or finiteness of the integral, so we can assume, without loss of generality, $f(0) = 0$. Let $B(r) = \min_{x\in[-r,r]} f(x)$, which is well defined because f is continuous – convexity implies continuity in this context. If B is lower bounded, then so is f. Suppose B is not lower bounded. Continuity of f implies that B is also continuous, and $f(0) = 0$ implies that $B(0) = 0$. Further, B is a non-increasing function in its domain.

For $M \geq 0$, we can define $y(M) = \min\{r|B(r) = -M\}$, because the range of B contains all non-positive numbers. Fix some $M > 1$. Then, the continuity of B and f imply that either $f(y(M)) = -M$ or $f(-y(M)) = -M$. Without loss of generality, we assume $f(y(M)) = -M$ (the other case is similar), and write

$$\forall x \in [0, y(M)] : \ f(x) \leq \left(1 - \tfrac{x}{y(M)}\right) \times f(0) + \tfrac{x}{y(M)} \times f(y(M)) = -\tfrac{Mx}{y(M)}.$$

Using this fact, we integrate e^{-f} to get

$$\int_{\mathbb{R}} e^{-f(x)}dx \geq \int_0^{y(M)} e^{-f(x)}dx \geq \int_0^{y(M)} e^{\frac{Mx}{y(M)}}dx = y(M) \times \tfrac{e^M-1}{M}.$$

Monotonicity of B implies $y(M) > y(1) > 0$ since we also have $B(0) = 0$. Hence, the previous inequality yields

$$\int_{\mathbb{R}} e^{-f(x)}dx \geq y(1) \times \tfrac{e^M-1}{M} \quad \text{for every } M > 1.$$

This inequality contradicts $\int_{\mathbb{R}} e^{-f(x)}dx < \infty$. ■

We use the previous one dimensional result to show that, in the general case, not only the potential is lower bounded but also it has at least linear growth along every direction. The method is to first prove the potential is unbounded along every direction and then use that to prove linear growth.

Lemma 19 *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a convex potential (i.e. $\int e^{-f} < \infty$) and $u \in \mathbb{R}^d$ is unit vector. Then, f is coercive satisfying*

$$\sup_{t\geq 0} f(tu) = +\infty.$$

Proof Without loss of generality, let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex potential satisfying $\int e^{-f(x)}dx = 1$. Assume, for the sake of contradiction, that there is a direction $u_1 \in \mathbb{R}^d$ such that

$$\sup_{t\geq 0} f(tu_1) < M < \infty,$$

and let $\{u_1, u_2, ..., u_d\}$ be an orthonormal basis for $\mathbb{R}^d$. Using convexity, we have

$$f(tu_1) \geq f(x) + t\langle \nabla f(x), u_1\rangle - \langle \nabla f(x), x\rangle.$$

Taking supremum in both sides with respect to t yields $\langle \nabla f(x), u_1\rangle \leq 0$ for every $x \in \mathbb{R}^d$. Let $x_1 = \langle x, u_1\rangle$ and write $x = x_1u_1 + x_{-1}$ where $\langle u_1, x_{-1}\rangle = 0$. By convexity, we have

$$-f(x) \geq -f(0) - \langle \nabla f(x), x\rangle.$$

We can write

$$\begin{aligned}1 = \int_{\mathbb{R}^d} e^{-f(x)}dx &\geq \int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} e^{-f(0)-\langle\nabla f(x),x\rangle}dx_1dx_{-1}\\ &= \int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} e^{-f(0)-x_1\langle\nabla f(x),u_1\rangle-\langle\nabla f(x),x_{-1}\rangle}dx_1dx_{-1}\\ &\geq \int_{\mathbb{R}^{d-1}}\int_{x_1\geq 0} e^{-f(0)-x_1\langle\nabla f(x),u_1\rangle-\langle\nabla f(x),x_{-1}\rangle}dx_1dx_{-1}\\ &\geq \int_{\mathbb{R}^{d-1}}\int_{x_1\geq 0} e^{-f(0)-\langle\nabla f(x_1u_1+x_{-1}),x_{-1}\rangle}dx_1dx_{-1}.\end{aligned}$$

If we have $\sup_{x_1\geq 0}\langle\nabla f(x_1u_1+x_{-1}),x_{-1}\rangle < \infty$, then the inner integral diverges, therefore

$$\sup_{x_1\geq 0}\langle\nabla f(x_1u_1+x_{-1}),x_{-1}\rangle = \infty,$$

for almost every $x_{-1}\in\text{span}\{u_2,...,u_d\}$. Using finiteness of the integral once again, we write

$$1 = \int_{\mathbb{R}^d} e^{-f(x)}dx = \int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} e^{-f(x_1u_1+x_{-1})}dx_1dx_{-1}.$$

The inner integral should converge for almost every $x_{-1}\in\text{span}\{u_2,...,u_d\}$. Since a convex function restricted to a line is still convex, Lemma 18 implies $e^{-f(x_1u_1+x_{-1})}$ is lower bounded for almost every x_{-1}. Fix some $x_{-1}\in\text{span}\{u_2,...,u_d\}$ such that $g(x_1)=e^{-f(x_1u_1+x_{-1})}$ is lower bounded and $\sup_{x_1\geq 0}\langle\nabla f(x_1u_1+x_{-1}),x_{-1}\rangle=\infty$, which happens for almost every x_{-1}. By convexity, we have

$$f(x_1u+2x_{-1}) \geq f(x_1u+x_{-1}) + \langle\nabla f(x_1u_1+x_{-1}),x_{-1}\rangle.$$

