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

Murat A. Erdogdu[*]   Rasa Hosseinzadeh[†]

May 28, 2020

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 convergence rate, 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 starts to have an effect on the established rate 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.

We establish the convergence rate of LMC by first proving a moment dependent modified log-Sobolev inequality with explicit constants for a class of target distributions that have convex degenerate potentials at infinity. Then, we prove linearly diverging estimates for any order moments of the Markov chain defined by the LMC algorithm, and show that this is sufficient to obtain the above convergence rate. Our framework also allows for finite perturbations, and any order of smoothness $\beta \in (0, 1]$; consequently, our results are applicable to a wide class of non-convex potentials that are weakly smooth and exhibit at least linear tail growth.

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 continuous-time diffusion process with invariant measure as the target. When the target distribution 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),

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

---

[*]Department of Computer Science and Department of Statistical Sciences at the University of Toronto, and Vector Institute

[†]Department of Computer Science at the University of Toronto, and Vector Institute

where B_t is a d-dimensional Brownian motion. It is straightforward to show that the Langevin diffusion (1.1) admits the target Gibbs measure ν_* as its invariant distribution [MCF15]. In general, simulating a continuous-time diffusion such as (1.1) is impractical; thus, a numerical integration scheme is needed to approximate it. Due to its simplicity, efficiency, and well-understood theoretical properties, algorithms based on Langevin diffusion have found numerous applications in sampling and optimization literature [WT11, Dal17a, RRT17, XCZG18, LWME19]. 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) and relies on the following update rule

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

where W_k is an isotropic Gaussian vector independent from W_m and x_m for $m < k$, and η is the step size. LMC defines a Markov chain which has an invariant measure that is different than the target ν_*, and this difference is often termed as the bias which is due to the numerical integration. The bias of a discretized diffusion such as LMC can be generally controlled with a smaller step size, where in the limit case $\eta \downarrow 0$, the iteration (1.2) scaled with $1/\eta$ reduces to the SDE (1.1).

Convergence rate of LMC has been the focus of recent research. Rates are established under structural assumptions on the potential function, and they quantify the number of iterations required 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 rates under the global curvature assumptions on the potential function. For example, for strongly convex and smooth potentials, the convergence rate of $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ has been shown [Dal17b], whereby smooth function is a function with Lipschitz continuous gradient. We note that higher-order smoothness on the potential function may improve the convergence rate [MFWB19b]; however, we consider only the first-order smoothness in the current paper. For convex and smooth potentials with growth rate α, a convergence rate of $\widetilde{\mathcal{O}}(d^{1+4/\alpha}\epsilon^{-3})$ is known to hold for LMC [CB18]. More recently, however, it has been observed that tail growth structure is the determinant factor in sampling [CCAY+18, Ebe16, EMS18, EGZ19], 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 finite perturbations, which in turn allows for sampling from non-convex potentials with a wide range of modern applications in statistics. A condition on the target distribution ν_* that fits in this framework is the *log-Sobolev inequality* (LSI) of Bakry and Emery [BE85], which can be written as

$$\forall \rho, \quad \mathrm{H}\big(\rho|\nu_*\big) \le \lambda \mathrm{I}\big(\rho|\nu_*\big), \tag{1.3}$$

where $\mathrm{H}\big(\rho|\nu_*\big)$ denotes the KL-divergence (relative entropy) and $\mathrm{I}\big(\rho|\nu_*\big)$ denotes the relative Fisher information between ρ and ν_*, 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 [BE85], and it allows for finite perturbations due to Holley-Stroock perturbation lemma [HS87]; thus, potentials that have quadratic growth can be shown to satisfy LSI (this will be made precise later). Denoting the distribution of Langevin diffusion (1.1) at time t with ρ_t, it is known that $\frac{d}{dt}\mathrm{H}\big(\rho_t|\nu_*\big) = -\mathrm{I}\big(\rho_t|\nu_*\big)$ which, combined with the LSI condition (1.3) entails a differential inequality of the form $\frac{d}{dt}\mathrm{H}\big(\rho_t|\nu_*\big) \le -\frac{1}{\lambda}\mathrm{H}\big(\rho_t|\nu_*\big)$, which in turn yields an exponential contraction in KL-divergence, i.e., $\mathrm{H}\big(\rho_t|\nu_*\big) \le e^{-t/\lambda}\mathrm{H}\big(\rho_0|\nu_*\big)$ for the diffusion process.

An important implication of a condition like LSI (1.3) in the context of sampling with LMC is that fast mixing properties of the continuous-time diffusion induced by LSI are inherited by the discrete algorithm LMC. That is, LSI coupled with the smoothness condition on the potential is sufficient to obtain the fast convergence rate $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ [VW19], which is the best known rate for

2

LMC 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 perturbations and consequently sampling from non-convex potentials.

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

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

where $r : [0, \infty) \to [0, 1]$ is a monotone decreasing function which is typically 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 order of smoothness of the potential. Elementary algebra reveals that, one can iterate the inequality (1.4) and achieve convergence as long as $r(\eta) < 1$. Recent literature focused on exponential decays $r_{\exp}(t) = e^{-\alpha t}$ which are usually established under conditions like LSI (1.3) or strong convexity that correspond to potentials exhibiting quadratic growth (see, for example [Dal17b, VW19]). Nevertheless, the inequality (1.4) by no means benefits from the exponential decay, as $r(t)$ is only evaluated at short time horizons $t = \eta$. Indeed, any decreasing rate function $r(t)$ satisfying $r(0) = 1$ and $r'(0) < 0$ would achieve the same rate of convergence as exponential decay $r_{\exp}(t)$. For example, consider the algebraic rate $r_{\mathrm{alg}}(t) = 1/(1 + \alpha t)$ 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

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

However, in contrast to exponential decay, algebraic rates can be obtained under much weaker tail growth 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 [BGL13], and can be used to explain different convergence behavior. For example in the seminal work by [TV00], 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)

$$\forall \rho, \quad H(\rho|\nu_*) \leq \lambda I(\rho|\nu_*)^{1-\delta} M_s(\rho + \nu_*)^\delta \quad \text{with} \quad \delta \in [0, 1/2), \tag{1.6}$$

where $M_s(\rho) = \int (1 + \|x\|^2)^{s/2} \rho(x) dx$ is the s-th moment of any function ρ. This inequality entails a decay with the desired properties (1.5) under mild conditions on the potential. By further assuming that the gradient of the potential is β-Hölder continuous and carefully tuning the moment order $s = \mathcal{O}(\log(d/\epsilon))$ in mLSI (1.6), we can prove that, even with linearly diverging moment estimates for the LMC iterates, the algorithm is guaranteed to reach the ϵ-neighborhood of a d-dimensional target ν_* in KL-divergence after taking the advertised number of steps $N = \widetilde{\mathcal{O}}\big(d^{\frac{1}{\beta} + \frac{1+\beta}{\beta}(\frac{2}{\alpha} - \mathbb{1}_{\{\alpha \neq 1\}})} \epsilon^{-\frac{1}{\beta}}\big)$. In moderate dimensions $d \ll \epsilon^{-1}$, this convergence rate does not depend on the tail growth rate α, and it is controlled solely by the order of smoothness β, whereas in high dimensions $d = \mathcal{O}(\epsilon^{-1})$, the rate is determined by an interplay between the growth rate and the order of smoothness. The above rate also recovers the best known rate which was established for smooth potentials ($\beta = 1$) under the LSI condition (1.3) where the tail growth is quadratic ($\alpha = 2$).

Our contributions can be summarized as follows.

- For a potential function f whose tails behave like $\|x\|^{\alpha}$, and has β-Hölder continuous gradient, i.e.,

$$f(x) \sim \|x\|^{\alpha} \quad \text{for} \quad \alpha \in [1,2], \quad \text{and} \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|^{\beta} \quad \forall x,y, \qquad (1.7)$$

  we prove that LMC achieves the convergence rate $\widetilde{\mathcal{O}}\big(d^{\frac{1}{\beta} + \frac{1+\beta}{\beta}(\frac{2}{\alpha} - \mathbb{1}_{\{\alpha \neq 1\}})}\epsilon^{-\frac{1}{\beta}}\big)$ in KL-divergence. In moderate dimensions when $d \ll \epsilon^{-1}$, the tail growth rate α does not impact the convergence rate, whereas in high dimensions where $d = \mathcal{O}(\epsilon^{-1})$, tail growth enters the convergence rate through dimension dependency.
- As a key step in deriving the above convergence rate, we establish a modified log-Sobolev inequality (mLSI) (1.6) with explicit constant λ, and a target dependent moment function $M_s(\rho + \nu_*)$ for any order $s \geq 2$. Both of these are crucial in deriving a convergence rate with correct dependence on the dimension d as well as the accuracy ϵ. The final convergence result is obtained by employing the mLSI condition (1.6) for the optimal moment order $s = \mathcal{O}(\log(d/\epsilon))$.
- In lieu of (1.7), we are mainly interested in potentials exhibiting weak dissipativity, i.e.,
$$\langle x, \nabla f(x) \rangle \geq a\|x\|^{\alpha} - b \quad \text{with} \quad \alpha \in [1,2), \quad a,b > 0. \qquad (1.8)$$

  In order to use the condition mLSI (1.6), we establish linearly diverging moment estimates for the LMC iterates under (1.8). Somewhat surprisingly, this is sufficient to establish the convergence of LMC in KL-divergence.
- Our convergence results are valid under finite perturbations of the potential; consequently, they cover sampling from non-convex potentials with at least linear growth. Furthermore, our results also cover the case $\beta < 1$ for which the potential function is not smooth; more specifically, it does not have a Lipschitz gradient. To the best of our knowledge, this is the first convergence result for the LMC algorithm for weakly smooth potentials that exhibit subquadratic growth, which does not rely on regularization or Gaussian smoothing.
- Finally, using Csiszár-Kullback-Pinsker inequalities, the above convergence rates obtained under KL-divergence can be translated to estimates in 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. Section 1.1 reviews our notation, and Section 2 surveys the related work with a detailed comparison on the existing convergence rates. In Section 3, we establish the main technical results on the convergence of LMC for potentials with certain growth and smoothness properties. Section 4 discusses further implications of the tools developed in Section 3. We give concrete examples in Section 5, by applying these tools to non-convex sampling problems that are also weakly smooth. Proofs of the main theorems and corollaries are provided in Sections 6, 7, 8, 9 and 10 in order of appearance of their statement in the main text. Finally, we conclude in Section 11 with brief remarks on future work.

1.1 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. For a matrix $A \in \mathbb{R}^{d \times k}$, A_{ij} denotes its entry in the i-th row and j-th column, and whenever $d = k$, its trace is denoted by $\mathrm{Tr}(A) = \Sigma_{i=1}^{d} A_{ii}$. 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 $\|f\|_{\infty} = \sup_{x \in \mathbb{R}^d} |f(x)|$. $M_s(f)$ is used to denote the modified s-th moment of the function f (which is not necessarily a distribution),

4

defined as $M_s(f) = \int f(x)(1 + \|x\|^2)^{s/2}dx$. The gradient and the Hessian of f are denoted by $\nabla f(x)$ and $\nabla^2 f(x)$, respectively, where the derivatives are with respect to x. For a statement A, the indicator function is denoted with $\mathbb{1}_{\{A\}}$ and defined as

$$\mathbb{1}_{\{A\}} = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

We use $\mathbb{E}[x]$ to denote the expected value of a random variable x, where expectations are over all the randomness inside the brackets. For probability densities p, q on $\mathbb{R}^d$, we use $H(p|q)$ and $I(p|q)$ to denote their KL-divergence (or relative entropy) and relative Fisher information, respectively, which are defined as

$$H(p|q) = \int p(x) \log \frac{p(x)}{q(x)} dx, \quad \text{and} \quad I(p|q) = \int p(x) \left\| \nabla \log \frac{p(x)}{q(x)} \right\|^2 dx.$$

Similarly, we denote the entropy of p with $H(p) = -\int p(x) \log p(x) dx$. 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

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

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.

Finally, $\mathcal{O}$ and $\widetilde{\mathcal{O}}$ notations are frequently used to describe the dependence of a function f on another function g, and defined in the following sense

$$f(x) = \mathcal{O}(g(x)) \implies \limsup_{x \to \infty} \frac{f(x)}{g(x)} < \infty, \quad \text{and} \quad f(x) = \widetilde{\mathcal{O}}(g(x)) \implies \limsup_{x \to \infty} \frac{f(x)}{g(x) \log(g(x))^k} < \infty,$$

for some $k \geq 0$, where $\widetilde{\mathcal{O}}$ simply ignores the logarithmic factors. We use $f(x) \lesssim g(x)$ instead of $f(x) \leq \mathcal{O}(g(x))$ to improve readability.

2 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 [Mil94, MT13], and the convergence guarantees were mostly asymptotic [GM91, MT12]. Non-asymptotic analysis of LMC has drawn a lot of interest recently [DT12, Dal17a, Dal17b, DMM19, CB18, CCAY+18, VW19, DK19, BDMS19] where the focus was on potentials exhibiting strong tail growth properties. These papers were mostly influenced by the pioneering works by Dalalyan [Dal17b], and Durmus and Moulines [DM16, DM17] 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, $\widetilde{\mathcal{O}}(d\epsilon^{-2})$ steps are sufficient to reach ϵ accuracy under the L_2-Wasserstein distance [DM19], which can be further improved to $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ under an additional second-order smoothness assumption on the potential function.

In this paper, we establish guarantees under KL-divergence (relative entropy) which can be easily translated to TV and Wasserstein metrics using Csiszár-Kullback-Pinsker (CKP) [BV05] and/or Talagrand inequalities [Tal96, OV00]. For strongly convex and smooth potentials, it is known that $\widetilde{\mathcal{O}}(d\epsilon^{-1})$ steps of LMC yield an ϵ accurate sample in KL-divergence [CB18, DMM19].