Since that $\sup_{x_1\geq 0} f(x_1u_1+2x_{-1})=\infty$ and $\sup_{x_1\geq 0} f(2x_1u_1) < M$, once again by convexity

$$\tfrac{1}{2}f(4x_{-1}) + \tfrac{1}{2}f(2x_1u) \geq f(x_1u+2x_{-1}).$$

Taking supremum with respect to x_1 results in a contradiction. So the assumption was incorrect and no direction like u_1 exists. ■

In the light of the previous lemma, convexity implies a growth that is at least linear. This is established in the following proof.

Proof [Proof of Proposition 6] Let the function B from unit sphere to real numbers be defined as

$$B(u) = \inf\{t>0 | f(tu)\geq 1+f(0)\},$$

which is well defined because of Lemma 19. Convexity (and therefore continuity) of f implies B is continuous. Since unit sphere is compact, B attains its maximum on it. Let us call this maximum $t_0>0$. We have $f(t_0u)\geq 1+f(0)$ for all unit vectors $u\in\mathbb{R}^d$. For any $t>t_0$ and any unit vector u, because of convexity, we write

$$\left(1-\tfrac{t_0}{t}\right)f(0) + \tfrac{t_0}{t}f(tu) \geq f(t_0u) \geq 1+f(0).$$

Therefore, for $t>t_0$, we have

$$f(tu) \geq \tfrac{t}{t_0} + f(0),$$

for all unit directions u.

When $t \in [0, t_0]$, the function $t \to f(tu)$ is lower bounded by some constant, i.e.

$$\inf_{t\in[0,t_0]} f(tu) \coloneqq g(u) > -\infty$$

by Lemma 18. Since f is continuous in both t and u, $g(u)$ is also continuous. Further, since its domain is compact, by the extreme value theorem, g attains its infimum in its domain; thus, it is also lower bounded, say by $-M < 0$. Therefore, whenever $t \in [0, t_0]$, $f(tu) \geq -M$ for all unit directions u. Combining this with the previous result, we obtain that for $t \in [0, \infty)$,

$$f(tu) \geq \tfrac{t}{t_0} - |f(0)| \vee (M + 1).$$

For the second part, by convexity, we write

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$$

for all $x, y \in \mathbb{R}^d$. By choosing $y = 0$, we obtain

$$\begin{aligned}\langle \nabla f(x), x\rangle \geq& f(x) - f(0)\\ \geq& a\|x\|^\alpha - b - f(0)\end{aligned}$$

where in the last step, we used $f(x) \geq a\|x\|^\alpha - b$. This completes the proof. ■

Appendix E. Gap of Degenerate Convexity Degree for Linear Growth

In this section, we show that any function with finite difference from a linear function can not be convex degenerate of degree $\theta \leq 2$. Note that we previously showed functions with linear growth are convex degenerate for any $\theta > 2$. In the next lemma, we consider the single dimensional case. Restricting higher dimensional potentials to a single dimensional subspace, yields the proof for $d > 1$.

Lemma 20 *Suppose $g : \mathbb{R} \to \mathbb{R}$ is a function such that*

$$|a|x| - g(x)| < C, \qquad\qquad g''(x) \geq \frac{\mu}{(1 + x^2)^{\theta/2}}, \tag{E.1}$$

for $a, c, \mu \in \mathbb{R}$. Then $\theta > 2$.

Proof Without loss of generality assume $g(0) = 0$. (Shifting changes C at most by $g(0)$.) We restrict our attention to positive real numbers, from (E.1) we have for $x > 0$

$$|a - g(x)/x| < C/x \implies \lim_{x\to+\infty} g'(x) \overset{1}{=} \lim_{x\to+\infty} \tfrac{g(x)}{x} = a,$$

where step 1 follows from L'Hôpital's rule. Since $g''(x) > 0$ we have $g'(x) \leq a$. Let $D(x) = ax - g(x)$, from E.1 we have $|D(x)| \leq C$. We argue $D(x) \geq 0$ when $x \geq 0$. First using $g(0) = 0$, we calculate $D(0) = 0$. For the derivative we have $D'(x) = a - g'(x) \geq 0$. This implies $D(x) \geq 0$,

therefore $0 \leq D(x) \leq C$ for non-negative x. The derivative of D shows that it is increasing, therefore $\lim_{x\to\infty} D(x)$ exists and by the previous upper-bound we get the following

$$C \geq \lim_{x\to\infty} D(x) = \int_0^\infty D'(x) = \int_0^\infty (a - g'(x))dx,$$

where we used $D(0) = 0$. We substitute the following two equalities in the previous one

$$\begin{aligned} a &= \lim_{x\to\infty} g'(x) = g'(0) + \int_0^\infty g''(t)dt, \\ g'(x) &= g'(0) + \int_0^x g''(t)dt. \end{aligned}$$

The mentioned substitution reads

$$C \geq \int_0^\infty (\int_x^\infty g''(t)dt)dx \overset{1}{=} \int_0^\infty (\int_0^t g''(t)dx)dt = \int_0^\infty tg''(t)dt \geq \int_0^\infty \frac{t\mu}{(1+t^2)^{\theta/2}}dt \overset{2}{=} \frac{\mu}{2}\int_1^\infty \frac{du}{u^{\theta/2}},$$

where in step 1 we changed the order of integration and in step 2 we used the substitution $u = 1+t^2$. The finiteness of the last expression implies $\theta > 2$. ■

Appendix F. Proofs of Corollaries and Lemmas

Proof of Lemma 7. Let the bounds on ϕ and $\nabla\phi$ be κ_1 and κ_2, respectively. Since ϕ is bounded, $\int e^{-f-\phi}$ is finite, therefore it can be normalized to be a probability distribution. We ignore the normalizing constant since it does not change the gradient and the Hessian.