This is still the best known rate in this setup, and recovers the best known rates in TV [DM17, Dal17b] as well as in L_2-Wasserstein metrics [DM19]. However, for convex and smooth potentials that grow like $\|x\|^\alpha$, the rate drops to $\widetilde{\mathcal{O}}\big(d^{1+\frac{4}{\alpha}}\epsilon^{-3}\big)$ due to lack of strong convexity [CB18]. Among various contributions of [DMM19], 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 [Ebe16, EGZ19]. For example, [CCAY$^+$18] 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, [VW19] established convergence guarantees for target distributions that satisfy a log-Sobolev inequality. This corresponds to potentials with quadratic tails [BE85, BG99] up to finite perturbations [HS87]; thus, this result is able to deal with non-convex potentials that are not limited to a compact set, while achieving the same convergence rate of $\widetilde{\mathcal{O}}\big(d\epsilon^{-1}\big)$ 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 [Atc15, LFC17, DMP18, MFWB19a, DMM19], Gaussian smoothing [CDJB19, DDN20], or mirror mapping [HKRC18]. 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 convex function with β-Hölder continuous gradient. This assumption enforces a quadratic tail growth on the potential, in which case, [CDJB19] established the convergence rate of $\widetilde{\mathcal{O}}\big(d^{2+1/\beta}\epsilon^{-2/\beta}\big)$ in total variation distance. Furthermore, we focus on the last LMC iterate which characterizes the practical performance of this algorithm, in contrast to [DMM19] which provided guarantees for the average of the distributions of the LMC iterates.

Our analysis draws heavily on the theory of diffusion processes [BGL13, TV00] – more specifically, logarithmic Sobolev inequalities. These inequalities were first established for the Gaussian density [Gro75], and later generalized to Gibbs measure with a strongly convex potential by Bakry and Émery [BE85]. Combined with the Holley and Stroock's perturbation lemma [HS87], this theory covers a wide range of 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 [JKO98]. Building on this, sampling with a diffusion can be seen as an optimization algorithm in the space of probability distributions [Wib18, VW19, MCC$^+$19]; similarly, LSI can be interpreted as a gradient domination condition in this space, which is commonly referred to as the PL-inequality [Pol63] in the optimization theory. LSI and PL-inequality both yield exponential convergence in their corresponding space [Pol63, KNS16, Tos99, CS91]. Further promoting this analogy, PL-inequality is a special case of Łojasiewicz inequality [Loj63], and their counterparts are considered recently in [BB18] in the space of functionals. Thus, the modified LSI introduced in [TV00], can be viewed as a modified version of the Łojasiewicz inequality in the space of probability distributions. For a survey about the convergence properties of diffusion processes with the Fokker-Planck equation governing their evolution (including overdamped Langevin dynamics (1.1)) and several inequalities from functional analysis, we refer the reader to [MV99]. Finally, the analogy between optimization and sampling provided invaluable insights, in many cases improving our understanding, and ultimately the performance of various algorithms [ZLC17, BDMP17, BDM18, CFM$^+$18, BMD$^+$19, HKRC18, MCJ$^+$19].

It is worth mentioning that the rates we discussed in this section can be further improved by

WORK	CONVERGENCE RATE	SMOOTHNESS	CURVATURE	PERTURBATION	DISTANCE
[CB18, DMM19]	$\widetilde{\mathcal{O}}(d\epsilon^{-1})$	Lipschitz gradient	Strongly Convex	None	KL
[VW19]	$\widetilde{\mathcal{O}}(d\epsilon^{-1})$	Lipschitz gradient	Strongly Convex	Bounded difference	KL
[CB18]	$\widetilde{\mathcal{O}}(d^{1+\frac{4}{\alpha}}\epsilon^{-3})$	Lipschitz gradient	Convex Growth rate α	None	KL
This work	$\widetilde{\mathcal{O}}(d^{\frac{1}{\beta}+\frac{1+\beta}{\beta}(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq1\}})}\epsilon^{-\frac{1}{\beta}})$	β-Hölder gradient	Tail growth $\sim\|x\|^{\alpha}$	Bounded difference	KL
[Dal17b, DM17]	$\widetilde{\mathcal{O}}(d\epsilon^{-2})$	Lipschitz gradient	Strongly convex	None	TV
[Dal17b]	$\widetilde{\mathcal{O}}(d^3\epsilon^{-4})$	Lipschitz gradient	Convex	None	TV
[CDJB19]	$\widetilde{\mathcal{O}}(d^{2+\frac{1}{\beta}}\epsilon^{-\frac{2}{\beta}})$	Lipschitz+β-Hölder gradient	Strongly Convex	None	TV
This work	$\widetilde{\mathcal{O}}(d^{\frac{1}{\beta}+\frac{1+\beta}{\beta}(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq1\}})}\epsilon^{-\frac{2}{\beta}})$	β-Hölder gradient	Tail growth $\sim\|x\|^{\alpha}$	Bounded difference	TV

Table 1: List of convergence rates in KL-divergence and TV distance for the LMC (1.2) algorithm in various papers and their accompanying assumptions. Comparison is made with results relying only on first order smoothness. For additional information, refer to Section 2.1.

making higher order smoothness assumptions on the potential function [MFWB19b], or by considering higher order numerical integrators [LWME19, SL19], or by certain adjustments [DRV+17, GLR18, DCWY19]. The overdamped Langevin diffusion (1.1) considered in this work is first order, and its higher order versions such as underdamped [CCBJ18, MCC+19], or third-order schemes [MCF15, MMW+19] may also provide additional improvements.

2.1 Comparison

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, [Dal17b, DM17, CB18, CDJB19, VW19, DMM19] are in the quadratic growth regime, and achieve the best rates known to authors. Our results recover the convergence rate of [VW19] for smooth potentials ($\beta = 1$) satisfying the LSI condition ($\alpha = 2$). [CB18, Dal17b] establish guarantees for convex and smooth potentials; however, these rates drop significantly under lack of strong convexity, 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 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 convergence loses an additional factor in dimension dependency. The results of [CDJB19] 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 [CB18] is $\widetilde{\mathcal{O}}(d\epsilon^{-3} \times \mathcal{W}_2^4(\rho_0, \nu_*))$, 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 Table 1. For details of the initializations when $\alpha = 2$, we refer to [CCBJ18].

3 Main Results

Convergence rates of diffusion-based algorithms have been the subject of growing attention recently with many applications related to sampling with MCMC and non-convex optimization. Algorithms based on Langevin diffusion have been particularly of interest where the fast convergence of the algorithm has been frequently linked to the quadratic growth of the potential [EMS18, CCAY$^+$18, VW19]. Conversely and somewhat surprisingly, we prove that the rate of convergence, in terms of its dependence on the accuracy ϵ, is not directly influenced by the growth behavior of the potential function f as long as the growth is at least linear. Therefore in moderate dimensions, the smoothness properties of the potential entirely determines the rate. Furthermore, our results show that the tail growth rate of the potential impacts the performance in high dimensions, as it enters the convergence rate only through dimension dependency, in which case the convergence is determined by the interplay between the tail growth rate and the weak smoothness degree.

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 $\xi \geq 0$*

$$\|f - \tilde{f}\|_\infty \leq \xi,$$

where $\tilde{f}$ satisfies,

$$\nabla^2 \tilde{f}(x) \succeq \frac{\mu}{\left(1 + \frac{1}{4}\|x\|^2\right)^{\theta/2}} I_d, \tag{3.1}$$

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

The above condition allows for finite perturbations, and consequently permits sampling from non-convex potentials; thus, the determinant factor is the tail growth properties of the potential function. The boundary case $\theta = 0$ corresponds to quadratic tail growth, and whenever $\theta > 0$, due to the decaying nature of the lower bound on the Hessian, the potential function exhibits no convexity at infinity. For example, consider 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 $\theta > 2$.

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. [BG99]); 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 $s \geq 2$,*

$$\forall \rho, \quad \mathrm{H}(\rho|\nu_*) \leq \lambda \mathrm{I}(\rho|\nu_*)^{1-\delta} \mathrm{M}_s(\rho + \nu_*)^\delta, \tag{3.2}$$

where $\mathrm{M}_s(\rho) = \int (1 + \|x\|^2)^{s/2}\rho(x)dx$ is the s-th moment of any function ρ, and δ and λ are constants that depend on s, and defined as

$$\delta \triangleq \frac{\theta}{s - 2 + 2\theta} \in [0, 1/2),$$

$$\lambda \triangleq 4e^{2\xi}\mu^{-\frac{s-2}{s-2+2\theta}}.$$

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. Notice that the moment term $\mathrm{M}_s(\rho + \nu_*)$ depends on both ρ and ν_*, which is crucial in deriving a convergence rate with correct dimension and accuracy dependence. Modified LSI-type inequalities such as (3.2) as well as weak Poincaré inequalities appear in the analysis of diffusion operators [BGL13]. The mLSI condition (3.2) is similar in nature to the modified LSI of [TV00]; 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 convergence rate that is worse than what will be established below in Corollary 4. It also cannot recover the existing rates (e.g. [VW19]) in the limit case $\alpha \to 2$. Our proof builds on the construction made in [TV00] and uses the results of [BE85, HS87], which we defer to Section 6.

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 should be seen as a relaxation to the standard 2-dissipativity condition, $\langle \nabla f(x), x \rangle \geq a\|x\|^2 - b$ for some $a, b > 0$ [MSH02, MT12].

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

$$\langle \nabla f(x), x \rangle \geq a\|x\|^\alpha - b \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{3.3}$$

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

$$\|\nabla f(x)\| \leq M(1 + \|x\|^\zeta) \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{3.4}$$

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 [EMS18], whereas in the latter case, obtaining such a uniform bound is an open problem. 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 convergence rate for 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 (3.1) implies (3.3) in Assumption 2 (cf. Lemma 25); 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 diffusion processes $\{\widetilde{x}_{k,t}\}_{k\in\mathbb{N}, t\geq 0}$ where each iteration x_{k+1} of the LMC algorithm (1.2) can be written as $\widetilde{x}_{k,\eta}$ where

$$d\widetilde{x}_{k,t} = -\nabla f(x_k)dt + \sqrt{2}dB_t \quad \text{with} \quad \widetilde{x}_{k,0} = x_k, \tag{3.5}$$

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, $d\mathrm{H}(\widetilde{\rho}_{k,t}|\nu_*)/dt$, reduces to the negative relative Fisher information $-\mathrm{I}(\widetilde{\rho}_{k,t}|\nu_*)$ 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 [VW19, Proof of Lemma 3]), which yields the inequality

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

Combining this with mLSI (3.2) for $\rho = \widetilde{\rho}_{k,t}$, one obtains the following differential inequality for the interpolating diffusion process (3.5),

$$\frac{d}{dt} \mathrm{H}\big(\widetilde{\rho}_{k,t} | \nu_*\big) \leq -\frac{3}{4\lambda} \mathrm{H}\big(\widetilde{\rho}_{k,t} | \nu_*\big)^{\frac{1}{1-\delta}} \mathrm{M}_s\big(\widetilde{\rho}_{k,t} + \nu_*\big)^{-\frac{\delta}{1-\delta}} + \mathbb{E}\big[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\big]. \qquad (3.6)$$

The convergence rate of LMC can be derived by analyzing the differential inequality (3.6), which requires appropriate estimates on the Markov chain moments $\mathrm{M}_s(\widetilde{\rho}_{k,\eta} + \nu_*)$ defined by the iterates of LMC as well as the additive error $\mathbb{E}\big[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\big]$. The following proposition establishes the former with a linearly growing upper bound in the number of iterations. We defer to proof to Section 7.

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*

$$\mathrm{M}_s(\rho_k + \nu_*) \leq \mathrm{M}_s(\rho_0 + \nu_*) + C_s k\eta, \quad \text{for even integer } s \geq 2, \qquad (3.7)$$

where

$$C_s \triangleq \left(\frac{3a + 2b + 3}{1 \wedge a}\right)^{\frac{s-2}{\alpha}+1} s^s d^{\frac{s-2}{\alpha}+1}. \qquad (3.8)$$

Although the bound (3.7) 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 (3.8) is of order $\mathcal{O}(d^{\frac{s-2}{\alpha}+1})$ which is the same order as in the continuous-time case (cf. Lemma 12), and will help obtain an accurate dimension dependency in the final convergence rate.

In the following, 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 (3.6). In this context, order of smoothness refers to the Hölder exponent of the gradient of the potential, which is defined explicitly below.

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

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|^\beta \quad \text{for all} \quad x, y \in \mathbb{R}^d, \qquad (3.9)$$

where the order of smoothness β satisfies $\zeta \leq \beta \leq 1$ for the constant ζ in (3.4).

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 [CDJB19, Nes15], a term that is borrowed from optimization theory. Our results cover potentials satisfying (3.9) for any $\beta \in (0,1]$. In the case of weakly smooth potentials, existing results on vanilla LMC can only explain the convergence in the setting $\alpha = 2$, but even for this case, the conclusions that will be made in Corollary 4 are new and improves the best known rates.

We reiterate that the main use of the above assumption is to control the additive error term due to discretization, the second term on the right hand side of (3.6), with a bound that behaves like η^β (cf. Lemma 17). This term determines the accuracy ϵ dependence of the convergence rate.

β-Hölder continuity already imposes a growth condition on the gradient (3.4) 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 $f(x) = (1 + x^2)^{1/2}$ 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. Hence, keeping the gradient growth and the order of smoothness separate allows us to cover a wider range of potentials. The relationship among these parameters can be summarized as given below,

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

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 $\beta \geq 1$, which limits the applicability of the results to only smooth potentials with Lipschitz gradient.

Before we present the main technical result of this paper, we note that when $\alpha > 1$, all these assumptions are satisfied for potentials of the form $f(x) = \|x\|^\alpha + c$, and their bounded perturbations, e.g.

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

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.

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 (8.10), and for some $\Delta_0 > 0$ upper bounding the error at initialization, i.e. $\mathrm{H}(\rho_0|\nu_*) \leq \Delta_0$, if the step size satisfies*

$$\eta = (\sigma c_\gamma)^{-\frac{1}{1+\beta}} d^{-\frac{\alpha+\theta}{\alpha\beta} - \frac{\gamma}{\beta+1}} (1 + (1 - \alpha/2)\log(d))^{-\frac{1}{\beta}} \log\left(\frac{\Delta_0}{\epsilon}\right)^{-\frac{\gamma}{1+\beta}} \left(\frac{2}{\epsilon}\right)^{-\frac{1}{\beta} - \frac{\gamma}{1+\beta}}, \tag{3.10}$$

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

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

where γ is given by

$$\gamma \triangleq \gamma(s) = \frac{(1+\beta)\theta}{\beta(s-2)} \quad \text{for any even integer } s \geq 4, \tag{3.11}$$

and σ and c_γ are constants given as

$$\sigma = 4L^2\left(1 + 2a^\beta\left[1 + \tfrac{2\alpha}{a}\left(\log(16\pi/a) + M(2 + 2b/a)^2 + b + |f(0)|\right)\right]\right),$$

$$c_\gamma = \sigma^{\frac{1}{\beta}}(16\lambda)^{1 + \frac{1}{\beta} + 2\gamma}\left(\frac{\mathrm{M}_s(\rho_0 + \nu_*)}{16d^{\frac{s-2}{\alpha}+1}} \vee \frac{s^s}{16}\left(\frac{3a + 2b + 3}{1 \wedge a}\right)^{\frac{s-2}{\alpha}+1}\right)^\gamma.$$

11

The above theorem, proved in Section 8, implies that for smooth potentials that satisfy the LSI condition (1.3), i.e. $\alpha = 2$ and $\beta = 1$, we have $\gamma = \theta = 0$; thus, LMC achieves the convergence rate of $\widetilde{\mathcal{O}}(d\epsilon^{-1})$, recovering the rate established by [VW19]. In the general case, Theorem 3 implies the convergence rate $\widetilde{\mathcal{O}}(\gamma^{-1} d^{\frac{\alpha+\theta+\beta\theta}{\alpha\beta}+\gamma} \epsilon^{-\frac{1}{\beta}-\gamma})$ where $\gamma > 0$ is given in (3.11) and can be arbitrarily small. However, note that 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 bound on accuracy (8.10) is $\mathcal{O}(1)$, depending on the fixed problem parameters and the bound on the initial KL-divergence Δ_0. When initialized with a Gaussian, Δ_0 can also be characterized with the fixed problem parameters (cf. Lemma 26). More importantly, the upper bound on ϵ, as stated in (8.10), does not depend on the moment order s, which enables us to choose $s = \mathcal{O}(\log(d\epsilon^{-1}))$ and accordingly $\gamma = \mathcal{O}(1/\log(d\epsilon^{-1}))$, which in turn yields the optimal convergence rate. 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 (cf. Lemma 26). Then, for a sufficiently small ϵ satisfying $\epsilon \leq \psi$ where ψ is defined in (8.10), if the step size satisfies (3.10) for $s = 2 + 2\lceil \log(\frac{6d}{\epsilon}) \rceil$, the iterates of LMC reaches ϵ-accuracy of the target, i.e. $\mathrm{H}(\rho_N|\nu_*) \leq \epsilon$, after N steps satisfying*

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

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

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

The above corollary that is proved in Section 10, implies that the LMC algorithm achieves ϵ accuracy of the target in KL-divergence in $\widetilde{\mathcal{O}}(d^{\frac{\alpha+(1+\beta)\theta}{\alpha\beta}}\epsilon^{-\frac{1}{\beta}})$ 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. Therefore, in this case the convergence rate is simply $\widetilde{\mathcal{O}}(d^{\frac{2}{\alpha}(1+\frac{1}{\beta})-1}\epsilon^{-\frac{1}{\beta}})$. On the other hand, when the potential has linear tail growth (i.e. $f(x) \sim \|x\|$), by setting $\tilde{f} = (1+\|x\|^{1+\tau})^{1/(1+\tau)}$ 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 a convergence rate of $\widetilde{\mathcal{O}}(d^{2+\frac{3}{\beta}}\epsilon^{-\frac{1}{\beta}})$. Putting this all together, one can simply use $\theta = 2 - \alpha\mathbb{1}_{\{\alpha\neq 1\}}$, which yields the advertised convergence rate $\widetilde{\mathcal{O}}(d^{\frac{1}{\beta}+(1+\frac{1}{\beta})(\frac{2}{\alpha}-\mathbb{1}_{\{\alpha\neq 1\}})}\epsilon^{-\frac{1}{\beta}})$. We emphasize that, in moderate dimensions where $d \ll \epsilon^{-1}$, the rate only depends on the order of smoothness β, whereas in high dimensions where $d = \mathcal{O}(\epsilon^{-1})$, the tail growth rate α enters the rate through the dimension dependence.

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. 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, $\theta = 1 \neq 2 - \alpha\mathbb{1}_{\{\alpha\neq 1\}}$ 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 convergence rate stays the same. See Section 5 for a more detailed discussion with examples demonstrating these observations.

12

4   Further Implications

4.1   Convex potentials

The next proposition shows that convex potentials have at least linear growth.

Proposition 5. *For any differentiable convex potential $f : \mathbb{R}^d \to \mathbb{R}$, there exist constants $a, b > 0$ such that*

$$f(x) \geq a\|x\| - b, \quad \textit{for all} \quad x \in \mathbb{R}^d.$$

The proof of the above Proposition is deferred to Section 9, and it follows from showing that any convex potential is in fact coercive.

If a convex potential has tail growth rate α, i.e. $f(x) \geq a\|x\|^\alpha - b$ for some $a, b > 0$, then it is straightforward to show that α-dissipativity holds using the Taylor's theorem. Hence, one can argue that for convex potentials, the limiting factor for the applicability of our results is the order of smoothness of the potential function f. For example, consider $f_1(x) = \sqrt{1 + \|x\|^2}$ and $f_2(x) = \|x\|$. They are both convex with linear growth and 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. Therefore, our results, in particular Corollary 4, are applicable to f_1, but not to f_2.

4.2   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 perturbations. The following lemma formalizes this statement.

Lemma 6. *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 $\alpha = 1$.

Proof of Lemma 6.   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 still correct because $\kappa_2 < a$. 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. □

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 ξ and θ), which means that the convergence rate of LMC for the perturbed potential is the same as that for the original potential.

Adding a non-convex perturbation to a convex potential might result in a non-convex potential; consider for example the potential $f(x) = \|x\|^\alpha + 5\cos(\|x\|)$. The previous lemma implies that the convergence rate remains the same which means that the main factor in the performance in high dimensions is the tail growth rate rather than global curvature properties like (strong) convexity. It is worth noting that the order of smoothness (cf. Assumption 3) is a global assumption unlike the tail growth condition. For example, $\|x\|$ has a discontinuous gradient at the origin; therefore, Assumption 3 fails for any β.

4.3   Other probability metrics

In this section, we use Csiszár-Kullback-Pinsker (CKP) inequalities [BV05] as well as Talagrand's inequality to translate our result from KL-divergence to other measures of distance such as total variation (TV) and L_α-Wasserstein metrics. The proofs are straightforward, and postponed to Section 10. 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, which we refer to an explicit statement in their respective results.

Corollary 7 (Total variation distance)**.** *Instantiate the assumptions and notation of Theorem 3. For a sufficiently small ϵ satisfying $\epsilon \leq \sqrt{\psi/2}$ where ψ is defined in* (8.10)*, if LMC* (1.2) *is initialized with $x_0 \sim \mathcal{N}(x, I_d)$ for any $x \in \mathbb{R}^d$, then,*

$$N = \widetilde{\mathcal{O}}\big(d^{\frac{\alpha+\theta+\beta\theta}{\alpha\beta}}\epsilon^{-\frac{2}{\beta}}\big)$$

steps are sufficient to obtain $\mathrm{TV}(\rho_N, \nu_*) \leq \epsilon$.

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 above corollary recovers the convergence rate $\widetilde{\mathcal{O}}(d\epsilon^{-2})$ in TV distance, which was established in [DM17].

The next corollary establishes the convergence rate in terms of L_α-Wasserstein distance.

Corollary 8 (L_α-Wasserstein distance)**.** *Instantiate the assumptions and notation of Theorem 3. For a sufficiently small ϵ satisfying* (10.1)*, if LMC* (1.2) *is initialized with $x_0 \sim \mathcal{N}(x, I_d)$ for any $x \in \mathbb{R}^d$, then,*

$$N = \widetilde{\mathcal{O}}\big(d^{\frac{3\alpha+\theta+\beta\theta}{\alpha\beta}}\epsilon^{-\frac{2\alpha}{\beta}}\big)$$

steps are sufficient to obtain $\mathcal{W}_\alpha(\rho_N, \nu_*) \leq \epsilon$.

For functions that have quadratic growth, the above corollary is not optimal, because our result relies on the CKP inequality (cf. Lemma 24) which does not recover Talagrand's inequality when $\alpha = 2$ [BV05]. Therefore, the case $\alpha = 2$ is handled separately, where Talagrand's inequality is available, which ultimately yields the following improved rate.

Corollary 9 (L_2-Wasserstein distance)**.** *Suppose f is a smooth potential with quadratic growth i.e. a potential satisfying Assumption 1 with $\theta = 0$, Assumption 2 with $\alpha = 2$ and Assumption 3 with $\beta = 1$. For a sufficiently small ϵ satisfying* (10.2)*, and for a step size satisfying* (3.10)*, if the LMC algorithm* (1.2) *is initialized with an error upper bounded by Δ_0, i.e.* $\mathrm{H}(\rho_0|\nu_*) \leq \Delta_0$*, then,*

$$N = \widetilde{\mathcal{O}}\big(d\epsilon^{-2}\big)$$

14

steps are sufficient to obtain $\mathcal{W}_2(\rho_N, \nu_*) \le \epsilon$.

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 above corollary 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.

5   Applications

In this section, we apply the results of Sections 3 and 4 to various illustrative potential functions. We begin with a few basic examples in order to demonstrate the effect of tail growth and order of smoothness on the convergence of LMC.

5.1   Pedagogical examples

Example 1 (Weakly smooth potential with subquadratic tails) Consider the potential function $f(x) = \|x\|^\alpha$ 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. In the following, we show that $f(x)$ satisfies Assumptions 1, 2 and 3; hence, the results of Section 3 can provide a convergence rate for sampling from this particular potential using the LMC algorithm. For Assumption 1, consider the following function $\tilde{f}(x) = \left(1 + \|x\|^2\right)^{\alpha/2}$. Since $\nabla^2 \tilde{f}(x) \succeq \frac{\alpha(\alpha-1)}{2}(1 + \|x\|^2/4)^{\alpha/2-1} I_d$ and $\|f - \tilde{f}\|_\infty \le 1$, the function f satisfies Assumption 1 with $\theta = 2 - \alpha$. Simple calculation shows that (3.3) in Assumption 2 is also satisfied. For Assumption 3, we have $\nabla f(x) = \alpha x \|x\|^{\alpha-2}$, which is β-Hölder with $\beta = \alpha - 1$ (cf. Lemma 34), which also implies that the growth condition in Assumption 2 holds with $\zeta = \alpha - 1$. Therefore, Corollary 4 implies that we can reach ϵ accuracy in KL-divergence after taking $\widetilde{\mathcal{O}}\left(d^{\frac{3-\alpha}{\alpha-1}} \epsilon^{-\frac{1}{\alpha-1}}\right)$ steps.

It is important to highlight the impact of the order of smoothness on the convergence rate. More specifically, if one has the smooth potential $f(x) = (1 + \|x\|^2)^{\alpha/2}$ which has the same tail growth as $\|x\|^\alpha$, then the convergence rate becomes $\widetilde{\mathcal{O}}\left(d^{\frac{4-\alpha}{\alpha}} \epsilon^{-1}\right)$.

A finite and bounded perturbation should not change the above convergence rate. Indeed, 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 33 implies that it is also β-Hölder continuous. By Lemma 6, the rate obtained from Corollary 4 is applicable to $g(x) = \|x\|^\alpha + 10\cos(\|x\|) + \xi$, and the convergence rate $\widetilde{\mathcal{O}}\left(d^{\frac{3-\alpha}{\alpha-1}} \epsilon^{-\frac{1}{\alpha-1}}\right)$ still holds. Figure 1a demonstrates the behavior of this potential and its gradient when $\alpha = 1.5$, in 1 dimension.

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. The Hessian of this potential can be computed as

$$\nabla^2 f(x) = \left(1 + \|x\|^2\right)^{-3/2} \left[\left(1 + \|x\|^2\right) I_d - x x^\top\right].$$

If we let $\tilde{f} = f$ in Assumption 1, one can easily verify that the assumption holds for $\theta = 3$. Further, computing its gradient reveals that Assumption 2 is satisfied with $\zeta = 0$ and $\alpha = 1$. Finally, since its Hessian is bounded, Assumption 3 is satisfied with $\beta = 1$. Plugging these parameters in Corollary 4, we obtain the convergence rate $\widetilde{\mathcal{O}}\left(d^7 \epsilon^{-1}\right)$ in KL-divergence.
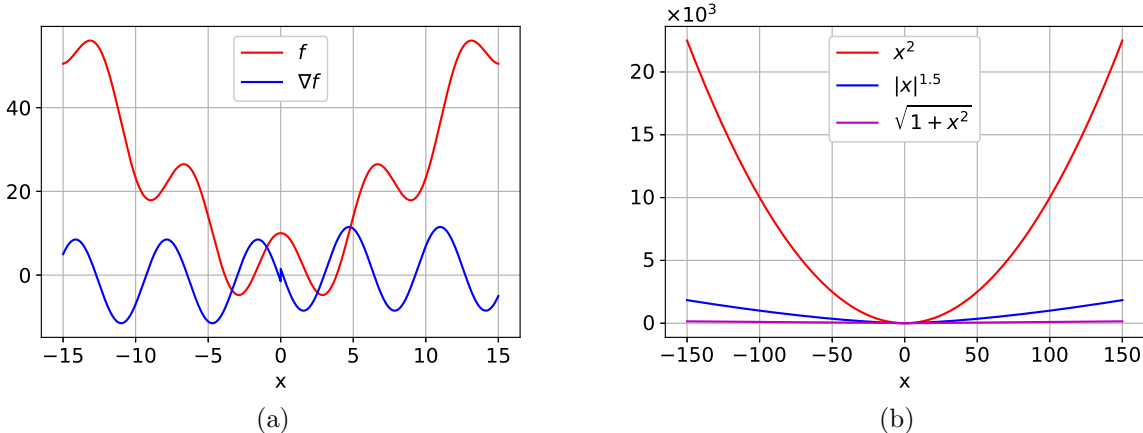
15

Figure 1: (a) Potential $f(x) = |x|^{1.5} + 10 \cos(x)$ and its gradient. The normalizing constant is ignored. (b) Comparison of tail growth needed for LSI and modified LSI. Note that both settings can work with perturbations which are ignored here for clarity.