Assumption 1 holds for f, meaning that there exists a $\tilde{f}$ such that $\|f - \tilde{f}\|_\infty < \xi$, and $\tilde{f}$ satisfies the conditions in Assumption 1. Since $|\phi| \leq \kappa_1$, we have

$$\|f + \phi - \tilde{f}\|_\infty < \xi + \kappa_1,$$

which proves that Assumption 1 also holds for $f + \phi$. For Assumption 2, we write

$$\langle \nabla f(x) + \nabla\phi(x), x\rangle \geq a\|x\|^\alpha - b - \langle \phi(x), x\rangle \geq a\|x\|^\alpha - b - \kappa_2\|x\| \geq a'\|x\|^\alpha - b',$$

for some $a', b' > 0$, where in the last step we used $\alpha > 1$. When $\alpha = 1$ this step is correct when $\kappa_2 < a$. Growth part remains true since the perturbation has bounded gradient

$$\|\nabla f + \nabla\phi\| \leq \|\nabla f\| + \|\nabla\phi\| \leq (\kappa_2 + M)\left(1 + \|x\|^\zeta\right),$$

which implies that g satisfies Assumption 2. Finally, for Assumption 3, since both $\nabla\phi$ and ∇f are β-Hölder continuous, so is their summation for the same order of smoothness β. ■

Proof of Corollary 4. Initializing with a Gaussian random vector provides us with

$$\mathrm{M}_s(\rho_0) = \mathbb{E}\left[\left(1 + \|x\|^2\right)^{s/2}\right] \leq 2^{s/2}\mathbb{E}\left[1 + \|x\|^s\right] \leq 2^{s/2}(1 + d^{s/2}(s-1)!!) \leq (2ds)^{s/2}.$$

We state a lemma to bound the moments of the target distribution.

Lemma 21 *Let f satisfy Assumption 2 then we have the following bound on the moment*

$$\mathrm{M}_s(\nu_*) \leq \left(\tfrac{a+b+3}{a}\right)^{s/\alpha} s^{s/\alpha} d^{s/\alpha} \quad \textit{for all} \quad s \geq 2.$$

Proof of Lemma 21. We utilize a method, similar to the method used in proof of Lemma 10, in order to bound the moments of target. From the proof of Lemma 10, we have the following inequality for $s \geq 2$.

$$\tfrac{d}{dt}\mathrm{M}_s(p_t) \leq (b+d+a+s-2)s\mathrm{M}_{s-2}(p_t) - \tfrac{as}{2}\mathrm{M}_{s+\alpha-2}(p_t).$$

If we let $p_0 = \nu_*$, then $p_t = \nu_*$ which means that the left hand side of the above inequality is zero. The derivative is well defined because Lemma 11 shows that $\mathrm{M}_s(\nu_*)$ is finite. By rearranging the previous inequality, we get

$$\mathrm{M}_{s+\alpha-2}(p_t) \leq \tfrac{2(b+d+a+s-2)}{a}\mathrm{M}_{s-2}(p_t).$$

Using the above inequality inductively from $s = 2$, we get

$$\mathrm{M}_{k\alpha}(p_t) \leq \left(\tfrac{2}{a}\right)^k (a+b+d+(k-1)\alpha)^k.$$

For every s there is an integer k such that $k\alpha \leq s < (k+1)\alpha$. We have the following bound

$$\mathrm{M}_s(p_t) \leq \mathrm{M}_{(k+1)\alpha}(p_t)^{\frac{s}{(k+1)\alpha}} \leq \left(\tfrac{2}{a}\right)^{s/\alpha} (a+b+d+k\alpha)^{s/\alpha} \leq \left(\tfrac{a+b+3}{a}\right)^{s/\alpha} s^{s/\alpha} d^{s/\alpha}.$$

■

Combining the Gaussian moment bound with the previous lemma yields

$$\mathrm{M}_s(\rho_0 + \nu_*) \leq 2\left(\tfrac{3a+b+3}{a}\right)^{s/\alpha} s^{s/\alpha} d^{s/\alpha}.$$

Using $s = 2 + 2\lceil \log(\frac{6d}{\epsilon}) \rceil$ implies d^γ and $(2/\epsilon)^\gamma$ are bounded with $\exp\left(\frac{(1+\beta)\theta}{2\beta}\right)$. By plugging this upper bound back in Theorem 3 and using the inequalities,

$$\gamma < \frac{(1+\beta)\theta}{2\beta}, \qquad \epsilon < 2 \vee 2\Delta_0/e, \qquad \lambda \leq \frac{4e^{2\xi}}{1 \wedge \mu},$$

the advertised rate is obtained. ■

Proof of Corollary 5. We prove the rate in each row separately. We start with total variation and state Pinsker's inequality, which bounds total variation with KL-divergence.