The dimension dependency in the previous convergence rate can be improved by changing $\tilde{f}$ to a function that is different than f. Observe that the difference between $\sqrt{1 + \|x\|^2}$ and $(1 + \|x\|^{1+\tau})^{1/(1+\tau)}$ is bounded for any $\tau \in (0, 1)$. Thus, if we set $\tilde{f}(x) = (1 + \|x\|^{1+\tau})^{1/(1+\tau)}$, 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 a convergence rate of $\widetilde{\mathcal{O}}(d^5 \epsilon^{-1})$ in KL-divergence, 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.

Example 3 (Smooth potential with linear tails up to log factor): We consider the potential function $f(x) = \|x\| \log(1 + \|x\|^2)$ which is smooth and has a superlinear tail growth. By choosing $\tilde{f}(x) = \log(1 + \|x\|^2) \sqrt{1 + \|x\|^2}$, one can verify Assumption 1 for $\theta = 1$. Simply calculating the gradient of f, one can also verify α-dissipativity (3.3) for $\alpha = 1$. Moreover, ∇f has logarithmic growth, which means that for any $\zeta > 0$, gradient growth condition (3.4) holds; thus, Assumption 2 is satisfied. Since ∇f is Lipschitz, Assumption 3 is satisfied with $\beta = 1$. Finally, invoking Corollary 4, one obtains a convergence rate of $\widetilde{\mathcal{O}}(d^3 \epsilon^{-1})$ in KL-divergence.

Comparing the above rate to that in the previous example, we observe that the extra log factor improves the dimension dependence from $\widetilde{\mathcal{O}}(d^5)$ to $\widetilde{\mathcal{O}}(d^3)$, while the ϵ-dependence remains the same. Note that in here $\alpha = 1$, same as in the previous example; yet, the convergence in this case is faster than the advertised rate of $\widetilde{\mathcal{O}}(d^{\frac{1}{\beta} + (1 + \frac{1}{\beta})(\frac{2}{\alpha} - \mathbb{1}_{\{\alpha \neq 1\}})} \epsilon^{-\frac{1}{\beta}})$ because the potential is not growing like $\|x\|^\alpha$ for any α due to the additional log factor. We note that including an extra log factor will only improve the convergence when $\alpha = 1$; when $\alpha > 1$, the convergence rate simply stays the same. One intuitive reason for this phenomenon is that for $\alpha > 1$ the tails are already convex but when $\alpha = 1$, the tails are linear, in which case the Hessian is almost equal to zero, and the extra log injects the much needed convexity to the tails.

Figure 1b compares the growth rate of these examples with that of the quadratic function, in which case the standard LSI (1.3) holds. Perturbations are ignored to make the figure clear.

16

5.2 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 be still obtained by using the explicit formulas presented in Theorem 3 and Corollary 4.

Example 4 (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 [Fu98, FF93], one assumes the following linear model

$$Y = Vx + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, I_n),$$

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 following 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 setting $\alpha = 2$ and $\zeta = 1$ in Assumption 2. This potential lacks smoothness and 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 $\beta = 1$. In other words, Assumptions 3 does not hold for a single β for all x. Initially, it might seem that our results are non-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 17, our results can be applied to 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 rate $\widetilde{\mathcal{O}}\left(d^{\frac{1}{q-1}} \epsilon^{-\frac{1}{q-1}}\right)$ in KL-divergence.

This form of Assumption 3 enables us to deal with composite potentials that are sum of two components each of which is Hölder continuous but with a different exponent.

Example 5 (Bayesian logistic regression): In Bayesian logistic regression, we are given n samples $V = \{v_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$, $Y = \{y_i\}_{i=1}^n \in \{0, 1\}^n$ according to the logistic regression model $\mathbb{P}(y_i = 1|v_i) = 1/(1 + \exp(-\langle x, v_i \rangle))$ for some parameter $x \in \mathbb{R}^d$, and we would like to make inference about the parameter x under some prior distribution $p(x)$. In this framework, it is common to use LMC to generate samples from the posterior distribution $p(x|V, Y)$ with the following potential function

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

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 [Dal17b, LWME19]. 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 $p(x) \propto \exp\left(1 - \sqrt{1 + \|x\|^2}\right)$ [GDV+19, HZ03], which results in a similar

setting as in Example 2 in the sense that simply choosing $\tilde{f}$ as the potential yields an $\widetilde{\mathcal{O}}(d^7\epsilon^{-1})$ convergence rate in KL-divergence. As before, the dimension dependency can be improved by a more careful choice of $\tilde{f}$, i.e.,

$$\tilde{f}(x) = \left(1 + \|x\|^{1+\tau}\right)^{1/(1+\tau)} - \langle Y, Vx \rangle + \sum_{i=1}^{n} \log(1 + \exp(\langle x, v_i \rangle)),$$

which yields a convergence rate of $\widetilde{\mathcal{O}}(d^5\epsilon^{-1})$. In here the parameters are set as follows $\theta = 2 + \tau$, $\alpha = 1$, $\zeta = 0$ and $\beta = 1$.

Next, consider the prior $p(x) \propto \exp\left(\sum_{i=1}^{d} |x_i|^q\right)$ for $q \in (1, 2)$ which is similar to the Bridge linear regression setting. The resulting potential is not smooth in this case, and the potential lacks quadratic growth. Our analysis shows that LMC reaches ϵ-accuracy in KL-divergence after $\widetilde{\mathcal{O}}\left(d^{\frac{3-q}{q-1}}\epsilon^{-\frac{1}{q-1}}\right)$ steps. Here, the parameters are given as $\theta = 2 - q$, $\alpha = q$, $\zeta = q - 1$, and $\beta = q - 1$.

Example 6 (Huberized Regression with log-linear prior): Let V and Y denote the matrix of covariates and the response vector respectively, as in Example 4. We consider the Bayesian analog of Huberized loss [PC08] with a log-linear prior, which corresponds to sampling from the posterior $p(x) \propto e^{-f(x)}$ where

$$f(x) = \sum_{i=1}^{n} \sqrt{1 + (y_i - \langle x, v_i \rangle)^2} + \sqrt{1 + \|x\|^2}.$$

By considering

$$\tilde{f}(x) = \sum_{i=1}^{n} \sqrt{1 + (y_i - \langle x, v_i \rangle)^2} + \left(1 + \|x\|^{1+\tau}\right)^{\frac{1}{1+\tau}},$$

for $\tau \in (0, 1)$ and applying an argument similar to the one in Example 2, one can obtain the convergence rate $\widetilde{\mathcal{O}}(d^5\epsilon^{-1})$ for the LMC algorithm in KL-divergence. In here, the parameters are given as $\theta = 2 + \tau$, $\alpha = 1$, $\zeta = 0$ and $\beta = 1$.

6 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 10. *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*

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

where $\tilde{f}_\varepsilon$ 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 10. Let $\tilde{f}_\varepsilon(x) = \tilde{f}(x) + \varepsilon\|x\|^{3/2}e^{-\|x\|^2}$, and notice that f satisfies

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

For its Hessian, we write

$$\nabla^2 \tilde{f}_\varepsilon(x) = \nabla^2 \tilde{f}(x) + \varepsilon e^{-\|x\|^2}\left\{ \left(\frac{3}{2}\|x\|^{-1/2} - 2\|x\|^{3/2} \right)I_d - \left(6\|x\|^{-1/2} - 4\|x\|^{3/2} + \frac{3}{4}\|x\|^{-5/2} \right)xx^\top \right\}$$

and choosing $\alpha_\theta = 8\sup_{r\ge0} r^{1.5}(1 + r^2/4)^{\theta/2}e^{-r^2}$, 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 $x \le 0.1$. Now by selecting

$$m(r) = (\mu - \alpha_\theta\varepsilon)(1 + r^2/4)^{-\theta/2} \vee \frac{\varepsilon}{2}r^{-1/2}\mathbb{1}_{\{r\le0.1\}},$$

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

Proof of Theorem 1. We follow a similar construction developed in [TV00], and define the functions h and $\tilde{h}_\varepsilon$ as

$$h(x) = f(x) + m(2r)\left(\|x\| - r\right)^2\mathbb{1}_{\{\|x\|\ge r\}} + C_r \quad \text{and} \tag{6.1}$$

$$\tilde{h}_\varepsilon(x) = \tilde{f}_\varepsilon(x) + m(2r)\left(\|x\| - r\right)^2\mathbb{1}_{\{\|x\|\ge r\}} + C_r,$$

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

$$e^{C_r} = \int_{\|x\|<r} e^{-f(x)}dx + \int_{\|x\|\ge r} e^{-f(x)}e^{-m(2r)(\|x\|-r)^2}dx. \tag{6.2}$$

Using Assumption 1 and Lemma 10, we notice that h and $\tilde{h}_\varepsilon$ satisfy $\left|h(x) - \tilde{h}_\varepsilon(x)\right| \le \xi + \varepsilon/2$, and also the growth of the function $\tilde{h}_\varepsilon$ can be characterized in the following three regions.

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

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

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

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

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

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

19

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

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

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 [BE85] that the distribution $e^{-\tilde{h}_\varepsilon}$ satisfies (1.3). Combining this with the Holley-Stroock perturbation lemma [HS87], we obtain

$$\forall \rho, \quad \mathrm{H}(\rho|e^{-h}) \leq \frac{e^{2\xi+\varepsilon}}{2m(2r)} \mathrm{I}(\rho|e^{-h}). \tag{6.3}$$

We will convert the above inequality on the perturbed potential h to an inequality on the potential function f. Using the definition in (6.1), we can obtain an upper bound on the relative entropy for all $r > 0$,

$$\mathrm{H}(\rho|e^{-f}) = \mathrm{H}(\rho|e^{-h}) + \int \rho(x)(f(x) - h(x))dx \tag{6.4}$$
$$= \mathrm{H}(\rho|e^{-h}) - m(2r) \int_{\|x\| \geq r} \rho(x)(\|x\| - r)^2 dx - C_r.$$

For the normalizing constant C_r explicitly given in (6.2), one can obtain a lower bound using the Jensen's inequality,

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

Combining this with (6.4), we obtain

$$\mathrm{H}(\rho|e^{-f}) \leq \mathrm{H}(\rho|e^{-h}) + m(2r) \int_{\|x\| \geq r} \left(e^{-f(x)} - \rho(x) \right)(\|x\| - r)^2 dx$$
$$\leq \mathrm{H}(\rho|e^{-h}) + m(2r) \int_{\|x\| \geq r} e^{-f(x)} \|x\|^2 dx$$
$$\leq \mathrm{H}(\rho|e^{-h}) + m(2r) \frac{\mathrm{M}_s(\nu_*)}{(1 + r^2)^{s/2-1}} \tag{6.5}$$

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 11 below.

Lemma 11. *For a given distribution ρ and for a constant $r > 0$, we have*

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

Proof of Lemma 11. For positive constants $p, q, s > 0$ satisfying $1/p + 1/q = 1$, we apply the Hölder's inequality and get

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

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

Similarly for the Fisher information, we write

$$
\mathrm{I}\left(\rho|e^{-h}\right) \leq 2\mathrm{I}\left(\rho|e^{-f}\right) + 2\int \rho(x)\|\nabla h(x) - \nabla f(x)\|^2 dx. \tag{6.6}
$$

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

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

where in the last step we applied Lemma 11. Plugging this back in (6.6), we get

$$
\mathrm{I}\left(\rho|e^{-h}\right) \leq 2\mathrm{I}\left(\rho|e^{-f}\right) + \frac{8m(2r)^2}{(1+r^2)^{s/2-1}}\mathrm{M}_s(\rho). \tag{6.7}
$$

Combining the inequalities (6.3), (6.5), and (6.7), we obtain

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

Finally, using the Lemma 27 and optimizing over $m(2r)$, we get

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

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. $\square$

21

7 Moment Bounds on the LMC Iterates

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

$$M_s(\rho_k) \le M_s(\rho_0) + C_s k\eta.$$

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.

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

for any C_2 satisfying

$$C_2 \ge 3a + 2b + 2d. \tag{7.1}$$

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

$$M_2(\widetilde{\rho}_{k,t}|x_k) \le g_2(x_k) + C_2 t,$$

where $g_s(x) = \big(1 + x^2\big)^{s/2}$ and $M_s(\widetilde{\rho}_{k,t}|x_k)$ 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.)

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

where

$$N_2 \triangleq 2\sqrt{2\eta}(1.5a + b). \tag{7.2}$$

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 $s \ge 4$ 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. For all even s, we have

1. $M_s(\widetilde{\rho}_{k,t}|x_k) \le g_s(x_k) + C_s t$.

2. $\mathbb{E}[-\langle\nabla f(x_k), Z\rangle g_s(\widetilde{x}_{k,t})|x_k] \le N_s$.

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

$$
\frac{\partial}{\partial t}\mathrm{M}_s(\widetilde{\rho}_{k,t}|x_k)
$$

$$
= \mathbb{E}[-s\langle\nabla f(x_k),\widetilde{x}_{k,t}\rangle g_{s-2}(\widetilde{x}_{k,t}) + s(d+s-2)g_{s-2}(\widetilde{x}_{k,t}) - s(s-2)g_{s-4}(\widetilde{x}_{k,t})|x_k]
$$

$$
\leq s\mathbb{E}\Big[\Big(-\big\langle\nabla f(x_k), x_k - t\nabla f(x_k) + \sqrt{2t}Z\big\rangle + (d+s-2)\Big)g_{s-2}(\widetilde{x}_{k,t})|x_k\Big]
$$

$$
\leq s\big(-\langle\nabla f(x_k), x_k\rangle + t\|\nabla f(x_k)\|^2 + (d+s-2)\big)\mathbb{E}[g_{s-2}(\widetilde{x}_{k,t})|x_k]
$$

$$
\qquad + s\sqrt{2t}\mathbb{E}[\langle-\nabla f(x_k), Z\rangle g_{s-2}(\widetilde{x}_{k,t})|x_k]
$$

$$
\leq s\Big[-a\|x_k\|^\alpha + b + 2\eta M^2(1+\|x_k\|^{2\zeta}) + (d+s-2)\Big]_+ (g_{s-2}(x_k) + C_{s-2}t) + s\sqrt{2\eta}N_{s-2}
$$

$$
\leq s\Big[-a(1+\|x_k\|^2)^{\frac{\alpha}{2}} + 2\eta M^2(1+\|x_k\|^2)^\zeta + (2\eta M^2 + a + b + d + s - 2)\Big]_+ (g_{s-2}(x_k) + C_{s-2}t)
$$

$$
\qquad + s\sqrt{2\eta}N_{s-2}
$$

$$
\overset{1}{\leq} \max_{u\geq 1} s\Big(-au^\alpha + 2\eta M^2 u^{2\zeta} + (2\eta M^2 + a + b + d + s - 2)\Big)(u^{s-2} + C_{s-2}t) + s\sqrt{2\eta}N_{s-2}
$$

$$
\leq s\max_{u\geq 1}\Big(-\frac{a}{2}u^{\alpha+s-2} + 2\eta M^2 u^{2\zeta+s-2}\Big)
$$

$$
\qquad + s\max_{u\geq 1}\Big(-\frac{a}{2}u^{\alpha+s-2} + (2\eta M^2 + a + b + d + s - 2)u^{s-2}\Big)
$$

$$
\qquad + sC_{s-2}\eta\max_{u\geq 1}\Big(-au^\alpha + 2\eta M^2 u^{2\zeta} + (2\eta M^2 + a + b + d + s - 2)\Big) + s\sqrt{2\eta}N_{s-2}
$$

$$
\overset{2}{\leq} s\Big[(2\eta M^2 + a + b + d + s - 2)\Big(\frac{2(2\eta M^2 + a + b + d + s - 2)(s - 2)}{a(\alpha + s - 2)}\Big)^{\frac{s-2}{\alpha}}
$$

$$
\qquad + C_{s-2}\eta(2\eta M^2 + a + b + d + s - 2) + \sqrt{2\eta}N_{s-2}\Big],
$$

in which substitution $u = \sqrt{1+\|x_k\|^2}$ is used in step 1 and Lemma 29 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

$$
\frac{C_s}{s} \geq (2\eta M^2 + a + b + d + s - 2)\Big(\frac{2(2\eta M^2 + a + b + d + s - 2)}{a}\Big)^{\frac{s-2}{\alpha}} \tag{7.3}
$$
$$
+ C_{s-2}\eta(2\eta M^2 + a + b + d + s - 2) + \sqrt{2\eta}N_{s-2}.
$$

For proving the second part of the induction step, we use Stein's lemma [Ste81] (the version we use is stated in Lemma 30) in the first equality below.

$$\mathbb{E}[-\langle \nabla f(x_k), Z\rangle g_s(\widetilde{x}_{k,t})|x_k]$$

$$= \mathbb{E}\Big[-s\sqrt{2t}\Big\langle \nabla f(x_k), g_{s-2}(\widetilde{x}_{k,t})\Big(x_k - t\nabla f(x_k) + \sqrt{2t}Z\Big)\Big\rangle|x_k\Big]$$

$$\leq s\sqrt{2t}\big(-\langle \nabla f(x_k), x_k\rangle + t\|\nabla f(x_k)\|^2\big)\mathrm{M}_{s-2}(\widetilde{\rho}_{k,t}|x_k) + 2stN_{s-2}$$

$$\leq s\sqrt{2\eta}\Big[-a\|x_k\|^\alpha + b + 2\eta M^2(1 + \|x_k\|^{2\zeta})\Big]_+ (g_{s-2}(x_k) + C_{s-2}\eta) + 2s\eta N_{s-2}$$

$$\leq s\sqrt{2\eta}\Big[-a(1 + \|x_k\|^2)^{\frac{\alpha}{2}} + 2\eta M^2(1 + \|x_k\|^2)^\zeta + (b + a + 2\eta M^2)\Big]_+ (g_{s-2}(x_k) + C_{s-2}\eta)$$

$$\quad + 2s\eta N_{s-2}$$

$$\leq s\sqrt{2\eta}\max_{u\geq 1}\Big(-au^\alpha + 2\eta M^2 u^{2\zeta} + (b + a + 2\eta M^2)\Big)(u^{s-2} + C_{s-2}\eta) + 2s\eta N_{s-2}$$

$$\leq s\sqrt{2\eta}\max_{u\geq 1}\Big(-\frac{a}{2}u^{\alpha+s-2} + 2\eta M^2 u^{2\zeta+s-2}\Big)$$

$$\quad + s\sqrt{2\eta}\max_{u\geq 1}\Big(-\frac{a}{2}u^{\alpha+s-2} + (b + a + 2\eta M^2)u^{s-2}\Big)$$

$$\quad + s\sqrt{2\eta}C_{s-2}\eta\max_u\Big(-au^\alpha + 2\eta M^2 u^{2\zeta} + (b + a + 2\eta M^2)\Big) + 2s\eta N_{s-2}$$

$$\leq s\Big[(b + a + 2\eta M^2)\sqrt{2\eta}\Big(\frac{2(b + a + 2\eta M^2)(s-2)}{a(\alpha + s - 2)}\Big)^{\frac{s-2}{\alpha}}$$

$$\quad + C_{s-2}\eta\sqrt{2\eta}(b + a + 2\eta M^2) + 2\eta N_{s-2}\Big] \leq N_s,$$

where

$$\frac{N_s}{s} = (b + a + 2\eta M^2)\sqrt{2\eta}\Big(\frac{2(b + a + 2\eta M^2)}{a}\Big)^{\frac{s-2}{\alpha}} + C_{s-2}\eta\sqrt{2\eta}(b + a + 2\eta M^2) + 2\eta N_{s-2}. \quad (7.4)$$

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

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

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

setting $t = \eta$ yields

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

and finally, induction on k gives

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

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 (7.1) and (7.2) to get $N_2 \leq C_2\sqrt{2\eta}$, then we use (7.3) and (7.4) inductively, to establish $N_s \leq C_s\sqrt{2\eta}$. By combining the previous inequality with $4\eta M^2 \leq a$, we can strengthen the bound (7.3) to

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

C_s, as defined in (3.8), satisfies the previous inequality and (7.1), which in turn implies that it also satisfies (7.3) and (7.1). □

The next proposition is the analog moment bound for the continuous-time process, and is adapted from [TV00] for the sake of comparison with the bound for the discrete time process.

Lemma 12. *Let f satisfy Assumption 2 and p_t be the distribution of Z_t, then*

$$\mathrm{M}_s(p_t + \nu_*) \le \mathrm{M}_s(p_0 + \nu_*) + K_s t,$$

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

Proof. Because of linearity of integral, it is sufficient to prove

$$\mathrm{M}_s(p_t) \le \mathrm{M}_s(p_0) + K_s t.$$

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

$$\frac{d}{dt}\mathrm{M}_s(p_t) = \int p_t(x)\Big[\Delta(1 + \|x\|^2)^{\frac{s}{2}} - \Big\langle \nabla f(x), \nabla(1 + \|x\|^2)^{\frac{s}{2}}\Big\rangle\Big]$$

$$= (ds + s(s-2))\mathrm{M}_{s-2}(p_t) - s(s-2)\mathrm{M}_{s-4}(p_t) - s\int p_t(x)\langle\nabla f(x), x\rangle(1 + \|x\|^2)^{\frac{s-2}{2}}$$

$$\le (b + d + s - 2)s\mathrm{M}_{s-2}(p_t) - s\int p_t(x)a\|x\|^\alpha(1 + \|x\|^2)^{\frac{s-2}{2}}$$

$$\le (b + d + a + s - 2)s\mathrm{M}_{s-2}(p_t) - \frac{as}{2}\mathrm{M}_{s+\alpha-2}(p_t)$$

$$\le (b + d + a + s - 2)s\mathrm{M}_{s+\alpha-2}(p_t)^{\frac{s-2}{s+\alpha-2}} - \frac{as}{2}\mathrm{M}_{s+\alpha-2}(p_t)$$

$$\overset{1}{\le} (b + d + a + s - 2)s\left(\frac{2(b + d + a + s - 2)(s - 2)}{a(s + \alpha - 2)}\right)^{\frac{s-2}{\alpha}}$$

$$\le (b + d + a + s - 2)\left(\frac{b + d + a + s - 2}{a/2}\right)^{\frac{s-2}{\alpha}} s,$$

where step 1 follows from Lemma 29. □

We utilize a method similar to the previous proof in order to bound the moments of target
Proof of Lemma 22. From the proof of Lemma 12, we have the following inequality for $s \ge 2$.

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

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 13 shows that $\mathrm{M}_s(\nu_*)$ is finite. By rearranging the previous inequality, we get

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

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

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

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

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

$\square$

8 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.

8.1 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 13. Using this growth, we bound the α-th exponential moment of the target ν_* in Lemma 14. Finally, using the exponential moment bound, in Lemma 15, we upper bound the α-th moment of a given distribution with its KL-divergence from the 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 13. *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.

$$f(x) = f(0) + \int_0^{\frac{R}{\|x\|}} \langle \nabla f(tx), x \rangle dt + \int_{\frac{R}{\|x\|}}^1 \langle \nabla f(tx), x \rangle dt$$

$$\geq f(0) - \int_0^{\frac{R}{\|x\|}} \|\nabla f(tx)\| \|x\| dt + \int_{\frac{R}{\|x\|}}^1 \frac{1}{t} \langle \nabla f(tx), tx \rangle dt$$

$$\geq f(0) - \left(\frac{M(2a+2b)}{a} \right) R + \int_{\frac{R}{\|x\|}}^1 \frac{1}{t} (a\|tx\|^\alpha - b) dt$$

$$\geq f(0) - M \left(\frac{2a+2b}{a} \right)^2 + \frac{a}{2} \|x\|^\alpha \int_{\frac{R}{\|x\|}}^1 t^{\alpha-1} dt$$

$$\geq f(0) - M \left(\frac{2a+2b}{a} \right)^2 + \frac{a}{2\alpha} \|x\|^\alpha \left(1 - \frac{R^\alpha}{\|x\|^\alpha} \right)$$

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

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 13 to prove that the α-th exponential moment of the target ν_* is bounded.

Lemma 14. *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,

$$\begin{cases} \tilde{\mu} &= \log\left(\frac{16\pi}{a}\right) + M\left(\frac{2a+2b}{a}\right)^2 + b + |f(0)|, \\ \tilde{d} &= d(1 + (1 - \alpha/2)\log(d)). \end{cases} \tag{8.1}$$

Proof. Using Lemma 13 we get

$$\int e^{\frac{a}{4\alpha}\|x\|^\alpha - f(x)} dx \leq e^{-f(0) + M\left(\frac{2a+2b}{a}\right)^2 + b} \int e^{-\frac{a}{4\alpha}\|x\|^\alpha} dx$$

$$= \frac{2\pi^{d/2}}{\alpha} \left(\frac{4\alpha}{a} \right)^{d/\alpha} e^{-f(0) + M\left(\frac{2a+2b}{a}\right)^2 + b} \frac{\Gamma(d/\alpha)}{\Gamma(d/2)}.$$

Next, using an inequality for the ratio of Gamma functions [JDK71], we obtain

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

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

$$\log \left(\int e^{\frac{a}{4\alpha}\|x\|^\alpha - f(x)} dx \right) \leq \frac{d}{2} \log(\pi) + \frac{d}{\alpha} \log\left(\frac{4\alpha}{a}\right) + \left(\frac{d}{\alpha} - \frac{d}{2} \right) \log\left(\frac{d}{2e}\right)$$

$$+ \left(\frac{d}{\alpha} + \frac{1}{2} \right) \log\left(\frac{2}{\alpha}\right) + M\left(\frac{2a+2b}{a}\right)^2 + b + |f(0)|$$

$$\leq \frac{d}{\alpha} \left(\log\left(\frac{16\pi}{a}\right) + \left(1 - \frac{\alpha}{2}\right) \log\left(\frac{d}{2e}\right) \right) + M\left(\frac{2a+2b}{a}\right)^2 + b + |f(0)|$$

$$\leq \tilde{d}\tilde{\mu}.$$

27

$\square$

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

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

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 14 implies $\log z \leq \tilde{d}\tilde{\mu}$. Using this bound on z we get

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

Rearranging this yields the desired inequality. $\square$

8.2 Single step bound

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

$$d\widetilde{x}_{k,t} = -\nabla f(x_k)dt + \sqrt{2}dB_t \quad \text{with} \quad \widetilde{x}_{k,0} = x_k, \tag{8.2}$$

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 Theorem 1 and the linear moment bounds 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 [VW19].