Lemma 22 (Pinsker's inequality) *For distributions p and q*

$$\mathrm{TV}\left(p, q\right) \leq \sqrt{\frac{1}{2}\mathrm{KL}\big(p|q\big)}.$$

If for given ϵ we use Corollary 4 with accuracy $2\epsilon^2$, Pinsker's inequality implies

$$\mathrm{TV}\left(\rho_N, \nu_*\right) \leq \epsilon.$$

Note that the upper bound on ϵ is changed and $2\epsilon^2$ needs to be smaller than upper bound in (C.10). In other words $2\epsilon^2 \leq \psi$, where ψ is defined in (C.10).

Now we prove the convergence rate bound for $\mathcal{W}_\alpha$. We start by stating a result, which is adapted from Corollary 3 in Bolley and Villani (2005), that bounds L_α-Wasserstein distance with KL-divergence.

Lemma 23 (Bolley and Villani (2005)) *For probability measure p on $\mathbb{R}^d$, if $\int e^{\theta\|x\|^\alpha} p(x)dx < \infty$, then*

$$\mathcal{W}_\alpha(p,q) \leq B\left[\mathrm{KL}\big(p|q\big)^{\frac{1}{\alpha}} + \left(\frac{\mathrm{KL}\big(p|q\big)}{2}\right)^{\frac{1}{2\alpha}}\right],$$

where

$$B \triangleq 2\inf_{\kappa}\left(\frac{1}{\kappa}\left(1.5 + \log\int e^{\kappa\|x\|^\alpha} p(x)dx\right)\right)^{\frac{1}{\alpha}}.$$

Lemma 12 proves an upper bound on B, namely $B < 2(4(\tilde{d}\tilde{\mu}+1.5)/a)^{1/\alpha}$. By plugging this upper bound back in the previous lemma we get

$$\mathcal{W}_\alpha(\rho_N,\nu_*) \leq 2\left(\frac{4\alpha}{a}(1.5 + \tilde{\mu}\left(1 + (1-\alpha/2)\log(d)\right)d)\right)^{\frac{1}{\alpha}}\left(\mathrm{KL}\big(\rho_N|\nu_*\big)^{\frac{1}{\alpha}} + \mathrm{KL}\big(\rho_N|\nu_*\big)^{\frac{1}{2\alpha}}\right).$$

If $\epsilon \leq 4\left(4\alpha a^{-1}(1.5+\tilde{d}\tilde{\mu})\right)^{1/\alpha}$, using Corollary 4 with accuracy $(\epsilon/4)^{2\alpha}(4\alpha a^{-1}(1.5+\tilde{\mu}\tilde{d}))^{-2}$, implies the convergence rate bound. In order to obtain the upper bound on the accuracy, first let ψ denote the bound in (C.10). Since we used $(\epsilon/4)^{2\alpha}(4\alpha a^{-1}(1.5+\tilde{\mu}\tilde{d}))^{-2}$ as the accuracy in terms of KL-divergence we need

$$(\epsilon/4)^{2\alpha}(4\alpha a^{-1}(1.5+\tilde{\mu}\tilde{d}))^{-2} \leq \psi,$$

by rearranging we get

$$\epsilon \leq 4(4\alpha a^{-1}(1.5+\tilde{\mu}\tilde{d}))^{-\frac{1}{\alpha}}\psi^{\frac{1}{2\alpha}}.$$

Collecting these upper bound together we get

$$\epsilon \leq 4(4\alpha a^{-1}(1.5+\tilde{\mu}\tilde{d}))^{-\frac{1}{\alpha}}\psi^{\frac{1}{2\alpha}} \wedge 4\left(4\alpha a^{-1}(1.5+\tilde{d}\tilde{\mu})\right)^{\frac{1}{\alpha}}, \tag{F.1}$$

where ψ is defined in (C.10) and $\tilde{d}$ and $\tilde{\mu}$ are defined in (C.1).

Finally, we give the proof for $\mathcal{W}_2$ convergence rate bound in the quadratic growth and smooth setting. When $\theta = 0$, Theorem 1 implies LSI with constant $\frac{4e^{2\xi}}{\mu}$. LSI implies Talagrand's inequality with the same constant (Otto and Villani, 2000).

$$\mathcal{W}_2(\rho_N,\nu_*) \leq 4e^{\xi}\sqrt{\mathrm{KL}\big(\rho_N|\nu_*\big)/\mu}.$$

Theorem 3 with accuracy $\frac{\epsilon^2\mu}{16e^{2\xi}}$ implies the convergence rate bound. Note that we do not need to choose any s since $\gamma = 0$ and Theorem 3 is independent of s. The upper bound on ϵ changes and $\frac{\epsilon^2\mu}{16e^{2\xi}}$ needs to be smaller than ψ. In other words

$$\epsilon \leq \frac{4e^{\xi}}{\sqrt{\mu}}\psi, \tag{F.2}$$

where ψ is defined in (C.10). ■

Appendix G. Additional Related Work

The LMC algorithm has been extensively studied in the context of sampling from a log-concave target distribution. Earlier results focused on characterizing its bias which is also referred to as the integration error (Milstein, 1994; Milstein and Tretyakov, 2013), and the convergence guarantees were mostly asymptotic (Gelfand and Mitter, 1991; Meyn and Tweedie, 2012). Non-asymptotic analysis of LMC has drawn a lot of interest recently (Dalalyan and Tsybakov, 2012; Dalalyan, 2017a,b; Durmus et al., 2019a; Cheng and Bartlett, 2018; Cheng et al., 2018a; Vempala and Wibisono, 2019; Dalalyan and Karagulyan, 2019; Brosse et al., 2019) where the focus was on potentials exhibiting strong tail growth properties. These papers were mostly influenced by the pioneering works by Dalalyan (2017b), and Durmus and Moulines (2017); Durmus et al. (2019b) where it was shown that for strongly convex and smooth potentials, LMC reaches ϵ accuracy in terms of total variation (TV) distance after $\widetilde{\mathcal{O}}(d\epsilon^{-2})$ steps. Similarly, $\widetilde{\mathcal{O}}\left(d\epsilon^{-2}\right)$ steps are sufficient to reach ϵ accuracy under the L_2-Wasserstein distance Durmus et al. (2019b), which can be further improved to $\widetilde{\mathcal{O}}\left(d\epsilon^{-1}\right)$ under an additional second-order smoothness assumption on the potential function.

In this paper, we established guarantees under KL-divergence (relative entropy) which can be easily translated to TV and Wasserstein metrics using Csiszár-Kullback-Pinsker (CKP) Bolley and Villani (2005) and/or Talagrand inequalities (Talagrand, 1996; Otto and Villani, 2000). For strongly convex and smooth potentials, it is known that $\widetilde{\mathcal{O}}\left(d\epsilon^{-1}\right)$ steps of LMC yield an ϵ accurate sample in KL-divergence (Cheng and Bartlett, 2018; Durmus et al., 2019a). This is still the best known rate in this setup, and recovers the best known rates in TV (Durmus and Moulines, 2017; Dalalyan, 2017b) as well as in L_2-Wasserstein metrics (Durmus et al., 2019b). However, for convex and smooth potentials that grow like $\|x\|^\alpha$, the bound on the rate drops to $\widetilde{\mathcal{O}}\big(d^{1+\frac{4}{\alpha}}\epsilon^{-3}\big)$ due to lack of strong convexity (Cheng and Bartlett, 2018). Among various contributions of Durmus et al. (2019a), LMC was also analyzed for convex potentials, but their result does not yield a convergence guarantee for the last LMC iterate.

Existing results that establish the fast convergence of LMC require strong curvature conditions on the potential function; therefore, their applicability is limited. Recently, it has been observed that global curvature assumptions can be relaxed to the tails of the potential (Eberle, 2016; Eberle et al., 2019). For example, Cheng et al. (2018a) extended these results to sampling from smooth potentials that are strongly convex outside of a compact set, obtaining the same dimension and ϵ dependency in the strongly convex case at the expense of an exponential dependence in the radius of the compact set. Similarly, Vempala and Wibisono (2019) established convergence guarantees for target distributions that satisfy a log-Sobolev inequality. This corresponds to potentials with quadratic tails (Bakry and Émery, 1985; Bobkov and Götze, 1999) up to finite perturbations (Holley and Stroock, 1987); thus, this result is able to deal with non-convex potentials that are not limited to a compact set, while establishing the same bound of $\widetilde{\mathcal{O}}\left(d\epsilon^{-1}\right)$ on the convergence rate in KL-divergence.

Convergence of the LMC algorithm is very little understood when the potential is weakly smooth. Contrary to previous work, our focus is on the convergence of vanilla LMC (1.2) without requiring any modifications on the algorithm such as methods based on proximal mapping (Atchadé, 2015; Luu et al., 2020; Durmus et al., 2018; Mou et al., 2019a; Durmus et al., 2019a), Gaussian smoothing (Chatterji et al., 2020; Doan et al., 2020), or mirror mapping (Hsieh et al., 2018). We also do not assume a composite structure on the potential, in which case the potential is given by $f(x) = U(x) + \psi(x)$ where $\psi(x)$ is a strongly convex and smooth function, and $U(x)$ is a con-

vex function with β-Hölder continuous gradient. This assumption enforces a quadratic tail growth on the potential, in which case, Chatterji et al. (2020) established the convergence rate bound of $\widetilde{\mathcal{O}}\left(d^{2+1/\beta}\epsilon^{-2/\beta}\right)$ in total variation distance. Furthermore, we focus on the last iteration of the LMC algorithm, in contrast to Durmus et al. (2019a) which provided guarantees for the average of the distributions of the LMC iterates.