Lemma 16 ([VW19]). *Suppose $\widetilde{x}_{k,t}$ is the interpolation of the discretized process (8.2). Let $\widetilde{\rho}_{k,t}$ denote its distribution. Then*

$$\frac{d}{dt}\mathrm{H}(\widetilde{\rho}_{k,t}|\nu_*) = -\mathrm{I}(\widetilde{\rho}_{k,t}|\nu_*) + \mathbb{E}\left[\left\langle \nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k), \nabla \log\left(\frac{\widetilde{\rho}_{k,t}(\widetilde{x}_{k,t})}{\nu_*(\widetilde{x}_{k,t})}\right)\right\rangle\right]$$

$$\leq -\frac{3}{4}\mathrm{I}(\widetilde{\rho}_{k,t}|\nu_*) + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k)\|^2\right]. \tag{8.3}$$

Proof. The following proof is included for reader's convenience. For further notational convenience, we denote with $\widetilde{\rho}_{kt}(x_k, \widetilde{x}_{k,t})$, the joint distribution of random variables x_k and $\widetilde{x}_{k,t}$, and similarly, we denote with $\widetilde{\rho}_{k|t}(x_k)$ and $\widetilde{\rho}_{t|k}(\widetilde{x}_{k,t})$, the conditional distributions of x_k conditioned on $\widetilde{x}_{k,t}$, and $\widetilde{x}_{k,t}$ conditioned on x_k, respectively.

The distribution of $\widetilde{x}_{k,t}$ conditioned on x_k can be described by the following Fokker-Planck equation.

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

28

Taking expectation with respect to x_k yields

$$\frac{\partial \widetilde{\rho}_{k,t}(x)}{\partial t} = \int \frac{\partial \widetilde{\rho}_{t|k}(x)}{\partial t} \rho(x_k) dx_k$$

$$= \int (\nabla \cdot (\widetilde{\rho}_{kt}(x_k,x)\nabla f(x_k)) + \Delta \widetilde{\rho}_{kt}(x_k,x)) dx_k$$

$$= \nabla \cdot \left(\rho(x_k) \int \widetilde{\rho}_{k|t}(x_k)\nabla f(x_k) dx_k \right) + \Delta \widetilde{\rho}_{k,t}(x)$$

$$= \nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\mathbb{E}_{\widetilde{\rho}_{k|t}}[\nabla f(x_k)|\widetilde{x}_{k,t} = x] \right) + \Delta \widetilde{\rho}_{k,t}(x).$$

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

$$\frac{d}{dt}\mathrm{H}\big(\widetilde{\rho}_{k,t}|\nu_*\big) = \frac{d}{dt}\int \widetilde{\rho}_{k,t}(x)\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)dx$$

$$= \int \frac{\partial \widetilde{\rho}_{k,t}}{\partial t}(x)\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)dx$$

$$= \int \left(\nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\mathbb{E}_{\widetilde{\rho}_{k|t}}[\nabla f(x_k)|\widetilde{x}_{k,t} = x]\right) + \Delta \widetilde{\rho}_{k,t}(x)\right)\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)dx$$

$$\stackrel{1}{=} \int \left(\nabla \cdot \left(\widetilde{\rho}_{k,t}(x)\mathbb{E}_{\widetilde{\rho}_{k|t}}[\nabla f(x_k)|\widetilde{x}_{k,t} = x] + \nabla\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right) - \nabla f(x)\right)\right)\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)dx$$

$$\stackrel{2}{=} -\int \widetilde{\rho}_{k,t}(x)\left\langle \mathbb{E}_{\widetilde{\rho}_{k|t}}[\nabla f(x_k)|\widetilde{x}_{k,t} = x] + \nabla\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right) - \nabla f(x), \nabla\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)\right\rangle dx$$

$$= -\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \int \widetilde{\rho}_{k,t}(x)\left\langle \nabla f(x) - \mathbb{E}_{\widetilde{\rho}_{k|t}}[\nabla f(x_k)|\widetilde{x}_{k,t} = x], \nabla\log\left(\frac{\widetilde{\rho}_{k,t}(x)}{\nu_*(x)}\right)\right\rangle dx$$

$$= -\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}_{\widetilde{\rho}_{kt}}\left[\left\langle \nabla f(\widetilde{x}_{k,t}) - \nabla f(x_k), \nabla\log\left(\frac{\widetilde{\rho}_{k,t}(\widetilde{x}_{k,t})}{\nu_*(\widetilde{x}_{k,t})}\right)\right\rangle\right]$$

$$\stackrel{3}{\leq} -\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}_{\widetilde{\rho}_{kt}}\left[\|\nabla f(\widetilde{x}_{k,t}) - f(x_k)\|^2\right] + \frac{1}{4}\mathbb{E}_{\widetilde{\rho}_{kt}}\left[\|\nabla\log\left(\frac{\widetilde{\rho}_{k,t}(\widetilde{x}_{k,t})}{\nu_*(\widetilde{x}_{k,t})}\right)\|^2\right]$$

$$= -\frac{3}{4}\mathrm{I}\big(\widetilde{\rho}_{k,t}|\nu_*\big) + \mathbb{E}\left[\|\nabla f(\widetilde{x}_{k,t}) - f(x_k)\|^2\right],$$

in which equality 1 is follows from $\Delta\widetilde{\rho}_{k,t} = \nabla \cdot (\nabla\widetilde{\rho}_{k,t})$, equality 2 follows from the divergence theorem, inequality 3 follows from $\langle u,v \rangle \leq \|u\|^2 + \frac{1}{4}\|v\|^2$, and in the last step, the subscript of the expectation is removed and indicates that the expectation is taken with respect to both x_k and $\widetilde{x}_{k,t}$. $\qquad\square$

Next, using Lemma 16, we bound the time derivative of the KL-divergence $\frac{d}{dt}\mathrm{H}\big(\widetilde{\rho}_{k,t}|\nu_*\big)$, and obtain a useful differential inequality.

Lemma 17. *If the potential f satisfies Assumptions 1, 2 and 3, then*

$$\frac{d}{dt}\mathrm{H}\big(\widetilde{\rho}_{k,t}|\nu_*\big) \leq -\frac{3}{4}\lambda^{-\frac{1}{1-\delta}}(\mathrm{M}_s(\rho_0 + \nu_*) + C_s(k+1)\eta)^{-\frac{\delta}{1-\delta}}\mathrm{H}\big(\widetilde{\rho}_{k,t}|\nu_*\big)^{\frac{1}{1-\delta}}$$

$$+ \frac{16\alpha L^2 M^{2\beta}}{a}\mathrm{H}\big(\rho_k|\nu_*\big)\eta^{2\beta} + 4L^2\left(1 + M^{2\beta}\left(1 + \frac{2\alpha\widetilde{\mu}}{a}\right)\right)\widetilde{d}\eta^\beta, \tag{8.4}$$

when $t \leq \eta \leq \frac{1}{2}\big(1 \wedge \frac{a}{2M^2}\big)$. The constants $\widetilde{d}$ and $\widetilde{\mu}$ are defined in (8.1).

29

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

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

where step 1 follows from Lemma 28, step 2 from Assumption 2, step 3 from the fact that $2\zeta\beta \leq \alpha$, and step 4 from Lemma 15 and $\eta < 1$. Plugging the above inequality back in (8.3) and using Theorem 1 and Proposition 2 results in (8.4). $\qquad\square$

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 18. *Suppose f satisfies Assumptions 1, 2 and 3, then*

$$
\begin{aligned}
\mathrm{H}(\rho_{k+1}|\nu_*) \leq &\mathrm{H}(\rho_k|\nu_*)\left(1 - \frac{3\eta}{8\lambda^{\frac{1}{1-\delta}}}\left(\frac{\mathrm{H}(\rho_k|\nu_*)}{\mathrm{M}_s(\rho_0 + \nu_*) + C_s(k+1)}\right)^{\frac{\delta}{1-\delta}} + \frac{16\alpha L^2 M^{2\beta}\eta^{2\beta+1}}{a}\right) \\
&+ \sigma\tilde{d}\eta^{\beta+1},
\end{aligned} \tag{8.5}
$$

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

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

Proof. Let

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

where κ_1 and κ_2 are constants independent of t. We can rewrite (8.4) as

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

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

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

We will use the following basic comparison lemma for differential inequalities; see, for example [McN86] for a simple proof.

Lemma 19. *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*

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

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

For positive t, we have

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

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

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

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

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

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

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

$$+ \frac{16\alpha L^2 M^{2\beta}}{a} \mathrm{H}\big(\rho_k | \nu_*\big) \eta^{2\beta+1} + \sigma \tilde{d} \eta^{\beta+1}.$$

We rewrite the previous inequality to get

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

$$+ \sigma \tilde{d} \eta^{\beta+1}.$$

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

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

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

$$\mathrm{H}\big(\rho_{k+1} | \nu_*\big) \leq \mathrm{H}\big(\rho_k | \nu_*\big) \left(1 - \frac{3}{8\lambda^{\frac{1}{1-\delta}}} \left(\frac{\mathrm{H}\big(\rho_k | \nu_*\big)}{\mathrm{M}_s(\rho_0 + \nu_*) + C_s(k+1)\eta} \right)^{\frac{\delta}{1-\delta}} \eta \right)$$

$$+ \frac{16\alpha L^2 M^{2\beta}}{a} \mathrm{H}\big(\rho_k | \nu_*\big) \eta^{2\beta+1} + \sigma \tilde{d} \eta^{\beta+1}.$$

Rearranging the above inequality yields the desired result. $\qquad\qquad\square$

8.3 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 18 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 convergence rate leading to the main corollary.

Proof of Theorem 3. We simplify the recurrence relation for the single step bound in (8.5). For notational convenience, let $A = \frac{\lambda^{-1/(1-\delta)}}{16}\left(\frac{\sigma\tilde{d}}{M_s(\rho_0+\nu_*)\vee C_s}\right)^{\delta/(1-\delta)}$. 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 18, if $k < N$ and $\mathrm{H}(\rho_k|\nu_*) \geq \epsilon/2$, then

$$\mathrm{H}(\rho_{k+1}|\nu_*) \leq \left(1 - \frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}\right)\mathrm{H}(\rho_k|\nu_*) + \sigma\tilde{d}\eta^{\beta+1}. \tag{8.6}$$

The above expression depends on the choice of step size η and number of steps N; thus, given (8.5), we verify the inequality (8.6) for

$$\eta^{-1} = (\sigma\tilde{d})^{\frac{1}{\beta}}(16\lambda)^{\frac{1}{\beta(1-2\delta)}}\left(\frac{M_s(\rho_0+\nu_*)\vee C_s}{16}\right)^{\frac{\delta}{\beta(1-2\delta)}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\frac{\delta}{\beta(1-2\delta)}}\left(\frac{2}{\epsilon}\right)^{\frac{1-\delta}{\beta(1-2\delta)}},$$

$$N = (\sigma\tilde{d})^{\frac{1}{\beta}}(16\lambda)^{\frac{1+\beta}{\beta(1-2\delta)}}\left(\frac{M_s(\rho_0+\nu_*)\vee C_s}{16}\right)^{\frac{(1+\beta)\delta}{\beta(1-2\delta)}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{1+\frac{(\beta+1)\delta}{\beta(1-2\delta)}}\left(\frac{2}{\epsilon}\right)^{\frac{1-\delta(1-\beta)}{\beta(1-2\delta)}}. \tag{8.7}$$

For the above choices of η and N, using (8.5) together with the fact that $k < N$ and $\mathrm{H}(\rho_k|\nu_*) \geq \epsilon/2$, in order for (8.6) to hold, it suffices to prove the following inequality

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

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

$$\begin{cases} \frac{3\lambda^{-1/(1-\delta)}}{8}\left(\frac{\epsilon/2}{M_s(\rho_0+\nu_*)+C_s(N+1)\eta}\right)^{\delta/(1-\delta)} \geq \frac{2A\eta^{\delta\beta/(1-\delta)}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}}, \\[2mm] \frac{A\eta^{\delta\beta/(1-\delta)+1}}{\log\left(\frac{2\Delta_0}{\epsilon}\right)^{\delta/(1-\delta)}} \geq \frac{16\alpha L^2 M^{2\beta}}{a}\eta^{2\beta+1}. \end{cases} \tag{8.8}$$

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

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

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 (8.8) holds.