Our analysis draws heavily on the theory of diffusion processes (Bakry et al., 2013; Toscani and Villani, 2000) – more specifically, logarithmic Sobolev inequalities. These inequalities were first established for the Gaussian density (Gross, 1975), and later generalized to Gibbs measure with a strongly convex potential by Bakry and Émery (1985). Combined with the Holley and Stroock's perturbation lemma (Holley and Stroock, 1987), this theory covers potentials that can be represented as a finite perturbation of a strongly convex function. It is well-known that the overdamped Langevin diffusion (1.1) follows the gradient flux or the steepest descent of KL-divergence, with respect to the L_2-Wasserstein metric (Jordan et al., 1998). Building on this, sampling with a diffusion can be seen as an optimization algorithm in the space of probability distributions (Wibisono, 2018; Vempala and Wibisono, 2019; Ma et al., 2019a); similarly, LSI can be interpreted as a gradient domination condition in this space, which is commonly referred to as the PL-inequality (Polyak, 1963) in the optimization theory. LSI and PL-inequality both yield exponential convergence in their corresponding space (Polyak, 1963; Karimi et al., 2016; Toscani, 1999; Carlen and Soffer, 1991). Further promoting this analogy, PL-inequality is a special case of Łojasiewicz inequality (Lojasiewicz, 1963), and their counterparts are considered recently in Blanchet and Bolte (2018) in the space of functionals. Thus, the modified LSI introduced in Toscani and Villani (2000), can be viewed as a modified version of the Łojasiewicz inequality in the space of probability distributions, note that, this modified LSI differs from the one introduced in Gentil et al. (2005) which interpolates between LSI and Poincaré inequality. Functional inequalities in Bertini and Zegarlinski (1999); Zegarlinski (2001) are similar in nature to the mLSI (1.6), yet their main focus is infinite dimensional semigroups (except (Zegarlinski, 2001, Sec. 2)). Specifically, the log-Nash inequality in (Bertini and Zegarlinski, 1999, Theorem 1.1) shares the same characteristics as (A.7). This result and mLSI-type results in general, to our knowledge, are stated by absorbing various important constants (in our context) into a leading constant, thus they cannot provide a sharp rate for LMC. For a survey about the convergence properties of diffusion processes with the Fokker-Planck equation governing their evolution (including overdamped Langevin dynamics (1.1)) and several inequalities from functional analysis, we refer the reader to Markowich and Villani (1999); Gentil et al. (2005). Finally, the analogy between optimization and sampling provided invaluable insights, in many cases improving our understanding, and ultimately the performance of various algorithms (Zhang et al., 2017; Brosse et al., 2017, 2018; Chatterji et al., 2018; Bhatia et al., 2019; Hsieh et al., 2018; Ma et al., 2019b).

It is worth mentioning that the rates we discussed in this section can be further improved by making higher order smoothness assumptions on the potential function (Mou et al., 2019b), or by considering higher order numerical integrators (Li et al., 2019; Shen and Lee, 2019; Dalalyan et al., 2019; He et al., 2020), or by certain adjustments (Durmus et al., 2017; Ge et al., 2018; Dwivedi et al., 2019). The overdamped Langevin diffusion (1.1) considered in this work is first order, and its higher order versions such as underdamped (Cheng et al., 2018b; Ma et al., 2019a), or third-order schemes (Ma et al., 2015; Mou et al., 2019c) may also provide additional improvements.

Appendix H. Useful Lemmas

Lemma 24 *For the potential function f, assume that there exists a function $\tilde{f}$ satisfying*

$$\left\|\nabla f - \nabla \tilde{f}\right\|_\infty \leq \xi.$$

If $\tilde{f}$ satisfies (2.1) in Assumption 1 for $\theta < 1$. Then α-dissipativity in Assumption 2 is satisfied for $\alpha = 2 - \theta$ with the following constants

$$a = \frac{\mu}{2(\alpha-1)} \quad \text{and} \quad b = \left(2(\|\nabla \tilde{f}(0)\| + \mu + \xi)^\alpha/\mu\right)^{1/(\alpha-1)}.$$

Remark 25 *The additional assumption about bounded perturbation of gradient is to prevent cases when the perturbation is bounded but its gradient is not, for example, $(1 - 2\sin(x))^{\frac{1}{3}}$.*

Proof Using the fundamental theorem of calculus we have

$$\begin{aligned}
\langle \nabla \tilde{f}(x), x\rangle &= \langle \int_0^1 \nabla^2 \tilde{f}(tx)x dt + \nabla \tilde{f}(0), x\rangle \\
&= \langle \nabla \tilde{f}(0), x\rangle + \int_0^1 x^\top \nabla^2 \tilde{f}(tx)x dt \\
&\geq -\|\nabla \tilde{f}(0)\|\|x\| + \int_0^1 \mu\left(1 + \|tx\|\right)^{\alpha-2}\|x\|^2 dt \\
&\geq -\|\nabla \tilde{f}(0)\|\|x\| + \frac{\mu\|x\|}{\alpha-1}\left((1+\|x\|)^{\alpha-1} - 1\right) \\
&= -\left(\|\nabla \tilde{f}(0)\| + \mu\right)\|x\| + \frac{\mu}{\alpha-1}\|x\|^\alpha.
\end{aligned}$$