For the first inequality in (8.8), we consider two cases. In the first case $N\eta \geq 1$, since we have $M_s(\rho_0 + \nu_*) + C_s(N + 1)\eta \leq 3(M_s(\rho_0 + \nu_*) \vee C_s)N\eta$, the following condition implies the desired inequality

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

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 $M_s(\rho_0 + \nu_*) + C_s(N + 1)\eta \leq 3(M_s(\rho_0 + \nu_*) \vee C_s)$; hence, the following condition suffices

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

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 (8.6) 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

$$\mathrm{H}\big(\rho_k|\nu_*\big) \leq \epsilon \implies \mathrm{H}\big(\rho_{k+1}|\nu_*\big) \leq \epsilon, \text{ for } k < N. \tag{8.9}$$

We split this into two cases. First, consider the case $\epsilon/2 \leq \mathrm{H}\big(\rho_k|\nu_*\big) \leq \epsilon$. In this case, using (8.6) and $\mathrm{H}\big(\rho_k|\nu_*\big) \leq \epsilon$, 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 $\tilde{d}$. The second case is when $\mathrm{H}\big(\rho_k|\nu_*\big) \leq \epsilon/2$. Using Lemma 18, we need to show

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

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 $M_s \geq 1$, we obtain the following two conditions on the accuracy ϵ to be combined together later,

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

$$\epsilon \leq 32(\tilde{\lambda} \wedge \tilde{\lambda}^2)(1 \wedge \sigma d) \leq 2^{5+\frac{3\beta(1-2\delta)}{1-\delta+\delta\beta}}\lambda^{\frac{1+\beta}{1-\delta+\delta\beta}}(\sigma\tilde{d})^{\frac{1-2\delta}{1-\delta+\delta\beta}}.$$

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 $\mathrm{H}(\rho_k|\nu_*) \leq \epsilon$, then (8.9) shows $\mathrm{H}(\rho_N|\nu_*) \leq \epsilon$ and there is nothing to prove. For the second case, we have $\mathrm{H}(\rho_k|\nu_*) > \epsilon$ for all $k < N$; therefore, (8.6) combined with Lemma 31 and the fact that $\mathrm{H}(\rho_0|\nu_*) \leq \Delta_0$ imply

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

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 18, to a condition on the accuracy ϵ. That is, we have

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

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

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

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 18 shows that the first step is decreasing and $\mathrm{H}(\rho_1|\nu_*) \leq \Delta_0$. Continuing inductively from there, we get either $\mathrm{H}(\rho_k|\nu_*)$ is decreasing or it is less than ϵ, in both of the cases, we have $\mathrm{H}(\rho_k|\nu_*) \leq \Delta_0$. 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

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

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

$$\psi = \min\left\{2, \frac{2\Delta_0}{e}, 2\Delta_0 e^{\frac{-1}{16(\tilde{\lambda}\wedge\tilde{\lambda}^2)}}, 32(1 \wedge \sigma d)(\tilde{\lambda} \wedge \tilde{\lambda}^3)(1 \wedge \Delta_0^{-1})^{\frac{\beta}{4}}, 16(\tilde{\lambda} \wedge \tilde{\lambda}^2)(1 \wedge \sigma d)\left(1 \wedge \frac{a}{2M^2}\right), \right.$$

$$\left. 2\left(\tilde{\lambda}^{0.5} \wedge \tilde{\lambda}^2\right)\left(1 \wedge \frac{2a\sigma d^2}{M^{2\beta}}\right)^{0.5}, 8(\tilde{\lambda}^2 \wedge \tilde{\lambda})\left(1 \wedge \frac{a}{\alpha L^2 M^{2\beta}}\right)^{\frac{1}{3}}(1 \wedge \sigma d), 32(\tilde{\lambda} \wedge \tilde{\lambda}^2)(1 \wedge \sigma d)\right\},$$

$$(8.10)$$

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 (8.7) to get

$$\eta = \sigma^{-\frac{1}{\beta}}(16\lambda)^{-\frac{s-2+2\theta}{\beta(s-2)}}\left(\frac{\mathrm{M}_s(\rho_0 + \nu_*)}{16d^{(s-2+\alpha)/\alpha}} \vee \left(\frac{3a+2b+3}{1 \wedge a}\right)^{\frac{s-2+\alpha}{\alpha}}\frac{s^s}{16}\right)^{-\frac{\theta}{\beta(s-2)}}$$

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

$$N = \sigma^{\frac{1}{\beta}}(16\lambda)^{\frac{(1+\beta)(s-2+2\theta)}{\beta(s-2)}}\left(\frac{\mathrm{M}_s(\rho_0 + \nu_*)}{16d^{(s-2+\alpha)/\alpha}} \vee \left(\frac{3a+2b+3}{1 \wedge a}\right)^{\frac{s-2+\alpha}{\alpha}}\frac{s^s}{16}\right)^{\frac{(1+\beta)\theta}{\beta(s-2)}}$$

$$d^{\frac{1}{\beta}+\frac{(s-2+\alpha)(1+\beta)\theta}{\alpha\beta(s-2)}}(1 + (1-\alpha/2)\log(d))^{\frac{1}{\beta}}\log\left(\frac{2\Delta_0}{\epsilon}\right)^{1+\frac{(\beta+1)\theta}{\beta(s-2)}}\left(\frac{2}{\epsilon}\right)^{\frac{1}{\beta}+\frac{(1+\beta)\theta}{\beta(s-2)}}.$$

$$\square$$

9 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.

Lemma 20. *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.

Since the range of B contains all non-positive numbers, we can define $y(M) = \min\{r|B(r) = -M\}$, for $M \geq 0$. 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

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

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

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

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

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

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

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 21. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a convex potential and $u \in \mathbb{R}^d$ is unit vector. Then, f is coercive satisfying*

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

Proof. 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 $x = x_1 u_1 + x_{-1}$ 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

$$
1 = \int_{\mathbb{R}^d} e^{-f(x)} dx \geq \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} e^{-f(0)-\langle \nabla f(x), x \rangle} dx_1 dx_{-1}
$$

$$
= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} e^{-f(0)-x_1\langle \nabla f(x), u_1 \rangle - \langle \nabla f(x), x_{-1} \rangle} dx_1 dx_{-1}
$$

$$
\geq \int_{\mathbb{R}^{d-1}} \int_{x_1 \geq 0} e^{-f(0)-x_1\langle \nabla f(x), u_1 \rangle - \langle \nabla f(x), x_{-1} \rangle} dx_1 dx_{-1}
$$

$$
\geq \int_{\mathbb{R}^{d-1}} \int_{x_1 \geq 0} e^{-f(0)-\langle \nabla f(x_1 u_1 + x_{-1}), x_{-1} \rangle} dx_1 dx_{-1}.
$$

If we have $\sup_{x_1 \geq 0} \langle \nabla f(x_1 u_1 + x_{-1}), x_{-1} \rangle < \infty$, then the inner integral diverges, so for almost every $x_{-1} \in \mathrm{span}\{u_2, ..., u_d\}$,

$$
\sup_{x_1 \geq 0} \langle \nabla f(x_1 u_1 + x_{-1}), x_{-1} \rangle = \infty.
$$

Using finiteness of the integral once again, we write

$$
1 = \int_{\mathbb{R}^d} e^{-f(x)} dx = \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} e^{-f(x_1 u_1 + x_{-1})} dx_1 dx_{-1}.
$$

The inner integral should converge for almost every $x_{-1} \in \mathrm{span}\{u_2, ..., u_d\}$. Since a convex function restricted to a line is still convex, Lemma 20 implies $g(x_1) = e^{-f(x_1 u_1 + x_{-1})}$ is lower bounded for almost every x_{-1}. Fix some $x_{-1} \in \mathrm{span}\{u_2, ..., u_d\}$ such that $g(x_1) = e^{-f(x_1 u_1 + x_{-1})}$ is lower bounded and $\sup_{x_1 \geq 0} \langle \nabla f(x_1 u_1 + x_{-1}), x_{-1} \rangle = \infty$, which happens for almost every x_{-1}. By convexity, we have

$$
f(x_1 u + 2x_{-1}) \geq f(x_1 u + x_{-1}) + \langle \nabla f(x_1 u_1 + x_{-1}), x_{-1} \rangle.
$$

Taking supremum over x_1, we have $\sup_{x_1 \geq 0} f(x_1 u_1 + 2x_{-1}) = \infty$. But since $\sup_{x_1 \geq 0} f(2x_1 u_1) < M$, we can write

$$
\frac{1}{2} f(4x_{-1}) + \frac{1}{2} f(2x_1 u) \geq f(x_1 u + 2x_{-1}).
$$

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

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

Proof of Proposition 5. 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 21. 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 $f(t_0 u) \geq 1 + f(0)$ for all unit vectors $u \in \mathbb{R}^d$. For any $t > t_0$ and any unit vector u, because of convexity, we write

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

Therefore, for $t > t_0$, we have

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

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.

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

by Lemma 20. 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 $t \in [0, \infty)$,

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

This completes the proof. $\qquad\square$

10   Proofs of Corollaries

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

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

We state a lemma to bound the moments of the target distribution. The proof is in to section 7.

Lemma 22. *Let f satisfy Assumption 2 then we have the following bound on the moment*

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

Combining the Gaussian moment bound with the previous lemma yields

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

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. $\qquad\square$

Proof of Corollary 7.   First we state Pinsker's inequality, which bounds total variation with KL-divergence.

Lemma 23 (Pinsker's inequality). *For distributions p and q*

$$\mathrm{TV}(p, q) \leq \sqrt{\frac{1}{2}\mathrm{H}(p|q)}.$$

38

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

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

Note that the upper bound on ϵ is changed and $2\epsilon^2$ needs to be smaller than upper bound in (8.10). In other words $2\epsilon^2 \leq \psi$, where ψ is defined in (8.10). $\qquad\square$

Proof of Corollary 8. First we state a result, which is adapted from Corollary 3 in [BV05], that bounds L_α-Wasserstein distance with KL-divergence.

Lemma 24 ([BV05]). *For probability measure p on $\mathbb{R}^d$, if $\int e^{\theta\|x\|^\alpha} p(x)dx < \infty$, then*

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

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 14 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

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

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. In order to obtain the upper bound on the accuracy, first let ψ denote the bound in (8.10). 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

$$\epsilon \leq 4(4\alpha a^{-1}(1.5 + \tilde{\mu}\tilde{d}))^{-\frac{1}{\alpha}}\psi^{\frac{1}{2\alpha}} \wedge 4\left(4\alpha a^{-1}(1.5 + \tilde{d}\tilde{\mu})\right)^{\frac{1}{\alpha}}, \qquad (10.1)$$

where ψ is defined in (8.10) and $\tilde{d}$ and $\tilde{\mu}$ are defined in (8.1). $\qquad\square$

Proof of Corollary 9. When $\theta = 0$, Theorem 1 implies LSI with constant $\frac{4e^{2\xi}}{\mu}$. LSI implies Talagrand's inequality with the same constant[OV00].

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

Theorem 3 with accuracy $\frac{\epsilon^2\mu}{16e^{2\xi}}$ implies the convergence rate. 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

$$\epsilon \leq \frac{4e^\xi}{\sqrt{\mu}}\psi, \qquad (10.2)$$

where ψ is defined in (8.10). $\qquad\square$

11 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 a wide range of non-convex 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 presented results show that the convergence rate of LMC can be described as a function of the tail growth rate and the order of smoothness in high dimensions.

There are several important future directions that one can consider, among which we highlight a few here. Verifying the tightness of convergence rates established for the LMC algorithm (1.2) is important; thus, one needs to derive lower bounds in this framework, similar to those in [GLL19, CBL20, CLW20]. Moreover, in practice, higher order variants of LMC is commonly used for sampling. Among these, algorithms that are based on the underdamped Langevin diffusion received a lot of interest. Therefore, generalizing the results of this paper to higher order and/or general Itô diffusions is important.

Our results explain the behavior of LMC in the case where $\beta \in (0,1]$, and they do not cover the case $\beta = 0$. Indeed, there is no known result on vanilla LMC for this case; thus, exploring the behavior of LMC in this regime may be of interest. We note that many of the bounds in the paper can be improved, at the expense of introducing some additional complexity into the results. Moreover, our results hold only for the last iterate of the LMC algorithm; therefore, investigating the behavior of the subsequent iterates is also an interesting direction left for another study.

Acknowledgements

References

[Atc15]   Yves F Atchadé, *A moreau-yosida approximation scheme for a class of high-dimensional posterior distributions*, arXiv preprint arXiv:1505.07072 (2015).

[BB18]   Adrien Blanchet and Jérôme Bolte, *A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions*, Journal of Functional Analysis **275** (2018), no. 7, 1650–1673.

[BDM18]   Nicolas Brosse, Alain Durmus, and Eric Moulines, *The promises and pitfalls of stochastic gradient langevin dynamics*, Advances in Neural Information Processing Systems, 2018, pp. 8268–8278.

[BDMP17]   Nicolas Brosse, Alain Durmus, Éric Moulines, and Marcelo Pereyra, *Sampling from a log-concave distribution with compact support with proximal langevin monte carlo*, arXiv preprint arXiv:1705.08964 (2017).

[BDMS19]   Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis, *The tamed unadjusted langevin algorithm*, Stochastic Processes and their Applications **129** (2019), no. 10, 3638–3663.

[BE85]   Dominique Bakry and Michel Émery, *Diffusions hypercontractives*, Séminaire de probabilités de Strasbourg **19** (1985), 177–206 (fr). MR 889476

[BG99]   Sergej G Bobkov and Friedrich Götze, *Exponential integrability and transportation cost related to logarithmic sobolev inequalities*, Journal of Functional Analysis **163** (1999), no. 1, 1–28.

[BGL13]   Dominique Bakry, Ivan Gentil, and Michel Ledoux, *Analysis and geometry of markov diffusion operators*, vol. 348, Springer Science & Business Media, 2013.

[BMD+19]   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).

[BV05]   François Bolley and Cédric Villani, *Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities*, Annales de la Faculté des sciences de Toulouse: Mathématiques, vol. 14, 2005, pp. 331–352.

[CB18]   Xiang Cheng and Peter L Bartlett, *Convergence of langevin mcmc in kl-divergence*, PMLR 83 (2018), no. 83, 186–211.

[CBL20]   Niladri S Chatterji, Peter L Bartlett, and Philip M Long, *Oracle lower bounds for stochastic gradient sampling algorithms*, arXiv preprint arXiv:2002.00291 (2020).

[CCAY+18]   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 (2018).

[CCBJ18]   Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, Proceedings of the 31st Conference on Learning Theory (COLT2018), Proceedings of Machine Learning Research, vol. 75, PMLR, 2018, pp. 300–323.

[CDJB19]   Niladri S Chatterji, Jelena Diakonikolas, Michael I Jordan, and Peter L Bartlett, *Langevin monte carlo without smoothness*, arXiv preprint arXiv:1905.13285 (2019).

[CFM+18]   Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan, *On the theory of variance reduction for stochastic gradient monte carlo*, arXiv preprint arXiv:1802.05431 (2018).

[CLW20]   Yu Cao, Jianfeng Lu, and Lihan Wang, *Complexity of randomized algorithms for underdamped langevin dynamics*, arXiv preprint arXiv:2003.09906 (2020).

[CS91]   EA Carlen and Avraham Soffer, *Entropy production by block variable summation and central limit theorems*, Communications in mathematical physics **140** (1991), no. 2, 339–371.

[Dal17a]   Arnak Dalalyan, *Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent*, Proceedings of the 2017 Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 65, PMLR, 07–10 Jul 2017, pp. 678–689.

[Dal17b]   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** (2017), no. 3, 651–676.

[DCWY19]   Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu, *Log-concave sampling: Metropolis-hastings algorithms are fast*, Journal of Machine Learning Research **20** (2019), no. 183, 1–42.

[DDN20]   Anh Duc Doan, Xin Dang, and Dao Nguyen, *Weakly smooth langevin monte carlo using p-generalized gaussian smoothing*, arXiv preprint arXiv:2002.10071 (2020).

[DK19]   Arnak S Dalalyan and Avetik Karagulyan, *User-friendly guarantees for the langevin monte carlo with inaccurate gradient*, Stochastic Processes and their Applications **129** (2019), no. 12, 5278–5311.

[DM16]   Alain Durmus and Eric Moulines, *Sampling from strongly log-concave distributions with the unadjusted langevin algorithm*, arXiv preprint arXiv:1605.01559 **5** (2016).

[DM17]   Alain Durmus and Éric Moulines, *Nonasymptotic convergence analysis for the unadjusted langevin algorithm*, The Annals of Applied Probability **27** (2017), no. 3, 1551–1587.

[DM19]   ———, *High-dimensional bayesian inference via the unadjusted langevin algorithm*, Bernoulli **25** (2019), no. 4A, 2854–2882.

[DMM19]   Alain Durmus, Szymon Majewski, and Blazej Miasojedow, *Analysis of langevin monte carlo via convex optimization.*, Journal of Machine Learning Research **20** (2019), no. 73, 1–46.

[DMP18]   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** (2018), no. 1, 473–506.

[DRV+17]   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** (2017), no. 4, 2195–2237.

[DT12]   Arnak S Dalalyan and Alexandre B Tsybakov, *Sparse regression learning by aggregation and Langevin Monte-Carlo*, Journal of Computer and System Sciences **78** (2012), no. 5, 1423–1443.

[Ebe16]   Andreas Eberle, *Reflection couplings and contraction rates for diffusions*, Probability theory and related fields **166** (2016), no. 3-4, 851–886.

[EGZ19]   Andreas Eberle, Arnaud Guillin, and Raphael Zimmer, *Couplings and quantitative contraction rates for langevin dynamics*, The Annals of Probability **47** (2019), no. 4, 1982–2010.

[EMS18]   Murat A Erdogdu, Lester Mackey, and Ohad Shamir, *Global non-convex optimization with discretized diffusions*, Advances in Neural Information Processing Systems, 2018, pp. 9671–9680.

[FF93]   LLdiko E Frank and Jerome H Friedman, *A statistical view of some chemometrics regression tools*, Technometrics **35** (1993), no. 2, 109–135.

[Fu98]   Wenjiang J Fu, *Penalized regressions: the bridge versus the lasso*, Journal of computational and graphical statistics **7** (1998), no. 3, 397–416.

[GDV+19]   Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, Lester Mackey, et al., *Measuring sample quality with diffusions*, The Annals of Applied Probability **29** (2019), no. 5, 2884–2928.

[GLL19]   Rong Ge, Holden Lee, and Jianfeng Lu, *Estimating normalizing constants for log-concave distributions: Algorithms and lower bounds*, arXiv preprint arXiv:1911.03043 (2019).

[GLR18]   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).

[GM91]   Saul B Gelfand and Sanjoy K Mitter, *Recursive stochastic algorithms for global optimization in r^d*, SIAM Journal on Control and Optimization **29** (1991), no. 5, 999–1018.

[Gro75]   Leonard Gross, *Logarithmic sobolev inequalities*, American Journal of Mathematics **97** (1975), no. 4, 1061–1083.

[HKRC18]   Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher, *Mirrored langevin dynamics*, Advances in Neural Information Processing Systems, 2018, pp. 2878–2887.

[HS87]   Richard Holley and Daniel Stroock, *Logarithmic sobolev inequalities and stochastic ising models*, Journal of Statistical Physics **46** (1987), no. 5, 1159–1194.

[HZ03]   Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[JDK71]   Petar M. Vasić Jovan D. Kečkić, *Some inequalities for the gamma function*, Publications de l'Institut Mathématique **11(25)** (1971), no. 31, 107–114 (eng).

[JKO98]   Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the fokker–planck equation*, SIAM journal on mathematical analysis **29** (1998), no. 1, 1–17.

[KNS16]   Hamed Karimi, Julie Nutini, and Mark Schmidt, *Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 795–811.

[LFC17]   Duy Tung Luu, Jalal M Fadili, and Christophe Chesneau, *Sampling from non-smooth distribution through langevin diffusion*.

[Loj63]   Stanislaw Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les équations aux dérivées partielles **117** (1963), 87–89.

[LWME19]   Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu, *Stochastic runge-kutta accelerates langevin monte carlo and beyond*, Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 7748–7760.

[MCC⁺19]   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 (2019).

[MCF15]   Yi-An Ma, Tianqi Chen, and Emily Fox, *A complete recipe for stochastic gradient mcmc*, Advances in Neural Information Processing Systems, 2015, pp. 2917–2925.

[MCJ⁺19]   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** (2019), no. 42, 20881–20885.

[McN86]   Alex McNabb, *Comparison theorems for differential equations*, Journal of mathematical analysis and applications **119** (1986), no. 1-2, 417–428.

[MFWB19a]   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 (2019).

[MFWB19b]   _____, *Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity*, arXiv preprint arXiv:1907.11331 (2019).

[Mil94]   Grigorii Noikhovich Milstein, *Numerical integration of stochastic differential equations*, vol. 313, Springer Science & Business Media, 1994.

[MMW⁺19]   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 (2019).

[MSH02]   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** (2002), no. 2, 185–232.

[MT12]   Sean P Meyn and Richard L Tweedie, *Markov chains and stochastic stability*, Springer Science & Business Media, 2012.

[MT13]   Grigori Noah Milstein and Michael V Tretyakov, *Stochastic numerics for mathematical physics*, Springer Science & Business Media, 2013.

[MV99]   P. A. Markowich and C. Villani, *On the trend to equilibrium for the fokker-planck equation: An interplay between physics and functional analysis*, Physics and Functional Analysis, Matematica Contemporanea (SBM) 19, 1999, pp. 1–29.

[Nes15]   Yu Nesterov, *Universal gradient methods for convex optimization problems*, Mathematical Programming **152** (2015), no. 1-2, 381–404.

[OV00]   Felix Otto and Cédric Villani, *Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality*, Journal of Functional Analysis **173** (2000), no. 2, 361–400.

[PC08]   Trevor Park and George Casella, *The bayesian lasso*, Journal of the American Statistical Association **103** (2008), no. 482, 681–686.

[Pol63]   Boris Teodorovich Polyak, *Gradient methods for minimizing functionals*, Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki **3** (1963), no. 4, 643–653.

[RRT17]   Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky, *Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis*, Proceedings of the 2017 Conference on Learning Theory, vol. 65, 2017, pp. 1674–1703.

[SL19]   Ruoqi Shen and Yin Tat Lee, *The randomized midpoint method for log-concave sampling*, Advances in Neural Information Processing Systems, 2019, pp. 2098–2109.

[Ste81]   Charles M Stein, *Estimation of the mean of a multivariate normal distribution*, The annals of Statistics (1981), 1135–1151.

[Tal96]   Michel Talagrand, *Transportation cost for gaussian and other product measures*, Geometric & Functional Analysis GAFA **6** (1996), no. 3, 587–600.

[Tos99]   Giuseppe Toscani, *Entropy production and the rate of convergence to equilibrium for the fokker-planck equation*, Quarterly of Applied Mathematics **57** (1999), no. 3, 521–541.

[TV00]   Guiseppe Toscani and Cédric Villani, *On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds*, Journal of Statistical Physics **98** (2000), no. 5-6, 1279–1309.

[VW19]   Santosh Vempala and Andre Wibisono, *Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices*, Advances in Neural Information Processing Systems, 2019, pp. 8092–8104.

[Wib18]   Andre Wibisono, *Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem*, Conference On Learning Theory, 2018, pp. 2093–3027.

[WT11]   Max Welling and Yee W Teh, *Bayesian learning via stochastic gradient langevin dynamics*, Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 681–688.

[XCZG18] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu, *Global convergence of langevin dynamics based algorithms for nonconvex optimization*, Advances in Neural Information Processing Systems, 2018, pp. 3122–3133.

[ZLC17]   Yuchen Zhang, Percy Liang, and Moses Charikar, *A hitting time analysis of stochastic gradient langevin dynamics*, arXiv preprint arXiv:1702.05575 (2017).

A Useful Lemmas

Lemma 25. *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 (3.1) in Assumption 1 for $\theta < 1$. Then α-dissipativity in Assumption 2 is satisfied for $\alpha = 2 - \theta$ with the following constants

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

Remark. 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

$$\langle \nabla \tilde{f}(x), x \rangle = \langle \int_0^1 \nabla^2 \tilde{f}(tx) x dt + \nabla \tilde{f}(0), x \rangle$$

$$= \langle \nabla \tilde{f}(0), x \rangle + \int_0^1 x^\top \nabla^2 \tilde{f}(tx) x dt$$

$$\geq -\|\nabla \tilde{f}(0)\|\|x\| + \int_0^1 \mu(1 + \|tx\|)^{\alpha - 2}\|x\|^2 dt$$

$$\geq -\|\nabla \tilde{f}(0)\|\|x\| + \frac{\mu\|x\|}{\alpha - 1}\left((1 + \|x\|)^{\alpha - 1} - 1\right)$$

$$= -\left(\|\nabla \tilde{f}(0)\| + \mu\right)\|x\| + \frac{\mu}{\alpha - 1}\|x\|^\alpha.$$

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

$$\langle \nabla f(x), x \rangle \geq -\left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)\|x\| + \frac{\mu}{\alpha - 1}\|x\|^\alpha$$

$$\geq \frac{\mu}{2(\alpha - 1)}\|x\|^\alpha - \left(-\frac{\mu}{2(\alpha - 1)}\|x\|^\alpha + \left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)\|x\|\right)$$

$$\overset{1}{\geq} \frac{\mu}{2(\alpha - 1)}\|x\|^\alpha - \left(\frac{2\left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)^\alpha}{\mu} \times \frac{\alpha - 1}{\alpha}\right)^{1/(\alpha - 1)}$$

$$\geq \frac{\mu}{2(\alpha - 1)}\|x\|^\alpha - \left(\frac{2\left(\|\nabla \tilde{f}(0)\| + \mu + \xi\right)^\alpha}{\mu}\right)^{1/(\alpha - 1)},$$

where step 1 follows from Lemma 29. $\qquad\square$

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 $\nu_* \propto e^{-f}$ is bounded as follows*

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

Remark. 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.

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

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

Using the previous formula, we bound the KL-divergence

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

Using the previous bound and the formula for the Gaussian entropy concludes the proof. $\qquad\square$

Lemma 27. *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*

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

Proof. Taking derivative and setting it equal to zero yields the value for x_*. $\qquad\square$

Lemma 28. *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 $h_2(x) = 2(x^\gamma + 1) - (1$ are non-negative when $\gamma \in [0, 1]$ and $\gamma \in [0, 2]$, respectively. $\qquad\square$

Lemma 29. *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. $\qquad\square$

Lemma 30 (Stein's lemma [Ste81])**.** *Suppose $x \sim \mathcal{N}(\mu, \sigma^2 I_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is weakly differentiable. For $a \in \mathbb{R}^d$*

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

Lemma 31. *If $x_k \leq (1-a)x_{k-1} + b$ for $0 < a < 1$ and $0 \leq b$, then*

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

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 $1 - a \leq e^{-a}$, (A.1) is achieved. $\qquad\square$

A.1 Some Properties of Hölder Continuity

Lemma 32. *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. $\qquad\square$

Lemma 33. *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 $\|x - y\| > 1$,

$$\|f(x) - f(y)\| \leq \|f(x) - g(x)\| + \|g(x) - g(y)\| + \|f(y) - g(y)\| \leq B + h_g^\alpha \|x - y\|^\beta + B \leq (2B + h_g^\alpha)\|x - y\|^\beta.$$

Taking the maximum of constants in the two cases completes the proof. $\qquad\square$

Lemma 34. *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 $\|x - y\| \leq \|x\| + \|y\| \leq 2\|x\|$, which in turn implies $\|x\|^{\alpha-2} \leq 2^{2-\alpha}\|x - y\|^{\alpha-2}$. Therefore,

$$\|f(x) - f(y)\| \leq \|\|x\|^{\alpha-2}x - \|y\|^{\alpha-2}y\| \leq \|\|x\|^{\alpha-2}x - \|x\|^{\alpha-1}\frac{y}{\|y\|} + \|x\|^{\alpha-1}\frac{y}{\|y\|} - \|y\|^{\alpha-2}y\|$$

$$\leq \|x\|^{\alpha-1}\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\| + |\|x\|^{\alpha-1} - \|y\|^{\alpha-1}|$$

$$\overset{1}{\leq} \|x\|^{\alpha-1}\|\frac{x}{\|x\|} - \frac{y}{\|x\|} + \frac{y}{\|x\|} - \frac{y}{\|y\|}\| + \|x - y\|^{\alpha-1}$$

$$\leq \|x\|^{\alpha-2}\|x - y\| + \|x\|^{\alpha-1}\|\frac{y}{\|y\|}(\frac{\|y\|}{\|x\|} - 1)\| + \|x - y\|^{\alpha-1}$$

$$\leq 2\|x\|^{\alpha-2}\|x - y\| + \|x - y\|^{\alpha-1} \leq (1 + 2^{3-\alpha})\|x - y\|^\alpha \leq 5\|x - y\|^\alpha,$$

where inequality 1 follows from Lemma 28. $\qquad\square$