Since $\left\|\nabla f - \nabla \tilde{f}\right\|_\infty \leq \xi$, we get

$$\begin{aligned}
\langle \nabla f(x), x\rangle &\geq -\left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)\|x\| + \frac{\mu}{\alpha-1}\|x\|^\alpha \\
&\geq \frac{\mu}{2(\alpha-1)}\|x\|^\alpha - \left(-\frac{\mu}{2(\alpha-1)}\|x\|^\alpha + \left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)\|x\|\right) \\
&\overset{1}{\geq} \frac{\mu}{2(\alpha-1)}\|x\|^\alpha - \left(\frac{2\left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)^\alpha}{\mu} \times \frac{\alpha-1}{\alpha}\right)^{1/(\alpha-1)} \\
&\geq \frac{\mu}{2(\alpha-1)}\|x\|^\alpha - \left(\frac{2\left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)^\alpha}{\mu}\right)^{1/(\alpha-1)},
\end{aligned}$$

where step 1 follows from Lemma 30. ■

Lemma 26 *Under Assumption 3, the KL-divergence between distribution $\rho = \mathcal{N}(x, I_d)$ for $x \in \mathbb{R}^d$ and the target distribution $\nu_* = e^{-f}$ is bounded as follows*

$$\mathrm{KL}\big(\rho|\nu_*\big) \leq f(x) + \frac{L}{\beta+1}d^{\frac{\beta+1}{2}} + \frac{d}{2}\log\left(2\pi e\right).$$

Remark 27 *The RHS depends on $f(x)$, so if it is possible to find a minimizer (or an almost minimizer) of f, it is preferred to generate initial point from a Gaussian distribution centered around the minimizer. Moreover, the value of $f(x)$ is for the normalized distribution, therefore in this lemma we used $\nu_* = e^{-f}$ instead of $\nu_* \propto e^{-f}$.*

Proof First we bound $\mathbb{E}_{y\sim\rho}[f(y) - f(x)]$ as follows.

$$\begin{aligned}
\mathbb{E}_{y\sim\rho}[f(y) - f(x)] &= \mathbb{E}_{y\sim\rho}\left[\int_0^1 \langle \nabla f(ty + (1-t)x), y - x\rangle \, dt\right] \\
&= \mathbb{E}_{y\sim\rho}\left[\int_0^1 \langle \nabla f(ty + (1-t)x) - \nabla f(x), y - x\rangle \, dt\right] \\
&\quad + \mathbb{E}_{y\sim p}\left[\int_0^1 \langle \nabla f(x), y - x\rangle \, dt\right] \\
&= \int_0^1 \mathbb{E}_{y\sim\rho}\left[\langle \nabla f(ty + (1-t)x) - \nabla f(x), y - x\rangle\right] dt \\
&\quad + \int_0^1 \langle \nabla f(x), \mathbb{E}_{y\sim\rho}[y - x]\rangle \, dt \\
&\leq \int_0^1 \mathbb{E}_{y\sim\rho}\left[t^\beta L\|y - x\|^{\beta+1}\right] dt \\
&\leq \frac{L}{\beta+1}\mathbb{E}_{y\sim\rho}\left[\|y - x\|^{\beta+1}\right] \leq \frac{L}{\beta+1}\mathbb{E}_{y\sim\rho}\left[\|y - x\|^2\right]^{\frac{\beta+1}{2}} \leq \frac{L}{\beta+1} d^{\frac{\beta+1}{2}}.
\end{aligned}$$

Using the previous formula, we bound the KL-divergence

$$\mathrm{KL}\big(\rho|\nu_*\big) = \int \rho(y)\log\left(\rho(y)\right)dy + \int \rho(y)f(y)dy = -\mathrm{H}\left(\rho\right) + \mathbb{E}_{y\sim\rho}\left[f(y) - f(x)\right] + f(x).$$

Using the previous bound and the formula for the Gaussian entropy concludes the proof. ■

Lemma 28 *For $a, b > 0$, the function $x \to a/x + bx^\theta$ is minimized at $x_* = (a/(\theta b))^{\frac{1}{1+\theta}}$ and the minimum value and an upper bound is given as*

$$\tfrac{1+\theta}{\theta^{\theta/(1+\theta)}} a^{\frac{\theta}{1+\theta}} b^{\frac{1}{1+\theta}} \leq 2a^{\frac{\theta}{1+\theta}} b^{\frac{1}{1+\theta}}.$$

Proof Taking derivative and setting it equal to zero yields the value for x_*. ■

Lemma 29 *If $0 \leq \gamma \leq 2$, then following inequality holds*

$$\|u + v\|^\gamma \leq 2(\|u\|^\gamma + \|v\|^\gamma).$$

Further, when $\gamma \leq 1$ the factor 2 on the right hand side can be omitted.

Proof The inequality follows from the fact that functions $h_1(x) = (x^\gamma + 1) - (1 + x)^\gamma$ and $h_2(x) = 2(x^\gamma + 1) - (1 + x)^\gamma$ are non-negative when $\gamma \in [0, 1]$ and $\gamma \in [0, 2]$, respectively. ■

Lemma 30 *Suppose $A, B, \alpha, \beta > 0$ and $\alpha > \beta$ and $f(x) = -Ax^\alpha + Bx^\beta$. The following upper bound on f holds when $x > 0$*

$$\sup_{x \geq 0} f(x) \leq B \left(\frac{B\beta}{A\alpha} \right)^{\frac{\beta}{\alpha - \beta}}.$$

Proof Setting the derivative equal to zero implies $x^{\alpha-\beta} = \frac{\beta B}{\alpha A}$. Plugging this into $f(x)$ we get $f(x) \leq Bx^\beta = B \left(\frac{B\beta}{A\alpha} \right)^{\frac{\beta}{\alpha-\beta}}$. Since $\alpha > \beta$ this function has a maximizer not a minimizer. ■

Lemma 31 (Stein's lemma Stein (1981)) *Suppose $x \sim \mathcal{N}(\mu, \sigma^2 I_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is weakly differentiable. Then, for $a \in \mathbb{R}^d$*

$$\mathbb{E}\left[\langle x - \mu, af(x) \rangle\right] = \sigma^2 \mathbb{E}\left[\mathrm{Tr}(\nabla\left[af(x)\right])\right] = \sigma^2 \mathbb{E}\left[\langle a, \nabla f(x) \rangle\right]$$

Lemma 32 *If $x_k \leq (1-a)x_{k-1} + b$ for $0 < a < 1$ and $0 \leq b$, then*

$$x_k \leq e^{-ak} x_0 + \frac{b}{a}. \tag{H.1}$$

Proof Recursion on $x_k \leq (1-a)x_{k-1} + b$ yields

$$x_k \leq (1-a)^k x_0 + b(1 + (1-a) + (1-a)^2 + \cdots + (1-a)^{k-1}) \leq (1-a)^k x_0 + \frac{b}{a}.$$

Using the fact that $1 - a \leq e^{-a}$, (H.1) is achieved. ■

H.1. Some Properties of Hölder Continuity

Lemma 33 *Let f be α-Hölder continuous with constant h_f^α and β-Hölder continuous with constant h_f^β and $0 < \beta < \alpha \leq 1$, then f is γ-Hölder with constant $h_f^\alpha \vee h_f^\beta$ when $\beta < \gamma < \alpha$.*

Proof We consider two cases based on $\|x - y\|$. First, when $\|x - y\| \leq 1$,

$$\|f(x) - f(y)\| \leq h_f^\alpha \|x - y\|^\alpha \leq h_f^\alpha \|x - y\|^\gamma \|x - y\|^{\alpha - \gamma} \leq h_f^\alpha \|x - y\|^\gamma.$$

For the second case, when $\|x - y\| > 1$,

$$\|f(x) - f(y)\| \leq h_f^\beta \|x - y\|^\beta \leq h_f^\beta \|x - y\|^\gamma \|x - y\|^{\beta - \gamma} \leq h_f^\beta \|x - y\|^\gamma.$$

Taking the maximum of constants in two cases completes the proof. ■

Lemma 34 *Let f be α-Hölder continuous with constant h_f^α and g be β-Hölder continuous with constant h_g^β and $\beta < \alpha \leq 1$. If the difference of f and g is bounded i.e. $\|f - g\|_\infty < B$ then f is β-Hölder with constant $h_f^\alpha \vee (2B + h_g^\beta)$. In a specific case, every bounded and Lipschitz function is τ-Hölder for $\tau \in (0, 1)$.*

Proof We consider two cases based on $\|x-y\|$. First, when $\|x-y\| \leq 1$,

$$\|f(x)-f(y)\| \leq h_f^\alpha \|x-y\|^\alpha \leq h_f^\alpha \|x-y\|^\beta \|x-y\|^{\alpha-\beta} \leq h_f^\alpha \|x-y\|^\beta.$$

For the second case, when $\|x-y\| > 1$,

$$\begin{aligned}\|f(x)-f(y)\| &\leq \|f(x)-g(x)\| + \|g(x)-g(y)\| + \|f(y)-g(y)\| \\ &\leq B + h_g^\alpha \|x-y\|^\beta + B \\ &\leq (2B + h_g^\alpha)\|x-y\|^\beta.\end{aligned}$$

Taking the maximum of constants in the two cases completes the proof. ■

Lemma 35 *The function $\|x\|^{\alpha-2}x$ is $\alpha-1$-Hölder for $1<\alpha<2$.*

Proof Without loss of generality, assume $\|y\| \leq \|x\|$ which implies $\|x-y\| \leq 2\|x\|$, which in turn implies $\|x\|^{\alpha-2} \leq 2^{2-\alpha}\|x-y\|^{\alpha-2}$. Therefore,

$$\begin{aligned}\|f(x)-f(y)\| &\leq \|\|x\|^{\alpha-2}x - \|y\|^{\alpha-2}y\| \\ &\leq \|\|x\|^{\alpha-2}x - \|x\|^{\alpha-1}\frac{y}{\|y\|} + \|x\|^{\alpha-1}\frac{y}{\|y\|} - \|y\|^{\alpha-2}y\| \\ &\leq \|x\|^{\alpha-1}\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\| + |\|x\|^{\alpha-1} - \|y\|^{\alpha-1}| \\ &\overset{1}{\leq} \|x\|^{\alpha-1}\|\frac{x}{\|x\|} - \frac{y}{\|x\|} + \frac{y}{\|x\|} - \frac{y}{\|y\|}\| + \|x-y\|^{\alpha-1} \\ &\leq \|x\|^{\alpha-2}\|x-y\| + \|x\|^{\alpha-1}\|\frac{y}{\|y\|}(\frac{\|y\|}{\|x\|}-1)\| + \|x-y\|^{\alpha-1} \\ &\leq 2\|x\|^{\alpha-2}\|x-y\| + \|x-y\|^{\alpha-1} \leq (1+2^{3-\alpha})\|x-y\|^\alpha \leq 5\|x-y\|^\alpha,\end{aligned}$$

where inequality 1 follows from Lemma 29. ■